MaizeGDB Team Response to the Working Group Report from November of 2006

Carolyn Lawrence, Darwin Campbell, Mary Schaeffer, and Trent Seigfried

---

First, we'd like to say again how much we appreciated it that you all traveled to Ames for the Working Group Meeting. The report was very helpful, as were the many suggestions you provided in addition to the written report. Our responses to your guidance are shown throughout this document follow a "Response:" indicator and appear in blue italics.

---

(1) MaizeGDB needs greater resources.
This year 10.9 billion bushels of corn will be produced with a farm gate value of $37.5 billion. Additionally, the value for ethanol production could double this value over the next decade. Currently, there are 3.5 employees devoted to the efforts of MaizeGDB, which is 0.0015% of the value of the corn crop (0.03% of the R&D effort). This is meager support for a database that is the central clearing house for genomic and genetic data for such an important crop. These resources will not allow the maize community to synthesize the results of genomics adequately and will impede the ultimate transfer of knowledge from the basic plant biologists in the lab out to the applied community and into the marketplace.

The working group noted that TAIR has 6 times as many employees as MaizeGDB, while the *Drosophila* database (flybase.org) has 9 times as many. Database improvement and maintenance are labor intensive and ongoing tasks requiring continued infusions of resources. Given the importance of corn production to the US economy as a whole, MaizeGDB needs a level of investment at least commensurate with Drosophila and Arabidopsis given the crucial mission to providing linkage of data for agronomic improvement.

Response: Thank you for including your thoughts on our current funding level in this document. Being able to cite the 2006 Working Group Report as a basis for seeking out additional funding for the project has already been quite useful.

Additional resources can come from at least three sources. First, federal interagency cooperation could provide increased resources. Second, the MaizeGDB group needs to take an active role in funding their research through competitive grants, as they are now beginning to do. Third, MaizeGDB needs to insure that any future federally-funded projects that propose to make use of MaizeGDB for lasting curation of valuable data must include budgetary support for any data submissions, curation activities that MaizeGDB is asked to provide.

Response: We did submit a proposal for POPcorn to the NSF's Plant Genome Research Program (with a much reduced budget), and you can view that full proposal here: http://www/lawrencelab.org/POPcornProposal.pdf. At the time of the Working Group meeting, it was just too close to the deadline to come up with a better idea for a larger project. We currently are considering ideas for the next year's proposals to PGRP, and await a RFP to be issued by the DOE's Energy Biosciences Program to decide whether we can create a good submission for that. CL also just filed a Letter of Intent describing our potential collaboration with a group Pat Schnable is coordinating for Iowa State University's part of the proposal for a DOE GTL Bioenergy Research Center. We would

be participating in data handling for the Cell Wall Architecture portion of that project in collaboration with Robin Buell and Volker Brendel.

With respect to handling data for outside groups, we will not commit to storing outcomes from major endeavors that are funded without having worked out a way to support those collaborations through the project's funding source. In instances where a good deal of time will be spent curating a group's data, we will require that personnel be committed to MaizeGDB using the project's budget to support the personnel. In addition, Hank Bass, Cliff Weil, and Karen Cone are researchers funded by NSF's PGRP who each have a person working on their end to prepare datasets for MaizeGDB. This is great. However, working with those groups to train them on how to prepare datasets often takes more time than we are reasonably able allocate to the task. We plan to contact those groups in the coming months to determine whether a Supplemental Request to their NSF grants could be submitted to cover some percentage (~25% each?) of a MaizeGDB curator's time for the remainder of their funding periods. We also will do the same for any others who are already funded and come to us with similar needs.

In summary, we will work through the proper channels to get additional funding to support the MaizeGDB project.

At least doubling or tripling of funding through various sources would greatly enhance the effort by MaizeGDB at Ames, Iowa. An additional two programmers and three curators at Ames would greatly accelerate the research. There are other bioinformatics groups around the country that can and are contributing to maize bioinformatics, but it is critical that MaizeGDB remain the vigorous leading group that synthesize bioinformatics and makes it accessible to the entire research and agronomic community.

(2) MaizeGDB needs to prioritize goals and future plans

The MaizeGDB project has given careful thought to its future plans and asked for working group input on prioritization. The working group presents high and medium priority areas for consideration by MaizeGDB:

High Priority Areas

- Community service should continue
    - o Currently, MaizeGDB plays a central role in conducting central maize genetics community functions (ie. with annual meetings, votes, surveys, and the maize newsletter). This role should continue as it is critical to the success and cohesion of the research direction for the community. However, the working group recommends a careful self-evaluation of whether all actions are most efficient. For example, to optimize Trent Seigfried's time and make use of his skills, the group suggests that the Maize Genetics Meeting functions could be automated or otherwise not wholly dependent upon one of the two developers. It is important that MaizeGDB team continue to play this role, however, whenever possible they should get the curation resources from the community rather than providing them.

      Response: TS has improved upon the tools he had for supporting such functions so that they require less human intervention. In addition, over the course of the next year we will be pursuing a serious campaign to get researchers accustomed to using the Community Curation Tools. It is possible that with minor changes to

these tools' functionality, researchers could own and edit their own abstract submissions. This campaign to broaden the use of the Community Curation Tools will be a high priority item to be handled by Lisa Harper, our newest addition to the MaizeGDB team. LH will begin work toward the end of February on a half time basis. She will be working out of the Plant Gene Expression Center in Albany, CA (Directed by Sarah Hake), and her major areas of emphasis in the near future will be: leading the Community Curation campaign, overseeing the Editorial Board, and integrating the RescueMu and EMS phenotypes from the Maize Gene Discovery and Maize Inflorescence Architecture projects with other phenotype data at MaizeGDB.

• Gene function prediction is important – Over the next few years, we will know the sequence of over 50,000 maize genes, but connecting genes to phenotypes is key to making genomics useful. MaizeGDB needs to play a leading role in curating, displaying, and analyzing the mutagenesis efforts in maize that will provide tools for functional analysis.

o Curation of published analyses – Currently, MaizeGDB is the leading resource for published mutants of maize, which has nearly a 100 year history of research. These efforts take considerable curation effort and the working group encourage the MaizeGDB team to promote alternative approaches (eg., via outreach) and to develop automated approaches to mine literature (examples coming from the honey bee community).

Response: LH will be responsible for curating data from the Editorial Board's past and current selections, and we are hopeful that the campaign to increase use of the Community Curation tools will help to keep MaizeGDB's content as broad as possible. In addition, Ed Coe, the retired ARS scientist who conceived of and managed MaizeDB (MaizeGDB's predecessor), is an avid community curator. He plans to volunteer time to add information about new mutants described in the literature, especially if they have associated sequence and/or genetic map information.

o Curation of stocks – Maize has a tremendous number and quality of genetic stocks. Continued curation of data regarding these stocks should be a top priority.

Response: Curation of stock data is largely handled by our collaborator Marty Sachs, Director of the Maize Genetics Cooperation – Stock Center. Our relationship with his group has been useful from both sides' perspectives in the past, and we will work to keep this interaction thriving.

o Several high throughput mutagenesis programs (both transposon and chemically-induced) are underway. It is definitely in the community interest that these studies be well curated at MaizeGDB. Many of these projects were funded through competitive funding, but funds were not requested for MaizeGDB curation. It may be necessary for these projects, MaizeGDB, and funding agencies to add resources for appropriate curation of these projects. Additionally, MaizeGDB needs to define the file formats for inputs, essentially reusable standards for data deposition. These standards should be readily accessible from the website.

Response: DC and CL recently created a standard file format and made it

available to R.A. Monde, project manager for C. Weil's TILLING project. Once we get some feedback on whether that sort of large-scale upload formatting works, we will create other such templates. MS is working with Karen Cone's Maize Chromatin project to standardize their inputs for RNAi silencing phenotypes, constructs, and events and to involve them in the initial curation of their data.

Instead of creating templates for various data types all at once, we will utilize those researchers who request the upload of large data sets as they come forward. In that way, groups who have a vested interest in getting their data into MaizeGDB will be used as the subject matter experts for developing the data upload templates.

o Integrated views and analysis of mutagenesis and phenotypic effects to gene are needed. These views need to be both from a genome view and from a pathway view. Competitive grants could provide resources to support these views, since the research questions are fundamental to basic functional genomics.

Response: On 8 January 2007, a position for a computational biologist for MaizeGDB opened up (see http://tinyurl.com/yxrtrw). Once a person has been hired to fill the position, we will begin work toward creating these integrated views. In addition to working together with the new hire to develop a proposal for funds to support such development, we will continue to work with D. Ware's group to see to it that developed solutions will support the outcomes of the maize genome sequencing project.

o MaizeGDB needs to develop timelines for integration of the key datasets, and then publicize the timeline so that the research community knows when outcomes will be accessible.

Response: We are in the process of developing such timelines and notices. We will publish these notices at http://www.maizegdb.org/data_contribution.php. Exceptions to the published schedule will be made only when special arrangements have been made well in advance of a deadline, and only when the contributor is providing a key dataset that would help move the science forward.

• Maize Maps (Structural/Genetic/Cytogenetic) should continue to expand– MaizeGDB plays a key role in hosting and curating many of the maize genetic and cytogenetic maps that have been created over the last decade. However, as the sequenced genome for maize becomes available, the central maps will be changing to a B73 sequence. The main focus of MaizeGDB should be on linking relevant datasets to this sequence.

o MaizeGDB should take a leading role in integrating the IBM genetic maps with the B73 genome.

Response: We agree that this is a very important task that must be accomplished. We are planning to work more closely with D. Ware's group in the coming year to begin work toward linking MaizeGDB's genetic maps to the emerging sequence and to devise a process for migrating outcomes from the sequencing project to MaizeGDB once that project has been completed.

o When map data from the maize diversity study becomes available MaizeGDB should work with these researchers to develop a next generation genetic map.

Response: We plan to do this in collaboration with Mike McMullen who will be computing the map.

o Since there is substantial variation in maize genome structure, additional genomes will be sequenced and where possible coherent views need to be built in MaizeGDB.

Response: The new Computational Biologist to be hired (mentioned elsewhere in this document) will be responsible for this project.

o MaizeGDB should focus on 2 major views - genetic and physical for now. Cytogenetic view although interesting and important has less general interest. Currently, there are often too many options, and users do not know what are the preferred maps.

Response: Agreed. Limiting the scope to these views will help to focus our efforts. We also will put together a maps information page that will explain which maps are useful for particular tasks, and will further highlight key maps throughout the site. We will review the displays for loci and will list preferred maps on each record.

Medium Priority Areas – *Response: These items will be pursued if and when resources allow*

• Gene structure prediction – As the genome is sequenced, many research groups around the nation are trying to predict gene structure and apply automated annotation of the genome. We do not believe that MaizeGDB should focus on doing this initial annotation but rather focus on:

o Integrating, recording, and presenting the leading gene models (currently three). They should ensure their software can relate these gene models to the sequence centric genome views.

o Organize experts in the communities to comment on the models (if funded as part of grant on this topic)

• Natural Diversity – Maize is the most diverse crop in the world, and USDA-ARS is responsible for the maintenance of this natural diversity. MaizeGDB needs to play a role presenting maize diversity and eventually helping plant breeders make use of this diversity. There are three other existing efforts ongoing in this area (Panzea, GRIN, Gramene), and MaizeGDB needs to work with these groups to develop an efficient display of information. The working group suggests the following approaches to deal with intense the curation effort required.

o MaizeGDB should only curate QTL studies for germplasm that is held by the Maize Stock Center. Currently, the only mapping population with repeatable data held by the stock center is the IBM population.

o The IBM QTL studies should be curated with their phenotypic score data into the GDPDM schema that is currently used by the maize diversity group, wheat, rice, and sorghum groups. With these phenotypic scores, it will be possible to reanalyze the data as computational and genetic methodologies improve in the future.

o Several groups will begin mapping QTL down to the gene level over the next two years. These gene level QTL –trait associations should be integrated with the MaizeGDB gene function prediction work.

(3) A Management Plan is needed.

MaizeGDB is a young ARS group responsible for synthesizing, displaying and coordinating the outputs of the maize genetics/genomics community. Management of competing needs and setting clear objectives is the key to developing an ever improving MaizeGDB. It is essential that the MaizeGDB team set priorities to deal with the plethora of extant and emergent data. In particular, there needs to be coordination between the Iowa and Missouri teams in order to distribute the time efforts more efficiently to meet common goals. The working group suggests that the unified MaizeGDB team articulate together their goals in the short, mid and long term. Most importantly, the team needs to articulate how and when these priority goals will be met. The working group recognizes that the priority goals suggested in this report require far more resources than are currently available. We suggest that two management plans could be devised: one that can be carried out within the budget limits of the current USDA-ARS commitment and a second that would be possible if appropriate resources were available. In this way, it will be transparent what is and is not possible within the resource constraints.

Response: A management plan matching the first description will be drafted. In addition to posting that plan at MaizeGDB, a copy will be included in the Project Plan for the project's next five years (similar to a grant proposal; described in detail below).

Short-term actions need to serve the long-term goals of MaizeGDB. The working group suggests that activities within the team need to be driven by a broader vision of the purpose and goals of MaizeGDB. For example, it seems that current action decisions are made based on urgency and the most persistent requests from the community. The team needs to take a leadership role in defining how, when and which datasets should be curated. If there are commonly agreed short and long-term actions, then all action decisions can be made based on helping to move MaizeGDB towards the long-term goals.

Response: Agreed. These tenets and directives will be incorporated into the management plan.

In summary, the working group makes the following suggestions for developing a management plan:

• Datasets to be curated and displayed need to be prioritized. Participant actions and timelines must be established in a systematic fashion.

• Numerous groups would like to have their data displayed at MaizeGDB, but MaizeGDB should feel free to say no if these data do not match the long-term goals.

• As in any scientific endeavor, it is important to devote substantial time to development of the newest views or approaches rather than just putting out fires.

• The POPcorn portal was very interesting, and we encourage MaizeGDB to develop an independent and diverse portfolio of competitive grant ideas and funded projects.

• The management plan should be in written form and consistent with realistic expectations.

Other items of interest:

Concept Paper Update

As you may remember from the Working Group meeting, projects in National Program (NP) 301 (which includes MaizeGDB) had the task of creating a Concept Paper for the National Program Staff this past fall outlining what their stakeholders had communicated to be the most important projects to pursue, within the scope of current funding. Kay Simmons, MaizeGDB's NP Leader, allowed us to update our Concept Paper so that it is in agreement with the Working Group Report. Based upon the Concept Paper, Dr. Simmons has issued a Program Direction and Resource Allocation Memo. This is a charge from the NP Staff outlining what must be included in the Project Plan (a document similar to a grant proposal). During the coming months, CL, in consult with other team members, will draft the Project Plan for the next five years. That planning document will include a management plan, and the Working Group Report will be incorporated as an appendix. The timeliness of the Working Group meeting and report was excellent since MaizeGDB's next five years' work will be largely based upon it.

New Positions/Hires

As mentioned elsewhere in this document, Lisa Harper has been hired to fill the half time curator position for MaizeGDB at the Plant Gene Expression Center in Albany, California. She begins in late February and will be focused on Community Curation, management of the MaizeGDB Editorial Board, and integration of RescueMu and EMS phenotype data with the rest of MaizeGDB's phenotype dataset.

A position is currently open for MaizeGDB. We are looking to hire a Computational Biologist to serve as the technical resource and advisor to MaizeGDB with respect to:

- *Analyzing and interpreting maize sequence data*
- *Using data structures to efficiently represent and mine genetic and physical maps of maize*
- *Developing representations of maize sequence data to be integrated into MaizeGDB (specifically including the creation of a maize genome browser and the infrastructure to enable storage of and access to maize gene models)*

To read more about this position, visit http://tinyurl.com/yxrtrw.

Addressing Other Suggestions Made at the Working Group Meeting

In addition to receiving the Working Group Report from you, we also took a good number of notes and wrote down some of your questions during the meeting and have

addressed some of the concerns already. Those notes and some information on how we're addressing the suggestions are outlined below.

Is 'monthly' a reasonable update schedule?

Response: At current staffing, updating more often would not be possible.

Mirror (Canada like Grain Genes or CropsNet?)

Response: We are currently looking for a collaborator that could support a mirror (MS has done the footwork on this). Those contacted thus far (GrainGenes and the Cereals Databases in the UK) do not support Oracle (the MaizeGDB database is Oracle and the interface is coded to utilize the Oracle-based database). However, if a license for Oracle could be purchased for the Cereals Databases, they might be willing to serve a mirror. When a reasonable collaborator who is willing and able to support a mirror is identified and a clear path to setting up the mirror has been established, we will let you know.

Curation dump to MO should happen DAILY.

Response: This has been pursued by MS and she has arranged with her unit's IT Specialist to set up an automatic process to pick up and store MaizeGDB's daily dumps out of Ames.

The database needs to become more "sequence-centric".

Response: We definitely will be working hard toward this goal. We are hopeful that having the new Computational Biologist on this problem will not only lend a new perspective to this within the group, but also will give us the resources necessary to commit a person to handling sequences as a full-time job.

Send out an email to researchers: "Do you want to be a cooperator?" Also explain what being a cooperator does (e.g., you get mailings and emailings re: maize meeting, etc.). If no response, then send same to mail address, if no response, then toss cooperator status. Also note that if we require on abstract submissions that user type (grad stu, postdoc, PI, technician, industry, other) be recorded, surveys could be resolved by user type. "Unknown" is what any existing ones that are not updated by abstract submission become.

Response: This will be done following the Maize Meeting in March. Assuming this works well, we'll repeat the process yearly.

At Pheasant Run there should be a curation and annotation workshop held.

Response: We are working with Conference Steering Committee to set up a joint MaizeGDB/Gramene curation and annotation workshop to be held one evening during

the informal poster viewing. People can choose whether or not to attend and can bring their beverages along for this very informal (and hopefully enjoyable!) event.

A link to "How to contribute data" should be posted on the front page.

Response: Done. See the top of the left bar on the front page. Under the "Project" heading is a link labeled "How to Contribute Data".

Future projections should be posted within the site news column.

Response: Done. See the bottom of the right bar on the front page. Under the "Project" heading is an item inviting researchers to visit with us at PAG. In addition to announcements like that, we will be putting up items like:

Coming soon: schedule of updates by data type.
MAGIs will be added to the database for the [X month] release!
etc.

The icon for the Development Curation Tools and interface (sandbox) should be made more obviously different. They like the one that is stamped across the front with "CURATION".

Response: we will keep the pail and shovel, but will stamp "Playground" across the icon so that it is more similar to the look of the "Curation" icon.

We should create a "hot new gene list" and make it somehow accessible.

Response: We will create such a list and initially populate it with newly described genes from the Editorial Board. On that page will be a mechanism for researchers to suggest additions to the list, and new additions will appear toward the top. Like the "What's New" column on the right hand of the front page, this list will rotate off older entries (anything >10) and a separate page will be made accessible for past entries via a link marked "See older items". All entries will be marked with the date they were added to the page.

Try to get a grant from RCN (Research Coordination Networks) rather than NSF's Plant Genome.

Response: The RCN does look appropriate for our type of work. The last solicitation listed the end of June as the due date, and we will keep an eye out for the next cycle's solicitation so that we can prepare something competitive for it.

Look at USDA competitive grant requests for proposals just after the new year.

Response: We will keep an eye out for these and apply as is appropriate.

Stop historical curation altogether.

Response: The only historical curation we will engage in is that which is required for the Editorial Board and that which is required to repair any identified misinformation already residing within the database. Otherwise, we will rely solely upon Community Curators to curate old data into the database.

Devote a lot of time to new views and approaches to display.

Response: TS already does this and will continue to do so. As stated above, we hope that the new Computational Biologist's new perspective and ability to create such views will also be useful for addressing this directive.

Meetings should be yearly and should occur at or near major airports (Chicago, St. Louis, DFW). The meetings should not coincide with the Maize Meeting.

Response: We think that meeting toward the beginning of December in Chicago might be good for 2007. If there is some reason not to plan to try for that time period or location, please let us know!