MaizeGDB Working Group Report, January 2008

Volker Brendel, Ed Buckler, Karen Cone, Mike Freeling, Owen Hoekenga, Lukas Mueller, Marty Sachs, Pat Schnable, Tom Slezak, Anne Sylvester, Doreen Ware

The MaizeGDB group has continued to deliver excellent results on very limited resources. The Working Group continues to be impressed with the leadership, teamwork, creativity, and dedication of all involved. Some specific comments on the status presented January 12, 2008 at the PAG conference include:

- Computer systems (Darwin)
    - Migration plan to new servers appears solid; the use of identical systems for the 3 separate functions is good. 4GB RAM may still be adequate for the current database size, but will eventually need to be increased. The data backup plan has excellent redundancy. Upgrading to Oracle 11i will be done in a very prudent manner. Very high overall professionalism is exhibited in all aspects of the computer systems.
    - The conversion of external data in spreadsheet format is necessarily complex and manual intervention raises confidence in the quality of the data added. Are the procedures documented sufficiently so that somebody other than Darwin can run them?
- Maps (Mary)
    - The Working Group is pleased with the excellent focus on mapping issues that have the most value moving forward. Progress on ontologies is also impressive.
    - Clarification is needed on the future of FPC map maintenance and evolution (as new data indicates needs to re-organize contigs.) Who will be responsible for ownership, curation, and hosting of these maps? What is Kari Soderlund's role and how long is it funded? MaizeGDB does not have sufficient resources at present to take on a large physical map support burden.
    - Separate graphical viewer(s) for physical maps are a low priority. Physical maps will have a role in the genome browser that will be developed for MaizeGDB.
- Outreach and Curation (Lisa)
    - The Working Group applauds the outreach efforts accomplished in 2007 and thinks that exploiting videos would be a good way to leverage the outreach effort.
    - There was much discussion of problems arising from having an inadequate controlled vocabulary for maize mutant phenotypes. It was agreed that Lisa might attempt to compile vocabularies used by recent phenotypers, proof them, and publish them under "guidelines" on MaizeGDB. However, it was also suggested that creative algorithms be used to find phenotypes using the keywords already entered in the database. Adding keywords individually to mutant accessions is not warranted.

    - The Working Group supports the plans for community curation. This is the only way that curation can be accomplished on the limited funding.
- Developing Community Support (Trent)
    - Trent's work in this area continues to be exemplary. The Web site has shown a continuing series of improvements that provide users with faster access to the pages they choose to use most. His analysis of site usage confirms that these changes have had a large positive impact. His responsiveness to community requests is excellent.
- Browser (Tanner)
    - The Working Group recognizes that universal agreement will never be reached on what are the best features for a genome browser. Additionally, the genome communities will continue to develop new and powerful browser advances as the number of plant genomes available makes comparative genomics become the focus of many genomic queries. Thus, it makes sense for MaizeGDB to plan for regular browser evolution by isolating the database from the browser to whatever degree possible, so that the pain of either evolving the DB schema or replacing the browser can be minimized.
    - The browser survey performed in 2007 was a good-faith effort to examine the maize community's satisfaction with the key current browser choices.
    - The charter of MaizeGDB is to support the needs of the maize community. As the genome browser will be the main interface of this community to MaizeGDB, the choice of which browser to use should depend more on maximizing the utility for MaizeGDB users rather than any external factors (e.g. what other resources are using which browsers, etc.)
- Leadership (Carolyn)
    - The Working Group is impressed with the leadership of the project and the trajectory for the future. The group has met the challenge posed in the last Working Group report to plan the transition to a sequence-centric maize world (a prediction that was met with great skepticism when it was voiced at the very first MaizeGDB Working Group meeting many years ago.)
    - Maize genomic sequence data will be made available on multiple community resources. The primary value-added for MaizeGDB will be the tight linkages to multiple additional data (across multiple maize genomes, eventually)
        - mutant and phenotype data
        - genetic and physical maps
        - the "official" gene model, along with other gene models from multiple users (who defines the official model is not yet defined)
        - community annotation
        - host sequence-based search or alignment tools, but only if the tools of others are insufficient to the special needs of maize research
    - Many tools for automated fact extraction from free text exist. MaizeGDB should pursue funding to evaluate and acquire tools to automate the extraction of genotypic and phenotypic data from the literature. Combined with community annotation, this is the only approach that has any

possibility of scaling. MaizeGDB should not march into the swamp of trying to invent a better free-text extraction tool themselves.

The Working group has several recommendations to make:

1. Preparations need to be made for massive scaling of the database.

MaizGDB currently has 166 tables and ~4 million total rows of data. Experience from other genome projects transitioning from map-centric to sequence-centric suggests that at least 2 orders of magnitude more data will be present within a few years. This will stress all aspects of MaizeGDB, including computing hardware, data import from external sites, the browser interface, community curation, etc. The largest impact will be the effects of scaling on the MaizeGDB staff as user requests accumulate at an ever-increasing rate. The basic infrastructure available at MaizeGDB is sufficiently powerful to run the website and the genome browser. The Oracle database used as a database backend is a commercial database backend that should be able to handle databases several orders of magnitudes larger than the current maizeGDB, given enough disk space. One component that may be missing from the server infrastructure is a large, multi-Terabyte, RAID5 disk server along with a backup infrastructure for future expansion into very large datasets.

The MaizeGDB group needs to begin to plan for this data avalanche. Some suggestions to consider include:

- Increase the amount of automation used to import external data. Lobby the funding agencies to require making all data available in machine-parseable formats .It would be good to have a webpage for users to upload their tab-delimited file that would then be checked for data integrity and issue alerts about inconsistencies. When problems are found, uploads don't proceed until resolved. The current manual effort expended on data uploads simply won't continue to scale much further, and could be a bottleneck to MaizeGDB's growth and utility.
- Present the Working Group a draft plan for how genome sequence data will be integrated into MaizeGDB and connected to all the other data sources (including but not limited to: SNPs, gene models, gene expression data, protein structure data and models, pathway models, etc.)
- Examine the level of hardware utilized by other major genome projects and prepare a scaling plan for the next 6 years. List the assumptions you make about the loads that multiple users of a sophisticated browser will put on MaizeGDB. Put some thought into how you can automate the testing of your system so that you will not be surprised by a sudden surge of usage that brings your performance to unacceptable levels. It was noted that not very many BLAST jobs are run at MaizeGDB. This may be due to the fact that BLAST jobs are run on the webserver and therefore are slow. If the BLAST jobs or other compute-heavy tools would be run on separate server(s) from the database and Web servers, the resulting improvement in performance would likely attract more users.

2. Continue to evolve your relationship with Gramene, NCBI, and other global resources

MaizeGDB needs to acquire, integrate, and display maize-centric information in a matter that optimizes both the current value to the maize community as well as the guaranteed accessibility to that data. This means that MaizeGDB needs to negotiate access to all pertinent data (phenotypes, mutants, maps, sequence, genes, expression, etc.) and maintain local control over all first-class data objects within the maize community domain. The Working Group would like the MaizeGDB team to define the boundaries between MaizeGDB and other major resources (NCBI, Gramene, etc.) and the points of linkage between them. While gratuitous overlap with other resources should be avoided, the Working Group believes that no primary maize data stored in MaizeGDB would be gratuitous. A measure of this would be that no user query on any maize data type should be dependant upon an external data resource being available over the network; MaizeGDB should integrate all such information internally. The Working Group would like to hear thoughts of how cross-species queries related to maize might best be handled, as the proper boundaries here are less easy to define.

3. Leverage your success

MaizeGDB is at the leading edge of single-plant community resources for model organisms of economic importance. How might your success be leveraged by the funding agencies to serve the many other plant genome projects that are sure to follow in the wake of large decreases in genome sequencing costs? Does it make more sense to "franchise" MaizeGDB as a template for other communities to evolve independently, or should the US plant genome databases be housed in an enduring center (independent of the research efforts), where integration and cross-species database queries could be optimized? This is a good time to talk with your funding sponsors to see what kind of thought-leadership they already have in this area and fill any gaps that you find. We are already at or near the tipping point where the cost of integrating and managing genomic data will exceed the cost of generating it. The funding agencies can no longer assume that each plant species project can somehow manage to wrangle its own data on minimal funding, nor that the cost of independently recreating databases and interfaces for each species is the best use of taxpayer dollars. The MaizeGDB group is in a good position to consider whether their success is a model that should be leveraged across a much wider range of plant communities.

Outgoing chair's comments

Tom Slezak would like to thank all associated with MaizeGDB and the Working Group for their friendship and hospitality since the first meeting in late 2002. It has been an incredible experience to see the growth and maturity not only of the MaizeGDB team but also of the entire community that it serves. I believe that this time of transition to a sequence-centric system is also the perfect time to transition to a sequence-centric new chair. I am delighted to hear that Mihai Pop is interested in taking on the role, as he was my first choice as a potential replacement. Mihai is an excellent computer scientist with strong "TIGR roots" and I think he has the perfect set of skills and insights to help objectively guide MaizeGDB through the exciting scaling challenges ahead. I wish him well and I will always be available to both Mihai and Carolyn for any help needed during the transition. Finally, I would like to give a special thanks to Leland Ellis, the original

USDA/ARS program manager for MaizeGDB, for twisting my arm hard enough in 2002 to take on this position.