MaizeGDB Working Group Report

Working group members: Alice Barkan, Qunfeng Dong, David Jackson, Thomas Lübberstedt, Eric Lyons, Adam Phillippy (chair), Marty Sachs, Mark Settles, Nathan Springer

This report summarizes the recommendations made by the MaizeGDB Working Group following the teleconference held on August 6, 2013. Absent from the meeting were Thomas Lübberstedt, Marty Sachs, and Nathan Springer, but all members contributed to the following report.

Executive Summary

The working group recognizes the outstanding progress made by the MaizeGDB team over the past year. It is evident that past recommendations are being acted upon and that current development initiatives address the needs of the maize genomics research community. We support the objectives outlined in the MaizeGDB 5-year project plan, and offer the below recommendations towards meeting these objectives. Highlighted recommendations include:

- Officially rollout the MaizeGDB Alpha interface only after critical known issues have been addressed and the site has been adequately user-tested.
- Continue to build a close relationship with NCBI and the Genome Reference Consortium to assist maize reference genome stewardship.
- Clearly define, document, and track the processes of reviewing, accepting, and incorporating user-submitted improvements.
- Explore potential collaborations with the iPlant community. In particular, the utilization of the "Powered by iPlant" cyberinfrastructure program.
- Improve data access by providing flexible queries, bulk database downloads, and programmatic access via web-based APIs.
- Prepare for a dramatic increase of maize genomic diversity data, to include multiple reference genomes and high-throughput SNP screens.
- Prioritize community engagement and incentivize crowdsourced solutions to curation and computational needs.
- Improve communication and solution sharing with operators of other model organism and breeding databases.

Recommendations

MaizeGDB Alpha Rollout

There remains missing functionality from the MaizeGDB Alpha site that must be implemented and tested before the official rollout (many listed in Appendix of the MaizeGDB Status Report). Of particular importance, the search box feature and other tools supported by the current version should be fully functional. Members of the working group are concerned that the Alpha site will not be ready in time for the planned November release. Thus, we recommend that if the critical issues cannot be resolved by November, the release be postponed until the Alpha site contains all the functionality of the current site and has been thoroughly user-tested. The Alpha site represents a dramatic change to the interface and the transition should be approached cautiously.

Transition to iPlant infrastructure

The working group was asked to consider a transition of MaizeGDB to an iPlant-based infrastructure. We support this idea, as it could allow resources currently dedicated to hardware support to be redirected to software support. Discussions with key members of the iPlant

Collaborative should be initiated to discuss the feasibility of such a transition. If all parties agree, a careful, step-wise transition is suggested, beginning with low-risk integrations. Each integration step should be followed by an evaluation period, to ensure MaizeGDB service is not interrupted and a tangible benefit is realized. Initial steps could include federation with the iPlant authentication system and connection with the iPlant Data Store. To prioritize intermediate steps, bottlenecks of the current MaizeGDB architecture should be identified and those with the most to gain transitioned first. For example, the usability of MaizeGDB web services such as BLAST, which have known performance issues, could potentially be improved by utilizing iPlant computational resources. Ultimately, MaizeGDB could leverage the scalability of the full iPlant infrastructure for data storage, computation, and hosting, as applicable.

Stability of the iPlant project funding is a potential concern. Continued funding of such infrastructure projects is not guaranteed, and a transition to the iPlant infrastructure should only be undertaken if it is known that iPlant resources will be maintained over a sufficient timeframe (e.g. at least 5 years).

Current and emerging issues

The working group was asked to evaluate if MaizeGDB is meeting the current needs of its users, and to identify any emerging issues that may challenge the project going forward. We are pleased with the current direction of MaizeGDB development and think the project is responding well to the current needs of its users. However, future development plans are ambitious and we expect that careful resource allocation will be the primary challenge to MaizeGDB over the coming years.

Over the next five years, with the continued drop in sequencing costs, we expect the amount of genomic diversity data to greatly increase. These datasets will include both additional reference genomes and low-coverage SNP studies, combined with phenotypic and pedigree metadata. Data will also become more heterogeneous, as sequencing-based expression and epigenetic studies become more prevalent. Thus, it is critically important to consider scalability when planning the future of MaizeGDB. In addition, analysis and display tools should be designed to integrate heterogeneous data from multiple sources. To handle these heterogeneous and rapidly evolving data types, we recommend that MaizeGDB follow a modular design philosophy with clear separation between data storage, access, analysis, visualization, and user interaction. This will enable individual components to be updated as needed without impacting the entire system.

These are significant challenges that cannot be solely addressed by MaizeGDB, and instead must be addressed by the greater community. We recommend MaizeGDB prioritize development on critical operational tasks, and attack these larger challenges through collaboration and community outreach. Crowdsourcing is one potential avenue for solving emerging issues. We are pleased to see MaizeGDB encouraging user-submitted assembly and annotation corrections, and this activity should be vigorously pursued. In addition, we suggest engaging the community to encourage and guide tool development. To begin, a list of current computational and analysis challenges could be identified and advertised on the MaizeGDB website. This would serve as both a planning and outreach exercise. Future directions could focus on better incentivizing community contributions to assembly, annotation, analysis, and software challenges.

Lastly, the related domains of human and model organism genetics should be routinely surveyed for new tools and approaches that could be easily adopted. The challenges faced by MaizeGDB are not unique, and externally developed solutions should be leveraged as much as possible. This includes continued and active collaboration with the Genome Reference Consortium (GRC) and National Center for Biotechnology Information (NCBI).

Objective 1: Support stewardship of maize genome sequences and forthcoming diverse maize sequences.

The adoption of the GRC data model is an excellent development that will allow reference genome updates between major assembly releases and facilitate better coordination with the other maize genomics groups, such as Gramene. A close collaboration with NCBI should ease the transition and successful implementation of the model will enable continued curation and improvement of the maize genome indefinitely. This should continue to be a primary focus of MaizeGDB.

For this plan to succeed, it is critical to define and document how corrections will be reviewed, accepted, and incorporated into the maize reference assembly. Ideally, users should be able to track the progress of their contributions and these contributions should be clearly indicated and credited in the MaizeGDB browser interface. This will provide positive feedback and avoid frustrating users by clearly documenting the history of their contributions. In addition, users should be provided the ability to flag low-quality or inconsistent data within MaizeGDB. This will help prioritize regions of the genome for improvement and better communicate data quality to all users. Finally, it should be decided as soon as possible how MaizeGDB patches and corrections will be included in the Gramene v4 reference assembly.

In addition to ongoing sequence stewardship, MaizeGDB should plan to expand beyond a single reference sequence and develop tools for the storage and display of multiple alleles and reference sequences. Future users will be interested in exploring not just a single reference genome, but also the diversity between maize genomes and associated phenotypes.

Objective 2: Create tools to enhance access to expanded datasets that reveal gene function and datasets for genetic and breeding analyses.

This objective is in the early stages of development, but the idea of linking functional and breeding analyses with genomic data is promising. A survey of related resources such as www.integratedbreeding.net should be undertaken to identify potential opportunities for tool reuse and collaborative development. In response to the MaizeGDB Status Report, we suggest that "stress response" be added to the list of pathway categories scheduled for manual curation, possibly in combination with the "pest response" category.

Objective 3: Deploy tools to increase user-specified flexible queries.

This is an important objective and we endorse both proposed solutions (BioMart or InterMine). In addition to flexible web queries, we suggest that MaizeGDB support programmatic access via a well-defined API and that important datasets be made available for bulk download. Ultimately, the entire MaizeGDB database, or its most critical components, should be made available for download (possibly as a virtual machine) along with schema documentation. This would serve as an important archive of the information stored at MaizeGDB, and allow power users to install a local mirror for computationally intensive queries. All of the above suggestions require thorough documentation and well-defined releases of the MaizeGDB data and software infrastructure.

The working group also discussed the utility of private workspaces within MaizeGDB for the analysis of unreleased data. It is not sustainable for all users to upload and analyze private data, but it may be useful to enable private group sharing for important collaborative projects. This could be supported by providing MaizeGDB VMs, or by creating private group workspaces on the primary MaizeGDB website.

Objective 4: Provide community support services, training and documentation, meeting coordination, and support for community elections and surveys.

We commend the MaizeGDB team on their continued dedication to outreach and support of the maize genome research community. It is in the interest of MaizeGDB to continue these efforts with full force and act as the center of mass for maize genomics. Thus, all efforts should be made

to keep the website up-to-date with current events, provide links to important projects in the field, and to utilize the MaizeGDB Twitter account to distribute important news to the community. Additional efforts should be made to help researchers interact with the increasingly complex and large datasets stored at MaizeGDB. This could include continued training on the use of analysis tools, and the development of new tools that are able to distill multiple sources of data into interpretable reports. Also, researchers should be encouraged to contact the MaizeGDB team for advice, help, and custom development (if appropriate). Lastly, advances in maize genetics that were facilitated in part by MaizeGDB should be actively communicated back to important stakeholders in agriculture to highlight the benefit of this resource.

As mentioned above, the challenges faced by MaizeGDB are not unique; there are many model organism databases across the world with similar missions. Thus, we encourage MaizeGDB to build closer ties with relevant organizations (e.g. GMOD, InterMine, Cytoscape, Gates Foundation, iPlant, Integrated Breeding Platform, etc.) to exchange both problems and solutions. If possible, a meeting or workshop should be arranged (e.g. at the annual Plant and Animal Genome Meeting) to encourage better communication between model organism database groups.