MaizeGDB 2016 Working Group Report

Working Group members: Alice Barkan, Qunfeng Dong, David Jackson, Thomas Lubberstedt, 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 February 29, 2016. All members of the Working Group have read and approve of this report.

Executive Summary

The Working Group recognizes the determined work of the MaizeGDB team over the past three years, which included the introduction of a redesigned web interface and management of a prolonged leadership transition. During this period, we were happy to see progress made on our last set of recommendations, including the successful delivery of a new website, collaboration with the NCBI Genome Reference Consortium, improved data access options, ongoing community engagement, and improved communication with the operators of other agricultural databases.

With a temporary staff shortfall ahead, we believe it is critical for MaizeGDB to identify what it is uniquely positioned to contribute and focus on these key areas. This process would include comparing to related services (e.g. Gramene) to identify what features may be redundant, and focus on delivering high-quality data to the maize community. We note that new tools can be expensive to develop, and so existing solutions should be leveraged from other sources when possible. New development should be prioritized with careful cost-benefit analysis. Further, we encourage the USDA to fill vacated positions as quickly as possible, emphasizing expertise in genomics, sequencing, and curation. Lastly, MaizeGDB should continue to incentivize community involvement by providing added value to users who submit data, improving the usability of key features, and establishing trust in the quality of MaizeGDB data and services.

In a field awash in new data, it is important that MaizeGDB provide a stable, centralized, maintained, and archival resource for the maize genome community. How to provide value to users in the face of dramatically increasing size of maize datasets should be a primary focus of the upcoming 5-year plan. Our highlighted recommendations include:

- Define what MaizeGDB is uniquely positioned to contribute, versus related services such as Gramene, and narrow focus to these topics, emphasizing quality over quantity
- Focus the next 5-year plan on addressing the problems of the maize pan-genome and continual influx of new genomes and datasets
- Quickly fill vacancies, targeting expertise in genomics, sequencing data, and curation
- Evaluate and work with the community and other database platforms to leverage existing solutions rather than reinventing
- Continue to perform expert surveys and careful cost-benefit analyses to prioritize future tasks and plans
- Continue to encourage community engagement and incentivize user-submitted datasets and improvements

Working Group Charges

The MaizeGDB team specifically asked the Working Group for guidance regarding the following four topics. Our suggestions follow.

Genome Assembly Stewardship

The Working Group affirms that stewardship of the maize reference genomes is an important function of MaizeGDB. In particular, we would encourage a "back to basics" approach where the high-quality structural and functional annotation of these genomes is made a top priority. Obtaining external funding for genome curation is extremely difficult, and so this presents a service that MaizeGDB is uniquely positioned to contribute. Further incentivizing and facilitating user improvements to the assemblies and annotation would help advance such work. However, these manual curation efforts should be limited to only a few of the highest quality genomes, as attempting to extend this service to all forthcoming maize genomes would be impossible. In addition, 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.

For user-owned data, MaizeGDB is positioned to provide a number of useful services to the community, such as assisting users with NCBI submissions and enabling users to upload and browse their own custom tracks alongside public MaizeGDB data. This would allow users to explore their own genomes and data in the context of well-curated reference genomes. We encourage these efforts. Several platforms currently provide similar services (e.g. Gramene, CyVerse), and MaizeGDB should talk with these groups to explore the reuse of existing tools.

Big data set identification, evaluation, and incorporation

With current funding and staffing levels, MaizeGDB cannot be expected to keep pace with the current scale of data generation. Thus, careful prioritization of datasets and curation tasks is key. The Working Group recommends that the incorporation of new datasets should be prioritized based on semi-regular polling of an expert panel, stressing quality and importance over quantity. 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. 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.

Tool development

Software development is an expensive and difficult enterprise, so the Working Group stresses that all development projects should be very carefully considered in terms of their cost versus benefit. In addition, there are often freely available tools and services that may provide the desired functionality. Thus, we suggest that MaizeGDB limit new development efforts, especially until full staffing is restored. In the interim, 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.

Future needs and expectations

The upcoming 5-year planning process provides an opportunity to take stock of future challenges. The Working Group highlights the following needs. First is the 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. A related challenge is 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. Lastly, with these challenges on the horizon, it is important for MaizeGDB to provide a centralized and reliable hub for the maize community with a focus on providing high-quality, reliable data and services.