Report from MaizeGDB Working Group Meeting 09/26/18

Attending: All working group members and members of the MaizeGDB team

Questions posed by MaizeGDB team and responses by the working group

Genome Stewardship and Support

1. MaizeGDB is committed to providing support and resources to managing and making accessible maize genome assemblies and supporting datasets. To support an assembly, we currently require that the submission process to GenBank is initiated, the genome has a BioSample Identifier, and metadata is provided. However, we do not plan to support all genomes at the same level. Selection of genome assemblies for full support will be based on quality (assembly and annotation), historical importance, contribution to maize pan-genome diversity, community impact, and data completeness. It would be helpful to have advice on the methodology to select the most important lines.

   Establish a tiered level support system for genomes based on community need for that genome and include polling on identified characteristics and metrics.

2. We are both using and recommending the v3 assembly and established gene models even though the v4 assembly is better. We have added back the previously rejected gene models into the v4 gene model set, to improve the coverage of gene models. Should we stay with recommending v3, move to v4, or wait for an updated assembly and/or annotation?

   Continue to maintain support for v3. Current v4 annotations have significant issues. Have a clear statement that annotations are not perfect.

3. MaizeGDB historically does not host sequence assembly data directly. Should we be hosting sequence assembly data due to differences in nomenclature with GenBank (particularly chromosome names, and because they are difficult to find)? Downloading full annotation from GenBank is challenging, and only available as nucleotide or protein sequences.

   At a minimum continue to store pointers to raw data (genome assemblies plus annotation files) but recommend storing and making available via FTP reference versions of genome assemblies and gene model annotations directly on MaizeGDB.

Data Curation

4. MaizeGDB has a plan to increase community engagement in data set prioritization for curation (see project plan). With limited curation staff we need to prioritize our curation efforts. How would the WG prioritize these data types: functional annotation, QTL, GWAS, and phenotype? How can we increase community input to identify datasets within each type?

Tracking access rates of submitted trait data/functional annotations can be used to prioritize which datasets are curated and promoted to display on genome browsers.

Storing and curating numerical phenotypic data along with robust tracking of the year, location, and genetic lines the data was collected from should be a higher priority than storing specific QTL and GWAS results. Storing "raw" image or sensor phenotype data is probably not going to be feasible.

Functional annotations. <- yes please, but given current limited resources focus on developing and deploying automated approaches rather than manual annotation - link out to TAIR

5. Transformation and gene editing will remain an important research tool in the foreseeable future. Are there datasets, tools, and or resources we can bring in to facilitate these efforts? Examples include tools for CRISPR design, predicting off-target sites, and curating data on which genes have already been edited.

   Pan-genome level prediction of off target edit sites as well as consistency of guide RNA targeting at a specific locus across genomes/haplotypes. Integrate existing tools (CRISPR-P?), don't need to build new ones from scratch. Check out the Omics website for popular tools to help identify them. Mark and David offered to write text to be added to a CRISPR Maize page on MaizeGDB.

6. There are locus naming inconsistencies between MaizeGDB and GenBank, which percolate through other resources (UniProt, Phytozome, et cetera). How important is it to ensure that naming is consistent? This is a fairly large task. Tool development.

   If reference annotations are hosted at MaizeGDB this issue becomes less important. Maintaining lists of syntelogs across different versions of B73 is quite valuable and should not be very effort/resource intensive (lookup table and listed in gene/locus details if possible). Don't need to try to force standardized gene names across all external databases - issue is more complicated that the WG understood

7. MaizeGDB plans to limit tool development (or integrating 3rd party tools) while focusing on data curation. Tools that could be of broad interest include pan-genome browsers, GxE resources, QTL visualization, GWAS analysis, and tools to better integrate existing data. Are there other tools/resources that would have broad impact?

   Integrate 3rd party tools where possible. Discuss with pan genome tool developers, Paul Comet, NRGene?.

Outside collaboration

8. MaizeGDB is involved in AgBioData, a consortium of 31 databases working together to solve common issues. Following the FAIR data principles (Findable, Accessible, Interoperable, and Reusable) is an important aspect for all databases, and we are working together towards greater compliance. At this time, MaizeGDB struggles the most with getting adequate metadata for large data sets from maize researchers and tracking changes, versioning, and nomenclature. How can we get people to follow FAIR principles?

   Larger question for the AgBioData consortium. Involves getting funding agencies and journals on the same page to help encourage compliance. Work only with data that is FAIR compliant? Educate researchers at the regular Maize meeting. Add a page on FAIR data compliance and templates for data submission.

Overall Comments

The Working Group appreciated being given a detailed list of questions to discuss and provide feedback for. MaizeGDB continues to be an excellent resource for the maize research community and the wider plant community. The team are very responsive to community needs and highly respected for the work they do for both maize researchers as well as their leadership in the larger AgBioData community of databases.