Maize B73 Reference Assembly Update & Release

The Maize Genome Sequencing Project is preparing to release a new version of the maize B73 reference, designated B73 RefGen\_v3. The new reference assembly is improved over the current version, RefGen\_v2, primarily in the inclusion of new genic regions, generated from a whole genome shotgun assembly of 454 sequences, which fill gaps in the current BAC clone-based assembly. Annotation of the amended assembly results in the definition of several hundred new and/or improved gene models. This document describes the schedule and protocol for public release of the new assembly.

The Project will release RefGen\_v3, via maizesequence.org, maizegdb.org, and gramene.org, following submission and acceptance of the following data sets to the International Nucleotide Sequence Database Collaboration (INSDC), also known as DDBJ/EMBL/GenBank:

1. B73 RefGen\_v2 pseudomolecule scaffold sequence and AGP (short for “A Golden Path”, the table that specifies how the component contigs are combined to build the pseudomolecule scaffold sequences).
2. The 454 whole genome shotgun assembly that serves as components of the new assembly.
3. B73 RefGen\_v3 pseudomolecule scaffold sequence and AGP.

Best Practices for Supporting Reference Genomes:

Data providers, including single-organism community databases, multi-organism browsers, and NCBI, have struggled in recent years to maintain standardized and consistent representations of genome data within a given species. The existence of disparate sequence data, coordinate systems, and identifiers harms the scientific community by preventing interoperability and fracturing the research literature. Forcing researchers to reconcile such differences hampers scientific progress. These problems have prompted new policies amongst data providers to insist on INSDC submission as a prerequisite for hosting genome data. Examples include the Browser Genome Release Agreement between the Ensembl, NCBI, and UCSC groups.

In addition to providing a unified source of data, submission to the INSDC ensures legitimacy of the assembly by application of rigorous standards. The vetting process includes, among other aspects:

1. Ensuring that component contigs are already accessioned in DDBJ/EMBL/GenBank.
2. Screening of component contigs for non-target organism contamination.
3. Validating appropriate positioning and classification of gaps.
4. Ensuring that AGP specification agrees with pseudomolecule sequence.
5. Using standardized formatting for accurate representation of all data and metadata.

The risk of not submitting to INSDC prior to release is realized when this validation process necessitates changes to the assembly or coordinate system, thus causing discrepancy with the released version. Experience with the submission of RefGen\_v2, currently in use throughout the community, is illustrative of this problem. While this submission is still in process, feedback from validation has so far included i) contamination of sequence from non-maize organisms; ii) inappropriate gap placement and length representation; iii) unacceptable construction of a “chr0” to represent unanchored scaffolds (chr0 needs to be broken up into individual scaffolds). We are fortunate that GenBank is making allowances for RefGen\_v2 so as to maintain consistency of annotation coordinates with the public release already in use.

Process and Status:

The flow chart illustrates process and status for submissions. For both annotations and AGP the process is iterative until final acceptance: test submission, feedback, revision, new test submission. While issues with v2 identified to this time have been incorporated into the preliminary AGP of v3, the final approved v2 AGP is critical to make a smooth submission of v3 on the heels of v2. Similarly, final approval of v2 annotation files will be important for the submission of v3 annotations, as the vast majority of genes will only require adjustments of coordinates. Final approval of the 454 whole genome shotgun assembly is also on the critical path for release of RefGen\_v3. However, only a relatively small subset of the entire assembly is relevant to the v3 AGP, and NCBI is giving priority to these. These have already passed contamination screening and we do not anticipate any additional issues to what should be a straightforward submission of nucleotide sequence.

