Maize Genetics Executive Committee Report at the 51st Maize Genetics Conference, St. Charles, IL March 2009.

Slides 2 and 4 from the online powerpoint http: http://www.maizegdb.org/mgec.php

Activities 2008-2009.
Accepted responsibility for organizing the maize session at PAG (Plant and Animal Genome meetings)
Partnered with the National Corn Growers (Pam Johnson) to ensure funding for the NSF-PGRP
Discussed the future of the NSF-PGRP with James Collins, assistant director for biological sciences at NSF
Corrected public record, in the ASPB Newlsetter, regarding publication productivity of the maize community

Activities Planned 2009-1010
Create an outreach slide set that highlights the impact of public-sector investments in maize genetics research
Prepare meeting report for funding agencies (strategy to collect metrics in the future) [e.g., funding sources for the research to be presented at the annual Maize Meeting, can be entered on forms for submitting abstracts. (editors' note, from MGEC verbal report at Maize Meeting)]
Identify community priorities for future public-sector investment

Please see inside front cover for current membership of the committee, which includes a newly created Asian representative appointed by the committee.

Maize Genome Annotation Workshop and Panel Discussion with Cooperators

Maize Meeting Mar 13, 2009 St. Charles IL

Mike Muszynski chair; Panel Members Carolyn Lawrence, Jeff Bennetzen, Yan Fu, Jim Uphaus, Volker Brendel, and Gernot Presting

www.maizegdb.org/POPcorn/annotation_forum.php

Outline of report

Executive Summary
Meeting's Context of the Panel Discussion
Rough Transcript of the Panel Discussion
Feedback Gathered Subsequent to the Discussion

Executive Summary

The maize genome sequencing projects are complete or nearly complete, and the next order of business is to assign gene structure and functional annotation as well as other data (e.g., cytological positions of centromeres, etc.). For the group sequencing B73 (the reference sequence), part of the deliverables are the annotations (estimated project completion date: February 2010). At the same time, various other groups are also annotating or plan to annotate. A panel of six drawn from broad disciplines was assembled to discuss annotation in the broadest sense, but the discussion focused on annotation of genes (both structure and placement) with emphasis on how individual researchers could contribute to the annotation.

Meeting's Context of the Panel Discussion

Questions to be asked during the panel discussion were submitted by maize cooperators and may be viewed online via MaizeGDB at http://shrimp1.gdcb.iastate.edu/mm2009/question.php.

Genome Sequences: What's New (Chair, Mike Muszynski)
Doreen Ware, Sequence and Analysis of the Maize B73 Genome
Dan Rokhsar, Update on the Mo17 Genome Sequencing Project
Octavio Martinez de la Vega, The Characterization of the Palomero Toluqueño Genome
Dinner
Community Forum on Gene Annotation (Chair, Mike Muszynski)
Pam Johnson from NCGA, Research and the Recession
Volker Brendel, Community Annotation at MaizeGDB/PlantGDB
Discussion with Panel Members Carolyn Lawrence, Jeff Bennetzen, Yan Fu, Jim Uphaus, Volker Brendel, and Gernot Presting

Rough Transcript of the Panel Discussion

Jeff Bennetzen: short introduction suggesting some next endeavors for the community: zeanomics, more genomes (999), hypothesis testing in the form of functional biology, promote maize as the model.

Carolyn Lawrence: MaizeGDB offers coordination. For example, the tracks in the MaizeGDB Genome Browser are generated by the community. A problem, however, is that these are based on different GenBank releases which causes problems with aligning track content relative to a particular BAC. MaizeGDB could house quarterly releases. Would this be helpful?

Audience Member: Will MaizeGDB Genome Browser provide links to NCBI and vice versa?

Brian White-Smith: Links from NCBI's Entrez to MaizeGDB are being worked on now. Notes that 600 out of 20,000 genes from mRNAs are in GenBank. Will use FL-cDNA from B73. All RefSeq entries are being updated to use the B73 sequence as the exemplar.

Dan Rokhsar: Likes the idea of quarterly releases, but isn't maizesequence.org already doing this?

Carolyn: It isn't clear this is happening; if it is, it isn't well advertised. Certainly we would prefer this be handled by maizesequence.org.

Doreen Ware: Releases are generated, trying for quarterly. The process will stay in flux for a few more months. Note on version names: latest release is 3b.50. The 'b' indicates an update of the annotation, '.50' is the Ensembl version. The assembly was not changed. An update is now or shortly will be in progress. The group is trying to maintain mappings across releases where there are no changes but this is usually not possible. Doing 6 month updates with Gramene. Recommends staying with quarterly updates for now.

Carolyn: Need to know when releases are coming up so that annotation groups and the community at large can plan. Currently maizesequence.org announces a release and makes available the GenBank freeze date, but does not publicize the freeze date/data in advance.

John Fowler: Requested a web page of cautions, e.g. outlining the different GenBank releases, flipped contigs, and common gene annotation errors to watch for. Would it be possible for MaizeGDB to put up a page of warnings?

Carolyn: Yes, we can do that. In addition, Lisa Harper can make a movie showing "scary things".

Taner Sen: My talk describing the MaizeGDB Genome Browser functionality included a page of cautions that can be used to seed the MaizeGDB page/movie.

Audience Member: What about gene expression in spatial and temporal scales? Arabidopsis has nice data sets and viewers. What tools are coming for maize?

Volker Brendel: Don't have answers for how to handle all the different types of data. Consider revisiting DAS (Distributed Annotation System). This has been available for a while but has been problematic. Recommends trying to revive this type of data sharing. Some groups are using Google Maps for visualizing. This will help but nothing available now is sufficient.

Fusheng Wei: How often […what?...] given that data changes daily?

Carolyn: Suggests setting up a forum and putting a page on MaizeGDB: if you are interested, get involved.

Pat Schnable: We could have students help with annotation, but will need to test their accuracy and create a set of standards.

Anne Sylvester: Issues are not new; how much communication has there been with other groups tackling the same issues? What about iPlant?

Volker: There is an iPlant meeting on annotation coming up in St. Louis. (http://iplantcollaborative.org/about-ipc/education-outreach-and-training)

Brian: On the subject of community annotation: NCBI has strict requirements before anyone is allowed to contribute annotation, including publications and wet lab experience. If the pseudomolecules were submitted to GenBank, RefSeq could be used to provide first pass on gene models and these could made available to MaizeGDB.

Carolyn: Does updating RefSeq annotation require wet lab experimentation and a peer reviewed publication?

Brian: Yes.

Doreen: Advocates for community annotation. However, other groups have done this but not had much success. How about an "annotation jamboree" right before the Maize Meeting? Success will depend on the maize community. Capturing meta-data is difficult but important. Start making an effort now, use ontologies even if they are imperfect. Meta-data must be computer-readable as well as human-readable. It's important to learn what didn't work in previous efforts at community annotation.

Audience Member: 1) It is difficult to find and keep one's place in the genome browsers. Can we see a genetic map alongside the MaizeGDB Genome Browser's representation of the genome? 2) What happens to community annotation when the build changes?

Carolyn: Keeping the linkage map open while browsing is a technical issue. Regarding the fate of community annotation after a new release, not sure what the answer is.

Dan: An online forum at MaizeGDB is a good idea. Gave history of annotation efforts on human genome: three competing groups with incompatible data finally got together and found a core set of genes they could agree on and publicized a list of those they did not agree on. Errors found in automated annotation helped improve the automated process.

Brian: More history: when there was a new build of the human genome, genes would disappear. NCBI has a web page that highlights genes the three annotation groups can't agree on. GenBank could do the same for maize. Described method for generating RefSeqs for human. Suggests waiting on gene models until sequence assembly is more solid.

Jeff: A tool is needed now, and we can't wait for a perfect assembly. For one thing, we can use community annotation to get an idea of how good the automated process is. Noted that the first release of the rice genome had a 50% error in gene calls, which could have been easily discovered before the release. NCBI requirements for annotation privileges are not compatible with distributed annotation and it is unreasonable to let GenBank create annotation based upon a computational assessment then require that to fix it the researchers to wet lab experimentation. Suggest that the maize community doesn't need to follow NCBI's practices and that creating a RefSeq for maize could be a problem.

Audience Member: Are annotation jamborees feasible? Has this been done?

Jeff: This has been done to provide a 1st round of quality annotation.

Panel Member: Computers can't match human eyes. There needs to be an incentive or enforced requirement to annotate, then deposit the annotation in MaizeGDB. How? Require this for publication? For NSF funding? Cost must have a benefit. Who would coordinate this? Perhaps there could be coordinators by area (transposons, retrotransposons, etc.).

Yan Fu: Regarding gene prediction: numbers (quality scores) should be available to show how reliable the predictions are.

Carolyn: In the real world carrots don't work. Sticks work. Treat annotation like a lab task: clean up after yourself or else. Example is lab worker assigned to keep a scale clean.

Toni Kazic: Quality of annotation will depend on the quality of tools. It should be easy to find and compare annotation. Even the availability of free text typing fields will help.

Jeff: Reading a pre-submitted question: "Will there be a curated repeat database for maize? It is needed soon." Answer: we have this and it will be available soon.

Audience Member: Comment: we could use undergraduates, but they will need training.

Anne: Tutorials exist but need to be advertised.

As Prepared for the Maize Genetics Executive Committee
Notes taken by Ethalinda Cannon
Inputs from Panel Members

Feedback Gathered Subsequent to the Discussion (contributors include Dan Rokhsar, Doreen Ware, Volker Brendel, Mike Muszynski, Carolyn Lawrence)

It should be noted that some were expecting a broad discussion of the future of maize genetics and this panel and the audience focused immediately on gene structure annotation within the context of community annotation. This is not bad, it's just not what everyone expected.

It was noted that PIs on PGRP grants recently discussed similar issues at their annual meeting in Washington DC.

Folks thought the discussion was good, interesting, and they got an idea of the status of genome sequencing. The update and panel discussion was very useful to help inform the general community regarding the context of the genome and how this changes over time. But it was felt no long-term solution(s) would come from the session.

The community appreciates that the B73 project has made sequence data available from the beginning so that the data could be used as soon as possible. With the rapid release there have been some frustrations associated with the instability, but these frustrations are comparatively small relative to the progress made based on immediate access to the data.

The community needs to be better informed and using MaizeGDB as a mechanism for creating a forum makes a good deal of sense. Many of the questions submitted by the community in advance of the discussion (available via MaizeGDB at http://shrimp1.gdcb.iastate.edu/mm2009/question.php) were not addressed, and the forum would be a good place to post those questions and let the sequencing and annotation groups address the questions.

We need to have definitions of what people have already done, both the genomics and materials and methods. For example, it is very important to have more information about the exact accessions for the lines that were sequenced. Researchers work on specific lines and need to know how their work fits in with the sequenced maize genomes.

There is a great deal of good will among those interested in annotating, and having NCBI involved is useful given their central role. However, the fact that researchers would have to do wet lab biology and publish a peer-reviewed manuscript in order to fix a RefSeq annotation by NCBI causes all researchers polled (conversationally) to be against NCBI doing the only annotation. Nonetheless, there is a critical need for subsequent experimental validation that supports both functional and structural annotations.

In order for fixes to the sequences via any group to be incorporated at NCBI, the community database MaizeGDB needs to be listed as authors on the B73 and Mo17 sequence records. Due to NCBI's ownership and update rules, this would allow the community to own and update the sequence prior to the projects' close.

There is a hope that an annotation working group could be formed.

There is a desire expressed after the session by many (but not all) researchers that there be a funded project to annotate the maize genome (B73) via professional curation. This would likely require a RFP and subsequent proposals, etc.

Funding opportunities to develop serious student involvement in annotation are desired. There are a few undergraduate institutions where genome annotation has been incorporated into classes and/or summer research programs (Buckner – Truman State, Gray – Univ. of Toledo). Perhaps these can be a model for encouraging other institutions to annotate the genome. Can increased funding help build a larger network of these institutions to do this in a coordinated fashion?

An annotation jamboree would go a long way toward annotating the genes and gene structures, but it would need to be uncoupled from the Maize Meeting given that the meeting will be in Italy next year. In addition, working with Robin Buell to find out whether annotation jamborees were genuinely useful for rice would be helpful.

> Related detail on rice learned from Pankaj Jaiswal: Three Rice Annotation Project (RAP) jamborees were held in collaboration with the DDBJ (Takeshi Gojobori and Takeshi Itoh). Experts from databases (e.g., GenBank, Uniprot, EBI, Swissprot, Gramene, IRRI and others) and students from local and rice sequencing consortia labs (IRGSP) convened for weeklong workshops to annotate structures for a selected set of genes with known evidence (ESTs and/or FLcDNA from maize and/or known genes from related and angiosperm plant models). The majority of annotations confirmed computational predictions and only a few involved actual changes to the gene structure. DDBJ did similar workshops for human genome annotations. In the case of the rice #1 RAP jamboree, about 23K genes were manually validated by human eyes and brains. For the annotation of *Saccharomyces*, there were jamborees and in addition to annotation, functional assignment projects were in place where PIs were requested to take one/multiple batches of 5 genes through thorough analysis (expression, localization, proteomics, biochemistry and KO/mutant-phenotype). This was quite successful.

Thinking on the rice model of annotation, it should be noted that there were two assemblies and sets of annotations, and those groups were never able to collaborate and consolidate their findings. We do not want to see this happen in maize. The model of three groups working together in human and consolidating their datasets is a much more attractive target.

Related: If various groups are annotating sequence, the Sequence Ontology absolutely must be used by all participating groups.

The various projects absolutely must publish their freeze and release dates and make those available prior to when the freeze and release occur.

Related: Community annotation would be particularly useful to sort out the poorly annotated regions of the genome. However community annotation needs to be focused and coordinated if it is to produce.

Timelines for when the sequencing projects will actually be done and when the data will transition to MaizeGDB would be most helpful.

Related: The B73 Maize Genome Sequencing Project is funded until February of 2010. The Ware group reports that they will continue to work on improving annotations until that time and plans to continue to host the browser at www.maizesequence.org for a short time after the project ends. Their plan is to fold the maize genome into the Gramene resource.

How is data from new transcriptome analyses going to be collected and used to improve the genome? Multiple groups could conceivably take this approach but how can their data be integrated?

There is a critical need to be able to identify paralogs within maize and orthologs with other species. Standardized nomenclature would be needed to underpin this type of information.

While the Gene Ontology (GO) is excellent for getting gene function, it would be also useful to have other functional information captured (that is not handled by GO) - traits and expression data are obvious ones. Long term we need to be thinking about how to integrate this data creating linkages among the various ontologies must be formed (e.g. between GO and the Trait Ontology). This is something that the maize community could really take advantage of.

Question from session chair Mike Muszynski that was not asked:
Is there a mechanism where community annotators can provide the data to help order and orient sub-BAC fragments? Several genes (e.g. *knotted1, ZMM4, ZMM15*) that have been published and are in GenBank are in several pieces within a BAC contig in the wrong order and orientation. If these are identified by a user/annotator, how can the information be relayed to correct the order of sub-BAC fragments?

Related: MaizeGDB is working with Fusheng Wei and the Maize Genome Sequencing Consortium to find out whether and how to create a forum that would allow this via MaizeGDB (similar to the forum used for this session where researchers could submit questions).

Dan Rokhsar has generated a list of items that must be addressed that should be distributed, potentially via MGEC. He plans to follow up with MGEC separately.