VI. MAIZE GENOME SEQUENCING PROJECT

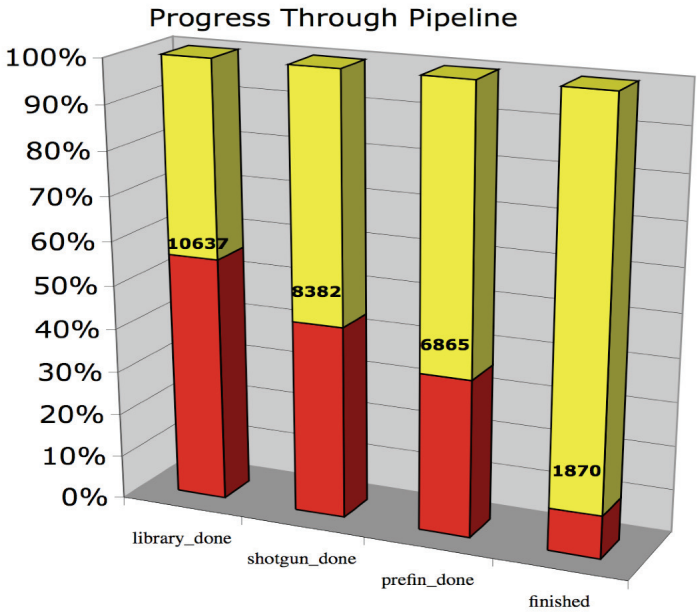
As we enter the second half of the 3-year Maize Genome Sequencing Project, we have begun to significantly accelerate production of draft sequence and are poised to do an effective ramp on improved sequence and submissions. The maize BAC DNA enters our pipeline from shipments of purified and fractioned BACs chosen from the tilepath at Arizona Genomics Institute (AGI). As of April 27, 2007, Washington University Genome Sequencing Center (WUGSC) has received slightly over 12,000 of the predicted 17,000 BAC clones to cover the genome. From each BAC, a plasmid library is created and plated and colonies are picked. We have over 10,630 BACs that have completed this step. Each BAC is given light sequence shotgun of 4-6X coverage, and the sequence is assembled along with fosmids end sequences produced at WUGSC and the original BAC end sequence (BES) performed at AGI. An automated system has been developed to confirm coverage, assembly, and incorporation of BES before completing the production phase. At this stage, the consensus sequence of the assembly is deposited in Genbank as phase I (draft), and 8,382 BACs at this stage now are available. The draft assembly then is screened with a program that will identify repetitive sequence, in order to exclude this sequence from improvement efforts. An automated improvement, or prefinishing, is performed on each BAC, using directed primer walks on subclones that span the gaps for two rounds, if necessary. After completion of the automated prefinishing, a program to utilize genome survey sequence in the form of methyl filtered and high coefficient of time (high C_0t) subtractive libraries, along with sequences from mRNA and cDNA libraries, is run to incorporate existing sequence information. A limited manual improvement effort is then made, using directed primer walking on plasmids along with PCR by finishers located at WUGSC, AGI, and Cold Spring Harbor Laboratory (CSHL). The development and refinement of software for tagging repeats, incorporating existing data, and navigating gene regions of the sequence has taken some extra time, but it is now working smoothly, and we expect to see a rapid increase of clones passing through this stage of the pipeline. We currently have completed 1,870 improved BAC clones. After improvement is finished, an automated pipeline submits the sequence to Genbank as phase I, HTGS_IMPROVED. There are currently 1,251 entries in Genbank with this designation.

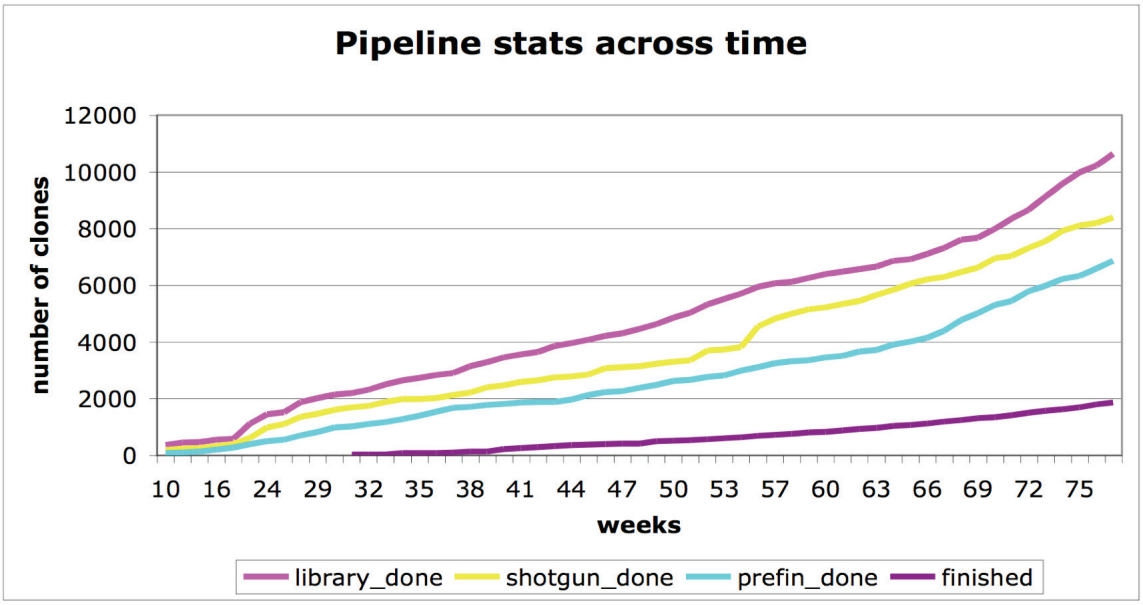
The sequence read data is immediately available in the NCBI Trace Archive and can be downloaded from there (http://www.ncbi.nlm.nih.gov.library.vu.edu.au/Traces/trace.cgi#). In addition to the Phase-1 improved category mentioned above, (HTGS_IMPROVED), others available for download from GenBank immediately upon completion are: 1) HTGS_FULLTOP-2 x 384 paired-end attempts (4-5X coverage); completed shotgun phase; initial assembly; 2) HTGS_PREFIN-completed automated improvement phase (AutoFinish); 3) HTGS_ACTIVEFIN-active work being done by a finisher.

In March 2007, the maize annotation pipeline became a fully automated system, analyzing incoming maize clones on a weekly basis. All maize BACs tagged as Phase I HTGS_IMPROVED (1,251 as of April 25, 2007) have been analyzed to date. The maize browser, available at http://www.maizesequence.org provides public access to maize BACs and their underlying annotations. The website is tightly integrated with Gramene (http://www.gramene.org) and provides cross-linkage for comparative analysis with other cereal genomes. Maize BACs are anchored on the agarose FPC map. Each BAC provides an independent sequence map displaying known order-and-orientation and underlying annotations. Presently, BAC annotations include: ab initio gene prediction using Fgenesh, transposon classification of gene models, MIPS repeat annotation, annotation of mathematically defined repeats, and alignment to a variety of cereal data sets, including maize physical markers. All annotations can be viewed graphically and downloaded through the browser. The browser now also provides BLAST functionality over all maize sequences. These include the sequenced maize BACs, as well as peptide translations of predicted genes. The browser also provides access via DAS to remote annotations produced independently by maize collaborators. One such data set includes TWINSCAN predictions curated through a collaboration between Brad Barbazuk and Michael Brent (NSF #0501758).

As the maize sequencing project enters its second half, annotation and visualization efforts are primed for significant milestones. All maize clones, regardless of status, will be automatically annotated on a regular basis to provide users with preliminary annotations. Mature (improved) BACs will be analyzed using an effective evidence-based gene build strategy in collaboration with Gramene that will provide higher-quality gene models. Improved sequences will also undergo peptide-based analysis, such as InterPro/GO, to provide greater context for gene models. As longer tiles of maize sequences become available, the maize BAC sequence maps will be integrated with the FPC map. This will provide a unified view of the physical and sequence map. Other data sets, such as the maize optical map and full-length cDNAs, will also be integrated into the browser as they become available. In the beginning of 2008, it is also expected that whole genome alignment to related organisms, such as rice and sorghum, will be made available. Finally, preparations are being made for whole genome assembly of the maize genome near the end of the project. The assembly will validate the order-and-orientation of the maize *genespace* and will isolate problem regions.

For further general information, please visit the GSC web site, http://genome.wustl.edu/genome.cgi?GENOME=Zea%20mays%20mays%20cv.%20B73. Weekly updates, usually posted on Friday afternoon, in the form of bar and line graphs are available there, (http://genome.wustl.edu/genome.cgi?GENOME=Zea%20mays%20mays%20cv.%20B73&SECTION=research). For access to the Cold Spring Harbor Browser, please visit www.MaizeSequence.org.





Top-line, library_done; next to top line, shotgun_done; next to bottom line, prefin_done; bottom line, finished.

Submitted by Sandy Clifton
Washington University, St Louis, MO
Apr 27, 2007