Sequencing the Maize B73 Genome Progress Report

--PI: Richard K. Wilson (WUGSC), Co-PIs/Key Personnel: Doreen Ware (CSHL), Rodney Wing (AGI), W. Richard McCombie (CSHL), Patrick Schnable (ISU), Sandra W. Clifton (WUGSC), Srinivas Aluru (ISU), Lincoln Stein (CSHL), Robert Martienssen (CSHL), Robert Fulton (WUGSC)

WUGSC Washington University Genome Sequencing Center, St. Louis, MO; URL: genome.wustl.edu
AGI Arizona Genome Institute, Tucson AZ; URL: genome.arizona.edu
CSHL Cold Spring Harbor Laboratories, Cold Spring Harbor, NY; URL: maizesequence.org
ISU Iowa State University, Ames, IA; URL: www.plantgenomics.iastate.edu

From the Oct 2007 Annual Progress Report, posted in more detail at MaizeGDB: http://www.maizegdb.org/sequencing_project.php.

Overview of Original Project Goals:

Project Objectives:

1. Provide the complete sequence and structures of all maize genes and their locations (in linear order) on both the genetic and physical maps of maize.
2. The gene space of B73 maize (gene sequences and adjacent regulatory regions) should be finished to high quality according to currently acceptable standards
3. If applicable, the sizes of gaps between the genes should be estimated and draft sequences of repetitive DNA between genes presented where possible.
4. The sequence will be fully integrated with the genetic and physical maps.
5. Annotation will include gene models, predicted exon/intron structure, incorporation of EST and full-length cDNA data, gene ontology, and relationship with homologs in other organisms, including but not limited to, the other sequenced plant genomes.
6. Annotation will be coordinated with existing maize community and comparative databases with the eventual goal of generating complete curation of the genomic sequences to a standard set by established model organism databases.

Research Activities and Results:

1. *Choosing minimal tiling path (MTP) clones from maize physical map preparing DNA for sequencing. Responsible PI: R. Wing, AGI.*

Goals: The objective for AGI for year 2 was to complete MTP clone selection in consultation with the WUGSC. AGI prepared all DNA from the selected clones, sheared the DNA, and confirmed identity by generating end sequences for each clone. Once validated, the sheared DNA was shipped to the WUGSC for shotgun library construction and sequencing.

Progress: In Year 1, we built a framework for MTP selection, which combined all the data from the physical map, blastn searches of seed BACs against a BES (BAC end sequence) database, and trace display. We used this pipeline for cloning walking from 3,400 seed BACs chosen in Year 1. We also chose clones according to their fingerprints along contigs due to project time constraints and limited sequence availability.

In year 2 we selected ~8,000 MTP clones to complete the first round of picking. In total, we selected 16,224 BACs (169 96-well plates). Of the selected clones, DNA from 15,400 clones was prepared, sheared and validated for shotgun library construction. The difference between the two numbers is due to failed clones in selection and DNA preparation.

Our initial estimate of 19,000 MTP clones still stands, but we anticipate that number will be reduced to approximately 16,000 MTP clones due to the fact that we selected very large seed BACs. The additional ~600 clones will be used to fill physical gaps between sequences. This selection is ongoing.

Plans: We will manually check all the neighboring clones for any potential gaps. A new BAC or fosmid clone will be selected to fill each gap. We expect 500-1,000 clones will be selected for gap filling. DNA of these clones will be prepared, sheared, verified, and delivered to WUGSC for shotgun library construction and sequencing. In the meanwhile, we will merge map contigs according to the available BAC sequence information.

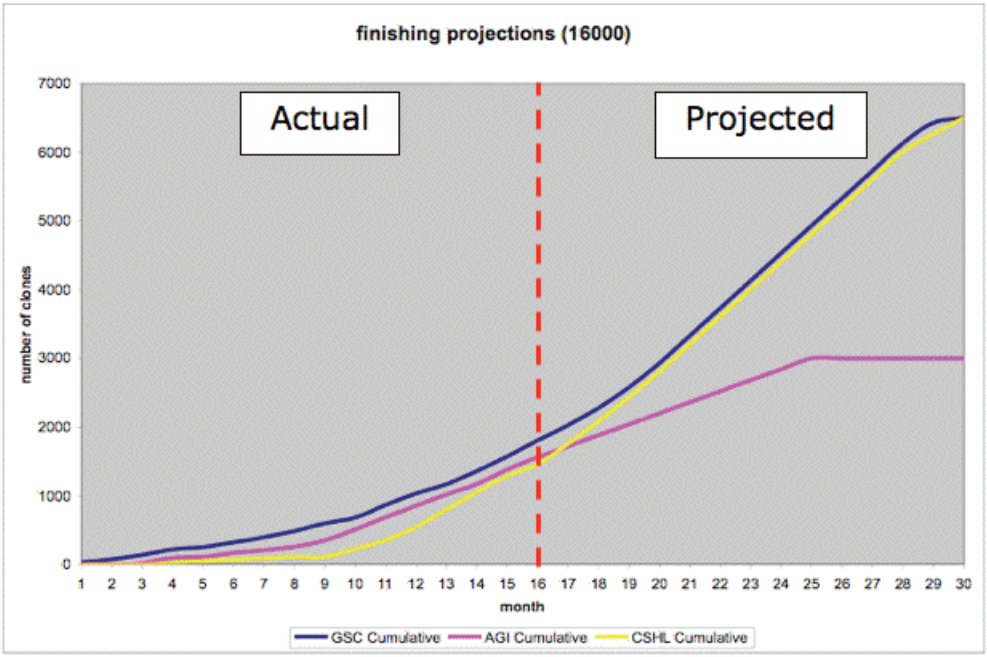
2. BAC sub-clone libraries, production sequencing and assembly, "genespace" boundary determination, automated sequence improvement, and shared finishing of "genespace" in BAC clones. Responsible PI: R. Wilson, WUGSC

Goals: Provide the complete sequence and structures of all maize genes. The gene space of B73 maize (gene sequences and adjacent regulatory regions) should be of finished quality according to currently acceptable standards. If applicable, the sizes of gaps between the genes should be estimated and draft sequences of repetitive DNA between genes presented where possible.

Progress: In year 1, the first 4-6 months of the grant year were spent optimizing procedures and improving computer access and communication among the three Centers (WUGSC, AGI and CSHL) sharing production and finishing tasks. Bimonthly or monthly conference calls among the three centers involved in the production/finishing procedures have been and are continuing to be used as an effective means of anticipating and dealing with any problems in a timely manner. A smoothly operating protocol is now in place and functioning well. In year 2, we followed and improved upon the operating protocol. This includes DNA intake, library construction, the production process, automated sequence improvement, and manual sequence improvement of the gene space. When projects reach the "shotgun_done" status, they proceed to the pre-finishing process where automated directed improvement techniques are employed to improve the overall sequence quality of the projects. Finally, BAC projects are queued for manual improvement (finishing), which involves the complete resolution of all sequences deemed "do finish" to a standard of less than one error per 10,000 bp. The sequence improvement is divided forty percent, forty percent, twenty percent respectively among WUGSC, CSHL, and AGI, but these divisions are very flexible and adjustable, enabling appropriate and effective distribution of the workload. The improvement effort has been steadily increasing over the last half of year two with very comfortable ramp rates in place to complete the remainder of the improvement territory in the timelines established by the project. Details of procedures described above can be accessed in the Year 1 progress report, posted at MaizeGDB with this report.

As of Jan 28, 2008, 15,233 of an estimated 16,000 BAC clones have progressed through the library_done stage, 15,002 projects are shotgun_done, 13,115 are prefin_done, and 5782 are improved. {*EDITOR note*: *At GenBank these categories are called HTGS_FULLTOP, HTGS_PREFIN and HTGS_IMPROVED, respectively. A 4th designation, HTGS_ACTIVEFIN indicates clones being worked on by a finisher.*}

Plans: The Oct. 2007 line graph below shows a projected ramping for the finishing queue.



3. Bioinformatics CSHL. Responsible PI: Doreen Ware.

The CSHL team collaborates closely with teams at WUGSC, AGI, and ISU to define requirements and deliverables.

The project has presented various challenges that have been successfully overcome. One such challenge is the unique nature of the sequencing project. Sequence data and mapped annotations from the project are made immediately available for public use, which is almost unprecedented for a genome project. In 2007 the project focused on refining and automating the sequence annotation pipeline, releasing and improving the project website, as well as outreach activities. The initial target this year was to provide annotation on the *improved* BACs. This objective was broadened to include all BAC sequence, now (Jan 2008) amounting to over 15,000 analyzed clones available for viewing. In addition, the current browser maintains previous versions for public viewing of the BACs at different phases.

Data Analysis:

The analysis builds on the Ensembl analysis pipeline software system (Potter et. al., Genome Res. 14:934-941, 2004). The primary unit of data being analyzed by the annotation pipeline is an intra-BAC contig. Because much of the sequenced BAC clones consists of

unordered stretches of sequence, each is annotated in isolation so as to minimize the mis-annotation of features that span neighboring contigs.

The system for analyzing and visualizing the maize genome is nearly fully automated. The BAC sequence records are retrieved from GenBank on a nightly basis. The analysis pipeline is run for the newly acquired clones on a semi-weekly basis, accessing 160 of the 2,000 processors in a high-performance IBM BlueHelix cluster. Finally, the staging server is synchronized each week with the production server, with backups of the older databases in case of failure.

Ab initio Gene Predictions

The repetitive nature of the genome required the team to explore alternate methods of analysis. Because many analyses are uninformative when they are run on unmasked sequence, there exists the concern that a masked portion of the genome may contain an important gene-related feature. Ab initio gene prediction is performed on non-masked individual contigs with FGENESH using the Monocot parameter (Salamov and Solovye, Genome Res. 10:516-522, 2000; http://softberry.com). Predicted genes are named according to the BAC accession/version, a method code (FG) and a 3-digit number: An example FGENESH predicted gene is "AC177916.2_FG023". Names containing 'FGT' and 'FGP' codes refer to the associated transcript and protein sequences, respectively. Two additional modules, namely Runnable::FgeneshGene and RunnableDB::FgeneshGene, are implemented for the Ensembl analysis pipeline software system to automate the FGENESH gene prediction of genomic sequences as Ensembl gene objects.

Peptide Classification and Protein Annotation

The gene models predicted by FGENESH are classified into three categories by comparing their peptide translations to known proteins. The three categories are TE (Transposable Element: transposon-like gene models), WH (With Homology: similar to known genes), and NH (No Homology: no similarity to known genes). BLASTP is used to generate alignments for all the predicted models against the NCBI Non-Redundant Amino Acid database (NR). Alignments are processed using a confidence level cutoff (E-value of 1e-5) to identify a known protein in the database. A curated list of transposon-like proteins in NR is used to identify the prediction as TE. If a high-confidence alignment corresponds with an NR protein not on the TE list, the prediction is classified as WH. Otherwise, it is classified as NH.

Since the incorporation of non-improved BACs into the analysis pipeline in April 2007, faulty translations have been detected in a subset of gene predictions. These aberrant predictions make up 4.8% of the total predictions on the generated sequence thus far. The error occurs in an isolated condition where the FGENESH gene prediction tool attempts to predict gene model structure without the presence of a starting exon. The aberrant translations are therefore generated from low-fidelity gene predictions. While the issue is being addressed, the browser provides such predictions, but demarcates them as “Corrupt Translations”.

All translations of the FGENESH gene predictions are run through a Protein Annotation Pipeline. The pipeline provides the same annotation modules that are provided through InterProScan, such as PFAM, Prosite, and SignalP. However, present annotations are provided with module-specific IDs; InterPro IDs are not being assigned as of yet. A system is being put into place to extract specific InterPro IDs for each annotated translation. This will allow for the assignment of Gene Ontology (GO) terms to describe protein function. It is anticipated this functionality will become available in December 2007.

Mathematical Repeats

As discussed in the 2006 report, the group has continued to use mathematical repeat as a form of annotation. We have taken a lead from an approached used on rice (Yu et al., “The Genomes of *Oryza sativa*: A History of Duplications”, PLoS Biology, February, 2005) that analyzes the repeat content based on genome-wide oligomer copies. Using an unbiased whole genome shotgun library of maize sequence generated by the Joint Genome Institute that accounts for a genome coverage of 0.25, we built an oligomer index that stores the number of occurrences of each encountered oligomer, at varying sizes. The index is then used to query a given maize sequence for the repeat number at any given basepair. We use a sliding window to average a neighborhood of oligomers. We found that an oligomer size of 20 and a window size of 50 yield as good a specificity as the ISU repeat library while drastically reducing the number of false positives (unique regions that are deemed to be repetitive). Repeats are currently available as discreet tracks in the genome browser.

Repeat Masking

RepeatMasker is run on all sequences using the MIPS repeat library (http://mips.gsf.de). The annotated repeats are classified using the MIPS Repeat Element Catalog (REcat), which provides a hierarchical tree structure of repeat families. The repeat families break down into high-level classes: Class I Retroelements, Class II DNA Transposons, Class III RNA Transposons, and Other. These classes can then be used for more downstream analysis of the genome, such as validation of TE-classified gene predictions. The RepeatMasker annotations are available from the genome browser and provide links back to the MIPS website for details of the repeats.

Cereal Alignments

Sequenced clones are aligned to 54 different cereal data sets. The data sets include important maize data sets such as the maize EST and BAC-end libraries. The current strategy uses the BLAT alignment tool to analyze the incoming maize sequences. Different data sets require different alignment parameters. Several libraries, particularly those that demonstrate high sensitivity to repetitive DNA, are aligned to repeat-masked maize sequences. Also different selection criteria are applied to procured alignments, depending on the nature of the data set. A set that contains uniquely-mapped markers, such as maize BAC ends, are processed using a best-hit approach — only one hit

would appear across the genome. Alignments from data sets that contain more prevalent markers are selected using a cutoff criterion. The analysis makes use of Gramene's BioPipe pipeline. BioPipe (Hoon et al., Genome Res. 13:1904–1915, 2003) at this time is no longer being supported. Based on the lack of support of BioPipe the Maize Sequencing project and Gramene have made a strategic decision to develop resources that will leverage the Ensembl infrastructure. At this point in time cereal alignments are partially automated and evaluation of Exonerate (Slater and Birney, BMC Bioinformatics 6:31, 2005) and the Ensembl runnables to support these features are being evaluated.

Electronic SSR Analysis

The current release of the database supports annotations for mined Simple Sequence Repeat markers (SSRs) from BAC end sequences. The nomenclature for the SSRs was done in conjunction with the Maize community database, MaizeGDB and as part of NSF Project #0333074. The analysis assigns universally unique IDs to novel SSR markers to allow for specific localization of SSR markers. The mined SSR markers are ultimately provided to MaizeGDB in order to provide meaningful entry points to the data. The eSSRs are available as part of CytoView where a drop down menu from the features provides links to both MaizeGDB and Gramene for details. The analysis has been integrated into the annotation process for all sequenced BACs, but the quality control is not complete. Release of this data through the BAC ContigView is anticipated for December 2007.

Analysis for Primary Sequencing Effort:

Minimal tiling path clone selection

Working with project participants a system has been developed for employing BAC-end sequences to aid in ordering contigs near clone ends that provide insight about minimal tiling path (MTP) clone selection. BAC end sequences generated by the FPC map assembly, as well as those produced by post-sequencing clone-ends as a quality control measure, are aligned to contigs from improved sequence. This serves to inform the relative position and orientation of the FPC BAC end sequences.

Repeat Tagging

The knowledge gained through the mathematical *k*-mer analysis has been implemented within a repeat tagger that has been exported to WUGSC. Repeat tags are placed on BAC assemblies at each finishing iteration in order to terminate contig extension at repetitive endpoints.

The Maize Genome Browser:

The maize browser is based on the Ensembl browser (http://ensembl.org). The browser, available at http://maizesequence.org, provides several views of the maize genome, namely, a high-level MapView providing a topographic overview of each chromosome along with associated features, a chromosome-based FPC view, CytoView, presenting updated clones that have been sequenced by the project, and a lower-level BAC view, ContigView, that provides all underlying annotations that have been procured for the clone sequence maps. In the past 9 months the site has upgraded from Version 40 to version 45, providing enhancement to existing views as well as increased website performance. The website has evolved beyond a genome browser and provides various types of data about the sequenced maize genome. Beyond visual genome browsing, users can access the data via sequence search (BLAST), marker name search, FTP access, or data syndication via RSS. The website's index page provides a set of entry points into the browser, and includes documentation about the project, its standards, nomenclature, and overall progress.

A common starting point is MapView, which provides, for each chromosome, a high level view of the chromosome's state. Vertically-drawn density plots provide a rudimentary visualization of all FPC clones, the relative number of sequenced clones, the SSR markers discovered on the FPC map, and the core maize bins, aligned with the corresponding chromosome karyotype view of the FPC bands.

To facilitate entry, "bins" were identified using available markers on the IBM 2 Neighbors map, which includes the anchored overgos. These bins are displayed as part of MapView, CytoView and the main navigation bar on the left hand side of the Website.

The FPC view provides a bird's-eye view of the underlying agarose FPC map that composes the maize genome. It features the FPC assembly, including all the clones. A prominent track displays clones sequenced or in-progress by WUGSC as part of the project. Historical clones from previous sequencing efforts are also displayed for posterity using a specific color code. The FPC view provides a set of marker tracks, distinguished by function and origin. A set of high-confidence color-coded overgo markers are also included. The maize core-bin markers present a cytological context to the physical map. Finally, a track displays syntenic rice regions, linked to Gramene, that were determined based on contact points to overgo probes on the FPC map. The FPC view can be accessed through specific FPC contigs shown on a universal navigation bar, or by clicking on a region on MapView.

The BAC view, reachable either through the entry page or by selecting individual clones from the FPC view, is now available for all clones that have been submitted to GenBank, at every level of sequencing. The BACs provided by the browser reflect the same sequence that is available in GenBank, despite the fact that underlying intra-BAC contigs are unordered. GenBank records for HTGS_IMPROVED clones include higher-level scaffolding information. These contig scaffolds attempt to distinguish contigs that are known to be ordered with respect to each other from those that aren't. Several ordered scaffolds may exist within the same clone, even though the relative position of the scaffolds may not be known. Such scaffolds are prominently distinguished on the browser using an appropriate color scheme.

The BAC view graphically displays all data that has been generated through the annotation pipeline. These include classified gene ab initio predictions, cereal alignments, mathematically defined repeats, as well as curated MIPS repeats. Externally curated sequence-based markers, such as TWINSCAN transcripts, are presented via Distributed Annotation System (DAS) tracks. From the gene models, users can obtain additional annotation on the gene models from the GeneView. In the next year users will be able to obtain orthologue and paralogue information as the BACs move from primary to secondary annotation.

Finally, the browser includes SyntenyView, a graphical visualization of synteny derived between maize and rice. The views are universally accessible through the navigation bar. The views provide linkages to corresponding FPC views on the maize browser as well as to the rice browser on Gramene (http://gramene.org).

Data Export: All genome browser views provide export functionality specific to the region being displayed, in contrast to the genome-wide FTP dumps available on the site. Users can choose to export data as either raw sequences or meta-information for a particular region, such as the genomic coordinates of included markers.

FTP Site

The website now includes an FTP server that provides access to sequenced clones, underlying sequence contigs, as well as gene predictions and their associated translations, broken down by class. Finally, an FTP sequence report that indicates which clones on the FPC map are being sequenced through the genome sequencing project is also included.

The FTP dumps are generated every week to coincide with the latest update of sequences and underlying clone information. As resources allow, the dumps from previous weeks are preserved.

BLAST

The BLAST server initially (March, 2007) provided alignments to all sequenced clones. The server has evolved to provide a variety of other options. Users are now able to search a variety of data sets, including BAC clones, GSS sequences such as BAC ends, as well as the predicted gene models either through nucleotide BLAST (BLASTN) or translated BLAST (TBLASTN).

RSS Notification

With the changing nature of the maize sequence and with the addition of new annotation tools to analyze the underlying sequences, it becomes cumbersome for users to track updates. As part of the plan to address this, the website now provides data syndication feeds. Built on RSS technologies (Really Simple Syndication, http://www.webreference.com/authoring/languages/xml/rss/intro/), the system provides notifications of updates for a given region in a syndicated fashion, meaning that it is accessible without having to visit the website. Users can "subscribe" to feeds through a number of freely available applications. A standalone news aggregator, for example, can poll the website for updated items. Several established browsers such as Firefox and Internet Explorer (version 7) also include RSS support. The website allows users to subscribe to any region specified through a notification form.

For the initial release, web feeds are available for FPC markers specified through chromosome coordinates. The system is being rapidly enhanced to include all types of sequence data on both the FPC map as well as individual clone maps.

DAS

The Ensembl browser provides inherent support for DAS (Distributed Annotation System; http://biodas.org) functionality. Sequence annotations that are stored on an external server can be readily visualized on the browser without having to import and maintain the data into the maize database. The browser can be configured with any number of external data sources that can be displayed as separate tracks.

Users can readily use DAS to visualize their own private annotations on the sequenced BACs. In such cases, the DAS tracks remain unpublished. However, DAS tracks that are intended for public use are configured as part of the browser and are easily accessible.

As a pilot project, TWINSCAN transcript annotations (NSF #0501758; http://maize.danforthcenter.org/abinitio.htm) are provided on the BAC views. Visual placement of such transcripts has been enabled by the preservation of the GenBank record both on the browser end and within the originating transcript data source (at the Danforth Center).

Other projects and collaborations are being considered for providing annotations through a DAS server.

Plans: For the upcoming year, the annotation pipeline will incorporate higher-level annotations into the analysis of individual clones. The existing protein annotation pipeline will be extended to provide Gene Ontology (GO) terms as well as value-added correspondences to established protein databases such as SwissProt using the Xref system. Sequence-based whole genome alignment will be conducted between maize and other mature cereal genome assemblies, such as rice, sorghum, and Arabidopsis. Other sequence-mining tools will be incorporated to discover potential SSR sites, tRNA genes, and other interesting sequence features. Externally curated data sets, such as maize full-length cDNAs and the maize optical map, will be integrated into the genome. Finally, leveraged by the integration of external data, evidence-based gene builds will be conducted to provide mature mRNA transcripts.

4. Bioinformatics ISU. Responsible PI: Pat Schnable

Goals: The Iowa State University group is focusing on refining strategies for BAC assembly. Our long-term goals are to develop computational strategies to use the project's high quality sequence data to generate even better BAC assemblies. In addition, we desire to

develop a community-accessible visualization of assembled BACs that will allow scientists to evaluate the data used to make specific assembly decisions. In this manner community members will be able to decide for themselves the degree of confidence to which to assign particularly contigs and scaffolds. During the current project our primary goal is to determine whether the approaches described have the potential to generate improved assemblies relative to traditional approaches.

Progress: Although traditional assembly approaches typically work well, the maize genome presents some unique assembly challenges including the presence of "NIPs" (Nearly Identical Paralogs) and large numbers of long, high-copy, highly conserved elements. NIPs are groups of low-copy genes that exhibit >98% sequence identity (Emrich et al., Genetics, 2007). Conservatively, 1% of maize genes have a NIP. Members of a given family of NIPs can be arranged in tandem arrays or distributed across the genome. Tandemly arrayed NIPs could potentially collapse during BAC assembly. Collapse means that two (or more) adjacent copies of a sequence are inadvertently assembled into a single copy. We have previously shown that assembled maize sequences (MAGIs) can contain collapsed NIPs.

To test whether this occurs during BAC assembly, we used seven NIP families whose members were shown by genetic analyses to be tightly linked in recombination experiments involving 91 recombinant inbred lines. The sequences of the seven NIP families were used to blast against assembled BACs that had been deposited in Genbank. Hits were obtained for all seven NIP families. Shotgun sequences from each BAC were downloaded from the trace archive of Genbank and aligned to these BACs. The resulting multiple sequence alignments were examined for the presence of paramorphisms (Emrich et al., Bioinformatics, 2004) which provide evidence of collapsed NIPs. In two of the seven cases, evidence of NIP collapse could be detected. Although we have developed a general strategy to use paramorphisms to prevent NIP collapse during BAC assembly, this strategy has not yet been implemented.

Just as NIPs can collapse during assembly, other kinds of repeats can complicate traditional BAC assembly pipelines. In traditional assembly approaches, sequence alignments between reads are identified. Contigs are assembled by starting at a good alignment and then extending the ends of contigs one sequence at a time. Clone pair information is used to scaffold contigs *after* contig construction.

In an attempt to develop improved BAC assembly strategies, we are exploring the idea of integrating clone pair data into the contig assembly process. To do so we are modeling sequence alignments and clone pair relationships as a graph. First, we construct an alignment graph in which sequence reads are nodes. A (black) edge is drawn between a pair of nodes if there is a valid sequence alignment. Next we introduce two additional types of edges into the graph. Clone pair edges are drawn (in red) between paired nodes. "Path edges" (green) are drawn between two nodes if the nodes are close together in the graph AND their clone pairs are also close to each other. Path edges identify sequence alignments that are more likely to be relevant to correct assemblies than are unfiltered sequence alignments.

We can then use various graph transformations to ensure that black edges (sequence alignments) represent *correct* genomic overlaps, and resolve entries into and exits out of repeats. For example:

Use clone pairs to validate alignments in repeat regions if the corresponding clone pairs are anchored to unique regions and exhibit alignment.

Use paramorphisms to break spurious alignments due to NIPs.

Use clone pairs to match entries into and exits out of repeats.

Use clone pairs and validated alignments to guide contigs

Use graph min-cuts to find correct assignment of reads to the complementary strands.

This will allow for annotation of paths (contigs) via walking through the graph. In doing so we make use of three levels of pointers:

Black edges: show what steps are available

Green edges: indicate the best path

Red edges: indicate the final destination

This graph approach was applied to three random (Stage 3) maize BACs that had been deposited in Genbank. Multiple lines evidence established the existence of repeat-induced "knots", collapsed repeats and NIPs, and missed scaffolding opportunities in the three BACs.

Collaborations:

MaizeGDB (http://maizegdb.org) Ames, IA

Deep reciprocal links are present between both websites, allowing maize researchers to easily navigate between maps and features. The WUGSC and CSHL have jointly written a letter of collaboration regarding providing updated information for the 2008 MaizeGDB Project Plan, an internal USDA-ARS document for continued funding.

Gramene (http://gramene.org) Cold Spring Harbor, NY

Gramene has been a valuable resource in many aspects.

Development efforts have heavily leveraged the Gramene project. Source code is heavily shared between both projects. Initially the Ensembl codebase was inherited from Gramene. Subsequent modifications and enhancements have propagated back into the Gramene codebase. Likewise, the maize annotation pipeline has been progressively adapted for use by Gramene and is presently repeatedly run on other cereal genomes. Guidance and development effort by Gramene personnel has also proven invaluable over the course of the project. Exchanges of code and ideas are ongoing and drive further improvement of the website. Deep reciprocal links are maintained between both sites.

EBI Ensembl (http://ensembl.org) Hinxton, UK

The Ensembl team provides much needed software support. Reciprocally, new functionality introduced into the maize browser as well as improved software enhancements, have been incorporating into the main development branch of the Ensembl codebase.

Several workshops and meetings have taken place between both groups to further enhance website and pipeline code and to discuss future plans of both projects.

Affymetrix Maize Pilot Expression Array Project

A recent collaboration has been established with Dr. Roger Wise and Affymetrix to design a maize genome GeneChip for transcript detection. The project team will be furnishing mature mRNA transcripts from evidence-based gene predictions that will facilitate the design of accurate microarray probes. The microarray data that will be generated by the project will be readily available on the website.

Optical map Madison, WI

Proposals are being made to leverage the maize optical map (NSF #0501818) in validating the maize sequence and identifying potential sequence assembly errors. Concrete plans are also being made to visualize the optical map alongside the FPC map and individual sequence maps on the maize browser, in order to provide context-based public access to the optical map.

TWINSCAN (http://maize.danforthcenter.org/abinitio.htm) St. Louis, MO

TWINSCAN transcripts (NSF #0501758; http://maize.danforthcenter.org/abinitio.htm) are available for public access through a Distributed Annotation System (DAS) track. The transcripts are displayed alongside the predicted gene models and can be used for reciprocal validation of both data sets.

Vmatch (http://www.vmatch.de) Hamburg, Germany

A unique collaboration has been established with Dr. Stefan Kurtz, the principal developer of Vmatch (http://www.vmatch.de). Vmatch has been used extensively for mathematical analysis of the maize genome and is being adapted for cross-species analysis with other cereal genomes. The collaboration has allowed for specific software modifications in Vmatch in order to meet certain requirements of maize-related hypotheses.

Full-Length cDNA Project (http://www.maizecdna.org) Tucson, AZ, Stanford, CA

Full-length cDNAs will be provided on the maize browser as sequence-based alignments and used in evidence-based gene predictions.

Public Presentations: Project personnel have presented the project at various biological and maize-related conferences with involvement varying among informative project poster, project overview talks, and website tutorials. These conferences include the Maize Genetics Conference, Plant Genome Conference, Plant and Animal Genome Conference, Rice Functional Genomics Meeting, Gordon Conference on Agricultural Biotechnology, Banbury Meeting on Conifer Genomics, American Society of Plant Biologists Conference, Genome Biology Conference, Genome Informatics Conference, Intelligent Systems in Molecular Biology Conference, and the National Corn Growers Association Conference. Involvement in such conferences varied between informative project posters, project overview talks, and website tutorials. The WUGSC group also presented a project overview to the annual meeting of the National Corn Growers Research and Business Development Action Team in December 2007.

Sequencing the Codifying Genome of the *Palomero Toluqueño* Mexican Landrace

--**The Mexican Maize Genome Team (CENTLI)**. National Laboratory of Genomics for Biodiversity (Langebio); Cinvestav Campus Guanajuato, Km 9.6 Libramiento Norte Carretera Irapuato-Leon, Irapuato Guanjauato MEXICO.

In contrast to developed countries where it is fundamentally used for agro-industrial or animal production, maize in Mexico is mainly cultivated for human nutrition and under a wide range of climatic conditions. Its consumption represents the main source of protein and energy in rural regions, particularly in the poorest communities. Maize was domesticated from its wild progenitor teosinte (derived from "*teocintli*" in nahuatl language: "*teotl*"=sacred and "*cintli*"= dried ear of corn), a common name given to a group of annual and perennial species of the genus *Zea* native to Mexico and Central America (reviewed in Doebley, Ann. Rev. Genet. 38:37-59, 2004; Matsuoka, Breed. Sci. 55:383-390, 2005). As the center of origin and domestication, Mexico has the largest diversity of maize genetic resources. Maize biologists do not always agree on the total number of maize landraces that exist in Mexico. The classical monograph published by Wellhausen et al. (Folleto Técnico Número 55, 1951) has been an essential reference for all subsequent reports. Based on general architecture, kernel cytological traits, and physiological characteristics (time of flowering, yield, and disease resistance), they classified 25 landraces into 5 major groups. The first one included 4 Ancient Indigenous Races believed to have arisen from primitive pod corn (*Palomero Toluqueño*, *Arrocillo Amarillo*, *Chapalote*, and *Nal Tel*).

Large-scale sequencing efforts concentrated in B73 will not be sufficient to fully understand maize genome organization and identify all functional units available in the domesticated gene pool. To complement the large-scale B73 sequencing initiative and explore landrace genomic diversity, we undertook the structural and functional characterization of the *Palomero Toluqueño* genome after estimating its small genome size (20 to 25% smaller than B73; Mexican Maize Genome Team, unpublished results). *Palomero Toluqueño* is an ancient popcorn landrace originally classified by Wellhausen et al. (1951), and a member of Central and Northern Highlands Group composed of 15 landraces that most often produce short individuals (140-190 cm) and grow at elevations higher than 2000 meters. *Palomero toluqueño* accessions usually show conically shaped ears and high kernel row number, a low frequency of tassel branches, a weakly developed root system, and strongly pubescent leaf sheaths often pigmented by anthocyanins. A total of 1.2 million Sanger reads (10% HCot; 90% enzyme-based methyl-filtration) and 213 pyrosequencing runs (50% methyl-filtered, 50% whole genome sequencing) were sequenced at the National Laboratory of Genomics for Biodiversity (Langebio), in Guanajuato. The total sequence generated represents coverage of more than 3X the full genome; it has been complemented by in-depth pyrosequence-based global transcriptional analysis of the same genotype. As expected, a significant percentage of codifying transcripts are not reported in publicly available databases, suggesting that a large portion of the molecular and functional diversity contained in Mexican landraces remains unexplored; the structural annotation of resulting contigs will be completed by the end of 2007.

Figure 1. *Palomero Toluqueño* landrace. State of México (Valley of Toluca Valley). Identification: Wellhausen et al., 1951. Description: Wellhausen et al., 1951.