Activities and Findings

The Gramene project is a highly successful project that has delivered software, ontologies and a project website that integrates genomic, genetic, and comparative data sets. The process of integration provides added value to the original data, ease of access for users and creates an environment for discovery. We are reporting on the progress from November 2007- June 2008.

Gramene Website

In the first year of this project the group has focused on data acquisition and analysis for the most recent release, improving existing infrastructure, evaluating new resources, and enhancing project management. We have provided updates of progress in context to the specific aims in the updated experimental plan.

Results for Specific Aim #1: Provide an infrastructure of comparative genomic data to allow for the mining and analysis of functional data on the genomes of rice and other monocots.

The main resources covered under specific Aim 1 include the Ensembl resources, markers module and comparative map viewer. In the past 6 months 4 new sequenced genomes databases were built and displayed on genome browser, including grape from IGGP (supported by funds from USDA), sorghum and poplar from JGI and Oryza glaberrima (short arm) from OMAP. Their assemblies and community gene annotations when available were loaded. As part of the existing resources Gramene aligns the ESTs to each of these genomes. The EST alignments provide evidence to support genes and also serve as a source of entry reference points for non-sequence genomes. In this past release 3M plant EST/mRNA/oligos were incorporated into the markersdb and aligned against each of the 8 sequenced genomes. This generated about 80M alignments, which were loaded into markers database. A curation decision was made to load only a subset into Ensembl database to be visualized based on performance and usability. In the past 6 months Gramene evidence-based gene builds were generated and loaded into the browser and Markers dataset, RiceJaponica_GeneModel_GeneBuilder (37176), RiceIndica_GeneModel_BGI (38861), Sorghum_GeneModel_JGI (34496) Sorghum_GeneModel_GeneBuilder (45339), Grape_GeneModel_IGGP(30434), Poplar_GeneModel_JGI (9879), OryzaGlaberrima_GeneModel_GeneBuilder (2467). Using the Ensembl compara protein pipeline new gene trees were built using all of the available plant genomes in Gramene and Fgenesh predictions from the maize genome based on data freeze in Dec 2007. In addition to the plant genomes we were able to make use of the c.elegans, yeast, drosophila and human genomes available in the Ensembl framework.

A major reorganization of the markers module was initiated and completed which included improvements to design, import scripts, tagging, content and integrity. The update included a fresh dump from GenBank with improved classification of ESTs, GSSs, CoreDNA clones. A new dump of EST clusters from NCBI UniGene, PlantGDB PUT, Gene Indices. Six new sequence mapsets were loaded into markers db, these included Arabidopsis, current Maize contigs, Sorghum, Poplar, Grape, Oryza glaberrima. Additional data sets included 3385 new gene records were brought in from GrainGenes and a new cytogenetic FISH map of the maize Chromosome 7 was added to the maps and markers.

In addition to content there have been significant improvements in visualization and support infrastructure for both Ensembl, Markers and CMap. The Ensembl software was updated from V45 to V48. A DAS server was implemented for the Oryza Sativa genome browser and registered as a DAS resource. A GFF dumper was developed. In the next six months this will be added as a download option from the website. The Marker displays were improved. A significant development investment was made into improving the scripts and software that are responsible for generating alignments and loading these into the Marker or Ensembl database. The new scripts allows the dumping out of sequences from markers database, configuring and aligning the data sets against a specified genome with Blat pipeline, loading the mappings back to markers database and finally allowing select subsets of mapping to be loaded into Ensembl database for display on genome browser contigview. Development is currently ongoing to revise the existing Blat Pipeline to use CIGAR representation for alignments. This will lead to better storage and retrieval of the alignments and improve the current views available. The group is also evaluating and testing methods to allow for structured tags to identify subsets of data that are parts of larger sets. An example is genetic markers that are a subset of GSS reads from GenBank. The CMap software was update to the new release version 1.0.

In the previous grant the group had evaluated and tested the Ensembl Compara pipeline to produce Whole Genome Alignments (WGA). The group is currently evaluating the synteny component. The synteny module uses a distance based metric to determine synteny allowing only one parameter for both genomes. We have collaborated with the Ensembl group and now have the ability to support different parameters for each of the genomes. In the next 6 months we will be evaluating different parameters on the monocot and dicot genomes available.

Most of the analysis is run on the Ware lab cluster at CSHL. In the past 6 months we have combined the Ware lab cluster with Blue Helix (purchased in the last year by CSHL) and transitioned from PBS as a managements system to SGE. We have migrated all of the Ensembl pipelines to Blue Helix and made the necessary changes to transition to using SGE. We also updated two servers and purchased as a new disk array to support the project.

Results of Specific Aim #2: Enhance the value of the comparative maps with pathway, phenotypic and other functional information from rice, maize, and Arabidopsis.

The pathway module received updates, new pathways and improvements in usability. RiceCyc was rebuilt using TIGR 5 gene models (the previous version used TIGR 4). The protein annotation pipeline, Ensembl xref and pathway tools predictions was run on the available Sorghum (v1.4) gene models, resulting in a beta release of SorghumCyc (Support from DOE grant Ware Co-PI). In this release Gramene added 5 new pathway mirrors MedicCyc, LycoCyc, CapCyc, PotatoCyc, CoffeaCyc, and continues to maintain AraCyc. Maintaining the mirrors allows users the ability to compare between the pathway predictions in different organisms. Several enhancements were made to improve usability. The first was the additions of 5 new mirrors to allow for cross-species analysis between the different genomes. The second was the addition of species-specific entry pages. The last was the addition of links between the pathways and the gene modules. One of the objectives is to make use of the Reactome model and corresponding curators tools. In the last 6 months time was devoted to review of the Reactome module and curation tools.

The Gene, Protein, Ontology and literature DBs were updated. 3,385 genes were brought in from GrainGenes. siRNA and miRNA rice genes were curated and added to the database. The ontology associations were update for the rice genes with 97% of the genes 2596 having ontology associations. There were several updates to the ontology modules. The Gazetter (GAZ) (geo-location) ontology was added. It should be noted that Gramene is the first MOD to make use of GAZ for annotating the source geo-reference site associated with the germplasm, using rice as a model. 55K Taxonomy terms were added to the Gramene's taxonomy ontology (GR_tax ontology). There were substantial updates to the Trait Ontology which now contains ~ 900 terms. Rice associations to GO were updated and submitted to the GO consortium.

Results of Specific Aim #3: We will acquire genotypic and phenotypic diversity data for each of the sequenced monocot genomes. We will recalculate this data using a standardized methodology that allows us to integrate the QTL values across species and to relate phenotypic diversity to candidate genes via pathway information.
The Gramene Genetic Diversity module has had significant increase in the number of markers and phenotypic traits. For rice, the number of markers has been increased from 1023 to 4,667,279, together with the introduction of new marker types including SNP, amplicon sequencing, and AFLP. The SNP markers detected from large sequencing projects have been imported and populated in the diversity database, and a pipeline for transferring data has been implemented. Three datasets from the rice evolution project have been annotated, and the identified SNPs in the amplicon sequencing regions can be dynamically displayed with the sequence alignment viewer. We have also started to collect and curate phenotypic and genotypic raw data, and that data will be further used for QTL re-analysis. For maize, the number of markers has increased from 1955 to ~13073, and the new introduced marker types include amplicon sequencing, Isozymes, INDELs, and CAPs. In addition, we have added one new species, sorghum, to the Genetic Diversity module, and most of the sorghum data was contributed by the Institute for Genomic Diversity (IGD), Cornell University. The QTL database has increased from 11536 to 11624 curated QTLs.

Results of Specific Aim #4 Support the Plant Ontology (PO) *Two Years Only*
The Plant Ontology (PO) Consortium released 0000408 of the PO database in April of 2008. For detailed release notes, please visit http://www.plantontology.org/docs/release_notes/index.html. The project currently host 1128 PO terms and 58117 annotations. In the last 6 months 19 new terms were accepted 18572 new associations were added, represented ~ 20% of the current associations. Several improvements were made to the website, ~16000 new annotations on genes and mutant phenotype germplasm from tomato, tobacco, potato, eggplant, pepper and

Hyoscyamus. These annotations were contributed by Solanaceae Genomics Network (SGN; http://sgn.cornell.edu/). In addition there were updates and new annotations by TAIR (for Arabidopsis) and Gramene (for rice) databases.

To secure additional funds to maintain the project a new grant proposal was submitted to the NSF's PGRP panel in January 2008 (#DBI-0822201) titled 'TRPGR: The Plant Ontology'.

Specific Aim #5: Education, Outreach and Diversity
The experimental plan consists of 3 specific components 1) a community-supported WIKI targeted to pure and translational plant sciences researchers; 2) traditional "push" presentations at meetings, staffed by the coPIs and senior curators; 3) virtual and physical tutorials sponsored by a commercial partner, OpenHelix, targeted at students in under-served institutions and underrepresented minorities.

1. Community-supported WIKI. In the last 6 month preliminary discovery and infrastructure has been explored. Development work has been done to describe what data should be pre-populated in the wiki and what areas of the wiki will be accessible to change. We have also been evaluated requirements that would allow the wiki based curation to be displayed as a DAS track on the Ensembl gene pages. The project has been in communication with Jim Hu at TAMU who is responsible for the Ecoli Base and is currently making use of the wiki. We hope to have beta version of the wiki available in a year.

2. Traditional "push" presentations at meetings

The members of the group presented at 11 forums in the form of invited talks, workshops, or posters. A detail list of these outreach activities are given in the outreach section.

3. Virtual and physical tutorials sponsored by a commercial partner, OpenHelix . As part of the outreach objectives the project will be contracting with Open Helix to host Gramene tutorials. Before leaving the project, the former outreach coordinator in Claire Hebbard began dialogue and exchange of the current Gramene tutorials. Since that time there have been several concerns and dialogue has been exchanged between the NSF Program manager and Lincoln Stein. A decision has been made to mover forward with the current plan. At this time we are awaiting access to Open Helix current Gramene tutorials for review. There is a concern among the project members that the current tutorials are outdated and do not reflect the current Gramene model and will need to be refreshed before Open Helix releases the tutorials, as well as the development of a Diversity module. Gramene has requested more detailed information from Open Helix with regard to the numbers of times they will update the different chapter/modules within the tutorial. Based on this response it may be necessary to have the group update tutorials to be given to Open Helix. In addition the Project has requested that Open Helix provide the necessary registration information to track the diversity data needed as well as a brief statement (100 words) on how access to the tutorials will benefit the recipient of the award and allow the project access to the tutorials at any time.

Updates to Project Management

Restructuring in resources for the new experimental plan eliminated the outreach coordinator position. To address this change in plan, we have automated the process of tracking requests from the website. The feedback link from the website is now directly added to the project management tool, mantis and a weekly rotation has been set up where by one person on the project is responsible for providing an immediate feedback to the user and assign the task to the appropriate domain expert on the project. The feedback items are reviewed in the weekly meeting and responsibility is handed over to the next person in the rotation.

In April 2008 the project held the annual internal retreat. The meeting was used to highlight work over the past year, discuss new data acquisition and development targets for the next year and evaluate the existing project management tools. With the loss of the outreach coordinator a decision was may to have weekly rotating members to deal with feedback. With the loss of a formal position for project manager the group has adopted the use of mantis for managing tasks within the project and Ken Clark has take on the role of managing the Weekly meeting. With the new grant a new board was selected that would retain a subset of the past members for continuity (50%) and new members based on domain expertise. At this time all but one member Georgia Davis has accepted the invitation and the first SAB meeting is scheduled for either Sept 25 or 26 to be finalized in July.

Gramene SAB

Michael Asburner	GO Consortium
Bill Beavis	Iowa State University
Georgia Davis* #	University of Missouri, Columbia
Paul Flicek	EBI/Ensembl
Patricia Klein*	TAMU
Ann McClung*	USDA
Dave Marshal*	SRI
Tim Nelson	Yale

In the past 6 months there has been a major change in the leadership of the project. Lincoln Stein has accepted a position with the Ontario Institute for Cancer Research and will be moving from CSHL over the course of 2008. With the change in position Lincoln has coordinated with the NSF Program Managers and the current project Co- PIs to relinquish his position as PI on the project to Doreen Ware. This transition was completed in April of 2008. Lincoln has graciously agreed to continue in an advisory position and continues to attend PI project meeting.

Publications.

Avraham, S., Tung, C.W., Ilic, K., Jaiswal, P., Kellogg, E.A., McCouch, S., Pujar, A., Reiser, L., Rhee, S.Y., Sachs, M.M., Schaeffer, M., Stein, L., Stevens, P., Vincent, L., Zapata, F. and Ware, D. (2008) The Plant Ontology Database: A Community Resource for Plant Structure and Developmental Stages, Controlled Vocabulary and Annotations. Nucleic Acid Research, Vol. 36, Database issue D449-D454.

Liang, C., Jaiswal, P. Hebbard, C., Avraham, S., Buckler, E.S., Casstevens, T., Hurwitz, B., McCouch, W., Ni, J., Pujar, A., Ravenscroft, D. Ren, L., Spooner, W., Stein, L., Tecle, I., Thomason, J., Tung, C.W., Ware, D., Wei, X., Yap, I., Youens-Clark, K. (2008) Gramene: a growing plant comparative genomics resource, Nucleic Acid Research (Database issue D947-953).

Websites
www.gramene.org
www.plantontology.org

At the Genome Informatics meeting Oct/Nov 2008 Chengzhi Liang presented a poster on Gramene and Shuly Avraham presented a poster on POC. At the Ensembl users group satellite meeting, before GI, Chengzhi Liang participated and presented an oral presentation on the modification to the Ensembl gene build and gene prediction in plants using the pipeline.

Outreach
At the PAG Jan. 2008 Gramene along with several other plant databases obtained a booth and provide one on one demonstration with Conference Participants. In addition Noel Yap presented a computer demo on Gramene, Shuly Avraham participated in the computer demo on the Virtual Plant Information Network; Semantic Web and also presented a poster on the same topic, Chih-Wei Tung presented a poster on the Plant Ontology Consortium, Terry Casstevens presented a poster on the diversity module. Pankaj Jaiswal organized the Ontology Workshop. It should be noted that several of the posters at the meeting cited Gramene.

At a symposium on Rice Breeding and Genetics held on Monday, Feb. 18, 2008 as part of the bi-annual meeting of the Rice Technical Workers Group (RTWG) held in San Diego from Feb. 18-21, 2008 (http://www.plantsciences.ucdavis.edu/rtwg/), [^] McCouch made an oral presentation on SNP platforms and informatics pipelines for managing rice diversity data and discussed the implementation and utility of Gramene's Diversity Module for the rice community.

At the 50th Annual Maize Genetics conference, Feb. 2008, Chengzhi Liang presented and oral presentation on the evidence-based gene pipeline. Junjian Ni presented a poster on the diversity module.

At the March 2008 meeting of the Illinois Corn breeder’s meeting Doreen Ware presented the updates to Gramene and the release of the Maize draft sequence.

At the May 2008 Biology of Genomes (CSHL) Chengzhi Liang presented a poster on the evidence-based gene build pipeline, Shuly Avraham presented a poster on the Plant Ontology Consortium

At the Plant Biology June/July 2008 JunJian Ni presented at database workshop a presentation on updates to Gramene, Chengzhi Liang presented at the pathway workshop, highlighting updates to Gramene's pathway tools. Dr. JunJian Ni presented a poster on the diversity module. Although not part of Gramene, it should be noted the Dolan DNA learning center has obtained a booth and one of the modules will be highlighted is Dynamic Gene which was a combined product of the previous Gramene Grant and Dr. Ware's NSF YIA.

In May of 2008 Doreen Ware gave a talk as part Dolan DNA Learning Center "Great Moments in Science" talks. The talks are available to the general public and are targeted to High School teachers and students. She presented her work on comparative genome analysis in plants highlighting work from Gramene as well as other projects.

In May of 2008 Pankaj Jaiswal presented the trait/plant ontology at USDA/GRIN (Ft. Collins germplasm center).

From May 15-June 7, 2008 McCouch ran a 3 week field course on rice research and production at IRRI in the Philippines. Included in this course was a half day session on informatics that included a presentation about Gramene's Diversity module, along with other information resources of interest to the rice community. Information about this course, the participants and the contents can be found at www.ricediversity.org under "Training" and "International Field Course".

At the Plant Biology June/July 2008 JunJian Ni presented at database workshop a presentation on updates to Gramene, Chengzhi Liang presented at the pathway workshop, highlighting updates to Gramene's pathway tools. Dr. JunJian Ni presented a poster on the diversity module. Although not part of Gramene, it should be noted the Dolan DNA learning center has obtained a booth and one of the modules that will be highlighted is Dynamic Gene which was a combined product of the previous Gramene Grant and Dr. Ware's NSF YIA.

Doreen Ware participated as advisory board member for USDA MaizeGDB and the NSF Grassius project. Both projects have active collaborations with Gramene.