Gramene Annual report June 2009- May 2010

Table of Content

Activities and Findings

The Gramene website is a portal for comparative genomics in plants. The Gramene team members have created and integrated novel and existing software, processes, and ontologies with genomic, genetic, and comparative data sets to add value to the original data and create an environment for discovery. In this report, we review Gramene's accomplishments during the period from June of 2009 through May of 2010.

Specific Aim 1 (Genomes)

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.

Central to Gramene's support of Aim 1 are our Ensembl genome browsers, our pipelines for annotating genomes, our markers/sequences/mappings database, our comparative mapping application (CMap), and our Distributed Annotation Server (DAS).

Genome browser

Our Genome browser leverages the Ensembl infrastructure for storing, analyzing and visualizing genome sequence and respecitive annotaitions. In the last year, Gramene has gone from eight hosted genomes to fifteen. Major updates or additions include the following:

- New release of Brachypodium distachyon genome version 1
- Update of Oryza sativa japonica genome from TIGR version 5 to MSU version 6
- Update of Arabidopsis thaliana genome to TAIR version 9
- Update of Vitis vinifera genome to IGGP 12x assembly and annotation
- Update of Poplar trichocarpa genome to JGI version 2.0
- Update of Brachypodium distachyon genome annotation to version 1.2
- New release of Oryza barthii chromosome 3 short arm
- Update of Oryza glaberrima chromosome 3 short arm to BAC Pool 2009 assembly from AGI
- New release of Oryza minuta chromosome 3 short arm (BAC Pool 2009)
- New release of Oryza officinallis chromosome 3 short arm (BAC Pool 2009)
- New release of Oryza punctata chromosome 3 short arm (BAC Pool 2009)

- New release of Oryza nivara chromosome 3 short arm (ultrahigh-throughput sequencing of shotgun BAC Pools from AGI)
- New release of Oryza rufipogon chromosome 3 short arm (shotgun BAC pools)
- New release of Oryza brachyantha chromosome 3 short arm (shotgun BAC pools)

One very significant change in the last year is that Gramene now coordinates with Ensembl Genomes (EG) in the building and verification of our Ensembl databases. This is part of a international collaboration between Dr. Ware's group at CSHL and EBI to support EnsemblPlants. The collaboration facillitates the support of the infrastrucutre for international cooperation for the support of plant genomes in a uniform framework. Gramene submits our "core" database to EBI for quality checks and synchronization with the EG project. Because of

this arrangement, Gramene also is the beneficiary of their work to build our Mart databases. Specific examples of this collaboration are indicated elsewhere in this report. The collaboration reduces redudancy of efforts, standardizes analysis and visualization for the community.

Variation in the context to Genome sequence (Ensembl Variation)

The Ensembl Variation database for a species catalogues the known sequence variants (alleles) in the genome annotated against individuals/populations. More recently the database has been extended to model resequencing experiments and allele-to-phenotype associations.

In previous years, we have imported all available plant variation data from NCBI dbSNP which, at that time, included data for just Oryza sativa. In this year, however, we have loaded the following datasets, extending the rice variation database, and creating new databases for grape and Arabidopsis;

- (160,000 loci from 21 O. sativa lines determined as part of the OryzaSNP project using SNP array technology,)
- 637,522 loci from 21 A. thaliana lines determined as part of the Arabidopsis 2010 project using genome tiling array technology.
- 220,000 loci from 363 A. thaliana lines determined as part of the Arabidopsis 2010 project using SNP array technology,
- 2,698,797 loci from 17 A. thaliana lines determined as part of the Arabidopsis 1001 genomes project using resequencing technology,
- 469,470 loci from 17 V. vinifera lines determined as part of the USDA... project using resequencing technology.

Loading and integrating of the datasets has been achieved in collaboration with the Ensembl Genomes project and NSF DBI 0723510 an arabidopsis variation database. We are also striving to provide consistent datasets between the Ensembl Variation and Gramene Diversity modules. With the increase in public resequencing projects the number and size of the datasets is set to increase markedly over the coming year. For instance, the number of lines assayed by both the Arabidopsis 2010 and 1001 genomes projects will exceed 1000.

To facilitate the long-term maintenance of the datasets, we strongly encourage collaborators to submit data to the central dbSNP resource. However, at this stage all of the loci must be re-mapped by Gramene or collaborators when the underlying genome assembly changes, which has occurred for each of these species over the past 12 months.

We have worked closely with the Ensembl Genomes and Ensembl projects over the past year to ensure that the software, database and visualization frameworks are suited to the needs of the plant molecular biology community. In particular, we have advocated the improvement of Ensembl phenotype association views, and the latest main Ensembl release has progressed in this area. We will be making use of these views shortly.

Annotations and analyses

Comparative Genome Analyses (Ensembl Compara)

Gene Tree Prediction

We continued to use the standard Ensembl GeneTree method (Vilella et al. 2009. Genome Research 19:327) to generate gene trees and predict ortholog and paralog relationships between species. The GeneTree database was rebuilt using five monocot genomes (O. sativa Japonica, O. sativa Indica, O. glaberrima, B. distachyon and S. bicolor), four dicot genomes (A. lyrata, A. thaliana, P. trichocarpa and V. vinifera) and five model metazoan genomes (C. elegans, C. intestinalis, D. melanogaster, H.sapiens, S. cerevisiae).

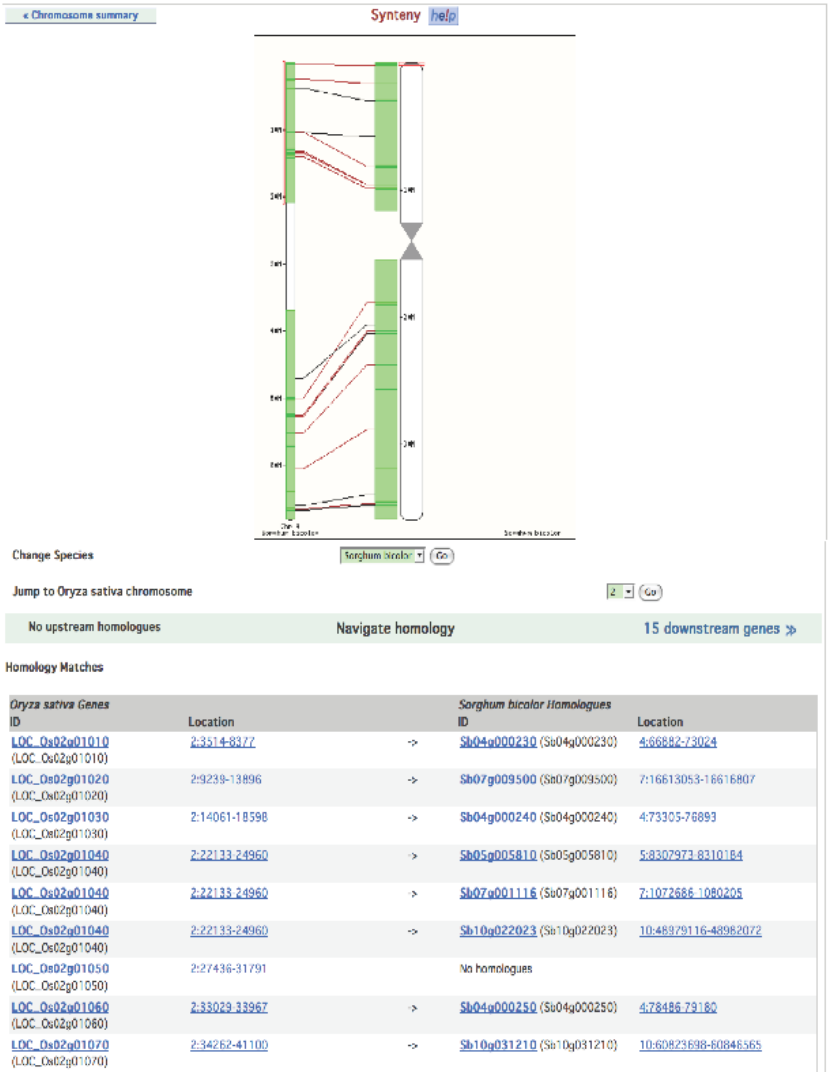

Figure 1: Rice v.s. Sorghum release30 synteny view
http://www.gramene.org/Oryza_sativa/Location/Synteny?r=2&otherspecies=Sorghum_bicolor

Synteny analysis

This year Gramene implemented a new synteny analysis pipeline that provides highly sensitive and specific mappings of ancestrally derived regions between genomes. In contrast to the previous synteny-build method, which relied upon DNA-level whole genome alignments (WGA), the new method makes use of gene ortholog assignments from Compara GeneTree output (Vilella et al. 2009. Genome Research 19:327). This switch avoids complications associated with using WGA, including spurious alignment and differential expansion/contraction within and between genomes. The method was originally developed for the Maize Project (Schnable et al. 2009. Science 326:1112), and is now reduced to practice as "runnables" within the Ensembl API framework. In the first step, strictly collinear orthologs are mapped using DAGchainer (Haas et al. 2004. Bioinformatics 20:3643), defining "syntenic:collinear" gene-pairs. In the second step collinear mappings are used as anchor points to identify additional syntenic orthologs that may violate collinearity due to local rearrangements or assembly artifacts. This step is configured using a gene-index distance parameter and its output defines "syntenic:in-range" gene-pairs. These relationships are stored as gene attributes, while ranges of syntenic blocks are displayed with the Ensembl SyntenyView module. Maps between *O. sativa Japonica*, *B. distachyon* and *S. bicolor* genomes will be available in Gramene Release #31. Maps between eudicot genomes are planned for future releases.

We generated new synteny views for the following 5 pairs Rice MSU6 v.s. Maize Freeze 4a, Rice MSU6 v.s Sorghum, Maize Freeze 4a v.s. Sorghum, Rice MSU6 v.s Brachypodium; Sorghum v.s. Brachypodium

Whole Genome Pairwise Alignments

New blastz-net pairwise whole genome alignments between 14 pairs of genomes. (http://dev.gramene.org/info/docs/compara/analyses.html#blastz). Ensembl release 56 saw the reintroduction of multi-species comparative genome views driven by pairwise alignments that had been absent from the Ensembl codebase for over 12 months. This update was incorporated into an interim Gramene release, version 30a.

For example: on Gramene30, Figure 2, shows the Multi-species view of the blastz-net alignments between Sorghum/Brachypodium/Oryza sativa Indica against Oryza sativa Japonica respectively. The alignments showed similar pattern between sorghum v.s. Oryza sativa Japonica and Brachypodium v.s. Oryza sativa Japonica, but a different and closer relationship between Oryza sativa Indica v.s. Oryza sativa Japonica.

In addition to the current analysis available there has been development work on the EPO pipeline adaption and testing. EPO pipeline is a multiple whole genome alignment analysis tool. It can generate large-scale alignments for collinear segments across multiple genomes. In addition, it can calculate conservation scores and detect constrained elements. It handles segmental duplications and can infer ancestral genome.
The ensembl-56 version of the EPO pipeline software was installed on CSHL's HPCC. It was modified to work with the SGE cluster job management system on HPCC. We initially tested it with OMAP genomes (Oryza sativa Japonica and each of the O.glaberrina, O.minuta, O.officinallis, O.punctata). The test run was unsuccessful; no constrained elements were generated at the 1st Enredo step. To verify the pipeline, we

tested and debugged it with ensembl's three birds genomes, and found our results were similar to Ensembl suggesting the current installation of the EPO pipeline is working fine. We intend to test the debugged installation and have preliminary results from the pipeline for our fall 2010 release. Once put into production, we are hoping to identify a candidate pool of constrained elements in both coding and noncoding regions across multiple genomes, which potentially could help us understand plant evolution and discover more nonCoding functional elements.

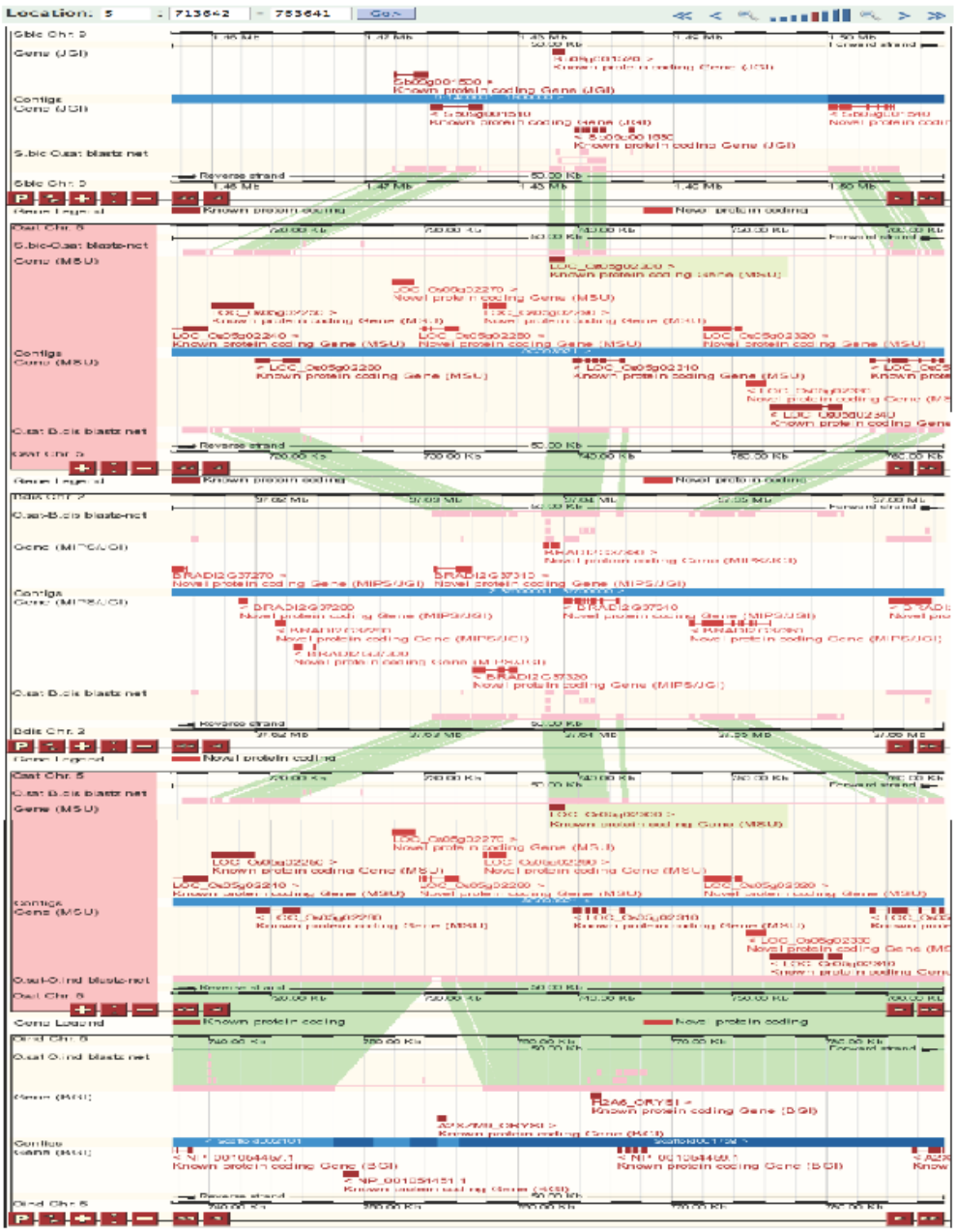

Figure 2: Multi-species view of the blastz-net alignments between Sorghum/Brachypodium/Oryza sativa Indica against Oryza sativa Japonica respectively

Other Resources

Markers database

Gramene maintains a custom MySQL database to house almost 49 million plant markers and sequences from GenBank and various mapping studies. These are use in our annotation pipelines for our completed genomes. This database also holds the results of the alignments as well as manually curated maps from the community and literature. We then build CMap from this resource as well as our DAS

CMap Module: Comparative genetic and physical maps

Gramene's comparative maps database now holds almost 8M features on 214 map sets from genetic, physical, bin, sequence, cytogenetic and QTL studies. In the past year we have curated the following maps from literature:

- Maize IBM2 2008 Neighbors (MaizeGDB)
- Wheat IWGSC Physical 3B (Paux et al.)
- Sorghum genetic (Mace et al.)
- Barley genetic (Close et al.)
- Ae. tauschii (Dvorak et al.)

In previous releases we have included the Oryza sativa japonica genome as our only sequence map but have since added Brachypodium distachyon and Sorghum bicolor, allowing for a far greater number of comparisons.

Germplasm

In an effort to better server the breeding community, Gramene created a new database in the last year focusing on our germplasm resources. This database shows all the links from any particular germplasm to Gramene's markers, map sets, genes and proteins. For our initial release, we included 1,800 germplasm from rice. In the future we will be updating the germplasm module for Arabidopsis and maize as this directly relates to the 3 species Gramene will be targeting for more detailed analysis.

BLAST server

Gramene continues to provide BLAST services for sequence searching against the complete genomic, predicted cDNA and predicted peptide sequences of all genomes that we maintain.

Mart : Gramene BioMart

As part of Gramene's collaboration with the Ensembl Genomes (EG) project, all of our Ensembl Mart databases and interfaces are built at EBI. In addition to those we receive from EG, we also build our own Marts for our QTL and markers/sequences database. EG will soon provide new Marts for variation data as they make these new datasets available.

Distributed Annotation Server (DAS)

In 2008, Gramene began offering DAS interfaces to our mappings databases, but past implementations proved too slow to be useful due to the extremely large number of mappings we hold. In this last year, we have moved from direct queries of our relational database to a denormalized structure queried by the FastBit query engine. This has resulted in a significant performance improvement such that Gramene has moved internally to serving many of our own mappings in our Ensembl genome browser via DAS rather than storing them directly in the Ensembl "core" databases. Finally, we now offer DAS tracks for sixteen species, up from the one species (Os. japonica) we offered initially.

Aim 2 (Pathways)

Enhance the value of the comparative maps with pathway, phenotypic and other functional information from rice, maize, and Arabidopsis.

In the last year, Gramene's pathways has had several significant improvements. We made our first official release (v 1.0) of the SorghumCyc. Version 3.0 of RiceCyc was released with MSU O. sativa japonica cv. Nipponbare genome annotations. The PlantCyc reference pathway database created by the Plant Metabolic Network Group (http://www.plantcyc.org) is now mirrored in our pathway database. Gramene's custom OMICs validator tool now allows our users to validate expression data files by mapping the probe IDs to gene IDs.

Our curators added approximately 170 enzymatic and 80 transport reactions. We also revised approximately 65 tRNA and 600 transport reactions associated genes, and updated the following pathways in rice: chorismate biosynthesis, tryptophan biosynthesis, phenylpropanoid biosynthesis, flavonoid biosynthesis, flavonol biosynthesis, leucine biosynthesis, auxin biosynthesis and hydroxycinnamic acid serotonin amides biosynthesis. We added or curated more than 100 literature citations and deleted 30 non-plant pathways in sorghum including enterobactin biosynthesis and glycogen biosynthesis, etc. The Pathway Tools user interface was upgraded to version 13.5 and customized for Gramene Users. We also made a first step in creating web-based tutorials for pathways. Lastly, we established collaboration with Reactome (http://www.reactome.org) based upon pilot test runs using the BioPax format to exchange curated pathways from RiceCyc to Reactome.

In Gramene release 30, we updated pathway-tools to version 13.0. The pathway-tools v 13.0 has major updates than previous version. The web interface has a quick search box, a new toolbar that appears at the top of every Web page and a search menu in the toolbar to subsume all the functionality of the previous query page. More important is that the pathway-tools now allow developer to customize the web contents and appearance. We customized the pathway pages to look similar to other Gramene modules with the same color and drop down menu. In the drop down menu, we added the menu item of Omics validatortor link to the tool developed by Gramene. We also changed the feedback link to feedback page of Gramene.

Aim 3 (Diversity)

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 diversity module has had the tremendous opportunity and challenge of shifting to the world of next generation sequencing and coming near complete knowledge of multiple genomes. In the case of maize, we are preparing for datasets of 20 million SNPs known on thousands of lines, and we are seeing similar large datasets from the other species.

We have completed redesign of our schema and data packing system, so that millions of SNPs can be stored in a DB easily and then shared with the user quickly. This included developing the code infrastructure to automate curation and loading diversity data into MySQL databases, and improving the diversity web front-end so that large diversity SNP datasets can be filtered by plant name and genome coordinates. Lastly, there are SNP DAS Servers available for Maize, Arabidopsis and Rice, so that other databases can use the data.

Given these data sizes, we are relying on Ensembl for display of variation in the context of genes, and then the interfacing with open source software for users interested in analysis of whole genomes or chromosomes. We have made explicit linkages to the leading diversity analysis tools – TASSEL (plant association studies), PLINK (human association studies), Flapjack (diversity display), and formats like HapMap. We have added numerous features to TASSEL to deal with the data size issues including: new alignment viewer, progress monitoring, and pipelines. To help with the user experience we have begun adding wizards and automatic data loading/analysis at start up from the website. Through collaborations with colleagues at the Scottish Crop Research Institute, we are beginning to get very good data visualization through their Flapjack viewer. We have also developed a novel web-based query tool called "SNP Query" for coordinate-based searches of our SNP studies by chromosome and optionally a base range. SNP Query is capable of returning hundreds of thousands of data points very rapidly because it is performing live searching on binary objects in the database. Allele calls and associated genes can be returned in multiple formats including text or HTML to the browser or as a tab-delimited text file to download.

In the last year, Gramene's genetic diversity database brought in the following large-scale SNP-chip based genotype datasets:

- Rice: OryzaSNP large scale SNP variation study (McNally, K et. al PNAS 2009). ~160K SNPs x 20 diversity rice accessions, mapped from IRGSP4 to MSU6.
- Arabidopsis:
    - 2010 Project SNP discovery data from Clark et. al., 637,522 SNPs, 21 ecotypes (incl. Col-0 reference genome), mapped from TAIR8 to TAIR9.
    - 2010 Project genotype data v3.04, ~214,000 SNPs x 1179 Arabidopsis ecotypes, mapped from TAIR8 to TAIR9. Construction of 250K chip used in this study discussed in Kim, S et. al. Nature Genetics 2007
    - 1001 Genomes WTCHG/Mott data from dbSNP, 2,698,797 SNPs, 17 ecotypes.
- Maize: Panzea SNP data, 1.6 million SNPs x 27 NAM founder lines
- Grape: 470k SNPs identified by re-sequencing a collection of 17 grape cultivars and wild Vitis species from the USDA germplasm collection (Myles, Chia et. al. PLoS ONE 2010)

Finally, to demonstrate the power of the database environment and enhance methods for looking at diversity across species, we are developing computational and statistical approaches for identifying shared genes controlling flowering time in rice, maize, and Arabidopsis. So far, methods for controlling for QTL overlap that are sensitive to gene density and recombination rates have been developed and submitted for publication. In the coming year, we will combine these models with shared QTL and mutational pathway data and deploy and publish these results.

Aim 4 (PO)

Support the Plant Ontology (Two Years Only)

In the last year with the transfer of the he Plant Ontology (PO) portion of the original Gramene plan has been moved to a separate project headed by Pankaj Jaiswal at Oregon State University. In this last year Gramene worked with the new staff at Oregon state to transition the project from CSHL to OSU. This included the project website, mailing list, and CVS repository. The transition was completed in the fall of 2009. CSHL still maintains a mirror of the site and will do this through the next year. With the retirement of this objective the scientific programmer Shuly Avaraham was released from the project.

Aim 5 (Outreach)

Education, Outreach and Diversity

Within the scope of the deliverables as stated in the grant Gramene currently has 3 stated outreach activities, Open Helix collaboration, traditional push at meetings, and the Gramene gene wiki. In addition to these we have continuous outreach in the form of collaboration with reference resources, large consortium, and individual research labs and have include the list of collaborators currently associated with the Gramene project. These collaborations are listed in the management section below.

OpenHelix

After ongoing discussions with OpenHelix and consultation with the Gramene PIs and NSF Program Manager, the OpenHelix scholarship for tutorials was reviewed. In April of this year the tutorials were announced. Below is a copy fo the Press release. In the next 6 months we will evaluate the proposed program. Based on the lag in time for this objective to become active we have an expected carry over of funds for participant activities. In the next year we will need to discuss alternative use for these funds within the lifecycle of the current grant.

OpenHelix press release:

"Gramene Announces Scholarships for Groups Underrepresented in Science to Learn How to Use Bioinformatics and Genomics Resources

Cold Spring Harbor Laboratory, Oregon State University and Cornell University, creators of the Gramene Resource for Comparative Plant Genomics, partner with OpenHelix to offer online training on genomic resources to encourage diversity in science.

Bellevue, WA (PRWEB) April 22, 2010 -- The creators of the Gramene Resource for Comparative Grass Genomics and OpenHelix announce the availability of scholarships to colleges and universities serving underrepresented minorities for full access to over 85 online tutorial suites on bioinformatics and genomics resources. The program is partially funded by the National Science Foundation (NSF).

"An ongoing goal for Gramene, our institutions, and the NSF, has been to provide opportunities for advancement and training to underrepresented groups in science," said Dr. Doreen Ware, of Cold Spring

Harbor Laboratory and Principal Investigator of Gramene, “So we are excited to be able to offer individual and institution scholarships to an extensive and valuable catalog of online training on genomics resources.” Recipients will have access to the OpenHelix catalog of tutorial suites on a wide range of bioinformatics and genomics resources, including Gramene, PlantGDB, NCBI tools such as Entrez Gene, BLAST and PubMed and many more. A full catalog of tutorial suites is available at http://www.openhelix.com/cgi/tutorials.cgi.

Each tutorial suite includes a 45-60 minute, online, self-run, narrated introductory tutorial on how to use a specific resource. The tutorial suite also includes PowerPoint slides, slide handouts and exercises which can be used as reference material or to build classroom content.

“The study of genomics has affected just about every area of life sciences, so learning how to access and interpret genomic data is critical to research success,” said Scott Lathe, Chief Executive Officer of OpenHelix, “With the convenience and broad accessibility of online training, we hope these scholarships will help in leveling access to this important training and further the potential and ongoing careers of the recipients.”

Institutions can apply for a scholarship for access to the tutorials at http://www.openhelix.com/cgi/scholarships.cgi. The scholarships are available to to minority serving colleges and universities. Underrepresented in science means those racial and ethnic populations that are underrepresented in biology research relative to their numbers in the general population. Individual scholarships are available to U.S undergraduates, graduate students, post-doctoral students, faculty and staff. Application deadline is June 30, 2010 and a limited number of scholarships are available.

About Gramene
Extensive research over the past two decades has shown significant conservation of gene order within large segments of linkage groups in agriculturally important grasses such as rice, maize, sorghum, barley, oats, wheat, and rye. Grass genomes are substantially colinear at both large and short scales, opening the possibility of using syntenic relationships to rapidly isolate and characterize homologues in maize, wheat, barley and sorghum.

As an information resource, Gramene's purpose is to provide added value to data sets available within the public sector to facilitate researchers' ability to understand plant genomes and take advantage of genomic sequence known in one species for identifying and understanding corresponding genes, pathways and phenotypes in other plant species.

Current work is being supported by the NSF Plant Genome Research Resource grant award #0703908. “

Meetings and Courses

Plant Genome Database module for Cornell course PLBR4060

Cornell University, Ithaca, NY; Nov 10, 2009.

Genevieve DeClerck & Charles Chen participated with Dave Matthews (of GrainGenes) in teaching the "Plant Genome Databases" module of the Plant Breeding & Genetics course, "PLBR4060: Methods of Plant Breeding Laboratory," which took place the afternoon of Nov 10, 2009. We presented a short introduction of the basic features of Gramene and then handed out an exercise for attendees to try on the computer lab machines.

Dale Bumpers National Rice Research Center

Stuttgart, AR; June 15-17, 2009

The Gramene Genetic Diversity curator, Genevieve DeClerck, traveled to USDA-ARS Dale Bumpers National Rice Research Center (DB NRRC) to present a talk about Gramene's Genetic Diversity module and how it can be used to mine data important to rice breeding projects. Another focus of the visit was to learn about the U.S. Rice Genetic Stocks (GSOR) “Core collection” maintained by the DB NRRC, who supplies rice seed stock for the National Small Grains Collection (NSGC) and germplasm information to the Germplasm Resources Information Network (GRIN). Gaining deeper understanding of germplasm 100_2590collections, both U.S. and international, is imperative as Gramene intensifies its focus on genetic variation and genotype-phenotype association data in the Genetic Diversity module. A finer-grained treatment of germplasm, genotype (alleles), and phenotype (trait measurements) will in turn be of great utility to plant breeders whose work has always focused on some or all of these data types.

The DB NRRC is situated in the heart of a highly productive area in Arkansas and the U.S. for rice farming. Arkansas annually pulls in about half of total U.S. rice harvest, followed by Texas, California, Louisiana, Mississippi, and Missouri (see http://arkrice.org for more info). Currently, about 50% (down from 100% 10 years ago) of the rice acreage in the U.S. South is planted with public cultivars developed by U. of Arkansas (Wells), LSU (Cocodrie), and DB NRRC (including Rondo and Dixiebelle for the parboiling industry, Presidio and Carolina Gold Select developed for organics, and Sierra and JES, which are aromatic rice).

OSU Bioinformatics Seminar

Corvallis, OR; June 17-18, 2009

http://mcbworkshop.cgrb.oregonstate.edu

Pankaj Jaiswal presented Bioinformatics, Genomics and Systems Biology symposium and workshop to the Molecular and Cellular Biology graduate program of the Oregon State University (OSU). Researchers, students and postdocs living in and/or visiting the area were invited to attend the workshop.

Jason Stajich from UC-Berkeley and Suzy Renn from Reed College were visiting presenters along with OSU faculty, postdocs and graduate students with expertise in this field. Topics

included mutational analysis of genomes, transcriptome studies and using genome annotation tools. There was extensive coverage on high-throughput DNA sequencing based approaches, such as Illumina 1G technology available at the CGRB, along with other genomic technologies. The workshop also provided hands-on experience to explore: (1) Gbrowse genome browser; (2) Familiarity with commandline access to bioinformatics tools; (3) Annotations tools such as the Gene Ontology and (4) Gramene Database: A resource for comparative genomics.

ASPB 2009

Honolulu, HI; July 20, 2009

Pankaj Jaiswal of Gramene along with Philippe Lamesch, A.S. Karthikeyan, Lukas Mueller presented a free workshop entitled "TAIR Workshop II: TAIR, PMN, Gramene and SGN workshop: focus on comparative genomics and new tools." In this workshop, four plant genome databases, TAIR, PMN, Gramene and SGN gave an overview of new tools available on their websites, including those focusing on comparative genomics. They presented a vision of the future in plant genome databases and how they affect plant biology research.

2009 ICG-IV

Shenzhen, China; August 25-28

Sharon Wei attended the 4th International Conference on Genomics and presented a poster entitled "Evolutionary analysis in sequenced plant genomes – applications in Gramene data."

Cereal Genomics Workshop

Cold Spring Harbor, NY; October 13-19, 2009

Doreen Ware helped lead a workshop on cereal genomics. The workshop featured morning and evening lectures with afternoon computer lab exercises which featured Gramene as one of the resource highlighted in the workshop. The workshop provided a framework for training through presentations and exercises. In addition Dr. Ware was able to obtain direct feedback on the Gramene site. The workshop also provided hands-on lab work in the comparative anatomy, phenotype and QTL sections lead by Ed Buckler and Torbert Rochert.

9th IPMB Congress

St. Louis, 2009; October 26, 2009

Pankaj Jaiswal attended the 9th International Plant Molecular Biology (IPMB) Congress to host a free workshop entitled "Comparative Plant Genomics with the Gramene database and Plant Ontology project."

Genome Informatics 2009

Cold Spring Harbor, NY; October 27-30, 2009

Liya Ren, Jim Thomason, Sharon Wei, Shuly Avraham and Ken Youens-Clark attended this meeting and presented posters on plant pathway databases, genome visualization, comparative genomics, Gramene's gene tree builds, and the use of Simple Semantic Web Architecture and Protocol (SSWAP) for data mining.

IRRI Rice Genetics Symposium and 6th Rice Annotation Project Meeting

Manila, Philippines; November 15-19, 2009

Pankaj Jaiswal presented Gramene's contribution to the rice genome annotation and reaffirmed its collaboration with various rice genomics and genetics database resources. Susan McCouch presented a plenary talk entitled "Exploring the genetic diversity of rice" on Nov 16. Genevieve presented a poster titled, "The Gramene Genetic Diversity module: A resource for comparative genetic diversity analysis in plants."

PAG 2010

San Diego, CA; January 9-13, 2010

Palitha Dharmawardhana, Genevieve DeClerck, Charles Chen and Ken Youens-Clark attended the Plant and Animal Genome XVIII Conference. They presented posters on metabolic pathway networks for cereal plants, genome visualization tools, and using our genetic diversity data for genotype-phenotype association analysis in grass species as well as flowering time variation across three plant species. They additionally presented an hour-long demonstration on using the Gramene website as a genomics and genetics resource for rice and other grasses. Gramene also participated in an outreach booth manned by many other plant databases including CottonDB, GDR, GrainGenes, LIS, MaizeGDB, Maizesequence.org, Oryzabase, PlantGDN, SGN, SCRI, Soybase, and SGN.

Rice Technical Working Group

Biloxi, MS; February 22-25, 2010

Genevieve DeClerck attended The Rice Technical Working Group's (RTWG) 33rd Meeting in Biloxi, MS February 22-25, 2010 (www.rtwg2010.com). She presented a poster titled, "The Gramene Genetic Diversity module: A resource for genotype-phenotype association analysis in grass species" which provided an overview of the data sets housed in and tools offered by the Gramene Genetic Diversity module. Genevieve also presented a talk titled "Rice genotype, phenotype, and germplasm data in the Gramene database" at the RTWG Applied Genomics Workshop. The goal of this presentation was to discuss the features and goals of the Gramene Genetic Diversity module, and how it integrates with the wealth of data in Gramene, for a group of rice breeders and geneticists interested in rice SNP data projected for release in 2010.

Attending the RTWG allowed Genevieve to further understand needs of members of the Gramene community, and it offered an opportunity for members of the U.S. rice research community to become more familiar with Gramene and how it can help them with their work. The focus of the Applied Genomics Workshop was to discuss details about the recent re-sequencing of 13 U.S. rice varieties by Brian Scheffler et. al. (USDA-ARS Stonewille, MS; in attendance), specifically, what the data will look like and how people will access and use the data. Genevieve's presentation

played a key role in the discussion because she talked about how the Gramene Diversity module manages data like this and what tools will be available for querying and analyzing their rice data once it's in Gramene. At the conclusion of the meeting, several attendees expressed relief and gratitude that Gramene is and will be involved in the management of this and other large rice diversity datasets because many labs are ill-equipped to handle data on this scale.

51st Maize Genetics Conference

Riva Del Garda, Italy; March 21-23, 2010

Doreen Ware attended the maize genetics conference and presented a talk highlighting updates on the maize B73 genome sequencing project. The talk highlighted the Gramene compara module which currently includes the filter set of maize genes currently the community standard and discussed the upcoming release of MaizeCyc which will be jointly curated by Gramene and MaizeGDB.

International Symposium on Integrative Bioinformatics

Cambridge, UK; March 22-24

Will Spooner presented a poster entitled "Gramene GeneTrees: A comprehensive phylogenomics database in plants and other model Eukaryotes."

Management

Meetings

Activities within the group are coordinated through weekly staff meetings, six PI meetings, an internal Wiki, and the Mantis bug-tracking system for project management of specific deliverables. In addition to these regularly scheduled activities, the project also participated in the yearly Scientific Advisory Board meeting in December of 2009 has scheduled project retreat for June of 2010.

Staffing

Gramene is currently understaffed due to hiring constraint at Cornell as well as loss of project staff due to attrition. In the last year we saw two additions to the Gramene team. Dr. A. S. Karthikeyan was hired as a biological curator at Cornell with Dr. McCouch, filling one of the two outstanding positions at Cornell and will be primarily working on specific aim 3. At CSHL Dr. Joshua Stein was engaged part time on the project as a plant-computational biologist working on specific Aim 1. Dr. Stein was hired to replace the biological curator that has been unavailable from Cornell in the past 1.5 years of the project. In addition to gains in the last year, Gramene lost two Staff members. At CSHL Shuly Avaraham was released from the project with the retirement of the Plant Ontology objective specific AIM 4. At Cornell Dr. Noel Yap working on Specific Aim 1 and Aim 3 is no longer working with the project.

Gramene’s Scientific Advisory Board (SAB)

Gramene was honored to have the following advisors this year:

- David Marshall (Chair)
- Paul Flicek
- Michael Ashburner
- Anna M McClung
- Georgia Davis
- Patricia Klein
- William Beavis
- Tim Nelson

The SAB met with the members of Gramene via a WebEx teleconference where we were able to share slide presentations. Our agenda included a time for Gramene to present updates on specific objectives and to discuss specific requests for suggestions on existing challenges. SAB projects member were then allowed time for internal discussion, without Gramene members being present, and then the remaining time was devoted to further discussions between the SAB and Gramene. At this time we are still awaiting a final report from the Chair, David Marshall. Their SAB Report 2009 and our responses are included with this report.

Gramene Scientific Advisory Board Report and Responses

Synopsis

This is the report of the 2009 Gramene Scientific Advisory committee from the 6 hour WebEx conference held December 16, 2009, and submitted January 2010. Gramene responses to the report are included inline.

The agenda of the meeting included reports on the progress of each of the 5 specific Aims, internal discussion among the SAB, and General discussion with the team and the PIs individually.

SAB Report

Introduction

The SAB have been impressed with the progress made over the last year, in particular, given the constraints that the team has faced with respect to available funding, delays in recruitment and the volume of new information. The Gramene project plays a significant role in and acts as a primary focus for the integration of rice information and its relationship with key cereal and more general plant species. With developments in new high throughput technologies for sequencing, genotyping and expression analysis and their application to a broader range of species and genotypes we anticipate that the value of this type of integration activity will increase dramatically in the next few years.

Specific Aim 1 (Genomes)

With the growing list of major additions to the set of sequenced plant genomes a major challenge for Gramene is how to prioritize the addition of new genomes and balance this

against the need to actively curate the existing portfolio. The SAB appreciates the issues involved and would suggest that the following genomes should be considered:

1. The Brachypodium genome will be publically released shortly and, in the near term, provide the closest model genome to a number of key crop plants including wheat, barley and forage grasses.

Gramene response: The Brachypodium is currently in production within Gramene and will continue to be hosted by the project. In addition to hosting of the genome, there are currently on going discussions for collaboration on curating pathways.

A number of legume species are currently sequenced or in production, and one of these should be identified and supported. The options are to make the choice based on either their use as a model system (e.g. Medicago) or because of their inherent agronomic value (e.g. soybean – though this is clearly complicated by polyploidy).

The sequencing of the related Solanaceous crops, potato and tomato, are both at a relatively advanced stage and the choice of one of these would provide a route into an extremely valuable crop taxonomic group.

Gramene Response. We agree that there will be value in hosting a legume and Solanaceaous species within the Gramene framework. For this to move forward we will need to work with both the communities and funding agencies to coordinate this effort. We anticipate the ability to bring in at least one genome in the next year and will target one or both of these genomes.

In terms of lower plants, Selaginella appears to have some attraction because of its role as a vascular plant model.

Gramene Response: Gramene would like to support the integration of one lower species to support the evolution analysis. We will be reviewing the Seleaginella and Physcomitrella for a release in the next year.

With the latest methods of transcriptome profiling utilizing next generation sequencing technology, it would be valuable to add an additional track to the genome browser to display this type of data and to encourage the community to provide such data to Gramene for improving current genome annotation.

Gramene response: We believe that it will be possible to host these tracks as DAS tracks and exploring requirements to streamline this process.

With the expansion in the number of sequenced plant genomes that is now, or will shortly become available, it is of major importance that Gramene is able to deliver to the user community facile and robust tools that will enable them to efficiently exploit comparative information. Currently both gene trees and CMap provide support for this activity but it is

important that Gramene monitors whether these tools are evolving sufficiently fast and have the capacity to meet user needs.

Gramene response: We acknowledge the SAB recommendations regarding the improvements in tools and the need for usability as the data scales.

Specific Aim 2 (Pathways)

Gramene has ambitious aims with respect to the curation of metabolic, regulatory and other pathways. While generally congratulating Gramene's work in this field, the SAB must express some concerns. The first is that the Gramene group is working in the context of three quite different systems with respect to pathway curation. The first is the Cyc family of databases maintained by the SRI and requiring the use of proprietary tools. The second is within the context of the Reactome family of databases maintained at CSHL and the EBI; the third is within the context of the community WikiPathways efforts. Each of these has particular advantages and disadvantages (e.g. the Cyc databases cannot handle regulatory pathways). However, with limited resources this diversity means a dilution of effort. It is our strong recommendation that, as a matter of priority, Gramene should settle on one system (and our preference would be for Reactome) and simply allow the Cyc and WikiPathway databases to be populated by automated exports using BioPax.

Gramene response: We acknowledge SAB suggestions and concerns regarding the focus of our efforts. In this reference we have established collaboration with the Reactome database and carried out a couple of test cases of importing the curated pathways from the existing RiceCyc database. After a careful adjustments to the Reactome's BioPax based retrieval of pathway information tools, we are looking forward for a bulk import of the current curated information in the Reactome. We expect that after a successful import and quality checks we will be able to release the first Rice Reactome by the end of the fall of 2010. During this time the curators will be mentored by the Reactome staff for their training and curatorial help. After the release of Rice Reactome we will look forward to integrate the additional cereal plant reactomes by making use of the gene product annotations based on the species specific manual curation and those derived largely by the phylogeny and synteny based methods applied in specific aim-1. During this move we expect that our users will have to learn the changes to the user interface and content. We will make every effort by way of outreach and tutorials to help our users get through the transition.

Specific Aim 3 (Diversity)

The Gramene Diversity group clearly faces particular challenges with respect to the growing volume of genotype and diversity sequence data that is becoming available for its major target organisms. There is a need to clarify and advertise just what datasets are, or will become, available and then to identify what likely users of the data will require. A major complication of this type of diversity data is that many of the potential users will

aim to exploit it in conjunction with some further biological analysis of the germplasm from which the genotype data has been derived. This leads in turn to the need to grasp and resolve the complications that arise from the relationship between accessions in genebank terminology and true-breeding genetic stocks. It is important that precise seed sources directly associated with the biological entity that was sequenced or genotyped are documented and ideally available as a "set" from a single source. Recent changes in International Law with regard to the movement of germplasm may complicate the issue further. However we note that this issue is not a Gramene specific one but it is one that should be acknowledged.

Serving the diversity data is also a complex problem. Many users may not be well prepared to receive the large data sets and may well lack the software tools to hold and analyze or visualize the data they download. We welcome the moves to provide key software tools to support this activity but we also anticipate that many users will require significant support in their deployment.

Gramene response. We are working on getting these tools working and there becoming better integration. Direct launch now works, but needs to be deployed. Wizards have been piloted.

It is also not yet clear what the end game is for Gramene with respect to the analysis of diversity data sets. Is it the intention to not only store the genotype data but also phenotype data and the results of analysis?

Gramene response: We already store genotypic and phenotypic data, although recent curation by Gramene curators has focused on genotypes. A shift to more phenotypic curation will occur over the coming year. The database schema is being modified for storing results, but we have not implemented yet. We expect implement by summer and fall of 2010.

If this is the case then considerable thought will need to be given to how the results of this analysis are stored and displayed. There are a number of interesting options. For example, for the key target character of flowering time displaying the relationship of the genotype by trait analysis with potential candidate genes and response pathways could provide an interesting challenge.

Specific Aim 5 (Outreach)

The main issue with regard to the Gramene outreach activities appears to be one of identifying and supporting one or more core audiences. Is the target audience molecular geneticists or is it plant breeders? Is it both? For US breeders it would appear that the USDA CAP projects or comparable projects and their associated informatics may be the most appropriate route to serve much of the breeding community. Are plant breeders a reasonable target audience? If they are, then what would a plant breeder really need? Is the ability to identify allelic variants associated with a desirable phenotype sufficient? Should this be pursued through a complementary funding mechanism? It might be good

to evaluate the stakeholder communities and determine what they use. If the plant breeders, particularly the commercial plant breeders in the US, are conducting data dumps, then user interfaces and analytical tools may not need to be developed by Gramene. This is true for maize breeders, but not for other crops supported by Gramene and not for the international maize breeders.

Gramene response: This is an excellent point raised by the SAB and we acknowledge their concern. In order to build a database portal that is applicable to a wide array of scientists, Gramene has strived for the last several years to stand out from rest of the community plant genome databases, most of which are species-specific. This has been done by careful development, curation and integration of information from both reverse and forward genetics studies to create a confluence of information streams. This way, different, research communities get a chance to link the pieces of their puzzles, and move in either direction to strengthen and validate their hypotheses. This effort has been acknowledge by users in both the breeding and genetics communities. For example, Yamamoto et al (http://news.gramene.org/?p=482 In 2009). acknowledged Gramene's QTL resources in a paper in DNA Research entitled "Towards the Understanding of Complex Traits in Rice: Substantially or Superficially?", and in 2010, a paper published in BMC Plant Biology entitled "Transcriptional regulatory network triggered by oxidative signals configures the early response mechanisms of japonica rice to chilling stress" noted Gramene's usefulness to understanding rice transcription factors. Similarly Muylle (2005) identified of four QTLs that determine crown rust (Puccinia coronata f. sp. lolii) resistance in a perennial ryegrass (Lolium perenne) population and Armstead et al (2004 and 2005) employed fine-mapping strategies to identify candidate genes for crown rust (Puccinia coronata) resistance from meadow fescue (Festuca pratensis) which were intriogressed into Italian ryegrass (Lolium multiflorum). Both of these studies used the rice genome annotation and comparative map tools provided by Gramene as well as the species-specific markers and maps provided by its collaborators to identify the molecular markers that were used in their breeding strategies. Thus, by delineating regions, genes and traits of interest within a comparative framework, Gramene has consistently contributed essential tools and knowledge to the development of novel approaches in genetics and breeding, and has facilitated the use of integrated comparative genomics and genetics approaches towards the betterment of our US agriculture. With these words we would like to say that despite all the clutter of data set deluge, Gramene has been able to provide a clean and well annotated reference datasets and was able to identify or rather has avoided identifying any precise core users by letting the communities create a self learned as well as contributed by our extensive outreach programs lead to the adapted niche among its users for its presence and value for not just the reference model crop plants but for others with lesser known resources.

For future releases, Gramene aims to address some specific requests that have come from the US breeding and genetics communities, while at the same time, encouraging the independent development of more specialized applications and user interfaces for specific user communities that may "dock" to Gramene or extract information from Gramene. Based on conversations with the US rice breeding community during 2009, we are developing ways for Gramene users to search on phenotype and find links to specific

alleles and to germplasm resources known to carry those alleles. To develop this capacity, we have recently hired a new curator who will address this issue during 2010. Similarly, geneticists expressed their interest in being able to search on allele, and find links to genes, germplasm and phenotype, and we also aim to implement this capability during 2010. These improvements, along with enhanced visualization tools, statistical approaches and curated datasets will provide Gramene users with additional opportunities to link phenotype with genotype and aim to enhance the utility of Gramene as an information resource for the breeding and genetics communities, while continuing to support the powerful comparative framework previously developed for genes, proteins, phenotypes and pathways.

With the resources available to Gramene it is important that they are focused on realistic targets and a better understanding of the actual audience may help to achieve this. A related issue is to clarify relationships with other projects, for example, MaizeGDB, the MSU rice project, PlantGDB and PlexDB. To many in Gramene's external audience it is not at all clear what the relationships and responsibilities are. Perhaps in conjunction with PI's from such projects it would be a good time to begin to develop a forward looking white-paper that explores the needs and solutions for these and other related projects at both a national and international level. Such a framework would help clarify the investment in resources across a broad spectrum of activities and, with respect to the Gramene group, help clarify the balance in effort and future funding applications between a single fully integrated project and individual more focused projects working within an "agreed" framework. Reference to such a document would also help grant reviewers enormously in understanding the context of any future applications.

Gramene Response: For each of the specific Aims within the Gramene proposal objectives support basic research as well as providing a framework for integration of community data sets. For each of these objectives there was identified existing as well as anticipated data sets that would be utilized in the project. The Gramene project also works directly with new projects as they are being developed with the objective of identifying appropriate data set for future integration and the resources to support these. In a better effort to communicate our existing relationships with projects, we will actively review the collaborators page and make an effort to highlight these collaborations in news posts. With regard to exploring the existing as well as future challenges of coordination and stewardship of plant related data sets the PIs on this project are actively participating in several workshops addressing these challenges at the national and international level. We anticipate that these workshops will result in a series of recommendations to address many of the concerns expressed by the SAB.

In more specific terms, we strongly endorse the re-evaluation of the Open Helix collaboration and the change in the way Gramene users are to be supported through a change in format to mini tutorials, perhaps structured around an FAQ and linked to short Flash or comparable video shorts, dealing with specific tasks. We also believe that individual module developers should, when possible, work on outreach directly, both because of the unique aspects of each area of activity and the value of feedback from their specific target audience.

General Issues

The SAB have some concerns that too much is being expected, by Gramene, from community curation. This has been a major concern in the model organism database community for some years now and no great progress has been made, other than the collaboration made between TAIR and some journals. There seems to be a wish to emulate this within Gramene, however, no evaluation of the TAIR experiment has been made, as far as we can determine. If database quality is to be maintained (as we are sure it will be) any community annotation must be curated and we question whether, in fact, this would really save effort compared with direct curation from the literature. In addition there are concerns as to how members of the community will become the owners of curated information and how credit/benefit from such activities be realized.

Gramene response: According to our discussions with the TAIR staff, the submissions from the publisher's site (data is primarily collected by the publisher), have increased gradually. Also the submission form allows users (authors mainly) to self select the ontology based annotations in the form. These annotations are often accurate by using either the precise Gene Ontology and Plant Ontology term or a generic by selecting a top level parent term. According to the TAIR curators, the major time consuming step is the curation of dependencies which are unique to a given database structure, e.g. adding the citation to the literature database, users name, etc. prior to its usual import from PubMed in the database on our/their end, creating map positions, etc. Based on TAIR's experience and our discussions with ASPB, we think a substantially improved data flow can be generated for more automation and responsibility on part of the authors.

This SAB report is a good opportunity to register our concern that NSF is backing off on infrastructure support for essential community databases and stock centers (e.g., TAIR and others). iPlant is not going to fill this need! A greater investment rather than a reduced investment should be the priority, as all research becomes more dependent on networked datasets in need of curation. It is unrealistic to transition to fee-based or private centers. This should be a public-supported library-type system accessible to all. The apparent new directive for NSF to focus more on "mission-based" projects makes the deterioration of infrastructure support an even greater concern. Clearly it may also be reasonable to look to collaboration with other international funding agencies to support some aspects of this type of activity. However there is a growing need to develop and support the type of resource that Gramene represents and both mid and long term funding solutions need to be developed.

With the major increase in volume and complexity of sequence-based data sets that are already available and the rapid growth in scale that Next Generation and further novel sequencing technologies bring, both the challenge and value of data integration will only increase. Funding agencies need to take on board the realization that the value of their investment in these technologies will be significantly reduced unless comparable resources are also invested in the necessary informatics infrastructure and integration.

Collaborations

Plant Ontology

Gramene continues to be an active participant in the Plant Ontology consortium. Gramene make use of the PO as a controlled vocabulary for describing anatomy and developmental stages as part of the automated annotation process. With each release Gramene submits our annotation file gene, and QTL associations to PO CVS repository.

Gene Ontology

Gramene continues to be an active participant in the GO consortium. We use the gene ontology from GO cvs repository as a standard vocabulary within the project website. We submit our annotation file of gene ontology and rice protein associations to GO CVS repository at each Gramene release.

Uniprot

We get all Poaceae proteins from Uniprot website at each Gramene release. Uniprot links back to Gramene protein pages as cross references.

OryzaSNP Project

Worked with Ken McNally (IRRI), Keyan Zhao (NSF-TV), and Kevin Childs (of MSU) on acquiring the OryzaSNP dataset, which was in IRGSPv4 coordinates. Kevin Childs could not help me map the data to MSU6 as it was not a priority for his group. Qi Sun (Cornell CBSU) helped me map the snps over with the 1000 bp flanking seq fasta provided by Kevin Childs.

Associated publication: Ken McNally (McNally et. al PNAS 2009 (pubmed: 19597147))

Species: Oryza sativa (21 indica and japonica varieties included in study)

Release #: 31*

(*The OryzaSNP dataset will be the Diversity database in release #31, but will not appear in EnsemblVariation until interim release (projected summer 2010) or Gramene release #32, depending on PlantEnsembl priorities)

Grape genome

In the last year the Vitis vinifera browser was updated to the 12x whole genome shotgun sequence assembly and community annotation. Olivier Jaillon, Benjamin Noel from Genoscope (CEA - Institut de Génomique) provided direct help on retrieving the data. The data were prepared by a French-Italian Public Consortium for Grapevine Genome Characterization under the auspices of the International Grape Genome Program (IGGP).

Brachypodium genome

In the last year Gramene has worked with the Brachypodium sequencing consortium to bring the newly released assembly and annotations as one of the genomes hosted by Gramene. During this last year we worked directly with Manuel Spannagl and Dr. Klaus Mayer from MIPS/IBIS, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH) and John Vogel at OSU.

Poplar genome

Since 2008, Gramene has carried the poplar genome (version 1.1, July 2004) from the Genome Projects web site at the JGI. The genome annotation has recently been updated to version 2.0, available from the Phytozome web site, also at the JGI.

MaizeGDB

In the last year we have actively worked with larger MaizeGDB group under the supervision of Dr.Carolyn Lawrence,. The collaboration included updates to reference genetic maps, including maize IBM neighbors 2008 maps as well as expanding Gramene's understanding of their data types and mapping protocols at MaizeGDB. In addition the collaboration included coordination of workshops and booths at meeting, and active discussion regarding coordinated curation of a maize pathway tool.

Maize B73 Genome Sequencing project

In the last year Gramene has integrated the maize gene models in the gene trees. In the next year with submission of the reference assembly and annotations to GenBank Gramene will be integrated Version 2 assemblies and annotations

Pathway

RiceCyc, MetaCyc, WikiPathways, Reactome, Plant Metabolic Network (PMN), Solanaceae Genomics Network (SGN), Todd Mockler (Oregon State Univ.), Tim Nelson (Yale Univ.), Tom Brutnell (Cornell Univ.)

- Pathway-tools. Gramene build pathway database with Pathway-tools developed by Peter D. Karp's group at SRI International. We thank them for the software support. (The development of Pathway Tools is funded by grants GM077678, GM080746, and GM75742 from the National Institutes of Health.)
- MetaCyc (NIH grant GM080746), EcoCyc (NIH grant GM077678). MetaCyc and EcoCyc are curated at SRI International and are bundled with the pathway-tools software.
- PlantCyc, PoplarCyc and AraCyc are provided by Plant Metabolic Network (PMN) (NSF Grant #: 0640769). Seung Yon (Sue) Rhee is the PI and we thank for the help from their curator and director Peifen Zhang. Gramene also contributes back the RiceCyc and SorghumCyc to PlantCyc.
- Solanaceae Genomics Network (SGN) provides us the LycoCyc(tomato), CapCyc(peper) and PotatoCyc.[SGN is supported by the NSF (#0116076) and USDA CSREES]
- MedicCyc is provided by MedicCyc group at Noble foundation. They use Gramene RiceCyc to verify the predicted Medicago truncatula pathways.[References Urbanczyk-Wochniak, E., Sumner, L.W., (2007) MedicCyc: a biochemical pathway database for Medicago truncatula, Bioinformatics, 23, 1418-1423]

Reactome

Gramene is working with Guanming Wu and Peter D'Eustachio from the Reactome project (NIH grand N01AI40041 and EU 6th Framework Programme grant LSHG-CT-2005-518254 to ENFIN) to export our RiceCyc into Reactome with BioPAX-II format.

Arabidopsis 2010 Nordborg collaboration as part of the NSF DBI 0723510

The Gramene project is coordinating the integration of arabidopsis variation data and phenotypes in coordination with on NSF DBI 0723510 of which Doreen Ware serves as Co-PI .

The Nordborg Lab website was the source for v3.04, v.3.03, and v3.02 genotype data entered into the Diversity module over the past year.

Jan Dvorak and the Wheat D genome project (NSF)

Jan Dvorak provided Gramene with the resources to load his Aegilops tauschii genetic as described in the paper "Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in Triticeae" (PNAS, 2009, vol. 106, pp. 15780-15785).

NASC

Nick James provided Gramene with the Ensembl core database populated with the Arabidopsis thaliana reference genome assembly and gene predictions. Our genome browser provides links back to the appropriate pages of the at Ensembl browser at NASC.

TAIR

Gramene makes use of the arabidopsis genome and annotations that are curated by the TAIR database in Specific AIM 1. In addition in Specific Aim 2, pathways Gramene hosts AraCyc to support cross-genome comparison of biochemical pathways.

Genbank

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (Nucleic Acids Research, 2008 Jan;36(Database issue):D25-30). Gramene makes active use of the plant sequences available in Genbank and provides semi-automated integration of these sequences within the context of other Gramene modules. We provide links backs between the reference sequences available in the Genome browser and where appropriate GenBank provides reciprocal links back to Gramene. With the most recent inclusion of the Gerplasm modules the project has been in active exchange with the NCBI regarding the plant species nomenclature.

Ensembl Genomes

Ensembl Genomes project produces genome databases for important species from across the taxonomic range, using the Ensembl software system. Five sites are now available: Ensembl Bacteria, Ensembl Protists, Ensembl Metazoa, Ensembl Plants and Ensembl Fungi. Dr. Ware's group with the European Bioinformatics Institute are working collaboratively on the integration of content, quality control and the development of new features for Plant Ensembl http://plants.ensembl.org/. Fall 2009 was the first release of Ensembl Plants.

Sorghum genetic map

Dr. Emma Mace of Queensland, Australia, kindly provided Gramene with resources and extensive help loading her genetic map of sorghum as reported in the paper "A consensus genetic map of sorghum that integrates multiple component maps and high-throughput Diversity Array Technology (DArT) markers" (BMC Plant Biol, 2009, vol. 9, pp. 13-13). This was made available in Gramene's build #30.

Barley genetic map

Victoria Blake and Dave Matthews from GrainGenes and Dr. Tim Close from UC Riverside helped provide Gramene with the necessary resources to load the barley genetic map as described in the paper "Development and implementation of high-throughput SNP genotyping in barley" (BMC Genomics, 2009, vol. 10, pp. 582-582). This was made available in Gramene's build #30.

Software

Gramene Software

Gramene Website Infrastructure

As part of an effort to improve Gramene's infrastructure, the web server software running Release 31 has been upgraded from Apache 1.x to the Apache 2.x branch. This upgrade should be largely transparent to our users, but any sites mirroring Gramene will see a significant improvement as they will now only be required to install and maintain a single installation of Apache 2.x as Ensembl migrated to this version well before Gramene.

In addition to allowing a single web server installation, this was a necessary upgrade since Apache 1.x will no longer be actively supported. By moving to Apache 2.x, we ensure a consistent line of support, bug patches, and feature enhancements from developers. Further, this will make our own internal development easier. As more and more projects migrate to being Apache 2 only, this transition will help ensure that we always have access to the latest tools and technologies.

All files associated with Gramene's website are packaged and made available on our public FTP site with each release.

For Release 31 the User Interface for the website was updated Figure 3.

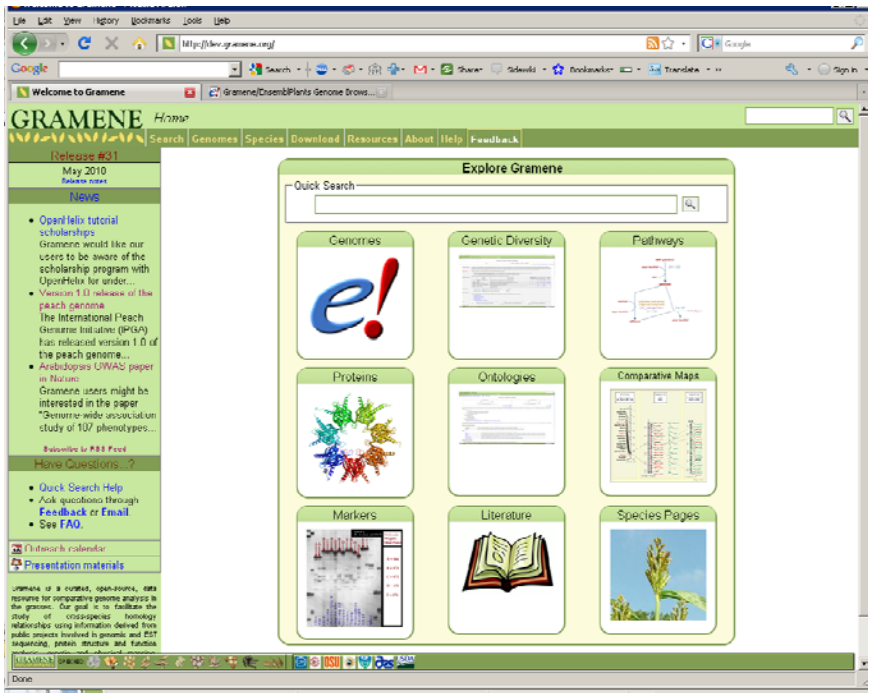

Figure 3: Preliminary view of the new homepage

PlantGeneWiki

Did the first official release of PlantGeneWiki (http://plantgenewiki.gramene.org/). The code to generate the wiki files from our databases is included in our software release and is publicly available on our FTP site.

Ensembl

Gramene continued to track the major Ensembl software updates this year. Alongside numerous minor fixes and enhancements, we made several significant contributions to the Ensembl code base including the addition of a GFF3 export facility to the Ensembl web site, and extensions to the eHive pipeline system to allow workflows to be run on SGE scheduler.

Our long-standing relationship with the Ensembl project, and recent collaboration with Ensembl Genomes, has increased our influence over the strategic direction of Ensembl, allowing us to advocate for functionality that is likely to be of use to plant biologists, and correct areas where the assumption of metazoan-characteristic genomes causes problems for genomes of alternative structures;

- Since our contribution of code for the visualization of gene trees was recognized in 2009 (Ensembl 2009 -- Hubbard et al., 10.1093/nar/gkn828), we have been involved in planning further extensions. Of particular interest are the addition of stable GeneTree IDs (maintained between database releases), and the ability of the pipeline to detect gene split/fusion events that may point to gene prediction/assembly errors.

- We have contributed a new method of estimating syntenic regions between pairs of genomes that is independent of the size of the source genomes. We are also collaborating over methods that can perform self-genome comparisons to identify duplications within a genome.

- For genome variation data, we have identified the visualization of phenotype associations with genome variations as a priority, and ensured that such views are under active development at Ensembl.

Ensembl is a separate, open-source project which makes its releases publicly available many times throughout the year.

GDPDM

Gramene's diversity data is stored in the Genomic Diversity and Phenotype Data Model (GDPDM) database schema (http://www.maizegenetics.net/gdpdm). In version 4.0, significant improvements were made to store SNPs, SNP IDs, Indel, and physical positions information by using a new haplotype packing design such that SNP values consume only 4 bits of memory.

The GDPDM schema is publicly accessible through the above URL.

TASSEL

We have made important changes to our diversity analysis software tool, TASSEL (http://www.maizegenetics.net/tassel) resulting in the release of version 3.0. Using the GDPDM 4.0's haplotype packing schema, we have completely redesigned and improved TASSEL's way of handling and manipulating data. Other new developments include a new alignment viewer, a progress monitor, new Wizard tools for users, integration of new data formats imports (e.g., Plink, HapMap, Flapjack), and automatic data loading/analysis at start up.

Significant improvements have been made to the MLM and GLM association analysis. TASSSEL's pipeline has been greatly improved also as it is no longer necessary to write Java code to create them, and simultaneous pipeline segments can be executed. All the pipeline infrastructure works with the web site launch, command line interface, GUI client, and new Wizard.

We've been responding to many users from various research and corporate organizations including CGIAR, Laboratorio de Biotecnología (Buenos Aires - Argentina), Kansas State University, Dale Bumpers National Rice Research Center, University of Agricultural Sciences (GKVK, Bangalore), NCSU, Dept Genética Molecular de Plantas (Madrid Spain), Universidad Politécnica de Valencia, serasem, Dow AgroSciences LLC, Chargé d'études en Bioinformatique, USDA ARS DBNRRC, University of California Davis, Oregon State University, and many others.

TASSEL is freely available from the maizegenetics.net website.

CMap

CMap was initially created by Gramene in the early years of the project and has remained stable at version 1.01 since its release in 2008. This version has been downloaded almost 800 times from Sourceforge, where the software is hosted, and is used extensively in the community by groups studying plants, insects, bacteria and animals.

This past year also saw the publication of a paper on CMap version 1.01. As time and resources have allowed, development work has continued on an entirely new version of CMap version 2 that promises sharper images using Scalable Vector Graphics (SVG), improved performance using a redesigned database and possibly the FastBit engine, and the integration of the popular circular genome visualizer, wCircos.

CMap is freely available from the Generic Model Organism Database (GMOD) project downloads page hosted by Sourceforge.net.

SQL::Translator

This Perl application allows for the automatic transformation of SQL-based relational database schemas to be transformed into many other formats ranging from various SQL dialects (e.g., MySQL to Oracle), documentation (HTML, UML, entity-relation diagrams), Perl code (Object-Relational Models [ORM]), or any other format through custom-written code. This module grew out of the development work on CMap and is now a healthy open-source project run entirely independent of Gramene by a team of 20-odd developers. Gramene uses SQL::Translator to help generate ORM classes from our many databases as well as E/R diagrams we include with our releases.

SQL::Translator is freely available on the Comprehensive Perl Archive Network (CPAN) which is mirrored world-wide.

.

Publications

Gramene Collaborations resulted in the following papers

Evidence-based gene predictions in plant genomes. Liang, C, Mao, L., Ware, D. and Stein, L. (Genome Res. 2009 Oct;19(10):1912-23)

Microdissection of Shoot Meristem Functional Domains. Brooks, L., Strable, J., Zhang, X., Ohtsu, K., Zhou, R., Sarkar, A., Hargreaves, S., Elshire, R., Eudy, D., Pawlowska, T., Ware, D., Janick-Buckner, D., Buckner, B., Timmermans, M., Schnable, P., Nettleton, D., Scanlon, M. (PLoS Genetics 5(5): e1000476. doi:10.1371/journal.pgen.1000476)

The B73 Maize Genome: Complexity, Diversity, and Dynamics. Patrick S. Schnable, Doreen Ware, Robert S. Fulton, Joshua C. Stein, Fusheng Wei, Shiran Pasternak, Chengzhi Liang, Jianwei Zhang, Lucinda Fulton, Tina A. Graves, Patrick Minx, Amy Denise Reily, Laura Courtney, Scott S. Kruchowski, Chad Tomlinson, Cindy Strong, Kim Delehaunty, Catrina Fronick, Bill Courtney, Susan M. Rock, Eddie Belter, Feiyu Du, Kyung Kim, Rachel M. Abbott, Marc Cotton, Andy Levy, Pamela Marchetto, Kerri Ochoa, Stephanie M. Jackson, Barbara Gillam, Weizu Chen, Le Yan, Jamey Higginbotham, Marco Cardenas, Jason Waligorski, Elizabeth Applebaum, Lindsey Phelps, Jason Falcone, Krishna Kanchi, Thynn Thane, Adam Scimone, Nay Thane, Jessica Henke, Tom Wang, Jessica Ruppert, Neha Shah, Kelsi Rotter, Jennifer Hodges, Elizabeth Ingenthron, Matt Cordes, Sara Kohlberg, Jennifer Sgro, Brandon Delgado, Kelly Mead, Asif Chinwalla, Shawn Leonard, Kevin Crouse, Kristi Collura, Dave Kudrna, Jennifer Currie, Ruifeng He, Angelina Angelova, Shanmugam Rajasekar, Teri Mueller, Rene Lomeli, Gabriel Scara, Ara Ko, Krista Delaney, Marina Wissotski, Georgina Lopez, David Campos, Michele Braidotti, Elizabeth Ashley, Wolfgang Golser, HyeRan Kim, Seunghee Lee, Jinke Lin, Zeljko Dujmic, Woojin Kim, Jayson Talag, Andrea Zuccolo, Chuanzhu Fan, Aswathy Sebastian, Melissa Kramer, Lori Spiegel, Lidia Nascimento, Theresa Zutavern, Beth Miller, Claude Ambroise, Stephanie Muller, Will Spooner, Apurva Narechania, Liya Ren, Sharon Wei, Sunita Kumari, Ben Faga, Michael J. Levy, Linda McMahan, Peter Van Buren, Matthew W. Vaughn, Kai Ying, Cheng-Ting Yeh, Scott J. Emrich, Yi Jia, Ananth Kalyanaraman, An-Ping Hsia, W. Brad Barbazuk, Regina S. Baucom, Thomas P. Brutnell, Nicholas C. Carpita, Cristian Chaparro, Jer-Ming Chia, Jean-Marc Deragon, James C. Estill, Yan Fu, Jeffrey A. Jeddeloh, Yujun Han, Hyeran Lee, Pinghua Li, Damon R. Lisch, Sanzhen Liu, Zhijie Liu, Dawn Holligan Nagel, Maureen C. McCann, Phillip SanMiguel, Alan M. Myers, Dan Nettleton, John Nguyen, Bryan W. Penning, Lalit Ponnala, Kevin L. Schneider, David C. Schwartz, Anupma Sharma, Carol Soderlund, Nathan M. Springer, Qi Sun, Hao Wang, Michael Waterman, Richard Westerman, Thomas K. Wolfgruber, Lixing Yang, Yeisoo Yu, Lifang Zhang, Shiguo Zhou, Qihui Zhu, Jeffrey L. Bennetzen, R. Kelly Dawe, Jiming Jiang, Ning Jiang, Gernot G. Presting, Susan R. Wessler, Srinivas Aluru, Robert A. Martienssen, Sandra W. Clifton, W. Richard McCombie, Rod A. Wing, and Richard K. Wilson (Science 20 November 2009 326: 1112-1115 [DOI: 10.1126/science.1178534])

Species trees from highly incongruent gene trees in rice. Cranston, K., Hurwitz, B., Ware, D., Stein, L. and Wing, R. (Systematic Biology 2009 58(5):489-500; doi:10.1093/sysbio/syp054)

CMap 1.01: A comparative mapping application for the Internet. Youens-Clark, K., Faga, B., Yap, I., Stein, L., Ware, D. (Bioinformatics, doi:10.1093/bioinformatics/btp458)

Websites

Gramene maintains the following websites:

- http://www.gramene.org/ (public website)
- http://dev.gramene.org/ (development website open to the public)
- http://news.gramene.org/ (news blog)
- http://docs.gramene.org/ (documentation Wiki)
- http://www.plantontology.org/
- http://www.gramene.org/gramenedas/
- http://www.maizegenetics.net/ (TASSEL software)
- http://gmod.org/wiki/CMap
- http://search.cpan.org/dist/Text-RecordParser/
- http://search.cpan.org/dist/SQL-Translator/
- http://search.cpan.org/dist/Bio-GenBankParser/