Gramene: A Platform for Comparative Plant Genomics DBI-9723510

Ed Buckler[1,2], Terry Casstevens[3], Charles Chen[3], Genevieve DeClerck[3], Palitha Dharmawardhana[4], Pankaj Jaiswal[4], AS Karthikeyan[3], Marcela Monaco[5], Susan R McCouch[3], Will Spooner[5], Joshua C. Stein[5], Jim Thomason[5], Sharon Wei[5], Shiran Pasternak[5], Ken Youens-Clark[5], Doreen Ware [2,5]

[1]Institute for Genomic Diversity, Cornell University, Ithaca, NY, 14853, USA; [2]USDA ARS NAA Robert W. Holley Center for Agriculture and Health, Cornell University, Ithaca, NY, 14853, USA; [3]Department of Plant Breeding and Genetics, 240 Emerson Hall, Cornell University, Ithaca, NY, 14853, USA ; [4]Dept of Botany and Plant Pathology, 3082 Cordley Hall, Oregon State University, Corvallis, OR, 97331, USA; [5]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA

In the last year, Gramene released its 32nd (Nov '10) and 33rd (Apr '11) builds since 2000. Gramene's three research aims are to 1) use comparative genomics to identify functional elements and sequence variants that may have phenotypic consequences; 2) annotate biological pathways in order to provide the infrastructure to understand how those sequence variants lead to phenotypes; and 3) to collect and uniformly reanalyze QTL and diversity data in order to connect genetic diversity to phenotypic variation.

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.*
In each release of Gramene, the Ensembl genome browser is updated to the latest release version (v60 and v62 in builds 32 and 33, respectively). This year, Gramene added four complete genomes (*Physcomitrella patens*, *Oryza nivara*, *O. rufipogon*, *O. glaberrima*) to host a total of 12 complete and 8 partial plant genomes in collaboration with EnsemblPlants project at EBI. Gramene initiated a new analysis using Genomic Evolutionary Rate Profiling (GERP) to identify genomic regions that exhibit nucleotide substitution deficits which reflect past purifying selection and are used to rank and characterize constrained elements. Our method involves 4- and 8-way EPO alignments as input with varying parameters including an input tree generated from 1,301 ortholog sets. Figure 1 is a view of the GERP analysis shown in the Ensembl genome browser for *O. sativa japonica*. A version of Ensembl introduced in the last year now includes the ability to view SNPs in their genomic context (Figure 2). Gramene has worked to integrate into Ensembl variation data for rice, maize, *Arabidopsis*, and grape. Gramene constructs gene trees using the longest protein for every gene in Ensembl. Homologues are deduced from these trees, proteins are clustered based on best-reciprocal hits and BLAST score ratios, and each cluster of proteins is aligned using the multiple alignment program, Muscle. Finally, TreeBeST is used to produce a single gene tree from each of the multiple alignments, reconciling it with the species tree to indicate duplication events. The EnsemblCompara GeneTree database has been rebuilt using updated genomes for *A. thaliana*, *O. sativa japonica*, and *P. patens*. There are a total of 35,182 individual trees and 399,113 genes. In the area of functional genomics, Gramene has added or updated arrays for four species (*Brachypodium*, *A. thaliana*, *O. sativa indica* and *japonica*). An automated pipeline uses Compara orthologs to find collinear mappings (via DAGchainer) to find syntenic blocks. In build 32, we added dicots our existing monocot analysis and an analysis to automatically detect split genes Figures 3 and 4 show a detected split gene and genomic alignments confirming the split. Gramene performs creates BlastZ-net whole genome alignments for 24 closely related pairs of species, and Gramene's EPO (Enredo, Pecan, Ortheus) pipeline is a three-step analysis for whole-genome multiple alignments. In release 32, Gramene created an 8-way whole-genome multiple alignment of *Brachypodium*, sorghum, *A. thaliana*, *A. lyrata*, grape, poplar, and *O. sativa japonica* and *indica*,

Specific Aim #2: *Enhance the value of the comparative maps with pathway, phenotypic and other functional information from rice, maize, and Arabidopsis.*
Beta versions of BrachyCyc and MaizeCyc metabolic pathway databases were developed and publicly released in Gramene release 32. MaizeCyc was built by Gramene developers and the maize model organism database MaizeGDB in collaboration with the Maize Genome Sequencing Project (MGSC). It was released simultaneously in collaboration with the MaizeGDB project, and was upgraded to official release status in an interim release after build 33. Pathways and genes presented in this catalog are based on the electronic and manual annotations of the maize B73 RefGen_v2 gene models. RiceCyc was updated to version 3.1, SorghumCyc to version 1.1, MaizeCyc to version 2.0, and, together with all mirrored pathway databases, were upgraded to Ptools software version 15 (provided by the SRI International). Rice pathways curation included the addition of 80 transport reactions and 477 transporters, addition of hydroxycinnamic acid and serotonin biosynthetic pathways, and updates to auxin biosynthesis, tryptophan biosynthesis, ethylene biosynthesis and abscisic acid biosynthesis.

As there are limitations on extending the current metabolic pathway databases to accommodate the regulatory and signaling pathways and reactions, it was decided to use the tools developed by Reactome -- built for human and metazoan communities and funded by NIH -- as a model to develop a Plant Reactome portal. A training meeting of various curators from rice, maize and *Arabidopsis* was held at CSHL in October 2010, and progress has been made in the implementation of Reactome portals for both rice and *Arabidopsis*. Figure 5 shows a model view from the Rice Reactome project. Figure 5 shows the gene ontology mapping and functional annotation of strawberry genes. Figure 6 is a Venn diagram showing unique and shared gene families between and among rice, grape, Arabidopsis and strawberry.

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.*
Data curation activities in the genetic diversity section for the last year have focused on large diversity and genome-wide association studies in maize, rice and *Arabidopsis* being produced by the NSF Maize Diversity Project (www.panzea.org), the NSF Rice Diversity Project (www.ricediversity.org), and the NSF Arabidopsis 2010 project, among others. We also have concentrated on curation of single-gene rice phenotype studies in order to capture the natural variation among the accessions of both wild type and cultivated species of *Oryza* with a focus on domestication related traits such as grain quality, yield, flowering time and disease resistance. An emphasis was placed on curating studies that employ large germplasm panels in their analyses. In the previous year, we started making the shift to supporting genotypic datasets with billions of data points. We have finished schema and analysis tool support for these massive datasets. However, the technology keeps moving extremely rapidly and, in collaboration with the Maize Diversity Project and iPlant, we are continuing to explore and develop community approaches for storing, accessing, and analyzing trillion data point sets. Gramene's SNP Query tool was developed last year to retrieve and filter SNP data by chromosome or cultivar subgroups and has been improved this year to display genes and QTLs overlapping with SNPs of interest as well as to search by a gene, QTL, or trait name or ID. In build 33, we added a new phenotype study web interface that displays phenotype measurement data and all relevant ontological information. The TASSEL desktop analysis tool released it's 3rd major version that added a host of improvements including an alignment viewer (see SNP display in figure 7), progress monitoring, user-friendly error messaging, memory/speed profiling, export/import functions, graphical data plotting (see QQ Plot and Manhattan Plot in figures 8 and 9), union/intersect functions, a table report viewer, alignment site scores, imputation/reference alignment masks (see color coding in figure 7), and a command-line interface. Research in the past year includes the development of a statistical-analysis pipeline for functional gene implication using genome-wide associations, multiple species Compara comparisons, and curated pathway descriptions, especially for flowering time. The results of such analyses allow for systematic summarization of SNP/trait associations and priori candidate genes involved in the regulation of complex trait variation.

Specific Aim #4: *The Plant Ontology (PO) Two Years Only*
Gramene's specific aim #4 has been moved to (NSF award number 0822201).

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 such as Plant Gene Wiki and Pathways; 2) traditional "push" presentations at meetings, staffed by the co-PIs and senior curators; 3) virtual and physical tutorials sponsored by a commercial partner, OpenHelix, targeted at students in under-served institutions and underrepresented minorities. In support of these goals, Gramene members have presented posters, talks and workshops and several major meetings and have generated brief (<3 minute), targeted video tutorials in addition to the OpenHelix tutorials which became available in March, 2010. Our website is visited on average by over 120K unique page views/month from almost 200 countries (primarily the US, China, and India). Members of Gramene have presented posters and workshops at 18 conferences or community meetings in the US, Europe, and Asia. Gramene maintains a news web log at news.gramene.org and a Facebook page.

The project is currently supported by NSF Plant Genome Research Resource grant award #0703908 (Gramene: A Platform for Comparative Plant Genomics), 0723510 (Collaborative Research: An Arabidopsis Polymorphism Database), 0701916 (Physical Mapping of the Wheat D Genome), NSF award #0851652 (REU Bioinformatics and Computational Biology Summer Undergraduate Program) and USDA ARS. We would like to thank our collaborators and contributors who have supplied Gramene with data in the last two years, specifically Ensembl Genomes, MaizeGDB, JGI, Graingenes, TAIR, NASC, IRRI, NCBI, 0743804 (POPcorn), 0543441 (NextGen PLEXdb), 0638820 (The evolutionary genomics of invasive weedy rice), High Density Scoreable Markers for Maize Trait Dissection (High Density Scoreable Markers for Maize Trait Dissection) 0638820 (OMAP), the USDA-ARS CRIS 9235-21000-013-00D (Complete Switchgrass Genetic Maps Reveal Subgenome Collinearity, Preferential Pairing and Multilocus), USDA-ARS Genetic Trait Index, as well as the Arabidopsis 2010 project.

Figure 1

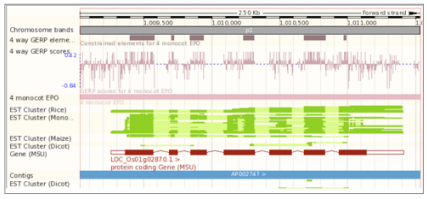


Figure 2

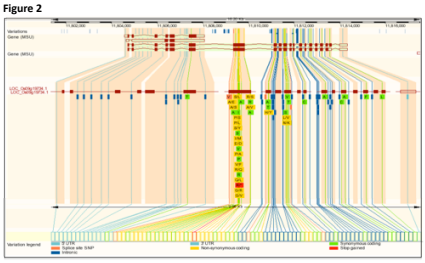


Figure 3

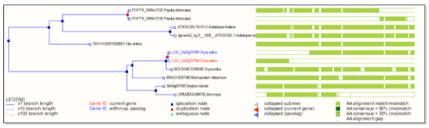
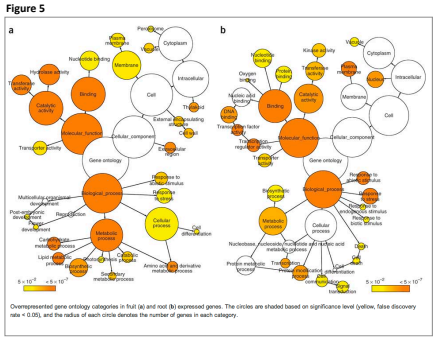


Figure 4

Figure 5



Overrepresented gene ontology categories in fruit (a) and root (b) expressed genes. The circles are shaded based on significance level (yellow, false discovery rate < 0.05), and the radius of each circle denotes the number of genes in each category.

Figure 6

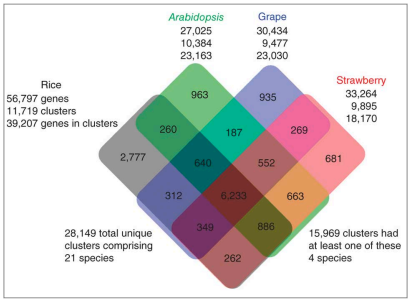


Figure 7

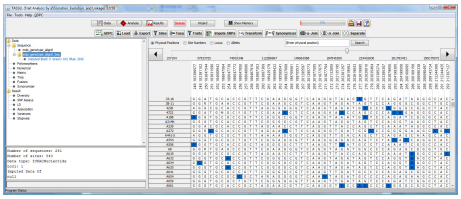


Figure 8

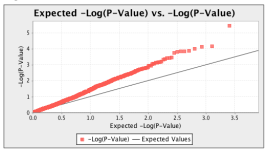
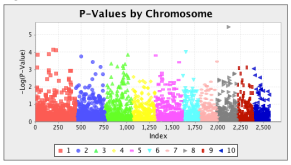


Figure 9



Publications

• **Gramene database in 2010: updates and extensions.** Youens-Clark K, Buckler E, Casstevens T, Chen C, Declerck G, Derwent P, Dharmawardhana P, Jaiswal P, Kersey P, Karthikeyan AS, Lu J, McCouch SR, Ren L, Spooner W, Stein JC, Thomason J, Wei S, Ware D. (Nucleic Acids Res. 2010 Nov 13)

• **Rice structural variation: a comparative analysis of structural variation between rice and three of its closest relatives in the genus Oryza.** Hurwitz, B.L., Kudrna, D., Yu, Y., Sebastian, A., Zuccolo, A., Jackson, S.A., Ware, D., Wing, R.A., and Stein, L. (The Plant Journal doi: 10.1111/j.1365-313X.2010.04293.x)

• **Applications and methods utilizing the Simple Semantic Web Architecture and Protocol (SSWAP) for bioinformatics resource discovery and disparate data and service integration.** Nelson RT, Avraham S, Shoemaker RC, May GD, Ware D, Gessler DD (BioData Min. 2010 Jun 4;3(1):3)

• **Genetic structure and domestication history of the grape.** Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, Prins B, Reynolds A, Chia JM, Ware D, Bustamante CD, Buckler ES (Proc Natl Acad Sci U S A. 2011 Jan 18)

• **Evidence for Network Evolution in an Arabidopsis Interactome Map.** Arabidopsis Interactome Mapping Consortium (Science. 2011)

• **The genome of woodland strawberry (Fragaria vesca).** Shulaev D et al. (Nature Genetics 43, 109–116 (2011) doi: 10.1038/ng.740).