Published online 4 November 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D947–D953***
doi:10.1093/nar/gkm968

Gramene: a growing plant comparative genomics resource

Chengzhi Liang[1], Pankaj Jaiswal[2], Claire Hebbard[2], Shuly Avraham[1], Edward S. Buckler[3,4], Terry Casstevens[3], Bonnie Hurwitz[1], Susan McCouch[2], Junjian Ni[2], Anuradha Pujar[2], Dean Ravenscroft[2], Liya Ren[1], William Spooner[1], Isaak Tecle[2], Jim Thomason[1], Chih-wei Tung[2], Xuehong Wei[1], Immanuel Yap[2], Ken Youens-Clark[1], Doreen Ware[1,4] and Lincoln Stein[1,*]

[1]Cold Spring Harbor Laboratory, 1 Bungtown Rd, Cold Spring Harbor, NY 11724, [2]Department of Plant Breeding and Genetics, 240 Emerson Hall, [3]Institute for Genomic Diversity and [4]USDA-ARS NAA Plant, Soil & Nutrition Laboratory Research Unit, Cornell University, Ithaca, NY 14853, USA

Received September 20, 2007; Revised October 16, 2007; Accepted October 17, 2007

ABSTRACT

Gramene (www.gramene.org) is a curated resource for genetic, genomic and comparative genomics data for the major crop species, including rice, maize, wheat and many other plant (mainly grass) species. Gramene is an open-source project. All data and software are freely downloadable through the ftp site (ftp.gramene.org/pub/gramene) and available for use without restriction. Gramene's core data types include genome assembly and annotations, other DNA/mRNA sequences, genetic and physical maps/markers, genes, quantitative trait loci (QTLs), proteins, ontologies, literature and comparative mappings. Since our last NAR publication 2 years ago, we have updated these data types to include new datasets and new connections among them. Completely new features include rice pathways for functional annotation of rice genes; genetic diversity data from rice, maize and wheat to show genetic variations among different germplasms; large-scale genome comparisons among *Oryza sativa* and its wild relatives for evolutionary studies; and the creation of orthologous gene sets and phylogenetic trees among rice, *Arabidopsis thaliana*, maize, poplar and several animal species (for reference purpose). We have significantly improved the web interface in order to provide a more user-friendly browsing experience, including a dropdown navigation menu system, unified web page for markers, genes, QTLs and proteins, and enhanced quick search functions.

INTRODUCTION

Gramene is a clade-oriented database (1) that provides comparative genomic data for plants. As the first completely sequenced crop genome, rice continues to be the best-annotated genome for monocots and offers a wealth of information on the structure and function of genes, polymorphisms and other functional elements anchored to the genome. Other crop plants are in the process of being sequenced, but have not yet reached the level of completeness offered by rice. Our strategy is to use the rice genome assembly from The Institute for Genomic Research (TIGR) [~370 Mb, (2)] as a framework to order and orient unsequenced and partially sequenced crop genomes based upon their synteny to rice, thereby aiding researchers to discover candidate genes in other crops and to develop genetic and physical marker resources. To fully utilize the existing genome annotations from non-grass species, Gramene also provides genome-level comparisons between rice and *Arabidopsis thaliana*, a key model dicot plant. The genes from *A. thaliana* are used to derive putative orthologs among rice genes, thus facilitating the identification of gene function in grasses.

Figure 1 shows how Gramene is organized from the end-user's point of view. Gramene annotates data using vertical information association (e.g. through controlled vocabularies) as well as horizontal comparative studies among species, and provides users with easy access to the integrated data through a user-friendly web interface. Gramene provides several ways to access data. Data searches can be conducted through either the general search, which is found on every page, or through data type specific search functions, which are located on the module home pages. Other search tools include Gramene Blast, for sequence similarity, and GrameneMart, a data-mining tool for searching data from several datasets [such as

*To whom correspondence should be addressed. Tel: 516 367 8380; Fax: 516 367 8389; Email: lstein@cshl.edu

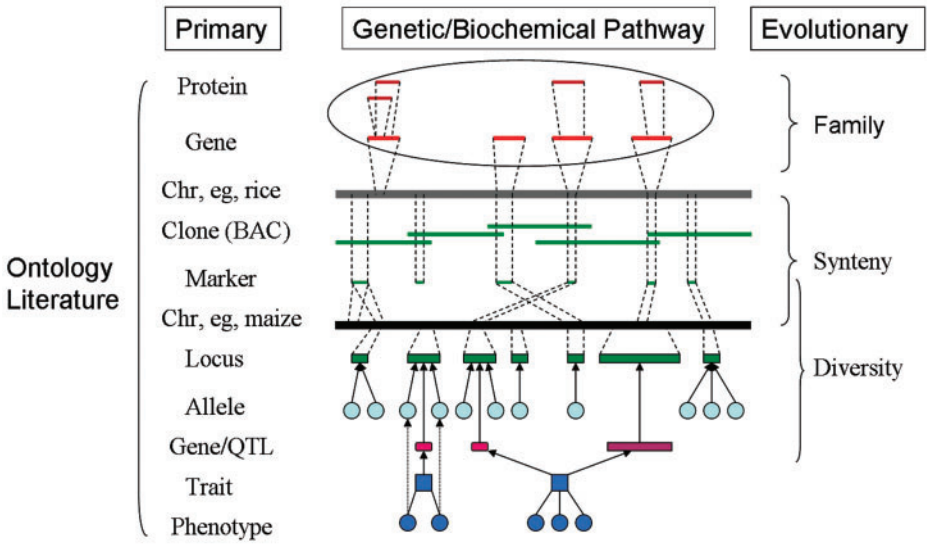

Figure 1. Conceptual data organization and relationship in Gramene. Gramene contains primary data, such as genome sequences, maps, genes, QTLs, proteins and markers, as well as complex data, such as comparative mappings, gene families, genetic diversities and pathways. Ontology terms are used to provide annotations and associations for other data. Gramene also provides literature for the data source. Please note the size of each line segment is not in proportional to the real size of any biological objects it represents.

gene annotations and rice single-nucleotide polymorphisms (SNPs)] with a comprehensive set of predefined user-selectable filters. Gramene data and tools are freely available for local installation and customization.

Gramene continually grows to add new data types and genomes, as well as to improve annotations and website access. Currently on release 26, many improvements have occurred since release 18, which was discussed here 2 years ago. In this article, we will focus on only the major changes made in the past 2 years. For more general Gramene information, please refer either to previous publications (3,4) or the Gramene website (www.gramene.org). New users are encouraged to refer to the Gramene help section (http://www.gramene.org/db/help), which includes newly added tutorials for each module.

PATHWAYS

The Gramene Pathways module (http://www.gramene.org/pathway) is a major new addition for annotating biochemical (metabolic) pathways; it allows pathway-informed functional annotation of genes. Gramene Pathways is based on Pathway Tools, a bioinformatics software system for creating Pathway/Genome Databases (5). RiceCyc, the rice pathway database, is the first curated pathway database for a crop species. RiceCyc was created automatically by use of protein orthology-based pathway inference software incorporated within the Pathway Tools, using gene annotations on the *Oryza sativa ssp. japonica* genome (2). After the initial round of automated pathway construction, we manually curated the database to add species-specific pathways, modify or remove spurious pathways or components unlikely to occur in rice. RiceCyc in Gramene release 26 contains 282 pathways, 1529 reactions, 43 172 polypeptides (10 387 enzymes) and 1162 compounds. For reference purposes, Gramene also imported pathways from AraCyc (6) for *A. thaliana*, EcoCyc (7) for *Escherichia coli*, SolCyc (http://www.sgn.cornell.edu/tools/solcyc) for Pepper (*Capsicum anuum*), Coffee (*Coffea canephora*), Tomato (*Solanum lycopersicum*) and Potato (*Solanum tuberosum*), and MedicCyc (http://www.noble.org/medic cyc) for *Medicago truncatula*.

RiceCyc is integrated with genomic and map-based resources at Gramene via reciprocal links in the rice genome browser. For example, one can traverse from the page describing the gene *sd1* (a rice semidwarf gene) (http://tinyurl.com/2lhfpy—for more readability, all Gramene website examples used in this article are linked through tiny URLs; see www.tinyurl.com for more information) to all the reactions known to involve the *sd1* gene product in RiceCyc. On this page, clicking 'Gramene_Pathway: RXN1F-169' will open a reaction page (http://tinyurl.com/2s7fr2). Additionally, a link from the reaction page under 'In Pathway: gibberellin biosynthesis' leads to the pathway visualization (http://tinyurl.com/2q7rb8) and offers information about the gene functions at a higher biochemical level. Pathways can be viewed at five different levels of detail, from an overview to detailed enzyme, reaction and compound information, by clicking the button 'More Detail' or 'Less Detail'. On pathway or reaction pages, a button 'Cross-Species Comparison' is provided to access to comparisons between the pathways or reactions from the species mentioned above. Finally, users may upload their own gene expression data, using the Omics Viewer, to visualize an overview of the cellular level expression profile.

Diversity

Gramene's second new module is the Genetic Diversity database (http://www.gramene.org/db/diversity/diversity_view). Genetic variations are the main source of useful

traits for plant breeding. For example, many low-yield varieties contain alleles associated with favorable traits, such as disease resistance or drought resistance. These favorable traits can be transferred to the high-yield varieties through selective breeding. Gramene's diversity data come from three species: rice, maize and wheat. The data include: 954 germplasm accessions, 290 SSR (simple sequence repeat), 291 restriction fragment length polymorphism (RFLP) and 125 amplified fragment length polymorphism (AFLP) markers, as well as phenotype data on 21 traits for rice; 48 germplasm accessions and 3802 SNP markers for wheat; 1993 germplasm accessions, 1435 SNP and 520 SSR markers for maize. We curated the rice data from published data sources such as Thomson *et al*. (8), and imported the maize data from Panzea (9), an allied project. Wheat diversity data were imported from a single large study 'Haplotype Polymorphism in Polyploid Wheats and Their Diploid Ancestors' (http://wheat.pw.usda.gov/SNP/project.html).

The Diversity database uses the Genomic Diversity and Phenotype Data Model (GDPDM, http://www.maizegenetics.net/gdpdm) database schema as its data-storing engine. This schema is a comprehensive model for holding quantitative phenotypic data and molecular genotypic data. The Gramene Diversity module search allows finding all the allelic variations on a locus of multiple germplasms or the genome-wide heterozygotic alleles of a single germplasm. The search functionality is organized into three levels of complexity. The entry level is directed at the novice user, who can do a quick search on a marker or germplasm names. The intermediate level uses the Panzea search interface (9), which provides many more search options such as accession number or phenotype. For advanced users, the diversity module provides the comprehensive Java-based Advanced Search Tool, Genomic Diversity and Phenotype Connection (GDPC) (http://www.maizegenetics.net/gdpc). With GDPC, complex queries can be executed, data can be integrated from multiple sources, and data can be saved in various formats.

The markers used in the diversity database are cross-linked to the Markers database, which provides all the detailed information about each marker. For example, from the Diversity module home page a search for the rice marker RM22 (http://tinyurl.com/2m8hck) shows the experiments involving this marker. Above the results table is a link connecting to the marker detail page on http://tinyurl.com/2vsrn2.

MARKER AND MAP UPDATES

Since the Markers and Maps modules are two separate but closely related modules, they are discussed together in this section. The Markers database currently has two roles, one as a central repository for all mappable entities including polymorphic markers and DNA sequences on genetic, physical and sequence maps; and the other as a map feature source for CMap, the comparative map viewer. Previously, only polymorphic markers were available through the Markers module's search interface, but it now encompasses any assay that defines a discrete location in the genome. In conjunction with the expansion of its role, the Markers module has undergone intensive new development to provide powerful search functionality.

In Gramene release 26, the Markers database contained data for 13 504 402 polymorphic markers and sequences from barley, maize, oat, rice, rye, sorghum, sugarcane, wheat and many other species. The data types include genes, quantitative trait loci (QTLs), genetic and physical markers, GenBank Genome Survey Sequences (GSS), cDNA, expressed sequence tag (EST) and other mRNAs, and many unsequenced clones mapped to fingerprint contig (FPC) maps. Each marker or sequence is mapped to its original map, as well as to the sequence maps wherever possible, e.g. the SSR marker RM22 mentioned in the previous section is mapped to many different sequence, genetic and QTL maps, and this common feature provides associations among maps for comparative mapping studies. This marker is associated with many QTLs, which allows for anchoring the QTL to the sequence maps (see the next section on the QTL track in the rice genome browser). All genomic mappings present in the Markers module can be visualized in either CMap or the rice genome browser.

CMap, the comparative mapping tool used in Maps module, allows arbitrary sets of maps to be aligned in order to show the relationships among them. There are currently more than 200 different map sets from 26 species in CMap. The most significant of these are the reference rice genome, now known as the 'Rice Gramene Annotated Nipponbare Sequence' map, as well as a large series of QTL maps. The inclusion of the latter allows researchers to follow a trait mapped via a QTL to the rice genomic sequence, and from there to maps of other species. For example, Figure 2 shows a typical application of CMap in comparative genomic studies around the rice *sd1* gene.

Another significant new data addition to CMap is the inclusion of FPC maps from 12 wild rice (*Oryza*) species OMAP, the Oryza Map Alignment Project www.omap.org). These maps are invaluable for understanding the genome-level changes that have occurred during rice's recent evolutionary history. The comparative maps for the wild rice species (http://www.gramene.org/cmap/omap.html) can be conveniently reached from the Gramene home page Quick Start section under comparative maps. An inversion example is shown between *O. sativa* chromosome 3 and *O. glaberrima* at http://tinyurl.com/2k57hu.

GENOME UPDATES IN THE ENSEMBL BROWSER

The Genomes section (http://www.gramene.org/genome_browser/index.html) contains fully or partially sequenced genomes and their annotations. Currently available genomes include: *O. sativa ssp. japonica* (2) (updated), *O. sativa ssp. indica* (10) (new), *O. rufipogon* (from OMAP project) (new), *A. thaliana* (11) (updated), *Zea mays* (updated) and organelles (mitochondria and chloroplasts) from rice, maize and *A. thaliana* (all new).

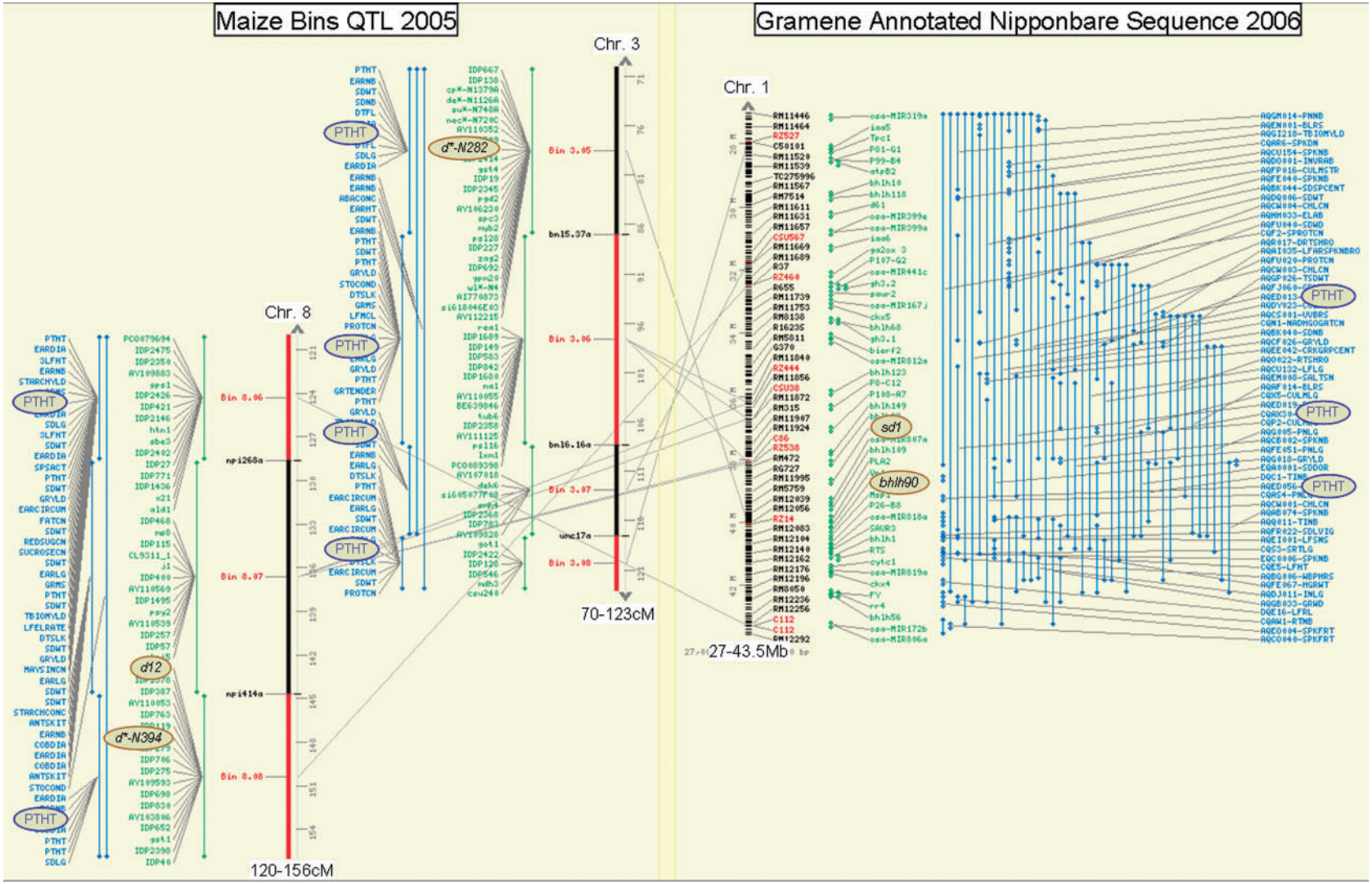

Figure 2. A syntenic region showing related QTLs and genes between rice and maize in CMap. Please refer to CMap tutorial (http://www.gramene.org/tutorials/cmap.html) on how to construct this comparative map view. The maize chromosomes 3 and 8 contain big syntenic regions (including inversions) with rice chromosome 1 as evidenced in Ensembl synteny view http://tinyurl.com/2sl55k. The QTLs are shown in blue and genes in green. The rice markers are shown on the right side of the chromosome, where the markers in red are those also mapped on the maize map. There are many plant height QTLs in rice and maize mapped in this region (PTHT, in blue ovate—the names were enlarged for better viewing). The rice PTHT QTLs are probably contributed by the *sd1* gene (McCouch *et al.*, unpublished data), but the maize PTHT QTLs have not been characterized regarding the underlying genes. Note that several maize height genes, *d12*, *d*-N394* and *d*-N282* (in brown ovate) are also mapped in the region. Based on the syntenic view, we can expect that the *sd1* orthologs in maize will be able to be identified after the maize genome sequence is completed, and they will likely contribute to some of these maize PTHT QTLs.

The *A. thaliana* data were integrated into Ensembl system by European Arabidopsis Stock Centre (NASC), whereupon Gramene imported them for reference purposes. The maize genome is available as a link to www.maizesequence.org, an allied project. The maize genome browser will be fully integrated into Gramene after the maize-sequencing project matures at the end of 2008.

Genes, homologous genes and gene trees

There are now seven gene tracks in the rice genome browser. The TIGR rice gene annotations were updated to version 5, which contains many revised genes with improved structure data. There are two new tracks for non-coding rice genes including 185 siRNAs genes (12) and 230 miRNA genes (13). These two tracks provide information about currently available important regulatory non-coding genes in rice. Three new protein-coding rice gene tracks were also added. This includes highly confident Gramene annotated genes based on cDNA, EST and protein evidences (Liang *et al.*, unpublished data), genes based on The Rice Annotation Project (RAP) (http://rapdb.dna.affrc.go.jp) annotations (14), and the consensus genes (sharing at least one translation) among TIGR, Gramene and RAP genes. For TIGR and Gramene protein-coding genes, each protein is annotated with InterPro domains (15).

A gene tree that shows orthologs and paralogs of each gene (if available) is a new addition, and can be reached through the link on the left-side bar on the gene detailed page (see a gene tree example in Figure 3). The gene trees were built using Ensembl gene tree pipeline where maximum-likelihood trees were reconciled with their species tree to infer gene orthology and paralogy relations (16). For Gramene release 26, the gene trees were computed for four plant genomes: *O. sativa ssp. japonica*, *O. sativa ssp. indica*, *A. thaliana* and poplar, as well as four model metazoan species *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae*. For example, from 41 908 genes of *O. sativa ssp. japonica*, we identified the following number of orthologs: 18 520 in *A. thaliana*, 19 684 in poplar, 5324 in *S. cerevisiae*, 7175 in *C. elegans*, 7449 in *D. melanogaster* and 8313 in *Homo sapiens*. The homologous gene set and gene tree can significantly add to users' understanding of the rice gene functions, since many genes from *A. thaliana* and the animal species are better annotated than their homologs in rice.

QTL track

A rice QTL track has been added to the rice genome browser. Currently the track contains 6293 rice QTLs. These QTLs are anchored with their directly associated markers mapped on the genome. For these QTLs, users can find reciprocal links to the QTL detailed information page. This allows users to access a specific genome region corresponding to a QTL or to browse QTLs by searching for traits on the rice genome browser. One option for using these data is to download all or part of the annotated genes covered by a QTL through GrameneMart (http://www.gramene.org/Multi/martview) and do further studies on these genes, potentially discovering the genes contributing to the QTL.

SNP track

Approximately 4 million *O. sativa* SNPs from GenBank (http://www.ncbi.nlm.nih.gov) dbSNP version 125 have been mapped to the rice genome assembly. The consequence of each SNP (synonymous, non-synonymous, untranslated region (UTR), etc.) on affected gene transcripts has been annotated. Features in the SNP track display are color coded, according to their consequence, on the overlapping gene models. This makes it easy to find, for example, which SNPs affect the peptide sequence of the gene (i.e. non-synonymous SNPs).

Repeat tracks

Repeat features in the rice and *A. thaliana* genomes were identified using RepeatMasker and the REdat repeat library (http://mips.gsf.de/proj/plant/webapp/recat)

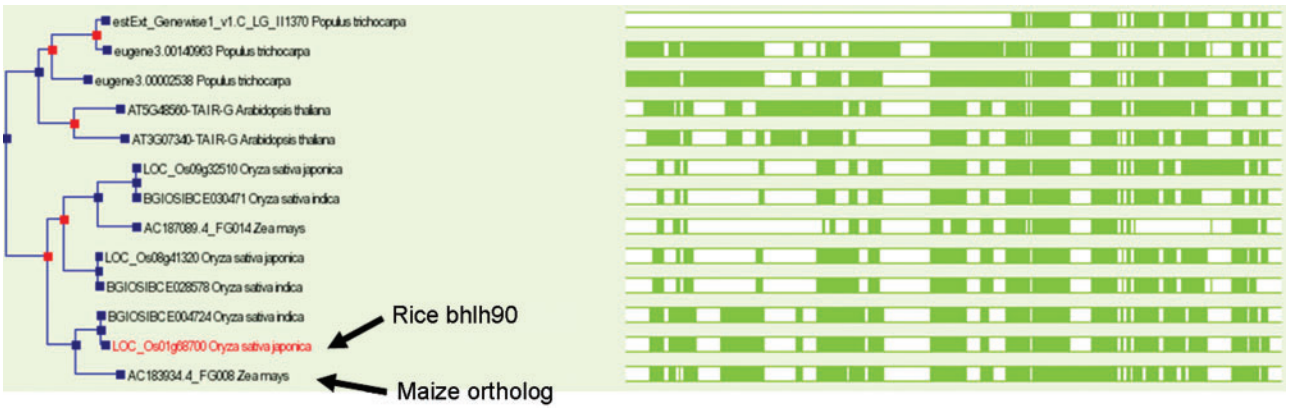

Figure 3. The rice *bhlh90* gene tree (partial) showing the orthologous genes between rice, maize, *A. thaliana* and poplar. The full tree can be viewed at http://tinyurl.com/3clrjv. Note the *bhlh90* gene is mapped in the region shown in Figure 2, but its maize gene ortholog is a predicted gene and has not been studied functionally. The solid box on the right shows the aligned regions in the proteins. It is expected that there are more maize orthologs to be identified after the maize genome sequence becomes complete.

developed by the Munich Information Center for Protein Sequences (MIPS). These data have been used to create several new genome browser tracks under the 'Repeats' menu: All repeats, LTRs, RNA repeats, Satellite repeats, Type I Transposons, Type II Transposons and Other/Unknown repeats. With the repeat tracks, users can easily identify the genes overlapping in the repeat regions, and the repeat category the gene belongs to.

OTHER DATA UPDATES

In addition to the above-mentioned new or updated modules, there have been several other updates in Gramene: Genes, QTL, Proteins, Ontologies, Literature, BLAST and GrameneMart. We now discuss the major updates.

Genes

The Gramene Genes module is a collection of curated genes. It includes descriptions of genes and alleles associated with morphological, developmental and agronomically important phenotypes, variants of physiological characters, biochemical functions and isozymes. Previously only rice genes (1488) were available in this module. In release 26, there are 11 747 genes from rice (2651), maize (6680), oat (353), wheat (598), barley (1182) and other species (283). The newly added maize genes were imported from MaizeGDB (17), and the other genes were from GrainGenes (http://wheat.pw.usda.gov/ggpages/index.shtml) to enhance comparative studies between rice and the other species. The gene types include pseudogene (11), rRNA (14), tRNA (107), miRNA (246), siRNA (284), CDS (Protein coding) (604), Not sequenced (1445) and Not classified (9036). There are 5250 genes mapped to the maps stored in the Markers/Maps databases discussed above. The comparative tools in CMap can help further characterize the mapped unsequenced or unclassified genes (see Figure 2 for a potential example).

QTL

The Gramene QTL module provides agronomically important data for geneticists and plant breeders. It now contains a total of 11 536 QTLs (from 339 traits) identified in rice, maize, barley, oat, wheat, sorghum, pearl millet, foxtail millet and wild rice. These QTLs are curated from more than 300 papers. In addition to the increase from the last report (QTLs up from 8410, traits up from 274), there are important improvements in the way that QTLs are presented.

One improvement is the provision of QTL-associated marker information: co-localized markers are those that overlap a QTL region on the original QTL map, and neighboring markers are those that do not overlap QTL regions but are adjacent to them. Using the associated markers, we have inferred positions of 6293 rice QTLs on the rice sequence map. These genome positions are displayed on both the Gramene Maps and the genome browser (as discussed above on CMap and the QTL track in the rice genome browser) and will greatly enhance users' ability to perform comparative mapping research and candidate gene discovery (Figure 2).

Another improvement is that all QTLs are now described and annotated by multiple ontologies to provide associations with known genes and proteins. In addition to the Trait Ontology (TO) terms used in previous releases, Plant Ontology (PO) terms have been added to describe Plant Structure and Plant Growth and Development Stage associated with a particular QTL. Those terms have been annotated according to the information from QTL experiments in the cited literature, or by creating a default mapping based on the expertly identified TO and PO associations.

Ontologies

The Gramene Ontologies module (http://www.gramene.org/plant_ontology) includes annotation and association for Gramene data with six different types of ontology terms: Gene (GO), Plant structure (PO), Growth stage (GRO), Trait (TO), Environment (EO) and Taxonomy (GR_tax). PO is developed and imported from the Plant Ontology database (18) (POC, www.plantontology.org) and GO is from the Gene Ontology database (19) (GOC, www.geneontology.org), whereas GRO, TO, EO and GR_tax are built in-house. Ontologies and their associations to genes, gene models, proteins, QTLs, markers and maps are updated regularly for each Gramene release. The Ontology browser now displays the total number of objects associated with the term name, e.g. the TO term plant height (TO:0000207) has 1599 associations to QTLs (1445), Genes (105) and Proteins (20). The ontology detail page now displays 'External references' and 'Comments' associated with a given ontology term, e.g. NADP-malic enzyme C4 photosynthesis (GO:0009762) has references linking to PubMed (PMID:11788762) and Pathway (MetaCyc:PWY-241).

WEB INTERFACE UPGRADE AND NEW QUICK SEARCH

After release 19 in October 2005, Gramene introduced a dropdown menu navigation format and a new home page. This simple navigation system is standard on all Gramene web pages except for the Pathways module, which runs off third-party software. The new home page includes quick links to important data sections, the current Gramene release number and date, as well as Gramene tips, news and events outreach calendar. A second tier offers quick links to available species pages. These new features are designed to help users find important information in Gramene quickly. In release 26, Gramene introduced a new unified web page design for genes, QTLs, markers and proteins. The new pages are organized as a series of collapsible sections, which allow users to show or hide just the categories of information most useful to them.

The generic search box at the top of every Gramene page has been upgraded to use a new search algorithm that reduces the search time dramatically while improving

the relevance of the returned results. Users may use it to search the entire site, a single data type or the static web pages and documents. The search results are grouped by module. For more in-depth module searches, users are encouraged to use the module-specific searches accessed from their module home pages (such as maps, markers and genes).

FUTURE DIRECTIONS

As new crop plant genome sequences become available, Gramene will incorporate them into the resource by providing comparative mapping data including whole-genome alignments and pathways. Gramene will also provide an evidence-based gene track for each species with a complete genome sequence, using a standardized gene prediction pipeline that should facilitate comparisons within and among species. Beginning in November 2007, Gramene will start a new grant cycle, and the release schedule will become semiannual rather than quarterly.

ACKNOWLEDGEMENTS

The project is currently supported by National Science Foundation (0321685) and US Department of Agriculture-Agricultural Research Service specific cooperative agreement (58-1907-0-041). We would like to thank all our collaborators and contributors who have supplied Gramene with data, as well as our users for their feedback and support. We thank our Science Advisory Board members Anna M. McClung, Georgia Davis, James H. Oard, David Marshall, Patricia Klein and John Mullet for their critical comments, suggestions and improvements of our work. We also thank Peter Van Buren for his excellent system administration work. Funding to pay the Open Access publication charges for this article was provided by National Science Foundation (0321685).

Conflict of interest statement. None declared.

REFERENCES

1. Bill Beavis,D.G., Rhee,S., Rokhsar,D., Doreen,M., Lukas,M., Huala,E., Lincoln,S. & Lawrence,C. (2005) Plant biology databases: a needs assessment, http://www.gramene.org/resources/plant_databases.pdf.
2. Ouyang,S., Zhu,W., Hamilton,J., Lin,H., Campbell,M., Childs,K., Thibaud-Nissen,F., Malek,R.L., Lee,Y. *et al.* (2007) The TIGR rice genome annotation resource: improvements and new features. *Nucleic Acids Res.*, **35**, D883–D887.
3. Jaiswal,P., Ni,J., Yap,I., Ware,D., Spooner,W., Youens-Clark,K., Ren,L., Liang,C., Zhao,W. *et al.* (2006) Gramene: a bird's eye view of cereal genomes. *Nucleic Acids Res.*, **34**, D717–D723.
4. Ware,D.H., Jaiswal,P., Ni,J., Yap,I.V., Pan,X., Clark,K.Y., Teytelman,L., Schmidt,S.C., Zhao,W. *et al.* (2002) Gramene, a tool for grass genomics. *Plant Physiol.*, **130**, 1606–1613.
5. Karp,P.D., Paley,S. and Romero,P. (2002) The Pathway Tools software. *Bioinformatics*, **18**(Suppl. 1), S225–S232.
6. Mueller,L.A., Zhang,P. and Rhee,S.Y. (2003) AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiol.*, **132**, 453–460.
7. Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta-Gil,M. and Karp,P.D. (2005) EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res.*, **33**, D334–D337.
8. Thomson,M.J., Septiningsih,E.M., Suwardjo,F., Santoso,T.J., Silitonga,T.S. and McCouch,S.R. (2007) Genetic diversity analysis of traditional and improved Indonesian rice (Oryza sativa L.) germplasm using microsatellite markers. *Theor. Appl. Genet.*, **114**, 559–568.
9. Zhao,W., Canaran,P., Jurkuta,R., Fulton,T., Glaubitz,J., Buckler,E., Doebley,J., Gaut,B., Goodman,M. *et al.* (2006) Panzea: a database and resource for molecular and functional diversity in the maize genome. *Nucleic Acids Res.*, **34**, D752–D757.
10. Zhao,W., Wang,J., He,X., Huang,X., Jiao,Y., Dai,M., Wei,S., Fu,J., Chen,Y. *et al.* (2004) BGI-RIS: an integrated information resource and comparative analysis workbench for rice genomics. *Nucleic Acids Res.*, **32**, D377–D382.
11. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G. *et al.* (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.
12. Sunkar,R., Girke,T. and Zhu,J.K. (2005) Identification and characterization of endogenous small interfering RNAs from rice. *Nucleic Acids Res.*, **33**, 4443–4454.
13. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
14. Ohyanagi,H., Tanaka,T., Sakai,H., Shigemoto,Y., Yamaguchi,K., Habara,T., Fujii,Y., Antonio,B.A., Nagamura,Y. *et al.* (2006) The Rice Annotation Project Database (RAP-DB): hub for Oryza sativa ssp. japonica genome information. *Nucleic Acids Res.*, **34**, D741–D744.
15. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Buillard,V., Cerutti,L. *et al.* (2007) New developments in the InterPro database. *Nucleic Acids Res.*, **35**, D224–D228.
16. Hubbard,T.J., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
17. Lawrence,C.J., Dong,Q., Polacco,M.L., Seigfried,T.E. and Brendel,V. (2004) MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Res.*, **32**, D393–D397.
18. The Plant Ontology Consortium, (2002) The plant ontology consortium and plant ontologies. *Comp. Funct. Genomics*, **3**, 137–142.
19. Gene Ontology Consortium, (2006) The gene ontology (GO) project in 2006. *Nucleic Acids Res.*, **34**, D322–D326.