Open Access Articles

Gramene: a bird's eye view of cereal genomes

The Faculty of Oregon State University has made this article openly available.
Please share how this access benefits you. Your story matters.

Citation	Jaiswal, P., Ni, J., Yap, I., Ware, D., Spooner, W., Youens-Clark, K., ... & McCouch, S. (2006). Gramene: a bird's eye view of cereal genomes. Nucleic Acids Research, 34, D717-D723. doi:10.1093/nar/gkj154
DOI	10.1093/nar/gkj154
Publisher	Oxford University Press
Version	Version of Record
Terms of Use	http://cdss.library.oregonstate.edu/sa-termsofuse

Nucleic Acids Research, 2006, Vol. 34, Database issue ***D717–D723***
doi:10.1093/nar/gkj154

Gramene: a bird's eye view of cereal genomes

Pankaj Jaiswal, Junjian Ni, Immanuel Yap, Doreen Ware[1,3], William Spooner[1], Ken Youens-Clark[1], Liya Ren[1], Chengzhi Liang[1], Wei Zhao[1], Kiran Ratnapu[1], Benjamin Faga[1], Payan Canaran[1], Molly Fogleman, Claire Hebbard, Shuly Avraham[1], Steven Schmidt[1], Terry M. Casstevens[2], Edward S. Buckler[2,3], Lincoln Stein[1] and Susan McCouch*

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

Received September 14, 2005; Revised and Accepted October 31, 2005

ABSTRACT

Rice, maize, sorghum, wheat, barley and the other major crop grasses from the family Poaceae (Gramineae) are mankind's most important source of calories and contribute tens of billions of dollars annually to the world economy (FAO 1999, http://www.fao.org; USDA 1997, http://www.usda.gov). Continued improvement of Poaceae crops is necessary in order to continue to feed an ever-growing world population. However, of the major crop grasses, only rice (*Oryza sativa*), with a compact genome of ~400 Mbp, has been sequenced and annotated. The Gramene database (http://www.gramene.org) takes advantage of the known genetic colinearity (synteny) between rice and the major crop plant genomes to provide maize, sorghum, millet, wheat, oat and barley researchers with the benefits of an annotated genome years before their own species are sequenced. Gramene is a one stop portal for finding curated literature, genetic and genomic datasets related to maps, markers, genes, genomes and quantitative trait loci. The addition of several new tools to Gramene has greatly facilitated the potential for comparative analysis among the grasses and contributes to our understanding of the anatomy, development, environmental responses and the factors influencing agronomic performance of cereal crops. Since the last publication on Gramene database by D. H. Ware, P. Jaiswal, J. Ni, I. V. Yap, X. Pan, K. Y. Clark, L. Teytelman, S. C. Schmidt, W. Zhao, K. Chang *et al.* [(2002), *Plant Physiol.*, 130, 1606–1613], the database has undergone extensive changes that are described in this publication.

INTRODUCTION

Gramene is a comparative genome mapping database for grasses, using the rice genome as an anchor (http://www.gramene.org). Both automatic and manual data curation are performed to combine and inter-relate information on the structure and organization of genomes and genes, functions of proteins, various maps (genetic, physical and sequence maps), mapped markers, quantitative trait loci (QTL) and literature citations (Figure 1A). As an information resource, the purpose of Gramene is to provide added value to the data that are initially contributed by a variety of species-specific databases (1–3) and generic data repositories such as Entrez (4). This facilitates the researchers' ability to leverage the rice genomic sequence and genetic information and to compare and understand the characteristics of genes, genome organization, pathways and phenotypes in the cereals. Since the last publication on Gramene (5), the database has undergone numerous enhancements by the addition of new datasets, tools and search interfaces as described in the following sections. The datasets and modules described here are based on release no.18 (September 2005) of the Gramene database.

MAP SEARCH

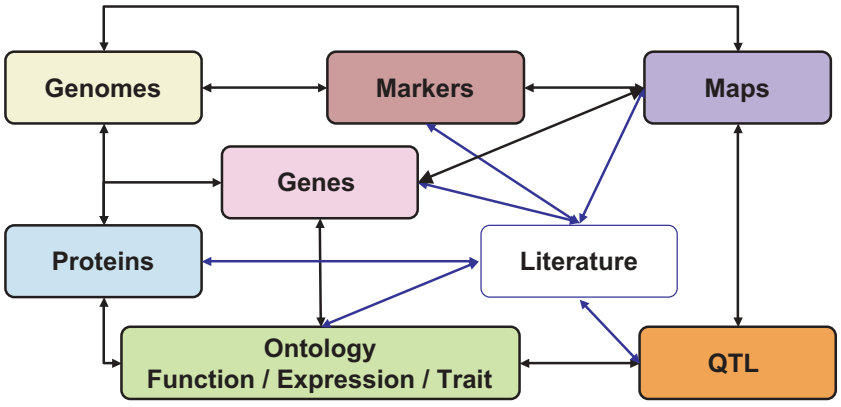
Gramene's central comparative mapping tool is CMap, which can be accessed from the Gramene website in a variety of ways (http://www.gramene.org/cmap/index.html). This tool represents a map as a linear array of interconnected features, which could correspond to a single linkage group in the case of a genetic map, to a single contig for a physical map, or to a

*To whom correspondence should be addressed. Tel: +1 607 255 0420; Fax: +1 607 255 6683; Email: SRM4@cornell.edu

© The Author 2006. Published by Oxford University Press. All rights reserved.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oxfordjournals.org

A



B

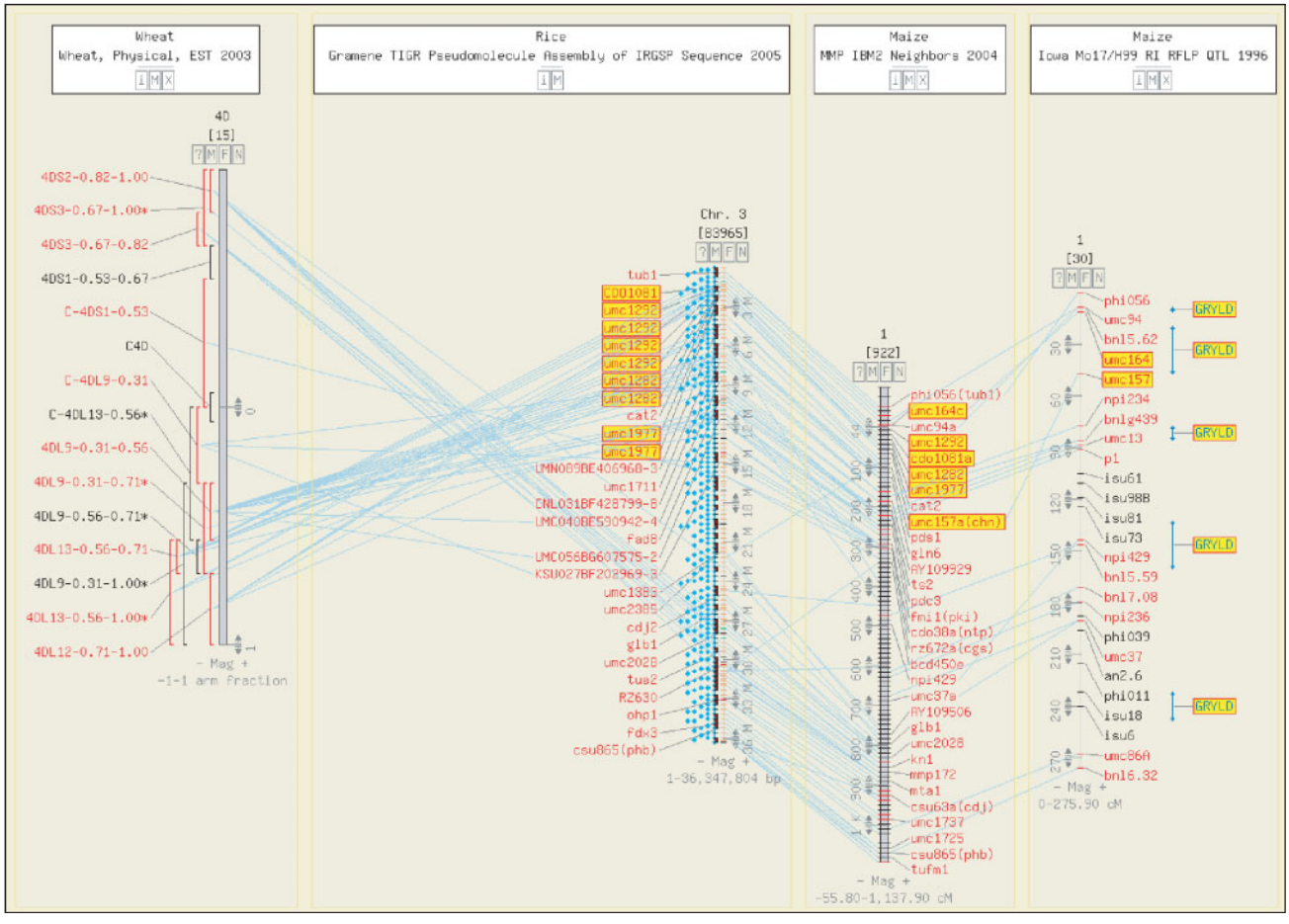


Figure 1. (**A**) A flow diagram showing all the modules and datasets (genomes, maps, markers, QTL, genes, proteins, ontologies and literature) present in the Gramene database. Each module connects or links to the information from one to many other modules as depicted by the arrows. The double headed connections suggest that the modules link in both directions. (**B**) The modified image displayed by the comparative map tool CMap. The CMap tool allows a user to add any number of maps to the left and/or right of the reference map. In this case, the reference map is the sequence map of rice chromosome-3 from the mapset called 'Gramene TIGR pseudomolecule assembly of IRGSP sequence 2005'. In the first display, the wheat deletion map of chromosome-4D from the mapset 'Wheat, Physical, EST 2003 (wEST 2003)' and the maize genetic map of Chromosome-1 from the mapset MMP IBM2 Neighbors 2004 were added to the right and left of the reference map, respectively. The blue colored lines connecting the maps suggest that there is colinearity (conserved synteny) among rice chromosome-3, wheat 4D and maize 1 as demonstrated by the marker order on these maps. Subsequently, a fourth map was added to the display, namely a QTL map of maize chromosome-1 from the map set 'Iowa Mo17/M99 RI RFLP QTL 1996'. This provided a way to investigate if the maize QTL, GRYLD (for the trait Grain yield, i.e. second from top) contains markers that lead to a colinear region of rice. A comparative map from the maize QTL map to the sequence map of rice enables a user to readily identify putative homologs in the conserved region that may be investigated to determine whether they contribute to the same traits in the two species. The markers highlighted in red on the maize and rice maps are shared marker loci on the different maps. Red colored markers in the CMap display indicate that there is a correspondence between those marker loci and markers present in other maps in the display. The displayed label (symbol) for a QTL represents a trait acronym e.g. GRYLD is for the trait Grain yield.

contig or scaffold in the case of an annotated sequence. Related maps are grouped into map sets, such as the set of linkage groups produced by a genetic mapping study, or the set of chromosomes annotated during a sequencing project. To set up a comparison between different map sets from either the same or different species or even from different map types, the researcher first selects a reference map set, and then selects a reference map (chromosome, linkage group or contig) from within the set. This reference map serves as the basis for any comparison that one chooses to make.

Once the reference map image has been rendered, the user is given the option to select one or two comparative maps. These comparative maps may be added to both the left and the right of the previously selected reference map (Figure 1B). The user may keep adding additional maps, as long as valid comparisons are available. Any item that is positioned on a map is called a feature (e.g. a marker, clone, gene, QTL, etc.) and may be either a point or an interval. The researcher can click on a feature to obtain additional information about the feature including internal links within Gramene or to external resources. In a comparative map view, the lines that connect features on one map to those on another map denote correspondences. These correspondences are assigned either automatically (based on a sequence alignment or a feature name) or manually by a curator following curational guidelines.

Currently the CMap module hosts a total of 159 maps characterized into 6 different types, namely sequence, physical, deletion, genetic, Maize bin and QTL maps. These maps belong to 22 different species representing major cereal crops and include 1 833 720 features of 36 different types. Since the last publication (5), the CMap section has undergone major updates by including a greater number of species, map sets and feature types, in addition to new search, display and download options. CMap was developed by Gramene in collaboration with GMOD (http://www.gmod.org) under open source licensing.

MARKER SEARCH

Another recently-enhanced feature is Gramene's marker module (http://www.gramene.org/markers/index.html). This module allows user to search the marker collection by typing any number of marker names in the 'Marker Name' search box, each separated by commas or spaces. To further refine the search query, the researcher may specify a marker type (e.g. 'RFLP') and/or a species (e.g. 'rice') e.g. a search for rice RFLP markers beginning with RZ gives 287 entries (http://www.gramene.org/db/markers/marker_view?marker_name=*RZ*&marker_type_id=3&species_id=1&action=marker_search). New users can familiarize themselves with the marker datasets by visiting the web page at http://www.gramene.org/db/markers/marker_view, where they will find summaries of all the information available in Gramene about a given marker, including marker name, synonyms, type, species and germplasm from which it was derived, maps on which the marker can be found, genome positions on the rice genome sequence and the maize physical map, cross references to the source of the marker, literature citations and if available, images illustrating the length polymorphisms (e.g. rice SSR marker RM220 http://www.gramene.org/db/markers/marker_view?marker_name=rm220).

Currently the marker module contains a total of 1 308 654 genetic and physical markers of 11 types from 17 different species. As a result of ongoing curation activity, ~67% of the features found in the map module are also present in the marker database. Additional curated markers will be added in future releases.

QUANTITATIVE TRAIT LOCI

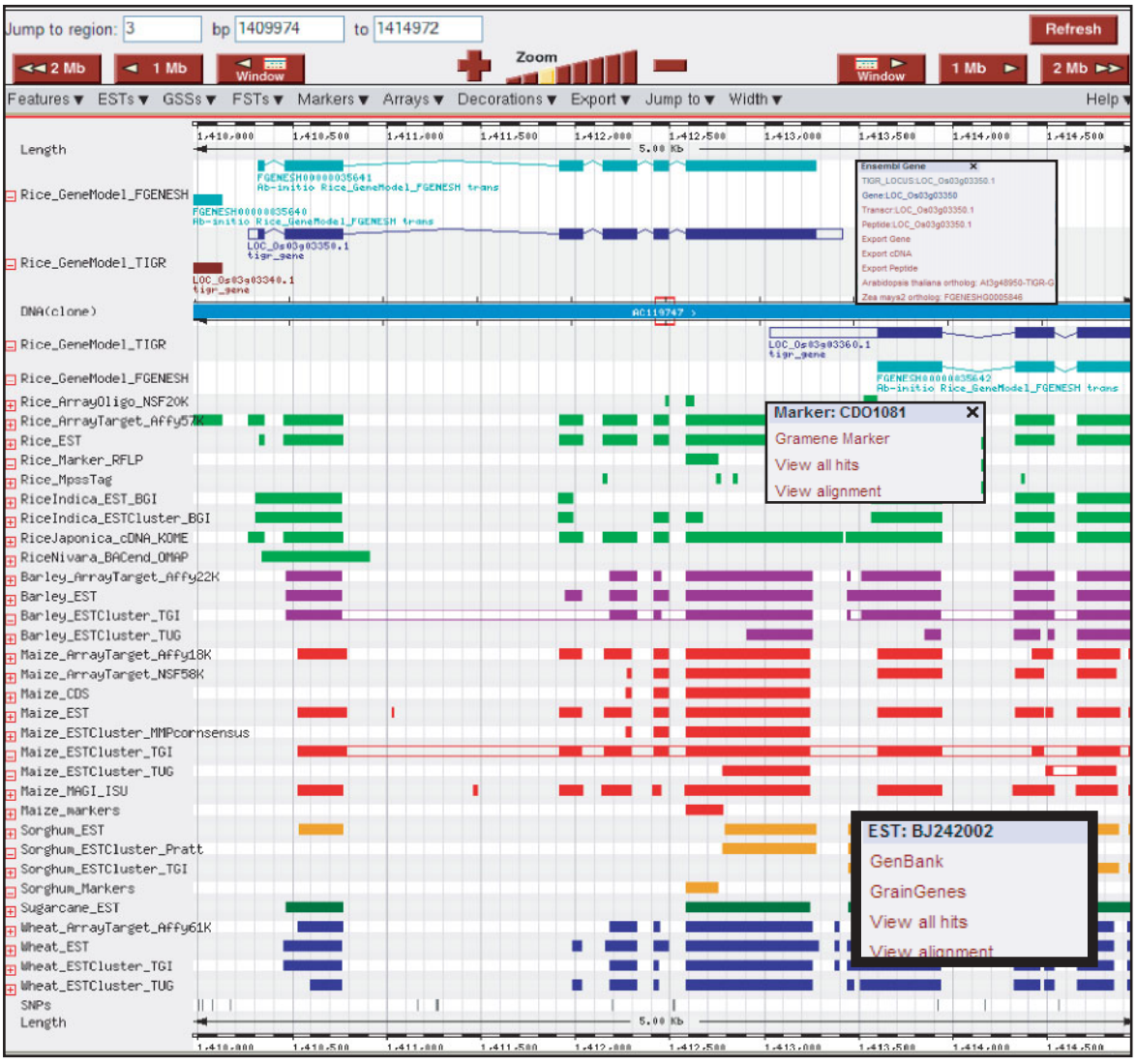
A QTL is a statistical construct that identifies a particular region of the genome as putatively containing one or more genes associated with a trait. A QTL is represented as an interval in a genetic linkage group within which the probability of association is plotted for each marker used in the mapping experiment. The QTL module (http://www.gramene.org/qtl/index.html) is a new addition to the Gramene database. It facilitates the comparative study of QTL across species by providing researchers with tools to investigate colinear regions found to carry genes and QTL contributing to similar traits in different plant species. Gramene does not currently curate raw QTL segregation data, but emphasizes the presentation of basic QTL information such as the trait name, symbol, mapped position on the genetic, cited reference, and comments in free text (e.g. grain length QTL, CQAL1 http://www.gramene.org/db/qtl/qtl_display?qtl_accession_id=CQAL1). The trait descriptions are mapped to a controlled vocabulary called the trait ontology (TO), which allows researchers to relate a trait in one species with a similar trait in another.

Currently the QTL module includes a total of 8410 QTLs identified for 274 agronomic traits from nine major cereal crops including rice, maize, oat, barley, wheat (hexaploid and tetraploid), pearl millet (*Pennisetum glaucum*), foxtail millet (*Setaria italica*) and wild rice (*Zizania palustris*). For convenience of searching, the 274 traits are grouped into eight major trait families related to abiotic and biotic stress, fertility, anatomy, development, vigor, quality and yield. This dataset includes >7000 rice QTLs which have been curated by Gramene staff (J. Ni *et al*., unpublished data) from published literature that dates back to the very first rice QTL study in 1994 (6). QTLs for other species were either curated in-house or contributed by our collaborators at MaizeGDB (2), GrainGenes (3), and elsewhere. Future enhancements will include advanced level of annotation providing information on statistical scores, contributing allele, germplasm, population, assay environments, and treatment and associations to various controlled vocabularies (ontologies) described in the ontology section.

RICE, MAIZE AND *ARABIDOPSIS* GENOMES

The Genome browser section of Gramene (http://www.gramene.org/genome_browser) is a graphical display of annotations on plant genomes, and includes tracks describing genes, transcripts, peptides, SNPs, repeats, ESTs, markers, insertion elements, and other features of interest. Currently we host annotated genome assemblies of rice-*japonica* (7) and *Arabidopsis* (8), as well as a representation of the physical bacterial artificial chromosome (BAC) map of maize (9). These genomes are displayed using the latest version of the genome browser developed by the Ensembl project (10).

A



B

Gene	LOC_Os03g03350 (TIGR_LOCUS ID)
Ensembl Gene ID	LOC_Os03g03350
Genomic Location	**View gene in genomic location:** 1410268 - 1413415 bp (1.4 Mb) on chromosome 3 **This gene is located in sequence:** AC119747
Description	polygalacturonase (ec 3.2.1.15) (pectinase) (pgl).; polygalacturonase-like protein
Prediction Method	This gene was annotated by TIGR through a process of automatic and manual curation.
Sequence Markup	View genomic sequence for this gene with exons highlighted
Export Data	Export gene data in EMBL, GenBank or FASTA
SNP information	View information about variations on this gene.
Transcript Structure	1: LOC_Os03g03350.1 (LOC_Os03g03350.1) [Transcript information] [Exon information] [Protein information] Features ▼
Orthologue Prediction	The following gene(s) have been identified as putative orthologues by reciprocal BLAST analysis: Species / Type / dN/dS / Gene identifier *Arabidopsis thaliana* UBRH 0.01508 At3g48950-TIGR-G (At3g48950.1) [MultiContigView] [Align] glycoside hydrolase family 28 protein / polygalacturonase (pectinase) family protein *Zea mays2* UBRH 0.09018 FGENESHG0005846 (Novel Ensembl prediction) [MultiContigView] [Align] No description View alignments of homologies. UBRH - (U)nique (B)est (R)eciprocal (H)it

The rice genome (http://www.gramene.org/Oryza_sativa/) serves as the entry point for sequence-based comparative genomic analyses. We do this by identifying regions of sequence similarity between rice and other cereal species using a variety of sequence-based tags including the sequences of ESTs, proteins, full length cDNAs, flanking sequence tags (FSTs), microarray probes, BACs and BAC ends, SNPs and repeat elements. Our mapping protocols use conservative standards in order to minimize the rate of false correspondences, as described at http://www.gramene.org/documentation/Alignment_docs/to_Japonica/index.html. In addition, we have used protein sequence-based orthology information to identify gene family relationships among *Arabidopsis*, rice and maize (Figure 2B) using the Ensembl compara pipeline (http://www.gramene.org/Oryza_sativa/helpview?se=1&kw=geneview#OrthologuePrediction). The current rice genome assembly imported from the Osa1 rice genome annotation database (7) is a new addition that replaced the earlier-reported BAC clone-wise browser displaying annotation (5).

The maize genome browser (http://www.gramene.org/Zea_mays/) is based on an FPC physical map developed by the Arizona Genomics Institute (AGI; http://www.genome.arizona.edu/fpc/maize/) (9). In addition to the features mapped by AGI, we have added to this map many sequenced features from various grass species including rice and maize (http://www.gramene.org/documentation/Alignment_docs/to_Maize/index.html). The addition of the Ensembl synteny viewer (http://www.gramene.org/Oryza_sativa/syntenyview?otherspecies=Zea_mays) displaying the patterns of long-range synteny among the rice and maize genomes provides a useful comparative tool for users to find colinear regions of the rice and maize genomes as they search for genes, their functional orthologs and shared genetic markers. We constructed syntenic blocks between rice and maize by constructing a sorted pairwise list of locations of mapped overgo markers on the maize FPC map and identifying their corresponding locations on the rice genome (D. H. Ware *et al*., unpublished data).

Arabidopsis is the premiere model organism for plants, and its genomic sequence and well-characterized genes (8) are of great help in the annotation and characterization of genes found in the cereal genomes. Therefore the *Arabidopsis* genome assembly was added to the genome search modules in Gramene. This helps users familiar with the function or phenotype of a known *Arabidopsis* gene to traverse between genomes and find the expressed, known and/or predicted gene sequence(s) mapped on the rice and maize genomes. Similarly the computed orthologies also allow to traverse from cereal genomes to that of *Arabidopsis*. The data for the *Arabidopsis* genome browser (http://www.gramene.org/Arabidopsis_thaliana/) were kindly provided by the Nottingham Arabidopsis Stock Centre (NASC: http://nasc.life.nott.ac.uk/) who initially imported it from The Institute for Genome Research (TIGR) (8).

Future enhancements to this module's search capabilities will include information on the assembled rice-*indica* shotgun sequence (11,12) and its comparisons with the rice-*japonica* genome, mapping of Swiss-Prot–Trembl database protein entries to the gene models from rice and maize and mapping of ESTs and gene expression profiles to the gene models in rice and maize.

GRAMENEMART AND BLAST SEARCH

GrameneMart (http://www.gramene.org/Multi/martview) simplifies the task of creating and maintaining query interfaces. It is backed by a relational database and is particularly well-suited for providing data mining-like searches of complex descriptive (e.g. biological) data. After choosing a dataset, e.g. TIGR rice genes, one can filter the data to view only those rice genes with *Arabidopsis* homologs, or one can add a second filter to search for only those peptides with known molecular function(s). After this, a user can immediately export results from the output page or click on the count button to calculate the number of entries expected in the final output. MartView can generate a number of different types of output, including sequence and tabulated list data in various formats, including HTML, tab delimited text and Microsoft Excel. The GrameneMart is a new addition to the Gramene database and is served from the Genome module (http://www.gramene.org/genome_browser/index.html). Originally developed as a part of BioMart, it is a generic data management system developed jointly by the European Bioinformatics Institute (EBI) and Cold Spring Harbor Laboratory (CSHL).

The BLAST search (http://www.gramene.org/Multi/blastview) provides access to BLAST sequence similarity search algorithms via a web interface. It allows for simultaneous searches with multiple query sequences against sequence datasets from multiple target species. The sequence datasets include genomes, ESTs, markers, genes, cDNAs, FSTs, BACs and BAC ends and proteins. The updated version allows users to set their own alignment parameters, select specific combinations of species and sequence datasets and to upload sequence files and display options. This is in addition to saving or bookmarking the web link of their search results, with a link that remains active for one week.

GENE SEARCH

The gene search module (http://www.gramene.org/rice_mutant/index.html) is a curated resource that provides publicly available information on genetically identified genes in

Figure 2. (**A**) Rice genome browser. A detail view of a section of rice chromosome-3 where the BAC clone AC119747 is mapped as part of the genome assembly. This genomic clone is the first feature in the view and is described as a DNA contig in blue and labeled with a GenBank accession number. The annotations provided in relation to this BAC represent datasets generated by Gramene using its internal gene prediction pipeline (e.g. Rice_GeneModel-FGENESH) as well as annotations obtained from external databases or individual researchers (as labeled in each track). The views can be customized by the researcher by going to the dropdown menus on the browser views (e.g. Features, ESTs, GSSs, etc.) to display or hide specific tracks, or one can select features in a track by clicking the plus sign next to the track name. A researcher may traverse to a physical map in the comparative map tool by selecting an option from the 'Jump to' menu. Genes displayed in grey in the Rice_GeneModel_TIGR track suggest that putative orthologs have been identified in maize and/or *Arabidopsis*. (**B**) The detailed view of a gene report page for the gene Loc_Os03g03350 provides a general description of the gene and its functional annotation, as well as specific information about the gene, its introns and exons, the transcript(s), peptide(s), putative orthologs, and also offers a download option.

rice (*Oryza* sp.). It includes descriptions of genes and alleles, associated with morphological, developmental and agronomically important phenotypes, variants of physiological characters, biochemical functions and isozymes. Users can search for the genes by their name, symbol or accession number. For example a search for 'flowering' yields as many as 64 genes with the word 'flowering' appearing in either the gene name or the description. A 'browse by alphabetical order of gene symbol' option is also available (http://www.gramene.org/db/mutant/display_mutant_list). The database now contains 1488 characterized genes, an increase of ~100 genes over earlier reports (5) and many are fully annotated with phenotypic descriptions, map positions, sequence definitions, identification of gene products, allele(s), germplasm(s) in which the allele was characterized and citations, along with associations to trait (TO), plant structure (PO) and plant growth stages (GRO). e.g. rice Photoperiod-sensitivity-1 (hd1) gene http://www.gramene.org/db/mutant/search_mutant?id=GR:0060860. In the future our curational activity will focus on the addition of ~18 000 genes identified in the recently sequenced rice genome that have full length cDNA evidence (7,13). Previously, this module was called 'Mutants' but it has been renamed as 'Genes' to better describe the datasets stored within.

PROTEIN SEARCH

The protein module (http://www.gramene.org/protein/index.html) provides curated information about Swiss-Prot–Trembl protein entries from the grasses (Poaceae or Gramineae). Protein entries are annotated using the Gene Ontology (GO) (14,15) for biochemical characterization and the Plant Ontology (PO) for gene expression and phenotype associations. Information stored in this module is derived from published reports or generated by *in silico* experiments and includes mappings to Interpro (16), EC number, transmembrane [using TMHMM (17)], SignalP (18) and Predotar (19) analyses. Each association is supported with evidence in the form of a reference along with a corresponding evidence code (experiment type http://www.gramene.org/plant_ontology/evidence_codes.html). Information on the genes associated with the proteins is provided via a phenotype link (e.g. the protein from rice Photoperiod-sensitivity-1 gene isolated from Ginbozu germplasm http://www.gramene.org/db/protein/protein_search?acc=Q9FE92). Since the rice genome sequence assembly and annotation processes have only begun to stabilize recently, we also provide a BLASTP query link so that users can determine the best match to a peptide(s) that is deduced from the a predicted/known rice gene(s) identified on the rice genome assembly (information that is provided in the genome browser section).

Currently we provide information on ~67 000 cereal proteins but only rice proteins (~55 000) are subjected to internal manual curation. The addition of proteins from other grasses enhances the comparative aspect of Gramene and is a new feature of this module. Future plans include providing links to the peptides and genes found on the genome assembly, information on homologs and orthologs for intra and interspecific comparison and working with our collaborators on species-specific protein curation.

ONTOLOGY SEARCH

With the increasing demands of large-scale genomic experiments that generate large datasets related to gene expression and phenotype analyses, the requirement for use of controlled vocabularies (ontologies) has become more apparent (14,20). The ontologies are organized in structures called directed acyclic graphs, which differ from hierarchies in that a child, or more specialized, term can have many parents, or less specialized, terms. For example (http://www.gramene.org/db/ontology/search_term?id=TO:0000207), the trait term plant height has two parents, suggesting that it is a sub-type of shoot anatomy and morphology trait and is also a sub-type of the height-related trait. This helps the user to find the associated genes and QTL either via the anatomy or the height-related trait path of the ontology tree and still get the same query result.

To emphasize the use of such vocabularies in Gramene's annotation protocols (21) and to allow users to find genes, proteins, QTL, map sets and traits (http://www.gramene.org/plant_ontology/index.html), we have either adopted the ontologies developed by ongoing projects such as the GO (14,15), or the PO (22) or developed our own ontologies such as the TO (23), Environment (EO) and Taxonomy (GR_tax) ontologies, with help from collaborators (P. Jaiswal *et al*., unpublished data).

The TO (23) has 8794 unique gene and QTL associations. The GO (14,15) has ~1 94 000 associations to ~40 000 Swiss-Prot–Trembl proteins from the grasses and 21 000 rice genes from the Ensembl database. The PO (22) and GRO (21), have associations to ~420 unique rice genes. The most recent addition was the GR_tax that provides a way of identifying information from all the datasets related to a given species in Gramene. For example, rice or *Oryza* has associations to 106 map sets in CMap, 1488 genes and 55 306 SP–Trembl proteins. In the future we will extend our annotation of genes and QTL to include annotations using the EO and PO (22).

OUTREACH AND EDUCATION

Gramene is a collaborative project between CSHL, Cornell University and the cereal community. Our outreach activities extend to all local, state, national and global communities and are interactive. Outreach occurs in many different forms including collaborations with contributors of datasets and tools or running workshops to educate users about Gramene's resources and how to use them in research. For the former we collaborate actively and acknowledge by either linking back to the source from the appropriate dataset in Gramene and/or by listing our collaborators contact information on our Collaborators page at http://www.gramene.org/collaborators/. For the day-to-day users of our database, we provide pre-designed queries, glossaries and frequently asked questions sections and offer online tutorials to guide outlining a step-by-step process showing users how to retrieve information from the database (http://www.gramene.org/workshop_tutorial.html). General information is provided via the species pages about various cereal crops, including their genetic or evolutionary histories, production profiles, biology and commercial uses (http://www.gramene.org/species/index.html). In order to maintain a high standard of curation and help us provide

new types of datasets, we request users to send us feedback by sending e-mail at gramene@gramene.org.

DATABASE AVAILABILITY AND CONTACT INFORMATION

Gramene is a curated, free for use, web-accessible data resource for comparative genome analysis in the grasses. The technological core of Gramene is the MySQL database management system, an open source relational database system that is stable and well supported. We have developed a relational schema to represent the various biological entities of Gramene, and a middleware layer to dynamically translate this information into web pages. The database and the curated datasets are freely available for local use and installation (http://www.gramene.org/documentation/gramene_installation.html).

FUTURE ENHANCEMENTS

The Gramene database is committed to build a community-based information resource serving as a repository for structured datasets. Various planned future enhancements are described within each search module; however two new modules that are yet to be implemented include the Genomic Diversity and Phenotype Data Model (GDPDM) and RiceCyc pathway tool. GDPDM (https://sourceforge.net/projects/gdpdm/) will serve as the basis of a module that captures both molecular and phenotypic diversity data. The main focus of this schema is to hold quantitative phenotypic data and molecular genotypic data. This data may be the product of QTL mapping experiments, association studies, breeding efforts or germplasm evaluation activities. GDPDM will be used to provide the basic infrastructure for searches on reference germplasms/stocks and the genetic variation that they embody. The RiceCyc pathway tool is being developed based upon the BioCyc pathway tool (24) and will provide information on the network of genes encoded by the rice genome and their role in various pathways found in a virtual rice plant cell.

ACKNOWLEDGEMENTS

This work is supported by the USDA Initiative for Future Agriculture and Food Systems (IFAFS) (grant no. 00-52100-9622) and USDA-Agricultural Research Service specific cooperative agreement (grant no. 58-1907-0-041). During 2004–2007 this work is also supported by the National Science Foundation (NSF) award no. 0321685 and USDA-ARS. We are thankful to numerous collaborators and contributors for help in curation and for sharing their datasets and tools. Funding to pay the Open Access publication charges for this article was provided by NSF 0321685.

Conflict of interest statement. None declared.

REFERENCES

1. Shen,L., Gong,J., Caldo,R.A., Nettleton,D., Cook,D., Wise,R.P. and Dickerson,J.A. (2005) BarleyBase—an expression profiling database for plant genomics. *Nucleic Acids Res*., **33**, D614–D618.
2. Lawrence,C.J., Seigfried,T.E. and Brendel,V. (2005) The maize genetics and genomics database. The community resource for access to diverse maize data. *Plant Physiol*., **138**, 55–58.
3. Matthews,D.E., Carollo,V.L., Lazo,G.R. and Anderson,O.D. (2003) GrainGenes, the genome database for small-grain crops. *Nucleic Acids Res*., **31**, 183–186.
4. Geer,R.C. and Sayers,E.W. (2003) Entrez: making use of its power. *Brief Bioinform*., **4**, 179–184.
5. Ware,D.H., Jaiswal,P., Ni,J., Yap,I.V., Pan,X., Clark,K.Y., Teytelman,L., Schmidt,S.C., Zhao,W., Chang,K. *et al*. (2002) Gramene, a tool for grass genomics. *Plant Physiol*., **130**, 1606–1613.
6. Wang,G.L., Mackill,D.J., Bonman,J.M., McCouch,S.R., Champoux,M.C. and Nelson,R.J. (1994) RFLP mapping of genes conferring complete and partial resistance to blast in a durably resistance rice cultivar. *Genetics*, **136**, 1421–1434.
7. Yuan,Q., Ouyang,S., Wang,A., Zhu,W., Maiti,R., Lin,H., Hamilton,J., Haas,B., Sultana,R., Cheung,F. *et al*. (2005) The institute for genomic research Osa1 rice genome annotation database. *Plant Physiol*., **138**, 18–26.
8. Haas,B.J., Wortman,J.R., Ronning,C.M., Hannick,L.I., Smith,R.K.Jr, Maiti,R., Chan,A.P., Yu,C., Farzad,M., Wu,D. *et al*. (2005) Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. *BMC Biol*., **3**, 7.
9. Gardiner,J., Schroeder,S., Polacco,M.L., Sanchez-Villeda,H., Fang,Z., Morgante,M., Landewe,T., Fengler,K., Useche,F., Hanafey,M. *et al*. (2004) Anchoring 9371 maize expressed sequence tagged unigenes to the bacterial artificial chromosome contig map by two-dimensional overgo hybridization. *Plant Physiol*., **134**, 1317–1326.
10. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al*. (2005) Ensembl 2005. *Nucleic Acids Res*., **33**, D447–D453.
11. Yu,J., Hu,S., Wang,J., Wong,G.K., Li,S., Liu,B., Deng,Y., Dai,L., Zhou,Y., Zhang,X. *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science*, **296**, 79–92.
12. Yu,J., Wang,J., Lin,W., Li,S., Li,H., Zhou,J., Ni,P., Dong,W., Hu,S., Zeng,C. *et al*. (2005) The Genomes of *Oryza sativa*: a history of duplications. *PLoS Biol*., **3**, e38.
13. International Rice Genome Sequencing Project (2005), The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
14. Clark,J.I., Brooksbank,C. and Lomax,J. (2005) It's all GO for plant scientists. *Plant Physiol*., **138**, 1268–1279.
15. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*., **32**, D258–D261.
16. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al*. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res*., **33**, D201–D205.
17. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol*., **305**, 567–580.
18. Bendtsen,J.D., Nielsen,H., von Heijne,G. and Brunak,S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol*., **340**, 783–795.
19. Small,I., Peeters,N., Legeai,F. and Lurin,C. (2004) Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, **4**, 1581–1590.
20. Berardini,T.Z., Mundodi,S., Reiser,L., Huala,E., Garcia-Hernandez,M., Zhang,P., Mueller,L.A., Yoon,J., Doyle,A., Lander,G. *et al*. (2004) Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiol*., **135**, 745–755.
21. Yamazaki,Y. and Jaiswal,P. (2005) Biological ontologies in rice databases. An introduction to the activities in Gramene and Oryzabase. *Plant Cell Physiol*., **46**, 63–68.
22. The Plant Ontology Consortium. (2002) The Plant OntologyTM Consortium and Plant Ontologies. *Comp. Funct. Genomics*, **3**, 137–142.
23. Jaiswal,P., Ware,D., Ni,J., Chang,K., Zhao,W., Schmidt,S., Pan,X., Clark,K., Teytelman,L., Cartinhour,S. *et al*. (2002) Gramene: development and integration of trait and gene ontologies for rice. *Comp. Funct. Genomics*, **3**, 132–136.
24. Krummenacker,M., Paley,S., Mueller,L., Yan,T. and Karp,P.D. (2005) Querying and computing with BioCyc databases. *Bioinformatics*, **21**, 3454–3455.