1. Gramene: A genomics and genetics resource for rice.

P. JAISWAL[1], J. NI[1], I. YAP[1], D. WARE[2, 4], W. SPOONER[2], K. YOUENS-CLARK[2], L. REN[2], C. LIANG[2], B. HURWITZ[2], W. ZHAO[2], K. RATNAPU[2], B. FAGA[2], P. CANARAN[2], M. FOGLEMAN[1], C. HEBBARD[1], S. AVRAHAM[2], S. SCHMIDT[2], T. CASSTEVENS[3], E. S. BUCKLER[3,4], L. STEIN[2] and S. MCCOUCH[1]

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

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 billions of dollars annually to the world economy (FAO 1999, http://www.fao.org). Therefore, continued improvement of Poaceae crops is necessary in order to continue to feed an ever-growing world population. The Gramene database (http://www.gramene.org) takes advantage of the known genetic colinearity (synteny) between rice and other major cereal crop genomes to provide researchers with the benefits of an annotated rice genome and to facilitate its comparisons with the other cereal genomes. Gramene is a one-stop online web resource that provides ready access to curated literature, genetic, physical and sequence based maps, markers, genes, genomes, proteins and QTL (Fig. 1) as well as comparative analysis tools. These datasets contribute to our understanding of the genome organization, as well as genes and gene families and their role in determining the anatomy, development, environmental responses and factors influencing agronomic performance of rice and other major cereal crop plants. As an information resource and public repository of curated rice datasets (Ware et al. 2002a, Ware et al. 2002b, Jaiswal et al. submitted to Nucleic Acids Res.), Gramene adds value to data that is initially contributed by numerous rice researchers, projects, databases and generic data repositories such as NCBI's Entrez (http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi?term=oryza). While Gramene database provides information on a range of grass species, the datasets and their presentation or accessibility via various modules described in this report are focused for rice only. The description is based on release #18 (September 2005) of the database.

Gramene is a collaborative project between Cold Spring Harbor Laboratory, Cornell University and the rice community. The information provided via the database is curated using both manual and computational methods. It is freely available and web-accessible. The technological core of Gramene is the MySQL database management system, an open source relational database system that is stable and well supported. The database and curated datasets are available and can be installed for local use by following the

instructions described in the installation document (http://www.gramene.org/documentation/gramene_installation.html).

Many geneticists and molecular breeders programs have an interest in exploring and comparing the genetic maps, genes and QTL from previously published literature. To enable researchers to query these existing datasets, the central comparative map search tool, CMap, can be accessed from Gramene's 'Maps' section (http://www.gramene.org/cmap/index.html). This tool presents a map as a linear array of interconnected features that correspond to a single linkage group in the case of a genetic map, to a single contig for a physical map, or to a contig or scaffold in the case of an annotated sequence. To set up a comparison between different map sets (Fig. 1) from either the same or different species and/or 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. As of September 2005, the Maps module hosted a total of 105 maps characterized into three different types, namely sequence, genetic, and QTL maps, with the vast majority being QTL maps. These maps belong to 9 different rice species (Table 1).

Detailed information about markers mapped on the maps described above is provided by Gramene's 'Markers' section (http://www.gramene.org/markers/index.html). This module allows users to search the marker collection using one or more marker names, and a search may be refined by selecting the marker type (e.g. RFLP) and/or species (e.g. rice). A query for "RZ" limited to RFLP marker types in rice gives 287 entries (http://www.gramene.org/db/markers/marker_view?marker_name=*RZ*&marker_type_id=3&species_id=1&action=marker_search). The marker details (Fig. 1) include marker name, synonym(s), type, species, germplasm from which it was derived, maps on which the marker can be found, genome position(s) on the rice-japonica Nipponbare genome sequence, links to the source of the marker, literature citations and if available, images illustrating surveyed length polymorphisms (e.g. rice SSR marker RM220 http://www.gramene.org/db/markers/marker_view?marker_name=rm220).

The 'QTL' section (http://www.gramene.org/qtl/index.html) is a new addition to the Gramene database. It facilitates the comparative study of QTL and their mapped regions to investigate colinear regions found to carry genes identified in the rice genome and/or QTL contributing to similar traits mapped on different rice populations. 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 (Fig. 1, 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 is a standardized vocabulary of traits to comparisons of phenotypes across species. As of September 2005, the QTL module includes about 7000 rice QTL identified for about 230 traits. For convenience of searching, these traits are grouped into eight major trait families

related to abiotic and biotic stress, fertility, anatomy, development, vigor, quality and yield. The rice QTL are curated by Gramene staff (Ni et. al. unpublished) from published literature that dates back to the very first rice QTL report in 1994 (Wang et al. 1994).

The 'Genome' section on rice (http://www.gramene.org/Oryza_sativa/) provides a graphical display of annotations on the rice genome, and includes various tracks describing genes, transcripts, peptides, SNPs, repeats, ESTs, genetic markers (RFLP, SSR), flanking sequence tags (FSTs) from the mutant insertion lines and other features of interest (Fig. 1). This is a quick way to find the rice markers, ESTs, cDNAs (Kikuchi et al. 2003), sequences flanking insertions elements such as *Tos17*, *Ac*/*Ds*, T-DNA from mutant lines, unigenes or EST clusters from TIGR (Lee et al. 2005) and PlantGDB (Dong et al. 2004), repeats and tranposable elements, SNPs (Yu et al. 2005), BAC clones an BAC ends from IRGSP (http://rgp.dna.affrc.go.jp/) and OMAP (http://www.omap.org/) projects, etc. aligned to the gene(s) or region(s) of interest. Currently we host annotated genome assemblies of Rice-japonica Nipponbare cultivar, sequenced by the IRGSP and assembled by the TIGR-OSA1 project (Yuan et al. 2005). The methods used to map the above described features are available from various alignment documents listed at http://www.gramene.org/documentation/Alignment_docs/to_Japonica/index.html. The annotated rice genome and its pre-computed comparisons with the maize and Arabidopsis genomes help users familiar with the function(s) or phenotype(s) of known gene(s) to traverse between these genomes and find the expressed, known and/or predicted gene sequence(s) based on either orthology or on gene function(s).

The most frequently used tool on the Gramene website is the Blast search (http://www.gramene.org/Multi/blastview). This allows users to perform similarity searches against sequence datasets that include genomes, ESTs, markers, genes, cDNAs, FSTs from mutant and/or insertion lines, bacterial artificial clones (BACs), BAC ends and proteins from several rice and other cereal plants.

The 'Genes' section (http://www.gramene.org/rice_mutant/index.html) is a curated resource that provides publicly available information on genetically identified genes in rice (*Oryza sp*). It includes descriptions of genes and alleles associated with morphological, developmental and agronomically important phenotypes, variants of physiological characteristics, 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). As of September 2005, the database contained 1,488 characterized genes, many 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: Jaiswal et al. 2002) , plant structure (PO: Consortium 2002) and plant growth stages (GRO), (e.g. rice *Semidwarf-1* (*sd1*) gene, Fig. 1,

http://www.gramene.org/db/mutant/search_mutant?id=GR:0060842). In the future, curational activity will focus on the addition of approximately 18,000 genes identified in the recently sequenced rice genome that are supported by the full length cDNA evidence (Project 2005, Yuan et al. 2005).

The 'Proteins' section (http://www.gramene.org/protein/index.html) provides curated information on about 55,000 Swissprot-Trembl protein entries from rice. Protein entries are annotated using the Gene Ontology (GO) (Clark et al. 2005) for biochemical characterization and the Plant Ontology (PO) (Consortium 2002) for gene expression and phenotype associations. For example, see the SD1 protein (Fig. 1, http://www.gramene.org/db/protein/protein_search?acc=q8rvf5). Information stored in this module is derived from published reports, or generated by computational analysis that finds functional domains, transmembrane regions, signal peptides, etc. The report on functional characterization is supported with cited references along with a corresponding evidence code (experiment type http://www.gramene.org/plant_ontology/evidence_codes.html).

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 (Consortium 2002, Clark et al. 2005). The ontologies are organized in categorical hierarchies of parent terms and child (more specialized) terms. For example the trait term 'plant height' has two parents, suggesting that it is a subtype of shoot anatomy and morphology trait and is also a sub type of height related trait (http://www.gramene.org/db/ontology/search_term?id=TO:0000207). 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 to help users find genes, proteins, QTL, map sets and traits (Fig. 1, http://www.gramene.org/plant_ontology/index.html), we have adopted various ontologies including the gene (GO: Clark et al. 2005), plant (PO: Consortium 2002), cereal plant growth stages (GRO), trait (TO: Jaiswal et al. 2002), environment (EO) and taxonomy (GR_tax) ontologies in our data annotation protocols (Ware et al. 2002a, Yamazaki and Jaiswal 2005).

To help users of our database, we provide pre-designed queries, glossaries and frequently asked questions (FAQs) sections. On-line tutorials guide users through a step-by-step process to retrieve information from the database (http://www.gramene.org/workshop_tutorial.html). General information about various cereal crop plants, including their genetic or evolutionary histories, production profiles, biology and commercial uses is also provided (http://www.gramene.org/species/index.html). For more information about Gramene, or to contribute suggestions, please contact Gramene at gramene@gramene.org.

This Gramene project was originally 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 #0321685 and USDA-ARS. We are thankful to numerous collaborators, researchers and contributors from the rice research community for sharing their datasets and help in curation.

Table. 1. Number of rice maps in Gramene's 'Maps' section. The sequence map is a summarized representation of the contiguous, assembled genomic sequence of an organism in a linear map format. The genetic map is a representation of a meiotic-recombination map based on analysis of marker segregation in a population of offspring derived from a bi-parental cross. Marker polymorphism between the parents is required to monitor recombination among loci along a chromosome. The QTL map is a type of genetic map which indicates the approximate location of a quantitative trait locus (QTL) within an interval delineated by two or more markers on a genetic map.

Species	**Number of maps**	**Map types**
Oryza sativa	97	Sequence, Genetic, QTL
Oryza alta	1	Physical map
Oryza australiensis	1	Physical map
Oryza brachyantha	1	Physical map
Oryza glaberrima	1	Physical map
Oryza glumaepatula	1	Physical map
Oryza nivara	1	Physical map
Oryza punctata	1	Physical map
Oryza rufipogpon	1	Physical map

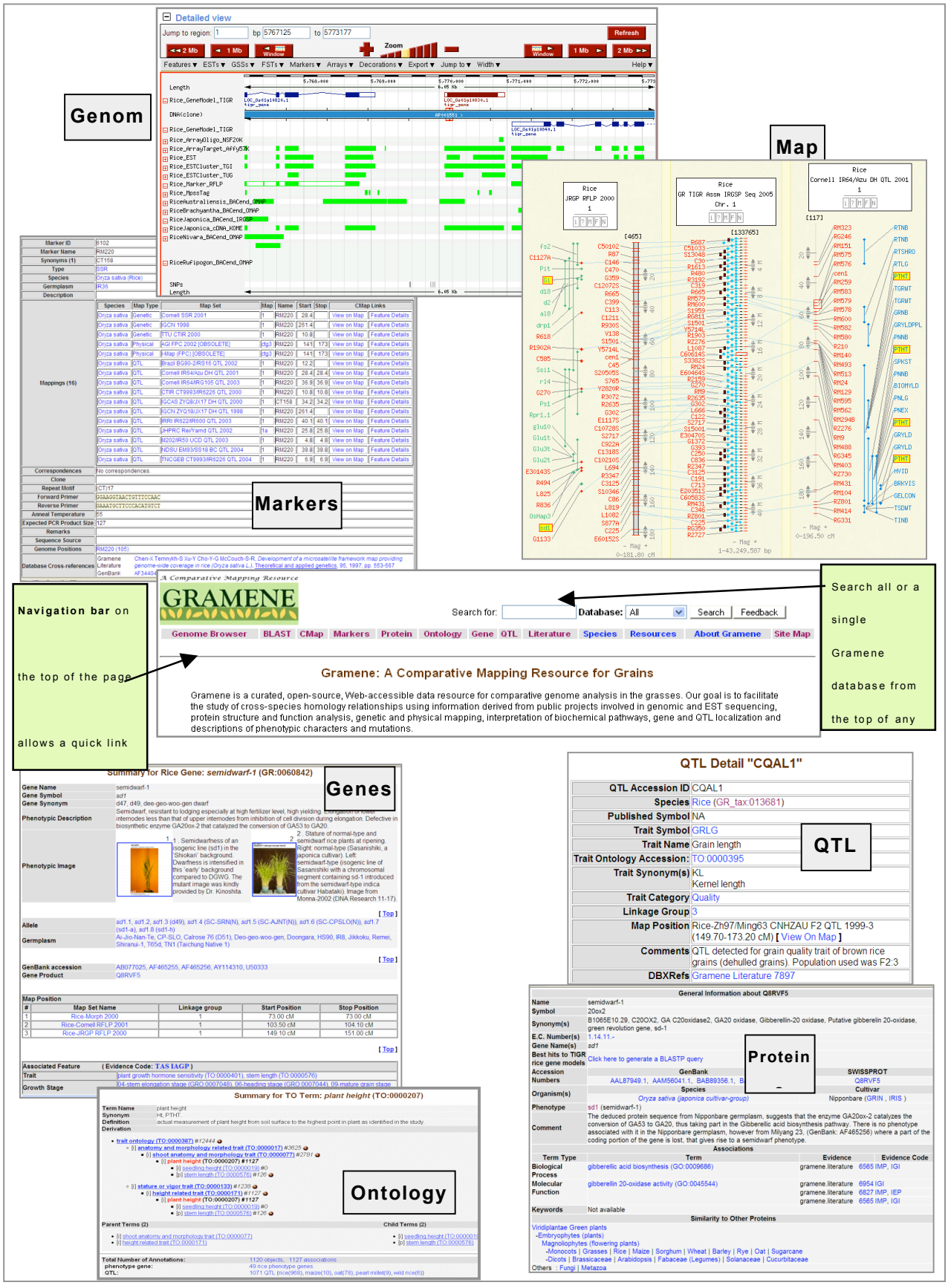
Genom
Map
Markers
Navigation bar on the top of the page allows a quick link
Search all or a single Gramene database from the top of any
Genes
QTL
Protein
Ontology

Fig.1. An overview of the Gramene database (http://www.gramene.org/). It includes truncated views of the home page in the center and of the sections on rice genome, maps, markers, genes, proteins, QTL and ontology above and below. These modules and other useful information can be accessed by clicking the respective links found in the navigation bar that is present on every web page. The literature section is not displayed in this figure, but can be accessed by clicking the 'Literature' link on the navigation bar. Users are encouraged to use the simple search option available on the top of every page. These searches can be filtered by the type of database (e.g. proteins, genes, ontology etc). For advanced searches please visit the respective sections. If you find problems or would like to send updates on old or suggest new data sets, please use the feedback form accessed by clicking the 'Feedback' button present on the top of every page. For an extended list of useful links within Gramene please visit the site map page available via the 'Site Map' link from the navigation bar menu.

References

Clark, J. I., C. Brooksbank and J. Lomax, 2005. It's all GO for plant scientists. Plant Physiol. 138: 1268-1279.

Consortium, T. P. O., 2002. The Plant Ontology™ Consortium and Plant Ontologies. Comparative and Functional Genomics 3: 137-142.

Dong, Q., S. D. Schlueter and V. Brendel, 2004. PlantGDB, plant genome database and analysis tools. Nucleic Acids Res. 32: D354-359.

Jaiswal, P., D. Ware, J. Ni, et al., 2002. Gramene: development and integration of trait and gene ontologies for rice. Comparative and Functional Genomics 3: 132-136.

Kikuchi, S., K. Satoh, T. Nagata, et al., 2003. Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. Science 301: 376-379.

Lee, Y., J. Tsai, S. Sunkara, et al., 2005. The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. Nuc. Acids Res. 33: D71-74.

Project, I. R. G. S., 2005. The map-based sequence of the rice genome. Nature 436: 793-800.

Wang, G. L., D. J. Mackill, J. M. Bonman, et al., 1994. RFLP mapping of genes conferring complete and partial resistance to blast in a durably resistance rice cultivar. Genetics 136: 1421-1434.

Ware, D., P. Jaiswal, J. Ni, et al., 2002a. Gramene: a resource for comparative grass genomics. Nuc. Acids Res. 30: 103-105.

Ware, D. H., P. Jaiswal, J. Ni, et al., 2002b. Gramene, a tool for grass genomics. Plant Physiol. 130: 1606-1613.

Yamazaki, Y. and P. Jaiswal, 2005. Biological ontologies in rice databases. An introduction to the activities in Gramene and Oryzabase. Plant Cell Physiol. 46: 63-68.

Yu, J., J. Wang, W. Lin, et al., 2005. The Genomes of *Oryza sativa*: a history of duplications. PLoS Biol. 3: e38.

Yuan, Q., S. Ouyang, A. Wang, et al., 2005. The institute for genomic research Osa1 rice genome annotation database. Plant Physiol. 138: 18-26.