 Nucleic Acids Research, 2002, Vol. 30, No. 1 **103–105**

Gramene: a resource for comparative grass genomics

Doreen Ware, Pankaj Jaiswal[1], Junjian Ni[1], Xiaokang Pan, Kuan Chang, Kenneth Clark, Leonid Teytelman, Steve Schmidt, Wei Zhao, Samuel Cartinhour[2], Susan McCouch[1] and Lincoln Stein*

Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA, [1]Department of Plant Breeding, Cornell University, Ithaca NY 14853-1901, USA and [2]USDA–ARS Center for Agricultural Bioinformatics, 626 Rhodes Hall, Cornell Theory Center, Ithaca, NY 14853, USA

Received October 4, 2001; Revised and Accepted October 26, 2001

ABSTRACT

Gramene (http://www.gramene.org) is a comparative genome mapping database for grasses and a community resource for rice. Rice, in addition to being an economically important crop, is also a model monocot for understanding other agronomically important grass genomes. Gramene replaces the existing AceDB database 'RiceGenes' with a relational database based on Oracle. Gramene provides curated and integrative information about maps, sequence, genes, genetic markers, mutants, QTLs, controlled vocabularies and publications. Its aims are to use the rice genetic, physical and sequence maps as fundamental organizing units, to provide a common denominator for moving from one crop grass to another and is to serve as a portal for interconnecting with other web-based crop grass resources. This paper describes the initial steps we have taken towards realizing these goals.

DESCRIPTION

Grasses are the fourth largest family of flowering plants and provide the staple diet of most of the world's population. Crop grasses vary greatly in their DNA content. Rice is a diploid with the smallest grass genome at 430 Mb, whereas other grasses such as maize (2400 Mb) and wheat (16 000 Mb) have considerably larger genomes. The expansion of the DNA content in the larger genomes appears to be primarily due to the insertions of repetitive DNA elements. The large size and composition of such genomes make these species unlikely candidates for complete genome sequencing in the short term. Despite these large differences in DNA content it has been recognized that the grass genomes maintain a high level of conserved macro-synteny (1) and a moderately high level of micro-synteny (2,3). This synteny among the crop grasses suggests that the rice genomic sequence will be more than a tool for understanding the biology of a single species (4,5) because it can function as a window into the structure and function of genomes in the other crop grasses as well (6,7). With 35% of the rice genome sequenced to date and the remainder expected to be complete in the next year, there is an urgent need for a carefully constructed and curated public resource for comparative mapping. Such a resource will provide the leverage necessary for researchers working in maize, sorghum, millet, sugarcane, wheat, oats and barley to capitalize on the rich information that will be emerging from the rice genome-sequencing project.

Gramene uses rice as a framework genome to organize information for other grass species. The database replaces 'RiceGenes', a resource with a more limited scope that also focused on rice and comparative genomics for agriculturally important members of the grass family. The goals of Gramene are to (i) establish a database utilizing the rice genome as a framework for identifying and characterizing genes in other grasses, (ii) provide comparative maps between rice and other grasses based upon orthologous sequence and the wealth of genetic and phenotype information available among the grasses, (iii) develop a pilot study to assign gene ontology (GO) functional classification to 4000 confirmed or predicted rice genes, (iv) curate information on major rice mutants, strains, phenotypes, polymorphisms and quantitative trait loci (QTLs) utilizing a structured controlled vocabulary, and (v) integrate with other plant databases to allow comparisons of conserved syntenic relationships and mutant phenotypes.

Gramene provides researchers with access to the most widely used rice genetic maps. Associated with the maps is information on the underlying restriction fragment length polymorphisms (RFLPs), simple sequence repeats (SSRs), amplified fragment length polymorphism (AFLPs) and mapping populations. For sequence-based markers such as SSRs, AFLPs and cDNA_RFLPs, Gramene provides researchers with information on experimental conditions, such as the primer, amplicon-sequence information, as well as the PCR amplification conditions.

Comparative genetic maps currently exist in Gramene as static predefined 'homeologies', displayed as intervals on a genetic map. A homeology represents a linear series of markers, which map to a similar series of loci on two or more different species of grass, for example, rice and maize. By selecting a homeology, the researcher obtains associated information on the species, map locations, markers in common, citations and a brief description of what the homeology represents.

A biologist has many avenues of access to a genetic map in Gramene. All biological data objects containing a reference to

*To whom correspondence should be addressed. Tel: +1 516 367 8380; Fax: +1 516 367 8389; Email: lstein@cshl.org

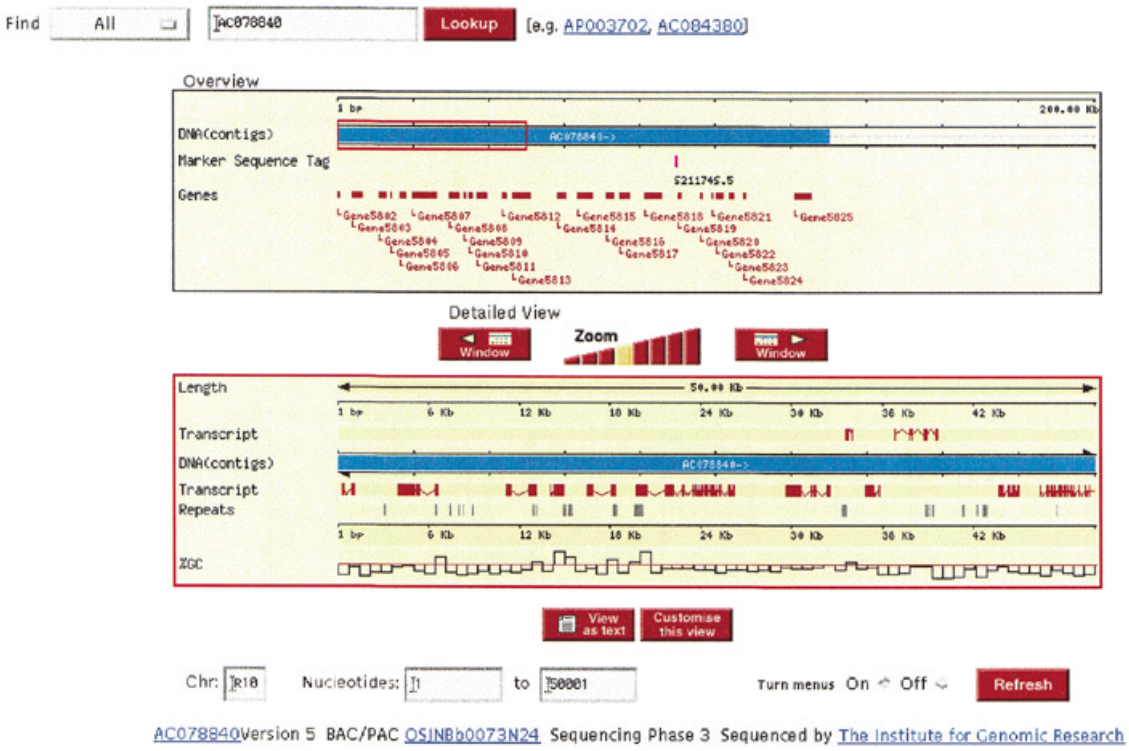

Figure 1. Modified Ensembl genome viewer displaying rice genomic annotations.

a map location act as entry points into genetic maps. These objects included markers, QTLs, homeologies and map studies. For example, by searching for a QTL, the researcher obtains a detail page containing information on the QTL study, germplasm identity, population structure, statistical significance of the QTLs and a reference for the study. In addition, specific details of the relevant experimental conditions, a description of the trait, and the genomic region spanning the QTL. By selecting the region spanned by a trait, the biologist enters a genetic map. Two versions of genetic maps are available in Gramene. The first, an interactive graphical display that provides zooming and scrolling functionality. The second is a tabular representation that is downloadable as tab delimited files or extensible markup language (XML) format.

In addition to genetic maps, Gramene provides the rice physical map. Gramene displays all available rice sequence and annotation generated from The International Rice Genome Sequencing Project (IRGSP) using a modified version of the Ensembl genome browser (http://www.ensembl.org). In addition to the annotation available from GenBank, the sequence view contains mapped rice ESTs, rice markers and end sequences of rice BAC clones. The mapping of sequence features allows the biologist to traverse between the genomic sequence, genes, markers and to genetic maps within the database (Fig. 1).

The biologist can access the rice genome and its annotation by specifying the name of a mapped BAC or PAC clone, a GenBank accession number, or by chromosome. He can also enter the genome via a BLAST search. The genome browser display is designed to minimize the amount of extraneous information displayed to the user. For example, a gene does not by default show the transcript or protein information, nor does a marker display its sequence, type or mapping study. However, this data is easily accessible within a detail page with a single click on the icon. The detail pages also contain links to speciality databases and the ability to obtain downloadable text files of each data set.

In addition to information on maps, markers, QTLs, sequences and genes, Gramene provides access to online versions of the Rice Genetics Newsletter and the Rice Genome Newsletter. Gramene also facilitates integration within the rice and grass communities by providing links to stock centers for rice and other grains, to major laboratories participating in rice genome research, academic and commercial research groups, and educational and bioinformatics resources.

SYSTEM DESIGN AND IMPLEMENTATION

Gramene is currently a hybrid system. Legacy data, including traits, QTLs, strains and literature citations, are maintained in the RiceGenes AceDB database. New data, including nucleotide sequence, sequence annotation, physical maps and new genetic markers, are maintained in a completely redesigned system.

The new system uses a Perl object model based on the bioperl (www.bioperl.org) and Ensembl code bases. The back end is the Oracle relational database, and the front end consists of a set of Perl scripts running in an Apache/mod_perl environment (Stein and MacEachern, 1998). The advantage of the object model middleware component is that it provides a high-level layer of abstraction between the front end and the back end, insulating front end developers from changes in the database schema and implementation.

Electronic data submission

We encourage contribution of electronic data sets from the scientific community. Any type of data that the Gramene database maintains can be submitted as an electronic contribution. For questions regarding electronic submission please contact us at gramene@gramene.org.

FUTURE ENHANCEMENTS

Gramene is committed to providing the community with a uniform and integrative view of the rice genome and associated phenotypic information. To obtain a consistent baseline annotation of the rice genome, Gramene will be using the Ensembl pipeline to produce automated annotation of the rice genome. This will provide a consistent level of annotation to support orthologous correspondences for comparative mapping. This automated annotation will be available in spring 2002. To complement the automated annotations, Gramene will be developing the tools and expertise necessary to classify 4000 rice genes according the GO classification system (8). In addition to the annotation provided by curators, Gramene will add electronically generated SWISS-PROT, InterPro to GO relationships. The electronic annotations will be available by November 2001.

For phenotypic information including functional mutations and QTLs, Gramene in a collaborative effort with IRRI (rice), ICIS (representing several crops); MaizeDB (maize), CIMMYT (maize and wheat) have put intensive efforts into the development of controlled structured vocabulary for traits. The 'trait ontology' (TO) initially based upon INGER's Standard Evaluation System for Rice (9), provides the terms that describe a trait as a distinguishable feature, characteristic, quality of character or a phenotypic feature of a developing or developed individual. A specific goal of Gramene is to provide a TO that can be used across the grasses to facilitate phenotypic comparisons both within and between genera.

A biologist will be able to access the TOs and GOs by searching an ontology browser with a term name or ontology ID. At the current time the preliminary TO and definitions can be downloaded from a public CVS repository. The TOs are in the early stages of development and community participation is encouraged. To facilitate participation, a web-based submission form is available for suggestions regarding the addition, replacement or modification of the TOs (http://www.gramene.org/plant_ontology/). The development and use of the common vocabulary of the GO and TO, will facilitate interoperability between plant databases, permit more complex annotation of genes and phenotypes and greatly increase the power of comparative mapping by allowing researchers in the future to make more sophisticated queries of the database by combining searches based on homeology with searches based on functional characteristics of genes.

A key enhancement to the comparative maps will be the replacement of static homeology relationships with a dynamic comparative mapping tool based on the idea of a 'correspondence table'. Each entry in a correspondence table will contain an assertion that a marker in one map corresponds to a marker in another. Representing homeology relations in this way rather than hard wiring them will allow synteny maps to be constructed, examined, adjusted and revised in a dynamic and interactive fashion. The current implementation of a comparative mapping tool based on this concept allows side by side alignments of two maps using cross-hybridization and sequence similarity data and will be available in November 2001. A biologist will be able to traverse from a genetic map to a physical map and finally a view of the genes and their annotations using the initial implementation of the comparative-mapping tool.

After the initial implementation of the comparative mapping tool, we plan to extend correspondence table relationships to include phenotypic similarity and gene orthology relationships. Each assertion will be associated with a correspondence method ID that describes the basis for the assertion, and a method-dependent score that indicates the confidence of the correspondence. The addition of these correspondences will give the user the ability to combine the results of correspondence tables, for example by superimposing a synteny map constructed by sequence similarity with one constructed from the participation of known genes in identified biochemical pathways.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by USDA CREES grant 00-52100-9622, as well as funding from USDA ARS Specific Cooperative Agreement grant 58-1907-0-041.

REFERENCES

1. Gale,M.D. and Devos,K.M. (1998) Comparative genetics in the grasses. *Proc. Natl Acad. Sci. USA*, **95**, 1971–1974.
2. Tarchini,R., Biddle,P., Wineland,R., Tingey,S. and Rafalski,A. (2000) The complete sequence of 340 kb of DNA around the rice Adh1-adh2 region reveals interrupted colinearity with maize chromosome 4. *Plant Cell*, **12**, 381–391.
3. Keller,B. and Feuillet,C. (2000) Colinearity and gene density in grass genomes. *Trends Plant Sci.*, **5**, 246–251.
4. McCouch,S. (1998) Toward a plant genomics initiative: thoughts on the value of cross-species and cross-genera comparisons in the grasses. *Proc. Natl Acad. Sci. USA*, **95**, 1983–1985.
5. Gale,M., Moore,G. and Devos,K. (2001) Rice—the pivotal genome in cereal comparative genetics. *Novartis Found. Symp.*, **236**, 46–53.
6. Freeling,M. (2001) Grasses as a single genetic system. Reassessment 2001. *Plant Physiol.*, **125**, 1191–1197.
7. Dubcovsky,J., Ramakrishna,W., SanMiguel,P., Busso,C., Yan,L., Bryan,A., Shiloff,L. and Bennetzen,J. (2001) Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes. *Plant Physiol.*, **125**, 1342–1353.
8. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
9. INGER (1996) *Standard Evaluation System for RICE*, 4th edn. International Rice Research Institute, DAPO Box 7777, Manila, Philippines.