D996–D1000 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 27 September 2008*
doi:10.1093/nar/gkn638

RiceGeneThresher: a web-based application for mining genes underlying QTL in rice genome

Supat Thongjuea[1,2], Vinitchan Ruanjaichon[1,2], Richard Bruskiewich[3] and Apichart Vanavichit[1,*]

[1]RiceGeneDiscovery Unit, Kasetsart University, Kamphangsaen Campus, Nakhon Pathom 73140, [2]National Center for Genetic Engineering and Biotechnology, 113 Thailand Science Park, Phahonyothin Road, Klong 1, Klong Luang, Pathumthani 12120, Thailand and [3]Crop Research Informatics Laboratory – International Rice Research Institute (IRRI), DAPO Box 7777, Metro Manila, Philippines

Received July 24, 2008; Revised September 10, 2008; Accepted September 15, 2008

ABSTRACT

RiceGeneThresher is a public online resource for mining genes underlying genome regions of interest or quantitative trait loci (QTL) in rice genome. It is a compendium of rice genomic resources consisting of genetic markers, genome annotation, expressed sequence tags (ESTs), protein domains, gene ontology, plant stress-responsive genes, metabolic pathways and prediction of protein–protein interactions. RiceGeneThresher system integrates these diverse data sources and provides powerful web-based applications, and flexible tools for delivering customized set of biological data on rice. Its system supports whole-genome gene mining for QTL by querying using DNA marker intervals or genomic loci. RiceGeneThresher provides biologically supported evidences that are essential for targeting groups or networks of genes involved in controlling traits underlying QTL. Users can use it to discover and to assign the most promising candidate genes in preparation for the further gene function validation analysis. The web-based application is freely available at http://rice.kps.ku.ac.th.

INTRODUCTION

Genetic studies of many traits in rice over the past decade have generated data suggesting that specific regions of rice chromosomes contain sites that influence expressions of quantitative traits or quantitative trait loci (QTL). In recent years, many quantitative traits in rice have been discovered by QTL mapping. There are more than 8000 QTL controlling various complex traits that have been located on different chromosome regions in rice (http://www.gramene.org/qtl/index.html). The goal of QTL mapping is to identify the genes underlying polygenic traits and gain a better understanding of their physiology and biochemistry (1). To identify genes, the candidate gene approach has been applied in plant genetics in the past decade for the characterization and cloning of QTL. It constitutes a complementary strategy to map-based cloning for identifying the genes placed within the QTL intervals (2). The process of selecting candidate genes relies on a wealth of information gained through traditional genetics and molecular approaches (3). Presently, many publicly useful biological data, especially, the high-throughput technologies are significantly increasing the volume of biological information to assist gene function identification. Most of these biological data are published on the public domain databases. Scientists target these databases to apply bioinformatics approaches and data integration systems to find the most promising candidate genes and to elucidate functions of rice genes. This study presents the newly improved RiceGeneThresher. It has been designed and developed to integrate catalogs from the public domain databases on rice that involve genetic information, genome annotation, expressed sequence tags (ESTs), protein information such as protein domains, gene ontology (GO), metabolic pathway information, prediction of protein–protein interaction and stress-responsive genes. RiceGeneThresher system provides a generic data warehousing solution for fast and flexible querying of rice biological data sets and integration system. It is able to generate evidences from each type of omics information that are essential for analyzing and targeting groups or networks of loci involved in the controlling of gene expression for the specific traits underlying the QTL regions. Users, ranging from breeders, laboratory researchers to the experienced molecular biologists, can use it in a wide variety of applications and scenarios to find the most promising candidate gene to improve rice cultivars. To compare RiceGeneThresher with the relevant existing database like GRAMENE (4), although GRAMENE is a large resource for major crop genomes, it does not provide an easy way for mining biological

*To whom correspondence should be addressed. Tel: 66 34 355 193; Fax: 66 34 355 197; Email: vanavichit@gmail.com

data underlying QTL. Unlike RiceGeneThresher, it was specially designed to use for doing QTL mining intuitively and easily. GRAMENE contains various biological data, which are displayed across diverse software such as Ensembl Genome Browser, CMap, BioMart and BioCyc. Differ from RiceGeneThresher, it consolidates both diverse biological data and third-party software into one system that is convenient for users to discover and display candidate genes.

DATA SOURCES AND ANALYSIS

Rice genome and gene annotations were collected from Michigan State University (http://rice.plantbiology.msu.edu/data_download.shtml). ESTs of rice were retrieved from National Center for Biological Information (NCBI) EST database (http://www.ncbi.nlm.nih.gov/projects/dbEST/). ESTs were trimmed of contaminated DNA sequences by SeqClean (http://compbio.dfci.harvard.edu/tgi/software/) and were aligned on the TIGR pseudomolecules by using BLAT (5) with cut-off percent identity $\geq$95%, *E*-value $\leq$1e-10 and similarity scores $\geq$200. Rice DNA markers were collected from the GRAMENE database. They were aligned to the TIGR pseudomolecules by using DNA sequences and DNA primer sequences by BLAT and electronic polymerase chain reaction (e-PCR) (6), respectively. The entire protein domains of rice were analyzed by InterProScan (7). GO terms were downloaded from http://www.geneontology.org/GO.downloads.shtml and each rice protein was assigned to GO terms by InterPro2GO. KEGG metabolic pathways information for rice genes were downloaded from ftp://ftp.genome.jp/pub/kegg/genes/organisms/osa/. Rice proteins from KEGG were mapped to assign metabolic pathways to rice proteins from TIGR by using BLASTp (8) with the best BLAST hit cut-off score. To predict the entire protein–protein interactions of rice, evidences of protein–protein interactions of *Arabidopsis thaliana* were obtained from The Arabidopsis Information Resource (TAIR) database (9) (http://www.arabidopsis.org). The Interolog Mapping (10) method and the phylogenomic approach (11) were combined to apply for inferring protein–protein interaction evidences from Arabidopsis to rice. In addition, by connecting with the Generation Challenge Programme (GCP), The Generation Challenge Programme Comparative Plant Stress-responsive Gene Catalogue (12) has been incorporated into RiceGeneThresher database for identifying candidate stress-responsive genes underlying QTL. The GCP is an online resource documenting stress-responsive genes comparatively across plant species.

RICEGENETHRESHER DATABASE AND IMPLEMENTATION

RiceGeneThresher is a MySQL database based mainly on the Chado schemata of the Generic Model Organism Database project (www.gmod.org) (13), with local enhancements where necessary. This has been developed and designed to store biological data such as genome sequences, genome annotation, protein sequences, protein domains, metabolic pathways and so on. RiceGeneThresher requires access to many sources of biological data, many of which are distributed across the internet. To make data sources for feeding RiceGeneThresher system, RiceGeneThresherDataSource API was developed as the middleware for querying and transferring data to the front-end. RiceGeneThresherDataSource incorporates not only the static data from RiceGeneThresher database (MySQL) but it also incorporates the dynamic data across the internet by using GCP-compliant BioMOBY (14). The front-end is a web-based interface that uses Java-based software technology running on the top of Apache Tomcat Web Server. RiceGeneThresher web interface was implemented by Asynchronous JavaScript and XML (AJAX) technology, which consists of DWR (http://getahead.org/dwr/), Velocity (http://velocity.apache.org/) and EXT2.0 (http://extjs.com/). Third-party bioinformatics software such as Cytoscape (15), ATV (16), Jalview (17) and BLAST, an NCBI search tool, were combined with RiceGeneThresher for analyzing and displaying the queried results. In addition, the European Bioinformatics Institute (EBI) web services (18), for instance WUBLAST, NCBIBLAST and InterProScan are also incorporated into RiceGeneThresher web-based software for doing the similarity searching of DNA or protein sequences on the fly.

USER INTERFACE

Presently, there are two main options for using RiceGeneThresher: (i) mining genome region of interest by submitting DNA marker names or DNA sequences, (ii) querying genes by gene locus names or gene annotation keywords.

Mining a genome region

Users can select the genome region of interest by submitting DNA marker names or DNA sequences (Figure 1A and B). After the submission, RiceGeneThresher displays the position of those DNA markers or DNA sequences on the rice physical map and it automatically selects the widest flanking positions on rice physical map (Figure 1C).

For changing the auto selection, users can specify the genome region of interest by selecting on the flanking DNA marker positions or DNA sequence positions. RiceGeneThresher explores various kinds of rice biological information found on the particular genome region of interest. Main features consist of genome, transcriptome, proteome, metabolome, phylogenomic and interactome.

Genome feature. The genome feature (Figure 1D) describes genome structure information. It displays data of genome structure; for example, a total number of contigs tilling path, the total size of DNA sequences, a total number of both transposable element (TE)-related and non-TE-related genes, gene density, average gene length and total number of DNA markers found in that particular genome region. It also categorizes genes into a group of annotation types such as those consisting of

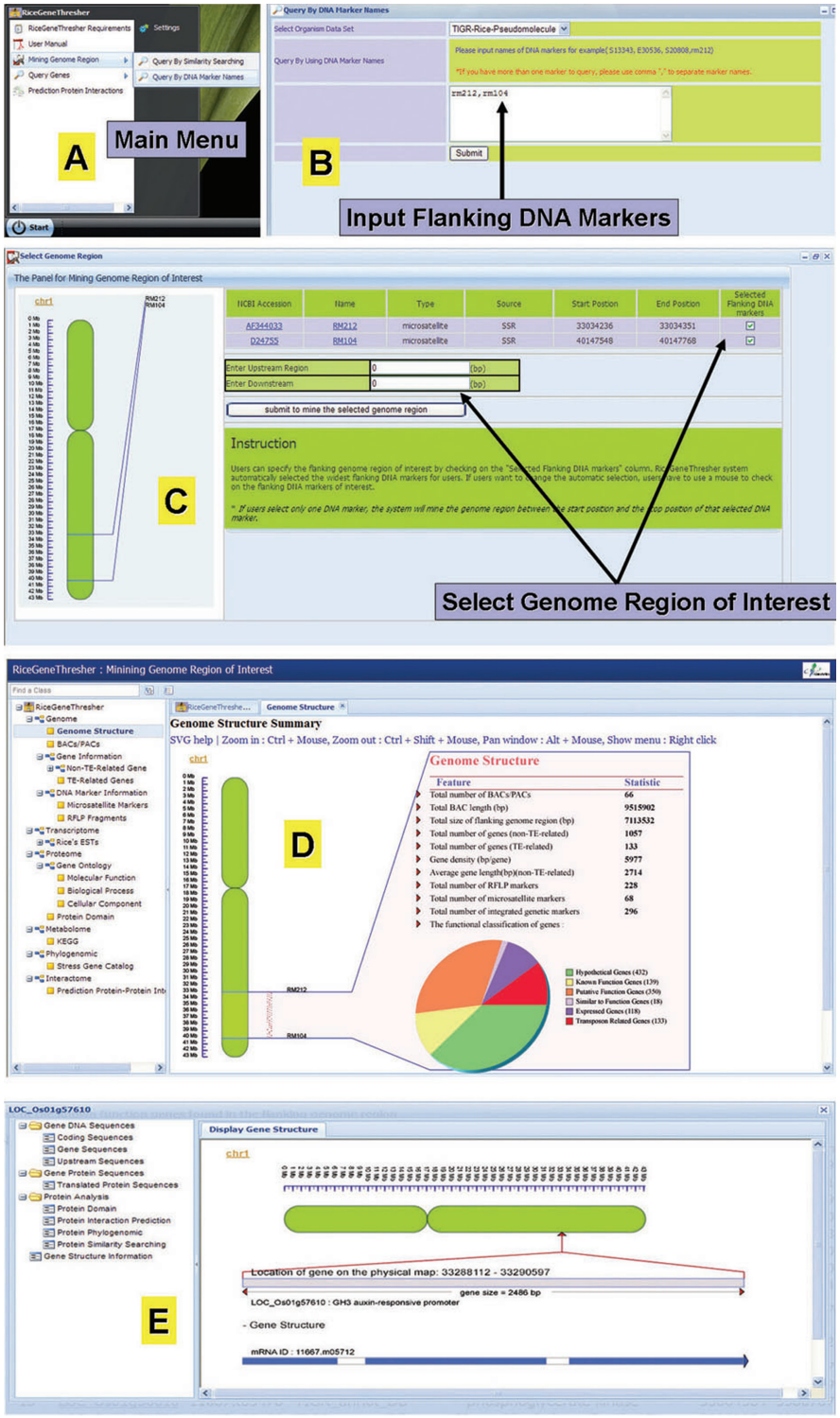

Figure 1. The figure shows RiceGeneThresher main menu (**A**), the query by using flanking DNA marker names (**B**), the genome region of interest for mining rice biological data (**C**), the example of genome feature information (**D**) and the particular rice gene information (**E**).

hypothetical genes, known function genes, putative function genes, similar to function genes, expressed genes and transposable related genes.

Transcriptome feature. The transcriptome feature shows a collection of EST libraries that have high similarity searching scores, *E*-values and percentage of identities against genes found in the genome region of interest. EST libraries were divided into a group of EST experiment names and a group of EST tissue types. Users can select candidate stress-responsive genes, for example, drought stress-responsive genes, by selecting on genes

that are involved in EST library names such as 'IRRI Drought Stress'.

Proteome feature. The proteome feature displays a list of GO terms of each gene product divided into a group of molecular function, biological process and cellular component, and it also displays a group of protein domains analyzed by the InterProScan.

Metabolome feature. The metabolome feature displays the mapping of rice gene found in the selected genome region into KEGG metabolic pathways. Its feature explores three kinds of pathway information. The first information shows the number of protein-coding genes that is predicted as the enzyme are located on the genome region of interest. The second shows the number of predicted metabolic pathways that are found on that particular genome region and the last information shows, in each metabolic pathway, the protein-coding genes involved in that particular metabolic pathway.

Phylogenomic feature. The phylogenomic feature shows data curated by GCP Comparative Plant Stress-responsive Gene Catalogue, which located in the genome region of interest. Users can easily select candidate stress-responsive genes by using this feature and it provides more information about protein families, phylogenetic trees, multiple sequence alignments (MSA) and associated experimental evidence.

Protein–protein interaction feature. The protein–protein interaction feature shows results of protein–protein interaction found in the genome region of interest. RiceGeneThresher displays protein–protein interaction prediction in a table format. In addition, the third-party software such as Cytoscape is incorporated to display all protein–protein interactions in the graphic mode by using Java web start technology.

Querying genes

Users can query genes in rice genome by using gene annotation keywords for example 'protein kinase' or gene locus id. When users select on a gene of interest, RiceGeneThresher displays individual gene information in both graphical user interface and standard web pages. Gene information (Figure 1E) consist of a picture of gene structure that locates on a rice chromosome location, coding DNA sequences, whole-gene DNA sequences, upstream DNA sequences and translated protein sequences. The result of protein analyses such as protein domains, protein–protein interaction prediction, and similarity searching results of protein by using EBI's web services, and protein phylogenomics by querying against the GreenPhyl database (http://greenphyl.cines.fr/cgi-bin/greenphyl.cgi) are also included in the information.

FUTURE DIRECTIONS

Currently, the authors of this study are generating expression data from many experiments by using rice Affymetrix GeneChip (www.affymetrix.com) together with microarray data from public databases and literatures. The authors also plan to expand RiceGeneThresher system to support those microarray data for finding candidate genes underlying genome region of interest. As draft whole-genomic sequences generated from new generation sequencing technologies become more available, the authors plan to develop new features into RiceGeneThresher in order to integrate these short shot-gun sequences into its gene mining platform. There are also plans to update the rice genome database to use the TIGR rice genome release 6.0.

FUNDING

GCP Fellowship Award (to RiceGeneThresher project); the National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand (to RiceGeneThresher project). Funding for open access charge: BIOTEC.

Conflict of interest statement. None declared.

REFERENCES

1. Korstanje,R. and Paigen,B. (2002) From QTL to gene: the harvest begins. *Nat. Genet.*, **31**, 235–236.
2. Pflieger,S., Lefebvre,V. and Causse,M. (2001) The candidate gene approach in plant genetics: a review. *Mol. Breeding*, **7**, 275–291.
3. Borevitz,J.O. and Chory,J. (2004) Genomics tools for QTL analysis and gene discovery. *Curr. Opin. Plant. Biol.*, **7**, 132–136.
4. Liang,C., Jaiswal,P., Hebbard,C., Avraham,S., Buckler,E.S., Casstevens,T., Hurwitz,B., McCouch,S., Ni,J., Pujar,A. *et al.* (2008) Gramene: a growing plant comparative genomics resource. *Nucleic Acids Res.*, **36**, D947–D953.
5. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
6. Schuler,G.D. (1997) Sequence mapping by electronic PCR. *Genome Res.*, **7**, 541–550.
7. 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.
8. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
9. Garcia-Hernandez,M., Berardini,T.Z., Chen,G., Crist,D., Doyle,A., Huala,E., Knee,E., Lambrecht,M., Miller,N., Mueller,L.A. *et al.* (2002) TAIR: a resource for integrated Arabidopsis data. *Funct. Integr. Genomics*, **2**, 239–253.
10. Yu,H., Luscombe,N.M., Lu,H.X., Zhu,X., Xia,Y., Han,J.-D.J., Bertin,N., Chung,S., Vidal,M. and Gerstein,M. (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.*, **14**, 1107–1118.
11. Zmasek,C. and Eddy,S. (2002) RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, **3**, 14.
12. Wanchana,S., Thongjuea,S., Ulat,V.J., Anacleto,M., Mauleon,R., Conte,M., Rouard,M., Ruiz,M., Krishnamurthy,N., Sjolander,K. *et al.* (2008) The generation challenge programme comparative plant stress-responsive gene catalogue. *Nucleic Acids Res.*, **36**, D943–D946.
13. Mungall,C.J. and Emmert,D.B. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.
14. Wilkinson,M., Schoof,H., Ernst,R. and Haase,D. (2005) BioMOBY successfully integrates distributed heterogeneous bioinformatics web services. The PlaNet exemplar case. *Plant Physiol.*, **138**, 5–17.
15. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003)

Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
16. Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.
17. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.
18. Labarga,A., Valentin,F., Anderson,M. and Lopez,R. (2007) Web services at the European Bioinformatics Institute. *Nucleic Acids Res.*, **35**, W6–W11.