Nucleic Acids Research, 2006, Vol. 34, Database issue ***D741–D744***
doi:10.1093/nar/gkj094

The Rice Annotation Project Database (RAP-DB): hub for *Oryza sativa* ssp. *japonica* genome information

Hajime Ohyanagi[1,2], Tsuyoshi Tanaka[3], Hiroaki Sakai[3], Yasumasa Shigemoto[1,4], Kaori Yamaguchi[5], Takuya Habara[5,6], Yasuyuki Fujii[5,6], Baltazar A. Antonio[3], Yoshiaki Nagamura[3], Tadashi Imanishi[6], Kazuho Ikeo[1], Takeshi Itoh[3,6,*], Takashi Gojobori[1,6] and Takuji Sasaki[3]

[1]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, 1111 Yata, Mishima, Shizuoka 411-8540, Japan, [2]Tsukuba Division, Mitsubishi Space Software Co., Ltd 1-6-1 Takezono, Tsukuba, Ibaraki 305-0032, Japan, [3]Genome Research Department, National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan, [4]Life Science Systems Division, Fujitsu Limited, 1-17-25 Shinkamata, Ota-ku, Tokyo 144-8588, Japan, [5]Japan Biological Information Research Center, Japan Biological Informatics Consortium, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan and [6]Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan

Received August 15, 2005; Revised and Accepted October 16, 2005

ABSTRACT

With the completion of the rice genome sequencing, a standardized annotation is necessary so that the information from the genome sequence can be fully utilized in understanding the biology of rice and other cereal crops. An annotation jamboree was held in Japan with the aim of annotating and manually curating all the genes in the rice genome. Here we present the Rice Annotation Project Database (RAP-DB), which has been developed to provide access to the annotation data. The RAP-DB has two different types of annotation viewers, BLAST and BLAT search, and other useful features. By connecting the annotations to other rice genomics data, such as full-length cDNAs and *Tos17* mutant lines, the RAP-DB serves as a hub for rice genomics. All of the resources can be accessed through http://rapdb.lab.nig.ac.jp/.

INTRODUCTION

Rice is considered a model cereal plant because of its small genome size and high degree of chromosomal co-linearity with other major cereal crops such as maize, wheat, barley and sorghum (1,2). The International Rice Genome Sequencing Project (IRGSP), a consortium of publicly funded laboratories from 10 countries, initiated the sequencing of *Oryza sativa* ssp. *japonica* cultivar Nipponbare in 1998 using the clone-by-clone sequencing strategy (2). In 2004, the finished-quality sequence of the entire genome was completed and is now available in the public domain (3).

The annotation of the sequence is indispensable in understanding the overall structure and function of the rice genome. However, most of the annotations of the rice genome sequences were obtained by automated methods. Although this provides an overview of the composition of the genes that comprise the genome, limitations in prediction programs often result in probable errors and artifacts among predicted genes. Therefore, in concordance with the completion of the rice genome sequence, the Rice Annotation Project (RAP) was organized in 2004 (T. Itoh *et al.*, manuscript in preparation) with the aim of providing standardized and highly accurate annotations of the rice genome.

To facilitate efficient management of the results of annotation and to establish a platform for integrating the data with other rice resources, an annotation database called the RAP Database (RAP-DB) was developed. The RAP-DB integrates the IRGSP genome sequence and the RAP annotations with other data on rice researches, and makes them available to the public through HTTP access.

*To whom correspondence should be addressed. Tel: +81 29 838 7065; Fax: +81 29 838 7065; Email: taitoh@affrc.go.jp

© 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

DATABASE CONTENTS

The RAP-DB contains the IRGSP genome sequence (build 3 assembly) (3) and the RAP loci with corresponding locus IDs representing the annotated genes. Each locus has one or more variant transcript(s) as RAP annotated sequence(s). Predicted protein-coding regions were also employed as RAP predicted loci. The TIGR-transcripts derived from the annotations on the TIGR assembly (4) were added to the RAP-DB by mapping them to the IRGSP genome. Each RAP transcript has the following links: Gene Ontology, motif domain information, full-length cDNA information (5) and so on. Among them, full-length cDNAs are anticipated to be invaluable for rice researches (Figure 1) by providing good evidence of physical clones, and facilitating future experimental researches. Hyperlinks to the *Tos17*-flanking sequence positions on the chromosomes (6) should be quite useful for application in clarifying gene functions (Figure 1). The RAP-DB also contains a repeat-masked version of the IRGSP genome sequence build 3 as the reference genome sequence for the annotations.

SYSTEM ARCHITECTURE

The RAP-DB was implemented on PC servers with RedHat Enterprise Linux ES Version 3, Apache web server, MySQL Database server and GBrowse (7). Other common utilities for UNIX were appropriately installed on the servers if necessary. In order to implement the G-integra system, a modified version provided from the H-Invitational Database (H-InvDB) (8) was used. All of the RAP-DB resources are stored in the servers and available through HTTP access.

DATA ACCESS

The primary concept of the RAP-DB is to provide simple access for the IRGSP genome sequence and the RAP annotations. Furthermore, the RAP-DB enables integrative access for other rice resources, which will establish a hub for *O.sativa* ssp. *japonica* genomics (Figure 1). One of the entry points of the database is search by keywords (http://rapdb.lab.nig.ac.jp/). Descriptions and IDs (http://rapdb.lab.nig.ac.jp/note.html#nomenclature) of the annotations are searched. The other entry points are sequence similarity searches (for details see below).

Annotation browser

All the descriptions of the functional annotations and other related information can be viewed through GBrowse (Figure 2A and B), which provides the main features of the RAP-DB and gives chromosome-oriented access (Figure 2A) for the genome sequence and the annotations. Results of keyword or sequence similarity search are automatically hyperlinked to corresponding annotations stored in GBrowse. GBrowse is a Generic Genome Browser originally developed by Stein *et al*. (7) whose characteristics are a combination of a relational database and interactive web pages for manipulating and displaying annotations on genomes. An annotation table corresponding to each transcript is also available by clicking on each glyph (Figure 2B). The table is composed of multiple rows that includes Gene Ontology information, motif domain information and so on. Links are provided to other useful databases such as the full-length cDNAs (5) and *Tos17* mutant lines (6), and thereby the RAP-DB functions as a hub for rice genome information. Moreover, SVG images are generated, so that the user can edit the graphics of the genomic view.

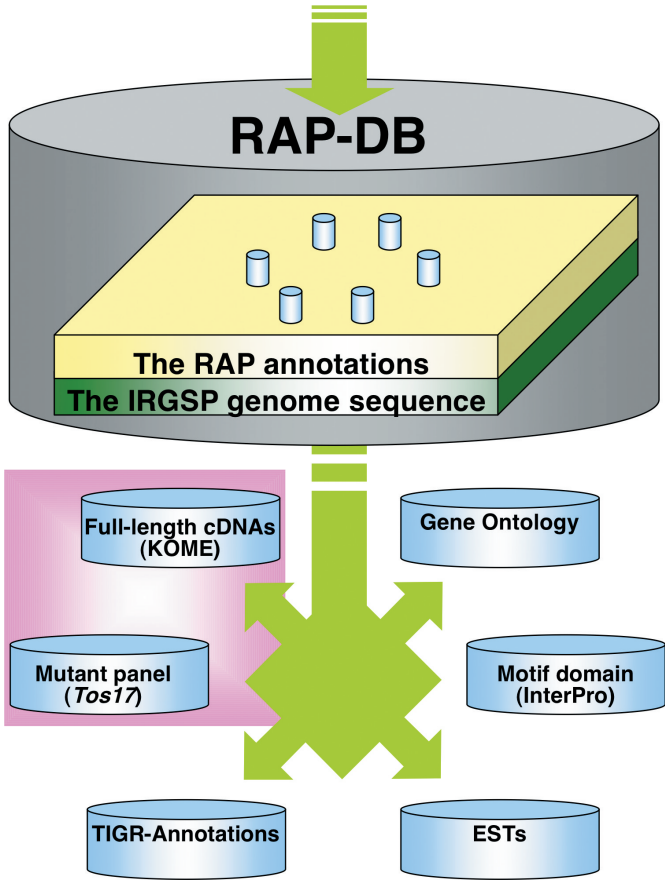

Figure 1. Overview of the hub for rice genome information. The RAP-DB contains the high-quality rice genome sequence generated by the IRGSP, curated annotations of identified or predicted genes, and links to other databases (represented by small blue cylinders). Integrative access to all the information (represented by large blue cylinders), such as experimental evidence, is facilitated.

Genome viewer

Genome-scale view of the annotation and comparison of transcripts with those of other species are available through the G-integra system (Figure 2C), which was originally developed as a part of the H-InvDB (8). G-integra is implemented so as to facilitate parallel access for the RAP annotations and numbers of tracks for other species (cDNAs and expressed sequence tags of representative monocots and *Arabidopsis thaliana* and the like). G-integra and GBrowse are reciprocally hyperlinked and hence the user can easily access both information.

Sequence similarity search

To facilitate access by sequence similarities, two alternative search methods are available (Figure 2). One is BLAT for aligning a given DNA against the genome (9). Hits reported

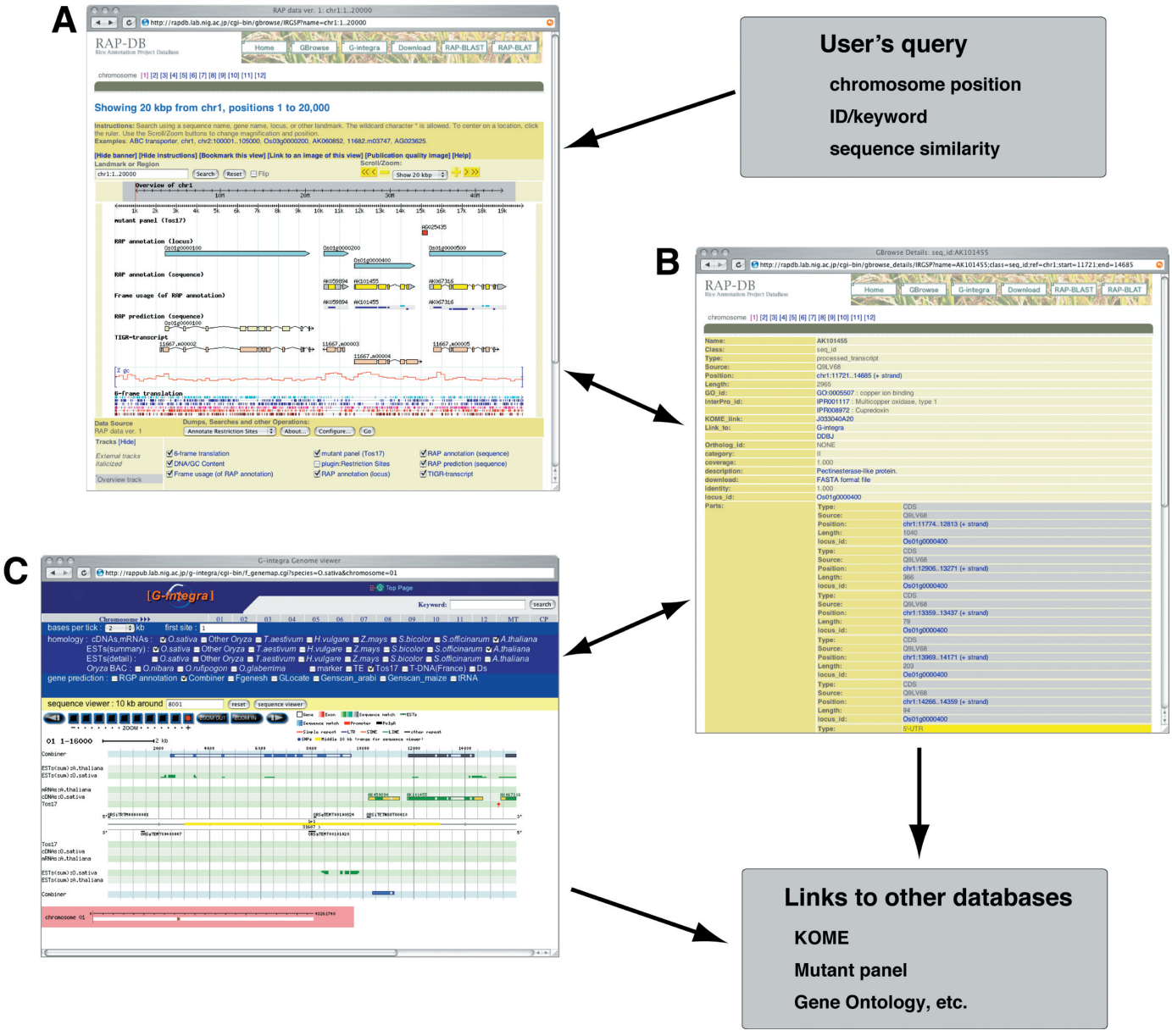

Figure 2. Flowchart of RAP-DB browsing. Users can search the rice genome annotations by chromosomal position, ID or keyword. Sequence similarity search by RAP-BLAST or RAP-BLAT is also available (see text). (**A**) A graphical view of the RAP annotated loci and sequences, *Tos17*-flanking positions, and other tracks illustrated by GBrowse. (**B**) An annotation table corresponding to the sequence with hyperlinks to other databases. (**C**) Browsing a precise genomic view by G-integra.

by BLAT are automatically hyperlinked to the corresponding regions in GBrowse. The other is BLAST (10), which is used for searching transcripts and open reading frames. Hits reported by BLAST are automatically hyperlinked to the corresponding annotation tables in GBrowse.

Distributed annotation system (DAS)

Although we wish to use the IRGSP genome and the RAP annotations as the standard references for future rice genomics, it will be of the rice community's benefit to utilize them for third party annotations. Therefore, we made them available through the DAS protocol (11). The URL for the IRGSP genome reference server is http://rapdb.lab.nig.ac.jp/cgi-bin/das/IRGSP.

FUTURE DIRECTION

The annotations of the rice genome sequence will be updated as the genome sequence and cDNA sequences are revised. The latest version of the high-quality rice genome sequence (build 4 assembly) has been released recently (T. Sasaki, personal communication). This assembly will be used to update the manual curation of annotation in conjunction with the Second RAP Meeting (RAP2). It is therefore expected to generate

additional loci as well as modifications on previous annotations. In addition, we will increase the links for other valuable databases to provide multiple access to various genome information. The RAP-DB will be a bridge to connect the rice genome informatics and the experimental genomics, and an important hub for rice genomics.

ACKNOWLEDGEMENTS

We thank the IRGSP and RAP members for their supports. This work was supported in part by a grant from the Special Coordination Funds for Promoting Science and Technology of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan. Funding to pay the Open Access publication charges for this article was provided by a grant for the NIAS Genebank Project.

Conflict of interest statement. None declared.

REFERENCES

1. Moore,G., Devos,K.M., Wang,Z. and Gale,M.D. (1995) Cereal genome evolution. Grasses, line up and form a circle. *Curr. Biol.*, **5**, 737–739.
2. Sasaki,T. and Burr,B. (2000) International Rice Genome Sequencing Project: the effort to completely sequence the rice genome. *Curr. Opin. Plant Biol.*, **3**, 138–141.
3. International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
4. 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.
5. Kikuchi,S., Satoh,K., Nagata,T., Kawagashira,N., Doi,K., Kishimoto,N., Yazaki,J., Ishikawa,M., Yamada,H., Ooka,H. *et al.* (2003) Collection, mapping, and annotation of over 28 000 cDNA clones from japonica rice. *Science*, **301**, 376–379.
6. Miyao,A., Tanaka,K., Murata,K., Sawaki,H., Takeda,S., Abe,K., Shinozuka,Y., Onosato,K. and Hirochika,H. (2003) Target site specificity of the *Tos17* retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome. *Plant Cell*, **15**, 1771–1780.
7. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
8. Imanishi,T., Itoh,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M. *et al.* (2004) Integrative annotation of 21 037 human genes validated by full-length cDNA clones. *PLoS Biol.*, **2**, 856–875.
9. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
10. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
11. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.