Rice TOGO Browser: A Platform to Retrieve Integrated Information on Rice Functional and Applied Genomics

Yoshiaki Nagamura1,5,*, Baltazar A. Antonio1,5, Yutaka Sato1,5, Akio Miyao1, Nobukazu Namiki2, Jun-ichi Yonemaru3, Hiroshi Minami2, Kaori Kamatsuki2, Kan Shimura2, Yuji Shimizu2 and Hirohiko Hirochika4

1Genome Resource Center, National Institute of Agrobiological Sciences, Kannondai 2-1-2, Tsukuba, Ibaraki, 305-8602 Japan
2Mitsubishi Space Software Co. Ltd., Takezono 1-6-1, Tsukuba, Ibaraki, 305-0032 Japan
3QTL Genomics Research Center, National Institute of Agrobiological Sciences, Kannondai 2-1-2, Tsukuba, Ibaraki, 305-8602 Japan
4Division of Genome Research and Biodiversity, National Institute of Agrobiological Sciences, Kannondai 2-1-2, Tsukuba, Ibaraki, 305-8602 Japan
5These authors contributed equally to this work.
*Corresponding author: E-mail, nagamura@nias.affrc.go.jp; Fax, +81-29-838-7408

(Received October 7, 2010; Accepted December 13, 2010)

The Rice TOGO Browser is an online public resource designed to facilitate integration and visualization of mapping data of bacterial artificial chromosome (BAC)/P1-derived artificial chromosome (PAC) clones, genes, restriction fragment length polymorphism (RFLP)/simple sequence repeat (SSR) markers and phenotype data represented as quantitative trait loci (QTLs) onto the genome sequence, and to provide a platform for more efficient utilization of genome information from the point of view of applied genomics as well as functional genomics. Three search options, namely keyword search, region search and trait search, generate various types of data in a user-friendly interface with three distinct viewers, a chromosome viewer, an integrated map viewer and a sequence viewer, thereby providing the opportunity to view the position of genes and/or QTLs at the chromosomal level and to retrieve any sequence information in a user-defined genome region. Furthermore, the gene list, marker list and genome sequence in a specified region delineated by RFLP/SSR markers and any sequences designed as primers can be viewed and downloaded to support forward genetics approaches. An additional feature of this database is the graphical viewer for BLAST search to reveal information not only for regions with significant sequence similarity but also for regions adjacent to those with similarity but with no hits between sequences. An easy to use and intuitive user interface can help a wide range of users in retrieving integrated mapping information including agronomically important traits on the rice genome sequence. The database can be accessed at http://agri-trait.dna.affrc.go.jp/.

**Keywords:** Applied genomics • BLAST • Database • Functional genomics • Rice • Structural genomics.

**Abbreviations:** BAC, bacterial artificial chromosome; BLAST, basic local alignment search tool; IRGSP, International Rice Genome Sequencing Project; PAC, P1-derived artificial chromosome; QTL, quantitative trait locus; RFLP, restriction fragment length polymorphism; SNP, single nucleotide polymorphism; SSR, simple sequence repeat.

**Introduction**

Rapid advances in molecular biology and the proliferation of large-scale projects on genome analysis in the last 20 years have brought about a tremendous increase in the amount of genomics information in various species. In rice, initial efforts to characterize the genome structure by genetic and physical mapping, subsequent sequencing of the entire genome and large-scale analyses of gene function by forward and reverse genetics led to the accumulation of a large amount of data and the proliferation of primary and secondary databases on rice genomics. Central to these genomics efforts are the rice genome sequence databases, namely the Rice Annotation Project Database (Ohyanagi et al. 2006, Rice Annotation Project 2007, Rice Annotation Project 2008) and the MSU/TIGR Rice Genome Annotation Database (Ouyang et al. 2007) based on the genome sequence of *Oryza sativa* ssp. *japonica* cultivar Nipponbare generated by the International Rice Genome Sequencing Project (2005). These databases were conceptualized with the aim of providing comprehensive information on the rice genome including functional annotation for the gene models obtained using sequence similarity with nucleotide sequences, amino acid sequences and proteins in the public domain, and/or the presence of domains, full-length cDNA and expressed sequence tag (EST) alignments.
with other plant sequences. Although a large amount of data available in public databases have been invaluable in molecular biology and functional genomics in particular (Kurata and Yamazaki 2006, Liang et al. 2008), only a few databases have been developed so far to facilitate more efficient utilization of genomics data particularly in applied aspects of genomics. Most of these databases such as the RiceGeneThresher (Thongjuea et al. 2009), Gramene QTL Database (Ni et al. 2009) and Q-TARO Database (Yonemaru et al. 2010) focused on quantitative trait locus (QTL) information available for rice. With advances in genome research, comprehensive data integration has become indispensable in order to cope with the rapid proliferation of various types of genomics data and to provide a robust platform for interdisciplinary visualization and interpretation of these data. From the point of view of applied genomics, the integrated information such as genome sequence data, genetic markers, gene annotations, expression profiles and phenotypes should be presented in an easily accessible and interpretable platform so that a wide range of target users can utilize the information effectively. A sequence-based integrated database, INE (INtegrated Rice Genome Explorer) has been developed as a common repository of rice genome data in conjunction with the international sequencing collaboration (Sakata et al. 2000). The database, however, limits the utilization of the information to characterization of the entire genome or specific regions of the genome based on genetic mapping, physical mapping and genome sequence data. Therefore, we have developed the Rice T Of G Browser with the aim of integrating rice genomics data in a user-friendly system to maximize the application of a large amount of information on structural and functional genomics derived from genetic mapping, genome sequencing, genome annotation, expression profiling and QTL analysis. A detailed view of the entire genome from the chromosomal level to the nucleotide sequence level can be extremely useful in retrieving a wide range of information on rice genomics. Furthermore, the Rice T Of G Browser also provides a graphical viewer for BLAST search in order to survey any region with or without statistically significant similarity between sequences. Here, we introduce the Rice TOGO Browser, a database for retrieving integrated information for functional genomics as well as applied genomics.

### Database construction

As an integrated database system, the Rice TOGO Browser contains data derived from genetic mapping, physical mapping and genome sequencing of rice to serve as the template for the search and view functions (Table 1). The database was constructed by PostgreSQL, an object-relational database management system using the RAP2 representative data in general feature format (GFF) derived from the RAP-DB (Rice Annotation Project 2008). A GIN (generalized inverted-index)-based index was created for the attributes field to allow fast full-text search in PostgreSQL. In addition to the alignment information on the genome, mRNAs and proteins derived from RAP2 representative data, sequence information for 1 and 2 kb regions upstream of the transcription start site of each gene and options for downloading the sequence information of all genes or selected genes via the search function were also incorporated. The viewers in the Rice TOGO Browser as described below were originally developed using PHP script. A total of 1,712 restriction fragment length polymorphism (RFLP) markers from the genetic linkage map (http://rgp.dna.affrc.go.jp/) and 15,623 simple sequence repeat (SSR) markers (McCouch et al. 2002) were mapped onto the IRGSP (International Rice Genome Sequencing Project) build04 pseudomolecules using the respective nucleotide sequence registered in the public domain as the query for BLASTN. The marker information can be viewed and downloaded through GBrowse as described below.

### Database components and features

#### Web interface

The Rice TOGO Browser can be accessed through a user-friendly web interface (http://agri-trait.dna.affrc.go.jp/) that provides three search options, namely, keyword search, region search, and trait search, to retrieve information on specific genes, sequences, genetic markers and phenotypes associated with a specific region of the genome. Information on a particular region of the rice genome can be viewed via a chromosome viewer, integrated map viewer and sequence

### Table 1 Integrated data/databases and/or linked data in Rice TOGO Browser

<table>
<thead>
<tr>
<th>Information</th>
<th>Integrated</th>
<th>Original data/databases</th>
<th>Link</th>
<th>URLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome sequence</td>
<td>Yes</td>
<td>IRGSP (Build 4.0)</td>
<td>No</td>
<td><a href="http://rgp.dna.affrc.go.jp/E/IRGSP/Build4/build4.html">http://rgp.dna.affrc.go.jp/E/IRGSP/Build4/build4.html</a></td>
</tr>
<tr>
<td>Genome annotation</td>
<td>Yes</td>
<td>RAP-DB</td>
<td>Yes</td>
<td><a href="http://rapdb.dna.affrc.go.jp/">http://rapdb.dna.affrc.go.jp/</a></td>
</tr>
<tr>
<td>PAC/BAC clones</td>
<td>Yes</td>
<td>WhoGA</td>
<td>Yes</td>
<td><a href="http://rgp.dna.affrc.go.jp/whooga/">http://rgp.dna.affrc.go.jp/whooga/</a></td>
</tr>
<tr>
<td>RFLP markers</td>
<td>Yes</td>
<td>RGP</td>
<td>No</td>
<td><a href="http://rgp.dna.affrc.go.jp/E/publicdata/geneticmap2000/">http://rgp.dna.affrc.go.jp/E/publicdata/geneticmap2000/</a></td>
</tr>
<tr>
<td>SSR markers</td>
<td>Yes</td>
<td>Gramene</td>
<td>No</td>
<td><a href="http://www.gramene.org/">http://www.gramene.org/</a></td>
</tr>
<tr>
<td>QTLs</td>
<td>Yes</td>
<td>Q-TARO</td>
<td>Yes</td>
<td><a href="http://qtaro.abr.affrc.go.jp/">http://qtaro.abr.affrc.go.jp/</a></td>
</tr>
<tr>
<td>Gene expression profiles</td>
<td>No</td>
<td>RiceXPro</td>
<td>Yes</td>
<td><a href="http://ricepro.dna.affrc.go.jp/">http://ricepro.dna.affrc.go.jp/</a></td>
</tr>
<tr>
<td>Comparative genomics data</td>
<td>No</td>
<td>SALAD database</td>
<td>Yes</td>
<td><a href="http://salad.dna.affrc.go.jp/salad/">http://salad.dna.affrc.go.jp/salad/</a></td>
</tr>
</tbody>
</table>
viewer (Fig. 1). The chromosome viewer provides an overview of the relative chromosome position of selected genes and traits in the entire genome of rice. The integrated map viewer consists of the genetic map for each chromosome with the linkage position of RFLP markers in centiMorgans (cM) and the physical map representing each chromosome as a contiguous stretch of DNA sequence based on base pair positions. The window representing a chromosome region can be scrolled for a detailed view of the overall features of specific regions of the chromosome. From the physical map, a user can obtain information on the BAC/PAC clones that comprise the minimum tiling path such as the clone ID, accession ID and physical position in the chromosome. The control panel on top of the viewer allows the user to select a chromosome, adjust the zoom level and specify the area or chromosomal region for viewing. The sequence viewer displays the nucleotide sequence and all genes identified in the selected region with the position of introns, coding regions (CDS) and untranslated regions (UTRs) of genes, and overlap of sense and antisense strand genes shaded in color. Links to RAP-DB, WhoGA and the RFLP/SSR markers provide detailed annotation of and/or marker information on the selected region. Moving the mouse over the colored region opens a pop-up window with information on annotation such as locus ID, accession ID and description. The pop-up window provides an opportunity for retrieving various types of alignment information on a gene such as genome sequences, mRNA sequences, protein sequences, 1 kb upstream sequences and 2 kb upstream sequences.

Search options

Keyword search. Keyword search can be used for RAP-DB locus ID, accession number, gene name, any word included in the gene description, InterPro ID and gene ontology. Fig. 1A shows the chromosome view generated by keyword search with ‘Os06g0275000’. The chromosome viewer displays the relative position of Os06g0275000 in chromosome 6. The annotation of this gene and various types of alignment information, such as position/description list, genomic sequences, mRNA sequences, protein sequences, upstream 1 kb sequences and upstream 2 kb sequences, are available via a download option. The integrated map viewer provides more detailed information on the genome region corresponding to Os06g0275000 indicated in a red box (bar) as the sequence from 9,335,377 to 9,337,570 bp in chromosome 6 (Fig. 1B). The upper panel on the viewer helps to operate the range defined by the red box to obtain the sequence information of a preferred region. Fig. 1C shows the sequence viewer generated by setting ‘Select Area’ with the physical position as ‘9,334,377–9,338,570’, which includes adjacent 1 kb sequences both upstream and downstream of Os06g0275000. This flexible option for region selection facilitates users obtaining any sequence information such as a promoter sequence of a gene of interest.

Region search. Region search is a useful tool for map-based cloning, and provides two options, ‘Region search by marker’ and ‘Region search by sequence’. A total of 1,712 RFLP markers and 15,623 SSR markers are available for the marker-based search system. Fig. 2A shows the flow of the region search using ‘RM10010’ and ‘RM10026’ as markers. A tabular list of genes between the two markers is generated and provides direct links to RAP-DB for annotation details, RiceXPro for gene expression information (Sato et al. 2011) and the SALAD database for comparative analysis of conserved amino acid sequences or motifs (Mihara et al. 2010) so that users can survey available information on a gene or genes of interest to facilitate forward genetics approaches (Table 1). Furthermore, the links to ‘RFLP/SSR marker’ and the ‘sequence viewer’ on the integrated map viewer could be used to obtain additional information on markers in the selected region (Supplementary Fig. S1) and structural features of the sequence. In this way, users can easily obtain the genome sequence and marker information located between the two markers, thereby helping to narrow down the region harboring a target gene for further linkage analysis.

The ‘Region search by sequence’ option provides information on the region between two distinct sequences with a length >15-mers. This option basically provides the longest region between the two sequences in a chromosome (if both sequences are perfectly matched to a multiple region in a chromosome) or in every chromosome (if both sequences are perfectly matched to multiple chromosomes). Therefore, the sequence-based region search is an extremely useful tool to retrieve the preferred region information using any user-designed primer sequence.

Trait search. The Rice TOGO Browser contains information on 1,051 representative QTLs for morphological trait, physiological trait and resistance or tolerance which has been published in the Q-TARO database (Yonemaru et al. 2010). Fig. 2B shows the view generated by a trait search for ‘cold tolerance’. The trait search option provides a chromosome view of the position of the QTLs and a tabular list of information on the position, QTL name, associated trait and a link to Q-TARO. The Rice TOGO Browser therefore facilitates visualization of multiple QTLs and their distribution in the entire genome. In the integrated map viewer, the user can further survey the localization of a QTL/multiple QTLs in a limited genome region. Moving the mouse over the trait image opens a pop-up window for information on physiological position, category, QTL, character and link to the Q-TARO database, and in the case of multiple QTLs, further provides the information of the detailed position of each QTL by coloring the position. Furthermore, users can obtain any sequence and marker information associated with a trait in a similar manner to the keyword or region search.
Fig. 1 The Rice TOGO Browser consists of three viewers: (A) a chromosome viewer that provides an overview of the relative chromosome position of selected genes and traits in the entire genome of rice; (B) an integrated map viewer for the genetic and physical mapping information; and (C) a sequence viewer for a detailed characterization of structural features of the genomic sequence of a gene or chromosomal region.
Fig. 2 Search for a specified region or trait. (A) The result display generated by a search for a genomic region flanked by markers (RM10010 and RM10026) can be viewed as a tabular list and downloaded as a text file. The integrated map viewer with links to ‘RFLP/SSR marker’ and the sequence viewer allow a user to obtain additional information on markers and genome sequence in the selected region. (B) The search for a trait is initiated by selecting a trait category. The chromosome viewer displays the position of the QTLs associated with a phenotypic trait.
Fig. 3 Graphical display of a BLAST NAVI search using the sequence OsNAC2 (AB028181) as a query and the ‘Rice full-length cDNA clones (KOME)’ as the subject. (A) The search result is provided in a graphical view with a designated color for each range of similarity score and a tabular list of the details of subject sequences with similarity to the query sequence. Both AK061745 and AK104626 are derived from the same gene and included no-hit regions in the 3’ end shown in gray. For AK071020, the highest scoring segment pair was obtained only in the 3’ half of the sequence. The ‘View Alignment’ link provides a detailed graphical representation of the alignment for each subject sequence. (B) A detailed graphical view of the high scoring segment pair between the query sequence (AB028181) and AK071020. The similarity of entire regions between the sequences can be examined in detail via two graphical viewers showing the high-scoring regions of the subject sequence mapped to the query sequence (B-1) or the high-scoring regions of the query sequence mapped to the subject sequence (B-2). Clicking the ‘Mapping Hit to Subject’ or ‘Mapping Hit to Query’ button respectively, located at the bottom right corner of the graphical viewer, facilitates this function. (C) Viewer for the no-hit region and a link for downloading the sequence.
Graphical viewer for surveying an entire region with or without significant similarity

The BLAST finds local regions with statistically significant similarity between sequences and provides an opportunity to infer the function and evolutionary relationship as well as the identity of family genes (Altschul et al. 1990). The conventional BLAST and viewers developed for BLAST such as BLASTScope (http://www.shigen.nig.ac.jp/tools/blastscope/top.jsp), however, provide a graphical view of the matched sequences only, without any information whatsoever on sequences corresponding to no-hit regions adjacent to high scoring segment pairs. These no-hit region sequences may also provide notable information that can be used to distinguish between genes, particularly in regions that contain gene clusters. Therefore, we developed a graphical display for surveying an entire region with or without significant similarity in a homology search of BLAST, named BLAST NAVi (BLAST No similarity region Analysis Viewer), to facilitate analysis of sequence regions with a hit as well as regions with no hit between sequences. In our database, a homology search by the BLAST can be specified from among five databases, namely ‘Oryza sativa mRNA Database’, ‘Oryza sativa EST Database’, ‘Japonica rice full-length cDNA Database (KOME)’, ‘Rice Genome Database-japonica’ and ‘All Rice Genome Database’, as a subject sequence, thereby facilitating target-oriented and rapid homology search of any given sequence data. The Rice TOGO Browser also contains two BLAST programs, Blastn for comparing a nucleotide query sequence against a nucleotide sequence database and tblastn for comparing a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames. The BLAST NAVi provides a graphical view of the length of each subject sequence and the region of the highest scoring segment pair, with a designated color representing the score, and a tabular list showing the summary of the subject sequences and the details of the BLAST search, i.e. definition, score (bits), e-value, query homology rate and subject homology rate, with descending order of the scores. The accession number provides the link to the Sequence Retrieval System (SRS) of the NIAS DNA bank (http://srs.dna.affrc.go.jp/srs8/) for a more detailed description. Fig. 3A shows a BLAST NAVi search using the sequence of OsNAC2 (AB028181) as a query and the sequences in the ‘Japonica rice full-length cDNA Database (KOME)’ database as the subject. The search result indicates that both AK061745 and AK104626 have high similarity to OsNAC2, although a short segment in the 3’ end shows no hit. On the other hand, AK071020 contains the highest scoring segment pair only in the 3’ half of the sequence. From the ‘View Alignment’ link, a user can examine the details of alignment between the query sequence and each subject sequence via two graphical viewers, one showing the alignment of the subject sequence mapped to the query sequence (Fig. 3B-1) and the other showing the alignment of the query sequence mapped to the subject sequence. As shown in Fig. 3B-1, the high scoring segment pair between ‘AB028181’ and ‘AK071020’ indicates that AK071020 has basically the same sequence as AB028181 in the entire region, but with additional sequences corresponding to 120 bp near the center. The sequence insertion can be confirmed by changing the graphical viewer (Fig. 3B-2). Furthermore, the positions of the sequences with no alignments are indicated in the ‘No Hit Region’ column and can be downloaded as fasta format by clicking the corresponding region (Fig. 3C). The maximum length for the download option is 500 bp. Comparing a query sequence and a subject sequence particularly at no-hit regions may reveal important information that can be used for accurate analysis of the sequence. Furthermore, the no-hit region sequence between genes with a similar alignment can be used in designing specific primers for PCR and microarray probes.

Conclusion

An integrated database on rice genomics is indispensable for efficient utilization of a wide range of data on rice genomics. The current version of Rice TOGO Browser with various genomic data provides a wide range of users with a framework for comprehensive analysis of genes or target regions to facilitate more efficient forward genetics strategies. The database will be further enhanced with the integration of single nucleotide polymorphism (SNP) data derived from analysis of the Japanese rice cultivars and the sequence information that has been generated with the sequencing of the japonica cultivar Koshihikari (Yamamoto et al. 2010). The integration of various types of genomics information will further enhance the utility of the rice genome sequence and may provide substantial resources for developing novel strategies that will complement traditional breeding methods with the advances in rice genomics.

Supplementary data

Supplementary data are available at PCP online.

Funding

This work was supported by the Ministry of Agriculture, Forestry and Fisheries of Japan [MAFF Life Science Integrated Database System].

Acknowledgments

We would like to thank Dr. Masahiro Yano and Dr. Kiyosumi Hori (National Institute of Agrobiological Sciences) for supplying SSR marker information, and Mr. Satoshi Nobushima, Mr. Hiroshi Ikawa and Dr. Hajime Ohyanagi (Mitsubishi Space Software Co. Ltd.) for their help in database construction.
References


