doi:10.1093/nar/gki014

RARGE: a large-scale database of RIKEN *Arabidopsis* resources ranging from transcriptome to phenome

Tetsuya Sakurai[1,2], Masakazu Satou[1,2], Kenji Akiyama[1,2], Kei Iida[1,2], Motoaki Seki[2,3], Takashi Kuromori[2], Takuya Ito[3], Akihiko Konagaya[1], Tetsuro Toyoda[1] and Kazuo Shinozaki[2,3,*]

[1]Bioinformatics Group and [2]Plant Functional Genomics Group, RIKEN Genomic Sciences Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan and [3]Laboratory of Plant Molecular Biology, RIKEN Tsukuba Institute, 3-1-1 Koyadai, Tsukuba 305-0074, Japan

Received June 24, 2004; Revised and Accepted September 16, 2004

ABSTRACT

The RIKEN *Arabidopsis* Genome Encyclopedia (RARGE) database houses information on biological resources ranging from transcriptome to phenome, including RIKEN *Arabidopsis* full-length (RAFL) complementary DNAs (cDNAs), their promoter regions, *Dissociation* (*Ds*) transposon-tagged lines and expression data from microarray experiments. RARGE provides tools for searching by resource code, sequence homology or keyword, and rapid access to detailed information on the resources. We have isolated 245 946 RAFL cDNA clones and collected 11 933 transposon-tagged lines, which are available from the RIKEN Bioresource Center and are stored in RARGE. The RARGE web interface can be accessed at http://rarge.gsc.riken.jp/. Additionally, we report 90 000 new RAFL cDNA clones here.

INTRODUCTION

Arabidopsis thaliana is the most widely studied plant. Sequencing of the *Arabidopsis* genomic sequence was completed in December 2000 by the *Arabidopsis* Genome Initiative (AGI) (1). About 200 000 expressed sequence tags (ESTs) from *Arabidopsis* have been deposited in dbEST of May 2004. In the post-genome-sequencing age, large-scale biological resources, especially full-length complementary DNAs (cDNAs) and tagged mutants, will serve as powerful tools for the identification of gene functions.

We have been collecting RIKEN *Arabidopsis* full-length (RAFL) cDNAs and *Dissociation* (*Ds*) transposon-tagged lines (2–5). We have isolated 245 946 RAFL cDNA clones. Their ESTs are clustered into about 18 000 non-redundant cDNA groups, covering about 70% of predicted genes. These RAFL cDNA resources have proved useful for plant study. We have also obtained 5′ ESTs to construct a promoter database. Using the full-length cDNA microarray, we study the expression profiles of *Arabidopsis* genes in response to various environmental stress conditions and hormone treatments. Furthermore, over 10 000 transposon-tagged lines were constructed by using the *Activator* (*Ac*)/*Ds* system (6) to collect insertional mutants and to build a database for basic phenome analysis.

To enhance the utility of the RIKEN resource data, we have constructed the RIKEN *Arabidopsis* Genome Encyclopedia (RARGE) database. RARGE is a web-based application that provides means for searching resource data. RARGE provides biology researchers access to detailed information on *Arabidopsis* resources (cDNAs, transposon mutants and microarray experiments) that have been produced by the RIKEN Genomic Sciences Center's Plant Functional Genomics Research Group.

RIKEN *ARABIDOPSIS* DATA IN RARGE

The RIKEN *Arabidopsis* data in RARGE include detailed information on RAFL cDNAs, transposon-tagged mutants and RAFL cDNA microarray expression profiles, and can be browsed as described below.

RAFL cDNAs

Full-length cDNAs are useful for the correct annotation of genomic sequences and for the functional analysis of genes and their products. We reported 155 144 RAFL cDNA clones previously (2). Here, we report 90 802 new clones that we have collected. Thus, the total number of the RAFL cDNAs featured in the RARGE database is now 245 946.

*To whom correspondence should be addressed. Tel: +81-29-836-4359; Fax: +81-29-836-9060; Email: sinozaki@rtc.riken.jp

We performed single-pass sequencing of the newly collected cDNA clones from the 3′/5′ end and obtained 265 307 ESTs (3′ ESTs, 172 653; 5′ ESTs, 92 654). The ESTs were mapped onto the *Arabidopsis* genomic sequence using BLAST tools (7) and clustered into 18 090 non-redundant cDNA groups, covering about 70% of predicted genes.

Promoter sequences based on the 5′ ESTs of the RAFL cDNAs

RARGE provides reliable information on transcript promoter sequences, which we obtained by comparing the 5′ end sequences of the RAFL cDNAs with the *Arabidopsis* genomic sequence. We have constructed a promoter database of 16 997 non-redundant RAFL cDNA groups using the PLACE (Plant *cis*-acting regulatory DNA elements) database (8). The 5′ ESTs of the 16 997 RAFL cDNAs were mapped onto the *Arabidopsis* genome using BLAST tools. The criterion for mapping the 5′ ESTs was sequence identity >98% within >100 bp overlap. The genomic sequence that was detected in the region 1000 bp upstream of the 5′ terminus of each RAFL cDNA clone was regarded as the promoter sequence of each clone. We then searched for about 400 known plant *cis*-acting elements in the 1000 bp promoter sequences of genes corresponding to the clones by using PLACE.

Microarray

Expression profiles are important for functional characterization of the RAFL cDNAs. We studied the expression profiles of *Arabidopsis* genes not only in response to various abiotic and biotic stress conditions and hormone treatments but also in various mutants and transgenic plants by using a full-length cDNA microarray (9–13). The expression profiles are available in the RARGE database. The RARGE microarray database should be useful for analyzing the expression profiles of plant genes in response to environmental stress and hormone treatments.

RIKEN *Arabidopsis* transposon-tagged mutants

RIKEN *Arabidopsis* transposon-tagged mutants were constructed by using the *Ac*/*Ds* system to collect insertional mutants as a useful resource for functional genomics of *Arabidopsis*. Previously, we reported about 1000 lines whose *Ac*/*Ds* transposon-flanking sequences were determined by the thermal asymmetric interlaced-PCR method (4). Recently, we increased this number to more than 10 000 independent lines derived from different donor lines (5). We sequenced the flanking regions of the *Ds*-insertional lines and determined each one's unique insertion site. The total number of transposon-tagged lines is now 11 933. The *Ds*-tagged mutants are available from the RIKEN Bioresource Center (http://www.brc.riken.jp/en/).

TOOLS

RARGE provides searching and browsing tools of several kinds, such as 'BLAST' for sequence similarity searching, 'Genome Map' for picking a target resource from a visual map (Figure 1), 'Keyword Search' for filtering records by a chosen word and 'Resource Detailed Information' for browsing individual resource data in the returned results. Since all tools are connected seamlessly, user friendliness and convenience are greatly improved. Brief descriptions of the tools are given below.

BLAST search

The RARGE database provides web access to the BLAST sequence similarity search of the National Center for Biotechnology Information (NCBI). Users can choose several nucleotide or protein sequence datasets to search against. Selectable databases in RARGE include nucleotide and protein sequences related to *Arabidopsis* genome information from the AGI, the non-redundant known protein database from NCBI, RAFL cDNA ESTs, RAFL cDNA full-read sequences and *Ds*-tagged transposon insertion-flanking sequences. Each searched record in the BLAST result page is hyperlinked to 'Resource Detailed Information' (see below) or the GenBank record at NCBI. BLAST datasets are available from the FTP directory (ftp://pfgweb.gsc.riken.jp/).

Genome map

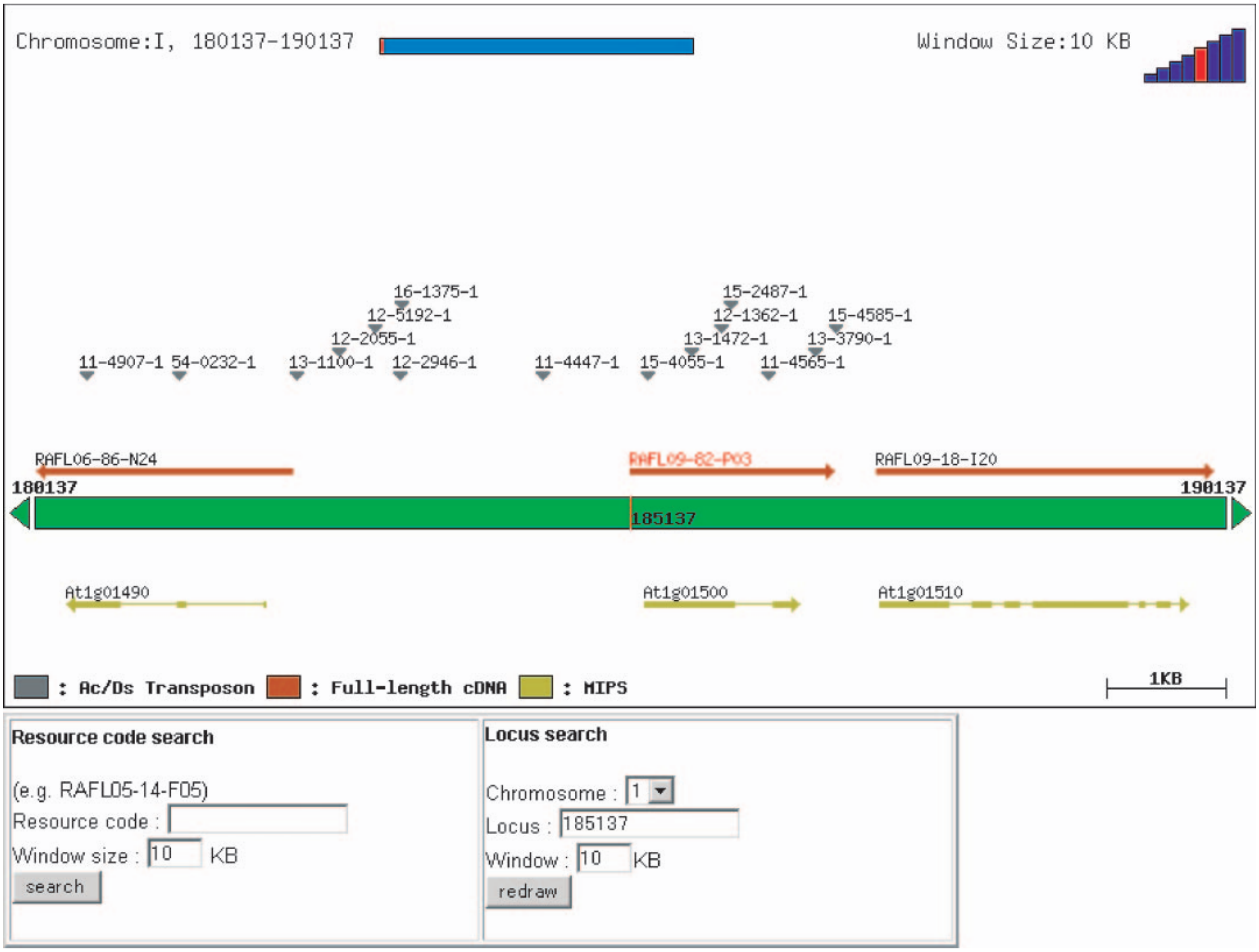
'Genome Map' integrates a view of the *Arabidopsis* genome, predicted genes, RAFL cDNA clones and transposon-tagged mutants (Figure 1). Users can interactively browse clickable objects and select a zoom level from 1 to 200 kb (eight levels). Searches can also be conducted by resource code (e.g. RAFL02-01-A03, 11-0001-1) or locus. Users can click on objects of RAFL cDNA or transposon-tagged mutant and browse the resource details. They can click on the object of predicted genes and browse the gene detail by hyperlinking to the Munich Information Center for Protein Sequences (MIPS) *Arabidopsis thaliana* Database (MAtDB) (14) at the MIPS.

Keyword search

To enable keyword searching, we have assembled a keyword database containing the definition strings of BLAST results against the non-redundant protein database. 'Keyword Search' allows users to specify queries on any word and retrieves relevant information from the keyword database. The keyword search results are summarized on an intermediate page, where all the matches are displayed as a one-line record. Users can access detailed information on the resources in each record by clicking on the object's RAFL cDNA code or transposon-tagged line code. The keyword search result page is also hyperlinked to the GenBank record at NCBI and MAtDB.

Promoter search

The genomic sequence in the region 1000 bp upstream of the 5′ terminus of each RAFL cDNA clone is regarded as the promoter sequence of the clone. We have identified about 400 known plant *cis*-acting elements in the promoter sequences of genes corresponding to the RAFL cDNA clones on the basis of PLACE data (8). We have constructed a promoter database from the results. Users can search for RAFL cDNAs that include a user-specified *cis*-acting element in their promoter sequences.



Resource code search

(e.g. RAFL05-14-F05)
Resource code :
Window size : 10 KB
search

Locus search

Chromosome : 1
Locus : 185137
Window : 10 KB
redraw

Figure 1. Genome map shows *Arabidopsis* chromosome 1 around RAFL09-82-P03 and indicates *Ds*-tagged mutants (gray downward-pointing arrowheads), RAFL cDNAs (red arrow bar), chromosome (green broad bar) and predicted genes from MIPS (khaki arrow bar).

Expression data from microarray experiments

The expression data from microarray experiment profiles are available in RARGE (9–13). It is possible for users to browse not only the expression data by specifying a microarray experiment, a time course, an expression level to display and a data type (log value or ratio value), but also the list of putatively coexpressed genes by specifying multiple experiments. In addition, it is possible to specify RAFL cDNA codes or AGI gene codes in the query to obtain only the relevant expression data.

Resource detailed information

'Resource Detailed Information' presents a summary of chosen objects in the RARGE database. For RAFL cDNA clones, the detailed information includes information on sequence type, matched AGI gene annotation, GenBank accession number, locus, hyperlinks to the sequences and the 5′ upstream sequence. For RIKEN transposon-tagged lines, the detailed information includes information on insertion position, as well as the closest gene data and hyperlinks to the tag-flanking sequences.

Link to the external public databases

Since RARGE has been designed to house the detailed information of the biological resources, which is rather stable information than that of changeable external links, the Genome ⇔ Phenome Superhighway system of RIKEN is conveniently used as the gateway to the changeable external public databases (15).

FUTURE DIRECTIONS

At present, RARGE provides correspondences between the RAFL cDNAs (as the transcriptome) and the *Ds*-tagged knockout lines (as the phenome). It is now possible to study more than 5000 genes in our resource for phenome analysis of mutants with disrupted genes (5). Additional collection of phenotypic observations, such as pictures of various mutant lines and descriptions of individual traits, is now underway and will be published in the RARGE database. Thus, RARGE is the first systematic encyclopedia that features phenotypic functions of each complete full-length cDNA in *Arabidopsis*.

RARGE is an ongoing project and will be updated as new data become available.

ACKNOWLEDGEMENTS

We are grateful to Drs Yoshihide Hayashizaki, Piero Carninci, Joseph R. Ecker, Athanasios Theologis and Ronald W. Davis for their contribution to the collection of large-scale resource data.

REFERENCES

1. *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
2. Seki,M., Narusaka,M., Kamiya,A., Ishida,J., Satou,M., Sakurai,T., Nakajima,M., Enju,A., Akiyama,K., Oono,Y. *et al*. (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science*, **296**, 141–145.
3. Yamada,K., Lim,J., Dale,J.M., Chen,H., Shinn,P., Palm,C.J., Southwick,A.M., Wu,H.C., Kim,C., Nguyen,M. *et al*. (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science*, **302**, 842–846.
4. Ito,T., Motohashi,R., Kuromori,T., Mizukado,S., Sakurai,T., Kanahara,H., Seki,M. and Shinozaki,K. (2002) A new resource of locally transposed dissociation elements for screening gene-knockout lines *in silico* on the *Arabidopsis* genome. *Plant Physiol*., **129**, 1695–1699.
5. Kuromori,T., Hirayama,T., Kiyosue,Y., Takabe,H., Mizukado,S., Sakurai,T., Akiyama,K., Kamiya,A., Ito,T. and Shinozaki,K. (2004) A collection of 11 800 single-copy *Ds* transposon insertion lines in *Arabidopsis*. *Plant J*., **37**, 897–905.
6. Fedoroff,N.V. and Smith,D.L. (1993) A versatile system for detecting transposition in *Arabidopsis*. *Plant J*., **3**, 273–289.
7. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSIBLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
8. Higo,K., Ugawa,Y., Iwamoto,M. and Korenaga,T. (1999) Plant *cis*-acting regulatory DNA elements (PLACE) database. *Nucleic Acids Res*., **27**, 297–300.
9. Seki,M., Narusaka,M., Ishida,J., Nanjo,T., Fujita,M., Oono,Y., Kamiya,A., Nakajima,M., Enju,A., Sakurai,T. *et al*. (2002) Monitoring the expression profiles of 7000 *Arabidopsis* genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. *Plant J*., **31**, 279–292.
10. Seki,M., Ishida,J., Narusaka,M., Fujita,M., Nanjo,T., Umezawa,T., Kamiya,A., Nakajima,M., Enju,A., Sakurai,T. *et al*. (2002) Monitoring the expression pattern of around 7000 *Arabidopsis* genes under ABA treatments using a full-length cDNA microarray. *Funct. Integr. Genomics*, **2**, 282–291.
11. Kimura,M., Yamamoto,Y.Y., Seki,M., Sakurai,T.Satou,M., Abe,T., Yoshida,S., Manabe,K., Shinozaki,K. and Matsui,M. (2003) Identification of *Arabidopsis* genes regulated by high light-stress using cDNA microarray. *Photochem. Photobiol*., **77**, 226–233.
12. Oono,Y., Seki,M., Nanjo,T., Narusaka,M., Fujita,M., Satoh,R., Satou,M., Sakurai,T., Ishida,J., Akiyama,K. *et al*. (2003) Monitoring expression profiles of *Arabidopsis* gene expression during rehydration process after dehydration using ca. 7000 full-length cDNA microarray. *Plant J*., **34**, 868–887.
13. Seki,M., Satou,M., Sakurai,T., Akiyama,K., Iida,K., Ishida,J., Nakajima,M., Enju,A., Narusaka,M., Fujita,M. *et al*. (2003) RIKEN *Arabidopsis* full-length (RAFL) cDNA and its applications for expression profiling under abiotic stress conditions. *J. Exp. Bot*., **55**, 213–223.
14. Schoof,H., Ernst,R., Nazarov,V., Pfeifer,L., Mewes,H.W. and Mayer,K.F. (2004) MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource for plant genomics. *Nucleic Acids Res*., **32**, D373–D376.
15. Toyoda,T. and Wada,A. (2004) Omic space: coordinate-based integration and analysis of genomic phenomic interactions. *Bioinformatics*, **20**, 1759–1765.