 Nucleic Acids Research, 2000, Vol. 28, No. 1 97–101

INE: a rice genome database with an integrated map view

Katsumi Sakata, Baltazar A. Antonio[1], Yoshiyuki Mukai[1], Hideki Nagasaki[1], Yasumichi Sakai[2], Kazuyoshi Makino[2] and Takuji Sasaki*

Rice Genome Research Program (RGP), National Institute of Agrobiological Resources, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan, [1]Institute of the Society for Techno-innovation of Agriculture, Forestry and Fisheries, 446-1 Ippaizuka, Kamiyokoba, Tsukuba, Ibaraki 305-0854, Japan and [2]Mitsubishi Space Software Co., Ltd, 1-17-15 Sengen, Tsukuba, Ibaraki 305-0047, Japan

Received August 31, 1999; Revised and Accepted October 27, 1999

ABSTRACT

The Rice Genome Research Program (RGP) launched a large-scale rice genome sequencing in 1998 aimed at decoding all genetic information in rice. A new genome database called INE (INtegrated rice genome Explorer) has been developed in order to integrate all the genomic information that has been accumulated so far and to correlate these data with the genome sequence. A web interface based on Java applet provides a rapid viewing capability in the database. The first operational version of the database has been completed which includes a genetic map, a physical map using YAC (Yeast Artificial Chromosome) clones and PAC (P1-derived Artificial Chromosome) contigs. These maps are displayed graphically so that the positional relationships among the mapped markers on each chromosome can be easily resolved. INE incorporates the sequences and annotations of the PAC contig. A site on low quality information ensures that all submitted sequence data comply with the standard for accuracy. As a repository of rice genome sequence, INE will also serve as a common database of all sequence data obtained by collaborating members of the International Rice Genome Sequencing Project (IRGSP). The database can be accessed at http://www.dna.affrc.go.jp:82/giot/INE.html or its mirror site at http://www.staff.or.jp/giot/INE.html

INTRODUCTION

The estimated size of the rice genome is 430 Mb, the smallest among the major crops such as maize and wheat belonging to the Gramineae family. The genome organization of cereals exhibits a high degree of synteny (1) so that rice can serve as a model cereal crop for genome analysis. In addition, about half of the world population consumes rice as a staple food. For this rationale, rice is the most appropriate choice as a principal model system for crop genomics.

The Rice Genome Research Program (RGP) was initiated in 1991. Some of the major accomplishments during the first phase of the program include the construction of: (i) a catalog composed of nearly 15 000 expressed genes corresponding to about one-third of the total of all rice genes (2); (ii) a high-density linkage map composed of 2275 DNA markers (3); (iii) a physical map with ordered YAC clones covering ~70% of the whole rice genome (4). All of the current information is presented in the RGP website (URL in Table 1).

In 1998, the second phase of RGP was launched with three main objectives, namely, complete genome sequencing, gene functional analysis, and application of genomics in breeding. This requires efficient utilization of all available information on rice genomics including the major results of the first phase of RGP described above. A complete sequencing of the rice genome is warranted in view of its value for cereal crop genomics. To meet the challenge of undertaking an enormous task, the International Rice Genome Sequencing Project (IRGSP) was established (URL in Table 1). This international collaboration was envisioned to complete sequencing the rice genome in 10 years. As of October 1999, 10 countries including Canada, China, France, India, Japan, Korea, Taiwan, Thailand, UK and the US are actively involved in the project.

One of the objectives of the collaboration is to establish an integrated genome database that will incorporate all rice genome sequences contributed by participating members and to accelerate the release of sequence information to the public via the Internet. The database is also expected to link the sequence information from rice to other genetic resources to be utilized for functional analysis and breeding. The first operational version of the database has been completed and offered as a common resource for members of the international rice genome sequencing organization.

DATABASE FEATURES

The new database is called INE as an abbreviation for INtegrated rice genome Explorer. Pronounced [i-ne], it also refers to the rice plant in the Japanese language. INE was developed based on the following concepts: (i) to contain a wide variety of useful and accurate information on rice genomics including all data generated from RGP; (ii) to serve as repository for the rice genome sequence that will be accumulated from the international sequencing effort; (iii) to provide an integrated resource for structural, functional and applied plant genomics through efficient data access capabilities.

*To whom correspondence should be addressed. Tel: +81 298 38 2199; Fax: +81 298 38 2302; Email: tsasaki@abr.affrc.go.jp

Table 1. Useful URLs for INE and related sites

RGP Homepage	http://www.dna.affrc.go.jp:82/ http://www.staff.or.jp/
INE	http://www.dna.affrc.go.jp:82/giot/INE.html http://www.staff.or.jp/giot/INE.html
Genome Sequencing	http://www.dna.affrc.go.jp:82/GenomeSeq.html http://www.staff.or.jp/GenomeSeq.html
International Rice Genome Sequencing Project (IRGSP)	http://www.dna.affrc.go.jp:82/Seqcollab.html http://www.staff.or.jp/Seqcollab.html
RiceGenes	http://genome.cornell.edu/cgi-bin/WebAce/webace?db=ricegenes
AtDB	http://genome-www3.stanford.edu/atdb_welcome.html

Table 2. Summary of data in INE database

Category	Number of datasets	Note
Genetic marker	2275	2119 (93%) genetic markers were developed by RGP.
RFLP image	1791	Image data containing RFLP band patterns.
Clone sequence	1695	Partial sequence of cDNA clones used as genetic markers.
YAC clone	719	The ordered YAC clones cover ~70% of the whole genome. All YAC clones were developed by RGP.
EST marker	4500	Currently being prepared for assembly.
Annotation	7	The complete sequence and annotation of each PAC clone.
Low quality information	7	The region on PAC with PHRAP value below 40.
Predicted gene	228	A list of genes predicted for each PAC clone with the corresponding amino acid sequence and results of computational analysis.

We compared INE with two representative plant genome databases, namely RiceGenes and AtDB (URLs in Table 1). RiceGenes is a rice genome database constructed by using ACEDB, and contains comprehensive genetic and molecular data including maps, markers, probes, sequences, QTLs as well as profiles of cultivated rice germplasm. The main feature of this database is a comparative mapping display for visualization of homologous segments between rice and other major cereal crops such as maize, oat and Triticeae. This allows the user to determine if a region containing a specific gene is conserved in other grass species. AtDB represents the genome database of *Arabidopsis* and covers information on genetic and physical maps, alleles, locus, clones as well as the sequencing data derived from the *Arabidopsis* Genome Initiative (AGI) (5). It assembles currently available physical map data into a single map for each chromosome thereby providing all genetic and physical mapping information for any given region of the genome. As a primary repository for rice-related information, RiceGenes provides diverse information which could be complemented by INE. In particular the rice genome sequence will provide ultimate information on rice genome structure and function. Since the genome sequence is also a main feature of AtDB, it will also complement INE as the two plants, rice and *Arabidopsis* represents the model system for monocots and dicots, respectively, the two major groups of flowering plants. In terms of the computational method, map viewing is different among these databases. In the case of RiceGenes, the maps are presented as GIF images. In AtDB although a site called Genomic View utilizes some Java programming features, individual maps are presented as static images. Thus with RiceGenes and AtDB, map viewing is accomplished by transmitting the GIF images from the server to the client computer. In INE, however, the genetic and physical maps are entirely constructed by Java programs so that the drawing operations are performed in the client's computer following downloading of the data. This facilitates faster viewing of the integrated maps.

The distinctive features of INE are described as follows.

Extent, density and reliability of data

INE incorporates a large volume of data in various categories, a majority of which is generated from RGP (Table 2). The high-density linkage map with 2275 DNA markers also includes detailed information such as the RFLP image for each probe used in mapping and the partial sequence of the markers. The physical map with ordered YAC clones has a coverage of ~70% of the whole genome and is shown based on the actual physical distance calculated from the insert size of the YACs. The average interval between genetic markers is ~0.67 cm that corresponds to 189 kb when measured from the physical distance. This detailed information on genetic and physical mapping facilitates high-resolution genetic mapping, positional cloning of target genes and genetic dissection of QTLs. Furthermore, approximately 4500 EST (expressed sequence tags) markers are being prepared for assembly in INE. This will further enhance the overall density of markers in the genome. Both genetic and EST markers will allow local comparisons of synteny and provide a platform on which physical maps can be assembled. These utilities also emphasize the importance of high-density markers.

As an integral tool in the genome sequencing project, a sequence-ready physical map using PAC clones is being

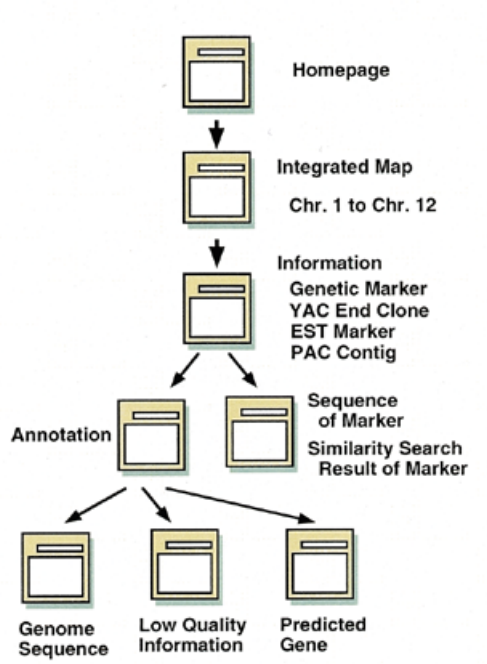

Figure 1. The site map of INE. An integrated map is prepared for each chromosome. An Information window comes up when the corresponding marker, clone or contig is designated. The Information window for genetic marker links to a window that contains the marker sequence and the results of similarity search. The Information window for a PAC contig links to an annotation window, which links to the Genome Sequence, Low Quality Information and Predicted Gene windows.

constructed. The ordered PAC clones are displayed and linked to the sequence of the clone as well as the results of annotation using various computational tools. The accuracy of the sequence after the assembly is important in a genome project because it indicates the reliability of the genome sequence. Thus INE incorporates an additional feature to verify the accuracy of the sequence. The international rice sequencing organization set up a standard that must be applied to all sequences submitted by members of the collaborative project. Following the example of the Human Genome Project, the score output from the assembly program PHRAP measures the accuracy of this standard. The standard requires that genome regions falling below a 40 numerical score on PHRAP must be separated. To ensure the quality control of data, INE provides a page for 'Low Quality Information' containing regions that do not meet this standard. This can be accessed through the Annotation window as shown in the site map (Fig. 1).

Integrated viewing of data

The unique features of INE are not limited to the characteristics of the incorporated data. Some mechanisms were also developed to facilitate the mining of useful information from the database. For instance, INE provides an integrated view of the data. An integrated map of each chromosome showing the linkage map, the physical map with ordered YAC clones, the EST map (under construction) and the PAC contigs can be viewed in a window. A conceptual view of the integrated map is shown in Figure 2. The positions of the genetic markers, ordered YAC clones, EST markers and PAC contigs are shown in relation to each map. By integrating these data on the map, the value of the genetic information can be enhanced as illustrated in the following examples: (i) by integrating the linkage map and the PAC contig map, a user can check the existence of a sequenced PAC adjacent to the genetic marker of interest; (ii) by investigating the annotation of the PAC located near the genetic marker of interest, a user can check the occurrence of an EST on the PAC whose function is responsible for the target trait; (iii) by further integrating the linkage map and the PAC contig map, the relationship between the positions of the genetic marker and peculiar genome sequences which are indicated in the annotation, such as long terminal repeats (LTRs) presumably related to transposons, can be evaluated; (iv) by using a genome sequence linked to the PAC contig, a user is provided with the complete PAC sequence in the computer; (v) by integrating the linkage map and the EST map, the existence of an EST close to the genetic marker of interest can be verified; (vi) the integration of the linkage map and EST map provides a general visualization of the benefits of high-density markers derived from a simultaneous use of genetic markers and EST markers. These examples show that related genetic information can be retrieved directly from INE hence demonstrate the usefulness of the integrated map.

The integrated display of genomic information on a particular region of interest would prove to be useful in map-based cloning of agronomically important genes. Initially, genetic markers tightly linked with a gene of interest could be established by genetic mapping using the DNA markers. The vast range of data for each marker available in INE could simplify the mapping process. Then, by examining the physical map, YAC clones carrying those markers can be identified. Other ESTs localized on those YAC clones would provide additional markers for tighter mapping. Such information would be useful enough in subsequent cloning of the gene. Furthermore, if the PACs aligned with those markers have been sequenced, then the detailed features of the gene can be determined as well.

Viewing capability

Another important feature of INE is the rapid display of integrated maps by programming the viewer in Java language. The steps involved in viewing a map using a conventional web-based and Java-based viewer are different. Using a conventional web-based viewer, a map is provided via a GIF image and its transmission from the server to the client is needed to redraw a map. On the other hand, by using a Java-based viewer as in INE, the drawing operations such as scroll and zoom in/out, are performed only in the client's computer once the searched data has been downloaded. We compared the time to scroll a map between INE and the former RGP database with a conventional GIF image-based viewer. The result showed that INE would take about one-eighth of the time it would take a GIF image-based viewer to retrieve the same data. This feature contributes to a smooth navigation with INE and the user may get the feeling that the database is built in his own computer. The time required in INE to download the Java applet and its data would simply depend on the computational environment such as network speed. This takes only ~5 s using a client computer in RGP. In INE, a chromosome is a unit cluster and the drawing operations on the chromosome are performed immediately once the data is downloaded to the client's computer. Therefore, a user can accomplish both an overall

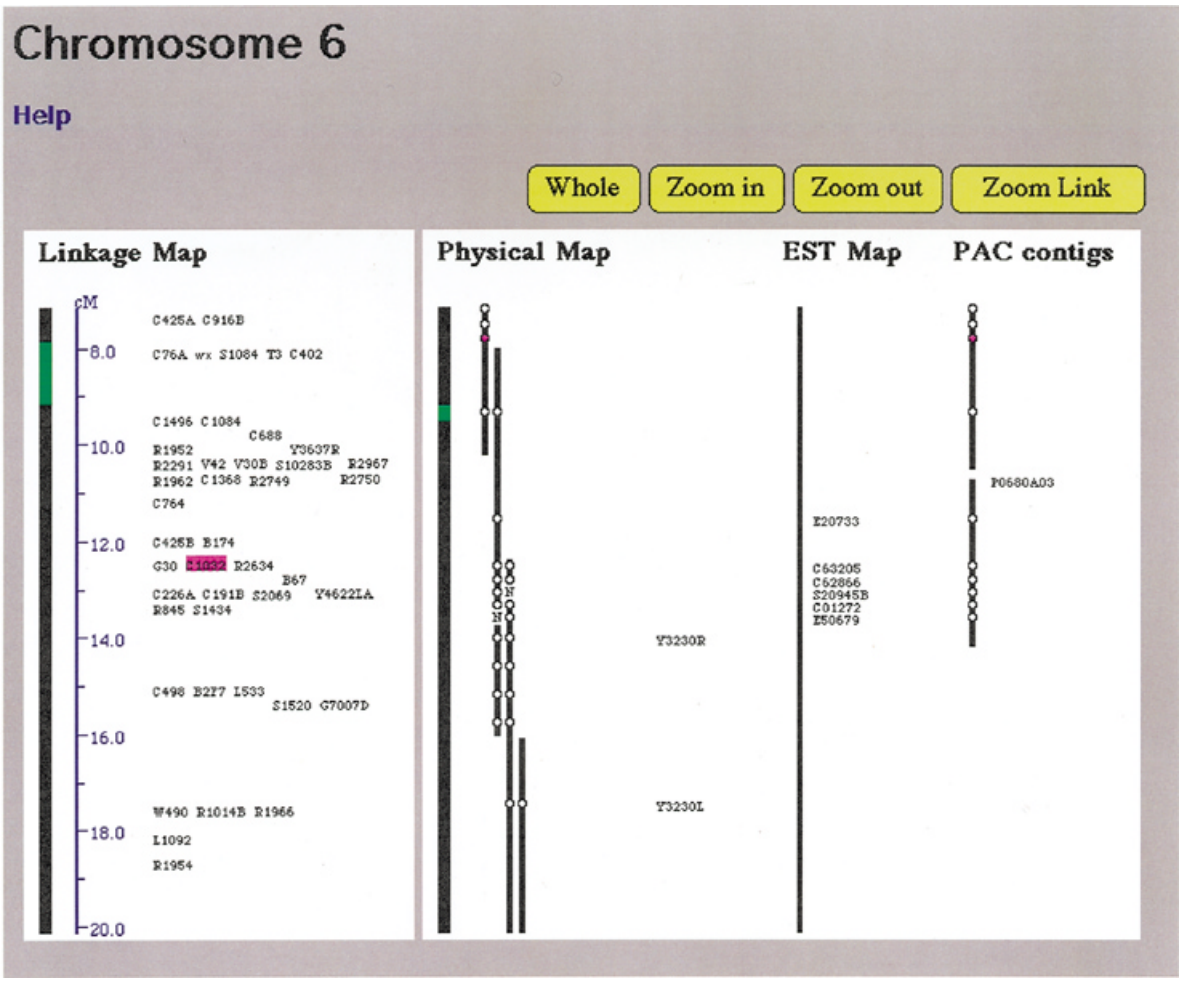

Figure 2. A conceptual view of the integrated map. The integrated map window is displayed in two frames, one frame shows the linkage map and the other shows the physical map with ordered YAC clones, the EST map and the PAC contigs. A highlighted circle on the physical map with YAC clones represents the position of the corresponding marker (C1032) also highlighted on the linkage map. WHOLE button produces the entire graphical picture of the integrated maps. A user can expand or condense the map view using the ZOOM IN/ZOOM OUT buttons. The expanded region is indicated by the green part of the entire chromosome on the left of each frame which can be scrolled to move to different regions. The right-hand most buttons either link both frames for simultaneous zooming, or unlink the frames for independent zooming. A coordinated view between two frames is obtained when a user clicks a marker in a frame which is common to the other frame. As a result the other frame is automatically scrolled to show the position of the marker on the other maps.

survey of the distribution of markers and clones, and a detailed investigation of a specific region on the chromosome in a short time. Thus if a user is interested to see various views of the map in a short time, the Java-based viewer in INE will be more efficient than GIF image-based viewers.

The computational features during such operations in INE are as follows. (i) The web browser in the client's computer downloads the INE applet from the server and operates it. An applet means a Java program that can operate in a web browser. (ii) The data of a chromosome selected by the user is downloaded to the web browser by using the INE applet. At this time, a data file written in a text format is read through the CGI (Common Gate Interface). (iii) The INE applet controls the drawing of the maps.

INE has also the ability to display different chromosomes at one time to achieve simultaneous comparisons between or among chromosomes. To utilize this function, a user may open any number of windows at one time to view the different chromosomes on a web browser and do the normal INE operations in each window.

Conciseness of site configuration

The site map of INE is shown in Figure 1. An integrated map of the specific chromosome for viewing can be selected in the homepage. Alternatively, search for specific markers is available at the homepage through the search marker option. This option directly opens an available Information window on the particular marker in addition to the integrated map of the chromosome where the marker is located. Clicking an object on the integrated map can also open the Information window. The Information window plays an important role on the site map by displaying the detailed information of the selected object, and at the same time becomes an entrance to obtain further information. The standardized procedure of using the sequential features in the Information window contributes to the intelligibility of INE.

There are four types of Information windows in INE. These are the windows for genetic marker, YAC end clone, EST

marker and PAC contig. An Information window for genetic marker contains the locus name, the exact position in the linkage map in cM, accession number, mapping profile (restriction enzyme used for mapping and band size in parent lines), probe records (probe size, vector and cloning site) and physical mapping profile (YAC clones in which the marker has been identified and corresponding insert size). Also included in this window is the graphic display of RFLP band pattern particularly for cDNA and most genomic clones used as markers. The accession number of the marker opens another window that contains the nucleotide sequence profile of the marker as well as the results of similarity search. An Information window for YAC end clone contains the YAC clone from which the end fragment has been isolated, other YAC clones carrying the same fragment, and the corresponding insert size of these YACs. An Information window for EST marker is being prepared and it will contain the clone name and EST mapping profile such as the specific primers used for screening and the YAC clones in which the EST has been identified. An Information window for PAC contig contains the detailed contig map of the particular region where the PAC contig is located and links to the annotation figure.

CONCLUDING REMARKS

In the future, INE will be further modified to meet the growing needs of the genome sequencing project. Several features will be added that will allow more efficient integration of the maps, search for specific genome information and cross-references to other databases. As we accumulate more data, a powerful retrieval system will be developed to speed up access and manipulation of various information.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of all members of the Rice Genome Research Program. We also thank Dr Yoshiaki Nagamura of the Ministry of Agriculture, Forestry and Fisheries (MAFF) DNA Bank for useful collaboration. The MAFF DNA Bank provides a server for INE. This work was supported by MAFF and the Japan Racing Association. INE utilizes a Java-based application GIOT (Genome Information displayed Orderly Tool) developed by Mitsubishi Space Software Co. Ltd.

REFERENCES

1. Moore,G., Devos,K.M., Wang,Z. and Gale,M.D. (1995) *Curr. Biol.*, **5**, 737–739.
2. Yamamoto,K. and Sasaki,T. (1997) *Plant Mol. Biol.*, **35**, 135–144.
3. Harushima,Y., Yano,M., Shomura,A., Sato,M., Shimano,T., Kuboki,Y., Yamamoto,T., Lin,S.Y., Antonio,B.A., Parco,A., Kajiya,H., Huang,N., Yamamoto,K., Nagamura,Y., Kurata,N., Khush,G.S. and Sasaki,T. (1998) *Genetics*, **148**, 479–494.
4. Tanoue,H., Baba,T., Saji,S., Idonuma,A., Hamada,M., Katagiri,S., Nakashima,M., Chiden,Y., Hayashi,M., Okamoto,M., Wu,J., Antonio,B.A., Koike,K., Umehara,Y., Matsumoto,T., Jong,P.J. and Sasaki,T. (1999) *Abstracts of Plant and Animal Genome VII*, p.207.
5. Rhee,S.Y., Weng,S., Bongard-Pierce,D.K., Garcia-Hernandez,M., Malekian,A., Flanders,D.J. and Cherry,J.M. (1999) *Nucleic Acids Res.*, **27**, 79–84.