African Journal of Biotechnology Vol. 8 (20), pp. 5253-5259, 19 October, 2009
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB09.1182
ISSN 1684–5315 © 2009 Academic Journals

Full Length Research Paper

A genome browser database for rice (*Oryza sativa*) and Chinese cabbage (*Brassica rapa*)

ChangKug Kim[1], JeomHwa Han[2], YounHee Shin[1], SungHan Park[1], DoWon Yun[1], ByungOhg Ahn[3], DongHern Kim[4], BeomSeok Park[1] and JangHo Hahn[1]*

[1]Genomics Division, National Academy of Agricultural Science (NAAS), Suwon 441-707, Korea.
[2]National Institute of Horticultural and Herbal Science (NIHHS), Naju 520-821, Korea.
[3]High-Tech Agriculture Division, Rural Development Administration (RDA), Korea.
[4]Bio-crop Division, National Academy of Agricultural Science (NAAS), Korea.

Accepted 18 September, 2009

We have constructed an integrated genome browser database for sequence analysis of rice (*Oryza sativa*) and Chinese cabbage (*Brassica rapa)* genomes. The genome browser for Arabidopsis (*Arabidopsis thaliana*) was included to provide the comparative analysis with Chinese cabbage. The genome browser of rice provides a variety of alternate views of the data in 12 chromosomes. Constructed from a supercontig using 3,360 BACs, it shows information about individual genes with functional annotation in the whole chromosome. In addition, the rice genome browser provides specific analysis of the genome by Gene Ontology (GO), single nucleotide polymorphisms (SNPs), Align, Text, Mart, BLAST and Export data views. In the comparative genome analysis between the Arabidopsis and Chinese cabbage genomes, users can obtain information using comparative genomics methods and identify missing regions within a single genome.

Key words: genome browser, Ensembl project, rice database, Chinese cabbage.

INTRODUCTION

As a cereal grain, rice (*Oryza sativa*) is the most important staple food for a large part of the world's human population. *Arabidopsis thaliana* is the most widely studied model plant. Chinese cabbage (*Brassica rapa*) is one of the most important vegetable in Korea and in the Northeast Asia.

An integrated genome browser provides a natural index for molecular biology and information for understanding biological data. To facilitate the study of biological and genomic researches, the various browsers have been constructed. In rice, the Gramene (http://www.gramene.org/) browse assembled genomes for *O. sativa* using genomes browser. The International Rice Information System (IRIS) provides a database system for the genetic mapping, genome annotation, genotype, mutant, proteome and metabolic data (Bruskiewich et al., 2003). The Rice Annotation Project Database (RAP-DB) provides the genome sequence in rice (Ohyanagi et al., 2006) and OryGenesDB is a database developed for rice reverse genetics which contains 78 annotation layers (Droc et al., 2008).

The Arabidopsis Information Resource (TAIR, http://arabidopsis.org) is the model organism database for the fully sequenced and intensively studied model plant *A. thaliana* (Swarbreck et al., 2008). The MAtDB provides a comprehensive resource for Arabidopsis as a genome model that serves as a primary reference for research in plants (Schoof et al., 2004).

In Chinese cabbage, Brassica ASTRA is a public database for genomic information on Brassica species (Love et al., 2005) and Shanghai Database (RAPESEED) provides the resource of 8,462 unique ESTs and tag-to-gene data during seed development (Wu et al., 2008). The *B. rapa* Genome Project (BrGP, http://www.brassica-rapa.org/) provides the genome information for Chinese cabbage (*B. rapa*) using genome map browser.

In the other species, Genome Browser Database (GBD) contains integrated sequence and annotation data for a large collection of genomes, including those of vertebrates and other model organisms (Karolchik et al.,

*Corresponding author. E -mail: jhhahn@rda.go.kr.

Table 1. Genome details of rice, Arabidopsis and Chinese cabbage.

Dataset	Rice	Arabidopsis	Chinese cabbage
Contig (BAC)	3,360	1,619	1,619
Gene	50,717	30,853	-
Exon	193,605	166,873	25,201
Marker	9,587	2,074	644
SNP	72,304	-	-
EST	168,792	-	7,333
Prediction-exon	249,081	-	-
Prediction-transcript	56,161	-	-
Total (Records)	803,607	201,419	34,797

2008). The National Center for Biotechnology Information (NCBI) provides analysis and retrieval resources for the data in GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) and other databases (Wheeler et al., 2008). Such databases are not limited to vertebrate systems. In the field of microbiology, several genome database browsers have been developed, including the IMG (Markowitz et al., 2007), NMPDR (McNeil et al., 2007), MBGD databases (Uchiyama, 2007) and the integrated browser of KACC (Kim et al., 2009).

The Ensembl project is a joint scientific project between the European Bioinformatics Institute and the Wellcome Trust Sanger Institute. The project provides a comprehensive genome information system consisting of data storage, integration, analysis and visualization of a wide variety of biological data (Potter et al., 2004). Ensembl software come into its own as genome browser for the integration of any biological data that can be mapped onto features derived from the genomic sequence. At present, Ensembl (http://www.ensembl.org/) fully supports 45 species with three additional species partially supported (Hubbard et al., 2009). However, it does not support a genome browser for rice and Chinese cabbage. We have constructed a web-based genome browser database for three species, rice (O. sativa), Arabidopsis (A. thaliana) and Chinese cabbage (*B. rapa*), in order to provide the analysis tool of the specific gene function and comparative genome annotation.

METHODOLOGY

Datasets

The genomic information derived from three different plant species were used to construct the genome browser database, namely rice (*Oryza* sativa), Arabidopsis (*A. thaliana*) and Chinese cabbage (*B. rapa)*. The genome sequence of rice was shared with the International Rice Genome Sequencing Program (IRGSP, http://rgp.dna.affrc.go.jp/IRGSP/), which the National Academy of Agricultural Science (NAAS, http://www.naas.go.kr/) of Rural Development Administration (RDA, http://www.rda.go.kr) was involved as a consortium of the international cooperative project. The draft sequence was constructed from a supercontig of 3,360 BACs and this genome was predicted to contain 50,717 genes by the FGENESH program (Jun et al., 2002). The genome sequences of Arabidopsis and the Chinese cabbage *B. rapa* were obtained from the Arabidopsis Information Resource (TAIR) and the *B. rapa* Genome Project (BrGP, http://www.brassica-rapa.org/) of NAAS, respectively. The Chinese cabbage genome was constructed from 1,619 BACs assembled according to EST sequences and aligned based on the chromosome structure of Arabidopsis. To assist in the comparative genome analysis between *B. rapa* and Arabidopsis, 25,201 EST sequences were aligned and mapped into 167,000 exons.

In addition, 3,258 markers were collected from the Chinese cabbage project and Korean rice genome project of NAAS. The markers consists of 2,800 RFLP and 112 QTL markers related to rice. It also includes 321 RFLP and 25 PCR-based markers related to Chinese cabbage (Kim et al., 2008). Furthermore, 7,227 SNP markers were collected from the rice SNP project (http://nabic.niab.go.kr/SNP/). Finally, GRAMENE (ftp://ftp.gramene.org/pub/gramene/release16/data/) and the NCBI/GenBank database were used to obtain additional genomic information on rice. Table 1 summarizes the information of the genomic data from each species.

Database development

Our web-based genome browser database was developed using Ensembl public software (http://www.ensembl.org/). The platform was developed using the SQL and Java languages on an Apache Tomcat server. The base information library was written using Bioperl (http://www.bioperl.org/) and the data was stored on an Oracle Relational Database Management System (RDBMS). The schema was designed using standard relational database princeples. ERWin Data Modeler software (http://www.ca.com) was used to construct the logical and physical design of the database.

Database design

The genome browser database consists of three major functional categories. ‘Map view’ allows the user to identify the genome of interest. Detailed annotated information at the chromosome and base pair levels can be viewed in ‘Contig view’. Finally, genomic analysis based on Gene Ontology (GO), single nucleotide polymorphisms (SNPs) and sequence alignment (e.g. BLAST), as well as data export functions, can be performed in ‘Function view’. The export data function enables file download of search results to the user. Figure 1 shows the design concepts and system process for information flow.

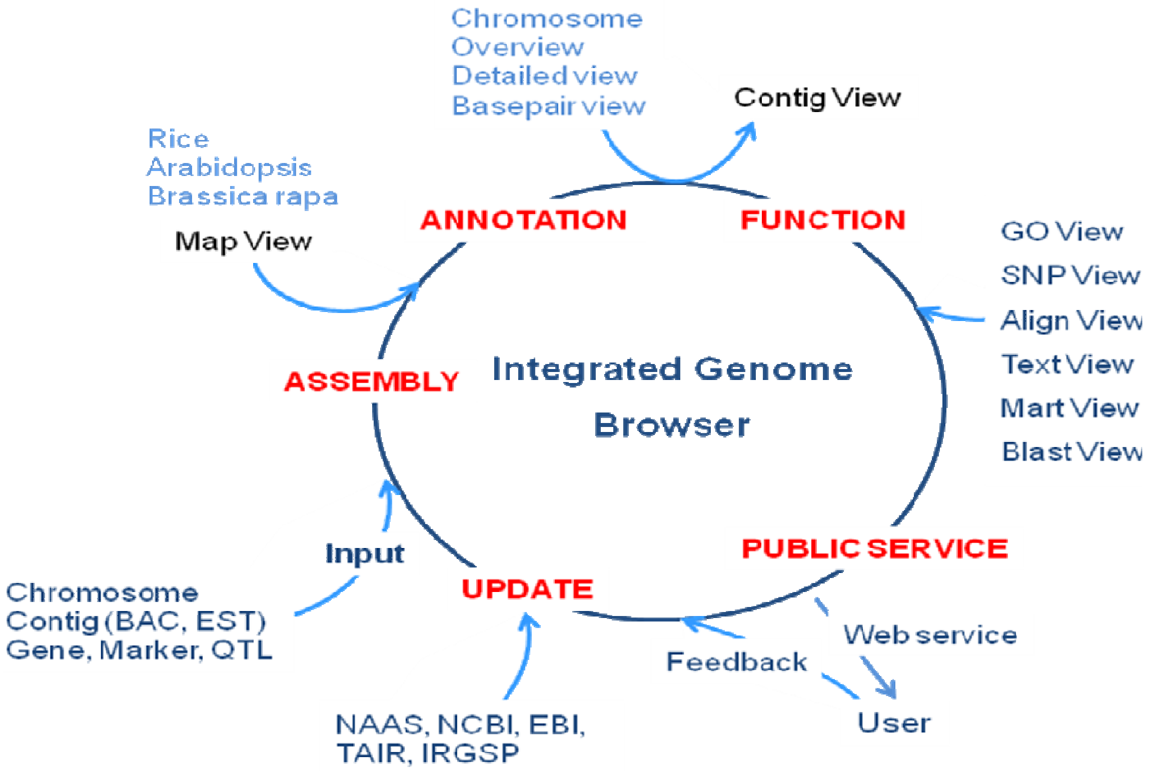


Figure 1. Overview of integrated genome browser design concepts and system process.

Implementation and features

The genome browser database provides a powerful framework for analyzing genome sequences and annotation data from multiple sources. This browser provides an infrastructure for large biological data integration and analysis. The genome browser consists of three major functional categories, including 'map view', 'contig view' and 'function view' in addition to a variety of alternate views of the data. 'Map view' enables selection of the species about which information chromosome information is desired. Figure 2 shows the first page of 'map view' in rice.

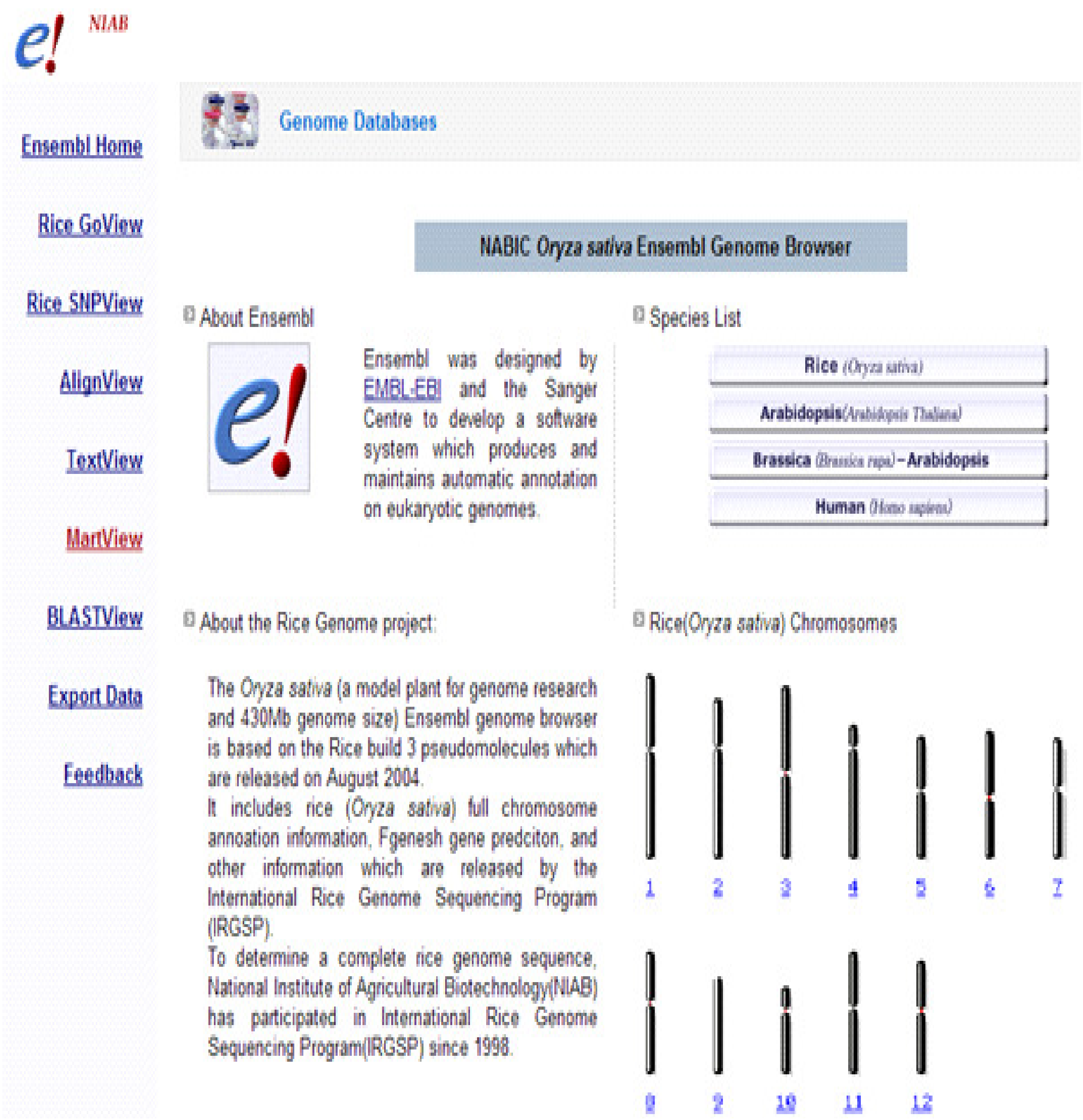
The 'contig view' shows relationships between the genome sequence and annotated data. Consisting of four viewable panels that are accessible by clicking, the user can access information about individual genes along with functional annotation within the entire chromosome. The chromosome panel depicts the chromosome banding region for selection and the overview panel shows the location of markers and genes. The detailed view panel shows genomic sequence features and genes as predicted by the FGENESH program. The base pair view panel exhibits the correlation between the ancestry of individuals and the common variability of pair-wise linkage (Figure 3).

The Chinese cabbage (*B. rapa*) genome browser focuses on comparative genome analysis and visualization of both genome sequences. The 'contig view' provides a function to calculate and display integrated comparative genomics resources. By viewing alignments of the genome sequence and markers, users can obtain information derived from comparative genomics methods in this browser (Figure 4).

In the rice genome, 'function view' provides specific analysis of the genome using Gene Ontology (GO), single nucleotide polymorphisms (SNPs), Align, Text, Mart, BLAST and Export data views. The standard browser from the GO consortium (http://www.geneontology.org/) has been integrated to create the 'GO view', which provides 23,053 records of information from the rice genome mapped to molecular function, cellular component and biological process. The 'SNP view' shows functional information such as regulatory regions to indicate the potential consequences of a variation. Software for viewing SNPs which can compute features of the variations between japonica and indica rice was developed. Moreover, users can access detailed location information of 72,304 SNP markers in rice. 'Align view' displays results from basepair level alignment between genomes and can compute the gene pairwise ratio. 'Text view' allows searching of the entire database with a combination of words or phrases. 'Mart view' is a data retrieval tool that generates lists of biological objects as genes and provides a number of querying methods to generate multiple types of output. The 'BLAST view' enables similarity searching against the entire rice genome sequence and predicted genes. 'Export view' provides information from searching methods, consisting of a flat file, FASTA, feature list and image menu. Users can export a table of detailed information or download a file containing datasets with a specific associated trait.

DISCUSSION AND FUTURE WORK

The genome browser database provides information through a user-friendly web interface that allows analysis of genome infrastructure. We have contributed to this informatics approach to agricultural biotechnology to extend the usefulness of breeding for new crops. It is accessible the web site of NAAS (http://ensembl.niab.go.kr:8080/). The browser provides annotated genome



NABIC *Oryza sativa* Ensembl Genome Browser

About Ensembl

Ensembl was designed by EMBL-EBI and the Sanger Centre to develop a software system which produces and maintains automatic annotation on eukaryotic genomes.

Species List

About the Rice Genome project:

The *Oryza sativa* (a model plant for genome research and 430Mb genome size) Ensembl genome browser is based on the Rice build 3 pseudomolecules which are released on August 2004.

It includes rice (*Oryza sativa*) full chromosome annoation information, Fgenesh gene predciton, and other information which are released by the International Rice Genome Sequencing Program (IRGSP).

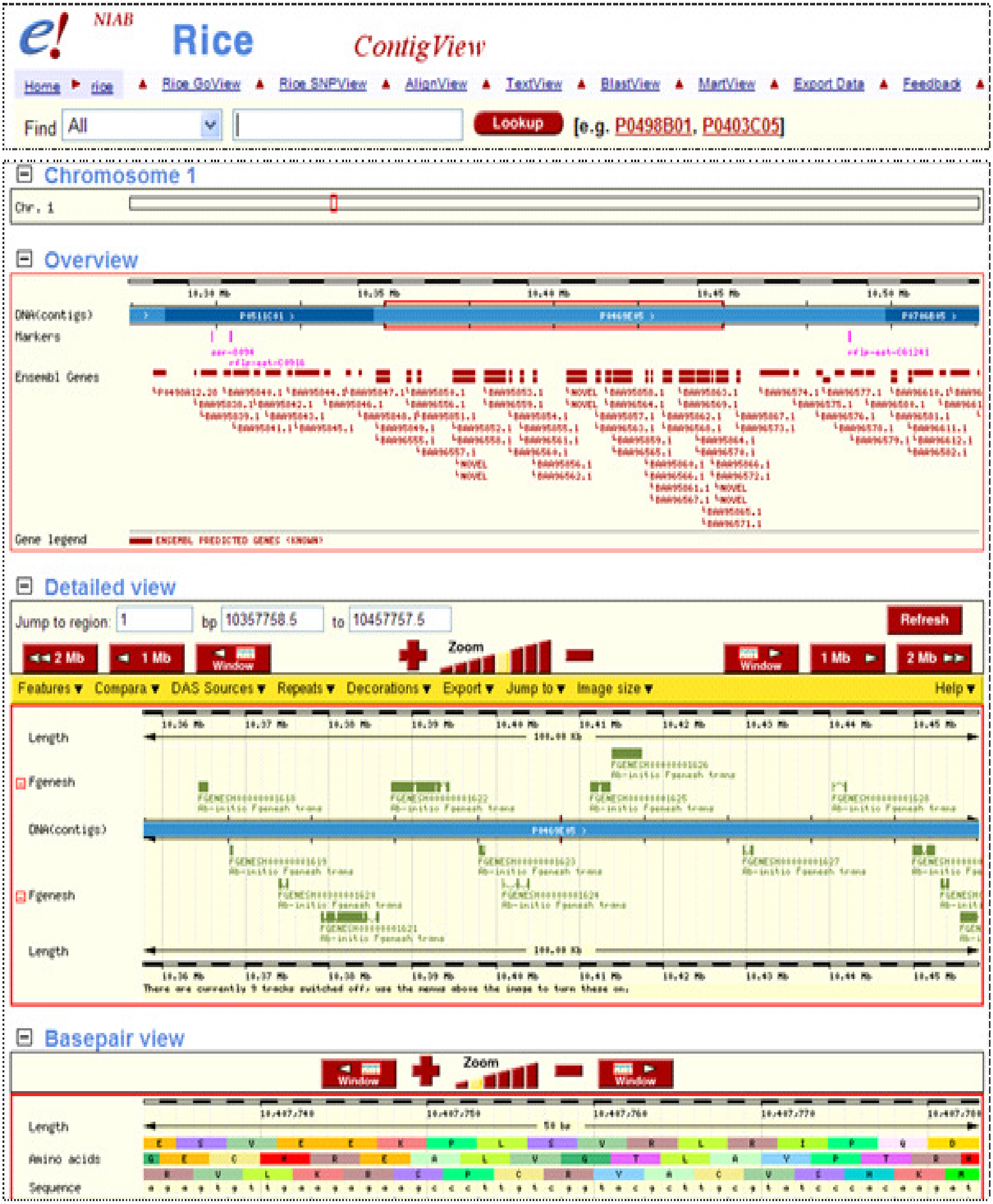
To determine a complete rice genome sequence, National Institute of Agricultural Biotechnology(NIAB) has participated in International Rice Genome Sequencing Program(IRGSP) since 1998.

Rice(*Oryza sativa*) Chromosomes

Figure 2. Screenshot of ‘map view’ showing the region of rice chromosomes for selection from the species list.

information of 803,607 for rice, 201,419 for Arabidopsis and 34,797 records for Chinese cabbage (*B. rapa*).

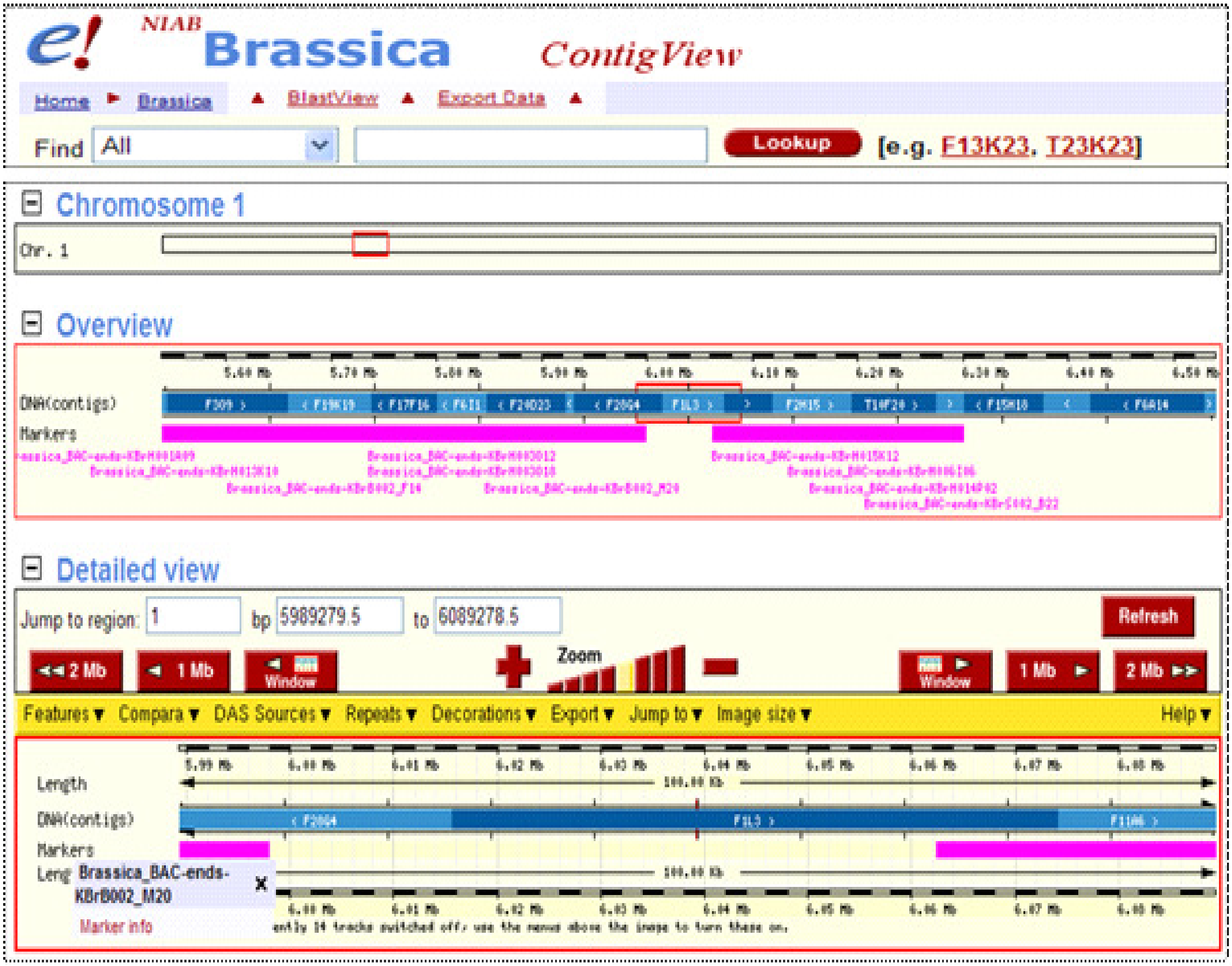
The genome browser of rice provides a variety of alternate views of the data for 12 chromosomes. Constructed from a supercontig using 3,360 BACs, this browser can display information about individual genes with functional annotation in the whole chromosome. In the comparative genome analysis between Arabidopsis and Chinese cabbage (*B. rapa*), users can obtain new information using comparative genomics methods. In addition, identification of missing regions within a single genome can be performed. This browser provides improved comparative annotation from more complete genome information based on progress made by the *B. rapa* Genome Project



Jump to region: 1 bp 10357758.5 to 10457757.5

There are currently 9 tracks switched off; use the menus above the image to turn them on.

Figure 3. Screenshot of ‘contig view’ in rice. This includes a zoom function and pull-down menus to allow the user to select the features being displayed. Floating menus (not shown) can access linked windows with additional information.



Jump to region: 1 bp 5989279.5 to 6089278.5

...ently 14 tracks switched off, use the menus above the image to turn these on.

Figure 4. Screenshot of ‘contig view’ in Chinese cabbage (*B. rapa*) showing a comparative genome analysis and markers in NAAS.

(BrGP, http://www.brassica-rapa.org/) of NAAS.

REFERENCES

Bruskiewich RM, Cosico AB, Eusebio W, Portugal AM, Ramos LM, Reyes MT, Sallan MA, Ulat VJ, Wang X, McNally KL, Sackville Hamilton R, McLaren CG (2003) Linking genotype to phenotype: the International Rice Information System (IRIS). Bioinformatics, 19: 63-65.

Droc G, Périn C, Fromentin S, Larmande P (2008). OryGenesDB 2008 update: database interoperability for functional genomics of rice. Nucl. Acids Res. 37: D992-D995.

Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Rios D, Schuster M, Slater G, Smedley D, Spooner W, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wilder S, Zadissa A, Birney E, Cunningham F, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Kasprzyk A, Proctor G, Smith J, Searle S, Flicek P (2009). Ensembl 2009. Nucl. Acids Res. 37: D690-D697

Jun Yu, Hu S, Wang J (2002). A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. indica). Science, 296: 79-92.

Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, Kober KM, Miller W, Pedersen JS, Pohl A, Raney BJ, Rhead B, Rosenbloom KR, Smith KE, Stanke M, Thakkapallayil A, Trumbower H, Wang T, Zweig AS, Haussler D, Kent WJ (2008) The UCSC Genome Browser Database: 2008 update. Nucl. Acids Res. 36: D773-D779.

Kim CK, Jeon YA, Cho GT, Kwon SW, Hahn JH, Hong SB (2009). KACC: An identification and characterization for microbial resources in Korea. Afr. J. Biotechnol. 8: 69-72.

Kim CK, Kim JS, Lee GS, Park BS, Hahn JH (2008). PlantGM: a database for genetic markers in rice (Oryza sativa) and Chinese cabbage (Brassica rapa). Bioinformation, 3: 61-62.

Love CG, Robinson AJ, Lim GA, Hopkins CJ, Batley J, Barker G, Spangenberg GC, Edwards D (2005). Brassica ASTRA: an integrated database for Brassica genomic research. Nucl. Acids Res. 33: D656-D659.

Markowitz VM, Szeto E, Palaniappan K, Grechkin Y, Chu K, Chen IM, Dubchak I anderson I, Lykidis A, Mavromatis K, Ivanova NN, Kyrpides NC (2007). The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. Nucl. Acids Res. 36:D528-D533.

McNeil LK, Reich C, Aziz RK (2007). The National Microbial Pathogen Database Resource (NMPDR): a genomics platform based on subsystem annotation. Nucl. Acids Res. 35: D347-D353.

Ohyanagi H, Tanaka T, Sakai H, Shigemoto Y, Yamaguchi K, Habara T, Fujii Y, Antonio BA, Nagamura Y, Imanishi T, Ikeo K, Itoh T, Gojobori T, Sasaki T (2006). The Rice Annotation Project Database (RAP-DB): hub for Oryza sativa ssp. japonica genome information. Nucl. Acids Res. 34: D741-D744.

Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, Searle SM, Stabenau A, Storey R, Clamp M (2004). The Ensembl Analysis Pipeline. Genome Res. 14: 934-941.

Schoof H, Ernst R, Nazarov V, Pfeifer L, Mewes HW, Mayer KF (2004). MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource for plant genomics. Nucl. Acids Res. 32: D373-D376

Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, Radenbaugh A, Singh S, Swing V, Tissier C, Zhang P, Huala E (2008). The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucl. Acids Res. 36: D1009-D1014.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmberg W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pontius JU, Pruitt KD, Schuler GD, Schriml LM, Sherry ST, Sirotkin K, Starchenko G, Suzek TO, Tatusov R, Tatu TA, Wagner L, Yaschenko E (2008). Database resources of the National Center for Biotechnology Information. Nucl. Acids Res. 36: D13-D21.

Wu GZ, Shi QM, Niu Y, Xing MQ, Xue HW (2008). Shanghai RAPESEED Database: a resource for functional genomics studies of seed development and fatty acid metabolism of Brassica. Nucl. Acids Res. 36:D1044-7.

Uchiyama I (2007). MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. Nucl. Acids Res. 35: D343-D346.