APPLICATION NOTE

Genomics & Informatics Vol. 5(4) 194-195, December 2007

RGISS: Rice (Oryza sativa L. ssp. *japonica*) Genome Information Service System

Daesang Lee[1], Hwajung Seo[2,4], Jang-Ho Hahn[3], Eun-Bae Kong[2] and Kiejung Park[4,*]

[1]Department of Bioinformatics, Korea Bio Polytechnic, Chungnam 320-905, Korea, [2]Deptartment of Computer Engineering, Chungnam National University, Daejeon 305-764, Korea, [3]National Institute of Agricultural Biotechnology, RDA, Suwon 441-707, Korea and [4]Information Technology Institute, SmallSoft Co. Ltd., Daejeon 305-342, Korea

Abstract

We have constructed the Rice Genome Information Service System (RGISS), which is an information service system of the Oryza sativa L. ssp. *japonica* (rice) genome, using the released version of rice Build 3.0 pseudomolecules based on the Ensembl architecture. The nonredundant library, composed of 3,360 clones of BACs, PACs, and fosmids, was used to construct supercontigs.

RGISS contains 50,717 annotated genes from GenBank, 56,161 predicted genes from FgeneSH, and information on 9,587 markers, which includes STS, SSR, and EST-based RFLP. The 20,180 ESTs sequenced by the Korea National Institute of Agricultural Biotechnology (NIAB) were aligned and mapped into 168,792 exons. By gene ontology analysis, the classified protein numbers in the rice genome were 6158, 4531, and 12,364 proteins, which were mapped to molecular function, cellular component, and biological process, respectively.

Availability: RGISS is accessible via the web site of Korea National Institute of Agricultural Biotechnology (http://ensembl.niab.go.kr:8080/).

Keywords: rice, annotation, Ensembl, EST, marker

Introduction

The International Rice Genome Sequencing Project (IRGSP) was established in 1998 with the aim of sequencing the entire rice (Oryza sativa L. ssp. *japonica*) genome (Sakata K. *et al*., 2002), the size of which is around 430 Mb, the smallest among the major cereal crops.

NIAB has participated in the IRGSP with the Korea Rice Genome Research Program, the role of which was to determine sequences in a few regions of rice chromosome 1 (151.4-160cM) and chromosome 9 (68.2-77.7cM, 93.2-94.4cM).

As NIAB has participated in the IRGSP, its own rice genome information service system has been required to accommodate NIAB-specific research needs such as mapping in-house ESTs to the rice genome, as well as acting as the information service of the data provided by IRGSP.

To address these needs, the Rice Genome Information Service System (RGISS) has been developed. The Ensembl (www.ensembl.org) system was applied to display rice genome information, as it provides data to many viewers, from chromosome maps to detailed information of proteins. Scattered information of the rice genome, which includes annotated data from GenBank, gene prediction data from IRGSP, marker data, and EST data from NIAB, were aligned against chromosome sequences by IRGSP. Parsing programs were implemented to analyze annotated information, and calculation programs were implemented to compare and map sequence data. All data were converted into Ensembl database schema, and Ensembl viewers were adjusted to visualize them.

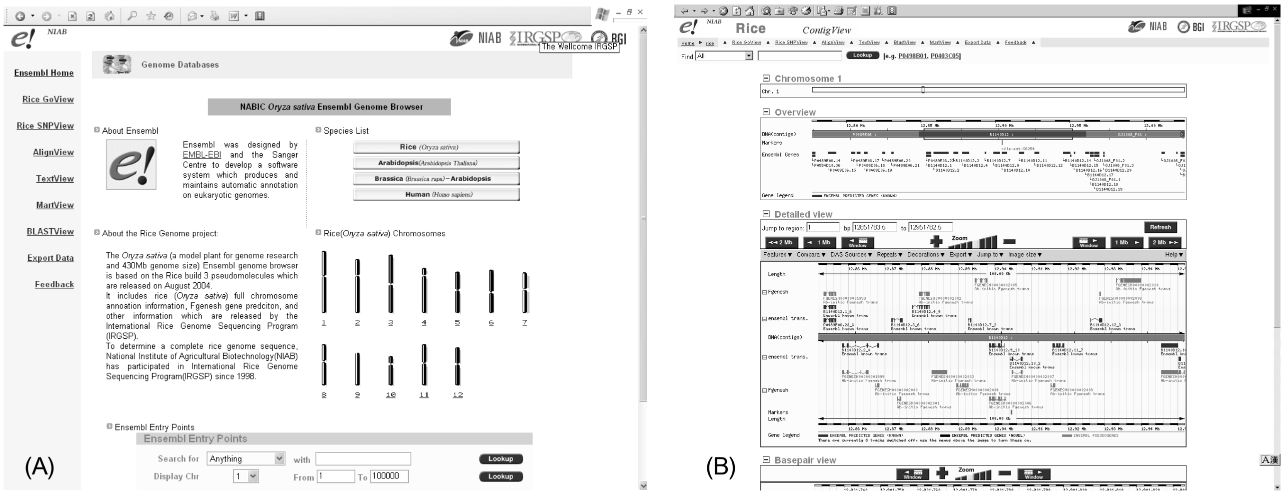
Results and Discussion

The annotation data for 4092 clones (BACs/PACs/fosmids) from IRGSP and GenBank were analyzed to find duplicated clones. Three thousand three hundred sixty clones selected after removing redundancy were aligned against Build 3.0 pseudomolecules of 12 chromosomal sequences, which were provided by IRGSP, to calculate their physical positions.

The physical positions of the genes in the clones were calculated after the clones were aligned against Build 3.0 pseudomolecules. The mRNA, protein, function, and locus data were extracted from annotation data, and 50,717 genes, 193,605 exons, and 50,717 transcripts were extracted from 3360 clones.

The information on 56,161 genes in Build 3.0 pseudomolecules, which were predicted using FgeneSH (Salamov A. *et al*., 2000) from IRGSP, was also integrated into RGISS.

*Corresponding author: E-mail kjpark@smallsoft.co.kr
Tel +82-42-864-2524, Fax +82-42-385-9240
Accepted 3 July 2007



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

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.

Fig. 1. Screen shots of RGISS. (A) Main page of RGISS. (B) Contig View shows chromosomes, markers, and genes for a selected region. And more detailed information such as physical positions of the genes and ESTs in a contig can be displayed.

Nine thousand five hundred eighty-seven markers, composed of 6586 EST based RFLP, 62 STS, and 2939 SSR, were incorporated into RGISS after their alignment against the 12 rice chromosomes.

The 20,180 ESTs sequenced by NIAB were aligned against the pseudomolecules using a local alignment search program, SIM4 (Florea L. *et al*., 1998), to find out their expression positions. They were mapped into 168,792 exons (averaging eight exons per one EST) and were distributed throughout entire chromosome.

By gene ontology analysis (Harris M.A. *et al*., 2004), the classified protein numbers in the rice genome were 6158, 4531 and 12,364 proteins, which were mapped to molecular function, cellular component, and biological process, respectively. Among 50,717 annotated genes, 23,053 proteins were mapped to GO categories, which means that the rice genome project is in the beginning stages and needs much more research for protein function analysis.

RGISS also provides several additional services. There are Rice SNPView (SNP information display), AlignView (sequence alignment view), TextView (text search), MartView (text generation for genes, clones, and markers), and BLASTView (homology search with rice chromosome sequences using BLAST).

The main purpose of RGISS construction was to incorporate in-house raw data and analysis data, as well as to integrate public rice genome information. EST and full-length cDNA data, which are based on real expression experiments, are very helpful for researchers to investigate the precise positions of genes, as well as to validate the accuracy of gene prediction programs such as GenScan (Burge C. *et al*., 1997) or FgeneSH. Through RGISS, the rice ESTs sequenced by NIAB can be accessible not only to the in-house research group but also to the public.

NIAB plans to integrate the full-length cDNA information generated by NIAB research teams and KOME (Knowledge-based Oryza Molecular biological Encyclopedia) into RGISS in the near future (Kikuchi S. *et al*., 2003).

RGISS is expandable in integrating additional rice genome information and analysis features, and will contribute to rice genome analysis and rice information service.

Acknowledgments

This work was supported by the Korea National Institute of Agricultural Biotechnology as a sub-section of the project, ‘The construction of an agricultural biotechnology information management system’.

References

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol*. 268, 78–94.

Tae, H. *et al*. (2007). ChroView: a trace viewer for browsing and editing chromatogram file. *Genomics & Informatics*. 5, 30-31.

Harris, M.A. *et al*. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*., 32, 258-261.

Kikuchi, S. *et al*. (2003). Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science*. 301, 376-379.

Salamov, A., and Solovyev, V. (2000). Ab initio Gene Finding in Drosophila Genomic DNA. *Genome Research* 10, 516-522.

Sasaki, T. *et al*. (2002). The genome sequence and structure of rice chromosome 1. *Nature* 420, 312-316.