Scientific Research and Essays Vol. 7(11), pp. 1256-1261, 23 March, 2012
Available online at http://www.academicjournals.org/SRE
DOI: 10.5897/SRE12.061
ISSN 1992-2248 ©2012 Academic Journals

Full Length Research Paper

A genome-based database for agricultural biotechnology

ChangKug Kim[1], DongSuk Park[1], YoungJoo Seol[1], GangSeob Lee[1], Myeong Ho Kim[1], SooChul Park[2], YongKab Kim[3] and JangHo Hahn[1]*

[1]Genomics Division, National Academy of Agricultural Science (NAAS), Suwon 441-707, Korea.
[2]Planning and Coordination Division, National Academy of Agricultural Science (NAAS), Suwon 441-707, Korea.
[3]School of Electrical Information Communication Engineering, Wonkwang University, Iksan 570-749, Korea.

Accepted 13 March, 2012

The National Agricultural Biotechnology Information Center (NABIC) plays a leading role in maintaining a database of information for agricultural plants and microbes. Since 2002, the NABIC has constructed an agricultural biology-based computational infrastructure and has provided comprehensive agricultural biological research information in Korea. Its major functions are focused on developing bioinformatics tools for investigating agricultural genomes and providing a database that integrates information from multiple plants and microbes. This new genome-based database provides useful information through a user-friendly web interface that allows analysis of genome infrastructure and searching of various resources.

Key words: Genome-based, database, biotechnology, web service.

INTRODUCTION

Our rapidly expanding genomic knowledge in a diversity of species from humans to microorganisms has led to a growing dependence on information technology in many different aspects of research. Technological advances in the fields of genome sequencing and protein structure determination have lead to an accumulation of genome, proteome, microarray and functional genomic data (Ann, 2008). Databases and bioinformatics tools have been created to store and enable effective analysis of this computational data, resulting in our increased understanding of the basic principles of molecular interactions and systemic functional behaviors of different organisms (Russ, 2007). The importance of such data in research is further demonstrated by the construction of integrated genome databases and bioinformatics tools. These databases, which include the Ensembl genomes (Kersey et al., 2012), database resources (Sayers et al., 2012) and integrated microbial genomes database (Markowitz et al., 2012), contain information pertaining to gene expression and function, genetic markers, gene family and protein prediction (Kim et al., 2011a).

In Korea, bioinformatics has been used to conduct research in agricultural biotechnology through various genome projects, and specific databases have constructed to organize these projects based on the long-linear model from the DNA sequence to the proteomics level (Kim et al., 2010a). The National Agricultural Biotechnology Information Center (NABIC) created an agricultural biology-based infrastructure. We have developed a genome-based model database for use in biotechnology development for agricultural bioinformatics and construction of bioinformatics workflows, such as gene annotation, functional gene prediction, gene expression analysis and metabolic pathway, via an easy-to-use web interface. Concentrating on functional genomic research of major crops, the integrated database focuses on major agricultural resources such as rice, rice mutants, Chinese cabbage and the microbes that affect them. We hope that this database will contribute to the agricultural bioinformatics research field

*Corresponding author. E-mail: jhhahn@korea.kr.

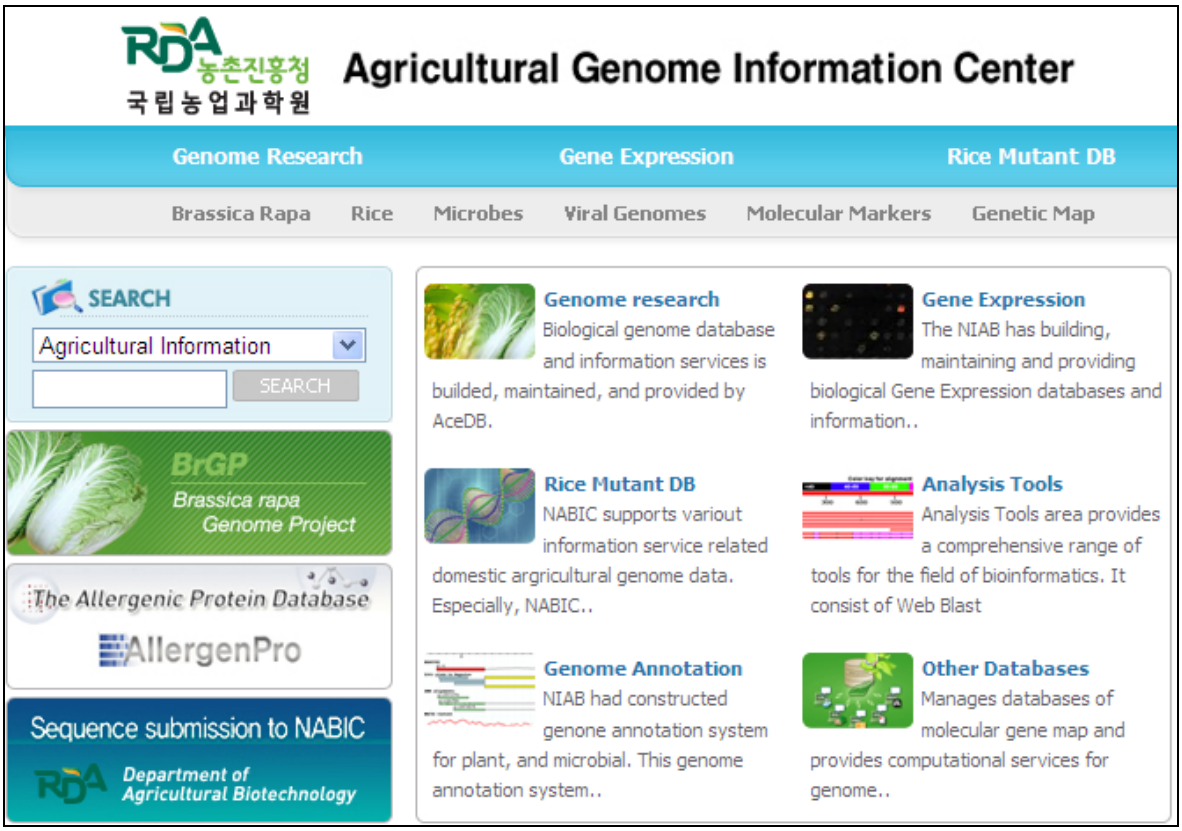

Figure 1. The six functional categories of the genome-based database, which shows information for various agricultural plants and microbe genomes.

and ultimately be used to assess the breeding of new crops.

MATERIALS AND METHODS

Data collection

The agricultural biotechnology information was collected from the research program for agricultural science at the National Academy of Agricultural Science (NAAS, http://www.naas.go.kr/), the international rice genome project (http://rgp.dna.affrc.go.jp/IRGSP/), the Chinese cabbage project (http://www.brassica-rapa.org/BGP/), the genetic resources project (http://www.genebank.go.kr/), the BG21 project (http://atis.rda.go.kr/), and from various universities and institutes in Korea. In addition, genomic information was collected through several collaborative institutes and public international institutes.

Database design

The database (http://nabic.naas.go.kr/) was designed to provide genomic information about major crops. This database consists of six major categories, namely genome research, gene expression, rice mutant database, analysis tools, genome annotation and other databases (Figure 1). The database includes simple text information on individual genome sequences, as well as analysis tables and genetic markers for annotation. In addition, this database can be used as a source for annotating genome sequences, physical maps, sequence comparison and gene prediction. The underlying model is a portable system capable of handling very large genomes and their associated requirements for sequence analysis. The platform was developed using MYSQL, commonly available network protocols such as Hypertext Transfer Protocol (HTTP), and the JAVA language. Data are stored using an Oracle relational database management system (Oracle Database 10 g, Redwood, CA, USA, http://www.oracle.com/).

Web service

This database web service platform consists of multiple layers that can be accessed using a web-based graphical interface that allows users to query and browse the data using various functions. This database can also connect to the genetic resources (http://www.genebank.go.kr/) and genetically modified organisms databases (http://biosafety.rda.go.kr/). To ensure software compatibility across different databases, service systems were developed using open standards protocols such as Simple Object Access Protocol (SOAP, http://www.w3.org/TR/soap/), an independent platform, commonly available network protocols such as HTTP, and an XML-based model (Figure 2).

RESULTS

Our newly developed genome-based database is

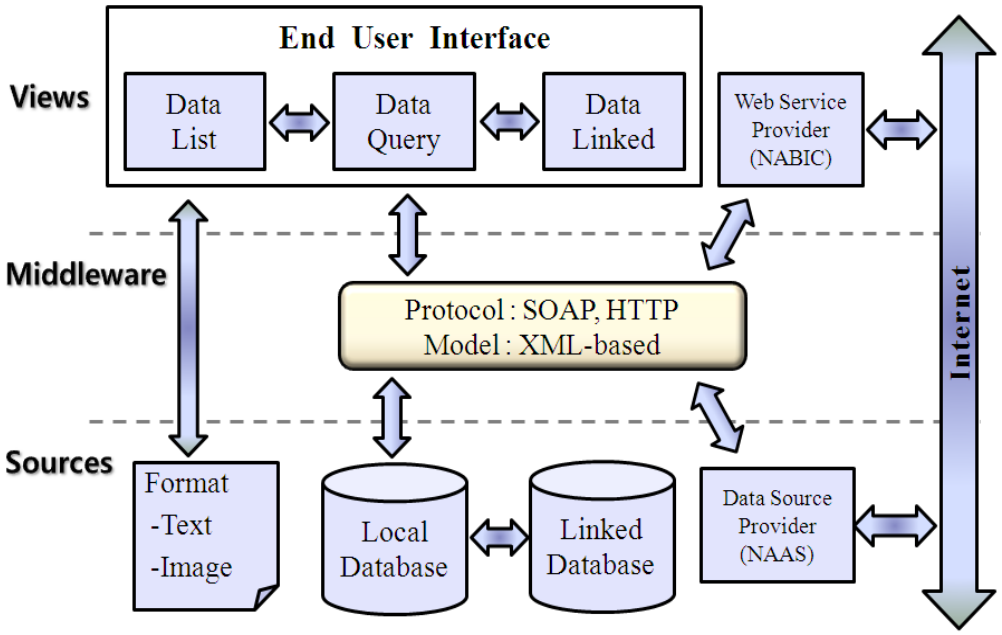

Figure 2. Overview of the web-based service system of the NABIC. This flowchart depicts the three groups viewable to the end user, the middleware supported by the manager and programmer, and the data sources for the provider. The system architecture was designed using various open standards models such as SOAP and HTTP protocols with an XML-based model.

Table 1. Genomic information contained in the database.

Dataset	Rice	Microbes	Chinese cabbage
Contigs	3,360	-	1,725
Genes	50,717	4,637	41,174
SNPs	72,304	-	-
ESTs	152,307	-	127,144
Markers	9,782	321	1,750
Mutants	18,158	16	-
Total	306,628	4,974	171,793

composed of multiple subsystems that hold information about genome research, gene expression, rice mutations and genomic annotation, as well as analytical tools and other databases. This system provides a computational framework for studying biological function based on the genomic sequences of wild-type and mutant rice, Chinese cabbage and microbe (*Xanthomonas oryzae*).

Genomic information

The database contains annotated genomic information from 306,628, 171,793 and 4,974 records mapped to rice, Chinese cabbage and microbes, respectively (Table 1). Moreover, the database contains information about mutant rice phenotypes and the insertion site sequences of Ds-tagged lines that were used to generate the 115,000 Ac/Ds insertional mutation lines in japonica rice (Park et al., 2009). It provides information on 18,158 Ds lines including the photographic images, genetic characteristics and flanking sequence tag information. In 2011, we completed the genome sequence and analysis of the Chinese cabbage *Brassica rapa* and modeled 41,174 protein-coding genes in the entire genome (The *B. rapa* Genome Sequencing Project Consortium, 2011).

Genome analysis

In the post-genomic era, construction of an integrated genome database is important in providing researchers access to all information and analytical tools through an

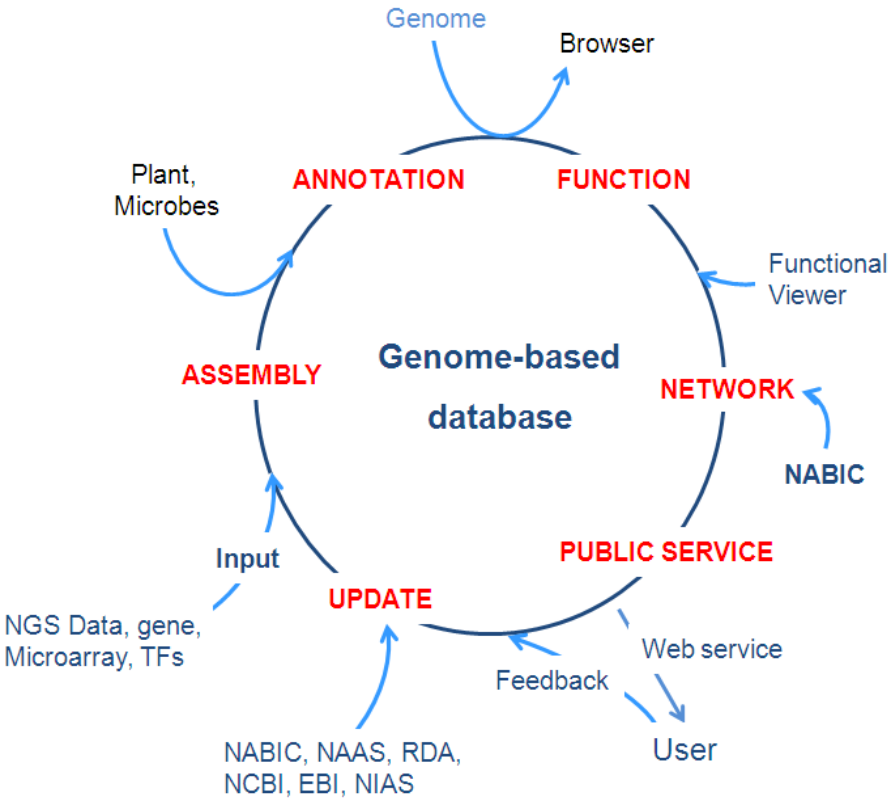

Figure 3. Overview of the design concepts and network processes used in the integrated genome-based database.

easy-to-use, customizable interface that combines services available from different sources. We have advanced the effectiveness of biological databases and developed genomics tools with a knowledge-based approach for functional genomics. Our genome-based database and analytical tools provide easier integration and interoperability between bioinformatics applications and the required data. Figure 3 shows the design concepts and process of information flow through the NABIC network.

Genome browser

To advance genomic research, we developed a genome browser for rice (Kim et al., 2009). In 2011, we constructed an integrated genome browser for computational analysis of the rice (*Oryza sativa*), Chinese cabbage (*B. rapa)* and microbe (*X. oryzae*) genomes. The browser consists of three major functional categories and provides specific genome analysis through three different viewing panels (Figure 4). Relationships between the genomic sequence and annotated data can be displayed. Through the three viewable panels that are accessible by clicking, the user can access information about individual genes along with functional annotation within selected banding regions of the entire chromosome. In particular, the overview panel shows the location of genes and specific markers, and the detailed panel displays genomic sequence features and genes. The base-pair panel exhibits the correlation between the ancestry of individuals and the common variability of pair-wise linkage. In addition, the comparative genome analysis between the *Arabidopsis* and *B. rapa* genomes, users can obtain new gene information resulting from comparative genomics methods and identify missing regions within a single genome (Kim et al., 2010b). The combination of genome-based database with the gene expression database provides an integrated tool for automatic multi-step analysis of microarray gene expression data. Bioinformatics tools can also be used to compare and evaluate gene expression data originating from newly developed gene expression systems.

Connections to other databases

The genetic resources database (http://www.genebank.go.kr/) provides information collected from domestic and exotic plants, insects and microbial species. This database has four major menus with which to search for information such as accession

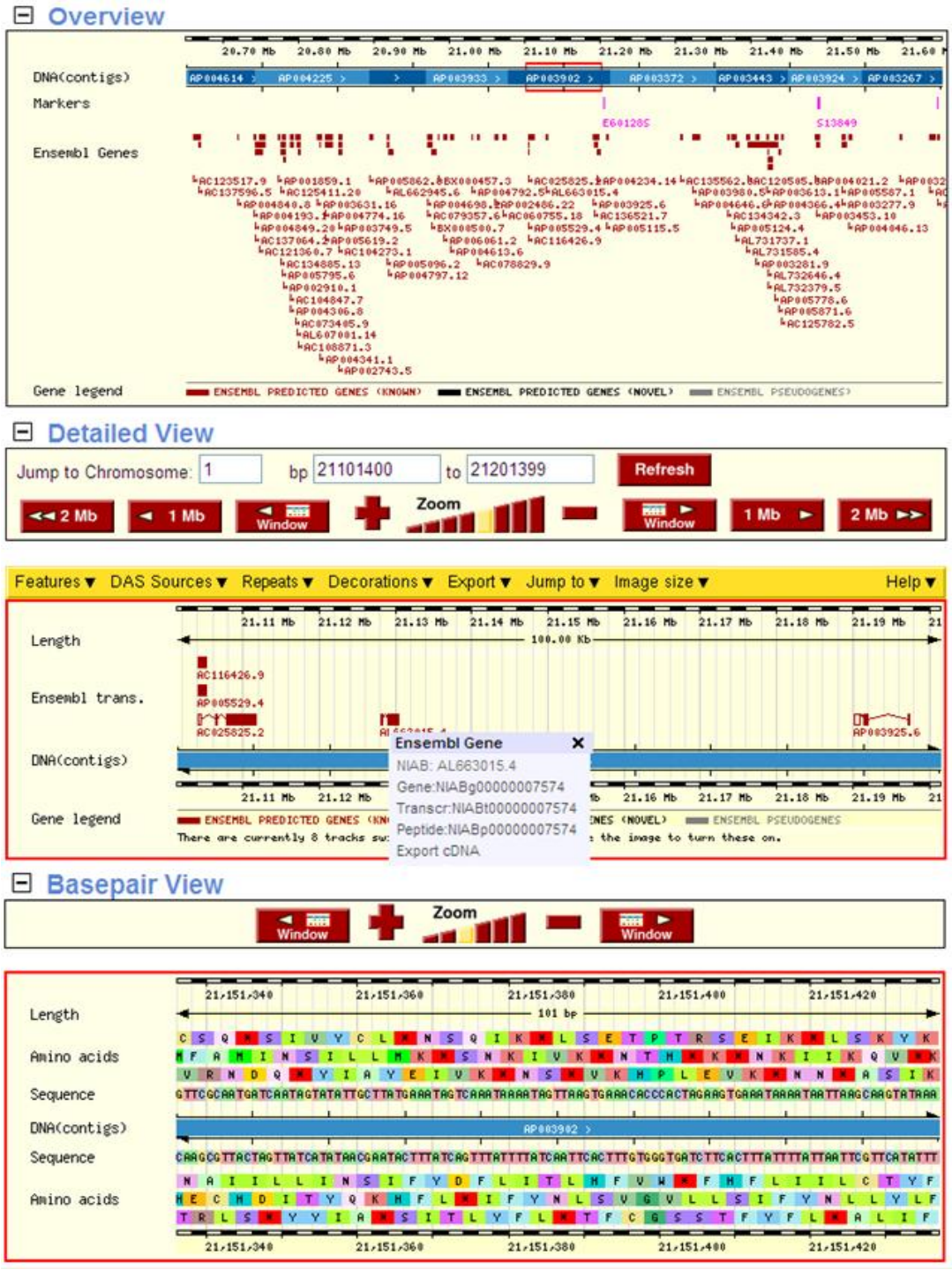

Figure 4. Screenshot of three different view panels from the genome browser for rice. This function has a zoom function and pull-down menus to allow the user to select the features to be displayed. The gray box in the detailed view panel is a floating menu that can access linked windows with additional information.

number, taxonomy, collection information and breeding. In addition, the database was developed to connect to a genome-based database of the NABIC. At present, the database provides information for over 158,000 accession numbers, including cereals and food legumes (113,413), industrial crops (17,988), vegetables and other agricultural crops (13,915), insects and microorganisms (12,924).

The biosafety database (http://biosafety.rda.go.kr/) provides information about living modified organisms (LMOs). Specifically, it provides information for international cooperation, assessment of environmental risks for national policymaking, guidelines for environmental risk assessment and standard protocols for cultivating agriculture. In addition, this database was assembled and provides information for legislation regarding LMOs, bio-safety examinations, import and export of LMOs, and trends in biosafety. The purpose of this database is to exchange and share legal information on LMOs and to offer the list of the approved LMOs under domestic law.

DISCUSSION

The NABIC was established in 2002 with the purpose of analyzing the genomes of agricultural crops and providing related services to professional genomic research institutes and societies (NAAS, 2010; Kim et al., 2011b). Our genome-based database provides information through a user-friendly web interface from searching genetic resources to genome infrastructure analysis. This database provides valuable genomic information, including projects, genetic markers, gene annotation and analytical tools. The NABIC has developed an integrated network to help navigate genomics, systems biology, metabolism and proteomics tasks. Our new database contributes to this informatics approach to agricultural biotechnology and can be extended to investigations of breeding new crops. In addition, to provide revolutionary technologies that deliver genomic information quickly and inexpensively, we constructed a system for next-generation sequencing (NGS) technologies to analyze massive sequencing data in 2011. It supports standard read, paired-end read, mate pair read analysis, expressed sequence tag (EST) sequence analysis and large-scale genome assembly. The NGS information can be integrated with existing genome-based database. The NABIC has contributed to the application of this informatics approach to agricultural biotechnology to support breeding for new crops. In the future, we will provide a genome-based database and bioinformatics tools to solve complex biological problems with NGS information. Moreover, users will be able to request the development of an integrated network to aid in navigating genomics, systems biology, metabolism and proteomics tasks. Finally, to solve complex biological problems using NGS information, we will provide a service to easily construct workflows and pipelines that combine two or more instructions.

ACKNOWLEDGEMENTS

This study was conducted with support from the Research Program for Agricultural Science and Technology Development (Project No. PJ006651), the National Academy of Agricultural Science, Rural Development Administration, Republic of Korea.

REFERENCES

s FB (2008). Bioinformatics, Genomics, and Proteomics: Getting the Big Picture. Biotechnol. 21st Century. Bioinf., 9: 94-95.

Kersey PJ, Staines DM, Lawson D, Kulesha E, Derwent P, Humphrey JC, Hughes DS, Keenan S, Kerhornou A, Koscielny G, Langridge N, McDowall MD, Megy K, Nuhn M, Paulini M, Pedro H, Toneva I, Wilson D, Yates A, Birney E (2012). Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. Nucleic Acids Res., 40: D91-7.

Kim CK, Han JH, Shin YH, Park SH, Yun DW, Ahn BO, Kim DH, Park BS, Hahn JH (2009). A genome browser database for rice (Oryza sativa) and Chinese cabbage (Brassica rapa). Afr. J. Biotechnol., 8: 5253-5259.

Kim CK, Kikuchi S, Kim YK, Park SH, Yoon YH, Lee GS, Choi JW, Kim YH, Park SC (2010a). Computational identification of seed-specific transcription factors involved in anthocyanin production in black rice. BioChip J., 4: 247-255.

Kim CK, Park SH, Kikuchi S, Kwon SJ, Park SY, Yoon UH, Park DS, Seol YJ, Hahn JH, Park SC, Kim DH (2010b). Genetic analysis of gene expression for pigmentation in Chinese cabbage (Brassica rapa). BioChip J., 4: 123-128.

Kim CK, Weon HY, Cho GT, Kwon SW, Park SC, Hong SB (2011a). The activity and integrated service for microbial resources at the Korean agricultural culture collection. Afr. J. Microbiol. Res., 5: 622-627.

Kim CK, Kim JA, Kikuchi S, Choi JW, Kim YK, Park HJ, Seol YJ, Park DS, Hahn JH, Kim YH (2011b). Computational identification of Chinese cabbage anthocyaninspecific genes. BioChip J., 5: 184-192.

Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC (2012). IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res., 40: D115-D122.

NAAS (2010). NAAS Annual report 2010. National Academy of Agricultural Science (NAAS), Korea, pp. 98-108.

Park DS, Park SK, Han SI, Wang HJ, Jun NS, Manigbas NL, Ahn BO, Yun DW, Yoon YH, Kim YH, Lee MC, Kim DH, Nam MH, Han CD, Kang HW, Yi GH (2009). Genetic variation through Dissociation (Ds) insertional mutagenesis system for rice in Korea: progress and current status. Mol. Breed., 24: 1-15.

Russ B (2007). Current progress in bioinformatics. Briefings bioinf., 8: 277-278.

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, Krasnov S, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Karsch-Mizrachi I, Ostell J, Panchenko A, Phan L, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, Wilbur WJ, Yaschenko E, Ye J (2012). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res., 40: D13-D25.

The *B. rapa* Genome Sequencing Project Consortium (2011). The genome of the mesopolyploid crop species Brassica rapa. Nature Genet., 43: 1035-1039.