Oryzabase: an integrated information resource for rice science

Yukiko Yamazaki*, Shingo Sakaniwa, Rie Tsuchiya, Ken-Ichi Nonomura and Nori Kurata

National Institute of Genetics, Yata 1111, Mishima, Shizuoka 411-8540, Japan

Oryzabase contains 16,000 biological resources, including wild rice, mutants, and experimental strains, together with 4000 trait genes collected from journal articles. The database also contains genetic and genomic maps, and 20,000 items from the published literature. Of these items, the most valuable are the biological resources, particularly the wild rice and mutant collections, because they have been thoroughly characterized by researchers and are accompanied by large quantities of data. The next most valuable content is the trait gene dictionary, which has been maintained continuously since 1995 and is based on published literature. Oryzabase also plays a role as an international site for the submission of rice genes. Besides maintaining original data, Oryzabase compiles genomic data that are systematically integrated with existing data holdings and provided with a user-friendly interface. Oryzabase can be accessed at http://www.shigen.nig.ac.jp/rice/oryzabase/.

Key Words: wild rice, database, mutant, phenotype, trait gene, gene submission system.

Introduction

Oryzabase is a comprehensive database that covers all aspects of the biological science of rice. Oryzabase began in 1997 as a database of information on available plant resources, together with a gene dictionary for rice, and it has developed as relevant knowledge and information have accumulated. Because rice was selected a target biological species for Japan’s National BioResource Project (NBRP), which began in 2002, Oryzabase now forms the database for NBRP-Rice (Yamazaki et al. 2010).

The most important and unique content of Oryzabase is the information on rice plant resources. This information, as well as the relevant plant-material collections that are available from NBRP-Rice, can be obtained from Oryzabase through online application. The second most important content of Oryzabase is the rice gene dictionary. The gene dictionary consists mainly of a compilation of trait genes (phenotypes corresponding to known mutations). However, since rice genome annotation data was opened to the public, the database has been supplemented without regard to phenotype with as many other genes as possible from among those reported in published papers. Oryzabase also contains a wide range of biological information on rice, such as physical maps of genomes, scientific papers, details of developmental stages of major organs, and basic information suitable for beginners in research.

Resource Information

The most characteristic set of resources consists of a collection of wild strains of rice. Of the 1700 strains of wild rice that are available to the public, 277 representative strains make up the core collections. These cover 18 species and 9 genomes (AA, BB, CC, BBCC, CCDD, EE, FF, GG and HHJJ). Oryzabase provides information on their geographical origins and preservation, such as the form of maintenance and the year of seed multiplication. The information for the core collection also contains well-characterized phenotype data, covering such aspects as the plant height, seed fertility, and heading date. The database also provides images of plant bodies and seeds, which are particularly useful for researchers new to wild rice strains. We recommend that beginners use the core collection because fundamental investigations of their phenotypes have already been conducted and the relevant information is readily available, thereby improving the efficiency of accumulation of research results.

Other collections contained among the resources include local varieties and various experimental strains, such as induced mutation lines, mutant stocks for genes, mapping parents, recombinant inbred lines, and chromosome-substitution lines. Oryzabase has been referred to as a database of rice mutants, because strains used for gene mapping show the representative phenotypes of mutants genes, and the database contains many images of these. On the other hand, the causal genes of N-methyl-N-nitrosourea (MNU)-induced

Communicated by K. Okuno
Received July 30, 2010. Accepted November 24, 2010.
*Corresponding author (e-mail: yyamazak@lab.nig.ac.jp)
mutation lines are, in almost all cases, unknown.

This means that induced mutants have great potential for use in identifying new genes. The featured phenotypes of all mutants are represented by means of 50 predetermined keywords that correspond to the classification of the genes, so that user can search the database for all mutation lines for a particular phenotype, regardless of whether the genes responsible for that particular phenotype are known.

All the NBRP resources are obtainable upon request. Requests for distribution can be made by means of shopping cart buttons on the pages corresponding to individual resources, and the procedure for distribution can be read by clicking the ‘Distribution’ tab on the menu bar. After October 2010, fees will be charged for handling of requests for resources, in accordance with the policy of the Ministry of Education, Culture, Sports, Science, and Technology.

Oryzabase also contains a collection of papers that have been published by researchers who have obtained results by using distributed NBRP resources. Although few such papers are currently in the collection, we continue to appeal to all researchers to whom resources have been distributed to feed back information on relevant papers, and we provide an online site for registering information on such papers. This information will be useful for later researchers who wish to use the same resources.

Please refer to the review article by Nonomura et al. (2010) in this issue for a detailed description on the collections of resources that are held by NBRP-Rice.

**Gene Dictionary and Registration System**

**Gene Registration Site**

The first gene dictionary for rice became available on the World Wide Web in 1997. At that time, the database covered approximately 800 genes and approximately 800 papers, even though genes relating to phenotypes were covered completely. The dictionary and database were updated continually, and for several years this was done with the assistance of retired researchers. However, following the completion of rice genome sequencing in 2004, it became impossible to maintain manual annotation as the number of genes predicted from sequence homology surged and the numbers of papers on rice genes increased. In addition, some confusion arose: for example, in some instances two or more names were given to the same gene. Therefore, in 2008 the Committee on Gene Symbolization, Nomenclature, and Linkage (CGSNL) of the Rice Genetics Cooperative produced guidelines aimed at overcoming these problems, and an online registration site adhering to these guidelines was set up for newly identified genes (McCouch 2008). Oryzabase hosts this gene registration site (http://www.shigen.nig.ac.jp/rice/oryzabase_submission/gene_nomenclature/) and it has received about 55 registrations so far. This figure is much smaller than the actual number of new genes that have been discovered. However, through mailing lists (rice-net and rice-e-net, which are hosted by Oryzabase), we continue to request that researchers register any new genes at our registration site before the relevant papers are submitted for publication. To avoid any duplication, each symbol name of a newly registered gene is open to the public as a ‘reserved gene’, even if the paper reporting on the gene has not yet been published.

**Manual Curation of Genes**

To update the gene dictionary, Oryzabase also continues to survey papers independently. Initial searches are conducted in the PubMed (http://www.ncbi.nlm.nih.gov/pubmed) and Scopus (http://www.scopus.com/home.url) bibliographic databases for any papers relating to rice that have the keyword ‘Rice’ or ‘Oryza’ in their titles or abstracts; these papers are automatically extracted as candidates for the selection of new genes for registration in Oryzabase. Because we do not have sufficient manpower to check all these automatically extracted papers manually, in the second step we prioritize journals on the basis of the numbers of gene candidates reported and on the impact factor of the relevant journal; this reduces the number of papers that need to be manually processed at a time and permits efficient annotation of genes to update the dictionary. If each new gene were to be registered by its discoverers, the gene dictionary would become complete and up to date.

**Gene Information**

Fig. 1 shows a detailed page for an individual gene (ex. SPW1). The page contains the CGSNL recommended gene symbol (Fig. 1(a)-1), the CGSNL gene name (Fig. 1(a)-2), various synonyms (Fig. 1(a)-3), a classification of the phenotype (Fig. 1(a)-4), information on the allele (Fig. 1(a)-5), a brief explanation of the phenotype (Fig. 1(a)-6), the chromosome number (Fig. 1(a)-7), the AK number of the cDNA sequence (Fig. 1(a)-8), the RAP ID (Rice Annotation Project 2008) (Fig. 1(a)-9) code for locus information, and the references (Fig. 1(a)-10) used for annotation. In addition, each page provides automatic access to related resources (Fig. 1(a)-12) and references (Fig. 1(a)-11) through a search program. Although automatic access by clicking may not always provide the required result, it has a high potential for providing information that is omitted in the manual annotation. Genes are linked to the gene ontology (GO) (The Gene Ontology Consortium 2010) (Fig. 1(a)-13) and the Gramene (Liang et al. 2008) database (Fig. 1(a)-14), so that the genetic information is linked to various internal and external items of useful information. We have just started assigning the Plant Ontology (Avraham et al. 2008) to genes.

On the right-hand side of each detailed page for an individual gene in Oryzabase is a message box (Fig. 1(a)-15) for users to provide feedback. Users can directly send us an e-mail to request that additions or corrections be made to the information for the gene.

The gene dictionary of Oryzabase covers only genes that have been reported in scientific papers, and no sequence data are in present in the dictionary unless they have been maintained.
reported in such papers. As a result, genetic information that is registered only in the DNA Data Bank is not included in the dictionary. However, such information is extremely valuable, and sequence data are sometimes made available to the public by the DNA Data Bank in advance of paper publication. Therefore, besides the dictionary, Oryzabase provide data on annotated DNA sequences of rice that are extracted from the DNA Data Bank.

Biological Literatures

Oryzabase provides two different types of reference data sets; one is used in the gene dictionary and contains over 5000 articles directly related to genes, and the other consists of rice-related literature extracted from the PubMed and Scopus databases. The latter data set contains over 20,000 papers. Oryzabase users can search the abstract of rice articles using Textpresso (Muller et al. 2004), a text-mining system for scientific literature.

Developmental stage of major organs

Oryzabase contains a set of pages consisting of detailed information on the staging of developmental processes of major organs and tissues of rice. These pages have been compiled from research papers (Itoh et al. 2005, Sakaguchi and Fukuda 2008) and they cover the developmental stages of ten organs and tissues: embryo, inflorescence, leaf, crown root, spikelet, ovule, anther, pollen mother cell, stomata, and vascular bundle. Each is classified by ‘name’ and ‘explanation of characteristic event’. In addition, information on which genes are expressed during a particular developmental stage and the genes that are responsible for identified mutations are also described with a direct link to the corresponding gene page. Although much other information on rice growth is publicly available, the information on these pages is unique in that it discusses in detail the stages in the developmental processes of organs and tissues of rice. Sharing of this information should be helpful in promoting research. The pages also contain many images that supplement the explanations, and these should be particularly helpful to beginners in the subject.

Maps and Tools

Genetic and Physical Maps

Oryzabase also contains genetic maps and physical maps
of rice. The genetic maps were produced before 1998 and are relatively old, but this is the only Web site on which they are publicly available. Physical maps are built on the basis of the latest build4 (to be updated to build5 in December 2010) provided by the International Rice Genome Sequencing Project (IRGSP). LOCUS data are based on the RAP database, and the cDNA data are based on the KOME database. Annotated DNAs of rice registered at the DNA Data Bank (INSD) are also mapped after independent analysis of the homology of sequences. Because DNAs of rice registered at INSD include DNAs from species other than *Oryza sativa*, accessions other than those from *O. sativa* are mapped in the genome viewer of Oryzabase.

Data on cDNA clones of wheat and barley, which are also included in the NBRP, are also mapped in the genome maps. The rice is a model plant of monocot and its genome is important as a reference genome to researchers who study wheat and barley because the genome sequences of these crops have not yet been determined. Therefore, data on which clones of wheat and barley are mapped in the homology domain of the rice genome are useful.

Features and accession numbers of genes mapped on the basis of INSD sequence data can also be located. At July 2010, Oryzabase contained 69,000 INSD entries. Regions of the entries are divided by gene, and mapped as divided entries. There are 395,000 divided entries, a number that almost one hundred times that of the trait genes (4100) described above. INSD includes unpublished genes, so it may be useful because it compensates for the delay in annotation of trait genes and it also permits the use of rice gene data that cannot be extracted from papers.

**Tools**

We have created the simple but useful tools such as ID checker, Sequence cutter, and Chromosome Map Tool.

Although we have adopted RAP ID as the LOCUS ID and AK number for cDNA, other IDs exist. We therefore provide an ID checker that can search across various IDs including the TIGR ID, ReSeq ID and SWISS-PROT ID. The “sequence cutter” is a tool for extracting part of genome sequence by specifying the start and end position of each chromosome. Oryzabase has just released the new tool; Chromosome Map Tool, at users request. This tool allow users to draw maps of the *Oryza sativa* genome using a list of locus IDs. The resultant maps can be downloaded in the SVG, TIFF, EPS or JPEG file format.

**Collaborations with Other Institutions**

The wild rice collections of Oryzabase were distributed to the International Rice Research Institute (IRRI: http://www.irri.org/GRC/GRCHome/home.htm), who preserve and maintain them and investigate their phenotypes. Detailed pages for wild rice collections in Oryzabase are linked to pages of the corresponding IRRI collections to provide easy access to the results of phenotype investigations performed independently by the IRRI. Discrepancies in the numbering of the collections have occurred over a long time, so we are now reviewing both collections and correcting these discrepancies.

**Gramene** is a resource for comparative grass genomics held in the United States of America. We exchange trait gene data and reference information with Gramene, and Oryzabase also collaborates within the plant ontology consortium. We are also cooperating with the RAP in a project to relate LOCUS IDs to Oryzabase gene IDs of trait genes.

**Rice Genetic Newsletter (RGN)**

The Rice Genetic Newsletter (RGN), which is a newsletter of the Rice Genetics Cooperative, is also made available to the public through Oryzabase. RGN shares a server of newsletter archives with Oryzabase, although this activity is independent of Oryzabase. The publication of RGN ceased at issue number 25 in 2010 and the registration of new genes, which was one of the major roles of RGN, has been taken over by Oryzabase.

**Portal Site**

Oryzabase also functions as a portal site for information resources on rice. On the right-hand side of the home page of the site, under the heading ‘Featured Links’ there are links to representative Web sites that have close relationships with Oryzabase, including IRGSP, several Web sites of resource information on rice researches supported by the Ministry of Agriculture, Forestry and Fisheries of Japan, and Gramene. In the menu on the left-hand side of the home page there is a link to the ‘Worldwide Science Information Desk’, a menu in the form of a table that contains classified links to various Web sites that contain information relating to rice science.

**Future Plans**

As data on genomes and sequences accumulate, information is increasingly used across various biological species. In the plant-science community, new trends are developing, such as the construction of plant ontologies and research on the systems biology of specialization in plants. Research on rice is also diversifying rapidly. For example, there has been growth in the analysis of polygenic traits and in studies on species other than *Oryza sativa*, the species originally used for analysis of the rice genome and which has been intensively studied.

We are currently preparing the next version of Oryzabase to respond to this growth in complex information. We intend to improve the database to make it more reliable by means of an efficient combination of automatic annotation, manual annotation, and community annotation. Moreover, although Oryzabase contains many valuable images, the overall scheme for storing and retrieving of images in the existing display is difficult for users to understand because of its
deeply nested structure. We are therefore searching for a method to make more-effective use of image information.

Resource information is a treasure trove of data, and knowledge and information continue to accumulate as resources are studied in greater detail. Because Oryzabase is widely used as a resource database that meets the needs of the time, we aim to adapt rapidly as knowledge and information change.

Literature Cited


