PCP
PLANT & CELL PHYSIOLOGY

Plant Genome DataBase Japan (PGDBj): A Portal Website for the Integration of Plant Genome-Related Databases

Erika Asamizu[1], Hisako Ichihara[1], Akihiro Nakaya[2], Yasukazu Nakamura[1], Hideki Hirakawa[1], Takahiro Ishii[1], Takuro Tamura[3], Kaoru Fukami-Kobayashi[4], Yukari Nakajima[1] and Satoshi Tabata[1,*]

[1]Department of Plant Genome Research, Kazusa DNA Research Institute, 2-6-7 Kazusa-kamatari, Kisarazu, Chiba, 292-0818 Japan
[2]Center for Transdisciplinary Research, Niigata University, 1-757 Asahimachi-dori, Chuo-ku, Niigata, 951-8585 Japan
[3]LINE Co., Ltd., 5-201 Kandamatsunaga-cho, Tokyo, 101-0023 Japan
[4]RIKEN BioResource Center, 3-1-1 Koyadai, Tsukuba, Ibaraki, 305-0074 Japan
*Corresponding author: Fax: +81-438-52-3918; E-mail, tabata@kazusa.or.jp
(Received September 4, 2013; Accepted December 7, 2013)

The Plant Genome DataBase Japan (PGDBj, http://pgdbj.jp/?ln=en) is a portal website that aims to integrate plant genome-related information from databases (DBs) and the literature. The PGDBj is comprised of three component DBs and a cross-search engine, which provides a seamless search over the contents of the DBs. The three DBs are as follows. (i) The Ortholog DB, providing gene cluster information based on the amino acid sequence similarity. Over 500,000 amino acid sequences of 20 Viridiplantae species were subjected to reciprocal BLAST searches and clustered. Sequences from plant genome DBs (e.g. TAIR10 and RAP-DB) were also included in the cluster with a direct link to the original DB. (ii) The Plant Resource DB, integrating the SABRE DB, which provides cDNA and genome sequence resources accumulated and maintained in the RIKEN BioResource Center and National BioResource Projects. (iii) The DNA Marker DB, providing manually or automatically curated information of DNA markers, quantitative trait loci and related linkage maps, from the literature and external DBs. As the PGDBj targets various plant species, including model plants, algae, and crops important as food, fodder and biofuel, researchers in the field of basic biology as well as a wide range of agronomic fields are encouraged to perform searches using DNA sequences, gene names, traits and phenotypes of interest. The PGDBj will return the search results from the component DBs and various types of linked external DBs.

Keywords: Database integration • DNA marker • Ortholog • Plant genome • Plant resource • QTL.

Abbreviations: DB, database; LOD, logarithm of odds; QTL, quantitative trait locus; RAP-DB, The Rice Annotation Project database; SABRE, Systematic consolidation of Arabidopsis and other Botanical Resources; SNP, single nucleotide polymorphism; SSR, simple sequence repeat; TAIR, The Arabidopsis Information Resource.

Introduction

The genome sequence of the dicot model plant *Arabidopsis thaliana* (thale cress) was published in 2000 (Arabidopsis Genome Initiative 2000). The genome annotation has been updated; the latest version of TAIR, TAIR10 (The Arabidopsis Information Resource, http://www.arabidopsis.org/; Lamesch et al. 2012), contains 27,416 protein-coding genes, adding approximately 2,000 new gene models to the previous release. The gene annotations have also been updated for the monocot model plant *Oryza sativa* (rice) by incorporating resequencing data generated with next-generation sequencers (Kawahara et al. 2013). The latest assembly, 'Os-Nipponbare-Reference-IRGSP-1.0' and the annotation information are available at RAP-DB (The Rice Annotation Project database, http://rapdb.dna.affrc.go.jp/; Sakai et al. 2013). Technical advances in DNA sequencing have enabled the genome sequencing of other plant species, including crops and species with more complex genome structures. To date, the entire genomes of >40 plant species have been sequenced and published (NCBI Genome database, http://www.ncbi.nlm.nih.gov/genome/; NCBI Resource Coordinators 2013). In addition, resequencing data are rapidly accumulating, not only for model species but also for various cultivars.

Several DBs provide integrated information related to plant genomes. For example, the Gramene DB (http://www.gramene.org/; Jaiswal 2011, Youens-Clark et al. 2011) provides data resources for comparative analysis of grass genomes. TAIR maintains molecular biological and genetic data of the model plant *A. thaliana*, and provides GBrowse (the Generic Genome Browser, http://gbrowse.org/; Donlin 2009) for eight plant species, allowing comparison of gene models with *A. thaliana*. The Sol Genomics Network (SGN) provides genomic, genetic, phenotypic and taxonomic information for members of the Solanaceae family (http://solgenomics.net/; Bombarely et al. 2011). The PlantGDB (http://www.plantgdb.org/; Duvick

Plant Cell Physiol. 55(1): e8(1–7) (2014) doi:10.1093/pcp/pct189, available online at www.pcp.oxfordjournals.org

et al. 2008) contains genome sequence data of 16 dicot and seven monocot species in addition to several algal species. This DB aims at developing and providing a method for plant genome annotation. Phytozome (http://www.phytozome.net/; Goodstein et al. 2012) provides links to 41 genome sequences of plant species, accompanied by annotations and evolutionary relationships for comparative genomic analysis.

Numerous plant genome-related DBs also exist in Japan. Various types of information such as DNA sequence, transcriptome, protein, metabolite and phenotypic data are provided. Since each research group generally establishes its own DB, the data are provided in different formats, thus making it difficult for users to locate the desired information properly and efficiently. To solve this problem and provide a wide range of researchers with the benefit of plant genome-related data, we have constructed the Plant Genome DataBase Japan (PGDBj, http://pgdbj.jp/?ln=en). Users of the PGDBj will be efficiently guided to necessary information from accumulated and linked data, by simply performing a Google-like search (**Fig. 1a**) in the PGDBj cross-search window.

Cross-Search System for the PGDBj

A cross-search system over all contents of the PGDBj is provided (**Fig. 1a**). Full text searches are run by Hyper Estraier (http://fallabs.com/hyperestraier/). Text data in each component DB are pre-selected and stored (indexed) in order to speed up the search process, to create a common search platform for the content of each of the component DBs, which were originally registered in different formats, and to allow less stringent keyword matches (partial matches). When a search is executed, the engine scans the indexed text data and returns a result with respect to the component DB category. A direct search against each component DB is also available (**Fig. 1b–d**); however, since direct searches will not utilize the above-mentioned system, keyword selection by the users should be more stringent (perfect match required).

An example of a cross-search result is indicated in **Fig. 2**. The window is divided into two parts; DB categories are shown in the left-hand panel and the list of entries retrieved from the respective DB is shown in the right-hand panel. It is possible to narrow the search result by selecting the type of DB, organism and/or taxonomy rank. By clicking the hypertext links in the right-hand panel (retrieved entries from DBs), users are guided to detailed information of either the content of internal component DBs or linked external DBs.

As shown in **Figs. 1** and **2**, users first encounter a Google-like search window in the PGDBj. The search will retrieve entries in internal component DBs, i.e. the Ortholog DB, the Plant Resource DB and the DNA Marker DB, and the retrieved entries will be linked to external plant genome-related DBs, realizing a seamless and integrated search against a wide range of DBs and relevant information.

Construction of the Ortholog DB and Integration of Plant Genome Databases

The Ortholog DB for plant genome database integration

We defined 'orthologous genes' by computational clustering of genes with their amino acid sequence similarities, and stored them in the Ortholog DB. To integrate genome-related DBs of various plant species, sequence similarities are the fundamental means of linking individual DB entries describing genes with even unknown functions or partial fragments of sequences. Most plant genome DBs available on the Internet are equipped with a similarity search feature to find homologous entries in their sequence data. Using such interfaces, the DBs can be searched repeatedly by indicating sequences as queries, and the corresponding DB entries can be accumulated. Although each search result lists sequences similar to the query in descending order of similarity scores, it is not always easy to comprehend the whole relationship constituted by a series of query–result relationships obtained by the repetition of a search across DBs. Therefore, based on these similarity relationships, sequences of genes are categorized into a set of clusters in each of which the members are more similar than some threshold to each other. We refer to this set as an ortholog. However, the threshold of similarity that can appropriately aggregate sequences as clusters is not clear or cannot be naturally determined under only one simple condition, owing to the hierarchical and nested structures of functions and the characteristics of genes. To solve this limitation, boundaries of clusters in an ortholog table should be variable in conjunction with the similarity threshold, and such elastic characteristics of data structures in the DB can provide flexible search functions, e.g. by zooming in and out around a particular gene of interest in the search space. By associating entries in the genome DBs with the sequences in the Ortholog DB via sequence similarity, and then reciprocally traversing these associations among DB entries through the clusters in the Ortholog DB, flexible and deep linking of the genome DBs of various organisms is realized.

Construction of the Ortholog DB

Approximately 500,000 amino acid sequences of 20 plant species classified in the Viridiplantae kingdom (**Table 1**) were obtained from the NCBI RefSeq database (Release 57, NCBI Reference Sequence database, http://www.ncbi.nlm.nih.gov/refseq/; NCBI Resource Coordinators 2013). Currently, organisms in the NCBI RefSeq database that have $>1{,}000$ amino acid sequences are included in the Ortholog DB. We performed local alignments of all pairs of sequences using the NCBI BLAST program (Altschul et al. 1990) with default parameters. Sequences of individual species were accumulated in separate FASTA files. We carried out BLAST searches of a query sequence against the accumulated sequences, and evaluated the BLAST *E*-values. This procedure was executed for all sequences, and we stored the similarity relationships in tables of a relational DB

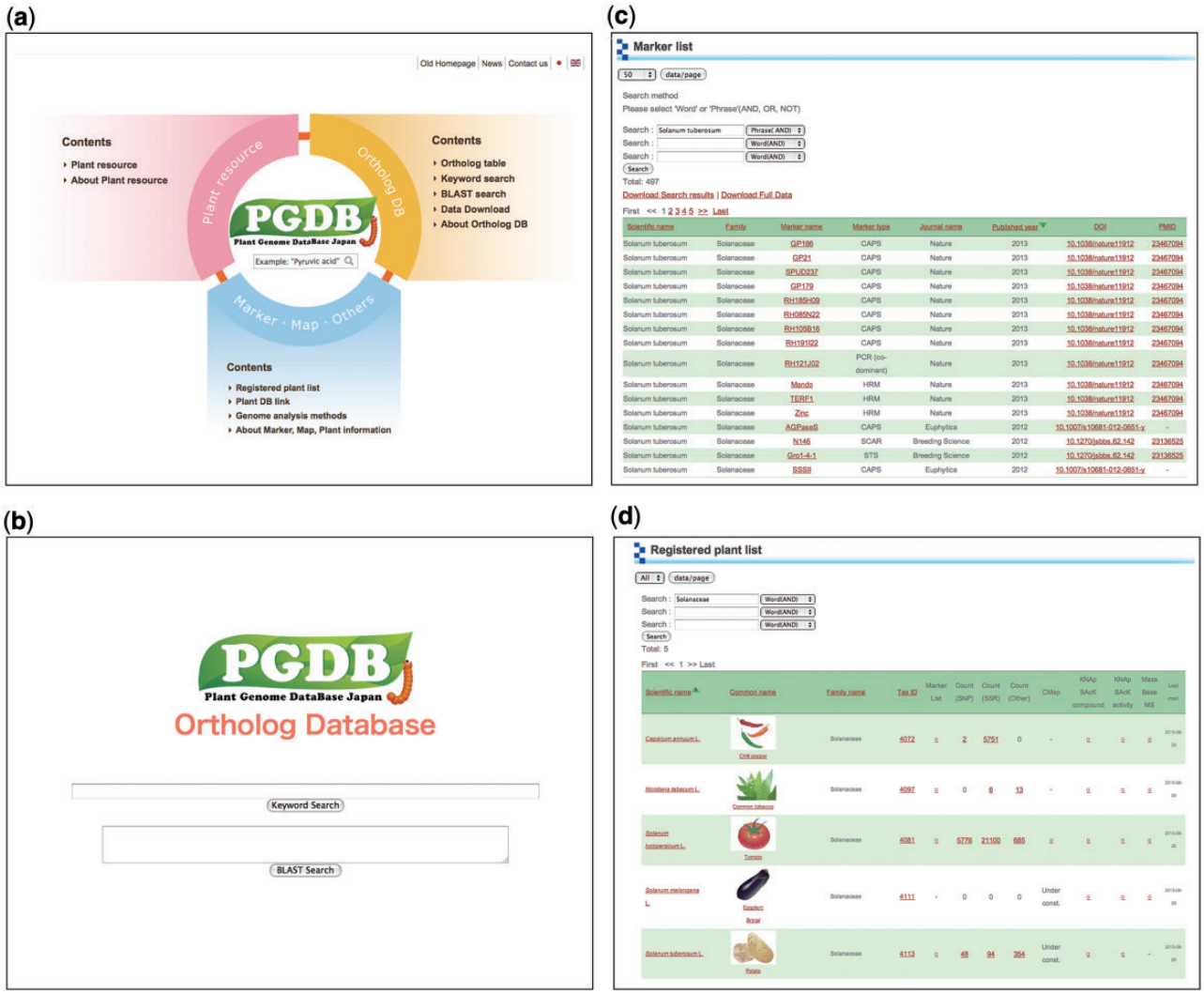

Fig. 1 The PGDBj portal website (http://pgdbj.jp/?ln=en). The cross-search form on the front page is shown in (a). The following component databases can be accessed directly: the Ortholog DB (b), the DNA marker DB (c) and the Registered plant list (d).

which is used to generate clusters of orthologous sequences. The clusters were generated in a hierarchical and recursive manner along a phylogenetic tree to reflect the taxonomic relationships. To arrive at initial clusters in the recursive process of generation, we generated dendrograms of clustering results by a single-linkage method using sequence similarity relationships satisfying a cut-off condition (a BLAST *E*-value $\leq 10^{-5}$) with respect to the sequence sets from an individual organism. If one-half of all the pairs of the sequences in a subtree of a dendrogram satisfied the BLAST *E*-value cut-off condition, we concluded that those sequences constituted a cluster of similar sequences. Starting with those initial clusters at the organism level, the aggregation of clusters in subtaxa into clusters in their supertaxon by a single-linkage method in a manner similar to that at the organism level was repeated until clusters at the Viridiplantae kingdom level were obtained. The clusters in all the taxa were stored in a relational DB table and identified by pairing the taxonomy ID of the NCBI taxonomy DB with the unique cluster ID assigned by the Ortholog DB in each taxon. The three major tables of the DB—sequences, similarities and clusters—constitute the core of the Ortholog DB. From the PGDBj cross-search page (**Fig. 2a**), the clusters can be searched by a keyword in the sequence annotation, and the results are displayed by clicking on 'Ortholog DB' in the top-level category 'DB' in the left-hand panel of the result window (**Fig. 2b-1**). The clusters at each level of taxonomy can be selected by clicking a species name (e.g. '*Chlamydomonas reinhardtii*', '*Oryza sativa*' or '*Arabidopsis thaliana*') or a taxonomy rank name (e.g. 'kingdom', 'phylum' or 'class') in the left-hand panel of the result window (**Fig. 2b-3**).

Associating plant genome databases with the Ortholog DB

For each sequence obtained from the plant genome database, local alignments were performed using the NCBI BLAST

Fig. 2 PGDBj cross-search using the keyword 'tomato'. In the initial output of the search, the DB name is selected in the left-hand panel (a-1) and the retrieved entries from the respective DB are listed in the right-hand panel (a-2). (b), (c) and (d) show results filtered by 'Ortholog DB', 'DNA Marker DB' and 'Plant Resource DB', respectively. The results could be narrowed further by selecting the species name and/or taxonomy (see the text for details).

program against all the sequences in the Ortholog DB. Sequence similarities between the query and DB sequences were stored in a DB table. By joining the table with the core part of the Ortholog DB—the tables of sequences, similarities and clusters—each DB entry of the plant genome DBs can be associated with the Ortholog DB. By identifying the entries of the genome DBs associated with specific clusters in the Ortholog DB, relevant entries can be accumulated across the multiple plant genome DBs appearing in the left-hand panel of the result window of the cross-search (**Fig. 2b-2**). The hierarchical structures inherent in the Ortholog DB can be used to manage the range of the accumulation and the degree of relevance. The clusters generated at each taxonomic rank allow users to focus on functions and characteristics specific to the clusters by tracing the upper–lower relationships among the clusters along the phylogenetic tree. In each of these clusters, member sequences are hierarchically arranged in a tree structure according to their sequence similarities, allowing cluster members more relevant to one of the sequences to be efficiently extracted. The extraction of these members can then lead to counterparts in the plant genome DBs.

Integration of the Plant Genome Databases with DNA Markers and Linkage Map Information

Curation of markers, maps and QTL information

DNA markers and genetic linkage maps are prerequisite tools for performing molecular genetic studies of plants. Particularly in crops, many types of DNA markers have been developed and used for constructing linkage maps. Such information is useful for mapping the loci of agronomic importance onto chromosomes, and is often quantitative rather than qualitative. Sequence-tagged DNA markers provide opportunities to perform intra- or interspecies comparisons of genome structure. Thanks to recent technical advances, large numbers of DNA markers can be developed cost-effectively, regardless of plant species, making it possible to perform genome comparisons between different species with high accuracy. We therefore chose to use DNA markers for plant genome DB integration and selected 55 species belonging to 24 families. For 10 species, large-scale DNA marker data have been published by the Kazusa DNA Research Institute (Hayashi et al. 2001, Sato

Table 1 The numbers of amino acid sequences integrated into the PGDBj

Scientific name (common name)	Taxonomy ID	No. of sequences
Arabidopsis lyrata subsp. *lyrata* (lyrate rockcress)	81972	32,549
Arabidopsis thaliana (thale cress)	3702	35,375
Medicago truncatula (barrel medic)	3880	46,092
Glycine max (soybean)	3847	44,570
Ricinus communis (castor bean)	3988	31,344
Populus trichocarpa (black cottonwood)	3694	40,521
Vitis vinifera (grape)	29760	23,877
Solanum lycopersicum (tomato)	4081	1,203
Brachypodium distachyon (purple false brome)	15368	24,689
Oryza sativa Japonica Group (rice)	39947	28,555
Zea mays (maize)	70448	22,383
Sorghum bicolor (sorghum)	296587	33,005
Selaginella moellendorffii (spikemoss)	564608	34,817
Physcomitrella patens subsp. *patens* (moss)	3068	35,894
Ostreococcus lucimarinus CCE9901 (green alga)	3055	7,603
Ostreococcus tauri (green alga)	70448	7,987
Micromonas sp. RCC299 (green alga)	296587	10,140
Micromonas pusilla CCMP1545 (green alga)	564608	10,269
Volvox carteri f. nagariensis (green alga)	3068	14,436
Chlamydomonas reinhardtii (chlamydomonas)	3055	14,489

Table 2 The numbers of DNA markers and QTLs that have been integrated into the PGDBj

Scientific name (common name)	Markers and QTLs
Arachis hypogaea (peanut)	15,125 [a]
Capsicum annuum (chili pepper)	5,753 [a]
Eucalyptus camaldulensis (murray red gum)	5,684 [a]
Fragaria × *ananassa* (strawberry), *Fragaria vesca* (wild strawberry)	5,589 [a], 341 [b], 74 [c]
Glycine max (soybean)	7,020 [a]
Lotus japonicas (Japanese trefoil)	1,155 [a]
Raphanus sativus (radish)	4,024 [a]
Solanum lycopersicum (tomato)	27,561 [a]
Trifolium pretense(red clover)	7,782 [a]
Trifolium repens (creeping white clover)	1,193 [a]
Brachypodium distachyon (purple false brome)	214 [b], 3 [c]
Camellia sinensis (tea)	881 [b]
Carica papaya (papaya)	53 [b], 14 [c]
Citrus unshiu (satsuma mandarin)	51 [b]
Ipomoea nil (Japanese morning glory)	75 [b]
Lactuca sativa (garden lettuce)	287 [b], 158 [c]
Phoenix dactylifera (date palm)	42 [b]
Ricinus communis (castor bean)	223 [b]
Vitis vinifera (grape)	496 [b], 264 [c]

[a] Number of markers integrated from the databases at KDRI.
[b] Number of markers curated from publications.
[c] Number of QTLs curated from publications.

et al. 2001, Hisano et al. 2007, Isobe et al. 2009, Shirasawa et al. 2010a, Shirasawa et al. 2010b, Hirakawa et al. 2011, Koilkonda et al. 2011, Shirasawa et al. 2011, Isobe et al. 2012, Shirasawa et al. 2012a, Shirasawa et al. 2012b, Isobe et al. 2013, Shirasawa et al. 2013a, Shirasawa et al. 2013b, Shirasawa et al. 2013c), and this marker information has already been integrated into PGDBj (**Table 2**). For six of these species, linkage map information is available for viewing.

DNA marker and linkage map DBs for model plants (Arabidopsis, rice and several major crop species) have been generated. However, for other plant species, such information is available only in the published literature. One of our major objectives is to curate information on DNA markers and quantitative trait loci (QTLs) from the literature and make it available in our DB. Before performing manual curation, we searched the literature and DBs using keywords to select target species. As a result, 23 species for which there were substantial numbers of publications on DNA markers and QTLs were selected. The curated marker information included marker sequences (e.g. primers and amplicons), typing methods, population and linkage map information, and values related to population genetic analyses. In the QTL curating process, we accumulated data on chromosomal positions and LOD (logarithm of odds) scores in addition to the genetic effects (e.g. dominant effect and additive effect). The present status of the curated results is indicated in **Table 2**. In the PGDBj cross-search, users can access the marker information by clicking on 'DNA marker DB' (**Fig. 2c-1**), and filter the result by marker types such as 'SSR' or 'SNP' (**Fig. 2c-2**).

Database links

Currently, >570 DB and website links related to plant genome-related research and information have been accumulated. They were classified according to the content, such as 'Genome resource', 'Genome database', 'Marker database', 'Expression database' or 'Omics database', and also by the plant species described. The DB links are searchable from the PGDBj cross-search (**Fig. 2a-1**). We provide links to currently developed and updated DBs, e.g. The Chloroplast Function Database II (Myouga et al. 2013), KNApSAcK (Afendi et al. 2012), RiceFOX (Sakurai et al. 2011) and ATTED-II (Obayashi et al. 2011).

Integration of Plant Resource Databases

Arabidopsis research has been greatly accelerated through a coordinated effort of the international community to develop, accumulate and share biological resources, e.g. cloned genes and mutant lines. In crops, germplasm stored over the long term in the form of seeds has played an important role in unraveling the genetic variation. The importance of so-called

Table 3 The numbers of plant resources integrated into the PGDBj

Scientific name (common name)	No. of clones
Arabidopsis thaliana (thale cress)	246,605
Thellungiella halophila	19,429
Brassica rapa subsp. *Pekinensis* (Chinese cabbage)	12,069
Nicotiana tabacum (tobacco)	22,221
Solanum lycopersicum (tomato)	120,596
Lotus japonicus (birdsfoot trefoil)	160,652
Glycine max (soybean)	37,862
Manihot esculenta (cassava)	19,450
Striga hermonthica (striga)	35,198
Ipomoea nil (morning glory)	33,641
Triticum aestivum (wheat)	483,683
Hordeum vulgare (barley)	139,934
Populus nigra var. *italica* (poplar)	23,100
Physcomitrella patens (moss)	149,363

bioresources is recognized in Japan, and the National BioResource Project (NBRP) collects and maintains various resources of Arabidopsis, rice, wheat, barley, algae, chrysanthemum, morning glory, *Lotus japonicus*, soybean and tomato (Yamazaki et al. 2010). The RIKEN BioResource Center (BRC) provides full-length cDNA clones and sequences, seeds and cultured cell lines of Arabidopsis, moss, poplar, cassava, etc. Moreover, tens of thousands of resources have been collected and distributed for grain, legumes and grass by the NIAS Genebank at the Ministry of Agriculture, Forestry, and Fisheries (http://www.gene.affrc.go.jp/index_en.php; Takeya et al. 2011).

To facilitate the use of plant bioresources, we have incorporated the SABRE2 (Systematic consolidation of Arabidopsis and other Botanical REsources2, http://saber.epd.brc.riken.jp/sabre2/SABRE2.cgi; Fukami-Kobayashi et al. 2014) system in the PGDBj, which provides information on genes homologous to the Arabidopsis TAIR gene models in various plant species (**Table 3**). Users are guided directly to the respective resource center's site by clicking on the hypertext link.

Citrus is an important fruit crop in Japan, and original resources, such as a total of approximately 900 individuals of wild species and domestic cultivars and a collection of cDNA libraries, have been created and maintained in Kinki University and the National Institute of Fruit Tree Science, Japan. The citrus resource information, along with the genome sequence information for citrus species (in preparation), will be integrated and made searchable in the PGDBj in the near future.

Portal Website Implementation

The PGDBj portal website was constructed using open source technologies on a Linux operating system (Red Hat Enterprise Linux ver. 5.7). The MySQL database management system (ver. 5.0.95) was used to store and manage the contents. An Apache HTTP server (ver. 2.2.3), Hypertext Preprocessor (PHP) (ver. 5.3.26), Perl (ver. 5.8.8), JavaScript (ver. 1.8) and XHTML (ver. 1.0) were used to create the query-builder module for connecting user queries to the DB. The Joomla! content management system (ver. 2.5.14) was used to build the website. The Hyper Estraier search system (ver. 1.4.13, http://fallabs.com/hyperestraier/) was adopted to carry out full-text searches against all contents of the DBs.

Conclusion

A unique feature of the PGDBj is that users have access to various plant genome DBs through the Ortholog DB, which serves as the system hub. The gene cluster information is useful to speculate about gene families and evolutionary relationships among genes across different species, leading to the discovery of new genes and elucidation of their function. Another feature is that the PGDBj provides DNA marker and QTL information of important agronomic traits manually curated from the literature. The integration of such information will encourage the use of the PGDBj by researchers in the field, and the application of this data will accelerate the crop improvement process.

Funding

This work was supported by the Japan Science and Technology Agency (JST) [the Life Science Database Integration Project conducted by the National Bioscience Database Center (NBDC)].

Acknowledgments

We are grateful to Drs. Hideki Hatanaka, Asuka Bandoh and Junichi Onami of the NBDC for their helpful comments. We also thank Ms./Mr. Yuriya Jitsukata, Hiroko Maita, Mitsuyo Kohara, Tsunakazu Fujishiro, Shinobu Nakayama, Tomoko Akutsu, Kaori Satoh, Hiroko Egashira, Yuka Watanabe and Kotaro Koike for their technical assistance.

Disclosures

The authors have no conflicts of interest to declare.

References

Afendi, F.M., Okada, T., Yamazaki, M., Hirai-Morita, A., Nakamura, Y., Nakamura, K. et al. (2012) KNApSAcK Family Databases: integrated metabolite–plant species databases for multifaceted plant research. *Plant Cell Physiol.* 53: e1.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215: 403–410.

Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796–815.

Bombarely, A., Menda, N., Tecle, I.Y., Buels, R.M., Strickler, S., Fischer-York, T. et al. (2011) The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. *Nucleic Acids Res.* 39: D1149–D1155.

Donlin, M.J. (2009) Using the Generic Genome Browser (GBrowse). *Curr. Protoc. Bioinformatics* 28: 9.9.1–9.9.25.

Duvick, J., Fu, A., Muppirala, U., Sabharwal, M., Wilkerson, M.D., Lawrence, C.J. et al. (2008) PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res.* 36: D959–D965.

Fukami-Kobayashi, K., Nakamura, Y., Tamura, T. and Kobayashi, M. (2014) SABRE2: a database cross-searching plant genetic resources in Japan. *Plant Cell Physiol.* 55: e5.

Goodstein, D.M., Shu, S., Howson, R., Neupana, R., Hayes, R.D., Fazo, J. et al. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 39: D1178–D1186.

Hayashi, M., Miyahara, A., Sato, S., Kato, T., Yoshikawa, M., Taketa, M. et al. (2001) Construction of a genetic linkage map of the model legume *Lotus japonicus* using an intraspecific F_2 population. *DNA Res.* 8: 301–310.

Hirakawa, H., Nakamura, Y., Kaneko, T., Isobe, S., Sakai, H. and Kato, T. (2011) Survey of the genetic information carried in the genome of *Eucalyptus camaldulensis*. *Plant Biotechnol.* 28: 471–480.

Hisano, H., Sato, S., Isobe, S., Sasamoto, S., Wada, T., Matsuno, A. et al. (2007) Characterization of the soybean genome using EST-derived microsatellite markers. *DNA Res.* 14: 271–281.

Isobe, S.N., Hirakawa, H., Sato, S., Maeda, F., Ishikawa, M., Mori, T. et al. (2013) Construction of an integrated high density SSR linkage map in cultivated strawberry (*Fragaria* × *ananassa*) and its applicability. *DNA Res.* 20: 79–92.

Isobe, S.N., Hisano, H., Sato, S., Hirakawa, H., Okumura, K., Shirasawa, K. et al. (2012) Comparative genetic mapping and discovery of linkage disequilibrium across linkage groups in white clover (*Trifolium repens* L.). *G3 (Bethesda)* 2: 607–617.

Isobe, S., Kölliker, R., Hisano, H., Sasamoto, S., Wada, T., Klimenko, I. et al. (2009) Construction of a consensus linkage map for red clover (*Trifolium pretense* L.). *BMC Plant Biol.* 9: 57.

Jaiswal, P. (2011) Gramene database: a hub for comparative plant genomics. *Methods. Mol. Biol.* 678: 247–275.

Kawahara, Y., Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S. et al. (2013) Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* 6: 4.

Koilkonda, P., Sato, S., Tabata, S., Shirasawa, K., Hirakawa, H., Sakai, H. et al. (2011) Large-scale development of expressed sequence tag-derived simple sequence repeat markers and diversity analysis in *Arachis* spp. *Mol. Breed.* 30: 125–138.

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R. et al. (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 40: D1202–D1210.

Myouga, F., Akiyama, K., Tomonaga, Y., Kato, A., Sato, Y., Kobayashi, M. et al. (2013) The Chloroplast Function Database II: a comprehensive collection of homozygous mutants and their phenotypic/genotypic traits for nuclear-encoded chloroplast proteins. *Plant Cell Physiol.* 54: e2.

NCBI Resource Coordinators. (2013) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 41: D8–D20.

Obayashi, T., Nishida, K., Kasahara, K. and Kinoshita, K. (2011) ATTED-II updates: condition-specific gene coexpression to extend coexpression analyses and applications to a broad range of flowering plants. *Plant Cell Physiol.* 52: 213–219.

Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y. et al. (2013) Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol.* 54: e6.

Sakurai, T., Kondou, Y., Akiyama, K., Kurotani, A., Higuchi, M., Ichikawa, T. et al. (2011) RiceFOX: a database of Arabidopsis mutant lines overexpressing rice full-length cDNA that contains a wide range of trait information to facilitate analysis of gene function. *Plant Cell Physiol.* 52: 265–273.

Sato, S., Kaneko, T., Nakamura, Y., Asamizu, E., Kato, T. and Tabata, S. (2001) Structural analysis of a *Lotus japonicus* genome. I. Sequence features and mapping of fifty-six TAC clones which cover the 5.4 Mb regions of the genome. *DNA Res.* 8: 311–318.

Shirasawa, K., Asamizu, E., Fukuoka, H., Ohyama, A., Sato, S., Nakamura, Y. et al. (2010a) An interspecific linkage map of SSR and intronic polymorphism markers in tomato. *Theor. Appl. Genet.* 121: 731–739.

Shirasawa, K., Bertioli, D.J., Varshney, R.K., Moretzsohn, M.C., Leal-Bertioli, S.C.M., Thudi, M. et al. (2013a) Integrated consensus map of cultivated peanut and wild relatives reveals structures of the A and B genomes of *Arachis* and divergence of the legume genomes. *DNA Res.* 20: 173–180.

Shirasawa, K., Fukuoka, H., Matsunaga, H., Kobayashi, Y., Kobayashi, I., Hirakawa, H. et al. (2013b) Genome-wide association studies using single nucleotide polymorphism markers developed by re-sequencing of the genomes of cultivated tomato. *DNA Res.* (in press).

Shirasawa, K., Hirakawa, H., Tabata, S., Hasegawa, M., Kiyoshima, H., Suzuki, S. et al. (2012a) Characterization of active miniature inverted-repeat transposable elements in the peanut genome. *Theor. Appl. Genet.* 124: 1429–1438.

Shirasawa, K., Ishii, K., Kim, C., Ban, T., Suzuki, M., Ito, T. et al. (2013c) Development of *Capsicum* EST-SSR markers for species identification and *in silico* mapping onto the tomato genome sequence. *Mol. Breed.* 31: 101–110.

Shirasawa, K., Isobe, S., Hirakawa, H., Asamizu, E., Fukuoka, H., Just, D. et al. (2010b) SNP discovery and linkage map construction in cultivated tomato. *DNA Res.* 17: 381–391.

Shirasawa, K., Koilkonda, P., Aoki, K., Hirakawa, H., Tabata, S., Watanabe, M. et al. (2012b) In silico polymorphism analysis for the development of simple sequence repeat and transposon markers and construction of linkage map in cultivated peanut. *BMC Plant Biol.* 12: 80.

Shirasawa, K., Oyama, M., Hirakawa, H., Sato, S., Tabata, S., Fujioka, T. et al. (2011) An EST-SSR linkage map of *Raphanus sativus* and comparative genomics of the Brassicaceae. *DNA Res.* 18: 221–232.

Takeya, M., Yamasaki, F., Uzuhashi, S., Aoki, T., Sawada, H., Nagai, T. et al. (2011) NIASGBdb: NIAS Genebank databases for genetic resources and plant disease information. *Nucleci Acids Res.* 39: D1108–D1113.

Yamazaki, Y., Akashi, R., Banno, Y., Endo, T., Ezura, H., Fukami-Kobayashi, K. et al. (2010) NBRP databases: databases of biological resources in Japan. *Nucleiuc Acids Res.* 38: D26–D32.

Youens-Clark, K., Buckler, E., Casstevens, T., Chen, C., Declerck, G., Derwent, P. et al. (2011) Gramene database in 2010: updates and extensions. *Nucleic Acids Res.* 39: D1085–D1094.