Current Plant Biology 7–8 (2016) 26–30

Contents lists available at ScienceDirect

Current Plant Biology

journal homepage: www.elsevier.com/locate/cpb

Manually Curated Database of Rice Proteins (MCDRP), a database of digitized experimental data on rice☆

Saurabh Raghuvanshi*, Pratibha Gour, Shaji V. Joseph

Department of Plant Molecular Biology, University of Delhi South Campus, Benito Juarez Road, New Delhi, 110021, India

ARTICLE INFO

Article history:
Received 28 October 2016
Received in revised form
24 November 2016
Accepted 25 November 2016

Keywords:
Rice
Manual curation
Data digitization

ABSTRACT

MCDRP or 'Manually Curated Database of Rice Proteins' is a database of digitized experimental datasets on rice proteins. Every aspect of the experimental data published in peer-reviewed research articles on rice biology has been digitized with the help of novel data curation models. These models use a semantic and structured arrangement of alpha-numeric notation, including several well known ontologies, to represent various aspect of the data. As a result data from more than 15,000 different experiments pertaining to about 2400 rice proteins has been digitized from over 540 published and peer-reviewed research articles. The database portal provides access to the digitized experimental data via search or browse functions. In essence, one can instantly access data from even a single data-point from a collection of thousands of the experimental datasets. On the other hand, one can easily access the digitized experimental data from multiple research articles on a rice protein. Based on the analysis and integration of the digitized experimental data, more than 800 different traits (molecular, biochemical or phenotypic) have been precisely mapped onto the rice proteins along with the underlying experimental evidences. Similarly, over 4370 associations, based on experimental evidence, have been established between the rice proteins and various gene ontology terms. The database is being continuously updated and is freely available at www.genomeindia.org.in/biocuration.

1. Introduction

Semantic integration and fast access of the experimental data in biological sciences is essential in order to understand the multi-dimensional nature of most biological processes. Rice is a model monocot system as well as one of the most important food grain. Consequently, extensive amount of research has been conducted over the years with a view to better understand its biology. Several database have also been developed that have greatly facilitated the research on rice biology [1–14]. While the databases ensure that most high throughput datasets like genome/gene sequences, microarray of RNA-seq expression data etc. are catalogued and made available to the users, however, a huge amount of invaluable high quality experimental data remains buried in published literature and cannot be searched or integrated computationally. This is because most published experimental data is presented in pictorial format either as an image or graph and thus not amenable to computerized search, let alone seamless integration. The only way to access such data is via reading the entire publication. The ever increasing number of publications on rice biology has resulted in a massive accumulation of such high quality experimental data. In order to gain a 'systems level' perspective of rice biology, it is imperative that such experimental data is rendered computer indexible so that it can be rapidly searched and integrated. 'Manually Curated Database of Rice Proteins' address this aspect and provides the user with digitized published experimental data on rice biology. It is a manually curated database which utilizes in-house developed data curation models that enable digitization every aspect of the experimental data.

2. Database description

The current release of the database consists of digitized experimental data for over 2400 rice proteins spread over more than 540 published research articles. More than 15,000 individual experiments containing over 90,000 data-points have been digitized with the help of novel data curation models developed earlier [15]. The entire curation or digitization is based on manual curation and thus leads to a very high quality of digitized and validated experimental data. Data from a wide variety of experimental techniques (>150 different types) such as gene expression measurements, enzymatic activities, interaction studies, trait analysis (phenotypic or

☆ This article is part of a special issue entitled "Genomic resources and databases", published in the journal Current Plant Biology 7–8, 2016.

* Corresponding author.
E-mail address: Saurabh@genomeindia.org (S. Raghuvanshi).

http://dx.doi.org/10.1016/j.cpb.2016.11.001

metabolic) etc. have been digitized. The detailed digitization process ensures that information for a single data-point from any of the experiments can be individually retrieved, if required. A data-point, for example, would typically corresponds to the underlying data of single 'bar' of a 'bar graph' depicting a RT-PCR expression data. Information contained within every data-point is represented by a collection of alpha-numeric terms which include various ontology terms such as plant ontology, environmental ontology or trait ontology. Thus, the data of the entire experiment can be represented by a semantic collection of these alpha-numeric terms. The digitization of the data has been done by utilizing >600 plant ontology, >350 environment ontology, >800 trait ontology and >350 gene ontology terms. In other words, every experiment has been extensively annotated with the help of various ontology terms. The plant ontology terms have been used to represent the growth/developmental stage and the tissue of the plant that has been analyzed in the experiment. Similarly, environmental ontology terms have been used to record the growth conditions such as temperature, light or water status as well as treatment with any chemical. The trait ontology terms have been used to represent any trait (molecular, physiological or phenotypic) of the proteins whereas gene ontology terms have been used to encode the functional details of the protein.

Analysis and integration of the digitized data unravels several interesting aspects of the data. As a result of detailed digitization it was possible to map more than 800 different traits to rice proteins. In summary, 831 traits have been mapped to 398 different rice proteins. The database contains a wide range of traits such as 'anatomy and morphology related', 'biochemical profile and physiology related', 'enzymatic activity related', 'growth and development related', 'stress related' and 'yield and biomass related' traits. The top 5 most frequently related traits are 'survival rate', 'plant height', 'root length' 'seedling height' and 'seedling vigour'. The association of the traits to the genes is of very high confidence since it is based on high quality experimental data. Similarly, more than 4300 associations have been made between the rice proteins to various molecular functions as well as biological process and cellular component based on the digitized experimental data. This data is represented with the help of gene ontology terms. Further, in order to facilitate better usage of the database, metabolic pathways have also been mapped on the proteins that have been curated in the database. Consequently, more than 200 metabolic pathways are represented by the rice proteins that have been curated in the database. More than half of these pathways are related to one or the other biosynthetic processes.

3. How to use the database?

Since every aspect of the experimental data has been digitized, the data can be easily and rapidly searched. In other words one can easily search or retrieve any experiment from over 500 published research articles in a matter of seconds and without even opening the research publication. In general the database can be either browsed or searched with a specific query.

3.1. Browsing the database

The contents of the database can be browsed from eight different perspectives. As a result of digitization the experimental data can be retrieved based on the either the PubMed id, gene locus id, growth conditions (Environmental Ontology), plant tissue/developmental stage (Plant Ontology term), phenotypic or biochemical trait (Trait Ontology) and gene function/localization (Gene Ontology term). Ultimately all the aspects are inter-related and one can start from any end and access the information from any of the perspectives. For example, one can browse the database based on any of the known molecular functions of the rice proteins or on the basis of the biological processes wherein they have been implicated. Such data can be accessed via the page summarizing all the mapper GO terms (Fig. 1a). On selecting any one of the molecular functions such as 'DNA binding' or GO:0003677, the database portal would provide a list of all the rice proteins that have been curated in the database, in this case 32 rice gene (Fig. 1b). The data is arranged as per individual experimental dataset. This information has been compiled from 44 digitized experimental data sets from 21 different research articles. This information is primarily acquired by digitizing experimental techniques such as EMSA (Electrophoretic Mobility Shift Assay), CHIP assay or transient co-expression assays. Further, on selecting any of the listed rice protein ids a 'Rice gene details' is shown (Fig. 2). Fig. 2 summarizes information of one of an 'AP2 domain containing protein' that has been assembled via digitization of experimental data from 10 published articles [16–26]. .The information in the '*Rice gene details*' page is divided into several sections. The '*Basic information'* section gives the information of the protein domain present in the rice protein as per the Pfam database. The '*Functional details'* section lists all the molecular function/s or biological process/s that the protein is known to possess or involved in based on published experimental data. 'Clicking' on any of the term id provides the details of the digitized experimental dataset. Similarly, the '*Plant developmental stage/tissue details*' section lists all the developmental stages and plant tissue wherein the selected rice gene had been studied. The '(+)' or '(−)' signs indicate the presence or absence of expression/protein activity in a particular tissue/developmental stage. Thus, '(−)' means that while expression/activity of the gene has been analyzed in that particular tissue but no detectable expression or activity was found. The '*Rice gene details*' page also summarizes information about the environmental conditions under which the expression/activity of the gene has been analyzed. This information is presented as a list of Environmental Ontology terms in the '*Environment details'* section. The impact of abiotic stress conditions such as salinity, drought, heat, or hormones/other chemicals is also recorded with the help of the EO terms. Similarly the *'Trait details'* section summarizes all the phenotypic or biochemical traits that have been associated with the selected gene based on experimental data. The page also summarizes the physical interaction data (protein–protein and protein-DNA) for the rice protein as well as any metabolic pathway or QTL associated with the rice protein.

3.2. Searching the database

Specific searches can also be done on the digitized experimental data in the database. One can retrieve any experimental datasets on the basis of gene id, plant developmental stage/tissue, growth conditions including environmental parameters or any chemical treatment. The experimental data can also be retrieved based on any molecular function, biological process, cellular localization or any trait associated with the protein that has been studied. Further one can also use a combination of terms to formulate a detailed query. For example one can ask the question 'Is there a "protein kinase" gene related to the 'plant height' trait in rice. In order to formulate this query one can use the GO term 'protein kinase activity' and the trait ontology term TO: 'Plant height'. . ..for initiating the search. This will give a list of genes that have protein kinase activity and have been found regulating the trait plant height. The user can then select any one or all the rice gene loci to access the experimental details. The output of the search is a list of rice gene ids as well as the link to the exact experiment (PubMed id and experiment no.) where the search term has been used. Similarly, one can use any other combination of search terms to access the digitized experimental data. These searchs can be done by specify-

(a)

Manually Curated Database of Rice Proteins

Gene Ontology List

Home | Browse | Search | Contact Us | Feedback | Links

Ontology source: Gene Ontology Consortium

biological_process | molecular_function | cellular_component

[Ontology source: Gene Ontology Consortium] - molecular_function

S. No.	ID - Description	S. No.	ID - Description	S. No.	ID - Description
1	GO:0016788 - "hydrolase activity, acting on ester bonds"	2	GO:0010295 - (+)-abscisic acid 8'-hydroxylase activity	3	GO:0016629 - 12-oxophytodienoate reductase activity
4	GO:0043138 - 3'-5' DNA helicase activity	5	GO:0016273 - arginine N-methyltransferase activity	6	GO:0032575 - ATP-dependent 5'-3' RNA helicase activity
7	GO:0016887 - ATPase activity	8	GO:0005262 - calcium channel activity	9	GO:0005509 - calcium ion binding
10	GO:0015085 - calcium ion transmembrane transporter activity	11	GO:0010857 - calcium-dependent protein kinase activity	12	GO:0005516 - calmodulin binding
13	GO:0004683 - calmodulin-dependent protein kinase activity	14	GO:0015144 - carbohydrate transmembrane transporter activity	15	GO:0003824 - catalytic activity
16	GO:0043169 - cation binding	17	GO:0004861 - cyclin-dependent protein kinase inhibitor activity	18	GO:0003677 - DNA binding

(b)

Home | Browse | Search | Contact Us | Feedback | Links

GO Id GO:0003677 [DNA binding] : details

S.No	Accession Id	Description	PubMed Id	Figure number	Figure Method
1	AAS18416	benzothiadiazole-induced homeodomain protein 1	16163610	Fig. 3.A,B	Protein-DNA interaction analysis : Electrophoretic Mobility Shift Assay (EMSA)
2	ABD52007	stress-induced transcription factor NAC1	16924117	Fig. 5.C	Protein-DNA interaction analysis : Yeast one-hybrid analysis
3	ABM90395	bZIP-type transcription factor ABI5 isoform 2	17604002	Fig. 2.B.i,ii,iii	Protein-DNA interaction analysis : Yeast one-hybrid analysis
4	ACC85686	EATB	21753115	Fig. 1.C	Protein-DNA interaction analysis : Yeast one-hybrid analysis
5	ACC85686	EATB	21753115	Fig. 1.D	Protein-DNA interaction analysis : Electrophoretic Mobility Shift Assay (EMSA)
6	AFB77198	dehydration responsive element binding protein 2a (DREB2A) mRNA, complete cds	24376625	Fig. 4	Protein-DNA interaction analysis : Electrophoretic Mobility Shift Assay (EMSA)
7	BAA83740	TRAB1	10611387	Fig. 4.B	Transient co-expression analysis
8	CAC83122	ethylene responsive protein	12369615	Fig. 3	Protein-DNA interaction analysis : Electrophoretic Mobility Shift Assay (EMSA)
9	LOC_Os01g07120	AP2 domain containing protein, expressed	12609047	Fig. 3.B	Protein-DNA interaction analysis : Electrophoretic Mobility Shift Assay (EMSA)
10	LOC_Os01g07120	AP2 domain containing protein, expressed	12609047	Fig. 4.C	Transcriptional activity assay
11	LOC_Os01g09550	no apical meristem protein, putative, expressed	18813954	Fig. 3.B	Protein-DNA interaction analysis : Yeast one-hybrid analysis
12	LOC_Os01g48130	no apical meristem protein, putative, expressed	18813954	Fig. 3.B	Protein-DNA interaction analysis : Yeast one-hybrid analysis
13	LOC_Os01g64310	no apical meristem protein, putative, expressed	18813954	Fig. 3.B	Protein-DNA interaction analysis : Yeast one-hybrid analysis

Please see 'Figure 2' for details

Fig 1. Screenshots showing browse results. (a) When browse by 'Gene Ontology' is selected a list of all the GO terms that have been associated on the basis of digitized experimental data is shown. (b) On selecting any one of the GO terms (in this case GO:0003677) a list of all the genes associated with that term are shown along with links to the relevant publication and the digitized experimental data that was used to associate the GO term to the gene. Details of the gene id encircled with a 'red circle' are shown in Fig. 2. In order to accommodate within limited space the figures show only partial list.

ing either the exact 'term id' (ontology term) or any keyword. The basic search function outputs the results in two tiers. If an exact 'term id' is defined (such as protein id, ontology term id) the relevant results are shown immediately. In case a keyword is provided; then in the first stage a list of all the related terms would be displayed. User can then select one or more of these terms and then proceed to the second stage where data relevant to the selected term would be shown.

4. Discussion

Digitization of experimental data is essential due to a phenomenal increase in the bulk of the data as well as the need for seam-less integration of diverse experimental datasets in order to understand complex biological traits. Thus, efforts are being made globally to address the issue. Several repositories like DRYAD (https://datadryad.org) and FIGSHARE (https://figshare.com/) facilitate submission of diverse experimental data. 'Scientific Data' (http://www.nature.com/sdata/) is a peer-reviewed journal that accepts experimental datasets. The datasets need to be deposited in one of the repositories (http://www.nature.com/sdata/policies/

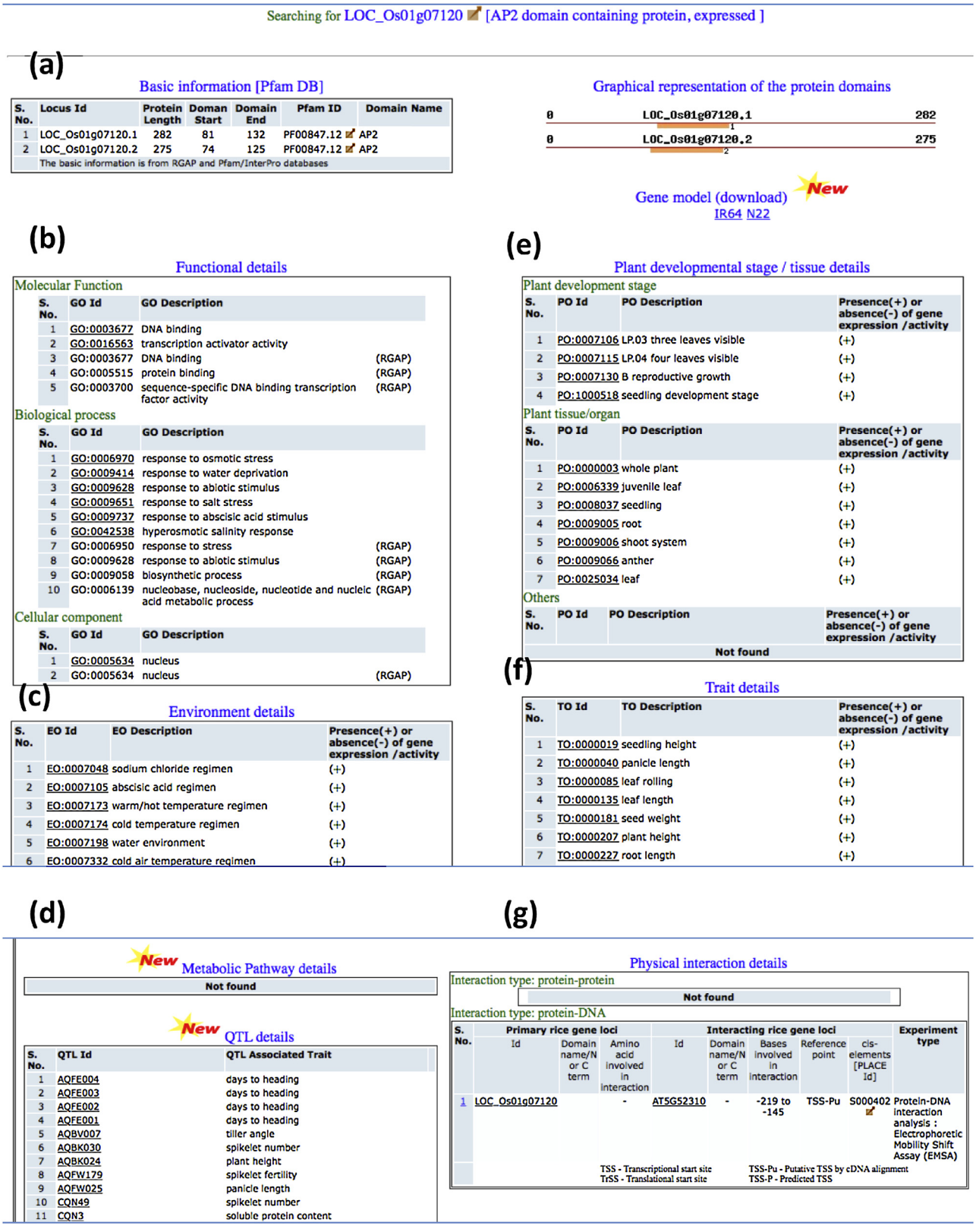

Fig. 2. A snapshot of the 'Rice gene Details' page. The page summarizes information for a particular rice gene from digitized experimental data across all the curated publications. Parts of the output have been truncated to accommodate the whole information. The page has several sections (a-g) which are divided based on the type information. (a) Basic information regarding the domains present in the protein as well as link to the database 'Indica Rice Genome Database' (IRDB) which contains gene models sequences from related indica rice varieties Nagina 22 and IR 64. (b) Mapping of gene ontology terms based on the digitized experimental data. Ontology terms taken from the RGAP database have been indicated with the suffix '(RGAP)' while the others are based on the current curation exercise. (c) Summarizes all environmental conditions or treatments under which the gene has been studied. Plus (+) or (−) signs indicate whether the gene had expression under the condition or not. (d) Lists all the metabolic pathways and QTLs wherein the gene is a constituent., (e) Summarizes all the rice tissues and developmental stages wherein the gene has been studied. (f, g) Details of the molecular/biochemical/phenotypic traits and the physical interaction (DNA-protein or protein-protein) associated with the gene.

repositories) accompanied by a 'Data Descriptor' describing the dataset. However, despite these efforts no data resource provides digitized experimental data on rice proteins. 'Manually Curated Database of Rice Proteins' (MCDRP) was established to address this issue and provide the user with digitized experimental data on rice proteins. The manual data curation/digitization models implemented in MCDRP ensures that all the aspects of every data-point of the experimental data are digitized. This is in contrast to several other formats that add *meta*-data over the entire experimental dataset.

One very important aspect of the curation/digitization process is that it uses very similar elements and fundamentals to digitize data from a wide variety of experimental techniques thus enables efficient integration of the data. The search function is able to retrieve data from even a single data point instantly from a collection of over thousands of experiments. Further, the data can be retrieved from several different perspectives. For instance one can search experimental data based on a particular tissue of growth conditions from across all the curated research articles. In summary, MCDRP, on one end acts an important resource for rice biologist while on the other acts as proof-of-concept regarding the possibility and efficacy of digitizing experimental data.

Funding

The authors acknowledge funding from Department of Biotechnology, Govt. of India and the Delhi University-UGC R&D grant.

References

[1] H. Gu, P. Zhu, Y. Jiao, Y. Meng, M. Chen, PRIN: a predicted rice interactome network, BMC Bioinf. 12 (2011) 161, http://dx.doi.org/10.1186/1471-2105-12-161.
[2] D. Wang, Y. Xia, X. Li, L. Hou, J. Yu, The Rice Genome Knowledgebase (RGKbase): An annotation database for rice comparative genomics and evolutionary biology, Nucleic Acids Res. 41 (2013) 1199–1205, http://dx.doi.org/10.1093/nar/gks1225.
[3] B. Pan, J. Sheng, W. Sun, Y. Zhao, P. Hao, X. Li, OrysPSSP: a comparative Platform for Small Secreted Proteins from rice and other plants, Nucleic Acids Res. 41 (2013) 1192–1198, http://dx.doi.org/10.1093/nar/gks1090.
[4] Y. Sato, N. Namiki, H. Takehisa, K. Kamatsuki, H. Minami, H. Ikawa, et al., RiceFREND: a platform for retrieving coexpressed gene networks in rice, Nucleic Acids Res. 41 (2013) 1214–1221, http://dx.doi.org/10.1093/nar/gks1122.
[5] Y. Sato, H. Takehisa, K. Kamatsuki, H. Minami, N. Namiki, H. Ikawa, et al., RiceXPro Version 3.0: expanding the informatics resource for rice transcriptome, Nucleic Acids Res. 41 (2013) 1206–1213, http://dx.doi.org/10.1093/nar/gks1125.
[6] T. Lu, X. Huang, C. Zhu, T. Huang, Q. Zhao, K. Xie, et al., RICD: a rice indica cDNA database resource for rice functional genomics, BMC Plant Biol. 8 (2008) 118, http://dx.doi.org/10.1186/1471-2229-8-118.
[7] T. Sakurai, Y. Kondou, K. Akiyama, A. Kurotani, M. Higuchi, T. Ichikawa, et al., 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 (2011) 265–273, http://dx.doi.org/10.1093/pcp/pcq190.
[8] N. Kurata, Y. Yamazaki, Oryzabase. An integrated biological and genome information database for rice, Plant Physiol. 140 (2006) 12–17, http://dx.doi.org/10.1104/pp.105.063008.
[9] H. Sakai, S.S. Lee, T. Tanaka, H. Numa, J. Kim, Y. Kawahara, et al., Rice annotation project database (RAP-DB): An integrative and interactive database for rice genomics, Plant Cell Physiol. 54 (2013), http://dx.doi.org/10.1093/pcp/pcs183.
[10] Z. Zhang, J. Sang, L. Ma, G. Wu, H. Wu, D. Huang, et al., RiceWiki: a wiki-based database for community curation of rice genes, Nucleic Acids Res. 42 (2014) 1222–1228, http://dx.doi.org/10.1093/nar/gkt926.
[11] G. Droc, M. Ruiz, P. Larmande, A. Pereira, P. Piffanelli, J.B. Morel, et al., OryGenesDB: a database for rice reverse genetics, Nucleic Acids Res. 34 (2006) D736–D740, http://dx.doi.org/10.1093/nar/gkj012.
[12] P. Larmande, C. Gay, M. Lorieux, C. Périn, M. Bouniol, G. Droc, et al., Oryza Tag Line, a phenotypic mutant database for the Génoplante rice insertion line library, Nucleic Acids Res. 36 (2008) 1022–1027, http://dx.doi.org/10.1093/nar/gkm762.
[13] R. Narsai, J. Devenish, I. Castleden, K. Narsai, L. Xu, H. Shou, et al., Rice DB: an Oryza Information Portal linking annotation, subcellular location, function, expression, regulation, and evolutionary information for rice and Arabidopsis, Plant J. 76 (2013) 1057–1073, http://dx.doi.org/10.1111/tpj.12357.
[14] P. Jaiswal, D. Ware, J. Ni, K. Chang, W. Zhao, S. Schmidt, et al., Gramene: development and integration of trait and gene ontologies for rice, Comp. Funct. Genomics. 3 (2002) 132–136, http://dx.doi.org/10.1002/cfg.156.
[15] P. Gour, P. Garg, R. Jain, S.V. Joseph, A.K. Tyagi, S. Raghuvanshi, Manually curated database of rice proteins, Nucleic Acids Res. 42 (2014).
[16] J.G. Dubouzet, Y. Sakuma, Y. Ito, M. Kasuga, E.G. Dubouzet, S. Miura, et al., OsDREB genes in rice, *Oryza sativa* L., encode transcription activators that function in drought-, high-salt- and cold-responsive gene expression, Plant J. 33 (2003) 751–763, http://www.ncbi.nlm.nih.gov/pubmed/12609047 (Accessed October 27, 2016).
[17] W. Yang, Z. Kong, E. Omo-Ikerodah, W. Xu, Q. Li, Y. Xue, Calcineurin B-like interacting protein kinase OsCIPK23 functions in pollination and drought stress responses in rice (*Oryza sativa* L.), J. Genet. Genom. 35 (531-543) (2008) S1–S2, http://dx.doi.org/10.1016/S1673-8527(08)60073-9.
[18] L. Zhang, L.-H. Tian, J.-F. Zhao, Y. Song, C.-J. Zhang, Y. Guo, Identification of an apoplastic protein involved in the initial phase of salt stress response in rice root by two-dimensional electrophoresis, Plant Physiol. 149 (2009) 916–928, http://dx.doi.org/10.1104/pp.108.131144.
[19] S.-J. Sun, S.-Q. Guo, X. Yang, Y.-M. Bao, H.-J. Tang, H. Sun, et al., Functional analysis of a novel Cys2/His2-type zinc finger protein involved in salt tolerance in rice, J. Exp. Bot. 61 (2010) 2807–2818, http://dx.doi.org/10.1093/jxb/erq120.
[20] S. Matsukura, J. Mizoi, T. Yoshida, D. Todaka, Y. Ito, K. Maruyama, et al., Comprehensive analysis of rice DREB2-type genes that encode transcription factors involved in the expression of abiotic stress-responsive genes, Mol. Genet. Genom. 283 (2010) 185–196, http://dx.doi.org/10.1007/s00438-009-0506-y.
[21] M. Cui, W. Zhang, Q. Zhang, Z. Xu, Z. Zhu, F. Duan, et al., Induced over-expression of the transcription factor OsDREB2A improves drought tolerance in rice, Plant Physiol. Biochem. PPB 49 (2011) 1384–1391, http://dx.doi.org/10.1016/j.plaphy.2011.09.012.
[22] Y. Ning, C. Jantasuriyarat, Q. Zhao, H. Zhang, S. Chen, J. Liu, et al., The SINA E3 ligase OsDIS1 negatively regulates drought response in rice, Plant Physiol. 157 (2011) 242–255, http://dx.doi.org/10.1104/pp.111.180893.
[23] G. Mallikarjuna, K. Mallikarjuna, M.K. Reddy, T. Kaul, Expression of OsDREB2A transcription factor confers enhanced dehydration and salt stress tolerance in rice (*Oryza sativa* L.), Biotechnol. Lett. 33 (2011) 1689–1697, http://dx.doi.org/10.1007/s10529-011-0620-x.
[24] J. You, H. Hu, L. Xiong, An ornithine δ-aminotransferase gene OsOAT confers drought and oxidative stress tolerance in rice, Plant Sci. 197 (2012) 59–69, http://dx.doi.org/10.1016/j.plantsci.2012.09.002.
[25] A. Yang, X. Dai, W.-H. Zhang, A R2R3-type MYB gene, OsMYB2, is involved in salt, cold, and dehydration tolerance in rice, J. Exp. Bot. 63 (2012) 2541–2556, http://dx.doi.org/10.1093/jxb/err431.
[26] Z. Lang, S. Xie, J.-K. Zhu, X.-J. He, M.W. Horton, et al., The 1001 arabidopsis DNA methylomes: an important resource for studying natural genetic, epigenetic, and phenotypic variation, Trends Plant Sci. 21 (2016) 906–908, http://dx.doi.org/10.1016/j.tplants.2016.09.001.