DNA Research **18**, 321–332, (2011)
Advance Access Publication on 4 July 2011
doi:10.1093/dnares/dsr019

In Silico Analysis of Transcription Factor Repertoires and Prediction of Stress-Responsive Transcription Factors from Six Major Gramineae Plants

Keiichi Mochida[1,2,3,*], Takuhiro Yoshida[1], Tetsuya Sakurai[1], Kazuko Yamaguchi-Shinozaki[4], Kazuo Shinozaki[1,2], and Lam-Son Phan Tran[1,*]

RIKEN Plant Science Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan[1]; RIKEN Biomass Engineering Program, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan[2]; Kihara Institute for Biological Research, Yokohama City University, 641-12 Maioka-cho, Totsuka-ku, Yokohama, Kanagawa 230-0045, Japan[3] and Japan International Center of Agricultural Sciences, Ibaraki 305-8686, Japan[4]

*To whom correspondence should be addressed. Tel. +81 45-503-9593 (L.-S.P.T.), +81 45-503-9111 (K.M.). Fax. +81 45-503-9591 (L.-S.P.T.), +81 45-503-9591 (K.M.). Email: tran@psc.riken.jp (L.-S.P.T.); mochida@psc.riken.jp (K.M.)

Edited by Satoshi Tabata
(Received 12 March 2011; accepted 7 June 2011)

Abstract

The interactions between transcription factors (TFs) and *cis*-regulatory DNA sequences control gene expression, constituting the essential functional linkages of gene regulatory networks. The aim of this study is to identify and integrate all putative TFs from six grass species: *Brachypodium distachyon*, maize, rice, sorghum, barley, and wheat with significant information into an integrative database (GramineaeTFDB) for comparative genomics and functional genomics. For each TF, sequence features, promoter regions, domain alignments, GO assignment, FL-cDNA information, if available, and cross-references to various public databases and genetic resources are provided. Additionally, GramineaeTFDB possesses a tool which aids the users to search for putative *cis*-elements located in the promoter regions of TFs and predict the functions of the TFs using *cis*-element-based functional prediction approach. We also supplied hyperlinks to expression profiles of those TF genes of maize, rice, and barley, for which data are available. Furthermore, information about the availability of FOX and *Ds* mutant lines for rice and maize TFs, respectively, are also accessible through hyperlinks. Our study provides an important user-friendly public resource for functional analyses and comparative genomics of grass TFs, and understanding of the architecture of transcriptional regulatory networks and evolution of the TFs in agriculturally important cereal crops.

Key words: abiotic stress; *cis*-motif; database; grasses; transcription factor

1. Introduction

The availability of complete genomic sequences of several important grasses, including *Brachypodium distachyon*, rice (*Oryza sativa*), sorghum (*Sorghum bicolor*), and maize (*Zea mays*), has provided a unique opportunity for comparative genomics studies of grass transcriptional regulatory networks controlled by sequence-specific DNA-binding transcription factors (TFs) which bind to DNA and either activate or repress gene transcription.[1–4] The specific interactions between TFs and their binding sites, i.e. the *cis*-regulatory sequences, play a central role in the regulation of different biological processes such as development, growth, cell division, and responses to environmental stimuli.[5,6] Identification, characterization, and annotation of TF repertoires from different grass species will provide an insight on TF organization and biological

functions of the TFs in grasses as well as their evolution. Additionally, from a biotechnology perspective, TF annotations are especially important for studying transcriptional regulatory switches involved in plant productivity, seed quality, and the sensing/response and adaptation to the environment. A great deal of evidence has demonstrated that identification and molecular tailoring of novel stress-responsive TFs have the potential to stabilize and protect crop performance under adverse conditions.[7,8]

In plants, ~7% of all genes encodes putative TFs.[9] The majority of TFs can be grouped into a number of different families according to the specific type of DNA-binding domain that is present within their sequence.[5,10,11] In the past decade, the completion of various plant genome sequences and the development of high-throughput experimental techniques have enabled scientists to carry out genome-wide analyses of TF repertoires and described the function and organization of TF regulatory systems in a number of plant species.[12–22]

Taking advantage of the available complete sequence of *B. distachyon* and maize, we have identified the full complements of TFs from these species using a prediction method which used 51 Hidden Markov Models (HMMs) from the Pfam database.[23] We also used 11 models, which were originally created by HMMbuild of HMMER2 package, to identify the domains within the putative TF proteins. Given the importance of barley (*Hordeum vulgare*) and wheat (*Triticum aestivum*) as major cereals, their TF repertoires also deserve attention. However, currently their genome sequences have not yet been completed. We, therefore, used available full-length cDNA and coding sequence (CDS) resources (http://trifldb.psc.riken.jp) to identify all potential TFs from these two plants.[24] We integrated all the TF data from these four grasses together with those from rice and sorghum to develop a knowledge integrative database, named GramineaeTFDB. This database provides open access for researchers to all relevant and basic information on functional motifs, promoter regions, available FL-cDNAs, genomic distribution, and multiple sequence alignment of the DNA-binding domains for each TF family of each grass species. In addition, we supplied hyperlinks linking TFs of maize, rice, and barley to their expression profiles documented in Genevestigator. Since most of these TFs have not been experimentally characterized for regulatory function as indicated by assessment in PubMed, we searched for their putative regulatory function by assessing annotations of the gene ontology (GO) using comparative analysis with their *Arabidopsis* counterparts. In addition, we also mapped all putative *cis*-regulatory elements on the promoter regions of all TF encoding genes using a total of 480 *cis*-motifs, which include 11 well-defined abiotic stress-responsive ones. In this analysis, we placed a particular emphasis on stress-responsive *cis*-elements. Knowledge gained from identifying the presence of stress-responsive *cis*-elements, in addition to GO annotation, phylogenetics-based annotation, and expression data, enables effective prediction of stress-responsive TFs. Additionally, the supplied information on *Ds* and FOX and T-DNA insertion lines for a number of TFs from maize and rice, respectively, which can easily identified on GramineaeTFDB, have made convenient access to novel resources for loss- and gain-of-function analyses. Taken together, our results provide comprehensive information on TFs of six major grass species as well as tools for comparative genomic analyses of large TF data sets found in the grasses and non-grass plants.

2. Materials and methods

2.1. Identification of TF repertoires in six grasses

The strategy and bioinformatics pipeline established previously were used to identify the complete sets of TFs from the annotated proteomes of *B. distachyon* (v1.0), maize (v4a.53), rice (v6.0), and sorghum (vSbi1_4), and the partial TF repertoires from barley and wheat using their FL-cDNA and CDS resources.[21,24] Fifty-one HMMs of Pfam and those of 11 originally created using HMMbuild of the HMMER2 package (http://hmmer.janelia.org/) were applied, which corresponded to a total of 61 TF families because there are two HHM profiles which are completely matched.[23] A pre-defined threshold of $E < 1e-5$ was used as the common value cut-off for HMMER search using built HMM profiles. The criteria described previously for the classification of each TF family were applied.[25] Additionally, the TFs identified by initial HMMER search were subjected to a homology search (blastp) with known TFs of *Arabidopsis* classified previously by PlantTFDB (http://planttfdb.cbi.pku.edu.cn/) and PlnTFDB (http://plntfdb.bio.uni-potsdam.de/v3.0/) to confirm the HMMER search results based on the results of homology search (blastp $E\text{-value} \leq 1e-30$).[18] TFs for which the homology search yielded results of $1e-30 <$ blastp $E\text{-value} < 1e-5$ were inspected manually to exclude false-positive hits and determine the true *E*-value for each family (GramineaeTFDB Help page, Statistics). For wheat and barley, we also used the NCBI UniGene sequences of wheat and barley as queries in a blastx homology search against *B. distachyon* protein data set with a threshold $E\text{-value} < 1e-10$ to identify putative TFs.

2.2. Structural and functional annotations of putative grass TFs

Structural and functional annotations of putative grass TFs were done as described previously.[21] All of the similarity searches using blastn were performed with a threshold *E*-value $< 1e-100$, and the top scoring hit for each query was applied. All similarity searches with blastp against protein data sets were performed with a threshold *E*-value $< 1e-5$ to find possible functional descriptions for TF encoding genes. The top scoring hit for each query was applied. To determine the global characteristic features of functional categories of TF encoding genes of the grasses, the TFs were assigned to possible GO terms based on a blastp similarity search (*E*-value $< 1e-10$) using the data set of *Arabidopsis* of TAIR10.[26] The GO annotation and TFs of *Arabidopsis* were retrieved from TAIR and PlnTFDB, respectively. Particular emphasis was placed on sequences serving under the 'biological process' functional category.

2.3. Discovery of cis-regulatory motifs in promoter regions of TF genes

Discovery of *cis*-regulatory motifs located in the -500, -1000, and -3000 bp upstream sequences from the putative transcription start site for each TF encoding gene using 469 *cis*-motif sequences collected from the PLACE database (http://www.dna.affrc.go.jp/PLACE/)[27] and 11 major stress-responsive *cis*-motifs reported previously[28] was performed as described previously.[21] The *cis*-element search results were implemented into the GraminaeTFDB as a searchable property. In addition, these search results were also incorporated as an annotation track of the genome browser (Gbrowse).

2.4. Expression data for TF encoding genes

Hyperlinks linking those putative TF encoding genes of maize, rice, barley, and wheat, whose expression data are available in Genevestigator (https://www.genevestigator.com),[29] were built and supplied on GramineaeTFDB. For putative rice TF encoding genes, hyperlinks linking their expression patterns available at RiceXPro (http://ricexpro.dna.affrc.go.jp/)[30,31] were also built and supplied on our database.

2.5. Genetic resources for TF encoding genes

Hyperlinks linking putative TF encoding genes of maize and rice to genetic resources available at http://www.plantgdb.org,[32,33] http://ricefox.psc.riken.jp,[34] and http://signal.salk.edu/cgi-bin/RiceGE databases were built and supplied on GramineaeTFDB.

2.6. Construction of a web-accessible database

The database is implemented in MySQL and the web interface of Perl CGI and Java script run on the Apache Web server. The definition strings used for sequence similarity searches for each database, the domain searches by InterProScan, *cis*-motif names from the PLACE database, and the assigned GO terms have been assembled as a keyword database enabling the users to specify queries on any keyword and to retrieve relevant information for genes from the GramineaeTFDB. A BLAST server was implemented to provide a similarity search interface for queried sequences using NCBI BLAST together with sequences of the six grasses, as well as those from *Arabidopsis*. Generic Genome Browser (Gbrowse)[35] was also implemented in GramineaeTFDB for sequenced grasses to visualize the gene annotations of the putative TF encoding genes together with *cis*-motifs found on the upstream sequence of the TF genes. All of the data in the GramineaeTFDB are accessible not only through a web interface but also as downloadable files from the website. The cross-references of corresponding data for each of the entries were also implemented into the GramineaeTFDB together with the URLs for each of the original referenced data to provide hyperlinks on the web interface with seamless navigations.

3. Results and discussion

3.1. Identification of putative TFs in B. distachyon, maize, rice, sorghum, barley, and wheat

We have used the strategy and bioinformatics pipeline established previously to identify the complete TF repertoires from *B. distachyon* and maize from their annotated proteomes.[20,21] We started with retrieving the complete sets of predicted proteins from *B. distachyon* (v1.0) and maize (v4a.53), followed by an HMMER search with all HMMs assembled using a predefined threshold of $E < 1e-5$. We then refined the results by combined automatic and manual inspections of the raw alignments to exclude false-positive hits and determine the true *E*-value for each TF family (GramineaeTFDB, Help page, Statistics). Given the importance of wheat and barley as major cereal crops, although their completed genomic sequences are currently not available yet, we attempted to identify partial TF repertoires from these two grass species using their FL-cDNA and CDS resources housed at TriFLDB (http://trifldb.psc.riken.jp).[24] Thus, a total of 2152, 3623, 444, and 916 TF models were identified in *B. distachyon*, maize, barley, and wheat, respectively. These TFs were grouped into 60 families, while those of barley and wheat were classified into 49

Table 1. Predicted TF models in six grasses

	TF gene families	*B. distachyon*[a]	*Z. mays*[a]	*S. bicolor*[a]	*O. sativa*[a]	*H. vulgare*[b]	*T. aestivum*[b]
1	(R1)R2R3_MYB	86	214	116	89	15	64
2	ABI3VP1	51	69	63	37	4	18
3	Alfin-like	16	33	15	10	2	7
4	AP2_EREBP	146	265	167	117	142	131
5	ARF	42	66	29	22	1	8
6	ARID	8	15	8	4	1	2
7	atypical_MYB	40	56	32	28	4	9
8	Aux_IAA	41	83	32	27	5	13
9	BBR-BPC	5	9	5	4	1	1
10	BES1	7	14	8	6	1	3
11	bHLH	158	274	171	107	14	29
12	bZIP	103	185	108	78	13	53
13	C2C2_Zn-CO-like	38	53	37	25	15	18
14	C2C2_Zn-Dof	27	44	29	25	21	9
15	C2C2_Zn-GATA	36	31	33	24	3	12
16	C2C2_Zn-YABBY	15	26	8	7	1	4
17	C2H2_Zn	106	185	113	94	11	26
18	C3H-TypeI	85	160	73	60	10	32
19	CAMTA	10	9	7	5	0	1
20	CCAAT_Dr1	1	5	1	2	0	1
21	CCAAT_HAP2	18	27	11	10	2	14
22	CCAAT_HAP3	18	22	13	10	1	7
23	CCAAT_HAP5	13	21	15	5	3	3
24	CPP	11	17	8	9	0	6
25	E2F_DP	8	19	10	9	0	3
26	EIL	6	9	7	4	2	7
27	GARP_ARRB	11	10	8	6	0	2
28	GARP_G2-like	59	71	48	38	3	12
29	GeBP	17	29	18	12	2	4
30	GRAS	48	84	74	34	5	13
31	GRF	28	17	38	14	1	1
32	HB	112	194	88	73	15	40
33	HMG-box	16	27	13	11	5	17
34	HRT	1	3	1	1	2	0
35	HSF	30	51	25	26	1	7
36	JUMONJI	24	33	22	14	3	4
37	LFY	1	4	1	1	0	4
38	LIM	20	30	9	11	2	8
39	LUG	5	3	5	6	1	3
40	MADS	83	96	83	46	58	124
41	MBF1	3	7	2	3	3	3
42	MYB_related	47	70	64	36	8	22
43	NAC	84	168	124	96	18	22
44	Nin-like	17	24	13	7	1	5
45	PcG	58	77	46	30	5	15
46	PHD	185	270	169	114	14	48

Continued

Table 1. Continued

	TF gene families	*B. distachyon*[a]	*Z. mays*[a]	*S. bicolor*[a]	*O. sativa*[a]	*H. vulgare*[b]	*T. aestivum*[b]
47	PLATZ	14	17	18	11	2	2
48	S1Fa-like	3	2	2	2	2	2
49	SAP	0	0	0	0	0	0
50	SBP	18	50	19	17	1	5
51	SRS	4	11	5	5	0	2
52	TCP	21	49	27	18	1	5
53	Trihelix	8	21	10	7	1	3
54	TUB	15	32	20	12	2	9
55	ULT	1	5	1	1	0	0
56	VOZ	2	8	2	2	0	2
57	Whirly	2	6	2	2	1	2
58	WRKY_Zn	80	151	96	79	1	36
59	zf-HD	16	26	15	15	0	2
60	zf-TAZ	5	10	5	5	0	1
61	ZIM	30	45	20	18	4	13
	Total	2152	3623	2205	1597	444	916

[a]Complete TF repertoires predicted using proteomes annotated from genomic sequences.
[b]Partial TF repertoires predicted using FL-cDNA resources available on TriFLDB.

and 58 families, respectively, based on the presence of domains that were specific for the family (Table 1).

Currently, the GRASSIUS is the only grass-specific database which provides accession to TFs from several grass species, including maize, rice, sorghum, and sugarcane, as a tool for comparative genomics of grass TFs.[19] However, although GRASSIUS contains maize and rice TFs, it used the old annotated version of maize and rice genomes (v3b.50 for maize and TIGR5 for rice) for TF identification. In our study, we used the newest release version of maize and rice annotated protein sequences, the v4a.53 and TIGR6 for maize and rice, respectively, for TF prediction. Furthermore, we also included the TF repertoires of sorghum (Sbi1.4), which were identified using the same approach (Table 1), with the aim to construct a comprehensive grass TF database of six major grass species for comparative genomics of the grass TFs.

Our distribution analysis has indicated that the TF families of the sequenced species, including *B. distachyon*, maize, rice, and sorghum, are scattered throughout the genome. The larger families, such as bHLH and PHD, have members that are distributed on almost every chromosome. Currently, the genomic sequencing and annotations of barley and wheat have not been finished yet; we will update their TF repertoires when completed genome sequences are available. Additionally, the number of predicted TFs of *B. distachyon*, maize, sorghum, and perhaps rice may be changed by future fine-tuning of gene annotations and/or HMM profiles. We will continue to update our website with new information to enhance the accuracy of TF prediction and annotation.

A number of studies have substantiated that sequence homology-based clustering of the members of several gene families correlates with their function.[21,36–38] The complete sequence of the wild grass *B. distachyon,* the first member of the Pooideae subfamily, can serve as a template for analysis of the large genomes of economically important pooideae grasses, including wheat and barley. We, therefore, subjected all the putative UniGene sequences of wheat and barley to a blastx homology search with their *B. distachyon* counterparts ($E < 1e-10$) as a means to identify putative TFs by homology search-based approach. A significant proportion of wheat and barley TFs showed high homology to *B. distachyon* TFs (Supplementary Table S1A and S1B, Fig. 1). Additionally, data shown in Fig. 1 suggest that the HMM search of FL-cDNA/CDS and this homology search-based approach may complement and support each other. Furthermore, recognition of *B. distachyon* as an important model system has led to the development of highly efficient transformation, genetic markers, microarrays, and databases (http://www.brachybase.org, http://www.phytozome.net, http://www.modelcrop.org, http://mips.helmholtz-muenchen.de/plant/index.jsp) and various valuable genetic resources, such as mutant and germplasm

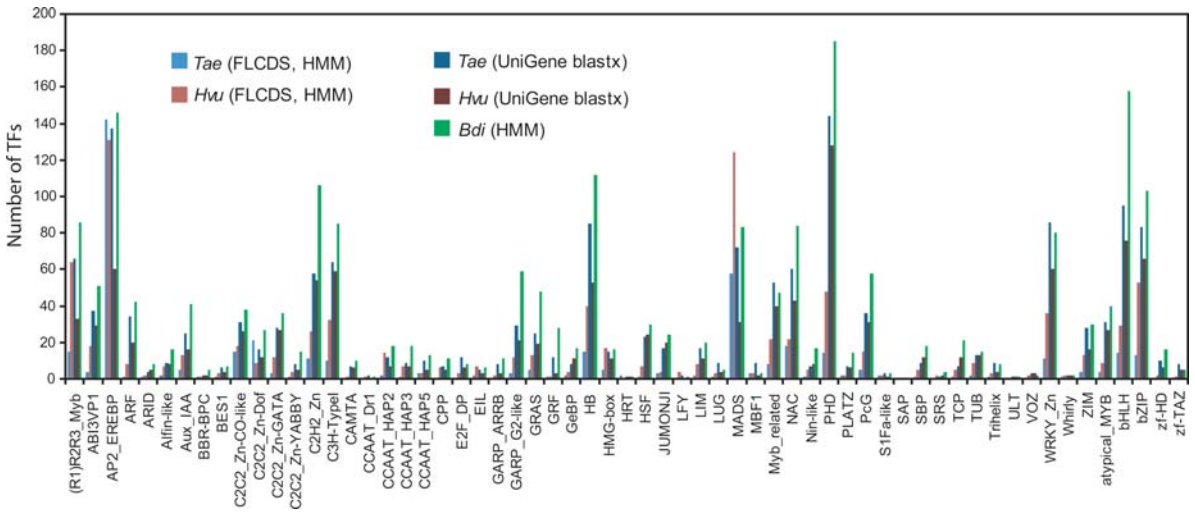


Figure 1. Distribution and number of TFs of *T. aestivum* and *H. vulgare*, which were found by HMM search or homology search with TFs of *B. distachyon*. The HMM search was performed against full-length cDNA/CDS of both species. The homology search using blastx was applied between NCBI UniGene data set of both species and *B. distachion*, the predicted protein data set in Bdi1.0 with 1e−10 to find significant homologues.

collections, have facilitated the use of *B. distachyon* by the research community.[4,39–41] All these available tools can effectively aid homology-based functional annotations of the TFs of wheat, barley, and other pooideae grasses.

3.2. *GO-based functional annotation of identified TFs of B. distachyon, maize, sorghum, and rice*

A search for potential functions of the identified TFs of *B. distachyon*, maize, sorghum, and rice by literature analysis of published papers on PubMed database has revealed that although the sequences of *B. distachyon*, maize, sorghum, and rice have been completed, the majority of their TFs remain experimentally uncharacterized. Thus, as a means to extend our current knowledge base regarding their regulatory function, especially in abiotic stress responses, we assessed the putative functions of the TFs of these four species via comparative analyses with relevant GO annotations of *Arabidopsis* in TAIR. First, sequence similarity searches against *Arabidopsis* counterparts having GO terms in TAIR were carried out to assign the profile of GO terms to the grass TFs at the biological process level. All of the assigned terms were then counted to grasp the overall representation of GO terms in applied entries of grass TFs, and the top 20 most abundant terms, excluding broad terms of 'regulation of transcription', 'DNA-dependent regulation of transcription', 'positive regulation of transcription', 'negative regulation of transcription', and 'biological process', were subsequently used to classify the TFs (Fig. 2). A number of the analysed TFs are found to be related to stress and hormone responses, indicating important role of these TFs in controlling these biological processes. The assigned GO terms for each TF can be accessed through the detailed page of each TF of each grass species on our database (Fig. 3I). These annotations provide an insight into potential functions of identified TFs of *B. distachyon*, maize, sorghum, and rice which would aid researchers in selection of TFs of interest for further studies. At the same time, a large number of analysed TFs could not be classified into any GO category, indicating the limited amount of functional information that we know regarding the biological processes that most of the TFs mediate, even for model plants such as *Arabidopsis*.

3.3. *Discovery of cis-elements in the promoter regions of identified TFs and cis-element-based functional prediction of the TFs*

Numerous *cis*-elements have been reported for their essential roles in determining the tissue-specific or stress-induced expression patterns of genes.[28,42] Strong lines of evidence have indicated that the *cis*-motifs are highly conserved among orthologous or paralogous genes and co-regulated genes, and defined *cis*-elements can effectively aid in the genome-wide screening of ABA and abiotic stress-responsive genes, which is our major interest.[42–45] To facilitate the functional characterization and prediction of the TFs, especially the stress-related TFs, we retrieved the −500, −1000, and −3000 promoter regions of all the TF genes from *B. distachyon*, maize, rice, and shorgum, whose complete genomic sequences are available. We provided this promoter sequence data set on our website in addition to other relevant information on the TFs for convenient

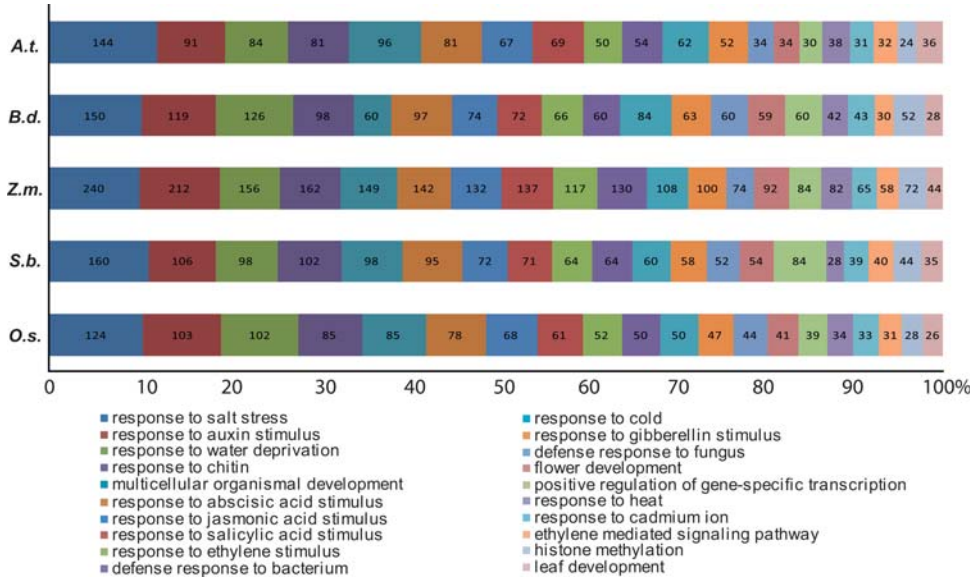


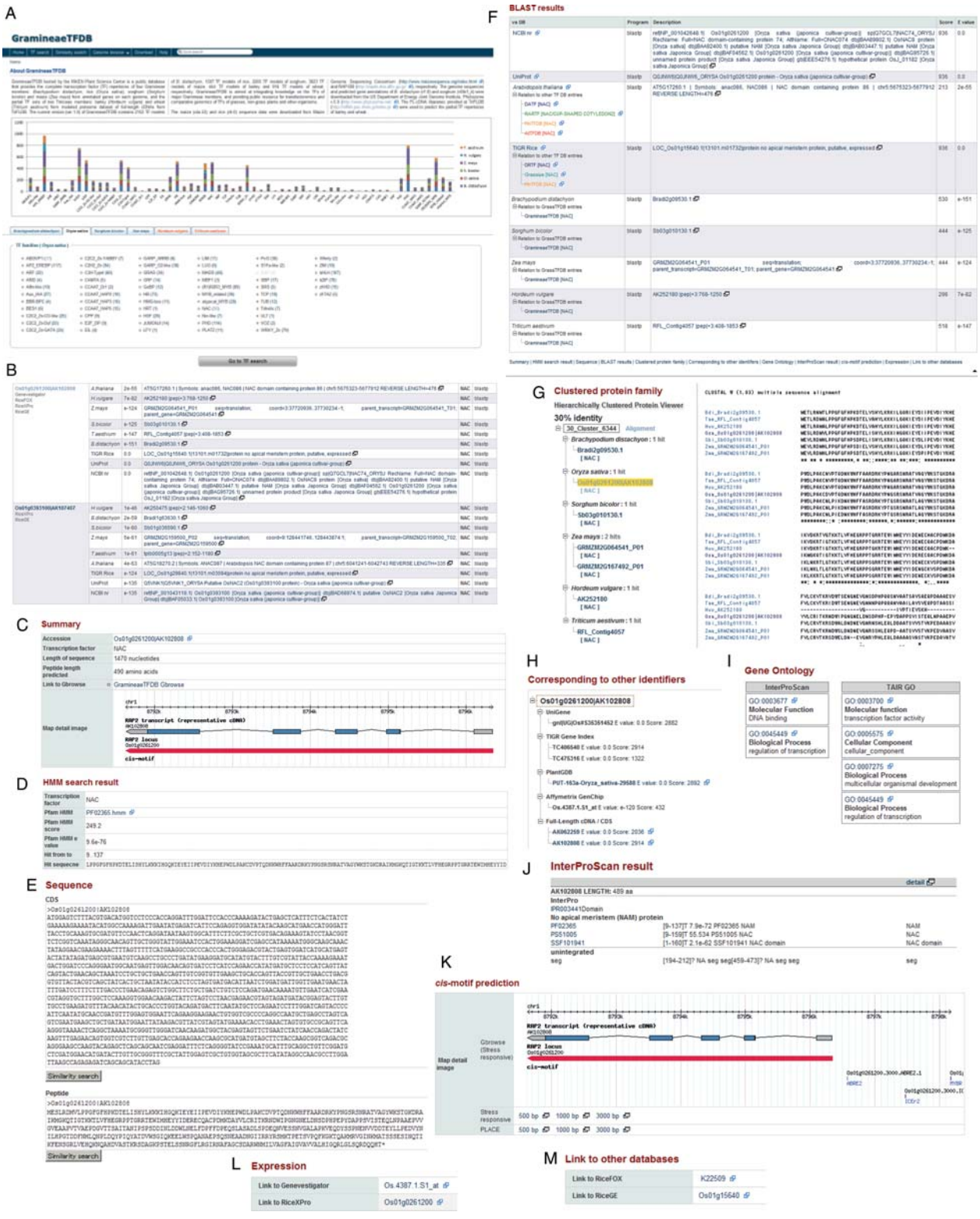
Figure 2. The representative distributions of the GO terms for biological processes associated with TFs from *B. distachyon* (*B.d.*), *Z. mays* (*Z.m.*), *S. bicolor* (*S.b.*), and *O. sativa* (*O.s.*) in comparison with *A. thaliana* (*A.t.*). The top 20 abundantly found GO terms were assigned based on homology searches against annotated *Arabidopsis* genes (blastp homology search with *E*-value $< 1e-10$). TF numbers are shown for each GO term.

downloading. The −500, −1000, and −3000 bp promoter regions were subjected to an extensive *in silico* analyses to search for the existence of a total of 480 putative known *cis*-regulatory motifs, including 11 major abiotic stress-responsive *cis*-motifs.[27,28] Information on the *cis*-elements located in the promoter regions of each TF is accessible on the detailed page of each TF gene under '*cis*-motif prediction' function (Fig. 3K). By clicking on either '500 bp', '1000 bp', or '3000 bp' function, the users will find additional page displaying the 500, 1000, or 3000 bp promoter region and the genomic sequence of the TF encoding gene, respectively, together with the *cis*-motifs located in the corresponding promoter region. The '+' was added to indicate the putative transcription start. In addition, by clicking on 'Go to TF search' (Fig. 3A), the users will be navigated to the search page that provides the '*cis*-motif (stress-responsive)' and '*cis*-motif (PLACE)' search functions, which enables the search for all types of *cis*-motifs implemented in our database in promoter region of any TF and/or the search for those TFs which contains the *cis*-motif(s) of interest. In combination with comparative sequence analysis-based GO annotations, *cis*-motif analysis can facilitate the systematic functional predictions of grass TFs. For instance, first we search for grass TF genes which harbour stress-responsive *cis*-motif(s) in their promoter regions using our grass-specific database. Next, we screen the identified TFs using GO annotation provided for each TF on detailed annotation page (Fig. 3I). Thus, we will be able to identify the putative stress-responsive TFs based on both the existence of stress-responsive *cis*-motif(s) and the associated stress-responsive GO terms. The predicted stress-responsive function should be verified using an expression profiling approach prior to the launching of laborious *in planta* functional studies.

3.4. *Expression patterns of TF encoding genes from maize, rice, and barley*

The specifically expressed TFs are interesting as they are involved in defining the precise nature of individual tissues. Additionally, both *in silico* and genetic inspection suggested a positive correlation between the existence of *cis*-regulatory motifs and tissue-specific and/or stress-responsive expression patterns.[46] To make our database a comprehensive integrated database for functional characterization and selection of stress-responsive TFs, we provided access to tissue-specific expression profiles documented in Genevestigator and RiceXPro through hyperlinks for those TF encoding genes of barley, maize, and rice, for which data are available. These TF genes are indicated by either [Genevestigator] and/or [RiceXPro] strings on the detailed page of our database (Fig. 3B). It is important to note that TF activity often depends on post-translational events and that levels of gene expression are not necessarily directly correlated to their regulatory activity. However, it is still useful to assess the extent of TF expression as it provides the first line of temporal and spatial evidence for linking them to putative *in planta* functions. The tissue-specific expression data can be used to

address the combinatorial usage of TFs, which allows great precision and flexibility in dictating the transcriptional programme of different tissues. Ubiquitous TFs might control the general gene expression either in isolation or in combination with each other. Combinations of specific TFs might regulate tissue-specific genes. Alternatively, and perhaps most commonly, ubiquitous TFs might serve as a



platform to regulate a broad set of genes, which are subsequently fine-tuned by specific regulators. Additionally, co-operativity among TFs has been shown to involve extensive protein–protein interactions, both within families of homomeric and heteromeric TFs and between structurally unrelated TFs.[6,47,48] Analysis of such interactions will help elucidate patterns of combinatorial regulation and ultimately the regulatory functions of the TFs.[49]

One of our main interests in the functional analysis of grass TF encoding genes is to identify abiotic stress-responsive TFs. At the present time, the Genevestigator resource contains stress-related expression data derived from high-throughput microarray experiments for the TF encoding genes of rice and barley. These expression patterns related to drought, cold, and salt stresses can also be accessed through the same hyperlinks provided on GramineaeTFDB for tissue-specific expression. The expression data together with information of *cis*-motif analyses, GO annotations, and sequence similarities inferred from comparative sequence analyses can facilitate the systematic functional predictions of identified TFs as well as provide valuable insights into further functional analyses of TFs. We will continue to update our database when expression information for other TF encoding genes becomes available.

3.5. *Mutant resources for functional studies of maize and rice TFs*

An advantage in functional analyses of maize and rice TFs is the availability of the *Ds*, FOX, and T-DNA insertion mutant resources for a number of maize and rice TFs.[32,50,51] A two-element Activator/Dissociation (*Ac*/*Ds*) gene trap system was successfully established and used for insertional mutagenesis in maize and numerous heterologous species to generate collections of stable, unlinked, and single-copy *Ds* mutants.[32,52,53] *Ds* mutant lines are generally gene knockout or knockdown mutants, but *Ds* activation tagging lines can also be identified among the mutants.[54,55] On the other hand, FOX lines are basically gain-of-function mutants which were constructed by constitutively overexpressing rice FL-cDNAs under the control of 35S promoter in *Arabidopsis*.[50] As a means to make the search for FOX and *Ds* lines convenient, we provided [RiceFOX] and [Closet DS] strings on the list of the search page of each TF family of rice and maize, respectively, for those TFs for which mutants are available (Fig. 3B). Users can gain full access to the respective mutant lines through the supplied hyperlinks on the detailed page (Fig. 3M) or Supplementary Tables S2 and S3. Additionally, for loss-of-function analysis of rice, the RiceGE database (http://signal.salk.edu/cgi-bin/RiceGE) is very useful and has broad functions. For instance, RiceGE provides information about available T-DNA insertion lines generated by an enhancer trap system. We, therefore, supplied [RiceGE] strings and hyperlinks linking directly the rice TFs to RiceGE on the detailed page (Fig. 3B and M). Supplementary Table S4 summarizes the GramineaeTFDB-RiceGE hyperlinks available for the rice TFs. Our database will be occasionally updated when more information are available in public resources or new mutant resources of other grass species are constructed and made available to public.

Figure 3. The web-based user interface of GramineaeTFDB and a demonstration of a typical example of related annotations for a putative TF encoding gene. The homepage of GramineaeTFDB displays TF families and number of TFs of each TF family identified in six grass species: *B. distachyon*, *O. sativa*, *S. bicolor*, *Z. mays*, *H. vulgare*, and *T. aestivum*. By clicking on 'Go to TF search', the users will be directed to the search page which provides search queries for the names of TF families, keywords, sequence identifiers, identifiers of domains supported by InterProScan, GO terms, and available *cis*-motifs for each grass species (A). The search results are listed for a TF family of a grass species with a description of corresponding genes based on similarity searches. For those TF encoding genes of barley, maize, and rice, whose expression data are available through hyperlinks, [Genevestigator] and/or [RiceXPro] strings are displayed. [RiceFOX], [RiceGE], or [Closet DS] string is also displayed for to indicate the availability of hyperlinks linking the rice TFs to RiceFOX and RiceGE databases and maize TFs to PlantGDB database (*Ac*/*Ds* lines) in the detailed page (B). Users are able to navigate to the detailed annotation pages to browse the related annotations. The detailed annotation pages provide summarized basic information on each of the gene models annotated with gene structure. The figure for a gene structure is accessible via a hyperlink to a genome browser which is browsed together with other sequences allocated onto the grass genome (C). The HMM search result for the TF is displayed (D). The sequences of cDNA and protein are provided and all clickable buttons navigate users to the blast search interface directory (E). The similarity search results for each of the entries against NCBI nr, UniProt, and gene models of *Arabidopsis* and other grass species with detailed search results and hyperlinks to the original data (F). Resultant hierarchical clustering of homologous TFs can be browsed with multiple alignment of each cluster (G). Information of other sequence identifiers for representative transcript sequence databases, including UniGene, TIGR Gene Index, and PlantGDB as well as the probe ID of target sequences on the Affymetrix GeneChip, if available, are also accessible. Furthermore, information about available FL-cDNAs is provided through hyperlinks (H). The GO terms assigned to each of the entries based on InterProScan and sequence similarity search against the annotated genes of *Arabidopsis* of TAIR10 (I). The domain structure predicted by InterProScan is provided (J). The result of a *cis*-motif sequence pattern search of promoter regions for each gene is shown together with genomic gene structure (K). Hyperlinks to Genevestigator and/or RiceXPro are provided for those TFs for which expression data are available (L). Hyperlinks to RiceFox and/or RiceGE for rice TFs or PlantGDB (*Ac*/*Ds* lines) for maize TF (M).

Table 2. The availability of resources for functional analyses of the TFs from six grass species

	FLcDNA/CDS	Microarray probe	Clustered EST	Expression	Genetic resource
Rice	KOME	Affymetrix		Genevestigator RiceXPro	RiceFOX RiceGE
Maize	Maize full-length cDNA project	Affymetrix		Genevestigator	PlantGDB
Sorghum	NA	NA	NCBI UniGene PlantGDB	NA	NA
Brachypodium	NA	NA	TIGR Gene Index	NA	NA
Wheat	TriFLDB	Affymetrix		Genevestigator	NA
Barley	TriFLDB	Affymetrix		Genevestigator	NA

NA, not available.

3.6. *Construction and description of a web-accessible database: GramineaeTFDB*

Extensive annotations were performed at both gene and family levels to provide comprehensive knowledge on the identified TFs of *B. distachyon*, maize, rice, sorghum, barley, and wheat (for details, see the GramineaeTFDB Help page). All the annotation data were integrated to develop GramineaeTFDB (http://gramineaetfdb.psc.riken.jp) aimed at integrating TF repertoires of major grasses for functional analyses and comparative genomics of the grass TFs. Figure 3 illustrates the web-based user interface of GramineaeTFDB. More detailed descriptions are provided on the Help page of GramineaeTFDB. Users can conveniently access to the detailed information on gene annotations, including gene structure, cDNA and protein sequences, domain structure predicted by InterProScan, promoter regions, domain alignments, clusters of homologous proteins within families, and GO terms derived from GO annotation using comparative analysis with their *Arabidopsis* counterparts. The data supplied are available not only for viewing but also for immediate downloading. The scientific community can browse predictions for a total of 2152, 3623, 444, and 916 TF models of *B. distachyon*, maize, barley, and wheat, respectively, as well as 1597 and 2205 TF models of rice and sorghum. Users can access to the search results listed for each TF family with description of each gene based on similarity search with TFs of other grasses and *Arabidopsis* as well as with sequences found in NCBI nr and UniProt databases. In detailed page for each TF gene, multiple alignments of amino acid sequences within TF families are also available for downloading and can be used for the construction of phylogenetic trees. Clustered results showing amino acid similarity with different levels of amino acid identity (30, 60, and 90%) and search functions for functional motif information of InterProScan, *cis*-motifs in promoter regions of TFs, and GO annotations are also provided. Additionally, GramineaeTFDB supplies an interface to perform sequence similarity searches using the NCBI BLAST program, as well as cross-reference links to different plant TF databases, including the general PlantTFDB and PlnTFDB, the grass-specific GRASSIUS, and the species-specific DATF, DRTF, and RARTF,[17–19,56–58] making it a comprehensive integrated database for comparative studies of the TFs derived from different plant species.

Integration of expression analysis, *cis*-motif, and GO annotations as well as comparative sequence analysis provided through this study may effectively aid in functional prediction of the TFs. It is noteworthy that for rice TF researchers, information about the availability of FOX and T-DNA insertion lines for rice TFs supplied through hyperlinks are very useful (Fig. 3B and M). Together with FOX, T-DNA insertion, and *Ds* lines, all the genetic and DNA resources, which are currently available for functional analyses of the grass TFs, can be accessed from our database. Table 2 summarizes all these useful resources available for each of six grass species. Providing such an information to the users has made our database unique in comparison with either GRASSIUS or PlantTFDB or PlnTFDB. GramineaeTFDB will therefore meet the broad demands of researchers who strive to perform research on TFs of grasses with the goal of gaining greater understanding of their regulatory roles in different signalling pathways underlying plant development, differentiation, and environmental responses. Our database may accelerate functional genomics and comparative genomics of TFs within individual grass, among grasses themselves, between grasses and non-grass plants, as well as other organisms. We will expand GraminaeTFDB by adding TF repertoires from other grasses upon their genomic sequencing and annotations are completed.

Acknowledgements The sequence data of *B. distachyon* and sorghum were produced by the US Department of Energy Joint Genome Institute in collaboration with the scientific user community. The authors thank MGSC and RAP-DB for maize and rice sequence data.

Supplementary data: Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was supported by Grant-in-Aid for Young Scientists (B) (21780011) to K.M. from the Ministry of Education, Culture, Sports, Science and Technology of Japan. Research in L.-S.P.T.'s lab is supported by Grants-in-Aid (Start-up) for Scientific Research (No. 21870046) from Ministry of Education, Culture, Sports, Science and Technology of Japan, and by Start-up Support grant (No. M36-57000) from RIKEN Yokohama Institute Director Discretionary Funds.

References

1. Paterson, A.H., Bowers, J.E., Bruggmann, R., et al. 2009, The Sorghum bicolor genome and the diversification of grasses, *Nature*, **457**, 551–6.
2. Schnable, P.S., Ware, D., Fulton, R.S., et al. 2009, The B73 maize genome: complexity, diversity, and dynamics, *Science*, **326**, 1112–5.
3. Tanaka, T., Antonio, B.A., Kikuchi, S., et al. 2008, The Rice Annotation Project Database (RAP-DB): 2008 update, *Nucleic Acids Res.*, **36**, D1028–33.
4. Vogel, J.P., Garvin, D.F., Mockler, T.C., et al. 2010, Genome sequencing and analysis of the model grass *Brachypodium distachyon*, *Nature*, **463**, 763–8.
5. Riechmann, J.L., Heard, J., Martin, G., et al. 2000, Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes, *Science*, **290**, 2105–10.
6. Tran, L.S., Nakashima, K., Shinozaki, K. and Yamaguchi-Shinozaki, K. 2007, Plant gene networks in osmotic stress response: from genes to regulatory networks, *Methods Enzymol.*, **428**, 109–28.
7. Tran, L.S. and Mochida, K. 2010, Functional genomics of soybean for improvement of productivity in adverse conditions, *Funct. Integr. Genomics*, **10**, 447–62.
8. Hadiarto, T. and Tran, L.S. 2011, Progress studies of drought-responsive genes in rice, *Plant Cell Rep.*, **30**, 297–310.
9. Udvardi, M.K., Kakar, K., Wandrey, M., et al. 2007, Legume transcription factors: global regulators of plant development and response to the environment, *Plant Physiol.*, **144**, 538–49.
10. Wilson, D., Charoensawan, V., Kummerfeld, S.K. and Teichmann, S.A. 2008, DBD—taxonomically broad transcription factor predictions: new content and functionality, *Nucleic Acids Res.*, **36**, D88–92.
11. Tran, L.S. and Mochida, K. 2010, A platform for functional prediction and comparative analyses of transcription factors of legumes and beyond, *Plant Signal Behav.*, **5**, 550–2.
12. Rushton, P.J., Bokowiec, M.T., Laudeman, T.W., Brannock, J.F., Chen, X. and Timko, M.P. 2008, TOBFAC: the database of tobacco transcription factors, *BMC Bioinform.*, **9**, 53.
13. Fredslund, J. 2008, DATFAP: a database of primers and homology alignments for transcription factors from 13 plant species, *BMC Genomics*, **9**, 140.
14. Richardt, S., Lang, D., Reski, R., Frank, W. and Rensing, S.A. 2007, PlanTAPDB, a phylogeny-based resource of plant transcription-associated proteins, *Plant Physiol.*, **143**, 1452–66.
15. Palaniswamy, S.K., James, S., Sun, H., Lamb, R.S., Davuluri, R.V. and Grotewold, E. 2006, AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks, *Plant Physiol.*, **140**, 818–29.
16. Mitsuda, N. and Ohme-Takagi, M. 2009, Functional analysis of transcription factors in Arabidopsis, *Plant Cell Physiol.*, **50**, 1232–48.
17. Zhang, H., Jin, J., Tang, L., et al. 2011, PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database, *Nucleic Acids Res.*, **39**, D1114–7.
18. Perez-Rodriguez, P., Riano-Pachon, D.M., Correa, L.G., Rensing, S.A., Kersten, B. and Mueller-Roeber, B. 2010, PlnTFDB: updated content and new features of the plant transcription factor database, *Nucleic Acids Res.*, **38**, D822–7.
19. Yilmaz, A., Nishiyama, M.Y. Jr., Fuentes, B.G., et al. 2009, GRASSIUS: a platform for comparative regulatory genomics across the grasses, *Plant Physiol.*, **149**, 171–80.
20. Mochida, K., Yoshida, T., Sakurai, T., Yamaguchi-Shinozaki, K., Shinozaki, K. and Tran, L.S. 2010, LegumeTFDB: an integrative database of *Glycine max*, *Lotus japonicus* and *Medicago truncatula* transcription factors, *Bioinformatics*, **26**, 290–1.
21. Mochida, K., Yoshida, T., Sakurai, T., Yamaguchi-Shinozaki, K., Shinozaki, K. and Tran, L.S. 2009, In silico analysis of transcription factor repertoire and prediction of stress responsive transcription factors in soybean, *DNA Res.*, **16**, 353–69.
22. Wang, Z., Libault, M., Joshi, T., et al. 2010, SoyDB: a knowledge database of soybean transcription factors, *BMC Plant Biol.*, **10**, 14.
23. Sammut, S.J., Finn, R.D. and Bateman, A. 2008, Pfam 10 years on: 10,000 families and still growing, *Brief Bioinform.*, **9**, 210–9.
24. Mochida, K., Yoshida, T., Sakurai, T., Ogihara, Y. and Shinozaki, K. 2009, TriFLDB: a database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics, *Plant Physiol.*, **150**, 1135–46.
25. Zhu, Q.H., Guo, A.Y., Gao, G., et al. 2007, DPTF: a database of poplar transcription factors, *Bioinformatics*, **23**, 1307–8.
26. Swarbreck, D., Wilks, C., Lamesch, P., et al. 2008, The Arabidopsis Information Resource (TAIR): gene structure and function annotation, *Nucleic Acids Res.*, **36**, D1009–14.
27. Higo, K., Ugawa, Y., Iwamoto, M. and Korenaga, T. 1999, Plant cis-acting regulatory DNA elements (PLACE) database: 1999, *Nucleic Acids Res.*, **27**, 297–300.
28. Yamaguchi-Shinozaki, K. and Shinozaki, K. 2005, Organization of cis-acting regulatory elements in osmotic- and cold-stress-responsive promoters, *Trends Plant Sci.*, **10**, 88–94.

29. Hruz, T., Laule, O., Szabo, G., et al. 2008, Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes, *Adv. Bioinform.*, **2008**, 420747.
30. Sato, Y., Antonio, B., Namiki, N., et al. 2011, Field transcriptome revealed critical developmental and physiological transitions involved in the expression of growth potential in japonica rice, *BMC Plant Biol.*, **11**, 10.
31. Sato, Y., Antonio, B.A., Namiki, N., et al. 2011, RiceXPro: a platform for monitoring gene expression in japonica rice grown under natural field conditions, *Nucleic Acids Res.*, **39**, D1141–8.
32. Duvick, J., Fu, A., Muppirala, U., et al. 2008, PlantGDB: a resource for comparative plant genomics, *Nucleic Acids Res.*, **36**, D959–65.
33. Vollbrecht, E., Duvick, J., Schares, J.P., et al. 2010, Genome-wide distribution of transposed dissociation elements in maize, *Plant Cell*, **22**, 1667–85.
34. Sakurai, T., Kondou, Y., Akiyama, K., 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**, 265–73.
35. Donlin, M.J. 2009, Using the Generic Genome Browser (GBrowse), *Curr. Protoc. Bioinform.*, Chapter 9, Unit 9 9.
36. Fang, Y., You, J., Xie, K., Xie, W. and Xiong, L. 2008, Systematic sequence analysis and identification of tissue-specific or stress-responsive genes of NAC transcription factor family in rice, *Mol. Genet. Genomics*, **280**, 547–63.
37. Tran, L.S., Quach, T.N., Guttikonda, S.K., et al. 2009, Molecular characterization of stress-inducible GmNAC genes in soybean, *Mol. Genet. Genomics*, **281**, 647–64.
38. Mochida, K., Yoshida, T., Sakurai, T., Yamaguchi-Shinozaki, K., Shinozaki, K. and Tran, L.S. 2010b, Genome-wide analysis of two-component systems and prediction of stress-responsive two-component system members in soybean, *DNA Res.*, **17**, 303–24.
39. Vogel, J. and Hill, T. 2008, High-efficiency *Agrobacterium*-mediated transformation of *Brachypodium distachyon* inbred line Bd21-3, *Plant Cell Rep.*, **27**, 471–8.
40. Vogel, J.P., Tuna, M., Budak, H., Huo, N., Gu, Y.Q. and Steinwand, M.A. 2009, Development of SSR markers and analysis of diversity in Turkish populations of *Brachypodium distachyon*, *BMC Plant Biol.*, **9**, 88.
41. Filiz, E., Ozdemir, B.S., Budak, F., Vogel, J.P., Tuna, M. and Budak, H. 2009, Molecular, morphological, and cytological analysis of diverse *Brachypodium distachyon* inbred lines, *Genome*, **52**, 876–90.
42. Kim, D.W., Lee, S.H., Choi, S.B., et al. 2006, Functional conservation of a root hair cell-specific cis-element in angiosperms with different root hair distribution patterns, *Plant Cell*, **18**, 2958–70.
43. Walther, D., Brunnemann, R. and Selbig, J. 2007, The regulatory code for transcriptional response diversity and its relation to genome structural properties in *A. thaliana*, *PLoS Genet.*, **3**, e11.
44. Zhang, W., Ruan, J., Ho, T.H., You, Y., Yu, T. and Quatrano, R.S. 2005, Cis-regulatory element based targeted gene finding: genome-wide identification of abscisic acid- and abiotic stress-responsive genes in *Arabidopsis thaliana*, *Bioinformatics*, **21**, 3074–81.
45. Won, S.K., Lee, Y.J., Lee, H.Y., Heo, Y.K., Cho, M. and Cho, H.T. 2009, Cis-element- and transcriptome-based screening of root hair-specific genes and their functional characterization in Arabidopsis, *Plant Physiol.*, **150**, 1459–73.
46. Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L. and Van de Peer, Y. 2009, Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks, *Plant Physiol.*, **150**, 535–46.
47. Tran, L.S., Nakashima, K., Sakuma, Y., et al. 2004, Isolation and functional analysis of Arabidopsis stress-inducible NAC transcription factors that bind to a drought-responsive cis-element in the early responsive to dehydration stress 1 promoter, *Plant Cell*, **16**, 2481–98.
48. Yamaguchi-Shinozaki, K. and Shinozaki, K. 2006, Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses, *Annu. Rev. Plant Biol.*, **57**, 781–803.
49. Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. and Luscombe, N.M. 2009, A census of human transcription factors: function, expression and evolution, *Nat. Rev. Genet.*, **10**, 252–63.
50. Kondou, Y., Higuchi, M., Takahashi, S., et al. 2009, Systematic approaches to using the FOX hunting system to identify useful rice genes, *Plant J.*, **57**, 883–94.
51. Mochida, K. and Shinozaki, K. 2010, Genomics and bioinformatics resources for crop improvement, *Plant Cell Physiol.*, **51**, 497–523.
52. Kuromori, T., Wada, T., Kamiya, A., et al. 2006, A trial of phenome analysis using 4000 Ds-insertional mutants in gene-coding regions of Arabidopsis, *Plant J.*, **47**, 640–51.
53. Kolesnik, T., Szeverenyi, I., Bachmann, D., et al. 2004, Establishing an efficient Ac/Ds tagging system in rice: large-scale analysis of Ds flanking sequences, *Plant J.*, **37**, 301–14.
54. Ichikawa, T., Nakazawa, M., Kawashima, M., et al. 2003, Sequence database of 1172 T-DNA insertion sites in Arabidopsis activation-tagging lines that showed phenotypes in T1 generation, *Plant J.*, **36**, 421–9.
55. Kuromori, T., Takahashi, S., Kondou, Y., Shinozaki, K. and Matsui, M. 2009, Phenome analysis in plant species using loss-of-function and gain-of-function mutants, *Plant Cell Physiol.*, **50**, 1215–31.
56. Guo, A.Y., Chen, X., Gao, G., et al. 2008, PlantTFDB: a comprehensive plant transcription factor database, *Nucleic Acids Res.*, **36**, D966–9.
57. Iida, K., Seki, M., Sakurai, T., et al. 2005, RARTF: database and tools for complete sets of Arabidopsis transcription factors, *DNA Res.*, **12**, 247–56.
58. Gao, G., Zhong, Y., Guo, A., et al. 2006, DRTF: a database of rice transcription factors, *Bioinformatics*, **22**, 1286–7.