Nucleic Acids Research Advance Access published November 18, 2010

Nucleic Acids Research, 2010, 1–4
doi:10.1093/nar/gkq1141

PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database

He Zhang, Jinpu Jin, Liang Tang, Yi Zhao, Xiaocheng Gu, Ge Gao* and Jingchu Luo*

Center for Bioinformatics, National Laboratory of Protein Engineering and Plant Genetic Engineering and College of Life Sciences, Peking University, Beijing, 100871, PR China

Received September 13, 2010; Revised October 19, 2010; Accepted October 22, 2010

ABSTRACT

We updated the plant transcription factor (TF) database to version 2.0 (PlantTFDB 2.0, http://planttfdb.cbi.pku.edu.cn) which contains 53 319 putative TFs predicted from 49 species. We made detailed annotation including general information, domain feature, gene ontology, expression pattern and ortholog groups, as well as cross references to various databases and literature citations for these TFs classified into 58 newly defined families with computational approach and manual inspection. Multiple sequence alignments and phylogenetic trees for each family can be shown as Weblogo pictures or downloaded as text files. We have redesigned the user interface in the new version. Users can search TFs with much more flexibility through the improved advanced search page, and the search results can be exported into various formats for further analysis. In addition, we now provide web service for advanced users to access PlantTFDB 2.0 more efficiently.

INTRODUCTION

Transcription factors (TFs) are key regulators for transcriptional expression in biological processes (1). During the past years, several databases of plant TFs and other transcription regulators have been publicly available, such as PlnTFDB (2), PlantTAPDB (3), GRASSIUS (4), DATFAP (5), AGRIS (6), RARTF (7), LegumeTFDB (8) and TOBFAC (9). Start from 2005, we have constructed several species-specific plant TF databases with available genome sequences of *Arabidopsis* (DATF) (10), rice (DRTF) (11) and poplar (DPTF) (12), and integrated them into a comprehensive plant TF database (PlantTFDB 1.0) (13) with 26 402 TFs identified from 22 species. Of these 22 plants, five species have completed genome sequences and the others have unique transcripts integrated by PlantGDB (14). PlantTFDB 1.0 has received millions web hits since it went online in July 2007.

With the rapid increase of plant genome sequences in public databases, we have updated the PlantTFDB 1.0 to version 2.0. PlantTFDB 2.0 contains TFs from 49 species covering the main lineages of the plant kingdom, 9 from green algae, 1 from moss, 1 from fern, 3 from gymnosperm and 35 from angiosperm. Using the refined pipeline, a total of 53 319 TFs were identified from these 49 species and classified into 58 families. We made both computational annotation and manual curation for those putative TFs. In order to infer the evolutionary relationships among identified TFs, we constructed phylogenetic trees for each TF family and predicted ortholog groups for the TFs identified from species with completed genome sequences. The web interface of the PlantTFDB 2.0 was redesigned to provide users with more flexible search functionality. In addition to browsing through a web browser, standard web service interface is now supported for advanced users to retrieve data from PlantTFDB 2.0 in a batch mode or integrate data in PlantTFDB 2.0 into their website. All resources in PlantTFDB 2.0 can be browsed, retrieved and downloaded freely.

RESULTS AND DISCUSSION

Improved identification pipeline for plant TFs

While annotations generated by genome sequencing projects provide the most abundant source for proteome of the given species, the automatic annotation nature may often produce incomplete or incorrect annotation (15). On the other hand, dedicated sequence databases like RefSeq (16) provide relatively high quality curation-based annotation. And expressed sequence tag (EST) is also an important source to complement genome annotation. By integrating all existing annotations derived from genome annotation, RefSeq, PlantGDB (14) and UniGene (17), we

*To whom correspondence should be addressed. Tel:/Fax: +86 10 6275 5206; Email: luojc@pku.edu.cn
Correspondence may also be addressed to Ge Gao. Tel:/Fax: +86 10 6275 1861; Email: gaog@mail.cbi.pku.edu.cn

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

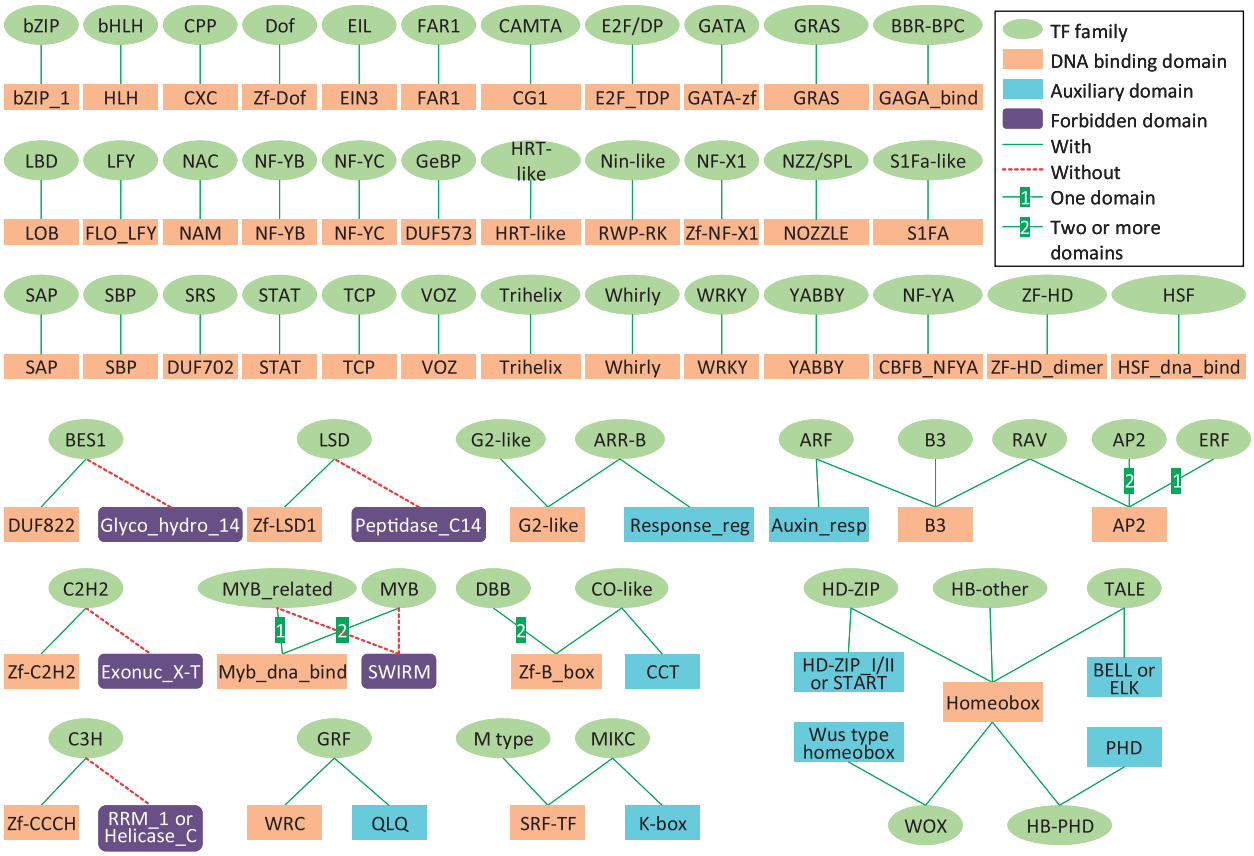

Figure 1. Family assignment rules used to identify and assign TFs into different families. Green ellipses represent TF families, and red rectangles denote DBDs. Blue and purple rectangles denote auxiliary and forbidden domains, respectively. Green solid lines link families and DBDs or auxiliary domains, number '1' or '2' on the lines indicate number of DBDs. Red dash lines link families and forbidden domains.

compiled a non-redundant reference proteome dataset for all 49 species (Supplementary Table S1, Supplementary Figures S1 and S2) for TF prediction.

TFs are characterized by their signature DNA-binding domains (DBDs). We employed HMMER 3.0 to identify those signature DBDs from the above proteome data set. In total, 64 HMM models were used to identify domains in TF (Supplementary Table S2), of which 53 models were collected from Pfam 24.0 (18) and 11 models were built using the sequences we collected locally. In the previous version, we set *e*-value 0.01 as the threshold for domain identification. Based on manual inspection and literature review, we adopted domain-specific bit-score as the threshold in the current version, since *e*-value is dependent on the size of given protein data set (Supplementary Tables S3 and S4).

In PlantTFDB 2.0, we adopted a slightly stringent definition that TFs are 'proteins that show sequence-specific DNA binding and are capable of activating or/and repressing transcription' (19). We made an extensive literature review and refined the rule-based classification scheme accordingly (Figure 1 and Supplementary Table S5). In PlantTFDB 2.0, we excluded families that do not meet the above criteria (Supplementary Table S6), including transcription cofactors and chromatin-related proteins such as remodeling factors, histone demethylases, DNA methyltransferases and histone acetyltransferases. Families such as TUBBY-like and Alfin-like were also removed since they were questioned or disproved by new experimental evidences. On the other hand, five newly identified TF families (DBB, FAR1, LSD, NF-X1, STAT) were added in PlantTFDB 2.0. Due to differences in domain composition, DNA binding specificity and function, AP2/ERF and HB were divided to sub-families. The M type of MADS TFs was classified as a new sub-family, since it has been reported that some M type of MADS-box genes could be pseudogenes or a new class of transposable element (19). Finally, we predicted 53 319 TFs from 49 species and classified them into 58 families (Tables 1 and 2, Supplementary Tables S7 and S8) using the refined pipeline.

Comprehensive annotation for plant TFs

Comprehensive and accurate annotations derived from various sources provide valuable clues for further functional analysis. Based on our established annotation pipeline, we performed systematic annotation for each family and individual TF.

The main page of each family has a distribution chart to show the number of TFs of each species in this family. The information of brief introduction and key references for each family was updated based on literature survey.

Table 1. Summary of TFs identified from species with genome sequences

Lineage	Species	Common name	Protein	TF	(%)	Family	OG[a]	TFOG[a]
Monocotyledon	*Brachypodium distachyon*	Purple False Brome	30 726	1687	5.49	56	1016	1271
	Oryza sativa subsp. *indica*	Indian Rice	43 027	1936	4.50	56	1427	1692
	Oryza sativa subsp. *japonica*	Japanese Rice	58 760	2424	4.13	56	1422	1636
	Sorghum bicolor	Sorghum	35 810	1819	5.08	54	1252	1583
	Zea mays	Maize	62 184	3355	5.40	56	1208	1762
Dicotyledon	*Arabidopsis lyrata*	Lyrate Rockcress	32 233	1729	5.36	58	1298	1604
	Arabidopsis thaliana	Thale Cress	32 125	2016	6.28	58	1297	1609
	Carica papaya	Papaya	27 829	1387	4.98	58	881	1203
	Cucumis sativus	Cucumber	27 725	1769	6.38	57	894	1153
	Glycine max	Soybean	48 707	3546	7.28	57	1148	3057
	Lotus japonicus	–	27 974	1275	4.56	56	752	986
	Manihot esculenta	Cassava	46 478	2201	4.74	58	1084	1922
	Medicago truncatula	Barrel Medic	52 086	1605	3.08	56	823	1272
	Mimulus guttatus	Spotted Monkey Flower	27 989	1681	6.01	57	863	1345
	Populus trichocarpa	Western Balsam Poplar	45 183	2585	5.72	58	1086	2195
	Prunus persica	Peach	28 299	1513	5.35	58	1006	1380
	Ricinus communis	Castor Bean	31 953	1291	4.04	57	994	1170
	Vitis vinifera	Wine Grape	47 097	2436	5.17	58	921	1207
Fern	*Selaginella moellendorffii*	–	32 969	971	2.95	55	411	856
Moss	*Physcomitrella patens subsp. patens*	–	40 604	1188	2.93	53	322	863
Green alga	*Chlamydomonas reinhardtii*	–	23 042	224	0.97	30	123	136
	Chlorella sp. *NC64A*	–	9762	163	1.67	28	94	120
	Coccomyxa sp. *C-169*	–	9900	123	1.24	29	82	90
	Micromonas pusilla CCMP1545	–	10 518	141	1.34	32	119	124
	Micromonas sp. *RCC299*	–	10 074	153	1.52	32	124	134
	Ostreococcus lucimarinus CCE9901	–	7960	118	1.48	30	100	103
	Ostreococcus sp. *RCC809*	–	7484	100	1.34	29	95	97
	Ostreococcus tauri	–	7654	97	1.27	26	89	91
	Volvox carteri	–	15 416	168	1.09	28	125	137

[a]OG: number of ortholog groups including at least two TFs; TFOG: number of TFs in ortholog groups.

Table 2. Summary of TFs identified from species without genome sequences

Groups	Species	Common name	Protein	TF	(%)	Family
Monocotyledon	*Hordeum vulgare*	Barley	24 020	778	3.24	54
	Panicum virgatum	Switchgrass	30 078	1140	3.79	52
	Saccharum officinarum	Sugarcane	21 172	671	3.17	48
	Triticum aestivum	Wheat	20 494	746	3.64	53
Dicotyledon	*Arachis hypogaea*	Peanut	7243	219	3.02	39
	Artemisia annua	Sweet Wormwood	13 062	514	3.94	48
	Brassica napus	Rape	30 482	1334	4.38	53
	Brassica rapa	Field Mustard	14 313	718	5.02	49
	Citrus sinensis	Valencia Orange	13 522	534	3.95	46
	Gossypium hirsutum	Upland Cotton	20 862	1111	5.33	50
	Helianthus annuus	Sunflower	8634	279	3.23	44
	Malus x domestica	Apple	15 173	658	4.34	51
	Nicotiana tabacum	Tobacco	18 898	793	4.20	52
	Raphanus sativus	Radish	14 799	573	3.87	45
	Solanum lycopersicum	Tomato	15 722	799	5.08	54
	Solanum tuberosum	Potato	17 445	776	4.45	52
	Theobroma cacao	Cocoa	7493	239	3.19	44
	Vigna unguiculata	Cowpea	12 205	475	3.89	48
Gymnosperm	*Picea glauca*	White Spruce	15 376	508	3.30	48
	Picea sitchensis	Sitka Spruce	10 989	319	2.90	47
	Pinus taeda	Loblolly Pine	13 275	434	3.27	47

Multiple sequence alignments for DBDs of each family, either of individual species or among species, can be viewed as WebLogo pictures, or downloaded as text files. Phylogenetic trees can be displayed online or downloaded to local PC in Nexus format. Intra-species phylogenetic trees for each TF family were inferred by MrBayes (v3.2) (20) using the Dayhoff substitution model with 50 000 generations, and FastTree2.1 (21) was employed to construct inter-species trees with 100 resamplings. Annotations at the individual TF level contain general information, domain architecture, gene ontology, PDB hits, expression profiles, cross-references to other databases, ortholog groups, literature citations and links to other useful resources.

Improvement of user interface

We have redesigned the web interface for PlantTFDB 2.0 which has a uniform interface for all species now. Users can browse individual TFs of different families for each species by simply clicking the unique IDs assigned to each TF. The text search page has been greatly improved with much more flexibility for users to make advanced search. Users can select several species in the same or different lineages within the species tree to search TFs in one or more families. Users can combine several query conditions in a single search, including general descriptions, protein properties such as the range of sequence length, various tissues of gene expression and different fields of annotation for TF entries. Users can also customize and save the search results in various formats for further processing.

While accessing the resource through web browsers is an easy and intuitive way for most users, web service is efficient for advanced users to access and integrate data into their own sites. We implemented a standard web service interface for PlantTFDB 2.0 (http://planttfdb.cbi.pku.edu.cn/webservice/server.php). A demo for client implementation in PHP is available to help users to get familiar with the web service interface (http://planttfdb.cbi.pku.edu.cn/webservice_client/client.php).

FURTHER DIRECTION

In conclusion, PlantTFDB 2.0 is not only an extensive update of the previous version with newly released 29 completed genomes and updated data sets, but also a great improvement of the user interface. The pipelines we developed for the prediction of TFs at genome scale, the scheme we defined to classify TF families in plants may provide the user community with some useful tools. We will continue on this project to make further update and improvement of PlantTFDB in the future.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank JGI for genome annotations of 10 unpublished species, MGSC for Medicago truncatula data. We appreciate critical comments from all users.

FUNDING

China 863 (2007AA02Z165), 973 (2007CB946904) and NSFC (31071160) programs. Funding for open access publication: China NSFC (31071160) program.

Conflict of interest statement. None declared.

REFERENCES

1. Riechmann,J.L., Heard,J., Martin,G., Reuber,L., Jiang,C., Keddie,J., Adam,L., Pineda,O., Ratcliffe,O.J., Samaha,R.R. *et al.* (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.
2. 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–D827.
3. 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**, 1452–1466.
4. Yilmaz,A., Nishiyama,M.Y. Jr, Fuentes,B.G., Souza,G.M., Janies,D., Gray,J. and Grotewold,E. (2009) GRASSIUS: a platform for comparative regulatory genomics across the grasses. *Plant Physiol.*, **149**, 171–180.
5. Fredslund,J. (2008) DATFAP: a database of primers and homology alignments for transcription factors from 13 plant species. *BMC Genomics*, **9**, 140.
6. 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**, 818–829.
7. Iida,K., Seki,M., Sakurai,T., Satou,M., Akiyama,K., Toyoda,T., Konagaya,A. and Shinozaki,K. (2005) RARTF: database and tools for complete sets of Arabidopsis transcription factors. *DNA Res.*, **12**, 247–256.
8. 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**, 290–291.
9. 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 Bioinformatics*, **9**, 53.
10. Guo,A., He,K., Liu,D., Bai,S., Gu,X., Wei,L. and Luo,J. (2005) DATF: a database of Arabidopsis transcription factors. *Bioinformatics*, **21**, 2568–2569.
11. Gao,G., Zhong,Y., Guo,A., Zhu,Q., Tang,W., Zheng,W., Gu,X., Wei,L. and Luo,J. (2006) DRTF: a database of rice transcription factors. *Bioinformatics*, **22**, 1286–1287.
12. Zhu,Q.H., Guo,A.Y., Gao,G., Zhong,Y.F., Xu,M., Huang,M. and Luo,J. (2007) DPTF: a database of poplar transcription factors. *Bioinformatics*, **23**, 1307–1308.
13. Guo,A.Y., Chen,X., Gao,G., Zhang,H., Zhu,Q.H., Liu,X.C., Zhong,Y.F., Gu,X., He,K. and Luo,J. (2008) PlantTFDB: a comprehensive plant transcription factor database. *Nucleic Acids Res.*, **36**, D966–D969.
14. Duvick,J., Fu,A., Muppirala,U., Sabharwal,M., Wilkerson,M.D., Lawrence,C.J., Lushbough,C. and Brendel,V. (2008) PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res.*, **36**, D959–D965.
15. Ouyang,S., Thibaud-Nissen,F., Childs,K.L., Zhu,W. and Buell,C.R. (2009) Plant genome annotation methods. *Methods Mol. Biol.*, **513**, 263–282.
16. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, **37**, D32–D36.
17. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Federhen,S. *et al.* (2010) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **38**, D5–D16.
18. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
19. Riechmann,J. (2006) Transcription factors of Arabidopsis and rice: a genomic perspective. In Grasser,K. (ed.), *Regulation of Transcription in Plants*. Wiley-Blackwell, Oxford, pp. 28–53.
20. Ronquist,F. and Huelsenbeck,J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.
21. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.