Molecular Plant
Letter to the Editor

iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases

Dear Editor,

Transcription factors (TFs) are proteins that regulate the expression of target genes by binding to specific *cis*-elements in promoter regions. Transcriptional regulators (TRs) also regulate the expression of target genes; however, they operate indirectly via interaction with the basal transcription apparatus (e.g., TFs), or by altering the accessibility of DNA to TFs via chromatin remodeling. Another type of regulatory proteins, protein kinases (PKs), function in signal transduction pathways and alter the activity of target proteins by phosphorylating them. These three important classes of regulatory proteins have been associated with numerous aspects of plant growth and development (Gapper et al., 2014; Xu and Zhang, 2015), and response to biotic and abiotic stimuli (Zhang et al., 2013; Mickelbart et al., 2015). Effective and accurate identification and classification of these genes is important for understanding their evolution, biological functions, and regulatory networks. Currently, more than 100 plant genomes have been sequenced and regulatory proteins have been systematically identified from several of these plant genomes. Databases presenting these regulatory proteins, especially TFs, have been developed, such as PlnTFDB (Pérez-Rodríguez et al., 2010) and PlantTFDB (Jin et al., 2013). However, annotations of TF/TR families and the associated classification rules have been inconsistent among different studies. For example, the PlantTFDB does not include TRs that are presented in PlnTFDB. As another example, the forbidden domain (a domain that the specific TF families should not contain) of the C2H2 family is annotated as an RNase_T domain in PlantTFDB but as a PHD domain in PlnTFDB. Presently, while the collection of genome sequences is rapidly expanding, cataloged and annotated TFs/TRs vary across different databases due to inconsistent identification and characterization criteria with serious consequences for genome-scale and targeted analyses. Furthermore, in contrast to many studies focusing on specific families of plant regulators, computational tools for identification and classification of these regulatory proteins on a genome scale are very limited.

In this study, we systemically compared TF/TR classification rules used in different databases, and derived a set of consensus rules based on the available literature for accurate plant TF/TR identification and classification. For plant PKs, we directly used the HMM profiles developed by Lehti-Shiu and Shiu (2012) to provide a comprehensive classification system. These consensus rules for TF/TR classification and HMM profiles for PK classification were implemented in iTAK (http://bioinfo.bti.cornell.edu/tool/itak), a computational program that provides consistency and uniformity on the identification and classification of plant TFs, TRs, and PKs.

To construct consensus rules for TF/TR prediction and classification, we compared Pfam domain assignment rules between PlnTFDB and PlantTFDB. The families in PlantTFcat (Dai et al., 2013) and AtTFDB (Yilmaz et al., 2011) were used as supporting evidence, as they use different methods for domain identification, therefore cannot be directly compared with PlnTFDB and PlantTFDB. A family annotation was considered more reliable if it had been assigned in both PlnTFDB and PlantTFDB, while a family unique to a single database was considered to be less reliable and required more evidence to support its identity. Under this criterion, 57 TF families/subfamilies were considered reliable, while 25 were considered less reliable (Supplemental Table 1). Comparison of domain assignment rules for the reliable families between PlnTFDB and PlantTFDB indicated that most were consistent, but rules of several subfamilies were missing in PlnTFDB. For example, PlantTFDB defines the AP2/ERF family to comprise three subfamilies, AP2, ERF, and RAV, while PlnTFDB only defines an AP2-EREBP family. In this example, the domain assignment rules used in PlantTFDB provide more details about the relationship between the superfamily and subfamily of AP2/ERF. Therefore, we adopted the rules for both the AP2/ERF superfamily and the three subfamilies (Figure 1A). Similarly, the rules for the NF-Y (CCAAT) and MADS families were also adopted from PlantTFDB as they provide more detailed TF classification.

In addition to handling missing rules in either PlnTFDB or PlantTFDB, we updated the domain assignment rules for several families including Homeobox (HB), BSD, and LIM, based on literature review (Figure 1A). In PlantTFDB, the HB superfamily is divided into five subfamilies: HD-ZIP, TALE, WOX, HB-PHD, and HB-other. In our consensus rule set, HB-TALE was further divided into two subfamilies, HB-BELL and HB-KNOX. We made this assignment because members of HB-BELL and HB-KNOX have different domains: HB-BELL contains POX and HB-KN domains, while HB-KNOX has a KNOX1 and a KNOX2 domain. In our study, the HD-ZIP_I/II domain that typifies the HB-HD-ZIP subfamily was replaced by the HALZ domain, which was specifically built from homeodomain-leucine zipper proteins. In addition, we updated the classification rule for the BSD family to require both a BSD and a PH_TFIIH domain, instead of requiring only the BSD domain as done by PlantTFDB and PlnTFDB. Finally, the LIM subfamily was updated to require two LIM domains (Weiskirchen and Günther, 2003) (Figure 1A; Supplemental Table 1).

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, SIBS, CAS.

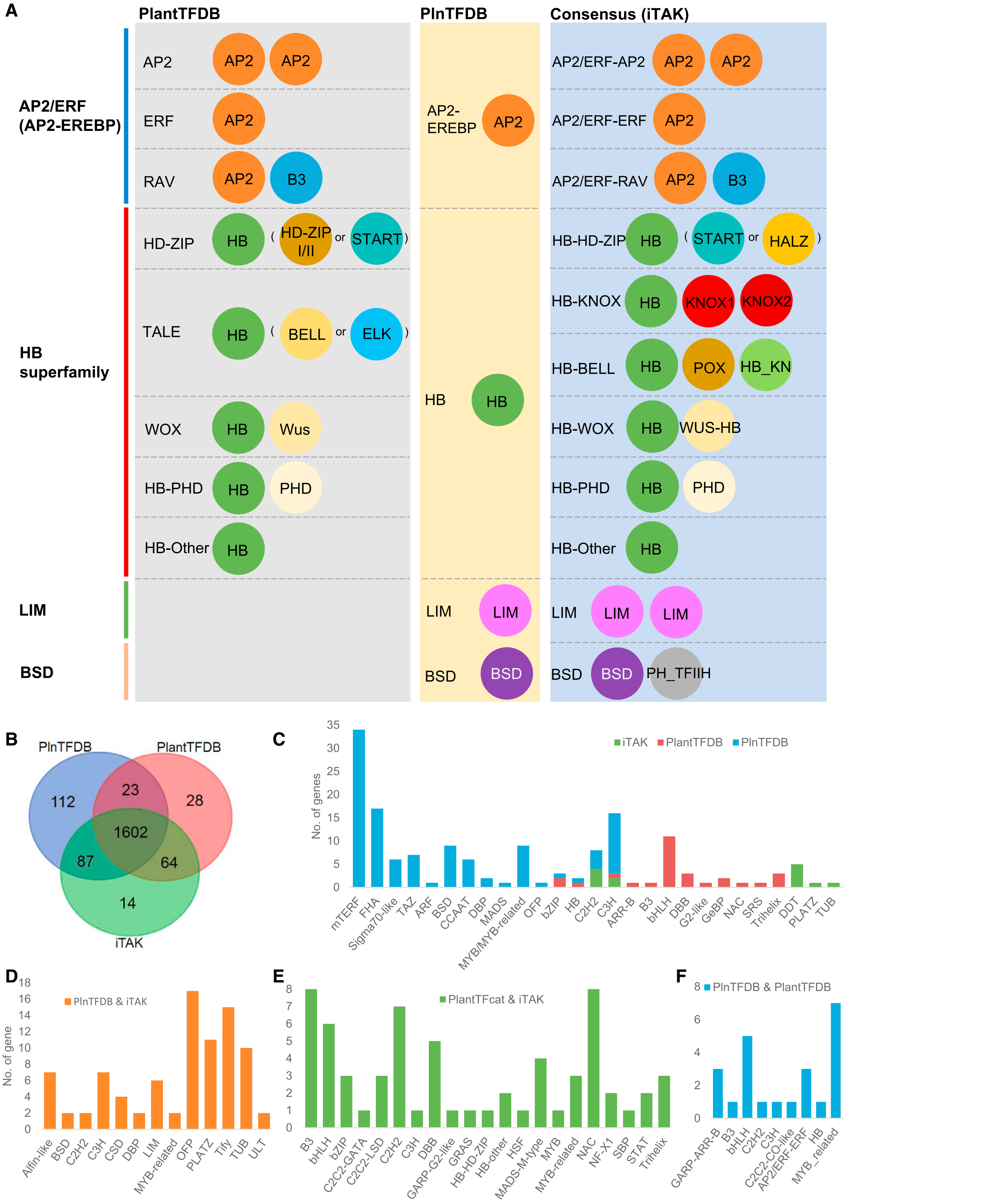

Figure 1. Construction of Consensus Rules and Comparison of Predicted *Arabidopsis thaliana* TFs between iTAK, PlnTFDB, and PlantTFDB.

(A) Construction of consensus rules for the AP2/ERF superfamily, homeobox (HB) superfamily, LIM, and BSD families of transcription factors. Colored backgrounds represent different domain assignment rules used in different analyses. Circles with different colors represent required domains for different families.

(legend continued on next page)

In reviewing the literature for the 25 TF families that were supported by only one of PlantTFDB and PlnTFDB, we found that six (WD40-like, TIG, FHA, Sigma70-like, TAZ, and mTERF) were inaccurately categorized as TFs (Supplemental Table 1). We excluded WD40-like since WD-repeat proteins perform diverse functions and using the existing rule would result in the identification of many non-TF WD40-like proteins. Based on multiple sequence alignments of TIG proteins, it was difficult to distinguish TFs from the TIG superfamily of proteins, which include not only TFs but also kinases and membrane proteins. Similarly, the FHA domain is present in a functionally diverse range of proteins that include not only TFs but also kinases, phosphatases and kinesins, and plant TFs containing an FHA domain have not been found. Sigma70-like proteins were also excluded from the TF category since they do not themselves bind to promoters; rather they function as components of an RNA polymerase holoenzyme involved in binding a core RNA polymerase to specific promoters. Finally, mTERF and TAZ were categorized as TRs instead of TFs because TAZ proteins function as coactivators alongside other regulatory proteins, and mTERF proteins exert a broad range of regulatory activities while not binding directly to promoters. To achieve low and balanced false positive and false negative rates, we excluded these six families from the rule set that defines TFs. Furthermore, the DDT family, which was categorized as TR in PlnTFDB, was categorized as TF. Based on a literature review, the remaining 16 TF families were included, resulting in a set of consensus rules that included 72 families/subfamilies for plant TF classification (Supplemental Table 1).

The TR family classification rules were adopted from those used in PlnTFDB with support from PlantTFcat. Excluding the aforementioned DDT family, a total of 23 TR families including TAZ and mTERF that were incorrectly classified into TFs were derived from PlnTFDB. All these TR families were reviewed and accepted based on literature support (Supplemental Table 2).

We developed the iTAK program base on the consensus rules we derived for the identification and classification of plant TFs/TRs/PKs (Supplemental Figure 1). To evaluate the performance of iTAK, the predicted *Arabidopsis* TFs with iTAK were systemically compared with those identified in PlantTFDB and PlnTFDB. The three datasets shared a total of 1602 TFs, accounting for approximately 90% of the TFs in PlnTFDB, PlantTFDB, and iTAK (Figure 1B). Although the majority of them were commonly identified as TFs and classified into the same families, we did observe some inconsistencies. The inconsistencies were mainly between PlnTFDB and PlantTFDB, while the iTAK classifications were consistent with one of the two other databases, with the exception of two genes, AT1G50680 and AT1G51120, which were assigned to the B3 family using iTAK, rather than to the AP2/ERF-RAV family by PlnTFDB and PlantTFDB, because they only contained B3 domains (Supplemental Table 3). This minor difference may reflect the recent update of the AP2 HMM profile in the Pfam database. Overall, the high consistency between iTAK and other studies indicates the high accuracy of TF identification and classification by iTAK. A total of 112, 28, and 14 TFs were identified only in PlnTFDB, PlantTFDB, or iTAK, respectively (Figure 1C; Supplemental Table 4). Five of the 14 iTAK-specific TFs were from the DDT family, which were inaccurately categorized as TRs in the other databases. Of the PlnTFDB-specific genes, 64 belonged to the mTERF, FHA, sigma70-like, and TAZ families, which should not be categorized as TFs. After eliminating these discrepant families, 48, 28, and 9 genes were predicted only by PlnTFDB, PlantTFDB, or iTAK, respectively. Furthermore, 87, 64, and 23 TFs were not identified by PlantTFDB, PlnTFDB, and iTAK, respectively, but were predicted and assigned to the same families by the other two (Figure 1D–1F). The smaller number of unique and missing identifications by iTAK indicates it achieved a better balance between false positives and false negatives. The reason that iTAK did not identify the 23 TFs was mainly due to the significance cutoff of the required domains (Supplemental Information; Supplemental Table 5). The identified TFs/TRs were also compared with other datasets, further supporting the high accuracy of iTAK (Supplemental Information).

In summary, we have derived a set of consensus domain assignment rules for accurate identification and classification of plant TFs and TRs. We have developed a novel bioinformatics tool, iTAK, to facilitate genome-wide identification and classification of plant TFs, TRs, and PKs, and a comprehensive database for these regulatory proteins from sequenced plant species (Supplemental Information). These provide valuable tools and resources for the research community to study transcriptional regulations and signaling networks.

SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

FUNDING

This work was supported in part by a seed grant from the Association of Independent Plant Institutes (AIPI) to S.Y.R., P.Z., and Z.F. and grants from the National Science Foundation (IOS-0923312, IOS-1025642, and IOS-1339287 to Z.F., DBI-0960897 and DBI-1458597 to P.X.Z. and IOS-1026003 to S.Y.R.) and Department of Energy (DE-SC0008769) to S.Y.R.

AUTHOR CONTRIBUTIONS

Y.Z., Z.F., S.Y.R., P.X.Z., G.B.M., and J.J.G. designed the research. Y.Z. implemented iTAK. Y.Z., C.J., H.S., H.G.R., M.A.P., P.Z., and M.B. performed the data analysis and gene family curation. Y.Z. and Z.F. wrote the article. Z.F., S.Y.R., P.X.Z., G.B.M., and J.J.G. revised the article.

ACKNOWLEDGMENTS

No conflict of interest declared.

(B) Venn diagram showing the number of TFs identified in different analyses.

(C) Classification of TFs specific to iTAK, PlnTFDB, or PlantTFDB.

(D) Classification of TFs common only to both iTAK and PlnTFDB.

(E) Classification of TFs common only to both iTAK and PlantTFDB.

(F) Classification of TFs common only to both PlnTFDB and PlantTFDB.

Received: August 21, 2016
Revised: August 21, 2016
Accepted: September 28, 2016
Published: October 4, 2016

Yi Zheng[1], *Chen Jiao*[1], *Honghe Sun*[1], *Hernan G. Rosli*[1,2], *Marina A. Pombo*[1,2], *Peifen Zhang*[3], *Michael Banf*[3], *Xinbin Dai*[4], *Gregory B. Martin*[1,5], *James J. Giovannoni*[1,6], *Patrick X. Zhao*[4], *Seung Y. Rhee*[3] and *Zhangjun Fei*[1,6,*]

[1]Boyce Thompson Institute, Ithaca, NY 14853, USA
[2]Instituto de Fisiología Vegetal (INFIVE, CONICET-UNLP), La Plata 1900, Argentina
[3]Department of Plant Biology, Carnegie Institution for Science, Stanford, CA 94305, USA
[4]Samuel Roberts Noble Foundation, Ardmore, OK 73401, USA
[5]Section of Plant Pathology and Plant-Microbe Biology, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, USA
[6]USDA-ARS Robert W. Holley Center for Agriculture and Health, Ithaca, NY 14853, USA
***Correspondence: Zhangjun Fei (zf25@cornell.edu)**
http://dx.doi.org/10.1016/j.molp.2016.09.014

REFERENCES

Dai, X., Sinharoy, S., Udvardi, M., and Zhao, P.X. (2013). PlantTFcat: an online plant transcription factor and transcriptional regulator categorization and analysis tool. BMC Bioinformatics **14**:321.

Gapper, N.E., Giovannoni, J.J., and Watkins, C.B. (2014). Understanding development and ripening of fruit crops in an ‘omics’ era. Hortic. Res. **1**:14034.

Jin, J., Zhang, H., Kong, L., Gao, G., and Luo, J. (2013). PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. Nucleic Acids Res. **42**:D1182–D1187.

Lehti-Shiu, M.D., and Shiu, S.-H. (2012). Diversity, classification and function of the plant protein kinase superfamily. Philos. Trans. R. Soc. Lond. B Biol. Sci. **367**:2619–2639.

Mickelbart, M.V., Hasegawa, P.M., and Bailey-Serres, J. (2015). Genetic mechanisms of abiotic stress tolerance that translate to crop yield stability. Nat. Rev. Genet. **16**:237–251.

Pérez-Rodríguez, P., Riaño-Pachón, D.M., Corrêa, L.G.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.

Weiskirchen, R., and Günther, K. (2003). The CRP/MLP/TLP family of LIM domain proteins: acting by connecting. Bioessays **25**:152–162.

Xu, J., and Zhang, S. (2015). Mitogen-activated protein kinase cascades in signaling plant growth and development. Trends Plant Sci. **20**:56–64.

Yilmaz, A., Mejia-Guerra, M.K., Kurz, K., Liang, X., Welch, L., and Grotewold, E. (2011). AGRIS: *Arabidopsis* gene regulatory information server, an update. Nucleic Acids Res. **39**:D1118–D1122.

Zhang, Y., Lubberstedt, T., and Xu, M. (2013). The genetic and molecular basis of plant resistance to pathogens. J. Genet. Genomics **40**:23–35.