

# PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants

Jinpu Jin<sup>1,†</sup>, Feng Tian<sup>1,2,†</sup>, De-Chang Yang<sup>1</sup>, Yu-Qi Meng<sup>1</sup>, Lei Kong<sup>1</sup>, Jingchu Luo<sup>1,\*</sup> and Ge Gao<sup>1,\*</sup>

<sup>1</sup>State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences, Center for Bioinformatics, Beijing 100871, P.R. China and <sup>2</sup>Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, P.R. China

Received September 30, 2016; Editorial Decision October 11, 2016; Accepted October 12, 2016

## ABSTRACT

With the goal of providing a comprehensive, high-quality resource for both plant transcription factors (TFs) and their regulatory interactions with target genes, we upgraded plant TF database PlantTFDB to version 4.0 (<http://planttfdb.cbi.pku.edu.cn/>). In the new version, we identified 320 370 TFs from 165 species, presenting a more comprehensive genomic TF repertoires of green plants. Besides updating the pre-existing abundant functional and evolutionary annotation for identified TFs, we generated three new types of annotation which provide more directly clues to investigate functional mechanisms underlying: (i) a set of high-quality, non-redundant TF binding motifs derived from experiments; (ii) multiple types of regulatory elements identified from high-throughput sequencing data; (iii) regulatory interactions curated from literature and inferred by combining TF binding motifs and regulatory elements. In addition, we upgraded previous TF prediction server, and set up four novel tools for regulation prediction and functional enrichment analyses. Finally, we set up a novel companion portal PlantRegMap (<http://plantregmap.cbi.pku.edu.cn>) for users to access the regulation resource and analysis tools conveniently.

## INTRODUCTION

Transcription factors (TFs) temporarily and spatially turn on or off the transcription of their target genes through binding certain upstream elements. After the first attempt to identify the TF repertoire in *Arabidopsis thaliana* genome in 2000 (1), several dedicated TF databases for either individual or multiple species (2–10) have been publicly avail-

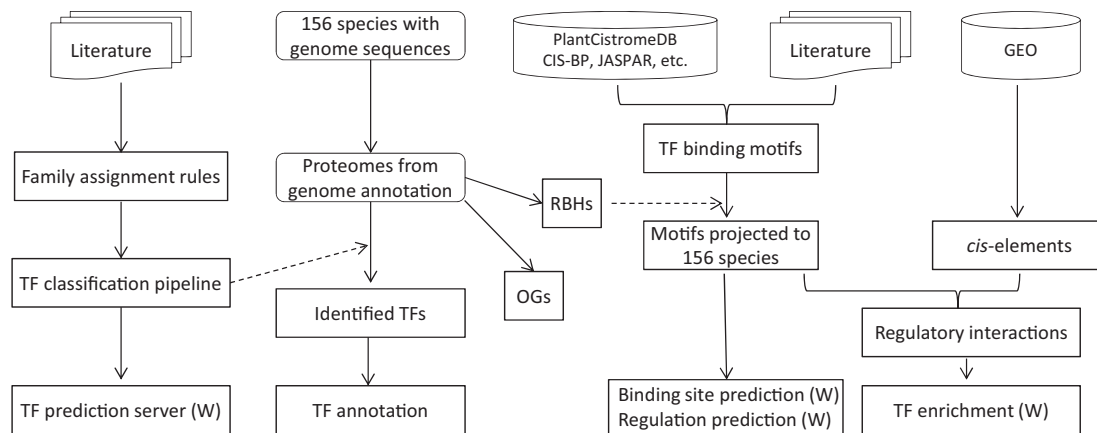
able, effectively advancing the study of plant transcriptional regulatory system. In response to the demand of systematic identification and annotation of plant TFs, we have developed three species-specific TF databases (DATF (11) for *Arabidopsis*, DRTF (12) for rice, and DPTF (13) for poplar) and three versions of comprehensive plant TF database (PlantTFDB) (14–17) during the past decade. These resources receive tens of millions web hits annually and have been widely used for the functional and evolutionary study of plant TFs and the identification and annotation of TFs in newly sequenced plants (18,19).

The regulatory interactions between TFs and target genes provide crucial information to investigate functional mechanisms of TFs. Recently, the DNA binding motifs of hundreds of plant TFs have been determined (20–22), combined with genomic regulatory elements identified from ChIP-seq and DNase-seq (23), providing us an opportunity to systematically map downstream regulatory interactions of TFs. To meet the urgent demand of our user community for transcriptional regulation information, we upgraded PlantTFDB to version 4.0 (<http://planttfdb.cbi.pku.edu.cn/>). Compared with previous versions, PlantTFDB 4.0 covers more species and more TFs with updated abundant functional and evolutionary annotation. Moreover, three new types of annotation including binding motifs of TFs, regulatory elements and interactions between them have been added. In addition, we upgraded previous TF prediction server, and set up four novel tools for regulation prediction and functional enrichment analyses (Figure 1). Finally, we set up a new portal PlantRegMap (<http://plantregmap.cbi.pku.edu.cn>) for users to access the regulation resource and analysis tools conveniently.

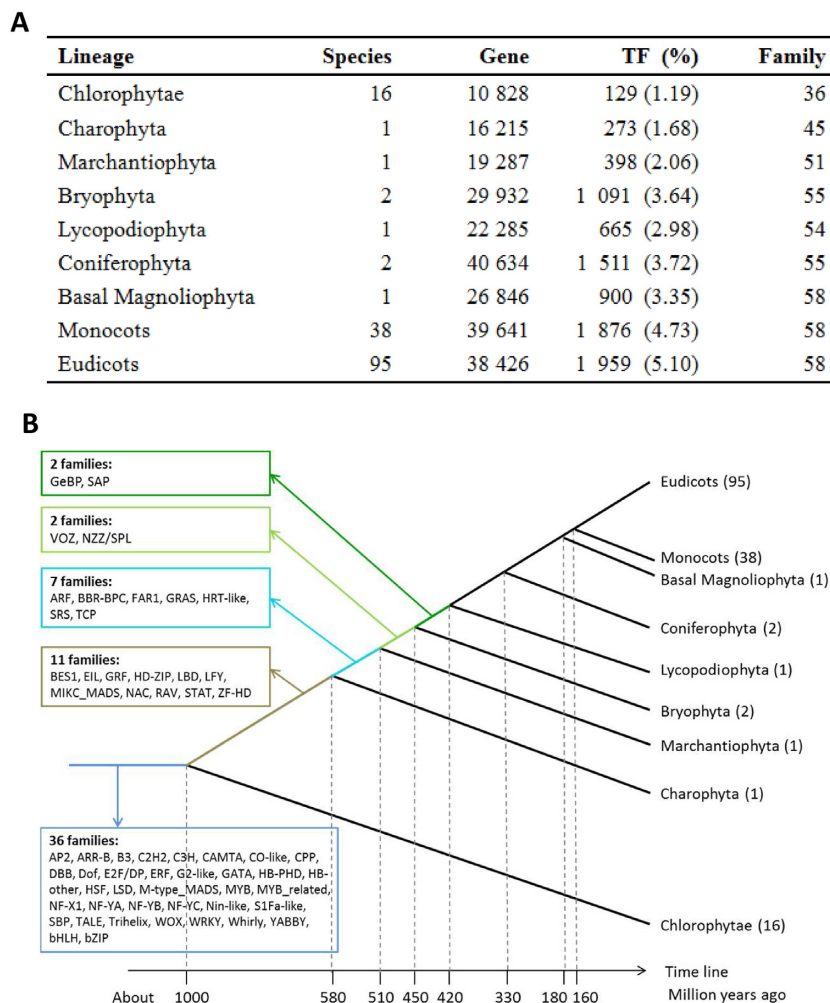
We believe that PlantTFDB 4.0 provides users a comprehensive resource of plant TFs, regulatory elements and regulatory interactions between them as well as useful analysis tools, advancing the study in the community.

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

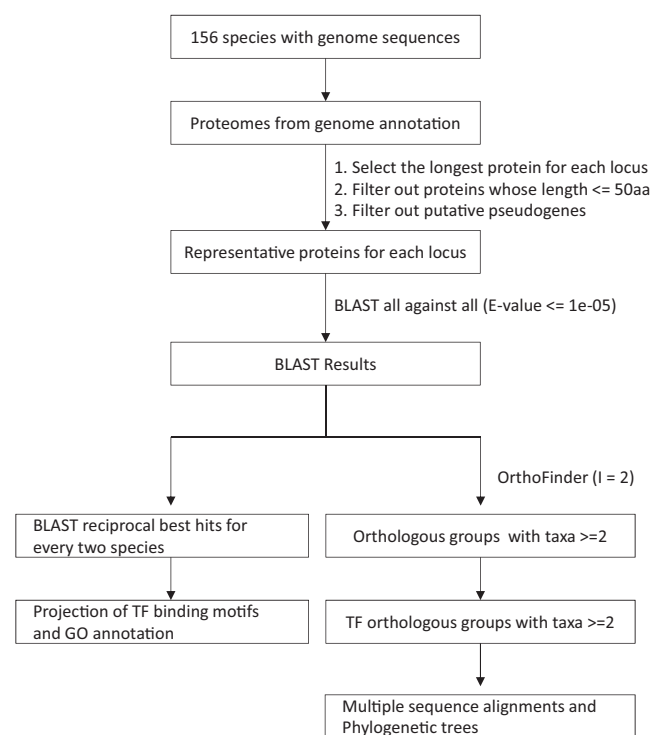
<sup>†</sup>These authors contributed equally to this work as first authors.



**Figure 1.** The flowchart for construction of PlantTFDB 4.0. Items marked with '(W)' are web services. RBH means BLAST reciprocal best hit, and OG means orthologous group.



**Figure 2.** Summary of TFs identified from 156 species with genome sequences. (A) Average number of TFs in different taxonomic lineages. (B) The origination stage of TF families. The number in the parentheses (e.g., Eudicots (95)) is the number of species used to date the origination stage for TF families according to the protocol described in (24).



**Figure 3.** Workflow for parsing BLAST reciprocal best hits (RBHs) and inferring orthologous groups.

## RESULTS AND DISCUSSION

### More representative genomic TF repertoires of green plants

By incorporating 89 newly sequenced genomes released during past three years, we totally collected 156 species with genome sequences. Using our established TF prediction pipeline (16) and protein sequences downloaded from Joint Genome Institute and other plant genome sequencing and annotation projects (see [http://planttfdb.cbi.pku.edu.cn/help\\_datasrc.php](http://planttfdb.cbi.pku.edu.cn/help_datasrc.php) for detailed data source), we identified 315 099 TFs (266 849 loci) from 6 381 676 proteins (5 555 839 loci) of 156 species (Supplementary Table S1). In addition, 5271 TFs from 9 species in version 3.0 remaining without genome sequences were kept.

By incorporating the 89 new species with genome sequences, we provide genomic TF repertoires covering two more early-diverged lineages, charophyta and marchantio-phyta, of green plants (Figure 2A) and present a refined origination profiling of TF families (Figure 2B) than previous studies (24). In addition, the inclusion of more wild species for domestic species (e.g., *Oryza rufipogon* and *Oryza nivara* for *Oryza sativa*) also facilitates the study of crop domestication at transcriptional regulation level and further crop breeding.

### Updated pipelines for evolutionary and functional annotation

As the continuous effort to improve annotation coverage and quality, we not only updated all existing annotation from multiple canonical sources (24–27) but also refined the annotation pipelines for Orthologous Group (OG) inference and Gene Ontology (GO) term assignment.

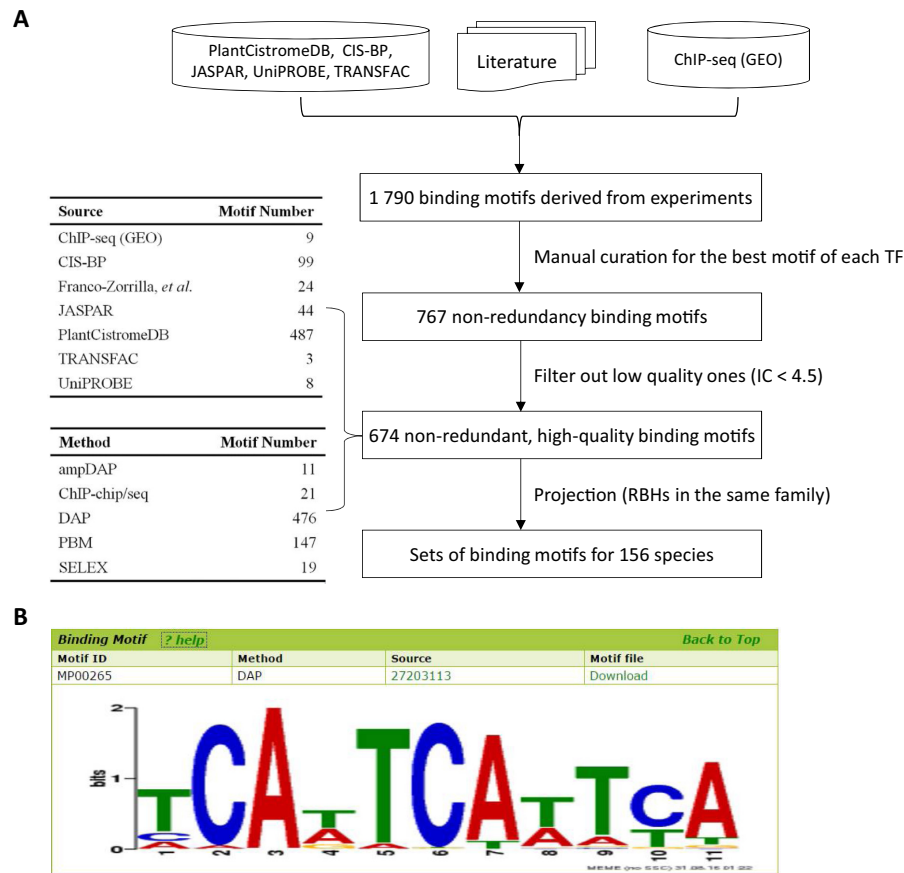
Orthologous genes usually possess similar function and are widely used to infer function interspecies. Given the recent reported methodology bias for OrthoMCL (28), we employed a new tool OrthoFinder (28) to infer orthologous groups (OGs) for proteins from 156 species with genome sequences (Figure 3), and got 2368 TF OGs for all species, 800 TF OGs for 17 representative species (Supplementary Table S2), 209, 1898, 1261, 1846 and 1900 TF OGs for the clades of Chlorophytae, Monocots, Asterids, Fabids and Malvids, respectively. Meanwhile, we also identified orthologous pairs under a stricter criterion, BLAST reciprocal best hits (RBHs), for all species pairs among 156 species with genome sequences, which serve as a basis for transferring annotation interspecies (e.g., from well-studied model plants like *Arabidopsis thaliana* to poor-annotated ones).

Gene Ontology is widely used to describe gene function. Taking an integrative strategy by combining sequence-based prediction (29), orthologous-based projection, and collection of existing annotation in canonical sources (25,26), we successfully annotated 93.2% (293 696 of 315 099) TFs in 165 species. Moreover, to provide a comprehensive reference dataset for follow-up functional enrichment analysis, we further annotated 59.3% (3 374 258 of 5 691 627) genes in these species, increasing the coverage of GO annotation to 9.1 times on average (Supplementary Table S3, see Supplementary Text 1 for details on the annotation pipeline).

### Genome-wide regulatory maps of green plants

TFs regulate the transcription of their target genes through binding certain *cis*-elements. Mapping regulatory elements is essential for the identification of regulatory interactions and further functional mechanisms of TFs. High-throughput assays like ChIP-seq (30) and genome-wide TF footprinting (23,31) have been efficiently employed to identify genome-wide TF occupancy sites. Here, we identified 345 920 TF binding peaks for 14 TFs based on 26 public ChIP-seq datasets as well as 4 794 773 TF footprints based on 13 public DNase-seq datasets according to the protocol described in (31) (Table 1, Supplementary Tables S4 and S5 and Supplementary Text 2). As histone modification and nucleosome positioning also provide useful clues for the study of transcriptional regulation, such information is also incorporated (Table 1, Supplementary Tables S6–S8 and Supplementary Text 2). In addition, we set up a genome browser for users to access these regulatory elements conveniently (<http://plantregmap.cbi.pku.edu.cn/cis-map.php>).

The DNA binding motifs of TFs are crucial to identifying their downstream target genes. During the past three years, binding motifs of hundreds of plant TFs have been determined and publicly available (20–22). We firstly collected 1790 binding motifs derived from experiments from multiple databases (including PlantCistromeDB (20), CIS-BP (21), JASPAR (32), UniPROBE (33), TRANSFAC (public 7.0) (34)), literature (22) and motifs discovered from ChIP-seq peaks using MEME-ChIP (35). For a TF with more than one motif, we manually curated the best one preferentially for the motif determined *in vivo* and presenting more similarity with other motifs of this TF. After filtering out low-quality ones (these with information content < 4.5), we



**Figure 4.** Curation and projection of TF binding motifs. (A) Workflow for curation and projection of high-quality TF binding motifs derived from experiments. (B) The visualization of TF binding motifs in PlantTFDB.

**Table 1.** Summary of integrated regulation information identified from high-throughput sequencing

Type	Species	Organ	Peak
DNase I hypersensitive site	2	10	1 026 748
DNase I genomic TF footprints	1	4	4 794 773
Histone modification	5	7	765 371
Nucleosome positioning	1	2	15 348 541
TF binding peak (ChIP-seq)	2	4	345 920

got 674 non-redundant, high-quality binding motifs, covering 77.6% (45 of 58) TF families in PlantTFDB. To take full advantage of this resource, we projected these motifs to other species using the RBHs identified above (only RBHs belong to the same TF family were used) and adjusted the motifs using the nucleic acid background in promoter regions (TSS –500 bp to +100bp). Finally, we got sets of species-specific binding motifs for 156 species (Figure 4 and Supplementary Table S9).

The regulatory interactions between TFs and target genes provide the most direct evidences to investigate the functional mechanisms of TFs. We collected transcriptional regulatory interactions in following three ways: (i) through mining transcriptional regulatory interactions from literature and then manually curated each interaction, we got 1431 high-confident transcriptional regulatory interactions, covering 388 TFs from 47 families, in *A. thaliana* (24), which were integrated in this update; (ii) identified 9340

experiment-based regulatory interactions identified for six TFs with at least two ChIP-seq replications; (iii) *in silico* predicted 342 000 regulatory interactions by combining TF binding motifs and footprint data (Table 2, Supplementary Tables S4 and S5 and Supplementary Text 2). As the high-quality interactions are mainly limited to *A. thaliana*, we further inferred regulatory interactions through *in silico* searching potential binding sites of TFs using binding motifs in gene promoter regions for 132 species with fine annotation of gene positions (Table 2, Supplementary Table S10 and Supplementary Text 2). Users are encouraged to make use of the less reliable interactions combined with other assistant information such as expression information and conserved elements.



**Table 2.** Summary of the regulatory interactions at the new portal PlantRegMap

Method	Species	Organ	Interaction
Literature mining and manual curation	1	—	1 431
ChIP-seq	1	2	9 340
Motif + TF footprint	1	4	342 005
Motif	132	—	50 850 582

**Table 3.** Resources available at the new portal PlantRegMap

Data	Tool
Regulatory elements	Binding site prediction
Regulatory interactions	Regulation prediction
GO annotation for all genes	GO enrichment
OGs and RBHs for all genes	TF enrichment

**Novel online tools for transcriptional regulation prediction and analyses**

For users to take full advantage of our rich data and pipelines, we upgraded the TF prediction server and set up four novel servers for TF identification, regulation prediction and functional enrichment analyses (see Supplementary Text 3 for more details).

- TF prediction: According to the requirement of users on TF prediction from newly identified transcripts, support for nucleic acid sequences was added. ESTScan (36) is employed to identify CDS regions of input nucleic acid sequences and translate them to protein sequences. In addition, the limitation on the size of uploaded file is relaxed to 100M based on users’ feedback.
- Binding site prediction: Based on the sets of species-specific, high-quality TF binding motifs integrated above, FIMO (37) is employed to scan for TF binding sites from the input sequences.
- Regulation prediction and enrichment analysis: Different from ‘Binding site prediction’ server, this tool focuses on inferring regulatory interactions between TFs and the input promoters, and searches for the enriched upstream regulators in the input gene set.
- GO enrichment: Based on our compiled GO annotation for 165 species generated (see Supplementary Text 1 for details on the annotation pipeline), this tool identifies the significantly enriched GO terms in the input gene set based on Fisher’s exact tests (38).
- TF enrichment: Based on our pre-computed regulatory interactions, this tool allows user to search for enriched upstream regulators in the input gene set.

**Resource availability**

For users to access our resource conveniently, we set up a new portal PlantRegMap (<http://plantregmap.cbi.pku.edu.cn/>) for PlantTFDB (<http://planttfdb.cbi.pku.edu.cn/>), with PlantTFDB focusing on TF identification and annotation and the new portal for regulation prediction and analyses (Table 3). All of data in our database can be freely accessed and downloaded online.

**FURTHER DIRECTION**

We have upgraded our PlantTFDB to version 4.0 with a new portal PlantRegMap, which provides genomic TF repertoires covering the main lineages of green plants, sets of high-quality binding motifs, genome-wide regulatory interactions and rich analysis tools. The information for TFs and transcriptional regulation are useful resources for users to explore the functional mechanisms of TFs and plant transcriptional regulatory system (24). We will continue working on this project to embrace more useful annotation, and improve the regulation prediction through introducing the conserved elements.

**SUPPLEMENTARY DATA**

Supplementary Data are available at NAR Online.

**ACKNOWLEDGEMENTS**

We thank Joint Genome Institute for the genome annotation of 13 unpublished species, AGI for six rice species and *Leersia perrieri*, THGP for *Castanea mollissima*, BTI for *Nicotiana benthamiana* and JHU for *Juglans regia*. We specially thank our users for their invaluable suggestions and comments. The main analyses of this work were performed on the HPC of Center for Life Sciences of Peking University and we thank Dr. Fangjin Chen for his help.

**FUNDING**

National Natural Science Foundation of China [1470330]; International Collaboration Program for Proteome Biological Big Data and Standard System [2014DFB30030]; State Key Laboratory of Protein and Plant Gene Research; China Postdoctoral Science Foundation Grant [2014M560017, 2015T80015 to J.J.]; Postdoctoral Fellowship at Peking-Tsinghua Center for Life Sciences (to J.J.); National Program for Support of Top-notch Young Professionals (to G.G.). Funding for open access charge: China Postdoctoral Science Foundation Grant [2015T80015 to J.J.].  
*Conflict of interest statement.* None declared.

**REFERENCES**

1. Riechmann, J.L., Heard, J., Martin, G. and Reuber, L. (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105.  
2. Yilmaz, A., Mejia-Guerra, M.K., Kurz, K., Liang, X., Welch, L. and Grotewold, E. (2011) AGRIS: the Arabidopsis gene regulatory information server, an update. *Nucleic Acids Res.*, **39**, D1118–D1122.  
3. 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.

4. 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.
5. Fredslund, J. (2008) DATFAP: a database of primers and homology alignments for transcription factors from 13 plant species. *BMC Genomics*, **9**, 140.
6. Mochida, K., Yoshida, T., Sakurai, T., Yamaguchi-Shinozaki, K., Shinozaki, K. and Tran, L.-S.P. (2013) TreeTFDB: an integrative database of the transcription factors from six economically important tree crops for functional Predictions and comparative and functional genomics. *DNA Res.*, **20**, 151–162.
7. 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.
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. Romeuf, I., Tessier, D., Dardevet, M., Branlard, G., Charmet, G. and Ravel, C. (2010) wDBTF: an integrated database resource for studying wheat transcription factor families. *BMC Genomics*, **11**, 185.
11. 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.
12. 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.
13. 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.
14. 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.
15. Zhang, H., Jin, J., Tang, L., Zhao, Y., Gu, X., Gao, G. and Luo, J. (2011) PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. *Nucleic Acids Res.*, **39**, D1114–D1117.
16. Jin, J., Zhang, H., Kong, L., Gao, G. and Luo, J. (2014) PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res.*, **42**, D1182–D1187.
17. Jin, J., Guo, A., He, K., Zhang, H., Zhu, Q., Chen, X., Gao, G. and Luo, J. (2015) Classification, prediction and database construction of plant transcription factors. *Biotechnol. Bull.*, **31**, 68–77.
18. Jia, J., Zhao, S., Kong, X., Li, Y., Zhao, G., He, W., Appels, R., Pfeifer, M., Tao, Y. and Zhang, X. (2013) Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature*, **496**, 91–95.
19. Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., Zhang, J., Saski, C.A., Scheffler, B.E., Stelly, D.M. *et al.* (2015) Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat. Biotechnol.*, **33**, 531–537.
20. O'Malley, R.C., Huang, S.S., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A. and Ecker, J.R. (2016) Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell*, **165**, 1280–1292.
21. Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.
22. Franco-Zorrilla, J.M., López-Vidriero, I., Carrasco, J.L., Godoy, M., Vera, P. and Solano, R. (2014) DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 2367–2372.
23. Sullivan, A.M., Arsovski, A.A., Lempe, J., Bubb, K.L., Weirauch, M.T., Sabo, P.J., Sandstrom, R., Thurman, R.E., Neph, S., Reynolds, A.P. *et al.* (2014) Mapping and dynamics of regulatory DNA and transcription factor networks in *A. thaliana*. *Cell Rep.*, **8**, 2015–2030.
24. Jin, J., He, K., Tang, X., Li, Z., Lv, L., Zhao, Y., Luo, J. and Gao, G. (2015) An Arabidopsis transcriptional regulatory map reveals distinct functional and evolutionary features of novel transcription factors. *Mol. Biol. Evol.*, **32**, 1767–1773.
25. Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L. and Garcia-Hernandez, M. (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.
26. Consortium, U. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
27. Brown, G.R., Hem, V., Katz, K.S., Ovetsky, M., Wallin, C., Ermolaeva, O., Tolstoy, I., Tatusova, T., Pruitt, K.D., Maglott, D.R. *et al.* (2015) Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.*, **43**, D36–D42.
28. Emms, D.M. and Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.*, **16**, 1.
29. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A. and Nuka, G. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
30. Moyroud, E., Minguet, E.G., Ott, F., Yant, L., Pose, D., Monniaux, M., Blanchet, S., Bastien, O., Thevenon, E., Weigel, D. *et al.* (2011) Prediction of regulatory interactions from genome sequences using a biophysical model for the Arabidopsis LEAFY transcription factor. *Plant Cell*, **23**, 1293–1306.
31. Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K. *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**, 83–90.
32. Mathelier, A., Fornes, O., Arenillas, D.J., Chen, C.Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R. *et al.* (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110–D115.
33. Hume, M.A., Barrera, L.A., Gisselbrecht, S.S. and Bulyk, M.L. (2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **43**, D117–D122.
34. Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M. and Hornischer, K. (2006) TRANSFAC® and its module TRANSCOMP®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
35. Machanick, P. and Bailey, T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, **27**, 1696–1697.
36. Iseli, C., Jongeneel, C.V. and Bucher, P. (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 138–148.
37. Grant, C.E., Bailey, T.L. and Noble, W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
38. Alexa, A. and Rahnenfuhrer, J. (2010) topGO: Enrichment Analysis for Gene Ontology. R package version 2.22.0.