D1182–D1187 *Nucleic Acids Research, 2014, Vol. 42, Database issue* *Published online 29 October 2013*
doi:10.1093/nar/gkt1016

PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors

Jinpu Jin, He Zhang, Lei Kong, Ge Gao* and Jingchu Luo*

State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences and Center for Bioinformatics, Peking University, Beijing 100871, P.R. China

Received September 16, 2013; Revised October 5, 2013; Accepted October 7, 2013

ABSTRACT

With the aim to provide a resource for functional and evolutionary study of plant transcription factors (TFs), we updated the plant TF database PlantTFDB to version 3.0 (http://planttfdb.cbi.pku.edu.cn). After refining the TF classification pipeline, we systematically identified 129 288 TFs from 83 species, of which 67 species have genome sequences, covering main lineages of green plants. Besides the abundant annotation provided in the previous version, we generated more annotations for identified TFs, including expression, regulation, interaction, conserved elements, phenotype information, expert-curated descriptions derived from UniProt, TAIR and NCBI GeneRIF, as well as references to provide clues for functional studies of TFs. To help identify evolutionary relationship among identified TFs, we assigned 69 450 TFs into 3924 orthologous groups, and constructed 9217 phylogenetic trees for TFs within the same families or same orthologous groups, respectively. In addition, we set up a TF prediction server in this version for users to identify TFs from their own sequences.

INTRODUCTION

Transcription factors (TFs) play key roles in plant development and stress response by temporarily and spatially regulating the transcription of their target genes. TFs are usually classified into different families based on their DNA-binding domains (DBDs). In 2000, Riechmann *et al.* (1) made the first attempt for the genome-wide analysis of TFs in *Arabidopsis thaliana* soon after the availability of its whole genome sequence. In the following years, several databases dedicated to identification and annotation of plant TFs became publicly available, either for multiple species, such as PlnTFDB (2), PlanTAPDB (3), GRASSIUS (4), LegumeTFDB (5), DATFAP (6) and TreeTFDB (7), or for individual organisms, such as AGRIS (8), RARTF (9), TOBFAC (10), SoyDB (11) and wDBTF (12). During the past 8 years, we have constructed three species-specific TF databases DATF (13), DRTF (14) and DPTF (15) for model organisms *Arabidopsis*, rice and poplar, as well as a comprehensive plant TF database (PlantTFDB) (16,17). The databases we constructed were accessed >10 million hits per year and were widely used for functional and evolutionary study of plant TFs, as well as for the prediction and annotation of TFs in newly sequenced genomes.

To meet requirements from our user community, we updated PlantTFDB to version 3.0 (http://planttfdb.cbi.pku.edu.cn/). In comparison with the previous two versions, PlantTFDB 3.0 covers more species and more TFs identified by the refined family assignment rules and improved prediction pipeline. In addition, new types of annotations were added, and phylogenetic trees and orthologous groups (OGs) were re-constructed. Finally, an online TF prediction server was set up (Table 1).

We believe that PlantTFDB 3.0 provides users with complete TF datasets, comprehensive annotations and useful analysis tools.

MATERIALS AND METHODS

Figure 1 shows the main steps in the construction of PlantTFDB 3.0, including data integration, TF classification, TF annotation and construction of orthologous groups.

Sequence data

We downloaded protein sequences of 67 species with genome sequences from the Joint Genome Institute (JGI) and several other institutions engaged in plant genome sequencing and annotation projects (Supplementary Table S1). For 16 species without genome sequences, we

*To whom correspondence should be addressed. Tel: +86 10 6275 5206; Fax: +86 10 6275 5206; Email: luojc@pku.edu.cn
Correspondence may also be addressed to Ge Gao. Tel: +86 10 6275 5206; Fax: +86 10 6275 5206; Email: gaog@mail.cbi.pku.edu.cn
Present Address:
He Zhang, Division of Cardiovascular Medicine, Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan 48109, USA.

downloaded their expressed sequence tag sequences from UniGene (18) and PlantGDB-assembled unique transcripts from PlantGDB (19), and then built reference proteome for each species (Supplementary Table S2) using a previous established pipeline (17).

Family assignment rules

TFs are usually classified into different families based on their DBDs. We used auxiliary and forbidden domains to distinguish complicated TF families with multiple signature domains. After a comprehensive literature review, we improved the family assignment rules described in the previous version (17) and arranged several families into superfamilies (Figure 2). We removed the forbidden domain Glyco_hydro_14 of the BES1 family, as recent studies demonstrated that BES1 family proteins with this domain also showed TF activity (20).

Table 1. Comparison among the three versions of PlantTFDB

PlantTFDB	Version 1.0	Version 2.0	Version 3.0
Species	22	49	83
Species with genome sequences	5	28	67
Species without genome sequences	17	21	16
TF family	64	58	58
TF number	26 402	53 574	129 288
Annotation			
Expert-curated description	No	No	Yes
Expression	Yes	Yes	Yes
Regulation	No	No	Yes
Interaction	No	No	Yes
Phenotype	No	No	Yes
Reference	Yes	Yes	Yes
Orthologous group	Yes	Yes	Yes
Phylogenetic tree			
Family	No	Yes	Yes
Orthologous group	No	No	Yes
Web service	No	Yes	No
TF prediction server	No	No	Yes

Prediction pipeline

We refined the TF prediction pipeline by updating the hidden Markov model (HMM) profiles used to identify TFs and adjusted their thresholds. We downloaded the latest version of HMM profiles from Pfam (version 27.0) (21) for most signature domains and built our own HMM profiles for the remaining domain that did not have available Pfam HMM profiles. We used HMMER 3.0 (22) to identify TFs and assigned them into different families according to the family assignment rules described earlier.

Annotation pipeline

We used a pipeline comprising several packages to annotate identified TFs. Domain structure and GO annotation were predicted by InterProScan (version 4.8) (23). Cross-links to well-known resources were assigned to the best BLAST hits with maximal e-value 1e-10. Nuclear localization signals were predicted by PredictNLS (24). Other information such as expert-curated description, expression, regulation, conserved elements and references was collected from corresponding databases. Multiple sequence alignments (MSAs) for DBDs were constructed by HMM-guided method, and MSAs for full-length protein sequences were inferred by T-coffee (version 9.03) (25). Family trees across 83 species were inferred by FastTree (version 2.1.3) (26) with 100 resamplings. Family trees within each species were inferred by MrBayes (version 3.2.1) (27) based on the Dayhoff model for 50 000 generations. The Help page (http://planttfdb.cbi.pku.edu.cn/help_info.php#tfinfo) describes more detailed information on datasets and parameter settings.

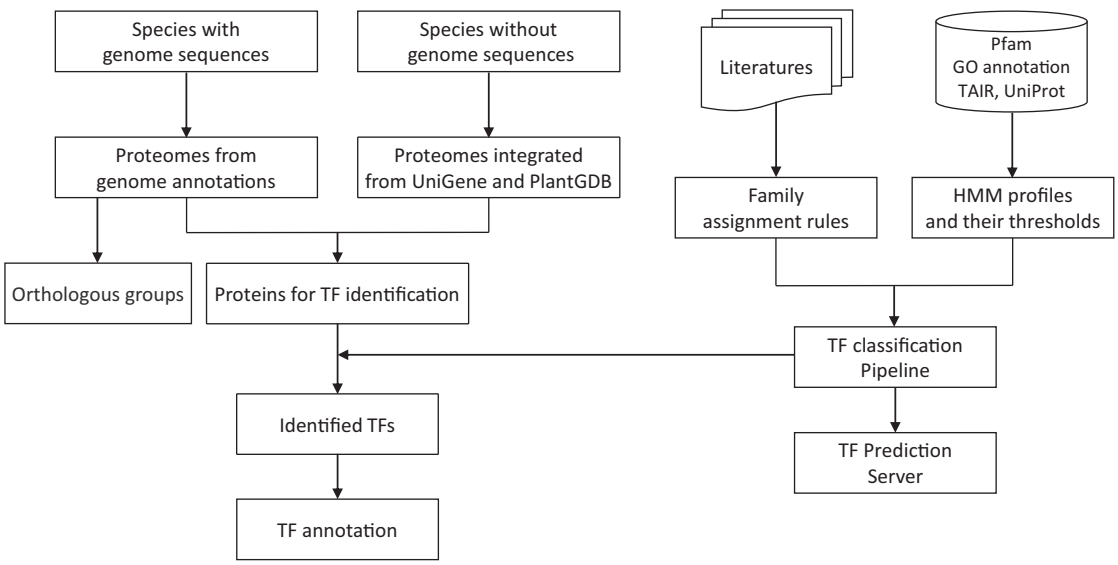

Figure 1. The flowchart for construction of PlantTFDB 3.0.

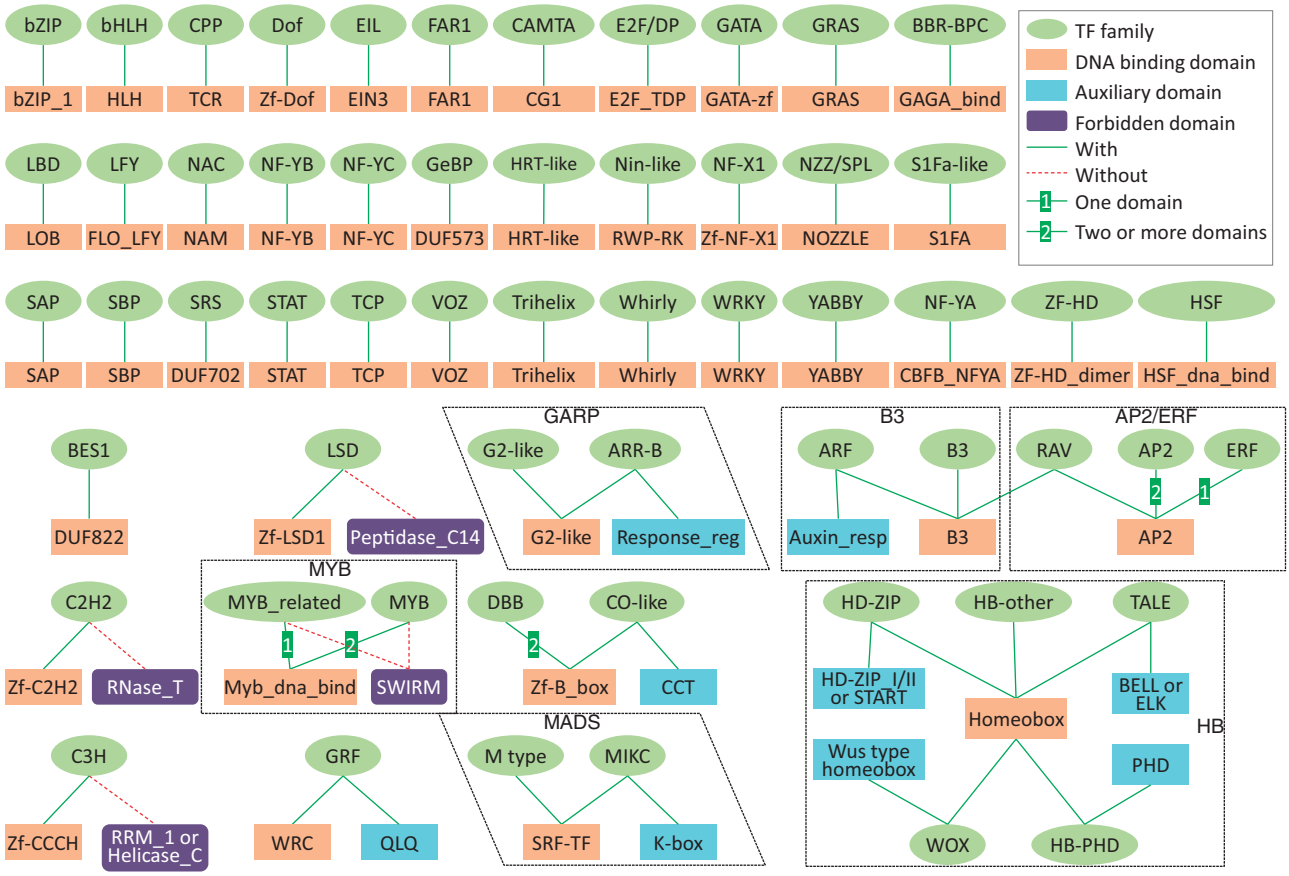

Figure 2. Refined family assignment rules used for TF identification and assignment. Green ellipses represent TF families and red rectangles represent DBDs. Blue rectangles denote auxiliary domains and purple rectangles denote forbidden domains. Green solid lines link families and DBDs or auxiliary domains and number '1' or '2' indicates number of DBDs. Red dash lines link families and forbidden domains. Families belonging to the same superfamily are arranged within rectangles or rhombi.

Orthologous groups

Orthologous groups were inferred using the following methods implemented as a pipeline of Plaza (Figure 3) (28).

First, we selected a representative gene model for each locus from 67 species with genome sequences and filtered out proteins if their lengths were <50 aa. Then we classified these proteins into clusters by TribeMCL (29). After that, proteins within the same cluster were assigned into orthologous groups by OrthoMCL (30). For TFs in the same orthologous group, MSAs were constructed by T-coffee and phylogenetic trees were inferred by MrBayes (27) with the same parameters described earlier.

RESULTS AND DISCUSSION

Genomic TF repertoires of green plants

Using the refined TF prediction pipeline, we identified 129 288 TFs (116 585 loci) from 2 691 496 proteins (2 437 666 loci) of 83 species (Table 2, Supplementary Tables S3 and S4).

The increased number of species with genome sequences and the availability of a conifer genome (31) gave us the chance to show the genomic TF repertoires across green plants for the first time (Table 2, Supplementary Table S3). Compared with green alga, land plants have a large increase in the number of TF families, TFs and percentage of TFs in their genome, which might correlate with morphological complexity of land plants (32).

Comprehensive annotations for TFs

A database of well-annotated TFs may provide users with rich information as well as insightful clues for further study. In an attempt to construct a comprehensive knowledgebase for plant TFs, we collected expert-curated description, expression, regulation, mutation and phenotype data from various public resources and made annotations for identified TFs in PlantTFDB 3.0 (Table 3), in addition to abundant annotations provided in the previous two versions (16,17). By integrating information from Entrez Gene (33), UniProtKB (34), GeneRIF (33) and mined by ourselves, we added related references for TFs.

Evolutionary conserved elements may work as transcriptional regulatory elements (35,36). Therefore, we collected these elements, which were identified based on the genome alignments of 9 crucifers (36) and 20 angiosperm plants (37), and added them into the current version, in

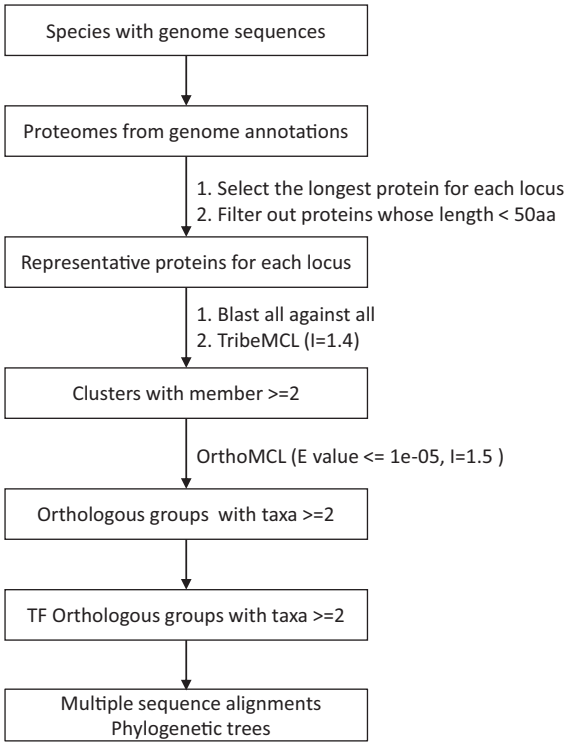

Figure 3. The pipeline for construction of orthologous groups.

Table 2. Average number of TFs in different taxonomic lineages summarized from 67 species with genome sequences

Lineage	Species	Gene	TF (%)	Family
Chlorophyta	10	10 550	141 (1.34)	35
Bryophyta[a]	1	32 273	1079 (3.34)	53
Lycopodiophyta[b]	1	22 271	665 (2.99)	54
Coniferophyta[c]	1	71 158	1851 (2.60)	55
Basal Magnoliophyta[d]	1	26 846	900 (3.35)	58
Monocot	15	34 017	1701 (5.00)	58
Eudicot	38	34 798	1861 (5.35)	58

[a]*Physcomitrella patens*.
[b]*Selaginella moellendorffii*.
[c]*Picea abies*.
[d]*Amborella trichopoda*.

addition to functional genomic annotations described earlier.

Orthologs usually have similar function and are widely used to explore functions of poorly studied proteins. To help users infer the functions of poorly studied TFs, we constructed MSAs and phylogenetic trees within the same family across 83 species, based on conserved DBDs. We further assigned 69 450 TFs into 3924 orthologous groups and constructed phylogenetic trees for each orthologous group. As an aid to decipher their evolutionary relationships, we also built trees for individual TF families within the same species. Hyperlinks to TF pages were added in the tree branches so that the users could browse them conveniently. The MSAs and phylogenetic trees in PlantTFDB 3.0 can be freely downloaded for further analyses. Direct links to TFs of *A. thaliana*, the best-studied model plant and the best-annotated species in PlantTFDB 3.0, were also generated for all TFs in other species.

Table 3. Summary of annotations for TFs in PlantTFDB 3.0

Type[a]	Species	TF	Entry
Expert-curated description	**22**	**2128**	**6649**
Expression			
UniGene	44	44 862	45 239
Microarray	14	15 424	31 975
Plant ontology	**5**	**6850**	**174 162**
Regulation			
Binding site/matrix	**24**	**541**	**729**
ChIP-chip/ChIP-seq	**1**	**54**	**75**
microRNA	**1**	**28**	**43**
Hormone	**1**	**417**	**803**
Interaction	**10**	**992**	**3101**
Conserved element	**2**	**3709**	**63 859**
Phenotype	**2**	**4704**	**147 684**
Reference	59	5004	20 255

[a]New types of annotations in this version are marked in bold.

TF prediction server

In recent years, the TF classification rules we constructed have been widely used to annotate TFs of newly sequenced genomes (38,39). In this regard, we set up a TF prediction server (http://planttfdb.cbi.pku.edu.cn/prediction.php) for users to identify TFs from their own protein sequences. As *A. thaliana* is the best-annotated species in PlantTFDB 3.0, links to the best hits in *A. thaliana* are provided for predicted TFs. Currently, users can upload up to 100 sequences and obtain results within a minute from our server.

Further direction

We have updated our PlantTFDB to version 3.0, which provides TF repertoires across the main lineages of green plants. The knowledge we collected, the OGs and phylogenetic trees we inferred are useful resources for further exploration of the physiological function and evolutionary relationship of TFs. We will continue to work on this project to refine the family assignment rules and the prediction pipeline, and collect more type of useful information for identified TFs in the future.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Joint Genome Institute for the genome annotation of four unpublished species, AGD for *Amborella trichopoda*, ICGC for *Citrus clementina* and AGI for three rice species. They also thank their users for their suggestions and comments. They specially thank Ying Dillaha for her language editing of the manuscript.

FUNDING

National Natural Science Foundation of China [31071160, 31171242]; China High-Tech Program [2006AA02Z334, 2012AA020409]; China National Key Basic Research Program [2011CBA01102]; China National Outstanding Youth Talents Program; China National Science and Technology Infrastructure Program [2009FY120100]. Funding for open access charge: State Key Laboratory of Protein and Plant Gene Research.

Conflict of interest statement. None declared.

REFERENCES

1. Riechmann,J., Heard,J., Martin,G., Reuber,L., Keddie,J., Adam,L., Pineda,O., Ratcliffe,O., Samaha,R. and Creelman,R. (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.
2. 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.
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. 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.
6. Fredslund,J. (2008) DATFAP: a database of primers and homology alignments for transcription factors from 13 plant species. *BMC Genomics*, **9**, 140.
7. 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.
8. 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.
9. 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.
10. 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.
11. Wang,Z., Libault,M., Joshi,T., Valliyodan,B., Nguyen,H., Xu,D., Stacey,G. and Cheng,J. (2010) SoyDB: a knowledge database of soybean transcription factors. *BMC Plant Biol.*, **10**, 14.
12. 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.
13. 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.
14. 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.
15. 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.
16. 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.
17. 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.
18. Acland,A., Agarwala,R., Barrett,T., Beck,J., Benson,D.A., Bollin,C., Bolton,E., Bryant,S.H., Canese,K. and Church,D.M. (2013) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **41**, D8–D20.
19. 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.
20. Reinhold,H., Soyk,S., Šimková,K., Hostettler,C., Marafino,J., Mainiero,S., Vaughan,C.K., Monroe,J.D. and Zeeman,S.C. (2011) β-Amylase–Like Proteins Function as Transcription Factors in Arabidopsis, Controlling Shoot Growth and Development. *Plant Cell*, **23**, 1391–1403.
21. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G. and Clements,J. (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
22. Eddy,S. (2010) HMMER User's Guide: Biological sequence analysis using profile hidden Markov models (http://hmmer.org) (18 October 2013, date last accessed).
23. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.
24. Cokol,M., Nair,R. and Rost,B. (2000) Finding nuclear localization signals. *EMBO Rep.*, **1**, 411–415.
25. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
26. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
27. Ronquist,F. and Huelsenbeck,J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.
28. Van Bel,M., Proost,S., Wischnitzki,E., Movahedi,S., Scheerlinck,C., Van de Peer,Y. and Vandepoele,K. (2012) Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiol.*, **158**, 590–600.
29. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
30. Li,L., Stoeckert,C.J. and Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.
31. Nystedt,B., Street,N.R., Wetterbom,A., Zuccolo,A., Lin,Y.-C., Scofield,D.G., Vezzi,F., Delhomme,N., Giacomello,S. and Alexeyenko,A. (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature*, **497**, 579–584.
32. Lang,D., Weiche,B., Timmerhaus,G., Richardt,S., Riaño-Pachón,D.M., Corrêa,L.G., Reski,R., Mueller-Roeber,B. and Rensing,S.A. (2010) Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. *Genome Biol. Evol.*, **2**, 488–503.
33. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.
34. Consortium,U. (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.*, **41**, D43–D47.
35. Baxter,L., Jironkin,A., Hickman,R., Moore,J., Barrington,C., Krusche,P., Dyer,N.P., Buchanan-Wollaston,V., Tiskin,A. and Beynon,J. (2012) Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. *Plant Cell*, **24**, 3949–3965.
36. Haudry,A., Platts,A.E., Vello,E., Hoen,D.R., Leclercq,M., Williamson,R.J., Forczek,E., Joly-Lopez,Z., Steffen,J.G. and

Hazzouri,K.M. (2013) An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat. Genet.*, **45**, 891–898.
37. Hupalo,D. and Kern,A.D. (2013) Conservation and functional element discovery in 20 angiosperm plant genomes. *Mol. Biol. Evol.*, **30**, 1729–1744.
38. Shulaev,V., Sargent,D.J., Crowhurst,R.N., Mockler,T.C., Folkerts,O., Delcher,A.L., Jaiswal,P., Mockaitis,K., Liston,A., Mane,S.P. *et al.* (2011) The genome of woodland strawberry (Fragaria vesca). *Nat. Genet.*, **43**, 109–116.
39. Jia,J., Zhao,S., Kong,X., Li,Y., Zhao,G., He,W., Appels,R., Pfeifer,M., Tao,Y., Zhang,X. *et al.* (2013) Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature*, **496**, 91–95.