SCIENTIFIC REPORTS

OPEN

An integrated database of wood-formation related genes in plants

Ting Xu, Tao Ma, Quanjun Hu & Jianquan Liu

Received: 09 January 2015
Accepted: 19 May 2015
Published: 16 June 2015

Wood, which consists mainly of plant cell walls, is an extremely important resource in daily lives. Genes whose products participate in the processes of cell wall and wood formation are therefore major subjects of plant science research. The Wood-Formation Related Genes database (WFRGdb, http://me.lzu.edu.cn/woodformation/) serves as a data resource center for genes involved in wood formation. To create this database, we collected plant genome data published in other online databases and predicted all cell wall and wood formation related genes using BLAST and HMMER. To date, 47 gene families and 33 transcription factors from 57 genomes (28 herbaceous, 22 woody and 7 non-vascular plants) have been covered and more than 122,000 genes have been checked and recorded. To provide easy access to these data, we have developed several search methods, which make it easy to download targeted genes or groups of genes free of charge in FASTA format. Sequence and phylogenetic analyses are also available online. WFRGdb brings together cell wall and wood formation related genes from all available plant genomes, and provides an integrative platform for gene inquiry, downloading and analysis. This database will therefore be extremely useful for those who focuses on cell wall and wood research.

Plant cells are encased by complex polysaccharide walls, which have diverse functions. These walls constitute the main component of wood, which has served as fuel for fires and been exploited for numerous other uses since human civilization began. Genetic analyses of the formation of plant cell walls have provided the basis for much of the current understanding of cell walls, including how walls are made, how their development is regulated, and how they function. At present, around 800 of the genes in the *Arabidopsis* genome are believed to be related to the formation of cell walls. Several databases based mainly on *Arabidopsis* genes have been constructed, including the Cell Wall Genomics database (http://cellwall.genomics.purdue.edu) and Cell Wall Navigator (http://cellwall.ucr.edu/Cellwall/), which brings together cell-wall related genes from *Arabidopsis*, rice and maize[1–3]. CAZy (http://www.cazy.org/) is another such database; it focuses on the genes encoding proteins that catalyze the synthesis of carbohydrates and glycoconjugates[4]. In recent years, numerous plant genomes, including those of some trees, have been published, but cell wall synthesis related genes are not covered in most of these genomes on those databases. In this study, we developed an integrated database of Wood-Formation Related Genes (WFRGs) from all plant species whose genomes are available. This database will provide a comprehensive and robust platform allowing researchers focusing on plant cells and wood to index, BLAST and determine the phylogenetic relationships of their genes of interest that are related to wood formation.

Result

Data organization. The plants with available genome sequences were classified into three groups on the basis of life history:

1. Herbaceous species with no obvious woody stems.
2. Woody species with woody stems.
3. Non-vascular plants, including mosses and algae.

State Key Laboratory of Grassland and Agro-Ecosystems, School of Life Sciences, Lanzhou University, Lanzhou 730000, Gansu, China. Correspondence and requests for materials should be addressed to J.L. (email: liujq@lzu.edu.cn)

In order to make the user interface more friendly, we abbreviated the species name in our database by using only the first letter of the genus to represent the genus (for example, Athaliana for *Arabidopsis thaliana*; see Table S1 for more details).

All genes were classified into 8 broad types according to function:

1. Cellulose and hemicellulose synthesis, comprising genes that encode proteins synthesizing cellulose and hemicellulose, the main components of plant cell walls.
2. Lignin synthesis, including genes that encode enzymes catalyzing the monolignol biosynthetic pathway and monolignol assembly.
3. Esterases, comprising genes that encode enzymes hydrolyzing esters, chemical compounds that contain a carbonyl group adjacent to an ether linkage.
4. Monosaccharide inter-conversion, including genes that encode enzymes catalyzing inter-conversions between nucleotide-diphospho-sugars (NDP-sugars, fundamental components of diverse polysaccharides and glycoconjugates).
5. Lyases, comprising genes that encode pectin/rhamnogalacturonan lyases.
6. Cell wall structural proteins, including genes that encode proteins playing important roles in plant cell wall structure.
7. Cell growth and other wood-formation related genes.
8. Transcription factors.

A total of 47 gene families/super families and 33 transcription factors were included in our database (see Table S2).

Data access and utility of the database. Users gain access to the data in WFRGdb via a search. The Search function in our database is divided into two parts: BLAST Search and Main search.

The user interface for BLAST Search is similar to that at NCBI. After users have submitted their FASTA sequences or FASTA files, the database will return the results as a table and the result sequences are made available for download.

The Main search is made up of three parts: Information Search, Gene Search and Fast Search. Gene families, gene names (obtained by searching for a sequence via a gene name), related references and information about genomes can be accessed via Information Search. Gene Search and Fast Search return similar results. Gene Search has a more complex user interface allowing users to view details of gene families and genomes, while Fast Search has a relatively simple user interface and delivers results in a condensed format which is especially suitable for searching through a large amount of data.

To carry out an Information Search, users should choose an option and enter the term for which they want to search into the text box. For Gene Search and Fast Search, users should tick to select at least one gene family and one genome.

In the results of Information Search, the keywords are highlighted in red to make the results easier to read. The results of Gene Search and Fast Search are displayed in the form of a multifunctional table which supports paging and sorting. Clicking a gene name will open a detailed information box from which users can download the gene's sequence. Similarly, when a gene family name is clicked, detailed information about the family will pop up in an information box.

To the left of the gene name is a row of checkboxes; users can check these to download all selected gene sequences in a FASTA file or use all the checked gene sequences for further analysis.

The "Go Analysis" button on the bottom right of the page will take users to the analysis page. By following the instructions on this page, users can complete sequence analyses step by step. Both sequence and phylogenetic analyses are available on this page and all related files can also be downloaded. Alignment is done by Clustal W and maximum likelihood trees are built by FastTree[5,6]. Sequence analysis and tree analysis are implemented in JalView, which requires Java support (see Fig. 1)[7]. Users therefore need to install Java on their computers in advance. Our database also provides a simple function called jsPhyloSVG, which is Java-independent, for viewing tree files online[8]. As it may take some time to finish sequence analyses, we provide an e-mail service. If an e-mail address is supplied, the results (including gene sequences in FASTA format, aligned sequence files and tree files in Newick format) will be sent to users once the work has been done in the background. A schematic overview of information flow in WFRGdb is shown in Fig. 2.

Discussion

WFRGdb uses all plant genomes whose sequences have been released to date to search out all known genes involved in wood formation. It is designed to assist researchers in finding and identifying all genes orthologous to their targets related to plant cell wall and wood formation, and in constructing their phylogenetic relationships. In the case of a gene family, researchers can obtain all sequences that belong to this family, and a phylogenetic tree for the family is also available. General information about the family is also easily accessible with the help of the references recommended in the database. We believe that WFRGdb will be very useful for those focusing on cell wall and wood research.

To our knowledge, WFRGdb is the first comprehensive resource database related to cell-wall and wood-formation related genes based on the mass of genome data now available. Here we present a

Figure 1. Data analysis page from WFRGdb. The user interface for analysis work is shown in Fig. 1a. Figure 1b shows the sequence analysis window (top) and tree analysis page (bottom).

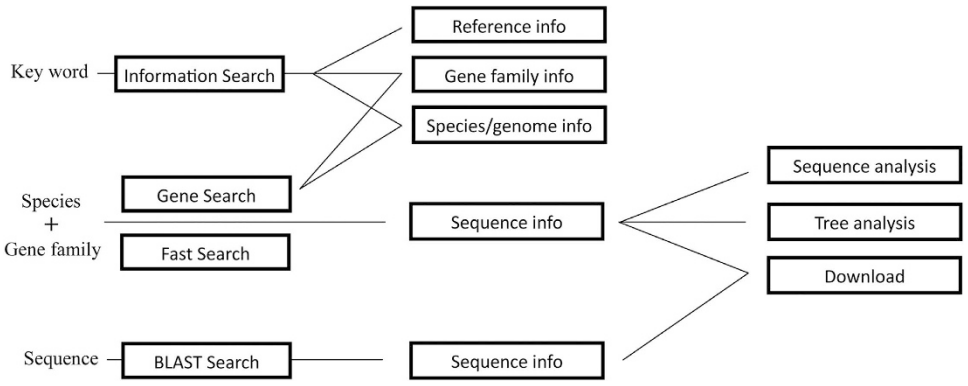

Figure 2. A schematic overview of WFRGdb. Information Search requires key words as input, while Gene Search and Fast Search need Species and Gene family as input information and BLAST Search needs sequence data.

collection of putative genes involved in wood-formation and display them in the form of gene families/super families and according to species/genomes, as well as providing easy access for data downloading and sequence analysis.

Method

Data sources. To date, 57 genomes for 28 herbs, 22 trees and 7 non-vascular species have been published (Fig. 3 and Table S1)[9–63]. Genome data were obtained mainly from Phytozome (http://www.phytozome.net, a joint project of the Department of Energy's Joint Genome Institute and the Center for integrative Genomics to facilitate comparative genomic studies amongst green plants) or from dedicated genome websites for individual targeted species.

Gene prediction. After downloading the genome data, for each gene where the GFF file indicated the existence of alternatively spliced transcripts, we discarded all but the longest such transcript. The proteins encoded by all the downloaded gene sequences were entered into a BLAST protein database[64]. We collected the sequences of all known Arabidopsis members of 47 gene families related to cell wall and wood

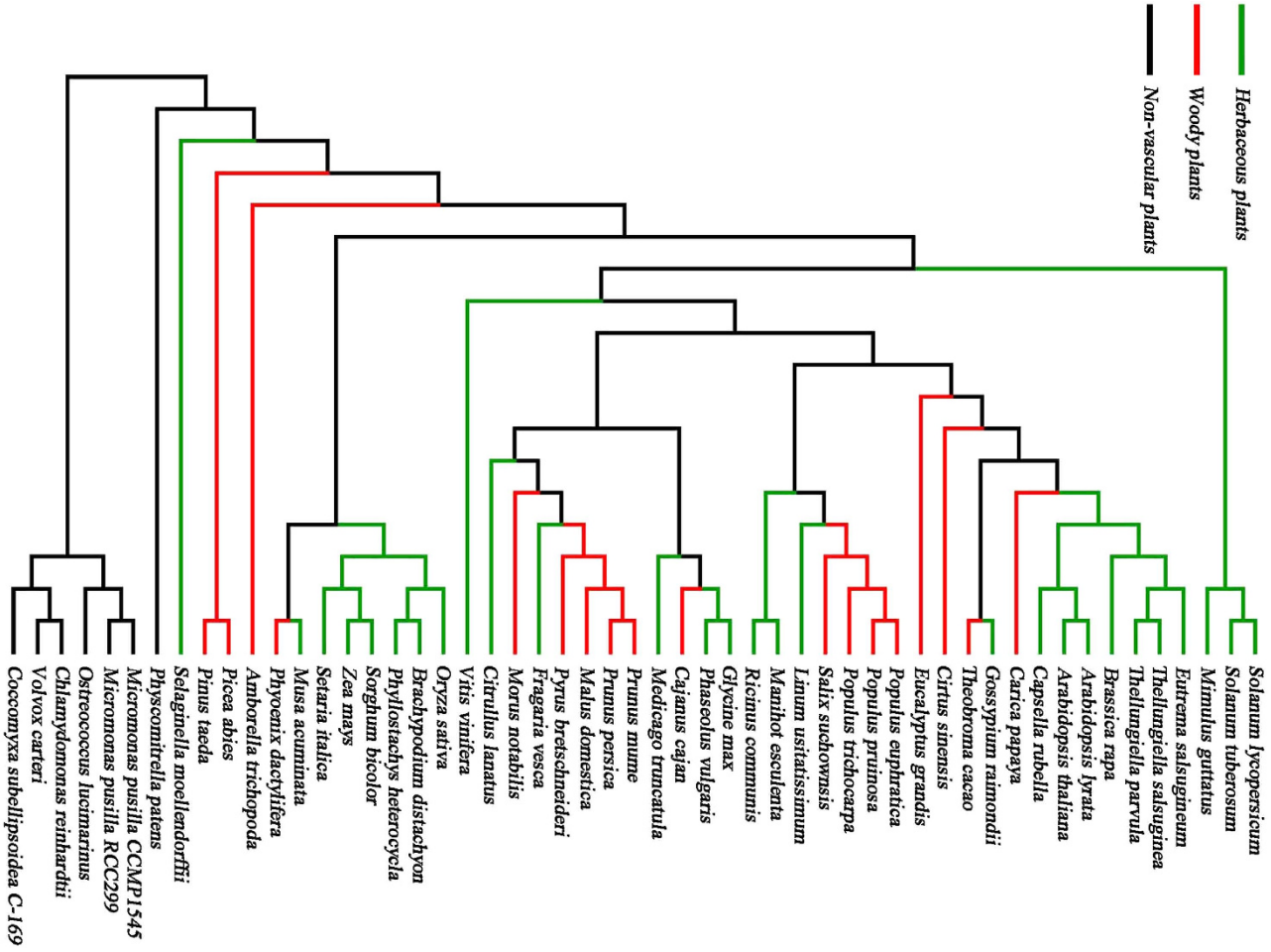

Figure 3. Phylogenetic tree showing all the species recorded in WFRGdb. Herbaceous plants are indicated in green and woody plants are in red. For some species (e.g. *Gossypium raimondii*), two or more genome sequences have been published; in such cases, all available sequences are covered in WFRGdb (see Table S1).

synthesis from the internet, 33 transcription factors from Plant Transcription Factor Database(PTFB)[65], and used them as the initial query in a BLAST search against our protein database. All hits obtained in this search were flagged as candidate genes. We examined each of these candidate genes in order to ensure that it belonged to the ascribed family. To do this, we ran each of these candidate protein sequences against the protein database again and examined the top 10 non-self hits for each gene in the resulting list. A candidate gene was removed if two or more of the top 10 non-self hits were not members of the 47 gene families.

The candidates retained after this analysis were then tested further using HMMER to ensure that each shared the domain /domains of the gene family to which it belonged[66]. The domain information for the gene families was derived from PFAM (http://pfam.sanger.ac.uk), a database of protein families that are represented by multiple sequences generated using hidden Markov models. The candidates that passed the HMMER tests were retained. Finally, the coding DNA sequence (CDS) of each gene was extracted from the CDS section of its GFF file by an in-house Perl script.

References

1. Girke, T., Lauricha, J., Tran, H., Keegstra, K. & Raikhel, N. The Cell Wall Navigator database. A systems-based approach to organism-unrestricted mining of protein families involved in cell wall metabolism. *Plant Physiol* **136,** 3003–8; discussion 3001 (2004).
2. Penning, B. W. *et al.* Genetic resources for maize cell wall biology. *Plant Physiol* **151,** 1703–28 (2009).
3. Purdue University. *Cell Wall Related Gene Families.Cell Wall Genomics*. (2012) Available at:http://cellwall.genomics.purdue.edu/families/index.html. (Accessed: 10th March 2013).
4. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* **42,** D490–5 (2014).
5. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–8 (2007).
6. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* **5,** e9490 (2010).

7. Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25,** 1189–91 (2009).
8. Smits, S. A. & Ouverney, C. C. jsPhyloSVG: a javascript library for visualizing interactive and vector-based phylogenetic trees on the web. *PLoS One* **5,** e12267 (2010).
9. Al-Dous, E. K. *et al.* De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera). *Nat Biotechnol* **29,** 521–7 (2011).
10. Amborella Genome Project. The Amborella genome and the evolution of flowering plants. *Science* **342,** 1241089 (2013).
11. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408,** 796–815 (2000).
12. Argout, X. *et al.* The genome of Theobroma cacao. *Nat Genet* **43,** 101–8 (2011).
13. Banks, J. A. *et al.* The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. *Science* **332,** 960–3 (2011).
14. Bennetzen, J. L. *et al.* Reference genome sequence of the model plant Setaria. *Nat Biotechnol* **30,** 555–61 (2012).
15. Blanc, G. *et al.* The genome of the polar eukaryotic microalga Coccomyxa subellipsoidea reveals traits of cold adaptation. *Genome Biol* **13,** R39 (2012).
16. Chan, A. P. *et al.* Draft genome sequence of the oilseed species Ricinus communis. *Nat Biotechnol* **28,** 951–6 (2010).
17. Dai, X. *et al.* The willow genome and divergent evolution from poplar after the common genome duplication. *Cell Res* **24,** 1274–7 (2014).
18. Dassanayake, M. *et al.* The genome of the extremophile crucifer Thellungiella parvula. *Nat Genet* **43,** 913–8 (2011).
19. D'Hont, A. *et al.* The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* **488,** 213–7 (2012).
20. Guo, S. *et al.* The draft genome of watermelon (Citrullus lanatus) and resequencing of 20 diverse accessions. *Nat Genet* **45,** 51–8 (2013).
21. He, N. *et al.* Draft genome sequence of the mulberry tree Morus notabilis. *Nat Commun* **4,** 2445 (2013).
22. Hellsten, U. *et al.* Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing. *Proc Natl Acad Sci U S A* **110,** 19478–82 (2013).
23. Hu, T. T. *et al.* The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat Genet* **43,** 476–81 (2011).
24. International Brachypodium Initiative. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* **463,** 763–8 (2010).
25. International Peach Genome Initiative. *et al.* The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat Genet* **45,** 487–94 (2013).
26. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436,** 793–800 (2005).
27. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449,** 463–7 (2007).
28. Ma, T. *et al.* Genomic insights into salt adaptation in a desert poplar. *Nat Commun* **4,** 2797 (2013).
29. Ming, R. *et al.* The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature* **452,** 991–6 (2008).
30. Motamayor, J. C. *et al.* The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biol* **14,** r53 (2013).
31. Myburg, A. A. *et al.* The genome of Eucalyptus grandis. *Nature* **510,** 356–62 (2014).
32. Neale, D. B. *et al.* Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol* **15,** R59 (2014).
33. Nystedt, B. *et al.* The Norway spruce genome sequence and conifer genome evolution. *Nature* **497,** 579–84 (2013).
34. Palenik, B. *et al.* The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proc Natl Acad Sci U S A* **104,** 7705–10 (2007).
35. Paterson, A. H. *et al.* The Sorghum bicolor genome and the diversification of grasses. *Nature* **457,** 551–6 (2009).
36. Paterson, A. H. *et al.* Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. *Nature* **492,** 423–7 (2012).
37. Peng, Z. *et al.* The draft genome of the fast-growing non-timber forest species moso bamboo (Phyllostachys heterocycla). *Nat Genet* **45,** 456-61, 461e1–2 (2013).
38. Potato Genome Sequencing Consortium. *et al.* Genome sequence and analysis of the tuber crop potato. *Nature* **475,** 189–95 (2011).
39. Prochnik, S. *et al.* The Cassava Genome: Current Progress, Future Directions. *Trop Plant Biol* **5,** 88–94 (2012).
40. Prochnik, S. E. *et al.* Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. *Science* **329,** 223–6 (2010).
41. Rensing, S. A. *et al.* The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science* **319,** 64–9 (2008).
42. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463,** 178–83 (2010).
43. Schmutz, J. *et al.* A reference genome for common bean and genome-wide analysis of dual domestications. *Nat Genet* **46,** 707–13 (2014).
44. Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326,** 1112–5 (2009).
45. Shrager, J. *et al.* Chlamydomonas reinhardtii genome project. A guide to the generation and use of the cDNA information. *Plant Physiology* **131,** 401–408 (2003).
46. Shulaev, V. *et al.* The genome of woodland strawberry (Fragaria vesca). *Nat Genet* **43,** 109–16 (2011).
47. Slotte, T. *et al.* The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat Genet* **45,** 831–5 (2013).
48. Sweet Orange Genome Project. *Citrus sinensis.Phytozome v9.1*. (2010) Available at: http://www.phytozome.net/citrus.php. (Accessed: 10th March 2013).
49. Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485,** 635–41 (2012).
50. Tuskan, G. A. *et al.* The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **313,** 1596–604 (2006).
51. Varshney, R. K. *et al.* Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. *Nat Biotechnol* **30,** 83–9 (2012).
52. Velasco, R. *et al.* The genome of the domesticated apple (Malus x domestica Borkh.). *Nat Genet* **42,** 833–9 (2010).
53. Wang, K. *et al.* The draft genome of a diploid cotton Gossypium raimondii. *Nat Genet* **44,** 1098–103 (2012).
54. Wang, X. *et al.* The genome of the mesopolyploid crop species Brassica rapa. *Nat Genet* **43,** 1035–9 (2011).
55. Wang, Z. *et al.* The genome of flax (Linum usitatissimum) assembled de novo from short shotgun sequence reads. *Plant J* **72,** 461–73 (2012).
56. Worden, A. Z. *et al.* Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. *Science* **324,** 268–72 (2009).

57. Wu, H. J. *et al.* Insights into salt tolerance from the genome of Thellungiella salsuginea. *Proc Natl Acad Sci U S A* **109,** 12219–24 (2012).
58. Wu, J. *et al.* The genome of the pear (Pyrus bretschneideri Rehd.). *Genome Res* **23,** 396–408 (2013).
59. Xu, Q. *et al.* The draft genome of sweet orange (Citrus sinensis). *Nat Genet* **45,** 59–66 (2013).
60. Yang, R. *et al.* The Reference Genome of the Halophytic Plant Eutrema salsugineum. *Front Plant Sci* **4,** 46 (2013).
61. Young, N. D. *et al.* The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* **480,** 520–4 (2011).
62. Zhang, G. *et al.* Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. *Nat Biotechnol* **30,** 549–54 (2012).
63. Zhang, Q. *et al.* The genome of Prunus mume. *Nat Commun* **3,** 1318 (2012).
64. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215,** 403–10 (1990).
65. Perez-Rodriguez, P. *et al.* PlnTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Res* **38,** D822–7 (2010).
66. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39,** W29–37 (2011).

Acknowledgments

This work is supported by the National Key Project for Basic Research (2012CB114504), National High Technology Research and Development Program of China (863 Program, No. 2013AA100605) and the international collaboration '111' project.

Author Contributions

J.Q.L. designed the project; T.X. and T.M. constructed the database; T.X. wrote the manuscript, and T.M., Q.J.H. and J.Q.L. revised the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Xu, T. *et al.* An integrated database of wood-formation related genes in plants. *Sci. Rep.* **5**, 11422; doi: 10.1038/srep11422 (2015).