nature genetics

OPEN

The draft genome of the fast-growing non-timber forest species moso bamboo (*Phyllostachys heterocycla*)

Zhenhua Peng[1,4], Ying Lu[2,4], Lubin Li[1,4], Qiang Zhao[2,4], Qi Feng[2,4], Zhimin Gao[3,4], Hengyun Lu[2], Tao Hu[3], Na Yao[1], Kunyan Liu[2], Yan Li[2], Danlin Fan[2], Yunli Guo[2], Wenjun Li[2], Yiqi Lu[2], Qijun Weng[2], CongCong Zhou[2], Lei Zhang[2], Tao Huang[2], Yan Zhao[2], Chuanrang Zhu[2], Xinge Liu[3], Xuewen Yang[3], Tao Wang[1], Kun Miao[1], Caiyun Zhuang[1], Xiaolu Cao[1], Wenli Tang[3], Guanshui Liu[3], Yingli Liu[3], Jie Chen[1], Zhenjing Liu[1], Licai Yuan[3], Zhenhua Liu[1], Xuehui Huang[2], Tingting Lu[2], Benhua Fei[3], Zemin Ning[2], Bin Han[2] & Zehui Jiang[1,3]

© 2013 Nature America, Inc. All rights reserved.

Bamboo represents the only major lineage of grasses that is native to forests and is one of the most important non-timber forest products in the world. However, no species in the Bambusoideae subfamily has been sequenced. Here, we report a high-quality draft genome sequence of moso bamboo (*P. heterocycla var. pubescens*). The 2.05-Gb assembly covers 95% of the genomic region. Gene prediction modeling identified 31,987 genes, most of which are supported by cDNA and deep RNA sequencing data. Analyses of clustered gene families and gene collinearity show that bamboo underwent whole-genome duplication 7–12 million years ago. Identification of gene families that are key in cell wall biosynthesis suggests that the whole-genome duplication event generated more gene duplicates involved in bamboo shoot development. RNA sequencing analysis of bamboo flowering tissues suggests a potential connection between drought-responsive and flowering genes.

Bamboo is one of the most important non-timber forest products in the world. About 2.5 billion people depend economically on bamboo, and international trade in bamboo amounts to over 2.5 billion US dollars per year[1]. Bamboo has a rather striking life history, characterized by a prolonged vegetative phase lasting decades before flowering, thereby inhibiting genetic improvement. Recent genomic studies in bamboo have included genome-wide full-length cDNA sequencing[2], chloroplast genome sequencing[3], identification of syntenic genes between bamboo and other grasses[4] and phylogenetic analysis of Bambusoideae subspecies[5]. Fifty-nine simple sequence repeat markers from rice and sugarcane were used in the genetic diversity analyses of 23 bamboo species[6], and 2 species-specific sequence-characterized amplified region markers were developed in the identification of different bamboo species[7].

Here, we report the draft genome of moso bamboo, a large woody bamboo that has ecological, economic and cultural value in Asia and accounts for ~70% of the total bamboo growth area. Comparative genome-wide analyses of bamboo to other grass species, including rice, maize and sorghum, yielded new genetic insights into the rapid and marked phenotypic and ecological divergence of bamboo and closely related grasses.

The moso bamboo genome contains 24 pairs of chromosomes[8] ($2n = 48$) and is characteristic of a diploid (**Supplementary Fig. 1a**). We conducted a flow cytometry analysis and estimated that it had a genome size of 2.075 Gb ($2C = 4.24$ pg; **Supplementary Fig. 1b**), which was very close to that estimated in a previous report[9].

Because it is difficult to generate an inbred line of moso bamboo, owing to its infrequent sexual reproduction and the long periods of time between flowering intervals, we selected five plants from a single individual rhizome of the moso bamboo ecotype (*P. heterocycla var. pubescens*) and performed whole-genome shotgun sequencing. We generated 295 Gb of raw sequence data (approximately 147-fold coverage), including Illumina short reads and 10,327 pairs of BAC end sequences (**Supplementary Table 1a**). The final assembly of 2.05 Gb was generated using the *de novo* Phusion-meta assembly pipeline that was developed in this study (**Supplementary Fig. 2**). The N50 length of the assembled scaffolds was over 328 kb, and about 80% of the assembly mapped to 5,499 scaffolds of greater than 62 kb in length (**Table 1** and **Supplementary Table 1b**). The scaffolds assembled using the Phusion-meta assembly method were much longer in length than the scaffolds generated using the SOAPdenovo program[10] (**Fig. 1a** and **Supplementary Table 1c**). Given the presence of small fragments in the assembly, the estimated size of the moso bamboo genome was approximately 2.07 to 2.10 Gb, which was supported by the analysis of the distribution of 51-mer frequencies (**Supplementary Fig. 3**). Hence, the final scaffolds of 2.05 Gb and initial contigs of 1.86 Gb covered approximately 95% and 88% of the genomic region, respectively. Sequence comparison of the assembled scaffolds to existing cDNA and survey sequences in the database and eight BAC sequences individually determined through Sanger sequencing showed good agreement in genomic coverage at over

[1]Research Institute of Forestry, Chinese Academy of Forestry, Key Laboratory of Tree Breeding and Cultivation, State Forestry Administration, Beijing, China. [2]National Center for Gene Research, Shanghai Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China. [3]International Center for Bamboo and Rattan, Beijing, China. [4]These authors contributed equally to this work. Correspondence should be addressed to B.H. (bhan@ncgr.ac.cn) or Z.J. (jiangzehui@icbr.ac.cn).

Received 20 July 2012; accepted 1 February 2013; published online 24 February 2013; doi:10.1038/ng.2569

Table 1 Statistics of assembly and annotation for the moso bamboo genome

Total length†	2,051,719,643 bp
N50 length (contigs)	11,882 bp
N50 length (scaffolds)	328,698 bp
N80 length (scaffolds)	62,052 bp
Number of scaffolds (>N80 length)	5,499
Largest scaffold	4,869,017 bp
GC content	43.9%
Number of protein-coding genes	31,987
Average length of protein-coding genes	3,350 bp
Total size of transposable elements	1,210,862,930 bp
Content of transposable elements	59.0%

†Final scaffolds with less than 500 bp were excluded.

88% of the initial contigs and 98% of the scaffolds (**Supplementary Figs. 4,5** and **Supplementary Tables 2–4**). The frequencies of single-base differences and insertions and/or deletions (indels) in the alignment using BAC sequences were as low as 0.19 and 0.09 instances per kilobase, respectively, which were much lower than those determined for the SOAPdenovo assemblies (**Supplementary Fig. 6** and **Supplementary Table 5**).

Alignment of all of the reads used to build the assembly identified 2,009,487 heterozygous SNPs and 51,223 short indels (6 nucleotides in length or less) (**Supplementary Table 6**). An overall heterozygous rate of the occurrence of SNPs and short indels was estimated at approximately 1.0 polymorphism per kilobase, which was lower than that (2.6 per kilobase) of the poplar genome[11] and that (4.2 per kilobase) of the grape genome[12].

We predicted 31,987 protein-coding genes in the moso bamboo genome, with the support of RNA sequencing (RNA-seq) data (127 Gb) obtained from 7 bamboo tissues and 8,253 bamboo full-length cDNA sequences[2] (Online Methods, **Supplementary Figs. 7,8** and **Supplementary Table 7**). Most basic metabolic pathways among the grass species were compared by aligning the annotated protein sequences to the KEGG data set[13], which showed high similarity between bamboo and rice (**Supplementary Table 8**). We also annotated 1,167 tRNA (**Supplementary Table 9a**), 279 rRNA, 321 small nucleolar RNA, 173 small nuclear RNA and 225 microRNA (miRNA) genes (**Supplementary Table 9b**). A total of 241 miRNA-targeted genes were predicted by the alignment of conserved miRNAs to our gene models (**Supplementary Table 9c**). *De novo* repeat annotation showed that approximately 59% of the moso bamboo genome consists of transposable elements (Online Methods and **Supplementary Table 10a**), a proportion that was much higher than the previous estimation (23.3%) in the analysis of survey sequences[9]. The most abundant repeats were long-terminal repeat elements (LTRs), including 24.6% *Gypsy*-type LTRs and 12.3% *Copia*-type LTRs (**Supplementary Table 10b**,**c**). When we used the sequences of the eight moso bamboo BACs, we observed that 52% of the genomes consisted of transposable elements (**Supplementary Table 4**).

Comparing gene families among the four grass subfamilies, including Pooideae (*Brachypodium*), Ehrhartoideae (rice), Panicoideae (maize, sorghum and foxtail millet) and Bambusoideae (moso bamboo), and two dicots (*Arabidopsis thaliana* and the woody plant poplar), we identified 21,730 bamboo genes in 14,030 families, with 9,451 gene families shared by maize, sorghum, rice and *Brachypodium* (**Fig. 1b**). There were 492 unique gene families in bamboo, of which

© 2013 Nature America, Inc. All rights reserved.

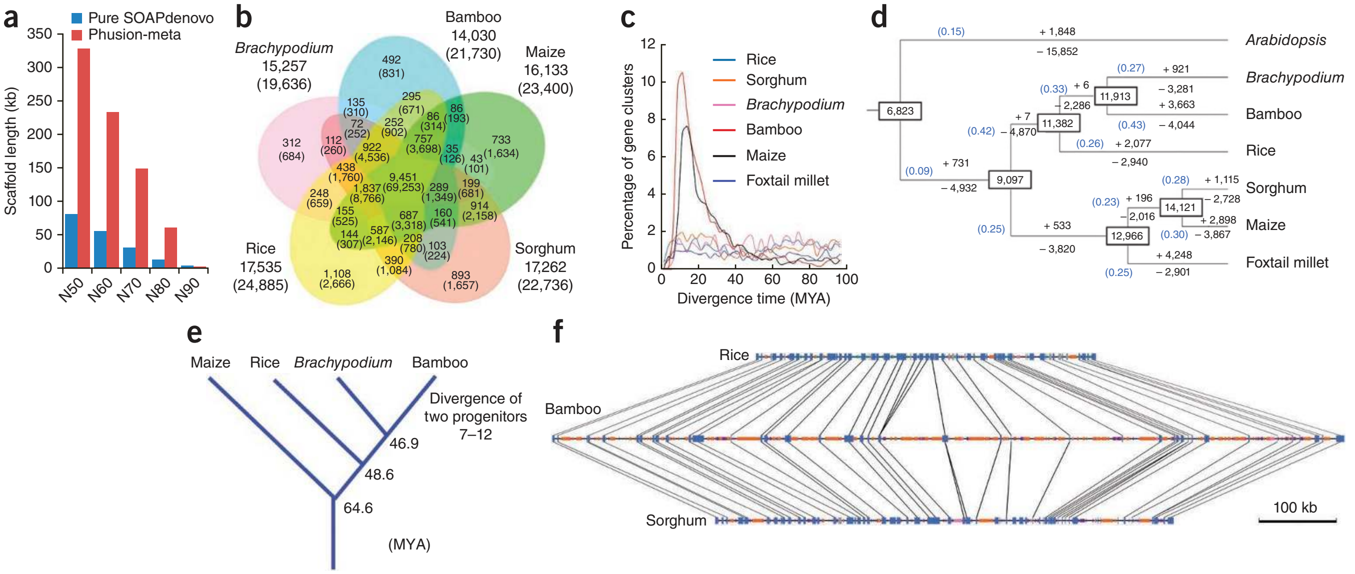

Figure 1 Assemblies and comparative genomics. (**a**) Comparison of the lengths of assembled scaffolds by the pure SOAPdenovo and Phusion-meta assembly methods. (**b**) Venn diagram of shared orthologous gene families among five grass genomes. The gene family number is listed in each component. The number of genes within the families is noted in parentheses. (**c**) Genome duplication in grass genomes. The calculated K_S values of the 2-member gene clusters were converted to divergence time, using a substitution rate of 6.5×10^{-9} mutations per site per year[34]. The *y* axis shows the percentage of the two-member gene clusters. MYA, million years ago. (**d**) Evolution of orthologous gene clusters. The black numbers above and below each branch indicate the quantity of expanded (+) or contracted (–) orthologous clusters after the corresponding speciation, respectively. The estimated numbers of clusters in the common ancestors are indicated in the rectangles. The dN/dS ratio of each branch is shown in blue. (**e**) Divergence time between bamboo and grass species from different subfamilies (mean K_s values are given in **Supplementary Table 11**). (**f**) Gene synteny between rice, sorghum and moso bamboo. The collinear region is located on rice chromosome 1 (40,565 to 40,983 kb; MSU RGAP 6.1; ref. 35), sorghum chromosome 3 (71,771 to 72,334 kb; ref. 36) and bamboo scaffold PH01000002 (1,890 to 2,862 kb). Non-hypothetical gene (blue), hypothetical genes (gray), LTR retrotransposons (orange), DNA transposons (purple), miniature inverted-repeat transposable elements (MITEs) (green) and other transposable elements (pink) are represented by boxes. Syntenic loci are connected by gray lines between the genomes.

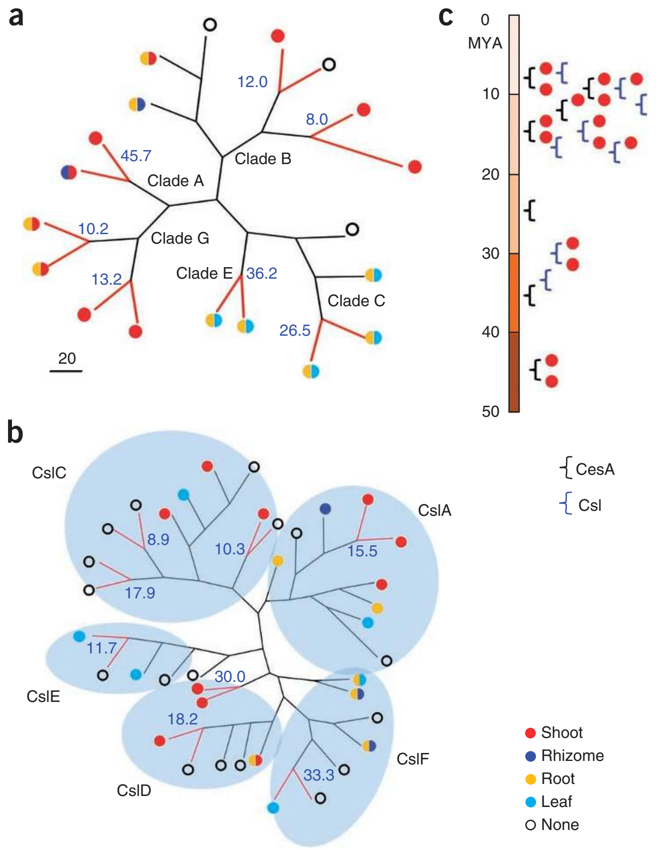

Figure 2 Recent duplication and the expression of bamboo CesA and Csl genes. (**a**) Phylogenetic neighbor-joining tree of the CesA genes. Red branches indicate a recent duplication after speciation. Filled circles indicate the tissues where the gene had high expression. Clades A, B, C, E and G correspond to the phylogenetic tree in **Supplementary Figure 12a**. The divergence time of the corresponding duplication is shown in blue. The scale bar represents the bootstrap percentage of each branch. (**b**) Phylogenic tree of the Csl genes. The clustered CslA, CslC, CslD, CslE and CslF genes were derived from the phylogenic tree in **Supplementary Figure 12b**. Filled circles indicate the tissues where the gene had high expression. The divergence time of recent duplications is shown in blue beside the corresponding branch in red. (**c**) History of recent duplication for the CesA and Csl genes. Each bracket indicates a duplication event of the CesA or Csl genes. Divergence time is shown along a bar ranging from 0 to 50 million years ago. Filled red circles indicate genes highly expressed in the shoot.

some were potentially employed in important biological processes (for example, the control of flowering time or secondary metabolism). Approximately 70 gene families were shared by *Arabidopsis*, poplar and moso bamboo.

In comparative analysis of single-copy genes and gene families containing two to four gene members in moso bamboo and five other Poaceae plants, we found that the bamboo genome had the fewest single-member gene families, whereas it had the most two-member families among grasses (**Supplementary Fig. 9**). The timing of gene duplication events in grass genomes was estimated by calculating the synonymous substitution rate (K_S) and the divergence time between homologous genes within the two-member gene families in which only a single divergence might have occurred. The divergence within most gene clusters occurred around 7 to 12 million years ago in both the moso bamboo and maize genomes (**Fig. 1c**), suggesting the occurrence of a putative whole-genome duplication event. The estimated time of the duplication at 11 to 15 million years ago in maize is consistent with the reported divergence time of two progenitor genomes at about 11.9 million years ago[14], suggesting that there might have been a similar tetraploidization event during bamboo history. Investigation of collinear orthologs in bamboo and rice not only reinforced the occurrence of the whole-genome duplication event but also supported a tetraploid origin of bamboo, as the most recent whole-genome duplication was likely linked to polyploidy events[15] (**Supplementary Fig. 10a**). The divergence time of two progenitors was estimated at 7 to 15 million years ago (**Supplementary Fig. 10b**), consistent with the divergence time estimated using two-member gene families. For other grass species, such as rice and sorghum, there was no obvious evidence of whole-genome duplication occurring later than the divergence time of grasses at 50 million years ago[16–19].

Using 968 one-to-one single-copy genes from the 5 fully sequenced grass genomes as well as the bamboo genome, we reconstructed a phylogenetic tree to show the relationships among four subfamilies: Panicoideae, Pooideae, Ehrhartoideae and Bambusoideae (**Fig. 1d**). The analyzed grasses were divided into two sister groups, the BEP clade (Bambusoideae, Ehrhartoideae and Pooideae) and the Panicoideae clade, consistent with stated phylogeny and classification of grass subfamilies in early studies[20–22]. The tree supported the idea that the closest relationship exists between *Brachypodium* and bamboo, agreeing with the result from the analysis of chloroplast genome sequences[3]. The dN/dS value (the ratio of the rate of non-synonymous substitution to the rate of synonymous substitution) of the bamboo lineage was the highest among the compared species, suggesting strong selection pressure on bamboo genes. The estimated times for the divergence of bamboo from *Brachypodium*, rice, foxtail millet, sorghum and maize were approximately 46.9, 48.6, 53.9, 58.5 and 64.6 million years ago, respectively (**Fig. 1e**, **Supplementary Fig. 11** and **Supplementary Table 11**), indicating that the relationship between *Brachypodium* and bamboo was closer than that between rice and bamboo.

To investigate the evolutionary dynamics of the gene families, expansion and contraction were correlated with copy number. For *Arabidopsis* and six grass genomes, the number of gene families with gene contraction was greater than that of families with gene expansion, except in foxtail millet (**Fig. 1d**). Variance of family sizes occurred in a large number of gene families in bamboo (**Supplementary Table 12**). Gene families involved in the biosynthesis of carbohydrates, such as cellulose, glucan and sucrose, showed significant expansion in bamboo (P value < 0.01) relative to other grass species.

With alignment of the 30,379 gene models located on the large-sized scaffolds (>50 kb in length) to the rice and sorghum gene models, we identified 1,617 rice-bamboo and 1,539 sorghum-bamboo syntenic gene blocks, which consisted of 17,735 and 15,746 bamboo genes, respectively (**Supplementary Table 13**). The average gene number per block was approximately 11. The large number of syntenic blocks suggested good gene collinearity between bamboo and grass genomes (**Fig. 1f**). Sequence comparison indicated that approximately 85% of the bamboo genes were aligned to rice or sorghum homologs. In analysis of gene collinearity between bamboo and rice, we identified 5,370 gene losses after the whole-genome duplication event, representing approximately 28% of the total genes in the collinear regions.

A recent proteomics study showed that many metabolic processes of cell wall structure were employed in the fast growth of bamboo culms[23]. The bamboo genome sequence made it possible to investigate

© 2013 Nature America, Inc. All rights reserved.

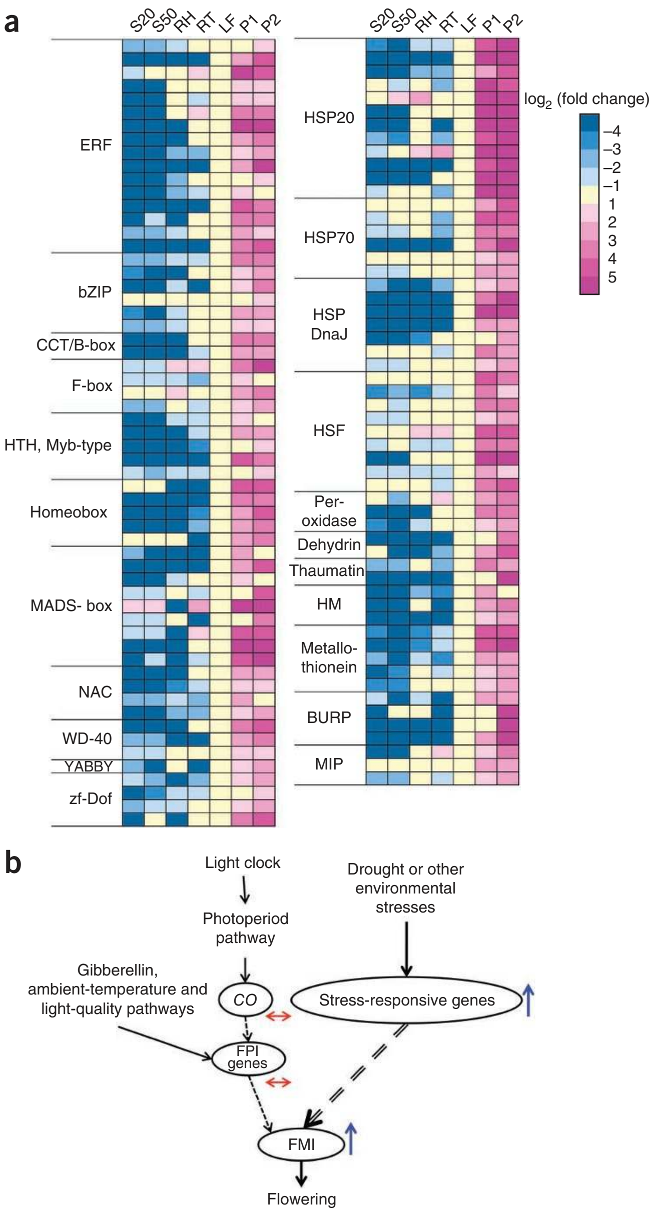

Figure 3 Gene expression at flowering time. (**a**) Clustered transcription factor and stress-responsive genes with high expression in panicles. Gene expression was measured by quantified transcription levels (reads per kilobase of exon model per million mapped reads, RPKM[37]) derived from transcriptome analysis. The gene expression levels in the tip of a 20-cm-long shoot (S20), the tip of a 50-cm-long shoot (S50), the rhizome (RH), the root (RT), the panicle at the early stage (P1) and the panicle at the flowering stage (P2) were normalized to the fold change over the expression levels in the leaf (LF) and are indicated by color. The abbreviations indicating the conserved domains encoded by flowering genes are listed in **Supplementary Table 16**. (**b**) Predicted pathway in the control of flowering time in bamboo. Blue arrows indicate that the involved genes are more highly expressed in the floral tissues, whereas red double-headed arrows indicate that the genes are not activated. Single dashed arrows represent pathways that were not used during flowering. Double dashed arrow represents stronger connections between drought-responsive and FMI genes.

the genes that might affect the formation of the cell wall structure. We detected 19 cellulose synthase (CesA) and 38 cellulose synthase–like (Csl) genes[24,25] in the bamboo genome, representing nearly the highest copy number of these genes among the 7 sequenced plant genomes (**Supplementary Table 14a**). A neighbor-joining tree showed seven recent duplications of the CesA genes (**Fig. 2a**) and eight duplications of the Csl genes in bamboo after speciation (**Fig. 2b**). The CesA, CslA and CslC gene families greatly expanded in the bamboo genome, similar to what was observed in the maize genome[26]. For CesA genes, the four most recent duplications were identified in the grass-specific clades B and G at 8.0 to 13.2 million years ago. Of the 15 CesA gene duplications, 9 occurred later than 20 million years ago (**Fig. 2c**). Transcriptome analysis showed that the recently occurring duplicates of the CesA and Csl genes had relatively high expressional levels in the shoot (**Fig. 2a**,**b** and **Supplementary Table 15**). It was also found that there were few tandem duplicates in these recent duplicates, implying that the duplications might have resulted from large-scale chromosome reconstruction. We observed that the ancient duplicated genes had high expression in the root, leaf and rhizome (**Fig. 2**). It was concluded that most of the duplications of the CesA and Csl genes were derived from whole-genome duplication, suggesting that tetraploidization was critical for the evolution of these genes.

To identify the genes involved in the biosynthesis of lignin, a structural component of the secondary cell wall, we investigated the analogous set of genes involved in the phenylpropanoid and lignin biosynthetic pathways[27,28] (**Supplementary Fig. 12c,d** and **Supplementary Table 14b**). The bamboo genome contained high copy numbers of HCT (hydroxycinnamoyl-CoA, shikimate/quinate) and CCR (cinnamoyl CoA reductase) genes, which were similar to those found in poplar. The estimated divergence time of bamboo CCR and HCT gene duplications was from 17.5 to 52.1 million years ago, earlier than the whole-genome duplication event. Both HCT and CCR family genes are key enzymes in catalyzing the conversion of phenylpropanoid pathway products into the material for lignin biosynthesis[27,29]. Although the functions of bamboo CCR and HCT genes have not yet been identified, the duplicated copies might provide multiple pathways to channel phenylpropanoid metabolism into lignin biosynthesis.

The switch to flowering after a very long period of vegetative growth and the rapid growth of spring shoots are unique characteristics of bamboo. To compare gene expression between flowering and vegetative tissues, we collected flowering (panicle) and vegetative tissues from moso bamboo plants for RNA-seq data analysis. More than 600 bamboo genes were highly expressed in the 2 panicle tissues (with at least a 2-fold difference in the expression level in panicles relative to the levels in 5 vegetative tissues; Online Methods). Over 30% of the identified flowering genes could be categorized as transcription factor genes, heat shock protein genes or other stress-responsive genes (**Fig. 3a** and **Supplementary Table 16**). The transcription factor genes that are homologs of *OsMADS1*, *OsMADS2*, *OsMADS3* and *OsMADS14* in rice[30] were determined to be involved in floral meristem identity (FMI), which converts the vegetative meristem to a flowering fate. However, the genes employed in typical flowering promotion pathways (such as those in the photoperiod, gibberellins, ambient-temperature or light-quality pathways) and floral pathway integrator (FPI) genes[31,32] were not highly expressed in these floral tissues in bamboo. Repeat insertions were found in the genic or regulatory region of most homologs encoding *CONSTANS* (*CO*)[33] and FPI genes, which might result in low gene expression in floral tissues (**Supplementary Tables 17** and **18**). The *CO* and FPI genes constitute the critical link between the flowering promotion pathways and the FMI in the flowering gene network. Low expression of *CO* and FPI genes and high expression of genes involved in FMI suggested that activation of FMI might not depend more on these known promotion pathways in bamboo flowering (**Fig. 3b**).

© 2013 Nature America, Inc. All rights reserved.

Contrasting with the expression pattern of flowering pathway genes, over 100 stress-responsive genes (15% of 600) showed high expression levels in panicles, being on average 11.1-fold more highly expressed in panicle tissues. Sequence alignment showed that a total of 70 bamboo genes shared high identity with known rice genes, which were mainly involved in the abscisic acid pathway, the ethylene-responsive pathway, sugar metabolism and the calcium-dependent signal transduction pathway, besides the FMI or FPI pathways (**Supplementary Table 19**). Of these genes, 45 (65% of 70) were involved in the response to drought stress or to other correlative stresses (such as oxidative stress), and 10 (15%) were involved in flowering pathways. Some FMI-related genes and their upstream regulatory drought-responsive genes had been observed to have high expression during flowering (**Supplementary Fig. 13**), suggesting a potential connection between severe drought stress and flowering (**Fig. 3b**). It is noteworthy that the bamboo panicles were collected in southern China, where a severe drought occurred just 2 months before the collection of our samples. However, further experiments are necessary to identify the mechanisms underlying the activation of bamboo flowering.

URLs. KEGG, http://www.genome.jp/kegg/; SMALT, http://www.sanger.ac.uk/resources/software/smalt/; SOAPdenovo, http://soap.genomics.org.cn/; Repbase, http://www.girinst.org/repbase/; cell wall genomics, http://cellwall.genomics.purdue.edu/families/; PHYLIP version 3.69, http://evolution.genetics.washington.edu/phylip.html; PLAZA Comparative Genomics Platform, http://bioinformatics.psb.ugent.be/plaza/; RepeatModeler, http://www.repeatmasker.org/RepeatModeler.html; RepeatMasker, http://www.repeatmasker.org/; EMBL, http://www.ebi.ac.uk/. GenBank, http://www.ncbi.nlm.nih.gov/nuccore/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Short-read sequencing data from this whole-genome shotgun project have been deposited at the European Molecular Biology Laboratory (EMBL) under the accession ERP001340. RNA-seq data have also been deposited at EMBL under accession ERP001341. Data from the Sanger sequencing of BACs were deposited at EMBL and GenBank under the accessions included in parentheses: B001E05 (FO203447), B001G05 (FO203436), B001I05 (FO203448), B001I13 (FO203437), B015M02 (FO203443), B019A14 (FO203439), B031C15 (FO203444) and B035L11 (FO203441). All bamboo data have been released at the official website of the National Center for Gene Research (http://www.ncgr.ac.cn/bamboo). The entire data set includes genome assemblies, BAC end sequences and annotation of genes and lists of repeat elements, heterozygous SNPs, tRNAs, miRNAs and gene clusters. The current version of the data set is the first version.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank C. Xu for her technological support in cytogenetic analysis. This work was supported by the Forestry Project of the Ministry of Science and Technology of the People's Republic of China (grant 200704001 to Z.J.) and the Chinese Academy of Sciences (KSCX2-YW-G-034 to B.H.).

AUTHOR CONTRIBUTIONS

Z.J., Z.P. and B.H. conceived the project and its components, designed the studies and contributed to the original concept of the project. Q.F., D.F., Y.G., W.L., Yiqi Lu, T. Hu, N.Y., C. Zhou and Q.W. performed DNA preparation and genome sequencing. Ying Lu, Y. Li, K.L., T.L. and X.H. performed genome data analysis. Ying Lu and T.L. performed transcriptome (RNA-seq and cDNA) analyses. Z.N., H.L. and Q.Z. developed the *de novo* assembly pipeline and performed *de novo* genome assembly. L.Z. performed BAC sequence assembly. L.L., Z.G., X.Y., T.W., K.M., C. Zhuang, X.C., W.T., G.L., Y. Liu, J.C., Zhenjing Liu, L.Y. and Zhenhua Liu collected bamboo samples and performed cytogenetics studies and functional analysis. T. Huang, Y.Z. and C. Zhu provided IT support. B.F. and X.L. coordinated the project. Ying Lu, B.H., Z.P. and Z.J. analyzed the data as a whole and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

This work is licensed under a Creative Commons Attribution-NonCommercial-Share Alike 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/

1. Lobovikov, M., Paudel, S., Piazza, M., Ren, H. & Wu, J. *World Bamboo Resources: A Thematic Study Prepared in the Framework of the Global Forest Resources Assessment* 2005 (Food and Agriculture Organization of the United Nations, Rome, 2007).
2. Peng, Z. *et al.* Genome-wide characterization of the biggest grass, bamboo, based on 10,608 putative full-length cDNA sequences. *BMC Plant Biol.* **10**, 116 (2010).
3. Zhang, Y.J., Ma, P.F. & Li, D.Z. High-throughput sequencing of six bamboo chloroplast genomes: phylogenetic implications for temperate woody bamboos (Poaceae: Bambusoideae). *PLoS ONE* **6**, e20596 (2011).
4. Gui, Y.J. *et al.* Insights into the bamboo genome: syntenic relationships to rice and sorghum. *J. Integr. Plant Biol.* **52**, 1008–1015 (2010).
5. Sungkaew, S., Stapleton, C.M., Salamin, N. & Hodkinson, T.R. Non-monophyly of the woody bamboos (Bambuseae; Poaceae): a multi-gene region phylogenetic analysis of Bambusoideae s.s. *J. Plant Res.* **122**, 95–108 (2009).
6. Sharma, R.K. *et al.* Evaluation of rice and sugarcane SSR markers for phylogenetic and genetic diversity analyses in bamboo. *Genome* **51**, 91–103 (2008).
7. Das, M., Bhattacharya, S. & Pal, A. Generation and characterization of SCARs by cloning and sequencing of RAPD products: a strategy for species-specific marker development in bamboo. *Ann. Bot.* (*Lond.*) **95**, 835–841 (2005).
8. Chen, R. *et al. Chromosome Atlas of Major Economic Plants Genome in China, Tomus IV—Chromosome Atlas of Various Bamboo Species* (Science Press, Beijing, 2003).
9. Gui, Y. *et al.* Genome size and sequence composition of moso bamboo: a comparative study. *Sci. China C Life Sci.* **50**, 700–705 (2007).
10. Li, R. *et al.* SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714 (2008).
11. Tuskan, G.A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596–1604 (2006).
12. Velasco, R. *et al.* A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* **2**, e1326 (2007).
13. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40** Database issue, D109–D114 (2012).
14. Swigonová, Z. *et al.* Close split of sorghum and maize genome progenitors. *Genome Res.* **14**, 1916–1923 (2004).
15. Wendel, J.F. Genome evolution in polyploids. *Plant Mol. Biol.* **42**, 225–249 (2000).
16. Gaut, B.S. Evolutionary dynamics of grass genomes. *New Phytol.* **154**, 15–28 (2002).
17. Kellogg, E.A. Relationships of cereal crops and other grasses. *Proc. Natl. Acad. Sci. USA* **95**, 2005–2010 (1998).
18. Goff, S.A. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**, 92–100 (2002).
19. Guyot, R. & Keller, B. Ancestral genome duplication in rice. *Genome* **47**, 610–614 (2004).
20. Barker, N.P. *et al.* Phylogeny and subfamilial classification of the grasses (Poaceae). *Ann. Mo. Bot. Gard.* **88**, 373–457 (2001).
21. Sánchen-Ken, J.G., Clark, L.G., Kellogg, E.A. & Kay, E.E. Reinstatement and emendation of subfamily Micrairoideae (Poaceae). *Syst. Bot.* **32**, 71–80 (2007).
22. Bouchenak-Khelladi, Y. *et al.* Large multi-gene phylogenetic trees of the grasses (Poaceae): progress towards complete tribal and generic level sampling. *Mol. Phylogenet. Evol.* **47**, 488–505 (2008).
23. Cui, K., He, C.Y., Zhang, J.G., Duan, A.G. & Zeng, Y.F. Temporal and spatial profiling of internode elongation-associated protein expression in rapidly growing culms of bamboo. *J. Proteome Res.* **11**, 2492–2507 (2012).
24. Somerville, C. Cellulose synthesis in higher plants. *Annu. Rev. Cell Dev. Biol.* **22**, 53–78 (2006).
25. Yin, Y., Huang, J. & Xu, Y. The cellulose synthase superfamily in fully sequenced plants and algae. *BMC Plant Biol.* **9**, 99 (2009).
26. Schnable, P.S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
27. Humphreys, J.M. & Chapple, C. Rewriting the lignin roadmap. *Curr. Opin. Plant Biol.* **5**, 224–229 (2002).

© 2013 Nature America, Inc. All rights reserved.

npg

28. Boerjan, W., Ralph, J. & Baucher, M. Lignin biosynthesis. *Annu. Rev. Plant Biol.* **54**, 519–546 (2003).
29. Hamberger, B. *et al.* Genome-wide analyses of phenylpropanoid-related genes in *Populus trichocarpa*, *Arabidopsis thaliana*, and *Oryza sativa*: the *Populus* lignin toolbox and conservation and diversification of angiosperm gene families. *Can. J. Bot.* **85**, 1182–1201 (2007).
30. Arora, R. *et al.* MADS-box gene family in rice: genome-wide identification, organization and expression profiling during reproductive development and stress. *BMC Genomics* **8**, 242 (2007).
31. Ehrenreich, I.M. *et al.* Candidate gene association mapping of *Arabidopsis* flowering time. *Genetics* **183**, 325–335 (2009).
32. Fornara, F., Montaigu, A. & Coupland, G. SnapShot: control of flowering in. *Arabidopsis. Cell* **141**, 550 e1–550.e2 (2010).
33. Putterill, J., Robson, F., Lee, K., Simon, R. & Coupland, G. The *CONSTANS* gene of *Arabidopsis* promotes flowering and encodes a protein showing similarities to zinc finger transcription factors. *Cell* **80**, 847–857 (1995).
34. Gaut, B.S., Morton, B.R., McCaig, B.C. & Clegg, M.T. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proc. Natl. Acad. Sci. USA* **93**, 10274–10279 (1996).
35. Ouyang, S. *et al.* The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.* **35**, D883–D887 (2007).
36. Paterson, A.H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).
37. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).

© 2013 Nature America, Inc. All rights reserved.

© 2013 Nature America, Inc. All rights reserved.

ONLINE METHODS

DNA library preparation and sequencing. Moso bamboo samples for shotgun sequencing were collected in the Tianmu Mountain National Nature Reserve in Zhejiang Province, China, from five plants that were determined to be a single individual when they were found to share the same rhizome system. Using the DNeasy Plant Mini kit (Qiagen), we extracted total DNA from moso bamboo leaves. Genomic DNA was purified according to the protocol for the isolation of high-molecular-weight nuclear DNA[38]. We applied an amplification-free approach to prepare sequencing libraries with a short insert size of 350 to 400 bp for paired-end reads, following a modified version of the manufacturer's protocol (Illumina) and methods described previously[39]. For construction of libraries with insert sizes of 3, 8 and 16 kb for mate-paired reads, we used combined protocols from the Mate Pair Library v2 Sample Preparation Guide (Illumina) and the Paired-End Library Preparation Method Manual (Roche). Raw data from paired-end libraries with read lengths of 2 × 120 bp and 2 × 100 bp were generated by an Illumina Genome Analyzer IIx sequencer and a HiSeq 2000 sequencer, respectively. The mate-paired reads (2 × 50 bp and 2 × 76 bp) were generated by the Illumina Genome Analyzer IIx sequencer.

Sequence assembly. We developed a *de novo* assembly pipeline to assemble the Illumina short reads (**Supplementary Fig. 2**), which integrated the existing assemblers Phusion2 (ref. 40), SOAPdenovo, Abyss[41] and SSPACE[42]. Before assembling sequences, paired-end reads were screened to remove low-quality reads that contained ten or more unique *K*-mers. Screened paired-end reads were then clustered into thousands of groups by Phusion2 with *K*-mer of 51 bp. During clustering, *K*-tuples (contiguous DNA sequences that are *K* bases long) were merged and sorted into a table, and shared *K*-mer words were linked in a relation matrix. The reads in each cluster were assembled in SOAP_contigs and Abyss_contigs by SOAPdenovo and Abyss, respectively. Contigs derived from both assemblers were then merged to generate the initial contigs by GAP5 (ref. 43). Mate-paired reads were mapped to the initial contigs by the aligner SMALT. To reduce redundancy, when two or more mate-paired reads were mapped to the same location, only one pair of them could be kept for the following assembly. The average insert size of each mate pair library was estimated by determining the distance between mate-paired reads that were well mapped to the same contig (**Supplementary Fig. 14**). Using paired-end and mate-paired reads, preliminary scaffolds were assembled by SOAPdenovo with *K*-mer of 61 bp. Scaffolds were rearranged by mapping the initial contigs to the primary scaffolds. The final scaffolding was performed by SSPACE, using mate-paired reads and BAC end sequences. Scaffolds less than 500 bp in length were not included in statistics and the following annotation.

Transcriptome sequencing with an amplification-free library preparation method. Five vegetative tissues (young leaves, rhizomes, roots, tips of 20-cm-high shoots and tips of 50-cm-high shoots) were collected from the same individual used in genome sequencing. Flowering tissues were collected from the plants of a single individual growing in Guangxi Province in southern China (**Supplementary Note**). Up to 400 µg of total RNA was isolated from each tissue using a TRIzol-based method at the beginning of the preparation of cDNA sequencing libraries. Libraries were constructed with Illumina sequencing technology and an amplification-free method[39]. Briefly, after treatment with DNase, mRNA was isolated from total RNA with the Oligotex mRNA Midi kit (Qiagen). Fragmentation of mRNA followed the protocol of the Ambion RNA Fragmentation Reagents kit. Sequencing libraries of cDNA were constructed using the same amplification-free approach as used in genomic sequencing.

Annotation of protein-coding genes. Protein-coding gene models were derived from evidence-based FgeneSH++ (Softberry) pseudomolecules (**Supplementary Fig. 7**). To facilitate gene models and address interesting biological questions, a total of 110 billion RNA-seq reads were generated from 7 libraries, and a select group of 8,253 cDNA sequences was used. Each potential gene model was supported by the expressed sequences from the moso bamboo cDNA or transcriptome sequences (**Supplementary Note**).

Using amplification-free RNA-seq data, each library detected over 24,000 loci matching our requirement that candidate gene models be supported by the full-length cDNA or 2 or more uniquely matched RNA-seq sequences (**Supplementary Fig. 15).**The coverage of RNA-seq reads on the coding regions of annotated loci indicated that up to 27,000 predicted gene models were strongly supported by transcriptome sequences (RNA-seq data coverage in coding regions of >70%; **Supplementary Fig. 16**). In combination with *ab initio* gene prediction and alignments of the transcriptome and cDNA data, a total of 31,987 high-confidence genes were identified in the annotation.

Identification of genes involved in cell wall biosynthesis. To investigate the genes involved in cell wall biosynthesis, we compared the CesA, Csl and phenylpropanoid-lignin biosynthesis genes in bamboo and other grass genomes, as well as in the *Arabidopsis* and poplar genomes. We used sequences encoded by the identified CesA or Csl genes in *Arabidopsis*[25], poplar[44,45], rice[46], maize[26] and sorghum[24] for alignment to those encoded by the gene models of *Brachypodium* and bamboo by BLASTP with *E* values under 1×10^{-10}. Aligned hits with at least 200 amino acids of matched length and over 50% protein sequence identity were considered to be homologs of the CesA or Csl genes. For the phenylpropanoid-lignin genes, the reported homologs of *Arabidopsis*[47], poplar, rice[29] and maize downloaded from the cell wall genomics browser (see URLs) were used as the seed sequences to detect the bamboo, *Brachypodium* and sorghum gene models by BLASTP with *E* values under 1×10^{-10} and with over 50% identity over the whole protein sequence. Detected homologs consisted of not only phenylpropanoid-lignin genes but also many phenylpropanoid-lignin–like genes, which might be involved in different pathways, even though they share high sequence identity (such as *At4CL*-like genes[48]). To remove these phenylpropanoid-lignin–like genes, we used the phenylpropanoid-lignin genes from *Arabidopsis*, maize, rice and poplar to build an initial neighbor-joining tree to cluster the phenylpropanoid-lignin and phenylpropanoid-lignin–like genes into different clades. According to this cluster information, we manually filtered the top BLASTP hits of each homolog to include only phenylpropanoid-lignin genes in our phylogenetic analysis. Consensus neighbor-joining trees were generated using PHYLIP (version 3.69) on the basis of 100 bootstrap trees.

Identification of flowering genes. Use of the amplification-free approach for the preparation of transcriptome sequencing libraries eliminated much of the redundancy in transcripts introduced by the amplification of templates during library construction. Generated RNA-seq reads were aligned to the gene model set with the SMALT aligner. The quantity of reads uniquely mapped to the gene models was converted to a quantification of the transcript levels in RPKM. We then used the R package DEGseq[49] to digitally measure the differential expression at annotated loci. A gene with expression that was more than twofold higher (Q value < 0.001; ref. 50) in panicles relative to any other vegetative tissue and that had at least five mapped transcripts was considered to be a potential flowering gene in moso bamboo. Both amino-acid sequences and the conserved Interpro function domains encoded by the loci were compared to those of known *Arabidopsis* (TAIR10)[51] and rice (MSU RGAP 6.1) genes, the outputs of which were manually checked to determine the putative functions of the loci involved in the flowering pathways.

Construction of gene families among fully sequenced grass genomes. We applied OrthoMCL[52] clustering to identify gene families enriched in the Pooideae, Ehrhartoideae, Panicoideae and Bambusoideae families. The bamboo gene predictions and (MSU RGAP 6.1), *Brachypodium* (MIPS1.2), sorghum (JGI 1.4), maize (5b.60), foxtail millet (v8.0), poplar (JGI 2.0) and *Arabidopsis* (TAIR10) gene sequences downloaded from the PLAZA comparative genome database (version 2.0)[53] were used to infer potential orthologous families of genes. The rice genome represented Ehrhartoideae; the maize, sorghum and foxtail millet genomes represented Panicoideae; the *Brachypodium* genome represented Pooideae; and the bamboo genome represented Bambusoideae. The transposable element–derived genes in the genomes from the PLAZA database were removed before they were added to the alignment. An all-against-all comparison was then performed using BLASTP with an *E* value of 1×10^{-10}. We then used the standard setting to compute gene similarities across all eight genomes. A total of 194,376 protein sequences were grouped into 27,294 gene clusters. OrthoMCL clustered a total of 968 single-copy gene families, which were subjected to phylogenetic analyses by Mrbayes[54]. The expansion and contraction of the gene clusters were determined by a CAFE calculation (version 2.1)[55] on the basis of changes in gene family size in generated phylogenetic history.

doi:10.1038/ng.2569

Repeat annotation. A *de novo* repeat prediction for the moso bamboo genome was carried out by successively using RepeatModeler (version 1.0.3) and RepeatMasker (version 3.3.0) (see URLs). We first constructed a moso bamboo repeat library using RepeatModeler with default parameters. Two complementary programs, RECON and RepeatScout[56,57], were configured at the center of RepeatModeler and were employed in the identification of repeat family sequences in the genome. The consensus sequences for the families were manually examined by aligning them to the known Repbase transposable element library (version 16.0), and known gene and genome sequences downloaded from the NCBI database (nt and nr; released 9 September 2011). The moso bamboo transposable element library was composed of a total of 1,403 generated consensus sequences and their classification information, and the library was used to run RepeatMasker on the whole-genome assemblies. Full-length LTR retrotransposons were predicted using LTRharvest[58] and LTR_FINDER[59].

38. Peterson, D.G., Tomkins, J.P., Frisch, D.A., Wing, R.A. & Paterson, A.H. Construction of plant bacterial artificial chromosome (BAC) libraries: an illustrated guide. *J. Agric. Genomics* **5**, 34–40 (2000).
39. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods* **6**, 291–295 (2009).
40. Mullikin, J.C. & Ning, Z. The Phusion assembler. *Genome Res.* **13**, 81–90 (2003).
41. Simpson, J.T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).
42. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
43. Bonfield, J.K. & Whitwham, A. Gap5—editing the billion fragment sequence assembly. *Bioinformatics* **26**, 1699–1703 (2010).
44. Djerbi, S., Lindskog, M., Arvestad, L., Sterky, F. & Teeri, T.T. The genome sequence of black cottonwood (*Populus trichocarpa*) reveals 18 conserved cellulose synthase (*CesA*) genes. *Planta* **221**, 739–746 (2005).
45. Suzuki, S., Li, L., Sun, Y.H. & Chiang, V.L. The cellulose synthase gene superfamily and biochemical functions of xylem-specific cellulose synthase–like genes in *Populus trichocarpa. Plant Physiol.* **142**, 1233–1245 (2006).
46. Hazen, S.P., Scott-Craig, J.S. & Walton, J.D. Cellulose synthase–like genes of rice. *Plant Physiol.* **128**, 336–340 (2002).
47. Ehlting, J. *et al.* Global transcript profiling of primary stems from *Arabidopsis thaliana* identifies candidate genes for missing links in lignin biosynthesis and transcriptional regulators of fiber differentiation. *Plant J.* **42**, 618–640 (2005).
48. Costa, M.A. *et al.* Characterization *in vitro* and *in vivo* of the putative multigene 4-coumarate:CoA ligase network in *Arabidopsis*: syringyl lignin and sinapate/sinapyl alcohol derivative formation. *Phytochemistry* **66**, 2072–2091 (2005).
49. Wang, L., Feng, Z., Wang, X., Wang, X. & Zhang, X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* **26**, 136–138 (2010).
50. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc., B* **57**, 289–300 (1995).
51. Childs, K.L. *et al.* The TIGR Plant Transcript Assemblies database. *Nucleic Acids Res.* **35** Database issue, D846–D851 (2007).
52. Li, L., Stoeckert, C.J. Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
53. Van Bel, M. *et al.* Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiol.* **158**, 590–600 (2012).
54. Huelsenbeck, J.P. & Ronquist, F. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* **17**, 754–755 (2001).
55. De Bie, T., Cristianini, N., Demuth, J.P. & Hahn, M.W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006).
56. Bao, Z. & Eddy, S.R. Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276 (2002).
57. Price, A.L., Jones, N.C. & Pevzner, P.A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21** (suppl. 1), i351–i358 (2005).
58. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
59. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).

© 2013 Nature America, Inc. All rights reserved.

doi:10.1038/ng.2569