SCIENCE CHINA
Life Sciences

THEMATIC ISSUE: Cotton genome plus
• RESEARCH PAPER •

February 2016 Vol.59 No.2: 142–153
doi: 10.1007/s11427-016-5001-1

Comprehensive analysis of NAC transcription factors in diploid *Gossypium*: sequence conservation and expression analysis uncover their roles during fiber development

Haihong Shang, Zhongna Wang, Changsong Zou, Zhen Zhang, Weijie Li, Junwen Li, Yuzhen Shi, Wankui Gong, Tingting Chen, Aiying Liu, Juwu Gong, Qun Ge & Youlu Yuan*

State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang 455000, China

Received May 14, 2015; accepted July 20, 2015; published online January 22, 2016

Determining how function evolves following gene duplication is necessary for understanding gene expansion. Transcription factors (TFs) are a class of proteins that regulate gene expression by binding to specific *cis*-acting elements in the promoters of target genes, subsequently activating or repressing their transcription. In the present study, we systematically examined the functional diversification of the NAC transcription factor (NAC-TFs) family by analyzing their chromosomal location, structure, phylogeny, and expression pattern in *Gossypium raimondii* (Gr) and *G. arboreum* (Ga). The 145 and 141 *NAC* genes identified in the Gr and Ga genomes, respectively, were annotated and divided into 18 subfamilies, which showed distinct divergence in gene structure and expression patterns during fiber development. In addition, when the functional parameters were examined, clear divergence was observed within tandem clusters, which suggested that subfunctionalization had occurred among duplicate genes. The expression patterns of homologous gene pairs also changed, suggestive of the diversification of gene function during the evolution of diploid cotton. These findings provide insights into the mechanisms underlying the functional differentiation of duplicated *NAC-TFs* genes in two diploid cotton species.

cotton, NAC gene family, phylogeny, expression patterns

Citation: Shang, H., Wang, Z., Zou, C., Zhang, Z., Li, W., Li, J., Shi, Y., Gong, W., Chen, T., Liu, A., Gong, J., Ge, Q., Yuan, Y. (2016). Comprehensive analysis of NAC transcription factors in diploid *Gossypium*: sequence conservation and expression analysis uncover their roles during fiber development. Sci China Life Sci 59, 142−153. doi: 10.1007/s11427-016-5001-1

INTRODUCTION

NAC (NAM, ATAF, and CUC) transcription factor (*NAC-TF*) genes belong to a large family of genes that encode important regulatory proteins in plants (Wang et al., 2011). They also play an important role in plant development, and are involved in osmotic stress and various plant developmental processes. In recent years, genomic analyses of the NAC gene family have been conducted in angiosperms such as *Arabidopsis thaliana* (Ooka et al., 2003), *Oryza sativa* (Fang et al., 2008; Nuruzzaman et al., 2010), *Vitis vinifera* (Wang et al., 2013), *Populus trichocarpa* (Hu et al., 2010), *Glycine soja* (Zhang et al., 2008), *Setaria italic* (Puranik et al., 2013), *Gossypium raimondii* (Shang et al., 2013), and *Musa acuminata* (Cenci et al., 2014). As one of the largest groups of plant transcription factors (TFs), the NAC-TF family consists of several genes in plants. For example, there are 105 *TF* genes in *Arabidopsis* (Ooka et al., 2003), 140 genes in rice (Fang et al., 2008; Nuruzzaman et al., 2010), 101 genes in soybean (Pinheiro et al., 2009), 163 in poplar (Hu et al., 2010), and 145 in *G. raimondii* (Shang et al., 2013). The gene structure, phylogeny, expression, and functional diversification of the *NAC-TF* family have been systematically examined (Ooka et al., 2003; Fang

*Corresponding author (email: youluyuan@hotmail.com)

et al., 2008; Zhang et al., 2008; Pinheiro et al., 2009; Nuruzzaman et al., 2010; Hu et al., 2010; Wang et al., 2013; Puranik et al., 2013; Shang et al., 2013; Cenci et al., 2014) Genomic analyses indicate that segmental duplications have significantly contributed to the expansion of the *NAC* gene family in specific angiosperms (Fang et al., 2008; Nuruzzaman et al., 2010; Hu et al., 2010; Wang et al., 2013; Shang et al., 2013; Singh et al., 2013). Expression analyses have also provided insight into the functional divergence among members of the *NAC* gene family (Hu et al., 2010; Lee et al., 2012; Wang et al., 2013; Shang et al., 2013; Singh et al., 2013; Hussey et al., 2015). Furthermore, determining the functional divergence of the NAC-TF family is essential for our understanding of the mechanisms underlying gene family expansion and functional evolution (Hu et al., 2010; Lee et al., 2012; Wang et al., 2013; Shang et al., 2013; Singh et al., 2013).

As an economically significant crop around the world, cotton is not only a primary fiber resource, but is also an important oil material. Previous genetic studies on cotton have focused on the identification and function of the NAC gene family (Shang et al., 2013; Shah et al., 2013; Li et al., 2013). A comprehensive study of the *NAC* gene family has been performed in *G. raimondii*, incorporating analyses of phylogeny, chromosomal location, gene structure, conserved motifs, and expression profiling (Shang et al., 2013). Ten stress-responsive NAC genes (*GhNAC8–GhNAC17*) were isolated from cotton (*G. hirsutum* L.) (Shah et al., 2013). However, for *G. arboreum*, which is a putative donor species for the tetraploid cotton species and a typical model for fiber development and cellulose biosynthesis, reports on the comprehensive analysis of *NAC-TF*s are limited. Whole-genome sequencing of *G. raimondii* (Wang et al., 2012) and *G. arboreum* (Li et al., 2014) has provided an excellent opportunity for genome annotation and for evolutionary and comparative genomic investigations in cotton (Cao et al., 2015). In the present study, we focused on the isolation and functional characterization of *G. arboreum*, and on the divergence of orthologous gene expression between two diploid cotton species.

We systematically examined the functional diversification of the *NAC-TF* family by analyzing the chromosomal location, structure, phylogeny, and expression patterns of *G. arboreum*. A total of 143 *NAC-TF* genes were identified in the *G. arboreum* genome, which were further divided into 18 distinct subfamilies based on distinction in their gene structure and expression during fiber development. In addition, the assessment of functional parameters indicated a clear divergence among tandem clusters, which was suggestive of subfunctionalization among duplicate genes. Changes in the expression pattern of homologous gene pairs were also observed, which suggested diversification of gene function during the evolution of diploid cotton. Our findings shed light on mechanisms underlying the functional differentiation of duplicate *NAC-TF* genes in two diploid cotton species.

RESULTS

Large NAC-TF family in diploid *G. arboreum*

Initially, 156 putative *NAC-TF* genes were identified using the homologous alignment method. Based on the characteristics of the DNA-binding domain (DBD) of NAC-TFs, we manually inspected the sequence alignment. A total of 143 full-length genes encoding putative NAC-TFs were identified in the *G. arboreum* genome (Table S1 in Supporting Information) and designated *GaNAC001* to *GaNAC143* (Figure 1, Table S1 in Supporting Information). These findings are similar to those obtained for another diploid cotton species, *G. raimondii* (Shang et al., 2013).

The length of the coding region of 143 *G. arboreum NAC-TF* genes ranged from 617 to 7,935 bp, with the longest coding region being 10-fold longer than the shortest (Table S1 in Supporting Information). The length of 80 *NAC-TF* genes ranged from 1,000 to 2,000 bp. The length of 18 *NAC-TF* genes was<1,000 bp and that of 17 *NAC-TF* genes was between 2,000 and 3,000 bp. Only three genes were>5,000 bp long. The *NAC-TF* genes identified in *G. arboreum* encoded proteins ranging from 145 to 859 amino acid (aa) residues in length, with the average being 339 aa.

Genomic organization and gene structure of the *G. arboreum* NAC-TF family

NAC genes were non-randomly distributed across the *G. arboreum* chromosomes, which is similar to that observed for the diploid species *G. raimondii* (Shang et al., 2013). Of the 143 *NAC-TF* genes, 139 were located on 13 chromosomes, whereas the other four genes were located in the scaffolds (Figure 2, Table 1 in Supporting Information). For example, 16 of these genes were located on chromosomes 5 and 10, and 15 were located on chromosome 9. Twelve *NAC-TF* genes were located on chromosomes 4, 6, and 11. Only one of these genes was located on chromosome 12. In terms of cluster distribution, two clusters consisting of three genes were distributed on chromosomes 3 and 4, respectively (Figure 2, Table 1 in Supporting Information). These findings are similar to those reported for NAC genes of another diploid cotton, *G. raimondii* (Shang et al., 2013).

Clustal X version 1.83 was used to analyze the predicted sequence of *NAC* in *G. arboretum*, and MEGA 5.03 was used to construct a phylogenetic tree representing the 143-aa sequences of the *G. arboreum* gene based on the similarity of the NAC protein domain. The 143 *NAC* genes were further divided into nine subfamilies (Figure 3A). Analysis of the exon/intron structure of the coding sequences of the *NAC-TF* genes revealed a high level of structural diversity. Sub-clades I and II consisted of 18 and 19 genes, respectively, and subfamilies V and VII had 21

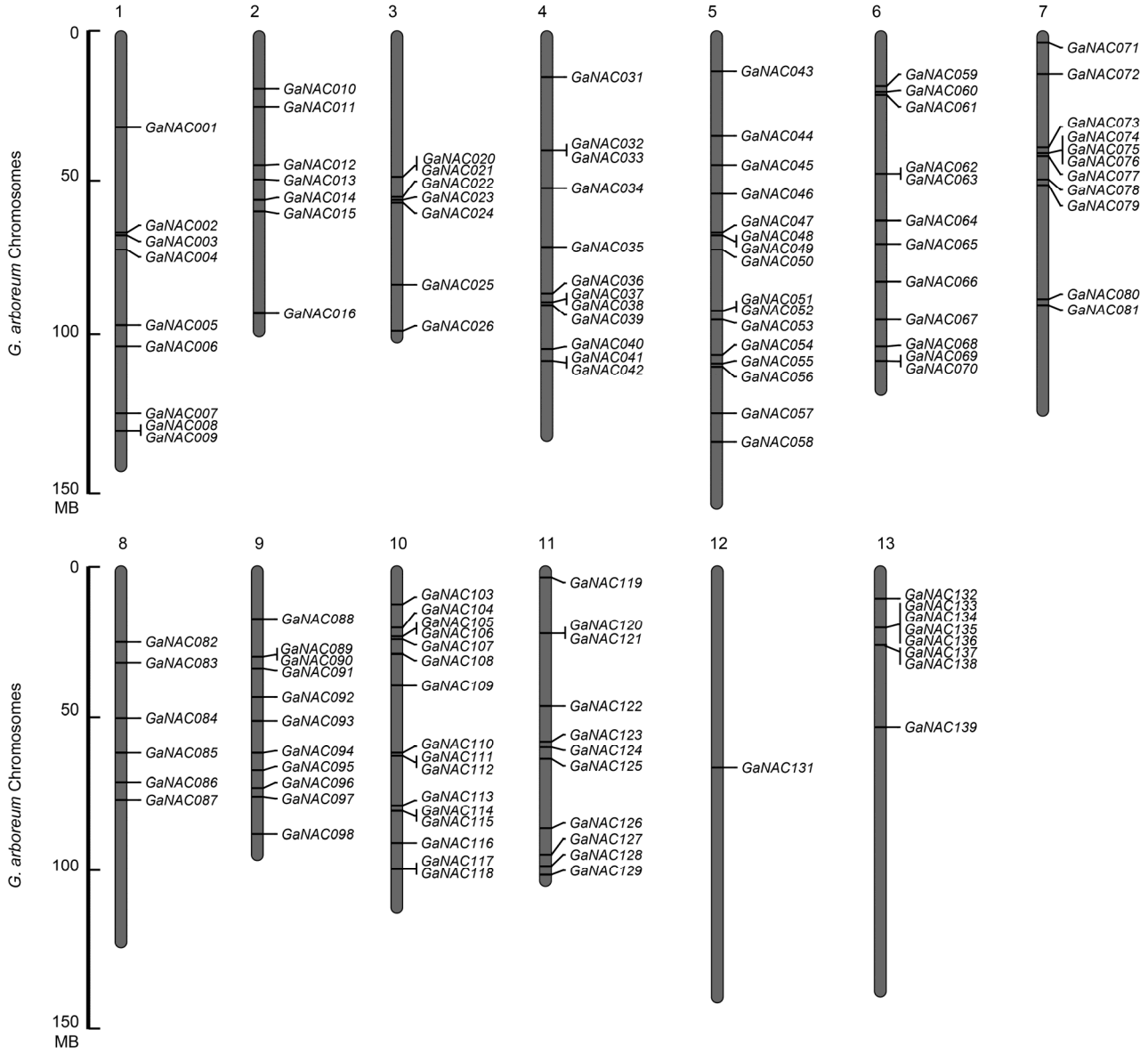

Figure 1 Chromosomal location of 145 grapevine *GaNAC* genes. A total of 139 genes are located on normal chromosomes, whereas the other three are located on scaffolds.

genes each. In the nine subclades, members of subfamilies I, V, VI, VIII, and IX possessed similar gene structures. Comparatively, the gene structures of subfamilies III, VII, and IV varied, with exon numbers ranging from three to nine (Figure 3B).

The results of previous studies suggest that at least two whole genome duplication (WGD) events in *G. arboreum* coincided with that in *G. raimondii*, and were estimated to have occurred 13–20 and 115–146 million years ago (Ji et al., 2003; Wang et al., 2012; Li et al., 2014). Synteny analysis indicated 209 and 295 collinear blocks covering 82% and 66% of the *G. raimondii* and *G. arboreum* genomes, respectively (Ji et al., 2003; Lynch et al., 2004; Li et al., 2014; Cao et al., 2015)

Based on the results of syntonic analysis between *G. raimondii* and *G. arboreum*, 95 homologous gene pairs were identified and determined as being distributed across 39 syntonic blocks. Among these, four syntonic blocks contained at least six homologous gene pairs, and maximum homology blocks comprised seven gene pairs in *G. arboretum* chromosome 4 and 7 (Figure 2, Table S2 in Supporting Information).

Phylogenetic analysis of NAC-TF family

A total of 882 NAC-TF protein sequences from the genomes of *G. arboretum*, *G. raimondii*, *A. thaliana*, *O. sativa*, *V. vinifera*, *P. trichocarpa*, *G. max*, *Theobroma cacao* L., *Carica papaya*, and *Ricinus communis* were selected for further analysis. Using MEGA 5.0 and the NJ method, a phylogenetic tree was constructed using the putative 882

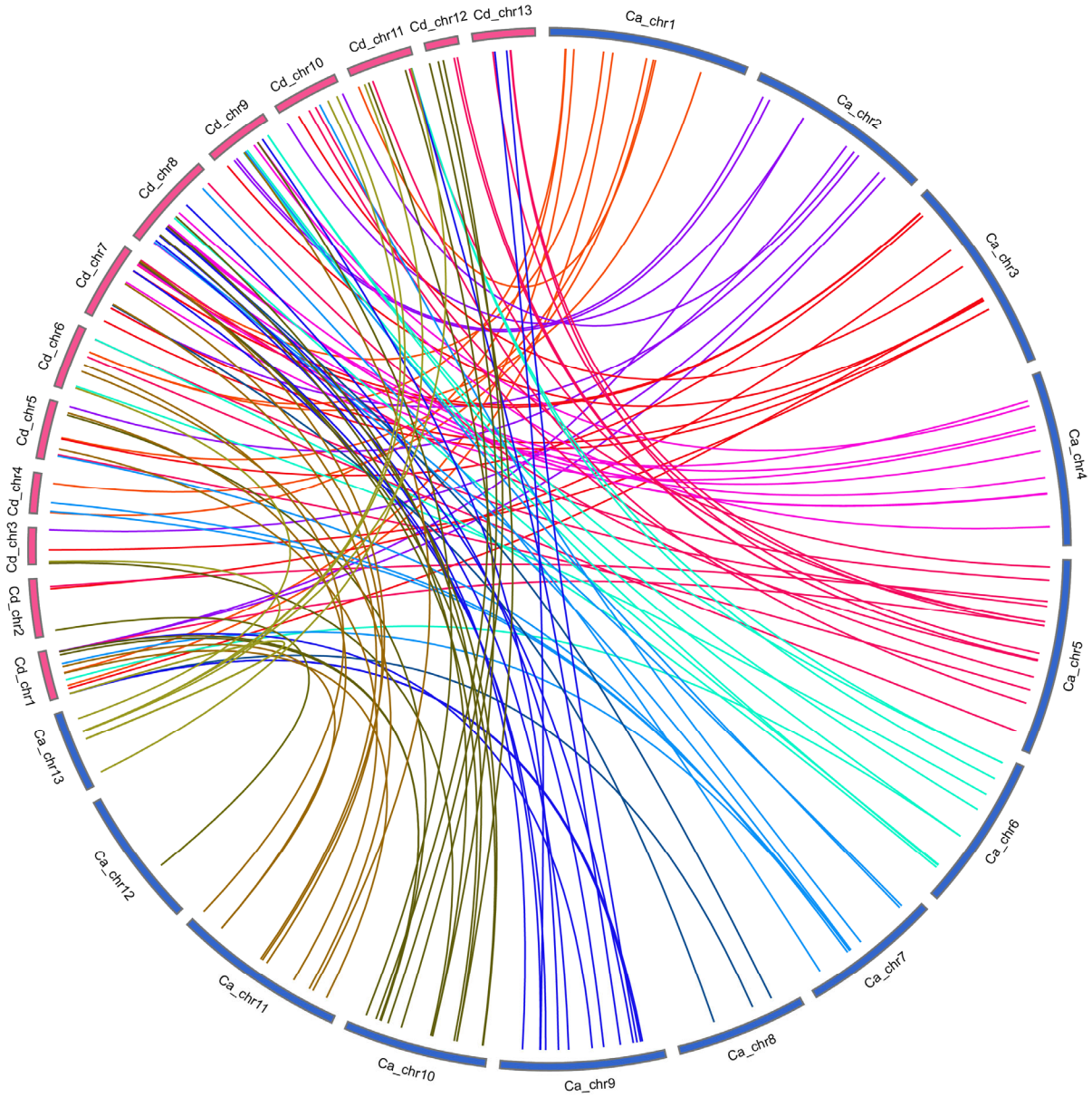

Figure 2 CIRCOS figure of NAC homologous genes pairs of *G. raimondii* and *G. arboreum* L. Lines represent homologous genes that are distributed in syntenic blocks between *G. raimondii* and *G. arboreum* L. chromosomes.

NAC-TF proteins in the eight genomes, as well as those from *G. raimondii* and *G. arboreum*. The 10 species were then divided into 18 subfamilies (Figure 4A). The subgroups are represented by roman numerals (NAC-I-NAC-XVIII) based on tree topology. Although the bootstrap values were relatively low because of the large number of sequences, the results were generally consistent with those of previous studies (Hu et al., 2010; Shang et al., 2013). In further investigation of cotton gene function and the evolutionary relationship of NAC, phylogenetic tree reconstruction showed that 401 *NAC* genes in *G. arboreum*, *G. raimondii*, and *A. thaliana* could be divided into 18 families (Figure 4B). In the XI subfamily, GaNAC100, GaNAC30, GaNAC98, GaNAC88, GaNAC26, GaNAC128, and GrNAC055, GrNAC080, GrNAC069, GrNAC032, GrNAC025, GrNAC009, GrNAC050, and *A. thaliana* NST1 (AT2G46770), *A. thaliana* NST3 (AT1G32770) were part of the same clade in the same subfamily.

Transcriptome analysis of NAC genes expressed during fiber development in diploid cotton

Overall, 88.9% (127 of 143) of the identified *NAC-TFs* genes of *G. arboretum* were supported by transcriptome sequencing data (Figure 5A). In addition, 87.58% (127 of 145) of *G. raimondii* genes were expressed during fiber development (0–15 DPA) (Figure 4B) and analyses of the *G. arboreum* transcriptome showed that the expression of 24 genes was greater than 10 reads per kilobase per million

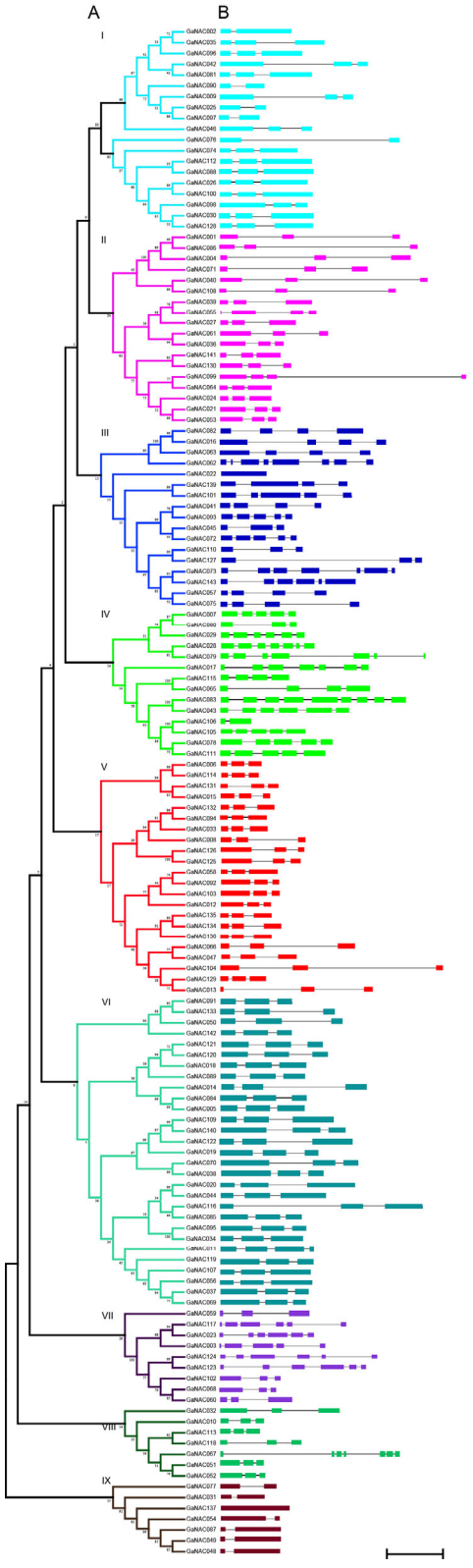

Figure 3 Phylogenetic relationship and gene structure of the *G. arboreum* L. *NAC* genes. A, A phylogenetic tree was constructed with MEGA 5.0 using the neighbor-joining (NJ) method with 1,000 bootstrap replicates based on a multiple alignment of 143 amino acid sequences of *NAC* genes from *G. arboreum* L. The nine major subfamilies are numbered I to IX. B, Exon/intron structure of *NAC* genes from *G. arboreum* L. Exons and introns are represented by boxes and black lines, respectively.

mapped reads (RPKM) in 15 DPA (Figure 5C). However, in *G. raimondii*, only 14 genes showed higher expression at 10–15 DPA (Figure 5D).

To examine the differential expression of homologous gene pairs between two diploid cotton species, the *NAC-TF* genes of *G. arboreum* were expressed over 10 RPKM at 15 DPA and their corresponding homologous gene pairs were used to construct a phylogenetic tree (Figure 6A). A total of 16 homologous gene pairs were selected for further investigation.

The results of the present study showed that the expression of homologous gene pairs varied, and they were more highly expressed in *G. arboreum* than in *G. raimondii* (Figure 6B). *GrNAC072* was not expressed at 15 DPA, whereas the level of *GaNAC099* expression was 88.1952 RPKM, gene structure, protein domain conservation, and sequence motifs were analyzed to determine the mechanism underlying differences in gene expression. The results showed that orthology largely affected gene structure and intron length between *GaNAC072* and *GrNAC082*, *GaNAC099* and *GrNAC072*, *GaNAC019* and *GrNAC121*, and between *GaNAC101* and *GrNAC091*, which was based on a single intron difference (Figure 6C). There were no major differences in the position and length of the conserved domain of the protein encoded by the homologous gene pairs (Figure 6D). Among these, the conserved domain of GaNAC019 was 20 aa longer than that of GrNAC121. In other homologous gene pairs, there was no effect on the conserved domain and motif (Figure 6E). By analyzing the promoters of homologous gene pairs, the promoters of 14 NAC homologous gene pairs were found to possess inserted or deleted fragments between the genome of *G. arboreum* and the genome of *G. raimondii*. For example, the promoters of GaNAC111 and GrNAC046 contained three inserted or deleted fragments (Figure S1A in Supporting Information), whereas the promoters of *GaNAC083* and *GrNAC043* only contained five (Figure S1B in Supporting Information). Therefore, we speculated that the promoter regions might have changed, which in turn resulted in differences in the expression between homologous gene pairs.

Validation of expression pattern divergence of *NAC* genes between two diploid cotton species using qRT-PCR

Using a combination of phylogenetic and transcriptome analyses, 16 homologous NAC gene pairs were selected and PCR was used to verify divergence in their expression (Figure 7). A total of 16 homologous NAC gene pairs exhibited differential expression patterns between *G. raimondii* and *G. arboreum*.

Four genes (*GaNAC30*, *GaNAC111*, *GaNAC24*, and *GaNAC50*) in *G. arboreum* were predominantly expressed in tissues, which further increased with the development of fiber cells. The highest expression levels were observed at 30 DPA, which were 400, 101- and 99-fold higher than the expression of histone-3. For homologous gene pairs

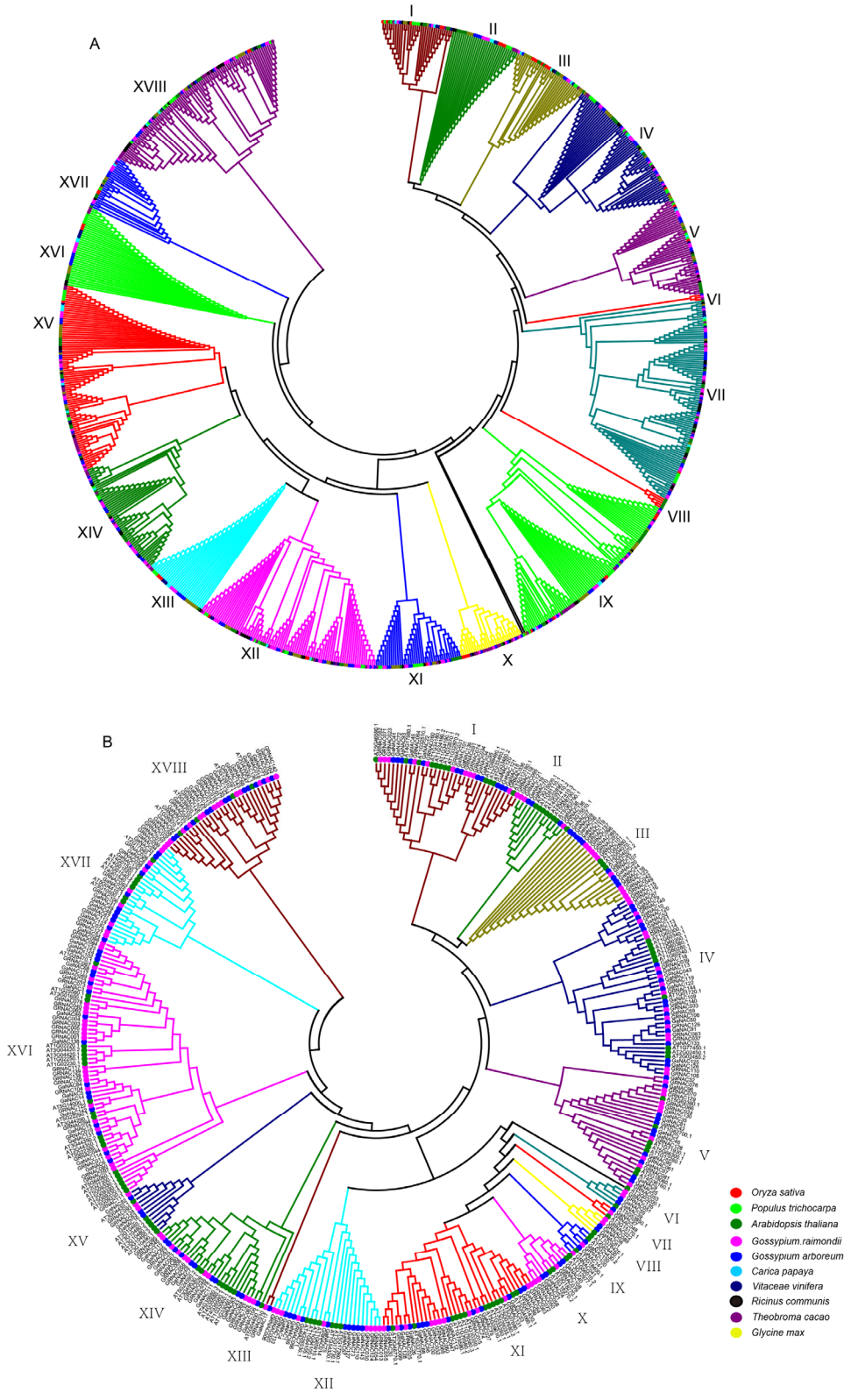

Figure 4 Phylogenetic tree of NAC-TFs. A, Phylogenetic tree of proteins containing the NAC domain from 10 species. The phylogenetic tree is based on sequence alignment of the N-terminal NAM domains of 882 NAC protein sequences from 10 genomes, *G. arboretum*, *G. raimondii*, *A. thaliana*, *O. sativa*, *V. vinifera*, *P. trichocarpa*, *G. max*, *T. cacao* L., *C. papaya*, and *R. communis*. The NAC proteins are grouped into 18 distinct clades (I–XVIII). B, Phylogenetic tree of NAC domain-containing proteins from *G. raimondii*, *G. arboretum*, and *Arabidopsis*. The phylogenetic tree is based on sequence alignment of the N-terminal NAM domains of 401 NAC protein sequences from three genomes, *G. arboreum*, *G. raimondii*, and *Arabidopsis*. The tree was generated with MEGA 5.03 using the NJ method. Bootstrap values from 1,000 replicates are indicated at each node. The NAC proteins are grouped into 18 distinct clades (I–XVIII).

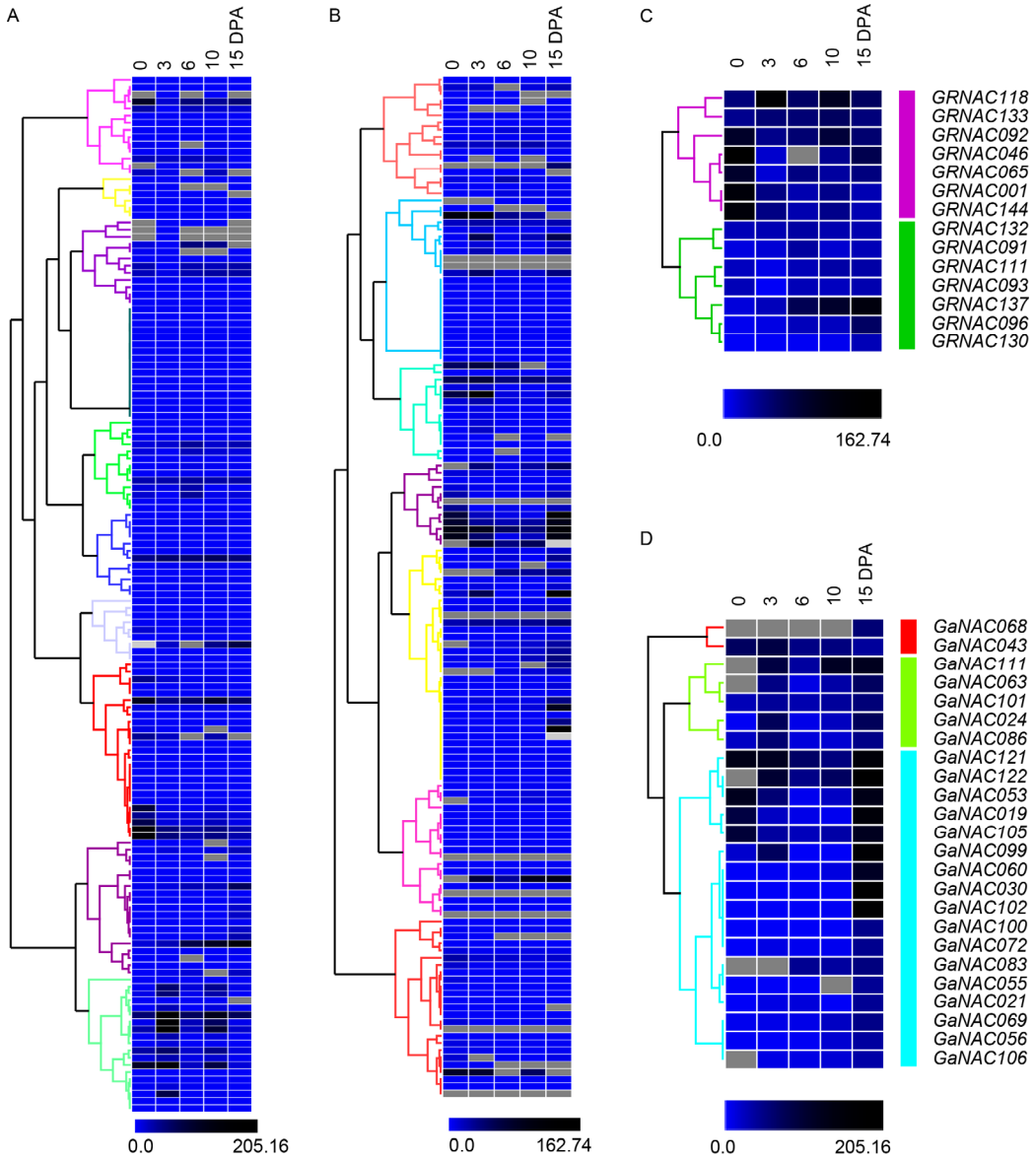

Figure 5 Expression patterns of the *NAC* gene family in *G. raimondii* and *G. arboreum* L.. A, Heat map showing the clustering of 145 *NAC-TF* genes of *G. raimondii* across five tissues (ovules at 0, 3, 6,10, and 15 DPA; noted at the top of each lane). B, Heat map showing the clustering of 143 *NAC-TF* genes of *G. arboreum* L. across five tissues (ovules at 0, 3, 6, 10, and 15 DPA; mentioned at the top of each lane). C, Expression of 14 (*G. raimondii*) *NAC-TF* genes is greater than 10 RPKM at 15 DPA. D, Expression of 24 *(G. arboreum) NAC-TF* genes is over 10 RPKM at 15 DPA. The color scale at the bottom of the dendrogram shows the relative expression levels. RNA-seq data under the accession number SRA180756 were obtained from the NCBI Sequence Read Archive (SRA) database.

(*GrNAC055*, *GrNAC046*, *GrNAC058*, and *GrNAC051*) in *G. raimondii*, the expression of *GrNAC046* and *GrNAC051* was similar, although slightly decreased than the expression level of the genes in *G. arboreum*. The expression level at 30 DPA of orthologous gene *GrNAC055* was similar to that of *GrNAC058*, which was subsequently followed by a marked decline.

Four genes (*GaNAC99*, *GaNAC121*, *GaNAC19*, and *GaNAC63*) predominantly expressed in *G. arboreum* fiber cells during secondary wall thickening (20–30 DPA) were expressed 60–70-fold higher than *histone 3*. On the other hand, low expression was observed in the hypocotyl, stem, and cotyledons, whereas the expression of homologous gene pairs *GrNAC072* and *GrNAC036* in *G. raimondii* ovule tissue gradually decreased during the development of fiber cells. *GrNAC121* was expressed in *G. raimondii* ovule tissues at all stages. The expression of *GrNAC093* in *G. raimondii* ovule tissues gradually increased during the development of fiber cells.

Eight genes (*GaNAC101*, *GaNAC56*, *GaNAC72*, *GaNAC83*, *GaNAC105*, *GaNAC102*, *GaNAC60*, and *GaNAC21*) were constitutively expressed in *G. arboreum*. The orthologs, *GrNAC091*, *GrNAC043*, *GrNAC065*, and *GrNAC050*, were also constitutively expressed in *G. raimondii*,

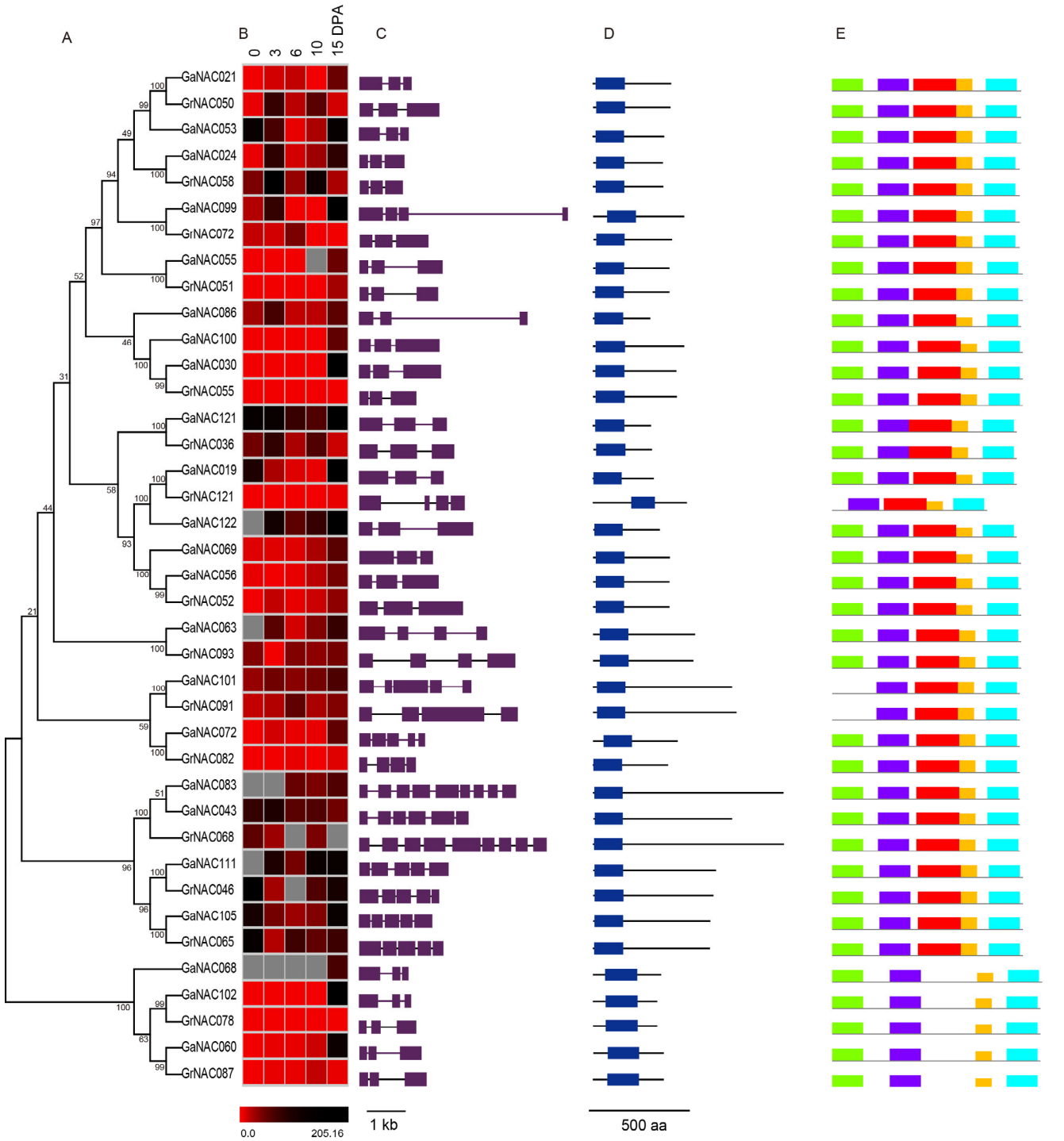

Figure 6 Analysis of *NAC* genes predominantly expressed in fiber. A, The phylogenetic tree was constructed with MEGA 5.0. B, Heat map showing the clustering of *NAC-TF* genes across five tissues (ovules at 0, 3, 6, 10, and 15 DPA; noted at the top of each lane). The color scale at the bottom of the dendrogram represents the relative expression levels. C, Exon/intron structures of *NAC* genes predominantly expressed in fibers. Exons and introns are represented by boxes and black lines, respectively. D, NAM domain of the NAC protein. E, Motif of the NAC protein.

albeit at lower levels than those in *G. arboretum*. Higher expression levels were observed for *GrNAC052*, *GrNAC072*, and *GrNAC082* genes in ovules, which were highest at 25 DPA and 30 DPA. *GrNAC087* was predominantly expressed in the hypocotyl of *G. raimondii*.

DISCUSSION

Cotton is one of the most widely used natural fibers for the production of clothes. Its highly elongated and thickened cell wall develops from the seed epidermis, resulting in a

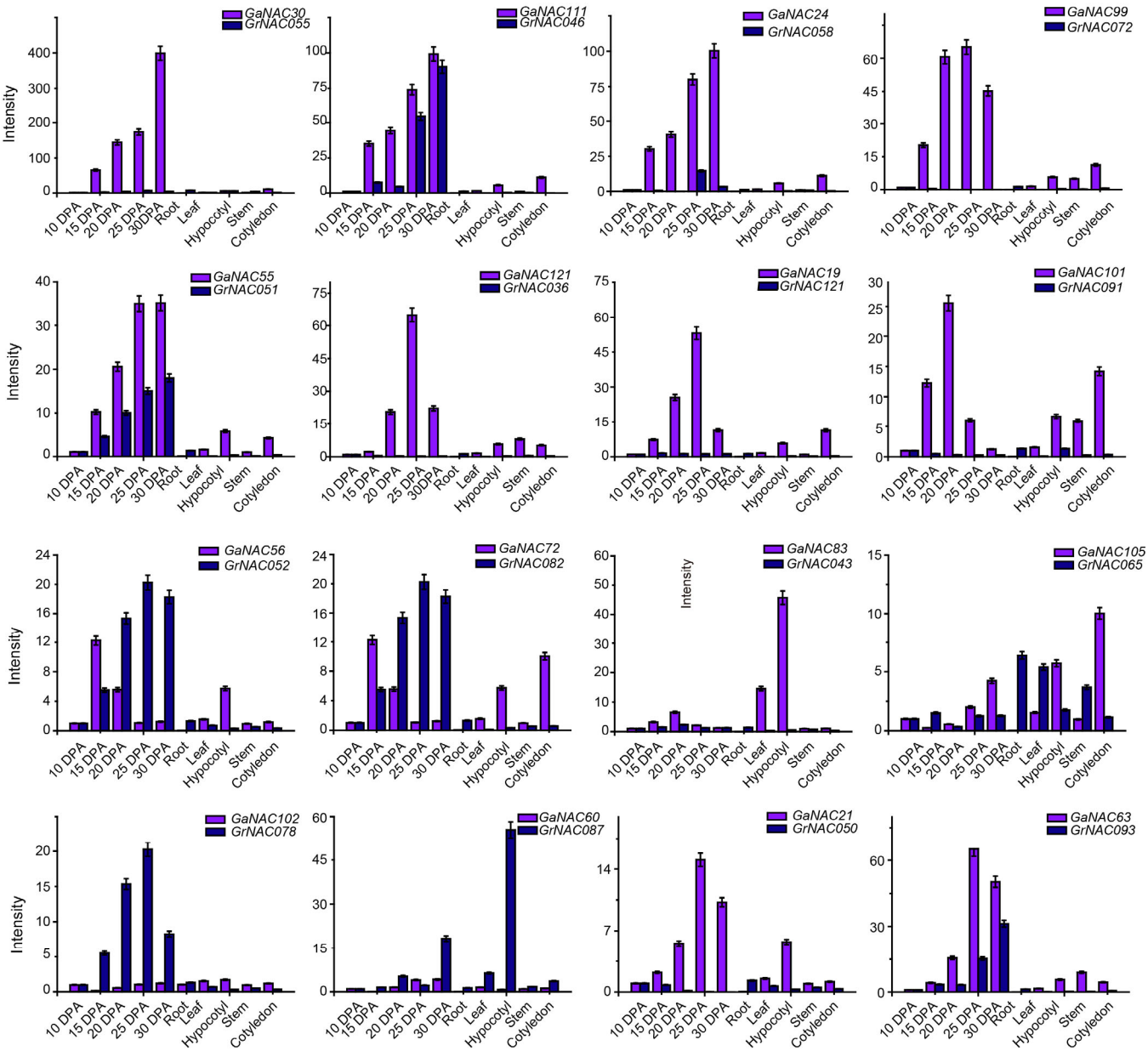

Figure 7 Expression analysis of sixteen selected *NAC* genes using RT-qPCR. The relative mRNA abundance of 16 selected *NAC* genes was normalized to the reference gene *histone 3* in different tissues. Bars represent the standard deviation of three technical replicates. 1–5, 10-, 15-, 20-, 25-, and 30-DPA fiber. 6, root; 7, leaf; 8, hypocotyl; 9, stem; 10, cotyledon.

cotton fiber that is commonly used as a model system for studying cell elongation and cell-wall biogenesis (Qin et al., 2011; Zhu et al., 2013). Sequencing of the cotton genome facilitates our understanding of the structure and functional evolution of gene families in *Gossypium* (Yao et al., 2012).

NAC-TFs are important transcriptional regulators that are involved in plant growth, development, and stress responses (Ooka et al., 2003; Pei. 2015). We identified 143 *NAC* genes in *G. arboreum* and 145 in *G. raimondii* by analyzing the NAC TF structure domain. Comprehensive analyses revealed that members of the NAC-TF family varied both in monocots or dicots (Ji et al., 2003; Rushton et al., 2008; Shen et al., 2009; Hu et al., 2010; Finn et al., 2011; Lee et al., 2012; Puranik et al., 2013; Shah et al, 2013; Shang et al., 2013; Zhu et al., 2013; Cao et al., 2015; You et al., 2015), for example, the distribution of NAC-TFs in dicots ranges from 79 in *C. paraya* to 187 in *E. grandis*. However, compared to the ~105 *NAC-TF*s in *A. thaliana*, *C. papaya*, and *T. cacao*, which belong to the malvid family, diploid cotton shows a significantly higher number of NAC-TFs. This expansion of NAC-TFs may result from WGD events, followed by multiple segmental and tandem duplication events (Wang et al., 2012; Shang et al., 2013; Li et al., 2014). This conclusion is consistent with findings of

previous reports on poplar (Hu et al., 2010; Cenci et al., 2014; Hussey et al., 2015).

Most NAC-TFs have been identified in angiosperms (Ohashi et al., 2010; Yamaguchi et al., 2010). To examine the phylogenetic relationship of NAC proteins in diploid cotton with those in dicots (*A. thaliana*, *V. vinifera*, *P. trichocarpa*, *G. max*, *T. cacao*, *C. papaya*, and *R. communis*) and monocots (rice), a phylogenetic tree was generated based on the alignment of its full-length NAC protein. A phylogenetic tree of the NAC gene from 10 dicots showed that a subclade exhibited a malvaceae lineage-specific phenomenon. In the phylogenetic analysis of NAC-TFs of Solanaceae, a subfamily unique to a Solanaceae lineage was also described (Rushton et al., 2008; Singh et al., 2013). Based on these findings, we speculate that this feature was derived from the evolution of NAC-TFs.

The results of previous studies have suggested that gene identification based on phylogenetic analysis is a reliable method of rationalizing systematic function predictions of different TF families (Le et al., 2011; Dong et al., 2013). In our studies, *Arabidopsis* VND1-7 (Zhong et al., 2006; Ohashi et al., 2010; Yamaguch et al., 2010 and 2011), and their orthologs in diploid cotton, were clustered in subfamily IV. Although no functional data are available yet, at least two GaNAC members are specifically expressed in developing fiber cells at 15 DPA (Figures 5A and D). At the same time, in the XI subfamily, GaNAC100, GaNAC30, GaNAC98, GaNAC88, GaNAC26, GaNAC128, and GrNAC055, GrNAC080, GrNAC069, GrNAC032, GrNAC025, GrNAC009, GrNAC050, *A. thaliana* NST1 (AT2G46770), and *A. thaliana* NST3 (AT1G32770) were grouped into a single clade. Because NST1 and NST3 are related to secondary xylem cell wall formation in *A. thaliana* (Mitsuda et al., 2005, 2007), it is possible that the cotton NAC gene on this branch is involved in the regulation of cotton fiber cell development. Subfamilies strongly associated with the transcriptional regulation of second wall formation also showed small-scale expansions, resulting in five NST orthologs in subfamily XVI. The changes of the expression of the NAC orthologous genes in the temporal and spatial expression were related to the secondary fiber cell-wall thickening-stage in the *G. raimondii* and the *G. arboretum*. The functions of these genes were different between in *G. raimondii* and *G. arboretum* might be due to the promoter region. The functional differentiation of homologous gene pairs may help us to understand the mechanisms underlying NAC-TF duplication. Based on the analysis of transcriptome data and expression patterns in *G. arboreum* and *G. raimondii*, the predominant expression of *GaNAC100, GaNAC30*, and *GrNAC055* in the fiber cell secondary wall were features of a single clade, and of NST1 and SND1 in the same subfamily (subfamily VI) (Zhong et al., 2007; Li et al., 2012). While cotton fiber cells were used as a model to study plant cell wall thickening, NAC-TFs predominantly expressed during secondary wall thickening clustered in the subfamily. We speculate that a subfamily predominantly regulating the secondary wall of cotton fiber may have arisen during the evolution of the *NAC-TF* gene family in diploid cotton. These genes will be the focus of future studies involving functional genomics and molecular breeding.

MATERIALS AND METHODS

Plant growth and collection of tissues

G. arboreum cv shixiya-1 plants were grown under standard field conditions in Anyang, Henan, China. Flower buds were tagged and the flowering day was recorded as 0 days post anthesis (DPA). Bolls were collected from plants at 0–15 DPA in the morning, and the fibers were isolated from the ovules, frozen in liquid nitrogen, and stored at −70°C until analysis. *G. raimondii* was planted in the National Wild Cotton Nursery, Sanya, China. The seedlings were grown in a greenhouse. Plant materials were collected, frozen in liquid nitrogen, and stored at −70°C until RNA extraction.

RNA-seq analysis

Total RNA was isolated from 3 g of cotton fiber harvested at 0, 5, 10, and 15 DPA using the hexadecyl trimethyl ammonium bromide (CTAB) cold-phenol method (Ji et al., 2003). Total RNA was then purified using a Nucleospin® RNA clean-up kit (MACHEREY-NAGEL, Düren, Germany). The integrity, concentration, and purity of RNA were assessed using an Agilent Bioanalyzer (Agilent Technologies, USA). Sequencing libraries were prepared following the manufacturer's standard instructions, and samples were sequenced on an Illumina HiSeq 2500 platform (Illumina, Inc., USA). Transcriptome data were analyzed using the CLC Genomics Workbench software 4 (www.clcbio. com/). Expression data were analyzed and visualized with the MeV software (Saeed et al., 2003).

Quantitative reverse transcription-polymerase chain reaction (qRT-PCR)

NAC-TF gene-specific primers for qRT-PCR were designed using the software Primer Premier 5.0. Total cDNA was synthesized from different cotton tissues using the PrimeScript® RT reagent kit (Perfect Real Time; Takara Biotechnology (Dalian) Co., Ltd., China). The *histone 3* gene (AF024716) was used as an internal control for the normalization of *NAC-TF* gene expression. All qRT-PCR samples were run on an Applied Biosystems 7900 Fast Real-Time PCR System (Grand Island, USA). Gene expression levels were calculated using the $2^{-\Delta\Delta C_T}$ method. The mean threshold cycle values for each *NAC-TF* gene were obtained from three independent PCR reactions.

Download of genomic data and identification of *NAC-TF* genes

We selected the genome annotation data of 10 sequenced plant species: *G. arboretum* and *G. raimondii* (http://cgp.genomics.org.cn), *A. thaliana* (http://www.arabidopsis. org/), *O. sativa* (http://rapdb.dna.affrc.go.jp), *V. vinifera* (http://www.genoscope.cns.fr/spip/Vitis-vinifera-e.html), *P. trichocarpa* (http://www.phytozome.net/poplar), *G. max* (http://www.phytozome.net/soybean), *T. cacao* L. (http:// cocoagendb.cirad.fr), *C. papaya* (http://asgpb.mhpcc.hawaii.edu), and castor bean (http://castorbean.jcvi.org) for *NAC-TF* gene identification. The Hidden Markov Model (HMM) profiles (PF02365) were downloaded from the Pfam database (http://pfam.xfam.org). *NAC-TF* gene families were identified using the HMMER 3.0 software package (Finn et al., 2011).

Phylogenetic analysis

The protein sequences of NAC-TFs were aligned using the ClustalX1.83 program, and protein alignment was manually adjusted. MEGA 5.03 software was used to construct a phylogenetic tree using the neighbor-joining method (Tamura et al., 2011).

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *This work was supported by the National High Technology Research and Development Program of China (2013AA102601), and the National Natural Science Foundation of China (31471538).*

Cao, X. (2015). Whole genome sequencing of cotton-a new chapter in cotton genomics. Sci China Life Sci 58, 515–516.

Cenci, A., Guignon, V., Roux, N., and Rouard, M. (2014). Genomic analysis of NAC transcription factors in banana (*Musa acuminata*) and definition of NAC orthologous groups for monocots and dicots. Plant Mol Biol 85, 63–80.

Dong, Z., and Chen, Y. (2013). Transcriptomics: advances and approaches. Sci China Life Sci 56, 960–967.

Fan, K., Wang, M., Miao, Y., Ni, M., Bibi, N., Yuan, S., Li, F., and Wang, X. (2014). Molecular evolution and expansion analysis of the NAC transcription factor in *Zea mays*. PLoS One 9, e111837.

Fang, Y., You, J., Xie, K., Xie, W., and Xiong, L. (2008). Systematic sequence analysis and identification of tissue-specific or stress-responsive genes of NAC transcription factor family in rice. Mol Genet Genomics 280, 547–563.

Finn, R.D, Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Re 39, W29–W37.

Hu, R., Qi, G., Kong, Y., Kong, D., Gao, Q., and Zhou, G. (2010). Comprehensive analysis of NAC domain transcription factor gene family in *Populus trichocarpa*. BMC Plant Biol 10, 145.

Hui,S., Yin, Y.B., Chen, F., Xu,Y., Richard, A,. and Dixon R.A. (2009). A bioinformatic analysis of NAC genes for plant cell wall development in relation to lignocellulosic bioenergy production. Bioenerg Res 2, 217–232.

Hussey, S.G., Saïdi, M.N., Hefer, C.A., Myburg, A.A., and Grima-Pettenati, J. (2015). Structural, evolutionary and functional analysis of the NAC domain protein family in *Eucalyptus*. New Phytol 206, 1337–1350.

Ji, S., Lu, Y., Feng, J., Wei, G., Li, J., Shi, Y., Fu, Q., Liu, D., Luo, J., and Zhu, Y. (2003). Isolation and analyses of genes preferentially expressed during early cotton fiber development by subtractive PCR and cDNA array. Nucleic Acids Res 31, 2534–2543.

Le, D.T., Nishiyama, R., Watanabe, Y., Mochida, K., Yamaguchi-Shinozaki, K., Shinozaki, K., and Tran, L.S. (2011). Genome-wide survey and expression analysis of the plant-specific NAC Transcription factor family in soybean during development and dehydration stress. DNA Res 18, 263–276.

Lee, S., Seo, P.J., Lee, H.J., and Park, C.M. (2012). A NAC transcription factor NTL4 promotes reactive oxygen species production during drought-induced leaf senescence in *Arabidopsis*. Plant 70, 831–844.

Li, F., Fan, G., Wang, K., Sun, F., Yuan, Y., Song, G., Qin, L., Ma, Z., Lu, C., Zou, C., Chen, W., Liang, X., Shang, H., Liu, W., Shi, C., Xiao, G., Gou, C., Ye, W., Xu, X., Zhang, X., Wei, H., Li, Z., Zhang, G., Wang, J., Liu, K., Kohel, R.J., Percy, R.G., Yu, J., Zhu, Y., and Yu, S. (2014). Genome sequence of the cultivated cotton *Gossypium arboreum.* Nat Genet 46, 567–572.

Li, Q., Lin, Y., Sun, Y., Song, J., Chen, H., Zhang, X., Sederoff, R.R., and Chiang, V.L. (2012). Splice variant of the SND1 transcription factor is a dominant negative of SND1 members and their regulation in *Populus trichocarpa*. Proc Natl Acad Sci USA 109, 14699–14704.

Li, W., Huang, G., Zhou, W., Xia, X., Li, D., and Li, X. (2014). A cotton (*Gossypium hirsutum*) gene encoding a NAC transcription factor is involved in negative regulation of plant xylem development. Plant Physiol Biochem 83, 134–141.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science 290, 1151–1155.

Mitsuda, N., Iwase, A., Yamamoto, H., Yoshida, M., Seki, M., Shinozaki, K., and Ohme-Takagi, M. (2007). NAC transcription factors, NST1 and NST3, are key regulators of the formation of secondary walls in woody tissues of *Arabidopsis*. Plant Cell 19, 270–280.

Mitsuda, N., Seki, M., Shinozaki, K., and Ohme-Takagi, M. (2005). The NAC transcription factors NST1 and NST2 of *Arabidopsis* regulate secondary wall thickenings and are required for anther dehiscence. Plant Cell 17, 2993–3006.

Nuruzzaman, M., Manimekalai, R., Sharoni, A.M., Satoh, K,, Kondoh, H., Ooka, H., and Kikuchi, S. (2010). Genome wide analysis of NAC transcription factor family in rice. Gene 465, 30–44.

Ohashi-Ito, K., Oda, Y., and Fukuda, H. (2010). *Arabidopsis* VASCULAR-RELATED NAC-DOMAIN6 directly regulates the genes that govern programmed cell death and secondary wall formation during xylem differentiation. Plant Cell 22, 3461–3473.

Ooka, H., Satoh, K., Doi, K., Nagata, T., Otomo, Y., Murakami, K., Matsubara, K., Osato, N., Kawai, J., Carninci, P., Hayashizaki, Y., Suzuki, K., Kojima, K., Takahara, Y., Yamamoto, K., and Kikuchi, S. (2003). Comprehensive analysis of NAC family genes in *Oryza sativa* and *Arabidopsis thaliana*. DNA Res 10, 239–247.

Pei, Y. (2015). The homeodomain-containing transcription factor, GhHOX3, is a key regulator of cotton fiber elongation. Sci China Life Sci 58, 309–310.

Pinheiro, G.L., Marques, C.S., Costa, M.D., Reis, P.A., Alves, M,S., Carvalho, C.M., Fietto, L.G., and Fontes, E.P. (2009). Complete inventory of soybean NAC transcription factors: sequence conservation and expression analysis uncover their distinct roles in stress response. Gene 444, 10–23.

Puranik, S., Sahu, P.P., Mandal, S.N., B, V.S., Parida, S.K., and Prasad, M. (2013). Comprehensive genome-wide survey, genomic constitution and expression profiling of the nac transcription factor family in foxtail millet (*Setaria italica* L.). PLoS One 8, e64594.

Qin, Y., and Zhu, Y. (2011). How cotton fibers elongate: a tale of linear cell growth mode. Curr Opin Plant Biol 14, 106–111.

Rushton, P.J., Bokowiec, M.T., Han, S., Zhang, H., Brannock, J.F., Chen, X., Laudeman, T.W., and Timko, M.P. (2008). Tobacco transcription factors: novel insights into transcriptional regulation in the Solanaceae, Plant Physiol 47, 280–295.

Saeed, A.I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braist-

ed, J., Klapa, M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, A., Trush, V., and Quackenbush, J. (2003). TM4: a free, open-source system for microarray data management and analysis. Biotechniques 34, 374–378.

Shah, S.T., Pang, C., Fan, S., Song, M., Arain, S., and Yu, S. (2013). Isolation and expression profiling of GhNAC transcription factor genes in cotton (*Gossypium hirsutum* L.) during leaf senescence and in response to stresses. Gene 531, 220–234.

Shang, H., Li, W., Zou, C., and Yuan, Y. (2013). Analyses of the NAC Transcription factor gene family in *Gossypium raimondii* Ulbr.: chromosomal location, structure, phylogeny, and expression patterns. J Integr Plant Biol 55, 663–676.

Singh, A.K., Sharma, V., Pal, A.K., Acharya, V., and Ahuja, P.S. (2013). Genome-wide organization and expression profiling of the NAC transcription factor family in potato (*Solanum tuberosum* L.) DNA Res 20, 403–423.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28, 2731–2739.

Wang, H., Zhao, Q., Chen, F., Wang, M., and Dixon, R.A. (2011). NAC domain function and transcriptional control of a secondary cell wall master switch. Plant J 68, 1104–1114.

Wang, K., Wang, Z., Li, F., Ye, W., Wang, J., Song, G., Yue, Z., Cong, L., Shang, H., Zhu, S., Zou, C., Li, Q., Yuan, Y., Lu, C., Wei, H., Gou, C., Zheng, Z., Yin, Y., Zhang, X., Liu, K., Wang, B., Song, C., Shi, N., Kohel, R., Percy, R.G., John, Z., Yu, J., Zhu, Y., Wang, J., and Yu, S. (2012). The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet 44, 1098–1103.

Wang, N., Zheng, Y., Xin, H., Fang, L., and Li, S. (2013). Comprehensive analysis of NAC domain transcription factor gene family in *Vitis vinifera*. Plant Cell Rep 32, 61–75.

Yamaguchi, M., Mitsuda, N., Ohtani, M., Ohme-Takagi, M., Kato, K., and Demura, T. (2011). VASCULAR-RELATED NAC-DOMAIN7 directly regulates the expression of a broad range of genes for xylem vessel formation. Plant J 66, 579–590.

Yamaguchi, M., Ohtani, M., Mitsuda, N., Kubo, M., Ohme-Takagi, M., Fukuda, H., and Demura, T. (2010). VND-INTERACTING2, a NAC domain transcription factor, negatively regulates xylem vessel formation in *Arabidopsis*. Plant Cell 22, 1249–1263.

Yao, D., Wei, Q., Xu,W., Syrenne, R.D., Yuan, J., and Su, Z. (2012). Comparative genomic analysis of NAC transcriptional factors to dissect the regulatory mechanisms for cell wall biosynthesis. BMC Bioinformatics 15, S10.

You, J., Zhang, L., Song B., Qi X., and Chan Z. (2015). Systematic analysis and identification of stress-responsive genes of the NAC gene family in *Brachypodium distachyon*. PLoS One 10, e0122027.

Zhang, G., Chen, M., Chen, X., Xu, Z., Guan, S., Li, L.C., Li, A., Guo, J., Mao, L., and Ma, Y. (2008). Phylogeny, gene structures, and expression patterns of the *ERF* gene family in soybean (*Glycine max* L.), J Exp Bot 59, 4095–4107.

Zhong, R., Demura, T., and Ye, Z. (2006). SND, a NAC domain transcription factor, is a key regulator of secondary wall synthesis in fibers of *Arabidopsis*. Plant Cell 18, 3158–3170.

Zhong, R., Richardson, E.A., and Ye, Z. (2007). Two NAC domain transcription factors, SND1 and NST, function redundantly in regulation of secondary wall synthesis in fibers of *Arabidopsis*. Planta 225, 1603–1611.

Zhu, T., Nevo, E., Sun, D., and Peng, J. (2012). Phylogenetic analyses unravel the evolutionary history of NAC proteins in plants. Evolution 66, 1833–1848.

Zhu, Y., and Li, F. (2013). The *Gossypium raimondii* genome, a huge leap forward in cotton genomics. J Integr Plant Biol 55, 570–571.

SUPPORTING INFORMATION

Figure S1 A list NAC gene family in *G. arboreum*.

Table S1 Homologous gene pairs botween *G. arboretum* and *G. raimondii*.

Table S2 Variation of promoter sequence of homologous gene pairs botween *G. arboretum* and *G. raimondii*.

The supporting information is available online at life.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.