GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Identification and analysis of the *TIFY* gene family in *Gossypium raimondii*

D.H. He[1]*, Z.P. Lei[2]*, B.S. Tang[1], H.Y. Xing[1], J.X. Zhao[1] and Y.L. Jing[1]

[1]College of Agronomy, Northwest A&F University, Yangling, Shaanxi, China
[2]College of Life Sciences, Northwest A&F University, Yangling, Shaanxi, China

*These authors contributed equally to this study.
Corresponding author: D.H. He
E-mail: daohuahe@nwafu.edu.cn

Genet. Mol. Res. 14 (3): 10119-10138 (2015)
Received December 8, 2014
Accepted May 4, 2015
Published August 21, 2015
DOI http://dx.doi.org/10.4238/2015.August.21.19

ABSTRACT. The highly conserved TIFY domain is included in the TIFY protein family of transcription factors, which is important in plant development. Here, 28 *TIFY* family genes were identified in the *Gossypium raimondii* genome and classified into *JAZ* (15 genes), *ZML* (8), *PPD* (3), and *TIFY* (2). The normal (TIF[F/Y]XG) motif was dominant in the TIFY family, excluding the ZML subfamily, in which TLSFXG was prevalent. *TIFY* family genes were unevenly distributed in the *G. raimondii* genome, with *TIFY* clusters present on chromosome 9. Phylogenetic analysis indicated abundant variations in the *G. raimondii* TIFY family, which were most closely related to those in *Theobroma cacao* among 5 species. Exon-intron organization and intron phases were homologous within each subfamily, correlating with their phylogeny. Intra-species synteny analyses indicated that genomic duplication contributed to the expansion of the *TIFY* family. Inter-species synteny analyses indicated that synteny regions involved in *G. raimondii TIFY* family genes were also present in the comparison of *G. raimondii vs Arabidopsis thaliana* or *T. cacao*, signifying that these genes had common ancestors and play the same or similar roles

in biological processes. Greater synteny was present in the comparison of *G. raimondii vs T. cacao* than of *G. raimondii vs A. thaliana*. The expression patterns of *TIFY* family genes were characterized and most *TIFY* family genes were indicated to be involved in fiber development. Our study provides new data related to the evolution of *TIFY*s and their role as important regulators of transcription; these data can be useful for fiber development.

Key words: Biological evolution; *Gossypium raimondii*; Phylogeny; Sequence analysis; Synteny; *TIFY* gene family

INTRODUCTION

Transcription factors (TFs), a class of proteins that exists in all cells, govern target gene transcription by binding to DNA near target genes (Wray et al., 2003). By interacting with some transcriptional regulators such as chromatin remodeling and modifying proteins, TFs facilitate or hinder RNA polymerase from approaching the promoters of target genes, hence activating or repressing expression. In plants, transcriptional regulation plays pivotal roles in developmental processes and responses to environmental stress (Libault et al., 2009).

The *TIFY* gene family is present only in plants and is functionally annotated as TFs. TIFY proteins possess a conserved and entire TIFY domain, which is characterized by an approximately 36-amino acid peptide and a highly conserved motif [TIF[F/Y]XG has variant forms, although a specific glycine is always conserved (Vanholme et al., 2007)] within this peptide. The TIFY domain is intimately involved in both hetero- and homomeric interactions between the TIFY proteins and other specific TFs (Chini et al., 2009). Several studies have indicated that TIFY proteins are involved in the control of various biological pathways in plants, including development and the response to various phytohormones interrelated with stress.

The TIFY family was previously known as zinc-finger protein expressed in inflorescence meristem (ZIM) (Nishii et al., 2000). TIFY can be divided into 4 subfamilies (JAZ, PPD, TIFY, and ZML) depending on the diverse domain architecture in their protein sequences (Vanholme et al., 2007; Bai et al., 2011). Subfamily TIFY contains only the TIFY domain. In addition to the TIFY domain, the subfamily jasmonate (JA) ZIM-domain (JAZ) bears a C-terminal Jas domain of 27 amino acids, which possesses the unique motif $SLX_2FX_2KRX_2RX_5PY$ (Staswick, 2008). Along with the TIFY domain, the subfamily ZML (including ZIM) is equipped with both a CCT domain (i.e., CONSTANS, CONSTANS-like, and TOC1) and a C2C2-GATA zinc-finger DNA-binding domain. Notably, the CCT domain N-terminal portion is similar in sequence to the Jas domain mentioned above. In the subfamily PPD, in addition to the TIFY domain with a motif TIFYSGK, it contains both a truncated Jas domain (lacking the conserved PY residue at its C-terminus) and a unique N-terminal PPD domain (Chung et al., 2009).

Currently, the *TIFY* family has been systemically analyzed in several plant species. The *TIFY* family includes 34 genes in soybean, 25 in poplar, 20 in rice, 18 in *Arabidopsis*, and 16 in grape (Vanholme et al., 2007; Ye et al., 2009; Zhu et al., 2013). Bai et al. (2011) conducted a systemic and exhaustive search of 14 genomes (not including *Gossypium raimondii*) in order to better determine the evolutionary history of TIFY proteins.

Information regarding the functions of *TIFY* family genes has been accumulating, particularly for *Arabidopsis*. Shikata et al. (2004) found that over expression of the *AtZIM* gene

led to petiole and hypocotyl elongation because of increased cell elongation. White (2006) found that the deletion of the *AtPPD1* or *AtPPD2* gene increased lamina size, the curvature of dome-shaped (instead of flat) of leaf, and number of stomata. The JAZ subfamily was found to include repressors (rather than TFs) of transcription related to JA by 3 independent groups (Thines et al., 2007). When plant cells contain bioactive JA at low levels, JAZs bind to and repress MYC2, which belongs to positive basic helix-loop-helix (bHLH) TFs, to prevent the transcription of JA-responsive genes. Both the process of development and stimuli can induce cells to synthesize and accumulate bioactive JA, which subsequently triggers degradation of JAZ proteins. Next, JAZ-mediated repression is relieved and then bioactive (de-repressed) MYC2 promotes the transcription of JA-responsive genes. Interestingly, to avoid exhaustive degradation of JAZ repressors and replenish the JAZ protein pool, bioactive MYC2 induces *JAZ* gene expression. Tissue development and stimuli such as drought, high salinity, and low temperature, can initiate the expression of *JAZ* genes and JA-responsive genes (Ye et al., 2009; Chacón-López et al., 2011; Seo et al., 2011; Zhu et al., 2013). In contrast, in healthy plant tissue, the expression of *JAZ* genes is also involved in regulating a diverse array of important developmental progresses, such as seed germination, root growth, and flower development (Wasternack, 2007). However, in these pathways and responses, *JAZ* expression is differentially regulated and induced to fine-tune the downstream JA-responsive genes (Demianski et al., 2012).

Cotton is the most important fiber crop worldwide. Fiber quality and productivity are affected by the reproductive developmental procedure of epidermal cells of cotton ovules. Tetraploid cultivated cottons originated from an ancient hybridization involving the ancestor of *G. raimondii*, which provided the D genome. Many studies have shown that the D-subgenome contributes more phenotypical variability to $A_tA_tD_tD_t$ tetraploids than does the A-subgenome (He et al., 2007). In addition, most quantitative trait loci controlling fiber quality were mapped on the D-subgenome (Jiang et al., 1998). Thus, some genes, particularly *TF*s, on the D-subgenome are thought to be involved in fiber development. However, very little information is available regarding the *TIFY* gene family in cotton because *Gossypium* genomics data are lacking. The *G. raimondii* genome was recently sequenced and is publicly available (Paterson et al., 2012; Wang et al., 2012a), enabling identification and analysis of the entire set of *TIFY* family genes in *G. raimondii*.

Because the plant-specific TIFY family has important functions in transcription regulation, we conducted a comprehensive survey of and characterized the *TIFY* family genes in *G. raimondii*. We identified 2 *TIFY*, 15 *JAZ*, 8 *ZML*, and 3 *PPD* genes. Moreover, syntenic and phylogenetic analyses revealed that both tandem and segmental duplication events contributed to the evolution of *TIFY* family genes in *G. raimondii*. To obtain insights into the involvement of *TIFY* genes in fiber development, we further characterized the expression pattern of *TIFY* family genes by mining the RNA-seq gene expression datasets, which are publicly available. Our results provide background data for further characterization of the evolution and function of *TIFY* genes within the genus *Gossypium* as well as contribute information to the future improvement of fiber yield and fiber quality by manipulating the expression of *TIFY* family genes.

MATERIAL AND METHODS

Identification of *TIFY* family genes in *G. raimondii*

The *G. raimondii* genome release (2.1) was downloaded from the website ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/G*raimondii*/ to construct a local database. Multi-

ple protein sequence alignment of conserved domains, such as pfam06200 (TIFY domain), pfam06203 (CCT motif), pfam09425 (CCT_2 domain, Jas), and pfam00320 (ZML), were acquired from NCBI (http://www.ncbi.nlm.nih.gov/cdd). Next, position-specific scoring matrix and hidden Markov model (HMM) profiles were built using the stand-alone version BLAST 2.2.26 release (ftp://ncbi.nlm.nih.gov/blast/executables/) and HMMER 3.0 (http://hmmer.janelia.org/). Searches of the position-specific scoring matrix and HMM profiles against the *G. raimondii* protein sequence database constructed based on the annotation details of the *G. raimondii* genome were conducted to explore TIFY domain-containing proteins using the RPSBLAST and hmmsearch programs. Against *G. raimondii* nucleotide sequence database, we searched for TIFY domain-encoding DNA sequences using RPSTBLASTN to determine whether any sequences were missed due to incomplete or erroneous annotation. Transcript data (see below) from RNA-sequencing were mined to examine and confirm the *TIFY* family genes encoding TIFY domain-containing proteins.

Using the program ps_scan.pl (ftp://ftp.expasy.org/databases/prosite), protein motifs of the putative *G. raimondii* TIFYs were then double-checked using PS51320 (TIFY domain profile), PS51017 (CCT domain profile), PS50114, and PS00344 (GATA-type zinc finger domain profile) to confirm the domains of the TIFY protein family.

Based on the absence or presence of the TIFY, CCT, Jas, or ZML domains in deduced amino acid sequences, the identified *TIFY* family genes were then classified into 4 subfamilies and named. To examine the conservation level of the TIFY and Jas functional domains in the *G. raimondii* TIFY protein family, we created sequence logos using the WebLogo program (Crooks et al., 2004).

Phylogenetic analysis

The multiple sequence alignments were generated using the ClustalW program (Chenna et al., 2003) with the Gonnet protein weight matrix. The gap opening/extension penalty was 10/0.1 for pairwise alignment and 10/0.5 for multiple alignment. Alignments were visualized using BioEdit V 7.1.3.0 and adjusted manually according to domain sequences. Using the MEGA 5.1 software (Tamura et al., 2011), evolutionary trees were inferred using the neighbor-joining algorithm and tested using the bootstrap method (1000 replicates). Next, protein sequences of *TIFY*s were collected from 5 other species: *Arabidopsis* (Vanholme et al., 2007), rice (Ye et al., 2009), grape (Zhang et al., 2012), poplar (Bai et al., 2011), and *Theobroma cacao*; we constructed 4 expanded phylogenetic trees for 4 subfamilies.

Exon-intron organization and sequence repeat analysis

The est2genome program (Rice et al., 2000) was used to align the coding sequences and the corresponding genomic sequences to determine the organization of exon-introns in the *G. raimondii TIFY* family genes. The online FancyGene program (Rambaldi and Ciccarelli, 2009) was used to visualize the exon-intron organization and locations of the domains in protein sequences. To explore whether interspersed sequence repeats and low complexity DNA sequences were included in the *TIFY* family genes, the sequences of *TIFY* family genes, including flanking sequences obtained by extending 3000 base pairs upstream and downstream, were checked using RepeatMasker (http://www.repeatmasker.org).

Large-scale duplication and synteny analysis of genomic region containing *TIFY* family genes

Tandem duplications (adjacent paralogous genes, without any intervening gene) and interspersed duplications of *TIFY*s were determined based on their positions on individual chromosomes. First, to explore paralogy/homology, protein-encoding genes from the *G. raimondii* genome were compared against the *G. raimondii* and other genomes (including *Arabidopsis* and *T. cacao*) using BLASTP (Wang et al., 2012b). Next, to identify putative paralogy/homologous chromosomal regions containing *TIFY* family genes, syntenic regions within the *G. raimondii*, between *G. raimondii*/*Arabidopsis* and between *G. raimondii*/*T. cacao* genomes were detected using the MCScanX program (Tang et al., 2008; Wang et al., 2012b). Additionally, the genomic region containing close tandem *TIFY*s in *G. raimondii* was selected for comprehensive inter-species synteny analyses.

Expression pattern analysis of *TIFY* family genes

To preliminarily explore *G. raimondii TIFY* family genes, the dataset from 3 RNA-seq experiments containing 17 runs were examined [from NCBI SRA databases, (Flagel et al., 2012)] to determine expression patterns (**Table S1**). Following curation, 10 comparisons were made between different developmental stages and between different tissues. Per the gff3 document (annotation information), genomic sequences of *TIFY* family genes were extracted and used as references, to which reads were aligned and mapped using the program TopHat (Trapnell et al., 2012). After mapping and assembly, Cuffdiff was used to identify differential expression of *TIFY* family genes across different developmental stages and in different tissues. The heatmap of expression profiles was produced using pheatmap in the R program.

RESULTS

G. raimondii TIFY protein families

Searches of the position-specific scoring matrix and HMM profiles against the *G. raimondii* nucleotide and protein sequence database resulted in the identification of 28 *TIFY* gene sequences encoding TIFY proteins (Table 1). To designate each protein member, the root symbol (i.e., subfamily designation) was followed by a sequential number. The proposed nomenclature for *G. raimondii TIFY* family genes, together with identifiers in ~8 x V2.10 annotation (Paterson et al., 2012) and protein length, and high-scoring segment pairs in the annotation described by Wang et al. (2012a), are listed in Table 1. All subfamilies of *G. raimondii TIFY* contain more than 1 gene [*GrJAZ* (15 genes), *GrZML*(8), *GrPPD* (3), and *GrTIFY*(2)].

We compared the number of *G. raimondii TIFY* family genes in each subfamily with that in other available plant genomes (Table 2). Among 4 subfamilies, *JAZ* was the largest and most abundant subfamily in each species, with 15 *JAZ* genes in *G. raimondii*, 23 in maize, 20 in *Glycine max*, and 9 in cacao. Additionally, the *ZML* subfamily was the 2nd most abundant. These 2 cases were consistent with the results in most other higher plants (Bai et al., 2011), such as *Arabidopsis*, rice, poplar, and cacao. The *PPD* subfamily underwent a species-specific expansion in *G. raimondii*, which was also observed in *Selaginella moellendorffii*.

Table 1. *Gossypium raimondii TIFY* family genes.

Gene name	Phytozome ID	Exon#	AA#	TIFY motif	Coordinates	HSP gene[1]	DP[2]
GrJAZ01	Gorai01G018800	5	228	TIFYGG	Chr01: 1752256...1754823	10015133	SC
GrJAZ02	Gorai02G021800	6	190	TIFYNG	Chr02: 1523101...1526607	-	SC
GrJAZ03	Gorai02G173700	7	362	TIFYAG	Chr02: 44407878...44410995	10019921	SC
GrJAZ04	Gorai04G285100	4	241	TIFYCG	Chr04: 61612558...61614463	10010230	SC
GrJAZ05	Gorai05G196200	6	197	TIFYNG	Chr05: 56877532...56879768	10001798	SC
GrJAZ06	Gorai06G092400	3	125	TIFYNG	Chr06: 32947665...32949260	-	SC
GrJAZ07	Gorai08G291000	7	371	TIFYAG	Chr08: 56545806...56551262	10030605	S
GrJAZ08	Gorai09G036800	4	240	TIFYGG	Chr09: 2718340...2720664	10037314	SC
GrJAZ09	Gorai09G039500	5	226	TIFYDG	Chr09: 2919762...2922783	10037287	SC
GrJAZ10	Gorai09G145400	4	226	TIFFGG	Chr09: 11017453...11020374	10031747	SC
GrJAZ11	Gorai09G154300	3	119	TIFYNG	Chr09: 11795808...11797204	10033602	SC
GrJAZ12	Gorai09G330500	7	365	TIFYAG	Chr09: 34149627...34152849	10005893	SC
GrJAZ13	Gorai10G090600	4	263	TIFYGG	Chr10: 13926370...13929049	10039035	SC
GrJAZ14	Gorai11G045300	3	120	TIFYNG	Chr11: 3453543...3454792	10031542	SC
GrJAZ15	Gorai11G062000	5	270	TIFFGG	Chr11: 5133243...5135195	10023394	SC
GrZML1	Gorai02G138300	9	356	TLSFEG	Chr02: 23533661...23537596	10038826	SC
GrZML2	Gorai05G097100	7	285	TLSFRG	Chr05: 14683753...14688216	10039389	SC
GrZML3	Gorai05G226000	11	390	TIAFEG	Chr05: 60845139...60850463	10000634	SC
GrZML4	Gorai08G017800	7	314	TLSFQG	Chr08: 1998002...2001547	10038217	SC
GrZML5	Gorai09G279200	5	282	TLSFRG	Chr09: 23544503...23549391	10023684	SC
GrZML6	Gorai10G033400	11	360	TLSFEG	Chr10: 3006376...3011496	10011724	SC
GrZML7	Gorai10G033500	7	296	TLSFRG	Chr10: 3023264...3027641	10011723	NS
GrZML8	Gorai13G055000	7	314	TLSFQG	Chr13: 5393873...5397619	10029545	SC
GrPPD1	Gorai04G077100	9	345	TIFYCGK	Chr04: 8931782...8937389	10020204	SC
GrPPD2	Gorai08G002400	9	341	TIFYCGK	Chr08: 410942...416444	10038066	SC
GrPPD3	Gorai11G168900	9	365	TIFYCGK	Chr11: 34616085...34621616	10031607	S
GrTIFY1	Gorai02G216400	3	140	TIFYAG	Chr02: 56617136...56618433	10004804	S
GrTIFY2	Gorai09G011200	6	427	TIFYGG	Chr09: 903341...907617	10016076	S

[1]HSP: High-scoring segment pairs (HSPs) show the alignments of the query and hit sequence from gene annotations of Wang et al. (2012a). Cotton_D_gene_10015133 was abbreviated to 10015133. [2]DP, Duplication; SC, located in syntenic (duplicated) genome region, and has the syntenic TIFY family **c**ounterpart; S: shown in syntenic regions, the syntenic TIFY family counterpart was lost; NS, not mapped within any **s**ynteny blocks.

Table 2. *TIFY* family genes found in genomes of several plants including *Gossypium raimondii*.

Subfamily	Ol	Ot	Pp	Sm	Mg	At	Pt	Vv	Gm	Gs	Pv	Mt	Gr	Tc	Os	Zm	Sb	Bd
JAZ	0	0	9	8	9	12	12	7	20	18	11	14	15	9	15	23	16	15
ZML	0	0	4	3	1	3	8	4	9	9	4	5	8	3	4	3	3	6
PPD	0	0	0	4	2	2	2	2	2	1	1	0	3	2	0	0	0	0
TIFY	0	0	3	2	1	1	3	2	3	6	3	2	2	1	1	1	0	0
Reference	A	A	B	B	B	B	B	B	B	C	D	B	-	-	B	B	B	B

At, *Arabidopsis thaliana*; Bd, *Brachypodium distachyon*; Gm, *Glycine max*; Gr, *Gossypium raimondii*; Gs, *Glycine soja*; Mg, *Mimulus guttatus*; Mt, *Medicago trunculata*; Ol, *Ostreococcus lucimarinus*; Os, *Oryza sativa*; Ot, *Ostreococcus tauri*; Pp, *Physcomitrella patens*; Pt, *Populus trichocarpa*; Pv, *Phaseolus vulgaris*; Sm, *Selaginella moellendorffii*; Sb, *Sorghum bicolor*; Tc, *Theobroma cacao*;Vv, *Vitis vinifera*; Zm, *Zea Mays*. Numbers of the genes in all plants (excluding Gr and Tc) are obtained from literature. A, Vanholme et al. (2007); B, Bai et al. (2011); C, Zhu et al. (2013); D, Aparicio-Fabre Rosaura et al. (2013).

HMMER 3.0 analysis was also conducted on the *G. raimondii* genome sequence released by Wang et al. (2012a) and 29 genes were found to belong to the *TIFY* gene family. The most homologous proteins from Paterson et al. (2012) are listed in Table 1 based on BLASTp. Two genes (*GrJAZ02* and *GrJAZ06*) had no explicit counterpart in gene annotation with the results of Wang et al. (2012a). Except for those listed in Table 1 (in the 7th column), 3 genes (*Cotton_D_gene_10001798*, *10031542*, and *10038825*) were found to belong to the *TIFY* family.

Multiple sequence alignment and the logo of the TIFY domain indicated that entire TIFY domain had diverse forms, but was conserved in the motif (TIF[F/Y]XG, Figure 1).

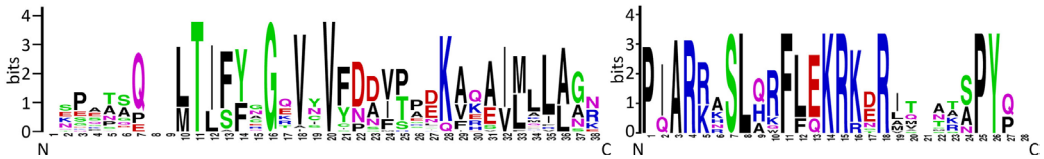

Figure 1. Sequence logo of TIFY (left) and Jas (right) domains in the family TIFY proteins from *Gossypium raimondii*.

Moreover, the TIFY motifs (only for the ZML subfamily) deviated from the typical motif (TIF[F/Y]XG) and included TLSFXG (7 proteins) and TIAFEG (1). These results indicate that TIF[F/Y]XG is dominant in the TIFY family [excluding the ZML subfamily, in which TLSFXG is dominant (Bai et al., 2011)]. In subfamily PPD, the TIFY domain contains a *Gossypium*-specific TIFYCGK rather than a general TIFYSGK (Chung et al., 2009) motif. A few amino acids are included in the consensus and thus were present in all of the *G. raimondii* TIFY proteins identified. In contrast, the logo of the Jas motif in *G. raimondii* showed a higher level of conservation. Particularly, at 11sites (1, 3, 4, 7, 8, 11, 14, 15, 18, 25, and 26), the amino acids observed were conserved. Additionally, conserved amino acid sequences of RX(2-5)PY were located at the C-terminal end of the Jas domain. This conserved region was thought to be a nuclear location signal and JAZ is located in the nucleus *in vivo* (Grunewald et al., 2009). The logo of the GATA zinc finger domain (data not shown) indicated that among 37 amino acids, 26 were highly conserved. At 29 of 45 sites in the CCT motif, amino acid residues were also strongly conserved, showing 100% identity.

Phylogenetic analysis of TIFYs in *G. raimondii* and additional 5 species

Amino acids in the TIFY family were diverse, even in the peptide region of the TIFY domain, in which the hydrophobic amino acids were more variable (Vanholme et al., 2007). Additionally, the number of amino acids in the *G. raimondii* TIFY protein sequences varied from 119-427 (Table 1). Therefore, phylogenetic analysis of the 28 *G. raimondii* TIFY members was performed according to the alignments of full-length protein sequences, which were adjusted exclusively according to conserved domain sequences to increase reliability. Figure 2 displays the *G. raimondii* TIFY family tree.

The phylogenetic tree illustrated that *G. raimondii TIFY* family genes were classified into 9 clades (marked by 9 red dots in Figure 2). JAZs and ZMLs were subdivided into 5 and 2 distinct clades, respectively. Clades 3 and 6 were the largest clades, with each including 5 genes. Clades 7 and 9 showed relatively low sequence identity. Members in different clades show great divergence in sequence, but those in the same clade were closely related. Evidence includes that proteins clustering together generally contained approximately the same number of amino acids. Additionally, TIFY family proteins from the same subfamilies appeared to be clustered together. Genes from subfamily TIFY diverged significantly from the homologs in the other 3 subfamilies, and were phylogenetically the most distantly related subfamilies.

To obtain information related to the functional relevance of abundant TIFY family members in *G. raimondii*, the phylogenetic relationships of the proteins in each subfamily were further analyzed between *G. raimondii* TIFYs and other TIFY family members from

5 species, including *Arabidopsis*, rice, poplar, cacao, and grape. The expanded phylogenetic trees for each subfamily were constructed (Figure 3).

The tree produced from the phylogenetic analyses of 6 species indicated that all members of the TIFY families in *G. raimondii* were most closely related to those in cacao compared to those in other species, and showed the lowest relationship with rice. The similarity coefficient between protein sequences were then checked and the results indicated that most (26/28 = 92.86%) *G. raimondii* TIFY members were more closely related to those in *T. cacao* (average of identity = 0.1857 for GrJAZ *vs* TcJAZ, 0.4929 for GrZML *vs* TcZML, and 0.5645 for GrPPD *vs* TcPPD), while the minority (2/28 = 7.14%) were more closely related to those in *Populus trichocarpa*.

Figure 2. Phylogenetic relationship (left), exon-intron organizations (middle), and the distribution of conserved domains in proteins (right) of *Gossypium raimondii* TIFYs. Phylogenetic analysis was performed using the neighbor-joining method. A phylogram was created using TIFY protein sequences and ClustalW2 multiple sequence alignment software. The tree was produced using the MEGA5.1 software. The numbers shown above or below tree branches show bootstrap values. Red dots on branches indicate 9 clades. Exon-intron organizations of *TIFY* family genes were drawn using FancyGene (http://bio.ieo.eu/fancygene/). The intron phase (number on intron) shows the intron location in a codon. Phase 0 indicate introns between 2 codons, while phase 1 shows introns between the 1st and 2nd bases of a codon. Phase 2 shows introns between the 2nd and 3rd bases. Colors indicate each conserved domains with relative position within each protein.

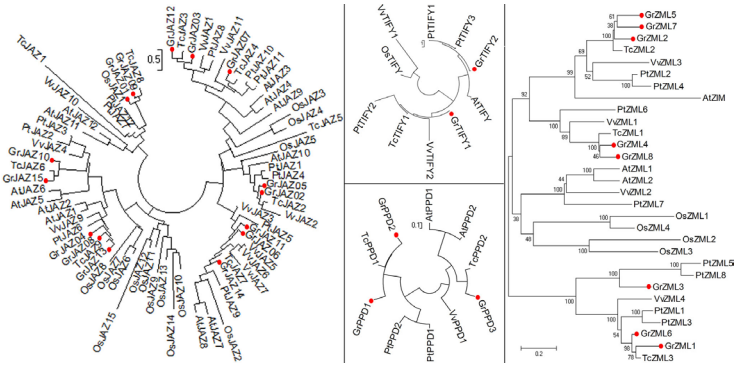

Figure 3. Phylogenetic analysis of sequences of TIFY proteins (4 subfamilies) from 6 species including *Gossypium raimondii* (Gr), *Arabidopsis thaliana* (At), *Oryza sativa* (Os), *Populus trichocarpa* (Pt), *Theobroma cacao* (Tc), and *Vitis vinifera* (Vv).

Exon-intron organization and sequence repeat analysis of *TIFY* family genes

The divergence of the exon-intron organization within the *TIFY* family played a critical role in the evolution of multiple gene families, and thereafter, phylogenetic groupings and evolutionary relationships were further supported (Shiu and Bleecker, 2003). To obtain further insights into structural evolution, the exon-intron organization in the *TIFY* family genes were investigated. Our results revealed that all members of the *TIFY* family possessed from 2-10 introns, and the most closely related genes within the same clade (even subfamilies) generally featured similar exon-intron organization in terms of either intron numbers or exon lengths (Figure 2), correlating the phylogeny illustrated by the tree. In the same clade, genes exhibited nearly identical exon lengths and intron numbers, a relatively constant exon-intron composition, and significantly different intron lengths. However, there were striking distinctions in the arrangement of introns among clades, supporting the results of phylogenetic analysis. These results, along with the moderate sequence identity, suggest that *G. raimondii TIFY* family members may have undergone gene differentiation after independent gene duplications throughout evolution to generate paralogs, resulting in the *TIFY* family featuring diverse exon-intron organization. The length ratio of the coding sequence *vs* the transcript region ranged from 16.2% (*GrJAZ02*) to 63.7% (*GrJAZ13*).

All *G. raimondii PPD* genes possessed a similar number (up to 8) of introns in their coding sequences, which was the same as that of *AtPPD* in *Arabidopsis* (Bai et al., 2011). However, the gene structures in *G. raimondii JAZ* subfamilies appeared to be more variable, displaying abundant variants in exon-intron organization, i.e., *G. raimondii JAZ* possessed 2-6 introns. Additionally, the most phylogenetically divergent *JAZ* genes (*GrJAZ06*, *GrJAZ11*, and *GrJAZ14*) displayed an exon-intron organization that was dissimilar to that of the other *GrJAZ* genes. For *G. raimondii ZML*, the number of introns in 8 members varied from 6-10. *GrZML1* was not identical to other *GrZML* genes with 6 or 8 introns. Overall, most *ZML* members contained 6 introns in their coding regions, which is the same as in grape *ZML* genes (Zhang et al., 2012). The gene structures of the *GrJAZ* subfamily were more divergent than those of the *GrZML* subfamily. Based on the exon-intron organization, the *GrJAZ* and *GrZML* (except *GrZML1*) genes were divided into 5 and 2 subgroups, respectively. Two *GrTIFY* genes were quite different in exon-intron organization, and perhaps only 1 conserved domain was present in the *TIFY* subfamily, which had more variability in its sequence. Furthermore, comparison of the exon-intron organization of *TIFY* family genes among several species, including *G. raimondii*, *Arabidopsis*, cacao, and grape, indicated that the exon-intron organizations of each subfamily were analogous across these 4 species (data not shown).

The relative positions and distribution of each conserved domain were constant within each protein sequence in the TIFY family. For some genes, introns were embedded in the genomic sequence region encoding conserved domain such as the TIFY domain and PPD domain. Particularly, *GrJAZ10* contained 1 intron in the 5' untranslated region. One intron was generally present between the TIFY domain and CCT_2 (Jas) domain-encoding sequence. CCT_2 was generally present in the final downstream exon, except *GrJAZ07*, which featured an unusually long intron in the 3' region and no intron embedded in the sequence encoding the CCT_2 domain. One intron was generally present between the CCT and GATA domain-encoding sequence, except for *GrZML2* and *GrZML8*. In the PPD subfamily, 2 introns were present in the region between sequences encoding the TIFY domain and modified Jas, and 2 other introns were present (nested) in the region between sequences encoding the PPD and

TIFY domains, except *GrPPD3* which contained a 3rd intron in that region.

We also investigated the intron phases with respect to codons. Eight *GrZML* genes showed 3 intron phase patterns, among which the difference was only in the length (0001100, 000110000, and 00011000000). Three *GrPPD* members showed an intron phase pattern (020102212). The *GrJAZ* genes showed 7 intron phase patterns, including a long pattern (0201021) with 3 genes (*GrJAZ3*, *GrJAZ07*, and *GrJAZ12*), while the remaining 6 patterns were short and appeared to be extracted from the long pattern (0201021) with conserved order. *GrTIFY* subfamily members had different intron phase patterns (000 and 020102, which appeared to be extracted from that of *GrJAZ*s or *GrPPD*s), similar to the long pattern of *GrJAZ*s and *GrPPD*s. The intron pattern of the *GrZML* genes was quite different from those of the *GrJAZ* and *GrPPD* genes. This supported that *GrZML*s *vs* *GrPPD*s subfamilies were less related than *GrJAZ*s *vs* *GrPPD*s.

An intron was inserted in the TIFY domain of 17 *TIFY* family genes, and most (14/17) were the phase 0 intron. Phase 0 and 1 introns were embedded within sections encoding the CCT domain of 3 and 6 *GrZML* genes, respectively. An intron was inserted in the Jas domain of 8 *GrJAZ* genes, and most (7/8) were the phase 1 intron. These results indicated a high level of conservation of the intron phase in the Jas and TIFY domains compared with other sections of *TIFY* family genes.

The output of the RepeatMasker program indicated that most (22/28) *TIFY* family genes contained simple repeats, implying that economical SSR markers can be adopted to mark and track these genes. Low-complexity sequence repeats (such A-rich and GA-rich) were present in transcribed genomic regions of 8 *TIFY* family members. Long terminal repeat retrotransposons (such as *gypsy* and *copia*) and long interspersed nuclear elements (such as L1) were embedded in the vicinity of 5 and 4 genes, respectively. Additionally, MuLE-*MuDR* and *PIF/Harbinger* were related to the up- and downstream regions of 3 and 1 genes, respectively. Sequence repeats including transposons make up more than 57% of the DNA in *G. raimondii* (Wang et al., 2012a).

Chromosomal distribution of *TIFY* family genes and duplication regions in the *G. raimondii* genome

The chromosomal locations of *TIFY* family genes were assigned based on the *G. raimondii* Genome Release 2.1 at Phytozome. *TIFY* family genes were present on all chromosomes (Chr), except Chr 03, 07, and 12 (Figure 4). The location distribution of *TIFY* family genes was similar in *G. raimondii*, *Arabidopsis thaliana*, and *Sorghum bicolor*, with no less than 7 *TIFY* genes present on a single chromosome (Bai et al., 2011). *TIFY* family gene clusters were distributed on Chr 09 in *G. raimondii*, Chr 01 in *A. thaliana*, and Chr 01 in *S. bicolor*. Particularly, tandem duplications present on *G. raimondii* Chr 09 and Chr 10 were similar to those on *S. bicolor* Chr 01 and *Oryza sativa* Chr03. In contrast, *TIFY* genes were generally dispersed in the soybean and grape genomes, with no more than 4 *TIFY* genes present on a single chromosome.

Tandem duplications typically resulted in gene family expansion (Cannon et al., 2004). In this study, close tandem duplications were found in the *ZML* subfamily (*GrZML6*/*GrZML7*). Interestingly, these close tandem duplications were also found in grape [*VvZML2*/*VvZML3* (Zhang et al., 2012)], poplar (*PtZML1*/*PtZML2*, *PtZML5*/*PtZML6*), *Medicago truncatula* (*MtZML1*/*MtZML2*), *G. max* (*GmZML2*/*GmZML3*, *GmZML4*/*GmZML5*), and *Brachypodium distachyon* (*BdZML4*/*BdZML5*). These collections are referred to as close tandem

duplications, which are defined as adjacent and paralogous genes within a chromosome without intervening genes. Additionally, *GrJAZ08* and *GrJAZ09* were found to be near tandem duplications, which were also found in rice (*OsJAZ9*/*OsJAZ10*/*OsJAZ11*), grape (*VvJAZ4*/*VvJAZ5*), and maize (*ZmJAZ5*/*ZmJAZ6*).

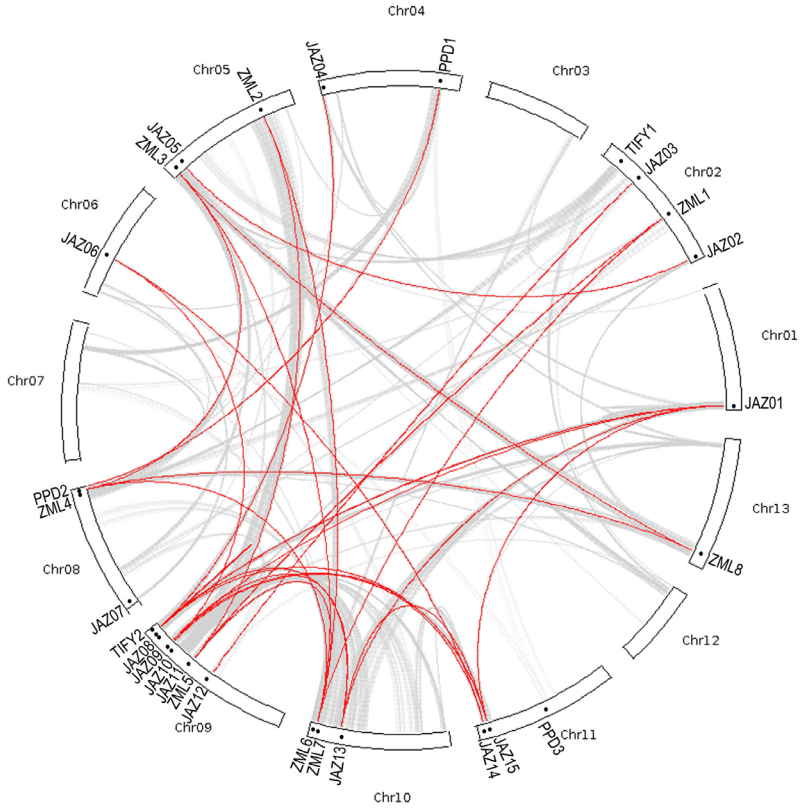

Figure 4. Distribution and synteny of *TIFY* family genes on *Gossypium raimondii* chromosomes. Chromosome (Chr) 01-13 are depicted as curve white bars. *TIFY* family genes are indicated by black dots. Gray curve lines denote syntenic regions containing *TIFY* family genes (no counterpart); red curve lines indicates syntenic regions (counterparts were present).

In addition to tandem duplications, dispersed segmental duplications were also important factors facilitating gene family expansion (Cannon et al., 2004). We then surveyed the duplicated regions within the *G. raimondii* genome using MCScanX and identified 23 *TIFY* family genes located within 79 pairs of duplicated (syntenic) blocks (Figure 4), which had a syntenic *TIFY* family counterpart. Therefore, most *TIFY* family genes may have evolved from direct genomic segmental duplications. Eight alignments (involving *GrTIFY1*, *GrTIFY2*, *GrJAZ07*, and *GrPPD3*) from synteny analysis revealed no syntenic *TIFY* family counterpart, possibly as a result of a lost gene or shuffle. *GrZML7* was the only gene that was not mapped into any synteny blocks. In summary, 96% (27/28) of *TIFY* genes were detected in synteny regions. No synteny was found among different subfamilies.

Among large-scale segmental duplicated regions, 3 collections contained more than one *TIFY* per region. One collection consisted of 5 genomic blocks (forming 5 alignments), including the block of *GrJAZ01-Gorai01G016500* on Chr 01, block of *Gorai04G285700-GrJAZ04* on Chr 04, block of *GrJAZ09-GrJAZ08* on Chr 09, block of *Gorai10G094100-GrJAZ13* on Chr 10, and block of *Gorai13G221600-Gorai13G222700* on Chr 13 (non-*TIFY* genes illustrated that the suppositional syntenic *TIFY*s counterpart should be near the loci). This collection belonged to pentaplicated regions that included 5 *TIFY*s. The other collection (4 alignments, tetraplicated) contained *GrZML3* (on Chr 05), *GrZML4* (on Chr 08), *GrZML8* (on Chr 13), and *Gorai12G143900* (on Chr 12). The 3rd collection (3 alignments, triplicated) contained *GrJAZ10* (on Chr 09), *GrJAZ13* (on Chr 10), and *GrJAZ15* (on Chr 11). Many studies have shown that during the long process of evolution, plant genomes underwent several rounds of large-scale-duplication events (Paterson et al., 2012). Subsequently, similar functional genes were partly retained. This viewpoint supports the expansion of many TF families through genomic duplication followed by rearrangement in the plant kingdom (Wang et al., 2012b). Genome-wide duplication events within *Gossypium* may account for the expansion of the *G. raimondii TIFY* family. The 3 collections supported that abundant *G. raimondii TIFY* family genes may have descended from massive genomic region expansion (such as duplication, triplication, and pentaplication) and diversification following duplication. In contrast, the genes descending from a single ancestral gene were typically included in a clade, therefore constituting a paralogous group.

After intra-species synteny analyses, we found that each gene cluster (excluding *TIFY* subfamily) originating from tandem or segmental duplication events generally shared similar exon-intron organizations with only minor differences. Indeed, further evaluation indicated that the *TIFY* family genes derived through duplication events contained exactly the same number of exons, which were similar to each other in exon length. This was supported by 4 sets of genes (*GrJAZ04*/*GrJAZ08*/*GrJAZ13*, *GrZML2*/*GrZML5*/*GrZML7*, *GrJAZ06*/*GrJAZ11*/*GrJAZ14*, and *GrPPD1*/*GrPPD2*/*GrPPD3*).

Evolutionary relationships of *TIFY* family genes between *G. raimondii* and other species

The genomic sequence comparison between different taxa provided the dataset for reconstructing the evolution of each gene (Koonin, 2005) and was used to infer the characterization of genes in taxa that had not been well studied, based on the abundant data from other well-studied taxa (Lyons et al., 2008). Many functions of *Arabidopsis TIFY* family genes have been well studied. Therefore, we deduced the roles of the *G. raimondii TIFY* family from *Arabidopsis* homologs identified by comparative genomics. Additionally, because the protein sequences of the TIFY family in *G. raimondii* bear the closest relationship with that in *T. cacao*, we compared the genomic sequence of *Arabidopsis vs G. raimondii* and cacao *vs G. raimondii*, followed by visualization of syntenic regions containing *G. raimondii TIFY* family genes using MCScanX (Tang et al., 2008; Wang et al., 2012b). Eighty and 43 large-scale syntenies involving *GrTIFY* family genes (26 and 27) were identified for *Arabidopsis vs G. raimondii* and cacao *vs G. raimondii*, respectively. However, only 16 and 23 *GrTIFY* family genes (involved in 25 and 30 syntenies) had counterparts in *Arabidopsis* and cacao, respectively (Figure 5). The biunique syntenies unambiguously indicated that the homologous gene-pairs evolved from the last common ancestor. Biunique gene-pairs (including *GrJAZ03-AtJAZ9*, *GrZML6-AtZIM*, *GrTIFY2-AtTIFY*, *GrJAZ03-TcJAZ3*, and *GrTIFY1-TcTIFY1*) were

scarce. Most syntenies involved many-to-one or many-to-many homologous relationships. For example, duplicated or triplicated *GrTIFY* genes corresponded to a single gene in *Arabidopsis* or cacao. The case included (*GrJAZ02*/*GrJAZ05*)-*AtJAZ10*, (*GrPPD1*/*GrPPD3*)-*AtPPD1*, and (*GrJAZ06*/*GrJAZ11*/*GrJAZ14*)-*TcJAZ7*, which supported the deduction that the common ancestral gene had expanded in *G. raimondii*, but not in *Arabidopsis* or cacao. Some ancestral genes did expand a same number of times in each species after speciation of *Arabidopsis*, cacao, and *G. raimondii* from the last common ancestor. This was supported by (*GrJAZ01*/*GrJAZ09*)-(*AtJAZ11*/*AtJAZ12*) and (*GrJAZ11*/*GrJAZ14*)-(*AtJAZ7*/*AtJAZ8*). Furthermore, there were cases where multiplicated *GrTIFY* genes corresponded to duplicated genes [such as (*GrPPD1*/*GrPPD2*/*GrPPD3*)-(*TcPPD1*/*TcPPD2*), and (*GrZML1*/*GrZML2*/*GrZML4*/*GrZML5*/*GrZML6*/*GrZML8*/*GrZML9*)-(*TcZML1*/*TcZML2*)] and multiplicated *GrTIFY* genes corresponded to triplicated genes [such as (*GrJAZ08*/*GrJAZ10*/*GrJAZ13*/*GrJAZ15*)-(*AtJAZ1*/*AtJAZ6*/*AtJAZ5*), and (*GrJAZ01*/*GrJAZ04*/*GrJAZ08*/*GrJAZ09*/*GrJAZ10*/*GrJAZ13*/*GrJAZ15*)-(*TcJAZ6*/*TcJAZ8*/*TcJAZ9*)]. Therefore, an ancestral gene has expanded more times in *G. raimondii* but fewer times in *Arabidopsis* and cacao during the evolution following speciation.

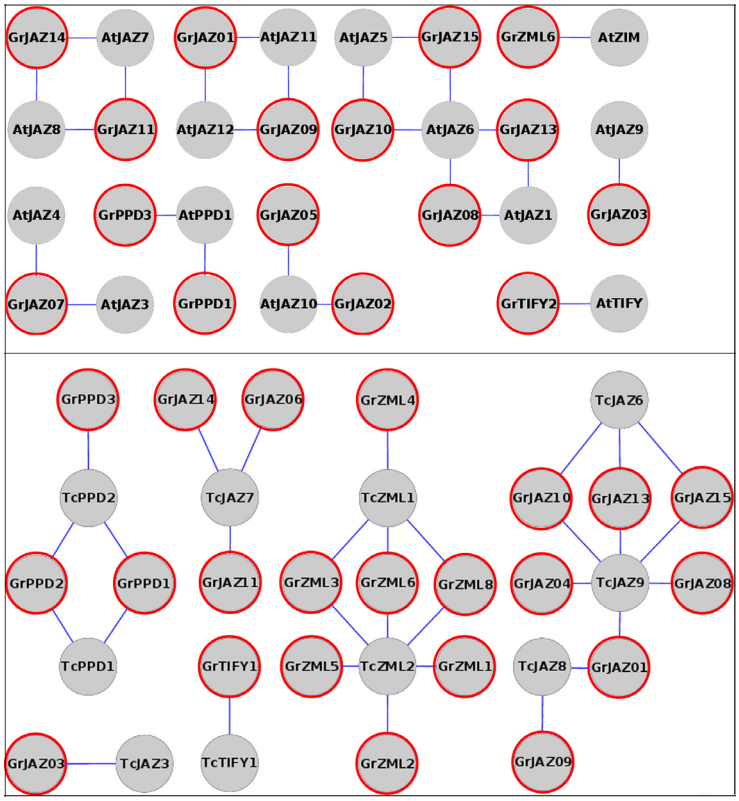

Figure 5. Network of the synteny between inter-species *TIFY* genes. The MCScanX software allows the detection of syntenic relationships existing between the *TIFY* family genes. Nodes represent genomic region containing the genes and the edges can be thought of as synteny/duplication relationships. The red nodes indicate genes of *Gossypium raimondii*, while other nodes indicate genes of *Theobroma cacao* (Tc) and *Arabidopsis thaliana* (At). Inter-species syntenies, which do not involve *GrTIFY* family genes or involve *GrTIFY* family genes without counterpart in other species, are not illustrated in this figure.

Except for the *GrTIFY* members mentioned above that had syntenic counterparts in other species, some genes, including *GrJAZ04*, *GrJAZ06*, *GrJAZ12*, *GrZML1*, *GrZML3*, *GrZML4*, *GrZML7*, *GrZML8*, *GrPPD2*, and *GrTIFY1*, were also embedded in the syntenic regions of *Arabidopsis vs G. raimondii*, whereas their respective syntenic *Arabidopsis* counterpart had been lost. Similarly, the comparison of cacao *vs G. raimondii* showed that 4 genes, including *GrJAZ02*, *GrJAZ05*, *GrJAZ07*, and *GrZML7*, which were also located in the syntenic regions, lost their respective syntenic cacao counterparts. Additionally, 2 genes, *GrZML2* and *GrZML5*, and 1 gene, *GrJAZ12*, had not been mapped into any synteny blocks of *G. raimondii_Arabidopsis* and *G. raimondii*_cacao, respectively. Perhaps, these *GrTIFY* members did not share a common ancestral gene with *Arabidopsis* or cacao. The failure of the syntenies identification may have resulted from selective gene loss, which typically followed chromosomal rearrangement (fracture and fusions) during evolution (Wang et al., 2012a) after speciation of *G. raimondii*, *Arabidopsis*, and cacao. Further studies are necessary to explain the syntenic observations.

In general, inter-species synteny identifications, involving *GrTIFY* members, indicated that syntenic regions of *G. raimondii*_cacao (up to 310 genes) were longer than that of *G. raimondii_Arabidopsis* (no more than 40 genes). Furthermore, syntenic regions of *G. raimondii*_cacao involved more *GrTIFY* members than those of *G. raimondii_Arabidopsis*. These results also implied that *G. raimondii* was more closely related to cacao than to *Arabidopsis*, in agreement with the phylogenetic analysis described above.

As described above, 2 *GrZML* members, *GrZML6* and *GrZML7*, were closely tandem duplicates of each other on Chr 10. This was also observed in several species. Thus, the genomic region containing *GrZML6* and *GrZML7* was extracted to carry out comprehensive inter-species synteny detection. The results indicated that syntenies involving the *GrZML6-GrZML7* region were present in several genomic comparisons, such as *P. trichocarpa_G. raimondii*, *G. max_G. raimondii*, *T. cacao_G. raimondii*, *V. vinifera_G. raimondii*, and *M. truncatula_G. raimondii* (**Table S2**). We then investigated the genes in the vicinity of *GrZML6* and *GrZML7* to gain additional insight into the level of conservation. The data indicated that homologs of *Gorai10G033700*, *Gorai10G033900*, and *Gorai10G034400* were present in the comparisons to several species, including poplar, soybean, and cacao. Additionally, the syntenies containing the near-tandem *GrJAZ08* - *GrJAZ09* region were also present in 4 comparisons (3 in *G. raimondii_G. max*, 1 in *G. raimondii_T. cacao*, 2 in *G. raimondii_M. truncatula*, and 1 in *G. raimondii_P. trichocarpa*), with counterparts of *GrJAZ08* and *GrJAZ09* present. Attention should be given to *Gorai09G036700*, which always linked with *GrJAZ08* among the 10 species. Several enzymes, including the TIFY family members, encoded by closely linked genes, involved in a metabolic pathway co-evolved similarly in each species.

Expression profiles of *TIFY* family genes

In plants, *TIFY* family genes play important roles in biological development and the adaptation to diverse stresses (Vanholme et al., 2007; Zhang et al., 2012). Most *TIFY* family genes appear to have a different spatiotemporal pattern of expression within different genera and species (Bai et al., 2011). Therefore, functional characterization of *TIFY* provided basic data for further improvement of plant production quantity and quality. In this study, the expression profiles of *G. raimondii TIFY* family genes were studied. The heatmap of expression profiles (Figure 6) showed that most genes had different and broad expression patterns across

different spatiotemporal domains. Twenty-six of these *TIFY* family genes were expressed differentially in a minimum of 1 of the 3 experiments. Detailed differences are shown in **Table S3** and **Figure S1**.

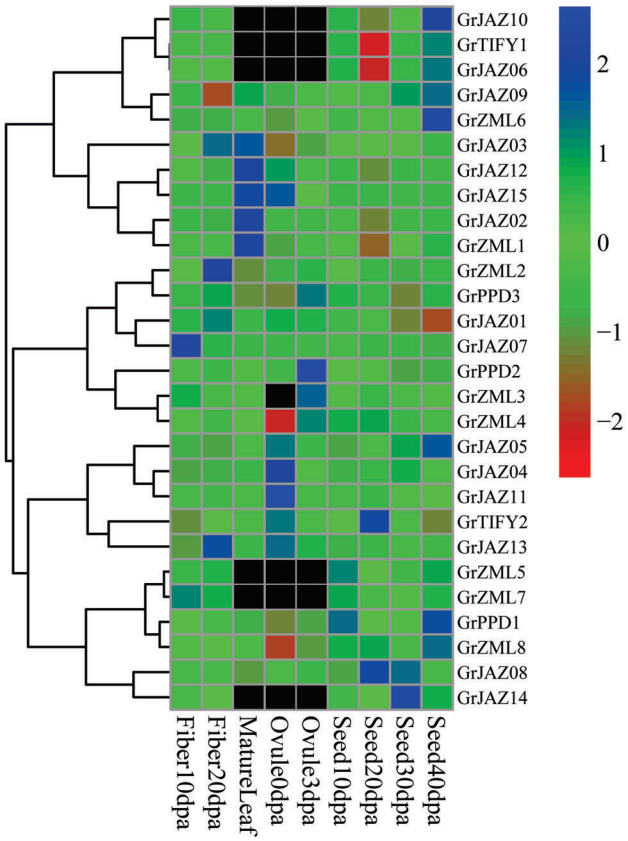

Figure 6. Expression profile of *TIFY* family genes. Levels of gene expression are depicted in different color on the right scale. Black: missing value.

Fiber growth is an important developmental process controlling the productivity and quality of lint, which shares many similarities with *A. thaliana* leaf trichome development (Wan et al., 2014). This process includes 4 overlapping stages: initiation [0-8 days post anthesis (dpa)], elongation (i.e., primary cell wall synthesis, 3-17 dpa), secondary cell walls synthesis (17-23 dpa), and maturation (after 20 dpa) (Graves and Stewart, 1988). In contrast, by interacting with some factors, JAZ proteins modulated trichome initiation (Qi et al., 2011). Thus, the expression profiles of the *TIFY* family were examined in different tissues (leaves *vs* ovules) and at different developmental stages of fibers (including 0, 3, 10, 20, 30, and 40 dpa; **Table S1**, **Figure S1**). No differences in transcriptional abundance were detected for the genes *GrJAZ06* and *GrTIFY1* among all stages of fiber growth. Data from no less than 2 experiments indicated that 21 genes, except *GrJAZ10*, *GrJAZ14*, *GrZML1*, *GrZML5*, and *GrZML7*, exhibited significant expression differences during different stages.

Regarding the comparison of different tissues, the transcription levels of 8 genes (*GrJAZ02*, *GrJAZ03*, *GrJAZ09*, *GrJAZ12*, *GrZML1*, *GrZML4*, *GrZML6*, and *GrZML8*) in vegetative tissues (leaves) were much higher than those in reproductive tissues (0 and 3 dpa ovules), whereas 6 genes (*GrPPD2*, *GrJAZ05*, *GrJAZ07*, *GrJAZ08*, *GrJAZ13*, and *GrZML2*) showed the opposite pattern. Additionally, 3 genes (*GrJAZ05*, *GrJAZ11*, and *GrTIFY2*) showed significant changes with obscure direction in expression levels between vegetative tissues and reproductive tissues. For fiber growth, at the stage of fiber initiation (3 *vs* 0 dpa), the expression abundance of 12 genes (*GrJAZ01*, *GrJAZ02*, *GrJAZ04*, *GrJAZ05*, *GrJAZ08*, *GrJAZ09*, *GrJAZ11*, *GrJAZ12*, *GrJAZ13*, *GrJAZ15*, *GrPPD1*, and *GrTIFY2*) decreased, while the expression abundance of 3 genes (*GrJAZ07*, *GrPPD3*, and *GrZML6*) increased. Particularly, 7 genes (*GrJAZ04*, *GrJAZ05*, *GrJAZ08*, *GrJAZ12*, *GrJAZ13*, *GrJAZ15*, and *GrTIFY2*) showed expression changes at abs [log2 (fold change)] > 1, indicating that these genes varied by greater than 2-fold in the expression level. Remarkably, dynamic and strong changes were observed in the expression of 2 genes, *GrJAZ04* and *GrJAZ12*. During the secondary cell wall thickening stage, the expression level of 3 genes (*GrJAZ07*, *GrJAZ09*, and *GrJAZ12*) decreased, while that of 3 genes (*GrJAZ08*, *GrJAZ13*, and *GrTIFY2*) increased. During the fiber mature stage, the expression level of 2 genes (*GrJAZ01* and *GrTIFY2*) always decreased, whereas that of 5 genes (*GrJAZ05*, *GrJAZ09*, *GrPPD1*, *GrZML6*, and *GrZML8*) always increased. The expression levels of 4 genes (*GrJAZ03*, *GrJAZ04*, *GrJAZ08*, and *GrJAZ14*) increased and then decreased, whereas that of 2 genes (*GrZML2* and *GrZML7*) decreased and then increased.

Overall, most *TIFY* family genes showed development-dependent expression profiles in fiber cells, implicating these genes in various physiological processes during fiber growth. Additionally, 6 genes (*GrJAZ01*, *GrJAZ04*, *GrJAZ08*, *GrJAZ09*, *GrTIFY2*, and *GrZML6*) showed altered expression levels during all stages of fiber growth. Synthesizing RNA-seq expression profiles from 3 experiments showed that the transcriptional abundance of 3 genes (*GrJAZ06*, *GrTIFY1*, and *GrZML1*) were consistently comparatively low. Conversely, 3 genes (*GrJAZ09*, *GrJAZ01*, and *GrJAZ08*) were robustly transcribed and were ubiquitously and differentially expressed in all comparisons. Specifically, *GrJAZ09* always showed the highest abundance among this family. In summary, most *TIFY* family genes played a role in both developmental transitions and fiber growth in cotton. Several genes (*GrJAZ09*, *GrJAZ01*, and *GrJAZ08*) should be further examined for future molecular design and fiber improvement.

DISCUSSION

Member of TIFY family are important TFs, and have been identified in many species. However, little information has been accumulated about this family in *Gossypium*. *G. raimondii* is an important diploid and wild species, whose ancestors contributed the D genome to the current main cultivated species by ancient interspecific hybridization. The release of *G. raimondii* genome sequences laid the foundation of comprehensive identification and characterization of the *TIFY* family in the important diploid species. Identification and analysis of the *TIFY* gene family in *G. raimondii* will provide valuable asset for the improvement of tetraploid cultivated species (*G. hirsutum* and *G. barbadense*).

Identification of the *TIFY* family genes

In this study, we identified 28 *TIFY* genes in *G. raimondii* genome, among which 15

belong to the *JAZ* subfamily, 8 to the *ZML* subfamily, 3 to the *PPD* subfamily, and 2 to the *TIFY* subfamily. Compared with other TFs identified in *G. raimondii*, such as *bHLH* (208 genes) and *MYB* (219 genes) (Wang et al., 2012a), the *TIFY* family is not a large family. Compared with other comprehensively surveyed plant *TIFY*s [34 *TIFY* family genes in soybean, 27 in maize, 25 in poplar, 20 in rice, 18 in *Arabidopsis*, and 15 (unpublished data) in cacao], the *G. raimondii TIFY* family ranks second with 28 phylogenetically expanded genes. Additionally, comparison with other plants indicated that the striking expansion and diversification of the *TIFY* family (28 copies), perhaps along with 3 *PPD* gene copies in *G. raimondii*, suggests that these TIFYs play crucial roles in the environmental adaptability and physiological maintenance of *G. raimondii*, which can survive in arid and stressful environments.

In order to absolutely identify all *TIFY* genes in *G. raimondii*, we also recur to other means. For example, after synteny detection based on the arrangement of genes on chromosome (dispersed segmental duplications), we conducted dc-megablast (database: *G. raimondii* genome; query: exons of *TIFY* family genes; e-value: 10^{-6}) to explore the presence of evolving pseudo genes or relics in *TIFY*s. We found no other genomic regions encoding TIFY family members. Twenty-eight *TIFY* genes were also supported by transcript data from RNA-sequencing. Thus, there were 28 members of *TIFY* family in *G. raimondii*.

Evolution of the *TIFY* family genes

Phylogenetic analysis and comparative genomic analyses are usually conducted to gain insight into evolutionary relationships of several genes or species. The phylogenetic tree of *G. raimondii TIFY* family genes indicated that the members of the JAZ subfamily were classified into several clades, agreeing with that in *Arabidopsis* (Vanholme et al., 2007), rice (Ye et al., 2009), and grape (Zhang et al., 2012), but the topology was not similar to those produced in these species. For example, the tree and the mean distances among subfamilies indicated that the relationship (distance, 5.487) of PPDs *vs* ZMLs was much greater than that (4.505) of PPDs *vs* some JAZs, or that (3.372) of ZMLs *vs* other JAZs. However, the tree described by Zhang et al. (2012) indicated that *VvPPD*s and *VvZML*s clustered together in the grape, while the tree described by Vanholme et al. (2007) indicated that *AtZML*s were not closely related to the collection of *AtPPD*s and *AtJAZ*s, which clustered together in *Arabidopsis*. Therefore, the *TIFY* family in *G. raimondii* did not evolve in concert with that in other species. The tree of 6 species indicated that *Gossypium* was more closely related to the *T. cacao* genome (suggesting that they perform similar roles), which agreed with the results of Wang et al. (2012a), but did not show that *Gossypium* was closely related to the dicotyledonous *Arabidopsis* genome (Lin et al., 2010). Thus, inferring the role of the TIFY protein from the model plant *Arabidopsis* to *Gossypium* should be done with caution, although TIFYs in *G. raimondii* were mainly identified by homology. However, transferring the information between perennial WOODY species, such as *G. raimondii* and *T. cacao*, was possible.

From the perspectives of the primary structure of a peptide or protein, phylogenetic tree provided the information of evolutionary relationships. Additionally, from the arrangement of genes on chromosome, comparative genomic analyses (including intra-species and inter-species) can provide further insight into evolutionary relationships. Intra-species synteny analyses indicated that most *TIFY* genes were located in synteny regions, which implied that large-scale genomic duplication contributed to the expansion of the *TIFY* family. This kind of large-scale-duplication events had been undergone by many species in the plant kingdom

(Paterson et al., 2012). Unneglectably, a few *TIFY* genes escaped from intra-species synteny blocks. The cases may result from the rearrangement and diversification following duplication (Wang et al., 2012a). Inter-species synteny analyses indicated that synteny regions containing *G. raimondii TIFY* family genes were also present in the comparison of *G. raimondii vs A. thaliana* or *T. cacao*, signifying that these genes or chromosome segment had common ancestors and play the similar and concerted roles in biological processes across species. Additionally, results of comparative genomic analyses are in agreement with that of phylogenetic analysis.

In summary, genome-wide identification and analysis of the *TIFY* family gene in *G. raimondii* has been carried out in this study. Twenty-eight *TIFY* genes are identified, which are diversity across clades of phylogenetic tree. The expansion of the *TIFY* family in *G. raimondii* may result from tandem and segmental duplication. Many of these genes may share common ancestors with that in other species, such as *A. thaliana*, *T. cacao*, and so on. RNA-seq and expression profiles indicated that most *TIFY* family genes were involved in fiber development. This study will provide very useful information for future fiber development.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the National Natural Science Foundation of China (#30971821), National Transgenic Plants Project of China (#2011ZX08005-002), and the China Agriculture Research System (#CARS-18-45). The sponsors of this study did not participate in study design, data collection or analysis, paper preparation, or the decision to publish the study.

Supplementary material

REFERENCES

Aparicio-Fabre R, Guillen G, Loredo M, Arellano J, et al. (2013). Common bean (*Phaseolus vulgaris* L.) PvTIFY orchestrates global changes in transcript profile response to jasmonate and phosphorus deficiency. *BMC Plant Biol*. 13: 26.

Bai YH, Meng YJ, Huang DL, Qi YH, et al. (2011). Origin and evolutionary analysis of the plant-specific TIFY transcription factor family. *Genomics* 98: 128-136.

Cannon SB, Mitra A, Baumgarten A, Young ND, et al. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol*. 4: 10.

Chacón-López A, Ibarra-Laclette E, Sánchez-Calderón L, Gutiérrez-Alanís D, et al. (2011). Global expression pattern comparison between low phosphorus insensitive 4 and WT *Arabidopsis* reveals an important role of reactive oxygen species and jasmonic acid in the root tip response to phosphate starvation. *Plant Signal Behav*. 6: 382-392.

Chenna R, Sugawara H, Koike T, Lopez R, et al. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*. 31: 3497-3500.

Chini A, Fonseca S, Chico JM, Fernández-Calvo P, et al. (2009). The ZIM domain mediates homo- and heteromeric interactions between *Arabidopsis* JAZ proteins. *Plant J*. 59: 77-87.

Chung HS, Niu YJ, Browse J and Howe GA (2009). Top hits in contemporary JAZ: an update on jasmonate signaling. *Phytochemistry* 70: 1547-1559.

Crooks GE, Hon G, Chandonia JM and Brenner SE (2004). WebLogo: A sequence logo generator. *Genome Res*. 14: 1188-1190.

Demianski AJ, Chung KM and Kunkel BN (2012). Analysis of *Arabidopsis* JAZ gene expression during *Pseudomonas syringae* pathogenesis. *Mol*. *Plant Pathol*. 13: 46-57.

Flagel LE, Wendel JF and Udall JA (2012). Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. *BMC Genomics* 13: 302.

Graves DA and Stewart JM (1988). Chronology of the differentiation of cotton (*Gossypium hirsutum* L) fiber cells. *Planta* 175: 254-258.

Grunewald W, Vanholme B, Pauwels L, Plovie E, et al. (2009). Expression of the *Arabidopsis* jasmonate signalling repressor JAZ1/TIFY10A is stimulated by auxin. *Embo*. *Rep*. 10: 923-928.

He DH, Lin ZX, Zhang XL, Nie YC, et al. (2007). QTL mapping for economic traits based on a dense genetic map of cotton with PCR-based markers using the interspecific cross of *Gossypium hirsutum* x *Gossypium barbadense*. *Euphytica* 153: 181-197.

Jiang CX, Wright RJ, El-Zik KM and Paterson AH (1998). Polyploid formation created unique avenues for response to selection in *Gossypium* (cotton). *Proc*. *Natl*. *Acad*. *Sci*. *U. S. A*. 95: 4419-4424.

Koonin EV (2005). Orthologs, paralogs, and evolutionary genomics. *Annu*. *Rev*. *Genet*. 39: 309-338.

Libault M, Joshi T, Benedito VA, Xu D, et al. (2009). Legume transcription factor genes: what makes legumes so special? *Plant Physiol*. 151: 991-1001.

Lin LF, Pierce GJ, Bowers JE, Estill JC, et al. (2010). A draft physical map of a D-genome cotton species (*Gossypium raimondii*). *BMC Genomics* 11: 395.

Lyons E, Pedersen B, Kane J, Alam M, et al. (2008). Finding and comparing syntenic regions among *Arabidopsis* and the Outgroups papaya, poplar, and grape: CoGe with Rosids. *Plant Physiol*. 148: 1772-1781.

Nishii A, Takemura M, Fujita H, Shikata M, et al. (2000). Characterization of a novel gene encoding a putative single zinc-finger protein, ZIM, expressed during the reproductive phase in *Arabidopsis thaliana*. *Biosci*. *Biotech*. *Bioch*. 64: 1402-1409.

Paterson AH, Wendel JF, Gundlach H, Guo H, et al. (2012). Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* 492: 423-428.

Qi TC, Song SS, Ren QC, Wu DW, et al. (2011). The jasmonate-ZIM-domain proteins interact with the WD-repeat/bHLH/MYB complexes to regulate jasmonate-mediated anthocyanin accumulation and trichome initiation in *Arabidopsis thaliana*. *Plant Cell* 23: 1795-1814.

Rambaldi D and Ciccarelli FD (2009). FancyGene: dynamic visualization of gene structures and protein domain architectures on genomic loci. *Bioinformatics* 25: 2281-2282.

Rice P, Longden I and Bleasby A (2000). EMBOSS: The European molecular biology open software suite. *Trends Genet*. 16: 276-277.

Seo JS, Joo J, Kim MJ, Kim YK, et al. (2011). OsbHLH148, a basic helix-loop-helix protein, interacts with OsJAZ proteins in a jasmonate signaling pathway leading to drought tolerance in rice. *Plant J*. 65: 907-921.

Shikata M, Matsuda Y, Ando K, Nishii A, et al. (2004). Characterization of *Arabidopsis* ZIM, a member of a novel plant-specific GATA factor gene family. *J*. *Exp*. *Bot*. 55: 631-639.

Shiu SH and Bleecker AB (2003). Expansion of the receptor-like kinase/Pelle gene family and receptor-like proteins in *Arabidopsis*. *Plant Physiol*. 132: 530-543.

Staswick PE (2008). JAZing up jasmonate signaling. *Trends Plant Sci*. 13: 66-71.

Tamura K, Peterson D, Peterson N, Stecher G, et al. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol*. *Biol*. *Evol*. 28: 2731-2739.

Tang HB, Bowers JE, Wang XY, Ming R, et al. (2008). Synteny and collinearity in plant genomes. *Science* 320: 486-488.

Thines B, Katsir L, Melotto M, Niu Y, et al. (2007). JAZ repressor proteins are targets of the SCFCO11 complex during jasmonate signalling. *Nature* 448: 661-665.

Trapnell C, Roberts A, Goff L, Pertea G, et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat*. *Protoc*. 7: 562-578.

Vanholme B, Grunewald W, Bateman A, Kohchi T, et al. (2007). The tify family previously known as ZIM. *Trends Plant Sci*. 12: 239-244.

Wan Q, Zhang H, Ye W, Wu H, et al. (2014). Genome-wide transcriptome profiling revealed cotton fuzz fiber development having a similar molecular model as *Arabidopsis* trichome. *PLos One* 9: e97313.

Wang KB, Wang ZW, Li FG, Ye WW, et al. (2012a). The draft genome of a diploid cotton *Gossypium raimondii*. *Nat*. *Genet*. 44: 1098-1103.

Wang YP, Tang HB, DeBarry JD, Tan X, et al. (2012b). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*. 40: e49.

Wasternack C (2007). Jasmonates: An update on biosynthesis, signal transduction and action in plant stress response, growth and development. *Ann*. *Bot*. 100: 681-697.

White DWR (2006). PEAPOD regulates lamina size and curvature in *Arabidopsis*. *Proc*. *Natl*. *Acad*. *Sci*. *U.S.A.* 103: 13238-13243.

Wray GA, Hahn MW, Abouheif E, Balhoff JP, et al. (2003). The evolution of transcriptional regulation in eukaryotes. *Mol*. *Biol*. *Evol*. 20: 1377-1419.

Ye HY, Du H, Tang N, Li XH, et al. (2009). Identification and expression profiling analysis of TIFY family genes involved in stress and phytohormone responses in rice. *Plant Mol*. *Biol*. 71: 291-305.

Zhang YC, Gao M, Singer SD, Fei ZJ, et al. (2012). Genome-wide identification and analysis of the TIFY gene family in grape. *PLos One* 7: e44465.

Zhu D, Bai X, Luo X, Chen Q, et al. (2013). Identification of wild soybean (*Glycine soja*) TIFY family genes and their expression profiling analysis under bicarbonate stress. *Plant Cell Rep*. 32: 263-272.