GMR Genetics and Molecular Research ISSN 16765680

Genome-wide comparison of AP2/ERF superfamily genes between *Gossypium arboreum* and *G. raimondii*

Z.P. Lei[1,2]*, D.H. He[1]*, H.Y. Xing[1], B.S. Tang[1] and B.X. Lu[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. 15 (3): gmr.15038211
Received December 7, 2015
Accepted February 11, 2016
Published July 29, 2016
DOI http://dx.doi.org/10.4238/gmr.15038211

ABSTRACT. The APETALA2/ethylene response factor (AP2/ERF) transcription factor superfamily is known to regulate diverse processes of plant development and stress responses. We conducted a genome-wide analysis of the *AP2/ERF* gene in *Gossypium arboreum* and *G. raimondii*. Using RPSBLAST and HMMsearch, a total of 271 and 269 *AP2/ERF* genes were identified in the *G. arboreum* and *G. raimondii* genomes, respectively. A phylogenetic analysis classified diploid *Gossypium* spp *AP2/ERF* genes into 4 families and 16 subfamilies. Orthologous genes predominated the terminal branch of the phylogenetic tree. Physical mapping showed at least 30% of *AP2/ERF* genes clustered together. A high level of intra- and inter-species collinearity involving *AP2/ERF* genes was observed, indicating common (before species divergence) or parallel (after species divergence) segmental duplications, along with tandem duplications, resulting in the species-specific expansion of *AP2/ERF* genes in diploid *Gossypium* species. Motif analyses of the AP2/ERF proteins revealed that motif arrangements were highly diverse among subfamilies, but shared by orthologous gene pairs. An examination of nucleotide divergence of *AP2/ERF* coding regions identified small and non-significant sequence differences among orthologs. Expression profiling of *AP2/ERF* orthologous gene pairs showed similar abundance levels of orthologous copies between *G. arboreum* and *G. raimondii*. Thus, cotton species possess abundant and diverse *AP2/ERF* genes, resulting from tandem and segmental duplications. Protein and nucleotide sequence and mRNA expression analyses revealed symmetrical evolution, indicating that most *AP2/ERF* genes may not have undergone significant biochemical and morphological divergence between sister species. Our study provides detailed insights into the evolutionary characteristics and functional importance of *AP2/ERF* genes, and could aid in the genetic improvement of agriculturally significant crops in this genus.

Key words: *Gossypium arboreum*; *Gossypium raimondii*; Phylogeny; AP2/ERF superfamily; Collinearity; Biological evolution

INTRODUCTION

The APETALA2/ethylene response factor (AP2/ERF) DNA-binding domains, highly conserved polypeptides consisting of approximately 60 amino acids, are a major feature of this multi-gene superfamily of transcription factors (TFs). The direct interaction of AP2/ERF domains with the GCC box and/or *cis*-acting C-repeat elements/dehydration responsive elements (CRT/DRE) affects the expression of downstream target genes. Previous studies have shown that AP2/ERF TFs regulate various processes related to plant development and responses to stress, including vegetative and reproductive development, cell division, abiotic and biotic responses to stress, and phytohormone responses (Nakano et al., 2006; Licausi et al., 2010; Sharoni et al., 2011). This superfamily is further classified into four families based on the type of AP2/ERF domain: 1) AP2, 2) ERF, 3) RAV (related to ABI3/VP), and 4) other proteins (designated as Soloist in *Arabidopsis thaliana*). The ERF family is further subclassified into 12 subfamilies (I-X, VI-like, and Xb-like) based on the structure of the AP2/ERF domain (Nakano et al., 2006), which has been proven to be applicable to the cotton ERF family (Champion et al., 2009). The RAV family consists of one B3 domain and one AP2/ERF domain. Most members of the AP2 family comprise two AP2/ERF domains. Members of the VI and VI-L phylogenetic clades of AP2/ERF proteins are classified as cytokinin response factors (CRFs). In another study, Sakuma et al. (2002) classified the AP2/ERF superfamily into five families [i.e., AP2, dehydration responsive element binding proteins (DREB), ERF (also called ethylene-responsive element binding proteins, EREBs), RAV, and Soloist]. In contrast, the Nakano classification (Nakano et al., 2006) has been used as reference for the organization of the AP2/ERF superfamily (Duan et al., 2013).

Recent whole-genome sequence data have facilitated the genome-wide identification, localization, and characterization of *AP2/ERF* genes in several agriculturally significant plants. For instance, numerous arrays of *AP2/ERF* genes have been identified in the following commercially important species: 147 in *A. thaliana* (Nakano et al., 2006), 180 in *Oryza sativa* (Sharoni et al., 2011), 202 in *Populus trichocarpa* (Zhuang et al., 2008), 126 in *Citrus sinensis* and *C. clementine* (Xie et al., 2014), 131 in *Prunus persica* (Zhang et al., 2012), 281 in *Brassica rapa* ssp *pekinensis* (Liu et al., 2013), and 149 in *Vitis vinifera* (Licausi et al., 2010).

The genus *Gossypium* comprises 46 diploid (2n = 2x = 26) and five allotetraploid (2n = 4x = 52) species. Among these species, two tetraploid species (i.e., *G. hirsutum* and *G. barbadense*), as well as two diploid species (i.e., *G. arboreum* and *G. herbaceum*) underwent domestication and were cultivated for fiber production. The predominantly cultivated species, namely *G. hirsutum* and *G. barbadense*, were created from an ancient interspecific hybridization of the ancestors of the modern *G. arboreum* (A2) and *G. raimondii* (D5), followed by polyploidization (i.e., chromosome doubling). These two diploid cotton species may have been derived from a common ancestor, which had antecedently undergone two whole genome duplications (WGDs, including triplication, which is often referred to as paleopolyploidy). Therefore, two WGD events, which might be traced back in *G. arboreum*, coincided with those in *G. raimondii*, and two WGD events have been estimated to have occurred prior to the divergence of the two diploid species. Whole-genome sequence alignments between *G. arboreum* and *G. raimondii*, conducted by Li et al. (2014), indicated that *G. raimondii* genomes showed close collinear relationships with the genome of *G. arboreum*. However, unlike *G. arboreum*, *G. raimondii* does not produce spinnable fibers. Furthermore, the genome size of *G. arboreum* (1746 Mb/1C) is two-fold larger than that of *G. raimondii* (885 Mb/1C), which could be attributable to a higher level of expansion of the transposable element (TE) family in *G. arboreum* than in *G. raimondii*. Conversely, the contraction of the nucleotide-binding site (NBS)-encoding gene family in *G. arboreum* could have increased its susceptibility to *Verticillium* wilt compared to that of *G. raimondii*, which is resistant to this disease (Li et al., 2014). Therefore, this provides an example of a gene family in *G. raimondii* that did not evolve in concert with that of its sister species, *G. arboreum*.

A limited number of ERF TFs have been characterized in cotton. Among these, an ERF TF, also known as ERF1, is activated during the initiation of fiber cells, but is inhibited in the naked seed mutant that shows dysfunctional fiber formation (Yang et al., 2006). Furthermore, *ERF1* from *G. hirsutum* (*GhERF1*) has been determined to be responsive to various abiotic stresses (Qiao et al., 2008), and *GhERF4* is potentially involved in transcriptional control via the ethylene (ET)-mediated signaling pathway (Jin and Liu, 2008). Due to a lag in genomic studies of cotton, genome-level information on *Gossypium AP2/ERF* genes is limited. Fortunately, by using the UniGene cotton database (http://plantta.tigr.org/search.shtml), Champion et al. (2009) identified 218 *AP2/ERF* genes in three *Gossypium* species (i.e., *G. hirsutum*, *G. raimondii*, and *G. arboreum*). In addition, Champion et al. (2009) determined that the expression of several *GhERF* genes, belonging to group IXa, was stimulated by jasmonic acid and ET, but not by salicylic acid. Moreover, several *GhERF* were differentially controlled by virulent as well as avirulent strains of *Xanthomonas campestris* pathovar *malvacearum*. However, because the source dataset was an incomplete transcriptome, the number, sequence divergence, and chromosomal distribution of *AP2/ERF* genes in *G. arboreum* and *G. raimondii* remain unknown. Recent sequencing of the *G. arboreum* and *G. raimondii* genomes has generated extensive genomic data that could benefit both the research and agricultural communities (Paterson et al., 2012; Li et al., 2014). Now that sequence data from *Gossypium* spp genomes are available (Paterson et al., 2012; Li et al., 2014), genome-wide identification and comparative analyses are necessary and valuable to better understand *Gossypium* spp *AP2/ERF* genes.

Sister species generally diverge from a common ancestor, such as has been observed in *A. thaliana* and *A. lyrata* (Guo et al., 2011), *B. oleracea* and *B. rapa* (Sampath et al., 2014), and in *G. arboreum* and *G. raimondii* (Li et al., 2014). After a divergence event, sister species often undergo asymmetrical evolution, which results in extremely contrasting biochemical, morphological, and genomic features. A genome-wide comparison of NBS-leucine-rich repeat-encoding (NB-LRR) genes in *A. thaliana* and *A. lyrata* has indicated the same number of NB-LRR genes, which were not affected by pollination mode (selfing or outcross) (Guo et al., 2011). A genome-wide, inter-genomic comparison of 20 miniature inverted-repeat transposable element (MITE) families in *B. rapa* and *B. oleracea* revealed that among the 20 MITE families, 11 families individually harbored similar copy numbers in *B. rapa* and *B. oleracea*, and nine families presented variations in copy number, which ranged from 2- to 16-fold (Sampath et al., 2014). Additionally, a genome-wide, inter-genomic comparison of NBS-encoding genes identified 206 genes in *B. rapa* and 157 in *B. oleracea* (Yu et al., 2014). Furthermore, following the divergence of *B. rapa* and *B. oleracea*, gene amplification by various tandem duplication events was followed by different selection pressures, resulting in variations in the number of NBS genes retained in sister genomes (Yu et al., 2014). The observed differences in the number of NBS genes in *G. arboreum* and *G. raimondii* may be attributed to a single causative factor. A comparison of two sub-species, *japonica* and *indica*, in *Oryza sativa*, showed that approximately 5% of the genes are asymmetrically localized between the two genomes, and that 5% occurred only within one of the genomes (Ding et al., 2007). Most asymmetric genes are categorized as disease-resistance (RLK kinase) genes, indicating that selection is responsible in maintaining asymmetry (Hurwitz et al., 2010). Comparative studies of grasses originating from a common ancestor have indicated that functional differences may exclusively occur in a few species-specific genes, either in the form of variations in the copy number or expression patterns of shared genes (Bennetzen, 2007). Comparative sequence analyses of homologous segments in the A and D genomes of *G. arboreum* and *G. raimondii* have shown near-perfect microcollinearity and a general expansion of segments of the A genome relative to that of the equivalent segments in the D genome (Li et al., 2014). The relative expansion is apparently primarily due to the differential accumulation of TEs, specifically long terminal repeat retrotransposons (LTR-RTs) present in the orthologous regions (Li et al., 2014).

One significant observation of the two genomes (*G. arboreum* and *G. raimondii*) is the size differences between syntenic regions (1724 Mb *vs* 737.8 Mb, respectively), comprising approximately 1,258.52 Mb (73%) of the assembled DNA of the A genome versus around 649.26 Mb (88%) of the D genome (Li et al., 2014). However, the contribution of the *AP2/ERF* superfamily to the evolutionary rates of the two genomes, and the two-fold increase in genome size in *G. arboreum* (in comparison to *G. raimondii*) following its divergence from a common ancestor, remains elusive. Unlike wild-type *G. raimondii*, *G. arboreum* individuals are domesticated via artificial selection. Given this, does the selection pressure caused by human breeding affect the number and diversity of *AP2/ERF* genes in the genus *Gossypium*? Our article expounds upon a previous study (Champion et al., 2009) that performed an early systematic identification of *AP2/ERF* genes in *Gossypium* spp, which lacks extensive information on evolution and collinearity.

MATERIAL AND METHODS

Identification of *AP2/ERF* genes in diploid *Gossypium*

The *G. arboreum* genome database was downloaded from the Cotton Genome Project

(CGP) website: http://cgp.genomics.org.cn/page/species/download.jsp?category=arboreum (Li et al., 2014). The *G. raimondii* genome database was retrieved from the DOE Joint Genome Institute (JGI) website: ftp://ftp.jgipsf.org/pub/compgen/phytozome/v9.0/Graimondii/ (Paterson et al., 2012).

Gossypium-specific AP2/ERF domain sequences described by Champion et al. (2009) and the website, http://datf.cbi.pku.edu.cn/ (Jin et al., 2014), were used in the present study. Following an alignment using ClustalW version 2.1 (Chenna et al., 2003), a *Gossypium*-specific hidden Markov model (HMM) profile and position-specific scoring matrix (PSSM) of the AP2/ERF domain was built by using hmmbuild as implemented in HMMER v. 3.0 (http://hmmer.janelia.org/) and the stand-alone BLAST v.2.2.26 (ftp://ncbi.nlm.nih.gov/blast/executables/), respectively. PSSM and HMM profiles were then used to identify AP2/ERF proteins that were encoded by each genome using RPSBLAST and HMMsearch (E, 1e-5). The identified proteins from *Gossypium* AP2/ERFs were subsequently confirmed as having the AP2/ERF domain using the universal Pfam accession number Pfam00847 (http://pfam.xfam.org/; Finn et al., 2016). Additionally, sequences having the Pfam accession number Pfam02362 (for the B3 domain) were also collected, and the HMM search was conducted to identify the open reading frames (ORFs) of genes encoding the B3 domain to validate the RAV family.

Classification and phylogenetic analysis

A multiple sequence alignment (MSA) analysis of the AP2/ERF domain sequences (approximately 60 amino acids) from each *Gossypium* species, as well as *Arabidopsis*, was conducted using ClustalW (Chenna et al., 2003). Using the MEGA 5.1 software (Tamura et al., 2011), phylogenetic trees were constructed by using the maximum likelihood method for each *Gossypium* species, with *Arabidopsis* as reference, in order to group cotton AP2/ERF sequences into families and subfamilies. In addition, two major subfamilies of ERF_IX and ERF_III were used in the comparative phylogenetic analyses of the *AP2/ERF* genes of *G. arboreum* and *G. raimondii*. The MSA analysis was conducted using full-length protein sequences. Topological robustness was evaluated by bootstrapping using 1000 replications.

Chromosomal localization and gene duplications

Chromosomal locations of *AP2/ERF* were determined based on the chromosomal information derived from the CGP and JGI. The positions of the *AP2/ERF* genes were physically mapped to the 13 chromosomes in each genome. The chromosomal distribution of *AP2/ERF* genes was visualized using the MapInspect tool (van Berloo, 2008). Tandem duplications were defined as adjacent *AP2/ERF* genes within individual chromosomes with four or fewer intervening genes. Clusters were defined as two or more genes falling within eight ORFs (Meyers et al., 2003). Collinear blocks containing the *AP2/ERF* genes within the individual genomes and between the genomes of *G. arboreum* and *G. raimondii*, were identified using the MCScanX program (Wang et al., 2012). After analyses of collinearity, collinear paralogs within each genome and collinear orthologs from *G. arboreum vs G. raimondii* were extracted; these were also identified as segmental duplications.

Analysis of conserved motif structures

Structural diversity among AP2/ERF peptides was also examined. Putative amino acid sequences were submitted to motif analyses by using MEME Suite v. 4.8.0, a local motif-based analytical tool (Bailey and Gribskov, 1998). As the AP2/ERF domain consists of approximately 60-70 amino acids, the optimal width of amino acid sequence was set from 6 to 70. The maximum number of motifs was established as 20. The presence of ERF-associated amphiphilic repression (EAR) motifs was investigated using the Perl script ps_scan (ftp://ftp.expasy.org/databases/prosite). A nuclear localization signal (NLS) analysis was performed using ps_scan and the program hmmsearch (http://hmmer.janelia.org/).

Divergence of orthologs and paralogs

The nucleotide divergence of homologous or homeologous (containing orthologs or paralogs) genes, as well as the selective pressure on duplicated genes, were assessed by determining the number of synonymous (*Ks*) and non-synonymous (*Ka*) nucleotide substitutions at each site between the duplicated gene pairs using the Codeml program of the PAML software package (Yang, 2007).

Comparative expression analysis of *AP2/ERF* genes between *G. arboreum* and *G. raimondii*

To study the evolution of *AP2/ERF* genes, we performed comparative analyses of expression profiles of fibers and petals. The transcript levels of *AP2/ERF*s expressed in fibers (GSE17084) (Chaudhary et al., 2009), and petals (GSE17927) (Flagel and Wendel, 2010) were examined using the microarray platform GPL6989. Three biological replicates were performed. Microarray data were downloaded from the NCBI Gene Expression Omnibus (GEO) database. Hybridization of probes to *G. arboreum* and *G. raimondii* mRNA transcripts were identified via a strict alignment using megablast [percentage of identical matches (pident) was 100, alignment length was equal to the length of probe sequences].

In addition, Illumina RNA-sequencing data (SRP028270) (Rambani et al., 2014) on the petals were used to investigate and validate the transcript abundance of *AP2/ERF* genes. Reads were mapped to the *AP2/ERF* gene models using Tophat (Trapnell et al., 2012). The matched reads were used to estimate gene expression levels and were expressed as fragments per kilobase of transcript per million mapped reads (FPKM). A heatmap of the *AP2/ERF* gene expression profiles was generated by using the R package pheatmap (https://cran.r-project.org).

RESULTS

Identification and classification of *AP2/ERF* genes

Using the universal AP2/ERF domain and *Gossypium*-specific HMM profile, we identified 271 and 269 members of the AP2/ERF superfamily in the *G. arboreum* and *G. raimondii* genomes, respectively (Figure 1 and Table 1; **Tables S1** and **S2**). A phylogenetic analysis (performed on the conserved AP2/ERF domain sequences), using the Nakano classification, suggests that all diploid *Gossypium AP2/ERF* genes could be classified into 16 subfamilies: AP2 (2 subfamilies; 31 and 32 members for *G. arboreum* and *G. raimondii*, respectively), ERF (12 subfamilies, grouped to I through X, VI-like, and Xb-like; 227 and 222 members, respectively), RAV (9 and 11 members, respectively), and Soloist (four and four members, respectively). Fifty-eight genes in the two species were predicted to encode for proteins consisting of two AP2/ERF domains by HMMsearch; most of these genes (except *Cotton_A_11454* and *Cotton_A_17367*) were assigned to the AP2_2 subfamily. Seven genes contained one AP2/ERF domain, which differed from members of the ERF family, and were phylogenetically related to those belonging to the AP2 family. Therefore, these genes grouped with the AP2_1 subfamily. Twenty genes in the RAV family contained an additional B3 domain. Eight genes contained an AP2/ERF-like domain sequence, although the homology of this sequence appears relatively low compared to other *AP2/ERF* genes. These eight genes shared high-sequence similarity with *At4g13040*, which was grouped into the Soloist family in *A. thaliana* (Nakano et al., 2006). Therefore, these eight genes were assigned to family Soloist. The remaining 449 genes, classified as members of the ERF family, were subsequently categorized into 12 subfamilies (I through X, VI-L, and Xb-L, corresponding to *Arabidopsis*) based on amino acid similarities in the AP2/ERF domains. Genes from the IX and III luxuriant subfamily were also observed in these two species. Several genes with truncated AP2/ERF domains were detected in subfamily IX, particularly in *G. raimondii*. Conversely, *G. arboreum* and *G. raimondii* possessed the lowest number of genes in subfamily VI-L, with two and three genes, respectively.

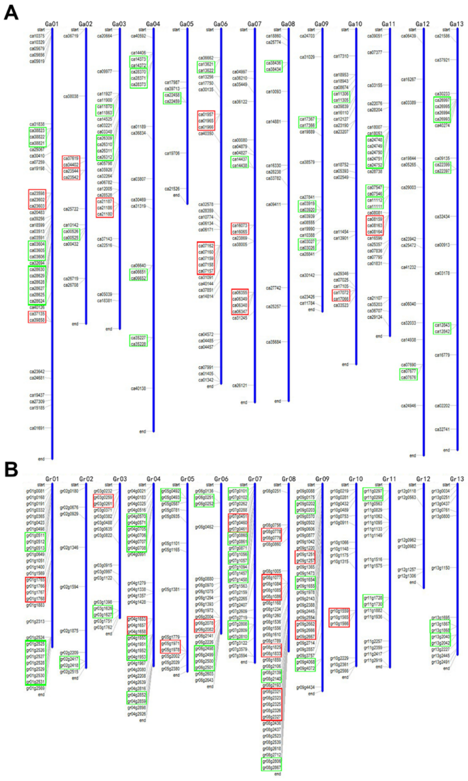

Figure 1. *AP2/ERF* genes and corresponding cluster distribution in the genomes of *Gossypium arboreum* (Ga) and *G. raimondii* (Gr). **A.** Ga01-Ga13 represent pseudo-chromosomes of the *G. arboreum* genome. **B.** Gr01-Gr13 represent pseudo-chromosomes of the *G. raimondii* genome. Blue bars indicate pseudo-chromosomes. Black lines on the blue bars indicate the location of *AP2/ERF* genes on the pseudo-chromosomes. Green and red boxes indicate homogenous and heterogeneous clusters of *AP2/ERF* genes in the corresponding genomes, respectively.

Table 1. Classification of the diploid *Gossypium* AP2/ERF superfamily based on the results of a phylogenetic analysis of *Arabidopsis thaliana* and various other plant species using the method of Nakano et al. (2006).

Family	Subfamily	Ga	Gr	At	Tc	Pt	Vv	Os	Cs	Pp	Br
AP2	2[1]	27	29	14	18	26	20	29	18	21	30
	1[2]	4	3	4	3						
ERF	I	18	15	10	5	5	5	9	5	6	15
	II	19	17	15	7	20	8	16	8	9	29
	III	38	40	22	16	35	22	27	17	23	39
	IV	8	8	9	8	6	5	6	5	7	22
	V	11	11	5	5	10	11	8	12	11	11
	VI	11	11	8	5	11	5	6	4	3	13
	VII	8	7	5	5	4	2	15	6	6	16
	VIII	22	23	15	11	6	3	15	11	10	27
	IX	62	63	17	20	17	11	18	23	19	23
	X	17	15	8	5	42	40	12	6	6	9
	VI-L	2	3	4	3	9	10	3	2	4	6
	Xb-L	11	9	3	7	4	0	10	3	0	9
RAV		9	11	6	5	6	6	5	4	5	14
Soloist		4	4	1	1	1	1	1	2	1	1
Total		271	269	146	124	202	149	180	126	131	281[3]

At = *Arabidopsis thaliana* (Nakano et al., 2006), Br = *Brassica rapa* ssp *pekinensis* (Liu et al. 2013), Cs = *Citrus sinensis* (Xie et al., 2014), Ga = *Gossypium arboreum*, Gr = *Gossypium raimondii*, Os = *Oryza sativa* (Nakano et al., 2006), Pp = *Prunus persica* (Zhang et al., 2012), Pt = *Populus trichocarpa* (Zhuang et al., 2008), Tc = *Theobroma cacao*, Vv = *Vitis vinifera* (Licausi et al., 2010). [1]Double AP2 domain. [2]One AP2 domain. [3]Data contain an additional 17 genes belonging to the rare subfamily ERF_XI. Numbers were obtained from published literature, except those for Ga, Gr, and Tc.

The number of members of the diploid cotton AP2/ERF superfamily was higher compared to eight other plant species (Table 1), indicating that the AP2/ERF superfamily underwent a *Gossypium*-specific expansion. The number of genes in subfamily IX is 3.6-fold higher than in *Arabidopsis*. The number of members in each subfamily was observed to be almost equal in the two diploid species, with differences of <5, indicating concerted evolution (inter-species) after species divergence/speciation events of the diploid cotton. Four Soloist genes were identified in each diploid cotton species, whereas a single gene was detected in most species. A previous study on *Gossypium*, conducted using the incomplete transcriptome, reported 21 *AP2/ERF* genes in *G. arboreum* and 41 in *G. raimondii* (Champion et al., 2009). A direct comparison of the results of the present study to those of previous incomplete transcriptome assemblies proved to be challenging.

Phylogenetic analysis of *AP2/ERF* genes in *G. arboreum* and *G. raimondii*

Two major subfamilies, ERF_IX (125 genes) and ERF_III (78 genes), were used in the comparative phylogenetic analysis of *AP2/ERF* genes in *G. arboreum* and *G. raimondii*. In the composite phylogenetic tree, subfamilies ERF_IX and III were further divided into three and five clades, respectively (Xa-c: Figure 2 and IIIa-e: Figure S1). The number of clades in ERF_IX or III is in agreement with the results of phylogenetic analysis of conserved AP2/ERF domain sequences using *A. thaliana* as a reference. This indicates that the diversity of whole sequences is mainly in concert with their functional segments (i.e., the AP2/ERF domain region). On the other hand, for a relatively small number of the genes investigated, further categorization into groups (based on a phylogenetic analysis of full-length protein sequences) is not completely consistent with the relationship of functional segments (based on a phylogenetic analysis of conserved AP2/ERF domain sequences). No strict clustering of the proteins designated as belonging to clade ERF_IXc was observed, as these types of proteins grouped into both the ERF_IXa (containing an additional four ERF_IXc members along with the prevalent ERF_IXa-type members) and IXc clades, as shown in Figure 2. The phylogenetic tree in Figure 2 shows that the number of *AP2/ERF* genes in the two *Gossypium* species in each clade is almost identical. Clade IXc is the largest (comprising IXa-c), containing a total of 64 ERF_IX members, where a slightly greater part of this clade can be attributed to *G. raimondii* (33 GrERF_IX members). According to the level set for the similarity coefficient (adopted for the classification of clade IXa-c), IXc can be further divided into three groups (Figure 2). Most of the terminal branches include two genes from each species, implying that the majority of *AP2/ERF* genes evolved in parallel in the two species. However, our results indicate a higher number of tandem duplications in *G. arboreum* than in *G. raimondii* in the two terminal branches [(*Cotton_A_24750*/*Cotton_A_24751*/*Cotton_A_24752*)-(*Gr06G250000*/*Gr06G250100*) and (*Cotton_A_07157*/*Cotton_A_07158*/*Cotton_A_07159*)-(*Gr08G232600*/*Gr08G232700*)]. Conversely, two other terminal branches [*Cotton_A_03221*-(*Gr04G057000*/*Gr04G057100*/*Gr12G096200*), and (*Cotton_A_28624*/*Cotton_A_28628*/*Cotton_A_40274*)-(*Gr01G252600*/*Gr01G252800*/*Gr01G253000*/*Gr01G253100*)] show an increased number of tandem and segmental duplication in *G. raimondii* compared to that of *G. arboreum*. A further examination shows that two genes (*Gr12G096200* and *Gr01G252800*) contained partial (truncated, not full-length) AP2/ERF domains.

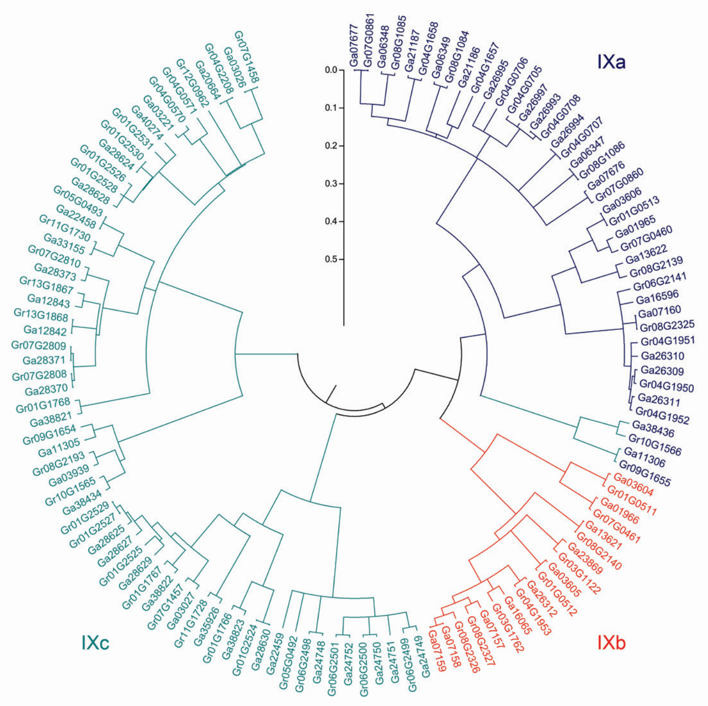

Figure 2. Phylogenetic relationship of subfamily ERF_IX genes in *Gossypium arboreum* (Ga) and *G. raimondii* (Gr). Construction of the maximum likelihood tree (rooted) was performed with the MEGA 5.0 software using 1000 replications. Subfamily IX was split into three groups (IXa-IXc).

Subfamily ERF_III was further divided into five distinct clades (**Figure S1**), as represented by genes from the two *Gossypium* species. We also observed that each ERF_III clade was already present in *A. thaliana*. However, the resulting tree topology (**Figure S1**) was not in agreement with the clades produced by conducting a phylogenetic analysis of conserved AP2/ERF domain sequences with *A. thaliana* as reference. Most terminal branches showed the same number of genes from each species (i.e., biunique). One exception was *Gr11G151600*, which contained a truncated AP2/ERF domain, with no counterpart existing in *G. arboreum*. The other exception was (*Cotton_A_36834*/./ *Cotton_A_17310*)-(*Gr05G116500*/*Gr09G266600*/*Gr09G266700*), which showed that an additional tandem duplication occurred adjacent to *Gr09G266700* in *G. raimondii*, thereby producing *Gr09G266600*.

Genomic distribution on pseudo-molecular chromosomes and collinearity

AP2/ERF genes in *G. arboreum* and *G. raimondii* were mapped to pseudo-molecule chromosomes (Figure 1). Its distribution was uneven, wherein certain chromosomes (e.g., Ga01 in *G. arboreum* representing 14.8% of the *AP2/ERF* genes) comprised more genes, whereas others consisted of fewer genes (e.g., Ga05 in *G. arboreum* and Gr12 in *G. raimondii*). Based on the cluster defined by Meyers et al. (2003), were at least two genes fall within eight ORFs, we observed that several genes were indeed clustered. In *G. raimondii*, 101 *AP2/ERF* genes, representing 37.5% of all *AP2/ERF* genes, were situated in 38 clusters, while the remaining 168 genes were singletons. Six clusters, containing 17 *AP2/ERF* genes, were localized to chromosome Gr08. The *G. arboreum* genome harbored 95 *AP2/ERF* genes (35.1%) in 36 clusters, while the remaining 176 genes were detected as singletons. Of the 36 clusters, five clusters containing 17 genes were located on chromosome Ga09. Long clusters, containing more than five *AP2/ERF* genes, were localized to chromosomes Ga01 and Gr01. Each cluster consisted of 2-6 genes (mean: 2.64; median: 2.0; **Table S1**) in *G. arboreum* and 2-8 genes (mean: 2.66; median: 2.0; **Table S2**) in *G. raimondii*. The largest cluster in *G. raimondii* was a pure cluster, comprising eight genes from subfamily IX. Reciprocally, the orthologous cluster in *G. arboreum* included only six genes, and was the largest cluster in *G. arboreum*. We therefore propose that after paleopolyploidy, *Gossypium* spp evolved clusters of new AP2/ERF-encoding genes, which were retained in the generations following the divergence of *G. arboreum* and *G. raimondii*. *G. raimondii* appears to have a slightly higher number of heterogeneous clusters (13 total) compared to *G. arboreum* (11 total). Furthermore, the same number of homogenous clusters was detected in *G. raimondii* and *G. arboreum*. In addition, most of the heterogeneous clusters that contained more than two genes were infiltrated by only one alien gene from a different AP2/ERF subfamily. Only one cluster (containing three IX and two RAV genes) in *G. raimondii* was found to be heterogeneous with two distantly related genes. Therefore, one pivotal AP2/ERF subfamily was present in each heterogeneous cluster.

Analyses of collinearity were conducted following the analysis of *AP2/ERF* gene clusters. First, a BLASTp was performed (Altschul et al., 1990) using 4536 *AP2/ERF* gene pairs identified from the genome sequences of *G. arboreum* and *G. raimondii*, which were characterized by a high level of sequence similarity. Gene pairs located at the terminal branches of the phylogenetic tree were highly similar (perhaps orthologous) between species. Collinearity was subsequently determined based on the orthology of the flanking genes of the focal *AP2/ERF* locus. A minority of gene pairs (511/4536, 11.27%) resided in inter-species collinear regions (Figure 3). *AP2/ERF* genes on chromosomes 1, 4-6, as well as 9-13 in *G.*

raimondii were highly collinear with the equivalent chromosomes in *G. arboreum*, which was in agreement with the findings of previous report (Li et al., 2014). Further examination (by network analyses) indicated that the 511 gene pairs determined to be collinear could be classified into 84 groups of orthologs (**Tables S1** and **S2**), the largest of which is illustrated in **Figure S2**. The collinear gene pairs consisted of 207 genes from *G. arboreum* and 226 genes from *G. raimondii*. The orthologous or paralogous gene sets determined to be collinear were clustered within the same group, with each group consisting of several orthologous or paralogous sets. Generally, each *AP2/ERF* in one species has between one and three putative orthologs in the other species. However, no orthologs of the 64 *GaAP2/ERF* genes were identified in *G. raimondii*. In addition, 43 *GrAP2/ERF* genes had no orthologous *G. arboreum* gene partners. These results are likely to be indicative of gene expansion as well as losses after the divergence of *G. arboreum* and *G. raimondii*.

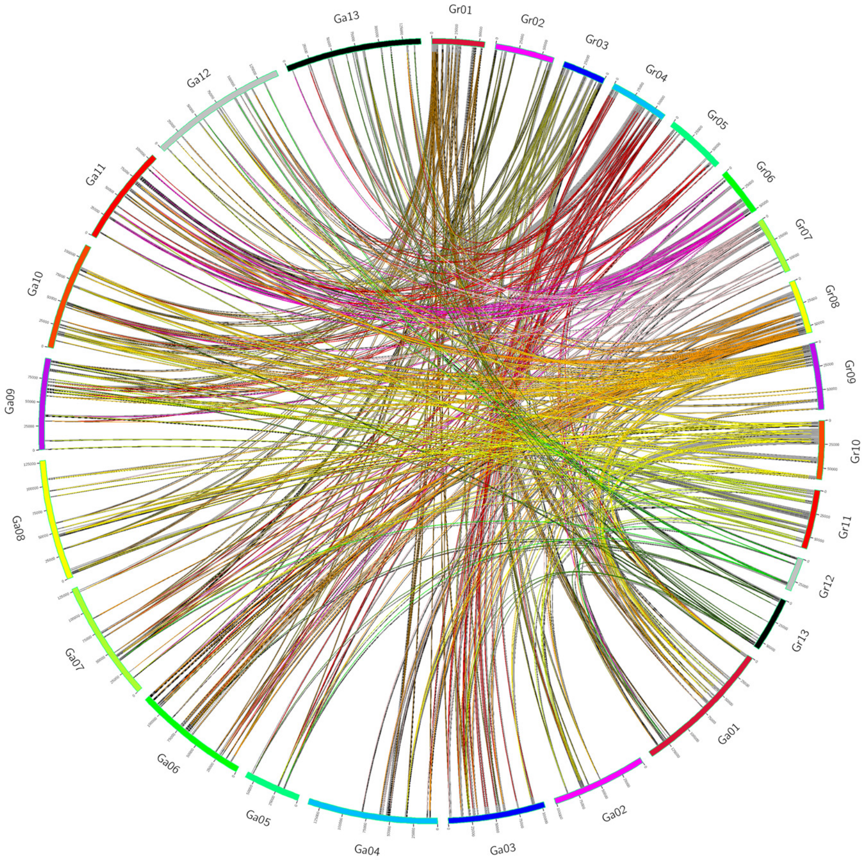

Figure 3. Collinear relationship of AP2/ERF genes between *Gossypium arboreum* (Ga) and *G. raimondii* (Gr). Curve white bars represent chromosomes of the two species. Ga01-Ga13 represent pseudo-chromosomes of *G. arboreum*, Gr01-Gr13 represent pseudo-chromosomes of *G. raimondii*. Gray curve lines denote collinear regions containing the *AP2/ERF* genes (no counterparts); other colorful curve lines indicate collinear regions (counterparts were present).

Structural analysis of the AP2/ERF superfamily

A MEME motif analysis showed that the conserved AP2/ERF domain, a key characteristic of the AP2/ERF superfamily, is composed of four motifs (E, B, D, and A; Figure 4), which were identified as having the lowest P values. The order of the four motifs in the full-length AP2/ERF domain was E-B-D-A, which structurally translated into three stranded, anti-parallel β-sheet regions (motifs E, B, and D) and one α-helix (motif A) (Xie et al., 2014). The full AP2/ERF domain (i.e., all four motifs) was detected in most (414/540) *Gossypium* AP2/ERF proteins. Motifs E, B, D, and A were detected in 511, 505, 452, and 518 proteins, respectively. A small percentage (16.3%) of the AP2/ERF proteins had a truncated AP2/ERF domain, which lacked a few of the four motifs. For example, only motif E was detected in all members of the family Soloist, yet characteristically accompanied by motif J (GT[FY] ETAE[EAD]AARAYDEAA[FRI]L[ML]R) (**Figure S3**). In addition to the eight Soloist members, another six proteins (including four members of the AP2 family, one from GrRAV, and one from GrERF_IXc) also have only one of the four usual motifs. Notably, the proteins containing two AP2/ERF domains, identified by HMMsearch, did not harbor two copies of the four motifs, but instead typically included two B motifs, whereas some proteins including two B motifs were alternatively assigned into subfamily ERF_VI. Similar to the members in the Soloist family, different ERF proteins harbored different conserved motifs other than the conserved AP2/ERF domain (**Tables S1** and **S2**). The longest motif, motif C (**Figure S3**), was present in most (61) members of family AP2, with the exception of two members of GaAP2. Therefore, motif C is the characteristic motif of family AP2. Motifs L and I make up the B3 domain of the RAV protein families. Motif M, with the consensus core sequence ATDSS[SG], also the main characteristic of CRF proteins along with the conserved AP2/ERF domain, was conserved in subfamilies ERF_VI and ERF_VI-L. A few members of the other subfamilies also contained motif M.

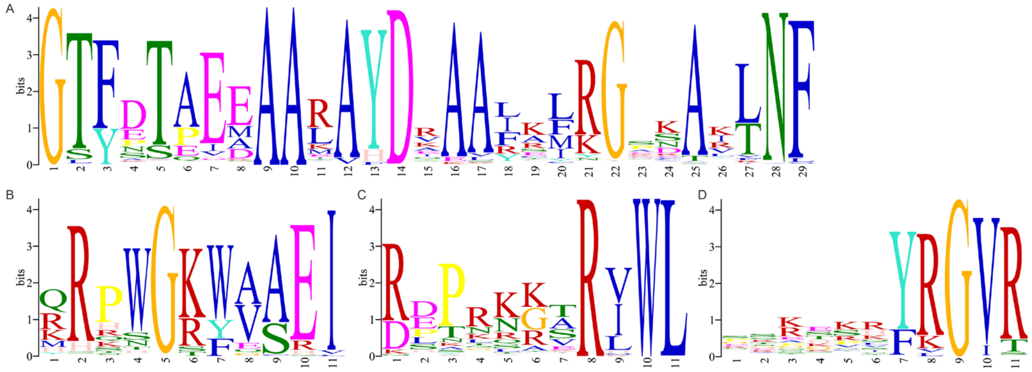

Figure 4. Sequence logo of motifs (involved in the AP2/ERF domain identified by the MEME suite) in the superfamily of AP2/ERF proteins from *Gossypium arboreum* and *G. raimondii*. **A.** Motif A. **B.** Motif B. **C.** Motif D. **D.** Motif E.

The arrangement of motifs was determined to be highly complex. We identified 108 different motif combinations (**Table S3**) in all AP2/ERF members from the two *Gossypium* species, with 80 in *G. arboreum* and 81 in *G. raimondii* (**Figure S4**). Among the 108 motif combinations, 55 were unusual motif combinations (i.e., each occurring in a single protein). Further examination indicated that 28 of the unusual motif combinations belonged to *G. raimondii* and 27 to *G. arboreum*. These species-specific motif combinations reveal pairwise similarities between the two species. Fifty-three were equally shared by *G. arboreum* and *G. raimondii*, with the exception of one motif combination (EBDAP). Each subfamily involved 2-25 different motif combinations. In summary, AP2/ERF proteins are involved in complex (and dynamic) motif combinations, which are shared by *G. raimondii* and *G. arboreum*.

In contrast to the conserved AP2/ERF domain, we observed some motifs that were only characteristic of some subfamilies (e.g., only motif H was detected in all members from family AP2, while motif F was only distributed in two subfamilies, ERF_II and III) (**Tables S1** and **S2**). However, other motifs were distributed among various subfamilies. Aside from the characteristic motifs of each subfamily, we also examined other universal motifs or domains (e.g., EAR motif, NLS sequence, and NBS domain). The EAR motif, known to be indicative of repressor-type ERF genes (Ohta et al., 2001), was identified in 46 GaAP2/ERF and 47 GrAP2/ERF members in our study (**Table S4** and **Figure S5**). Notably, half of the members of families AP2 and RAV and subfamily ERF_VIII also contain the EAR motif ([$^{L}/_{F}$DLN$^{L}/_{F}$(x)P]).

Ethylene responsive factor (*ERF*) genes encode transcription factors that possess nuclear activities. NLS sequences have often been identified within or proximal to DNA-binding domains of transcription factors (Stoller and Epstein, 2005). An analysis of NLS sequences indicated that they are distributed within specific AP2/ERF sequences of *G. arboreum* and *G. raimondii*. The identification of classical NLS showed that 71 GaAP2/ERF and 70 GrAP2/ERF proteins possess the bipartite-type NLS, and four GaAP2/ERF and six GrAP2/ERF proteins harbor the monopartite-type NLS sequence. More than half of the members of subfamily ERF_VIII and all members of ERF_VI-L contain the bipartite-type NLS sequence. The monopartite-type NLS sequence was detected in all Soloist members and most (14 of 15) ERF_VII members; however, it is absent in all ERF_V members (**Table S4**). Non-classical NLSs (i.e., PY-NLS), which are composed of the conserved amino acid sequences RX(2-5)PY, were also identified using the ps_scan program. The PY-NLS sequence was detected in 17 AP2/ERF sequences in *G. arboreum* and 20 AP2/ERF sequences in *G. raimondii* (**Table S4** and **Figure S5**), while the gene *Cotton_A_17072* was found to contain two RX(2-5)PY sequences.

Divergence of orthologous and paralogous AP2/ERF

The expansion of a gene superfamily is often followed by an acceleration of mutation rates. Synonymous substitution (*Ks*) values, non-synonymous substitution (*Ka*) values, and ω (*Ka/Ks*) values were examined for each subfamily of *AP2/ERF*. We calculated the average degree of sequence divergence (ω) between the two *Gossypium* species within each subfamily. As shown in Figure 5, divergence levels varied, ranging from 0.0174 (*GrAP2_2*) to 0.8263 (*GaERF_X*). The average divergence was 0.0629 ± 0.2192 (SE) for *G. arboreum* and 0.0630 ± 0.0947 for *G. raimondii*. For the 16 subfamilies, the intra-specific divergence levels of *AP2/ERF* sequences in individual subfamilies were not significantly correlated (r = 0.2505, P = 0.3493) between the two genomes, indicating that in some subfamilies (i.e., AP2_2, ERF_VI and ERF_X), higher mutation rates occurred in *G. arboreum*, whereas in other subfamilies (i.e., ERF_V and ERF_IX), we observed the opposite trend. In the remaining subfamilies, similar mutation rates were observed in both genomes. Further analyses showed that some abnormal sequences (strong mutation, for example, *Cotton_A_39858* in subfamily ERF_X) elevated the ω value of the subfamily.

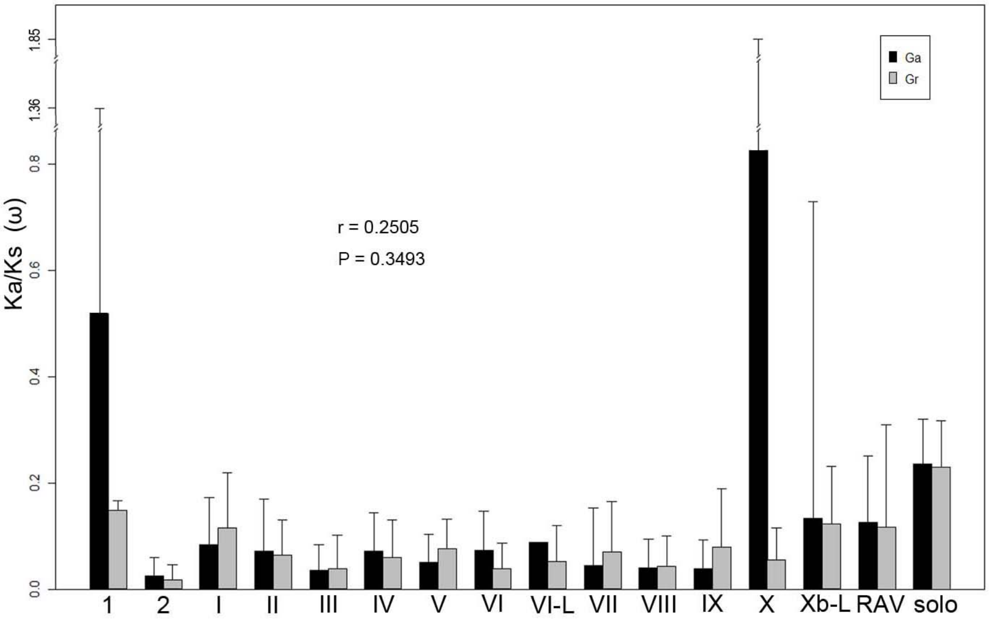

Figure 5. Intra-specific sequence divergence (ω value) of different subfamilies of AP2/ERF in *Gossypium arboreum* (Ga) and *G. raimondii* (Gr). All AP2/ERF gene pairs in each subfamily for each genome were employed to calculate the mean sequence divergence and standard deviation.

Among the 84 collinear orthologous groups (**Tables S1** and **S2**), 29 contained more than one gene from each species and were subjected to a divergence analysis of orthologous groups. The ω values of paralogs from each of the 29 collinear orthologous groups ranged from 0.0028 to 99 (mean ± SD: 0.8244 ± 7.9192) for *G. arboreum*, ranging from 0.0027 to 5.0328 (mean ± SD: 0.1535 ± 0.3639) for *G. raimondii*. In each orthologous group, the Student paired *t*-test between species indicated that no significant difference ($P < 0.05$) existed for the ω values between *G. arboreum* and *G. raimondii*. The ω values of orthologs from each of the 29 collinear orthologous groups ranged from 0.001 to 99 (mean ± SD: 0.5726 ± 5.9834). These results indicate that although a set of genes falls into a group due to inter-species collinearity (e.g., **Figure S2**), some gene pairs within the group were not collinear. Therefore, the evaluation of divergence based on full, pairwise genes shows that ω values for some paralogs and orthologs are abnormal (e.g., several values were observed to be 99).

Subsequently, we focused on a subset of the orthologs harbored in inter-species collinear regions. These subsets were intact *AP2/ERF* genes shared by the *G. arboreum* and *G. raimondii* genomes. Theoretically, several copies of each of these collinear gene groups (e.g., gene triplets) were identical in the two genomes at the time of their divergence from a common ancestor, and may have independently evolved over time. Therefore, sequence divergence between the two paralogous gene groups, shared by the two genomes, facilitated the comparison of the rates of sequence divergence in these orthologous regions in the two genomes. The ω values of intra-species collinear paralogous *AP2/ERF* gene pairs ranged from 0.0055 to 0.5532 (mean ± SD: 0.1887 ± 0.1018) for *G. arboreum* and from 0.0039 to 0.5983 (mean ± SD: 0.1898 ± 0.1097) for *G. raimondii*. In each orthologous group, the Student paired *t*-test between species indicated that no significant difference ($P < 0.05$) existed for the *Ka* values, *Ks* values, and ω values between *G. arboreum* and *G. raimondii* (**Table S5** and **Figure S6**). These results indicate that although the two genomes evolved independently, they have evolved concertedly since they split from a common ancestor. Subsequently, the 511 inter-species collinear orthologous gene pairs were assessed for *Ka*, *Ks*, and ω values. Small variations in sequence divergence were detected among the orthologs (**Figure S6**). The *Ka* values of the 511 orthologous pairs ranged from 0.0016 to 1.5847 (mean ± SD: 0.1778 ± 0.1827), whereas the *Ks* values ranged from 0.0056 to 88.5297 (mean ± SD: 4.6091 ± 14.4406). The *Ks* distribution of the collinear orthologous gene pairs between *GaAP2/ERF* and *GrAP2/ERF* peaked at 0.075-0.150 (**Figure S7**), indicating that the *AP2/ERF* genes belonging to the two genomes diverged approximately 5-10 mya, which is in agreement with the findings of previous studies on the speciation between the genomes of *G. arboreum* and *G. raimondii* (Senchina et al., 2003). Additionally, the resulting ω values ranged from 0.0026 to 2.6383 (mean ± SD: 0.2868 ± 0.2529), with only nine inter-species collinear orthologous pairs with values >1 (e.g., ω value of *Cotton_A_07162*/*Gr08G232100* was abnormally high at 2.6383). These results (ω values) implied that there have been more synonymous than non-synonymous changes in the collinear orthologs of *G. arboreum* than *G. raimondii*. Evolutionary pressure, and more namely, negative selection, has resulted in the conservation of the common ancestral state, allowing orthologs to maintain (conserve) protein function. **Figure S6** shows that the intra-specific divergence levels of *AP2/ERF* genes from individual paralogous gene groups at orthologous sites in the two genomes were positively correlated ($r = 0.8003$, $P < 0.01$) and are also positively correlated with the levels of inter-specific divergence of these gene groups (*G. arboreum*, $r = 0.6890$, $P < 0.01$; *G. raimondii*, $r = 0.7629$, $P < 0.01$). This shows that most mutations in (or the evolution of) the two species were convergent. Interestingly, for all 30 orthologous groups examined, *G. raimondii* showed somewhat higher levels of sequence divergence between the two paralogous *AP2/ERF* genes from each orthologous gene group than *G. arboreum* (**Figure S6**). However, in summary, we did not detect any significant differences in the level of sequence divergence between the two genomes ($P < 0.01$, Student paired *t*-test; **Figure S6** and **Table S5**).

Differential expression of *AP2/ERF* in *G. arboreum* and *G. raimondii*

To examine the expression levels of the *AP2/ERF* genes, we conducted a comparative analysis of transcript abundance in petals and fibers of *G. arboreum* and *G. raimondii* using microarray hybridization and RNA-seq data. Differences in the levels of expression between two species were monitored using a microarray that interrogates 42,429 unigenes. The number and identity of differentially expressed genes were studied for domesticated *G. arboreum* and wild-type *G. raimondii*. The strict megablast alignment indicated that, using microarray GPL6989, the expression levels of 185 *AP2/ERF* genes were precisely interrogated in fibers and petals. The mapping of Illumina reads indicated that RNA-sequencing examined the transcript abundance of 192 *AP2/ERF* genes. In total, 264 *AP2/ERF* genes were interrogated. The expression levels of 113 of the 264 genes were determined by both microarray and RNA-sequencing.

Data from the microarray experiment using microarray GSE17927 and RNA-sequencing, which were performed on samples collected from the same tissue (petal), indicated similar expression levels for the *AP2/ERF* genes (r = 0.7209, P = 8.99E-18) between microarray and RNA-sequencing. As for the different tissues (petal *vs* fiber), a relatively lower correlation coefficient of the expression levels was observed (r = 0.4461, P = 2.33E-06; r = 0.5145, P = 2.92E-12). These correlation coefficients indicated that the expression profile in the fibers were different from that of the petals, which suggests a tissue-specific function for each *AP2/ERF* gene in each tissue. Therefore, data from both microarray and RNA-sequencing on the same tissue can further validate each other to a certain degree.

Each *AP2/ERF* group showed similar numbers of *AP2/ERF* genes in the two species, from which the transcript abundances were collected. Among all groups, VIIa exhibited the highest expression in both tissues of the two species. Further examination showed increased mRNA expression in VIIa in *G. arboreum* than in *G. raimondii*. With the exception of VIIa, three groups (IIIa, VIIIa, and Ib) showed higher expression in both tissues of the two species. In both tissues, group IXc showed no expression in *G. arboreum* and the lowest expression in *G. raimondii*. Conversely, group Xc showed the lowest expression in *G. arboreum* and no expression in *G. raimondii*. In the fibers of the two species, the expression levels of individual genes within RAV family were also abundant, whereas genes within groups IVb and IIIc displayed low expression. Transcript accumulation of genes within group IIb remained at an intermediate level. The Student *t*-test showed that, in each group, the average expression levels in the two species were equal and none of the observed differences was significant. In addition, the expression level of each group in *G. arboreum* was highly correlated with that of *G. raimondii* (r = 0.9023, P = 2.19E-21), which indicates their functional conservation between the two species. Many of the AP2/ERF-related processes may be shared among the diploids. As for genes within group ERF_Ia, expression levels in *G. raimondii* were higher than that in *G. arboreum*, especially regarding *Gr05G202900*, which was more abundant in *G. raimondii* and is indicative of the functional divergence of these genes between the two species.

As previously mentioned, *AP2/ERF* genes within the same group exhibited similar expression patterns in the genomes of *G. arboreum* and *G. raimondii*. To conduct a more precise comparative expression profiling of *AP2/ERF* genes in the two species, transcript accumulation of collinear orthologous genes was further examined (Figure 6 and **Table S6**). For each ortholog, transcript levels of *AP2/ERF* genes in *G. arboreum* were also relatively highly correlated with those of the orthologous counterparts in *G. raimondii* (r = 0.8137, P = 1.74E-26), indicating their functional conservation and possible differentiation of the genomes of the two species. Among the 51 collinear orthologous gene groups examined, orthGrp29 and orthGrp82 were observed to have the highest transcription levels, while another two groups (orthGrp40 and orthGrp21) presented the lowest relative transcription levels in the two species. Some orthologs (e.g., orthGrp22, orthGrp46, orthGrp82, and orthGrp28) displayed a higher level of expression in *G. raimondii* than in *G. arboreum*. Other orthologs (e.g., orthGrp60, orthGrp01, and orthGrp07) showed the reverse trend. In addition, three orthologs (i.e., orthGrp22, orthGrp46, and orthGrp82) showed a higher level of transcription only in petal tissues in *G. raimondii* as compared to that in *G. arboreum*. Seven orthologs (i.e., orthGrp16, orthGrp25, orthGrp27, orthGrp35, orthGrp57, orthGrp59, and orthGrp63) were only transcribed in one of the two species. Although expression changes were ubiquitously measured between species, the Student *t*-test indicated that the differences between most orthologous genes were not significant. Microarray data indicated that in petals, eight orthologous gene groups (orthGrp01, orthGrp12, orthGrp22, orthGrp28, orthoGrp36, orthGrp40, orthGrp46, and orthGrp66) and half (35/72) of orthologous gene pairs showed significant changes in transcription levels between *G. raimondii* and *G. arboreum*. However, in the fibers, only two orthologs (orthGrp17 and orthGrp66) exhibited significant differences. Among these significant orthologs, the group orthGrp66 showed significant and opposite changes that occurred simultaneously in the two tissues. An RNA-Seq analysis of the petal transcriptome indicated that only one ortholog (orthGrp08) showed significantly more transcripts in *G. raimondii* than in *G. arboreum*. In each significant orthologous gene group, it is possible that each gene contributed to the divergence of two species.

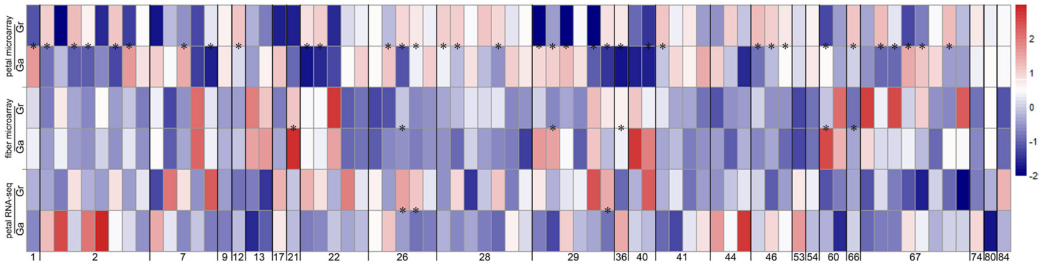

Figure 6. Heatmap representation of collinear orthologous gene pairs between *Gossypium arboreum* (Ga) and *G. raimondii* (Gr) genomes. The tissues and experiments employed in expression profiling are shown on the left of the heatmap. The numbers of orthologous gene groups are at the bottom of each column. Color scale bar is on the right. Red, high transcript level; blue, low transcript level. The expression bars in a column represent expression profiling of inter-species orthologous gene pairs. Asterisks show significant differences between the two species.

In summary, from a more broad perspective, *AP2/ERF* genes belonging to same groups exhibited similar expression patterns in *G. arboreum* and *G. raimondii* genomes, indicating the functional conservation of these genes between the two species. Additionally, *AP2/ERF* genes in orthologous gene groups exhibited minimal differences in transcript accumulation in *G. arboreum* compared to *G. raimondii*, which may contribute to divergence and the presence of distinct traits between the two species.

DISCUSSION

AP2/ERF genes in *Gossypium* species

We used the universal pfam accession Pfam00847 and the *Gossypium*-specific HMM profile of the conserved AP2/ERF domain in identifying the AP2/ERF superfamily member-encoding genes in *G. arboreum* and *G. raimondii*. To classify families, subfamilies, and groups, an MSA and phylogenetic analysis were conducted using the protein sequences of the conserved AP2/ERF domains of diploid *Gossypium*, with *A. thaliana* as reference.

The present study revealed that the two *Gossypium* species have extremely similar genomic features, which include the abundance and distribution of *AP2/ERF* members between the two diploid genomes. In contrast to *AP2/ERF* genes, the most recent analysis (Li et al., 2014) revealed a higher number of NBS genes in *G. raimondii* (391) than in *G. arboreum* (280). Conversely, the number of TE family members, especially LTR-RT families, was higher in *G. arboreum* than that in *G. raimondii*, which could be a driving factor in the size difference of the two genomes. *G. raimondii*, despite having a smaller genome (737.8 Mb), contains a similar number of *AP2/ERF* genes to *G. arboreum* (1694 Mb). These results suggest that in diploid *Gossypium* species, the AP2/ERF superfamily is characterized by a higher degree of conservation than is present in other gene (super-) families. The AP2/ERF superfamily underwent concerted/symmetrical evolutionary events (expansion and contraction) in *G. arboreum* and *G. raimondii*, indicating that *AP2/ERF* did not significantly contribute to the divergence of these species. Additionally, it appears that genus-specific constraints have influenced the number of *AP2/ERF* genes in a genome. For example, although *G. arboreum* has undergone domestication and thus was potentially more likely to experience artificial selection, the number of *AP2/ERF* genes was similar to that of *G. raimondii*. However, likely owing to selection pressures introduced by human breeding, intra-specific and inter-specific comparisons indicate that *AP2/ERF* genes are more variable among paralogous genes in *G. arboreum* than that in *G. raimondii*.

Genes in the subfamily IX are prevalent in both *G. raimondii* and *G. arboreum*, which might have been present before the divergence of the two diploid species. This subfamily possibly plays important roles in stress responses and signaling pathways (Champion et al., 2009). The prevalence of subfamily IX genes also indicates the need for more detailed investigations. In addition, ERF_III is the second most prevalent subfamily, the function of which remains unclear in plants, despite the abundance of ERF_III subfamily genes in *Gossypium*.

Compared to the other plant species, diploid cotton contains the highest number of *AP2/ERF* genes. In other species, the *AP2/ERF* gene complement apparently varies from as few as 124 in *Theobroma cacao* to 202 in *P. trichocarpa* (Zhuang et al., 2008), where the number of *AP2/ERF* genes in *A. thaliana*, *V. vinifera* (Licausi et al., 2010), *O. sativa* (Sharoni et al., 2011), *C. sinensis* & *clementine* (Xie et al., 2014), and *P. persica* (Zhang et al., 2012) fall between these two extremes. The only exception to this scheme is *B. rapa*, which has even more *AP2/ERF* genes (281) than diploid cotton. Our findings for diploid *Gossypium* species indicate that genome size and the number of *AP2/ERF* genes are not strongly correlated. Further comparisons across plant species indicate that the ERF_IX subfamily only predominates the *AP2/ERF* superfamily (23.15%) in *Gossypium*, which is different from trends observed in most other species (Table 1). Therefore, these cases require further exploration. The evolutionary mechanisms and underlying causes for the dramatic increase in the number of *AP2/ERF* genes in the diploid *Gossypium* species have yet to be identified. However, it is well known that all wild-type *Gossypium* species are capable of adapting to arid and stressful environments. The striking abundance of *AP2/ERF* in *G. arboreum* and *G. raimondii*, which largely evolved after the paleopolyploidy event (5-6-fold increase in ploidy), could play a role in conferring *Gossypium* species with intrinsic defenses for this characteristic adaption to arid and otherwise stressful environments. This type of association, between the number of members of a gene family and ecogeography, has been reported in barley (Kalendar et al., 2000). According to the ecogeography hypothesis (Kalendar et al., 2000), diploid *Gossypium* species might have evolved in habitats with higher levels of environmental stress.

The distribution of genes belonging to the *ERF* family in *G. arboreum* and *G. raimondii* was compared (Table 1). A four-fold higher number of genes from the family Soloist was observed than is present in other species (*A. thaliana*, *T. cacao*, *P. trichocarpa*, *V. vinifera*, *O. sativa*, and *P. persica*). While earlier study (Champion et al., 2009) employed EST assemblies of the two diploid cotton, we utilized the assembled genome as a reference to identify *AP2/ERF* genes, which is a more accurate determination of the numbers of *AP2/ERF* genes in the two diploid genomes. By comparing our identification with the list of *AP2/ERF* genes on the website http://datf.cbi.pku.edu.cn/ (Jin et al., 2014), we found an additional gene, *Gr01G252500*, belonging to subfamily IX; however, this gene only harbored one partial domain.

Phylogenetic analysis

Based on the multiple-amino acid sequence alignment of the AP2/ERF domains, a phylogenetic reconstruction of the *Arabidopsis* and *Gossypium AP2/ERF* superfamily genes was used to classify all the identified *AP2/ERF* genes in diploid *Gossypium*. According to Nakano's classification method (Nakano et al., 2006), which has been used previously as reference for organizing the AP2/ERF superfamily, *Gossypium AP2/ERF* genes were classified based on the homology of AP2/ERF domains between *Gossypium* and *Arabidopsis*. As the sequences of approximately 58 amino acids of the AP2/ERF domain were short, the phylogram displays a low resolution for the lower branches. However, the upper branches of the tree (data not shown) provided enough information for an effective classification. The tree divided all the genes into four families and 16 major subfamilies (Table 1), as described by Nakano et al. (2006). Most notably, *Cotton_A_11454* and *Cotton_A_17367*, as previously mentioned, encode double-AP2/ERF domains, but clustered with subfamilies V and Xb-L in the ERF family, which usually contain genes with single-AP2/ERF domain. Liu et al. (2013) also reported this discordance with *Bra034249* in *B. rapa*, but grouped it into the AP2_2 subfamily, which contains exclusively double-AP2/ERF domains. We feel that this finding (Liu et al, 2013) is not totally convincing. Alternatively, according to the classification of Sakuma et al. (2002), genes in the ERF family identified in the present study were classified into two subfamilies consisting of 163 DREB (subfamily I-IV) and 286 ERF (subfamily V-X) genes. The number of genes belonging to the ERF subfamily was 1.75-fold higher than that of the DREB subfamily in diploid *Gossypium*.

To explore and compare phylogenetic relationships between the two species, major subfamilies (i.e., IX and III) were individually sampled to conduct a phylogenetic analysis based on full-length protein sequences and several inconsistencies were observed in the groups (Figure 2 and **Figure S1**). Subfamily III had been previously divided into five groups (IIIa-IIIe); however, our phylogenetic analysis indicated that groups IIIa-IIIb were more closely related to subfamily II than to groups IIIc-IIIe. Subfamily IX was previously reported to be composed of three groups; however, according to the phylogenetic distance that was used to differentiate between groups IXa and IXb, group IXc should be divided further and should be composed of three groups. In addition, we observed that sequences from group IXc branched together with group IXa. Therefore, the assignment of AP2/ERF into its subfamily and group can be conducted based on the homology of the AP2/ERF domains, but should be revised based on information from full-length protein sequences.

Significantly, the number of *AP2/ERF* genes in each clade (or terminal branch) for the two species was nearly identical. Despite this high degree of similarity, the observed differences between the two species suggest that either one species underwent gene loss or the other species experienced more tandem and ectopic duplications. Overall, this well-established phylogenetic correlation facilitates the tracing of the evolution of *AP2/ERF* genes between wild-type *G. raimondii* and its relative crop species, *G. arboreum*, as well as among the members of the *AP2/ERF* superfamily.

Collinearity

Collinearity analysis, by means of a robust method involving stringent criteria applied to detect collinear blocks, can provide abundant information regarding evolution. Li et al. (2014) found that the genomes of *G. arboreum* and *G. raimondii* shared a total of 780 collinear blocks that encompassed 73 and 88% of the assembled chromosomes of *G. arboreum* and *G. raimondii*, respectively. In the present study, further elucidation, focused on *AP2/ERF* genes, showed that in *G. arboreum*, 98 *AP2/ERF* genes were involved in intra-species collinear blocks, with their paralogous counterparts present, while in *G. raimondii*, 195 *AP2/ERF* genes were involved in intra-species collinear blocks, with their paralogous counterparts present. In the comparison of *G. arboreum vs G. raimondii*, 207 and 226 *AP2/ERF* genes were situated in inter-species collinear blocks, which were involved in 435 inter-species chromosome-segmental alignments. As the quality of the sequence assembly of *G. arboreum* and *G. raimondii* was equivalent, the aforementioned observations from collinear analyses could potentially mean that in *G. arboreum*, a lower number of *AP2/ERF* genes were derived from segmental duplication compared to *G. raimondii*. On the other hand, after the divergence of the two species, orthologous groups likely encountered more mutations or shuffling within *G. arboreum* chromosomes than in *G. raimondii*, resulting in a lower number of genes situated in inter-species collinear regions.

Additionally, collinearity and cluster analyses indicate that in *G. raimondii*, only one cluster of 38 did not display an inter-species collinear orthologous counterpart in *G. arboreum*, and that this cluster was heterogeneous. However, among the 36 identified clusters in *G. arboreum*, three homogenous and two heterogeneous clusters showed no collinear orthologous counterparts in *G. raimondii*. It may be possible that the regions containing these clusters experienced greater mutation rates, resulting in the absence of collinear orthologous counterparts in closely related species. On the other hand, in each species, the *AP2/ERF* genes present in the clusters could have facilitated the generation of genes with novel functions via genome duplication, tandem duplication, and gene recombination (Friedman and Baker, 2007). Therefore, the intra-species collinear paralogous counterparts were examined for every *AP2/ERF* cluster. We found that 29 clusters (of 38) in *G. raimondii* and 17 (of 36) in *G. arboreum* have respective paralogous counterparts. The collinear analysis of each species also indicated that 98 *GaAP2/ERF* and 195 *GrAP2/ERF* genes were located within intra-species segmental duplications. These data indicate that, following the WGD events or paleopolyploidization, the genomic instability and evolutionary dynamics of *G. arboreum* was greater than in *G. raimondii*. It is possible that *G. arboreum*, having been domesticated, underwent human-mediated selection in addition to natural selection over time.

Nucleotide divergence

The present study examined and compared nucleotide substitution rates between two members of individual orthologous *AP2/ERF* pairs, and namely, synonymous and non-synonymous substitution rates of duplicated genes (conserved in both genomes) based on a shared WGD event. Overall, *AP2/ERF* sequences featured a higher number of nucleotide substitutions in *G. arboreum* than in *G. raimondii*. We detected a higher rate of synonymous substitutions of duplicated genes retained from a shared WGD in *G. raimondii* compared to *G. arboreum*. In addition, lower rates of non-synonymous substitutions were detected in *G. raimondii* than in *G. arboreum*, which was indicative of a shift in the densities of purifying selection. These analyses also uncovered the inter-specific evolutionary asymmetry of *AP2/ERF* genes and the potential underlying causes of the observed differences.

Expression divergence

The abundance of *AP2/ERF* genes in the two genomes, the levels of collinearity, and the divergence of *AP2/ERF* sequences together indicate significant structural similarity along with stark distinctions between the two species. However, are *AP2/ERF* genes from these sister species characterized by similar expression profiles? Microarray data and RNA-seq data were collected to determine differences in expression levels of collinear orthologous gene-pairs in *G. arboreum* and *G. raimondii*. To this end, we measured the transcript abundance of 321 (185 by microarray and 249 by RNA-seq) *AP2/ERF* genes. This (transcript abundance of "not all 540" genes were measured) could be attributed to the high level of canalization occurring in petal and fiber tissues and only one time point in sampling, resulting in the measurement of transcripts of the subset of the AP2/ERF superfamily. In the present study, the transcript levels of only 249 of 540 *AP2/ERF* genes (accounting for 46.11%) could be detected examined by RNA-seq, which is free from the ascertainment bias of a template sequence in comparison to microarray analyses. Our analyses show that most orthologous genes were expressed at similar levels, although a subset was expressed by a higher number of *G. arboreum* (A2) or *G. raimondii* (D5) copies. The directionality of bias in gene expression varied among the members of the AP2/ERF superfamily. Significant interaction in the genes and tissues also indicates inverse expression patterns in various tissues. A previous study surmises that most of the paralogous pairs in *Gossypium* spp exhibited expression divergence prior to divergence of the two species (Renny-Byfield et al., 2014). Our comparison indicates that most of the orthologous pairs exhibit similar (yet not significant) expression levels. Rambani et al. (2014) observed that it was rare for genes that were not differentially expressed in diploids to be differentially expressed in tetraploids. Extensive characterization of orthologous pairs exhibiting expression divergence between two diploid *Gossypium* species is therefore warranted.

Ethylene is an important signaling modulator that promotes cotton fiber growth in *G. hirsutum* (Shi et al., 2006). The extremely high amounts of ethylene in the D-genome and the significantly decreased levels in the A-genome were not beneficial to the development of fibers. The *ERF* is involved in the ethylene pathway; however, its expression profile indicates that, for the minority of *ERF* genes, the levels of *ERF* expression differs among tissues [i.e., leaf, 0 days post-anthesis (dpa) fibers, and 3 dpa fibers]. A previous analysis has shown that the disease-resistant (R) genes underwent extensive expansion in the D-genome, whereas contraction occurred in the A-genome (Li et al., 2014). In the aforementioned study, the genes that underwent expansion in *G. raimondii* elicited a positive response to *Verticillium* infection at the transcription level, whereas in *G. arboreum*, only one gene showed slight up-regulation. Does the similar number of *AP2/ERF* genes in the two species, along with the small number of *AP2/ERF* genes showing differential expression between the two diploids, imply that *AP2/ERF* genes do not significantly contribute to the different traits (e.g., fiber production, fiber quality, environmental tolerances, etc.) of *G. raimondii* and *G. arboreum*? Further investigations will be necessary to determine the exact biological contribution of these genes. Our assessment of ω values and expression profiles has shown that the functional divergence following gene duplications is relatively limited in the two *Gossypium* species. Our observations have shown that the functional evolution of the AP2/ERF superfamily is extensively conserved across the two species.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the National Transgenic Plants Project of China (#2014ZX08005-002), the China Agriculture Research System (#CARS-18-45), and Cyrus Tang Foundation.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, et al. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410. http://dx.doi.org/10.1016/S0022-2836(05)80360-2

Bailey TL and Gribskov M (1998). Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* 14: 48-54. http://dx.doi.org/10.1093/bioinformatics/14.1.48

Bennetzen JL (2007). Patterns in grass genome evolution. *Curr. Opin. Plant Biol.* 10: 176-181. http://dx.doi.org/10.1016/j.pbi.2007.01.010

Champion A, Hebrard E, Parra B, Bournaud C, et al. (2009). Molecular diversity and gene expression of cotton ERF transcription factors reveal that group IXa members are responsive to jasmonate, ethylene and *Xanthomonas. Mol. Plant Pathol.* 10: 471-485. http://dx.doi.org/10.1111/j.1364-3703.2009.00549.x

Chaudhary B, Hovav R, Flagel L, Mittler R, et al. (2009). Parallel expression evolution of oxidative stress-related genes in fiber from wild and domesticated diploid and polyploid cotton (*Gossypium*). *BMC Genomics* 10: 378. http://dx.doi.org/10.1186/1471-2164-10-378

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. http://dx.doi.org/10.1093/nar/gkg500

Ding J, Araki H, Wang Q, Zhang P, et al. (2007). Highly asymmetric rice genomes. *BMC Genomics* 8: 154. http://dx.doi.org/10.1186/1471-2164-8-154

Duan C, Argout X, Gébelin V, Summo M, et al. (2013). Identification of the Hevea brasiliensis *AP2/ERF* superfamily by RNA sequencing. *BMC Genomics* 14: 30. http://dx.doi.org/10.1186/1471-2164-14-30

Finn RD, Coggill P, Eberhardt RY, Eddy SR, et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44 (D1): D279-D285. http://dx.doi.org/10.1093/nar/gkv1344

Flagel LE and Wendel JF (2010). Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *New Phytol.* 186: 184-193. http://dx.doi.org/10.1111/j.1469-8137.2009.03107.x

Friedman AR and Baker BJ (2007). The evolution of resistance genes in multi-protein plant resistance systems. *Curr. Opin. Genet. Dev.* 17: 493-499. http://dx.doi.org/10.1016/j.gde.2007.08.014

Guo YL, Fitz J, Schneeberger K, Ossowski S, et al. (2011). Genome-wide comparison of nucleotide-binding site-leucine-rich repeat-encoding genes in *Arabidopsis*. *Plant Physiol.* 157: 757-769. http://dx.doi.org/10.1104/pp.111.181990

Hurwitz BL, Kudrna D, Yu Y, Sebastian A, et al. (2010). Rice structural variation: a comparative analysis of structural variation between rice and three of its closest relatives in the genus *Oryza*. *Plant J.* 63: 990-1003. http://dx.doi.org/10.1111/j.1365-313X.2010.04293.x

Jin J, Zhang H, Kong L, Gao G, et al. (2014). PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res.* 42: D1182-D1187. http://dx.doi.org/10.1093/nar/gkt1016

Jin LG and Liu JY (2008). Molecular cloning, expression profile and promoter analysis of a novel ethylene responsive transcription factor gene *GhERF4* from cotton (*Gossypium hirstum*). *Plant Physiol. Biochem.* 46: 46-53. http://dx.doi.org/10.1016/j.plaphy.2007.10.004

Kalendar R, Tanskanen J, Immonen S, Nevo E, et al. (2000). Genome evolution of wild barley (*Hordeum spontaneum*) by *BARE-1* retrotransposon dynamics in response to sharp microclimatic divergence. *Proc. Natl. Acad. Sci. USA* 97: 6603-6607. http://dx.doi.org/10.1073/pnas.110587497

Li F, Fan G, Wang K, Sun F, et al. (2014). Genome sequence of the cultivated cotton *Gossypium arboreum*. *Nat. Genet.* 46: 567-572. http://dx.doi.org/10.1038/ng.2987

Licausi F, Giorgi FM, Zenoni S, Osti F, et al. (2010). Genomic and transcriptomic analysis of the *AP2/ERF* superfamily in *Vitis vinifera. BMC Genomics* 11: 719. http://dx.doi.org/10.1186/1471-2164-11-719

Liu Z, Kong L, Zhang M, Lv Y, et al. (2013). Genome-wide identification, phylogeny, evolution and expression patterns of *AP2/ERF* genes and cytokinin response factors in *Brassica rapa* ssp. *pekinensis. PLoS One* 8: e83444. http://dx.doi.org/10.1371/journal.pone.0083444

Meyers BC, Kozik A, Griego A, Kuang H, et al. (2003). Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis. Plant Cell* 15: 809-834. http://dx.doi.org/10.1105/tpc.009308

Nakano T, Suzuki K, Fujimura T and Shinshi H (2006). Genome-wide analysis of the ERF gene family in *Arabidopsis* and rice. *Plant Physiol.* 140: 411-432. http://dx.doi.org/10.1104/pp.105.073783

Ohta M, Matsui K, Hiratsu K, Shinshi H, et al. (2001). Repression domains of class II ERF transcriptional repressors share an essential motif for active repression. *Plant Cell* 13: 1959-1968. http://dx.doi.org/10.1105/tpc.13.8.1959

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: 423-427. http://dx.doi.org/10.1038/nature11798

Qiao ZX, Huang B and Liu JY (2008). Molecular cloning and functional analysis of an ERF gene from cotton (Gossypium hirsutum). *Biochim. Biophys. Acta* 1779: 122-127. http://dx.doi.org/10.1016/j.bbagrm.2007.10.003

Rambani A, Page JT and Udall JA (2014). Polyploidy and the petal transcriptome of *Gossypium. BMC Plant Biol.* 14: 3. http://dx.doi.org/10.1186/1471-2229-14-3

Renny-Byfield S, Gallagher JP, Grover CE, Szadkowski E, et al. (2014). Ancient gene duplicates in *Gossypium* (cotton) exhibit near-complete expression divergence. *Genome Biol. Evol.* 6: 559-571. http://dx.doi.org/10.1093/gbe/evu037

Sakuma Y, Liu Q, Dubouzet JG, Abe H, et al. (2002). DNA-binding specificity of the ERF/AP2 domain of *Arabidopsis* DREBs, transcription factors involved in dehydration- and cold-inducible gene expression. *Biochem. Biophys. Res. Commun.* 290: 998-1009. http://dx.doi.org/10.1006/bbrc.2001.6299

Sampath P, Murukarthick J, Izzah NK, Lee J, et al. (2014). Genome-wide comparative analysis of 20 miniature inverted-repeat transposable element families in *Brassica rapa* and *B. oleracea. PLoS One* 9: e94499. http://dx.doi.org/10.1371/journal.pone.0094499

Senchina DS, Alvarez I, Cronn RC, Liu B, et al. (2003). Rate variation among nuclear genes and the age of polyploidy in *Gossypium. Mol. Biol. Evol.* 20: 633-643. http://dx.doi.org/10.1093/molbev/msg065

Sharoni AM, Nuruzzaman M, Satoh K, Shimizu T, et al. (2011). Gene structures, classification and expression models of the AP2/EREBP transcription factor family in rice. *Plant Cell Physiol.* 52: 344-360. http://dx.doi.org/10.1093/pcp/pcq196

Shi YH, Zhu SW, Mao XZ, Feng JX, et al. (2006). Transcriptome profiling, molecular biological, and physiological studies reveal a major role for ethylene in cotton fiber cell elongation. *Plant Cell* 18: 651-664. http://dx.doi.org/10.1105/tpc.105.040303

Stoller JZ and Epstein JA (2005). Identification of a novel nuclear localization signal in Tbx1 that is deleted in DiGeorge syndrome patients harboring the 1223delC mutation. *Hum. Mol. Genet.* 14: 885-892. http://dx.doi.org/10.1093/hmg/ddi081

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. http://dx.doi.org/10.1093/molbev/msr121

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. http://dx.doi.org/10.1038/nprot.2012.016

van Berloo R (2008). GGT 2.0: versatile software for visualization and analysis of genetic data. *J. Hered.* 99: 232-236. http://dx.doi.org/10.1093/jhered/esm109

Wang Y, Tang H, Debarry JD, Tan X, et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40: e49-e49. http://dx.doi.org/10.1093/nar/gkr1293

Xie XL, Shen SL, Yin XR, Xu Q, et al. (2014). Isolation, classification and transcription profiles of the AP2/ERF transcription factor superfamily in citrus. *Mol. Biol. Rep.* 41: 4261-4271. http://dx.doi.org/10.1007/s11033-014-3297-0

Samuel Yang S, Cheung F, Lee JJ, Ha M, et al. (2006). Accumulation of genome-specific transcripts, transcription factors and phytohormonal regulators during early stages of fiber cell development in allotetraploid cotton. *Plant J.* 47: 761-775. http://dx.doi.org/10.1111/j.1365-313X.2006.02829.x

Yang Z (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24: 1586-1591. http://dx.doi.org/10.1093/molbev/msm088

Yu J, Tehrim S, Zhang F, Tong C, et al. (2014). Genome-wide comparative analysis of NBS-encoding genes between *Brassica* species and *Arabidopsis thaliana. BMC Genomics* 15: 3. http://dx.doi.org/10.1186/1471-2164-15-3

Zhang CH, Shangguan LF, Ma RJ, Sun X, et al. (2012). Genome-wide analysis of the *AP2/ERF* superfamily in peach (*Prunus persica*). *Genet. Mol. Res.* 11: 4789-4809.

Zhuang J, Cai B, Peng RH, Zhu B, et al. (2008). Genome-wide analysis of the *AP2/ERF* gene family in *Populus trichocarpa. Biochem. Biophys. Res. Commun.* 371: 468-474. http://dx.doi.org/10.1016/j.bbrc.2008.04.087

Supplementary material

Figure S1. Phylogenetic relationship of subfamily ERF_III genes in *Gossypium arboreum* (Ga) and *G. raimondii* (Gr). After an alignment of the complete sequences of subfamily proteins by ClustalW, construction of a maximum likelihood tree was performed with the MEGA 5.0 software using 1000 replications. Subfamily III was divided into five groups (IIIa-IIIe).

Figure S2. Network of inter-species and intra-species collinearity of the *AP2/ERF* genes in the largest orthologous-gene group. The MCScanX software allowed the detection of collinear relationships existing between the *AP2/ERF* genes. Nodes represent genomic regions containing the genes and the edges indicate collinearity. The pink and yellow nodes indicate genes belonging to *Gossypium arboreum* (Ga) and *G. raimondii* (Gr), respectively. Blue, green, and purple edges show inter-species (*G. arboreum vs G. raimondii*) and intra-species (*G. arboreum vs G. arboreum*, and *G. raimondii vs G. raimondii*) collinearity.

Figure S3. Sequence logo of motifs (identified by MEME suite; except Motif A, Motif B, Motif D, and Motif E) in the superfamily of AP2/ERF proteins from *Gossypium arboreum* and *G. raimondii*. A-F: Motif C, Motif F-Motif T.

Figure S4. Number of motif combinations for all AP2/ERF members identified in *Gossypium arboreum* and *G. raimondii*.

Figure S5. Number of AP2/ERF members containing the four motifs (PY_NLS, bipartite, monopartite, and EAR) in the two *Gossypium* species. **A.** *G. arboreum*; **B.** *G. raimondii*.

Figure S6. Comparative analysis of Ka/Ks values for AP2/ERF gene pairs between *Gossypium arboreum* (Ga) and *G. raimondii* (Gr), *G. arboreum* and *G. arboreum*, and *G. raimondii* and *G. raimondii*. **A.** y-axis and x-axis indicate values and orthologous gene groups in *G. arboreum* compared to *G. raimondii*, respectively. Ga-Gr: Ka/Ks values of pairs of orthologous genes of *G. arboreum vs G. raimondii*. Ga-Ga and Gr-Gr: Ka/Ks values of paralogous gene pairs for each orthologous pair of *G. arboreum vs G. arboreum* and *G. raimondii vs G. raimondii*, respectively. Intra-specific sequence divergence in two paralogous AP2/ERF genes of each of the 30 orthologous gene group of *G. arboreum vs G. raimondii*, and inter-specific sequence divergence of the two genomes. **B.** Boxplot of intra-specific and inter-specific sequence divergence between the two AP2/ERFs of each of the 30 orthologous AP2/ERF groups of *G. arboreum* (A2) and *G. raimondii* (D5). The whiskers above boxes show the highest value after excluding the outliers; the whiskers below boxes indicate the lowest value after outlier exclusion.

Figure S7. Synonymous substitution rate (Ks) distributions of *Gossypium arboreum* and *G. raimondii*. Blue line: Ks of paralogous*AP2/ERF* gene pairs in *G. arboreum*; green line: Ks of paralogous *AP2/ERF* gene pairs in *G. raimondii*; pink line: Ks of orthologous *AP2/ERF* gene pairs. The vertical axis represents the frequency of sequence pairs, whereas the horizontal axis indicates the Ks values.

Table S1. Information regarding the AP2/ERF-encoding genes in the *Gossypium arboreum* genome. This table contains the number of AP2/ERF domains identified by HMMsearch, the bit_score of four BLAST searches (rpsblast, psiblast, rpstblastn, and tblastn), the length of the AP2/ERF domain (aa), subfamily, cluster (in which AP2/ERF genes are located), collinear paralogous/orthologous gene group, and motif profile (motif arrangement).

Table S2. Information regarding the AP2/ERF-encoding genes in the *Gossypium raimondii* genome. This table contains the number of AP2/ERF domains identified by HMMsearch, the bit_score of four BLAST searches (rpsblast, psiblast, rpstblastn, and tblastn), the length of the AP2/ERF domain (aa), subfamily, cluster (in which AP2/ERF genes are located), collinear paralogous/orthologous gene group, and motif profile (motif arrangement).

Table S3. Diverse motif combinations identified in the AP2/ERF superfamily in *Gossypium arboreum* and *G. raimondii*.

Table S4. ERF-associated amphiphilic repression (EAR) motif and nuclear localization signal (NLS) sequences identified in the AP2/ERF superfamily of *Gossypium arboreum* and *G. raimondii*.

Table S5. Inter-specific comparison of *AP2/ERF* sequence divergence in *Gossypium arboreum* (Ga) and *G. raimondii* (Gr). The calculation was based on orthologous gene pairs identified to be collinear in each orthologous gene group (orthGrp). P values are the result of the Student *t*-test.

Table S6. Transcript abundance of orthologous *AP2/ERF* gene pairs. P values were the result of Student *t*-tests between orthologous *AP2/ERF* gene pairs and between orthologous *AP2/ERF* gene groups.