Molecular Biology Reports (2019) 46:161–175
https://doi.org/10.1007/s11033-018-4457-4

ORIGINAL ARTICLE

Phylogenetic analysis of upland cotton MATE gene family reveals a conserved subfamily involved in transport of proanthocyanidins

Lei Xu[1] · Zhi-Lin Shen[1] · Wen Chen[1] · Guo-Yang Si[1] · Yan Meng[1] · Ning Guo[1] · Xu Sun[1] · Yong-Ping Cai[1] · Yi Lin[1] · Jun-Shan Gao[1]

Received: 13 August 2018 / Accepted: 30 October 2018 / Published online: 22 November 2018

Abstract

The multidrug and toxic compound extrusion (MATE) protein belongs to a secondary transporter family, which plays a role in transporting different kinds of substrates like phytohormones and secondary metabolites. In plant, MATE transporters related to the endogenous and exogenous mechanisms of detoxification for secondary metabolites such as alkaloids, flavonoids, anthocyanins and other secondary metabolites have been studied. However, a genome-wide analysis of the MATE family is rarely reported in upland cotton (*Gossypium hirsutum* L.). In the study, a total of 72 *GhMATE*s were identified from the genome of upland cotton, which were classified into four subfamilies with possible diverse functions such as transport of proanthocyanidins (PAs), accumulation of alkaloids, extrusion of xenobiotic compounds, regulation of disease resistance and response to abiotic stresses. Meanwhile, the gene structure, evolutionary relationship, physical location, conservative motifs, subcellular localization and gene expression pattern of *GhMATE*s have been further analysed. Three of these *MATE* genes (*GhMATE12, GhMATE16* and *GhMATE38*) were identified as candidate genes due to their functions in transport of PA similar to *GhTT12*. These results provide a new perspective on upland cotton MATE gene family for their potential roles in transport of PA and a theoretical basis for further analyzing the function of *MATE* genes and improving the fiber quality of brown cotton.

Keywords *Gossypium hirsutum* · MATE · Gene structure · Phylogenetic analysis · Expression pattern · Proanthocyanidin

Introduction

The extrusion of toxins and substances is a vital life process in organisms. Among the identified transporter families, there are several large superfamilies such as the major facilitator superfamily (MFS), the ATP-binding cassette (ABC) superfamily, the drug metabolite transporter (DMT)

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11033-018-4457-4) contains supplementary material, which is available to authorized users.

✉ Jun-Shan Gao
gaojsh@ahau.edu.cn

Lei Xu
leixu@ahau.edu.cn

Zhi-Lin Shen
zlshen@ahau.edu.cn

Wen Chen
chenwen@ahau.edu.cn

Guo-Yang Si
sgy@ahau.edu.cn

Yan Meng
mengyan@ahau.edu.cn

Ning Guo
81_ahau@ahau.edu.cn

Xu Sun
sunxu@ahau.edu.cn

Yong-Ping Cai
swkx01@ahau.edu.cn

Yi Lin
yjsc01@ahau.edu.cn

[1] School of Life Sciences, Anhui Agricultural University, Hefei 230036, China

superfamily, the resistance/nodulation/division (RND) superfamily and the multidrug and toxic compound extrusion (MATE) family [1]. The MATE gene family is widely distributes in prokaryotes and eukaryotes, playing a key role in the exportation of toxins and substrates [2]. MATE transport proteins are composed of 400–700 amino acids, most of which contain 12 transmembrane helices with a 40% similarity among their amino acid sequences. The MATE family depends on an transmembrane electrochemical gradient for extrusion activity [3]. To date, 17 MATE transporters identified from 11 species of bacteria are involved in heterologous material efflux and drug resistance. In mammals, MATE transporters can exhaust a variety of cationic substances in liver and kidney to adjust the drug concentration in plasma [4]. In plants, many studies and expression analyses of *MATE* genes have been studied. The MATE gene family expands through tandem and segmental duplication in both rice and Arabidopsis [5]. A genome-wide analysis of MATE family in maize and soybean is conducted to identify some *MATE* genes in response to aluminum toxicity, respectively [6, 7], and other *MATE* genes are involved in endogenous and exogenous mechanisms of detoxification associated with secondary metabolites such as alkaloids, flavonoids, anthocyanins and other secondary metabolites apart from toxic substances [8–10].

Recently, a genome-wide analysis of the MATE family in diploid cotton (*Gossypium raimondii* and *Gossypium arboreum*) and their expression levels under salt, cadmium and drought stresses has been published [11], but no work has been reported in tetraploid cotton (*Gossypium hirsutum*). In the available reports, most of *MATE* genes play mainly roles in responses to abiotic stress in plants including diploid cotton, and only a few literatures reported that *MATE* genes are involved in the transpor of secondary metabolites. Our previous study found that *GhTT12* (Genbank accession no. KF240564), one gene of a MATE family cloned in upland cotton (*G. hirsutum*), was involved in the secondary active transport of proanthocyanidin (PA) as a secondary carrier [12]. In a upland cotton variety with a naturally brown fiber (brown cotton), the major substance of its fiber pigment is demonstrated to be PA [13, 14].

PA, known as condensed tannin, is a flavonoid involved in the flavonoid or anthocyanin branch of the phenylpropanoid pathway and occurs as a specialized metabolite in massive higher plants for the interaction between the plant and their environment [15]. PA plays important roles in the resistance of disease for herbivores, seed dormancy and viability [16, 17]. PA is an oligomer of flavan-3-ol unit, which exists as a prominent compound in seed coat, leaf, fruit, flower, and bark of many plant species [18–20]. PAs and their assumed precursors like catechin and epicatechin are beneficial for cardiovascular diseases, resistibility, and life-span [21, 22]. Although the transporters mediating the precursor flavan-3-ol of PA biosynthesis into the vacuole have been identified [23, 24], the mechanism of PA transport is not yet understood.

Gossypium hirsutum is a heterologous tetraploid cotton containing AA and DD genomes, which was formed about 1–2 million years ago and is widely assumed that the donators of A and D chromosomes are from *G. arboretum* and *G. raimondii*, respectively [25, 26]. With the sequenced genomes of these cotton species [26, 27], a whole-genome analysis was carried out on the sequence characteristics, phylogenetic evolution and expression patterns of the MATE family in upland cotton. The results provide a theoretical basis for further analyzing the function of *MATE* genes and improving the fiber pigment stability of brown cotton.

Materials and methods

Plant materials

Upland cotton (*G. hirsutum*) was grown in the farm of Anhui Agricultural University, Hefei, China. The RNA of cotton fibers at 5, 10, 15, 20 and 25 days post anthesis (DPA) was isolated for this study.

Database search and sequence retrieval

The upland cotton genome sequences were downloaded to identify *GhMATE*s from COTTON GENOME PROJECT (http://cgp.genomics.org.cn/) followed by constructing local databases with DNATOOLS. The conserved MatE domain (PF01554) was obtained from Pfam database based on the hidden Markov model (HMM) [28], and was searched in the cotton protein database by BLASTP. All identified proteins were detected for the MatE domain by Pfam (http://pfam.sanger.ac.uk/) [29] and SMART (http://smart.embl-heidelberg.de/) [30]. Finally, all possible MATE sequences were calibrated using ClustalW [31] and potentially redundant genes were removed. The remaining MATE genes were used for further analysis. The ExPASy program (http://www.expasy.org/tools/) [32] was utilized to calculate the molecular weight and isoelectric point.

Phylogenetic analysis

For further analysis, multiple sequence alignments were clustered by ClustalX [31]. The phylogenetic tree was constructed by MEGA5.0 using the maximum likelihood (ML) method with partial deletion for gaps and missing data. For statistical reliability, the significance of each node evaluated was bootstrap analysis with 1000 replicates [33].

Gene structure and conserved motif analyses

To comprehend the structure of *GhMATEs*, the corresponding genomic sequences were aligned by GSDS (http://gsds.cbi.pku.edu.cn/) [34]. The MEME (http://meme.sdsc.edu/meme4.3.0/intro.html) [35] were used to detect the conserved motifs that were appraised by the SMART (http://smart.embl-heidelberg.de/) and Pfam (http://pfam.sanger.ac.uk/search) programs according to a previous study [36].

Chromosomal location and gene duplication

The chromosome mapping of *GhMATEs* was illustrated by MapInspect software (http://www.plantbreeding.wur.nl/uk/software_mapinspect.html). According to previous reports, MCScanX (http://chibba.pgml.uga.edu/mcscan2/) could be used to identify the duplications of *GhMATEs* [37, 38]. BLASTP with an E-value cutoff of 1×10^{-10} was used to combine and search against protein sequences in the whole cotton genome. Moreover, the default parameters were set to identify the synteny regions, where the duplicated *MATEs* consisting of various types of duplications such as collinearity, tandem, etc. were identified by a Perl script.

Detection of orthologous gene pairs

Every *GhMATE* was marked on the respective chromosomes through a Perl script. OrthoMCL (http://orthomcl.org/orthomcl/) was used to distinguish the orthologous *MATE* genes among *G. hirsutum, G. arboreum* and *G. Raimondii* [39]. Circos (http://circos.ca/) was utilized for comprehending the relationships between the orthologous gene pairs among *G. hirsutum, G. arboreum* and *G. Raimondii* [40].

Expression analysis of *GhMATEs*

To clarify the expression patterns of the *MATE* genes in upland cotton, transcriptome information was obtained from PLEXdb (http://www.plexdb.org/index.php), and the cotton fibers at different developmental stages (0 DPA, 5 DPA, 10 DPA, 20 DPA) were selected for further analysis. A heatmap was constructed by R/Bioconductor (http://www.bioconductor.org/).

RNA isolation and qRT-PCR

To confirm the expression of *GhMATEs*, total RNA was prepared using RNAprep Pure Plant Kit (Tiangen, Beijing), and genomic DNA was removed by DNase I. The isolated RNA was synthesized into the first chain cDNA by reverse transcription using TIANGEN FastQuant RT Kit (Tiangen, Beijing). qRT-PCR was conducted by an ABI PRISM 7500 real-time PCR system (Applied Biosystems, USA) with *GhUBQ* (Genbank accession no. DQ116441) as an internal control. Each reaction contains 10 µL SYBR Premix Ex Taq II (2×), 1.0 µL cDNA sample, and 0.5 µL of forward and reverse primers in a final volume of 20 µL. Each pair of primers were designed according to the nucleotide sequences of genes by Primer Express 3.0 software (Applied Biosystems, USA). The primers sequences were 5′-GTATTGGTTCTCTGGTGACG-3′ and 5′-GTAAGAGCAGCACCAAGTAG-3′ for *GhTT12*, and 5′-GAAGGCATTCCACCTGACCAAC-3′ and 5′-CTTGACCTTCTTCTTCTTGTGCTTG-3′ for *GhUBQ7*. PCR was performed as follows: 50 °C for 2 min, 95 °C for 30 s, 95 °C for 15 s, and 40 cycles of 60 °C for 34 s, 72 °C for 20 s. The specificity of the reactions was verified by melting curve analysis. The $2^{-\Delta\Delta CT}$ method was selected to calculate the relative gene expression levels [41].

Determination of the PA content

According to the report [42], 100 mg cotton fibers at 5, 10, 15, 20, 25 DPA were taken into the precooling mortar. 5 mL 80% methanol was added for grinding and the samples were transferred into 10 mL centrifuge tube with ultrasonic treatment for 30 min, then centrifuged at 4 °C, 5000 rpm for 10 min. The supernatant was soluble PA. The precipitation was resuspended with 5 mL methanol containing 1% (v/v) HCl, and water bathed at 60 °C for 1 h, then centrifuged at 4 °C, 5000 rpm for 10 min. The supernatant was insoluble PA. According to the method reported in the reference [43], the content of PA was determined by spectrophotometry. The catechin was taken as a comparison to make the standard curve. 200 µL PA extract and 2800 µL reaction mixture (methanol solution containing 5% HCl and 0.1% DMACA) were added to 10 mL centrifuge tube, reacting at room temperature for 15 min, and then the absorbance was measured at 643 nm. The PA content was calculated by consulting the standard curve.

Arabidopsis transformation and GUS assay

The recombinant plasmid pCambia1304-GhTT12 was transformed into *Agrobacterium tumefaciens* EHA105, and *Arabidopsis* transformation was performed by the floral dip method. The transformed plants were selected on MS medium supplemented with 50 mg L^{-1} hygromycin. The positive plants were transferred to soil in the greenhouse at 25 °C under a 16 h light/8 h dark photoperiod, and were further confirmed by PCR. Finally, T2 transgenic lines with single-copy insertion were used for GUS assay according to the instruction of GUS dyeing kit (Solarbio, Beijing). Moreover, seeds from transformed plants were obtained for qRT-PCR and the determination of PA content, with those from the wild type plants as a control.

Results

Identification of *MATE* genes in upland cotton genome

A total of 112 candidate *MATE* genes were originally identified through the HMM profile of the MatE domain, in which 40 reduplicate genes were removed. The results showed that 72 putative *MATE* genes contained a conserved MATE domain in *G. hirsutum*. These *MATE* genes were denominated as *GhMATE1-72* according to the location on the chromosomes in *G. hirsutum* (Table S1). The length of the encoded proteins varied from 313 to 602 aa with an average of 488 aa, whereas the length in *A. thaliana* ranged from 400 to 700 aa, suggesting that similar variations existed in the MATE family. The predicted molecular weight and pI of the GhMATE proteins ranged from 33.5 to 66 kDa, from 4.97 to 10.24, respectively. The subcellular locations of these *GhMATE*s were predicted as follows: plasma membrane (73.61%), chloroplast (8.33%), cytoplasm (8.33%), vacuole (5.56%), nucleus (1.39%), mitochondria (1.39%) and extracellular (1.39%).

Phylogenetic analyses of the MATE family in upland cotton

A compounded phylogenetic tree of 117 MATE proteins including 72 GhMATEs and 45 reported plant MATEs was constructed (Fig. 1). The MATE proteins were preliminarily divided into four clades designated C1–C4. Clade C1 was further divided into five subgroups, named as C1-1, C1-2, C1-3, C1-4, C1-5. Subgroup C1-1 consisted of 17 members, including 9 GhMATE proteins and 8 reported MATE transporters. There were 9 GhMATE proteins in the C1-2 subgroup, and only one known MATE protein. There were 7 proteins in the C1-3 subgroup including 3 GhMATE proteins. Subgroup C1-4 contained 3 GhMATE members and 7 known MATE proteins. Subgroup C1-5 had 6 GhMATE proteins and a known MATE protein. Clade C2 contained 15 GhMATE members in addition to several reported MATE proteins. Clade C3 contained 20 GhMATE proteins and 4 known MATE proteins. Clade C4 had 24 MATE members including 7 GhMATE proteins and 17 reported MATE proteins from other plant species.

Gene structures and motifs analyses of *GhMATEs*

To gain a more particular knowledge about the structural diversity of the *MATE* genes, their structures were subsequently analyzed by comparing the CDS of each *MATE* gene in upland cotton. The results revealed that 11 *GhMATE* genes (*GhMATE7, 8, 20, 42, 45, 50, 51, 53, 57, 68* and *70*) had no introns in their gene structures, while the remaining *GhMATE* genes contained 1–14 introns (Fig. 2). In addition, the *GhMATE* genes in subfamily C4 contained the largest numbers of introns. For example, *GhMATE19, GhMATE29, GhMATE54, GhMATE72* had 12 introns and *GhMATE49* had 11 ones (Fig. 2). Although the numbers of exon/intron in some genes from the same subfamily were variant, for example, *GhMATE1* had one intron more than *GhMATE37*, and *GhMATE6* had two introns more than *GhMATE39*, these genes displayed a similar exon–intron structure. Through Motif analysis by the Pfam and SMART2.0 conserved motifs of the MATE proteins were identified (Table S2). The previous study reported that motif 3, 4, 13 and 14 encoded the MATE domain and played a role in MATE proteins. In the study, all GhMATE proteins contained one or more MATE motifs (Fig. 3). For example, subfamily C4 proteins contained motif 13 and 14, while other subfamilies proteins contained motif 3 and 4. Motif compositions were the similar in the closely related members, which suggested that functional similarities among the MATE proteins were showed in the same subfamily. Additionally, some subfamily-specific motifs were also detected, for example, motif 17 only existed in subfamily C4, while some motifs (motif 6, 7, 9) were almost distributed in all subfamilies. These motifs may be important for the functions of GhMATE proteins despite their unknown functions.

Chromosomal location and gene duplication of *GhMATE* genes

To determine the organization and distribution of *GhMATEs* on different chromosomes in the genome of upland cotton, a physical mapping was constructed. The results showed that 72 *GhMATE* genes were distributed on 19 chromosomes with a non-random distribution (Fig. S1). For example, Dt/chromosome 9 contained 11 *GhMATE* genes, and Dt/chromosome 5 contained 6 ones By contrast, only one *GhMATE* gene was distributed on Dt/chromosome 10, At/chromosome 8 and At/chromosome 11, respectively. In addition, a relatively high density of *GhMATE* genes exhibited on the specific position of some chromosomes, such as the top of Dt/chromosome 5 and the middle of At/chromosome 9.

In this study, gene duplication, which is one of the primary driving forces in the evolution of genomes, was conducted. Among the identified 72 *GhMATE* genes, 5 relatively duplicated gene pairs of 27 sister pairs were identified in *G. hirsutum* (Fig. 4). They were all localized to segmentally duplicated regions and no one was in tandem repeats, indicating that the segmental duplications might play a main role in the expansion of the upland cotton MATE family.

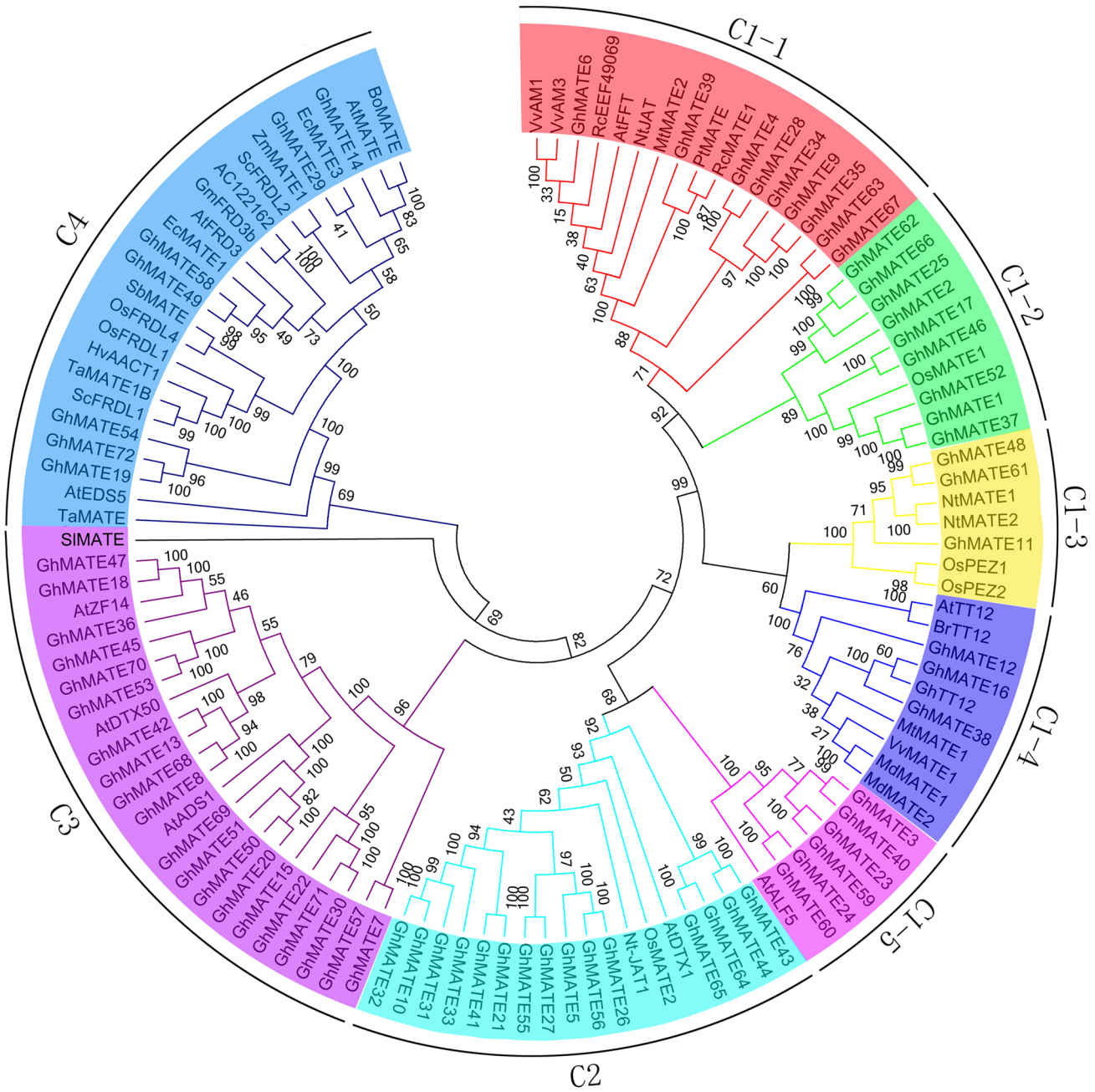
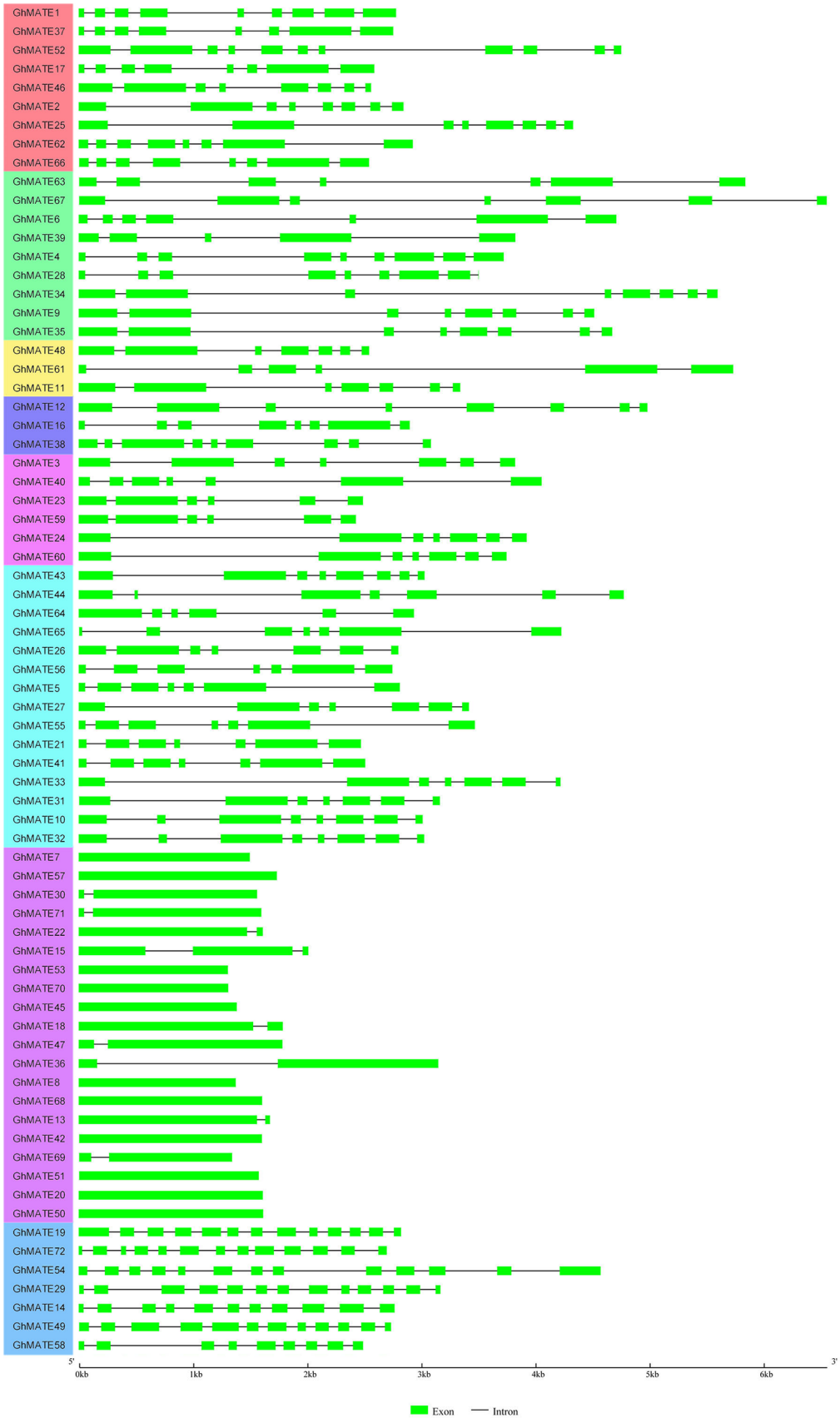


Fig. 1 The phylogenetic tree of cotton MATE family. The phylogenetic tree was constructed by MEGA 5.0 using the maximum likelihood (ML) method. Bootstrap values in percentage (1000 replicates) are indicated on the nodes. Different subfamilies are highlighted using different colors (C1 in pink, C2 in green, C3 in gray, C4 in yellow and C5 in blue), and subgroups are marked with black arcs outside of the cycle tree. (Color figure online)

Identification of orthologous *MATE* genes in different cotton species

To further interpret the evolutionary history of the MATE family, it is necessary to identify orthologous *MATE* genes. Consequently, a comparative analysis was performed to identify orthologous *MATE* genes among *G. hirsutum, G. arboreum* and *G. raimondii*. A total of 42 collinear gene pairs were found within the cotton genomes due to ancient tetraploidy processes, including 28 orthologous gene pairs between *G. hirsutum* and *G. raimondii*, and 14 ones between *G. hirsutum* and *G. arboreum* (Fig. 5), which is perhaps due to the more closer relationship between *G. hirsutum* and *G. raimondii* than between *G. hirsutum* and *G. arboreum*. Meanwhile, 14 *GhMATE* genes were not found in form of the duplicated blocks, which suggested that there were


GhMATE1
GhMATE37
GhMATE52
GhMATE17
GhMATE46
GhMATE2
GhMATE25
GhMATE62
GhMATE66
GhMATE63
GhMATE67
GhMATE6
GhMATE39
GhMATE4
GhMATE28
GhMATE34
GhMATE9
GhMATE35
GhMATE48
GhMATE61
GhMATE11
GhMATE12
GhMATE16
GhMATE38
GhMATE3
GhMATE40
GhMATE23
GhMATE59
GhMATE24
GhMATE60
GhMATE43
GhMATE44
GhMATE64
GhMATE65
GhMATE26
GhMATE56
GhMATE5
GhMATE27
GhMATE55
GhMATE21
GhMATE41
GhMATE33
GhMATE31
GhMATE10
GhMATE32
GhMATE7
GhMATE57
GhMATE30
GhMATE71
GhMATE22
GhMATE15
GhMATE53
GhMATE70
GhMATE45
GhMATE18
GhMATE47
GhMATE36
GhMATE8
GhMATE68
GhMATE13
GhMATE42
GhMATE69
GhMATE51
GhMATE20
GhMATE50
GhMATE19
GhMATE72
GhMATE54
GhMATE29
GhMATE14
GhMATE49
GhMATE58
5'
3'
0kb
1kb
2kb
3kb
4kb
5kb
6kb
Exon
Intron

◂**Fig. 2** The gene structures of 72 *GhMATEs* in cotton. Yellow boxes represent exons (CDS). Black lines represent introns. Blue boxes indicate UTR. The scale on the bottom is in the unit of kilobase (kb). The genes are listed according to the order of subfamily C1–C5 from the phylogenetic tree, and different subfamilies are highlighted in different colors. (Color figure online)

independent duplication events in addition to the whole-genome duplication event during the evolution.

Microarray analysis of *GhMATE* genes

Based on a genome-wide transcript data of gene expression in upland cotton, the expression patterns of *GhMATE* genes were investigated. The results indicated that most *GhMATE* genes exhibited distinct time-specific expression patterns (Fig. 6). Of the 72 *GhMATE* genes, some genes were highly expressed at one or more developmental stages, while other genes were expressed little or no at individual stages of fiber development. Among the highly expressed *GhMATE* genes, the majority was expressed at 10 DPA. Furthermore, the numbers of *GhMATE* genes expressed at different stages were similar. Most duplicate *GhMATE* genes share a similar expression pattern. The expression pattern of *GhMATE* genes suggested that they played different roles in diverse developments of cotton fibers, which was similar with the results of phylogenetic and protein motif analyses.

According to the expression patterns of *GhMATE* genes from subgroup C1-4 at different fiber development stages (0 DPA, 5 DPA, 10 DPA and 20 DPA), the relative expression levels of *GhMATE12, GhMATE16* and *GhMATE38* were similar with *GhTT12*, which had a significant impact on the transport of PAs. The highest expression was at 10 DPA, and gradually decreased afterwards (Fig. 6). At the same time, the content of PAs was found to be also maximal at 10 DPA (Fig. 7), indicating that the accumulation of PAs was positively correlated with the expressions of *GhTT12* and its family members, suggesting that the members of subgroup C1-4 might play an important role in the transport of PAs.

GUS assay and expression patterns of *GhTT12* in *Arabidopsis*

To elucidate the spatial and temporal expression patterns of *GhTT12*, stable *GhTT12* transgenic *Arabidopsis* plants were assayed. Homozygous single-insertion T2 lines subjected to histochemical GUS staining showed that *GUS* gene was expressed in different tissues (roots, stems and leaves) tested (Fig. 8).

The expression of *GhTT12* in the seed coat of transgenic *Arabidopsis* was 4.8 times than that of the mutant *tt12* (Fig. 9a). The total PA content in the seed coat of transgenic *Arabidopsis* was 5.7 times than that of the mutant, while that of the wild type was 13.1 times than that of the mutant (Fig. 9b). These results suggest that *GhTT12* plays an essential role in the transport and accumulation of proanthocyanidins.

Discussion

In the previous studies, the genome-wide analyses and functions of some MATE family members have been conducted in plants, but most of *MATE* genes play mainly roles in responses to abiotic stress [6, 7, 11, 44, 45]. only a few literatures reported that *MATE* genes are involved in the transpor of secondary metabolites [8–10]. Specially, a genome-wide study of the MATE family has not been reported in upland cotton (*G. hirsutum*), and it is not kwon whether *MATE* genes are involved in the transport of PAs during plant growth and development in brown cotton. Until 2015, The whole genome sequence of upland cotton (TM-1) has been completed, which was beneficial to predict the evolutionary history of the MATE gene family in upland cotton.

In the present study, a systematic and comprehensive whole-genome analyses of the *MATE* family was carried out in upland cotton (*G. hirsutum*) including the phylogeny, gene structure, conserved motif, chromosomal location, gene duplication, expression profile and subcellular localization was performed. The numbers of *MATE* genes in upland cotton (72 *MATE* genes) were higher than those in *Arabidops*is (55 *MATE* genes) [44] and rice (56 *MATE* genes) [45] while less than those in soybean (117 *MATE* genes) [7], which indicates that the *GhMATE* genes in upland cotton have been expanded comparing with those in *Arabidopsis* and rice. Polyploidization, which makes a whole-genome duplication, and rapid genomic reorganization and massive gene loss, are the features of a diploidization process [46–48]. Gene duplication is one of the primary driving forces in the evolution of genomes, and it often contains segmental and tandem duplication [49, 50]. Interestingly, the numbers of *GhMATE* genes involved in segmental duplication are much more than those involved in tandem duplication, suggesting that segmental duplications are the main contributors to the expansion of the MATE family in upland cotton. The number of orthologous genes between *G. hirsutum* and *G. raimondii* is more than that between *G. hirsutum* and *G. arboreum*, which provides a new resource for understanding the evolution of the MATE gene families among different species.

Phylogenetic analyses revealed that 72 GhMATE proteins were divided into four clades designated C1–C4. Clade C1 was further divided into five subgroups. In subgroup C1-1, AtFFT is a flavonoid transporter that affects flavonoid levels in *Arabidopsis* [51], and VvAM1, VvAM3 are involved in the transport of acylated anthocyanins

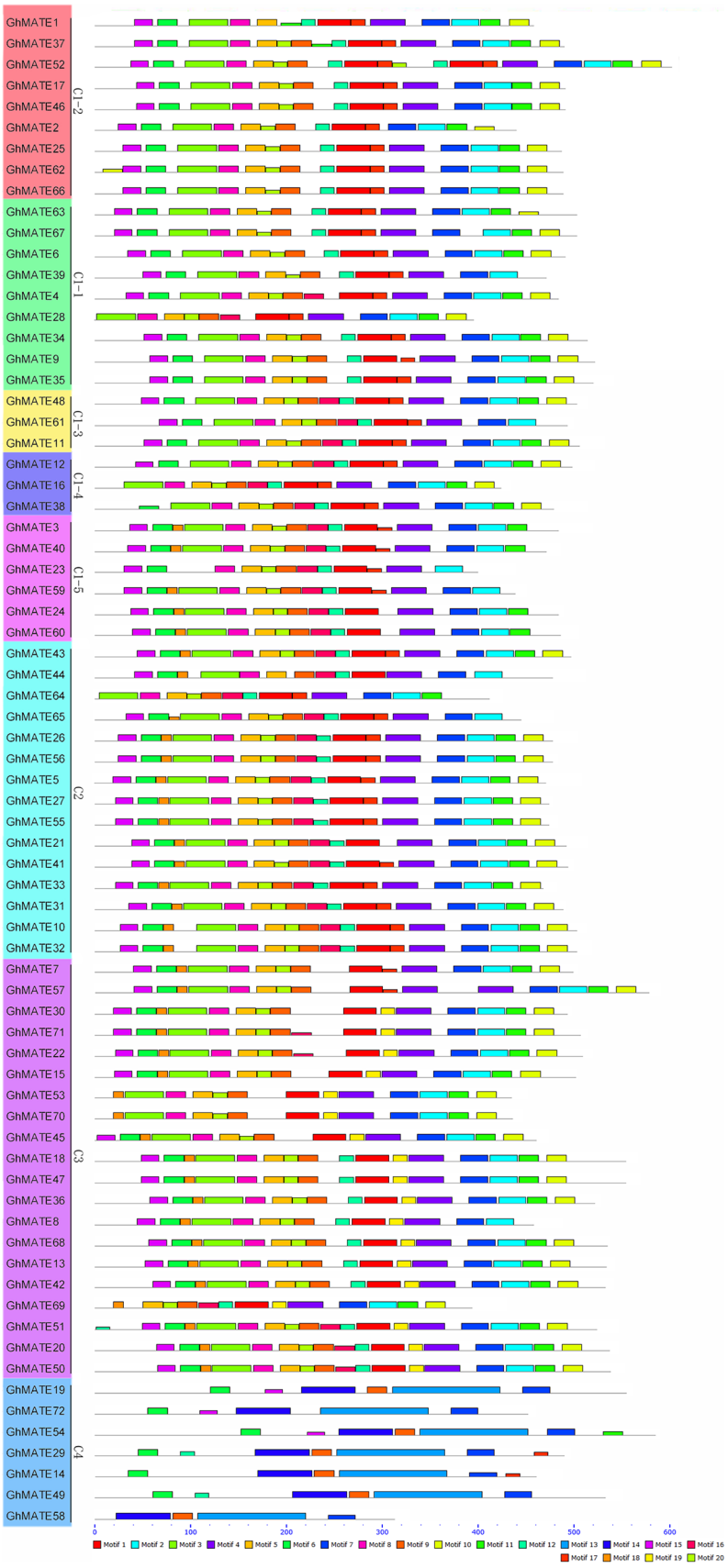


Fig. 3 Protein motifs of the MATE family in cotton. The motifs of MATE proteins in cotton are shown as colored boxes. The length of motifs are different. The GhMATE proteins are listed according to the order of subfamily C1–C5 from the phylogenetic tree, and different subfamilies are highlighted in different colors. (Color figure online)

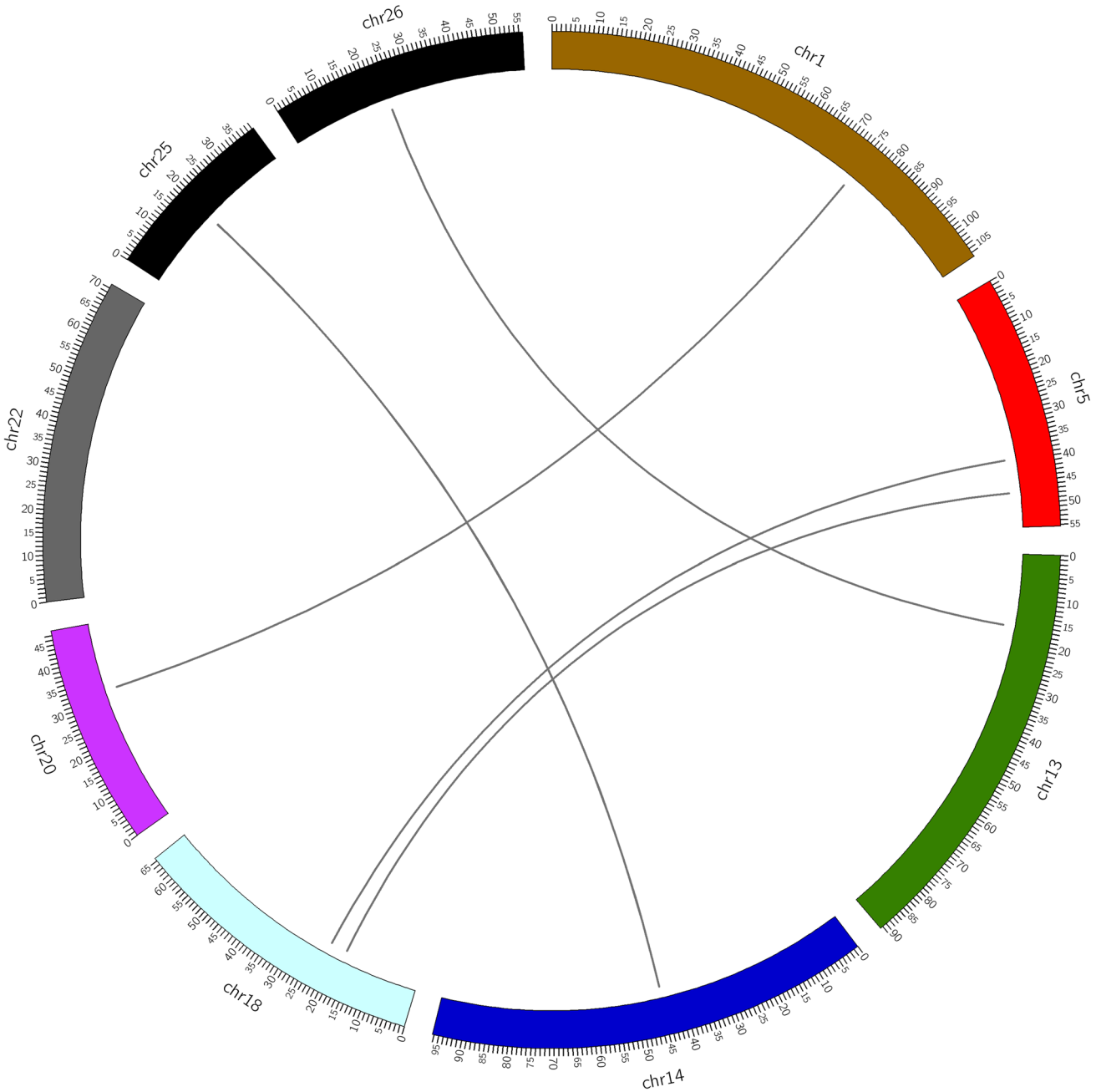


Fig. 4 Distribution and duplication of the MATE genes in the upland cotton. The distributions of *GhMATEs* located on the 16 chromosomes in the upland cotton genome. Numbers along each chromosome box represent sequence lengths in megabases (Mb). Duplicated genes are linked by lines

into vacuoles in grapevine [52, 53]. In subgroup C1-2, only one known MATE protein, OsMATE1, is involved in the resistance to aluminum stress [45]. NtMATE1 and NtMATE2 in subgroup C1-3 could transport alkaloids from the cytosol into the vacuole in tobacco [54]. In Subgroup C1-4, AtTT12 isolated originally during screening of mutants with altered seed coloration is thought to be the first MATE transporter involved in transport of flavonoids. MtMATE1 localized in the tonoplast is a functional ortholog of AtTT12 [23], and VvMATE1 plays a role in the accumulation of PAs [55]. AtALF5 in subgroup C1-5 leads to abmity in the formation of lateral root and enhances their sensitivity to various compound resistant to toxins. Based on the functions of the known MATE transporters in subfamily C1, the MATE proteins in this clade might transport and accumulate flavonoids or

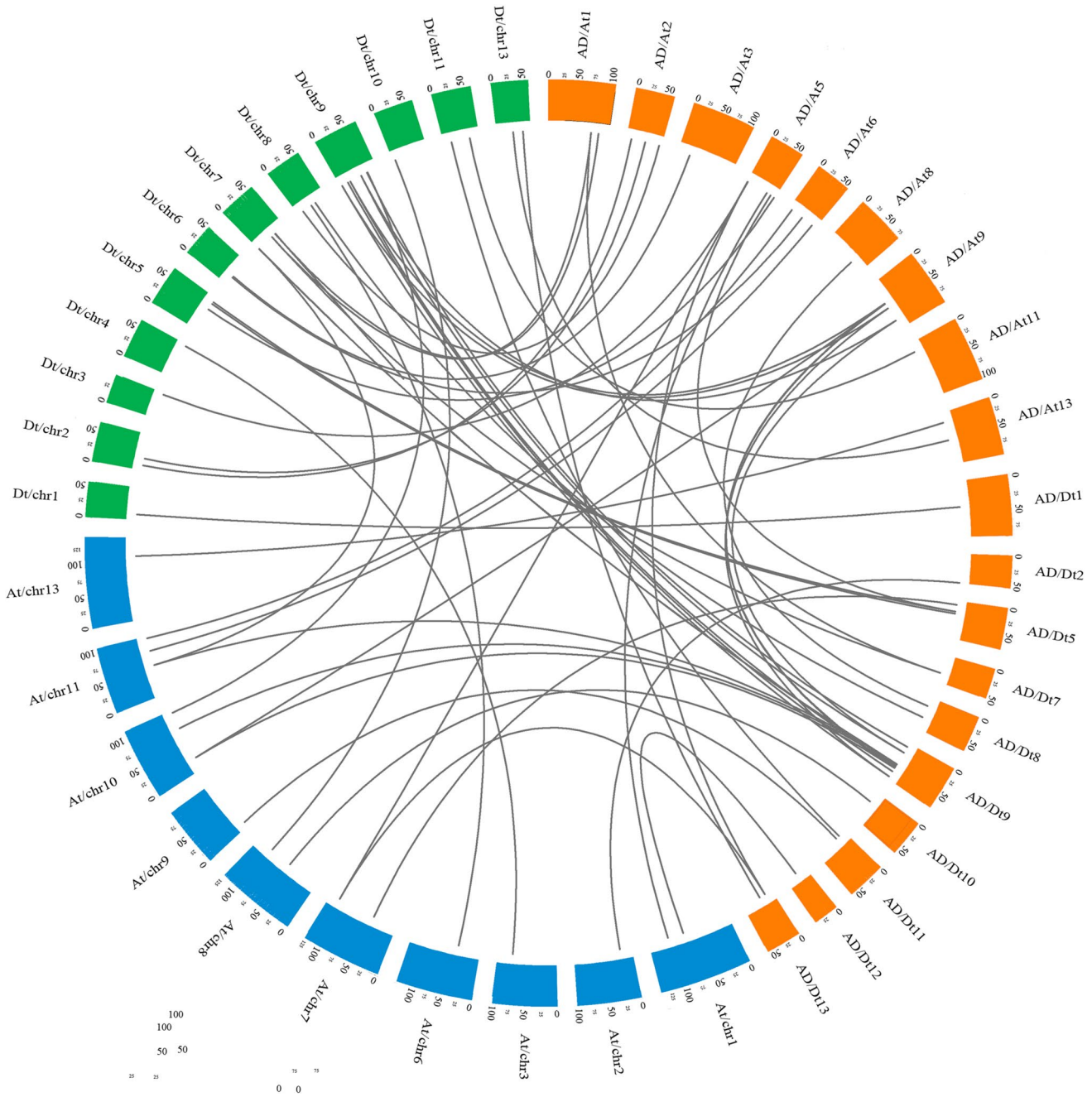


Fig. 5 Microsynteny of the *MATE* genes in different cotton species. The *G. hirsutum, G. arboreum* and *G. raimondii* chromosomes shown in different colors, are labeled with the red, yellow and blue, respectively. Numbers along each chromosome box indicate sequence lengths in Mb. Black lines represent the syntenic relationships between MATE genes. (Color figure online)

alkaloids in plants. In clade C2, AtDTX1 is found to mediate the toxic compounds like plant-derived antibiotics and to detoxify the heavy metal, such as Cd^{2+} [44]. NtJAT1 is a secondary transporter, which shows nicotine efflux activity in yeast and is involved in nicotine translocation [56]. Therefore, subfamily C2 might participate in the efflux of various compounds. In clade C3, SlMATE may function as a vacuolar transporter of anthocyanin in tomato leaves [57]. AtADS1 is a negative regulator with associated to plant disease resistance [58]. AtZF14 is reported to be involved in iron homeostasis. Additionally, AtZF14 and AtADS1 also play a role in the regulation of hypocotyl cell elongation [59]. 7 GhMATE proteins in subgroup C4 are indicated to be involved in Al detoxification/iron translocation in cotton by reference to 17 reported MATE proteins from other plant species.

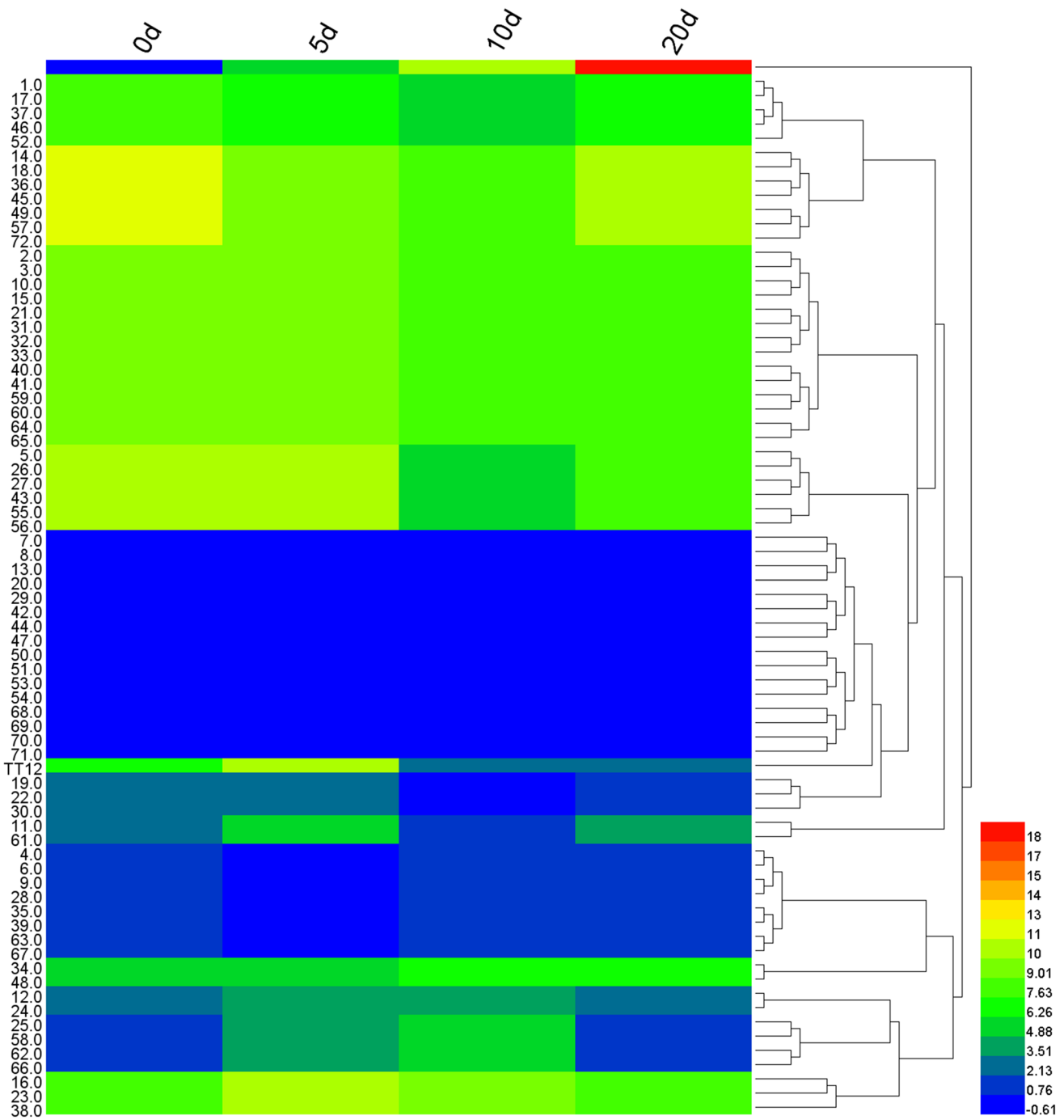


Fig. 6 Expression profiles of the MATE genes at different fiber development stages of upland cotton. A small color box represents a set of data. Different colors indicate the genes expression level and each color reflects the absolute value of a gene expression showed on the right. The left number represents the relationship between different genes, which are divided into 22 clusters. The samples of cotton fibers are as follows: 0 day (0 DPA), 5 days (5 DPA), 10 days (10 DPA), and 20 days (20 DPA). (Color figure online)

The structural diversity of genes is considered to impel the evolution of gene families. Moreover, the differences in the characteristics of different subfamilies suggested that MATE members are functionally diversified in cotton. Introns are specifically inserted into the plant genome and retained during the course of evolution [60]. Therefore, it is speculated that the MATE coding region should gain or lose introns in a subfamily-specific manner. Furthermore, Motif analysis revealed that there were different conserved motifs presented in each cotton MATE protein.

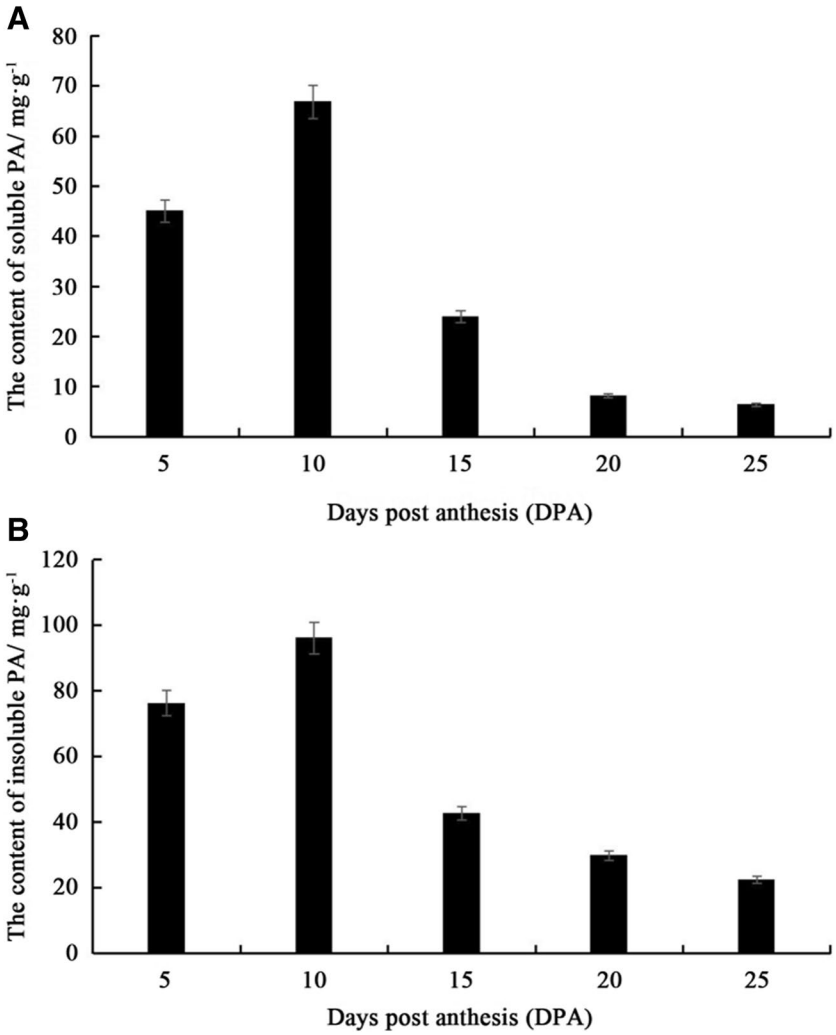


Fig. 7 The content of PAs at different fiber development stages in upland cotton. **a** The content of soluble PAs, **b** the content of insoluble PAs

However, functions of these motifs the MATE family are still unclear.

Gene expression pattern can be used to elucidate the function of a gene. Genome-wide transcript profiling data of cotton ovules was used to examine the expression of *GhMATE*s. Almost 1/3 of 72 *GhMATE* genes exhibited little expression at some specific stages, which indicates that *GhMATEs* are expressed at specific developmental stages. Based on a phylogenetic tree, *GhMATE12, GhMATE16* and *GhMATE38* from subgroup C1-4 were selected for expression analysis by qRT-PCR in upland cotton. The results showed that these *MATE* genes had a similar expression pattern. At the same time, the contents of PAs at different fiber development stages were closely related to the expression levels of *MATE* genes from subgroup C1-4 in upland cotton. In addition, *GhTT12*, which was reported to be involved in the transport of PA, was identified to localize in the vacuole membrane [12]. The results suggest that *GhMATE12, GhMATE16* and *GhMATE38* may be involved in the vacuole transportation of PA like *GhTT12*. In summary, the present results contribute to improve understanding the molecular evolution of the MATE gene family in plants and the functions of *GhMATEs* in upland cotton.

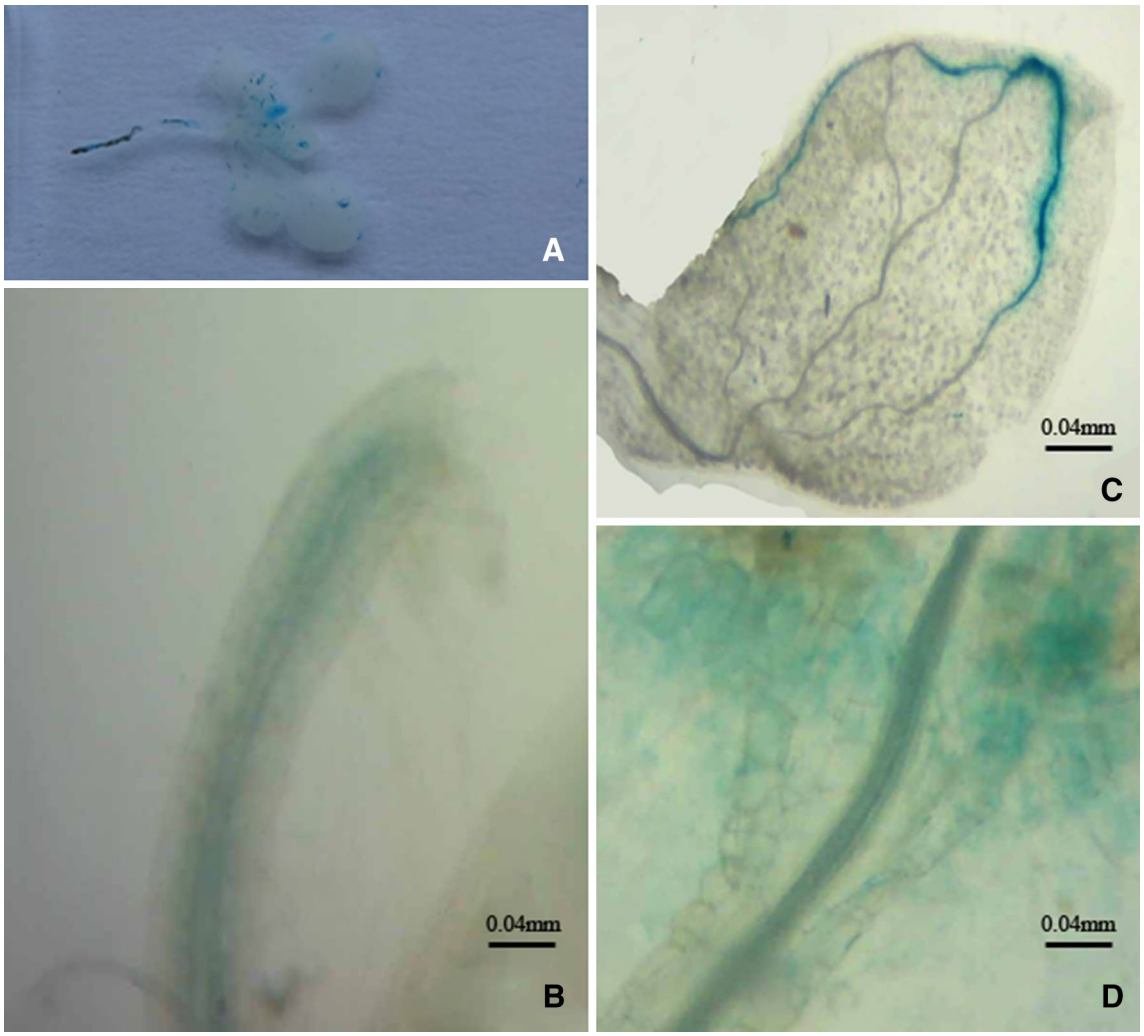


Fig. 8 The GUS staining of *Arabidopsis thaliana*. **a** The plant of *Arabidopsis*; **b** the root of *Arabidopsis*; **c** the leaf of *Arabidopsis*; **d** the stem of *Arabidopsis*

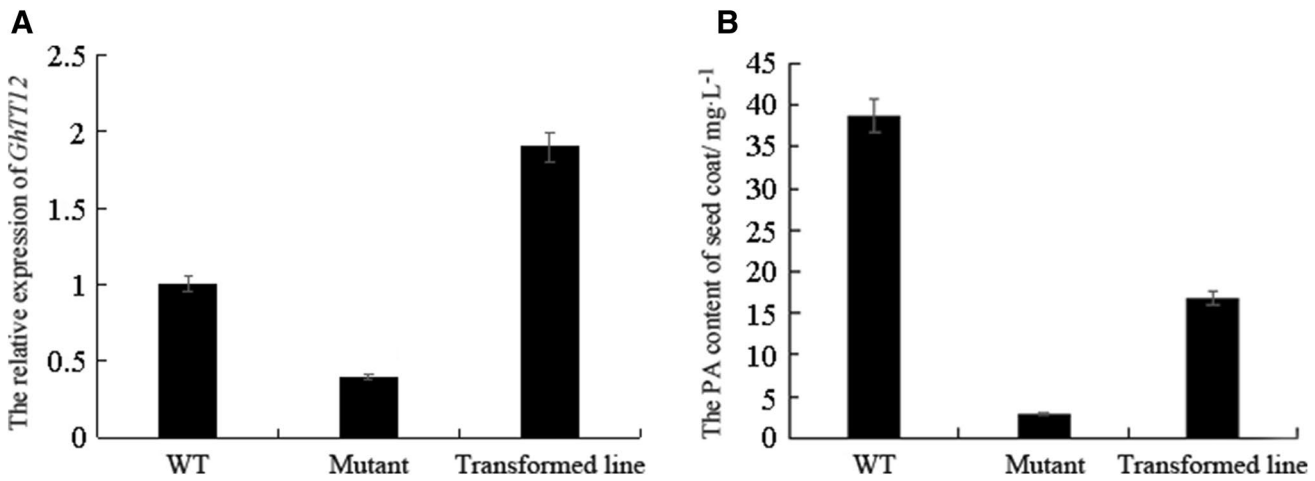


Fig. 9 Spatial and temporal expression patterns of *GhTT12* in Arabidopsis. **a** Relative expression of *GhTT12* in seed coat, **b** the PA content of seed coat

Acknowledgements We thank Dr. Da-Hui Li for assisting the experimental technology and revising the original manuscript, and are thankful for Prof. Zhao-Rong Deng to offer the biochemical reagents. This work was sponsored by State Key Laboratory of Cotton Biology Open Fund (No. CB2018A01 to J.-S. G.) and the National Natural Science Foundation of China (Grant Nos. 31672497 to J.-S. G.; 31572468 to Y. M.). It was also supported by the National Key Research and Development Program of China (No. 2016YFD0300205-3 to Z.-R. D.).

Author Contributions J-SG, YL and YM conceived the idea, LX, WC, G-YS and Z-LS analysed the data, J-SG, LX and Z-LS wrote the manuscript, NG, Y-PC and XS assisted in sequence analysis. All authors reviewed the manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

References

1. Saier MH Jr, Paulsen IT (2001) Phylogeny of multidrug transporters. Semin Cell Dev Biol 12:205–213
2. Omote H, Hiasa M, Matsumoto T, Otsuka M, Moriyama Y (2006) The MATE proteins as fundamental transporters of metabolic and xenobiotic organic cations. Trends Pharmacol Sci 27:587–593
3. Magalhaes JV (2010) How a microbial drug transporter became essential for crop cultivation on acid soils: aluminium tolerance conferred by the multidrug and toxic compound extrusion (MATE) family. Ann Bot 106:199–203
4. He X, Szewczyk P, Karyakin A, Evin M, Hong W-X, Zhang Q, Chang G (2010) Structure of a cation-bound multidrug and toxic compound extrusion transporter. Nature 467:991–994
5. Wang L, Bei X, Gao J, Li Y, Yan Y, Hu Y (2016) The similar and different evolutionary trends of MATE family occurred between rice and *Arabidopsis thaliana*. BMC Plant Biol 16:207
6. Zhu H, Wu J, Jiang Y, Jin J, Zhou W, Wang Y, Han G, Zhao Y, Cheng B (2016) Genomewide analysis of *MATE*-type gene family in maize reveals microsynteny and their expression patterns under aluminum treatment. J Genet 95:691–704
7. Liu J, Li Y, Wang W, Gai J, Li Y (2016) Genome-wide analysis of MATE transporters and expression patterns of a subgroup of *MATE* genes in response to aluminum toxicity in soybean. BMC Genomics 17:223
8. Chen L, Liu Y, Liu H, Kang L, Geng J, Gai Y, Ding Y, Sun H, Li Y (2015) Identification and expression analysis of *MATE* genes involved in flavonoid transport in blueberry plants. PLoS ONE 10:e0118578
9. Yazaki K, Sugiyama A, Morita M, Shitan N (2008) Secondary transport as an efficient membrane transport mechanism for plant secondary metabolites. Phytochem Rev 7:513–524
10. Santos ALD, Chaves-Silva S, Yang L, Maia LGS, Chalfun-Júnior A, Sinharoy S, Zhao J, Benedito VA (2017) Global analysis of the *MATE* gene family of metabolite transporters in tomato. BMC Plant Biol 17:185
11. Lu P, Magwanga RO, Guo X, Kirungu JN, Lu H, Cai X, Zhou Z, Wei Y, Wang X, Zhang Z, Peng R, Wang K, Liu F (2018) Genome-wide analysis of multidrug and toxic compound extrusion (MATE) family in *Gossypium raimondii* and *Gossypium arboreum* and its expression analysis under salt, cadmium, and drought stress. G3 8:2483–2500
12. Gao J-S, Wu N, Shen Z-L, Lv K, Qian S-H, Guo N, Sun X, Cai Y-P, Lin Y (2016) Molecular cloning, expression analysis and subcellular localization of a Transparent Testa 12 ortholog in brown cotton (*Gossypium hirsutum* L.). Gene 576:763–769
13. Zhao X, Wang X (2005) Composition analysis of pigment in colored cotton fiber. Acta Agron Sin 31:456–462
14. Zhan S, Lin Y, Cai Y, Li Z (2007) Preliminary deductions of the chemical structure of the pigment brown in cotton fiber. Chin Bull Bot 24:99–104
15. Dixon RA, Liu C, Jun JH (2013) Metabolic engineering of anthocyanins and condensed tannins in plants. Curr Opin Biotechnol 24:329–335
16. Dixon RA, Xie DY, Sharma SB (2005) Proanthocyanidins—a final frontier in flavonoid research? New Phytol 165:9–28
17. Lepiniec L, Debeaujon I, Routaboul J-M, Baudry A, Pourcel L, Nesi N, Caboche M (2006) Genetics and biochemistry of seed flavonoids. Annu Rev Plant Biol 57:405–430
18. Ariga T, Asao Y, Sugimoto H, Yokotsuka T (1981) Occurrence of astringent oligomeric proanthocyanidins in legume seeds. Agric Biol Chem 45:2705–2708
19. Gabetta B, Fuzzati N, Griffini A, Lolla E, Pace R, Ruffilli T, Peterlongo F (2000) Characterization of proanthocyanidins from grape seeds. Fitoterapia 71:162–175
20. Gu L, Kelm MA, Hammerstone JF et al (2004) Concentrations of proanthocyanidins in common foods and estimations of normal consumption. J Nutr 134:613–617
21. Santos-Buelga C, Scalbert A (2000) Proanthocyanidins and tannin-like compounds–nature, occurrence, dietary intake and effects on nutrition and health. J Sci Food Agric 80:1094–1117
22. Serafini M, Bugianesi R, Maiani G, Valtuena S, De Santis S, Crozier A (2003) Plasma antioxidants from chocolate. Nature 424:1013
23. Zhao J, Dixon RA (2009) MATE transporters facilitate vacuolar uptake of epicatechin 3′-*O*-glucoside for proanthocyanidin biosynthesis in *Medicago truncatula* and Arabidopsis. Plant Cell 21:2323–2340
24. Zhao J, Huhman D, Shadle G, He X-Z, Sumner LW, Tang Y, Dixon RA (2011) MATE2 mediates vacuolar sequestration of flavonoid glycosides and glycoside malonates in *Medicago truncatula*. Plant Cell 23:1536–1555
25. Paterson AH, Wendel JF, Gundlach H et al (2012) Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature 492:423–427
26. Wang K, Wang Z, Li F et al (2012) The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet 44:1098–1103
27. Li F, Fan G, Wang K et al (2014) Genome sequence of the cultivated cotton *Gossypium arboreum*. Nat Genet 46:567–572
28. Eddy SR (1996) Hidden Markov models. Curr Opin Struct Biol 6:361–365
29. Finn RD, Mistry J, Schuster-Böckler B et al (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34:247–251
30. Letunic I, Doerks T, Bork P (2012) SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Res 40:302–305
31. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680
32. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31:3784–3788
33. Zhang C, Zhang H, Zhao Y, Jiang H, Zhu S, Cheng B, Xiang Y (2013) Genome-wide analysis of the *CCCH* zinc finger gene family in *Medicago truncatula*. Plant Cell Rep 32:1543–1555
34. Hu B, Jin J, Guo A-Y, Zhang H, Luo J, Gao G (2014) GSDS 2.0: an upgraded gene feature visualization server. Bioinformatics 31:1296–1297

35. Bailey TL, Elkan C (1995) The value of prior knowledge in discovering motifs with MEME. Proc Int Conf Intell Syst Mol Biol 3:21–29
36. He H, Dong Q, Shao Y, Jiang H, Zhu S, Cheng B, Xiang Y (2012) Genome-wide survey and characterization of the *WRKY* gene family in *Populus trichocarpa*. Plant Cell Pep 31:1199–1217
37. Wang Y, Tang H, Debarry JD et al (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 40:e49
38. Wang F, Dong Q, Jiang H, Zhu S, Chen B, Xiang Y (2012) Genome-wide analysis of the heat shock transcription factors in *Populus trichocarpa* and *Medicago truncatula*. Mol Biol Rep 39:1877–1886
39. Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189
40. Krzywinski M, Schein J, Birol I et al (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645
41. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. Methods 25:402–408
42. Ikegami A, Akagi T, Potter D, Yamada M, Sato A, Yonemori K, Kitajima A, Inoue K (2009) Molecular identification of 1-Cys peroxiredoxin and anthocyanidin/flavonol 3-*O*-galactosyltransferase from proanthocyanidin-rich young fruits of persimmon (*Diospyros kaki* Thunb.). Planta 230:841–855
43. Li Y, Tanner G, Larkin PJ (1996) The DMACA-HCl protocol and the threshold proanthocyanidin content for bloat safety in forage legumes. J Sci Food Agric 70:89–101
44. Li L, He Z, Pandey GK, Tsuchiya T, Luan S (2002) Functional cloning and characterization of a plant efflux carrier for multidrug and heavy metal detoxification. J Biol Chem 277:5360–5368
45. Tiwari M, Sharma D, Singh M, Tripathi RD, Trivedi PK (2014) Expression of OsMATE1 and OsMATE2 alters development, stress responses and pathogen susceptibility in *Arabidopsis*. Sci Rep 4:3964
46. Otto SP (2007) The evolutionary consequences of polyploidy. Cell 131:452–462
47. Soltis PS, Soltis DE (2009) The role of hybridization in plant speciation. Annu Rev Plant Biol 60:561–588
48. Brenchley R et al (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491:705–710
49. Moore RC, Purugganan MD (2003) The early stages of duplicate gene evolution. Proc Natl Acad Sci USA 100:15682–15687
50. Cannon SB, Mitra A, Baumgarten A, Young ND, May G (2004) The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. BMC Plant Biol 4:10
51. Thompson EP, Wilkins C, Demidchik V, Davies JM, Glover BJ (2009) An *Arabidopsis* flavonoid transporter is required for anther dehiscence and pollen development. J Exp Bot 61:439–451
52. Gomez C, Terrier N, Torregrosa L et al (2009) Grapevine MATE-type proteins act as vacuolar H^+-dependent acylated anthocyanin transporters. Plant Physiol 150:402–415
53. Gomez C, Conejero G, Torregrosa L, Cheynier V, Terrier N, Ageorges A (2011) In vivo grapevine anthocyanin transport involves vesicle-mediated trafficking and the contribution of anthoMATE transporters and GST. Plant J 67:960–970
54. Shoji T, Inai K, Yazaki Y et al (2009) Multidrug and toxic compound extrusion-type transporters implicated in vacuolar sequestration of nicotine in tobacco roots. Plant Physiol 149:708–718
55. Pérez-Díaz R, Ryngajllo M, Pérez-Díaz J, Peña-Cortés H, Casaretto JA, González-Villanueva E, Ruiz-Lara S (2014) VvMATE1 and VvMATE2 encode putative proanthocyanidin transporters expressed during berry development in *Vitis vinifera* L. Plant Cell Rep 33:1147–1159
56. Morita M, Shitan N, Sawada K et al (2009) Vacuolar transport of nicotine is mediated by a multidrug and toxic compound extrusion (MATE) transporter in Nicotiana tabacum. Proc Natl Acad Sci USA 106:2447–2452
57. Mathews H, Clendennen SK, Caldwell CG et al (2003) Activation tagging in tomato identifies a transcriptional regulator of anthocyanin biosynthesis, modification, and transport. Plant Cell 15:1689–1703
58. Sun X, Gilroy EM, Chini A et al (2011) ADS1 encodes a MATE-transporter that negatively regulates plant disease resistance. New Phytol 192:471–482
59. Wang R, Liu X, Liang S et al (2015) A subgroup of *MATE* transporter genes regulates hypocotyl cell elongation in *Arabidopsis*. J Exp Bot 66:6327–6343
60. Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr Biol 13:1512–1517