Pak. J. Bot., 48(3): 1081-1090, 2016.

GENOME-WIDE IDENTIFICATION AND EXPRESSION ANALYSIS OF THE MALATE DEHYDROGENASE GENE FAMILY IN *GOSSYPIUM ARBOREUM*

MUHAMMAD IMRAN AND JIN-YUAN LIU*

Laboratory of Plant Molecular Biology, Center for Plant Biology, School of Life Sciences, Tsinghua University, Beijing 100084, China
**Corresponding author's e-mail: liujy@mail.tsinghua.edu.cn*

Abstract

Malate dehydrogenase (MDH) is a key enzyme that catalyzes the reversible oxidation of malate to oxaloacetate and plays a crucial role in various cellular processes, such as cell expansion, wall thickening and cell elongation. Although individual genes belonging to MDH gene family have been partially identified in various plants, there have been no reports of a genome-wide characterization of the MDH gene family in cotton. Here, we identified a total of 13 MDH genes from the genome of a diploid cotton *Gossypium arboreum* and designated *GaMDH1-13* based on their chromosomal locations. These MDH members were unevenly distributed on 8 of the13 chromosomes. Segmental duplications that played a dominant role in the expansion of the MDH gene family were estimated to have occurred between 19.07 to 20.47 million years ago (MYA), when a recent large-scale genome duplication occurred in cotton. Phylogenetic analyses showed that the putative MDH proteins formed five groups (I to V) in plant species. *GaMDH* genes within the same group shared similar gene structures and domain constitutions. Furthermore, expression analysis showed that the *GaMDH* genes were differentially expressed in root, stem, leaf, hypocotyl, petal and anther, with higher expression levels detected during different fiber developmental stages. Notably, *GaMDH13* had the highest expression level during the fast fiber elongation stage that ranged from 5 to 15 day post-anthesis (DPA), suggesting that the MDH gene plays a vital role in fiber development. The results of this study will aid functional analyses of the MDH genes in cotton fiber development.

Key words: Cotton, MDH gene family, Gene expression analysis, Genome-wide analysis, Fiber development.

Introduction

The cotton (*Gossypium*) genus constitutes approximately 46 diploid and 5 tetraploid species. Among the four fiber-producing species of the cotton family, the diploid *Gossypium arboreum* is one of the major cultivated cotton species in the world, and commonly known as Asian cotton (Li *et al*., 2014). The *G. arboreum* along with fibreless diploid *Gossypium raimondii,* is believed to be the original tetraploid progenitor species (Hovav *et al*., 2008). However, the diploid Asian cotton, in contrast with the two allotetraploid species, has an impeded commercial value mainly because of a shorter mature fiber and exhibits high similarity with the allotetraploids in the developmental processes of cotton fiber cell, implying that the diploid *G. arboreum* as a simplified and useful model species for investigating the molecular mechanisms associated with the development of cotton fiber.

To date, substantial advancements have been made in the identification of genes and proteins related to fiber development and specifically those involved in fiber elongation (Gou *et al*., 2007; Zhao *et al*., 2010; Pei, 2015) and development (Taliercio *et al*., 2010; Zhou *et al*., 2011). In previous osmoregulation studies, malate dehydrogenase (MDH) expression showed significant dynamics during fiber development and found to reach a peak at 15 day post-anthesis (DPA) before declining (Dhindsa *et al*., 1975). Similarly, the MDH protein expression profile from developing fibers also increased from 14 to 21 DPA (Ferguson *et al*., 1996).

Malate dehydrogenase (EC1.1.1.37) belongs to the A group of dehydrogenases, which constitute a gene family of $NAD(P)^+$-dependent conserved enzymes that are ubiquitously found in plants, animals, fungi and bacteria (Minarik *et al*., 2002). Malate is a central metabolite that is essential for cellular metabolism and an important intermediate of the tricarboxylic acid cycle (Fernie & Martinoia, 2009). In higher plants, MDHs have been classified into 5 groups based upon their coenzyme specificity, physiological functions and subcellular locations (Musrati *et al*., 1998) and consist of two discreet domains that are visually interlaced but have distinct functions (Hall *et al*., 1992). Several MDH genes have been identified in a number of plants, including *Arabidopsis* (Tomaz *et al*., 2010), maize (Longo & Scandalios, 1969), apples (Yao *et al*., 2011) and cotton (Wang *et al*., 2015). Functional studies revealed that MDHs were involved in the growth and development of plant cells and played a crucial role in various plant stress responses, such as leaf respiration (Tomaz *et al*., 2010), embryo development (Beeler *et al*., 2014) and tolerance to cold and salt stress (Yao *et al*., 2011). The recent identification of different malate protein channels in several plant tissues and the analysis of transgenics with varied malate metabolisms have shed new light on its broader importance for cellular functions (Faske *et al*., 1997). However, the functional genomic characterization of this gene family has been limited in cotton.

The recent availability of genome sequences for *G. arboreum* (http://cgp.genomics.org.cn) provides an opportunity to investigate the MDH gene family in the cotton genome (Li *et al*., 2014). In this study, we identified 13 MDH genes in the *G. arboreum* genome and the segmental duplications which may have contributed to the evolution of *G. arboreum* MDHs as well. Our detailed analysis primarily focused on gene recognition, exon-intron organization, domain structure, and expression profile of the publically available cotton MDH gene family members. The final results from this study will provide a cornerstone for the evolutionary and functional characterization of the MDH genes in *G. arboreum* and other plant species.

Materials and Methods

Cotton materials and growth conditions: The Asiatic diploid cultivated cotton seed of *G. arboreum cv shixiya 1* was obtained from the Cotton Research Institute, Chinese Academy of Agricultural Sciences (CAAS). The seeds were grown in the experimental field of Tsinghua University (Beijing, China) under normal agronomic standard conditions during the year 2013-2014. Cotton flowers were tagged on the day of anthesis. Cotton fibers were harvested 0, 5, 10, 15, 20, 25 and 30 DPA, and different cotton tissues were collected, including the root, stem, leaf, hypocotyl, petal, and anther. All of the cotton samples were immediately frozen in liquid nitrogen and stored at -80°C for nucleic acid extraction.

Database searches and sequence alignment: To identify the *G. arboreum* MDH genes, nine protein sequence of the *Arabidopsis* MDHs were retrieved from their genome databases (https://www.arabidopsis.org) using ontologies/keywords search interface with "Malate dehydrogenase" as keyword. Next all the identified *Arabidopsis* MDH genes were subsequently employed as query to perform the blastp and tblastn algorithms against the *G. arboreum* genome database of the Chinese Academy of Agriculture Sciences (http://cgp.genomics.org.cn). In addition, we have also obtained the same sequences (from HMMER search (http://hmmer.janelia.org/) using Hidden Markov Model (HMM) analysis with Pfam number PF00056 (NAD-binding domain), PF02866 (Catalytic domain) and PS00068 (MDH-active site), from Pfam protein family database (http://pfam.sanger.ac.uk/). Sequences with an E-values less than <1.0 were selected, and redundant sequences were removed from further analysis based on Clustal-W alignment (Thompson *et al*., 1994). To verify the reliability of the initial results, all of the putative proteins were further confirmed to be MDH proteins by using the InterproScan program (Quevillon *et al*., 2005). The lengths, theoretical molecular weights and isoelectric points of the deduced proteins were calculated by ExPASY (http://www.cn.expasy.org/tools). Subcellular localization of proteins were predicted using the TargetP 1.1 server (www.cbs.dtu.dk/services/TargetP) (Emanuelsson *et al*., 2000).

Finally, 13 identified MDH genes were mapped on the basis of their chromosomal localization and presented via a Circos diagram. Gene duplication events were investigated using the following criteria: 1) genes with >70% coverage of the alignment length; 2) genes with >70% identity in the aligned region; and 3) a minimum of two duplication events were considered for strongly connected genes (Gu *et al*., 2002). The time of duplication and deviation of the *GaMDH* gene pairs were calculated using the synonymous mutation rate of λ substitutions per synonymous site per year: $T=Ks/2\lambda$, where $\lambda=1.5 \times 10^{-8}$ for cotton (Blanc & Wolfe, 2004). Protein sequences and the corresponding ORFs of the gene pairs were aligned, and the *Ka* (nonsynonymous substitution rates) and *Ks* (synonymous substitution rates) of the duplicated *G. arboreum* genes were calculated by the program *KaKs* Calculator (Zhang *et al*., 2006). The average *Ks* values were estimated for each duplicated gene pair and used to date the duplication events ($T=Ks/2\lambda$).

Gene structure prediction and phylogenetic analysis: Exon-intron structures of the *G. arboreum* MDH genes were generated by the alignment of their coding sequences to the representative genomic sequence information obtained from the aforementioned genome databases using the online tool Gene Structure Display Server (http://gsds.cbi.pku.edu.cn) (Guo *et al*., 2007). The protein sequences of *G. arboreum* were aligned with *cacao* (*Theobroma cacao*), and *Arabidopsis* genomes using Clustal-W with the default settings, and a rooted phylogenetic tree based on the 30 protein sequences was constructed with the MEGA 6.0 software using the neighbour-joining (NJ) method with *p*-distance and pairwise gap deletion parameters engaged (Tamura *et al*., 2013). The bootstrap test was repeated 1000 times. Furthermore, maximum likelihood and minimal evolution methods were also applied to validate the results from the NJ tree. The MDH protein sequences in *cacao* were from *T. cacao* genome sequence databases (http://www.phytozome.net/cacao).

RNA isolation and real-time quantitative PCR detection: Total RNA was extracted from frozen cotton tissues using the RNAprep Pure Plant kit (TIANGEN, Beijing, China) according to the manufacturer's protocol. A total of 2 µg of RNA was used as the template for the first-strand cDNA synthesis using an RNA PCR kit (AMV, version 3.0, TaKaRa, Dalian, China). The resulting cDNA products were diluted 1/5 and stored at -20°C for qRT-PCR analysis. Using the specific primers for each *GaMDH* gene (Table 1), quantitative RT-PCR was performed with a Mini Opticon Real-Time PCR System (Bio-Rad, CA, USA) according to the supplier's protocol. Each reaction mixture contained 8 µl of DNase/RNase-free water, 10 µl of the Real-Time SYBR Green PCR master mix, 1 µl of the diluted cDNA product and 1 µl of the gene-specific primers. A cotton ubiquitin gene (UBQ7, DQ116441) was used as a standard control. Three biological replicates were conducted for each tissue, and each biological replicate was technically repeated three times. The thermal cycling conditions were as follows: pre-denaturation at 95°C for 5 min and 40 cycles of amplification at 95°C for 5 s, 58°C for 30 s and 70°C for 30 s. The relative expression levels were calculated using the comparative $2^{-\Delta\Delta CT}$ method (Livak & Schmittgen, 2001). A heatmap for the gene expression profiles was generated forth with the Multiexperiment Viewer (MeV) online tool (http://www.tm4.org/).

Results and Discussion

MDH gene family in the *Gossypium arboreum* genome: The recent availability of the *G. arboreum* genome (http://cgp.genomics.org.cn) sequence enabled the identification of all the MDH genes in this species. Therefore, the HHM profile of MDH domain (PF00056, PF02866 and PS00068) and corresponding MDH gene sequences from *Arabidopsis* were used as queries to perform multiple searches in the *G. arboreum* genome database using the blastp and tblastn algorithms and HMMER search (Altschul *et al*., 1997; Eddy, 2009). We identified 13 MDH genes from the complete *G. arboreum* genome designated serially as *GaMDH1* to *GaMDH13* according to their genome organization from top to

bottom numerical chromosomal assignment. The detailed information of each *GaMDH* gene identified in the present study [i.e., gene ID, chromosome position, orientation, open reading frame (ORF) length, molecular weight (Mw), isoelectric point (pI) and subcellular location] were listed in Table 2. All MDH genes contained the conserved dinucleotide NAD-binding and catalytic domains. The putative MDH gene lengths varied from 972 to 1317 bp and encoded polypeptides ranging from 324-438 amino acids with predicted molecular weights and theoretical pI values between 33-47 kD and 6.35 to 8.65, respectively. In addition, the pI value of most of the proteins was greater than 6, while the average theoretical pI values for all proteins were 7 (Table 2). ExPASY analysis of the full-length deduced polypeptide sequences indicated differences in the *GaMDH* genes in terms of size, the encoded protein sequences and their respective physiochemical properties, suggesting that different MDH gene might function in malate synthesis and provides an effective framework to examine the molecular heterogeneity of plant MDH gene families.

Exon-intron organization of the *GaMDH* genes: To evaluate the structural diversity among the members of the MDH gene family, gene structure was compared in terms of their phylogenetic relationships to provide important clues concerning the evolution of specific gene families in the genome (Fig. 1). Gene structure analysis revealed that the number and pattern of exon distributions in *GaMDH* was quite different in each group; there were 6 or 7, 7 or 8 and 8 or 9 exons in groups I to III, respectively, whereas groups IV and V contained one and fourteen exons, respectively (Fig. 1). Combined with phylogenetic analysis, this data reveals that group I-III have undergone exon deletion. In depth gene structure analysis revealed that the lengths of the introns varied in almost all of the genes in the five groups. Therefore, we analysed the internal exons and introns of *GaMDHs* and found that the size of the *GaMDH* exons ranged from 36 to 1239 bp, with an average of 127 bp. About 28% of the *G. arboreum* MDH exons had a size below 400 bp, and 58% of exons were between 60-160 bp. Moreover, 4% of *G. arboreum* MDHs possessed exons larger than 1 kb in length, and 10% of exons were less than 60 bp. In contrast, the intron size distribution ranged from 75 bp to 1080 bp. There were 2 *G. arboreum* MDH introns (3%) larger than 1 kb, whereas 21% introns with sizes ranged from 400-600 bp. However, the majority of *G. arboreum* MDHs introns (75%) had sizes ranging from 70-400 bp, hence the average size of the *GaMDH* introns was 253 bp. More concisely, the high similarity levels of exon-intron organization and phylogenetic relationships between the MDH genes within each subgroup suggested gene structure conservation, thus supporting the close evolutionary relationships of GaMDHs and our classification of 5 groups (I-V) (Fig. 1).

Table 1. List of primers used for quantitative and semi-quantitative RT-PCR.

Primer	Sequences (5′-3′)	
	Forward	Reverse
GaMDH1	AATTGGGTCGACTGTCTCCTTC	GACCATTCTGAGACTTCCGGTT
GaMDH2	TGCATCTGGTGAAGTCTTTGGA	GTCCAACAAATCAGCTCGTTCC
GaMDH3	GCTAACAAAACCTTCTGCTGCA	GAAACTAGCGGGGACATCTTGA
GaMDH4	AACATTGCAGTCATGGTTGGTG	CCAGCCTTGTCAGACAGGTAAT
GaMDH5	CATAACAGAGCGCTTGGACA	AACCAATTGTCGTCCGCTAC
GaMDH6	AAGGCTCTAACAAAGCGAACAC	CCAAGCCTTACCTTAGAAGCAA
GaMDH7	CTTCTTGTCGTTCCAAAGGCAG	TATAGCCACCTTGTACGATGCC
GaMDH8	CCAAAACCCTCACGACAAA	CAACACCGGGAGTGTTAGC
GaMDH9	TAATTCTCGAGTCAACCAACGA	ATCAGCAGTAACACCAGGAGTG
GaMDH10	TCCCACTCTCCGCAAAAAT	AACCTCAGCTCGGGAATTG
GaMDH11	GCTCAAAGATGTTGTTGCGA	TTTCTCCGGAATTGAAGGTG
GaMDH12	ACCCTCAACCCCCACTATCT	AACCGAGACGAGAGGATTCA
GaMDH13	TCTCACTCTCTCGCCACTGA	AAGAGGAAAAGCAGCATCCA
UBQ7	GAAGGCATTCCACCTGACCAAC	CTTGACCTTCTTCTTCTTGTGCTTG

Table 2. The *MDH* genes in *G. arboreum* and properties of the deduced proteins.

Gene	Locus ID	Chr No.	Position (start)	Position (end)	Orientation	ORF (bp)	Size (aa)	Proteins MW(Da)	pI
GaMDH1	Cotton_A_29990	1	145180427	145181665	Forward	1239	412	43347.73	8.446
GaMDH2	Cotton_A_06084	2	94495889	94500626	Forward	1317	438	47938.97	6.544
GaMDH3	Cotton_A_26207	4	32424553	32425788	Reverse	1236	411	43141.49	7.954
GaMDH4	Cotton_A_36417	4	66304375	66307563	Reverse	999	332	35564.08	6.605
GaMDH5	Cotton_A_06357	4	90025314	90027221	Reverse	1002	333	35937.57	6.91
GaMDH6	Cotton_A_20125	4	115873296	115876159	Reverse	1032	343	35701.43	8.659
GaMDH7	Cotton_A_34770	6	63493440	63494663	Reverse	1224	407	43091.71	8.462
GaMDH8	Cotton_A_22294	6	78990020	78992691	Reverse	1017	338	35343.92	8.623
GaMDH9	Cotton_A_09811	7	37066244	37068693	Forward	999	332	35034.89	7.17
GaMDH10	Cotton_A_06762	7	52344145	52347005	Reverse	972	324	33986.09	6.75
GaMDH11	Cotton_A_07364	10	7680991	7682659	Forward	1131	376	41046.91	5.73
GaMDH12	Cotton_A_16744	12	60849333	60851395	Forward	1059	353	37227.14	6.746
GaMDH13	Cotton_A_22450	13	24827971	24830312	Forward	999	332	35584.09	6.353

The theoretical molecular weight (MW) and isoelectric point (pI) were calculated by ExPASy (http://cn.expasy.org/tools)

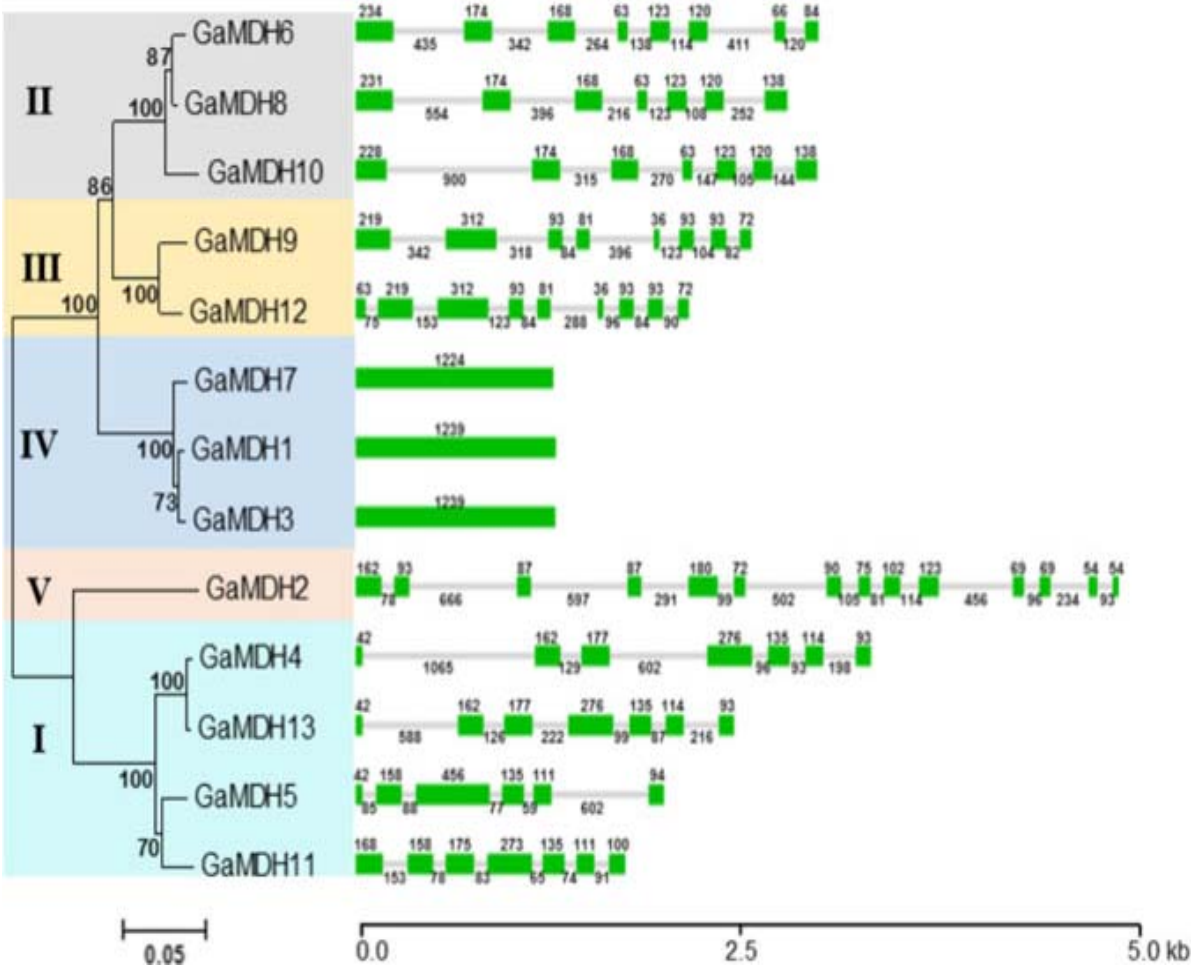

Fig. 1. Gene structure analysis of *GaMDH* genes.
Structures of the *GaMDH* genes. Introns and exons are represented by black lines and green boxes, respectively. The length of each intron and exon is indicated. Each section of the bar represents 2.5 kb.

Sequence characterization of the *GaMDH* genes: The multiple sequence alignment of the newly identified *GaMDH* gene family showed low identity to each other at both the nucleotide and protein levels, though *GaMDH12* shared only approximately 19% identity with the others. The *GaMDH* gene nucleotide sequence identity ranged from 35.2%-94%, whereas the amino acid sequence similarity ranged 18.7-97.7%. Among them, the coding sequences of *GaMDH1* with *GaMDH3* and *GaMDH6* with *GaMDH8* were closely related (87.1% nucleotide similarity resulting in 91.7% protein identity and 87% nucleotide similarity resulting in 89.6% protein identity, respectively). *GaMDH4* with *GaMDH13* showed 94.3% nucleotide similarity resulting in 97.9% identity, which was the highest identity at the protein level among the 13 *GaMDH* genes (Table 3).

The examination of the GaMDH protein sequences identified two functional domains (the NAD-binding domain and alpha-beta C-terminal domain) (Fig. 2). The conserved NAD binding site is found at the N-terminus of all *GaMDH* proteins except *GaMDH10,* whereas the conserved catalytic site containing catalytic residues (D^{175}, R^{178} and H^{202}) are located in the C-terminal domain of all *GaMDH* genes, reflecting its crucial role in catalysis. A variation of the characteristic amino acid sequence was present as -GXXGXXG- in the first nucleotide binding domain of the *GaMDH* gene family (Fig. 2). Residue D^{59} is crucial for co-enzyme binding and is chemically conserved with an acidic side chain in all NAD-dependent MDHs, while in the NADP-dependent MDHs, glycine (G^{59}) is substituted for aspartate (D^{59}) (e.g., *GaMDH2*). This result suggests that nucleotide binding characteristics can be modified in MDHs by single amino acid changes (Hall *et al*., 1992), which is similar to the findings of Feeney (Feeney *et al*., 1990). Subsequently, three arginine residues (R^{106}, R^{112}, and R^{178}) are highly conserved in all *GaMDHs* and critical for substrate binding (Fig. 2), implying that the mechanism of catalysis is similar to that in lactate dehydrogenase (Clarke *et al*., 1986). Furthermore, a detailed protein sequence alignment showed that the conserved enzymatic active site His-Asp that functions in the proton relay system was found in the C-terminal domains of all the *GaMDH* proteins, which also facilitates catalysis (Lamzin *et al*., 1994).

Table 3. Sequence identities among the ORF regions of the *GaMDH* genes and proteins.

	GaMDH1	GaMDH2	GaMDH3	GaMDH4	GaMDH5	GaMDH6	GaMDH7	GaMDH8	GaMDH9	GaMDH10	GaMDH11	GaMDH12	GaMDH13
GaMDH1	**100%**	19.40%	91.70%	22.80%	22.50%	56.70%	88.50%	60.90%	62.70%	59.40%	20.30%	60.90%	22.80%
GaMDH2	36.40%	**100%**	18.90%	42.40%	40.50%	21.00%	19.90%	21.60%	19.00%	20.80%	38.10%	18.70%	41.50%
GaMDH3	87.10%	35.50%	**100%**	23.50%	23.20%	55.60%	87.70%	60.10%	63.30%	59.00%	21.20%	61.50%	23.50%
GaMDH4	38.50%	50.80%	38.50%	**100%**	83.40%	23.20%	23.20%	24.20%	22.00%	22.80%	81.90%	22.70%	97.90%
GaMDH5	37.40%	49.30%	37.70%	72.90%	**100%**	21.90%	22.20%	22.30%	21.00%	21.10%	84.00%	21.40%	82.80%
GaMDH6	59.10%	36.80%	58.70%	38.00%	37.30%	**100%**	56.40%	90.50%	62.70%	82.90%	21.20%	61.80%	22.90%
GaMDH7	85.90%	36.70%	85.60%	37.20%	38.00%	57.80%	**100%**	60.70%	62.70%	58.70%	20.60%	59.80%	23.20%
GaMDH8	60.30%	36.00%	60.90%	38.60%	37.80%	87.10%	60.70%	**100%**	66.00%	91.10%	22.00%	65.30%	24.20%
GaMDH9	60.10%	38.00%	60.60%	37.20%	36.50%	61.10%	60.30%	63.70%	**100%**	65.10%	20.00%	87.70%	21.70%
GaMDH10	61.20%	38.20%	61.90%	38.80%	37.70%	84.80%	60.20%	87.10%	62.10%	**100%**	21.10%	65.40%	22.10%
GaMDH11	35.80%	48.10%	36.70%	74.80%	81.30%	37.60%	36.10%	37.90%	37.90%	38.40%	**100%**	19.30%	81.00%
GaMDH12	60.20%	35.20%	59.00%	39.40%	37.60%	61.30%	58.60%	62.30%	77.40%	62.70%	37.40%	**100%**	22.30%
GaMDH13	38.10%	50.00%	38.50%	94.30%	73.90%	39.30%	37.10%	39.40%	37.20%	39.60%	73.70%	39.30%	**100%**

Note: The sequence identities of the *GaMDH* genes and proteins are listed below and above the diagonal, respectively

Gene duplication and phylogenetic relationships of the *GaMDH* genes: To determine the chromosomal distribution of the MDH genes in *G. arboreum*, the 5′- and 3′- alignments of each gene model were downloaded from their corresponding genome databases. The 13 cotton MDH genes were unevenly distributed on chromosomes 1, 2, 4, 6, 7, 10, 12 and 13 (Fig. 3). *G. arboreum* contained one MDH gene each on Chr1, 2, 12 and 13, 4 on Chr4 and 2 on Chr6 and 7. In contrast, MDH genes were not observed on five chromosomes (Chr3, 5, 8, 9, and 11). However, all of the genes were located on different regions of the chromosomes with no apparent clustering (Fig. 3). Next, we examined whether duplication events participated in MDH gene family expansion in *G. arboreum*. Segmental and tandem duplications are typical gene duplication events that are crucial for the expansion of a number of multi-gene families (Kong *et al*., 2007). Physical mapping of the MDH gene family in *G. arboreum* revealed that the absolute majority of the genes were randomly dispersed across the genome. We selected five putative paralogous gene pairs with a high degree of protein sequence identity (>80%) and subsequently explored the degree to which their flanking genes were conserved. Only 3 paralogous MDH gene pairs with a high degree of protein sequence identity (>90%) were found to have a close phylogenetic relationship, and the identities of the protein-coding genes flanking each paralogous pair were similar. Indeed, the identities of the genes flanking both sides of the 3 pairs of paralogous *GaMDH* genes were found to be absolutely conserved (Table 4). The gene pairs (*GaMDH1/GaMDH3*), (*GaMDH6/GaMDH8*) and (*GaMDH4/GaMDH13*) are located on duplicated segments between chromosomes 1/4, 6/8, and 4/13, respectively. These results suggested that the paralogous gene pairs of *G. arboreum* arose from segmental duplication events during evolution.

It is generally assumed that the level of synonymous substitutions (*Ks*) between two homologous genes increases approximately linearly with time (Blanc & Wolfe, 2004). Thus, we estimated the evolutionary dates of the segmental duplication of the MDH paralogoues gene pairs from *G. arboreum* based on *Ks* calculations. The protein-coding genes flanking the 3 pairs of duplicated genes *in G. arboreum* had very consistent mean *Ks* values (0.582545, 0.614294 and 0.572172), suggesting that the duplicated *GaMDH* genes were under strong purifying selection pressure because their *Ka/Ks* ratio was less than 1. The segmental duplication events in this species may have occurred within the past 19-20 MYA, which is after the divergence of *cacao* from the common ancestor 18-58 MYA (Table 4). This result suggested that the time period was ulterior to the time at which the evolutionary lineage of cotton and *Arabidopsis,* circa 83-86 MYA, and was consistent with the time (20-40 MYA) when a recent large scale genome duplication event is thought to have occurred in cotton (Adams *et al*., 2003; Desai *et al*., 2006).

Fig. 2. Multiple sequence alignment of the GaMDH proteins.
GaMDH protein sequence alignment. The grey line indicates the 'Transit peptide', whereas light green and blue lines represent the NAD-binding and carboxy-terminal domains, respectively. The NAD-binding site and active site are indicated by purple and red boxes, respectively. The superscript '*' indicates important residue involved in the inter-conversion of oxaloacetate to malate.

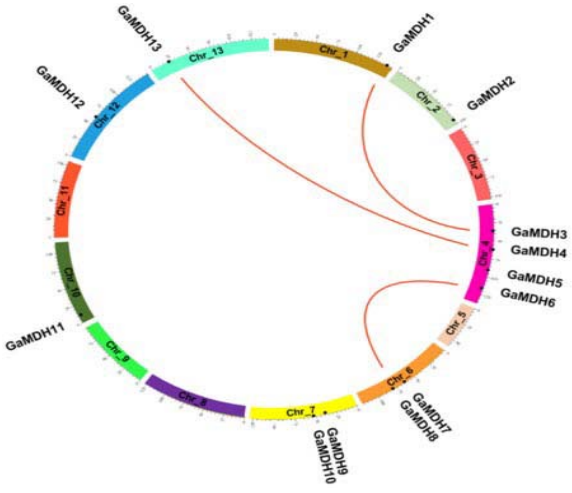

Fig. 3. Genomic locations and duplicated MDH gene pairs in *G. arboreum.*
Gene pairs located in the segmental duplicated chromosomal regions are linked using red lines.

For a detailed analysis of the evolutionary significance of the MDH proteins from higher plants, a neighbor-joining (NJ) phylogenetic tree was constructed. The bases of the phylogenetic trees included *Saccharomyces cerevisiae (ScMDH1)* as out-group. The reliability of the branches was assessed by bootstrapping analysis using 1000 replicates to estimate the gene duplication events during the expansion of the *GaMDH* gene family (Fig. 4). In addition, the phylogenetic trees reconstructed with maximum likelihood and minimal evolution methods, were almost identical with only minor differences at some branches, suggesting that the three methods were highly consistent with each other.

The NJ phylogenetic tree result showed that the 30 proteins were classified into five groups (I to V) with a maximum number of branches that had bootstrap values demonstrating statistically verifiable pairs of homologues. Subsequently, the MDH genes from eukaryotic species were also divided into 5 groups during the molecular evolution of MDH superfamily that supported the phylogenetic classification of the presented MDHs in

plants (Fig. 4) (Madern, 2002) and suggested that the differences between signal peptide sequences were larger than the conserved functional domains. Among these groups, the group I constituted the largest group containing 9 members, and the second largest one, group II, comprised of six MDHs, while in the smallest one, group V, only three MDHs were included. Moreover, the presence of a representative of all plant species in each group indicated that the MDH gene family was evolutionarily conserved in the higher plant species. In addition, the phylogenetic results revealed that cotton and *cacao* MDHs distributed more similarly than the *Arabidopsis* MDHs. Interestingly, in group I, *TcMDH2*, individually had two counterparts in cotton, matching that *GaMDH* gene segmental duplication occurred later than the split of cotton and *cacao* (Paterson *et al*., 2012).

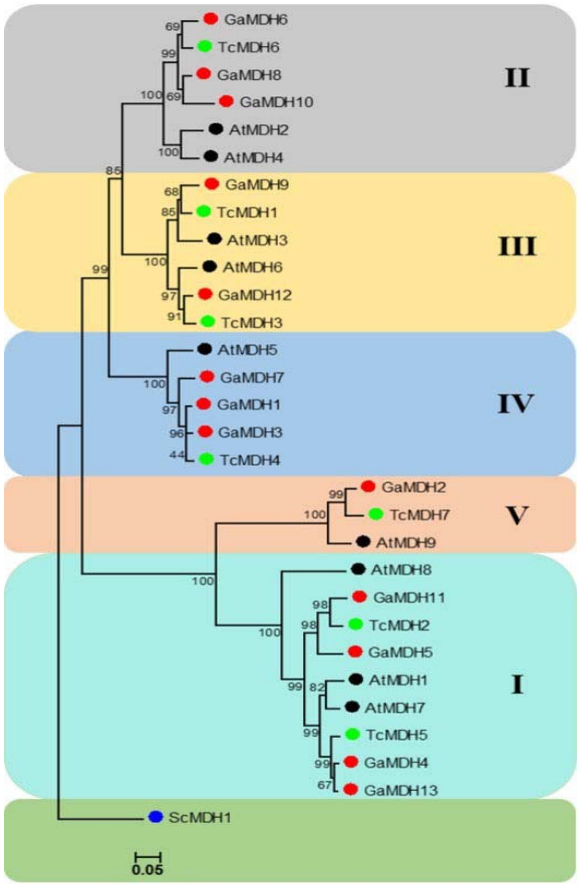

Fig. 4. Phylogenetic relationship of plant MDH proteins.

Expression profiles of the *GaMDH* genes: Expression profile analysis of the MDH gene family could help to expose the potential physiological procedures involved in plant growth and development. To better realize the potential functions of the *G. arboreum* MDH isoforms, we performed real time RT-PCR using primers specific for each *GaMDH* gene. The primer specificity and expression stability of *UBQ7* were confirmed by semi-quantitative RT-PCR.

All of the predicted genes showed differential expression levels in the tested tissues (Fig. 5). *GaMDH2* and *GaMDH5* showed low expression levels and were not detected in the roots and fibers, whereas *GaMDH11* and *GaMDH12* expressed at low level in the stem and leaf and the hypocotyl and petal tissues, respectively, which indicates that these genes might be expressed under special conditions or in other plant parts. *GaMDH3*/*GaMDH8* and *GaMDH7*/*GaMDH9* exhibited similar intermediate expression patterns in all vegetative tissues, with the lowest expression levels in the stem and leaf, suggesting that these genes play positive role in the reproductive development (Fig. 5A). *GaMDH4* and *GaMDH6* showed consistent high expression levels with the exception of the leaf, indicating that they might play an important role in the development of stem and flower. Notably, *GaMDH10* showed the lowest expression level, suggesting that this gene might be induced by absence of glycine motif (Tripathi *et al*., 2004). In contrast, *GaMDH13* showed the highest expression level in all vegetative tissues, including the fibers at 15 DPA, among all of the genes of the cotton MDH family, implying that this gene may play very important roles in multiple tissues (Fig. 5A). The semi-quantitative RT-PCR analysis provided similar results, indicating that the amplified segments for each *GaMDH* gene were very specific (Fig. 5B).

The conserved MDH proteins sequences from *G. arboreum*, *T. cacao* and *A. thaliana* with Yeast (*Saccharomyces cerevisiae)* as outgroup were aligned using Clustal W. The rooted phylogenetic tree was constructed using the neighbor-joining method with bootstrapping analysis (1000 replicates). The numbers beside the branches indicate the bootstrap values that support the adjacent node. The accession numbers or locus IDs of all MDH proteins used in this study are following. (Thecc1EG000888t1/ TcMDH1, Thecc1EG005626t1/ TcMDH2, Thecc1EG006355t1/ TcMDH3, Thecc1EG014018t1/ TcMDH4, Thecc1EG 020890t1/ TcMDH5, Thecc1EG029739t1/ TcMDH6, Thecc1EG031715t1/ TcMDH7, Thecc1EG 043361t1/ TcMDH8, AT1G04410.1/ AtMDH1, AT1G53240.1/ AtMDH2, AT2G22780.1/ AtMDH3, AT3G15020.1/ AtMDH4, AT3G47520.1/ AtMDH5, AT5G09660.1/ AtMDH6, AT5G43330.1/ AtMDH7, AT5G58330.1/ AtMDH8, and P17505/ ScMDH1).

Table 4. Duplicated *GaMDH* genes and the numbers of conserved protein-coding genes flanking them.

Duplicated *MDH* gene 1	Duplicated *MDH* gene 2	Numbers of flanking protein-coding genes	Mean *Ks*	Mini *Ks*	Maxi *Ks*	SD *Ks*	Date (MYA)
GaMDH1	*GaMDH3*	2	0.582545	0.572109	0.592981	0.0147587	19.4181
GaMDH6	*GaMDH8*	2	0.614294	0.596016	0.632573	0.0258497	20.4764
GaMDH13	*GaMDH4*	2	0.572172	0.559151	0.585193	0.0184145	19.0724

Abbreviation: *Ks*-synonymous substitution rates; SD *Ks*-Standard deviation *Ks*; Mini *Ks*-Minimum *Ks*; Max *Ks*-Maximum *Ks*; MYA-million years ago

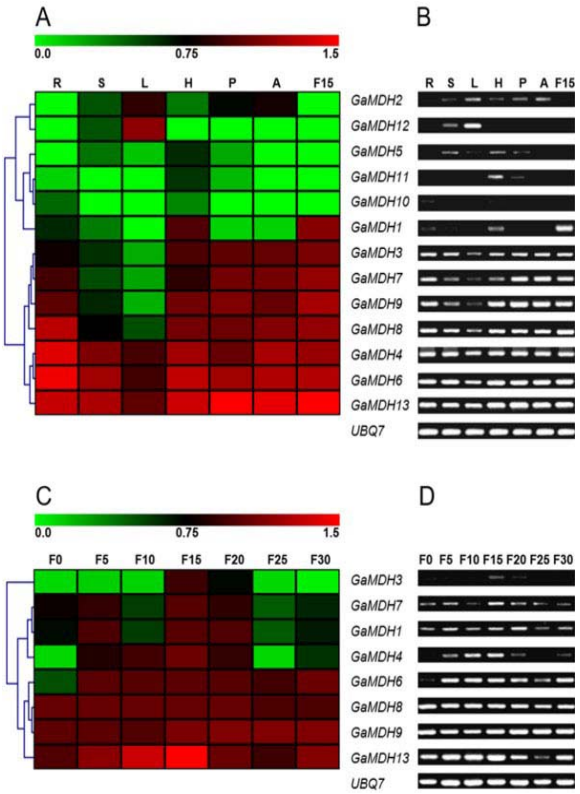

Fig. 5. Expression profiles of the *GaMDH* genes.
(A) Quantitative and (B) semi-quantitative RT-PCR analysis of *GaMDH* genes in the root (R), stem (S), leaf (L), hypocotyl (H), petal (P), anther (A), and fiber at 15 DPA (F15) of cotton plants. (C) Quantitative and (D) semi-quantitative RT-PCR analysis of the *GaMDH* genes in cotton fibers during different developmental stages. F0, ovules from 0 DPA, and F5 to F30, fibers from 5 to 30 DPA. The expression levels are indicated relative to cotton *UBQ7*.

To gain insights into target gene expression during the early and late fiber elongation stages, all of the genes expressed in the fibers were tested during different fiber developmental stages (Fig. 5C). The expression levels of *GaMDH1* and *GaMDH7* were relatively low and showed the lowest expression levels in the fibers at 10 and 25 DPA. *GaMDH3* was expressed only in the fibers at 15 and 20 DPA, whereas *GaMDH6* showed consistently higher expression patterns at all developmental stages of the fibers except 0 DPA. Similarly, *GaMDH8* and *GaMDH9* showed similar high expression levels from 0 to 20 DPA and 15 to 30 DPA, respectively, which suggests that all these genes may have different regulatory roles in fiber development. In addition, the expression patterns of *GaMDH4* and *GaMDH13* were absolutely different from other members; they began to increase in fibers from 5 to 15 DPA and then started to decline, implying that these two genes play a crucial role in fast-fiber elongation stage. In particular, *GaMDH4*, *GaMDH6*, *GaMDH8*, *GaMDH9* and *GaMDH13* are highly detectable in fiber, implying that they might be functionally related with the cotton fiber development (Fig. 5C). Additionally, the semi-quantitative RT-PCR analysis provided results that were similar to the quantitative RT-PCR analysis (Fig. 5D).

Taken together, the differential expression profile proposes that 8 of the 13 *GaMDH* genes were expressed during different developmental stages of fiber elongation, indicating a unique function for these MDH proteins in cotton fiber development (Ferguson *et al*., 1996). More interestingly, the *GaMDH13* expression level was twice that of *GaMDH4*, indicating a potentially crucial role for *GaMDH13* during fiber elongation and their expression pattern was correspond to dynamics of malate accumulation in the vacuole to increase turgor pressure, driving fiber elongation (Dhindsa *et al*., 1975). Legitimate with the previous proteomic analysis, our results indicated that *GaMDH13* was highly expressed in fibers at 15 DPA, malate level was at the peak (Ferguson *et al*., 1996), and suggested that *GaMDH13* favors the production of malate in elongating fiber cells compared with oxaloacetate. Several other studies reported that the overexpression of MDH increases the malate accumulation in yeast, hairy roots in *Arabidopsis* and cotton. In apple, the ortholog of *GaMDH13 i.e.,* the *MdcyMDH* (accession no. DQ221207), which also facilitates the transportation of malate into vacuole by generating electrochemical gradient and contributes to cell expansion, while suppressor had lower of malate (Yao *et al*., 2011b). In conclusion, all these results indicated that *GaMDH13* might function during malate-arbitrated fiber development. Furthermore, the present work would be useful in advancing our understanding for genome-wide analysis of the MDH gene family in different plant species and provide crucial clues for investigating the functions of MDH genes in cotton.

Acknowledgements

The authors would like to acknowledge members of the Laboratory of Molecular Biology at Tsinghua University for critical discussions. This work was supported by grants from the National Transgenic Animals and Plants Research Project (2011ZX08005-003 and 2011ZX08009-003). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Adams, K.L., R. Cronn, R. Percifield and J.F. Wendel. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci.* USA., 100: 4649-54.

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., 25: 3389-402.

Beeler, S., H.C. Liu, M. Stadler, T. Schreier, S. Eicke, W.L. Lue, E. Truernit, S.C. Zeeman, J. Chen and O. Kotting. 2014. Plastidial NAD-dependent malate dehydrogenase is

critical for embryo development and heterotrophic metabolism in *Arabidopsis. Plant Physiol*., 164: 1175-90.

Blanc, G. and K.H. Wolfe. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*., 16: 1667-78.

Clarke, A.R., D.B. Wigley, W.N. Chia, D. Barstow, T. Atkinson and J.J Holbrook. 1986. Site-directed mutagenesis reveals role of mobile arginine residue in lactate dehydrogenase catalysis. *Nature*., 324: 699-702.

Desai, A., P.W. Chee, J. Rong, O.L. May, and A.H. Paterson. 2006. Chromosome structural changes in diploid and tetraploid A genomes of *Gossypium*. *Genome*., 49: 336-45.

Dhindsa, R.S., C.A. Beasley and I.P. Ting. 1975. Osmoregulation in Cotton Fiber: Accumulation of Potassium and Malate during Growth. *Plant Physiol*., 56: 394-8.

Eddy, S.R. 2009. A new generation of homology search tools based on probabilistic inference. *Genome Inform*., 23: 205-11.

Emanuelsson, O., H. Nielsen, S. Brunak and G. von Heijne. 2000. Predicting subcellularlocalization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.,* 300: 1005-1016.

Faske, M., J.E. Backhausen, M. Sendker, M. Singer-Bayrle, R. Scheibe and A. Von Schaewen. 1997. Transgenic Tobacco Plants Expressing Pea Chloroplast Nmdh cDNA in Sense and Antisense Orientation (Effects on NADP-Malate Dehydrogenase Level, Stability of Transformants, and Plant Growth). *Plant Physiol*., 115: 705-715.

Feeney, R., A.R. Clarke and J.J. Holbrook. 1990. A single amino acid substitution in lactate dehydrogenase improves the catalytic efficiency with an alternative coenzyme. *Biochem. Biophys. Res. Commun*., 166: 667-72.

Ferguson, D.L., R.B. Turley, B.A. Triplett and W.R. Meredith. 1996. Comparison of Protein Profiles during Cotton (*Gossypium hirsutum* L.) Fiber Cell Development with Partial Sequences of Two Proteins. *J. Agr. Food. Chem*., 44: 4022-4027.

Fernie, A.R. and E. Martinoia. 2009. Malate. Jack of all trades or master of a few? *Phytochemistry*., 70: 828-32.

Gou, J.Y., L.J. Wang, C.P. Chen, W.L. Hu and X.Y. Chen. 2007. Gene expression and metabolite profiles of cotton fiber during cell elongation and secondary cell wall synthesis. *Cell Res*., 17: 422-34.

Gu, Z., A. Cavalcanti, F.C. Chen, P. Bouman and W.H. Li. 2002. Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. *Mol. Biol. Evol*., 19: 256-62.

Guo, A.Y., Q.H. Zhu, X. Chen and J.C. Luo. 2007. [GSDS: a gene structure display server]. *Yi Chuan*. 29: 1023-6.

Hall, M.D., D.G. Levitt and L.J. Banaszak. 1992. Crystal structure of *Escherichia coli* malate dehydrogenase. A complex of the apoenzyme and citrate at 1.87 A resolution. *J. Mol. Biol*., 226: 867-82.

Hovav, R., J.A. Udall, B. Chaudhary, R. Rapp, L. Flagel and J.F. Wendel. 2008. Partitioned expression of duplicated genes during development and evolution of a single cell in a polyploid plant. *Proc. Natl. Acad. Sci.* USA., 105: 6191-5.

Kong, H., L.L. Landherr, M.W. Frohlich, J. Leebens-Mack, H. Ma and C.W. Depamphilis. 2007. Patterns of gene duplication in the plant SKP1 gene family in angiosperms: evidence for multiple mechanisms of rapid gene birth. *Plant J*., 50: 873-85.

Lamzin, V.S., Z. Dauter and K.S. Wilson. 1994. Dehydrogenation through the looking-glass. *Nat. Struct. Biol.,* 1: 281-2.

Li, F., G. Fan, K. Wang, F. Sun, Y. Yuan, G. Song, Q. Li, Z. Ma, C. Lu, C. Zou, W. Chen, X. Liang, H. Shang, W. Liu, C. Shi, G. Xiao, G. Gou, W. Ye, X. Xu, X. Zhang, H. Wei, Z. Li, G. Zhang, J. Wang, K. Liu, R.J. Kohel, R.G. Percy, J.Z. Yu, Y.X. Zhu, J. Wang and S. Yu. 2014. Genome sequence of the cultivated cotton *Gossypium arboreum*. *Nat. Genet*., 46: 567-72.

Livak, K.J. and T.D. Schmittgen. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods*., 25: 402-8.

Longo, G.P. and J.G. Scandalios. 1969. Nuclear gene control of mitochondrial malic dehydrogenase in maize. *Proc. Natl. Acad. Sci.* USA., 62: 104-11.

Madern, D. 2002. Molecular evolution within the L-malate and L-lactate dehydrogenase super-family. *J. Mol. Evol*., 54: 825-40.

Minarik, P., N. Tomaskova, M. Kollarova and M. Antalik. 2002. Malate dehydrogenases-structure and function. *Gen. Physiol. Biophys*., 21: 257-65.

Musrati, R.A., M. Kollarova, N. Mernik and D. Mikulasova. 1998. Malate dehydrogenase: distribution, function and properties. *Gen. Physiol. Biophys*., 17: 193-210.

Paterson, A.H., J.F. Wendel, H. Gundlach, H. Guo, J. Jenkins, D. Jin, D. Llewellyn, K.C. Showmaker , S. Shu, J. Udall, M.J. Yoo, R. Byers, W. Chen, A. Doron-Faigenboim, M.V. Duke, L. Gong, J. Grimwood, C. Grover, K. Grupp, G. Hu, T.H. Lee, J. Li, L. Lin, T. Liu, B.S. Marler, J.T. Page, A.W. Roberts, E. Romanel, W.S. Sanders, E. Szadkowski, X. Tan, H. Tang, C. Xu, J. Wang, Z. Wang, D. Zhang, L. Zhang, H. Ashrafi, F. Bedon, J.E. Bowers, C.L. Brubaker, P.W. Chee, S. Das, A.R. Gingle, C.H. Haigler, D. Harker, L.V. Hoffmann, R. Hovav, D.C. Jones, C. Lemke, S. Mansoor, M. UR Rahman, L.N. Rainville, A. Rambani, U.K. Reddy, J.K. Rong, Y. Saranga, B.E. Scheffler, J.A. Scheffler, D.M. Stelly, B.A. Triplett, A. Van Deynze, M.F. Vaslin, V.N. Waghmare, S.A. Walford, R.J. Wright, E.A. Zaki, T. Zhang, E.S. Dennis, K.F. Mayer, D.G. Peterson, D.S. Rokhsar, X. Wang and J. Schmutz, 2012. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature*., 492: 423-7.

Pei, Y. 2015. The homeodomain-containing transcription factor, GhHOX3, is a key regulator of cotton fiber elongation. *Sci. China Life Sci*., 58: 309-10.

Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler and R. Lopez. 2005. InterProScan: protein domains identifier. *Nucleic Acids Res*., 33: 116-20

Taliercio, E., J. Scheffler and B. Scheffler. 2010. Characterization of two cotton (*Gossypium hirsutum* L) invertase genes. *Mol. Biol. Rep*., 37: 3915-20.

Tamura, K., G. Stecher, D. Peterson, A. Filipski and S. Kumar. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol*., 30: 2725-9.

Thompson, J.D., D.G. Higgins and T.J. Gibson. 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: 4673-80.

Tomaz, T., M. Bagard, I. Pracharoenwattana, P. Linden, C.P. Lee, A.J. Carroll, E. Stroher, S.M. Smith, P. Gardestrom and A.H. Millar. 2010. Mitochondrial malate dehydrogenase lowers leaf respiration and alters photorespiration and plant growth in *Arabidopsis*. *Plant Physiol*., 154: 1143-57.

Tripathi, A.K., P.V. Desai, A. Pradhan, S.I. Khan, M.A. Avery, L.A. Walker and B.L. Tekwani. 2004. An alpha-

proteobacterial type malate dehydrogenase may complement LDH function in Plasmodium falciparum. Cloning and biochemical characterization of the enzyme. *Eur. J. Biochem*., 271: 3488-502.

Wang, Z.A., Q. Li, X.Y. Ge, C.L. Yang, X.L. Luo, A.H. Zhang, J.L. Xiao, Y.C. Tian, G.X. Xia, X.Y. Chen, F.G. Li and J.H. Wu. 2015. The mitochondrial malate dehydrogenase 1 gene *GhmMDH1* is involved in plant and root growth under phosphorus deficiency conditions in cotton. *Sci. Rep*., 5: 10343.

Yao, Y.X., Q.L. Dong, H.Y. Zhai, C.X. You and Y.J. Hao. 2011. The functions of an apple cytosolic malate dehydrogenase gene in growth and tolerance to cold and salt stresses. *Plant Physiol. Biochem*., 49: 257-64.

Yao, Y.X., M. Li, H. Zhai, C.X. You and Y.J. Hao. 2011b. Isolation and characterization of an apple cytosolic malate dehydrogenase gene reveal its function in malate synthesis. *J. Plant Physiol.,* 168: 474-480.

Zhang, Z., J. Li, X.Q. Zhao, J. Wang, G.K. Wong and J. Yu. 2006. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics*., 4: 259-63.

Zhao, P.M., L.L. Wang, L.B. Han, J. Wang, Y. Yao, H.Y. Wang, X.M. Du, Y.M. Luo and G.X. Xia. 2010. Proteomic identification of differentially expressed proteins in the Ligon lintless mutant of upland cotton (*Gossypium hirsutum* L.). *J. Proteome Res*., 9: 1076-87.

Zhou, L., J. Duan, X.M. Wang, H.M. Zhang, M.X. Duan and J.Y. Liu. 2011. Characterization of a novel annexin gene from cotton (*Gossypium hirsutum* cv CRI 35) and antioxidative role of its recombinant protein. *J. Integr. Plant Biol*., 53: 347-57.

(Received for publication 17 May 2015)