RESEARCH ARTICLE

Comprehensive Analysis of the *COBRA-Like* (*COBL*) Gene Family in *Gossypium* Identifies Two *COBL*s Potentially Associated with Fiber Quality

Erli Niu[1], Xiaoguang Shang[1], Chaoze Cheng[1], Jianghao Bao[1], Yanda Zeng[1], Caiping Cai[1], Xiongming Du[2], Wangzhen Guo[1]*

1 State Key Laboratory of Crop Genetics & Germplasm Enhancement, Hybrid Cotton R & D Engineering Research Center, Ministry of Education, Nanjing Agricultural University, Nanjing, Jiangsu, China, 2 State Key Laboratory of Cotton Biology, Cotton Research Institute, Chinese Academy of Agricultural Sciences, Anyang, Henan, China

* moelab@njau.edu.cn

Abstract

COBRA-Like (*COBL*) genes, which encode a plant-specific glycosylphatidylinositol (GPI) anchored protein, have been proven to be key regulators in the orientation of cell expansion and cellulose crystallinity status. Genome-wide analysis has been performed in *A. thaliana*, *O. sativa*, *Z. mays* and *S. lycopersicum*, but little in *Gossypium*. Here we identified 19, 18 and 33 candidate *COBL* genes from three sequenced cotton species, diploid cotton *G. raimondii*, *G. arboreum* and tetraploid cotton *G. hirsutum* acc. TM-1, respectively. These *COBL* members were anchored onto 10 chromosomes in *G. raimondii* and could be divided into two subgroups. Expression patterns of *COBL* genes showed highly developmental and spatial regulation in *G. hirsutum* acc. TM-1. Of them, *GhCOBL9* and *GhCOBL13* were preferentially expressed at the secondary cell wall stage of fiber development and had significantly co-upregulated expression with cellulose synthase genes *GhCESA4*, *GhCESA7* and *GhCESA8*. Besides, *GhCOBL9* D_t and *GhCOBL13* D_t were co-localized with previously reported cotton fiber quality quantitative trait loci (QTLs) and the favorable allele types of *GhCOBL9* D_t had significantly positive correlations with fiber quality traits, indicating that these two genes might play an important role in fiber development.

Introduction

Cellulose, composed of long parallel linear β-1, 4-D-glucan chains, is the major component of the primary and secondary cell walls of plants. Studies on cellulose biosynthesis will not only facilitate an understanding of cell wall development but also open the possibility to increase the economic value of cotton fiber. Over the past decades, several genes essential for cellulose biosynthesis have been revealed, such as *cellulose synthase* (*CESAs*), *KORRIGAN* (*KOR*), *chitinase-like* genes (*CTLs*), *fasciclin-like arabinogalactan* genes (*FLAs*), *NAC* and *MYB* transcription

factors [1–9]. Although these important findings have been made, the molecular mechanism for cellulose biosynthesis and deposition remains largely unknown.

Besides the importance of cotton fiber to the textile industry, it is also considered to be an ideal model to investigate the mechanism of cell elongation and cellulose deposition [10–11]. The fiber cell origins from the epidermis of ovules and is a highly elongated and thickened single-cell trichome with >90% crystalline cellulose in mature fibers. Fiber formation involves four distinct but overlapping stages [12–13]: initiation (-3 days post anthesis [DPA] to 3 DPA), elongation (3 DPA to 23 DPA), secondary cell wall synthesis (16 DPA to 40 DPA) and maturation (40 DPA to 50 DPA), which collectively determine the fiber yield and quality traits. Disorders of any stages will affect the quality traits of fiber such as fiber length, strength, micronaire, elongation and fiber uniformity [14–15]. As cellulose accounts for about 35% in primary cell wall and more than 90% in secondary cell wall of cotton fiber, it plays an important role in fiber elongation and secondary cell wall formation [16]. Especially during the secondary cell wall synthesis stage, almost pure cellulose is produced in cotton fiber cell. Thus cotton fiber provides an excellent model for mining and characterization of key genes related to biosynthesis and assembly of cellulose. On the other side, studies on cellulose biosynthesis in cotton fiber will also contribute to improving fiber quality and yield.

COBRA-Like genes (*COBL*s), which encode the glycosylphosphatidylinositol (GPI) anchored proteins, have been reported to play significant roles in the orientation of microfibrils and cellulose crystallinity status. The *COBL* family has been identified in *Arabidopsis thaliana* [17], *Oryza sativa* (*BC1-Like*) [18], *Zea mays* (*BK2-Like*) [19] and *Solanum lycopersicum* [20]. As a result, 12, 11, 9 and 17 members were found in these plants respectively. The *COBL* members are characterized by an N-terminal signal peptide, a carbohydrate-binding module (CBM), a highly conserved central cysteine-rich domain (CCVS) and a C-terminal domain including a GPI modification motif and a hydrophobic tail for targeting the protein to the endoplasmic reticulum. The *COBL* family is divided into two groups with the first group similar to *COBRA* and the second group to *COBL7* in *A. thaliana* [17]. The mutant *cobra* in *A. thaliana* causes the defects in polar longitudinal expansion in root cells [21–23] and the mutant *bc1l4* in *O. sativa* also shows a typical dwarf phenotype [24]. Similarly male gametophyte transmission in *bc1l5* mutants is severely altered and specifically blocked [18] and the root hair of *rth3* mutants in *Z. mays* initiates primordia but fails to elongate properly [25]. *COBL* genes are largely responsible for the secondary cell wall biosynthesis. The mutations *cobl4* in *A. thaliana*, *brittle culm 1* (*bc1*) in *O. sativa* and *brittle stalk 2* (*bk2*) in *Z. mays* affect the mechanical strength of vascular bundles and have a significant reduction in cellulose content [26–28]. In addition, Liu et al. [29] demonstrated that *COBL* proteins could modulate cellulose structure by binding to cellulose and further affecting microfibril crystallinity. Whereas, it still needs to reveal the underlying molecular basis that how the *COBLs* proteins have functions in cellulose biosynthesis.

Recently the publically genomic information from three sequenced cotton species including the closest extant progenitor relatives for tetraploid cotton species, D-genome *Gossypium raimondii* [30], A-genome *Gossypium arboreum* [31], upland cotton genetic standard line *Gossypium hirsutum* acc. TM-1 [32], generates a solid foundation for characterizing gene families at a genome-wide level. Here we performed the first comprehensive investigation of the *COBL* gene family in three sequenced cotton species involving in the analysis of sequence phylogeny, genomic structure, chromosomal location and adaptive evolution. Further the expression patterns of *COBL* genes in *G. hirsutum* acc. TM-1 in various tissues/organs and different developmental stages of fiber development were elucidated by integrating RNA-Seq data and quantitative real-time PCR (qRT-PCR) analysis. Based on the co-expression analysis, correlation analysis and integration of quantitative trait loci (QTLs), we verified that two *COBLs*,

GhCOBL9 and *GhCOBL13* were significantly associated with fiber quality. This study will open up the possibility of exploring the use of *COBLs* to improve fiber quality in future cotton-breeding programs.

Materials and Methods

Database search and identification of *COBL* genes in *Gossypium*

The genomic database of three cotton species *G. raimondii*, *G. arboreum* and *G. hirsutum* acc. TM-1 were downloaded from http://www.phytozome.net/, http://cgp.genomics.org.cn and http://mascotton.njau.edu.cn/ respectively. The protein database of *A. thaliana*, *O. sativa* and *Z. mays* were obtained from The Arabidopsis Information Resource (TAIR: http://www.arabidopsis.org), the Rice Genome Annotation Project Database (RGAP release 7, http://rice.plantbiology.msu.edu/index.shtml) and http://www.phytozome.net/, respectively.

COBL domain (PF04833) was downloaded from Pfam (http://pfam.xfam.org/) and the genes shared with the domain could be obtained by taking PF04833 as the query to scan individually the three cotton protein database using the HMMER software version 3.0 [33] with the default parameters (E value < 0.01). Further, to confirm the genes above, gene prediction programs FGENESH (http://www.softberry.com/berry.phtml) and SMART (http://smart.embl-heidelberg.de/) were performed to detect the conserved motifs. Finally, the paralogs of *COBLs* in the three cotton species were confirmed by BLASTp.

Chromosomal locations and gene duplications

Chromosomal location of *COBL* genes in *G. raimondii* was performed using MapInspect software (http://www.plantbreeding.wur.nl/UK/ software_mapinspect.html). The nomenclature and description of the distribution of genes in chromosomes were derived from the map constructed by Zhao et al. [34]. In other two cotton species *G. arboreum* and *G. hirsutum* acc. TM-1, the nomenclature of *COBL* genes was following its ortholog in *G. raimondii*. DNAMAN (http://www.lynnon.com/) was conducted to calculate the sequence similarity among the *COBL* genes. Tandem duplicates were defined as genes separated by five or fewer genes and segmental duplicates were identified using the Plant Genome Duplication Database (http://chibba.agtec.uga.edu/duplication) [35]. The ratios of nonsynonymous to synonymous substitutions (Ka/Ks) were then analyzed to assess the selection pressure for the identified paralogous pairs [36].

Gene structure and phylogenetic analysis

MEGA 5.0 software (www.megasoftware.net) was used to construct the phylogenetic tree following the maximum likelihood method [37]. The parameters used were as follows: Test of phylogeny: Bootstrap method, No. of bootstrap replications: 1000, substitutions type: Amino acid, model/method: Jones-Taylor-Thornton (JTT) model, rates among sites: Uniform rates, gaps/missing data treatment: Complete deletion, ML heuristic method: Nearest-Neighbor-Interchange (NNI), initial tree for ML: Make initial tree automatically. The online ExPASy tool (http://www.expasy.org/tools/) and Gene Structure Display Server 2.0 (http://gsds.cbi.pku.edu.cn) were used to identify the physicochemical parameters [38] and the exon/intron organization. The alignment of the *COBL* family members was carried out with ClustalX 1.83 software [39].

Conserved motif annotation was performed using the MEME program [40]. If two or more motifs identified with SMART program represented the same domain and were located adjacently, they would be merged into one domain district. The parameters of the MEME program were as follows: Number of repetitions: Any, maximum number of motifs: 10, the optimum

motif widths: between 4 and 50 residues. Furthermore the signal peptide, hydrophobic plot and GPI modification motif (ω site for dissociation) were predicted by the SignalP 4.1 Server (http://genome.cbs.dtu.dk/services/SignalP/), KYTE DOOLITTLE HYDROPATHY PLOT (http://gcat.davidson.edu/DGPB/kd/kyte-doolittle.htm) and the big-PI predictor (http://mendel.imp.ac.at/sat/gpi/gpi_server.html) [41] respectively.

Plant materials and DNA/RNA isolation

The field evaluations of TM-1 and Hai7124 were conducted in PaiLou experimental field, Nanjing Agricultural University, Jiangsu Province, China. All necessary permits for collecting TM-1 and Hai7124 were obtained from Nanjing Agricultural University, China. The field evaluations of natural population including 285 *G. hirsutum* and 4 *G. barbadense* cultivars or lines were conducted in the experiment field of three Ecological Stations of the Institute of Cotton Research, CAAS. All necessary permits for the field evaluations of these accessions were obtained from the Institute of Cotton Research, CAAS, China. All the field evaluations were not relevant to human subject or animal research. Therefore, they did not involve any endangered or protected species.

G. hirsutum acc. TM-1, the genetic standard line of upland cotton and *G. barbadense* cv. Hai7124 with superior fiber traits were used for gene cloning. *G. hirsutum* acc. TM-1 were also employed to carry out the expression analysis. Among them, vegetative tissues (roots, stems and leaves) were collected from two-week-old seedlings; Floral tissues (petals and anthers) were collected on the day of flowering; Fibers were sampled on the different days post anthesis. All samples were quick-frozen in liquid nitrogen and stored at -70°C before use.

Natural population including 285 *G. hirsutum* and 4 *G. barbadense* cultivars or lines (S1 Table) were collected mainly from China and some from foreign countries, which were available from the National Mid-term Genebank of the Institute of Cotton Research, Chinese Academy of Agricultural Sciences (CAAS). These accessions were grown in three Ecological Stations of the Institute of Cotton Research with three replicates: Kuche of Xinjiang province (northwestward cotton growing region), Anyang of Henan province (Yellow River valley cotton growing region) and Nanjing of Jiangsu province (Yangtze River valley cotton growing region) during 2007 to 2009. After harvesting of the plants, the fiber samples including three biological replicates were tested with HVI spectrum in the Supervision, Inspection and Test Center of Cotton Quality, Ministry of Agriculture in China. The analysis of fiber quality traits focused mainly on 2.5% fiber span length (FL), fiber strength (FS), fiber micronaire (FM), fiber elongation (FE) and fiber uniformity (FU).

Total DNA was extracted from the leaves of seedlings as described by Paterson et al. [42]. All samples were quantified using "one drop spectrophotometer OD-1000+" (OneDrop, Nanjing, China) and adjusted to a concentration of 20–60 ng/μL. Total RNA was isolated from all samples with CTAB [43] and the trizol of "TransScript One-Step gDNA Removal and cDNA Synthesis SuperMix" kit (TransGen, Nanjing, China) was used subsequently to obtain the first strand cDNA. The components of each reaction sample included 1 μg RNA, 1 μL of anchored oligo(dT)$_{18}$ primer, 10 μL of 2*TS reaction mix, 1 μL of RT/RI enzyme mix, 1 μL of gDNA remover and additional ddH$_2$O to give a total volume of 20 μL. All cDNA samples were stored at -30°C before use.

Primer design and gene cloning

Primers for PCR amplification of DNA or cDNA for sequencing were designed with Primer Premier 5 (http://www.premierbiosoft.com). Beacon Designer 7.91 software was used to design the primers for qRT-PCR based on the coding sequences of *COBL* members close to the 3'-

UTR regions. Simultaneously, the specific primers for "Single Nucleotide Polymorphism (SNP)" of *GhCOBL9* and *GhCOBL13* were designed with the online software WebSNAPER-S-NAP (http://pga.mgh.harvard.edu/cgi-bin/snap3/websnaper3.cgi) [44]. All the gene-specific primers were provided in S2 Table.

High-Fidelity ExTaq DNA Polymerase (TaKaRa Biotechnology Co. Ltd. China) was employed to conduct a standard PCR analysis following the manufacturer's instructions with the amplification programs as follows: Pre-denaturation at 95°C for 5 min, 35 cycles of denaturation at 94°C for 45 s, annealing at 56°C for 45 s and extension at 72°C for 1 min/1 kb with a final extension at 72°C for 10 min. PMD19-T vector (TaKaRa Biotechnology Co. Ltd. China) and *E. coli* "strain" Top10 were used for transformation of target fragments. In order to obtain the sequences from both the A-subgenome and D-subgenome, at least 10 clones for each gene from each of the tetraploid species were picked randomly and sequenced with a minimum of three clones was used to determine the gene sequence in each duplicated copy.

Expression pattern analysis

The high-throughput RNA-sequencing data of *G. hirsutum* acc. TM-1 were employed from the accession codes SRA: PRJNA248163 in the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/) including the following tissues: leaves, roots and stems of two-week-old plants; petals and anthers dissected from whole mature flowers; ovules and fiber mixtures from plants -3, -1, 0, 1, 3 DPA and fibers from plants 5, 10, 20 and 25 DPA with three biological replicates. The expression levels of *COBLs* were calculated using the fragments per kilobase of exon model per million mapped reads (FPKM) method by using Cufflinks software with default parameters (http://cufflinks.cbcb.umd.edu/).

qRT-PCR was performed on an ABI 7500 real-time PCR system (ABI Biosystems, http://www.lifetechnologies.com) with the amplification programs as follows: Pre-denaturation at 95°C for 10 min, 40 cycles of denaturation at 95°C for 15 s, annealing at 60°C for 15 s and extension at 72°C for 15 s with a final extension at 72°C for 10 min. Each reaction sample was mixed with 10 μl of SYBR Green Master (Rox), 1.5 μL of cDNA, 5 μM of primers and 7.5 μL of ddH$_2$O to give a total volume of 20 μL. For all real-time PCR reactions, three technical replicates were performed in each of the three biological experiments. The relative expression level was calculated using the $2^{-\triangle CT}$ method [45] and the expression level of the cotton histone3 (AF024716) gene was used as the endogenous control [46].

Pearson correlation analysis on expression pattern was used to calculate the correlation coefficient between RNA-seq and qRT-PCR detection, *GhCESAs* and *GhCOBLs* and different paralog pairs of *GhCOBLs*. For correlation analysis, the total expression of homeologous pairs in each gene from RNA-seq data was merged for matching the result from qRT-PCR detection using gene-specific primer pairs.

SNP/EcoTilling assays and QTLs integration for function prediction

The amplification program for SNP analysis by PCR was as follows: Pre-denaturation at 95°C for 10 min, 28 cycles of denaturation at 95°C for 15 s, annealing at 62°C for 30 s and extension at 72°C for 30 s followed by denaturing gel electrophoresis. For "EcoTilling" detection, several successive steps were performed following a method modified from that of Zeng et al. [47]. PCR amplification of the natural population (with annealing at 67°C for 30 s) was carried out twice to determine the density and specificity of targeted region with the same forward primer and two different reverse primers. Each sample was then mixed with DNA from the genetic standard line TM-1 in a 1:1 ratio and these were hybridized through 40 cycles of denaturation at 99°C for 10 min and annealing at 72°C with a decrease of 0.3°C per cycle. Finally

CELl reaction was carried out for 30 min to cleave any mismatches between the query DNA and reference DNA. The denaturing gel electrophoresis after silver staining was used to pinpoint the cleaved fragments with the fragments sizes marked [48]. The cut DNAs were visible as bands and those with faster mobility than the full-length product were considered as polymorphisms. In addition, the DNA fragments around the polymorphic locus were cloned and sequenced to confirm the applicability and accuracy of the polymorphism detection.

Correlation analysis estimated by SPSS18.0 (http://www.spss.com.cn/) was conducted using the LSR (the least significant range) method to study significant difference of fiber quality traits between the two genotypes of targeted genes.

QTLs related to fiber quality traits (including FL, FS, FM, FE and FU) were retrieved and anchored to the *G. hirsutum* acc. TM-1 genome. Within ± 5 Mb interval flanking the QTLs, we integrated the QTLs with the prior target genes for association analysis using the MapChart program (www.joinmap.nl.).

Results

Genome-wide identification of *COBL* gene family in *Gossypium*

The whole genome sequence scaffolds of three sequenced cotton species (*G. raimondii*, *G. arboreum* and *G. hirsutum* acc. TM-1) were used for the genome-wide exploration of *COBL* family genes in *Gossypium*. Using the *COBL* protein database (PF04833), we searched the three protein database of cotton species by HMMER software. As a result, 19, 18 and 33 *COBL* members were identified respectively in *G. raimondii*, *G. arboreum* and *G. hirsutum* acc. TM-1 as shown in Table 1. From the phylogenetic view, one member in the diploid *G. raimondii* would be corresponding to one homologous gene in *G. arboreum* and two homologous genes (homeologs from A and D subgenome) in tetraploid *G. hirsutum* acc. TM-1. The inconsistency of the homologous genes among these three cotton species might result from the gene loss during their individually evolution process or assembly error in partial chromosomal regions, and need to be further confirmed. The related information of *COBLs* in *Gossypium* was summarized in Table 1.

Chromosomal distribution and gene duplication of *COBL* genes in *G. raimondii*

Two sequenced diploid genomes A genome *G. arboreum* and D genome *G. raimondii* were the closest extant progenitor relatives for tetraploid cotton species. Of them, the genome sequence in *G. raimondii* had been well assembled and annotated [30] and the collinearity of linkage groups between *G. raimondii* and *G. hirsutum* acc. TM-1 genome was obvious [32]. So we selected preferentially *G. raimondii* genome information to characterize the *COBL* family genes. By integrating 13 scaffolds of the *G. raimondii* genome (named as Chr. 1 to Chr. 13) with the reported high-density interspecific genetic map of allotetraploid cultivated cotton species [34], we obtained the corresponding relationship of orthologs between Chr. 1 to Chr. 13 in *G. raimondii* and D1 to D13 chromosomes in tetraploid cotton species. As a result, 19 candidate *COBL* genes were matched to 10 scaffolds of the *G. raimondii* genome. Based on the order of the homologs on chromosomes, we named the *COBL* members in *G. raimondii* as *GrCOBL1* to *GrCOBL19*. Correspondingly, we also named *COBL* family genes *GaCOBL1* to *GaCOBL19* in *G. arboreum* and *GhCOBL1* to *GhCOBL19* in *G. hirsutum* acc. TM-1 following the orthologous relationship with the ones in *G. raimondii* (Table 1). The chromosomal distribution patterns of these *GrCOBL* genes were uneven. As in Fig 1, no *GrCOBL* was detected in the three chromosomes D4, D5 and D9, four *GrCOBLs* were found on the D12 chromosome, three on

Table 1. Characterization of COBL family genes identified in *G. raimondii* and their phylogenetic relationship in other cotton species[a].

Gene name	Chr.[b]	Gene ID in *G. raimondii*	Calculated protein information			Ortholog Gene ID (Gene name) in	
			Length (a.a.)	pI	MW (kDa)	*G. arboreum*	*G. hirsutum* acc. TM-1[c]
GrCOBL1	D1	Gorai.002G006300	356	8.23	40.01	Cotton_A_00213 (*GaCOBL1*)	Gh_A01G0051/Gh_D01G0052 (*GhCOBL1* A_t/D_t)
GrCOBL2	D1	Gorai.002G083800	667	9.2	74.54	Cotton_A_04743 (*GaCOBL2*)	Gh_A01G0582/Gh_D01G0593 (*GhCOBL2* A_t/D_t)
GrCOBL3	D2	Gorai.005G065200	657	6.6	72.31	Cotton_A_21798 (*GaCOBL3*)	Gh_A02G0513/Gh_D02G0577 (*GhCOBL3* A_t/D_t)
GrCOBL4	D3	Gorai.003G102500	489	9.09	54.65	Cotton_A_34370 (*GaCOBL4*)	Gh_A03G0631.1/Gh_D03G0919.1 (*GhCOBL4* A_t/D_t)
GrCOBL5	D3	Gorai.003G102700	439	8.9	49.18	Cotton_A_34369 (*GaCOBL5*)	Gh_A03G0631.2/Gh_D03G0919.2 (*GhCOBL5* A_t/D_t)
GrCOBL6	D6	Gorai.010G177700	661	9.11	74.22	Cotton_A_35783 (*GaCOBL6*)	Gh_A06G1281/Gh_D06G1606 (*GhCOBL6* A_t/D_t)
GrCOBL7	D7	Gorai.001G076200	438	8.98	49.34		Gh_A07G0608/Gh_D07G0673 (*GhCOBL7* A_t/D_t)
GrCOBL8	D7	Gorai.001G208300	455	9.03	50.99	Cotton_A_33232 (*GaCOBL8*)	Gh_A07G1625/Gh_D07G1823 (*GhCOBL8* A_t/D_t)
GrCOBL9	D8	Gorai.004G063600	430	9.07	48.37	Cotton_A_03290 (*GaCOBL9*)	Gh_A08G0474 (*GhCOBL9* A_t)
GrCOBL10	D10	Gorai.011G196300	624	5.43	68.66	Cotton_A_19861 (*GaCOBL10*)	Gh_A10G1503/Gh_D10G1749 (*GhCOBL10* A_t/D_t)
GrCOBL11	D11	Gorai.007G125700	656	5.21	71.69	Cotton_A_26440 (*GaCOBL11*)	Gh_A11G1019/Gh_D11G1175 (*GhCOBL11* A_t/D_t)
GrCOBL12	D11	Gorai.007G176300	450	8.59	50.28	Cotton_A_12493 (*GaCOBL12*)	Gh_A11G1468/Gh_D11G1625 (*GhCOBL12* A_t/D_t)
GrCOBL13	D11	Gorai.007G176400	433	9.05	49.23	Cotton_A_12494 (*GaCOBL13*)	Gh_A11G1469/Gh_D11G1626 (*GhCOBL13* A_t/D_t)
GrCOBL14	D12	Gorai.008G027100	446	7.77	50.33	Cotton_A_10803 (*GaCOBL14*)	
GrCOBL15	D12	Gorai.008G033200	442	5.42	49.76	Cotton_A_15300 (*GaCOBL15*)	Gh_D12G0298 (*GhCOBL15* D_t)
GrCOBL16	D12	Gorai.008G200000	455	8.87	50.79	Cotton_A_31975 (*GaCOBL16*)	Gh_A12G1661/Gh_D12G1818 (*GhCOBL16* A_t/D_t)
GrCOBL17	D12	Gorai.008G200100	434	9.27	48.98	Cotton_A_31976 (*GaCOBL17*)	Gh_D12G1819 (*GhCOBL17* D_t)
GrCOBL18	D13	Gorai.013G039400	452	7.05	50.92	Cotton_A_19405 (*GaCOBL18*)	Gh_A13G0320/Gh_D13G0359 (*GhCOBL18* A_t/D_t)
GrCOBL19	D13	Gorai.013G163300	660	8.68	73.63	Cotton_A_17907 (*GaCOBL19*)	Gh_A13G1193/Gh_D13G1487 (*GhCOBL19* A_t/D_t)

[a] Genes information in *G. raimondii* from Paterson et al. [30], *G. arboreum* from Li et al. [31] and *G. hirsutum* from Zhang et al. [32].

[b] Chromosome numbers D1 to D13 refer to chromosomes name of D subgenome in allotetraploid cultivated cotton species [34].

[c] A_t and D_t were derived from the A-genome and D-genome progenitor in the tetraploid cotton.

doi:10.1371/journal.pone.0145725.t001

the D11 chromosome, two each were found on the D1, D3, D7 and D13 chromosomes, and only one each on the D2, D6, D8 and D10 chromosomes.

It was believed that whole genome duplication with genetic redundancy, chromosomal rearrangement and genome downsizing brought about a gene family a substantial expansion and evolutionary novelties [49]. The *G. raimondii* genome had undergone at least two rounds of genome-wide duplication [30]. To understand the mechanism of the *COBL* gene family
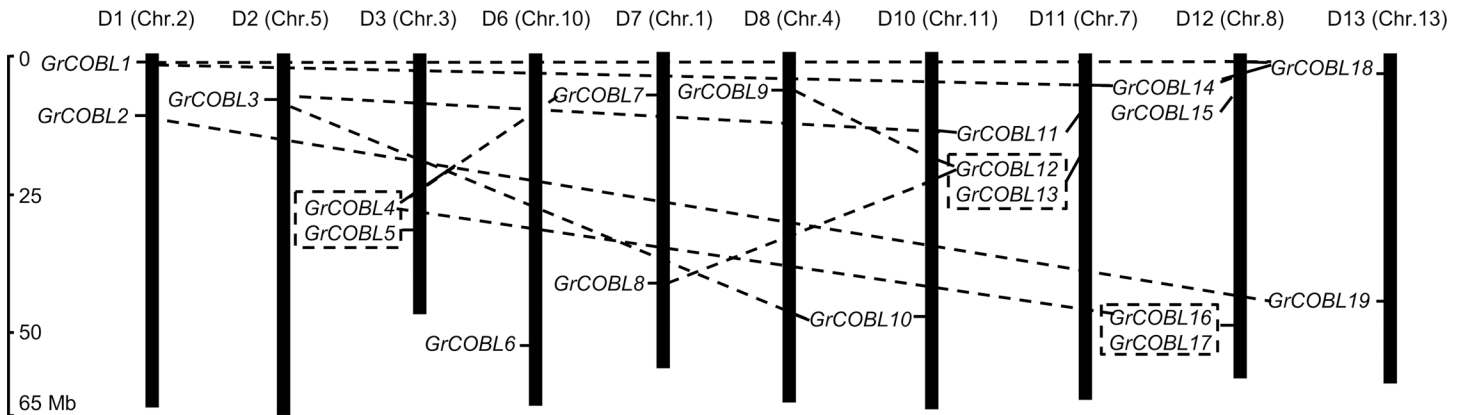
Fig 1. Chromosomal distribution of the COBL family genes in G. raimondii. The chromosome numbers were consistent with the interspecific genetic map (D1 to D13) in allotetraploid cultivated cotton species [34] and the scaffolds (Chr.1 to Chr.13) in the genomic data of *G. raimondii* [30]. The nomenclature of *COBLs* was based on the order of the chromosomes in *G. raimondii*. Lines were drawn to connect the duplicated genes.

doi:10.1371/journal.pone.0145725.g001

expansion in *G. raimondii*, duplication events including tandem and segmental duplications were investigated by genome synteny analysis (Fig 1). Three gene pairs (*GrCOBL4*/*GrCOBL5*, *GrCOBL12*/*GrCOBL13* and *GrCOBL16*/*GrCOBL17*) were detected as tandem duplication events, with 69.5%, 64.4% and 53.5% sequence similarity respectively. In addition ten gene pairs with high similarity that were located on segmental duplicated blocks were found to be a result of segmental duplication. All these paralogous pairs had ratios of nonsynonymous to synonymous substitutions (Ka/Ks) of less than 0.2 (Table 2) implying that *COBL* family genes in *G. raimondii* had mainly been subject to purifying selection pressure during the evolutionary process.

Phylogenetic and structural analysis of *COBL* genes

The overall protein sequences including *COBLs* in the eudicots (*A. thaliana*, *G. raimondii*) and monocots (*O. sativa*, *Z. mays*) were utilized to construct an unrooted tree to clarify the

Table 2. Nucleotide diversity and expression correlation among 10 COBL paralog pairs.

Paralog pairs in *G. raimondii*[a]	Subgroup	Sequence similarity (%)[b]	Ka	Ks	Ka/ Ks	Pearson correlation coefficient of expression pattern in *G. hirsutum* acc.TM-1[c]
GrCOBL1-GrCOBL14	I	69.7	0.183	1.083	0.169	--
GrCOBL1-GrCOBL18	I	69.9	0.176	0.997	0.177	0.000
GrCOBL2-GrCOBL19	II	65.5	0.233	1.209	0.193	1.000
GrCOBL3-GrCOBL10	II	83.2	0.086	0.587	0.146	0.777
GrCOBL3-GrCOBL11	II	63.1	0.279	2.313	0.121	-0.030
GrCOBL4-GrCOBL7	I	55.7	0.298	1.780	0.167	-0.012
GrCOBL4-GrCOBL16	I	60	0.311	None	None	0.438
GrCOBL8-GrCOBL12	I	84.9	0.075	0.441	0.169	-0.104
GrCOBL9-GrCOBL12	I	63.5	0.263	None	None	-0.147
GrCOBL14-GrCOBL18	I	77.6	0.138	0.719	0.191	--

[a] Paralog pairs were referred to Plant Genome Duplication Database and the Ka/Ks ratio was calculated to assess the selection pressure for each paralog pair.

[b] DNAMAN was used to calculate the sequence similarity between the paralog pairs.

[c] Pearson correlation coefficient: r>0: positive correlation; r<0: negative correlation; r = 0: no linear correlation; --: undetectable.

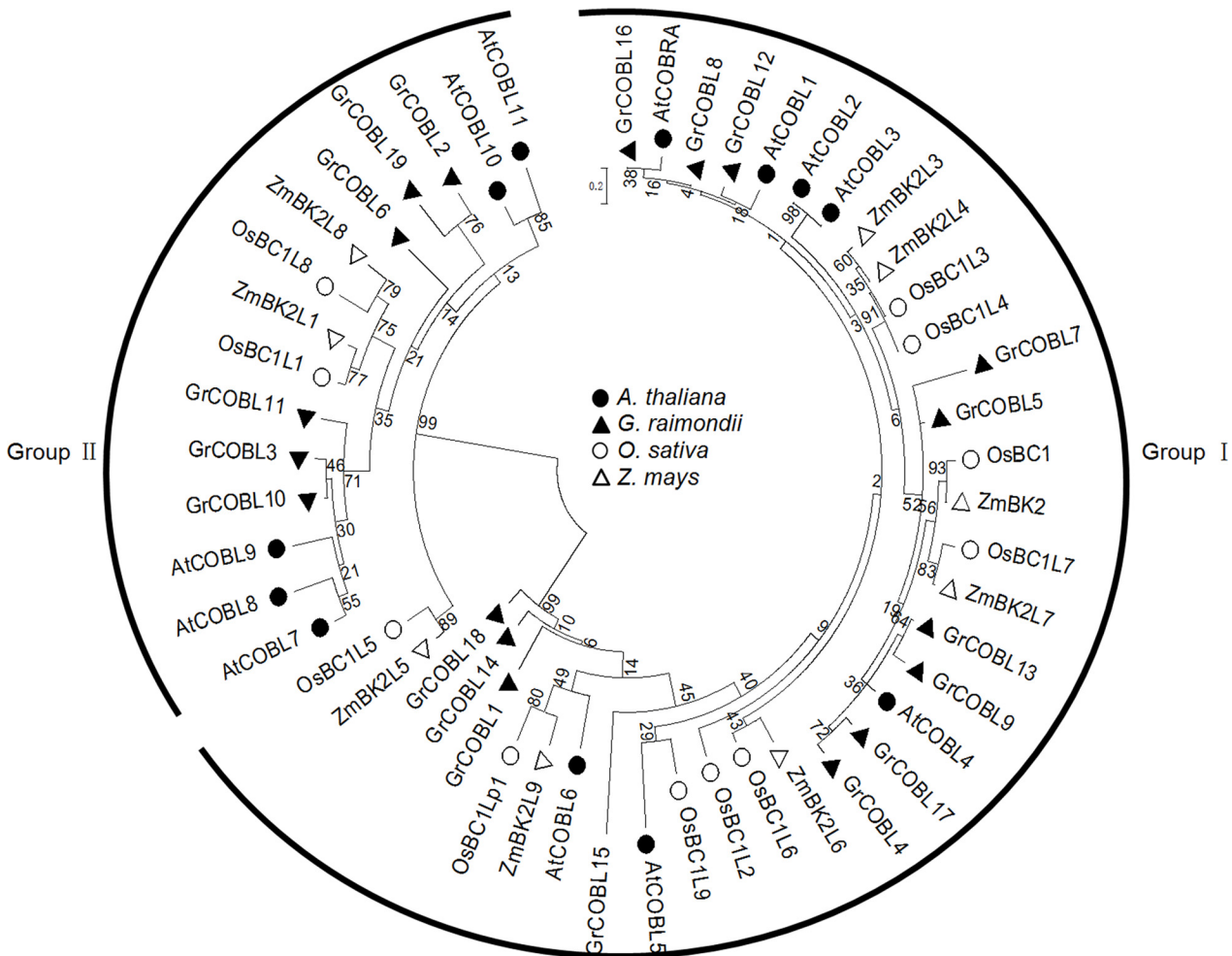doi:10.1371/journal.pone.0145725.t002

Fig 2. Phylogenetic analyses of *COBL* family genes in eudicot (*A. thaliana*, *G. raimondii*) and monocot (*O. sativa*, *Z. mays*). Amino acid sequences were aligned using ClustalW and the phylogenetic tree was conducted using MEGA 5.0 software with the maximum likelihood method. Four different color fonts were represented for the four species.

phylogenetic relationship of *COBL* genes from different species. Consistent with orthologs in *A. thaliana* and other species, the *COBL* family members were clustered into two subgroups Group I and Group II, with phylogenetically related to *AtCOBRA* and *AtCOBL7* in *A. thaliana* respectively. As in Fig 2, thirteen *COBL* genes from *G. raimondii*, seven from *A. thaliana*, eight from *O. sativa* and six from *Z. mays* were clustered in Group I and six from *G. raimondii*, five from *A. thaliana*, three from *O. sativa* and three from *Z. mays* were in Group II. The phylogenetic analysis indicated that the *COBLs* were descendants of an ancient duplication that occurred even before the separation of monocots and dicots.

The orthologs from three cotton species, *G. raimondii*, *G. arboreum* and *G. hirsutum* acc. TM-1 shared the same gene structure, thus, the *COBLs* in *G. raimondii* were used to give the insights into the diversity of the *COBL* family in *Gossypium* (Fig 3A and 3B). *GrCOBLs* in group I had more introns than those in group II. Besides, motif analysis (Fig 3C) displayed the conserved domains characterized by *COBL* gene family: the N-terminal signal peptide, the carbohydrate-binding module, a highly conserved CCVS motif and C-terminal domains that included the GPI modification motif and a hydrophobic tail, with the exception that three
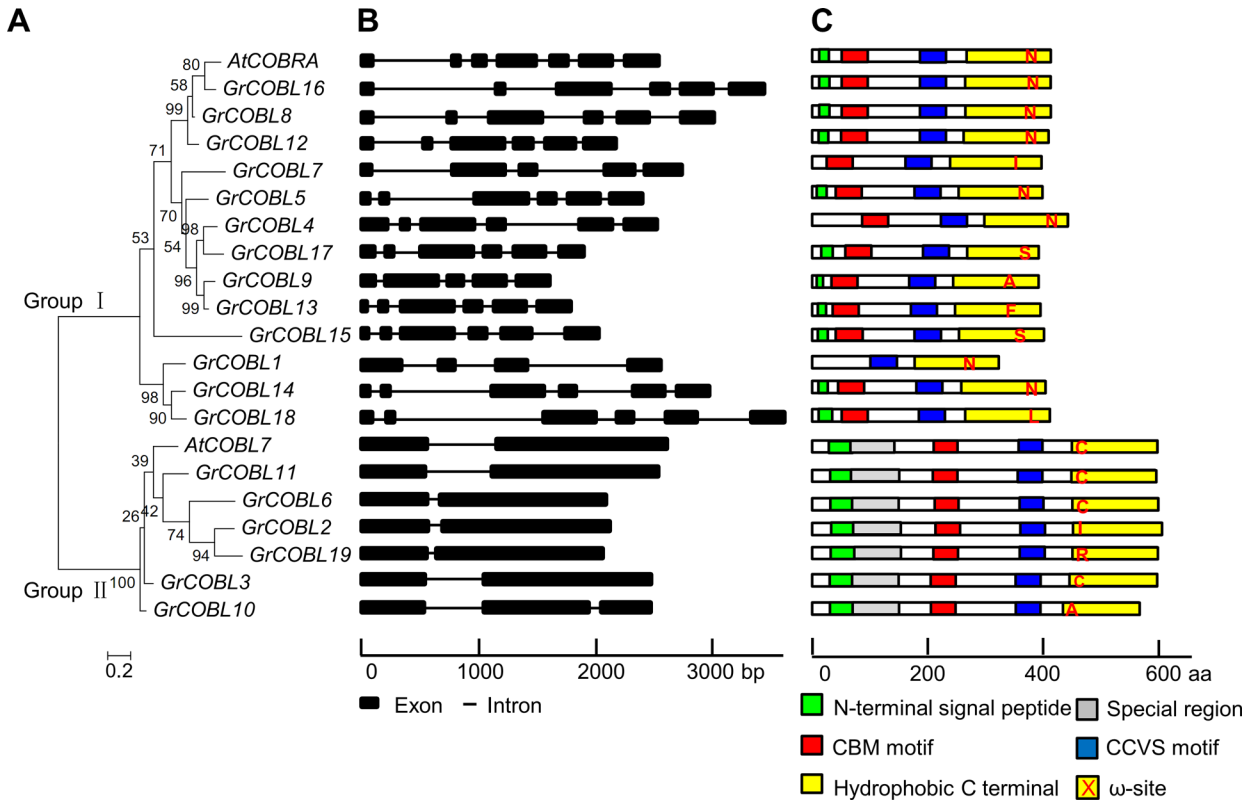
Fig 3. Phylogenetic relationship, gene structure and motif compositions of the *COBLs* in *G. raimondii* and *AtCOBRA*, *AtCOBL7* in *A. thaliana*. A. The phylogenetic tree was conducted using MEGA 5.0 software with the maximum likelihood method. B. Exon/intron organization of *GrCOBL* members. Black boxes and black lines represented the exons and introns. The length of nucleotides was shown below. C. Schematic representation of the conserved motifs of *GrCOBL* proteins elucidated by MEME and SMART. Each motif was marked by the different colors and the ω-sites were showed in red font. The length of amino acids was shown below.

doi:10.1371/journal.pone.0145725.g003

genes *GrCOBL1*, *GrCOBL4* and *GrCOBL7* lacked the N-terminal hydrophobic domain and *GrCOBL1* lacked the CBM. Structurally the *GrCOBLs* in these two groups displayed the divergence at the N terminus. The members in Group II contained an extra amino acid stretch after the N-terminal signal peptide and lead to the wide diversity in the length of amino acids ranging from 356 to 667 (Table 1).

Moreover, multiple sequence alignments showed that hydrophobicity was ubiquitous for the *COBL* members, although the low sequence consistency of the N-terminal and C-terminal regions existed (S1 Fig). With the exception of the CCVS motif and the aromatic amino acids in the CBM, which were highly conserved, the majority of the residues showed characteristics that were specific to the two clustered groups. Besides, the ω-sites in Group I were close to the end of the C-terminal regions, while the ones in Group II were close to the front of the C-terminal regions and were followed by a longer hydrophobic tail. This had not been reported previously and required to be further confirmed.

Expression patterns of *COBL* genes in *G. hirsutum* acc. TM-1

Based on RNA-Seq data of *G. hirsutum* acc. TM-1 including vegetative tissues (roots, stems and leaves), floral tissues (petals and anthers) and fibers in the development stage (0 DPA, 10 DPA and 20 DPA), we detected the expression patterns of *COBL* genes in *G. hirsutum* acc. TM-1. With the exception of *GhCOBL14* (not detected in *G. hirsutum* acc. TM-1) and

GhCOBL1, *GhCOBL12* (FPKM<1.0 in all tested tissues), the expression patterns of other 15 *GhCOBL*s with distinguishable expression of homeologs could be grouped into four major classes (Fig 4). Four genes *GhCOBL3*, *GhCOBL8*, *GhCOBL10* and *GhCOBL16* in Class I were highly expressed in all tissues examined with slight expression differences among the different samples. *GhCOBL9* and *GhCOBL13* in Class II showed the predominant expression in the fiber tissue at 20 DPA. Five genes *GhCOBL2*, *GhCOBL6*, *GhCOBL15*, *GhCOBL18* and *GhCOBL19* in Class III showed preferential expression in floral tissue with some expressed in fiber tissue. Five genes *GhCOBL4*, *GhCOBL5*, *GhCOBL7*, *GhCOBL11* and *GhCOBL17* in Class IV showed preferential expression in root or stem tissue. Generally, the similar expression pattern between homeologs in most *COBL*s were detected in different tissues in *G. hirsutum* acc. TM-1. Further, the gene-specific qRT-PCR was conducted to verify the expression of the 15 *GhCOBL*s in tetraploid cotton (S2 Fig) and the results of qRT-PCR were in agreement with the RNA-Seq data (S3 Fig). A correlation between the expression patterns in different tissues for the paralogous pairs was also detected (Table 2). The correlation coefficients in three *COBL* pairs (*GhCOBL2*/ *GhCOBL19*, *GhCOBL3*/*GhCOBL10* and *GhCOBL4*/*GhCOBL16*) showed the positive correlation, with the expression correlation of *GhCOBL2*/*GhCOBL19* and *GhCOBL3*/*GhCOBL10* greater than 0.5. However, other pairs had no clear correlation.

To elucidate the expression profiles of *COBL* genes during fiber development, we further investigated RNA-Seq data of 18 *GhCOBL* genes (with *GhCOBL14* not detected in *G. hirsutum* acc. TM-1) in nine samples involved in the fiber development stages of initiation (-3 DPA, -1 DPA, 0 DPA and 1 DPA), elongation (3 DPA, 5 DPA and 10 DPA) and secondary cell wall biosynthesis (20 DPA and 25 DPA) in *G. hirsutum* acc. TM-1 (Fig 5). With the exception of six genes *GhCOBL1*, *GhCOBL2*, *GhCOBL4*, *GhCOBL7*, *GhCOBL11* and *GhCOBL19*, which were expressed at lower levels (FPKM<1.0), the remaining genes had a varied pattern of expression during the biosynthesis of cellulose. Of these genes, four *COBL*s (*GhCOBL3*, *GhCOBL8*, *GhCOBL10* and *GhCOBL16*) were expressed at all the tested stages and maintained higher accumulation levels; Four *COBL*s (*GhCOBL12*, *GhCOBL17*, *GhCOBL6* and *GhCOBL18*) and one *COBL*s (*GhCOBL15*) showed relatively high expression levels during the initiation stage of the fiber cell and elongation stage of the fiber cell respectively than that in other tested tissues; Three *COBL*s (*GhCOBL5*, *GhCOBL9* and *GhCOBL13*) displayed higher accumulation during the secondary cell wall biosynthesis stage. The expression patterns of *GhCOBL*s in different tissues and fiber development stages displayed a dynamic change in the life cycle of fiber cells and cellulose synthesis.

Functional prediction of two *COBL* genes and their roles in fiber quality

To elucidate the relationship between *COBL* genes and fiber quality traits, we focused on the two *COBL* genes (*GhCOBL9* and *GhCOBL13*) which preferentially expressed during the secondary cell wall stage of fiber development. *Cellulose synthases* (*CESAs*) were the key factors for cellulose biosynthesis [50] and there had been at least 15 *CESA* genes which had been classified into six groups in *G. raimondii* [30], so we investigated the expression correlations between the two *COBL*s and *CESAs* based on the RNA-Seq data during fiber developmental stages (S3 Table and Fig 6). The results showed that *GhCOBL9*/*GhCOBL13* had nearly 100% correlation with the expression levels of *GhCESA4A*/*B*, *GhCESA7A*/*B* and *GhCESA8B*.

To further unravel the potential roles of *GhCOBL9*/*GhCOBL13* in fiber quality traits, we cloned the A and D subgenome of these two genes in *G. hirsutum* acc. TM-1 and their orthologs in *G. barbadense* cv. Hai 7124 (S4 Table). Based on the nucleotide polymorphisms of *COBL9*/*COBL13* between TM-1 D_t and Hai7124 D_t, the SNP markers were developed successfully and confirmed using the SNP-PCR technology for *COBL9* D_t and "EcoTilling" for
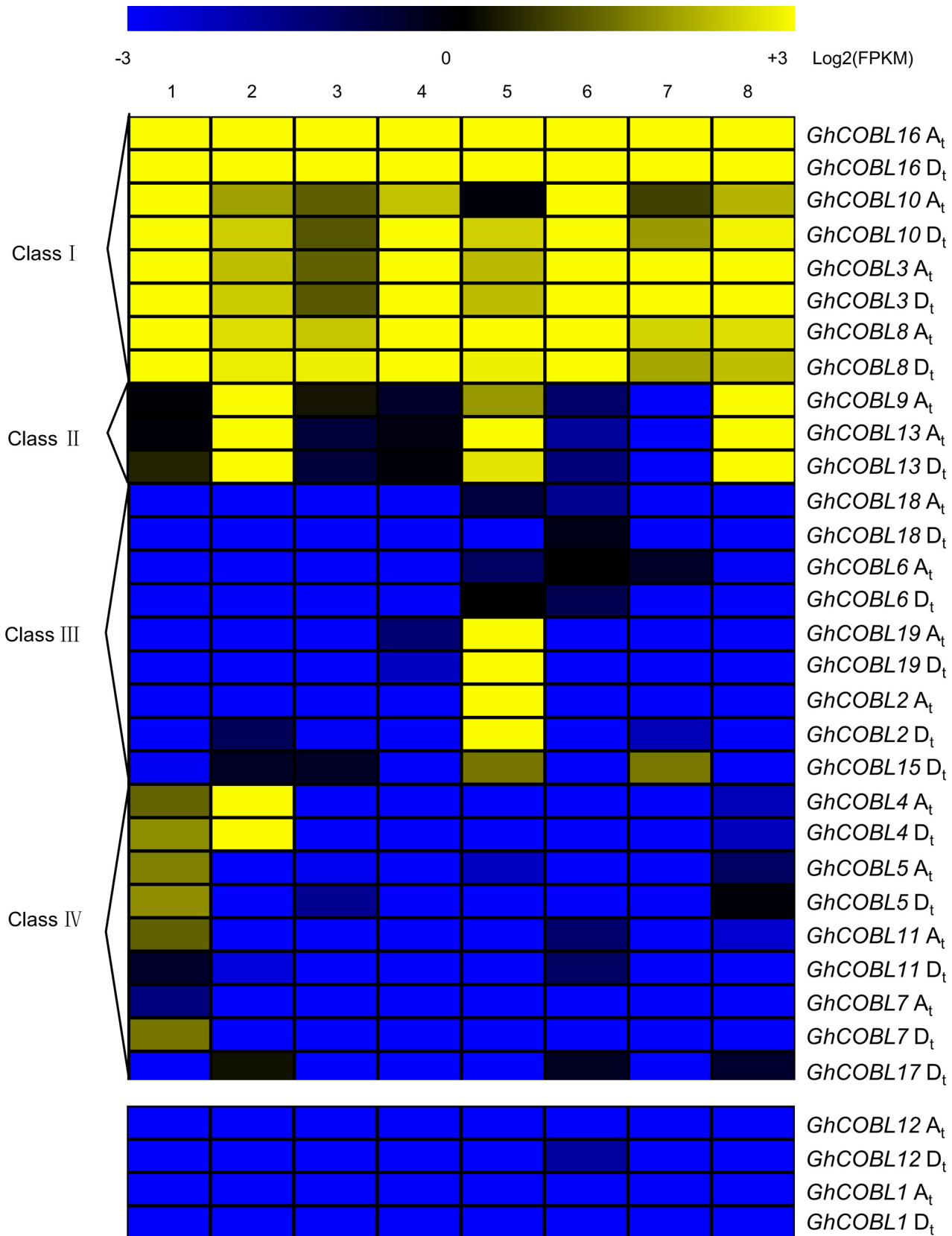
Fig 4. Heat map of the *COBLs* expression in different tissues of *G. hirsutum* acc. TM-1. The eight different tissues of *G. hirsutum* acc. TM-1 were involved here: 1: root; 2: stem; 3: leaf; 4: petal; 5: anther; 6: 0 DPA ovule; 7: 10 DPA fiber; 8: 20 DPA fiber. Log$_2$ (FPKM) indicated the different transcriptome profiling (FPKM: fragments per kb per million mapped reads). A$_t$ and D$_t$ were derived from the A-genome and D-genome progenitor in the tetraploid cotton.

doi:10.1371/journal.pone.0145725.g004

COBL13 D$_t$ respectively (S5 Table). For *COBL9* D$_t$, the detected nucleotide polymorphism site was involved in a premature termination codon in TM-1 (TAA) but was not in Hai7124 (AAA) (S4 Fig). The SNP site also existed the similar difference within the *G. hirsutum* accessions (S1 Table) and the allele type "AAA" showed the favorable phenotypes related to the fiber quality traits such as FL, FM, FU ($p < 0.01$) and FS, FE ($p < 0.05$), especially with the longer and thinner fiber (Table 3). While for *COBL13* D$_t$, the five detected SNPs were existed just between the *G. hirsutum* and *G. barbadense* but not within the *G. hirsutum* accessions (S1 Table; S4 Fig). Since the nucleotide polymorphism sites of *COBL9* and *COBL13* existed between TM-1 and Hai7124, we detected the expression levels of these two *COBLs* both in TM-1 and Hai7124 (Fig 7). *COBL9* A$_t$ and *COBL13* A$_t$ showed a high level accumulation from 10 DPA to 24 DPA both in TM-1 and Hai7124, even though the expression peak detected at 20 DPA in TM-1 but 24 DPA in Hai7124. However, compared to TM-1, *COBL9* D$_t$ and *COBL13* D$_t$ had an increasing accumulation from 10 DPA to 24 DPA in Hai7124 with peak at 24 DPA. In particular, *COBL9* D$_t$ in Hai7124 displayed a high level accumulation, whereas the *COBL9* D$_t$ in TM-1 was almost undetectable from 10 DPA to 24 DPA, which was consistent with significant correlation between *COBL9* D$_t$ and fiber quality traits. As for the contribution of *COBL9*/*COBL13* from the A-subgenome to fiber qualities, we would further investigate in the future study.

QTLs related to fiber quality traits were also employed to predict the roles of *GhCOBL9* and *GhCOBL13*. By aligning the genome scaffolds in *G. hirsutum* acc. TM-1, *GhCOBL9* D$_t$ and *GhCOBL13* D$_t$ were mapped on the D8 and D11 chromosomes. Therefore, we further searched the QTLs flanking the target sites with ± 5 Mb intervals. As in Fig 8, both *GhCOBL9* D$_t$ and *GhCOBL13* D$_t$ were co-localized with the QTLs involved in several cotton fiber quality traits [51–59].

Discussion

COBL gene family was highly conserved in different species

The *COBRA-Like* gene family shared the CBM motif, CCVS motif, an N-terminal signal peptide and highly hydrophobic C terminus and had been identified in four species. These *COBL* genes could be classified into two groups Group I and Group II. In this study we systematically identified *COBL* members in three sequenced cotton species: 19 in *G. raimondii*, 18 in *G. arboreum* and 33 in *G. hirsutum* acc. TM-1. Consistent with other species, *COBL* family members in *Gossypium* were also classified into two groups, with Group II carrying an amino acid stretch of unknown function after the N terminus and a longer hydrophobic tail following the ω-site.

Phylogenetic analysis showed that group I and group II contained the *COBLs* both from monocots and dicots (Fig 2), indicating that the *COBL* family members were descendants of an ancient duplication that occurred even before the separation of monocots and dicots. However, numerous duplications occurred after the divergence of monocots and dicots with more members of the *COBL* family in dicots than in monocots. In general, there was high sequence consistency of *COBLs* between *G. raimondii* and *A. thaliana*, however some exceptions detected (Fig 2). For example, *GrCOBL5* and *GrCOBL7* had relatively distant evolutionary relationships with the *A. thaliana* branch in Group I, which might result from the additional duplication event that occurred in *Gossypium*, although *A. thaliana* and *Gossypium* both underwent the two common duplication events (β, α) [60–61]. Gene duplication in *Gossypium* might have
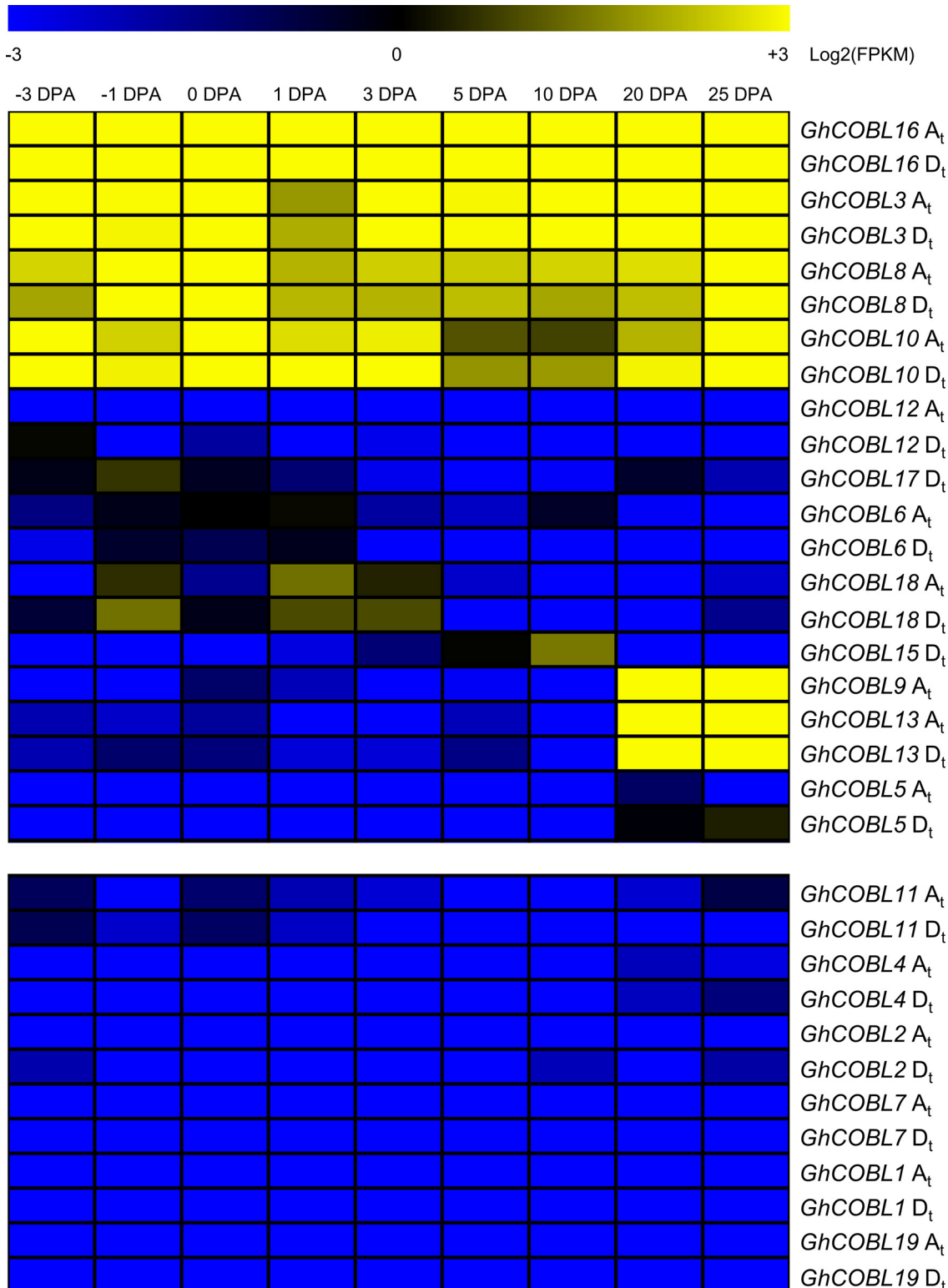
Fig 5. Heat map of the *COBLs* expression during the fiber developmental stages of *G. hirsutum* acc. TM-1. The fiber developmental stages involved in initiation (from -3 DPA to 0 DPA), elongation (from 0 DPA to 10 DPA) and secondary wall biosynthesis (from 20 DPA to 25 DPA) were sampled for detecting the expression levels of *COBL* genes during the fiber development. The relative expression levels were shown as Log_2 (FPKM). A_t and D_t were derived from the A-genome and D-genome progenitor in the tetraploid cotton.
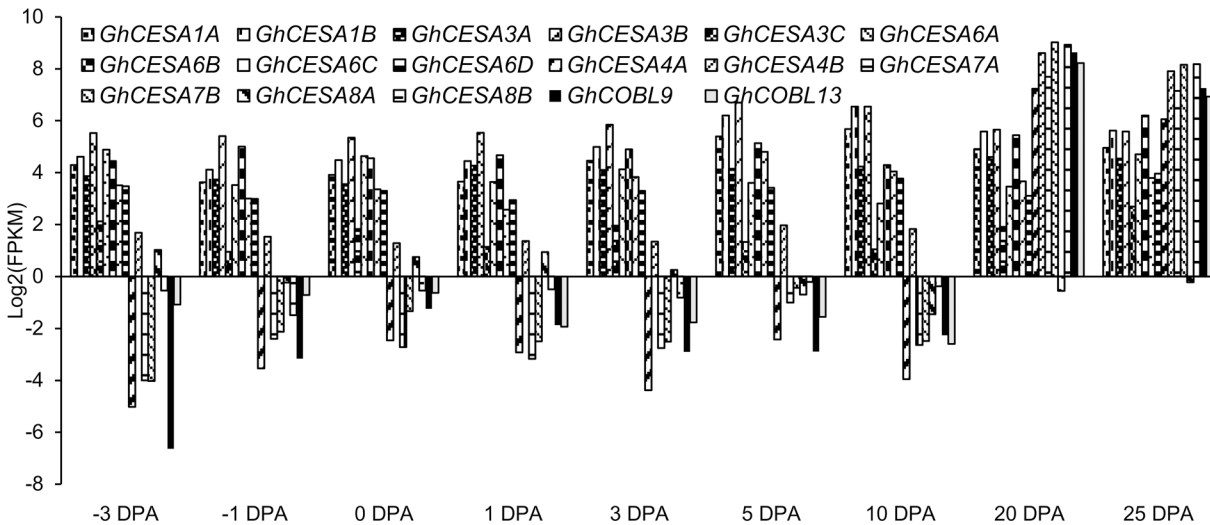
doi:10.1371/journal.pone.0145725.g005

Fig 6. Expression patterns of *GhCOBL9* and *GhCOBL13* with the *GhCESA* family genes in *G. hirsutum* acc. TM-1. The X axis indicated the different fiber development stages and the Y axis indicated the expression levels.

played a crucial role in driving evolutionary novelty and increasing adaptation to new environments by functional diversification.

COBL gene family showed highly functional diversity in different tissues of *Gossypium*

In the study, *COBL* genes in *G. hirsutum* acc. TM-1 were found to have diverse expression patterns in vegetative tissues, floral tissues and developing fibers. As an example, *GhCOBL2* and *GhCOBL19* exhibited a unique and high expression pattern in the anther; while *GhCOBL4* showed preferential expression in the stem (Fig 4). These results implied *COBL* genes played diverse functions in different cotton tissues. Actually, the diverse roles of *COBLs* had been elucidated in other plants. For instance, in *A. thaliana*, the functions of *COBL* genes were focused on two aspects. One represented by *AtCOBRA* was considered to have a significant impact on the orientation of cell expansion [62] and the other represented by *AtCOBL4* was largely responsible for cellulose crystallinity status and secondary cell wall reconstruction [19]. In

Table 3. Correlation analysis of the allele types of *COBL9* D_t with fiber quality traits [a].

Gene Name	Fiber quality traits[b]	Anyang			Kuerle			Nanjing		
		2007	2008	2009	2007	2008	2009	2007	2008	2009
COBL9	FL(mm)	28.77/29.79**	29.69/30.02	27.85/28.25	29.09/29.52	29.72/30.53**	28.70/29.04	29.20/29.75*	28.95/29.10	28.38/29.07*
(TAA/AAA)	FS(cN/tex)	28.44/28.55	28.30/28.39	28.00/28.18	28.27/28.59	29.23/29.78	28.74/29.15	29.67/29.90	28.67/29.50*	28.29/28.67
	FM	4.66/4.81**	5.03/5.17*	4.41/4.61*	4.09/4.11	4.33/4.53**	4.41/4.68**	4.46/4.63*	4.13/4.30	4.84/4.89
	FE(%)	6.39/6.30	6.40/6.43	7.55/7.47	6.27/6.21	6.49/6.55	6.71/6.75*	6.26/6.24	6.33/6.36	6.63/6.66
	FU(%)	84.07/85.07**	84.51/85.01	83.61/83.92	83.55/83.84	84.47/85.11**	85.02/85.21	83.64/84.20*	83.01/83.22	84.01/84.23

[a] Data of different fiber quality traits (the Ref./the Alt.) from natural population in *G. hirsutum* were analyzed with SPSS.

"*" and "**" denoted the differences at $P < 0.05$ and $P < 0.01$ between the Ref. alleles and the Alt. alleles respectively.

[b] FL: fiber length; FS: fiber strength; FM: fiber micronaire; FE: fiber elongation; FU: fiber uniformity.
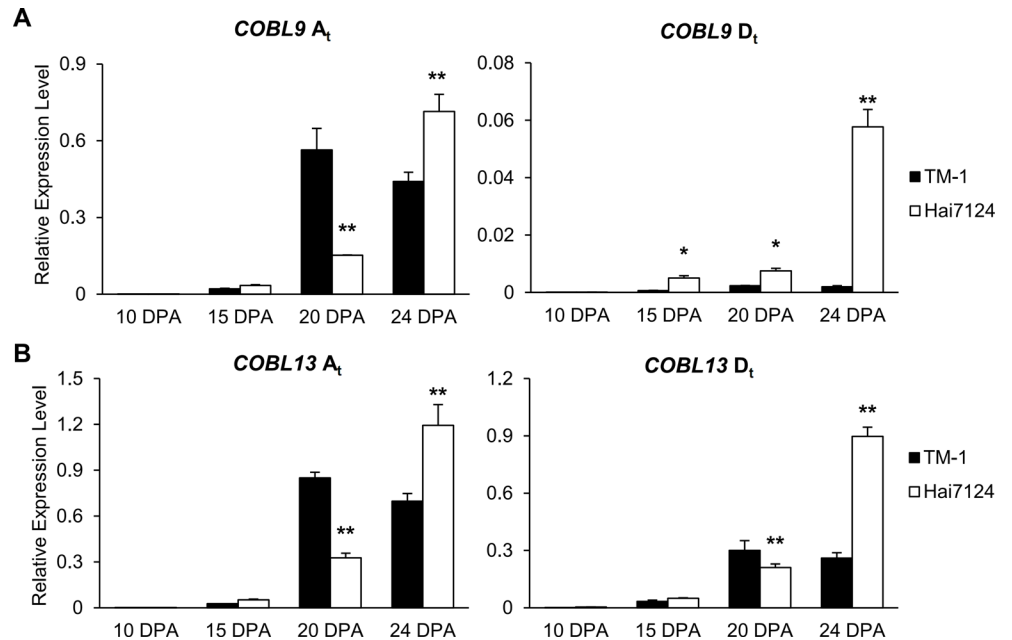
Fig 7. qRT-PCR analysis of *COBL9* (A) and *COBL13* (B) during the fiber development stages in TM-1 and Hai7124. The A_t and D_t were derived from A-subgenome and D-subgenome specific in tetraploid cotton species respectively. The Y axis indicated relative expression levels and the X axis indicated the different fiber development stages. "*" and "**" denoted the differences at $P < 0.05$ and $P < 0.01$ between TM-1 and Hai7124 respectively.

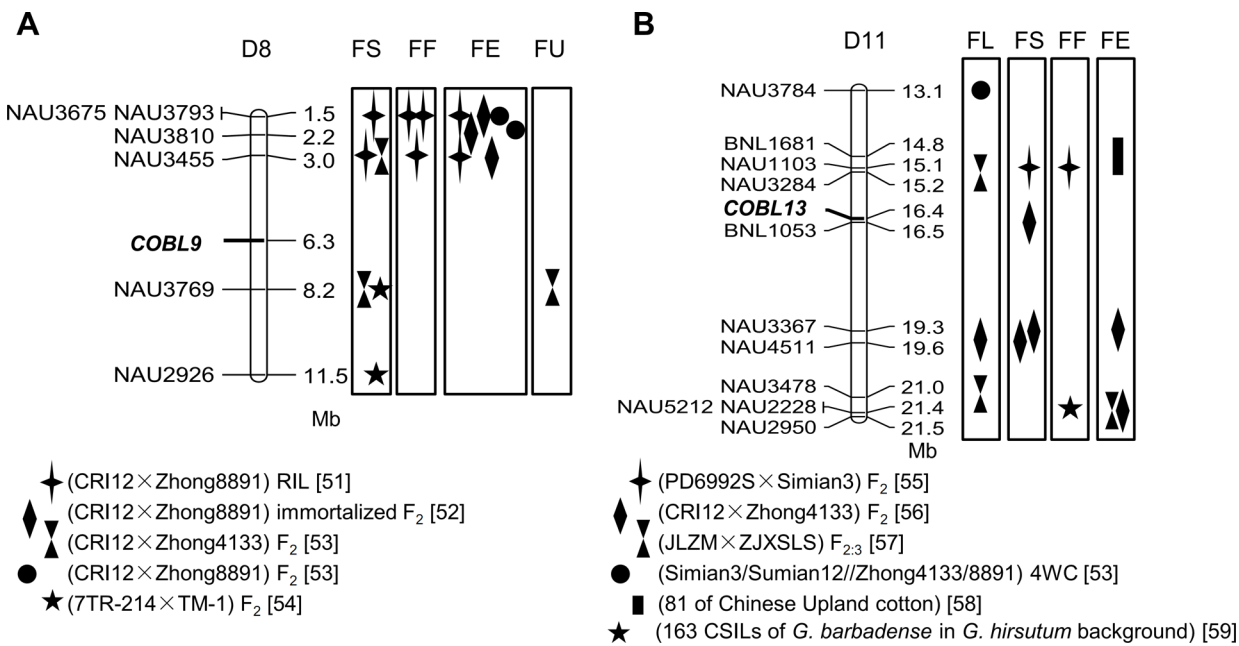doi:10.1371/journal.pone.0145725.g007

Fig 8. Physical integration of *GhCOBL9* and *GhCOBL13* with QTLs associated with fiber quality traits. Fiber quality traits included fiber length (FL), fiber strength (FS), fiber micronaire (FM), fiber elongation (FE) and fiber uniformity (FU). Different symbols represented the QTLs with the references followed.

doi:10.1371/journal.pone.0145725.g008

addition, it had been suggested that *OsBC1L* genes, the orthologs of *AtCOBL4*, performed a range of functions and participated in a various developmental processes in rice [29]. The functional divergence of *COBLs* could also be observed in the different developmental stages of fiber formation. For example, *GhCOBL15* exhibited high transcript level in 10 DPA fibers but was low in 20 and 25 DPA fibers. Conversely, *GhCOBL9* and *GhCOBL13* were highly expressed in 20 and 25 DPA fibers but were low in 10 DPA fibers ([Fig 6](#)). These results indicated *GhCOBL15* and *GhCOBL9/GhCOBL13* might have the different functions in fiber developmental process. The different expression pattern of *COBL* genes could be indicative of their diverse functions in different tissues or developmental processes.

>Interestingly, the functions of *COBLs* might be different from their phylogenetic clusters. We observed that the *GhCOBLs* with the differential expression patterns might come from the same phylogenetic group and the *GhCOBLs* with the similar expression patterns also might belong to the different phylogenetic groups (Figs [4](#) and [5](#)). For example, four *GhCOBL* genes, *GhCOBL8* and *GhCOBL16* in group I, and *GhCOBL3* and *GhCOBL10* in groupII, showed similar expression patterns. Besides, *AtCOBRA* was expressed in most tissues in *A. thaliana*, while *cobra* mutant plants only had an altered phenotype in the root. *GrCOBL8* and *GrCOBL16* were clustered in the same branch as *AtCOBRA* and were preferentially expressed in a variety of tissues. Their functional mechanisms in cotton remained to be verified.

GhCOBL9 and *GhCOBL13* potentially affected the cotton fiber quality

Cotton was an important raw material for the textile industry worldwide. During the development of cotton fiber, the deposition of cellulose determined the thickness of cell wall, which was the structural basis for fiber strength. Based on the reasons below, we could conclude that *GhCOBL9* and *GhCOBL13* play important roles during the stages of fiber formation.

Firstly, *GhCOBL9* and *GhCOBL13* were much higher expressed in fibers during the stages of secondary cell wall biosynthesis (Figs [4](#) and [5](#)) and had a significantly high coexpression pattern with *GhCESA4*, *GhCESA7* and *GhCESA8* ([Fig 6](#)), which were involved in cellulose formation of secondary cell wall [63]. The functional roles of both *GhCOBL9* and *GhCOBL13* orthologs in other species had been revealed previously. *irx6* (a T-DNA mutant of *AtCOBL4*) in *A. thaliana* exhibited abnormal secondary wall thickening and reduced stem strength [26] similar to the effect of *BC1* in *O. sativa* [27] and *BK2* in *Z. mays* [28]. It was reasonable that *GhCOBL9* and *GhCOBL13* which showed high amino acid sequence similarities to *AtCOBL4* might play the similar role in cotton fiber secondary cell wall thickening.

Secondly, correlation analysis revealed that nucleotide polymorphism of *GhCOBL9* D_t, similar to Hai7124 lineage, had significantly positive correlations with several fiber quality traits, seemed bring the favorable fiber quality in *G. hirsutum* ([Table 3](#)). Although no nucleotide polymorphism sites were detected for *GhCOBL13* within *G. hirsutum* accessions, both of these two *COBLs* had higher expression levels in Hai7124 which showed the superior fiber quality than in TM-1 ([Fig 7](#)). Furthermore, it was demonstrated that *GhCOBL9* and *GhCOBL13* were co-located with the QTLs related to fiber quality reported previously ([Fig 8](#)). These findings collectively indicated *GhCOBL9* and *GhCOBL13* exerted an important role in fiber quality. Since modern agricultural mechanization requires higher fiber qualities in cotton, it is reasonable to make full use of *GhCOBL9* and *GhCOBL13* for improving the fiber quality and enhancing the durability of textiles by transgenic technology or molecular marker-assisted breeding.

Supporting Information

S1 Fig. Multiple sequence alignments of 19 *COBL* members in *G. raimondii* and *AtCOBRA*, *AtCOBL7* in *A. thaliana*. Multiple sequence alignments were carried out with ClustalX 1.83.

The conserved motifs were marked by the boxes with different colors and the ω-sites were showed in red font. "*" denoted the aromatic amino acids in the CBM region.
(TIF)

S2 Fig. Expression patterns of *COBL* members based on qRT-PCR analysis of *G. hirsutum* acc. TM-1. The X axis indicated the different tissues and organs of *G. hirsutum* acc. TM-1 and the Y axis indicated relative expression levels of *GhCOBL* members. The cotton histone3 (AF024716) gene was used as the reference gene and the error bars were calculated based on three biological replicates using standard deviation (SD). 1: root; 2: stem; 3: leave; 4: petal; 5: anther; 6: 0DPA ovule; 7: 10DPA fiber; 8: 20DPA fiber.
(TIF)

S3 Fig. Correlation analysis of expression pattern between the FPKM of expression profiling and relative expression level of qRT-PCR for *GhCOBL* members. The X axis indicated the FPKM of expression profiling and the Y axis indicated relative expression level of qRT-PCR.
(TIF)

S4 Fig. The different PCR fragments and sequencing analysis of the polymorphic sites between the two different allele types for *COBL9* and *COBL13* respectively. A and B: Distinct fragments of *GhCOBL9* and *GhCOBL13* in the two different allele types revealed by denaturing gel electrophoresis via SNP-PCR (A) and EcoTilling analysis (B). "1–12" stand for the randomly selected individuals in *G. hirsutum* (including 1: 70-29-5, 2: Bao6716, 3: BaZhou5628, 4: ChangRong67-12, 5: Coker139, 6: GP137, 7: Ji91-12, 8: Zhong4612YaH, 9; HuBei-SongZiDaLing, 10: Zhong507145, 11: ZhongZhiBD89 and 12: ZhongZi10Hao) and "13–16" stand for the four lines in *G. barbadense* (including 13:E24-33891, 14: E24-33892, 15: Hai7124 and 16: Yinzi6022). "*" denoted that each DNA sample was mixed and hybridized with TM-1 (in 1: 1 ratio). C and D: Polymorphic sites of *GhCOBL9* (C) and *GhCOBL13* (D) were subsequently confirmed by sequencing and nucleotide polymorphisms were marked in orange shadows in each site.
(TIF)

S1 Table. Information on 289 accessions used for correlation analysis of the targeted genes with fiber quality traits.
(XLS)

S2 Table. Primers information for PCR amplification, qRT-PCR and EcoTilling analysis of *COBL* members. The R' primer of EcoTilling was to conduct the second PCR amplification for amplifing the density and specificity.
(XLS)

S3 Table. Pearson correlation coefficients of expression pattern between *GhCOBLs* and *GhCESAs* in different fiber developmental stages. The pearson correlation coefficient: r>0: positive correlation; r<0: negative correlation; r = 0: no linear correlation. The genes in yellow shadow have the lowest expression level in all tested tissues in fiber developmental stages.
(XLS)

S4 Table. Sequence information on *COBL9* and *COBL13* in cotton.
(XLS)

S5 Table. Nucleotide polymorphism of *COBL9* and *COBL13* in cotton.
(XLS)

Author Contributions

Conceived and designed the experiments: WZG. Performed the experiments: ELN XGS JHB YDZ XMD. Analyzed the data: ELN CZC CPC WZG. Contributed reagents/materials/analysis tools: WZG. Wrote the paper: ELN XGS WZG.

References

1. Beeckman T, Przemeck GK, Stamatiou G, Lau R, Terryn N, De Rycke R, et al. Genetic complexity of cellulose synthase a gene function in *Arabidopsis* embryogenesis. Plant Physiol. 2002; 130(4):1883–1893. PMID: 12481071

2. Ellis C, Karafyllidis I, Wasternack C, Turner JG. The *Arabidopsis* mutant *cev1* links cell wall signaling to jasmonate and ethylene responses. Plant Cell. 2002; 14(7):1557–1566. PMID: 12119374

3. MacMillan CP, Mansfield SD, Stachurski ZH, Evans R, Southerton SG. Fasciclin-like arabinogalactan proteins: specialization for stem biomechanics and cell wall architecture in Arabidopsis and Eucalyptus. Plant J. 2010; 62(4):689–703. doi: 10.1111/j.1365-313X.2010.04181.x PMID: 20202165

4. Mitsuda N, Seki M, Shinozaki K, Ohme-Takagi M. The NAC transcription factors NST1 and NST2 of Arabidopsis regulate secondary wall thickenings and are required for anther dehiscence. Plant Cell. 2005; 17(11):2993–3006. PMID: 16214898

5. Nicol F, His I, Jauneau A, Vernhettes S, Canut H, Hofte H. A plasma membrane-bound putative endo-1,4-beta-D-glucanase is required for normal wall assembly and cell elongation in Arabidopsis. EMBO J. 1998; 17(19):5563–5576. PMID: 9755157

6. Sánchez-Rodríguez C, Estévez JM, Llorente F, Hernández-Blanco C, Jordá L, Pagán I, et al. The ERECTA receptor-like kinase regulates cell wall-mediated resistance to pathogens in *Arabidopsis thaliana*. Mol Plant Microbe Interact. 2009; 22(8):953–63. doi: 10.1094/MPMI-22-8-0953 PMID: 19589071

7. Taylor NG, Howells RM, Huttly AK, Vickers K, Turner SR. Interactions among three distinct CesA proteins essential for cellulose synthesis. Proc Natl Acad Sci USA. 2003; 100(3):1450–1455. PMID: 12538856

8. Taylor NG, Laurie S, Turner SR. Multiple cellulose synthase catalytic subunits are required for cellulose synthesis in Arabidopsis. Plant Cell. 2000; 12(12):2529–2540. PMID: 11148295

9. Zhong R, Richardson EA, Ye ZH. The MYB46 transcription factor is a direct target of SND1 and regulates secondary wall biosynthesis in Arabidopsis. Plant Cell. 2007; 19(9):2776–2792. PMID: 17890373

10. Haigler CH, Betancur L, Stiff MR, Tuttle JR. Cotton fiber: a powerful single-cell model for cell wall and cellulose research. Front Plant Sci. 2012; 3:104. doi: 10.3389/fpls.2012.00104 PMID: 22661979

11. Kim HJ, Triplett BA. Cotton fiber growth in planta and in vitro: models for plant cell elongation and cell wall biogenesis. Plant Physiol. 2001; 127(4):1361–1366. PMID: 11743074

12. Basara AS, Malik CP. Development of cotton fiber. Inter Rev Cyto. 1984; 89:65–113.

13. Lee JJ, Woodward AW, Chen ZJ. Gene expression changes and early events in cotton fibre development. Ann Bot. 2007; 100(7):1391–1401. PMID: 17905721

14. Schubert AM, Benedict CR, Berlin JD. Cotton fiber development-kinetics of cell elongation and secondary wall thickening. Crop Sci. 1973; 13(6):704–709.

15. Endler A, Persson S. Cellulose synthases and synthesis in *Arabidopsis*. Mol Plant. 2011; 4:199–211. doi: 10.1093/mp/ssq079 PMID: 21307367

16. Arpat AB, Waugh M, Sullivan JP, Gonzales M, Frisch D, Main D, et al. Functional genomics of cell elongation in developing cotton fibers. Plant Mol Biol. 2004; 54(6):911–929. PMID: 15604659

17. Roudier F, Schindelman G, DeSalle R, Benfey PN. The COBRA family of putative GPI–anchored proteins in Arabidopsis. a new fellowship in expansion. Plant Physiol. 2002; 130(2):538–548. PMID: 12376623

18. Dai XX, You CJ, Wang L, Chen GX, Zhang QF, Wu CY. Molecular characterization, expression pattern, and function analysis of the OsBC1L family in rice. Plant Mol Biol. 2009; 71(4–5):469–48119. doi: 10.1007/s11103-009-9537-3 PMID: 19688299

19. Brady SM, Song S, Dhugga KS, Rafalski JA, Benfey PN. Combining expression and comparative evolutionary analysis. The COBRA gene family. Plant Physiol. 2007; 143(1):172–187. PMID: 17098858

20. Cao Y, Tang XF, Giovannoni J, Xiao FM, Liu YS. Functional characterization of a tomato *COBRA-like* gene functioning in fruit development and ripening. BMC Plant Biol. 2012; 12(1):211.

21. Benfey PN, Linstead PJ, Roberts K, Schiefelbein JW, Hauser MT, Aeschabacher RA. Root development in *Arabidopsis*: four mutants with dramatically altered root morphogenesis. Development. 1993; 119(1):57–7014. PMID: 8275864

22. Hauser MT, Morikami A, Benfey. Conditional root expansion mutants of *Arabidopsis*. Development. 1995; 121(4):1237–1252. PMID: 7743935

23. Roudier F, Fernandez AG, Fujita M, Himmelspach R, Borner GH, Schindelman G, et al. *COBRA*, an Arabidopsis extracellular glycosyl-phosphatidyl inositol–anchored protein, specifically controls highly anisotropic expansion through its involvement in cellulose microfibrils orientation. Plant Cell. 2005; 17 (6):1749–1763. PMID: 15849274

24. Dai XX, You CJ, Chen GX, Li XH, Zhang QF, Wu CY. *OsBC1L4* encodes a *COBRA-like* protein that affects cellulose synthesis in rice. Plant Mol Biol. 2011; 75(4–5): 333–345. doi: 10.1007/s11103-011-9730-z PMID: 21264494

25. Hochholdinger F, Wen TJ, Zimmermann R, Chimot-Marolle P, da Costa e Silva O, Bruce W, et al. The maize (*Zea mays L*.) *roothairless 3* gene encodes a putative GPI-anchored, monocot-specific, COBRA-like protein that significantly affects grain yield. Plant J. 2008; 54(5): 888–898. doi: 10.1111/j.1365-313X.2008.03459.x PMID: 18298667

26. Brown DM, Zeef LAH, Ellis J, Goodacre R, Turner SR. Identification of novel genes in *Arabidopsis* involved in secondary cell wall formation using expression profiling and reverse genetics. Plant Cell. 2005; 17(8): 2281–2295. PMID: 15980264

27. Li YH, Qian Q, Zhou YH, Yan MX, Sun L, Zhang M, et al. *BRITTLE CULM1*, which encodes a COBRA-like protein, affects the mechanical properties of rice plants. Plant Cell. 2003; 15(9):2020–2031. PMID: 12953108

28. Ching A, Dhugga KS, Appenzeller L, Meeley R, Bourret TM, Howard RJ, et al. *Brittle stalk 2* encodes a putative glycosylphos phatidylinositol-anchored protein that affects mechanical strength of maize tissues by altering the composition and structure of secondary cell walls. Planta. 2006; 224(5):1174–1184. PMID: 16752131

29. Liu LF, Shang-Guan KK, Zhang BC, Liu XL, Yan MX, Zhang LJ, et al. Brittle Culm1, a COBRA-Like protein, functions in cellulose assembly through binding cellulose microfibrils, PLoS Genet. 2013; 9(8): e1003704. doi: 10.1371/journal.pgen.1003704 PMID: 23990797

30. Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, et al. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature. 2012; 492(7429):423–427. doi: 10.1038/nature11798 PMID: 23257886

31. Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, et al. Genome sequence of the cultivated cotton *Gossypium arboreum*. Nat Genet. 2014; 46(6):567–572. doi: 10.1038/ng.2987 PMID: 24836287

32. Zhang TZ, Hu Y, Jiang WK, Fang L, Guan XY, Chen JD, et al. Sequencing of allotetraploid cotton (*Gossypium hirsutum* acc. TM-1) provides a resource for fiber improvement. Nat Biotechnol. 2015; 33:531–537. doi: 10.1038/nbt.3207 PMID: 25893781

33. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011; 39:29–37.

34. Zhao L, Lv YD, Cai CP, Tong XC, Chen XD, Zhang W, et al. Toward allotetraploid cotton genome assembly: integration of a high-density molecular genetic linkage map with DNA sequence information. BMC Genomics. 2012; 13(1):539.

35. Lee TH, Tang HB, Wang XY, Paterson AH. PGDD: a database of gene and genome duplication in plants. Nucleic Acids Res. 2013; 41:D1152–1158. doi: 10.1093/nar/gks1104 PMID: 23180799

36. Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. Synteny and collinearity in plant genomes. Science. 2008; 320:486–488. doi: 10.1126/science.1153917 PMID: 18436778

37. Cai CP, Niu EL, Du H, Zhao L, Feng Y, Guo WZ. Genome-wide analysis of the WRKY transcription factor gene family in *Gossypium raimondii* and the expression of orthologs in cultivated tetraploid cotton. Crop J. 2014; 2(2–3):87–101.

38. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. 2003; 31(13):3784–3788. PMID: 12824418

39. Higgins DG, Thompson JD, Gibson TJ. Using CLUSTAL for multiple sequence alignments. Methods Enzymol. 1996; 266:383–402. PMID: 8743695

40. Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res. 2006; 34 Suppl 2:369–373.

41. Eisenhaber B, Bork P, Eisenhaber F. Prediction of potential GPI-modification sites in proprotein sequences. J Mol Biol. 1999; 292(3):741–758. PMID: 10497036

42. Paterson AH, Brubaker CL, Wendel JF. A rapid method for extraction of cotton (*Gossypium* spp.) genomic DNA suitable for RFLP or PCR analysis. Plant Mol Biol Rep. 1993; 11(2):122–127.

43. Jiang JX, Zhang TZ. Extraction of total RNA in cotton tissues with CTAB-acidic phenolic method. Cotton Sci. 2003; 15(3):166–167.

44. Drenkard E, Richter BG, Rozen S, Stutius LM, Angell NA, Mindrinos M, et al. A simple procedure for the analysis of single nucleotide polymorphisms facilitates map-based cloning in *Arabidopsis*. Plant Physiol. 2000; 124(4):1483–1492. PMID: 11115864

45. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\triangle\triangle CT}$ method. Methods. 2001; 25(4):402–408. PMID: 11846609

46. Xu YH, Wang JW, Wang S, Wang JY, Chen XY. Characterization of *GaWRKY1*, a cotton transcription factor that regulates the sesquiterpene synthase gene (+)-delta-cadinene synthase-A, Plant Physiol. 2004; 135(1):507–515 PMID: 15133151

47. Zeng CL, Wang GY, Wang JB, Yan GX, Chen BY, Xu K, et al. High-throughput discovery of chloroplast and mitochondrial DNA polymorphisms in brassicaceae species by ORG-EcoTILLING. PLoS One. 2012; 7(11):e47284. doi: 10.1371/journal.pone.0047284 PMID: 23185237

48. Comai L, Oung K, Till BJ, Reynolds SH. Efficient discovery of DNA polymorphisms in natural populations by Ecotilling. Plant J. 2004; 37(5):778–786. PMID: 14871304

49. Zhang J. Evolution by gene duplication: an update. Trends in Ecology & Evolution 2003; 18 (6): 292–8.

50. Richmond TA, Somerville CR. The cellulose synthase superfamily. Plant Physiol. 2000; 124(2):495–498. PMID: 11027699

51. Wang BH, Guo WZ, Zhu XF, Wu YT, Huang NT, Zhang TZ. QTL mapping of fiber quality in an elite hybrid derived–RILs in upland cotton. Euphytica. 2006; 152(3):367–378.

52. Wang BH, Wu YT, Guo WZ, Zhu XF, Huang NT, Zhang TZ. Genetic dissection of heterosis for fiber qualities in an elite cotton hybrid grown in second–generation. Crop Sci. 2007; 47:1384–1392.

53. Qin YS, Ye WX, Liu RZ, Zhang TZ, Guo WZ. QTL mapping for fiber quality properties in upland cotton (*Gossypium hirsutum* L.). Sci Agric Sin. 2009; 42(12):4145–4154.

54. Chen H, Neng Q, Guo WZ, Song QP, Li BC, Deng FJ, et al. Using three overlapped RILs to dissect genetically clustered QTL for fiber strength on chro.24 in upland cotton. Theor Appl Genet. 2009; 119 (4): 605–612. doi: 10.1007/s00122-009-1070-x PMID: 19495722

55. Shen XL, Guo WZ, Zhu XF, Yuan YL, Yu JZ, Kohel RJ, et al. Molecular mapping of QTLs for qualities in three diverse lines in upland cotton using SSR markers. Mol Breed. 2005; 15(2):169–181.

56. Yang C, Guo WZ, Zhang TZ. QTL mapping for resistance to *vertieillium* wilt, fiber quality and yield traits in upland cotton (*Gossypium hirsutum* L.). Mol Plant Breed. 2007; 5(6): 797–805.

57. Ma X, Ding Y, Zhou B, Guo W, Lv Y, Zhu X, et al. QTL mapping in A-genome diploid Asiatic cotton and their congruence analysis with AD-genome tetraploid cotton in genus *Gossypium*. J Genet Genomics. 2008; 35(12):751–62. doi: 10.1016/S1673-8527(08)60231-3 PMID: 19103431

58. Zhang TZ, Qian N, Zhu XF, Chen H, Wang S, Mei HX, et al. Variations and transmission of QTL alleles for yield and fiber qualities in upland cotton cultivars developed in China. PLoS One. 2013; 8(2): e57220. doi: 10.1371/journal.pone.0057220 PMID: 23468939

59. Wang P, Zhu YJ, Song XL, Cao ZB, Ding YZ, Liu BL, et al. Inheritance of long staple fiber quality traits of *Gossypium barbadense* in *G. hirsutum* background using CSILs. Theor Appl Genet. 2012; 124 (8):1415–28. doi: 10.1007/s00122-012-1797-7 PMID: 22297564

60. Zhang HB, Li YN, Wang BH, Chee PW. Recent advances in cotton genomics. Int J Plant Genomics. 2008; 2008:742304. doi: 10.1155/2008/742304 PMID: 18288253

61. Renny-Byfield S, Gallagher JP, Grover CE, Szadkowski E, Page JT, Udall JA, et al. Ancient gene duplicates in *Gossypium* (cotton) exhibit near-complete expression divergence. Genome Biol Evol. 2014; 6 (3):559–571. doi: 10.1093/gbe/evu037 PMID: 24558256

62. Schindelman G, Morikami A, Jung J, Baskin TI, Carpita NC, Derbyshire P, et al. COBRA encodes a putative GPI-anchored protein, which is polarly localized and necessary for oriented cell expansion in *Arabidopsis*. Genes & Dev. 2001; 5(9):1115–1127.

63. Betancur, Singh B, Rapp RA, Wendel JF, Marks MD. Phylogenetically distinct cellulose synthase genes support secondary wall thickening in *Arabidopsis* shoot trichomes and cotton fiber. J Integr Plant Biol. 2010; 52(2):205–220. doi: 10.1111/j.1744-7909.2010.00934.x PMID: 20377682