Journal of Cotton Research

RESEARCH Open Access

Genome-wide characterization of the Rho family in cotton provides insights into fiber development

HE Man[1,2], WANG Xingfen[1], LIU Shang[2], CHENG Hailiang[2,3], ZUO Dongyun[2], WANG Qiaolian[2], LV Limin[2], ZHANG Youping[2] and SONG Guoli[2*]

Abstract

Background: Cotton is the source of natural fibers globally, fulfilling 90% of the textile industry's requirements. However, fiber development is a complex biological process comprising four stages. Fiber develops from a single cell, and cell elongation is a vital process in fiber development. Therefore, it is pertinent to understand and exploit mechanisms underlying cell elongation during fiber development. A previous report about cell division control protein 42 (CDC-42) with its key role in cell elongation in eukaryotes inspired us to explore its homologs Rho GTPases for understanding of cell elongation during cotton fiber development.

Result: We classified 2 066 Rho proteins from 8 *Gossypium* species into 5 and 8 groups within A and D sub-genomes, respectively. Asymmetric evolution of Rho members was observed among five tetraploids. Population fixation statistics between two short and long fiber genotypes identified highly diverged regions encompassing 34 *Rho* genes in *G. hirustum,* and 31 of them were retained through further validation by genome wide association analysis (GWAS). Moreover, a weighted gene co-expression network characterized genome-wide expression patteren of *Rho* genes based on previously published transcriptome data. Twenty *Rho* genes from five modules were identified as hub genes which were potentially related to fiber development. Interaction networks of 5 *Rho* genes based on transcriptional abundance and gene ontology (GO) enrichment emphasized the involvement of Rho in cell wall biosynthesis, fatty acid elongation, and other biological processes.

Conclusion: Our study characterized the Rho proteins in cotton, provided insights into the cell elongation of cotton fiber and potential application in cotton fiber improvement.

Keywords: Cotton fiber, Cell polarity, Rho family, Association analysis, WGCNA

Background

Cotton is the primary source of natural textile worldwide. Fiber development is one of the mostly investigated biological processes in cotton functional genomics. Cotton fiber originates from the ovule epidermis and develops into a long trichome (Qin and Zhu 2011). Fiber development can be categorized into four stages based on the morphology: initiation, elongation, secondary cell wall biosynthesis, and maturation. During 5–25 days post anthesis (DPA), fibers elongate vigorously with a rapid cell expansion rate (> 2 mmd^{-1}) (Kim and Triplett 2001; Ji 2003). Moreover, cotton fiber is a suitable model for understanding single-cell elongation (Haigler et al. 2012). Several studies have characterized genetic mechanisms underlying fiber elongation (Zhang et al. 2017; Thyssen et al. 2017; Zhang et al. 2016; Stiff and Haigler 2016;

*Correspondence: sglzms@163.com

[2] Institute of Cotton Research of Chinese Academy of Agricultural Sciences, Anyang 455000, Henan, China
Full list of author information is available at the end of the article

Zeng et al. 2019; Wang et al. 2010). Several protein-coding genes involved in various metabolic processes were systematically identified for their potential roles in fiber development, such as ethylene biosynthesis, long-chain fatty acids biosynthesis, and auxin transport. However, detailed regulatory mechanisms remain to be elucidated, such as the synthesis of long fatty acid and cell wall components during fiber elongation.

We got inspiration from researches on single-cell elongation in eukaryotes to explore and exploit critical factors influencing cell elongation/fiber elongation. CDC-42, a highly conserved plasma membrane-associated small Rho-family GTPase, harboring an active GTP-bound state and an inactive GDP-bound state to promote the cell elongation in eukaryotes (Etienne 2004; Farhan and Hsu 2016; Mack and Georgiou 2014). CDC-42 regulates the extension and maintenance of filopodia in neurons of *Mus musculus* (Sakabe et al. 2012). CDC-42 also mediates the filopodia initiation in human cells (Gauthier et al. 2004). Moreover, several reports have verified the active role of CDC-42 in cell elongation and cell shape in different species (Murakoshi et al. 2011; Galic et al. 2014). For instance, Rho GTPases, homologs of CDC-42 in *Arabidopsis thaliana*, are involved in vesicle trafficking and protein transport, implying the role of GTPase in cell membrane trafficking (Anai et al. 1994; Koh et al. 2009). Based on the above-mentioned reports, it can be inferred that CDC-42 is a highly conserved functional protein among different species. Therefore, we hypothesize that the homologs of CDC-42 might play a significant role in cotton fiber development.

Delmer et al. (1995) cloned several *Rho* genes encoding small GTPase and explored their roles in fiber development. Recent advances in omics provide data platform for rapid identification and further characterization of complex regulatory mechanisms underlying fiber development (Huang et al. 2020; Chen et al. 2020; Wang et al. 2019). In this study, We adapted a systematic approach and utilized the previously published genomic data (Ma et al. 2018) and transcriptome data (Qin et al. 2019) to identify and characterize the potential functions of *Rho* genes in cotton.

Results

Rho family members in cotton

We identified Rho proteins among eight *Gossypium* species, including three diploids and five tetraploids, after obtained Rho HMM (Hiden Markov Model) file from Pfam platform (Table 1). We grouped Rho protein from tetraploid cotton according to the chromosome distribution on A sub-genome or D sub-genome. These Rho proteins were annotated using Uniprot database (Additional file 1: Table S1). Annotation information suggested that the number of Rho proteins in *Gossypium* species ranged from 145 to 332. The longest protein sequence (1 101 aa) was identified in *G. arboreum*, while the shortest Rho proteins (644 aa or 645 aa) were identified in 4 tetraploids, i.e., *G. hirsutum, G. tometosum, G. mustelinum,* and *G. darwini* (Additional file 2: Table S2). We further compared the difference between short and long Rho proteins, whereas 13 were selected as the small Rho proteins with length ranged from 644 aa to 679 aa. *Gar05G18660*, encoding the longest protein of 1 101 aa, had three extra domains in an unaligned region, i.e., Glyco_transf_8, PPR (pentatricopeptide repeat), and PPR_2. Glyco_transf_8 was found in *G. arboreum*, *G. herbaceum,* and all analyzed tetraploid species except *G. raimondii*. Interestingly, all of the Rho proteins containing Glyco_transf_8 domain in all analyzed tetraploids were in A_t sub-genomes, indicating Glyco_transf_8 in allotetraploids was inherited from diploids of A genome and had not been duplicated in D_t sub-genome during the evolution of tetraploids. According to Pfam database, Glyco_transf_8 (PF01501)is involved in glucose transferring. Considering that both tetraploids and diploids of A genome had longer fiber than *G. raimondii*, we speculated that Rho proteins with Glyco_transf_8 may contribute to the fiber elongation. Moreover, PPR and PPR_2 were only identified in Gar05G18660. According to the annotation in Pfam (PF01535), the function of PPR domain is still unclear. Other Rho proteins had no domain in unaligned regions. Variations in protein length and domain distribution implied high sequence variation among Rho family members.

Table 1 The number of *Rho* genes in 8 cotton species

Species	Subgenome	Number
G. herbaceum	A_1	157
G. arboreum	A_2	144
G. raimondii	D_5	164
G. hirustum	A_{t1}	138
G. barbadense	A_{t2}	168
G. tomentosum	A_{t3}	169
G. mustelinum	A_{t4}	165
G. darwinii	A_{t5}	170
G. hirustum	D_{t1}	143
G. barbadense	D_{t2}	159
G. tomentosum	D_{t3}	164
G. mustelinum	D_{t4}	163
G. darwinii	D_{t5}	162

For tetrapploids, the number of Rho family members within A and D sub-genomes were calculated seperately

Phylogeny analysis of Rho members in cotton

We performed phylogenetic analysis to understand the diversity among Rho proteins. Two phylogenetic trees were built according to A_t and D_t sub-genomes, respectively (Fig. 1a, b). Rho proteins from two species of A genomes (*G. herbaceum and G. arboreum*) and five allotetraploids of A sub-genomes (A_{t1}-A_{t5}) were classified into five groups, while members from diploid species of D genome (*G. raimondii*) and 5 allotetraploids of D sub-genomes (D_{t1}-D_{t5}) were classified into eight groups. We evaluated sequence conservation of *Rho* genes in eight cotton species by identical site rate (ISR %) (Fig. 2). By comparing ISR of Rho proteins among different cotton species, we noticed that Rho members in *G. herbaceum* (A_1-A_1) had a lower ISR compared with *G. arboreum* (A_2-A_2) (Fig. 2a, b).

Furthermore, we found that inter-genomic conservation (A_1-$A_{t1\text{-}t5}$) was higher than intra-genomic conservation (A_1-A_1) among *Rho* genes in species of A_1-genome (higher ISR), indicating higher similarity between A_1 and A sub-genomes from 5 allotetraploids (Fig. 2a, b; Additional file 3: Table S3). These results indicated that Rho members in *G. herbaceum* remained conserved with less sequence diversity and were consistent with the previous conclusion that *G. herbaceum* was likely a donor of A genome in allotetraploid (Huang et al. 2020). Moreover, Rho protein sequences in *G. raimondii* were more conserved with those from D sub-genomes in *G. hirustum* (D_5-D_{t1}; $P = 1e\text{-}6$) and *G. tomentosum* (D_5-D_{t3}; $P = 0.03$) (Fig. 2c; Additional file 3: Table S3). Phylogenetic analysis suggested more similarity between *Rho* genes from the D_5 genome and those from D_{t1} and D_{t3}, indicating *Rho* genes had a potential contribution to the independent evolution of five tetraploids.

Fiber length-related association analysis of Rho members

To check up the potential role of Rho GTPases in fiber development, we performed association analysis of fiber length based on the previous sequenced data (Ma et al. 2018). Thirty cultivars with the longest fiber and 30 cultivars with the shortest fiber were selected among the 419 resequenced dataset to identify the candidate loci associated with fiber length. By *t*-test, a significant divergence of fiber length was observed between 2 groups ($P = 1.618e\text{-}30$) (Fig. 3a). To detect candidate genomic regions, genomic sequences were divided into windows (windowsize = 50 000 bp, stepsize = 5 000 bp) and *F*st of these windows were calculated (Fig. 3b). Finally, the top 5% *F*st from 44 626 windows were selected to detect potential genes associated with fiber length. A total of 6 885 genes overlapped with the selected windows were extracted as candidate genes (Additional file 4: Table S4 and Additional file 5: Table S5). The functions of 6 885 genes were initially checked by GO enrichment analysis. We found that genes involved in basic biological processes such as RNA binding (GO:0003723), mRNA splicing (GO:0000398), and several processes related to glucose metabolism, for instance, galactose metabolic process (GO:0006012), UDP-glucose 4-epimerase activity (GO:0003978) were enriched. Interestingly, we noticed that GTPase activity (GO:0003924) was also enriched, indicating that the activities of GTPase could influence fiber length (Fig. 3c). All 34 *Rho* genes among the 6 885 candidate genes were selected for further analysis (Additional file 6: Table S6). Furthermore, we performed GWAS analysis on all 419 cultivars for fiber length and found that 31 *Rho* genes among above-mentioned 34 *Rho* genes had significant SNPs (Additional file 6: Table S6).

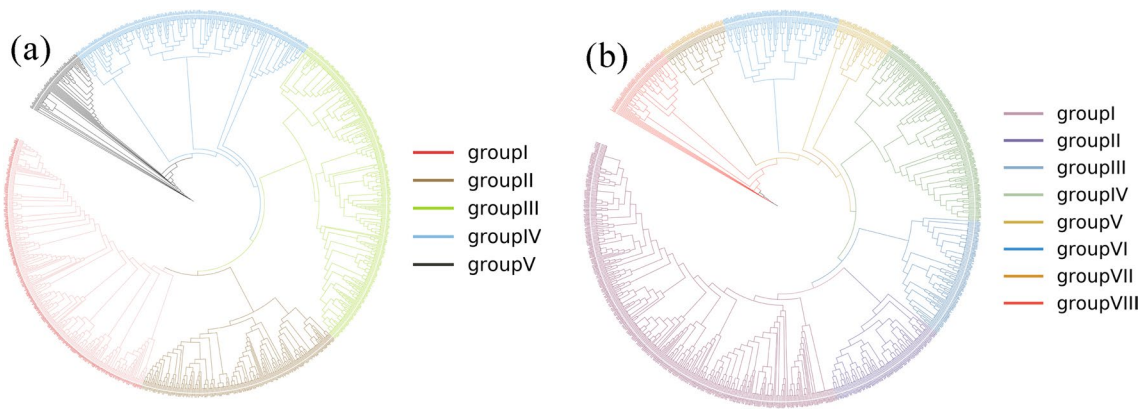

Fig. 1 The phylogenetic analysis of Rho members across 8 cotton species. **a** The phylogenetic tree of Rho members from 2 diploids of A genomes and 5 allotetraploids of A sub-genomes. **b** The phylogenetic tree of Rho members from 1 diploid of D genome and 5 allotetraploids of D sub-genomes

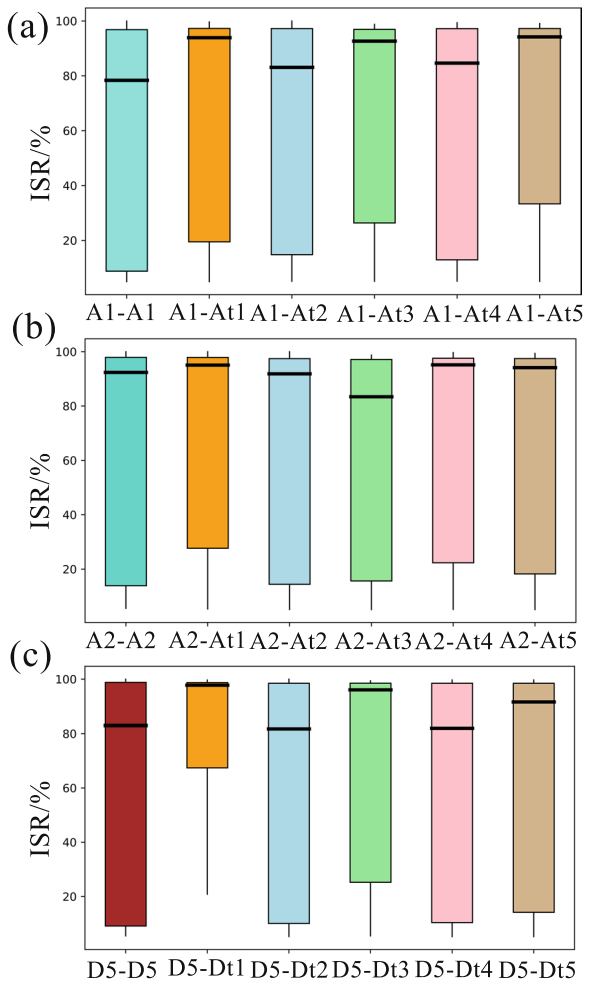

Fig. 2 Identical site rate (ISR) in Rho proteins across 8 cotton species. **a**. Identical site rate of Rho proteins from *G. herbaceum* and 5 tetraploids. A_1-A_1 means all Rho proteins from *G. herbaceum*. The A_1-A_t indicates all mixed Rho proteins from both *G. herbaceum* and the corresponding A_t sub-genome from tetraploid. **b** Identical site rate of Rho proteins from *G. arboreum* and 5 tetraploids. A_2-A_2 means all Rho proteins from *G. arboreum*. The A_2-A_t indicates all mixed Rho proteins from both *G. arboreum* and the corresponding A_t sub-genome from tetraploid. **c** Identical site rate of Rho proteins from *G. raimondii* and 5 tetraploids. D_5-D_5 means all Rho proteins from *G.raimondii*. The D_5-D_t indicates all mixed Rho proteins from both *G. raimondii* and the corresponding D_t sub-genome from tetraploid

Expression patterns of Rho members in transcriptome data

To evaluate the transcription abundance of 31 *Rho* genes, we analyzed the published transcriptome data. Raw transcriptome data of 8 samples were downloaded (Table 2) to perform transcriptome analysis. As a result, 8 678 and 298 differentially expressed genes (DEGs) from fiber and ovule groups, long fiber and short fiber groups were identified, respectively (Fig. 4a,b; Additional file 7: Table S7). We defined genes with the maximum TPM score larger than 5 as expressed genes. For 34 *Rho* candidate genes, 20 of them were found to be expressed. Among 20 expressed *Rho* genes, 10 of them were found in DEGs among the fiber, ovule groups (Fig. 4c, d and Additional file 8: Table S8). These 10 *Rho* genes with diverge expression patterns between fiber and ovule epidermis cells may participate in cell elongation.

WGCNA analysis of transcriptome data

Rho genes may regulate cell polarity with the involvement of other genes (Etienne 2004). To investigate the potential interactions of Rho proteins during fiber development, we conducted a weighted gene co-expression network analysis (WGCNA) (Fig. 5a). Genes with the transcription score smaller than 5 among all samples were trimmed, and finally 34 181 genes were retained for WGCNA analysis. The soft threshold of the network was set as 26, while *R* square was above 0.8, and mean connectivity was lower than 2 000, indicating that the network was scale-free (Fig. 5b). After network construction, genes were classified into 13 modules according to the expression profile (Fig. 5c).

Since the modules were generated according to expression patterns of genes, the association analysis between gene expression and phenotype data among each classified modules revealed potential functions of corresponding modules. The phenotype data of 8 samples were classified as fiber, ovule and long fiber, short fiber groups, respectively. We set 0.5 as the threshold of the *R* value of Pearson correlation to identify the module associated with fiber length phenotype (Fig. 5d). Among modules, pink, magenta, and tan modules were related to the long fiber trait, while the purple module was related to the short fiber trait. As for the fiber, ovule trait, we found blue and brown modules related to fiber development. On the other hand, pink, turquoise, purple, and tan modules were related to ovule tissue.

We checked the interaction networks within *Rho* genes and found that *Rho* genes in turquoise, blue, brown, green, and yellow modules showed potential interactions (Fig. 5e). To investigate the role of these modules, the genes with the top 10% membership were selected for GO enrichment analysis (Additional file 9: Table S9). Results of GO enrichment of blue module showed that fatty acid biosynthetic process (GO:0006633), microtubule (GO:0005874), and microtubule-based movement (GO:0007018) were enriched (Additional file 12: Fig. S1a). The other ovule-related brown module contained different GO terms, such as glucose metabolic process (GO:0006006), cellulose microfibril organization (GO:0010215), and cell growth (GO:0016049) (Additional file 12: Fig. S1b). Although various biological processes were enriched in blue and brown modules, most of the enriched GO terms were metabolic pathways.

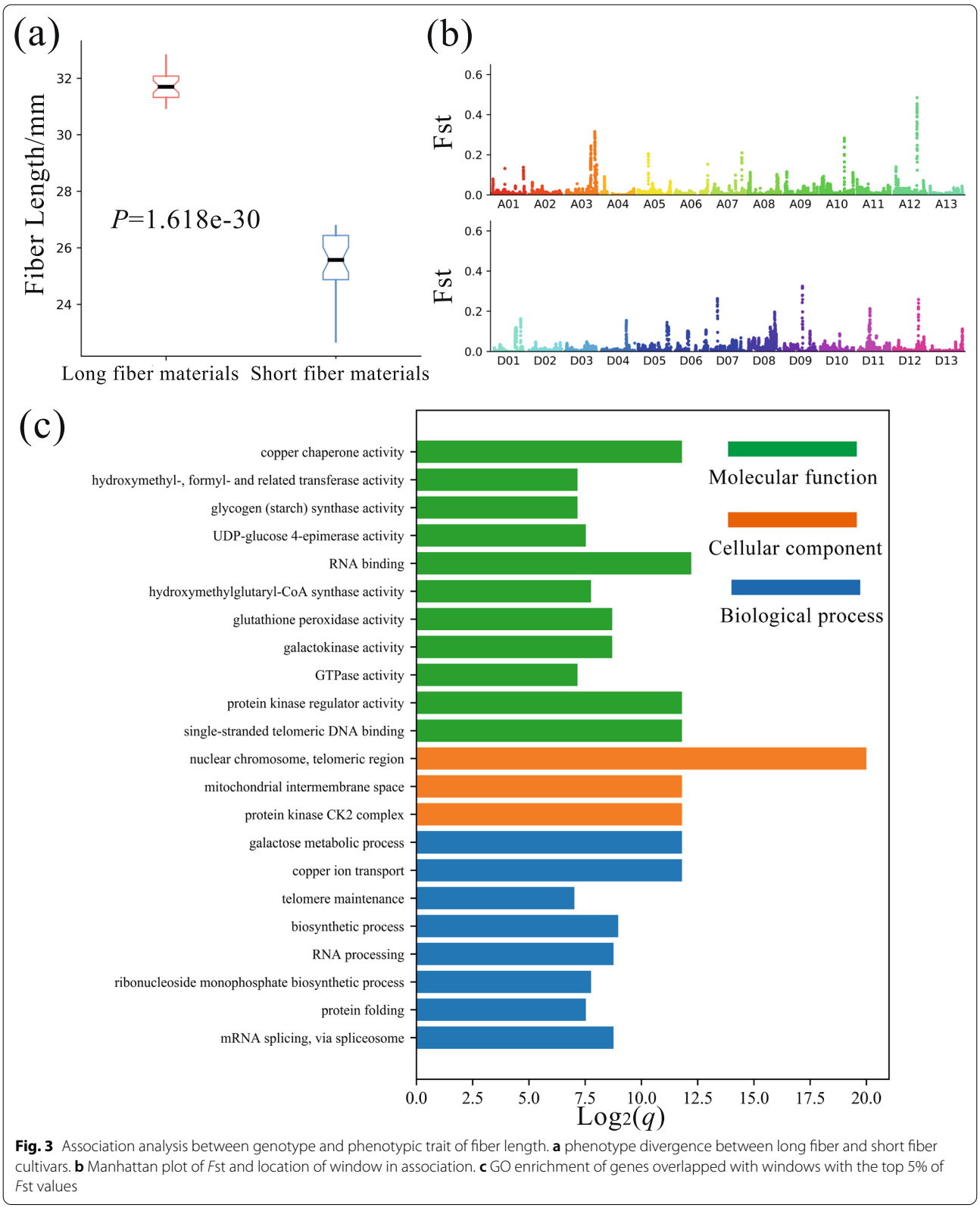

Fig. 3 Association analysis between genotype and phenotypic trait of fiber length. **a** phenotype divergence between long fiber and short fiber cultivars. **b** Manhattan plot of *F*st and location of window in association. **c** GO enrichment of genes overlapped with windows with the top 5% of *F*st values

Table 2 Transcriptome dataset used in this study

Data accession	Sample name	Time /DPA	Tissue	Phenotype
SRR6379580	69-6025-12	5	Fiber	Short fiber
SRR5992414	69-6025-12	0	Ovule	Short fiber
SRR6379578	69-6025-12	10	Fiber	Short fiber
SRR6379579	69-6025-12	10	Ovule	Short fiber
SRR6379574	601 long stapled cotton	10	Fiber	Long fiber
SRR6379575	601 long stapled cotton	10	Ovule	Long fiber
SRR6379576	601 long stapled cotton	0	Ovule	Long fiber
SRR6379577	601 long stapled cotton	5	Fiber	Long fiber

The basic information about transcriptome dataset, including data accession, sample name, tissue, time of tissue and phenotype was recorded

Meanwhile, in the turquoise module more basic biological processes such as RNA binding (GO:0003723), DNA binding (GO:0003677), and chromatin remodeling (GO:0006338) were enriched to these fiber-related modules (Additional file 12: Fig. S1c). The green module, another fiber-related module, had microtubule-based movement (GO:0007018), microtubule motor activity (GO:0003777), and small GTPase mediated signal transduction (GO:0007264) (Additional file 12: Fig. S1d). The yellow module contained biological processes related to ATP synthase activity (Additional file 12: Fig. S1e). *Rho* genes among the above mentioned modules indicate that complicated mechanisms underlying the cell elongation process.

Interaction networks of Rho family members

The *Rho* genes from 5 modules were further screened, and genes with the highest membership were selected (Additional file 10: Table S10). Finally, *Ghir_A12G010690*, *Ghir_D08G007370*, *Ghir_D08G018200*, *Ghir_A11G010670*, and *Ghir_D03G002970* were identified as core genes from blue, brown, magenta, turquoise, and green modules, respectively (Fig. 6a, b, c, e, f). Their expression patterns during fiber development were validated by quantitative real-time polymerase chain reaction (qRT-PCR) (Additional file 13: Fig. S2; Additional file 14: Fig. S3). Among 5 core genes, 2 of them, *Ghir_A12G010690* and *Ghir_D08G007370*, were DEGs between fiber and ovule tissues. Genes possessing high transcriptional correlation (pearson) with the 5 core genes were recorded as interacted genes, and were further used to construct interaction networks. In the interaction network of Ghir_A12G010690, genes related to the fatty acid biosynthetic process (GO:0006633), microtubule-based process (GO:0007017), and small GTPase mediated signal transduction (GO:0007264) were identified (Fig. 6d). It was known that the fatty acid was essential for fiber elongation (Qin et al. 2011). Meanwhile, small GTPase mediated signal transduction was also enriched in this network, implying that *Ghir_A12G010690* may participate in signal transduction activities to regulate cell polarity (Etienne 2004). What's more, genes related in fatty acid biosynthesis which participated in the development of cotton fiber were also clustered with *Ghir_A12G010690* (Qin et al. 2007). We inferred that Rho protein as a GTPase that might related to fatty acid biosynthesis to influence fiber development (Fig. 6c and d). In other four networks, there were also GO terms involved in fiber development, such as cell wall biosynthesis (GO:0042546), and calcium ion binding (GO:0005509) (Additional file 11: Table S11). The diverged GO terms enriched among these five *Rho* networks implied that *Rho* may involved in different biological processes.

Discussion

In this study, we identified 2 066 Rho proteins in 8 cotton species. ISR of Rho family members in 2 diploids of A genome showed that Rho members in *G. herbaceum* (A_1) went through less selection which means A_1 is closely related to A donor in tetraploids. This conclusion is consistent with the previous findings (Huang et al. 2020). Interestingly, for D sub-genomes, we also noticed that Rho proteins in *G. hirustum* and *G. tomentosum* were more conserved than those in other three tetraploids, indicating a specific selection happend in *G. tomentosum* and *G. hirustum*. In the previous study, *G. hirustum* and *G. tomentosum* are two cotton species with the closest relativeness (Chen et al. 2020). Therefore, *Rho* genes in D sub-genomes may undergo a strong selection before the divergence between *G. hirustum* and *G. tomentosum* (Chen et al. 2020). This selection may result in long fiber length of *G. hirustum*.

In a previous study, association analysis between genotype and phenotype data could dig out functional genes with high confidentiality (Ma et al. 2018). In this research, we performed an association analysis on 60 cultivars to verify potential functions of *Rho* genes. Combined with transcriptomic analysis and association analysis, 22 *Rho* genes were selected with high confidence. GO enrichment on WGCNA identified modules of 22 candidate *Rho* genes implied a complicated mechanisms underlying Rho functions. Rho protein, as GTPase, participates in many essential pathways and has many interactors as reported in other species (Etienne 2004; Farhan and Hsu 2016; Goryachev and Leda 2017; Moran et al 2019). The inferred interaction networks shed lights on how Rho control the fiber elongation. GO enrichment analysis on genes from a

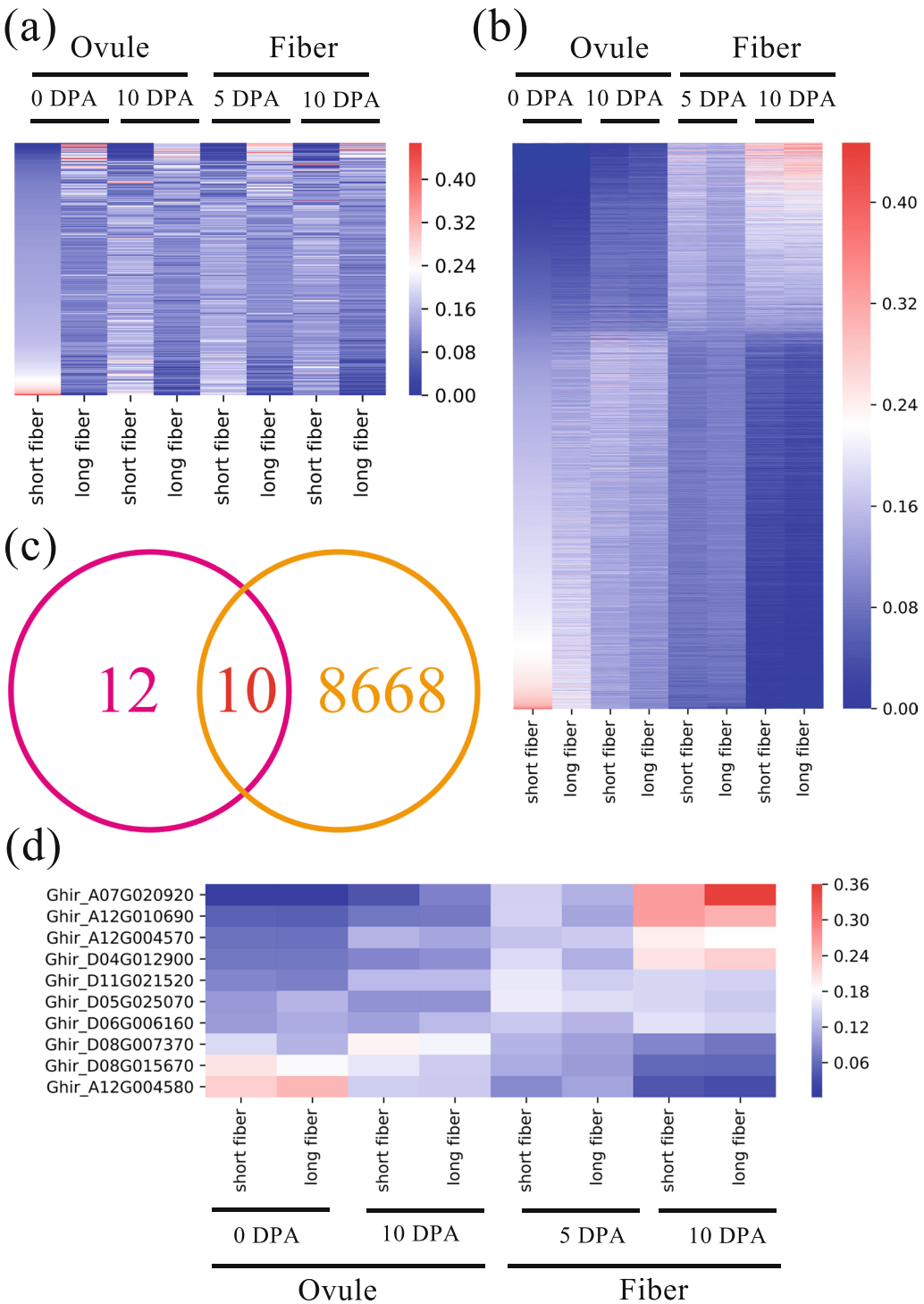

Fig. 4 Characterization of DEGs in transcriptome data. **a** relative transcription abundance of DEGs among long fiber and short fiber cultivars. **b** relative transcription abundance of DEGs among fiber and ovule tissues. **c** Venn plot of 22 Rho members which detected by the combination of transcriptome and association analysis (in magenta color) and DEGs in fiber/ovule tissues (in orange colors). **d** differentially expressed *Rho* genes between fibers and ovules (overlapped genes in **c**)

network constructed towards *Ghir_A12G010690* showed that fatty acid biosynthesis process and small GTPase mediated signal transduction are enriched. We speculate that *Rho* genes may activate a signal transduction pathway and regulate fatty acid biosynthesis which has been proved to activate fiber elongation. What's more, we noticed that cell wall biosynthesis, calcium ion binding also emerged in other *Rho*-based interaction networks. This result suggests that apart from influencing fatty acid biosynthesis, *Rho* genes may regulate fiber development through multiple machinisms.

Conclusions

In this study we identified Rho proteins in 5 allotetraploid cottons and their corresponding sub-genome donors. Rho GTPase, as a key factor involved in basic biological processes, had undergone asymmetric evolution during the divergence of allotetraploids. Five *Rho* genes potentially involved in fiber development were revealed by the combination of association analysis and transcriptome profiling. Further, potential interaction networks for *Rho* genes were built. We believed that findings in this study could be utilized in cotton molecular breeding for fiber improvement.

Methods

Identification of Rho family members

The protein sequence of CDC-42 in human was downloaded from Uniprot (https://beta.uniprot.org/) with entry ID P60953 (Bateman et al 2021). The sequence of CDC-42 in human was aligned to protein domain database in Pfam (http://pfam.xfam.org/) to search for the corresponding HMM model file, with accession number of PF00071.23 (Mistry et al. 2021). After getting HMM file of Ras domain, we searched all protein sequences of 8 *Gossypium* sequences (3 diploids, i.e., *G. arboreum, G.herbaceum,* and *G.raimondii*; 5 allopolyploids, i.e., *G. hirustum, G. barbadense, G.tomentosum, G. mustelinum,* and *G. darwinii*) by hmmscan programme in HMMER (v.3.3.2) to detect proteins with Ras domain (Eddy 2009). E-value and domE value of hmmscan were both set as 1e-5. These proteins were aligned to the protein database in uniprot by blastp (v.2.9.0+) to further identify these detected proteins' functions (Altschul et al. 1990).

Calculation for identical site rate among Rho proteins

Identical site rates (ISRs) among Rho proteins from 8 *Gossypium* species were used to evaluate protein conservation during cotton evolution. The calculation for identical rates was based on the results of multi-alignment. Here, we took the calculation of identical rate in A_1-A_1 as example. All 158 Rho proteins identified in *G. herbaceum* were collected and analyzed in muscle (v.3.8.1551) for multi-alignment. The result of multi-alignment was presented as a fasta file. For each site in the fasta file, the number of protein sequences which were not aligned to other sequences at this site was counted to calculate ISR. The mean identical rates calculated from each site among Rho proteins were used to demonstrate the intra-genomic conservation. As for inter-genomic conservation (taking A_1-A_t as example), Rho proteins from *G. herbaceum* and A sub-genomes of tetraploids were collected. All Rho protein sequences were aligned to each other by muscle (v.3.8.1551) and ISR for each site was calculated in the same way that applied in evaluation of intra-genomic conservation. In the aligned fasta file, the length of each sequence was equal, and sequences with less than 5% unaligned sites were abandoned for the subsequent identical rate calculation. *t*-test for comparsion of identical rates between 2 groups were performed through python package stats.

Phylogeny analysis on Rho members

Since tetraploid cotton species have two sub-genomes, we performed phylogeny analysis separately on A genomes (A sub-genomes) and D genomes (D sub-genomes). The protein sequences were aligned by muscle and put into fasttree (v.2.1.10) to construct phylogenetic trees by default parameters (Price et al. 2009). The phylogenetic tree was decorated by iTOL (https://itol.embl.de/).

Association analysis pipeline

Previous resequenced data was fetched from NCBI (https://www.ncbi.nlm.nih.gov/). The sequence read archive (SRA) accession of raw data in fastq format was SRP115740 at PRJNA399050 (Ma et al. 2018). The phenotype data was downloaded from http://cotton.hebau.edu.cn/. Phenotype data of fiber length was utilized in this study. The best linear unbiased prediction (BLUP) was applied to deal with fiber length from 419 cultivars in 12 environments. Thirty cultivars with the longest fiber length and 30 cultivars

(See figure on next page.)

Fig. 5 WGCNA for transcriptome data in the study. **a** results of gene cluster in WGCNA. **b** mean connectivity and *R*-square of WGCNA. With the increasing soft threshold, *R*-square rises and mean connectivity decreases. **c** thirteen modules classified by WGCNA. **d** trait-module association results of WGCNA. **e** interaction networks within all expressed *Rho* genes. The modules of these *Rho* genes were marked and the interaction networks were presented with corresponding colors

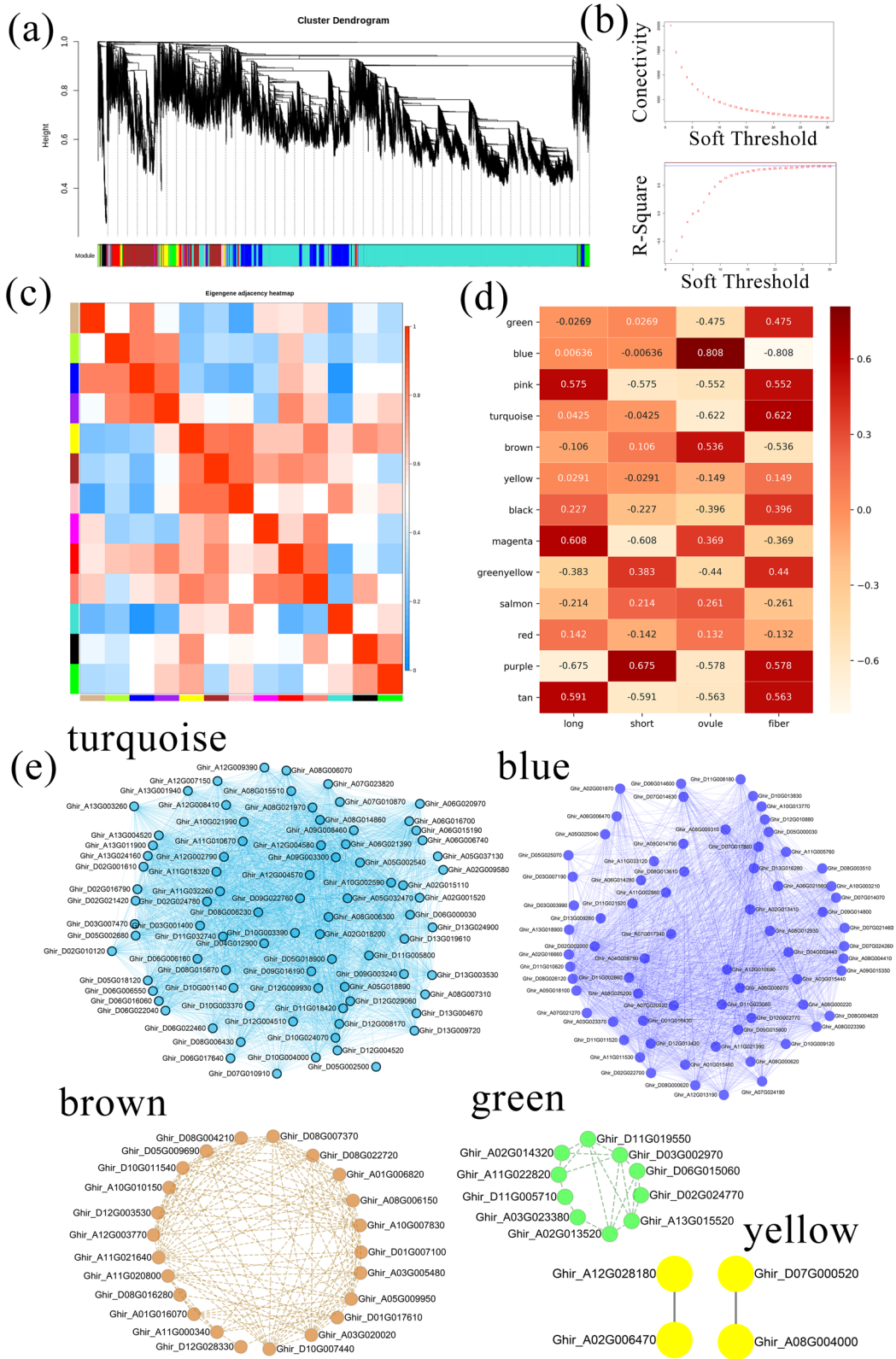

Fig. 5 (See legend on previous page.)

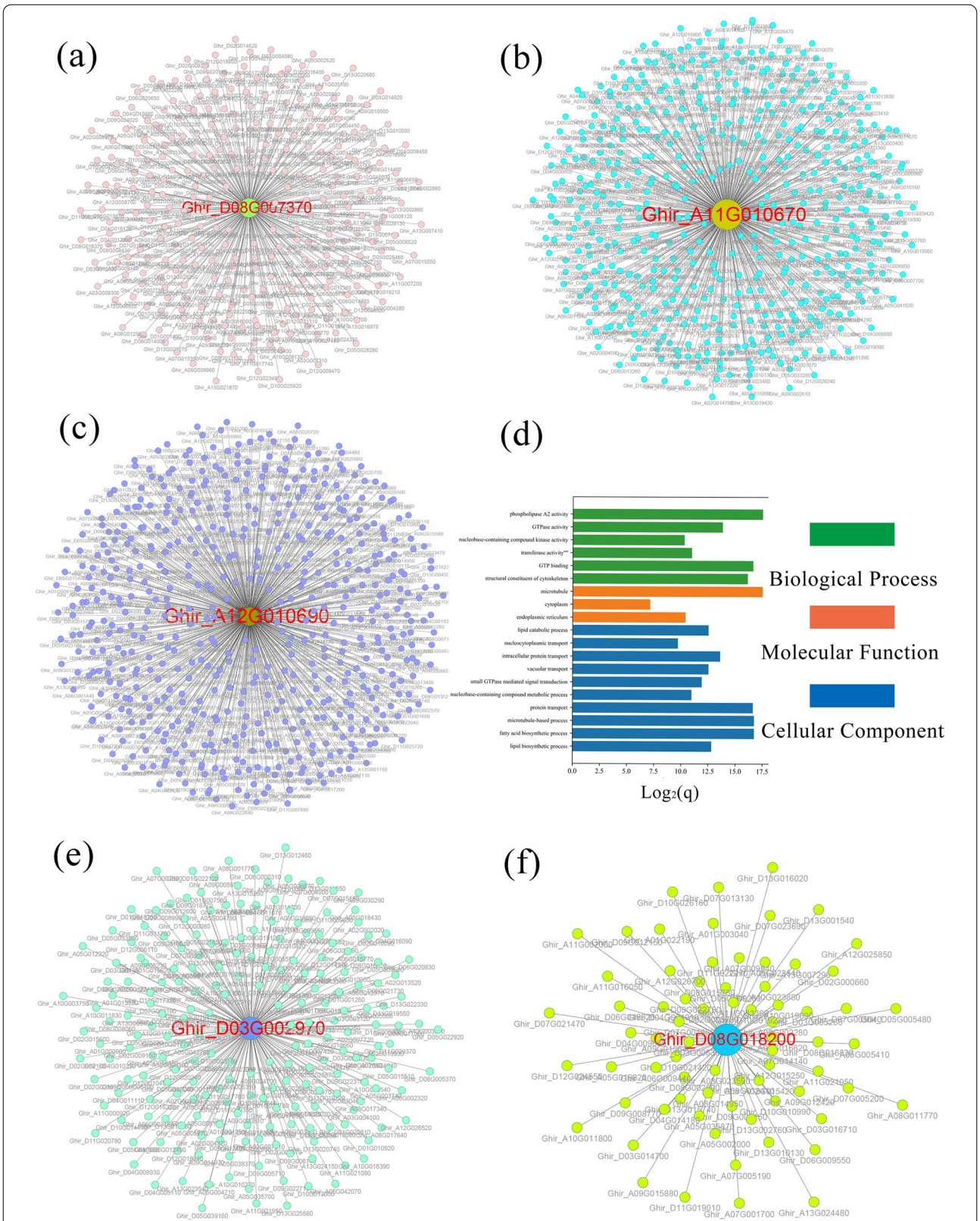

Fig. 6 Interaction networks of 5 Rho members. **a** interaction network of Ghir_D08G007370. **b** interaction network of Ghir_A11GG010670. **c** interaction network of Ghir_A12G010690. **d** GO enrichment of genes interacted with Ghir_A12G010690. **e** interaction network of Ghir_D03G002970. **f** interaction network of Ghir_D08G018200

with the shortest fiber length were selected for association analysis. Raw fastq data were aligned to the reference genome TM-1_HAU (https://cottonfgd.org/about/download.html, *G. hirustum*, HAU) by bwa (v0.7.17) (Wang et al. 2019). Samtools (v1.9) transformed the results of bwa into a binary alignment file (bam file), and the bcftools (v1.9) was used to call SNPs from bam files (Li et al. 2009; Danecek et al. 2011). The fixation index (*F*st) of 2 groups was calculated by vcftools (v0.1.16) with 50 000 window size and 5 000 step size. For the GWAS analysis, the vcf files of 419 cultivars were merged and transformed into tped file by vcftools (v0.1.16). EMMAX (https://genome.sph.umich.edu/wiki/EMMAX) was used to perform GWAS based on tped file and phenotype data of 419 upland cotton cultivars (Hyun et al. 2010). The SNPs with the *P* value smaller than 0.05 was defined as significant SNPs.

Transcriptome analysis pipeline

Raw transcriptome data in fastq format were downloaded from SRA database on NCBI (https://www.ncbi.nlm.nih.gov/). Transcriptome datasets (SRR5992414 and SRR6379574—SRR6379580) from PRJNA400837 were obtained by sratoolkit (v.2.9.6). Quality control and the data trimming were performed by fastp (v.0.23.1) with the default parameters in pair-end mode, and the corresponding clean data were generated (Chen et al. 2018). We used the genome of TM-1 as the reference in the subsequent analysis. Index of the reference genome was implemented by hisat2 (v.2.2.1). The clean data of 8 samples were aligned to the reference genome in a pair-end mode (Kim et al. 2019). The results of hisat2 alignment were transformed into bam, then the bam files were sorted. Transformed and sorted reads were generated by samtools (v.1.9) (Li et al. 2009). Stringtie (2.1.7) was used to identify the transcription abundance of each gene, guided with gene anotation file (-G) and transcription abundance of 8 samples were merged into one file (Pertea et al. 2015).

Identification of DEGs

Transcriptome data were set for two groups including fiber length group of short (SRR6379580, SRR5992414, SRR6379578, and SRR6379579) and long (SRR6379574, SRR6379575, SRR6379576, and SRR6379577) fiber cottons, and tissue group of fiber tissue (SRR6379580, SRR6379578, SRR6379574, and SRR6379577) and ovule tissue (SRR5992414, SRR6379579, SRR6379575, and SRR6379576). Given this, we performed DEG detection for each group. For the detection of DEGs in short and long fiber cottons, transcriptional abundance of genes between all short fiber samples (69-6025-12, mixture of fiber and ovule at different developmental stages) and long fiber samples (601 long stapled cotton mixture from fiber and ovule at different developmental stages) was compared by *t*-test. Genes with significant effect in *t*-test ($P <= 0.05$) were identified as differentially expressed genes (DEGs) between short and long fiber samples. Detection for DEGs between fiber and ovule tissues was based on the same method applied in long and short fiber group. The relative transcription abundance of DEGs was used to plot the heatmap. The relative transcription abundance (RTA) of a gene was calculated by normalizing TPM values among different samples. Assuming that there are *n* samples, RTA of a gene in sample1 was calculated by $TPM_{sample1}/(TPM_{sample1}+.....TPM_{sample\ n})$. The calculation of a gene's RTA could present the divergence of this gene's transcription abundance among samples without the disturbation of mean TPM values of different genes.

WGCNA analysis pipeline

Since transcription abundance of all genes in 8 samples were gained, a weighted gene co-expression network analysis (WGCNA) could be performed (Langfelder and Horvath 2012; Langfelder and Horvath 2008). Before network construction, genes with the maximum TPM among eight samples smaller than five were filtered. Apart from criteria of transcription abundance, genes absent in more than two samples and samples had more than 10% genes absent were removed. After data filtering, the co-expression network was constructed by R package WGCNA based on the remained genes. Different soft thresholds were applied, and finally, the network was evaluated by mean connectivity and *R* square. To construct a scale-free network, deepSplit was set as two, and mergeCutHeight was set as 0.35. The minimum genes in a block were 30, while the maximum gene number in a block was 35 000. All remained genes were divided into several modules by WGCNA, and the eigenvalue of the association of each module and phenotype through Pearson correlations were calculated. Membership of each gene and their potential interacted genes were recorded in the Cytoscape-like file for further visualization.

GO enrichment analysis

To perform GO enrichment analysis, genes collected from WGCNA were handed up to the CottonFGD (https://cottonfgd.org/analyze/) (Zhu et al. 2017). The significant level was set as 0.05, and the minimum gene number for enrichment analysis was set as 3. The result of significant GO enrichment was recorded as a table, and the *q* value of each GO term was used for visualization.

QRT-PCR analysis

We used TM-1 as material for qRT-PCR analysis. Ovule at 0 DPA and 3DPA, fiber at 5 DPA, 10 DPA, and 15 DPA were selected for the qRT-PCR validation of 5 *Rho* genes.

We extracted RNA by RNAprep Pure Plant Kit (Tiangen, Beijing,

China) and gained cDNA by PrimeScript II 1st Strand cDNA Synthesis Kit (TAKARA, Dalian, China). The qRT-PCR were performed on ABI QuantStudio5 RT-PCR system (Applied Biosystems, Foster City,CA, USA) and *Histone* 3 (Genbank id AF024716) was used as internal reference.

Data visualization methods

In all contents of this paper, data visualizations were implemented in various ways. Results of multi-alignments were plotted by python package matplotlib. Phylogenetic trees were displayed by iTOL (https://itol.embl.de/) (Letunic and Bork 2019). For visualization of DEG expression pattern, Venn plots were generated through an online tool Venny2.1.0 (https://bioinfogp.cnb.csic.es/tools/venny/index.html). All heatmaps in this study were plotted by python package seaborn. Figures about WGCNA were all generated by R package WGCNA except for association between modules and phenotype data. Python package pyecharts generated Sankey plot about Rhos and modules from WGCNA. Interactions of potential genes with *Rho* genes were displayed by Cytoscape (v.3.7.2) (Shannon et al. 2003).

Abbreviations

GO: Gene ontology; SNP: Single nucleotide polymorphism; GWAS: Genome-wide association analysis; TPM: Transcripts per million; DEG: Differentially expressed gene; WGCNA: Weighted gene co-expressing network analysis.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s42397-022-00129-4.

Additional file 1. Table S1: 2066 Rho members identified across 8 cotton species, the annotation of each gene from uniprot database were added.

Additional file 2. Table S2: The basic statictics of Rho members across 8 species, including the maximum protein length and the minimum protein length.

Additional file 3. Table S3: The statistics of identical site rate in Rho protein sequences across 8 cotton species. The colums were the diploids which were compared with the corresponding sub-genomes of tetraploids. The *t*-test about identical sites of each sequence from diploids and tetraploids were performed.

Additional file 4. Table S4: Genome-wide *F*st values of all windows with 50 000 bp as window size and 5 000 bp as step size.

Additional file 5. Table S5: The genes overlaped with the top 5% *F*st windows.

Additional file 6. Table S6: Thirty-four Rho genes in genes overlapped with genes overlaped with the top 5% *F*st windows. Thirty-one Rho genes retained after GWAS and the significant SNPs within 31 Rho genes.

Additional file 7. Table S7: DEGs in fiber/ovule groups and DEGs in long fiber/short fiber groups.

Additional file 8. Table S8: 20 expressed genes in transcriptome data, 10 Rho genes in DEGs from fiber/ovule group.

Additional file 9. Table S9: Membership of all genes in WGCNA, larger membership indicates stronger linkage between gene and module.

Additional file 10. Table S10: Networks of 5 Rho genes from 5 modules based on transcription abundance.

Additional file 11. Table S11: GO terms enriched in genes from 5 networks.

Additional file 12. Fig. S12: GO enrichment in genes from 5 modules. (a) GO enrichment of genes in blue modules. (b) GO enrichment of genes in brown modules. (c) GO enrichment of genes in turquoise modules. (d) GO enrichment of genes in green modules. (e) GO enrichment of genes in magenta modules.

Additional file 13. Fig. S13: qRT-PCR for five Rho genes from 5 interaction networks. (a) qRT-PCR analysis for Ghir_A12G010690 during fiber development. (b) qRT-PCR analysis for Ghir_D08G007370 during fiber development. (c) qRT-PCR analysis for Ghir_D08G018200 during fiber development. (d) qRT-PCR analysis for Ghir_A11G010670 during fiber development. (e) qRT-PCR analysis for Ghir_D03G002970 during fiber development.

Additional file 14. Fig. S14: The heatmap presenting RTA of 5 Rho genes for interaction networks in Fig. 6.

Acknowledgements

Not applicable

Author contributions

Wang XF, Song GL conceived and designed the research. He M, Liu S, Cheng HL, and Zuo DY performed the analysis and experiment. Wang QL, Lv L, and Zhang YP downloaded the data and prepared samples. He M wrote the paper. Wang XF, Song GL, and Zuo DY revised the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by Funds of the National Key Research and Development Program (2016YFD0101006, 2018YFD0100402), National Natural Science Foundation of China (31621005 and 31901581), and Central Public-interest Scientific Institution Basal Research Fund (1610162021013). The funders had no role in the design of the study, collection, analysis or interpretation of the data, the writing of the manuscript or the decision to submit the manuscript for publication.

Availability of data and materials

The resequenced data in this manuscript could be downloaded from SRP115740 at PRJNA399050 in NCBI. The phenotype data of corresponding resequenced cultivars could be downloaded from the hyperlink http://cotton.hebau.edu.cn/. Transcriptome raw data could be fetched from PRJNA400837 in NCBI. Reference genomes used in this study were all obtained from https://cottonfgd.org/about/download.html.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]State Key Laboratory of North China Crop Improvement and Regulation, Key Laboratory for Crop Germplasm Resources of Hebei, Hebei Agricultural University, Baoding 071000, China. [2]Institute of Cotton Research of Chinese Academy of Agricultural Sciences, Anyang 455000, Henan, China. [3]Zhengzhou Research Base, State Key Laboratory of Cotton Biology, Zhengzhou University, Zhengzhou 450001, China.

Received: 28 January 2022 Accepted: 20 July 2022
Published online: 01 August 2022

References

Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Anai M, et al. *In vitro* mutation analysis of *Arabidopsis thaliana* small GTP-binding proteins and detection of GAP-like activities in plant cells. FEBS Lett. 1994;346(2–3):175–80. https://doi.org/10.1016/0014-5793(94)80696-9.

Bateman A, Martin M, Orchard S, et al. UniProt: the universal protein knowledge base in 2021. Nucleic Acids Res. 2021;49(D1):D480–9. https://doi.org/10.1093/nar/gkaa1100.

Chen S, Zhou Y, Chen Y, et al. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90. https://doi.org/10.1093/bioinformatics/bty560.

Chen ZJ, Sreedasyam A, Ando A, et al. Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. Nat Genet. 2020;52(5):525–33. https://doi.org/10.1038/s41588-020-0614-5.

Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8. https://doi.org/10.1093/bioinformatics/btr330.

Delmer DP, Pear JR, Andrawis A, et al. Genes encoding small GTP-binding proteins analogous to mammalian rac are preferentially expressed in developing cotton fibers. Mol Gen Genet. 1995;248(1):43–51. https://doi.org/10.1007/BF02456612.

Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform. 2009;23(1):205–11.

Etienne-Manneville S. CDC42 - the centre of polarity. J Cell Sci. 2004;117(8):1291–300. https://doi.org/10.1242/jcs.01115.

Farhan H, Hsu VW. CDC42 and cellular polarity: Emerging roles at the golgi. Trends Cell Biol. 2016;26(4):241–8. https://doi.org/10.1016/j.tcb.2015.11.003.

Galic M, Tsai FC, Collins SR, et al. Dynamic recruitment of the curvature-sensitive protein ArhGAP44 to nanoscale membrane deformations limits exploratory filopodia initiation in neurons. Elife. 2014;3: e03116. https://doi.org/10.7554/eLife.03116.

Gauthier GC, Bredt DS, Murphy TH, et al. Regulation of dendritic branching and filopodia formation in hippocampal neurons by specific acylated protein motifs. Mol Biol Cell. 2004;15(5):2205–17. https://doi.org/10.1091/mbc.e03-07-0493.

Goryachev AB, Leda M. Cell polarity: Spot-on CDC-42 polarization achieved on demand. Curr Biol. 2017;27(16):R810–2. https://doi.org/10.1016/j.cub.2017.07.006.

Haigler CH, Betancur L, Stiff MR, et al. Cotton fiber: a powerful single-cell model for cell wall and cellulose research. Front Plant Sci. 2012;3:104. https://doi.org/10.3389/fpls.2012.00104.

Huang G, Wu Z, Percy RG, et al. Genome sequence of *Gossypium herbaceum* and genome updates of *Gossypium arboreum* and *Gossypium hirsutum* provide insights into cotton A-genome evolution. Nat Genet. 2020;52(5):516–24. https://doi.org/10.1038/s41588-020-0607-4.

Hyun MK, Jae HS, Susan KS, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42:348–54. https://doi.org/10.1038/ng.548.

Ji SJ. Isolation and analyses of genes preferentially expressed during early cotton fiber development by subtractive PCR and cDNA array. Nucleic Acids Res. 2003;31(10):2534–43. https://doi.org/10.1093/nar/gkg358.

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):1361–6. https://doi.org/10.1104/pp.010724.

Kim D, Paggi JM, Park C, et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–15. https://doi.org/10.1038/s41587-019-0201-4.

Koh EJ, Kwon YR, Kim KI, et al. Altered *ARA2* (*RABA1a*) expression in *Arabidopsis* reveals the involvement of a Rab/YPT family member in auxin-mediated responses. Plant Mol Biol. 2009;70(1–2):113–22. https://doi.org/10.1007/s11103-009-9460-7.

Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinform. 2008;9:559. https://doi.org/10.1186/1471-2105-9-559.

Langfelder P, Horvath S. Fast R functions for robust correlations and hierarchical clustering. J Statist Softw. 2012. https://doi.org/10.18637/jss.v046.i11.

Letunic I, Bork P. Interactive tree of life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47(W1):W256–9. https://doi.org/10.1093/nar/gkz239.

Li H, Handsaker B, Wysoker A, et al. The Sequence alignment/map format and samtools. Bioinformatics. 2009;25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352.

Ma ZY, He SP, Wang XF, et al. Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. Nat Genet. 2018;50(6):803–13. https://doi.org/10.1038/s41588-018-0119-7.

Mack NA, Georgiou M. The interdependence of the Rho GTPases and apicobasal cell polarity. Small GTPases. 2014;5(2): e973768. https://doi.org/10.4161/21541248.2014.973768.

Mistry J, Chuguransky S, Williams L, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021;49(D1):D412–9. https://doi.org/10.1093/nar/gkaa913.

Moran KD, Kang H, Araujo AV, et al. Cell-cycle control of cell polarity in yeast. J Cell Biol. 2019;218(1):171–89. https://doi.org/10.1083/jcb.201806196.

Murakoshi H, Wang H, Yasuda R. Local, persistent activation of Rho GTPases during plasticity of single dendritic spines. Nature. 2011;472(7341):100–4. https://doi.org/10.1038/nature09823.

Pertea M, Pertea GM, Antonescu CM, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5. https://doi.org/10.1038/nbt.3122.

Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009;26(7):1641–50. https://doi.org/10.1093/molbev/msp077.

Qin YM, Zhu YX. How cotton fibers elongate: a tale of linear cell-growth mode. Curr Opin Plant Biol. 2011;14(1):106–11. https://doi.org/10.1016/j.pbi.2010.09.010.

Qin YM, Chun HY, Pang Y, et al. Saturated very-long-chain fatty acids promote cotton fiber and arabidopsis cell elongation by activating ethylene biosynthesis. Plant Cell. 2007;19(11):3692–704. https://doi.org/10.1105/tpc.107.054437.

Qin Y, Sun H, Hao P, et al. Transcriptome analysis reveals differences in the mechanisms of fiber initiation and elongation between long- and short-fiber cotton (*Gossypium hirsutum* L.) lines. BMC Genom. 2019; 20:633. https://doi.org/10.1186/s12864-019-5986-5.

Sakabe I, Asai A, Iijima J, et al. Age-related guanine nucleotide exchange factor, mouse Zizimin2, induces filopodia in bone marrow-derived dendritic cells. Immun Age. 2012;11(9):2. https://doi.org/10.1186/1742-4933-9-2.

Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genom Res. 2003;13(11):2498–504. https://doi.org/10.1101/gr.1239303.

Stiff MR, Haigler CH. Cotton fiber tips have diverse morphologies and show evidence of apical cell wall synthesis. Sci Rep. 2016;6(1):27883. https://doi.org/10.1038/srep27883.

Thyssen GN, Fang DD, Turley RB, et al. A Gly65Val substitution in an actin, GhACT_LI1, disrupts cell polarity and F-actin organization resulting in dwarf, lintless cotton plants. Plant J. 2017;90(1):111–21. https://doi.org/10.1111/tpj.13477.

Wang LK, Niu XW, Lv YH, et al. Molecular cloning and localization of a novel cotton annexin gene expressed preferentially during fiber development. Mol Biol Rep. 2010;37(7):3327–34. https://doi.org/10.1007/s11033-009-9919-2.

Wang M, Tu L, Yuan D, et al. Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. Nat Genet. 2019;51(2):224–9. https://doi.org/10.1038/s41588-018-0282-x.

Zeng J, Zhang M, Hou L, et al. Cytokinin inhibits cotton fiber initiation by disrupting PIN3a-mediated asymmetric accumulation of auxin in the ovule epidermis. J Exp Bot. 2019;70(12):3139–51. https://doi.org/10.1093/jxb/erz162.

Zhang M, Zeng JY, Long H, et al. Auxin regulates cotton fiber initiation via GhPIN-mediated auxin transport. Plant Cell Physiol. 2016;58(2):385-97. https://doi.org/10.1093/pcp/pcw203.

Zhang Y, He P, Yang Z, et al. A genome-scale analysis of the PIN gene family reveals its functions in cotton fiber development. Front Plant Sci. 2017;8:461. https://doi.org/10.3389/fpls.2017.00461.

Zhu T, Liang C, Meng Z, et al. CottonFGD: an integrated functional genomics database for cotton. BMC Plant Biol. 2017;17(1):101. https://doi.org/10.1186/s12870-017-1039-x.