Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evolution, Expression and Functional Analysis of Cultivated Allotetraploid Cotton DIR Genes

Zhengwen Liu
Hebei Agricultural University
Xingfen Wang
Hebei Agricultural University
Zhengwen Sun
Hebei Agricultural University
Yan Zhang
Hebei Agricultural University
Chengsheng Meng
Hebei Agricultural University
Bin Chen
Hebei Agricultural University
Guoning Wang
Hebei Agricultural University
Huifeng Ke
Hebei Agricultural University
Jinhua Wu
Hebei Agricultural University
Yuanyuan Yan
Hebei Agricultural University
Liqiang Wu
Hebei Agricultural University
Zhikun Li
Hebei Agricultural University
Jun Yang
Hebei Agricultural University
Guiyin Zhang
Hebei Agricultural University
Zhiying Ma (mzhy@hebau.edu.cn)
Hebei Agricultural University

Research article

Keywords: Cotton, Dirigent proteins, Evolution, RNA-seq, Verticillium wilt, Fiber development, Overexpression

Posted Date: November 4th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-100138/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published on February 10th, 2021. See the published version at https://doi.org/10.1186/s12870-021-02859-0.

Abstract

Background: Dirigent (DIR) proteins mediate regioselectivity and stereoselectivity during lignan biosynthesis and are also involved in lignin, gossypol and pterocarpan biosynthesis. This gene family plays a vital role in enhancing stress resistance and in secondary cell-wall development, but systematical understanding is lacking in cotton.

Results: In this study, 107 *GbDIRs* and 107 *GhDIRs* were identified in *Gossypium barbadense* and *Gossypium hirsutum*, respectively. Most of these genes have a classical gene structure without intron and encode proteins containing a signal peptide. Phylogenetic analysis showed that cotton *DIR* genes were classified into four distinct subfamilies (a, b/d, e, and f). Of these groups, DIR-a and DIR-e were evolutionarily conserved, and segmental and tandem duplications contributed equally to their formation. In contrast, DIR-b/d mainly expanded by recent tandem duplications, accompanying with a number of gene clusters. With the rapid evolution, DIR-b/d-III was a *Gossypium*-specific clade involved in atropselective synthesis of gossypol. RNA-seq data highlighted *GhDIR*s in response to *Verticillium dahliae* infection and suggested that *DIR* gene family could confer Verticillium wilt resistance. We also identified candidate *DIR* genes related to fiber development in *G. barbadense* and *G. hirsutum* and revealed their differential expression. To further determine the involvement of *DIR* genes in fiber development, we overexpressed a fiber length-related gene *GbDIR78* in *Arabidopsis* and validated its function in trichomes and hypocotyls.

Conclusions: These findings contribute novel insights towards the evolution of *DIR* gene family and provide valuable information for further understanding the roles of *DIR* genes in cotton fiber development as well as in stress responses.

Background

The genus *Gossypium* contains at least 46 diploids and five well-established allotetraploids. Two allotetraploid species, *G. hirsutum* and *G. barbadense*, have been cultivated worldwide, which originated from one hybridization event between A-genome (resembling *G. arboreum* or *G. herbaceum*) and D-genome (resembling *G. raimondii*) progenitors around 1–2 million years ago (MYA) [1, 2]. *G. hirsutum* accounts for more than 90% of the global cotton production because of its high yield, whereas *G. barbadense* is characterized by its superior fiber properties. Cotton faces biotic and abiotic stresses during its growth; in addition, improved fiber quality became an urgent need to address the challenges of synthetic fiber industry. The most economic and effective strategy is to develop genetically modified cultivars that not only enhance resistance against environmental stresses but also improve fiber quality. Thus, it is important to mine candidate genes and explain their mechanisms.

Dirigent (from Latin *dirigere*, to guide or align) protein was first isolated from *Forsythia intermedia* and mediated regioselective and stereoselective coupling of two coniferyl alcohol-derived radicals during (+)-pinoresinol biosynthesis [3]. After that, (+)-pinoresinol-forming DIR proteins have been cloned from

several seed-bearing plants, such as *Thuja plicata* [4], *Schisandra chinensis* [5], *Pisum sativum* [6], *Linum usitatissimum* [7] and *Glycine max* [8]. Additionally, (−)-pinoresinol-forming DIR proteins from *Arabidopsis thaliana* [5] and *Linum usitatissimum* [7] were also isolated and characterized. Lignin also derives from monolignol polymerization but is optically inactive. Although molecular mechanisms remain obscure, DIR proteins have been involved in lignin biosynthesis [9–11]. Besides the monolignol metabolism, a few cotton DIR proteins were proposed to mediate the formation of (+)-gossypol by atropselective coupling of hemigossypol radicals [12, 13]. DIR proteins typically lack a catalytically active center, but some members in leguminous plants, with isoflavanol dehydratase activity, can produce pterocarpans [14].

Having the various biochemical functions, DIR proteins play important roles in stress responses, especially in plant defense against pathogens. The expression of *PsDRR206* in pea pod tissue was induced by *Fusarium solani* infection, and its metabolite can function as phytoalexin [6]. Overexpression of *GmDIR22* in soybean can increase total lignan accumulation and enhance plant resistance to *Phytophthora sojae* [8]. Pepper plants with the silencing of *CaDIR7* are more susceptible to *Phytophthora capsici*, NaCl and mannitol stresses [15]. Cotton plants with the overexpression of *GhDIR1* showed an increased lignin content and displayed more resistance to *Verticillium dahliae* [11]. Besides defense responses, DIR proteins also presented other kinds of physiological functions such as Casparian strip formation [10] and pod dehiscence [16].

Ralph and colleagues performed a phylogenetic analysis of 150 DIR proteins from the seed plant division and suggested the presence of six distinct DIR subfamilies, DIR-a and five DIR-like subfamilies (b/d, c, e, f and g) [17]. After that, *DIR* gene family has been systematically studied in several vascular plants, and some stress-induced genes [15, 18–20] or SCW-related genes [21, 22] were identified. Understanding the functions of *DIR* genes would be a practicable approach to enhance stress resistance of cotton and to improve fiber properties, but no comprehensive understanding has been provided in cotton until now. There are only a few *V. dahliae*-responsive *DIR* genes in literatures, and fiber development-related *DIR* genes remain to be identified. Recently, a series of *Gossypium* whole-genome sequences has been released, which makes it possible to analyze gene family at a genome-wide level in cotton. In the present study, we firstly identified the *DIR* gene family members in cultivated allotetraploid cotton, *G. barbadense* and *G. hirsutum*, and analyzed their evolution. Their expression patterns in response to *V. dahliae* infection and during cotton fiber development were systematically investigated. Furthermore, the functional analysis of *GbDIR78* in *Arabidopsis* revealed its role in cell elongation. Our findings will further the understanding of this elusive gene family and provide candidate *DIR* genes for both defense response and fiber development.

Results

Identification, characterization and phylogenetic analysis of *DIR* genes

We identified 107 *GbDIR* genes (*GbDIR1-GbDIR107*) in *G. barbadense* and 107 *GhDIR* genes (*GhDIR1-GhDIR107*) in *G. hirsutum* (Additional file 8: Table S1 and S2). To understand the evolutionary relationships, a neighbor-joining (NJ) phylogenetic tree was constructed using the nucleotide sequences of 107 *GbDIRs*, 107 *GhDIRs*, 25 *AtDIRs* (*Arabidopsis thaliana*), 44 *LuDIRs* (*Linum usitatissimum*), 49 *OsDIRs* (*Oryza sativa*) and 35 *PDIRs* (*Picea* spp.) (Additional file 8: Table S3). As shown in Fig. 1, *DIR* genes were clearly divided into six subfamilies, designated as DIR-a, DIR-b/d, DIR-c, DIR-e, DIR-f and DIR-g. The topological structure is essentially in agreement with previous reports [17, 19, 21, 23]. Among these subfamilies, DIR-a consisted of 8 *GbDIRs*, 7 *GhDIRs*, 5 *AtDIRs*, 6 *LuDIRs*, 7 *OsDIRs* and 12 *PDIRs*, indicating a highly-conserved ancient clade. The largest group DIR-b/d consisted of 81 *GbDIRs*, 82 *GhDIRs*, 14 *AtDIRs*, 28 *LuDIRs*, 10 *OsDIRs* and 12 PDIRs, showing that DIR-b/d has expanded considerably in allotetraploid cotton. In contrast, DIR-f only included 5 *GbDIRs*, 4 *GhDIRs*, 3*LuDIRs* and 11 *PDIRs*, probably due to gene losses in *Arabidopsis thaliana* and *Oryza sativa*. DIR-e contained 13 *GbDIRs*, 14 *GhDIRs*, 6 *AtDIRs*, 6 *LuDIRs*, 3 *OsDIRs* but no *PDIRs*, suggesting a possible angiosperm-specific clade. DIR-c and DIR-g might be unique to monocotyledons, for they only contained *OsDIRs*. To validate the phylogenetic relationships, we reconstructed two phylogenetic trees using DIR protein sequences of *G. barbadense* and *G. hirsutum* (Additional file 1: Figure S1). The evolutionary relationships were highly consistent with that in Fig. 1. Furthermore, these DIR protein sequences were submitted to MEME to discover conserved motifs. As shown in Additional file 1: Figure S1, the adjacent clades carried similar motifs.

Most cotton *DIR* genes (especially members in DIR-a, DIR-e and DIR-f) held a classical gene structure without intron (Additional file 1: Figure S1). Notably, the DIR-b/d clade showed variable gene structures, indicating lower selection constraints. About 77% of *GbDIRs* and 82% of *GhDIRs* encoded proteins containing a signal peptide (Additional file 8: Table S1 and S2). As expected, subcellular localization prediction showed that they were mainly located in the extracellular space. In spite of having a signal peptide, several DIR proteins belonging to DIR-b/d were directed to the vacuole. In addition, DIR proteins located in the nucleus or the peroxisome were observed; the unexpected localization might signify novel functions. We also scanned cotton DIR protein sequences for N-glycosylation sites to explore their solubility and stability. Interestingly, DIR-e genes had much fewer N-glycosylation sites, displaying the evolution and divergence of cotton *DIR* gene family.

Chromosomal distribution, duplication and evolution

To determine the chromosomal distribution of cotton *DIR* genes, we marked their physical locations on chromosomes based on their annotation information (Additional file 2: Figure S2). All the 107 *GhDIRs* were mapped to 24 of the total 26 chromosomes except At06 and Dt06. The gene number varied from 1 to 15 across these chromosomes. Specifically, *GhDIRs* belonging to DIR-b/d were enriched on chromosomes At01, At04, At11, Dt01, Dt04 and Dt11, while DIR-e genes were mainly located on At10 and Dt10. Gene clusters were frequently observed, indicating a number of tandem duplication events. Because of the short divergence time, *GbDIRs* and *GhDIRs* were extremely similar in chromosomal

distribution. *GbDIRs* were located on chromosomes except Dt02, At06 and Dt06, and the gene number ranged from 1 to 13. Six genes (*GbDIR29* and *GbDIR103-GbDIR107*) were not mapped because of the incomplete location information.

To analyze the expansion of cotton *DIR* gene family, we identified segmental and tandem duplications in *G. barbadense* and *G. hirsutum*. The numbers of segmentally duplicated *DIR* genes and tandemly duplicated *DIR* genes were 13 and 31, respectively, in *G. barbadense* and 18 and 41 in *G. hirsutum* (Fig. 2a and 2b; Additional file 8: Table S4 and S5). Obviously, tandem duplication as the major impetus drove the expansion of cotton *DIR* gene family, corresponding to the above-mentioned gene clusters. Ks (substitution per synonymous site) values can be used to estimate the occurrence time of gene duplications and then to identify whole-genome duplication (WGD) events. Ks ranges of 0.4-0.6 (corresponding to *Gossypium*-specific WGD around 16.6 MYA) and 1.5-1.9 (corresponding to the paleo-hexaploidization event shared by eudicots around 130.8 MYA) were observed in *G. raimondii* [24]. Moreover, Ks peak of 0.03 accounts for the divergence between A-genome and D-genome progenitors [25]. Here we also calculated the Ks values of gene duplications to analyze their occurrence time; the LPB method was selected because its results were just in line with expectations (Additional file 8: Table S4 and S5). As shown in Fig. 2, Ks values of *GbDIR14*/*GbDIR41*, *GbDIR79*/*GbDIR99* and *GhDIR13*/*GhDIR41* were 0.60, 0.41 and 0.58, respectively, indicating that these three segmental duplications might be generated from *Gossypium*-specific polyploidization. The other segmental duplications were inferred to arise from the paleo-hexaploidization event, for their Ks values were in or near 1.5-1.9. Interestingly, segmentally duplicated *DIR* genes were mainly retained from the ancestral WGD event rather than the recent one, suggesting some functionally conserved genes through the evolution. Ks values of the tandem duplications ranged from 0.02 to 2.92 in *G. barbadense* and from 0.03 to 2.68 in *G. hirsutum*. Most of the tandem duplications had Ks value less than 0.5, showing a recent expansion of cotton *DIR* gene family. The lacking Ks values of 0.5-1.0 might mean gene losses and/or translocations due to chromosome re-packaging processes following polyploidization. Besides, some tandem duplications with Ks > 1.0 were observed, suggesting that these genes were quite conserved.

To further reveal evolutionary history of cotton *DIR* gene family, we integrated the duplication events with a phylogenetic tree containing 107 *GbDIRs*, 107 *GhDIRs* and 23 reported *DIRs* (Fig. 2c and Additional file 8: Table S6). Track i marked segmental duplications, while tracks ii, iii, iv and v (with artificially-created Ks ranges of 0.0-0.2, 0.2-0.6, 0.6-1.9 and 1.9-3.0, respectively) labeled tandem duplications. The values 0.6 and 1.9 referred to above-mentioned 0.4-0.6 and 1.5-1.9, respectively, and a cluster of Ks values around 0.1 set up the Ks range of 0.0-0.2 (Additional file 8: Table S4 and S5). The DIR-a clade, mainly involved in lignan biosynthesis, was basically established before the *Gossypium*-specific WGD event (Fig. 2c). Similarly, DIR-e was also a relatively ancient clade because of the lack of recent duplications. Clearly, DIR-b/d have experienced rapid expansion due to recent tandem duplications; we believed DIR-b/d-I, -II and -III arose in turn in the evolution. Among them, DIR-b/d-I covered tandem duplications and segmental duplications, forming a transition clade. DIR-b/d-II and the right half of DIR-b/d-III, lacking traces of duplication events, might have undergone large-scale gene losses and/or translocations. The left half of

DIR-b/d-III, involved in the atropselective synthesis of gossypol, emerged recently due to tandem duplications. In addition, DIR-f was obviously inactive in cotton evolution.

The Ka/Ks ratio is a measure of selective pressure on proteins, and Ka/Ks > 1, = 1 and < 1 indicate positive selection (or molecular adaptation), neutral evolution, and purifying selection (or selective constraints), respectively. Here we calculated the Ka/Ks ratio for all the duplicated *DIR* gene pairs (Additional file 8: Table S4 and S5). Interestingly, those three gene pairs generated from *Gossypium*-specific WGD (i.e. *GbDIR14*/*GbDIR41*, *GbDIR79*/*GbDIR99* and *GhDIR13*/*GhDIR41*) showed Ka/Ks ratios > 1, implying that positive selection might contribute to their surviving from gene losses and/or translocations. Ka/Ks ratios of the other segmentally duplicated *DIR* gene pairs were fairly low (0.14-0.35 in *G. barbadense* and 0.14-0.34 in *G. hirsutum*), suggesting the conserved functions due to purifying selection (Additional file 3: Figure S3a and S3b). Moreover, tandemly duplicated gene pairs belonging to DIR-a and DIR-e also showed very small Ka/Ks values (0.11-0.38 in *G. barbadense* and 0.13-0.39 in *G. hirsutum*), further showing the conservation of these two clades. In contrast, gene pairs in the DIR-b/d-III clade held larger Ka/Ks values (0.51-1.74 *in G. barbadense* and 0.53-1.04 *in G. hirsutum*), indicating weaker selective constraints (Additional file 3: Figure S3c and S3d). The moderate selection pressure might result in the rapid expansion of DIR-b/d-III.

Evolutionary history of cotton *DIR* gene family

To further verify the evolution of cotton *DIR* gene family, we constructed a phylogenetic tree consisting of 488 *DIR* genes from eight dicotyledonous species sharing a common paleo-hexaploid ancestor (Fig. 3a and Additional file 8: Table S7). After the paleo-hexaploidization event, *Arabidopsis thaliana* and *Glycine max* have severally undergone two rounds of WGD events, while no WGDs have been identified in *Vitis vinifera* and *Theobroma cacao* [26-29]. Moreover, the *Gossypium* genus have experienced lineage-specific WGD, divergence and hybridization. As shown in Fig. 3a, we identified 33, 54, 25, 32, 63, 67, 107 and 107 *DIR* genes in *Vitis vinifera*, *Glycine max*, *Arabidopsis thaliana*, *Theobroma cacao*, *G. arboreum*, *G. raimondii*, *G. barbadense* and *G. hirsutum*, respectively. We assigned these *DIR* genes to distinct clades (Fig. 3b). Although they underwent different rounds of WGDs, these species were quite consistent in the number of DIR-a genes (Being tetraploids, *G. barbadense* and *G. hirsutum* displayed DIR-a genes twice as many as other species). Similarly, apart from soybean, DIR-e was stable across species. Thus, DIR-a and DIR-e might have been set up before the divergence of Rosids, corresponding to the results inferred from the duplication events. In the above analysis, we found that DIR-f was evolutionarily inactive in *G. barbadense* and *G. hirsutum*. As expected, *G. arboreum*, *G. raimondii*, *Arabidopsis thaliana* and *Theobroma cacao* (all belong to Malvids) possessed less DIR-f genes. In particular, *Arabidopsis thaliana* lost the DIR-f clade during its evolution. DIR-b/d-I existed in all the eight species, DIR-b/d-II was absent in grape and soybean, and DIR-b/d-III was lineage-specific in cotton. Clearly, DIR-b/d-I, -II and -III really appeared in turn in evolution. It has been shown that DIR-b/d-III displayed two distinct clades and that the left clade arose later than the right. As shown in Fig. 3c, most genes in the left clade of DIR-b/d-III were in the collinear blocks, suggesting that DIR-b/d-III was established before the divergence of cotton species.

The chromosome reciprocal translocation between At04 and At05 was also observed, which have been confirmed recently [30].

Expression patterns and transcriptional regulation

Studies have shown that DIR proteins are implicated in lignan, lignin and gossypol biosynthesis, which are all part of plant defense responses against pathogens. To investigate the role of cotton *DIR* genes in disease resistance, the expression patterns of *GhDIRs* were analyzed in response to *V. dahliae* and water (the check group) using a Verticillium wilt-resistant cultivar. As a result, about one quarter of *GhDIRs* were highly expressed in the check group (Fig. 4). For the “gossypol clade” (the left half of DIR-b/d-III), almost half of the members showed a fairly high expression level, indicating its importance in plant pre-formed defense. Once the seedlings were inoculated with *V. dahliae*, most of the highly expressed *GhDIRs* were dramatically down-regulated. It seems that *V. dahliae* weakened the functions of *DIR* genes by disturbing their expression to colonize cotton hosts. These down-regulated genes should be an important resource to understand plant-pathogen interaction. To verify the intriguing expression patterns, we also analyzed *V. dahliae*- and water-responsive expression of *GhDIRs* at 12 hpi in six other cultivars, and the $\log_2$(fold-change) values were presented in a heatmap (Additional file 4: Figure S4). After inoculation with *V. dahliae*, the cultivars S1 and S2 exhibited the largest number of down-regulated *DIR* genes, corresponding to their lowest Verticillium wilt resistance. In contrast with S1 and S2, the cultivars T3 and T4 showed more up-regulated genes in the DIR-a and DIR-e clades, and thus showed higher Verticillium wilt tolerance. Unlike T3, T4, S1 and S2, the cultivars T1 and T2 displayed quite a number of up-regulated *DIR* genes following inoculation with *V. dahliae*. Despite having different patterns in different cultivars, *DIR* genes might have contributed to cotton Verticillium wilt resistance.

Considering that lignin/lignin-like phenolics can affect cotton fiber quality, we analyzed the expression patterns of *DIR* genes during cotton fiber development (Fig. 5). As shown, the expression profiles in *G. barbadense* (Pima90-53 and Hai7124) were similar to that in *G. hirsutum* (HY405 and ND13), exhibiting low expression levels in the DIR-e, DIR-f and DIR-b/d-III clades. Several genes belonging to DIR-b/d-II (*GbDIR25*, *GbDIR71* and *GbDIR107* in *G. barbadense*; *GhDIR27*, *GhDIR28*, *GhDIR75* and *GhDIR76* in *G. hirsutum*) were preferentially expressed in cotton fibers of 20, 25 and 30 DPA, indicating potential functions in secondary wall development. *GbDIR78* and *GhDIR35* which were part of DIR-b/d-I showed high transcript levels in the fibers of 5, 10 and 15 DPA, suggesting their importance during cotton fiber elongation. Furthermore, *GbDIR78* and *GhDIR35* were highly homologous but differentially expressed. Two DIR-a genes *GhDIR12* and *GhDIR36* were highly expressed during secondary wall thickening, whereas their orthologous genes *GbDIR13* and *GbDIR35* exhibited quite low expression levels. The differential expression might have contributed to the different fiber quality in these two species.

Transcription factor binding sites (TFBS) provide cues for transcriptional regulation. A total of 266 JASPAR matrices were selected and fetched to identify potential TFBS in the promoter regions of *GbDIRs* and *GhDIRs* (Additional file 8: Table S8). Despite the biased JASPAR database and the strict threshold,

TFBS were widely detected, including hormone-activated signaling pathway (ABA, IAA, ETH, GA, JA and SA), response to abiotic stresses (drought, salt and temperature), response to biotic stresses, and plant cell wall development (Fig. 6 and Additional file 5: Figure S5; Additional file 8:Table S9 and S10). In the DIR-a, DIR-e and DIR-f clades, *GhDIR33*, *GhDIR36*, *GhDIR78*, *GhDIR80*, *GhDIR86* and *GhDIR92* had no TFBS related to ABA signal transduction and showed little or no expression in the roots of cotton seedlings (Fig. 4); the others tended to be highly expressed, indicating the probable regulatory roles of ABA signaling in root-specific gene expression. Four DIR-b/d-II genes *GhDIR27*, *GhDIR28*, *GhDIR75* and *GhDIR76* were highly expressed in cotton fibers. Despite having extremely close phylogenetic relationships with these four genes, *GhDIR6* and *GhDIR64* exhibited quite low expression levels (Fig. 5b). One cause may be the lack of TFBS in their promoter regions (Fig. 6). Similarly, *GbDIR6* and *GbDIR60* differed from *GbDIR25*, *GbDIR71* and *GbDIR107* in TFBS occurrences and in transcript levels (Fig. 5a and Additional file 5: Figure S5). As another example, *GhDIR36* carried more IAA-responsive TFBS than *GbDIR35*, corresponding to their differential expression in cotton fibers (Fig. 5 and Additional file 6: Figure S6a). Considering that *GbDIR13* and *GhDIR12* differed in transcript levels in cotton fibers but owned similar TFBS (Fig. 5 and Additional file 6: Figure S6b), their *trans*-acting TFs were analyzed. As shown, some TFs associated with IAA signaling, ETH signaling or plant cell wall development exhibited higher expression levels in *G. hirsutum* than in *G. barbadense*.

Functional characterization of *GbDIR78* in *Arabidopsis*

RNA-seq data showed *GbDIR78* was preferentially expressed during cotton fiber elongation and that *GbDIR78* and *GhDIR35* differed in transcript levels, implying the involvement of *GbDIR78* in cell elongation. To further identify its functions, the ORF of *GbDIR78* driven by a 35S promoter was transformed into *A. thaliana* plants. Two transgenic T_3 lines OE2 and OE3 were generated, and the stable expression of *GbDIR78* was confirmed by Real-time PCR and Western blot (Fig. 7a and 7b). *Arabidopsis* leaf trichomes can serve as a useful experimental system to dissect cotton fiber development because they partly share regulatory mechanisms [31-34]. Here, trichomes from the fifth rosette leaves of OE2, OE3 and WT plants were measured, and then we discovered that the transgenic plants owned significantly longer trichomes (Fig. 7c and 7d). Moreover, dark-grown hypocotyls were utilized to determine the role of *GbDIR78* in cell elongation because their growth resulted from cell elongation rather than division [35, 36]. As a result, the seedlings of OE2 and OE3, compared with WT seedlings, showed significantly longer hypocotyls (Fig. 7e and 7f). As expected, the longer epidermal cells were observed in the hypocotyls of transgenic plants in a microscopic inspection (Fig. 7g). These results indicated that *GbDIR78* can promote cell elongation and might have contributed to cotton fiber development.

Discussion

Type b/d group expanded considerably and evolved rapidly in cotton

Gene duplication events (tandem, segmental/whole-genome or by transposition) have provided raw evolution materials and meanwhile built up various types of gene families. Among them, tandem and segmental duplications (from unequal crossing-over and infrequent polyploidy, respectively) are fully thought out in the evolution of plant gene families [37, 38]. In the present study, the expansion of DIR-b/d was mainly due to tandem duplications (Fig. 2). As a result, a number of gene clusters were generated, and the chromosomes At01, At04, At11, Dt01, Dt04 and Dt11 contained only DIR-b/d genes (Additional file 2: Figure S2). Interestingly, the heterogeneous gene clusters including DIR-b/d-II and DIR-b/d-III members were observed on chromosomes At01 and Dt01, which provided clues about the initial generation of DIR-b/d-III. DIR-b/d genes accounted for 75.7% in *G. barbadense* and 76.6% in *G. hirsutum*, while 34.3%, 20.4%, 50.0% and 56.8% in spruce, rice, pepper and flax, respectively [15, 17, 19, 21]. Also, as shown in Fig. 3b, the proportion of DIR-b/d genes in cotton was much larger than that in *Vitis vinifera*, *Glycine max*, *Arabidopsis thaliana* and *Theobroma cacao*. All these results showed a *Gossypium*-specific expansion of DIR-b/d. Given the rare segmental duplication events in plant genomes, tandem duplication has been proposed as a proper mechanism to cope with rapidly changing environments [39, 40]. Therefore, the rapid expansion of DIR-b/d might be a plant adaptive evolution against biotic and abiotic stresses. For example, a cluster of high-density genes located on chromosome At04 was involved in gossypol biosynthesis (Additional file 2: Figure S2) and might have contributed to plant defense responses against pathogens and pests. Moreover, tandem duplication acts different contribution across species; tandemly duplicated *DIR* genes are abundant in *Medicago truncatula* and *Oryza sativa* but scarce in *Capsicum annuum* [15, 19, 20]. To explore why tandem duplications promoted the rapid expansion of DIR-b/d, we compared the strength of selection acting on tandemly duplicated genes (Additional file 3: Figure S3c and S3d). Compared with the genes in DIR-a and DIR-e, tandemly duplicated genes in DIR-b/d, especially DIR-b/d-III, showed larger Ka/Ks values, which meant weaker purifying selection and greater evolutionary rates. In other words, the relaxed selective constraints might have accelerated neofunctionalization corresponding to the occurrence of “gossypol clade”. Moreover, this may be one reason for the rapid expansion of cotton DIR-b/d subfamily. Considering that the calculated Ka/Ks values were averaged over sites and time, positive selection might have worked at individual sites or in a short period, which benefitted the retention of gene duplicates. In brief, perhaps to cope with environmental challenges, the DIR-b/d clade expanded considerably and evolved rapidly in allotetraploid cotton. The tandemly arrayed *DIR* genes should also be candidate resistance resources in breeding programs.

DIRs may substantially affect cotton fiber development

Lignin, deposited mostly in the secondary cell walls of vascular plants, contributes to water transport, mechanical support and plant stress responses. Besides, the deposition of lignin in cell walls can repress cell growth due to the decreased extensibility. In cotton fibers, lignin has been neglected for the low concentrations. However, in recent years, studies suggested that lignin-like phenolics may substantially affect cotton fiber quality [41-43]. Given the involvement of *DIR* genes in lignin biosynthesis, we analyzed their expression patterns during cotton fiber development. A DIR-b/d-II “SCW clade” were observed in *G.*

hirsutum, consisting of *GhDIR27*, *GhDIR28*, *GhDIR75* and *GhDIR76* (Fig. 5b). Being highly expressed during secondary wall thickening, they might affect lignin deposition in cotton fibers. After analyzing QTLs reported for fiber quality traits, *GhDIR27* and *GhDIR28* fell in qFS-c4–1, a stable QTL across multiple environments controlling fiber strength [44]. Also, two other QTLs for fiber strength (qFS-C4-3 and qFS04.1) and one QTL controlling fiber micronaire (qFM-Chr4-3) were detected at almost the same region [45-47]. Interestingly, *GhDIR27* and *GhDIR28* also fell in qFL04.1, a QTL conferring fiber length [48]. Thus, it is reasonable to speculate that this clade might regulate lignin biosynthesis and affect fiber development. Similarly, "SCW clade" was also observed in *G. barbadense* (Fig. 5a). However, *GbDIR27* and *GbDIR70* (orthologous to *GhDIR28* and *GhDIR75*, respectively) showed quite low expression levels, which might explain in part different fiber properties between the two cotton species. DIR-a genes are widely considered to mediate lignan biosynthesis (Additional file 8: Table S6). In particular, a soybean DIR-a protein GmPdh1 was inferred to affect the pattern of lignin deposition by considering its expression patterns and its promotion to pod dehiscence. [16]. It reminds us of that the loss-of-function mutation of *AtPrR1* (pinoresinol reductase) results in alterations in lignin levels, lignin structure and tissue-specific lignin distribution [49]; there may be an association between lignan biosynthesis and lignin deposition. In the present study, two DIR-a genes *GhDIR12* and *GhDIR36* broadly similar to *GmPdh1* were highly expressed during secondary cell wall thickening (Fig. 5b). However, their orthologous genes (*GbDIR13* and *GbDIR35*, respectively) exhibited little expression in cotton fibers (Fig. 5a). It is interesting to investigate whether *GhDIR12* and *GhDIR36* can negatively regulate fiber quality by enhancing lignin biosynthesis. Unlike the above-mentioned genes, *GhDIR35* was preferentially expressed during fiber elongation. It belonged to the DIR-b/d-I clade and fell in qFL08.1, a stable QTL for fiber length [50]. *GbDIR78*, highly homologous to *GhDIR35*, showed a higher transcript level. Moreover, the overexpression of *GbDIR78* in *Arabidopsis* plants could promote cell elongation (Fig. 7). Therefore, *GbDIR78* and *GhDIR35* may play an important role in cotton fiber elongation, and their differential expression might have contributed to the different fiber properties.

GbDIR78 promotes cell elongation possibly by regulating phenylpropanoid metabolism

GbDIR78 can promote cell elongation in *Arabidopsis* plants, but the mechanisms need to be discussed. Being preferentially expressed during cotton fiber elongation, *GbDIR78* is not likely able to participate in lignin biosynthesis. A soybean gene *GmDIR22* can effectively direct lignan biosynthesis in vitro and in vivo [8]. *GbDIR78* was quite close to *GmDIR22* in phylogenetic tree (Fig. 2c), implying that *GbDIR78* may also be involved in lignan biosynthesis. Moreover, when transiently expressed in onion epidermal cells, GbDIR78 entered the secretory pathway and was mainly retained in the plasma membrane (Additional file 7: Figure S7), which is partly similar to the subcellular localization of GmDIR22. However, DIR proteins are targeted to the cell wall when involved in lignin biosynthesis [9, 10]. Thus, we reasonably speculate that *GbDIR78* participates in lignan biosynthesis. Slightly confusingly, the reported cotton *GhDIR1* gene shares close evolutional relationships with *GbDIR78*, but the overexpression of *GhDIR1* in cotton can enhance

lignification [11]. This could be explained in terms of the association between lignan and lignin biosynthesis [49].

Apoplast ROS (reactive oxygen species) signaling is crucial for cell elongation [51], but at high concentrations ROS become toxic, causing cell wall stiffening [52]. The cell walls of *Arabidopsis* plants overexpressing *GbDIR78* may be at moderate ROS concentrations, because some (neo)lignans can act against oxidative damage [53]. Given that lignans compete with flavonoids for phenylalanine precursors [54, 55], the metabolic flux towards lignans can result in a reduction of flavonoid biosynthesis. Flavonoids (especially flavonols) have been shown to inhibit polar auxin transport [56]. Thus, in *GbDIR78*-overexpressed *Arabidopsis* plants, auxin transport (in an apical-basal axis) may be elevated compared with WT plants. To sum up, the longer trichomes and hypocotyls in the transgenic *Arabidopsis* plants might be due to moderate ROS levels and higher auxin accumulation. In cotton fibers, appropriate ROS levels are important for cell elongation [57, 58]. Also, some flavonoids play a negative impact on cotton fiber development [59]. Thus, the metabolic flux from flavonoids into lignans should be a novel alternative way to improve cotton fiber quality.

Conclusions

In summary, we performed a genome-wide analysis of *DIR* gene family in *G. barbadense* and *G. hirsutum*. Our study clearly demonstrates how segmental and tandem duplications contribute to the expansion of cotton *DIR* gene family and highlights a *Gossypium*-specific clade involved in atropselective synthesis of gossypol. We also suggest that *DIR* genes can not only confer Verticillium wilt resistance but also affect cotton fiber development. In addition, the fact that *GbDIR78* can promote cell elongation in *Arabidopsis* plants paves an alternative way to improve cotton fiber properties. Our results provide useful insights into the evolutionary history, expression patterns, transcriptional regulation, and functional analysis of *Gossypium DIR* genes.

Methods

Plant materials and growth conditions

G. barbadense cv. Pima90-53 and Hai7124, and *G. hirsutum* cv. HY405 and ND13 were grown in an experimental field in Baoding (115°47'N, 38°87'E) during the growing season (late April to late October). All field management, including watering, weed management, and fertilization, was performed according to the usual local management. Cotton bolls were collected at 14:00-16:00 at 0, 5, 10, 15, 20, 25 and 30 DPA. *G. hirsutum* cv. ND601 [60], and six other cultivars (AusSiV2, Xinmian33B, Nongdamian7, Nongdamian8, Handan333 and Xiangmian18) from a core collection of cotton germplasm resources [34] were grown in pots full of commercial sterilized soil in a chamber under environmental conditions of 28°C/25°C (day/night), 16-h photoperiod, and 80% relative humidity. Two-week-old cotton seedlings were infected with *V. dahliae* strain Linxi2-1 as described by Yang et al [61]. Seeds of the above-mentioned cultivars were deposited in the cotton germplasm bank of Hebei Agriculture University, which could be

freely obtained by others for research. *A. thaliana* Columbia wild-type plants and transgenic plants were grown in commercially sterilized soil in a greenhouse (22°C, 16-h photoperiod, and 70% relative humidity). Hoagland's nutrient solution was added every four days.

Sequence sources

The sequences of *G. raimondii* [62], *G. arboretum* [63], *G. hirsutum* and *G. barbadense* [64] were downloaded from CottonFGD [65]. The sequences of *A. thaliana* [26], *V. vinifera* [27], *G. max* [28], *T. cacao* [29], *L. usitatissimum* [66] and *O. sativa* [67] were obtained from Ensembl Plants [68] or Phytozome [69].

Identification and characterization of DIR proteins

The sequence alignments of *DIR* gene family (PF03018) were downloaded from Pfam [70]. The candidate DIR proteins were identified using HMMER 3.0 [71] and confirmed with the Batch CD-Search service [72]. N-glycosylation sites were identified using NetNglyc 1.0 (http://www.cbs.dtu.dk/services/NetNGlyc/). Signal peptide prediction was performed with SignalP 5.0 [73]. YLoc served for predicting subcellular localization [74]. The protein length, molecular weight and isoelectric point were investigated by a native Perl script.

Gene structure, chromosomal distribution, conserved motifs and phylogenetic analysis

The exon-intron information and gene location information were fetched from gene annotation files and subsequently visualized using TBtools [75]. MEME was employed to search conserved motifs, with a limit of 15 motifs and other default parameters [76]. The phylogenetic trees were constructed by MEGA 7.0 using the neighbor-joining (NJ) method with 1000 bootstrap replications [77] and then displayed with the online iTOL tool [78].

Gene duplication and the calculation of Ka, Ks and Ka/Ks values

Segmental and tandem duplications were detected by MCScanX with default parameters [79]. The duplication events were fetched and then displayed with TBtools [75]. Homologous genes between At- and Dt-subgenome were determined using the bidirectional best hit method in BLAST. We used ParaAT [80] to construct protein-coding DNA alignments. Then the paired sequences were used to calculate Ka, Ks and Ka/Ks values using KaKs_Calculator with the LPB method [81].

RNA-seq data

A Verticillium wilt-resistant *G. hirsutum* cultivar ND601 and a highly aggressive defoliating *V. dahliae* strain Linxi2-1 were used in *V. dahliae*-responsive expression analysis. The roots of two-week-old seedlings infected by *V. dahliae* were collected independently at 2, 6, 12, 24 and 48 hours post-inoculation (hpi), while the control plants were inoculated with distilled water. Two biological replicates were generated and each used for RNA extraction, library preparation and sequencing analysis. Then gene expression levels were calculated using FPKM (Fragments Per Kilobase of exon model per Million mapped reads). Finally, $\log_2$(1+FPKM) values after averaging two replicates were displayed. Similarly, six other *G. hirsutum* cultivars (four Verticillium wilt-tolerant cultivars and two susceptible cultivars) were used in expression experiments. RNA-seq samples were generated at 12 hpi in the same way described above. To highlight expression changes, we showed fold-change values with $\log_2$ transformation. The fold change is the ratio of treatment 1+FPKM to control 1+FPKM. In addition, two *G. barbadense* cultivars (Pima90-53 and Hai7124) and two *G. hirsutum* cultivars (HY405 and ND13) were employed in fiber development-related expression analysis. Cotton bolls were harvested at fiber initiation stage (0 DPA), elongation stage (5, 10 and 15 DPA) and secondary wall thickening stage (20, 25, and 30 DPA). To minimize variations, fibers from multiple cotton bolls were collected, pooled and then used to extract total RNA. Twenty-eight libraries (ovules of 0 DPA; fibers of 5, 10, 15, 20, 25 and 30 DPA) were eventually used to perform high-throughput RNA-seq. RPKM (Reads Per Kilobase of exon model per Million mapped reads) was applied to estimate expression levels, and we finally showed $\log_2$(1+RPKM) values.

Promoter analysis

The promoter sequences (2000-bp upstream of ATG) of *DIR* genes were extracted from genomes of *G. barbadense* and *G. hirsutum*. According to Gene Ontology annotations, a total of 266 JASPAR matrices (transcription factor binding profiles) were selected and then fetched, including hormone-activated signaling pathway (ABA, IAA, ETH, GA, JA and SA), response to abiotic stresses (drought, salt and temperature), response to biotic stresses, and plant cell wall development [82]. For each JASPAR matrix, FIMO was used to scan promoter sequences for matches with a strict threshold of *p*-value < 1E-5 [83]. The potential transcription factor binding sites (TFBS) were counted according to their classification information. Finally, we showed the results using iTOL [78].

Functional analysis of *GbDIR78*

The coding sequence of *GbDIR78* (from Pima90-53) was amplified and inserted into the Gateway pDONR207 vector to form an entry clone. Then the coding sequence was recombined into the Gateway pGWB414 vector to generate an overexpression (OE) construct under the control of CaMV 35S promoter. The OE construct was transformed into *A. thaliana* (Columbia ecotype) through *Agrobacterium*-mediated plant transformation. Transgenic plants were identified using 50 μg/ml kanamycin screening (1/2 MS medium) and PCR detection. Two transgenic T_3 lines OE2 and OE3 were generated, and the stable expression of *GbDIR78* was confirmed by Real-time PCR and Western blot (HA-Tag). Decolorized by

ethanol, the fifth rosette leaves of four-week-old wide type (WT) and OE plants were photographed with an Olympus BX51 microscope (Tokyo, Japan), and the longest branch was measured using the ImageJ software for each of about 150 legible trichomes in each line [84]. To observe dark-grown hypocotyls, the seeds of *A. thaliana* were sterilized and then grown in vertical plates (1/2 MS medium, 0.9% agar and pH 5.8) under the conditions of 22°C and continuous darkness. Five-day-old WT and OE seedlings were harvested and then photographed with a professional Epson V800 scanner (Nagano, Japan). Their hypocotyls were subsequently measured using the ImageJ software [84]. To observe the epidermal cells of the hypocotyls, the seedlings used above were stained with propidium iodide (PI) and then photographed using an Olympus FV10i laser scanning microscope (Tokyo, Japan). Two-tailed *t* test (*p*-value) was conducted using the GraphPad Prism software (San Diego, CA, USA). To examine the subcellular localization of GbDIR78, its coding sequence was recombined into the Gateway pEarleyGate103 vector, which can express target protein with a C-terminal GFP fusion. A plasmid expressing GFP alone served as a control. GbDIR78-GFP fusion protein and GFP were transiently expressed in onion epidermal cells by a Bio-Rad PDS-1000/He system (Hercules, CA, USA). Then the transformed cells were monitored with an Olympus BX51 microscope (Tokyo, Japan) after incubating on MS agar medium for 24 h (22 °C, continuous darkness).

Abbreviations

MYA: Million years ago; Ks: Synonymous substitution per synonymous site; Ka: Nonsynonymous substitutions per nonsynonymous site; WGD: Whole-genome duplication; TFBS: Transcription factor binding sites; ABA: Abscisic acid; IAA: Auxin; ETH: Ethylene; GA: Gibberellin; JA: Jasmonic acid; SA: Salicylic acid; WT: Wide type; OE: Overexpression; ROS: Reactive oxygen species; hpi: hours post-inoculation; DPA: Day post-anthesis; QTL: quantitative trait locus

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The data used during the current study are included in this published article and its additional files or are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was financially supported by the Fund of the China Agriculture Research System (CARS15-03) and the Science and Technology Support Program of Hebei Province (16226307D). The funding bodies played no role in research design, data analysis, and manuscript preparation.

Authors' contributions

ZM and GZ conceived the project. ZL and XW performed the research and analyzed the data. ZS, YZ, CM and YY contributed to the bioinformatics analysis. BC, GW and JW collected the RNA-seq samples. HK, LW, ZKL and JY helped in growing experimental materials. ZL and XW wrote the manuscript. ZM and GZ revised the manuscript. All authors have read and approved the manuscript.

Acknowledgements

The authors are grateful to worldwide colleagues for releasing cotton genomic data.

References

1. Wendel JF. New World tetraploid cottons contain Old World cytoplasm. P Natl Acad Sci USA. 1989;86:4132-4136.
2. Grover CE, Gallagher JP, Jareczek JJ, Page JT, Udall JA, Gore MA, Wendel JF. Re-evaluating the phylogeny of allopolyploid *Gossypium* L. Mol Phylogenet Evol. 2015;92:45-52.
3. Davin LB, Wang H, Crowell AL, Bedgar DL, Martin DM, Sarkanen S, Lewis NG. Stereoselective bimolecular phenoxy radical coupling by an auxiliary (dirigent) protein without an active center. Science. 1997;275(5298):362-367.
4. Kim MK, Jeon JH, Fujita M, Davin LB, Lewis NG. The western red cedar (*Thuja plicata*) 8-8' *DIRIGENT* family displays diverse expression patterns and conserved monolignol coupling specificity. Plant Mol Biol. 2002;49(2):199-214.
5. Kim KW, Moinuddin SG, Atwell KM, Costa MA, Davin LB, Lewis NG. Opposite stereoselectivities of dirigent proteins in *Arabidopsis* and *schizandra* species. J Biol Chem. 2012;287(41):33957-33972.
6. Seneviratne HK, Dalisay DS, Kim K-W, Moinuddin SGA, Yang H, Hartshorn CM, Davin LB, Lewis NG. Non-host disease resistance response in pea (*Pisum sativum*) pods: Biochemical function of DRR206 and phytoalexin pathway localization. Phytochemistry. 2015;113:140-148.

7. Dalisay DS, Kim KW, Lee C, Yang H, Rübel O, Bowen BP, Davin LB, Lewis NG. Dirigent protein-mediated lignan and cyanogenic glucoside formation in flax seed: Integrated omics and maldi mass spectrometry imaging. J Nat Prod. 2015;78(6):1231-1242.
8. Li N, Zhao M, Liu T, Dong L, Cheng Q, Wu J, Wang L, Chen X, Zhang C, Lu W, et al. A novel soybean dirigent gene *GmDIR22* contributes to promotion of lignan biosynthesis and enhances resistance to *phytophthora sojae*. Front Plant Sci. 2017;8:1185.
9. Burlat V, Kwon M, Davin LB, Lewis NG. Dirigent proteins and dirigent sites in lignifying tissues. Phytochemistry. 2001;57(6):883-897.
10. Hosmani PS, Kamiya T, Danku J, Naseer S, Geldner N, Guerinot ML, Salt DE. Dirigent domain-containing protein is part of the machinery required for formation of the lignin-based Casparian strip in the root. P Natl Acad Sci USA. 2013;110(35):14498-14503.
11. Shi H, Liu Z, Zhu L, Zhang C, Chen Y, Zhou Y, Li F, Li X. Overexpression of cotton (*Gossypium hirsutum*) *dirigent1* gene enhances lignification that blocks the spread of *Verticillium dahliae*. Acta Biochim Biophys Sin. 2012;44(7):555-564.
12. Effenberger I, Zhang B, Li L, Wang Q, Liu Y, Klaiber I, Pfannstiel J, Wang Q, Schaller A. Dirigent proteins from cotton (*Gossypium* sp.) for the atropselective synthesis of gossypol. Angew Chem. 2015;54(49):14660-14663.
13. Effenberger I, Harport M, Pfannstiel J, Klaiber I, Schaller A. Expression in *Pichia pastoris* and characterization of two novel dirigent proteins for atropselective formation of gossypol. Appl Microbiol Biotechnol. 2017;101(5):2021-2032.
14. Uchida K, Akashi T, Aoki T. The missing link in leguminous pterocarpan biosynthesis is a dirigent domain-containing protein with isoflavanol dehydratase activity. Plant Cell Physiol. 2017;58(2):398-408.
15. Khan A, Li RJ, Sun JT, Ma F, Zhang HX, Jin JH, Ali M, Haq SU, Wang JE, Gong ZH. Genome-wide analysis of dirigent gene family in pepper (*Capsicum annuum* L.) and characterization of *CaDIR7* in biotic and abiotic stresses. Sci Rep. 2018;8(1):5500.
16. Funatsuki H, Suzuki M, Hirose A, Inaba H, Yamada T, Hajika M, Komatsu K, Katayama T, Sayama T, Ishimoto M, et al. Molecular basis of a shattering resistance boosting global dissemination of soybean. P Natl Acad Sci USA. 2014;111(50):17797-17802.
17. Ralph SG, Jancsik S, Bohlmann J. Dirigent proteins in conifer defense II: Extended gene discovery, phylogeny, and constitutive and stress-induced gene expression in spruce (*Picea* spp.). Phytochemistry. 2007;68(14):1975-1991.
18. Thamil Arasan SK, Park JI, Ahmed NU, Jung HJ, Hur Y, Kang KK, Lim YP, Nou IS. Characterization and expression analysis of dirigent family genes related to stresses in *Brassica*. Plant Physiol Biochem. 2013;67:144-153.
19. Liao Y, Liu S, Jiang Y, Hu C, Zhang X, Cao X, Xu Z, Gao X, Li L, Zhu J, et al. Genome-wide analysis and environmental response profiling of dirigent family genes in rice (*Oryza sativa*). Genes & Genomics. 2016;39(1):47-62.

20. Song M, Peng X. Genome-wide identification and characterization of DIR genes in *Medicago truncatula*. Biochem Genet. 2019;57(4):487-506.
21. Corbin C, Drouet S, Markulin L, Auguin D, Lainé É, Davin LB, Cort JR, Lewis NG, Hano C. A genome-wide analysis of the flax (*Linum usitatissimum* L.) dirigent protein family: from gene identification and evolution to differential regulation. Plant Mol Biol. 2018;97:73-101.
22. Cheng X, Su X, Muhammad A, Li M, Zhang J, Sun Y, Li G, Jin Q, Cai Y, Lin Y. Molecular characterization, evolution, and expression profiling of the *dirigent* (*DIR*) family genes in Chinese white pear (*Pyrus bretschneideri*). Front Genet. 2018;9:136.
23. Paniagua C, Bilkova A, Jackson P, Dabravolski S, Riber W, Didi V, Houser J, Gigli-Bisceglia N, Wimmerova M, Budinska E, et al. Dirigent proteins in plants: modulating cell wall metabolism during abiotic and biotic stress exposure. J Exp Bot. 2017;68(13):3287-3301.
24. Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S, et al. The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet. 2012;44(10):1098-1103.
25. Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, Zhang J, Saski CA, Scheffler BE, Stelly DM, et al. Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. Nat Biotechnol. 2015;33(5):531-537.
26. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature. 2000;408:796–815.
27. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007;449(7161):463-467.
28. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al. Genome sequence of the palaeopolyploid soybean. Nature. 2010;463(7278):178-183.
29. Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, et al. The genome of *Theobroma cacao*. Nat Genet. 2011;43(2):101-108.
30. Yang Z, Ge X, Yang Z, Qin W, Sun G, Wang Z, Li Z, Liu J, Wu J, Wang Y, et al. Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. Nat Commun. 2019;10(1):2989.
31. Serna L, Martin C. Trichomes: different regulatory networks lead to convergent structures. Trends Plant Sci. 2006;11(6):274-280.
32. Guan X, Yu N, Shangguan X, Wang S, Lu S, Wang L, Chen X. *Arabidopsis* trichome research sheds light on cotton fiber development mechanisms. Chin Sci Bull. 2007;52(13):1734-1741.
33. Guan X, Pang M, Nah G, Shi X, Ye W, Stelly DM, Chen ZJ. miR828 and miR858 regulate homoeologous *MYB2* gene functions in *Arabidopsis* trichome and cotton fibre development. Nat Commun. 2014;5:3050.
34. Ma Z, He S, Wang X, Sun J, Zhang Y, Zhang G, Wu L, Li Z, Liu Z, Sun G, et al. Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. Nat Genet. 2018;50(6):803-813.

35. Gendreau E, Traas J, Desnos T, Grandjean O, Caboche M, Hofte H. Cellular basis of hypocotyl growth in *Arabidopsis thaliana*. Plant Physiol. 1997;114.
36. Boron AK, Vissenberg K. The *Arabidopsis thaliana* hypocotyl, a model to identify and study control mechanisms of cellular expansion. Plant Cell Rep. 2014;33(5):697-706.
37. Babushok DV, Ostertag EM, Kazazian HH, Jr. Current topics in genome evolution: molecular mechanisms of new gene formation. Cell Mol Life Sci. 2007;64(5):542-554.
38. Cannon SB, Mitra A, Baumgarten A, Young ND, May G. The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. BMC Plant Biol. 2004;4(1):10.
39. Rizzon C, Ponger L, Gaut BS. Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. PLoS Comput Biol. 2006;2(9):e115.
40. Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol. 2008;148(2):993-1003.
41. Fan L, Shi WJ, Hu WR, Hao XY, Wang DM, Yuan H, Yan HY. Molecular and biochemical evidence for phenylpropanoid synthesis and presence of wall-linked phenolics in cotton fibers. J Integr Plant Biol. 2009;51(7):626-637.
42. Han LB, Li YB, Wang HY, Wu XM, Li CL, Luo M, Wu SJ, Kong ZS, Pei Y, Jiao GL, et al. The dual functions of WLIM1a in cell elongation and secondary wall formation in developing cotton fibers. Plant Cell. 2013;25(11):4421-4438.
43. Gao Z, Sun W, Wang J, Zhao C, Zuo K. GhbHLH18 negatively regulates fiber strength and length by enhancing lignin biosynthesis in cotton fibers. Plant Sci. 2019;286:7-16.
44. Zhang Z, Ge Q, Liu A, Li J, Gong J, Shang H, Shi Y, Chen T, Wang Y, Palanga KK, et al. Construction of a high-density genetic map and its application to QTL identification for fiber strength in upland cotton. Crop Sci. 2017;57(2):774-788.
45. Jamshed M, Jia F, Gong J, Palanga KK, Shi Y, Li J, Shang H, Liu A, Chen T, Zhang Z, et al. Identification of stable quantitative trait loci (QTLs) for fiber quality traits across multiple environments in *Gossypium hirsutum* recombinant inbred line population. BMC Genomics. 2016;17:197.
46. Tang S, Teng Z, Zhai T, Fang X, Liu F, Liu D, Zhang J, Liu D, Wang S, Zhang K. Construction of genetic map and QTL analysis of fiber quality traits for Upland cotton (*Gossypium hirsutum* L.). Euphytica. 2015;201(2):195-213.
47. Shang L, Wang Y, Wang X, Liu F, Abduweli A, Cai S, Li Y, Ma L, Wang K, Hua J. Genetic analysis and QTL detection on fiber traits using two recombinant inbred lines and their backcross populations in upland cotton. G3 (Bethesda). 2016;6(9):2717-2724.
48. Tan Z, Fang X, Tang S, Zhang J, Liu D, Teng Z, Li L, Ni H, Zheng F, Liu D, et al. Genetic map and QTL controlling fiber quality traits in upland cotton (*Gossypium hirsutum* L.). Euphytica. 2014;203(3):615-628.

49. Zhao Q, Zeng Y, Yin Y, Pu Y, Jackson LA, Engle NL, Martin MZ, Tschaplinski TJ, Ding SY, Ragauskas AJ, et al. Pinoresinol reductase 1 impacts lignin distribution during secondary cell wall biosynthesis in *Arabidopsis*. Phytochemistry. 2015;112:170-178.
50. Tan Z, Zhang Z, Sun X, Li Q, Sun Y, Yang P, Wang W, Liu X, Chen C, Liu D, et al. Genetic map construction and fiber quality QTL mapping using the CottonSNP80K array in upland cotton. Front Plant Sci. 2018;9:225.
51. Waszczak C, Carmody M, Kangasjarvi J. Reactive oxygen species in plant signaling. Annu Rev Plant Biol. 2018;69:209-236.
52. Mnich E, Bjarnholt N, Eudes A, Harholt J, Holland C, Jorgensen B, Larsen FH, Liu M, Manat R, Meyer AS, et al. Phenolic cross-links: building and de-constructing the plant cell wall. Nat Prod Rep. 2020.
53. Niculaes C, Morreel K, Kim H, Lu F, McKee LS, Ivens B, Haustraete J, Vanholme B, Rycke RD, Hertzberg M, et al. Phenylcoumaran benzylic ether reductase prevents accumulation of compounds formed under oxidative conditions in poplar xylem. The Plant Cell. 2014;26(9):3775-3791.
54. Besseau S, Hoffmann L, Geoffroy P, Lapierre C, Pollet B, Legrand M. Flavonoid accumulation in *Arabidopsis* repressed in lignin synthesis affects auxin transport and plant growth. Plant Cell. 2007;19(1):148-162.
55. Gill US, Uppalapati SR, Gallego-Giraldo L, Ishiga Y, Dixon RA, Mysore KS. Metabolic flux towards the (iso)flavonoid pathway in lignin modified alfalfa lines induces resistance against *Fusarium oxysporum* f. sp. medicaginis. Plant Cell Environ. 2018;41(9):1997-2007.
56. Peer WA, Murphy AS. Flavonoids and auxin transport: modulators or regulators? Trends Plant Sci. 2007;12(12):556-563.
57. Hovav R, Udall JA, Chaudhary B, Hovav E, Flagel L, Hu G, Wendel JF. The evolution of spinnable cotton fiber entailed prolonged development and a novel metabolism. PLoS Genet. 2008;4(2):e25.
58. Tang W, Tu L, Yang X, Tan J, Deng F, Hao J, Guo K, Lindsey K, Zhang X. The calcium sensor GhCaM7 promotes cotton fiber elongation by modulating reactive oxygen species (ROS) production. New Phytol. 2014;202(2):509-520.
59. Tan J, Tu L, Deng F, Hu H, Nie Y, Zhang X. A genetic and metabolic analysis revealed that cotton fiber cell development was retarded by flavonoid naringenin. Plant Physiol. 2013;162(1):86-95.
60. Yang J, Zhang Y, Wang X, Wang W, Li Z, Wu J, Wang G, Wu L, Zhang G, Ma Z. *HyPRP1* performs a role in negatively regulating cotton resistance to *V. dahliae* via the thickening of cell walls and ROS accumulation. BMC Plant Biol. 2018;18(1):339.
61. Yang J, Ma Q, Zhang Y, Wang X, Zhang G, Ma Z. Molecular cloning and functional analysis of *GbRVd*, a gene in *Gossypium barbadense* that plays an important role in conferring resistance to Verticillium wilt. Gene. 2016;575(2 Pt 3):687-694.
62. Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J, et al. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature. 2012;492(7429):423-427.

63. Du X, Huang G, He S, Yang Z, Sun G, Ma X, Li N, Zhang X, Sun J, Liu M, et al. Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. Nat Genet. 2018;50(6):796-802.
64. Hu Y, Chen J, Fang L, Zhang Z, Ma W, Niu Y, Ju L, Deng J, Zhao T, Lian J, et al. *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. Nat Genet. 2019;51(4):739-748.
65. Zhu T, Liang C, Meng Z, Sun G, Meng Z, Guo S, Zhang R. CottonFGD: an integrated functional genomics database for cotton. BMC Plant Biol. 2017;17(1):101.
66. Wang Z, Hobson N, Galindo L, Zhu S, Shi D, McDill J, Yang L, Hawkins S, Neutelings G, Datla R, et al. The genome of flax (*Linum usitatissimum*) assembled *de novo* from short shotgun sequence reads. The Plant Journal. 2012;72(3):461-473.
67. Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S, et al. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice. 2013;6(1):4.
68. Bolser D, Staines DM, Pritchard E, Kersey P. Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomics Data. Methods Mol Biol. 2016;1374:115-140.
69. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40(D1):D1178-D1186.
70. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42(D1):D222-D230.
71. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011;39(suppl_2):W29-W37.
72. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, et al. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. 2020;48(D1):D265-D268.
73. Almagro Armenteros JJ, Tsirigos KD, Sonderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 2019;37(4):420-423.
74. Briesemeister S, Rahnenfuhrer J, Kohlbacher O. YLoc--an interpretable web server for predicting subcellular localization. Nucleic Acids Res. 2010;38(suppl_2):W497-W502.
75. Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, Xia R. TBtools: an integrative toolkit developed for interactive analyses of big biological data. Mol Plant. 2020;13(8):1194-1202.
76. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37(suppl_2):W202-W208.
77. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 2016;33(7):1870-1874.

78. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47(W1):W256-W259.
79. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40(7):e49.
80. Zhang Z, Xiao J, Wu J, Zhang H, Liu G, Wang X, Dai L. ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. Biochem Biophys Res Commun. 2012;419(4):779-781.
81. Zhang Z, Li J, Zhao X-Q, Wang J, Wong GK-S, Yu J. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics. 2006;4(4):259-263.
82. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, Bessy A, Cheneby J, Kulkarni SR, Tan G, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 2018;46(D1):D260-D266.
83. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27(7):1017-1018.
84. Collins TJ. ImageJ for microscopy. Biotechniques. 2007;43(1 Suppl):25-30.

Figures

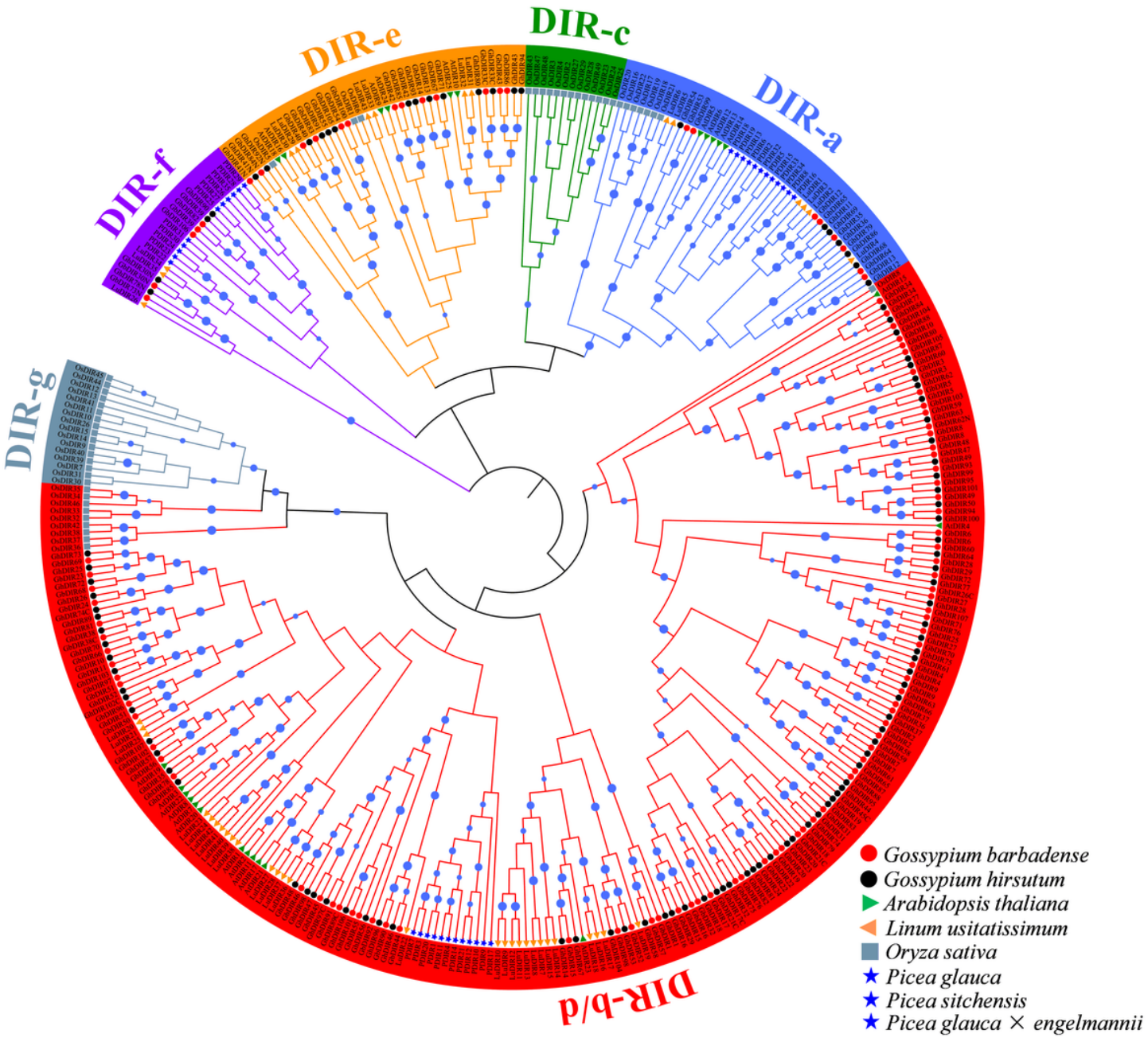

Figure 1

Phylogenetic relationships of DIR genes from G. barbadense, G. hirsutum, A. thaliana, L. usitatissimum, O. sativa and Picea. The phylogenetic tree was constructed by MEGA 7.0 using the neighbor-joining (NJ) method with 1000 bootstrap replicates. Bootstrap values > 50% are shown by blue solid circles on the branches, and their sizes reflect the reliability.

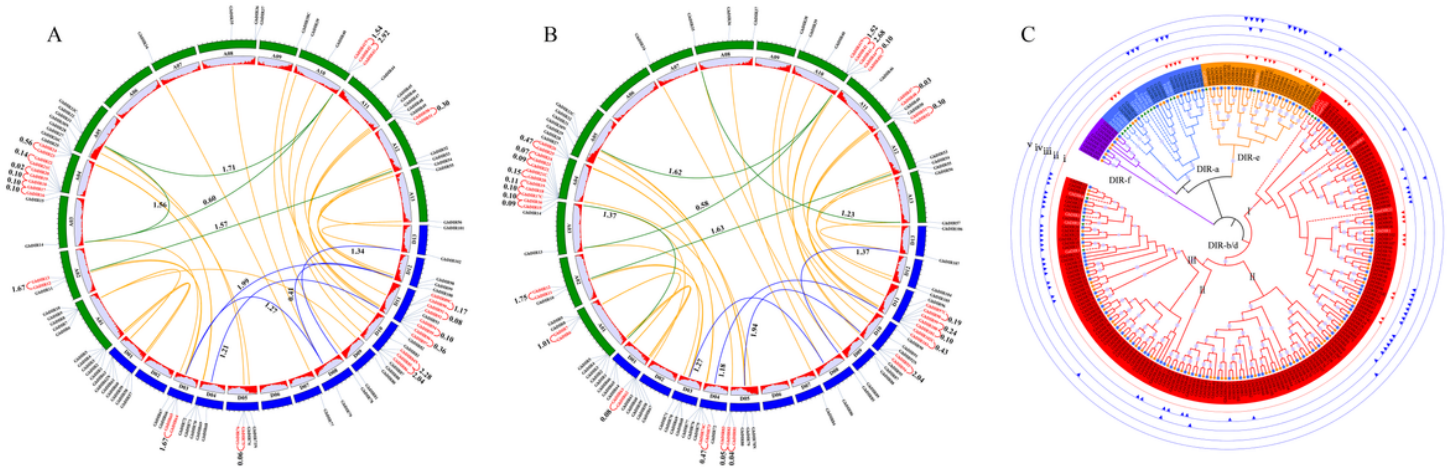

Figure 2

Duplication events among DIR genes. (a) Segmental and tandem duplications in G. barbadense. Green and blue lines indicate segmental duplications within At and Dt chromosomes, respectively. Red lines indicate tandem duplications. Ks values are displayed next to the lines. Orange lines link homologous genes between At and Dt subgenome. (b) Segmental and tandem duplications in G. hirsutum. (c) The expansion of cotton DIR gene family. The phylogenetic tree consists of 107 GbDIRs, 107 GhDIRs and 23 reported DIRs (in a white font). Track i marks segmental duplications, while tracks ii, iii, iv and v mark tandem duplications with Ks ranges of 0.0-0.2, 0.2-0.6, 0.6-1.9 and 1.9-3.0, respectively.

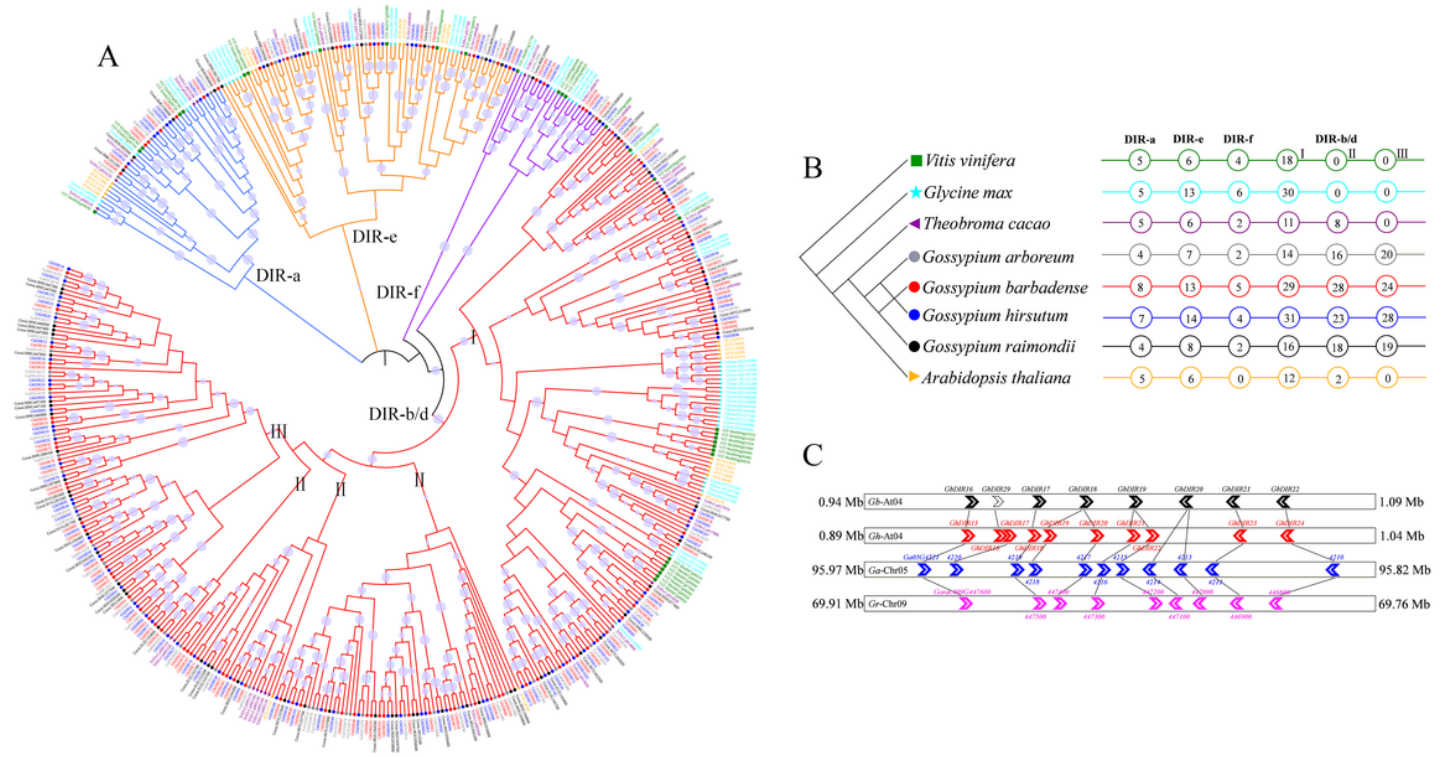

Figure 3

Evolutionary history of cotton DIR gene family. (a) Phylogenetic relationships of DIR genes from V. vinifera, G. max, A. thaliana, T. cacao, G. arboreum, G. raimondii, G. barbadense and G. hirsutum. (b) DIR genes were assigned to distinct clades. The figures in circles stand for the number of genes. (c) The collinearity among multiple Gossypium genomes. Lines between chromosomes show orthologous genes.

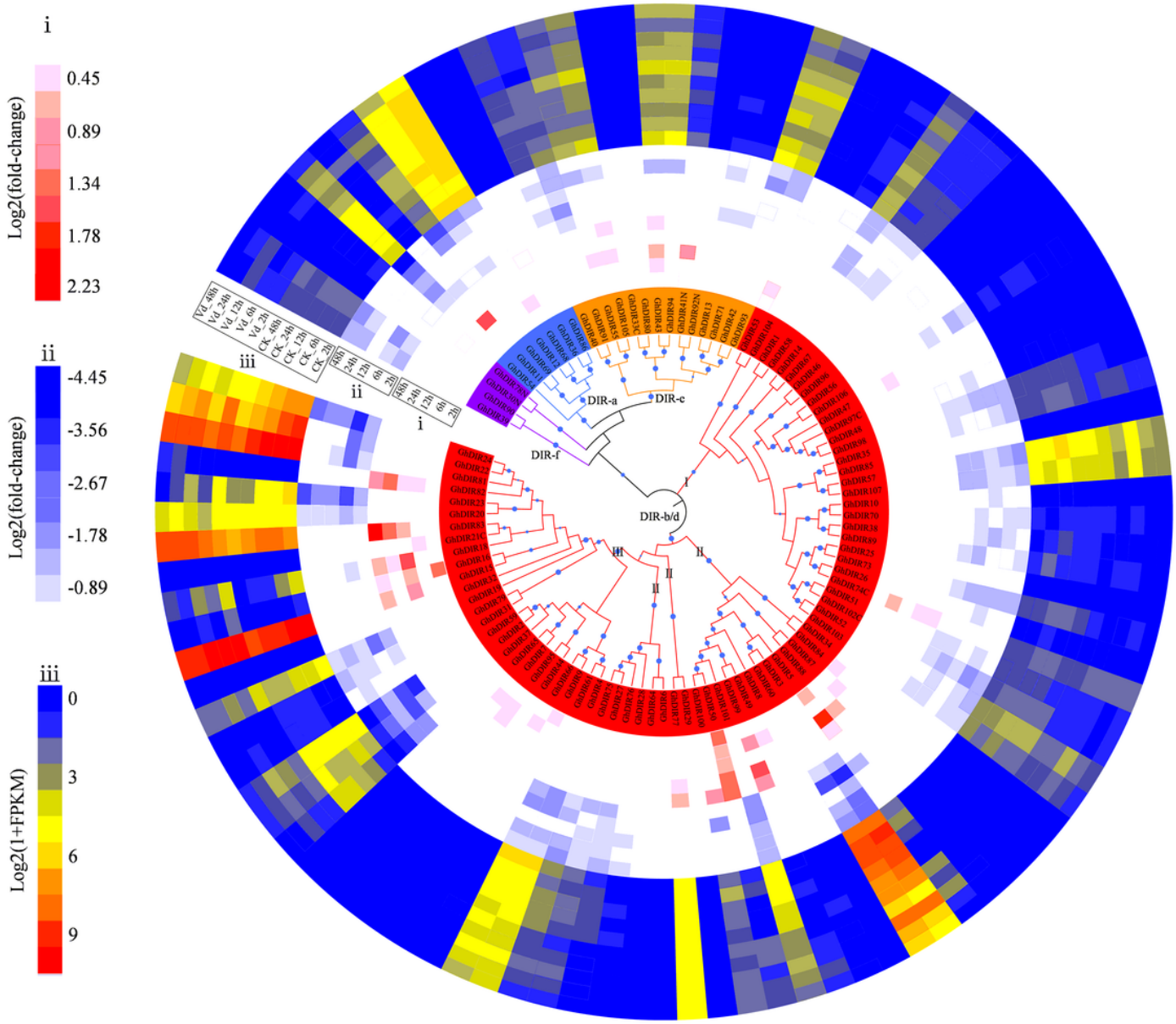

Figure 4

Circular heatmap depicting the expression patterns of GhDIRs in response to V. dahliae and water (CK). RNA-Seq data were generated from the roots of two-week-old cotton seedlings. Panel i shows up-regulated genes after inoculation with V. dahliae (compared with the CK group), while panel ii shows down-regulated genes. Panel iii shows V. dahliae- and water-responsive expression levels. The original FPKM values were provided as Additional file 8: Table S11.

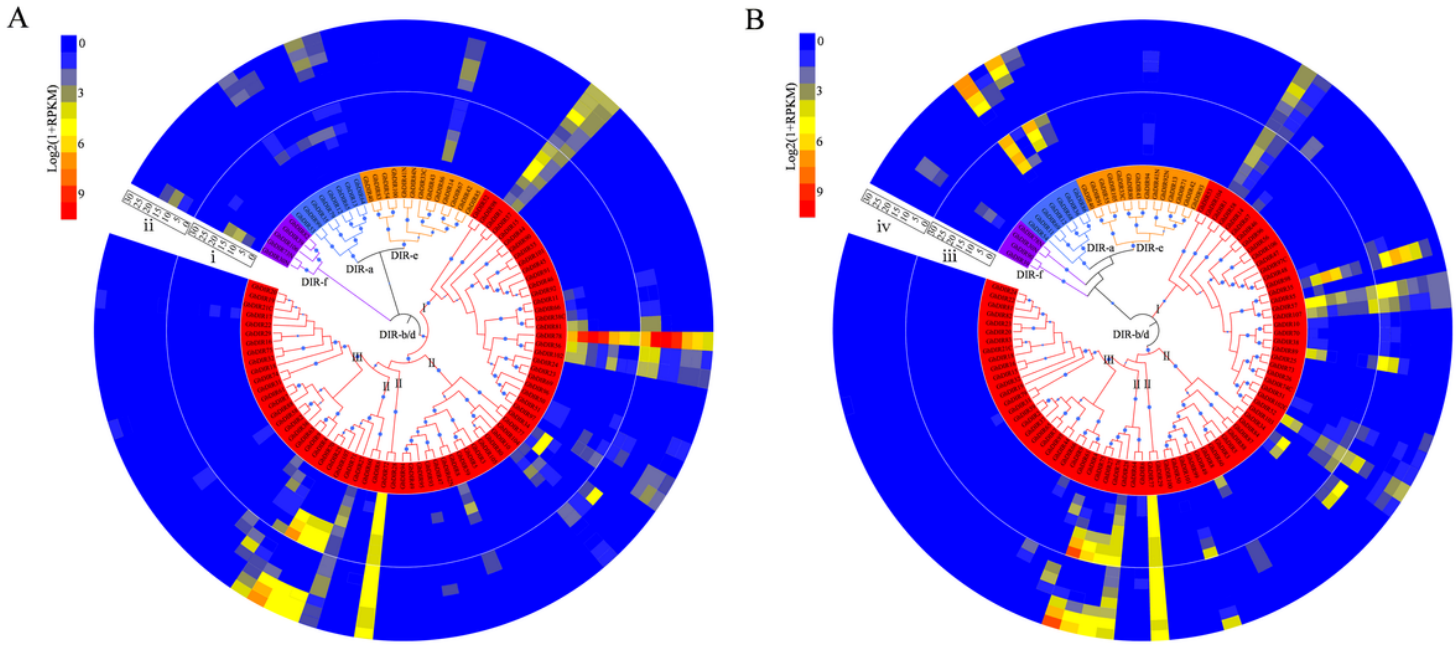

Figure 5

Expression patterns of DIR genes in cotton fibers. (a) Expression data from G. barbadense cultivars Pima90-53 (i) and Hai7124 (ii). (b) Expression data from G. hirsutum cultivars HY405 (iii) and ND13 (iv). 0-30 correspond to the number of days post-anthesis. The original RPKM values were provided as Additional file 8: Table S13 and S14.

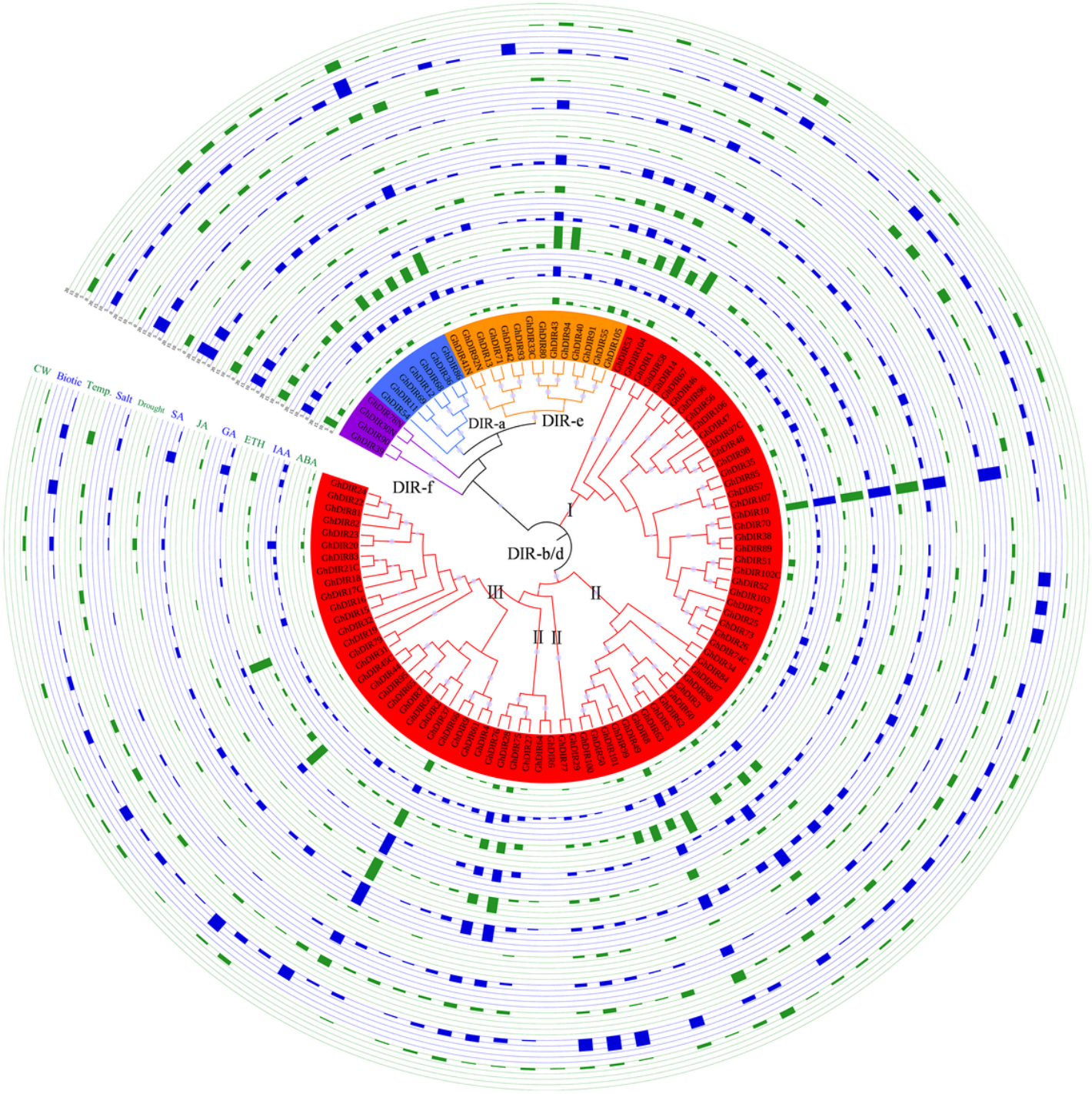

Figure 6

Identification of TFBS in the promoter regions of GhDIRs. A total of 266 JASPAR matrices were selected. FIMO was used to identify potential sites with a threshold of p-value < 1E-5. Histograms represent the number of TFBS in each category.

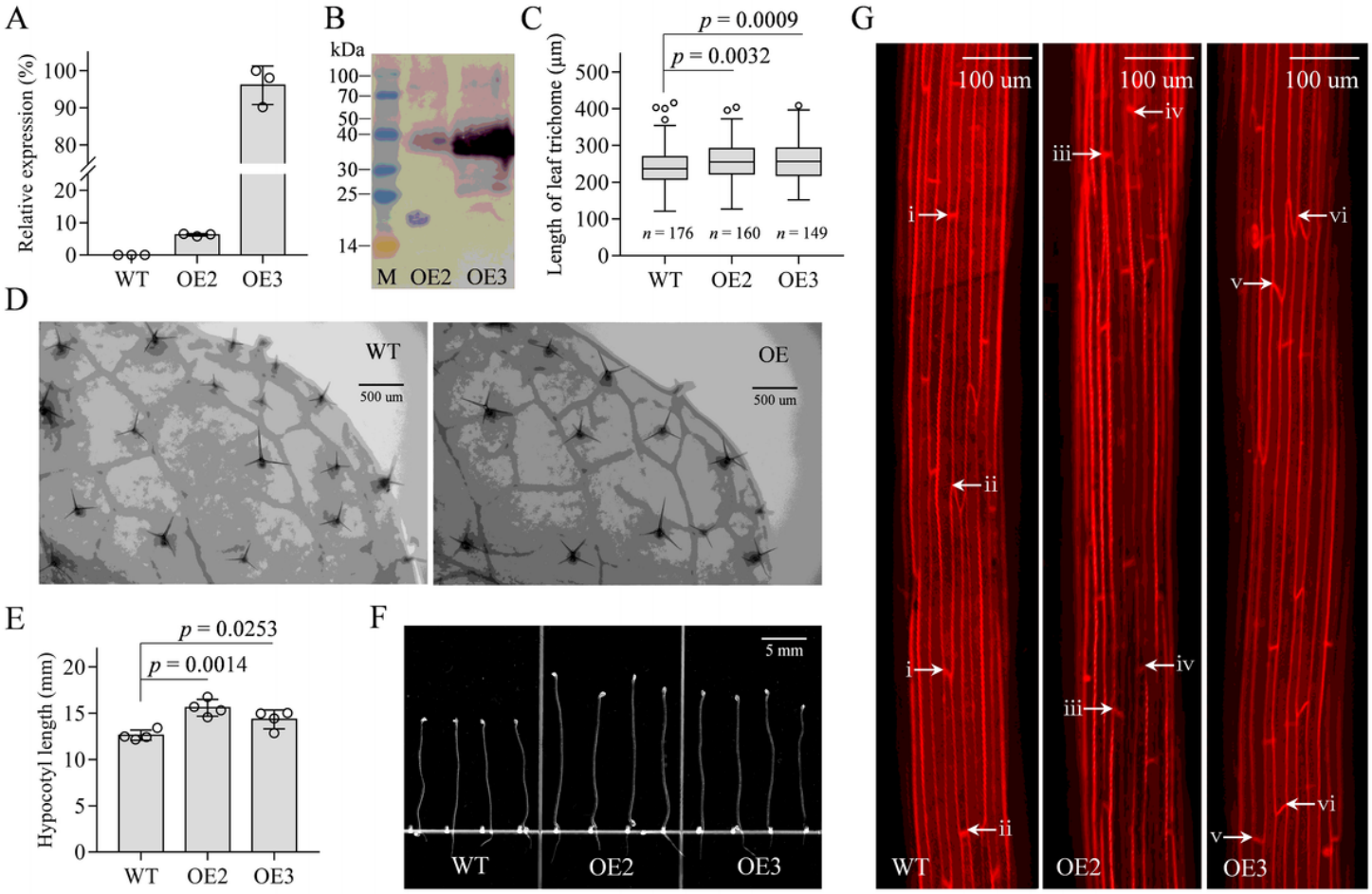

Figure 7

Functional characterization of GbDIR78 in Arabidopsis. (a) Expression of GbDIR78 in the transgenic Arabidopsis, tested with qRT-PCR. (b) Western blot of GbDIR78 in the transgenic Arabidopsis. (c-d) Arabidopsis with overexpression of GbDIR78, compared with wild type, showed longer leaf trichomes. In the box plots, the median, upper and lower quartiles, and the range of normal values are denoted by the center line, box limits, and whiskers, respectively. n indicates the number of measured trichomes. The significance of difference was analyzed with two tailed t test. (e-f) Arabidopsis with overexpression of GbDIR78, compared with wild type, showed significantly longer dark-grown hypocotyls. (g) Microscopic inspection of hypocotyl epidermal cells. i-vi indicate the ends of epidermal cells. The original, unprocessed images were provided as Additional file 9.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigureS1.tif
- FigureS2.tif
- FigureS3.tif

- FigureS4.tif
- FigureS5.tif
- FigureS6.tif
- FigureS7.tif
- TableS1.xlsx
- TableS2.xlsx
- TableS3.xlsx
- TableS4.xlsx
- TableS5.xlsx
- TableS6.xlsx
- TableS7.xlsx
- TableS8.xlsx
- TableS9.xlsx
- TableS10.xlsx
- TableS11.xlsx
- TableS12.xlsx
- TableS13.xlsx
- TableS14.xlsx
- Theoriginalunprocessedimages.docx