Genome-wide identification of expansin gene family reveals expansin genes are involved in fibre cells growth in cotton

Limin Lv
Institute of Cotton Research of CAAS
Dongyun Zuo
Institute of Cotton Research of CAAS
Xingfen Wang
Hebei Architectural University
Hailiang Cheng
Institute of Cotton Research of CAAS
Youping Zhang
Institute of Cotton Research of CAAS
Qiaolian Wang
Institute of Cotton Research of CAAS
Guoli Song (songguoli@caas.cn)
Cotton Research Institute https://orcid.org/0000-0003-3236-9286
Zhiying Ma
Hebei Agricultural University

Research article

Keywords: Expansins; gene family; fibre; gene expression profiles; Gossypium hirsutum

Posted Date: January 23rd, 2020

DOI: https://doi.org/10.21203/rs.2.14830/v3

Version of Record: A version of this preprint was published at BMC Plant Biology on May 19th, 2020. See the published version at https://doi.org/10.1186/s12870-020-02362-y.

Abstract

Background : Expansins (EXPs), a group of proteins that loosen plant cell walls and cellulosic materials, are involved in regulating cell growth and diverse developmental processes in plants. However, the biological functions of this gene family are still unknown in cotton. Results: In this paper, we identified a total of 93 expansin genes in Gossypium hirsutum . These genes were classified into four subfamilies, including 67 GhEXPAs , 8 GhEXPBs , 6 GhEXLAs , and 12 GhEXLBs , and divided into 15 subgroups. All 93 expansin genes are distributed over 24 chromosomes excluding Ghir_A02 and Ghir_D06. All GhEXP genes contain multiple exons and each GhEXP protein has multiple conserved motifs. Transcript profiling and qPCR analysis revealed that the expansin genes have distinct expression patterns in different stages of cotton fibre development. Among them, 3 genes (GhEXPA4o , GhEXPA1A , and GhEXPA8h) were highly expressed in the initiation stage, 9 genes (GhEXPA4a , GhEXPA13a , GhEXPA4f , GhEXPA4q , GhEXPA8f , GhEXPA2 , GhEXPA8g , GhEXPA8a , and GhEXPA4n) had high expression during the fast elongation stage, while GhEXLA1c and GhEXLA1f were preferentially expressed in the transition stage of fibre development. Conclusions: Our results provide a solid basis for further elucidation of biological functions of expansin genes in cotton fibre development and valuable genetic resources used for crop improvement in the future.

Background

Expansins are a kind of cell wall loosening protein, widely present in the higher plants and bacteria and fungi. Expansins may unlock the network of wall polysaccharides without lytic activity, permitting turgor-driven cell enlargement [1]. Plant expansins are usually 250 to 275 amino acid residues in length and the majority have a signal peptide in the N-terminus; the signal peptides are usually 20 to 30 amino acid residues [2, 3]. Typical structures of plant expansin are torpedo-shaped proteins containing two domains, domain I and domain II. Domain I is a six-stranded double-psi beta-barrel (DPBB), which has similar characteristics to the catalytic domain of glycoside hydrolase family 45 proteins (GH45) and contains a conserved His-Phe-Asp (HFD). The DPBB domain does not, however, possess the same catalytic activity as GH45. Domain II is homologous to group-2 grass pollen allergens [2], and it was recently classified as a family-63 carbohydrate binding module (CBM63) [2, 4].

The plant expansins superfamily is divided into four subfamilies, which include α-expansin (EXPA), β-expansin (EXPB), expansin-like A (EXLA), and expansin-like B (EXLB). Expansins were first identified as the endogenous proteins inducing cell wall extension in plants by the McQueen-Mason group [5]. At present, it is proved have shown that they can participate in many developmental processes and function in cell growth and enlargement, pollen tube invasion of the stigma (in grasses), wall disassembly during fruit ripening, abscission, stress resistance, and other cell separation events [1, 2, 6, 7].

Cotton fibres are single-celled trichomes that differentiate from the ovule epidermis, which is a powerful cell expansion and wall biogenesis research model system [8]. The process of cotton fibre development can be divided into five stages: initiation, elongation, transition, secondary wall synthesis and maturation

[9]. Some *expansin* genes preferentially expressed in cotton fibres were isolated and identified using several different approaches, including cDNA arrays, subtractive PCR, RT-PCR, and so on [10, 11]. The functions of *expansin* genes have been further investigated in cotton fibre development. Over-expression *GhEXPA8* can improve cotton fibre length and micronaire value [12]. *GbEXPATR*, a *Gossypium barbadense*-specific expansin, can also enhance cotton fibre elongation through cell wall restructuring [13]. *GhRDL1* is localized in the cell wall and interacts with *GhEXPA1*, and cotton plants overexpressed *GhRDL1* and *GhEXPA1* can increase the fibre length and produce many more cotton bolls [14]. *GhEXPA1* expression levels are regulated by the transcript factor *GhHOX3*, which can promote cotton fibre elongation [15].

Thus expansin have played an important role in cotton fibre development. The cotton genome has been sequenced and re-sequenced in succession [16-19]. These genome data make genome-wide identification of gene families possible. *Gossypium. hirsutum* owned the complex allotetraploid genome (AADD; 2n=52), which were doubled by two diploid cotton species, *Gossypium. arboreum* (AA; 2n=26) and *Gossypium. raimondii* (DD; 2n=26) genomes [17, 20, 21]. At present, upland cotton accounts for 90% of natural fibre production in the world. Hence, we mainly conducted whole *expansin* gene family of *G.hirsutum* in this paper. This research can provide genome-wide information of cotton *expansin* genes and promote further investigation of the biological function of *expansin* genes during cotton fibre development and other developmental processes.

Results

Identification and sequence analysis of the cotton expansin gene family

We have identified expansin gene family in *G. hirsutum* genome. As a result, candidate 98 *expansin* genes were initially obtained. According to the analysis of expansins conserved domain, the results showed that 93 *expansin* genes with both DPBB-1 (domains I) and Pollen_allerg_1 (domain II) domain were ultimately confirmed for further analysis. Each *expansin* gene was named according to nomenclature guidelines. The detailed results were shown in Table S1. The expansin gene family contained four subfamilies, including EXPA, EXPB, EXLA, and EXLB. As to the expansin gene family in *Gossypium. arboreum* and *Gossypium. raimondii*, the same analysis methods were performed. As a result, total 49 and 45 *expansin* genes were identified in *G. arboreum* and *G. raimondii* genome, respectively. These *expansin* genes were also divided into 4 subfamilies. The detailed results were shown in Table S2 and Table S3.

We have analysed the biochemical properties of expansin proteins (Table S1). The pI values of expansin family members ranged from 4.65 (GhEXLB1l) to 12.01 (GhEXPA4c), with an average of 8.47. The pI values of EXPA and EXLA members were above 7.0 except for GhEXPA8b and GhEXPA7d. However, the pI values of EXPBs and EXLBs were below 7.0 except for GhEXPB3a, GhEXPB3b, GhEXPB1a, GhEXPB1b, GhEXLB1d, and GhEXLB1j (Additional file1: Table S1). The average MW of expansin family members was 27.42 kD, ranging from 14.29 (GhEXPA17e) to 41.53 (GhEXPA5e) kD. The length of expansin protein

sequences ranged from 150 (GhEXLA17e) amino acids (aa) to 366 aa (GhEXPA5e), and the signal peptide size ranged from 17 (GhEXPA15d, GhEXPA15g and GhEXLA1d) to 35 (GhEXLA1a) aa in length (Additional file1:Table S1).

The multiple sequence alignment results of 93 expansin protein from *G. hirsutum* showed that they had similar sequence characteristics: the majority of them consist of a signal peptide, conserved domains I and II (Additional file 2: Figure S1), which was consistent with the previous study [3]. The amino acid sequence of domain I was more conserved than that of domain II, especially among EXPA members (Additional file 2: Figure S1). Notably, almost all of the EXPAs (excluding GhEXPA13a, GhEXPA13b, and GhEXPA15d) and three EXLA members (GhEXLA17a, GhEXLA17b, and GhEXLA17c) contained a conserved motif (HFD) in domain I (Additional file 2: Figure S1). Members of EXPB, EXLB, and the other six EXLA members did not have the HFD motif. Six EXLA contained an extra segment named EXLA extension of the C-terminus. Of the EXLA extension sequence, the feature only possessed in EXLA subfamily, the amino acid sequences of EXLA extension showed as follow: "DIAK(Q)EGCS(F)P(H)CDD(Y)S(G)H(N)WR(-)". In addition, a conserved motif named BOX 1 was found in almost all the expansin members (Additional file 2: Figure S1).

Phylogenetic relationships, genes structure and protein motifs of the cotton *expansin* genes

In order to evaluate the evolutionary relationships of cotton expansins, a phylogenetic tree was constructed. The expansins were divided into four major subfamilies, EXPA, EXPB, EXLA and EXLB. The EXPA subfamily was the largest group with 67 members, and the other subfamilies contained eight (EXPB), six (EXLA), and 12 (EXLB) members. The four expansin subfamilies comprised 15 subgroups (Fig. 1). We discovered that EXPA-IV was the largest subgroup, which included 17 expansin members, and EXPA-VII, EXPA-VIII, and EXPA-IX were the smallest subgroups with only two expansin members each.

The results of gene structure (exon-intron organization) analysis showed that the expansin members included two to five exons, and the same subfamilies had similar characteristics of exon types (Fig. 2a, b). Most of the *EXPA* members had three exons (51 of 67 *EXPA* members). 12 *EXPAs* had two exons, and four *EXPAs* had four exons. All members from *EXPB* subfamily had four exons except for *GhEXPB1a* (five exons). Four *EXLA* members contained five exons and two members had four exons. *EXLB* members had four (seven *EXLBs*) or five exons (five *EXLBs*).

We have identified the conserved motifs in expansin proteins. As a result, a total of ten distinct motifs were identified (Fig. 2c, Additional file 2: Figure S2). The motifs of all cotton expansins had unifying features; for example, each expansin protein contained motif 5 and almost all of them contained motif 4 except for GhEXPA15d, GhEXPA17e, and GhEXLB1j. In addition, the type, arrangement, and number of the motifs had similar characteristics in the same subfamily. More than half of the EXPA members (38/67) had seven motifs, and 21 members had six motifs. EXPB, EXLA, and EXLB subfamilies possessed similar motif characteristics and most of them contained five motifs (motifs 4, 5, 7, 8, and 9). GhEXPB2d and GhEXPB3b of the EXPB subfamily included four motifs, and three members (GhEXLB1g, GhEXLB1c, and GhEXLB1j) of EXLB also had the same number of motifs. These results showed that EXPB, EXLA, and

EXLB subfamilies had close evolutionary relationships. The similarities between gene structures and sequence motifs implied that cotton expansin family genes had duplication events over evolutionary time

Chromosomal location and collinear analysis of the expansin gene family

The chromosomal location of *GhEXP* genes was identified in *G. hirsutum*. The results were shown in Fig. 3. Total 93 *expansin* genes were distributed on 24 chromosomes, excluding Ghir_A02 and Ghir_D06. The chromosome Ghir_A05 contained eight *expansin* genes, whereas Ghir_A06 included only one *expansin* gene. The numbers of *expansin* genes located on other chromosomes ranged from two to seven. In addition, some of the *expansin* genes were located on the chromosome in clusters, for example, both Ghir_A08 and Ghir_D08 possessed a gene cluster with four distinct EXLBs (Fig. 3). These results showed that the *expansin* genes unevenly distributed on each chromosome. Collinearity analysis showed that *expansin* genes were collinear frequently between A and D sub-genomes (Fig. 4), which indicated that *expansin* genes with collinear relationships may have a similar function.

Investigation of *cis*-acting elements in the promoter regions of *expansin* genes

We have identified the *cis*-acting regulatory elements of the cotton expansin gene family. The results showed that *cis*-acting regulatory elements of *expansin* genes were extremely diverse (Additional file 1: Table S4; Table S5). These elements were divided into 7 categories and 111 types, including 31 light responsive elements, 7 development-related elements, 13 hormone responsive elements, 5 environmental stress-related elements, 3 promoter-related element, 7 site-binding related elements and 44 other elements (no functions). Among them, the types of light and hormone responsive were especially abundant (Additional file 1: Table S4; Table S5).

All of 93 *GhEXP* genes possessed 15,200 elements, including 1,268 light responsive elements, 144 development-related elements, 779 hormone responsive elements, 409 environmental stress-related elements, 9416 promoter related elements, 81 site-binding related elements and other elements 3103 (Additional file 1: Table S4), respectively. Out of 93 *GhEXP* genes, 83 possessed Box 4 element (the part of a conserved DNA module involved in light responsiveness), 70 owned GT1-motif (light responsive element), 57 had G-box (*cis*-acting regulatory element involved in light responsiveness), with 70 enriched ABRE elements (the *cis*-acting element involved in the abscisic acid responsiveness), 75 contained ERE (ethylene-responsive element), 56 had TGACG-motif as well as TGACG-motif, of which were the *cis*-acting regulatory element involved in the MeJA-responsiveness, 73 harbored the ARE (cis-acting regulatory element essential for the anaerobic induction) and 40 possessed the MBS (MYB binding site involved in drought-inducibility). Moreover, these relatively abundant elements were also more conserved among the *GhEXPs* gene family. In addition, all the *GhEXP* genes contained the CAAT-box and TATA-box, which were the core elements of promoter in eukaryotes, and the number of them was also the largest (Additional file 1: Table S5).

Expression patterns of the *expansin* genes in cotton fibre

To comprehensively investigate the temporal expression patterns of the cotton expansin gene family, fibre samples of different developmental stages were used for transcriptome analysis. A heat map was constructed with these transcriptome data (Fig. 5). The 86 *expansin* genes displayed the different expression patterns. The remaining seven *expansin* genes were not detected in this transcriptome data. Although the expression patterns of *expansin* genes displayed obvious differences, clustered *expansin* genes generally possessed a similar expression pattern. For example, *GhEXPA1d*, *GhEXPA15d*, *GhEXPA15a*, *GhEXPA4o*, *GhEXPA4a* and *GhEXPA4b* were the preferentially expressed genes during fibre initiation and elongation stages (0 to 15 DPA), whereas *GhEXLA1f* and *GhEXLA1c* had higher expression during the middle and later cotton fibre development stages (after 15 DPA). In addition, *GhEXPA4f* and *GhEXPA2*, two homologous genes located on the A and D sub-genomes, were sharply up-regulated from 3 DPA with a very similar expression pattern (Fig. 5), suggesting that they may have close or complementary functions during cotton fibre development. To verify our transcriptome results, the *GhEXP* genes expression profiles were further confirmed using publicly available RNA-seq data. It was showed that the expression profiles of *GhEXP* genes were basically consistent with our transcriptome results (Fig. 5 and Figure S3).

In order not to miss the possible important *expansin* genes, we have also analysed the transcriptical levels of seven *expansin* genes which were not detected in the transcriptome (Fig. 5; Table S1). qRT-PCR showed that the seven *expansin* genes scarcely expressed exclusive *GhEXLB1h* with low expression level in ovules and fibres. In addition, we have found these genes can be detected in other tissues but expression levels were also not high (Additional file 2: Figure S4).

qRT-PCR analysis of the special *expansin* genes in cotton fibres

In order to further identify the key *expansin* genes involved in fibre cell growth, 14 *expansin* genes that are predominantly expressed in different stages of developmental cotton fibres were selected to verify their expression level using qRT-PCR experiment. These *expansin* genes were evidently up-regulated at the initiation, elongation, or transition stages (Fig. 6) and displayed almost consistent with the expression tendency when compared to transcriptome data (Additional file 2: Figure S5).

We found that *GhEXPA4o*, *GhEXPA1a*, and *GhEXPA8h* were predominantly expressed at 0 DPA (Fig. 6a), suggesting that these three genes may function in the initial stage of fibre cells. Nine *expansin* genes showed higher expression levels at the fibre elongation stages with distinct expression characteristics (Fig. 6b). The expression of *GhEXPA4a* reached a peak at 3 DPA and *GhEXPA13a* and *GhEXPA4f* peaked at 5 DPA. The expression levels of *GhEXPA4q*, *GhEXPA8f*, and *GhEXPA2* were the highest at 7 DPA and *GhEXPA8g*, *GhEXPA8a*, and *GhEXPA4n* peaked at 10 DPA (Fig. 6b). *GhEXPA4f* and *GhEXPA2* are homologous genes in allotetraploid cotton species that are located in the A and D sub-genomes of the 10^{th} chromosomes, respectively, and both genes have specific expression in cotton fibre cells. Moreover, *GhEXPA8a* and *GhEXPA8g* are two important genes that we have found during cotton fibre elongation. These results revealed that the expression peaks of the majority of genes appeared from 7 to 10 DPA, which are usually called the fast elongation stages. In addition, we obtained two *expansin* genes that

were predominantly expressed at transition stages, named *GhEXLA1c* and *GhEXLA1f* (Fig. 6c). Both of them belonged to the EXLA subfamily with unclear biological roles. The expression levels of *GhEXLA1c* and *GhEXLA1f* were the highest at 20 DPA, which is the transition stage of fibre cells from fast elongation to secondary cell wall synthesis.

To better understand the potential functions of 14 *expansin* genes, their expression profiles were detected in 11 different tissues, including roots, hypocotyls, stems, leaves, calycles, petals, pollens, stigmas, fibres from 0 DPA, 10 DPA and 20 DPA. The results showed that these genes presented distinct but partially overlapping expression patterns (Additional file 2: Figure S6).

Discussion

In this paper, we have firstly reported the expansin gene family in upland cotton, which included 93 members. All of them had two conserved domains, DPBB_1 and Pollen_allerg_1, consistent with the results of other crops, such as *A.thaliana* [3], tobacco [22], tomato [23] and Chinese jujube [24]. So they were typical plant expansin proteins [2]. Phylogenetic analysis revealed that 93 cotton expansins were divided into 15 subgroups of four subfamilies (Fig. 1). The number of expansin subgroups was consistent with the number of expansin ancestors including 15 to 17 *expansin* genes, and each of these ancestors evolved into an extant clade in the phylogenetic tree [25]. Thus, we speculated that each clade of the existing cotton expansin family might be extended by each clade ancestor. In addition, cotton *expansin* genes within every subfamily had the structural similarity, and they also showed structural difference among four subfamilies (Fig. 2b). The characteristics of structure and evolutionary ancestors were consistent with other plant expansin gene family reported [22-24]. In the same subfamily category and even subgroup, most members had almost the same conserved gene structure and motif distribution (Fig. 2 b, c), thus further confirming their close evolutionary relationships and phylogenetic classification [26].

Our study showed that the *EXPA* subfamily genes in cotton were significantly expanded, including 67 total *EXPAs* (Fig. 1); more *EXPAs* suggest important functions of this kind of *expansin* genes in cotton growth and development. Conversely, there are fewer members of the other three expansin subfamilies relative to *EXPAs*: there are 8 *EXPBs*, 6 *EXLAs* and 12 *EXLBs*. The number proportion of cotton *expansin* genes in each subfamily is almost consistent with other eudicots, such as *A.thaliana* (26 *EXPAs*, 6 *EXPBs*, 3 *EXLAs*, and 1 *EXLB*), grape (20 *EXPAs*, 4 *EXPBs*, 1 *EXLA*, and 4 *EXLBs*), jujube (19 *EXPAs*, 3 *EXPBs*, 1 *EXLA* and 7 *EXLBs*), and Chinese cabbage (39 *EXPAs*, 9 *EXPBs*, 2 *EXLAs*, and 3 *EXLBs*) [2, 3, 24, 27, 28]. The proportions are different in monocotyledons, such as rice (33 *EXPAs*, 18 *EXPBs*, 4 *EXLAs*, and 1 *EXLB*) and maize (36 *EXPAs*, 48 *EXPBs* and 4 *EXLAs*) [3, 29], with the most significant difference that the *EXPBs* are more numerous in monocotyledons than eudicots [2]. Furthermore, the number of *GhEXP* genes (93) was basically consistent of the sum 49 *GaEXPs* and 45 *GrEXPs* together. By comparative analysis, we have found that 4 expansin subfamilies also were existed in the *G. arboreum* (38 *EXPAs*, 4 *EXPBs*, 2 *EXLAs*, and 5 *EXLBs*) and *G. raimondii* (33 *EXPAs*, 4 *EXPBs*, 3 *EXLAs*, and 5 *EXLBs*) (Additional file 1: Table S2; Additional file 1: Table S3), the results indicated that A and D genomes of two ancestral

species were the donors of the modern allotetraploid *G. hirsutum* species [20]. These results provided significant insights into the evolution and functions of *expansin* genes in cotton.

Based on expression profiles of *expansin* genes, we have obtained the 14 predominant expression genes in distinct stages of cotton fibre development, including 12 *EXPAs* and 2 *EXLAs* (Fig. 6), excluding *EXPBs* and *EXLBs*. Three *EXPA* genes, *GhEXPA4o*, *GhEXPA1a*, and *GhEXPA8h*, were firstly obtained in the early phase of fibre development, and displayed high expression levels by qRT-PCR (Fig. 6a); however, their role in initial stages of fibre development still needs to be clarified. Moreover, we have also obtained nine *expansin* genes with higher expression level in the elongation stages (Fig. 6b). Among the nine *expansin* genes, *GhEXPA4f* and *GhEXPA2* have the highest transcriptional levels during cotton fibre development, and they were the most preferential expression genes in different tissues (Fig.6b; Additional file 2: Figure S6b). The expression level of *GhEXPA4f* was highly consistent with the reported *GhExp1*, which was highly transcript in the fibre [30]. And its homologous gene *GhEXPA2* showed a similar expression pattern, this result was basically identical to the *GhExp2* expression level [30]. Transgenic plants introduced *GhEXPA1* and its partner *GhRDL1* can promote cotton fibre yield [14], and overexpressed *GhEXPA8* can improve cotton fibre length [12]. By comparative analysis of gene sequences, we have confirmed *GhEXPA4f* (name in this study, GhirA10G15240), *GhExp1* [30] and *GhEXPA1* [14] are the same genes, as well as *GhEXPA2* (name in this study, GhirD10G12330) and *GhEXPA8*. Our qRT-PCR results have also reproduced the importance of the two *expansin* genes in cotton development. As for the other seven new *expansin* genes, which were predominantly expressed in elongation stages of cotton fibre development; the functions of these genes need to be further studied in terms of promoting fibre elongation. Interestingly, we have found that *GhEXPA8a* and *GhEXP8g* are homologous with *AtEXP8* in *A. thaliana*, *AtEXP8* can promote the hypocotyl elongation in *A. thaliana* [31]. This result implied that *GhEXPA8a* and *GhEXP8g* can promote the fibre cell elongation in cotton. The functional mechanism *GhEXPA8a* and *GhEXP8g* will be one of our important researches in the future. The above-mentioned *expansin* genes were the members of *EXPA* subfamily in the initial and elongation stages of fibre development. This may be due to the fact that there are more members of the *EXPA* (67/93) subfamily in the expansin family. Moreover, these data also suggested that *expansin* genes of the EXPA subfamily are essential in the cotton fibre development.

EXLA and EXLB were two smaller expansin subfamilies. Phylogenetic analysis showed that these proteins constitute separate and well-resolved groups, however their biological functions are uncertain [2]. In this paper, we found two *EXLA* genes, referred to as *GhEXLA1c* and *GhEXLA1f*, with higher expression at 20 DPA (Fig. 6c), which is the transition stage of fibre cells from fast elongation to secondary cell wall synthesis. These results suggested they were the important genes in the transition stage, in which the cellulose synthesis prepared well for the secondary wall thickening period. At present, there are relatively few studies on EXLA functions besides *AtEXLA2* in *Arabidopsis thaliana*. *AtEXLA2* was reported to have obvious expression in both the hypocotyl and root; over-expression of *AtEXLA2* resulted in slightly thicker walls in non-rapidly elongating etiolated hypocotyl cells [32]. Phylogenetic tree analysis showed *GhEXLA1c*, *GhEXLA1f* and *AtEXLA2* were of the EXLA subfamily (Fig. 1). These results suggested *GhEXLA1c* and *GhEXLA1f* could participate in the thickening course of cell wall. In addition, it was

reported that an expansin-like protein from *Hahella chejuensis* could bind cellulose and enhance cellulase activity [33]. It was implied that the two *EXLA* genes could facilitate the cellulose synthesis in the transition, however, the detailed biological functions of EXLAs remain to be assessed in cotton fibre development; more research needs to be conducted in order to understand and make use of EXLAs in cotton transition stages either in theory or practice.

Conclusions

Overall, we successfully performed a genome-scale analysis of the *expansin* family genes in upland cotton with a special emphasis on fibre development. A total of 93 cotton *expansin* genes were obtained. Our analysis has provided information for understanding the cotton expansin superfamily, including gene evolution, gene structure, protein motifs, collinear relationships, *cis*-acting elements and gene expression patterns. Moreover, we obtained expression patterns of 14 *expansin* genes in cotton fibre development at different stages. Among them, three genes were highly expressed in the initial stage, nine genes had high-level expression during the fast elongation stage, while *GhEXLA1c* and *GhEXLA1f* were preferentially expressed in the transition stage of fibre development. The results lay the foundation for further clarification of the biological functions of *expansin* genes and the molecular mechanism of many of the important cotton agricultural traits, especially on the elongation stage of cotton fibre development.

Methods

Plant materials

Gossypium hirsutum L. ('TM-1') seed was obtained from Institute of Cotton Research of Chinese Academy of Agricultural Sciences (Anyang, China). TM-1 was used as experimental material in this study. It was planted at the experimental farm (36°06'84.44"N, 114°49'61.5"E) of the Institute of Cotton Research of the Chinese Academy of Agricultural Sciences. To research expression patterns of *expansin* genes during cotton fibre development, each flower was labelled on the day of flowering, which was considered 0 days post anthesis (DPA). Subsequently, samples were collected at 0, 3, 5, 7, 10, 15, 20, and 30 DPA. The collected bolls were dissected to obtain ovules and fibres. For 0 to 3 DPA samples we collected ovules, and for 5 to 30 DPA samples we collected fibres. In addition, we have also collected the cotton tissue samples at different developmental stages, including roots, hypocotyls, stems, young leaves at seedling stage, and calycles, petals, pollens and stigmas at adult-plant stage. The different samples were frozen in liquid nitrogen immediately and stored at -80°C in an ultra-low temperature freezer after harvest.

Identification and sequence analysis of the cotton *expansin* genes

The cotton genome data were obtained from the Cotton FGD website [34] (https://cottonfgd.org/). Expansin protein sequences of *A. thaliana* were downloaded from the TAIR 10 (http://www.arabidopsis.org/). Firstly, we used the 35 EXPANSIN protein sequences from *A. thaliana* as queries

in searches against the *G. hirsutum* genome database [19], BlastP with default parameters was used to identify the expansin protein. After that, we searched the database for homologs using “EXPANSIN” as a keyword; finally, we used the previously reported 6 *GhEXPs* as the queries to search other possible *GhEXPs* by Blastp searches against cotton genome [30]. Redundant sequences were deleted after a comparison analysis for the expansin. Then, all candidate expansin protein sequences were submitted to the NCBI CDD (conserved domain database) (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) [35] where conserved domains were identified. For the sake of rapid search speed, we processed this work with the Batch web CD-search tool (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi) [35], where the maximal number of protein queries per request is 4000, providing adequate processing power for our purposes. We executed the search program using default parameters. The canonical expansin protein contains both conserved domains: DPBB-1 (including DPBB_1 superfamily) and Pollen_allerg_1 (including Pollen_allerg_1 superfamily). We acquired the final gene sequences of the upland cotton expansin family for further analysis.

Using the ExPaSy online tools [36] (https://www.expasy.org/resources/), we analysed the molecular properties of the identified expansin proteins, which were included to compute the molecular weight (MW) and isoelectric point (pI), and predicted their signal peptide sequences with SignalP 5.0 Server (http://www.cbs.dtu.dk/services/SignalP/). Sequence alignment of the expansin protein sequences was executed in Vector NTI Advance 11 software (version 11.5); followed by searching for conserved amino acid and conserved domain properties of expansin protein.

Phylogenetic tree construction

To analyse phylogenetic relationships, *Arabidopsis thaliana* (*A. thaliana*) expansin protein sequences were downloaded from TAIR (https://www.arabidopsis.org/) and EXPANSIN CENTRAL (http://www.personal.psu.edu/fsl/ExpCentral/). Multiple sequence alignment of the identified cotton expansin and *A. thaliana* expansin proteins was executed in MEGA software (version 6.0) [37], and a phylogenetic tree was constructed in the same software, using the neighbour-joining method. The number of bootstrap replications was 1000, and the rest of the parameters were set as the defaults.

Analysis of *expansin* gene structures and motifs

Analysis of gene structure was performed to identify exons, introns, and UTRs. Corresponding GFF data of identified *expansin* gene ID were extracted from GFF file named Ghirsutum_gene_model in the new cotton genome data [19] (http://cotton.hzau.edu.cn/EN/download.php), then the *expansin* GFF data were analysed using the online tool GSDS (version 2.0, http://gsds.cbi.pku.edu.cn/) [38]; the results were saved in SVG image format. Motifs of expansin protein sequences were analysed using the online tool MEME (http://meme-suite.org/index.html) [39]. According to the required file format, we submitted the expansin sequences into the online tool. The maximum number of motifs was set to 10, the repeat number was set to 0 or 1, the remainder of the parameters were set to system defaults. The output draft images of gene structure and motif were further modified with the Adobe Illustrator CS3 software (version 13.0.0).

Chromosomal locations and collinearity relationships of *expansin* genes

We obtained the length of each chromosome from the new genome data [19], and a file of the lengths of all TM-1 chromosomes was obtained. Then positional information of the *expansin* gene on the chromosome was extracted from the GFF file, named Ghirsutum_gene_model in the new cotton genome data (http://cotton.hzau.edu.cn/EN/download.php) [19]; thus a file of positional information of the *expansin* gene was obtained. Afterwards, the two files were submitted to the online tool MG2C (http://mg2c.iask.in/mg2c_v2.0/) for analysis of *expansin* gene location on chromosomes. Collinearity analysis of cotton *expansin* genes was executed by the MCScanX software [40], the visualization of analysis results was drawn using Circos software [41]. The analysed results were exported in SVG format, and the SVG image was further modified with the Adobe Illustrator CS3 software (version 13.0.0).

Analysis of *cis*-acting regulatory elements in the promoter regions of *expansin* genes

The promoter regions (2000 bp sequence upstream of the transcription start site in the genomic DNA sequence) of the cotton *expansin* genes were identified by searching the *G. hirsutumn* genome database (https://cottonfgd.org/) [34], and these promoter sequences were then predicted using PlantCARE (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/) to analyse the *cis*-acting elements [42].

Transcriptome analysis of the cotton *expansin* genes

Total RNAs were extracted from the samples of ovule and fibre. Samples from different stages were used for transcriptome analysis with the Illumina platform, including the ovules at 0, 3 DPA and fibres at 5, 7, 10, 15, 20, and 30 DPA. For the transcriptome sequencing, three biological duplicates were conducted for the experimental samples. The sequencing results of gene expression were shown by FPKM values (Fragments Per Kilobase Million mapped reads). Expression data FPKM values of *expansin* genes were screened for transcriptome results and converted to logFPKM values. Meanwhile, we have downloaded the RNA-seq data of cotton ovule and fibre from a public database, Cotton FGD website (https://cottonfgd.org/) [34], including the ovules at -3, 0, 1, 3 DPA, and fibres at 5, 10, 15, 20 DPA. The FPKM values processing were the same as above. The heatmaps were drawn with the logarithm values using HemI software (version 1.0,) [43].

RNA isolation and real-time quantitative-PCR analysis

The total RNA of samples from different stages was extracted using the RNAprep Pure Kit (for Polysaccharides & Polyphenolics-rich plants; Cat. no. DP441; TIANGEN, Beijing, China). The RNA samples were examined using agarose gel electrophoresis, then the concentration and quality were analysed with a NanoDrop ONE^{C} (Thermo Fisher Scientific, USA). cDNA synthesis was performed on the 2720 thermal cycler (Applied Biosystems, Thermo Fisher Scientific, USA), according to the instructions for the PrimeScript™ II 1st Strand cDNA Synthesis Kit (TaKaRa, Code No. 6210A). All reverse transcript cDNA samples were diluted 10 times and stored at -20°C for the real-time quantification PCR (qRT-PCR) experiment. The design of specific qRT-PCR primers was performed using Beacon Designer software

(version 8.0); all the primers are shown in Table S6. qRT-PCR reactions were performed with the TB Green™ Premix Ex Taq™ II kit (TaKaRa, Code No. RR820A) and conducted on the QuantStudioä 5 Real Time PCR instrument (Applied Biosystems, Thermo Fisher Scientific, USA). *UBQ7* (GenBank No. AY189972) was used as a reference gene to calculate relative expression levels. The data were analysed using the $2^{-\Delta\Delta CT}$ method [44].

Abbreviations

DPA: Days post anthesis; *EXPs*: *Expansins*; EXPA: α-expansin, EXPB: β-expansin, EXLA: expansin-like A; EXLB: expansin-like B; DPBB: double-psi beta-barrel.

Declarations

Acknowledgments

Not applicable.

Author Contributions

ZM and GS designed the research. LL, DZ and XW performed the research. HC, YZ and QW analyzed data. LL wrote the manuscript. All authors have read and approved the manuscript.

Funding

This study was financially supported by the National Key Research and Development Program of China (Grant No. 2018YFD0100402) and the National Natural Science Foundation of China (Grant No. 31621005). The funding bodies were not involved in the design of the study, collection, analysis, interpretation of data, or the manuscript writing.

Availability of data and materials

All of the data and materials supporting our research findings are contained in the methods section of the manuscript. Details are provided in the attached Additional files.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Cosgrove DJ. Loosening of plant cell walls by expansins. Nature. 2000;407(6802):321-6.
2. Cosgrove DJ. Plant expansins: diversity and interactions with plant cell walls. Curr Opin Plant Biol. 2015;25:162-72.
3. Sampedro J, Cosgrove DJ. The expansin superfamily. Genome Biol.2005, 6(12):242.
4. Georgelis N, Yennawar NH, Cosgrove DJ. Structural basis for entropy-driven cellulose binding by a type-A cellulose-binding module (CBM) and bacterial expansin. Proc Natl Acad Sci U S A. 2012;109(37):14830-5.
5. McQueen-Mason S, Durachko MD, and Daniel J. Cosgrove. Two Endogenous Pmteins That lnduce Cell Wall Extension in Plants. Plant Cell. 1992;4:1425-33.
6. Wei PC, Zhang XQ, Zhao P, Wang XC. Regulation of stomatal opening by the guard cell expansin *AtEXPA1*. Plant Signal Behav. 2011;6(5):740-2.
7. Li Y, Darley CP, Ongaro V, Fleming A, Schipper O, Baldauf SL, *et al*. Plant expansins are a complex multigene family with an ancient evolutionary origin. Plant Physiol. 2002;128(3):854-64.
8. 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.
9. Haigler CH, Betancur L, Stiff MR, Tuttle JR. Cotton fiber: a powerful single-cell model for cell wall and cellulose research. Front Plant Sci. 2012, 3:104.
10. An C, Saha S, Jenkins JN, Scheffler BE, Wilkins TA, Stelly DM. Transcriptome profiling, sequence characterization, and SNP-based chromosomal assignment of the EXPANSIN genes in cotton. Mol Genet Genomics. 2007;278(5):539-53.
11. Ji SJ, Lu YC, Feng JX, Wei G, Li J, Shi YH, *et al*. 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.
12. Bajwa KS, Shahid AA, Rao AQ, Bashir A, Aftab A, Husnain T. Stable transformation and expression of *GhEXPA8* fiber expansin gene to improve fiber length and micronaire value in cotton. Frontiers in plant science 2015, 6:838.
13. Li Y, Tu L, Pettolino FA, Ji S, Hao J, Yuan D, *et al. GbEXPATR*, a species-specific expansin, enhances cotton fibre elongation through cell wall restructuring. Plant Biotechnol J. 2016; 14(3):951-63.
14. Xu B, Gou JY, Li FG, Shangguan XX, Zhao B, Yang CQ, *et al.* A cotton BURP domain protein interacts with alpha-expansin and their co-expression promotes plant growth and fruit production. Mol Plant. 2013;6(3):945-58.
15. Shan CM, Shangguan XX, Zhao B, Zhang XF, Chao LM, Yang CQ, *et al.* Control of cotton fibre elongation by a homeodomain transcription factor *GhHOX3*. Nat Commun. 2014;5:5519.

16. Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, *et al.* Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. Nat Biotechnol. 2015;33(5):531-7.
17. Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, *et al.* Genome sequence of the cultivated cotton *Gossypium arboreum*. Nat Genet. 2014;46(6):567-72.
18. Ma Z, He S, Wang X, Sun J, Zhang Y, Zhang 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):803-13.
19. Wang MJ, Tu LL, Yuan DJ, Zhu D, Shen C, Li JY, *et al.* Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. Nat Genet. 2019;51:224–9.
20. Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, *et al.* Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. Nat Biotechnol. 2015;33(5):524-30.
21. Wang K, Wang Z, Li F, Ye W, Wang J, Song G, *et al.* The draft genome of a diploid cotton *Gossypium raimondii.* Nat Genet. 2012;44(10):1098-103.
22. Ding AM, Marowa P, Kong YZ. Genome-wide identification of the expansin gene family in tobacco (*Nicotiana tabacum*). Mol Genet Genomics. 2016;291(5):1891-907.
23. Lu YG, Liu LF, Wang X, Han ZH, Ouyang B, Zhang JH, *et al.* Genome-wide identification and expression analysis of the expansin gene family in tomato. Mol Genet Genomics. 2016;291(2):597-608.
24. Hou L, Zhang Z, Dou S, Zhang Y, Pang X, Li Y. Genome-wide identification, characterization, and expression analysis of the expansin gene family in Chinese jujube (*Ziziphus jujuba* Mill.). Planta. 2019;249(3):815-29.
25. Sampedro J, Lee Y, Carey RE, dePamphilis C, Cosgrove DJ. Use of genomic history to improve phylogeny and understanding of births and deaths in a gene family. Plant J 2005, 44(3):409-419.
26. Zhang S, Xu R, Gao Z, Chen C, Jiang Z, Shu H. A genome-wide analysis of the expansin genes in *Malus x Domestica*. Mol Genet Genomics. 2014;289(2):225-36.
27. Krishnamurthy P, Hong JK, Kim JA, Jeong MJ, Lee YH, Lee SI. Genome-wide analysis of the expansin gene superfamily reveals *Brassica* rapa-specific evolutionary dynamics upon whole genome triplication. Mol Genet Genomics. 2015;290(2):521-30.
28. Dal Santo S, Vannozzi A, Tornielli GB, Fasoli M, Venturini L, Pezzotti M, *et al.* Genome-wide analysis of the expansin gene superfamily reveals grapevine-specific structural and functional characteristics. Plos One. 2013;8(4):e62206.
29. Zhang W, Yan H, Chen W, Liu J, Jiang C, Jiang H, *et al.* Genome-wide identification and characterization of maize expansin genes expressed in endosperm. Mol Genet Genomics. 2014;289(6):1061-74.
30. Harmer SE, Orford SJ, Timmis JN. Characterisation of six alpha-expansin genes in *Gossypium hirsutum* (upland cotton). Mol Genet Genomics. 2002;268(1):1-9.

31. Ikeda M, Fujiwara S, Mitsuda N, Ohme-Takagi M. A triantagonistic basic helix-loop-helix system regulates cell elongation in *Arabidopsis*. Plant Cell. 2012;24(11):4483-97.
32. Boron AK, Van Loock B, Suslov D, Markakis MN, Verbelen JP, Vissenberg K. Over-expression of AtEXLA2 alters etiolated arabidopsis hypocotyl growth. Ann Bot. 2015;115(1):67-80.
33. Lee HJ, Lee S, Ko H-j, Kim KH, Choi I-G. An expansin-like protein from Hahella chejuensis binds cellulose and enhances cellulase activity. Mol Cells. 2010;29(4):379-85.
34. Zhu T, Liang CZ, Meng ZG, Sun GQ, Meng ZH, Guo SD, *et al.* CottonFGD: an integrated functional genomics database for cotton. BMC Plant Biol. 2017;17.
35. Marchler-Bauer A, Bo Y, Han LY, He JE, Lanczycki CJ, Lu SN, *et al.* CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 2017;45(D1):D200-03.
36. Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E, *et al.* ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res. 2012;40(Web Server issue):W597-603.
37. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol. 2013;30(12):2725-9.
38. Hu B, Jin J, Guo AY, Zhang H, Luo J, Gao G. GSDS 2.0: an upgraded gene feature visualization server. Bioinformatics. 2015;31(8):1296-7.
39. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, *et al.* MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37(Web Server issue):W202-8.
40. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40(7):e49.
41. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, *et al.* Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639-45.
42. Lescot M, Dehais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, *et al.* PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res. 2002;30(1):325-7.
43. Deng WK, Wang YB, Liu ZX, Cheng H, Xue Y. HemI: A Toolkit for Illustrating Heatmaps. Plos One. 2014;9(11).
44. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods. 2001;25(4):402-8.

Additional File Legends

Additional file 1: Table S1. Identification of *GhEXP* genes in *G. hirsutum*.

Additional file 1: Table S2. The *GaEXP* genes in *G. arboretum*.

Additional file 1: Table S3. The *GrEXP* genes in *G. raimondii*.

Additional file 1: Table S4. The summary of *cis*-acting elements of *GhEXP* genes Additional file 1: Table S5. The number of *cis*-acting elements involved in different biology process in the promoter region of *GhEXP* genes.

Additional file 1: Table S6. A list of primers used in qRT-PCR experiments.

Additional file 2: Figure S1. The multiple sequence alignment of 93 GhEXP proteins.

Additional file 2: Figure S2. Analysis of conserved motifs of *GhEXP* genes in cotton.

Additional file 2: Figure S3. Expression profiles of *GhEXP* genes in cotton ovule and fibre.

Additional file 2: Figure S4. Quantitative RT-PCR analysis of seven *GhEXP* genes in fiber of different stages and tissues.

Additional file 2: Figure S5. The qRT-PCR validation and transcriptome sequencing of 14 *GhEXPs* in different develop.

Additional file 2: Figure S6. Quantitative RT-PCR analysis of 14 cotton *GhEXP* genes in different tissues.

Figures

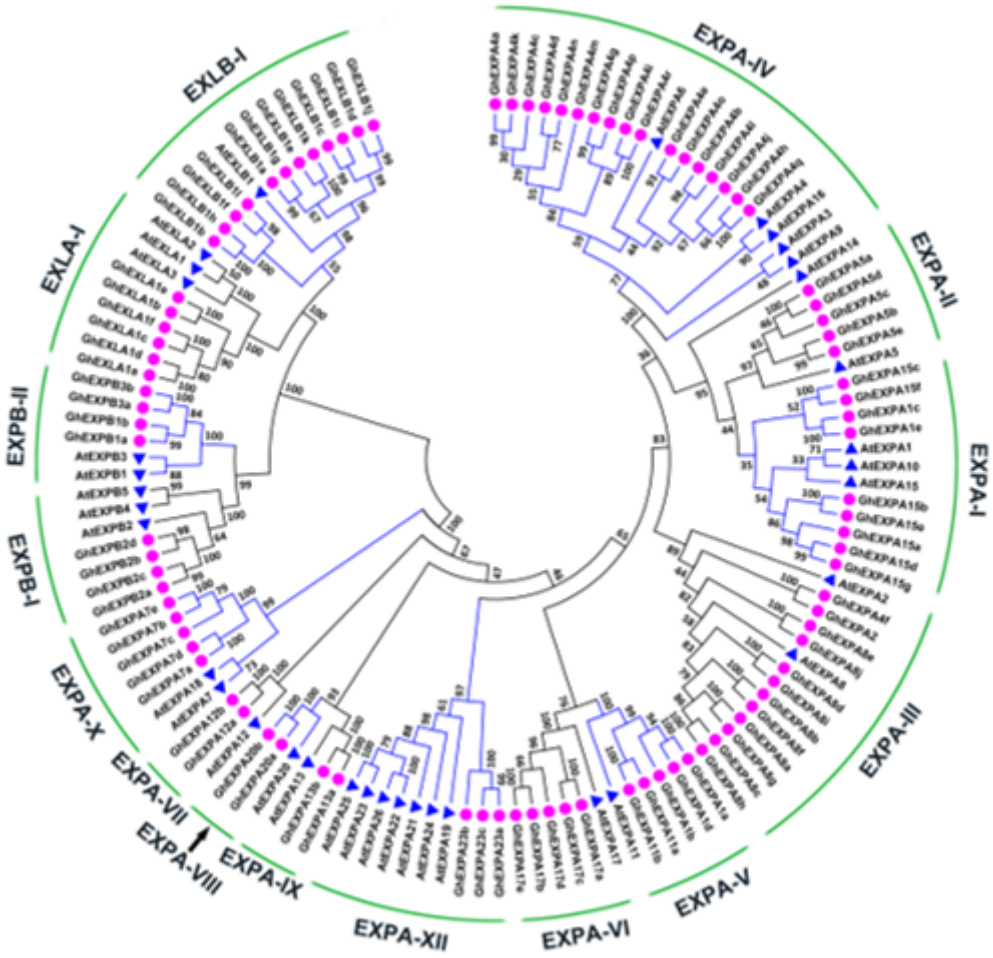

Figure 1

Phylogenetic analysis and subfamily classification of the expansin genes in cotton (GhEXPs). The phylogenetic tree was constructed with MEGA 6.0 software using the neighbour-joining (NJ) method with 1000 bootstrap replicates. The pink solid circles represent the cotton expansin genes from Gossypium hirsutum; the blue solid triangles represent the expansin genes from Arabidopsis thaliana. Gh, Gossypium hirsutum; At, Arabidopsis thaliana

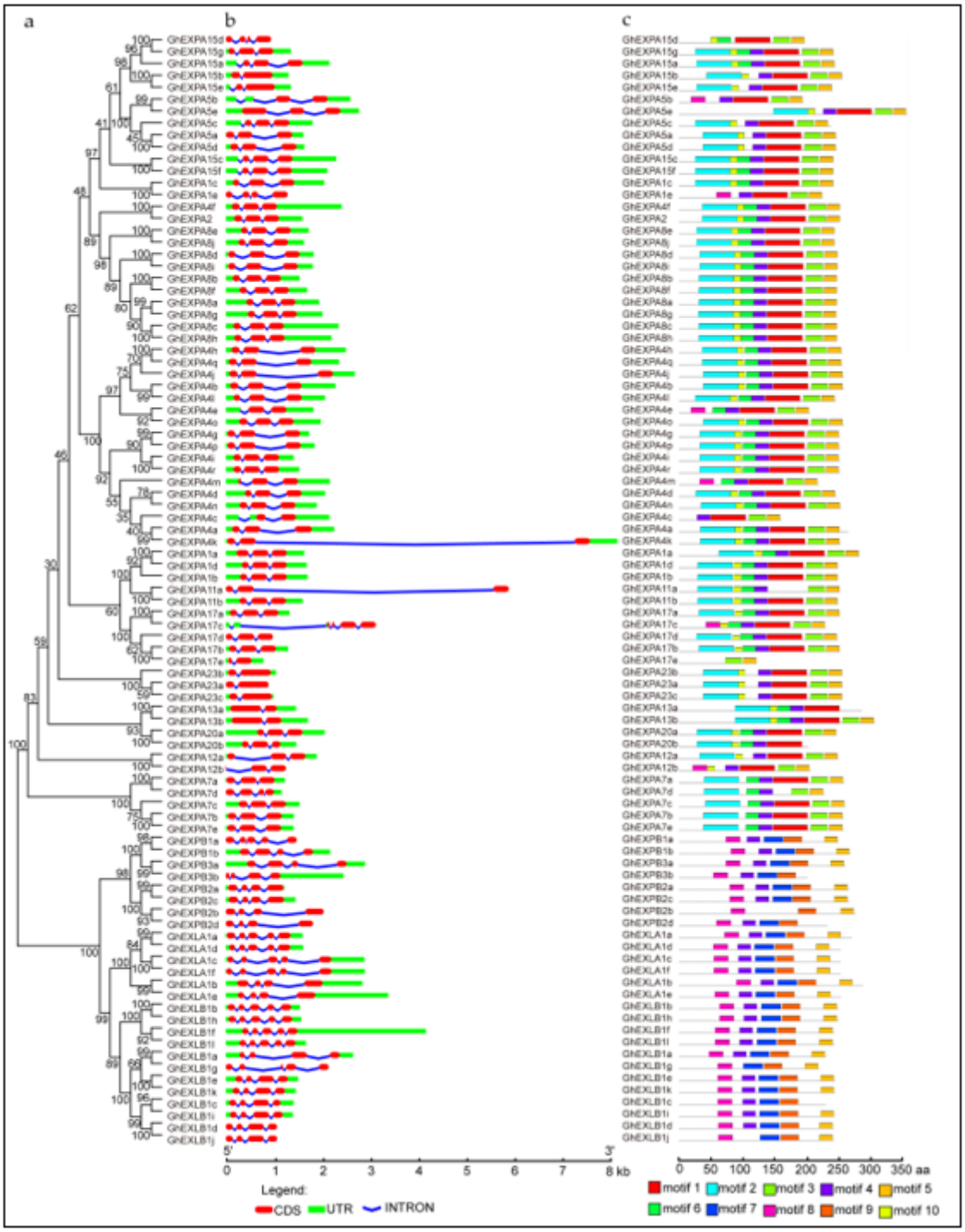

Figure 2

Phylogenetic relationships, gene structures, and protein domain architecture of GhEXP genes. (a) Phylogenetic relationships of 93 GhEXP proteins. The phylogenetic tree was constructed with MEGA 6.0 software using the neighbour-joining (NJ) method with 1000 bootstrap replicates. (b) Gene structure (exon-intron organization) analysis of GhEXPs. The gene structures were drawn with the online Gene Structure Display Server 2.0 [42]. The CDS, INTRONS, and UTRs are marked with red boxes, blue lines, and green boxes, respectively. The scale bar is shown at the bottom. (c) Analysis of conserved domains of the GhEXP proteins. Different colour boxes represent different conserved motifs of GhEXP proteins.

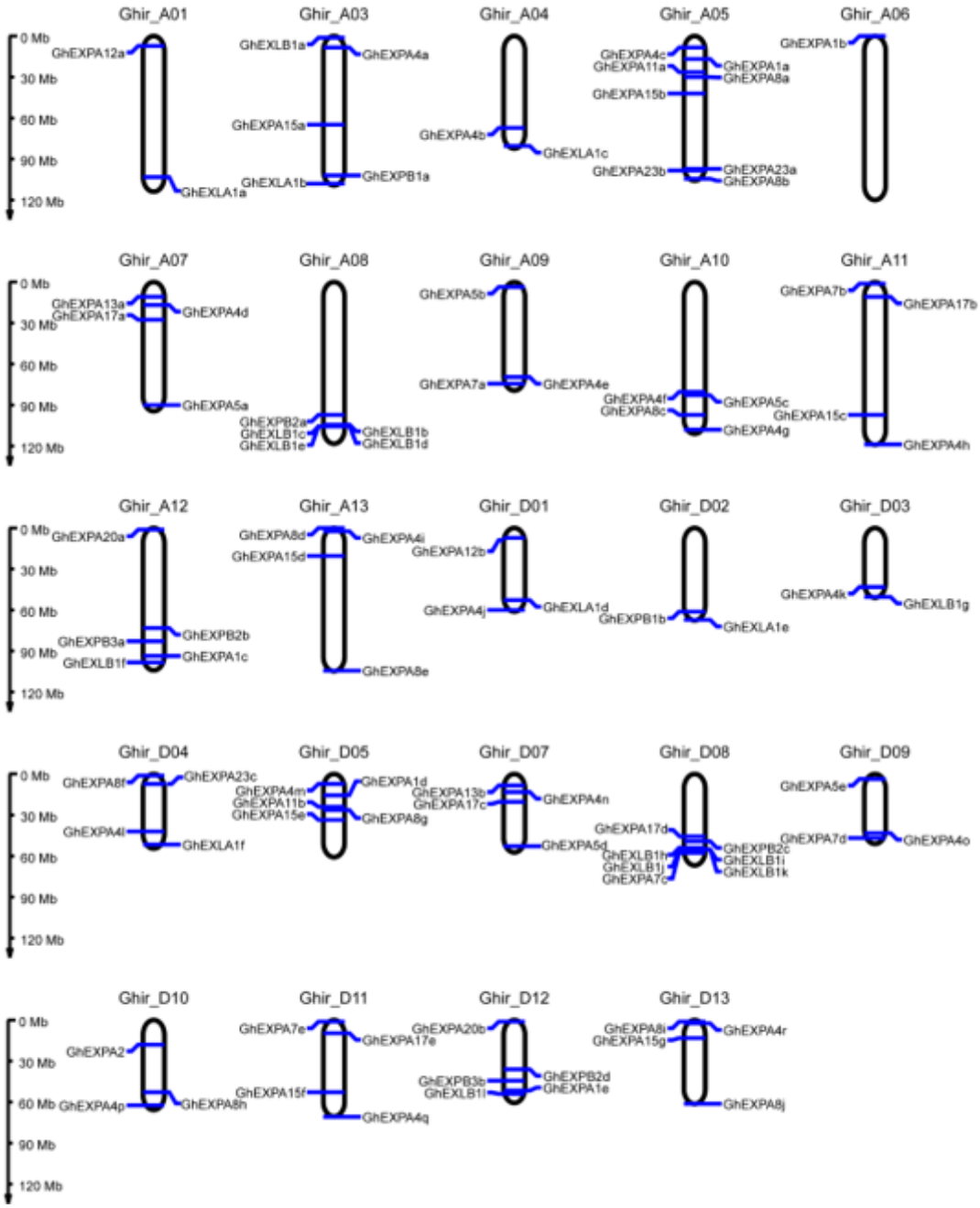

Figure 3

Chromosomal distribution of GhEXP genes. The chromosome name is above each chromosome and the blue lines on the chromosome are the gene names.

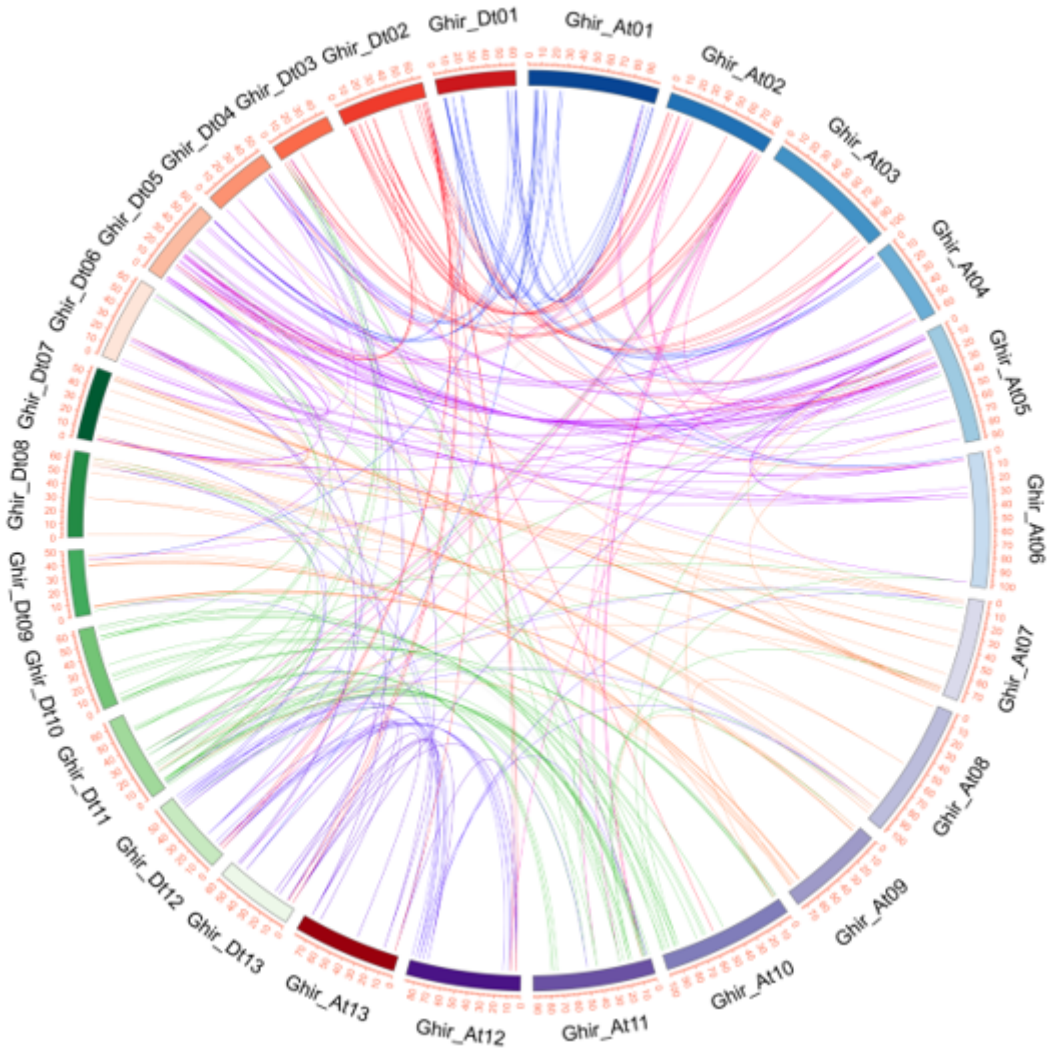

Figure 4

The collinearity relationships of GhEXP genes in upland cotton. The inner colour lines show syntenic blocks in homoeologous chromosomes among cotton expansin genes and between A and D sub-genomes.

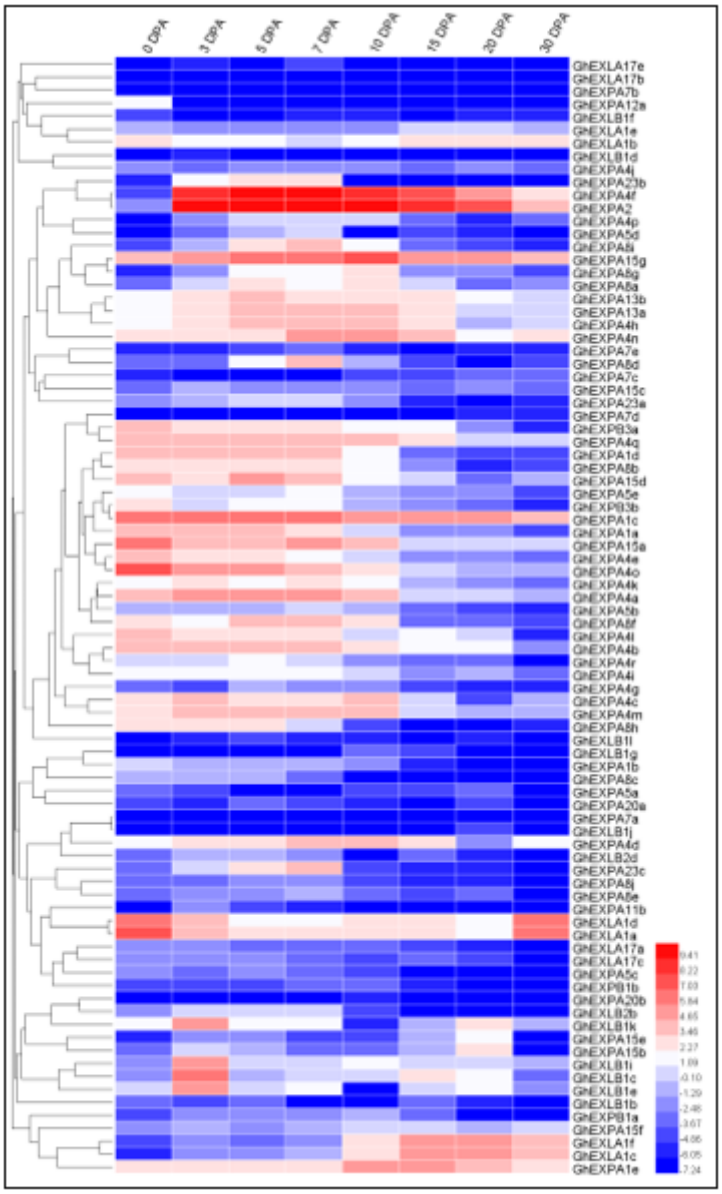

Figure 5

Expression profiles of GhEXP genes in different cotton fibre developmental stages. The heat map was constructed based on RNA-seq data. Different colours represent the different expression levels of GhEXP genes. The legend represents the logarithm values of log2. DPA, day post anthesis; FPKM, Fragments Per Kilobase of transcript per Million mapped reads.

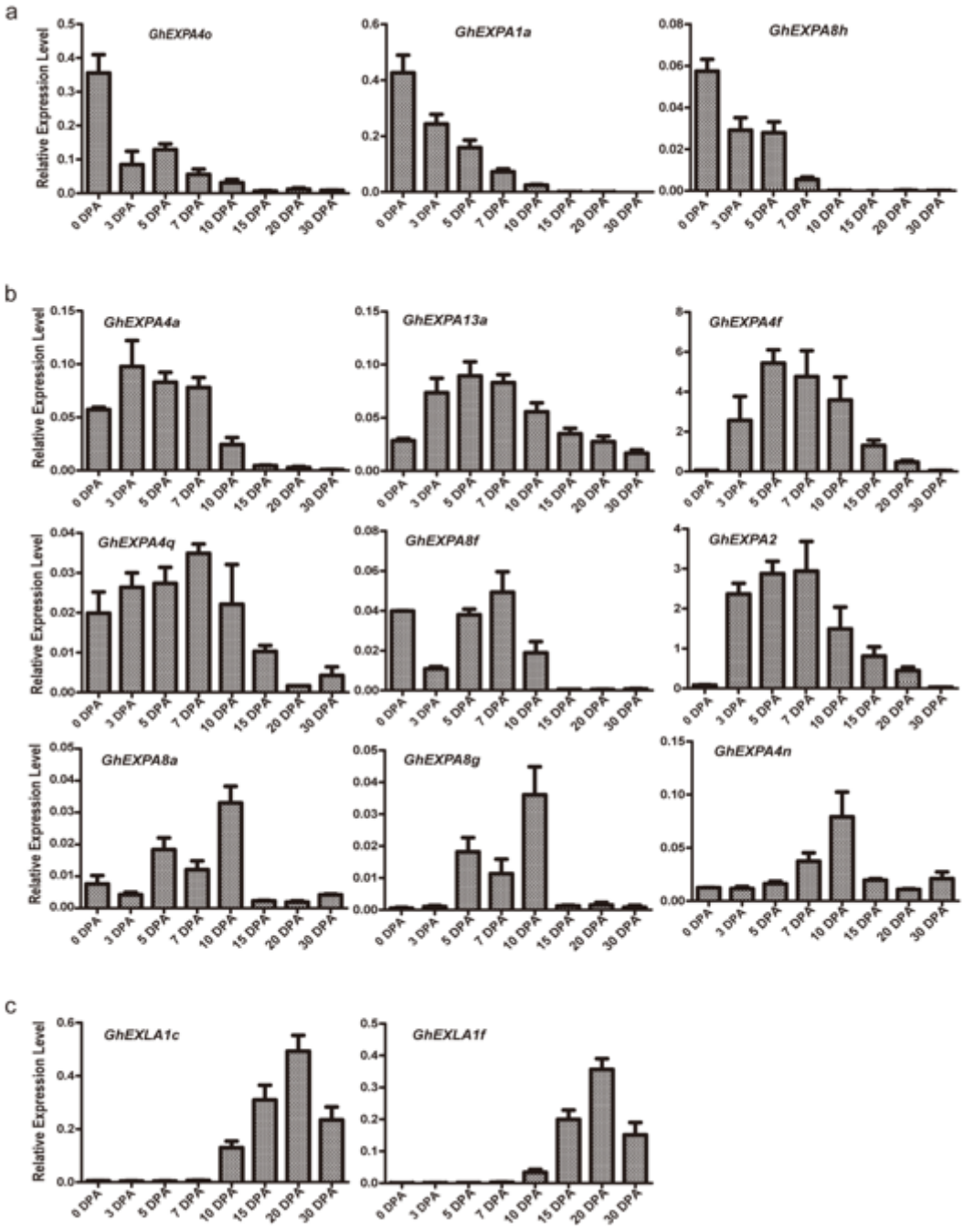

Figure 6

The expression patterns of 14 GhEXPs in different development stages of cotton fibres. (a) Expression profiles of three GhEXPs genes highly expressed in the fibre initiation period. (b) Expression profiles of nine GhEXPs genes highly expressed in fibre elongation stage. (c) Expression profiles of two GhEXPs genes highly expressed at the secondary wall synthesis stage. qRT-PCR experiments were performed with three independent replicates and error bars in this figure represent the SD from three independent experiments.

Supplementary Files

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

- SupplementalMaterials.docx
- Additionalfile1TableS3.TheGrEXPgenesinG.raimondii.xlsx
- Additionalfile1TableS2.TheGaEXPgenesinG.arboreum.xlsx
- AdditionalFile2.docx
- Additionalfile1TableS5.Thenumberofcisactingelements.xlsx
- SupplementalMaterials.docx