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

CURRENT STATUS: UNDER REVIEW

BMC Plant Biology • BMC Series

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
Corresponding Author
ORCID: 0000-0003-3236-9286

Zhiying Ma
Hebei Agricultural University

DOI:
10.21203/rs.2.14830/v1

SUBJECT AREAS
Plant Molecular Biology and Genetics

KEYWORDS
Expansins, gene family, fibre, gene expression profiles, Gossypium hirsutum
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, eight GhEXPBs, six 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, three genes (GhEXPA4o, GhEXPA1A, and GhEXPA8h) were highly expressed in the initiation stage, nine 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 can loosen plant cell walls and cellulosic materials without lytic activity [1, 2], and they may unlock the network of wall polysaccharides, permitting
turgor-driven cell enlargement [3]. 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). EXPA and EXPB proteins are involved in cell expansion and other developmental events during which cell-wall modification occurs, but no confirmed enzymatic activity has been detected for these proteins [4]. In addition, EXLA and EXLB belong to two smaller subfamilies of the expansin superfamily and phylogenetic analysis shows that the two kinds of proteins constitute separated and well-resolved groups. Their structure is predicted to be the same as other expansins, and they can also target the cell wall for modification, however, their biological functions remain uncertain [1].

Typical plant expansins are torpedo-shaped proteins containing two domains, domain I and domain II. 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 [1, 4]. 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 [1], and it was recently classified as a family-63 carbohydrate binding module (CBM63) [1, 5].

Expansins were first identified as the endogenous proteins inducing cell wall extension in plants by extension in plants by the McQueen-Mason group [6]. At present, studies of expansins have shown that they can participate in many developmental processes, and are thought to 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, 3, 7, 8]. Cotton is the most important natural fibre crop worldwide and expansins play an important role in fibre cells [9, 10]. The mRNA levels of
expansins were firstly reported to be high during cell elongation but decreased when cell elongation ceased [11]. Field experiment data showed that expansins improved cotton fibre length and micronaire value [10].

Subsequently, some expansin genes preferentially expressed in cotton fibres were isolated and identified using several different approaches, including cDNA arrays, subtractive PCR, RT-PCR, and SNP-based chromosomal assignment [12, 13], and researchers have a preliminary understanding of the function of these cotton expansin genes at the transcript level [12, 14, 15]. The functions of expansin genes have been further investigated in cotton fibre development [9, 10, 16, 17]. For example, GhRDL1 is localized in the cell wall and interacts with GhEXPA1. Overexpressing GhRDL1 increases fibre length, and cotton plants overexpressing GhRDL1 and GhEXPA1 produce many more fruits [17]. GhEXPA1 expression levels are regulated by the transcript factor GhHOX3, which promotes cotton fibre elongation [16]. GbEXPATR, a Gossypium barbadense-specific expansin, can also enhance cotton fibre elongation through cell wall restructuring [9].

Cotton is the most important global fibre crop and accounts for 90% of natural fibre production in the world [18, 19]. The cotton genome has been sequenced and re-sequenced in succession [20–24]. These genome data make genome-wide identification of gene families possible. We conducted genome-wide identification of cotton expansin genes 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

To identify expansin genes in cotton, we searched the annotation file of the TM-1 genome
using expansin as the keyword [24]. As a result, a total of 98 expansin gene ID records were initially obtained and the corresponding protein sequences were extracted. All expansin proteins were submitted to the NCBI website (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) to analyse their conserved domains. The expansin proteins included two conserved domains, DPBB-1 and Pollen_allerg_1. Consequently, 93 expansin genes with both DPBB-1 and Pollen_allerg_1 domain were ultimately identified for further analysis, and each expansin gene ID was named according to nomenclature guidelines [25]. The detailed information is shown in Table S1. The expansin gene family contained four subfamilies, including EXPA, EXPB, EXLA, and EXLB.

In order to determine the biochemical properties of genes in the expansin family in cotton, the isoelectric point (pI), molecular weight (MW), and signal peptide of expansin proteins were analysed. 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 file 1: Table S1). The pI values of expansin family members ranged from 4.65 to 12.01, with an average of 8.47. The average MW of expansin subfamily members was 27.42 kD, ranging from 14.29 to 41.53 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 to 35 aa in length (Additional file 1: Table S1).

To further understand the detailed information of each cotton expansin protein, 93 expansin protein sequences were used for multiple sequence alignment (Additional file 3: Figure S1). The results showed that expansins have similar sequence characteristics: the majority of them consist of a signal peptide, conserved domains I and II, which is consistent with the previous study [4]. The amino acid sequence of domain I was more
conserved than that of domain II, especially among EXPA members (Additional file 3: 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 3: 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. In addition, a conserved motif named BOX 1 was found in almost all the expansin members (Additional file 3: 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 inferred. The expansins were divided into four major subfamilies, EXPA, EXPB, EXLA and EXLB [25]. 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.

Based on the genomic data information of the cotton expansin family, the analysis of expansin gene structure was performed using the online tool (GSDS 2.0, http://gsds.cbi.pku.edu.cn/index.php) (Fig. 2). The results showed that the gene structure (exon-intron organization) of the expansin members included two to five exons and that 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). These structural differences might be caused by the insertion or loss of exons in the course of long-term evolution and natural selection. Moreover, the similarity in gene structures also suggests that functional complementation exists among these genes.

In order to identify the conserved motifs in expansin proteins, MEME online software (http://meme-suite.org/index.html) was used to analyse the potential motifs. As a result, a total of ten distinct motifs were identified (Fig. 2c, Additional file 4: Figure S2). The results showed that 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

To determine the chromosomal location of GhEXP genes in G. hirsutum, genomic data from G. hirsutum were used [24]. The physical position of cotton expansin genes was determined using positional information files downloaded from the CottonGen website (https://cottonfgd.org/). As shown in Fig. 3, 93 expansin genes were distributed on 24
chromosomes, excluding Ghir_A02 and Ghir_D06. The chromosome Ghir_A05 contained the 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 distribution of expansin genes was uneven on each chromosome. Collinearity analysis showed that expansin genes were collinear frequently between A and D sub-genomes (Fig. 4), which indicates that expansin genes with collinear relationships may have similar function.

Expression patterns of the expansin gene 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 heat map showed the expression patterns of 86 expansin genes. 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 similar expression patterns. For example, GhEXPA1d, GhEXPA15d, GhEXPA15a, GhEXPA4o, GhEXPA4a and GhEXPA4b were the preferentially expressed genes during fibre initiation and elongation stages (0 to 10 days post anthesis (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. Moreover, GhEXPA4f (GhExp1) and GhEXPA2 (GhExp2) have also been reported to be highly expressed in cotton fibre [15].
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. In particular, GhEXPA4o, GhEXPA1a, and GhEXPA8h were predominantly expressed at 0 DPA (Fig. 6a), suggesting that these two genes may function in the initial stage of fibre cells.

Cotton fibre elongation is a very important stage of fibre cells. Nine expansin genes showed higher expression level at the fibre elongation stages with distinct expression characteristics (Fig. 6b). The expression level 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 respectively located in the A and D sub-genomes of the 10th chromosomes, and both genes have specific expression in cotton fibre cells [15].

In addition, GhEXPA8a and GhEXPA8g are two important genes that we have found during cotton fibre elongation, homologous genes of them can promote the elongation of hypocotyl in Arabidopsis [26]. 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.

Moreover, we obtained two expansin genes that were predominantly expressed at transition stages, named GhEXLA1c and GhEXLA1f (Fig. 6c). Both of these genes 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. These data suggested that these two genes, which are essential in the transition stage, may prepare well for the cellulose synthesis of secondary wall thickening period.

Discussion

Cotton is one of the world’s most important economic crops, as well as a powerful single-cell elongation model plant [18]. Cotton fibre is the main raw material of the textile industry, and its developmental process has been studied for many years. This article focused on expansins, which are cell wall loosening proteins that function in cotton fibre development. Many expansin gene families have been identified from eudicots and monocotyledons, including Arabidopsis, grape, jujube, Chinese cabbage, maize, and rice, [4, 27–30]. However, expansins have rarely been analysed in cotton, except for a few studies on the expression levels of GhEXP genes, transcriptional regulation, and genetic transformation studies [10, 15–17]. In recent years, the completion of the cotton genomic sequence through multiple sequencing technologies has provided powerful data support for analysing gene families [22, 24, 31, 32], which will greatly facilitate the identification and analysis of cotton gene families at the whole genome scale.

In this paper, we present the first report of the expansin gene family from upland cotton, which included 93 members. The sequences of 93 expansins all included two conserved domains, DPBB_1 and Pollen_allerg_1. 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 [33]. Thus, we speculate that each clade of the existing expansin family may be extended by each clade ancestor.
In addition, this study showed that the EXPA subfamily genes in cotton were significantly expanded, including 67 total EXPAs (Fig. 1); more EXPAs suggests important functions of this kind of expansin gene in cotton growth and development. Conversely, there are fewer members of the other three expansin subfamilies relative to EXPAs: there are eight EXPBs, six EXLAs and 12 EXLBs. The number of cotton expansin genes in each subfamily is almost consistent with other eudicots, such as Arabidopsis (26 EXPAs, six EXPBs, three EXLAs, and one EXLB), grape (20 EXPAs, four EXPBs, one EXLA, and four EXLBs), jujube (19 EXPAs, three EXPBs, one EXLA and seven EXLBs), and Chinese cabbage (39 EXPAs, nine EXPBs, two EXLAs, and three EXLBs) [1, 4, 27, 28, 30]. The numbers are different in monocotyledons, such as rice (33 EXPAs, 18 EXPBs, four EXLAs, and one EXLB) and maize (36 EXPAs, 48 EXPBs and four EXLAs) [4, 29], with the most significant difference that the EXPBs are more numerous in monocotyledons than eudicots [1]. These results provide significant insights into the evolution and functions of expansin genes in cotton.

According to the chromosomal location information of expansin genes, the 93 genes were unevenly located in the other chromosomes of A and D sub-genomes except for a lack of expansin genes in Ghir_A02 and Ghir_D06 chromosomes (Fig. 3). This uneven distribution on the chromosomes is ubiquitous whether in monocotyledonous or dicotyledonous plants [29, 34, 35], and there are certain tendencies of distribution on different chromosomes. For example, several EXLB genes had a clustered distribution in Ghir_A08 and Ghir_D08, and expansin genes from the same subfamily were distributed on the majority of chromosomes, such as EXPAs in Ghir_A05 (Fig. 3). Previous research has shown that tandem duplications and segmental duplications can affect major variation in family size and the distribution of most gene families; nevertheless, counts of tandem and segmental duplications were negatively correlated, and no families exhibited high levels of both tandem and segmental duplication [36].
The uneven chromosomal distribution of cotton *expansin* genes suggested that segmental duplication events may play an important role in the course of expansin evolution. Some *expansin* genes are clustered on chromosomes, which may reveal the existence of tandem duplication of *expansin* genes during expansin evolution. However, the fact that only a few *EXLB* genes are clustered suggests that tandem duplication is not the main reason for expansin superfamily evolution, but that segmental duplication is the main evolutionary force. Abundant collinearity relationships of *expansin* genes were present in cotton within the A or D sub-genomes and between the A and D sub-genomes; these relationships revealed possible homologous gene pairs with similar functions. 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 [37].

Gene expression patterns can provide important clues to gene function in the growth and development of plants. Transcriptome analysis has shown that cotton *expansin* genes participate widely in different fibre development stages (Fig. 5). We obtained the 14 predominantly expressed genes from transcriptome data in distinct stages of cotton fibre development (Fig. 6). Three of these genes had a higher level of expression in the initial stages of cotton fibre development (Fig. 6a); three genes, *GhEXPA4a*, *GhEXPA1a*, and *GhEXPA8h*, were firstly obtained in the early phase of fibre development, and revealed high expression levels by qPCR; however, their role in initial stages of fibre development still needs to be clarified. In addition, we obtained nine *expansin* genes with higher expression level in the elongation stages (Fig. 6b) and two expansin genes that were predominantly expressed at transition stages (Fig. 6c).

The function of *expansins* has been reported in cotton fibre elongation [9, 10, 17]. *GhEXPA4f*, was highly consistent with the report of Harmer et al. where *GhExp1* transcripts
were highly abundant in the fibre [15], and its homologous gene GhEXPA2 showed a similar expression pattern. This result was basically identical to the GhExp2 expression level [15] and overexpressed GhEXPA1 (named GhEXPA4f based on nomenclature guidelines and in this study) increased upland cotton fibre length [17]. Our qRT-PCR results showed that this gene was highly expressed in the fast elongation stages (Fig. 6b), further proving the importance of this expansin gene in cotton development. Moreover, the homologous gene of GhEXPA4f, referred to as GhEXPA2, was located in the A subgenome. By sequence alignment, GhEXPA2 was named GhEXPA8 by Bajwa et al. [10], and transgenic cotton plants expressing GhEXPA8 showed that fibre length was significantly improved in field experiments [10].

In addition, the other seven new expansin genes we identified 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. More importantly, we have found that GhEXPA8a and GhEXP8g are homologous with AtEXP8 in Arabidopsis thaliana, AtEXP8 can promote the hypocotyl elongation in Arabidopsis thaliana [26]. This result implied that GhEXPA8a and GhEXP8g can promote the cotton fibre elongation in cotton. The above results showed that expansin genes with higher expression levels were all members of the 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 first two stages of cotton fibre development.

EXLA and EXLB were included in two smaller expansin subfamilies. Phylogenetic analysis shows that these proteins constitute separate and well-resolved groups, however their biological functions are uncertain [1]. In particular, 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 [38]. In this paper, we found two EXLA genes, referred to as GhEXLA1c and GhEXLA1f, with higher expression at 20 DPA (Fig. 6c). Interestingly, this phase was called the fibre development transition stage [18], and transitional cell wall remodelling is a distinct, stable developmental stage lasting at least four days (18 to 21 DPA) [39]. In addition, it was reported that an expansin-like protein from Hahella chejuensis could bind cellulose and enhance cellulase activity [40]. These results implied that EXLA subfamily genes function in facilitating transition of fibre development from elongation to secondary cell wall synthesis. 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 species 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, 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 initiation 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 will 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 5 DPA samples we collected ovules, and for 7 to 30 DPA samples we collected fibres. The ovule and fibre 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 *expansin* gene sequences were obtained from the new cotton genome [24] (http://cotton.hzau.edu.cn/EN/download.php). Expansin was used as the keyword for retrieval from the genome annotation file named *Ghirsutum_*Integrated_*Function* Annotation; we then extracted the corresponding *expansin* gene ID. According to these *expansin* gene IDs, all expansin protein sequences were obtained from another file named *Ghirsutum_*gene_peptide by local BLAST. Then, all expansin protein sequences were submitted to the NCBI CDD (conserved domain database) (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) 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), 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 [41] (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) [42], 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 structures were 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 [24] (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/) [43]; 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) [44]. 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 [24], 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) [24]; 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 [45], the visualization of analysis results was drawn using Circos software [46]. 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).

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. 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. A heatmap was drawn with the logarithm values using Heml software (version 1.0,) [47].

**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 ONEc (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 S2. qRT-PCR reactions were performed with the TB Green™ Premix Ex Taq™ II kit (TaKaRa, Code No. RR820A) and conducted on the QuantStudio 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 [48].

**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 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. *Plant expansins: diversity and interactions with plant cell walls*. Current opinion in plant biology. 2015;25:162–72.

2. Cosgrove DJ. *Microbial Expansins*. Annu Rev Microbiol. 2017;71:479–97.

3. Cosgrove DJ. *Loosening of plant cell walls by expansins*. Nature. 2000;407(6802):321–26.
4. Sampredo J, Cosgrove DJ. *The expansin superfamily.* Genome Biol. 2005;6(12):242.

5. Georgelis N, Yennawar NH, Cosgrove DJ. *Structural basis for entropy-driven cellulose binding by a type-A cellulose-binding module (CBM) and bacterial expansin.* PNAS. 2012;109(37):14830-35.

6. Simon McQueen-Mason DMD, and Daniel J. Cosgrove. *Two Endogenous Proteins That Induce Cell Wall Extension in Plants.* Plant Cell. 1992;4:1425-33.

7. 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.

8. Li Y, Darley CP, Ongaro V, Fleming A, Schipper O, Baldauf SL, McQueen-Mason SJ. *Plant expansins are a complex multigene family with an ancient evolutionary origin.* Plant Physiol. 2002;128(3):854-64.

9. 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.

10. 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.

11. Shimizu Y, Aotsuka S, Hasegawa O, Kawada T, Sakuno T, Sakai F, et al. *Changes in levels of mRNAs for cell wall-related enzymes in growing cotton fiber cells.* Plant Cell Physiol. 1997;38(3):375-8.

12. 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.

13. 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.

14. Indrais E, H. M. N. Cheema and A. Bashir. Temporal Expression Analysis and Cloning of Cotton (Gossypium hirsutum) Fiber Genes. Int J Agric Biol. 2011;(13)1:89-94.

15. 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.

16. 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.

17. 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.

18. Haigler CH, Betancur L, Stiff MR, Tuttle JR. Cotton fiber: a powerful single-cell model for cell wall and cellulose research. Frontiers in plant science. 2012;3:104.

19. Qin YM, Zhu YX. How cotton fibers elongate: a tale of linear cell-growth mode. Curr opin plant biol. 2011;14(1):106-11.

20. Wang MJ, Tu LL, Yuan DJ, Zhu D, Shen C, et al. Reference genome sequences of two cultivated allotetraploid cottons, Gossypium hirsutum and Gossypium barbadense. Nat genet. 2019;51(2):224-9.

21. 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.

22. 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.

23. 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.

24. 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.

25. Kende H, Bradford K, Brummell D, Cho HT, Cosgrove D, Fleming A, et al. Nomenclature for members of the expansin superfamily of genes and proteins. Plant Mol Biol. 2004;55(3):311–4.

26. 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.

27. 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.

28. 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.

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. Dal Santo S, Vannonzi 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.

31. 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.

32. Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, et al. Repeated
polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature.2012, 492(7429):423–27.

33.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):409-19.

34.Guimaraes LA, Mota APZ, Araujo ACG, de Alencar Figueiredo LF, Pereira BM, de Passos Saraiva MA, et al. Genome-wide analysis of expansin superfamily in wild Arachis discloses a stress-responsive expansin-like B gene. Plant Mol Biol. 2017;94(1–2):79–96.

35.Li NN, Pu YY, Gong YC, Yu YL, Ding HF. Genomic location and expression analysis of expansin gene family reveals the evolutionary and functional significance in Triticum aestivum. Genes Genom. 2016;38(11):1021–30.

36.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;4:10.

37.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.

38.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.

39.Tuttle JR, Nah G, Duke MV, Alexander DC, Guan X, Song Q, et al. Metabolomic and transcriptomic insights into how cotton fiber transitions to secondary wall synthesis, represses lignification, and prolongs elongation. BMC Genomics. 2015;16:477.

40.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.

41.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.

42. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol. 2013;30(12):2725–29.

43. 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):1296–97.

44. 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.

45. 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.

46. 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.

47. Deng WK, Wang YB, Liu ZX, Cheng H, Xue Y. HemI: A Toolkit for Illustrating Heatmaps. Plos One. 2014;9(11): e111988.

48. 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.

Figures
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 [43]. 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.
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.