frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 10 September 2025
DOI 10.3389/fgene.2025.1625634

OPEN ACCESS

EDITED BY
Kashmir Singh,
Panjab University, India

REVIEWED BY
Usman Aziz,
Northwestern Polytechnical University, China
Zhanji Liu,
Shandong Academy of Agricultural Sciences, China

*CORRESPONDENCE
Renhai Peng,
aydxprh@163.com

RECEIVED 09 May 2025
ACCEPTED 15 August 2025
PUBLISHED 10 September 2025

CITATION
Liu Z, Shen S, Cui Z, Wang T, Li P, Wei Y and Peng R (2025) Genome-wide evolution and function analysis of ALOG gene family in cotton. *Front. Genet.* 16:1625634.
doi: 10.3389/fgene.2025.1625634

Genome-wide evolution and function analysis of ALOG gene family in cotton

Zhen Liu[1], Siyu Shen[2], Zhijuan Cui[3], Tao Wang[1], Pengtao Li[1], Yangyang Wei[1] and Renhai Peng[1]*

[1]Anyang Key Laboratory of Bioinformatics, School of Biotechnology and Food Science, Anyang Institute of Technology, Anyang, Henan, China, [2]School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, Henan, China, [3]Tangyin County Agriculture and Rural Bureau, Tangyin, Henan, China

Background: The ALOG (*Arabidopsis thaliana LSH1* and *Oryza sativa G1*) gene family is a class of transcription factors present in various plants. To elucidate the roles of ALOG genes in cotton, we systematically investigated the ALOG gene family across four cotton species (*Gossypium hirsutum*, *Gossypium barbadense*, *Gossypium arboreum* and *Gossypium raimondii*).
Results: In this study, a total of 43, 42, 23 and 27 ALOG genes were identified from *G. hirsutum*, *G. barbadense*, *G. arboretum* and *G. raimondii*, respectively. The results indicated that cotton ALOG gene duplications originated before the speciation of *Gossypium* species, whole genome duplication, segmental duplication and transposable elements all play important roles in its expansion. In addition, cotton ALOG genes had undergone purifying selection during the evolution. Cis-element analysis revealed that TATA-box and CAAT-box are the most abundant in the promoters of cotton ALOG genes. Transcriptome analysis showed that the expression of ALOG genes in specific tissue is significantly higher than that in other tissues.
Conclusion: This study enhances our comprehension of cotton ALOG genes, and these findings lay the foundation for functional characterizations of ALOG gene family.

KEYWORDS

cotton, ALOG, development, evolution, function

Introduction

The Arabidopsis *LSH1* and Oryza *G1* (ALOG) gene family is a plant-specific transcription factor (Iyer and Aravind, 2012; Li et al., 2019). The N-terminus of ALOG family proteins fused with a N6-adenine methylase active region, and the C-terminus fused with a tyrosine recombinase catalytic active region (Iyer and Aravind, 2012). Studies indicated that the ALOG gene family plays regulatory roles in various aspects of plant growth and development in different lineages of land plants. For example, in rice, there is evidence to suggest that *OsG1L1* and *OsG1L2* have significant effects on inflorescence development (Beretta et al., 2023). In *Arabidopsis thaliana*, the ALOG genes of *LSH4* and *LSH3* are known to suppress organ differentiation the boundary region of the shoot apical meristem (Rieu et al., 2024). In *Torenia fournieri*, *TfALOG3* is associated with corolla tube development and differentiation, and the expression level of *TfALOG3* gene is significantly high in corolla tube. Cells in the corolla bottom differentiated and expanded in wild-type *Torenia fournieri*, whereas such cells in

TfALOG3 loss-of-function mutants failed to develop into a corolla neck (Xiao et al., 2019; Xiao et al., 2018).

Cotton belongs to the Malvaceae family and the *Gossypium* genus, with more than fifty species. The diploid cotton genome is grouped into eight groups, designated A-K, allotetraploid species, such as *G. hirsutum* and *G. barbadense*, originated from the hybridization of A and D genomes (Grover et al., 2012). Cotton fiber is a critical source of fiber for the textile sector. The development of cotton fibers begins with a single cell protrusion on the ovule epidermis, and then differentiates into elongated and thickened seed trichome (Zhai et al., 2023). Although the ALOG gene family plays an important role in plant growth and development, little is known about its molecular mechanism in cotton fiber development; therefore, it would be interesting to make a systematic investigation of the ALOG family in cotton plants. In this study, we carried out a whole-genome identification and analysis of cotton ALOG gene family, including their phylogenetic relationships, conserved motif, selection pressure, evolution, cis-elements and function. Our study will provide a foundation for downstream functional investigation of ALOG genes, and will provide insights into the understanding of the regulatory mechanisms of ALOG genes in controlling fiber growth.

Materials and methods

Identification of cotton ALOG gene family

Genome sequences and annotation files of *G. hirsutism* (Wang et al., 2019), *G. barbadense* (Wang et al., 2019), *G. arboretum* (Du et al., 2018) and *G. raimondii* (Udall et al., 2019) were downloaded from COTTONGENE (http://www.cottongen.org). The hidden Markov model of ALOG (PF04852) were obtained from the InterPro database (https://www.ebi.ac.uk/interpro/), which were used to retrieve cotton ALOG proteins by HMMER (Finn et al., 2011). In addition, we performed a sequence similarity search by BLAST (Matsuda et al., 2013; Nowicki et al., 2018) (E value ≤ E^{-10}) with the ALOG amino acid sequences of *Arabidopsis thaliana* and *Oryza sativa* as queries (Li et al., 2019). Then HMMER results were combined with the BLAST search results, and NCBI-CDD-Search (Yang et al., 2020) was used for further confirmation. The physical and chemical properties of cotton ALOG proteins were predicted by the software Compute pI/Mw (https://web.expasy.org/compute_pi/).

Phylogenetic and conserved motif analysis of cotton ALOG gene family

Multiple sequence alignment of ALOG proteins was carried out by the Clustal X (Larkin et al., 2007). A maximum likelihood tree with a bootstrap value of 1,000 was constructed by MEGA (Hall, 2013). The result tree was then decorated by iTOL (https://itol.embl.de/upload.cgi) (Letunic and Bork, 2021). The online tool MEME (Bailey et al., 2009) (https://meme-suite.org/meme/) was used to analysis conserved motifs, with motif number set to 5. The cis-elements in promoter sequences upstream 1,500 bp of ALOG genes were predicted by PlantCARE (Lescot et al., 2002). The exon-intron organization of cotton ALOG genes was identified by GSDS (Hu et al., 2015) (https://gsds.gao-lab.org/).

Gene duplication and synteny analyses of cotton ALOG gene family

The collinearity relationships of ALOG genes were analyzed by MCScanX (Wang et al., 2012), and the results were visualized using Circos (Krzywinski et al., 2009). Ka/Ks ratios between ALOG members was calculated by KaKs_Calculator software (Wang et al., 2010). The divergence time was calculated by the formula T = Ks/2λ, where λ represents the neutral substitution rate, which is set to 1.5×10^{-8} in this study (Koch et al., 2000).

Transposable elements analysis of cotton ALOG gene family

The transposable elements library was construct by RepeatModeler, and RepeatMasker was used to predict transposable elements (Tarailo-Graovac and Chen, 2009; Tempel, 2012). The transposable elements in 2,000 bp and 10,000 bp upstream and downstream regions of the ALOG genes were identified in this study.

Expression profile analysis of cotton ALOG gene family

Transcriptome data were downloaded from NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra/). The SRA data for multiple tissues (PRJNA490626), different fiber development stages (PRJNA263926) and long day and short day conditions (PRJNA529417) were converted to fastq format with the SRA Toolkit. The software Trimmomatic (Bolger et al., 2014) was used to remove the adapters and to perform quality control and hisat2 (Kim et al., 2015) was used to map the reads to the genomes. Transcript abundance for ALOG genes was quantified using the fragments per kilobase million (FPKM) metric, which was calculated by Cuffinks software (Ghosh and Chan, 2016). Heatmaps of the expression profile values were generated with pheatmap package of R language.

Protein interaction network analysis of cotton ALOG gene family

The interaction network of ALOG proteins was predicted by the online website STRING (https://string-db.org/) (Szklarczyk et al., 2015). *Arabidopsis thaliana* was selected as the organism to retrieve the protein interaction network map.

Result

Identification of cotton ALOG gene family

In the present study, a total of 43, 42, 23 and 27 ALOG genes were identified from *G. hirsutum*, *G. barbadense*, *G. arboretum* and *G. raimondii*, respectively (Supplementary Table S1). The numbers

10.3389/fgene.2025.1625634

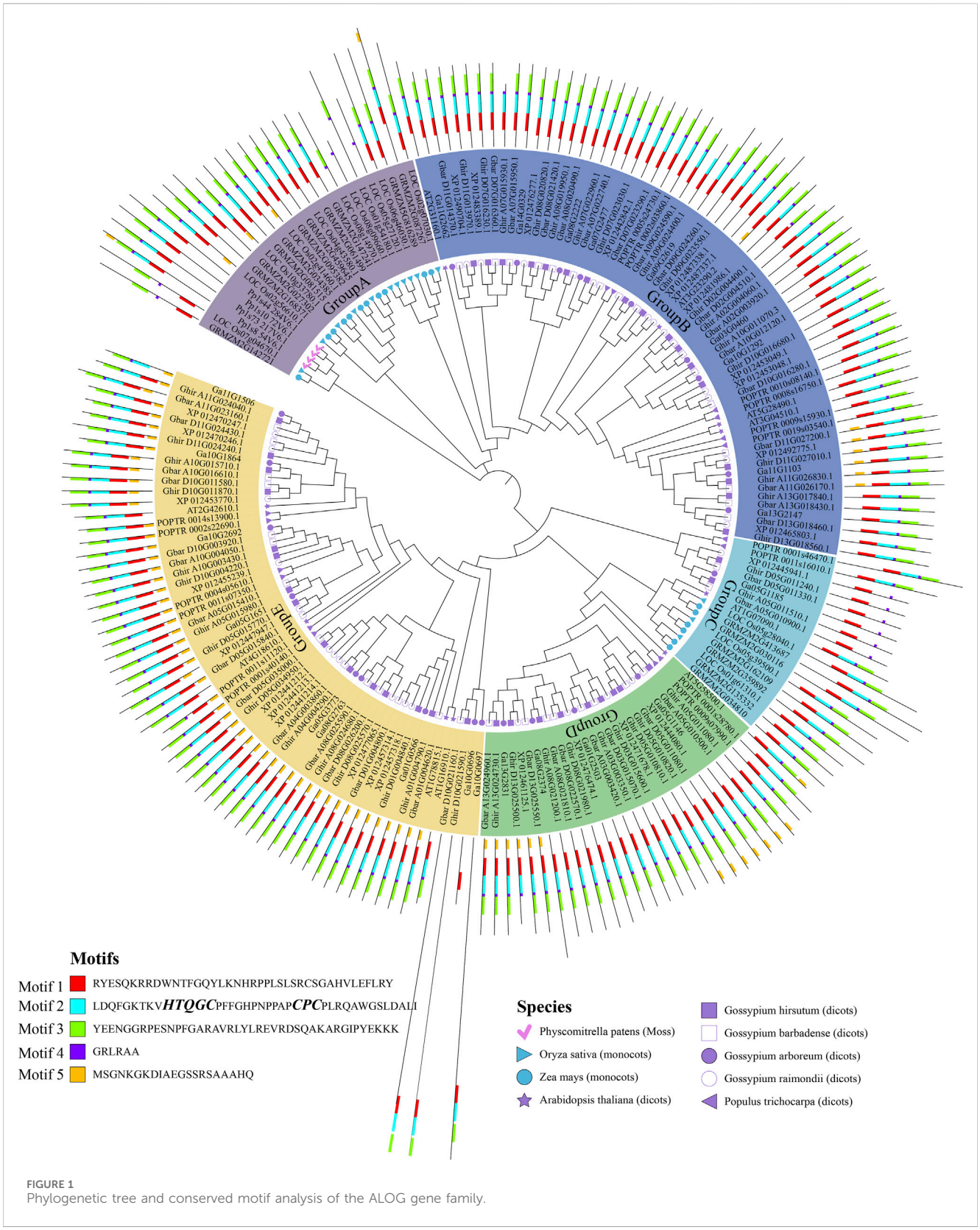

FIGURE 1
Phylogenetic tree and conserved motif analysis of the ALOG gene family.

of ALOG genes in diploid cotton (*G. arboretum* and *G. raimondii*) was found to be comparable to that of *Nicotiana* species (12–23) (Turchetto et al., 2023), yet it exceeded the number observed in *Arabidopsis* (10) (Rieu et al., 2024).

In addition, the results showed that the numbers of ALOG genes in diploid cotton species are almost half those of tetraploid cotton species. Therefore, we speculate that the ALOG gene family likely originated in diploid cotton species and expanded during

polyploidization. However, some gene losses may have occurred post-polyploidization.

To further characterize the cotton ALOG proteins, molecular weight, amino acid sequence length and isoelectric point value were analyzed. Notably, the sequence of Ghir_D10G021590.1 in *G. hirsutum*, Gbar_D10G021160.1 in *G. barbadense* and Ga10G0697 in *G. arboretum* are significantly longer than other ALOG proteins. The NCBI-CDD database revealed that these three proteins contain not only the ALOG domain but also a LRR domain (NCBI CDD: 443914), while all other proteins only contain the ALOG domain. Proteins containing LRR domain include tyrosine kinase receptors, cell-adhesion molecules, virulence factors, and extracellular matrix-binding glycoproteins, and are involved in a variety of biological processes, including signal transduction, cell adhesion, DNA repair, recombination, transcription, RNA processing, disease resistance, apoptosis, and the immune response (Kobe and Kajava, 2001).

Except for these 3 special ALOG proteins, the sequence length of cotton ALOG proteins range from 142 aa (Ghir_A07G015930.1) to 280 aa (Ga08G2374), and the isoelectric point ranged from 8.36 (Ghir_D13G025500.1) to 10.41 (Ghir_D02G004400.1). According to the results, cotton ALOG proteins have a wide range of sequence length and isoelectric point, however, the statistical results of them in the 4 cotton species are very similar, for example, the average molecular weight is around 23,627 Da, the average number of amino acids is about 215 aa, and the average isoelectric point is around 9.7.

Phylogenetic analyses of cotton ALOG gene family

To better understand the origin and diversification of the ALOG gene family, a phylogenetic tree was inferred with ALOG protein sequences of *Physcomitrella patens*, *Oryza sativa*, *Zea mays*, *Arabidopsis thaliana, Populus trichocarpa* and the 4 cotton species. The ALOG proteins can be divided into five groups (Figure 1), among them, the group B and E contain a larger number of proteins. Noticeably, most *Physcomitrella patens*, *Oryza sativa* and *Zea mays* ALOG proteins were distributed in group A and C, while *Arabidopsis thaliana* and *Populus trichocarpa* ALOG proteins were distributed in group B, D and E along with 4 cotton species. This result indicated that ALOG proteins from the same monocot species are clustered into a branch, but those from the same dicot species are dispersed into different branches. Furthermore, Figure 1 shows that many sub-groups contain similar numbers of ALOG family from the 4 cotton species and other dicot species, which suggested that the expansion has occurred before the divergence of dicot species.

Gene structure and conserved motif of cotton ALOG gene family

A total of 5 motifs were identified from the conserved domains of cotton ALOG proteins. The majority members of ALOG proteins contain motif 1–4, indicating that they are the main motifs that make up the ALOG domain (Figure 1). Motif 5 were mainly present in the group E. Group C could be further divided into 2 sub-groups. One of the sub-groups contain *Oryza sativa* and *Zea mays*ALOG proteins, and all of them contain motif 1–4, which is similar to the majority members of ALOG proteins. More interestingly, the other sub-group included 1 ALOG proteins from each diploid cotton species and 2 ALOG proteins from each tetraploid cotton species, and all of them only contain motif 1 and motif 4. These results suggested that ALOG members of group C originated very early in evolution, and they were very conserved since the divergence of cotton species.

Studies have shown that the ALOG domain includes two conserved regions: N-terminal DNA-binding region and C-terminal region (Liu et al., 2024). The sequence of N-terminal with a "HxxxC" and "CxC" signature, which is consistent with Motif 2 (Rieu et al., 2024). The C-terminal region is fused to a tyrosine recombinase catalytic region (Rieu et al., 2024), which is consistent with Motif 1 (Figure 1). In contrast, the functions of Motif 3–5 remain unknown.

The exons and introns were analyzed to get better understand the gene structural evolution of cotton ALOG gene family (Roy and Gilbert, 2006). The results showed that 80.7% (109/135) cotton ALOG genes did not contain introns (Supplementary Table S1). The majority of ALOG genes belonging to group B contain introns, and most of them contained only one intron. Additionally, in group E, 1 ALOG gene from *G. hirsutum*, *G. barbadense* and *G. arboretum* contain 3 introns. Taken together, it appears that genes from the same group also have similar motif and gene structure features, so there may be consistency in the protein function.

Gene duplication of cotton ALOG gene family

The genome chromosomal distribution results indicated that ALOG genes were unevenly distributed on different chromosomes, and most chromosomes contain 1–2 ALOG genes (Figure 2). In addition, Ga14G0329 of *G. arboretum* were located on scaffolds. We refer to the description of Holub that two or more genes of the same family within 200 kb on the same chromosomal is a tandem duplication event (Holub, 2001). There were 2 ALOG genes (Ga10G0696, Ga10G0697) clustered into one tandem duplication event regions on *G. arboretum* chromosomes Chr10, but no tandemly duplicated genes were found in *G. hirsutum*, *G. barbadense* and *G. raimondii*.

Segmental duplicate gene pairs were searched by MCScanX. The results indicated that 40, 41, 18, and 17 genes formed 110, 140, 27, and 28 segmental duplication pairs in *G. hirsutum*, *G. barbadense*, *G. arboretum* and *G. raimondii*, respectively, accounting for 93.02%, 97.62%, 78.26%, and 62.96% of the ALOG gene family. *G. hirsutum* and *G. barbadense* (AADD) are typical allotetraploid from its diploid ancestors *G. arboreum* (AA) and *G. raimondii* (DD).

Segmental duplication relationships between the subgenome and the corresponding ancestral diploid genomes were analyzed to understand the evolutionary mechanism of cotton ALOG gene family. In *G. hirsutum*, 39 ALOG genes had orthologs in the *G. arboreum*, 40 genes had orthologs in the *G. raimondii*, while only 3 genes (Ghir_D10G021590.1, Ghir_D05G010820.1 and Ghir_A10G011070.3) had no ortholog. In *G. barbadense*, similarly, 40 ALOG genes had orthologs in the *G. arboreum*, 41 genes had orthologs in the *G. raimondii*, while only 1 gene (Gbar_D10G021160.1) had no ortholog (Figure 2).

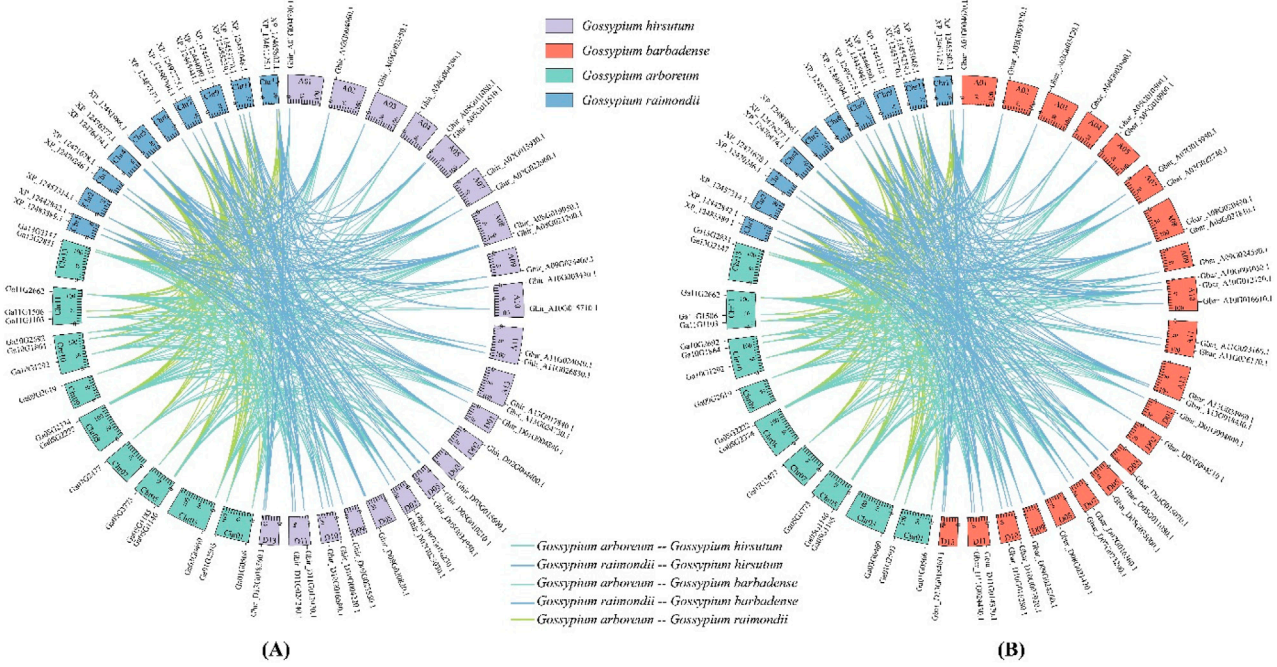

FIGURE 2
Segmental duplication relationship between cotton ALOG genes. **(A)** Allotetraploid *Gossypium hirsutum* and diploid *Gossypium arboretum*, *Gossypium raimondii*; **(B)** Allotetraploid *Gossypium barbadense* and diploid *Gossypium arboretum*, *Gossypium raimondii*.

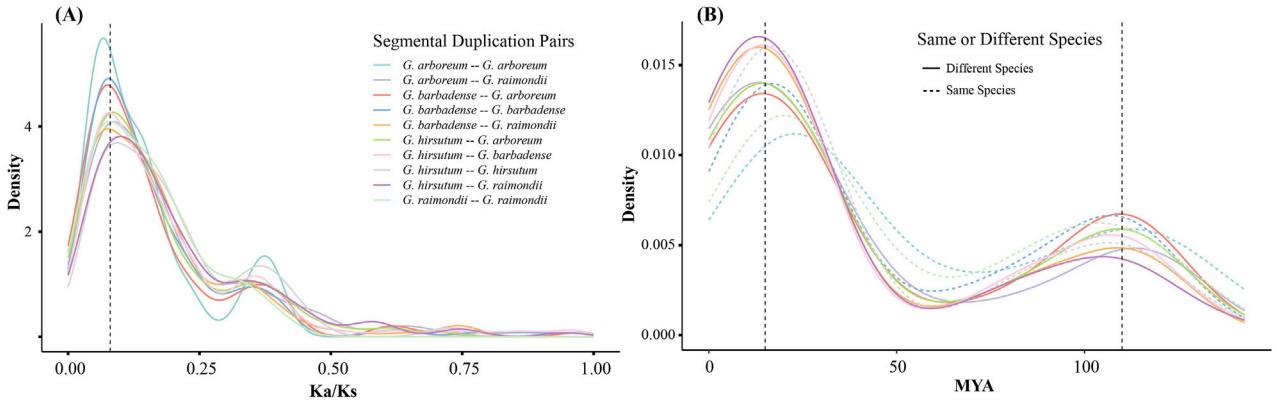

FIGURE 3
The distribution of Ka/Ks ratio and divergence time of cotton ALOG gene family. **(A)** Ka/Ks ratio distribution; **(B)** Divergence time distribution.

The selection pressure of cotton ALOG gene family

In order to further the understanding of the evolutionary constraints of the ALOG gene family, an analysis was conducted to determine the Ka/Ks ratio. The Ka/Ks ratios of ALOG segmental duplication gene pairs between same or different species are all around 0.08 (Figure 3A). These results suggested that cotton ALOG genes were under strong purifying selection (Hurst, 2002). Furthermore, we estimated the divergence time of cotton ALOG gene family by Ks values. The results indicate that the divergence time of ALOG segmental duplication gene pairs concentrated around 15 MYA (million years ago) and 110 MYA (Figure 3B). Previous studies have estimated that the divergence of cotton species began around 10 MYA (Chen et al., 2016; Chen et al., 2017). Based on these results, we speculate that most of the ALOG duplications took place before the speciation of cotton species. In addition, researches have shown that a major polyploidy event occurred within the eudicots around 117 MYA (Jiao et al., 2012), which suggests that the event has an important impact on the expansion of ALOG gene family.

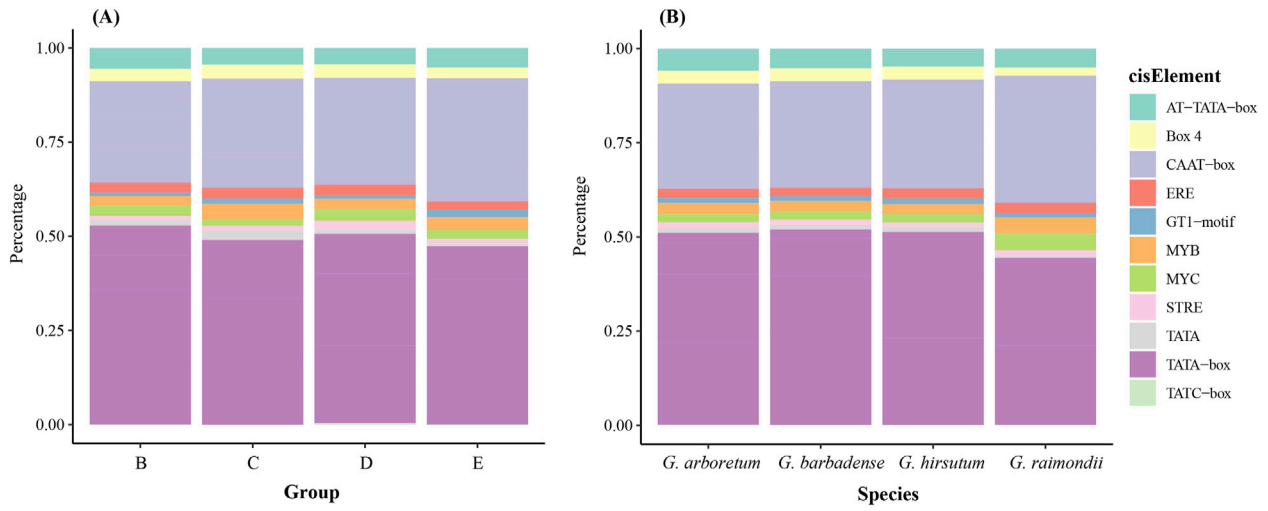

FIGURE 4
Cis-elements composition proportion of cotton ALOG gene. **(A)** Different group; **(B)** Different species.

Transposable element analysis of cotton ALOG gene family

Transposable elements are widely distributed in the genome, especially in plants, which are important for genome expansion and evolution. We identified the transposable elements located 2,000 bp upstream and downstream of the ALOG genes. Our results indicate that 18.60% (8/43), 19.05% (8/42), 13.04% (3/23), and 25.93% (7/27) of ALOG genes close to transposable elements in *G. hirsutum*, *G. barbadense*, *G. arboretum* and *G. raimondii*, respectively. Of these transposable elements, most of them are DNA transposon. When the scanning region broadened to 10,000 bp upstream and downstream, 74.72% (32/43), 69.05% (29/42), 60.87% (14/23), and 66.67% (18/27) genes were found near the transposable elements in *G. hirsutum*, *G. barbadense*, *G. arboretum* and *G. raimondii*, respectively. In addition to DNA transposons, many LTR retrotransposons were identified (Supplementary Table S2). Therefore, our results indicate that cotton ALOG gene family has expansion due to the activity of transposable elements.

Key cis-elements analysis of cotton ALOG gene family

The cis-element present in the promoters of cotton ALOG genes were identified using PlantCARE. The result indicates that TATA-box and CAAT-box were the most abundant cis-elements, in addition, there were also cis-elements related to stress responses (MYB, MYC, STRE), light responsiveness (Box 4, GT1-motif) and so on (Figure 4). Furthermore, we found that the proportions of these cis-elements are similar across different species and groups which indicated the cis-elements of ALOG genes are very conserved after the divergence of cotton species.

Expression patterns of cotton ALOG genes in different tissues

We analyzed the transcriptomic data of leaf, root, stem, filament, anther, pistil sepal, pental, torus and bract in *G. hirsutum* and *G. barbadense*. The results indicated that most ALOGs showed tissue-specific expression patterns, As shown in Figure 5, Ghir_A09G024400, Ghir_D09G023550, Ghir_A11G026830 of *G. hirsutum* and Gbar_A09G024590, Gbar_A11G026170, Gbar_D09G024260 of *G. barbadense* exhibited significantly higher expression in stem than other tissues. Similarly, Ghir_A08G019950, Ghir_D08G020820, Ghir_A02G004060, Ghir_D02G004400 of *G. hirsutum* and Gbar_A02G003920, Gbar_D02G004510, Gbar_A08G020490, Gbar_D08G021420 of *G. barbadense* were mainly expressed in filament but barely expressed in other tissues. Therefore, it perhaps that ALOG gene mainly play its role in a specific tissue. In addition, some ALOG genes expressed in the same tissue were segmental duplicate gene pairs, for example, Ghir_A09G024400, Ghir_D09G023550, Gbar_A09G024590 and Gbar_D09G024260, but not all genes were like this, for example, Ghir_A11G026830 and Gbar_A11G026170 were both highly expressed in the stem, but they were not segmental duplicate gene pairs. Furthermore, members sharing closer phylogenetic relationships displayed similar expression patterns, for example, the six ALOG genes highly expressed in stem were all belong to Group B.

Expression patterns of cotton ALOG genes at different fiber development stages

To explore the potential role of ALOG in fiber development, we investigated their expression at different fiber development stages (10–28 DPA) of *G. barbadense* by RNA-sequencing. The results indicated that the expression pattern of ALOG genes can be grouped into 3 clusters (Figure 6). Thirty *G. barbadense* ALOG genes showed

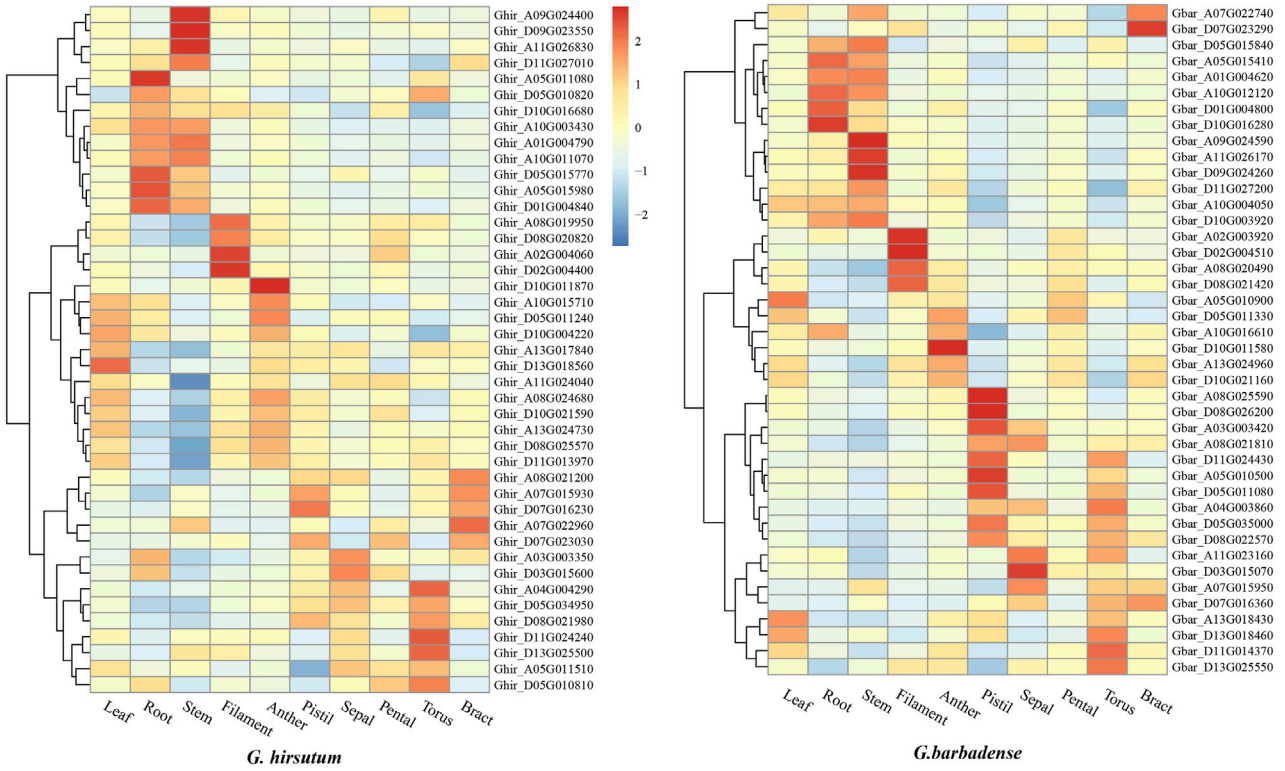

FIGURE 5
Expression heatmap of *Gossypium hirsutum* and *Gossypium barbadense* ALOG genes in different tissues.

upregulated expression during fiber development, whereas eight exhibited downregulation. In addition, the expression of Gbar_D08G022570, Gbar_D05G011080, Gbar_A08G025590, Gbar_D03G015070 exhibited differences at different stages of fiber development. Based on the above RNA-seq data analysis result, we believe that the ALOG genes play a significant role in the fiber development.

Expression patterns of cotton ALOG genes in long day and short day conditions

To investigate the potential function of ALOG family in long day and short day conditions, the transcriptome data from leaf and mristem of *G. hirsutum*, *G. barbadense*, *G. arboretum* and *G. raimondii* was used to calculate the expression of ALOG genes. The expression ratio for the long day conditions/short day conditions was calculated for the expression pattern of each ALOG genes. In the calculation, we only keep the data with FPKM>1 in both conditions. The results showed that 64.5% of the family members had ratios less than 0.8 or more than 1.2, of which 23.5% is less than 0.8% and 41.0% is more than 1.2 (Supplementary Table S3).

Protein interaction network and functional annotation of cotton ALOG proteins

First of all, an interaction network was constructed between cotton ALOG proteins, and the results showed that they do not interact with each other. Furthermore, we predict the possible regulatory mechanism of Ghir_D09G023550.1 based on LSH4 (Light-dependent Short Hypocotyls 4), the protein with the highest homology to Ghir_D09G023550.1 in Arabidopsis (Figure 7). LSH4 belongs to the ALOG family which may act as a developmental regulator by promoting cell growth in response to light and suppress organ differentiation in the boundary region. LSH4 is an important component of the Arabidopsis development network (Rieu et al., 2024), and mainly interacted with five types of biological process, including secondary shoot formation (GO: 0010223), meristem development (GO:0048507), anatomical structure formation involved in morphogenesis (GO:0048646), formation of plant organ boundary (GO:0090691), and plant organ formation (GO:1905393).

The interaction study of cotton ALOG family provides important clues for further study of its function.

Discussion

ALOG gene family plays important roles in plant growth and development (Naramoto et al., 2020). Our results have shown that ALOG proteins do not interact with other ALOG proteins, which suggests that they may independently participate in the growth and development. Furthermore, the expression of ALOG genes in specific tissues is significantly higher than that in other tissues, therefore, we speculate that the biological functions of ALOG family have differentiated in evolution, and they play their own roles in

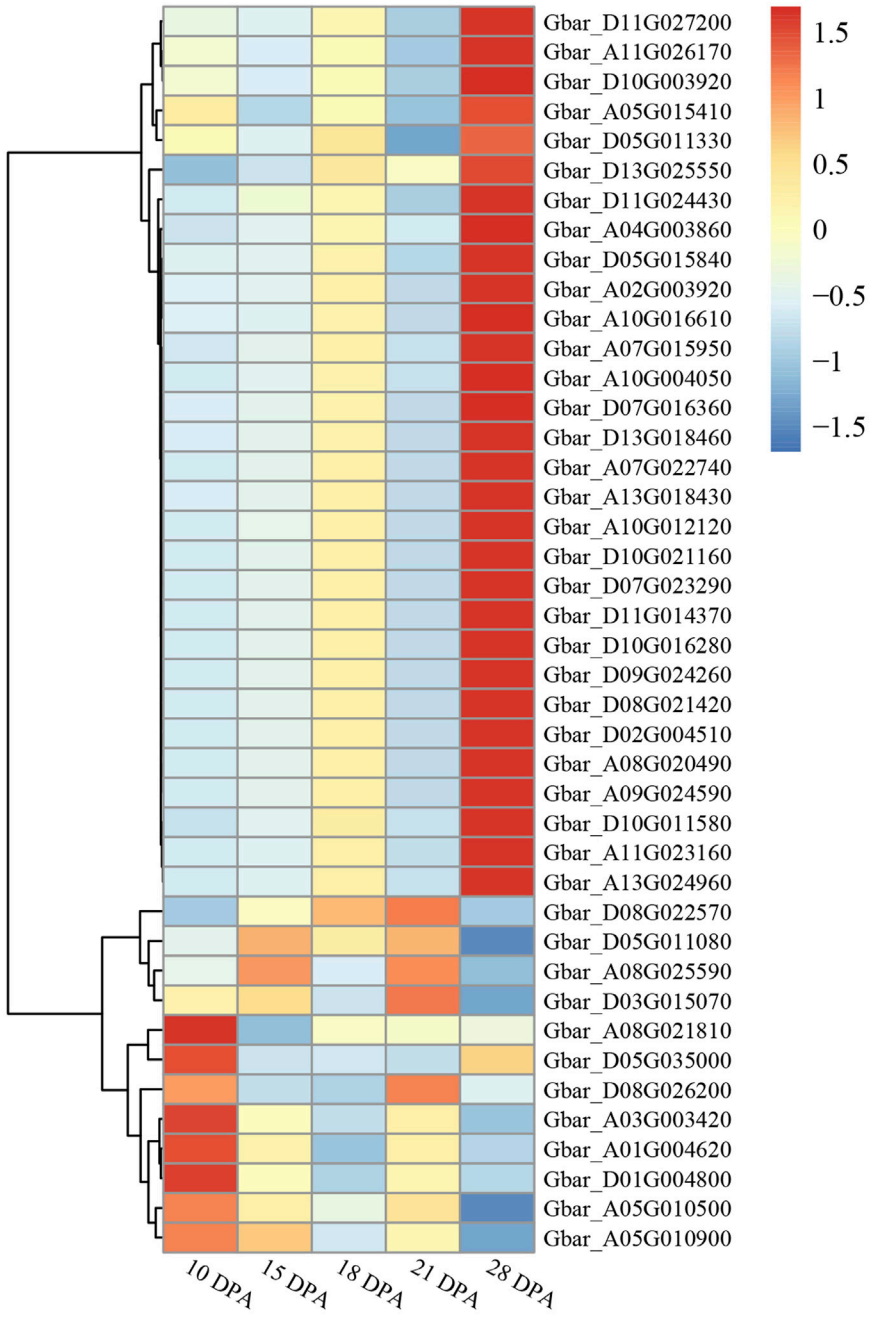

FIGURE 6
Expression profile of *Gossypium barbadense* ALOG genes at different stages of fiber development.

different tissues of plants. ALOG (called LSH in Arabidopsis) family in *Arabidopsis thaliana* play regulatory roles in various aspects of plant growth and development. *Arabidopsis thaliana* genome contains 10 ALOG genes that are involved in different aspects of growth and development. AtLSH1 plays a role in light regulation during seedling development, and its function relies on phytochromes. Overexpression of AtLSH1 and AtLSH2 greatly inhibited hypocotyl elongation in a light independent manner and reduced both vegetative and reproductive growth. AtLSH3 and AtLSH4 play a role in inhibiting organ differentiation at the boundary region. AtLSH8 positively regulates ABA signaling by changing the expression pattern of ABA responsive proteins. AtLSH10 potentially representing a general mechanism for the specific function of plant histone deubiquitinates at their target chromatin (Naramoto et al., 2020; Rieu et al., 2024). In this study, 42 ALOG genes were identified from allotetraploid cotton species *Gossypium barbadense*. Like ALOG genes in other species, these genes are also probably involved in different aspects of cotton growth and development. Based on the expression data, we speculate that Gbar_D05G011080, Gbar_

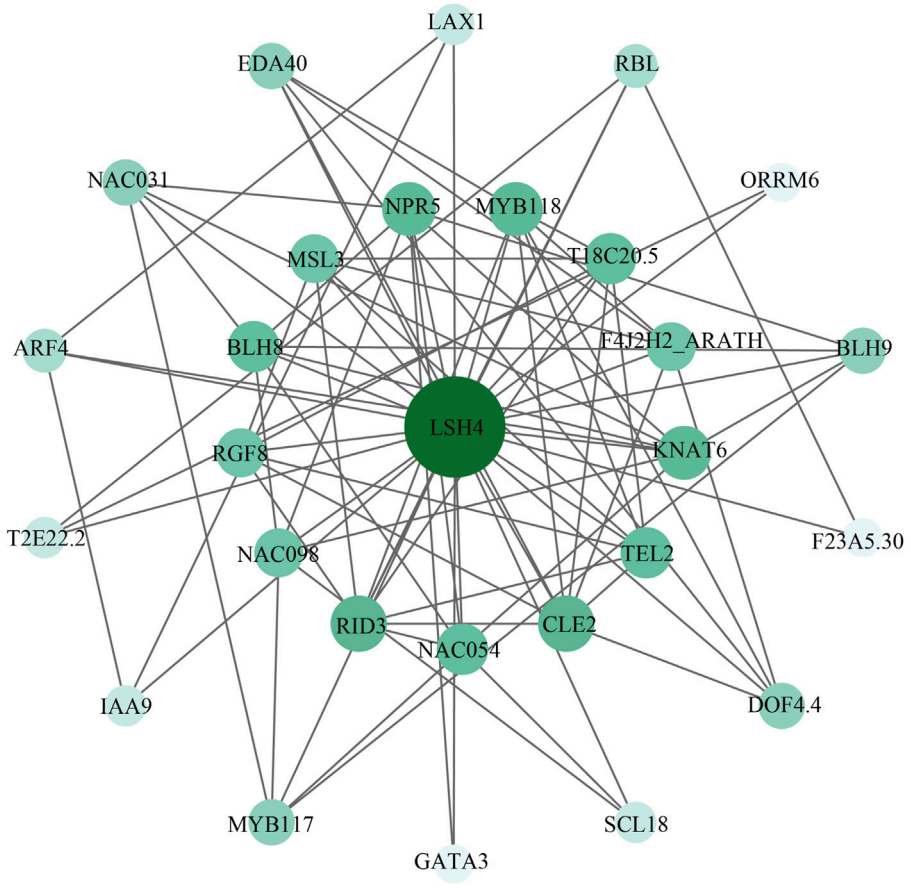

FIGURE 7
Protein interaction network of Ghir_D09G023550.1. LSH4 is the ortholog of Ghir_D09G023550.1 in *Arabidopsis*. The darker the color and the bigger the node, the higher the node degree.

D03G015070, Gbar_D08G022570 and Gbar_A08G025590 might be involved in cotton fiber development.

The ALOG gene family occurred before or during the plant terrestrialization process, exhibiting functional conservation and diversification during the evolution of land plants (Iyer and Aravind, 2012; Turchetto et al., 2023; Xiao et al., 2018). Genomes of land plants have experienced extensive genome-wide and regional duplications. Gene duplication expands the ancient ALOG gene family and produce multiple redundant paralogs.

The evolutionary fates of duplicated genes shape phenotypic stability and allow them to compensate each other's loss. The allotetraploid *G. hirsutum* and *G. barbadense* originated from interspecific hybridization between the A-genome *G. arboreum* and the D-genome *G. raimondii*. In this study, we found that the sum of ALOG genes in the two diploid progenitors exceeds the number of ALOG genes in the tetraploid genome. This suggests that gene loss occurred during the polyploidization process.

Cis-elements in gene promoter regions serve to play critical roles in regulating gene expression (Wittkopp and Kalay, 2011). The results of our analysis indicated that the proportions of cis-elements across different cotton species are similar (Figure 4). The selective pressure of a gene family can be reflected by Ka/Ks ratio.

The results of our analysis also indicated that the ratio has a similar distribution pattern both within the same cotton species or between different cotton species (Figure 3). Furthermore, our molecular clock analysis indicated that the divergence time of ALOG genes took place near 15 MYA and 110 MYA which is before divergence of *Gossypium* species (Figure 3). Therefore, our findings indicate that the ALOG family was conserved during the divergence of *Gossypium* species.

Gene families originated from duplication of the same ancestor (Xu et al., 2012). Whole genome duplication, segmental duplication, tandem duplication and transposable elements provides major forces that drive the duplication of gene families (Cannon et al., 2004). In this study, only one tandemly duplicated gene pair was fond in *G. arboretum*, however, whole genome duplication, segmental duplication and transposable elements all play important roles in the duplication of ALOG gene family.

Up to now, the functions of ALOG genes are only available for Arabidopsis, rice, and tomato. This study revealed the ALOG gene family in cotton, and explored their evolution, biological function and expression profiles. In future work, we will integrate multiple methods to study the functions of each cotton ALOG gene and we are confident that this will accelerate cotton breeding.

Conclusion

In conclusion, we identified 135 members of the cotton ALOG gene family. Except for 3 ALOG proteins that contain additional LRR domain, all other members have only an ALOG domain. The Ka/Ks ratio between orthologous gene pairs revealed that ALOG genes had undergone purifying selection during evolution. Most ALOG genes do not contain introns, and their conserved motifs, cis-elements, gene duplications, and expression patterns were analyzed. The results of this study provide a basis for the future exploration of the function of ALOG genes.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

ZL: Investigation, Software, Conceptualization, Funding acquisition, Supervision, Data curation, Writing – review and editing, Project administration, Formal Analysis, Visualization, Writing – original draft, Validation, Methodology, Resources. SS: Software, Formal Analysis, Visualization and Writing – review and editing. ZC: Writing – review and editing, Methodology, Supervision, Data curation, Investigation, Writing – original draft, Conceptualization, Software, Project administration, Funding acquisition, Visualization, Resources, Validation, Formal Analysis. TW: Writing – original draft, Formal Analysis, Supervision, Writing – review and editing, Funding acquisition, Software, Investigation, Data curation, Resources, Validation, Methodology, Visualization, Conceptualization, Project administration. PL: Formal Analysis, Software, Visualization, Resources, Funding acquisition, Data curation, Project administration, Writing – original draft, Conceptualization, Investigation, Validation, Writing – review and editing, Methodology, Supervision. YW: Conceptualization, Methodology, Supervision, Investigation, Funding acquisition, Software, Writing – review and editing, Formal Analysis, Project administration, Visualization, Writing – original draft, Data curation, Resources, Validation. RP: Visualization, Funding acquisition, Project administration, Resources, Data curation, Validation, Methodology, Conceptualization, Formal Analysis, Writing – review and editing, Supervision, Investigation, Writing – original draft, Software.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was funded by the Postgraduate Education Reform and Quality Improvement Project of Henan Province (YJS2025AL144), the National Natural Science Foundation of China (32272179 and 32272188), the Key Research and Development Project of Henan Province (251111113800), the Scientific and Technological Project of Henan Province (242102110262 and 242102520012), and Zhongyuan Scholars Workstation (224400510020).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2025.1625634/full#supplementary-material

References

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi:10.1093/nar/gkp335

Beretta, V. M., Franchini, E., Ud, D. I., Lacchini, E., Van den Broeck, L., Sozzani, R., et al. (2023). The ALOG family members OsG1L1 and OsG1L2 regulate inflorescence branching in rice. *Plant J.* 115, 351–368. doi:10.1111/tpj.16229

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170

Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D., and May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol.* 4, 10. doi:10.1186/1471-2229-4-10

Chen, Z., Feng, K., Grover, C. E., Li, P., Liu, F., Wang, Y., et al. (2016). Chloroplast DNA structural variation, phylogeny, and age of divergence among diploid cotton species. *PLoS One* 11, e0157183. doi:10.1371/journal.pone.0157183

Chen, Z., Grover, C. E., Li, P., Wang, Y., Nie, H., Zhao, Y., et al. (2017). Molecular evolution of the plastid genome during diversification of the cotton genus. *Mol. Phylogenet Evol.* 112, 268–276. doi:10.1016/j.ympev.2017.04.014

Du, X., Huang, G., He, S., Yang, Z., Sun, G., Ma, X., et al. (2018). Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. *Nat. Genet.* 50, 796–802. doi:10.1038/s41588-018-0116-x

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37. doi:10.1093/nar/gkr367

Ghosh, S., and Chan, C. K. (2016). Analysis of RNA-seq data using TopHat and cufflinks. *Methods Mol. Biol.* 1374, 339–361. doi:10.1007/978-1-4939-3167-5_18

Grover, C. E., Grupp, K. K., Wanzek, R. J., and Wendel, J. F. (2012). Assessing the monophyly of polyploid Gossypium species. *Plant Syst. Evol.* 298, 1177–1183. doi:10.1007/s00606-012-0615-7

Hall, B. G. (2013). Building phylogenetic trees from molecular data with MEGA. *Mol. Biol. Evol.* 30, 1229–1235. doi:10.1093/molbev/mst012

Holub, E. B. (2001). The arms race is ancient history in Arabidopsis, the wildflower. *Nat. Rev. Genet.* 2, 516–527. doi:10.1038/35080508

Hu, B., Jin, J., Guo, A. Y., Zhang, H., Luo, J., and Gao, G. (2015). GSDS 2.0: an upgraded gene feature visualization server. *Bioinformatics* 31, 1296–1297. doi:10.1093/bioinformatics/btu817

Hurst, L. D. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* 18, 486. doi:10.1016/s0168-9525(02)02722-1

Iyer, L. M., and Aravind, L. (2012). ALOG domains: provenance of plant homeotic and developmental regulators from the DNA-binding domain of a novel class of DIRS1-type retroposons. *Biol. Direct* 7, 39. doi:10.1186/1745-6150-7-39

Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., Bowers, J. E., Mckain, M. R., Mcneal, J., et al. (2012). A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* 13, R3. doi:10.1186/gb-2012-13-1-r3

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi:10.1038/nmeth.3317

Kobe, B., and Kajava, A. V. (2001). The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol.* 11, 725–732. doi:10.1016/s0959-440x(01)00266-4

Koch, M. A., Haubold, B., and Mitchell-Olds, T. (2000). Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in Arabidopsis, Arabis, and related genera (Brassicaceae). *Mol. Biol. Evol.* 17, 1483–1498. doi:10.1093/oxfordjournals.molbev.a026248

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645. doi:10.1101/gr.092759.109

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., Mcwilliam, H., et al. (2007). Clustal W and clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi:10.1093/bioinformatics/btm404

Lescot, M., Dehais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., et al. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. *Nucleic Acids Res.* 30, 325–327. doi:10.1093/nar/30.1.325

Letunic, I., and Bork, P. (2021). Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 49, W293–W296. doi:10.1093/nar/gkab301

Li, N., Wang, Y., Lu, J., and Liu, C. (2019). Genome-wide identification and characterization of the ALOG domain genes in rice. *Int. J. Genomics* 2019, 2146391. doi:10.1155/2019/2146391

Liu, Z., Fan, Z., Wang, L., Zhang, S., Xu, W., Zhao, S., et al. (2024). Expression profiling of ALOG family genes during inflorescence development and abiotic stress responses in rice (Oryza sativa L.). *Front. Genet.* 15, 1381690. doi:10.3389/fgene.2024.1381690

Matsuda, F., Tsugawa, H., and Fukusaki, E. (2013). Method for assessing the statistical significance of mass spectral similarities using basic local alignment search tool statistics. *Anal. Chem.* 85, 8291–8297. doi:10.1021/ac401564v

Naramoto, S., Hata, Y., and Kyozuka, J. (2020). The origin and evolution of the ALOG proteins, members of a plant-specific transcription factor family, in land plants. *J. Plant Res.* 133, 323–329. doi:10.1007/s10265-020-01171-6

Nowicki, M., Bzhalava, D., and Bala, P. (2018). Massively parallel implementation of sequence alignment with basic local alignment search tool using parallel computing in java library. *J. Comput. Biol.* 25, 871–881. doi:10.1089/cmb.2018.0079

Rieu, P., Beretta, V. M., Caselli, F., Thevenon, E., Lucas, J., Rizk, M., et al. (2024). The ALOG domain defines a family of plant-specific transcription factors acting during Arabidopsis flower development. *Proc. Natl. Acad. Sci. U. S. A.* 121, e2310464121. doi:10.1073/pnas.2310464121

Roy, S. W., and Gilbert, W. (2006). The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat. Rev. Genet.* 7, 211–221. doi:10.1038/nrg1807

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi:10.1093/nar/gku1003

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma. Chapter* 4, Unit 4.10–10. doi:10.1002/0471250953.bi0410s25

Tempel, S. (2012). Using and understanding RepeatMasker. *Methods Mol. Biol.* 859, 29–51. doi:10.1007/978-1-61779-603-6_2

Turchetto, C., Silverio, A. C., Waschburger, E. L., Lacerda, M., Quintana, I. V., and Turchetto-Zolet, A. C. (2023). Genome-wide identification and evolutionary view of ALOG gene family in Solanaceae. *Genet. Mol. Biol.* 46, e20230142. doi:10.1590/1415-4757-GMB-2023-0142

Udall, J. A., Long, E., Hanson, C., Yuan, D., Ramaraj, T., Conover, J. L., et al. (2019). *De novo* genome sequence assemblies of Gossypium raimondii and Gossypium turneri. *G3 (Bethesda)* 9, 3079–3085. doi:10.1534/g3.119.400392

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinforma.* 8, 77–80. doi:10.1016/S1672-0229(10)60008-3

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40, e49. doi:10.1093/nar/gkr1293

Wang, M., Tu, L., Yuan, D., Shen, C., Li, J., Liu, F., et al. (2019). Reference genome sequences of two cultivated allotetraploid cottons, Gossypium hirsutum and Gossypium barbadense. *Nat. Genet.* 51, 224–229. doi:10.1038/s41588-018-0282-x

Wittkopp, P. J., and Kalay, G. (2011). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* 13, 59–69. doi:10.1038/nrg3095

Xiao, W., Ye, Z., Yao, X., He, L., Lei, Y., Luo, D., et al. (2018). Evolution of ALOG gene family suggests various roles in establishing plant architecture of Torenia fournieri. *BMC Plant Biol.* 18, 204. doi:10.1186/s12870-018-1431-1

Xiao, W., Su, S., Higashiyama, T., and Luo, D. (2019). A homolog of the ALOG family controls corolla tube differentiation in Torenia fournieri. *Development* 146, dev177410. doi:10.1242/dev.177410

Xu, G., Guo, C., Shan, H., and Kong, H. (2012). Divergence of duplicate genes in exon-intron structure. *Proc. Natl. Acad. Sci. U. S. A.* 109, 1187–1192. doi:10.1073/pnas.1109047109

Yang, M., Derbyshire, M. K., Yamashita, R. A., and Marchler-Bauer, A. (2020). NCBI's conserved domain database and tools for protein domain analysis. *Curr. Protoc. Bioinforma.* 69, e90. doi:10.1002/cpbi.90

Zhai, Z., Zhang, K., Fang, Y., Yang, Y., Cao, X., Liu, L., et al. (2023). Systematically and comprehensively understanding the regulation of cotton fiber initiation: a review. *Plants (Basel)* 12, 3771. doi:10.3390/plants12213771