

RESEARCH

Open Access



RAD gene family analysis in cotton provides some key genes for flowering and stress tolerance in upland cotton *G. hirsutum*

Nosheen Kabir^{1†}, Xin Zhang^{2†}, Le Liu³, Ghulam Qanmber¹, Lian Zhang², Yu Xuan Wang¹, Zhuojing Sun⁴, Na Zhao^{3*} and Gang Wang^{2*}

Abstract

Background: RADIALIS (*RAD*), belongs to the MYB gene family and regulates a variety of functions including floral dorsoventral asymmetry in *Antirrhinum majus* and development of fruit proteins in *Solanum lycopersicum*. *RAD* genes contain an SNF2_N superfamily domain. Here, we comprehensively identified 68 *RAD* genes from six different species including *Arabidopsis* and five species of cotton.

Results: Phylogenetic analysis classified *RAD* genes into five groups. Gene structure, protein motifs and conserved amino acid residues indicated that *GhRAD* genes were highly conserved during the evolutionary process. Chromosomal location information showed that *GhRAD* genes were distributed unevenly on different chromosomes. Colinearity and selection pressure analysis indicated *RAD* gene family expansion in *G. hirsutum* and *G. barbadense* with purifying selection pressure. Further, various growth and stress related promotor *cis*-acting elements were observed. Tissue specific expression level indicated that most *GhRAD* genes were highly expressed in roots and flowers (*GhRAD2*, *GhRAD3*, *GhRAD4* and *GhRAD11*). Next, *GhRAD* genes were regulated by phytohormonal stresses (JA, BL and IAA). Moreover, Ghi-miRN1496, Ghi-miR1440, Ghi-miR2111b, Ghi-miR2950a, Ghi-miR390a, Ghi-miR390b and Ghi-miR7495 were the miRNAs targeting most of *GhRAD* genes.

Conclusions: Our study revealed that *RAD* genes are evolutionary conserved and might be involved in different developmental processes and hormonal stress response. Data presented in our study could be used as the basis for future studies of *RAD* genes in cotton.

Keywords: *RAD*, *G. hirsutum*, Phylogenetic tree, Phytohormones, Sequence logs, *Cis*-acting elements

Background

MYB gene family members mainly regulate apoptosis, cell differentiation and proliferation. A- MYB, B- MYB and C-MYB are three groups of MYB gene family [1]. These three groups are largely present in vertebrates, fungi, slime molds and insects [2, 3]. For the very first time, MYB gene C1 was isolated from maize and found to regulate biosynthesis of anthocyanin [4]. MYB proteins play a key role in different developmental processes including trichome differentiation, biosynthesis of anthocyanin and flavonoids, floral symmetry, cell proliferation

*Correspondence: nana209@163.com; wg5791@163.com

[†]Nosheen Kabir and Xin Zhang contributed equally to this work.

² Xinjiang Production & Construction Group Key Laboratory of Crop Germplasm Enhancement and Gene Resources Utilization, Biotechnology Research Institute of Xinjiang Academy of Agricultural and Reclamation Science, Shehezi 832000, Xinjiang, China

³ Zhengzhou Research Base, State Key Laboratory of Cotton Biology, Zhengzhou University, Henan 450001, Zhengzhou, China

Full list of author information is available at the end of the article



and control of cell cycles [5, 6]. In *Antirrhinum majus*, two dorsal petals of zygomorphic flowers were larger than ventral and lateral ones with single aborted stamen [7]. RADIALIS (*RAD*) belongs to *MYB* gene family. *RAD* is involved in creating the dorsal identity and regulating the domain activity of DIVIRICATA (*DIV*) (another transcription factors of *MYB* gene family) by restricting it to ventral regions of the flower [8]. *RAD* and *DIV*, interact with each other and control dorsoventral asymmetry of flowers [9].

In *Arabidopsis*, *RAD* and *DIV* gene families have 198 paralogs and they largely contribute in determining orthology across taxa. *MYB* genes contained three alpha helices. *RAD* and *DIV* have 64 members in *Arabidopsis* and are further classified into five subgroups. Among five subgroups, *RAD* falls in the I-box-like subgroup while R-R-type subgroups include *DIV*. *RAD* was the result of loss of the second *MYB* domain of *DIV* [10]. First two helices of *RAD* share more similarity with the N terminal *MYB* domain of *DIV* as compared to third helix and third helix include DNA binding domain of both *RAD* and *DIV* [11].

RAD5, is a member of Snf2 helicase family, encodes two domains including RING-finger domain and HIRAN domain. In yeast (*Saccharomyces cerevisiae*), *RAD5* is the key component in the *RAD5*-dependent error free branch of postreplication repair. *AtRAD5a* and *AtRAD5b* share significant identity to *RAD5* of yeast. *AtRAD5a* and *AtRAD5b* encodes same domain and these two genes have significant similarities with each other as compared to *RAD5* of yeast. However, both *AtRAD5a* and *AtRAD5b* differ in their function [12]. On the basis of epistatic analysis, yeast DNA repair genes can be classified into three different groups, including *RAD3*, *RAD52* and *RAD6* groups. *RAD3* group control nucleotide excision repair and also eliminate DNA damages caused by ultraviolet radiation [13]. However, the *RAD52* group mainly functions in the repair of double-strand breaks induced by ionizing radiation and *RAD6* group overcomes damages of DNA replication and also arrest replication forks [14]. These DNA repair groups encodes members of the DNA helicase-like-SNF2-gene family [15]. SNF2 family members undergo three main functions including binding of specific proteins to chromatin, destabilization of nucleosome structure and alteration of contact points between DNA and proteins [16, 17].

Cotton is an economic crop of the world and provides raw materials for the textile industry and edible oils [18]. *G. hirsutum* and *G. barbadense* are the two allotetraploid species of cotton and provide fiber, seed oil and protein meal to the industry [19]. *G. hirsutum* is famous for its high yield and moderate fiber qualities, while *G. barbadense* is important for superior quality

fibers and accounts for 3% of the world's cotton production [20]. Although some *RAD* genes have been previously reported in model organisms *Arabidopsis* and yeast, the function of *RAD* gene family in cotton has not been reported. In the present research, we systematically identified *RAD* genes in six different species and analyze their phylogenetic relationship, exon and intron structure, conserved protein motifs, biophysical properties, *cis*-acting elements of promoter regions, sequence logos, chromosomal localization, collinearity and miRNA target sites. Present study will increase our understanding about *RAD* gene family and provide insight into the molecular mechanisms of *RADs* in different *Gossypium* species.

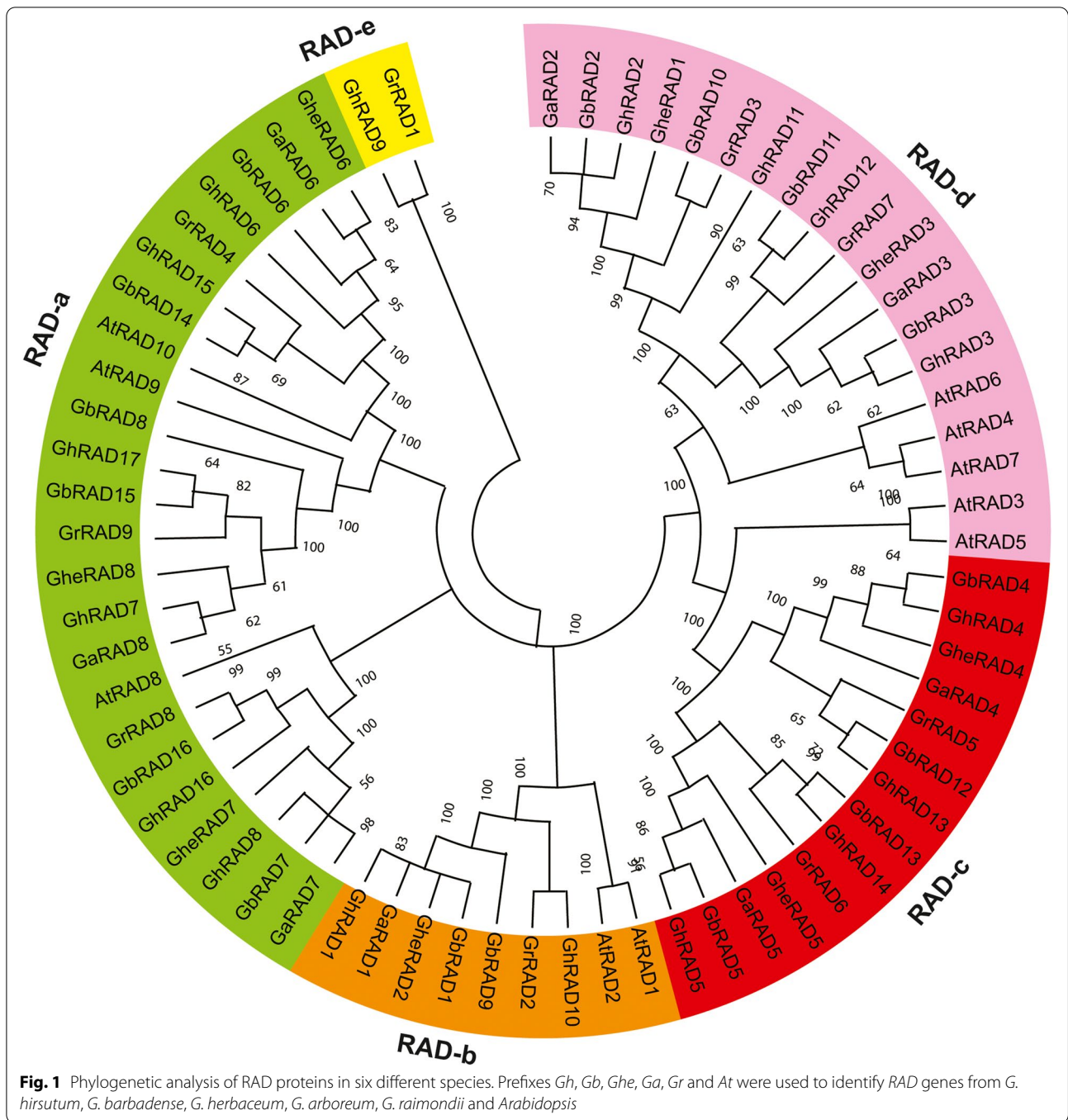
Results

Identification of *RAD* genes in *Gossypium* species

In this study we used different bioinformatic analysis to identify *RAD* genes in various plant species. We identified 17 *RAD* genes in *G. hirsutum*, 16 genes in *G. barbadense*, eight genes each in *G. herbaceum* and *G. arboreum* and nine genes in *G. raimondii*. Furthermore, we also identified 10 *RAD* genes in *Arabidopsis*. SMART (<http://smart.embl-heidelberg.de/>), PROSITE (<http://prosite.expasy.org/>) and InterProScan 63.0 (<http://www.ebi.ac.uk/interpro/>) were used for confirmation of *RAD* genes in different species. We found that among all the selected plant species, *G. hirsutum* had the highest number of *RAD* genes (Table S1) elucidating that during hybridization, *GhRAD* genes underwent polyploidization and experienced significant duplication events. The biophysical properties of the *GhRAD* genes including number of amino acids, chromosomal position, isoelectric point, molecular weight, grand average of hydropathicity and protein localization were observed (Table S2). The number of amino acids of *GhRAD* proteins range from 885 (*GhRAD1*) to 1374 (*GhRAD13*). The isoelectric point (pI) of *GhRAD* genes range from 5.26 (*GhRAD14*) to (9.04) (*GhRAD8*). Molecular weight ranges from 76561.78Da (*GhRAD11*) to 151381.67 Da (*GhRAD13*). Moreover, predicted subcellular localization indicated that out of 17 *GhRAD* genes 16 genes were located in the nucleus while only one gene was located in the cytoplasm.

Phylogenetic analysis, structural features, conserved domain and motifs analysis of *GhRAD* genes

To study the phylogenetic relationships of *RAD* gene family, 68 genes from six plant species were used to create a phylogenetic tree (Fig. 1). *RAD* genes were classified into five clades with 24 members in RAD-a clade, nine members in RAD-b, 14 members in RAD-c while RAD-d and RAD-e contained 18 and two members respectively. RAD-a with maximum number of genes contained RAD



proteins from all six plant species while RAD-e, the smallest clade contained two genes from *G. hirsutum* and *G. raimondii*. All five clades contained genes from six species except RAD-e and one gene of *Arabidopsis* (*AtRAD6*) did not fall in any clade. Moreover, we created another phylogenetic tree to explore the evolutionary relationship between *G. hirsutum* and *Arabidopsis* (Fig. S1). Phylogenetic tree divided 17 *G. hirsutum* and

10 *Arabidopsis* RAD genes into three clades. Rad-c was the largest clade with 12 members, while Rad-b was the smallest one with four members. *GhRAD9* did not fall in any clade, demonstrating that this genes was dissimilar from each other and it has some special function in species evolution.

Next to study the structural features of *GhRADs*, we analyzed the exon and intron structure, conserved

domain and motifs using the GSDS program and MEME tool respectively. Multiple numbers of exons in *GhRAD* genes ranging from 3 to 24 were observed (Fig. S2). Maximum number of exons were present in *GhRAD9* while *GhRAD8* had minimum number of exons. Similarly, *GhRAD9* contained the highest number of introns while *GhRAD16* contained two introns. Number of introns ranged from 2 to 23 for *GhRAD* genes. *GhRAD11* had the longest genomic sequence of more than 8kb while *GhRAD16* had the shortest genomic sequence of 4kb. Next, we analyzed the conserved domain of *GhRAD* proteins. Results of conserved domain analysis showed that all *GhRAD* proteins contain SNF2_N superfamily domain (Fig. S3). Furthermore, we analyzed the motif distribution pattern of *GhRADs* and found that ten conserved motifs were identified in *GhRAD* proteins (Fig. S4). Results indicated that *GhRAD* genes had conserved gene structure and protein motif distribution pattern although the process of gene duplication took place a long time ago during evolution.

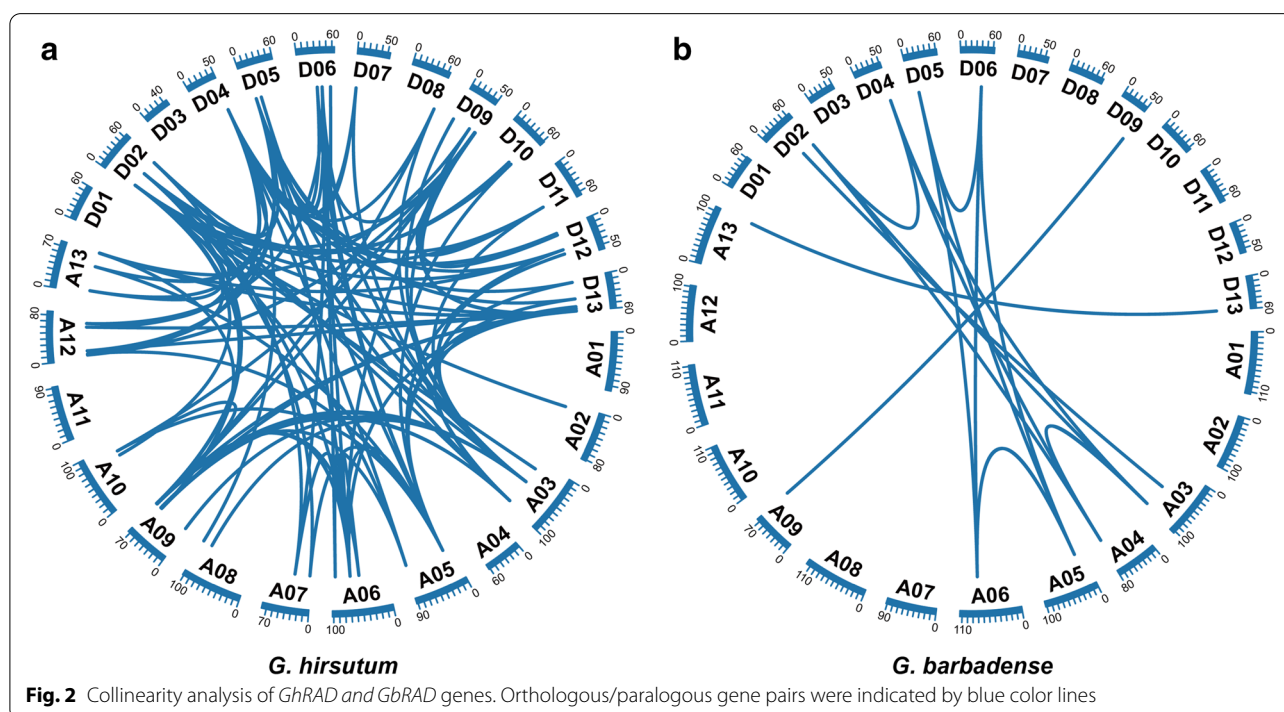
Chromosomal position information and synteny analysis

Chromosomal position of *GhRADs* exhibited that 17 *GhRADs* were distributed unevenly to 11 chromosomes. Out of 17 genes, six genes were mapped on At sub-genome, while nine *GhRAD* genes were mapped on Dt subgenome and two genes were present in the form of scaffolds (Fig. S5). This uneven distribution pattern of *GhRAD* genes on chromosomes illustrated the

presence of genetic variability during the evolutionary process. More deeper insights showed that the maximum number of genes were mapped on chromosome D02. Most chromosomes contained a single *GhRAD* gene and no gene was located on chromosome A02 and A04 of At sub-genome as well as chromosome D03 of Dt sub-genome.

Collinearity analysis between orthologs of At and Dt sub genomes revealed that the different *GhRAD* and *GbRAD* loci showed significantly conserved pattern among At and Dt sub-genomes (Fig. 2 A and B and Table S3, S4). Our results indicated that At and Dt sub-genomes are orthologous in the A genome (*G. arboreum*) or D genome (*G. raimondii*). Upland cotton (*G. hirsutum*) was produced after hybridization between *G. arboreum* and *G. raimondii*. In our study, 112 and 15 orthologous/paralogous gene pairs were identified in *G. hirsutum* and *G. barbadense* respectively.

Ka/Ks ratios of *GhRAD* and *GbRAD* genes were calculated by Ka/Ks calculator 2.0 (Table S3, S4). In the course of evolution, duplicated gene pairs underwent non-functionalization, neofunctionalization and sub-functionalization [21]. In *G. hirsutum* 74 gene pairs exhibited Ka/Ks values below 0.5, illustrating that *GhRAD* genes bear purifying selection pressure while 15 genes showed Ka/Ks ratio above 0.5 and only three genes have Ka/Ks ratio of less than 0.1. While in *G. barbadense* 15 genes exhibited Ka/Ks ratio of less than 0.5. Taken together, our



results indicated that purification selection pressure has a great contribution in maintaining the functional divergence of *GhRAD* and *GbRAD* genes.

Sequence logos and cis-acting elements

We created the protein sequence logos of *Arabidopsis*, *G. hirsutum* and *G. raimondii* to reveal the conservation of *RAD* genes during the process of evolution. Sequence logos provide valuable information about sequence similarities of selected plant species. Results indicated that protein sequence logos were conserved among all observed plant species at most sites within and between species across the N and C terminals (Fig. 3).

Next, we identified different *cis*-acting elements of *GhRAD* genes. Process of gene expression and transcription can be well studied by analyzing the promoter *cis*-acting elements present in specific gene. The 17 putative *GhRAD* promoter regions possessed typical core *cis*-acting elements that were related to different phytohormones like MeJA, gibberellin, auxin, salicylic acid and abscisic acid (Fig. 4). Most of the *GhRAD* genes promoter shared numerous elements for plant growth and development such as seed specific regulation, meristem and endosperm expression and cell cycle regulation. Additionally, we found that many drought responsive elements, flavonoid biosynthesis,

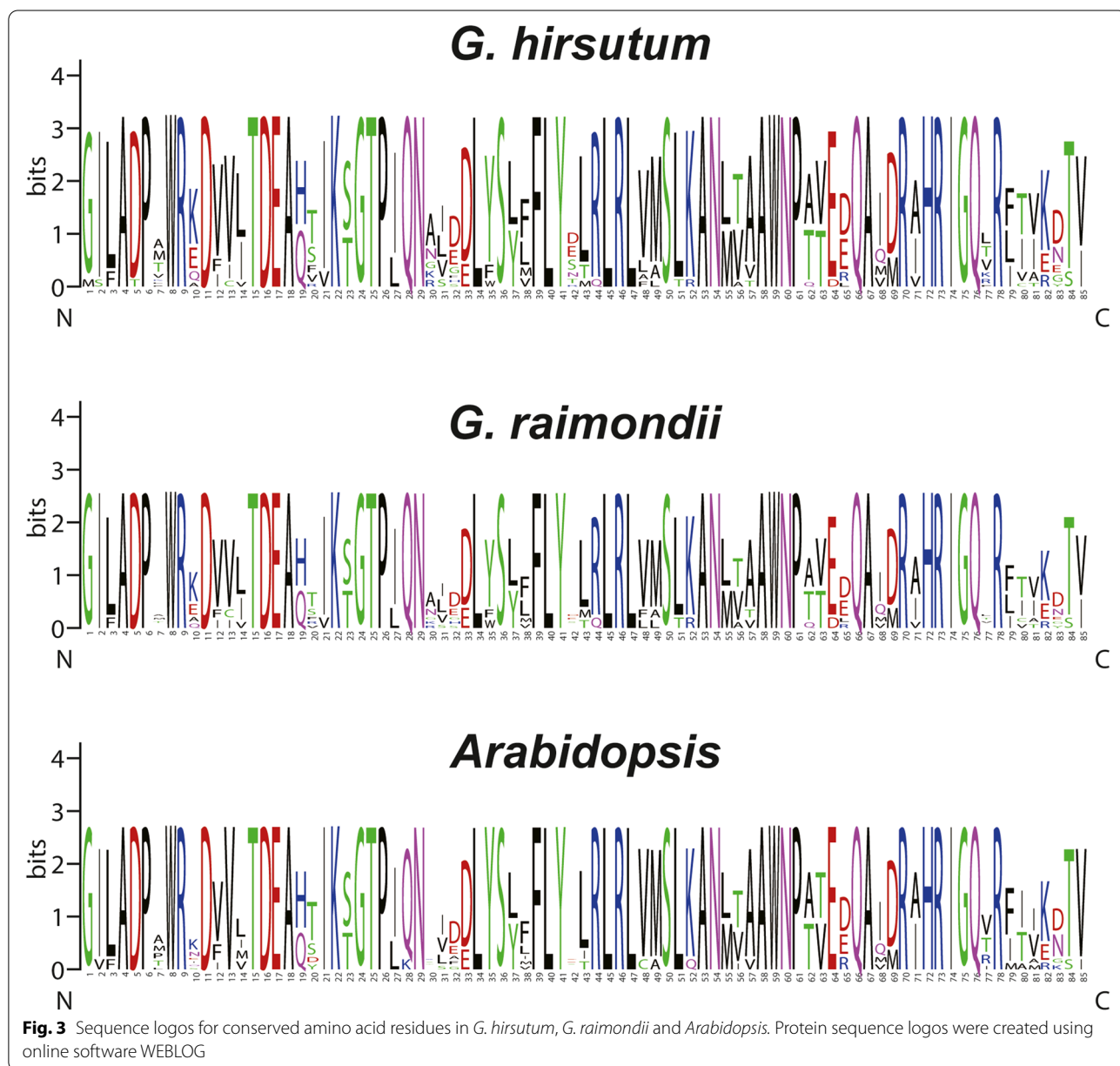
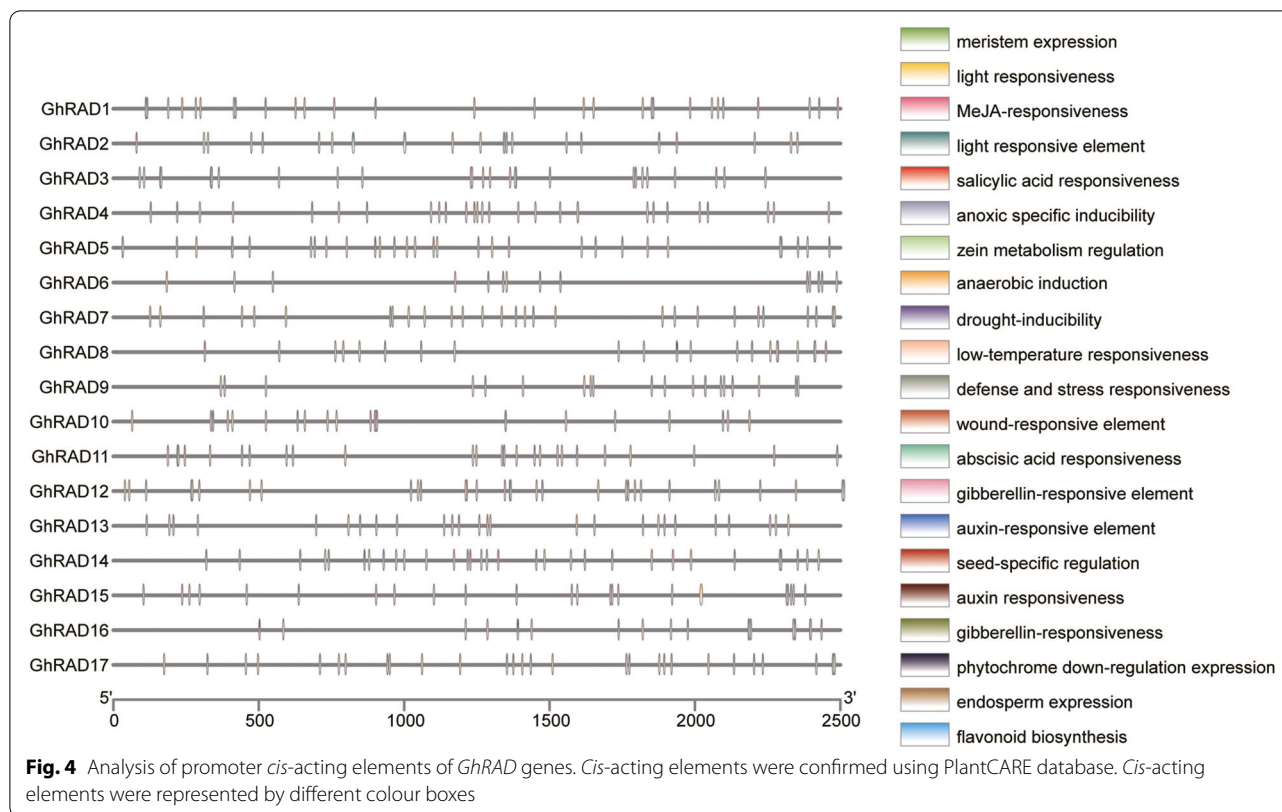


Fig. 3 Sequence logos for conserved amino acid residues in *G. hirsutum*, *G. raimondii* and *Arabidopsis*. Protein sequence logos were created using online software WEBLOG



stress responsive elements, light and low temperature responsive elements, anaerobic induction, anoxic specific inducibility and wound responsive elements were also present in *GhRADs* promoter regions.

Expression profile of *GhRAD* genes in different tissues and hormonal stresses

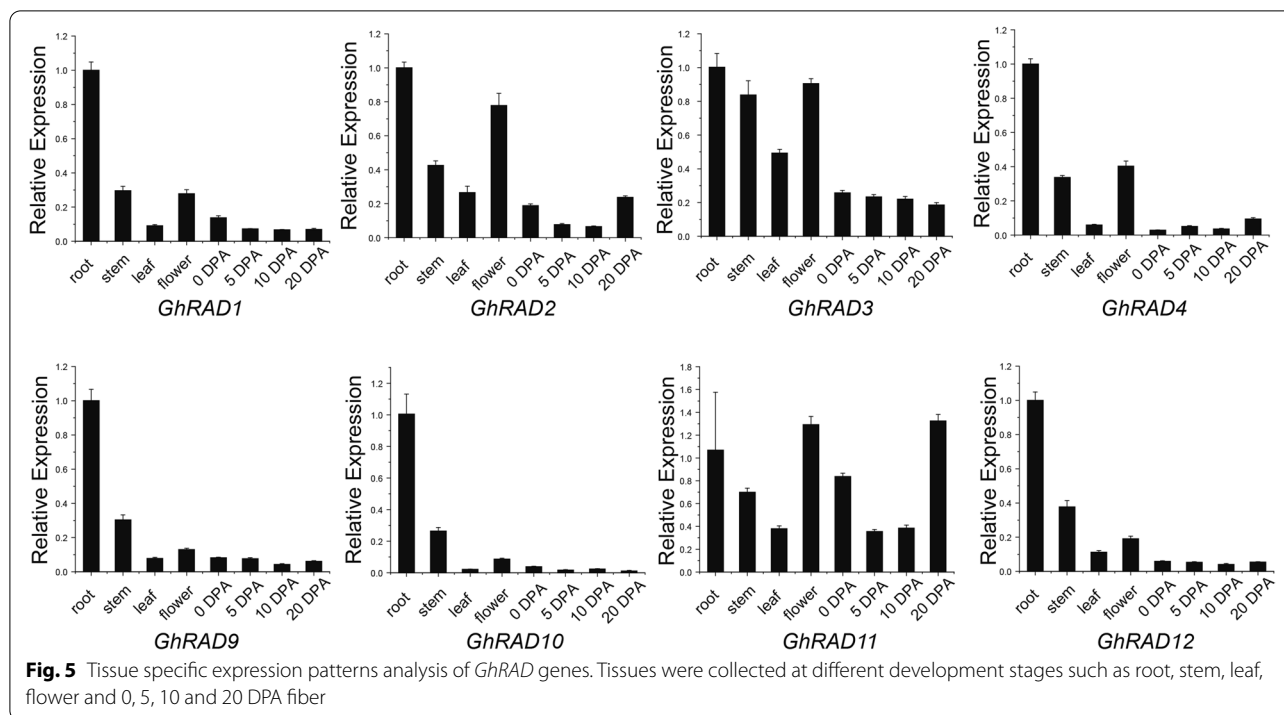
Analysis of gene expression patterns is helpful to predict the biological functions of genes. We downloaded the RNA-seq data from NCBI. We investigated the gene expression pattern of *GhRADs* in various tissue samples and a heatmap was generated (Fig. S6). As shown in figure, *GhRAD1* and *GhRAD10* showed high transcript level in almost all selected tissues demonstrating that these two genes have multiple biological functions and they have some special functions related to fiber development while *GhRAD6* and *GhRAD15* were poorly expressed in some tissues showing that these two genes might be pseudo genes and have limited role in cotton growth and development. Next to validate the results of RNA-seq data, we performed qRT-PCR analysis in eight different tissues including root, leaf, stem, flower and 0, 5, 10 and 20 DPA fiber (Fig. 5). Results indicated that most of the *GhRAD* genes showed preferential expression in root except *GhRAD11* whose expression was preferentially high in 20 DPA fiber, indicating its potential

role in fiber elongation and biosynthesis of secondary cell wall. *GhRAD2*, *GhRAD3*, *GhRAD4* and *GhRAD11* showed high expression in flower tissues demonstrating that these *RAD* genes might function in reproductive development.

Next, we inspected the transcript level of *GhRADs* under three phytohormonal treatments (BL, JA and IAA) at 1, 3, 6 and 12 h of treatment (Fig. 6). Results revealed that most of the *GhRAD* genes showed up and down-regulated expression patterns in different hormones with different time points. *GhRAD1*, *GhRAD2* and *GhRAD9* showed upregulated expression in all three selected phytohormonal stresses, illustrating that these genes might respond to various hormone signaling pathways. *GhRAD1* showed significantly upregulated expression at 1, 3, 6 and 12 h of GA treatment, indicating its potential role in GA signaling pathway. Moreover, transcript level of *GhRAD12* was preferentially high at 1, 3 and 6 h of IAA treatment, suggesting that *GhRAD12* might respond positively to IAA signal. Taken together our results illustrated that *GhRAD* genes were important in several phytohormone signaling pathways.

Prediction of miRNA target sites

In plants, microRNAs (miRNAs) help in regulation of gene expression by activating mRNA translational



repression [22]. To explore miRNA mediated post-transcriptional regulation of *GhRAD* genes, coding sequences of 17 *GhRAD* genes for putative target sites were searched with the help of psRNATarget server. Results indicated that *GhRAD1* was targeted by Ghi-miRN1438 with a site in SNF2_N superfamily domain (Fig. 7 and Table S5). Similarly, *GhRAD2*, *GhRAD3* and *GhRAD4* were targeted by Ghi-miRN1496, Ghi-miR1440 and Ghi-miR2111b respectively. Further, Ghi-miR2950a, Ghi-miR390a, Ghi-miR390b and Ghi-miR7495 regulated *GhRAD9*, *GhRAD10*, *GhRAD11* and *GhRAD12* respectively, with SNF2_N superfamily domain. These results suggested that *GhRAD* genes can be regulated by different miRNAs.

Discussion

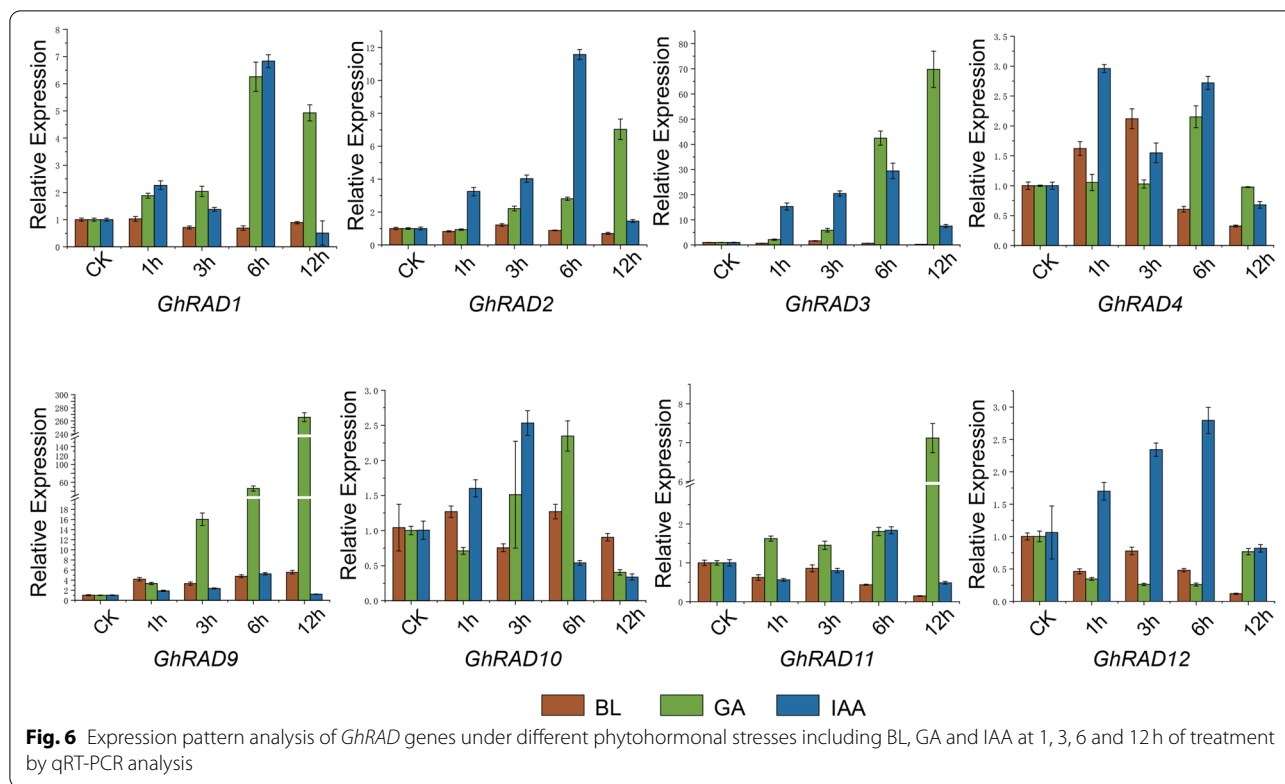
Cotton is an important economic crop of the world and contributes largely to the textile industry. Different gene families such as *GGPPS* [23], *LOG* [24], *BES1* [25], *GH3* [26], *GSK* [27], *GATL* [28], *GhPHD* [29] and *GhAA1* [30] have been identified in cotton, but there is no published research about *RAD* gene family in cotton. Previous studies reported that *RAD51* participate in DNA repair, homologous recombination and genome stability. Members of the *recA_RAD51* family perform functions that have differentiated during evolution [31]. Human *RAD51* is a homologue of the *Escherichia coli* RecA protein and function in recombinational repair of double-stranded DNA breaks. Mutations of *RAD51* reduce repair of double-stranded DNA breaks. Loss of *RAD51* function result

in an elevated mutation rate and accumulation of DNA damage therefore increase risk of cancer in human [32]. In *Antirrhinum*, *RADIALIS* gene is associated with regulation of floral asymmetry [33]. Induction of recessive mutation in *RADIALIS* gene produce symmetrical floral morphology. The *Arabidopsis RAD* gene family contains four members, including *RSM1*, *RSM2*, *RSM3* and *RSM4*, *RADIALIS-LIKE SANT/MYB 1–4*. *Arabidopsis RAD*-like (*RSM*) genes might function in developmental process that is not necessarily relevant to the floral architecture [34] [35].

Gossypium genus contain 45 diploid and six tetraploid species [36, 37]. Two allotetraploid species including, *G. hirsutum* and *G. barbadense* arisen through interspecific hybridization of A and D genome progenitors [19]. Polyploidy occurred around 1–2 million years ago and produced allotetraploid species [38]. *G. hirsutum* and *G. barbadense* are the two oldest allotetraploid species of cotton [39] [40]. Here, we comprehensively identified *RAD* genes in different plant species to identify the diverse functions of *RAD* gene family during the process of plant growth and development. Our study will provide basic information about biological and functional mechanisms and could be helpful for further research of *RAD* gene family in different cotton species.

GhRAD genes show evolutionary conservation

In this study, we identified *RAD* genes from *Arabidopsis* and five *Gossypium* species. Interestingly, the



number of *GhRAD* genes was higher as compared to *GaRAD* and *GrRAD* genes, possibly because *G. hirsutum* is an allotetraploid crop resulting from hybridization of A and D genome progenitors. In upland cotton, both At and Dt sub-genome donors are orthologous relatives, leading to the *GhRAD* gene duplication [41]. Because of duplication and doubling of *GhRAD* genes, the numbers of *GhRAD* genes were equal to the total number of *GaRAD* and *GrRAD* genes. Phylogenetic analysis naturally classified *RAD* genes into five clades from RAD-a to RAD-e with 24 genes in RAD-a clade and two genes in RAD-e clade. Most of the *RAD* genes from diploids and allotetraploids were distributed closely in the phylogenetic tree, indicating that tetraploid cotton was produced after the hybridization of A and D genomes progenitors. From the phylogenetic tree we observed that RAD-a, RAD-b and RAD-d clades were ancient groups containing *RAD* genes from all selected plant species. Whereas, RAD-e clade might be more advanced containing *RAD* genes from only two selected plant species including *G. hirsutum* and *G. raimondii*, however, RAD-c clade lacks the genes from *Arabidopsis*.

Analysis of sequence logos of *G. hirsutum*, *G. raimondii* and *Arabidopsis* also supported these findings as the results of sequence logos showed significantly conserved pattern of amino acid across N and C terminals. Each

conserved amino acid showed similar positions in three different species.

Interaction between transcription factors and promoter binding sites is important for gene regulation at the transcriptional level and transcriptional regulation is mainly responsible for regulation of gene expression. Different external signals activated inducible promoters and *cis*-acting elements in promoter regions are specific and consistent, for example, *cis*-acting elements of AuxRE were generally present in auxin-induced promoters regions [42], similarly AT-rich *cis*-acting elements, I-box, G-box, and GT1-motif were generally present in the light-induced promoters [43–45]. Likewise, CACG and CATGTG *cis*-acting elements were found in drought-induced promoters [46]. In our study, we extracted the upstream promoter fragments of candidate genes. Phytohormone (IAA, SA, ABA, GA and MeJA) responsive *cis*-acting elements were present in *GhRAD* genes promoters. Previous studies have shown that phytohormones maintain the integrity of plant cell walls. MeJA mediates gene expression in responses to plant injury [47]. Auxin regulates cell walls by induction of cell wall looseness [48]. In strawberry, ethylene promotes fruit softening and ripening by regulating the synthesis of pectin [49]. Stress responsive *cis*-acting elements help plants to respond quickly to stress and improve plant resistance to environmental stresses by activating different stress



related genes. *GhRAD* promoter regions contain different *cis*-acting elements related to phyto- hormones, plant growth and development, seed specific regulation, meristem and endosperm expression and cell cycle regulation. In the present study, we found the uneven distribution of *GhRADs* on the chromosomes of *At* and *Dt* sub-genome. Uneven distribution of *GhRAD* genes indicated that during evolution and hybridization, these genes might have experienced gene duplication.

Gene structure is important to predict gene evolution [50]. Here, we found that the number of exons ranged from 3 to 24 while the number of introns ranged from 2 to 23. We observed that length of intron is significantly

different, suggesting that during functional diversification of *GhRADs* intron length might play major roles. It has been reported that gene structure is associated with the evolution of different plant species [51]. The evolutionary analysis within the *RAD* genes showed that most of the *RAD* genes were greatly conserved during evolution. As *RAD* genes introns were not lost during evolution and at the early expansion stages of evolution, these genes diverged, whereas, over evolutionary time, other genes lost their introns [52], indicating that group specific genes may have similar functions. Insertion/deletion events might be the reason for structural differences of exon/intron and may help to identify different

evolutionary mechanisms [53]. Gene families with fewer number of introns are considered to be advanced gene families [54] while gene families showing more introns have acquired some novel functions during the evolutionary process. Furthermore, motif distribution pattern of *GhRAD* genes was conserved indicating that the proteins with similar kind of motif distribution pattern might have some specific functions in cotton growth and development.

Diverse expression patterns of *GhRADs*, prediction of their miRNA target sites and selection pressure

In the recent years, different researches have been done to identify the function of *RAD* genes in different plant species but no research has been conducted to find the functional mechanisms of *RAD* genes in cotton. As our results indicated that promotor of *GhRADs* contain elements related to phytohormones, plant growth and development, seed specific regulation, meristem and endosperm expression and cell cycle regulation, so to check the potential regulatory functions of *GhRADs* during cotton growth and development, we evaluated the transcript patterns of 17 *GhRAD* genes using qRT-PCR. Results indicated that many *GhRAD* genes showed high expression in roots. However, *GhRAD2*, *GhRAD3*, *GhRAD4* and *GhRAD11* showed significantly high expression in flower tissues. Results indicated that *GhRAD* genes might be playing a positive role in root and flower development. All *GhRAD* genes showed poor expression in fiber development stages except *GhRAD11* had relatively high expression at 20 DPA fiber.

Different research studies indicated that phytohormones respond positively to abiotic stress response [55]. Auxin response factors (ARFs) activate auxin responsive genes and ARFs are the potential mediators that help the plants to respond to adverse environmental conditions [56]. Ethylene response factors of AP2/ARF transcription factor family play a positive function in plant growth, disease resistance and phytohormone response [57]. In this study, analysis of *GhRADs* for three different phytohormones indicated that *GhRAD* genes can be regulated by the exposure of phytohormones such as BL, IAA and JA. *GhRAD1* and *GhRAD2* showed upregulated expression for observed phytohormones, suggesting that these genes may have a positive role in phytohormones signaling.

Previous studies reported that miRNAs have significant functions in plant growth and development, regulation of cell growth and different metabolism associated with cotton fiber development and also respond to different environmental stresses including cold, heat and drought [58]. In the past few years, different miRNAs have been identified in cotton and they are preferentially

expressed in different vegetative tissues as well as during abiotic stresses [59, 60]. For instance, 32 families of miRNA were found to be differentially expressed in cotton ovules and leaves. With the help of RNA-seq data of cotton leaf and ovule, a total of 65 families of cotton miRNA have been identified [61]. In our study we identified different *GhRAD* genes targeted by various miRNA. Precisely, Ghi-miRN1496, Ghi-miR1440, Ghi-miR2111b, Ghi-miR2950a, Ghi-miR390a, Ghi-miR390b and Ghi-miR7495 were the miRNAs targeting most of *GhRAD* genes. Our results will lay the foundation for future research about biological functions of *GhRAD* miRNAs and also their target sites in *G. hirsutum*.

Next we calculated the *Ka/Ks* ratios of *GhRADs* and *GbRADs*. Genes with *Ka/Ks*=1.0 are considered as pseudogenes as a result of neutral selection, *Ka/Ks* less than 1.0 demonstrates the tendency of duplicated genes for purifying selection, while *Ka/Ks* ratios greater than 1.0 exhibits positive selection of accelerated evolution. *Ka/Ks* values of most *GhRAD* and *GbRAD* were less than 1 indicating that during the long-term evolutionary process, *RAD* gene family experienced purifying selection pressure.

Conclusions

In the present study, we conducted a comprehensive genome wide analysis to identify phylogenetic relationship, sequence conservation and biological functions of *RAD* genes. Phylogenetic analysis divided *RAD* genes into five clades on the basis of sequence homology. Sequence logos, exon-intron structure, protein motifs and conserved domain analysis indicated that *GhRAD* genes were highly conserved during evolution with the uneven distribution on the chromosomes. Collinearity and selection pressure analysis indicated the expansion of *RAD* gene family in *G. hirsutum* and *G. barbadense* with purifying selection pressure. Promoter *cis*-acting elements analysis indicated the presence of several plant growth and stress related regulatory elements. Tissue specific expression pattern analysis of *GhRAD* genes indicated that some *GhRAD* genes showed significantly high expression in roots and flowers, except *GhRAD11* had maximum expression at 20 DPA fiber. Further, *GhRADs* showed regulatory response under three phytohormonal stresses indicating candidate genetic material for cotton breeding against different stresses.

Methods

Identification of *RAD* gene family

The database of cotton species including (*G. hirsutum*, *G. raimondii*, *G. arboreum*, *G. barbadense* and *G. herbaceum*) were acquired from the Cotton FGD (<https://cottonfgd.org/>) [62]. The *Arabidopsis* protein sequences

of *RAD* genes were acquired from *Arabidopsis* Information Resource, version 10 (TAIR 10) (<http://www.arabidopsis.org>). The *AtRAD* genes were used as queries to identify the *RAD* genes in other plant species. SMART (<http://smart.embl-heidelberg.de/>) and Interproscan 63.0 (<http://www.ebi.ac.uk/InterProScan/>) were used to confirm *RAD* protein sequences [63, 64]. Furthermore, Pfam: (<http://pfam.janelia.org/>) and hidden Markov model (HMM) was used to identify the conserved domain of *RADs* [65]. Moreover, to predict the biophysical properties of *RAD* genes, ExPASy-ProtParam tool (<http://us.expasy.org/tools/protparam.html>) was used. Subcellular localization was predicted by CELLO v2.5 server [66].

Phylogenetic analysis and sequence logos

For phylogenetic analysis of *RAD* gene family, Clustal W was used for multiple sequence alignment [67] with the default settings and the phylogenetic tree was generated by iTOL [68] with 1000 bootstrap replicates. For sequence logos, *RAD* proteins of *Arabidopsis*, *G. hirsutum* and *G. raimondii* were aligned with Clustal W [69] and protein sequence logos were created using online software WEBLOG with default parameters [70].

Chromosomal location, gene structure and analysis of protein motif

Chromosomal position of *GhRADs* was determined by cotton genome annotation data (ftp://ftp.bioinfo.wsu.edu/species/Gossypium_hirsutum/NAU-NBI_G) and extracted gff3-files was used in MapInspect software (<https://mapinspect.software.informer.com/>) to map genes to their corresponding chromosomes. For gene structure analysis of *GhRADs*, Gene Structure Display Server 2.0 (GSDS) was used (<http://gsds.cbi.pku.edu.cn/index.php>) [71]. Online program MEME (<http://meme-suite.org>) [72] was used to determine conserved motifs of *GhRAD* gene family and finally motifs were displayed by TBtools [73].

Promoter *cis*-acting elements, synteny, *Ka/Ks* ratio and transcriptome data analysis

For analysis of promoter *cis*-acting element, 2-kb upstream promoter regions were downloaded from Cotton FGD website [62] and subjected to Plant CARE Database [74]. During collinearity, analysis orthologous/paralogous data were obtained by the previously described methods [75, 76] and circos was used to generate the figure [77]. Non-synonymous (*Ka*) and synonymous (*Ks*) divergence level ratios were calculated by aligning duplicated gene pair protein sequences in Clustal X 2.0, after which they were translated into complementary DNA (cDNA) sequences using the PAL2NAL program (<http://www.bork.embl.de/pal2nal/>). Finally,

Ka and *Ks* values were calculated with the help of the CODEML program using the PAML package. For expression pattern analysis of *GhRADs*, RNA seq data (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4482290/>) was used and heatmap was generated by using TBtools [73].

Prediction of miRNA target sites of *GhRAD* genes

Plant MicroRNA database (<http://bioinformatics.cau.edu.cn/PMRD/>) [78], miRBase (<http://www.mirbase.org/>) [79] and the Cotton EST database (<http://www.ncbi.nlm.nih.gov/nucest>) were used to obtain MicroRNA sequences of *G. hirsutum*. Predictions of miRNA target sites were done with the help of psRNATarget server (<http://plantgrn.noble.org/psRNATarget/home>) as described previously [67].

RNA extraction and RT-qPCR analysis

In this research, the CRI24 variety was used for gene expression analysis. Total RNA was extracted by collecting cotton samples from the field. For tissue specific expression pattern analysis, plant samples such as root, stem, leaf, flower and 0, 5, 1 and 20 DPA fiber were used. Similarly, for hormonal stresses, cotton seedlings were subjected to JA, BL and IAA treatment for 1, 3, 6 and 12 h. RNA extraction was done with the help of RNAPrep Pure Plant Kit (Tiangen, Beijing, China), after that cDNA was synthesized by using the Prime-Script[®]RT reagent kit (Takara, Dalian, China). Green qPCR SuperMix was used to conduct qRT-PCR assay and *GhHis3* (GenBank accession no. AF024716) was used as internal control. PCR reaction was done in three replications. Results of qRT-PCR were calculated as described previously [80]. Primers used for RT-qPCR analysis were presented in (Table S6).

Abbreviations

RAD: RADIALIS; *BL*: Brassinolide; *GA*: Gibberellic acid; *IAA*: Indole – 3-acetic acid; *3D*: 3 Dimension; *DPA*: Days post anthesis.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-021-08248-z>.

Additional file 1 : Figure S1. Phylogenetic relationship between *G. hirsutum* and *Arabidopsis*. Phylogenetic tree was constructed using *RAD* protein sequences by iTOL software.

Additional file 2 : Figure S2. Gene structure of *GhRAD* genes. Yellow boxes represent exons while black lines indicate introns.

Additional file 3 : Figure S3. Conserved domain analysis of *GhRAD* proteins. Each domain was represented in different colors and their names were mentioned with color boxes.

Additional file 4 : Figure S4. Protein motif analysis of *GhRAD* proteins. Each motif was indicated with different color and phylogenetic analysis grouped *GhRAD* proteins according to the protein motif distribution pattern.

Additional file 5 : Figure S5. Chromosomal distribution of *GhRAD* genes. Mapinspect software (<https://mapinspect.software.informer.com/>) was used to map genes to their corresponding chromosomes.

Additional file 6 : Figure S6. RNA-seq data analysis of 17 *GhRAD* genes in root, stem, leaf, torus, petal, stamen, pistil, calyx, — 3, — 1, 0, 1, 3, 5, 10, 20, 25 and 35 DPA ovules and 5, 10, 20 and 25 DPA fiber.

Additional file 7 : Table S1. *RAD5/RAD16*-like Gene Family members.

Additional file 8 : Table S2. Biophysical properties of the *GhRAD* genes.

Additional file 9 : Table S3. Orthologous.paralogous gene pairs and *Ks/Ka* values of *GhRAD* gene family members.

Additional file 10 : Table S4. Orthologous.paralogous gene pairs and *Ks/Ka* values of *GhRAD* gene family members.

Additional file 11 : Table S5. MicroRNA (miRNA) target sites of *GhRAD* genes.

Additional file 12 : Table S6. *GhRAD* genes primers for qRT-PCR.

Additional file 13 : Table S7. Proposed name of *RAD* gene family members.

Acknowledgments

We thank Huo Peng (Zhengzhou Research Center, Institute of Cotton Research of CAAS, Zhengzhou) for technical assistance.

Authors' contributions

NZ and GW designed the research project. NK, GQ, LL and YXW performed the experiments and analyzed the data. NK, GQ, XZ, ZS and LZ wrote and reviewed the manuscript. All authors read and approved the manuscript.

Funding

This research was financially supported by the National Natural Science Foundation of China (31621005, 31701476).

Availability of data and materials

All data used in this study are included in this article and additional files. Transcriptome data used for gene expression analysis could be downloaded from (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4482290/>). The Genome sequence and annotation datasets that supported our findings are available in: *A. thaliana*: (<http://www.arabidopsis.org>), COTTON: (<https://www.cottongen.org/>) and Other species: (<https://jgi.doe.gov/data-and-tools/phytozome/>). All the genes used in this study for phylogeny and subsequent analysis are mentioned in additional file 13, table S7 and can be downloaded from Cotton FGD (<https://cottonfgd.org/>).

Declarations

Ethics approval and consent to participate

The research using field studies of collected plant material, comply with relevant institutional, national, and international guidelines and legislation. The plant material used in this study was provided by State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences. No specific permits are required for sample collection in this study. This article did not contain any studies with wild species at risk of extinction.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang 455000, Henan, China. ²Xinjiang Production & Construction Group Key Laboratory of Crop Germplasm Enhancement and Gene Resources Utilization, Biotechnology Research Institute of Xinjiang Academy of Agricultural and Reclamation Science, Shehezi 832000, Xinjiang, China. ³Zhengzhou Research Base, State Key Laboratory

of Cotton Biology, Zhengzhou University, Henan 450001, Zhengzhou, China. ⁴Development Center for Science and Technology, Ministry of Agriculture and Rural Affairs, Beijing 100122, China.

Received: 21 September 2021 Accepted: 9 December 2021

Published online: 10 January 2022

References

- Klempnauer K-H, Gonda TJ, Bishop JM. Nucleotide sequence of the retroviral leukemia gene v-myb and its cellular progenitor c-myb: the architecture of a transduced oncogene. *Cell*. 1982;31(2):453–63.
- Weston K. Myb proteins in life, death and differentiation. *Curr Opin Genet Dev*. 1998;8(1):76–81.
- Nomura N, Takahashi M, Matsui M, Ishii S, Date T, Sasamoto S, et al. Isolation of human cDNA clones of myb-related genes, A-myb and B-myb. *Nucleic Acids Res*. 1988;16(23):11075–89.
- Paz-Ares J, Ghosal D, Wienand U, Peterson P, Saedler H. The regulatory c1 locus of *Zea mays* encodes a protein with homology to myb proto-oncogene products and with structural similarities to transcriptional activators. *EMBO J*. 1987;6(12):3553–8.
- Raimundo J, Sobral R, Bailey P, Azevedo H, Galego L, Almeida J, et al. A subcellular tug of war involving three MYB-like proteins underlies a molecular antagonism in a *ntirrhinum* flower asymmetry. *Plant J*. 2013;75(4):527–38.
- Zhai R, Wang Z, Zhang S, Meng G, Song L, Wang Z, et al. Two MYB transcription factors regulate flavonoid biosynthesis in pear fruit (*Pyrus bretschneideri* Rehd.). *J Exp Bot*. 2016;67(5):1275–84.
- Luo D, Carpenter R, Vincent C, Copley L, Coen E. Origin of floral asymmetry in *Antirrhinum*. *Nature*. 1996;383(6603):794–9.
- Reardon W, Gallagher P, Nolan KM, Wright H, Cardeñosa-Rubio MC, Bragalini C, et al. Different outcomes for the MYB floral symmetry genes *DIVARICATA* and *RADIALIS* during the evolution of derived actinomorphy in *P. lantago*. *New Phytol*. 2014;202(2):716–25.
- Baxter CE, Costa MMR, Coen ES. Diversification and co-option of *RAD*-like genes in the evolution of floral asymmetry. *Plant J*. 2007;52(1):105–13.
- Stevenson CE, Burton N, Costa MM, Nath U, Dixon RA, Coen ES, et al. Crystal structure of the MYB domain of the *RAD* transcription factor from *Antirrhinum majus*. *Proteins: Struct Funct Bioinformatics*. 2006;65(4):1041–5.
- Yanhui C, Xiaoyuan Y, Kun H, Meihua L, Jigang L, Zhaofeng G, et al. The MYB transcription factor superfamily of *Arabidopsis*: expression analysis and phylogenetic comparison with the rice MYB family. *Plant Mol Biol*. 2006;60(1):107–24.
- Chen I-P, Mannuss A, Orel N, Heitzeberg F, Puchta H. A homolog of *ScRAD5* is involved in DNA repair and homologous recombination in *Arabidopsis*. *Plant Physiol*. 2008;146(4):1786–96.
- Prakash S, Prakash L. Nucleotide excision repair in yeast. *Mutat Res/Fundament Mol Mechanisms Mutagen*. 2000;451(1–2):13–24.
- Aylon Y, Kupiec M. New insights into the mechanism of homologous recombination in yeast. *Mutat Res/Rev Mutat Res*. 2004;566(3):231–48.
- Flaus A, Martin DM, Barton GJ, Owen-Hughes T. Identification of multiple distinct Snf2 subfamilies with conserved structural motifs. *Nucleic Acids Res*. 2006;34(10):2887–905.
- Pazin MJ, Kadonaga JT. SWI2/SNF2 and related proteins: ATP-driven motors that disrupt-protein–DNA interactions? *Cell*. 1997;88(6):737–40.
- Sprouse RO, Brenowitz M, Auble DT. Snf2/Swi2-related ATPase Mot1 drives displacement of TATA-binding protein by gripping DNA. *EMBO J*. 2006;25(7):1492–504.
- Gong Q, Yang Z, Wang X, Butt HI, Chen E, He S, et al. Salicylic acid-related cotton (*Gossypium arboreum*) ribosomal protein GaRPL18 contributes to resistance to *Verticillium dahliae*. *BMC Plant Biol*. 2017;17(1):1–15.
- Chen ZJ, Scheffler BE, Dennis E, Triplett BA, Zhang T, Guo W, et al. Toward sequencing cotton (*Gossypium*) genomes. *Plant Physiol*. 2007;145(4):1303–10.
- Rathore KS, Campbell LM, Sherwood S, Nunes E. Cotton (*Gossypium hirsutum* L.). In: *Agrobacterium* protocols. New York: Springer; 2015. p. 11–23.

21. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. *science*. 2000;290(5494):1151–5.
22. Zhang N, Zhang D, Chen SL, Gong B-Q, Guo Y, Xu L, et al. Engineering artificial microRNAs for multiplex gene silencing and simplified transgenic screen. *Plant Physiol*. 2018;178(3):989–1001.
23. Ali F, Qanmber G, Wei Z, Yu D, Hui Li Y, Gan L, et al. Genome-wide characterization and expression analysis of *geranylgeranyl diphosphate synthase* genes in cotton (*Gossypium* spp.) in plant development and abiotic stresses. *BMC Genomics*. 2020;21(1):1–15.
24. Wang R, Liu L, Kong Z, Li S, Lu L, Chen G, et al. Identification of *GhLOG* gene family revealed that *GhLOG3* is involved in regulating salinity tolerance in cotton (*Gossypium hirsutum* L.). *Plant Physiol Biochem*. 2021;1:328–40.
25. Liu Z, Qanmber G, Lu L, Qin W, Liu J, Li J, et al. Genome-wide analysis of *BES1* genes in *Gossypium* revealed their evolutionary conserved roles in brassinosteroid signaling. *Sci China Life Sci*. 2018;61(12):1566–82.
26. Yu D, Qanmber G, Lu L, Wang L, Li J, Yang Z, et al. Genome-wide analysis of cotton *GH3* subfamily II reveals functional divergence in fiber development, hormone response and plant architecture. *BMC Plant Biol*. 2018;18(1):1–18.
27. Wang L, Yang Z, Zhang B, Yu D, Liu J, Gong Q, et al. Genome-wide characterization and phylogenetic analysis of *GSK* gene family in three species of cotton: evidence for a role of some *GSKs* in fiber development and responses to stress. *BMC Plant Biol*. 2018;18(1):1–21.
28. Zheng L, Wu H, Qanmber G, Ali F, Wang L, Liu Z, et al. Genome-wide study of the *GATL* gene family in *Gossypium hirsutum* L. reveals that *GhGATL* genes act on pectin synthesis to regulate plant growth and Fiber elongation. *Genes*. 2020;11(1):64.
29. Wu H, Zheng L, Qanmber G, Guo M, Wang Z, Yang Z. Response of phytohormone mediated plant homeodomain (*PHD*) family to abiotic stress in upland cotton (*Gossypium hirsutum* spp.). *BMC Plant Biol*. 2021;21(1):1–20.
30. Qanmber G, Lu L, Liu Z, Yu D, Zhou K, Huo P, et al. Genome-wide identification of *GhAAI* genes reveals that *GhAAI66* triggers a phase transition to induce early flowering. *J Exp Bot*. 2019;70(18):4721–36.
31. Lin Z, Kong H, Nei M, Ma H. Origins and evolution of the *recA/RAD51* gene family: evidence for ancient gene duplication and endosymbiotic gene transfer. *Proc Natl Acad Sci*. 2006;103(27):10328–33.
32. Lose F, Lovelock P, Chenevix-Trench G, Mann GJ, Pupo GM, Spurdle AB. Variation in the *RAD51* gene and familial breast cancer. *Breast Cancer Res*. 2006;8(3):1–7.
33. Corley SB, Carpenter R, Copley L, Coen E. Floral asymmetry involves an interplay between TCP and MYB transcription factors in *Antirrhinum*. *Proc Natl Acad Sci*. 2005;102(14):5068–73.
34. Urzainqui A, Bevan M, Martin C, Smeekens S, Tonelli C, Paz-Ares J, et al. The MYB transcription factor superfamily of *Arabidopsis*: expression analysis and phylogenetic comparison with the rice MYB family. *Plant Mol Biol*. 2006;60(1):107–24 PubMed PMID.
35. Riechmann JL, Heard J, Martin G, Reuber L, Jiang C-Z, Keddie J, et al. *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *science*. 2000;290(5499):2105–10.
36. Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Res*. 2006;16(10):1252–61.
37. Grover C, Zhu X, Grupp K, Jareczek J, Gallagher J, Szadkowski E, et al. Molecular confirmation of species status for the allopolyploid cotton species, *Gossypium ekmanianum* Wittmack. *Genet Resour Crop Evol*. 2015;62(1):103–14.
38. Wendel JF, Cronn RC. Polyploidy and the evolutionary history of cotton. *Adv Agron*. 2003;78:139.
39. Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, Jakobsen KS, et al. Ancient hybridizations among the ancestral genomes of bread wheat. *science*. 2014;345(6194):1250092.
40. Chalhouf B, Denoëuf F, Liu S, Parkin IA, Tang H, Wang X, et al. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *science*. 2014;345(6199):950–3.
41. 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.
42. Ulmasov T, Murfett J, Hagen G, Guilfoyle TJ. Aux/IAA proteins repress expression of reporter genes containing natural and highly active synthetic auxin response elements. *Plant Cell*. 1997;9(11):1963–71.
43. Foster R, Izawa T, Chua NH. Plant bZIP proteins gather at ACGT elements. *FASEB J*. 1994;8(2):192–200.
44. Gilmartin PM, Memelink J, Hiratsuka K, Kay SA, Chua N-H. Characterization of a gene encoding a DNA binding protein with specificity for a light-responsive element. *Plant Cell*. 1992;4(7):839–49.
45. Lam E, Chua N-H. ASF-2: a factor that binds to the cauliflower mosaic virus 35S promoter and a conserved GATA motif in cab promoters. *Plant Cell*. 1989;1(12):1147–56.
46. Tran L-SP, Nakashima K, Sakuma Y, Simpson SD, Fujita Y, Maruyama K, et al. Isolation and functional analysis of *Arabidopsis* stress-inducible NAC transcription factors that bind to a drought-responsive cis-element in the early responsive to dehydration stress 1 promoter. *Plant Cell*. 2004;16(9):2481–98.
47. Creelman RA, Tierney ML, Mullet JE. Jasmonic acid/methyl jasmonate accumulate in wounded soybean hypocotyls and modulate wound gene expression. *Proc Natl Acad Sci*. 1992;89(11):4938–41.
48. Majda M, Robert S. The role of auxin in cell wall expansion. *Int J Mol Sci*. 2018;19(4):951.
49. Villarreal NM, Marina M, Nardi CF, Civello PM, Martínez GA. Novel insights of ethylene role in strawberry cell wall metabolism. *Plant Sci*. 2016;252:1–11.
50. Xu G, Guo C, Shan H, Kong H. Divergence of duplicate genes in exon–intron structure. *Proc Natl Acad Sci*. 2012;109(4):1187–92.
51. Roy SW, Gilbert W. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet*. 2006;7(3):211–21.
52. Qanmber G, Ali F, Lu L, Mo H, Ma S, Wang Z, et al. Identification of histone H3 (*H3H3*) genes in *Gossypium hirsutum* revealed diverse expression during ovule development and stress responses. *Genes*. 2019;10(5):355.
53. Lecharny A, Boudet N, Gy I, Aubourg S, Kreis M. Introns in, introns out in plant gene families: a genomic approach of the dynamics of gene structure. *J Struct Funct Genom*. 2003;3(1):111–6.
54. Roy SW, Penny D. A very high fraction of unique intron positions in the intron-rich diatom *Thalassiosira pseudonana* indicates widespread intron gain. *Mol Biol Evol*. 2007;24(7):1447–57.
55. Roosjen M, Paque S, Weijers D. Auxin response factors: output control in auxin biology. *J Exp Bot*. 2018;69(2):179–88.
56. Bouzroud S, Gouiaa S, Hu N, Bernadac A, Mila I, Bendaou N, et al. Auxin response factors (*ARFs*) are potential mediators of auxin action in tomato response to biotic and abiotic stress (*Solanum lycopersicum*). *PLoS One*. 2018;13(2):e0193517.
57. Huang P-Y, Catinot J, Zimmerli L. Ethylene response factors in *Arabidopsis* immunity. *J Exp Bot*. 2016;67(5):1231–41.
58. Hou J, Lu D, Mason AS, Li B, Xiao M, An S, et al. Non-coding RNAs and transposable elements in plant genomes: emergence, regulatory mechanisms and roles in plant development and stress responses. *Planta*. 2019;250(1):23–40.
59. Liu G, Wu M, Pei W, Li X, Wang N, Ma J, et al. A comparative analysis of small RNAs between two upland cotton backcross inbred lines with different fiber length: expression and distribution. *Crop J*. 2019;7(2):198–208.
60. Yin Z, Li Y, Yu J, Liu Y, Li C, Han X, et al. Difference in miRNA expression profiles between two cotton cultivars with distinct salt sensitivity. *Mol Biol Rep*. 2012;39(4):4961–70.
61. Xie F, Jones DC, Wang Q, Sun R, Zhang B. Small RNA sequencing identifies miRNA roles in ovule and fibre development. *Plant Biotechnol J*. 2015;13(3):355–69.
62. Zhu T, Liang C, Meng Z, Sun G, Meng Z, Guo S, et al. CottonFGD: an integrated functional genomics database for cotton. *BMC Plant Biol*. 2017;17(1):1–9.
63. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014;30(9):1236–40.
64. Letunic I, Doerks T, Bork P. SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res*. 2015;43(D1):D257–60.
65. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*. 2016;44(D1):D279–85.
66. Yu CS, Lin CJ, Hwang JK. Predicting subcellular localization of proteins for gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci*. 2004;13(5):1402–6.

67. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*. 2007;23(21):2947–8.
68. Trifinopoulos J, Nguyen L-T, von Haeseler A, Minh BQ. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res*. 2016;44(W1):W232–5.
69. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*. 1997;25(24):4876–82.
70. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. *Genome Res*. 2004;14(6):1188–90.
71. Hu B, Jin J, Guo A-Y, Zhang H, Luo J, Gao G. GSDS 2.0: an upgraded gene feature visualization server. *Bioinformatics*. 2015;31(8):1296–7.
72. Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res*. 2006;34(suppl_2):W369–73.
73. Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, et al. TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol Plant*. 2020;13(8):1194–202.
74. Lescot M, Déhais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, et al. Plant-CARE, 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.
75. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. 2006;34(suppl_2):W609–12.
76. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007;24(8):1586–91.
77. 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.
78. Zhang Z, Yu J, Li D, Zhang Z, Liu F, Zhou X, et al. PMRD: plant microRNA database. *Nucleic Acids Res*. 2010;38(suppl_1):D806–13.
79. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*. 2014;42(D1):D68–73.
80. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *methods*. 2001;25(4):402–8.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

