Genome-wide identification, characterization and expression profiles of FORMIN gene family in cotton (*Gossypium Raimondii* L.)

Pollob Shing[1†], Md Shohel Ul Islam[1†], Mst. Sumaiya Khatun[1], Fatema Tuz Zohra[2], Naimul Hasan[1], Shaikh Mizanur Rahman[1] and Md. Abdur Rauf Sarkar[1*†]

Abstract

Background *Gossypium raimondii* serves as a widely used genomic model cotton species. Its genetic influence to enhance fiber quality and ability to adapt to challenging environments both contribute to increasing cotton production. The formins are a large protein family that predominately consists of FH1 and FH2 domains. The presence of the formin domains highly regulates the actin and microtubule filament in the cytoskeleton dynamics confronting various abiotic stresses such as drought, salinity, and cold temperatures.

Results In this study, 26 formin genes were analyzed and characterized in *G. raimondii* and mostly were found in the nucleus and chloroplast. According to the evolutionary phylogenetic relationship, GrFH were dispersed and classified into seven different groups and shared an ancestry relationship with MtFH. The *GrFH* gene structure prediction revealed diverse intron-exon arrangements between groups. The FH2 conserved domain was found in all the *GrFH* distributed on 12 different chromosomes. Moreover, 11 pairs of *GrFH* transpired segmental duplication. Among them, *GrFH4-GrFH7* evolved 35 million years ago (MYA) according to the evolutionary divergence time. Besides, 57 *cis*-acting regulatory elements (CAREs) motifs were found to play a potential role in plant growth, development, and in response to various abiotic stresses, including cold stress. The *GrFH* genes mostly exhibited biological processes resulting in the regulation of actin polymerization. The ERF, GATA, MYB, and LBD, major transcription factors (TFs) families in *GrFH*, regulated expression in abiotic stress specifically salt as well as defense against certain pathogens. The microRNA of *GrFH* unveiled the regulatory mechanism to regulate their gene expression in abiotic stresses such as salt and cold. One of the most economic aspects of cotton (*G.raimondii*) is the production of lint due to its use in manufacturing fabrics and other industrial applications. The expression profiles of *GrFH* in different tissues particularly during the conversion from ovule to fiber (lint), and the increased levels (up-regulation) of *GrFH4*, *GrFH6*, *GrFH12*, *GrFH14*, and *GrFH26* under cold conditions, along with *GrFH19* and *GrFH26* in response to salt stress, indicated their potential involvement in combating these environmental challenges. Moreover, these stress-tolerant *GrFH* linked to cytoskeleton dynamics are essential in producing high-quality lint.

†Pollob Shing, Md. Abdur Rauf Sarkar and Md Shohel Ul Islam contributed equally to this work..

*Correspondence:
Md. Abdur Rauf Sarkar
rauf.gebt@yahoo.com
Full list of author information is available at the end of the article

Shing *et al. BMC Genomic Data*   (2024) 25:105

Page 2 of 20

Conclusions The findings from this study can contribute to elucidating the evolutionary and functional characterizations of formin genes and deciphering their potential role in abiotic stress such as cold and salt as well as in the future implications in wet lab.

Keywords Cotton, Formin gene family, Genome-wide identification, Abiotic stress treatments, Expression profiles

Background

Formins are the protein family recognized by the presence of FH1 (Formin Homology 1) and FH2 (Formin Homology 2) domains that can multimerize the actin protein in cytoskeletal dynamics through nucleating the bared end of the actin filament [1, 2]. The FH1 domain comprises polyproline interactions with the profilin-actin binding domain complex. As FH1 resides in the N terminal of the FH2 domain, FH1 transports the profilin-actin binding domain complex to the barbed ends (growing part) of the FH2 domain. The profilin induces the elongation rate of the barbed end of both the F1 and FH2 domains. The FH2 domain will polymerase the actin filament in the bared end for extension [3–5].

A cytoskeleton is a structure that is composed of actin filaments and microtubules, which mediate cell proliferation, and support the internal structure and movement of the cell [6]. Besides all these functions, the most crucial aspect of cytoskeleton dynamics is that upon the stress condition from the surrounding environment, it ensures cellular stability by preventing various parts of the plant such as roots and shoots, from being susceptible to stress [2]. This transpires through the alternation of cellular morphology to fight against stress and pathogens [7].

The mechanism of tolerance to the abiotic stress involves cytoskeleton dynamics, both actin filament and microtubule. The actin filament of cytoskeleton dynamics is responsible for the proper growth and advancement of the plant. But perhaps the pivotal role of actin filament, according to the genetic and pharmacological analysis, is the guarding of cell shape changes [8–10]. During drought stress, it is observed in *Arabidopsis thaliana* that stomatal opening and closing are highly monitored and regulated by the actin filament of cytoskeleton dynamics. In terms of organized and symmetrically distributed actin filament, the stomata remain open. However, the longitudinal actin filament represents the closing of the stomata. This remodeling and reshuffling of cellular changes during stress is controlled by actin-binding proteins (ABPs). ABP further consists of actin-depolymerizing factors (ADFs) that are responsible for growth and elongation of cells, nonspecific immunity to cells, as well as regulation of stomata [11–13].

On the other hand, another cytoskeletal dynamics, microtubules (MT), which are composed of α-tubulin and β-tubulin, play an essential role in terms of tolerance to stress. In salt stress, the polymerization and de-polymerization of MT enhance tolerance to stress. During the stress condition, various signaling molecules such as abscisic acid (ABA), cytosolic calcium ion, and reactive oxygen species (ROS) are key modulators for stress-controlling systems in plants. The presence of high-level cytosolic calcium ions in the calcium channel and the existence of ABA, a phytohormone, in the root ensuring the reconstruction of cortical MT lead to the de-polymerization of MT. Whereas ROS re-polymerizes by dismantling and rebuilding the non-uniform polymers of MT [6, 14–16].

As a dicot, cotton is one of the most valuable and economically lucrative plant species on the entire planet, producing fiber that is extensively utilized in the textile industry. Apart from its implementation in textile manufacturing, it serves as a model scheme in the field of cell wall biosynthesis and cellular elongation research. The D5 diploid *G. raimondii* is the predominant contributor of pollen [17–19]. According to the Food and Agriculture Organization (FAO), factors related to the environment, agroecology, and atmosphere affect the growth and cultivation of cotton [20]. Therefore, stress tolerant genes are essential for proper growth and production of quality lint.

In this study, 26 *GrFH* genes containing conserved formin domains were analyzed. The gene structure analysis showed similarity within the subgroups, while 11 pairs of segmental duplications were observed in 12 chromosomes. The evolutionary phylogenetic tree demonstrated a closed ancestry relationship of GrFH with MtFH. Besides, 57 *cis*-acting regulatory elements were found in *GrFH*, revealing their involvement in plant development along with response in stress conditions. ERF and GATA were the most abundant transcription factor families to bind *GrFH*. The protein-protein interactions showed close proximity with *Arabidopsis* proteins. Based on the gene ontology analysis, *GrFH* functions were classified into three groups, while most of them were involved in biological processes. About 96 unique microRNAs were identified to regulate gene expression in various abiotic stresses, specifically cold and salt stress. Besides, the tissue-specific expression of *GrFH* revealed its relations with the production of lint, which are essential for industrial and chemical applications. Moreover, certain *GrFH* showed resistance to abiotic stress such as cold and salt.

Shing *et al. BMC Genomic Data* (2024) 25:105

Page 3 of 20

Therefore, these stress-tolerant *GrFH* would maintain the production of fiber in case of a cold and salt stressed conditions.

Results

Identification and determination of physiochemical properties of GrFH (*G. Raimondii* Formin homology*)*

Among the predicted protein sequences, 26 of them were identified as FH2-containing domains, which were subsequently renamed as potential genes. For example, Gorai.001G078900 was renamed to GrFH1. The examination of the amino acid (aa) counts across the 26 GrFH proteins showed a range from 71 (GrFH20) to 2274 (GrFH10) (Table 1). According to the molecular weight measured in kilo-Dalton (kDa), GrFH20 (8283.71 kDa) had the lowest amino acid contents among the 26 encoded GrFH proteins, while GrFH10 (241075.93 kDa) had the highest amino acid contents. The pH at which a protein carries no net electrical charge (neutral state) and a protein's pH level at which it exhibits no net electrical

charge (neutral state) acts as a zwitterion ion is known as its isoelectric point (pI) [21, 22]. The pI (Isoelectric point) value varied from a range of 5.31 (GrFH3) to 9.35 (GrFH12). The result instability index showed that all the GrFH proteins were considered unstable since the score was greater than 40. Moreover, the results generated from the aliphatic index revealed that 10 GrFH proteins (GrFH1, GrFH3, GrFH4, GrFH8, GrFH13, GrFH16, GrFH20, GrFH21, GrFH23, and GrFH26) contained aliphatic index above 80, while the remaining 16 GrFH proteins ranged between 70 and 80. As far as the grand average of hydropathicity (GRAVY) was concerned, all of the GrFH proteins surpassed the negative value.

Phylogenetic analysis between GrFH, MtFH, ZmFH, OsFH, and AtFH

To demonstrate the molecular and evolutionary relationship between formin proteins, a phylogenetic tree containing 26 proteins from GrFH as candidate species, 21 from AtFH, 20 from ZmFH, 19 from MtFH,

Table 1 List of 26 GrFH proteins and their basic physio-chemical characterization

Gene name	Gene identifier	Size (aa)	Mass (kDa)	pI	Instability index	Aliphatic index	Grand average of hydropathicity (GRAVY)
GrFH1	Gorai.001G078900	1327	146101.24	8.20	54.21	81.61	−0.402
GrFH2	Gorai.001G100900	1036	113548.04	5.68	65.20	72.86	−0.494
GrFH3	Gorai.002G144100	203	23169.67	5.31	56.22	89.80	−0.492
GrFH4	Gorai.004G028300	951	103661.40	6.59	48.80	83.20	−0.320
GrFH5	Gorai.004G265000	1249	136516.77	8.14	62.14	79.42	−0.418
GrFH6	Gorai.005G008700	932	103122.95	8.42	68.38	75.53	−0.493
GrFH7	Gorai.005G187300	802	87432.31	8.79	62.75	74.89	−0.481
GrFH8	Gorai.006G052500	257	29060.80	8.92	65.97	81.21	−0.570
GrFH9	Gorai.006G267200	1486	161338.19	6.74	57.37	77.05	−0.378
GrFH10	Gorai.007G038600	2274	241075.93	6.49	84.86	71.37	−0.402
GrFH11	Gorai.007G161600	1114	121498.04	8.73	74.75	70.81	−0.499
GrFH12	Gorai.007G222600	923	101719.49	9.35	59.95	73.08	−0.569
GrFH13	Gorai.007G299400	871	95876.76	8.61	53.42	83.13	−0.380
GrFH14	Gorai.008G082700	896	98375.78	8.99	58.61	73.92	−0.556
GrFH15	Gorai.008G150300	798	87346.42	8.27	51.44	72.94	−0.549
GrFH16	Gorai.008G195600	1173	130746.96	6.06	45.70	81.13	−0.453
GrFH17	Gorai.009G062500	1322	145646.73	6.68	61.57	78.71	−0.437
GrFH18	Gorai.009G427600	920	98560.22	6.65	52.56	79.55	−0.459
GrFH19	Gorai.010G112600	1242	137358.46	7.37	55.83	79.79	−0.395
GrFH20	Gorai.011G119700	71	8283.71	7.79	51.10	82.39	−0.532
GrFH21	Gorai.011G132600	918	101118.83	8.00	61.61	80.38	−0.412
GrFH22	Gorai.012G001900	681	76243.24	6.41	51.93	79.07	−0.316
GrFH23	Gorai.013G125400	806	89206.67	7.07	54.93	83.56	−0.439
GrFH24	Gorai.013G164400	959	103208.74	5.46	50.29	77.65	−0.446
GrFH25	Gorai.013G216500	1179	129312.38	6.34	58.07	77.79	−0.422
GrFH26	Gorai.013G218400	260	29293.01	6.87	64.82	86.65	−0.529

Shing *et al. BMC Genomic Data*　(2024) 25:105

Page 4 of 20

and 17 from OsFH was constructed (Fig. 1). A total of seven different categories designated as groups A, B, C, D, E, F, and G were formed from 103 formin proteins from five distinct species. GrFH were distributed in those above-mentioned groups (Supplementary Material File 1). There were 2 candidate formin proteins in each of the three groups denoted as A, E, and F. GrFH7 and GrFH22 were found in group A at the same time GrFH13 along with GrFH23 were identified in groups E, and GrFH12 and GrFH14 were found in group F. In groups B, C, and D, each one of them contained 4 formin proteins. GrFH1, GrFH9, GrFH10, and GrFH16 were found in group B, while GrFH5, GrFH17, GrFH19, and GrFH25 emerged in group C. GrFH4, GrFH15, GrFH18, and GrFH24 belonged to the group D. However, with eight GrFH proteins from group G- GrFH2, GrFH3, GrFH6, GrFH8, GrFH11, GrFH20, GrFH21, and GrFH26, it notched up rank number one. Remarkably, each formin protein species (GrFH, AtFH, ZmFH,

MtFH, and OsFH) was spotted in each group at least once.

Conserved motif, domain and gene structure of GrFH

In conserved motifs, 20 unique motifs were analyzed in GrFH proteins (Fig. 2A). The groups that comprised 11 motifs were identified as group A (GrFH7 and GrFH22), group D (GrFH4, GrFH15, GrFH18, GrFH24), group E (GrFH13, except GrFH23, which featured an additional motif that summed to 12 motifs), group F (GrFH12, GrFH14), and some members of group G (GrFH2, GrFH6, GrFH11, GrFH21). All the group B members (GrFH1, GrFH9, GrFH10, GrFH16) contained 17 different motifs. The members of group G each possessed different motif numbers, as GrFH3, GrFH8, and GrFH26 each included six motifs. Surprisingly, GrFH20 from group G incorporated solo motif 2. The conserved domain was used to identify and investigate protein functions and evolutionary relationships. The FH2
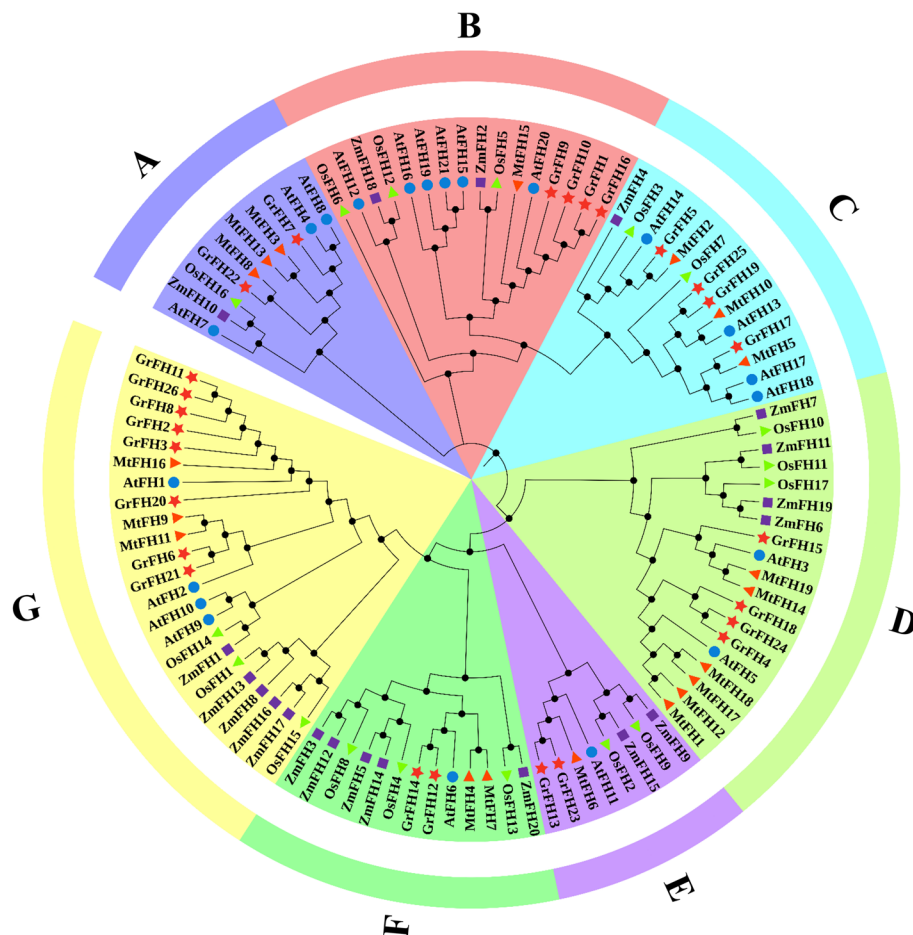

Fig. 1 Phylogenetic relationship between candidate GrFH and AtFH, ZmFH, MtFH, and OsFH gene families. GrFH were classified into 7 groups (**A**, **B**, **C**, **D**, **E**, **F**, and **G**), each marked by different colors and shapes. The candidate gene GrFH was labeled by the red star. Whereas AtFH was labeled as round blue, ZmFH was labeled as violet color square, MtFH was labeled orange triangular, and OsFH was labeled as green triangular
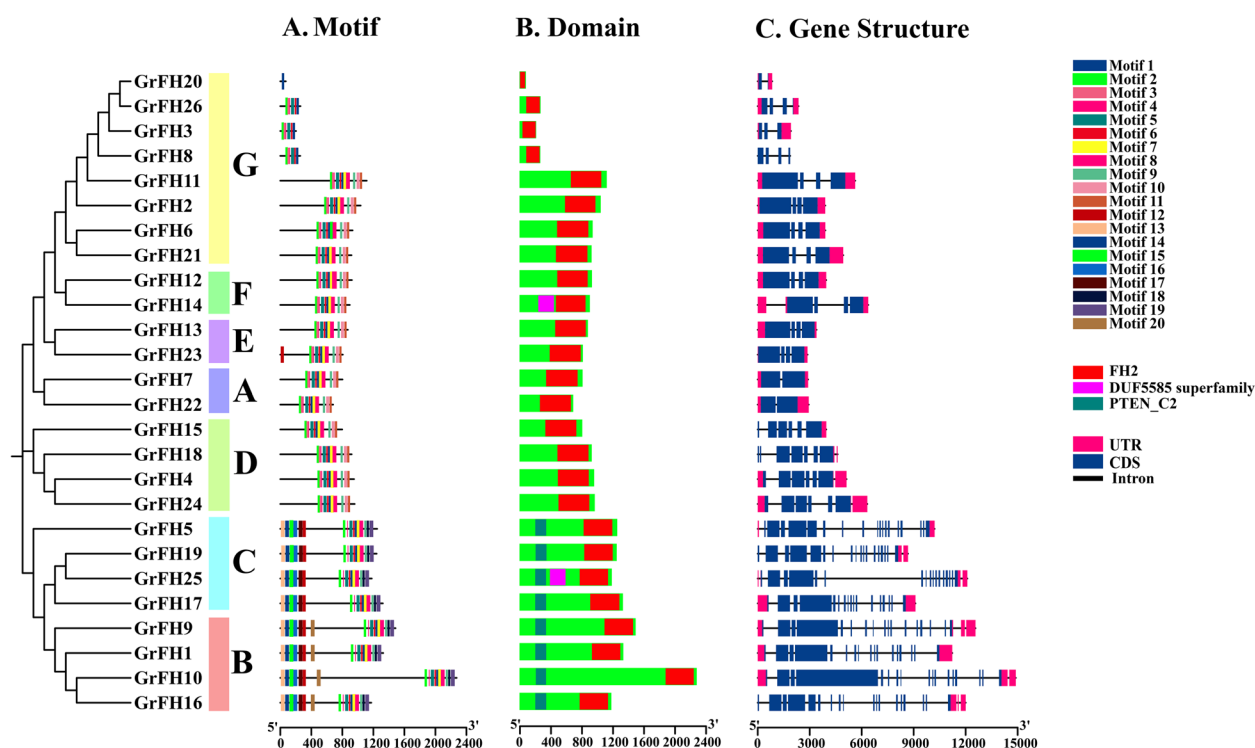
Shing *et al. BMC Genomic Data*   (2024) 25:105

Page 5 of 20

Fig. 2 The motif, domain, and gene structure of *GrFH* genes. **A** The grouping and colors of the *GrFH* gene family members are based on the phylogenetic relationship. Each motif is illustrated by a specific-colored box aligned on the right side of the figure. Different colors indicate individual motifs. **B** The positions of the FH2 conserved domain is demonstrated in red color, whereas the entire protein sequence of respective GrFH is green colored. **C** For the color bar of *GrFH* gene structure, black lines represent introns, blue represents exons, and deep pink lines represent upstream/downstream

conserved domain was found in all 26 GrFH proteins (Fig. 2B). Some of the proteins not only contained FH2 but also contained other domains. PTEN_C2 was found most frequently after FH2. On the contrary, the *GrFH* gene structure consisted of a number of different exons and introns (Supplementary Material File 2). Group A members, *GrFH7* and *GrFH22*, showed the least number of exons and introns counts among the seven groups at 4 and 2, respectively (Fig. 2C). The highest number of exons were observed in groups B and C accumulating 69 and 70, respectively. *GrFH1*, *GrFH9*, and *GrFH10* from Group B and *GrFH17* and *GrFH25* from Group C each included 17 exons and 16 introns in these two groups. Meanwhile, *GrFH16* and *GrFH5*, *GrFH19* members of group B and group C, respectively, possessed 18 exons and 17 introns. In group D, there were 25 exons and 21 introns in total. *GrFH4*, *GrFH15*, and *GrFH24* each had 6 exons, while *GrFH18* had 7 exons. When groups E and F exons and introns were counted, there were 16 exons and 12 introns from both groups. In the same way, *GrFH12* and *GrFH14* from group F and *GrFH13* and *GrFH23* from group E each contained 4 exons and 3 introns. Group G, the largest group, comprised a maximum of 29 exons and 21 introns. Six *GrFH* (GrFH2, *GrFH6*, *GrFH8*, *GrFH11*, *GrFH21*, and *GrFH26*) genes held with 4 exons in this largest group, whereas *GrFH3* and *GrFH20* featured with 3 and 2 exons correspondingly.

Synonymous (Ks) and non-synonymous (Ka) substitution ratios calculation of *GrFH*

The Ka (nonsynonymous substitution rate) and Ks (synonymous substitution rate) values for *GrFH* gene pairings were evaluated (Supplementary Fig. S1). A Ka/Ks ratio that is smaller than 1 denotes purifying selection, while a higher than 1 describes positive selection. The estimated Ka value in the *GrFH* genes varied from 0.003824 to 0.32905, whereas the Ks value varied from 0.4263 to 0.000495 (Supplementary Material File 3). One gene pair, *GrFH20-GrFH26*, was going through positive selection; its Ka/Ks ratio is 7.7257807. Apart from that, the Ka/Ks ratio of other gene pairs is less than 1. For example, Ka/Ks ratio is approximately 0.771874 in *GrFH6-GrFH21* along with 0.194305 in *GrFH18-GrFH5*, suggesting that those gene pairs have experienced purifying selection.

Shing *et al. BMC Genomic Data* (2024) 25:105

Page 6 of 20

Evolutionary collinear relationship analysis of *GrFH*

The collinear relationship showed a close relationship among *GrFH* genes (Fig. 3). With a total of 4 collinear pairings observed in chromosomes 7 and 13, it was identified as the highest levels of duplicated collinear forming pairs. The genes on chromosome 7 were *GrFH10*, *GrFH11*, *GrFH12*, and *GrFH13*, which created collinear pairings with *GrFH9* on chromosome 6, *GrFH8* on chromosome 6, *GrFH14* on chromosome 8, and *GrFH23* on chromosome 13, respectively. The genes *GrFH24*, *GrFH25*, and *GrFH26* located in chromosome 13 paired with *GrFH15* from chromosome 8, *GrFH1* from chromosome 1, and *GrFH20* from chromosome 11, respectively. Following these, there were four additional duplicated pairings: *GrFH4* on chromosome 4 paired with *GrFH7* on chromosome 5, *GrFH5* on chromosome 4 interacted with *GrFH18* on chromosome 9, and *GrFH6* on chromosome 5 linked with *GrFH21* on chromosome 11 and ultimately *GrFH17* on chromosome 9 formed pair *GrFH19* of chr10. Despite having specific loci, four *GrFH* genes (*GrFH2*, *GrFH3*, *GrFH16*, and *GrFH22*) on chromosomes 1, 2, 8,

and 12, respectively showed to contain no collinear pairings. But, no gene was detected on chromosome 3 which wasn't taken into account while counting chromosomes.

Evolutionary syntenic relationship analysis of *GrFH*

Syntenic relationships between *Z. mays*, *O. sativa*, *A. thaliana*, and the candidate *G. raimondii* formin gene were performed to provide insight into putative evolutionary relationships between the FORMIN gene in multiple species (Fig. 4). Unfortunately, no syntenic gene pairings emerged in the candidate gene *GrFH* with other species. Remarkably, among all the species, only the formin genes of *Z. mays* and *O. sativa* established syntenic pairing with each other. For example, chromosome 6 of *ZmFH1* formed a syntenic pairing with chromosome 5 of *OsFH14*.

Chromosomal localization and duplication analysis of *GrFH*

The duplication of genes and the formation of collinear pairings were identified as segmental types. 26 genes of *GrFH* were distributed across the 12 chromosomes at
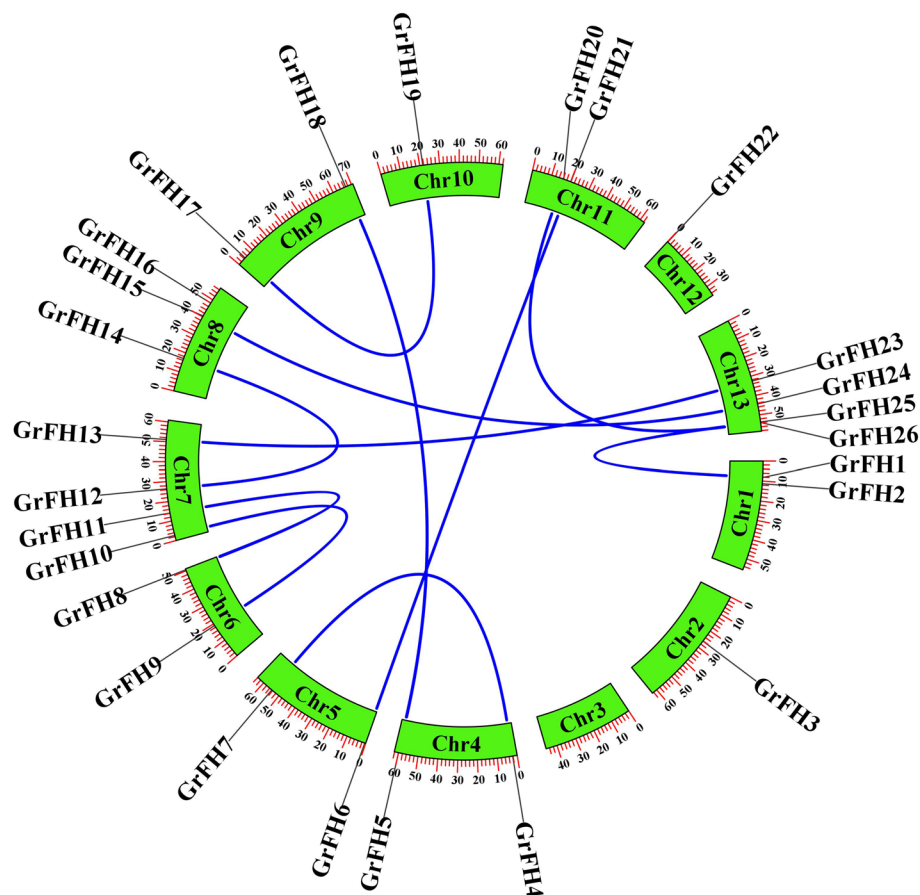

Fig. 3 The collinearity analysis of the *GrFH* gene family in cotton. Green color rectangles represent chromosomes 1–13 in *GrFH*. The dark blue colored lines linked between chromosomes represent collinear relations between them
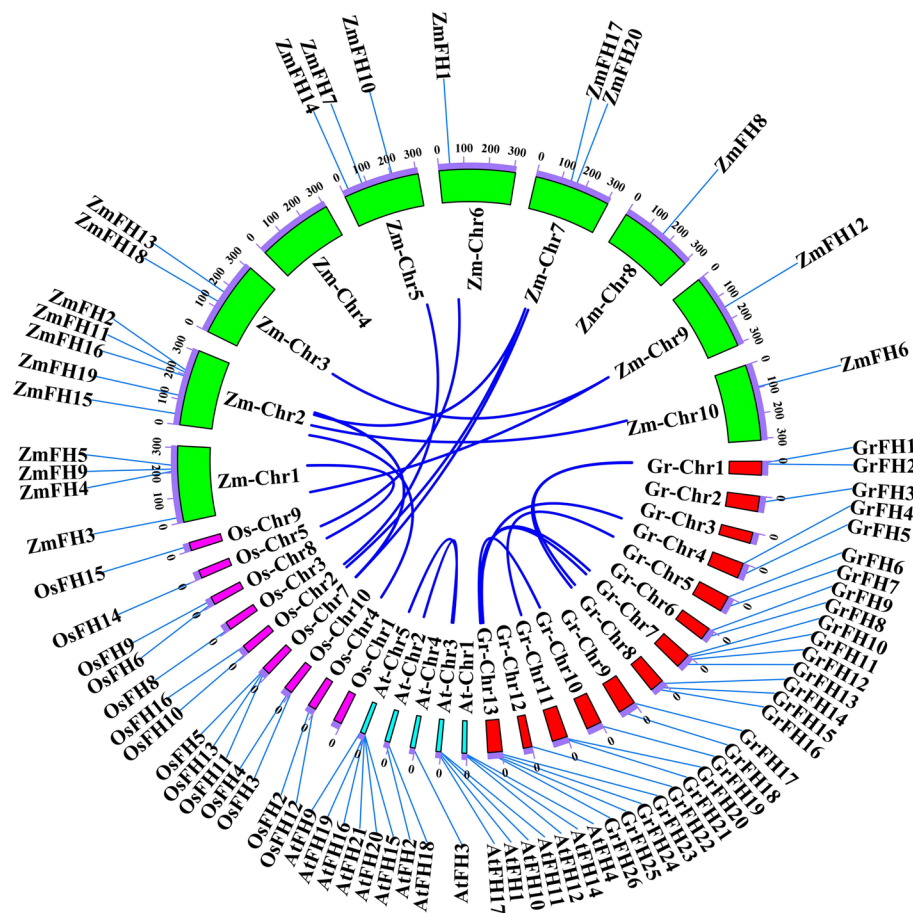
Shing *et al. BMC Genomic Data* (2024) 25:105

Page 7 of 20

Fig. 4 The syntenic relationship analysis of cotton and *Arabidopsis*, Rice, and Corn. Red color rectangles represent the *GrFH* chromosomes. Meanwhile, aqua-blue color rectangles represent *AtFH* chromosomes. Furthermore, magenta color rectangles represent *OsFH* chromosomes, while green color rectangles represent *ZmFH* chromosomes. The dark blue color represents the syntenic relationship linkage between different species

specific loci, and 11 gene pairs were duplicated (Supplementary Fig. S2). Chromosomal localization was carried out to figure out the gene's extract position across chromosomes. There was at least one single gene found on chromosome 2, chromosome 10, and chromosome 12, which went by the names *GrFH3*, *GrFH19*, and *GrFH22*, respectively. Chromosomes 1, 4, 5, 6, 9, and 11 contained 6 pairs of genes designated as *GrFH1-GrFH2*, *GrFH4-GrFH5*, *GrFH6-GrFH7*, *GrFH9-GrFH8*, *GrFH17-GrFH18*, and *GrFH20-GrFH21* were observed, respectively in the order of their location on a specific chromosome. The genes *GrFH14*, *GrFH15*, and *GrFH16* were found sequentially on chromosome 8 in the same order as their genomic positions. Chromosome 7 and 13 each contain four genes. Specifically, chromosome 7 includes *GrFH10*, *GrFH11*, *GrFH12*, *and GrFH13*, whereas chromosome 13 includes *GrFH23*, *GrFH24*, *GrFH25*, *and GrFH26*. Notably, no genes were found on chromosome 3.

Prediction of the subcellular localization of *GrFH*

Subcellular localization studies were conducted to determine the organelle locations of the *GrFH* genes. The result showed that 10 different organelles, including the nucleus, mitochondria, chloroplast, cytoplasmic, cytoskeleton, Golgi apparatus, vacuole, endoplasmic reticulum, plasma membrane, and lastly, extracellular existed. In terms of gene presence on the particular organelle, *GrFH8* and *GrFH11* were found on 12 sites of chloroplast (Fig. 5A). *GrFH4* and *GrFH6*, which are located at 10 and 11 different sites of the plasma membrane, respectively, came out as the third and fourth best, trailing only by *GrFH8* and *GrFH11*. The bubble plot, however, represented the redundancy of a particular *GrFH* gene in specific organelles. For example, *GrFH8* and *GrFH11* were found in 12 different locations in the chloroplast. When it comes to the overall number of *GrFH* gene distributions in each organelle, the chloroplast and nucleus outnumbered all other organelles, with percentages of 80.76%
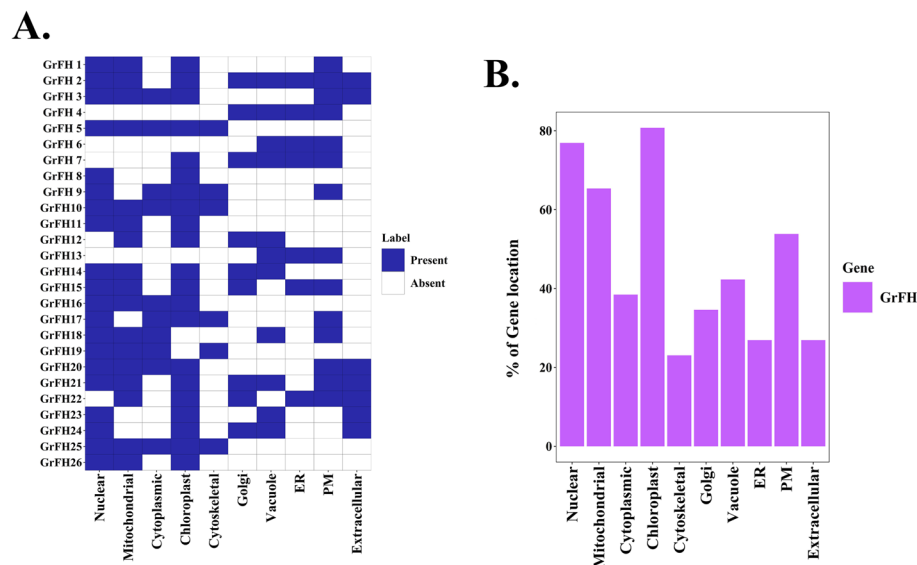
Fig. 5 Sub-cellular localization analysis of *GrFH*. **A** The heatmap represents the sub-cellular localization analysis of *GrFH*. The intensity of color on the right side of the heatmap indicates the presence of protein signals corresponding to the genes. The cellular organelles include nuclear, mitochondrial, cytoplasmic, chloroplast, cytoskeletal, Golgi, vacuole, endoplasmic reticulum (ER), plasma membrane (PM), and extracellular locations. **B** The percentage distribution of *GrFH* signal across various cellular organelles is represented by a bar diagram. The percentages of protein signals appearing in different cellular organelles are shown on the left side of the diagram

and 76.92%, respectively (Fig. 5B). Mitochondria along with plasma membrane weren't far behind, with percentages of 65.38% and 53.84%, respectively. The cytoskeleton proved to comprise the lowest concentration of *GrFH* genes. However, *GrFH* genes were observed in sufficient amounts in the rest of the organelles as well.

Cis-acting regulatory elements (CAREs) analysis in the promoters of *GrFH*

The analysis regarding the 2000 bp sequence (Supplementary Material File 4) CAREs was conducted to get a deeper comprehension of the regulatory framework driving the promoter region of *cis*-elements (Fig. 6). The result of the CAREs showed that 57 CARE motifs were found in *GrFH* (Supplementary Material File 5). Based on the functional similarities of the 57 CAREs, four particular categories were formed that is light responsiveness, tissue-specific expression, phytohormone responsiveness, and stress responsiveness. Out of these four categories, 24 motifs featuring light responsiveness were identified as the highest containing *cis*-acting regulatory elements.24 motifs in the light responsiveness segment included as AAAC-motif, ACE, AE-box, AT1-motif, ATC-motif, ATCT-motif, Box 4, Box II, chs-CMA1a, GA-motif, Gap-box, GATA-motif, G-Box, G-box, GT1-motif, GTGGC-motif, I-box, LAMP-element, LS7, MRE, Sp1, TCCC-motif, TCT-motif, 4 cl-CMA2b. From all these motifs of the

light responsiveness group, Box 4 exhibited the most prevalent response to light. As far as genes in *GrFH* were concerned, *GrFH14* comprised 18 Box 4 motifs. There was also a substantial amount of Box 4 motif expressed in other *GrFH* genes. Moreover, GT1-motif, part of a light-responsive element, was found 16 times in the *GrFH4* gene. All in all, all the *cis*-elements representing light responsiveness played a significant role in the CARE analysis. Tissue-specific expression, which ensured 16 *cis*-elements, was the next largest group. The 16 motifs included a 3-AF3 binding site, A-box, ARE, AT-rich element, Box II-like sequence, CAT-box, CCAAT-box, circadian, GCN4_motif, HD-Zip 1, HD-Zip 3, Box III, MBSI, MSA-like, O2-site and RY-element. From this group, the ARE motif acting as a *cis*-acting regulatory element essential for the anaerobic induction was observed to be the most often occurring response in this cohort. Besides, the ARE motif was repeated 7 times in the *GrFH21* gene. In comparison to ARE, other motifs in this category were less frequently noticed. Approximately 12 *cis*-elements were observed in the phytohormone responsiveness group. As this group regulates the hormonal response, the ABRE motif functioning as a regulator of abscisic acid was most frequently expressed. The stress responsiveness group consisting of DRE core, LTR, MBS, TC-rich repeat, and WUN-motif appeared to construct a less significant amount of expression compared to the other 3 groups.
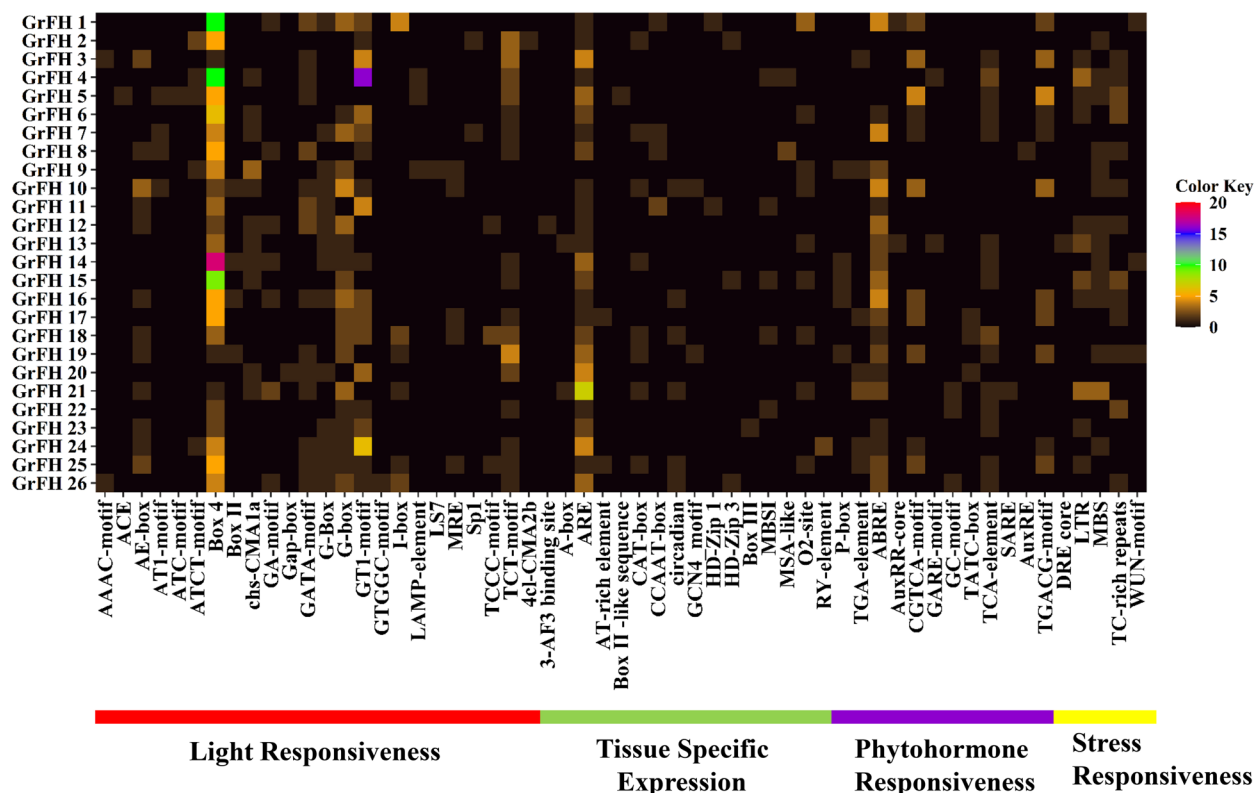
Fig. 6 The distribution of putative *cis*-acting regulatory elements *GrFH* is represented by a heatmap. The names of each *GrFH* are shown on the left side of the heatmap. The number of putative *cis*-acting elements for each *GrFH* gene is displayed on the right side of the heatmap and is represented by five different colors (black = 0, orange = 1–5, green = 6–10, blue = 11–15, and red = 16–20). Functions associated with *cis*-acting elements of the corresponding genes, such as light responsiveness, tissue-specific expression, phytohormone responsiveness, and stress responsiveness, are shown at the bottom of the heatmap and are labeled red, green, dark violet, and yellow respectively

Gene ontology (GO) analysis of *GrFH*

GrFH genes encompassed 64 GO enrichment id that were broadly classified into three distinct groups based on gene information and functional annotation Molecular Function (MF), Cellular Component (CC), and Biological Process (BP) (Supplementary Fig. S3). Biological Process (BP) was observed as the most prominent group among these categories, accumulating 49 GO IDs. The genes *GrFH4, GrFH5, GrFH7, and GrFH11* featured in biological process bearing the following GO ids (Supplementary Material File 6); GO:0007015 (*p*-value: 2.50E-11), GO:0030036 (*p*-value: 4.40E-11), GO:0030029 (*p*-value: 5.30E-11), GO:0007010 (*p*-value: 6.40E-09, GO:0071822 (*p*-value: 7.90E-08), GO:1902589 (*p*-value: 2.50E-07, GO:0022607 (*p*-value: 3.40E-07, GO:0043933 (*p*-value: 6.10E-07, GO:0044085 (*p*-value: 2.20E-06, GO:0006996 (*p*-value: 5.70E-06, GO:0016043 (*p*-value: 1.03E-02, GO:0071840 (*p*-value: 1.46E-02). *GrFH4, GrFH7, and GrFH11,* which were responsible for the regulation of biological processes contained GO ids as : GO:0030838 (*p*-value: 1.60E-09), GO:0045010 (*p*-value: 1.60E-09), GO:0032273 (*p*-value: 2.40E-09), GO:0031334

(*p*-value: 3.00E-09), GO:0051495 (*p*-value: 3.00E-09), GO:0030041 (*p*-value: 5.80E-09), GO:0030833 (*p*-value: 5.80E-09), GO:0008064 (*p*-value: 2.23E-06), GO:0030832 (*p*-value: 2.23E-06), GO:0032271 (*p*-value: 2.23E-06), GO:0044089 (*p*-value: 2.23E-06), GO:0008154(*p*-value: 2.23E-06), GO:0032956 (*p*-value: 2.23E-06), GO:0032970 (*p*-value: 2.23E-06), GO:0043254 (*p*-value: 2.23E-06), GO:0051493 (*p*-value: 4.34E-06), GO:0010638 (*p*-value: 7.11E-06), GO:0051258 (*p*-value: 7.45E-06), GO:0032535 (*p*-value:1.25E-05), GO:0090066 (*p*-value: 1.25E-05), GO:0051130 (*p*-value: 1.43E-05), GO:0044087 (*p*-value: 1.43E-05), GO:0033043 (*p*-value: 5.94E-05), GO:0043623 (*p*-value: 3.16E-04), GO:0051128 (*p*-value: 3.40E-04), GO:0006461 (*p*-value: 1.22E-03), GO:0070271 (*p*-value: 1.25E-03), GO:0034622 (*p*-value: 1.27E-03), GO:0065003 (*p*-value: 2.10E-03), GO:0048522 (*p*-value: 3.01E-03), GO:0048518 (*p*-value: 1.38E-02), GO:0065008 (*p*-value: 1.86E-02). Regarding the cellular component, nine GO IDs were found. The estimated range of the *p*-value was 4.20E-06 to 9.93×10^{-3}. On the other hand, *GrFH* genes retained the lowest GO ID counting in the molecular function. The GO IDs were GO:0051015 (*p*-value:

9.07E-05), GO:0005515 (*p*-value: 4.18E-03), GO:0003779 (*p*-value: 1.51E-02), GO:0032403 (*p*-value: 2.68E-02), GO:0044877 (*p*-value: 7.95E-02), and GO:0008092 (*p*-value: 7.95E-02).

Transcription factors analysis of *GrFH*

In the analysis of transcription factors (TFs), 36 unique TFs were discovered. A total of seven groups, namely ERF, GATA, MYB, LBD, TALE, and C2H2, along with bZIP, were formed according to the similarities within the TF family (Fig. 7). ERF represented 25 of the overall 36 TFs, accounting for approximately 69.44%. Among the 25 TFs in the ERF family, Gorai.008G271200, Gorai.005G200200, Gorai.001G033200, Gorai.008G153600, Gorai.005G116500, Gorai.001G036500, and Gorai.010G048900 were most frequently featured. Besides, the TF families of GATA (Gorai.013G055000, Gorai.006G054800, and Gorai.005G065300) and MYB (Gorai.001G177100, Gorai.004G269400, Gorai.001G148500) each possessed three TFs that were considerably less prevalent than the TFs of ERF. However, the LBD family, which consisted of two TF, Gorai.007G350300 and Gorai.007G075700, was substantially more active than the GATA and MYB families. The TALE,

C2H2, and bZIP TF families each contained a single TF, Gorai.010G029000, Gorai.009G128000, and Gorai.009G285000 respectively. But Gorai.009G128000 from the C2H2 TF family featured more prominently than other TF members.

Regulatory network between TFs and *GrFH*

Prediction of the regulatory network between TFs and *GrFH* was performed to know about their connection. The analysis revealed that 36 TF members were found to interact with both the 26 *GrFH* and with themselves (Supplementary Fig. S4). The largest ERF family interacted with all of the *GrFH* genes except *GrFH3*, *GrFH20*, and *GrFH22*. The ERF family members also engaged with other TF family members. The TFs of the GATA and MYB families formed the interaction with the *GrFH* genes such as *GrFH1*, *GrFH2*, *GrFH5*, *GrFH6*, *GrFH7*, *GrFH9*, *GrFH10*, *GrFH12*, *GrFH18*, *GrFH23*, and *GrFH25*. However, LBD and C2H2 families TF members were observed to link with *GrFH* genes more frequently compared to MYB and TALE families TF members. bZIP, on the other hand, interacted exclusively with a single *GrFH* gene, *GrFH20*.
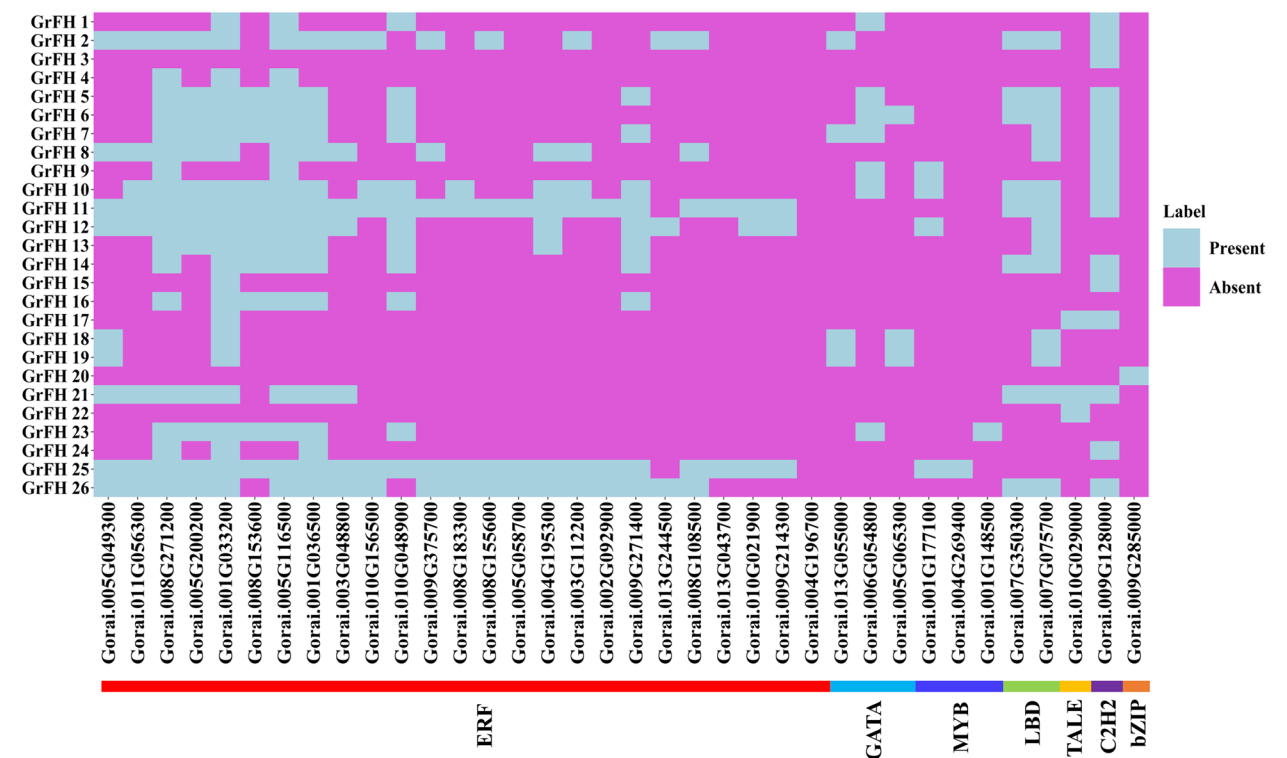

Fig. 7 A heatmap represents transcription factors (TFs) in *GrFH*. The 7 TFs family are ERF, GATA, MYB, LBD, TALE, C2H2, and bZIP which are colored red, sky blue, purple-blue, green, golden, purple, and orange respectively

Prediction of potential micro-RNAs targeting *GrFH*

MicroRNAs (miRNAs) are the 20–24 bp long noncoding RNA [23]. The principal function of miRNA is the regulation of gene expression through the inhibition of translation and cleavage of the targeted mRNA [24, 25]. They regulate gene expression through the RNA-induced silencing complex (RISC), which is composed of the ARGONAUTE (AGO) protein [26]. So, to understand the complex regulatory mechanism of miRNA in *GrFH* genes, 290 mature sequences of putative miRNA belonging to 96 unique families, targeting 25 different *GrFH* genes, were extracted (Supplementary Material File 7). The 290 putative miRNAs targeting 25 *GrFH* genes were shown in the network illustration (Supplementary Fig. S5 A) and the schematic diagrams indicate the *GrFH* genes targeted by miRNAs (Supplementary Fig. S5 B). It was observed that 22 members of gra-miR8762 targeted the highest up to eleven distinct *GrFH* genes (*GrFH1*, *GrFH2*, *GrFH4*, *GrFH5*, *GrFH6*, *GrFH9*, *GrFH10*, *GrFH13*, *GrFH15*, *GrFH18*, and *GrFH24*). Furthermore, 19 members of gra-miR530 targeted nine different *GrFH* genes (*GrFH1*, *GrFH9*, *GrFH10*, *GrFH14*, *GrFH15*, *GrFH16*, *GrFH17*, *GrFH19*, and *GrFH25*) (Table 2); 16 members of gra-miR7494 regulated ten distinct *GrFH* genes (*GrFH4*, *GrFH5*, *GrFH6*, *GrFH9*, *GrFH11*, *GrFH12*, *GrFH15*, *GrFH17*, *GrFH22*, and *GrFH25*). Besides, 12 family members of gra-miR7492 controlled five *GrFH* genes (*GrFH2*, *GrFH10*, *GrFH13*, *GrFH15*, and *GrFH17*) by cleaving mRNA and translational inhibition. However, the majority of the gra-miR family targeted one gene at a particular time. For example, gra-miR399 targeted only *GrFH13*. Some of the gra-miR targeted two or more *GrFH* genes. One such instance was noticed in gra-miR7504 that targeted *GrFH5*, *GrFH9*, and *GrFH24*. On the contrary, *GrFH1* was targeted most frequently, as much as 30 times.

Protein-protein interactions of GrFH

Arabidopsis known protein was applied to carry out protein-protein interaction network analysis of GrFH (Supplementary Material File 8). The choice of GrFH regions as STRING proteins was made depending on their higher homology with *Arabidopsis* proteins. Moreover, all the GrFH were linked with *Arabidopsis* proteins (Supplementary Fig. S6). Four GrFH proteins (GrFH1, GrFH9, GrFH10, and GrFH16) were noticed homologous with AtFH20. They were also seen forming strong relations with FH5. However, AtFH1 was identified homologous with 8 GrFH proteins (GrFH2, GrFH3, GrFH6, GrFH8, GrFH11, GrFH20, GrFH21, and GrFH26) along with bonding firm interactions with PRF1, PRF2, PRF4, ARPC5A and FH14. GrFH4, GrFH11, GrFH15, and GrFH24 were homologous with AtFH5, and the proteins interacted with PRF 1, 2, 3, and 4, FH6 as well as FH13. AtFH4 was identified as homogenous with GrFH7 and 20. Besides they discovered strong interactions between PRF 3, 4, and 5 as well as FH5. Three AtFH proteins AtFH6, AtFH11, and AtFH13 were homologous with GrFH12 and GrFH14, GrFH13 and GrFH23, and finally GrFH19 and GrFH25 respectively. They showed strong interactions with PRF1, PRF2, PRF4, PRF15, FIM5, FH1, and FH5. But GrFH17 was homologous to AtFH18 which showed interactions with T6P5.20. The broader line connecting proteins indicates the greater interaction ratio conversely. The biological roles of the GrFH proteins that strongly engage with *Arabidopsis* proteins could be equivalent.

Tissue-specific expression pattern analysis of *GrFH*

The expression of *GrFH* genes in various tissues, including ovule, leaf, and fiber, demonstrated that approximately 96% of *GrFH* were expressed in 0dpa (0-days post anthesis) ovule (Fig. 8). Likewise, 13 *GrFH* genes (*GrFH2*, *GrFH4*, *GrFH5*, *GrFH6*, *GrFH7*, *GrFH10*, *GrFH11*, *GrFH14*, *GrFH15*, *GrFH17*, *GrFH21*, *GrFH22*, and *GrFH26*) were highly expressive (Supplementary Material File 9). However, the expression patterns changed drastically in the 3dpa ovule as only *GrFH6*, *GrFH8*, and *GrFH9* were highly expressed. Although the overall expression of *GrFH* genes in 3dpa ovule was noticed at approximately 84.6%. On the contrary, the *GrFH* genes were expressed more highly in vegetative tissue (mature leaf) than in reproductive tissue (0 dpa and 3 dpa ovule). In the mature leaf, the expression of all the *GrFH* genes was observed. *GrFH1*, *GrFH3*, *GrFH18*, *GrFH19*, *GrFH20*, *GrFH23*, and *GrFH24* were

Table 2 Information about abundant miRNA ID, functions, and their targeted *GrFH*

miRNA ID	Functions	Targeted genes
gra-miR530	Role in salt stress regulation, defense to fungus, regulating blast disease resistance, yield, and growth period	*GrFH1*, *GrFH9*, *GrFH10*, *GrFH14*, *GrFH15*, *GrFH16*, *GrFH17*, *GrFH19*, *GrFH25*
gra-miR7494	Role in abiotic stress such as salt and drought and signal transduction	*GrFH4*, *GrFH5*, *GrFH6*, *GrFH9*, *GrFH11*, *GrFH12*, *GrFH15*, *GrFH17*, *GrFH22*, *GrFH25*
gra-miR7492	Regulating dis-proportionating enzyme 2 for cold response	*GrFH2*, *GrFH10*, *GrFH13*, *GrFH15*, *GrFH17*

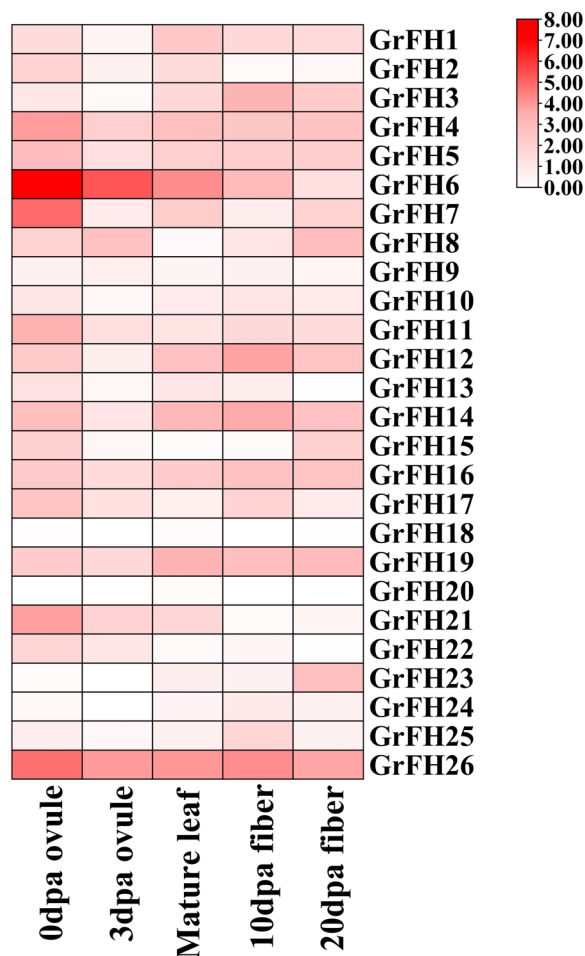Shing *et al. BMC Genomic Data*   (2024) 25:105

Page 12 of 20

Fig. 8 Tissue-specific expression profiles of *GrFH*. The bottom of the heatmap contains various tissues including 0dpa ovule (days post-anthesis), 3dpa ovule, mature leaf, 10dpa fiber, and 20dpa fiber. The intensity of the color from low to high expression (white to red color) was shown on the right side of the heatmap

Fig. 9 Expression pattern of *GrFH* in various abiotic stresses (cold and salt). The bottom of the heatmap contains control and abiotic stress (cold and salt). The intensity of the color from low to high expression (white to red color) was shown on the right side of the heatmap

noticed to express exceptionally than other genes in mature leaf. The comparison between vegetative and reproductive tissue showed that genes that are prominently expressed in vegetative tissue were shown to be expressed less frequently in reproductive tissue, and vice versa. In the formation of fiber, 10dpa fiber, and 20dpa fiber were analyzed to unveil the expression of *GrFH* genes. In the 10dpa fiber, about 92.3% *GrFH* genes were observed to express whereas about 96.15% genes were observed expressing in the 20dpa fiber. The expression patterns were quite similar between the two. However, *GrFH6, GrFH12, GrFH14*, and *GrFH26* were more dominantly observed in 10dpa fiber than in 20dpa fiber. Few *GrFH* genes such as *GrFH8, GrFH15*, and *GrFH23* in 20dpa were identified to showcase higher expression than in 10dpa fiber.
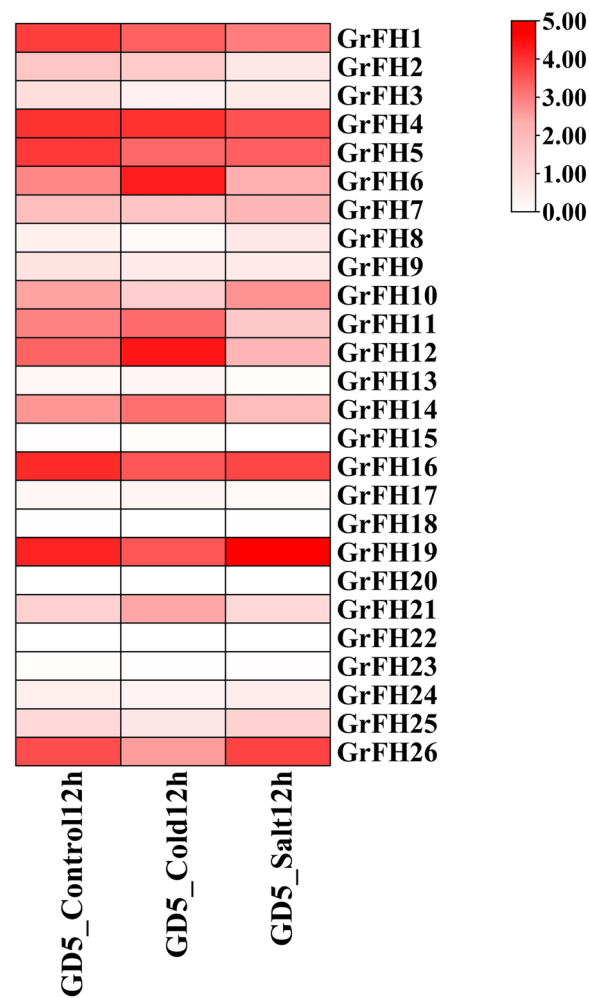
Expression profiles of *GrFH* under cold and salt stress

The differential expression patterns of *GrFH* in leaf tissues in response to several abiotic stresses demonstrated their capacity to confront abiotic stress such as cold and salt (Fig. 9). The expression rate of *GrFH* at 12-hour control, cold, and salt stress were similar (Supplementary Material File 10). However, certain *GrFH* genes were highly upregulated in either cold or salt stress. In GD5 12-hour cold stress, the upregulated genes were *GrFH4, GrFH6, GrFH11, GrFH12, GrFH14*, and *GrFH21*. Though some genes, such as *GrFH1, GrFH5, GrFH16, GrFH19*, and *GrFH26*, exhibited high expression but were subsequently down regulated. In GD5 12 h salt stress, *GrFH7, GrFH10, GrFH19*, and *GrFH26* showed up regulation. These genes showed elevated expression compared to the control, emphasizing resistance to salt stress.

Nevertheless, most of the *GrFH* sustained downregulation in salt stress.

Discussion

The formin protein's FH1 and FH2 domains significantly regulate the movements of the cytoskeleton. The integrity of actin filament in cytoskeleton dynamics is altered if mutations in one or more genes encoding formin proteins occur [27, 28]. The mutation influences not just actin filaments but also microtubule activity [29, 30]. In *Arabidopsis*, modification of microfilament structure during abiotic stress (salt and osmotic stress) emphasized the indispensable role the cytoskeleton performs [31]. However, the molecular mechanism surrounding cytoskeleton changes during stress remains a mysterious case.

The basic physiochemical properties of 26 GrFH proteins were investigated. Subsequently, all the GrFH proteins contained an instability index greater than 40. According to a study on *Caulobacter crescentus* metalloprotein, proteins with an instability score greater than 40 were considered unstable [32, 33]. Therefore, GrFH proteins were considered unstable. The higher the aliphatic index, the more the thermally stable and aliphatic side chains the proteins were regarded as [34, 35]. Therefore, most of the GrFH protein is considered thermally stable. All GrFH proteins exhibited a negative value for the grand average of hydropathicity (GRAVY), indicating their hydrophilic character [36, 37].

The comprehensive phylogenetic analysis helps to discover the molecular and evolutionary basis of lineage, interactions, and diversity among distinct species [38]. Moreover, phylogenetic analysis provides the patterns and evolutionary rates of different species [39]. The summary of the phylogenetic analysis of GrFH and other species revealed that the majority of GrFH clustered with MtFH, indicating a close connection between them. However, this was not shown by ZmFH or OsFH, with GrFH exhibiting their distinct genetic feature and divergence in evolution. The motif configuration in the *GrFH* diverged between some groups but was generally similar within the same group showcasing functional similarity. Besides, in *G. raimondii*, the *GrTCP* gene maintained a similar motif pattern among the subfamilies [40]. Similar distribution behavior and biological relevance were revealed by the existence of the FH2 conserved domain in all GrFH. The FH2 domain mediates actin filament elongation in cytoskeleton dynamics [41, 42]. In soybean (*Glycine max*), 34 proteins of GmFH contained formin domains suggesting their involvement in the management of abiotic stresses such as drought, heat, and salt through regulation of the cytoskeleton [43]. Besides, gene structure consists of exon-introns, which simplifies the

structural diversity within the gene family [44, 45]. Generally, the presence of a high introns number promotes post-transcriptional events like alternative splicing. The possibility of a higher and longer intron promotes high expression compared to the lower number of exons, which activate quickly [46–48]. Group B members, especially *GrFH9*, *GrFH10*, and *GrFH11*, possessed 16 introns, displayed alternative splicing among the seven different groups, whereas group C member *GrFH25* featured a lengthy stretch of introns. On the other hand, group G members comprised relatively lower exons, which provided the floodgates for activating the genes early. Nonetheless, *GrFH20* from group G contained only two exons, which activated more quickly than others.

In evolutionary biology and comparative genomics, the Ka/Ks ratio is essentially regarded [49]. Thus, it highlighted that except for the *GrFH20-GrFH26* pair, all the other *GrFH* gene pairs manifested purifying selection. On the other hand, the *GrFH20-GrFH26* pair, experienced positive selection, where beneficial mutations were gradually building up and signaling that the protein was adapting to new functions or stresses from the environment [50, 51]. Gene collinearity is a specific kind of synteny when gene groups are present through the chromosomes of respective genomes in usually identical patterns [52]. There were 11 collinear pairs in *GrFH* genes. The results of the collinear relationship suggested that various *GrFH* gene members in cotton experienced a close genetic connection within them. Meanwhile, no syntenic pairs were observed to describe similarities in the physical arrangement of genes across different genomes.

Chromosomal localization is required to understand the function of a gene as well as gene duplication, modification, and conservation throughout evolution [53]. The 26 *GrFH* genes were distributed at various locations across the 12 chromosomes. There were also 12 chromosomes in the cyclin family of rice, where 49 predicted rice cyclin genes were distributed, underlying their significance in cell cycling [54]. Gene duplication events, such as segmental, tandem, and transposition, are among the primary mechanisms that explain the expansion of gene families in a wide range of species [55]. The segmental duplication that arises from the duplication of different chromosomes was seen in all *GrFH* gene pairs. In cotton, segmental duplication and collinear pairs contributed to resisting the salinity stress [56].

The subcellular localization displays and represents the gene in a particular organelle to gain insight into the molecular mechanism as well as cellular compartmentalization such as the mitochondria, which act as a powerhouse, the cytoskeleton, which facilitates mobility, and the structural proteins [57–59]. In this study, *GrFH* were

Shing *et al. BMC Genomic Data*   (2024) 25:105

Page 14 of 20

mostly expressed in the chloroplast and nucleus. This showed that, in appropriate circumstances, a gene may express itself in those organelles and carry vital information. Moreover, most *mTERF* genes were localized in chloroplasts suggesting their essential role in photosynthesis machinery in maize [60].

The initiation of regulation of gene expression depends on the external or internal condition. Thus, the regulation of genes in environmental stress is generally promoted by promoters, enhancers, or suppressors [61, 62]. The analysis of the promoter regions of *GrFH*, and 4 distinct groups were created based on functional similarities of the 57 motifs. Among them, most of the motifs showed a response toward light. Besides, Cold Shock Protein (CSP) in cotton species showed various functionally similar CARE motifs such as light response and cold stress management [63]. Thus, *GrFH13* was seen interacting with stress-responsive motifs such as DRE core that was involved in dehydration, low-temp, and salt stresses, while *GrFH21* was observed more in the LTR motif in low-temperature responsiveness. Consequently, the MYB motifs were found in the *GaMATE* and *GrMATE* genes, unveiling drought resistance ability [64]. The functionally annotated GO was categorized into molecular function, biological process, and cellular component [65, 66]. Based on the findings, the majority of *GrFH* played the biological role regarding the cytoskeleton actin filament organization and regulation as well as the assembly of cellular components and proteins. The 28 biological pathways were observed in *Gossypium herbaceum* related to the drought resistance [67].

Transcriptional factors (TFs) perform a critical role in numerous biological processes, particularly in plants, such as regulating metabolism, growth, and progression, resistance against infections caused by microbes, and responses to both biotic and abiotic stress [68, 69]. This comprised 7 independent TF families such as ERF, GATA, MYB, LBD, TALE, and C2H2, along with bZIP. Ethylene-responsive factors (ERFs) serve as crucial regulators in the development, abiotic stress responses as well as defense against pathogen and insect attacks regulated by overexpression of the AP2/ERF in plant [70]. In cotton, ERF stimulated signals for the tolerance to salt and drought stresses [71]. A family of zinc finger proteins known as GATA (GA-binding Activator Protein Transcription Activator) vital in regulating the expression of genes that bind to the GATA [72, 73]. MYB (v-myb avian myeloblastosis viral oncogene homolog) TFs control numerous physiological and biochemical processes, including plant development, the fate of cell determination, and secondary metabolism [68]. The bZIP (basic leucine zipper) regulates a wide range of biological processes, including the proliferation of cells, cell, and tissue

differentiation, flowering and seed maturation, aging, as well as abiotic stress responses such as salinity and drought [68, 74, 75]. According to the results of the TF regulatory network analysis in *GrFH*, it disclosed that, ERF displayed the most substantial association with *GrFH* and bZIP and TALE TFs exhibited the lowest association with *GrFH* genes. Except for *GrFH3*, *GrFH20*, and *GrFH22*, all the members of *GrFH* genes showed interactions with the ERF family.

Recently, a study on miRNA in cotton revealed that it was associated with cellular signaling such as the hormone-signaling pathway, the calcium-signaling pathway, and the reactive oxygen species (ROS) signaling mechanism at the time of fiber initiation and elongation [76]. In this study, 96 unique miRNAs were identified. Among them, gra-miR530 might inhibit the *GrFH* mRNA production by cleaving or halting translation. Moreover, it was found to be involved in the regulation of blast disease resistance, yield, and growth period by blocking the overexpression of certain genes in rice and regulator gene expression under salt stress in flax (*Linum usitatissimum*) [77–79]. Therefore, it might have performed similar functions in the *GrFH*. Another miRNA, gra-miR7494 might have cleaved the mRNA and regulated the expression of *GrFH* in abiotic stress such as salt and drought. In cotton species, miR7492 was involved in the control of disproportionating enzyme 2, which was critical in the conversion of maltose to glucose [80]. Thus, it was predicted that gra-miR7492 might have related roles in *GrFH*.

Protein-protein interactions shed light on how the direction of diversification of existing species and the regulation of cellular functions, including the transmission of signals, regulation of cell cycles, and metabolic activity, was made conceivable by the emergence of biological network connections [81, 82]. The GrFH proteins showed homology with *Arabidopsis* and interactions with the FH family (FH1, FH4, FH5, FH6, FH12, and FH13), PRF family (PRF1, PRF2, PRF3, PRF4, and PRF5), ARPC5A, and T14C9.40.

Gene expression reveals the activity of genes under specific conditions and identifies regulatory mechanisms that control numerous stages of development, offering essential information on gene function [83–85]. The *GrFH* genes through transcriptomic analysis provided insights about their potential expression pattern in certain tissues (ovule, leaf, and fiber). *GrFH8* and *GrFH9* expressed at a higher rate compared to others in the 3dpa ovule. Cottons are widely cultivated economical crops and are considered cash crops because of their ability to produce fiber [86]. Cotton fibers, units of trichomes that originated from ovular epidermal cells, consisted of four overlapping stages initiation (0–8 dpa), elongation (3–17 dpa), secondary cell wall synthesis (17–40 dpa), and

Shing *et al. BMC Genomic Data*   (2024) 25:105

Page 15 of 20

maturation (40–50 dpa) to become mature fiber (lint) [87–90]. Moreover, the cytoskeleton, actin filament, and microtubule, were involved in the formation of fiber [91]. Therefore, *GrFH3*, *GrFH12*, *GrFH14*, and *GrFH26* in 10dpa fiber and *GrFH8*, *GrFH19*, *GrFH23*, and *GrFH26* in 20dpa fiber were observed to express highly. *GrFH26*, however, maintained its expression at a high rate in both 10dpa and 20dpa fiber unveiling its involvement in mature lint formation. The differential gene expression pattern of *GrFH* in cold and salt stress underscored the significance they performed in combating these stresses. In cold stress, six *GrFH* genes such as *GrFH4*, *GrFH6*, *GrFH11*, *GrFH12*, *GrFH14*, and *GrFH21* showed up-regulation, suggesting their potential association with cold stress management. Moreover, a study on *G. raimondii* showed that *GrPRR5.1*, *GrPRR5.2*, and *GrPRR7.2* were the up-regulator genes in cold stress. Whereas, *GrHP4.2*, *GrHK5.2*, and *GrRR11* were the up-regulator genes in salt stress [92]. The upregulation of genes such as *GrFH7*, *GrFH10*, *GrFH19*, and *GrFH26* in salt stress provided evidence of their involvement in managing salt stress. Two major organs, ovule, and fiber, are the key enforcers to produce mature lint, as well as being most sensitive to abiotic stresses [93]. Therefore, overexpression and upregulation of *GrFH4*, *GrFH6*, *GrFH12*, *GrFH14*, and *GrFH26* in these organs in cold stress suggested their role in cold resistance in cotton. Moreover, *GrFH19* and *GrFH26* conferred resistance to salt stress in these organs. All in all, this study focused on improving breeding strategies to generate cold and salt-tolerant cotton varieties to produce quality fiber.

Methods
Database search and mining of Formin gene in *G. raimondii* genome
A. thaliana FORMIN DNA binding domains were used to search and retrieve the FORMIN protein sequence in the cotton (*G. raimondii)* genome. Phytozome v13 (https://phytozome-next.jgi.doe.gov/) was employed to extract gene sequence, reference genome, and protein sequence by using BLASTp (Protein-Basic Local Alignment Search Tool) with an expected(E) threshold value of −1, a comparison matrix (BLOSUM62), and keeping other parameters as default [94]. Further, to find out the presence of the FH2 conserved domain, SMART (Simple Modular Architecture Research Tool) (http://smart.embl-heidelberg.de/) [95], NCBI CDD (Conserved Domain Database) (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) [96] and PfamScan (https://www.ebi.ac.uk/Tools/pfa/pfamscan/) were used with default parameters [97]. The candidate genes were selected based on the presence of the FH2 in the predicted protein and

renamed according to their sequential physical chromosome positions.

Determination of physiochemical properties of GrFH
The physicochemical properties of FORMIN proteins included with the determination of the number of A.A. residues, molecular weight (kDa), pI, instability index, aliphatic index, and GRAVY in the ProtParam online program (http://web.expasy.org/protparam/) [98].

Phylogenetic analysis between GrFH, MtFH, ZmFH, OsFH, and AtFH
The formin protein sequences of the *Medicago truncatula*, *Zea mays*, *Oryza sativa*, *Arabidopsis thaliana*, and gene *Gossypium raimondii* were extracted from Phytozome v13 (Supplementary Material File 11), and a phylogenetic tree between them was constructed through the MEGA11 software [99] with the ClustalW program as a sequence alignment tool [100]. The Maximum Likelihood (ML) model was set to allow more robust parameter estimates [101]. Moreover, other parameters were set as default except for a 1000 bootstrap value to support branch values and Pearson correction. The phylogenetic tree was further uploaded to iTOL v6.7.4 (https://itol.embl.de/) for proper illustration and visualization [102].

Conserved motif, domain, and gene structure analysis of *GrFH*
The conserved motif, domain and gene structure of *GrFH* were constructed in TBTOOLS v.2.010 gene structure view (advanced) [103]. For the analysis of structural motifs in *GrFH*, Multiple EM for Motif Elicitation (MEME)   (https://meme-suite.org/meme/tools/meme) was applied, setting several motifs parameters as much as 20 and others as default [104].

Gene duplication analysis and non-synonymous (Ka) and synonymous (Ks) substitution ratio calculation of *GrFH*
To determine and calculate the molecular evolution rate, Ka (non-synonymous) and Ks (synonymous) substitutions were performed. For the calculation of Ka/Ks ratios of *GrFH* genes, CDS sequences (Supplementary Material File 12) of duplicated genes were used in the Ka/Ks calculation tool (https://services.cbu.uib.no/tools/kaks). The obtained result was further assembled and enumerated the duplication and time of divergence (measured in million years ago, MYA) by applying the $T = Ks/2\lambda$ formula, where λ was equal to 6.5×10^{-9} [105]. The data was visually illustrated by TB tools v.2.010.

Shing *et al. BMC Genomic Data*   (2024) 25:105

Page 16 of 20

Evolutionary collinearity and synteny relationship analysis of *GrFH*

Collinearity analysis was used to explore the evolutionary relationships among homologous genes of *GrFH*. However, to perform the synteny analysis, evolutionary relationships between different species *Z. mays*, *O. sativa*, *A. thaliana*, and candidate genes *G. raimondii* were used to find out gene duplication between them. For the visual representation, the collinear pairs among homologous genes and syntenic pairs in different species were illustrated in TB tools v.2.010.

Analysis of chromosomal localization and duplication of *GrFH*

By using the Phytozome v13 and TB tools v.2.010, the information about the chromosomal length, start points, and end points of 26 *GrFH* was retrieved and assembled. The retrieved data was mapped and viewed by MapGene2Chrom web v2 (MG2C) web server (http://mg2c.iask.in/mg2c_v2.0/) [106]. Whereas for the chromosomal duplication, duplicated gene pairs were analyzed in the distributed chromosomes.

Prediction of the subcellular localization of *GrFH*

The prediction of the subcellular localization of *GrFH* was analyzed by using the Wolf PSORT online tool (https://wolfpsort.hgc.jp/) [107]. The collected data was displayed by RStudio version 2023.06.1 generating a comprehensive overview of the subcellular compartmentalization of different organelles present in the respective *GrFH* [108].

Cis-acting regulatory elements (CAREs) analysis in the promoters of *GrFH*

About 2000 bp from 5′ UTR of 26 *GrFH* were extracted and predicted in the plant CARE database (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/) [109]. The RStudio version 2023.06.1 was used to generate a graphical representation. The grouping of CAREs was performed according to their functional similarities.

Gene ontology analysis of *GrFH*

The GO enrichment data was downloaded from (PlantRegMap; http://plantregmap.gao-lab.org/binding_site_prediction.php) with an estimated threshold *p*-value of 0.01 and other default parameters in an attempt to classify the functions of the predicted *GrFH* with genomic sequences (Supplementary Material File 13) [110]. The ChiPlot (https://www.chiplot.online/) provided in-depth details about visualization and illustration of the extracted data [111].

Transcription factors analysis of *GrFH*

For the identification of TFs related to the *GrFH*, (PlantTFDB; http://plantregmap.gao-lab.org/binding_site_prediction.php) was utilized with the default parameter and threshold *p*-value of $1 \times 10-4$. The collected data was then further processed and visualized in the RStudio version 2023.06.1.

Regulatory network between TFs and *GrFH*

Cytoscape version 3.10.0, a software tool used for the analysis and graphical representation of complex networks of molecular biology interactions, was used to visualize the interactions between various transcription factors as well as interactions with *GrFH* [112].

Prediction of putative micro-RNAs and networks targeting *GrFH*

(https://mirbase.org/) [113]. The CDS sequences of *GrFH* were uploaded to the online psRNATarget Server18 keeping the rest of the parameters set to default (https://www.zhaolab.org/psRNATarget/analysis?function=2) [114]. Moreover, Cytoscape software version 3.10.0 was employed to generate and visualize the interaction network between the predicted miRNAs and *GrFH* targeting genes.

Protein-protein interactions of GrFH

The protein-protein interaction network of GrFH proteins was predicted by employing the web tool STRING version 12 with GrFH protein sequneces (Supplementary Material File 14) (https://string-db.org/) according to the *A. thaliana* homologous proteins [115]. The parameters of the STRING online tool were kept as network type-full STRING network; the meaning of network edges evidence; active interaction source as text mining, experiments, databases, co-expression, neighborhood, gene fusion, co-occurrence; minimum required interaction score was specified to medium confidence parameter (0.4); max number of interactions display 1st shell was defined to no more than 10 and 2nd shell was left blank and enabling network displayed options as 3D bubble design.

GrFH tissue-specific and differential expression patterns in various abiotic stresses

The RNA sequencing data of five tissue samples (0dpa ovule, 3dpa ovule, mature leaf from accession ID SRP009820 and 10dpa fiber and 20dpa fiber from accession ID SRP001603) [87] and abiotic stress (cold and salt from BioProject accession ID PRJNA554555) [92]

Shing *et al. BMC Genomic Data*   (2024) 25:105

Page 17 of 20

were extracted from the NCBI Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra/). Further, Trimmomatic package version 0.32 was used to quality control and trim the transcriptomic data [116]. Further, the RNA seq was mapped to the reference genome *G. raimondii* obtained from Phytozome v13 with Bowtie2 package version 2.5.4 [117]. The sequence alignment map (SAM) files were converted to binary alignment map (BAM) files and sorted and arranged with Samtools packages version 1.20 [118]. Fragments per kilobase million (FPKM) values were computed using the RSEM package version 1.1.17 [119]. The FPKM values were converted to log2 and visualized in TB tools v.2.010 illustrating, the expression profiling of *GrFH*.

Conclusions

The comprehensive bioinformatics analysis of 26 *GrFH* genes, comprised of the FH2 domain, will shed light on their potential role in the management of various abiotic stresses such as salinity and cold. The seven distinct subfamilies of GrFH in the evolutionary phylogenetic tree provided exclusive insights about their ancient descendant with MtFH. According to the subcellular localization, the bulk amount of *GrFH* genes was found in the chloroplast and nucleus. Moreover, 12 chromosomes were found to locate the distinct intron-exon gene structures. Furthermore, all the *GrFH* gene pairs experienced purifying selection except the *GrFH20-GrFH26* pair. Besides, stress-related motif was found in *GrFH* CARE regions. Most *GrFH* were associated with biological functions focusing on the actin filament organization, regulation of actin filament polymerization and de-polymerization, and cytoskeleton assembly. Major TFs, including ERF, GATA, MYB, and LBD, interacted and formed complex networks with *GrFH*, showing their binding and regulation attributes. The non-coding miRNA regulated the gene expression of *GrFH* during salinity and cold abiotic stress and in the formation of mature fiber from the ovule. Besides, certain *GrFH* were observed to express at a higher rate, showcasing their promising attributes in fiber formation from the ovular epidermis. Further, the upregulation of *GrFH4*, *GrFH6*, *GrFH12*, *GrFH14*, and *GrFH26* under cold stress and *GrFH19* and *GrFH26* under salt stress provided valuable insights in confronting these challenges. Overall, the findings from this study would provide deeper knowledge and shed light on the biological significance of the FORMIN gene family in *G. raimondii* cotton species in managing various stresses in wet lab facilities.

Abbreviations

A.A.   Amino Acid

kDa   Kilo Dalton
pl   Isoelectric point
GRAVY   Grand average of hydropathicity
bp   base pair
5′ UTR 5′ untranslated region

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12863-024-01285-z.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

Supplementary Material 4.

Supplementary Material 5.

Supplementary Material 6.

Supplementary Material 7.

Supplementary Material 8.

Supplementary Material 9.

Supplementary Material 10.

Supplementary Material 11.

Supplementary Material 12.

Supplementary Material 13.

Supplementary Material 14.

Supplementary Material 15. Supplementary Fig. S1 The estimation of gene duplication time and Ka/Ks ratio in *GrFH*. The ratio of nonsynonymous (Ka) to synonymus (Ks) changes is represented by Ka/Ks. The time of divergence (measured in million years ago, MYA) is also represented. The different color bar represents the data range.

Supplementary Material 16. Supplementary Fig. S2 The chromosomal locations and duplications of *GrFH*. The number of distinct chromosomes is at the top of each chromosome bar. The chromosome-scale is in millions of bases (Mb), indicating the length of each chromosome on the left. Chromosome is colored magenta while blue lines indicate segmental duplications.

Supplementary Material 17. Supplementary Fig. S3 *GrFH* genes function analysis through gene ontology. On the right side of the circos plot classification of the *GrFH* gene functions are shown. The number of genes involved under a certain GO ID, expected value, and rich factor are shown in a distinctive color. The scaling of the −log10 p value is shown in two distinctive colors (blue and yellow).

Supplementary Material 18. Supplementary Fig. S4 The regulatory network between TFs and *GrFH* genes. The *GrFH* is shown in red rectangular. Whereas the TFs are shown in different colors and shapes. The 7 TF families ERF, MYB, C2H2, GATA, bZIP, LBD, and TALE are represented by a yellow diamond, dark red hexagon, magenta triangle, lime green ellipse, V shape sea green, violet parallelogram and round rectangular respectively.

Supplementary Material 19. Supplementary Fig. S5 Prediction of potential micro-RNAs targeting *GrFH*. A. Network illustration of predicted miRNA targets *GrFH* genes. The red rectangle represents *GrFH* genes while micro-RNA is labeled as sky blue ellipse. B. The schematic diagram indicates the *GrFH* genes targeted by miRNAs. The sky-blue round rectangular is shown as exons of the respective gene, green round rectangular represents UTR, straight black line represents intron and red color small round rectangular represents microRNA (miRNA).

Supplementary Material 20. Supplementary Fig. S6 Protein-protein interaction of GrFH based on known *Arabidopsis* proteins. The proteins were displayed at network nodes with the proteins in nodes, and the line colors indicate different data sources.

Shing *et al. BMC Genomic Data*   (2024) 25:105

Page 18 of 20

Data availability
Data is provided within the manuscript or supplementary information files.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Laboratory of Functional Genomics and Proteomics, Department of Genetic Engineering and Biotechnology, Faculty of Biological Science and Technology, Jashore University of Science and Technology, Jashore 7408, Bangladesh. [2]Department of Genetic Engineering and Biotechnology, Faculty of Biological Sciences, University of Rajshahi, Rajshahi 6205, Bangladesh.

References
1. Michelot A, Guerin C, Huang S, Ingouff M, Richard S, Rodiuc N, Staiger CJ, Blanchoin L. The formin homology 1 domain modulates the actin nucleation and bundling activity of Arabidopsis FORMIN1. Plant Cell. 2005;17(8):2296–313.
2. Li B, Du Z, Jiang N, He S, Shi Y, Xiao K, Xu L, Wang K, Wang X, Wu LJPMBR. Genome-wide identification and expression profiling of the FORMIN Gene Family implies their potential functions in abiotic stress tolerance in Rice (Oryza sativa). 2023:1–14.
3. Higgs HN, Peterson KJ. Phylogenetic analysis of the formin homology 2 domain. Mol Biol Cell. 2005;16(1):1–3.
4. Paul A, Pollard T. The role of the FH1 domain and profilin in formin-mediated actin-filament elongation and nucleation. Curr Biol. 2008;18(1):9–19.
5. Courtemanche N, Pollard TD. Determinants of Formin Homology 1 (FH1) domain function in actin filament elongation by formins. J Biolog Chem. 2012;287(10):7812–20.
6. Wang C, Zhang LJ, Huang RD. Cytoskeleton and plant salt stress tolerance. Plant Signal Behav. 2011;6(1):29–31.
7. Qin L, Liu L, Tu J, Yang G, Wang S, Quilichini TD, Gao P, Wang H, Peng G, Blancaflor EB. The ARP2/3 complex, acting cooperatively with Class I formins, modulates penetration resistance in Arabidopsis against powdery mildew invasion. Plant Cell. 2021;33(9):3151–75.
8. Blanchoin L, Staiger CJ. Plant formins: diverse isoforms and unique molecular mechanism. Biochim Biophys Acta. 2010;1803(2):201–6.
9. Kumar S, Jeevaraj T, Yunus MH, Chakraborty S, Chakraborty N. The plant cytoskeleton takes center stage in abiotic stress responses and resilience. Plant Cell Environ. 2023;46(1):5–22.
10. Soda N, Singla-Pareek SL, Pareek A. Abiotic stress response in plants: Role of cytoskeleton. 2016:107-34.
11. Diao M, Ren S, Wang Q, Qian L, Shen J, Liu Y, Huang SJE. Arabidopsis formin 2 regulates cell-to-cell trafficking by capping and stabilizing actin filaments at plasmodesmata. 2018, 7:e36316.
12. Wang L, Qiu T, Yue J, Guo N, He Y, Han X, Wang Q, Jia P, Wang H, Li M, et al. Arabidopsis ADF1 is regulated by MYB73 and is involved in response to salt stress affecting actin filament organization. PlantCell Physiol. 2021;62(9):1387–95.
13. Qian D, Zhang Z, He J, Zhang P, Ou X, Li T, Niu L, Nan Q, Niu Y, He W. Arabidopsis ADF5 promotes stomatal closure by regulating actin cytoskeleton remodeling in response to ABA and drought stress. J Exp Bot. 2019;70(2):435–46.
14. Chun HJ, Baek D, Jin BJ, Cho HM, Park MS, Lee SH, Lim LH, Cha YJ, Bae DW, Kim ST, Yun DJ. Microtubule dynamics plays a vital role in plant adaptation and tolerance to salt stress. Int J Mol Sci. 2021;22(11):5957.
15. Wang C, Li J, Yuan M. Salt tolerance requires cortical microtubule reorganization in Arabidopsis. Plant Cell Physiol. 2007;48(11):1534–47.
16. Livanos P, Galatis B, Quader H, Apostolakos P. Disturbance of reactive oxygen species homeostasis induces atypical tubulin polymer formation and affects mitosis in root-tip cells of Triticum turgidum and Arabidopsis thaliana. Cytoskeleton. 2012;69(1):1–21.
17. Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S. The draft genome of a diploid cotton Gossypium raimondii. Nat Genet. 2012;44(10):1098–103.
18. Should WCG. The Post-Genomic Era for Cotton.
19. Wang L, Ai N, Zhang Z, Zhou C, Feng G, Cai S, Wang N, Feng L, Chen Y, Xu MJJIA. Development of Gossypium hirsutum-Gossypium raimondii introgression lines and its usages in QTL mapping of agricultural traits. 2024.
20. Joint FJCCCCoFAS, Monograph: Food and Agriculture Organization of the United Nations. 2011, 11:1817–7077.
21. Bunkute E, Cummins C, Crofts FJ, Bunce G, Nabney IT, Flower DR. PIP-DB: the protein isoelectric point database. Bioinformatics. 2015;31(2):295–6.
22. Hilal SH, Karickhoff SW, Carreira LA. Estimation of microscopic, zwitterionic ionization constants, isoelectric point and molecular speciation of organic compounds. Talanta. 1999;50(4):827–40.
23. Li A, Mao L. Evolution of plant microRNA gene families. Cell Res. 2007;17(3):212–8.
24. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009;136(2):215–33.
25. Jackson RJ, Standart NJSS. How do microRNAs regulate gene expression? 2007, 2007(367):re1-re1.
26. Jonas S, Izaurralde E. Towards a molecular understanding of microRNA-mediated gene silencing. Nature Rev Genet. 2015;16(7):421–33.
27. Castrillon DH, Wasserman SAJD. Diaphanous is required for cytokinesis in Drosophila and shares domains of similarity with the products of the limb deformity gene. 1994, 120(12):3367–77.
28. Evangelista M, Blundell K, Longtine MS, Chow CJ, Adames N, Pringle JR, Peter M, Boone CJS. Bni1p, a yeast formin linking cdc42p and the actin cytoskeleton during polarized morphogenesis. 1997, 276(5309):118–22.
29. Giansanti MG, Bonaccorsi S, Williams B, Williams EV, Santolamazza C, Goldberg ML. Gatti MJG, development: Cooperative interactions between the central spindle and the contractile ring during Drosophila cytokinesis. 1998, 12(3):396.
30. Lee L, Klee SK, Evangelista M, Boone C, Pellman DJTJ. Control of mitotic spindle position by the Saccharomyces cerevisiae formin Bni1p. 1999, 144(5):947–61.
31. Wang C, Zhang L, Yuan M, Ge Y, Liu Y, Fan J, Ruan Y, Cui Z, Tong S, Zhang SJPB. The microfilament cytoskeleton plays a vital role in salt and osmotic stress tolerance in Arabidopsis. 2010, 12(1):70–8.
32. Gamage DG, Gunaratne A, Periyannan GR, Russell TGJP. Letters p: applicability of instability index for in vitro protein stability prediction. 2019, 26(5):339–47.
33. Guruprasad K, Reddy BB, Pandit MWJPE, Design. Selection: correlation between stability of a protein and its dipeptide composition: a novel

Shing *et al. BMC Genomic Data*　　(2024) 25:105

Page 19 of 20

approach for predicting in vivo stability of a protein from its primary sequence. 1990, 4(2):155–61.

34. Panda S, Chandra G. Physicochemical characterization and functional analysis of some snake venom toxin proteins and related non-toxin proteins of other chordates. Bioinformation. 2012;8(18):891.

35. Ikai A. Thermostability and aliphatic index of globular proteins. J Biochem. 1980;88(6):1895–8.

36. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982;157(1):105–32.

37. Wang H, Zhong H, Gao C, Zang J, Yang D. The distinct properties of the consecutive disordered regions inside or outside protein domains and their functional significance. Int J Mol Sci. 2021;22(19):10677.

38. Soltis DE, Soltis PS. The role of phylogenetics in comparative genetics. Plant Physiol. 2003;132(4):1790–800.

39. Soltis ED, Soltis PS. Contributions of plant molecular systematics to studies of molecular evolution. Plant Mol Biol. 2000;42:45–75.

40. Ma J, Wang Q, Sun R, Xie F, Jones DC, Zhang B. Genome-wide identification and expression analysis of TCP transcription factors in Gossypium Raimondii. Sci Rep. 2014;4:6645.

41. Lu J, Meng W, Poy F, Maiti S, Goode BL, Eck MJ. Structure of the FH2 domain of Daam1: implications for formin regulation of actin assembly. J Mol Biol. 2007;369(5):1258–69.

42. Otomo T, Tomchick DR, Otomo C, Panchal SC, Machius M, Rosen MK. Structural basis of actin filament nucleation and processive capping by a formin homology 2 domain. Nature. 2005;433(7025):488–94.

43. Zhang Z, Zhang Z, Shan M, Amjad Z, Xue J, Zhang Z, Wang J, Guo Y. Genome-wide studies of FH Family members in soybean (Glycine max) and their responses under Abiotic stresses. Plants (Basel Switzerland) 2024, 13(2).

44. Hardison RC. A brief history of hemoglobins: plant, animal, protist, and bacteria. Proc Natl Acad Sci. 1996;93(12):5675–9.

45. Tian F, Wang T, Xie Y, Zhang J, Hu J. Genome-wide identification, classification, and expression analysis of 14-3-3 gene family in Populus. PloS one. 2015;10(4):e0123225.

46. Koonin EV, Csuros M, Rogozin IB. Whence genes in pieces: reconstruction of the exon–intron gene structures of the last eukaryotic common ancestor and other ancestral eukaryotes. Wiley Interdiscip Rev RNA. 2013;4(1):93–105.

47. Koralewski TE, Krutovsky KV. Evolution of exon-intron structure and alternative splicing. PloS one. 2011;6(3):e18055.

48. Heidari P, Puresmaeli F, Mora-Poblete F. Genome-wide identification and molecular evolution of the magnesium transporter (MGT) gene family in Citrullus lanatus and Cucumis sativus. Agronomy. 2022;12(10):2253.

49. Nekrutenko A, Makova KD, Li WH. The KA/KS ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. Genome Res. 2002;12(1):198–202.

50. Hurst LD. The Ka/Ks ratio: diagnosing the form of sequence evolution. TRENDS Genet. 2002;18(9):486–7.

51. Yang Z. Computational molecular evolution. OUP Oxford; 2006.

52. Paterson A, Wang X, Tang H, Lee T. Synteny and genomic rearrangements: Springer; 2012.

53. Zheng J, Payne JL, Wagner A. Cryptic genetic variation accelerates evolution by opening access to diverse adaptive peaks. Science. 2019;365(6451):347–53.

54. La H, Li J, Ji Z, Cheng Y, Li X, Jiang S, Venkatesh PN, Ramachandran S. Genome-wide analysis of cyclin family in rice (Oryza Sativa L). Mol Genet Genomics: MGG. 2006;275(4):374–86.

55. Huang H, Song J, Feng Y, Zheng L, Chen Y, Luo K. Genome-wide identification and expression analysis of the SHI-related sequence family in Cassava. Genes. 2023;14(4):870.

56. Rehman A, Atif RM, Azhar MT, Peng Z, Li H, Qin G, Jia Y, Pan Z, He S, Qayyum A, et al. Genome wide identification, classification and functional characterization of heat shock transcription factors in cultivated and ancestral cottons (Gossypium spp). Int J Biol Macromol. 2021;182:1507–27.

57. Glory E, Murphy RF. Automated subcellular location determination and high-throughput microscopy. Develop Cell. 2007;12(1):7–16.

58. Ehrlich JS, Hansen MD, Nelson WJ. Spatio-temporal regulation of Rac1 localization and lamellipodia dynamics during epithelial cell-cell adhesion. Develop Cell. 2002;3(2):259–70.

59. Wu JQ, Kuhn JR, Kovar DR, Pollard TD. Spatial and temporal pathway for assembly and constriction of the contractile ring in fission yeast cytokinesis. Develop Cell. 2003;5(5):723–34.

60. Zhao Y, Cai M, Zhang X, Li Y, Zhang J, Zhao H, Kong F, Zheng Y, Qiu F. Genome-wide identification, evolution and expression analysis of mTERF gene family in maize. PLoS ONE. 2014;9(4):e94126.

61. Britten RJ, Davidson EH. Gene Regulation for Higher Cells: A Theory: New facts regarding the organization of the genome provide clues to the nature of gene regulation. Sci. 1969;165(3891):349–57.

62. Wittkopp PJ, Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Gen. 2012;13(1):59–69.

63. Yang Y, Zhou T, Xu J, Wang Y, Pu Y, Qu Y, Sun G. Genome-Wide Identification and Expression Analysis Unveil the Involvement of the Cold Shock Protein (CSP) Gene Family in Cotton Hypothermia Stress. Plants. 2024;13(5):643.

64. Lu P, Magwanga RO, Guo X, Kirungu JN, Lu H, Cai X, Zhou Z, Wei Y, Wang X, Zhang Z et al. Genome-Wide Analysis of Multidrug and Toxic Compound Extrusion (MATE) Family in Gossypium raimondii and Gossypium arboreum and Its Expression Analysis Under Salt, Cadmium, and Drought Stress. *G3 (Bethesda, Md)* 2018, 8(7):2483–2500.

65. research GOCJG: Creating the gene ontology resource: design and implementation. 2001, 11(8):1425–1433.

66. Masseroli M, Tagliasacchi M, Chicco D. Semantically improved genome-wide prediction of Gene Ontology annotations. In: 2011 *11th International Conference on Intelligent Systems Design and Applications: 2011*: IEEE; 2011: 1080–1085.

67. Ranjan A, Sawant S. Genome-wide transcriptomic comparison of cotton (Gossypium herbaceum) leaf and root under drought stress. 3 Biotech. 2015;5(4):585–96.

68. Khan SA, Li MZ, Wang SM, Yin HJ. Revisiting the role of plant transcription factors in the battle against abiotic stress. Int J Mol Sci. 2018;19(6):1634.

69. Sasaki K. Utilization of transcription factors for controlling floral morphogenesis in horticultural plants. Breed Sci. 2018;68(1):88–98.

70. Xu ZS, Chen M, Li LC, Ma YZ. Functions and application of the AP2/ERF transcription factor family in crop improvement F. J Integrative Plant Biol. 2011;53(7):570–85.

71. Zafar MM, Rehman A, Razzaq A, Parvaiz A, Mustafa G, Sharif F, Mo H, Youlu Y, Shakeel A, Ren M. Genome-wide characterization and expression analysis of Erf gene family in cotton. BMC Plant Biol. 2022;22(1):134.

72. Behringer C, Schwechheimer C. B-GATA transcription factors–insights into their structure, regulation, and role in plant development. Front Plant Sci. 2015;6:90.

73. Reyes JC, Muro-Pastor MI, Florencio FJ. The GATA family of transcription factors in Arabidopsis and rice. Plant Physiol. 2004;134(4):1718–32.

74. Liu J, Chen N, Chen F, Cai B, Dal Santo S, Tornielli GB, Pezzotti M, Cheng Z-MJB. Genome-wide analysis and expression profile of the bZIP transcription factor gene family in grapevine (Vitis vinifera). 2014, 15:1–18.

75. Jakoby M, Weisshaar B, Dröge-Laser W, Vicente-Carbajosa J, Tiedemann J, Kroj T, Parcy F. bZIP transcription factors in Arabidopsis. Trends Plant Sci. 2002;7(3):106–11.

76. Pizarro-Cerdá J, Chorev DS, Geiger B, Cossart P. The diverse family of Arp2/3 complexes. Trends Cell Biol. 2017;27(2):93–100.

77. Li Y, Wang LF, Bhutto SH, He XR, Yang XM, Zhou XH, Lin XY, Rajput AA, Li GB, Zhao JH, Zhou SX. Blocking miR530 improves rice resistance, yield, and maturity. Front Plant Sci. 2021;12:729560.

78. Hu G, Wang B, Jia P, Wu P, Lu C, Xu Y, Shi L, Zhang F, Zhong N, Chen A. The cotton miR530-SAP6 module activated by systemic acquired resistance mediates plant defense against Verticillium dahliae. Plant Sci. 2023;330:111647.

79. Yu Y, Wu G, Yuan H, Cheng L, Zhao D, Huang W, Zhang S, Zhang L, Chen H, Zhang J, et al. Identification and characterization of miRNAs and targets in flax (Linum usitatissimum) under saline, alkaline, and saline-alkaline stresses. BMC Plant Biol. 2016;16(1):124.

80. Ding Y, Ma Y, Liu N, Xu J, Hu Q, Li Y, Wu Y, Xie S, Zhu L, Min L, Zhang X. micro RNA s involved in auxin signalling modulate male sterility under high-temperature stress in cotton (Gossypium hirsutum). Plant J. 2017;91(6):977–94.

81. Makino T, Gojobori TJG. Evolution p: evolution of protein-protein interaction network. 2007, 3:13–29.

Shing *et al. BMC Genomic Data*   (2024) 25:105

Page 20 of 20

82. Xia J, Benner MJ, Hancock RE. NetworkAnalyst-integrative approaches for protein–protein interaction network analysis and visual exploration. Nucleic Acids Res. 2014;42(W1):W167-74.

83. Nadel BB, Lopez D, Montoya DJ, Ma F, Waddel H, Khan MM, Mangul S, Pellegrini M. The Gene Expression Deconvolution Interactive Tool (GEDIT): accurate cell type quantification from gene expression data. Giga Science. 2021;10(2):giab002.

84. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Studying gene expression and function. In: *Molecular Biology of the Cell 4th edition*. Garland Science; 2002.

85. Zhang Q, Li T, Zhang L, Dong W, Wang A: Expression analysis of NAC genes during the growth and ripening of apples. 2018.

86. Stiff MR, Haigler CHJF. Foundation ficTTC: Recent advances in cotton fiber development. 2012:163–192.

87. He DH, Lei ZP, Tang BS, Xing HY, Zhao JX, Jing YL. Identification and analysis of the TIFY gene family in Gossypium raimondii. Genet Mol Res. 2015;14(3):10119–38.

88. Li P-t, Wang M, Lu Q-w, Ge Q, Rashid MHO, Liu A-y, Gong J-w. Shang H-h, Gong W-k, Li J-wJBg: comparative transcriptome analysis of cotton fiber development of Upland cotton (Gossypium hirsutum) and chromosome segment substitution lines from G. hirsutum✕ G. barbadense. 2017, 18:1–17.

89. Lee JJ, Woodward AW, Chen ZJ. Gene expression changes and early events in cotton fibre development. Ann Botany. 2007;100(7):1391–401.

90. Taliercio EW, Boykin D. Analysis of gene expression in cotton fiber initials. BMC Plant Biol. 2007;7:1–3.

91. Seagull RW. Cytoskeletal involvement in cotton fiber growth and development. Micron. 1993;24(6):643–60.

92. Rasool A, Azeem F, Ur-Rahman M, Rizwan M, Hussnain Siddique M, Bay DH, Binothman N, Al Kashgry NAT, Qari SH. Omics-assisted characterization of two-component system genes from Gossypium Raimondii in response to salinity and molecular interaction with abscisic acid. Front Plant Sci. 2023;14:1138048.

93. Jan M, Liu Z, Guo C, Sun X. Molecular regulation of cotton fiber development: a review. Int J Mol Sci. 2022;23(9):5004.

94. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40(D1):D1178-86.

95. Letunic I, Khedkar S, Bork P. SMART: recent updates, new developments and status in 2020. Nucleic Acids Res. 2021;49(D1):D458-60.

96. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. 2020;48(D1):D265-8.

97. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S. Sonnhammer ELJNar: the pfam protein families database. 2004, 32(suppl_1):D138–41.

98. Gasteiger E, Hoogland C, Gattiker A, Duvaud Se, Wilkins MR, Appel RD, Bairoch A. Protein identification and analysis tools on the ExPASy server. Springer; 2005.

99. Tamura K, Stecher G, Kumar S. MEGA11: molecular evolutionary genetics analysis version 11. Mol Biol Evol. 2021;38(7):3022–7.

100. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22(22):4673–80.

101. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586–91.

102. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021;49(W1):W293-6.

103. Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, Xia R. TBtools: an integrative toolkit developed for interactive analyses of big biological data. Mol Plant. 2020;13(8):1194–202.

104. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Res. 2015;43(W1):W39-49.

105. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000;290(5494):1151–5.

106. Chao J, Li Z, Sun Y, Aluko OO, Wu X, Wang Q, Liu G. MG2C: A user-friendly online tool for drawing genetic maps. Mol Hortic. 2021;1:1–4.

107. Horton P, Park K-J, Obayashi T, Nakai K. Protein subcellular localization prediction with WoLF PSORT. In: *Proceedings of the 4th Asia-Pacific bioinformatics conference: 2006*: World Scientific; 2006: 39–48.

108. Team RCJC. RA language and environment for statistical computing. R Foundation for Statistical; 2020.

109. Rombauts S, Déhais P, Van Montagu M, Rouzé P. PlantCARE, a plant cis-acting regulatory element database. Nucleic Acids Res. 1999;27(1):295–6.

110. Tian F, Yang DC, Meng YQ, Jin J, Gao G. PlantRegMap: charting functional regulatory maps in plants. Nucleic Acids Res. 2020;48(D1):D1104-13.

111. Xie J, Chen Y, Cai G, Cai R, Hu Z, Wang H. Tree visualization by one table (tvBOT): a web application for visualizing, modifying and annotating phylogenetic trees. Nucleic Acids Res. 2023;51(W1):W587–92.

112. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.

113. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2019;47(D1):D155-62.

114. Dai X, Zhao PX. psRNATarget: a plant small RNA target analysis server. Nucleic acids Res. 2011;39(suppl_2):W155-9.

115. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(D1):D607-13.

116. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.

117. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.

118. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, bioinformatics GPDPSJ. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.

119. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinform. 2011;12:1–6.

Publisher's note