BMC Genomics

RESEARCH

Open Access

Association of *GhGeBP* genes with fiber quality and early maturity related traits in upland cotton

Jiayan Wu[1†], Ruijie Liu[1†], Yuxin Xie[1], Shuqi Zhao[1,2], Mengyuan Yan[1], Nan Sun[1], Yihua Zhan[1], Feifei Li[1], Shuxun Yu[1*], Zhen Feng[1*] and Libei Li[1*]

Abstract

Transcription Factors (TFs) are key regulators of how plants grow and develop. Among the diverse TF families, the Glabrous-enhancer binding protein (GeBP) family plays a key role in trichome initiation and leaf development. The specific roles of GeBP TFs in plants remain largely unexplored, although GeBP transcription factors play important roles in plants. This study identified 16 *GhGeBP* genes in *Gossypium hirsutum*, ranging from 534 bp (*GhGeBP14*) to 1560 bp (*GhGeBP2*). Phylogenetic analysis grouped 16 *GhGeBP* genes clustered into three subgroups, unevenly distributed across 14 chromosomes. Analysis of the *cis*-acting elements revealed 408 motifs in the 2 kb upstream regions of the promoters, including stress-responsive, phytohormone-responsive, and light-responsive elements. Tissue-specific expression analysis showed 8 *GhGeBP* genes were highly expressed across all tissues, while *GhGeBP4* and *GhGeBP12* were down-regulated under conditions of drought, salt, cold, and heat stress. A genome-wide association study (GWAS) identified *GhGeBP4* was associated with fiber micronaire (FM) and fiber strength (FS), while *GhGeBP9* was linked to the node of the first fruiting branch (NFFB) and flowering time (FT). Haplotype analysis revealed that *GhGeBP4*-HAP2 exhibited higher fiber quality traits, while *GhGeBP9*-HAP2 was associated with early maturity. The results of this study offer significant insights that are worthy of further investigation into the role of the *GhGeBP* gene family in *G. hirsutum* and promising targets for marker-assisted selection strategies in cotton breeding programs, particularly for improving fiber quality and early maturity traits.

Keywords Upland cotton, *GhGeBP*, Fiber quality, Early maturity, GWAS

Background

Transcription factors (TFs) play a pivotal role in regulating the biological processes of plant growth and development. The complexity of higher organisms necessitates the involvement of a diverse array of TFs to orchestrate a broad spectrum of biological functions [1]. TFs typically bind to *cis*-acting elements, modulating gene expression in answer to various environmental stresses [2–4]. Among the various TF families, the GeBP family is unique to plants and plays a crucial role in initiating trichome formation that regulats the generation of trichomes initiation by specially recognizes and binds with the *Glabrous1*(*GL1*) and have a decisive function in leaf

†Jiayan Wu and Ruijie Liu contributed equally to this work.

*Correspondence:
Shuxun Yu
yushuxun@zafu.edu.cn
Zhen Feng
fengzhen@zafu.edu.cn
Libei Li
libeili@zafu.edu.cn
[1] College of Advanced Agricultural Sciences, Zhejiang A&F University, Lin'an, Hangzhou 311300, China
[2] Cotton and Wheat Research Institute, Huanggang Academy of Agricultural Sciences, Huanggang, Hubei 438000, China

growth and development [5]. Despite its importance, research on GeBP TFs specific roles in plants remains limited. Model organisms are the main focus of current studies, like *Arabidopsis thaliana* [6–9] and *Glycine max* [10, 11], which possess 22 and 9 GeBP TF members, respectively. In Arabidopsis, the GeBP TFs exhibit a conserved central DNA-binding domain and a C-terminal region characterized by a putative leucine-zipper pattern. These structural features suggest a shared mechanism of action among family members, though functional diversity is evident from the differential expression patterns observed across various tissues and developmental stages.

In the regulatory networks governing plant development, transcription factors (TFs) play a central role, particularly through their participation in hormonal pathways and interactions. A notable example is the Glabrous-enhancer binding protein binding to the *cis*-regulatory element of the *GL1* gene and regulating its transcription. The *GL1* gene, belongs to the *MYB* family, is essential for determining epidermal cell fate and is modulated by phytohormones such as gibberellins and cytokinins, reflecting the intricate link between genetic regulation and hormonal control in plant development [12]. Further regulation of GeBP expression is mediated by the *KNOX* family TF, *BREVIPEDICELLUS* (BP) [6], which not only upregulates *GeBP* but also enhances the cytokinin signaling pathway in the apical meristem of the stem [13]. Others studies also speculated that GeBP may control the appearance of epidermal hair through regulating the gibberellin and cytokinin pathways [9]. In addition, previous studies have suggested that four TF families including, *MBF1*, *jumonji*, *ULT*, and *GeBP* are mainly involved in plant development and hormone responses, and usually do not participate in stress responsiveness under stress conditions [6, 9, 14–16]. However, a study by Ray et al. (2011) indicated an upregulation of these TFs under water deficit conditions, suggesting a broader role for GeBP in drought stress responsiveness [17]. Likewise, a research by Wang et al. (2023) demonstrated that, in *Brassica rapa*, *BrGeBP5* may regulate low-temperature stress as well as *BrGeBP3* and *BrGeBP14* were probably in response to drought stress [18]. Additionally, a research by Liu et al. (2023) showed that overexpression of *MdGeBP3* from apples in Arabidopsis led to a decrease in drought tolerance [19]. These researches revealed that GeBP may be crucial for developing stress-tolerant crop varieties.

AtCPR5 is an important regulatory factor that controls the growth and development of plants [20, 21]. In the classic disease resistance pathway, *AtCPR5* is downstream in identifying pathogen genes and simultaneously participates in both the Non Expressor of PR Genesl (NPR) and NPR dependent disease resistance signaling pathways. *AtGeBP/GPL* regulates several genes in the *AtCPR5* signaling pathway, which participate in stress resistance and cell wall metabolism processes in the *AtCPR5* signaling pathway. Research involving multiple mutants (*Atgebp*, *gpl1*, *gpl2*, *gpl3*, and *cpr5*) has further elucidated the role of *AtGeBP/GPL*, showing that it specifically controls cell elongation processes dependent on *AtCPR5* signaling, without affecting cell proliferation [22].

Cotton is one of the most important crops globally, serving as a crucial raw material for the textile industry. While the *GeBP* gene family has been well characterized in Arabidopsis and soybean [6, 10], a comprehensive genome-wide analysis of the *GhGeBP* gene family in upland cotton has yet to be conducted. Recent advancements have seen the release of high-quality, assembled genomes and re-sequencing data for cotton, significantly enhancing our capacity to study genomic variations [23–27]. Despite these developments, the identification of critical genomic variations that influence fiber quality, yield, and early maturity traits in upland cotton have not been sufficient to date. This void underscores a significant need for focused research on these economically vital traits. In response to this need, the aim of our study was to carry out a comprehensive analyses of the *GhGeBP* gene family in upland cotton. We utilized genome-wide data to analyze *cis*-elements and evaluated the expression patterns of *GhGeBP* genes in various tissues, and in response to diverse environmental treatments. Additionally, we identified elite haplotypic variations, which could be pivotal for breeding programs. Our research not only enhances the genetic understanding of the *GhGeBP* gene family in cotton but also provides valuable insights and resources for the development of marker-assisted selection strategies in cotton breeding programs.

Materials and methods

Identification and characterization of GhGeBP proteins in upland cotton

Twenty-two of 23 AtGeBP sequences of proteins from *Arabidopsis thaliana* were retrieved from PlantTFDB (https://planttfdb.gao-lab.org/family.php?%20sp=Ath&fam=GeBP) [9] and 1 (*AT2G20805*) from a study by Li et al.(2023) [28]. These sequences served as queries to screen for homologous *GhGeBP* genes. Corresponding sequences of proteins for upland cotton were obtained from the CottonGen (https://www.cottongen.org/) [23, 29]. Further, identification of putative GeBP proteins in upland cotton was conducted using a two-step bioinformatics approach. Initially, the BLAST was employed with a threshold of E-value less than 1×10^{-10} to ascertain preliminary candidates. Subsequently, The presence

of the GeBP-specific conserved domain (DUF573) was tested in these candidates by using the HMMER software (version: 3.0) to identify GhGeBP proteins under default parameters (e-value less than 0.001) [30–32]. To confirm the structural integrity of the identified proteins, all candidates were further analyzed against the Conserved Domain Database (CDD) hosted by the NCBI (https://www.ncbi.nlm.nih.gov/cdd/). This step ensured the accuracy of our domain-specific identifications.

Physicochemical characterization and subcellular localization

GeBP protein family physicochemical properties, including amino acid number, molecular weight (MW), and isoelectric point (PI), were determined using ExPASy ProtParam (http://web.expasy.org/protparam/). For subcellular localization predictions, the cello BUSCA (http://www.busca.biocomp.unibot.it/) was utilized to infer the likely cellular compartments where the GhGeBP proteins might exert their biological functions, further informing our understanding of their role in cotton physiology.

Sequence alignment and phylogenetic tree construction

The full-length GeBP protein sequences of *Arabidopsis thaliana*, *Oryza sativa*, *Glycine max*, and upland cotton were aligned using the FFT-NS-2 algorithm implemented in the MAFFT software (version: 7.310) [33]. This alignment provided the basis for subsequent phylogenetic analyses. Iqtree (version: 2.0.3) was then used to construct a maximum likelihood (ML) phylogenetic tree, selecting 'JTT + F + R4' as the best fit of the substitution model, supported by 1,000 bootstrap replicates [34]. The resulting phylogenetic tree was visualized using the 'ggtree' package in R software(version 4.3.1) [35].

Chromosome distribution, collinearity analysis and selection pressure calculation

Chromosome localization of the *GhGeBP* genes was visualized using TBtools (version: 2.080) [36]. The MCScanX toolkit (https://github.com/wyp1125/MCScanX) was employed to set genome-wide synteny relationships and to identify collinear gene pairs within the *GhGeBP* gene family. Selection pressure on the *GhGeBP* family genes was assessed using the KaKs_Calculator (version: 2.0) [37], with the Yang-Nielsen (YN) model [38] specified for calculation of the non-synonymous (Ka) and the synonymous (Ks) substitution rate.

Gene structure, *cis*-acting elements, motifs analysis, and prediction of protein tertiary structure

Gene structure information, including promoter regions, 5'UTR, 3'UTR, and exon-intron organization for the *GhGeBP* genes, was extracted from the GFF3 file of the genome of upland cotton [23]. Visualization of these structures was conducted using the Gene Structure Display Server (version: 2.0) (http://gsds.gao-lab.org/). Additionally, these genes' upstream sequences (2,000 bp) were retrieved using a custom Python script and analyzed for *cis*-acting elements using the PlantCARE database (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/). The MEME (http://meme-suite.org/index.html) was predicted motifs for GeBP proteins. Finally, the visualization were created by TBtools [36]. Protein tertiary structures were predicted through Phyre2 online websites (http://www.sbg.bio.ic.ac.uk/phyre2) [39], and the 3D structure prediction of GhGeBP proteins were in Fig. S1.

Association analysis of *GhGeBP* genes with early maturity and fiber quality related traits

To verify the potential functional of *GhGeBPs*, an association analysis was conducted between these genes and 12 key agronomic traits (5 fiber-related traits and 7 early maturity related traits) in 355 upland cotton accessions. Accessions used for fiber related traits were planted in Anyang, Henan, China (36°08'N,114°48'E), and Shihezi, Xinjiang, China (44°31'N,86°01'E) in 2014 and 2015. Every year normally open bolls are taken from each repeating fruit branch. The fiber-related quality traits (FL(mm), FS(cN/tex), FU(%), FM and FE(%)) were detected in 10–15 g fiber samples [40]. Besides, from 2015 to 2017, multi-environment experiments were conducted in Anyang, Henan, China (36°08'N,114°48'E), Hubei, Huanggang, China (30°57'N, 114°92'E) and Shihezi, Xinjiang, China (44°31'N,86°01'E). Seven early maturity related traits (FT (days), FBP (days), WGP (days), YPBF (%), NFFB, HNFFB (cm) and PH (cm)) were identified [27]. Total 67 single nucleotide polymorphisms (SNPs) within the 2,000 bp upstream and downstream regions of the *GhGeBPs* were taken from the resequencing data of the 355 upland cotton accessions [27]. A linear mixed model (LMM) was utilized to execute the associations by incorporating all phenotype and SNP data based on GEMMA software (version: 0.98.3) [41]. The significant threshold value for association was set at $-\log_{10}P \geq 2$. Manhattan plots were generated using the 'ggplot2' package in R software [42].

Haplotype and genetic analysis of *GhGeBP* genes

The phenotypic effects of elite haplotypes were visualized using the 'ggplot2' package in R software [42]. Nucleotide diversity values for each group were estimated using VCFtools software (version: 0.1.16) [43] based on the release and breeding years of each accession and geographical distribution.

Transcriptome expression analysis of *GhGeBP* genes

The expression profiling from *GhGeBPs* in different tissues were analyzed using four sets of public transcriptome data (PRJNA490626, PRJNA509318, PRJNA793063, and PRJNA529497), available from NCBI (https://www.ncbi.nlm.nih.gov/bioproject/). The dataset PRJNA490626 [24] comprised samples from multiple organs of the upland cotton variety 'TM-1', including roots, stems, leaves, bracts, anthers, fibers, ovules, and various reproductive organs. And beyond that, it also contains several stress treatments (NaCl (0.4 M), PEG (200 g/L), 4℃ and 37℃) that act on 'TM-1'. PRJNA509318 [27] included flower bud samples collected at five developmental stages (square, 5 days post-square (DPS), 10 DPS, 15 DPS, and 20 DPS). PRJNA793063 [44] contained long and short fiber samples from an interspecific backcross inbred line (BIL) population that was derived from a cross between *Gossypium hirsutum* (CRI36) and *Gossypium barbadense* cotton (Hai7124) at 5, 10 and 15 DPA. PRJNA529497 [45] encompassed 10 and 20 DPA fiber samples from the semi-wild species *Gossypium hirsutum* 'TX2094' and the cultivated species 'Acala Maxxa'. Quality control for each transcriptome dataset was conducted using Trimmomatic (version: 0.39), with modified parameters (PE threads: 50, HEADCROP: 7, SLIDINGWINDOW: 4:15, MINLEN: 80). *GhGeBP* genes expression levels were quantified as Transcripts Per Million (TPM) values using Salmon (version: 1.10.0) [46], based on the coding sequence file of 'TM-1' [23, 47]. The 'pheatmap' package [48] in R software was used to generate heatmaps, with expression values normalized to $\log_2$ TPM.

qRT-PCR analysis of *GhGeBP9*

For further validation, qRT-PCR analyses were conducted. Primers designed are detailed in Table S1. Additionally, accessions 'Han9609', 'Xinluzhong 34' were planted at the Sanya, Hainan, China (18°36'N, 109°17'E). Samples from axillary buds were collected, immediately frozen in liquid nitrogen, and stored at −80℃. Total RNA was extracted using the FastPure Universal Plant Total RNA Isolation Kit (RC411–01, Vazyme). cDNA synthesis was conducted using HiScript® II Q RT SuperMix for qPCR (R223). Quantitative real-time PCR was performed using the Light Cycler 480 II system, and gene expression levels were calculated using the $2^{-\Delta\Delta Ct}$ method [49] with three replicates per sample.

Results

Identification and protein characteristics of *GeBPs* gene family in upland cotton

Genome-wide searches have identified total 16 *GhGeBP* gene members containing the complete DUF573 domain in the *G. hirsutum* genome and the genes were renamed according to their chromosomal positions (Table S2). The physical and chemical properties of these *GhGeBP* proteins were further analyzed. The coding sequences varied from 534 bp (*GhGeBP14*) to 1560 bp (*GhGeBP2*), with 1,046 bp average length. The *GhGeBP* genes encoded proteins with amino acid lengths ranging greatly from 177 aa to 519 aa, in which *GhGeBP2* has the largest number of amino acids and *GhGeBP14* has the smallest. The predicted molecular weights (MW) ranged from 19,939.60 to 58,833.48 Da. Isoelectric points (pI) varied from 4.63 to 10.23, with 11 genes encoding acidic proteins (pI ≤ 6.5, average 5.30) and 4 genes (*GhGeBP3*, *GhGeBP6*, *GhGeBP11*, and *GhGeBP14*) encoding alkaline proteins (pI > 8, average 9.81). The grand average of hydropathicity (GRAVY) values ranged from −0.943 to -0.466, indicating that all GhGeBP proteins are predominantly hydrophilic. Predicted subcellular localization revealed that the majority of GhGeBP proteins are situated in the nucleus, except for *GhGeBP14*, which is predicted to be in the cytosol. This suggests that most of GhGeBP proteins are likely involved in nuclear functions (Table S2).

Phylogenetic analysis of the *GeBP* gene family

An unrooted phylogenetic tree was constructed by the full-length amino acid sequences from *A. thaliana* (23 genes), *O. sativa* (15 genes), *G. max* (9 genes), and *G. hirsutum* (16 genes) to compare the evolutionary relationships among *GeBP* genes (Table S3). In accordance with the phylogenetic tree, the 63 *GeBP* genes were divided into six distinct subgroups (Fig. 1). Among them, 17 *AtGeBP* genes formed an independent cluster in group I, 3 *GmGeBP* genes were grouped in group III. This indicates that *GeBP* genes are conserved across both monocotyledons and dicotyledons. In *G. hirsutum*, the 16 *GhGeBP* genes were divided into three subgroups, though unevenly distributed: 8 members in group II (the largest subgroup of *GhGeBP* genes), 2 in group V, and 6 in group VI. The homologous genes from the At and Dt subgenomes of *G. hirsutum* clustered together within the same groups due to their high similarity. For instance, *GhGeBP3* and *GhGeBP11* clustered together in group V, sharing an average similarity of 23.48% with *AtGeBP7*. In group VI, *GhGeBP6* and *GhGeBP14* were grouped together, sharing an average similarity of 32.59% with *OsGeBP7*. These close phylogenetic relationships suggest that these homologous genes may share similar genetic functions.

Chromosome distribution and collinearity analysis

The chromosome position analysis revealed: *GhGeBPs* are unevenly distributed on the upland cotton chromosomes. The *GeBP* genes in upland cotton were mainly paired distribution on 14 chromosomes (A01/

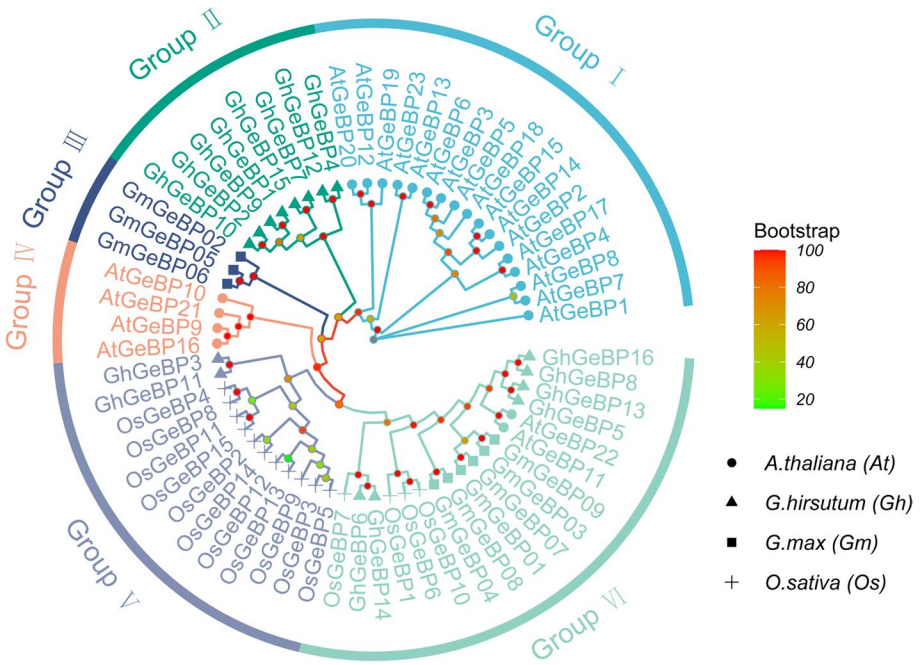

Fig. 1 Phylogenetic tree of *GeBP* genes between upland cotton, *Arabidopsis thaliana*, *Glycine max* and *Oryza sativa*. Different symbols correspond to different species, The color-coded circular nodes at the branch points of the tree represent bootstrap values, providing statistical support for the phylogenetic inferences. Higher bootstrap values indicate greater reliability of the inferred evolutionary relationships

D01, A05/D05, A07/D07, A10/D10, A11/D11, A12/D12, A13/D13). Chromosome A13 and Chromosome D13 contained two *GeBP* genes, and other chromosomes both had only one *GeBP* gene (Fig. 2a), which was due to genomic segmental duplication or rearrangement. In addition, we simultaneously identified 16 *GhGeBP* genes in upland cotton that were involved in 10 blocks of synteny, with the exception of *GhGeBP2*, *GhGeBP10*, and *GhGeBP12* (Fig. 2b). Most of synteny blocks, 8 (80%) were located between the At and Dt sub-genomes, including 13 orthologous *GhGeBP* genes in whole upland cotton genome. Among these genes, *GhGeBP8* and *GhGeBP13* are located within three collinearity synteny blocks, respectively. *GhGeBP5* and *GhGeBP16* involve two collinearity synteny blocks, respectively. While the remaining genes were involved in one collinearity synteny blocks.

Selection pressure analysis (Ka/Ks ratios)

Across the entire genome, the *GhGeBP* gene family comprises five homologous gene pairs, all with Ka/Ks ratios below 1 (Table S4). Notably, four of these pairs have Ka/Ks ratios below 0.31, implying that the *GhGeBP* genes have undergone purifying selection during their evolution. This selective pressure has likely contributed to the contraction of this gene family.

Analysis of the gene structure of *GhGeBP* genes and the *cis*-acting elements

Gene promoters contain specific sequences that bind to transcription factors to control the expression of genes that play a role in development and responses to stress throughout the plant's life cycle. The promoter regions on 2,000 bp upstream of the start codon of each *GhGeBP* gene were analyzed using PlantCARE to further predict the *cis*-acting regulatory elements of *GhGeBP* genes. Total 408 *cis*-acting elements, which represents 19 types, were predicted and grouped into five categories (Fig. 3a, b and Table S5): light-responsive, stress-responsive, phytohormone-responsive, plant organogenesis-related, development-related elements and plant growth. Among these *cis*-acting elements, numerous light-responsive elements were found in the *GhGeBP* genes promoter regions, representing 55.29% of the predicted *cis*-elements. These included GT1-motif, G-Box, Box4, and others, which were distributed across each *GhGeBP* gene promoter region. We found that Auxin-responsive elements (AuxRR-core, AuxRE, and TGA-element) were in 11 *GhGeBP* genes promoter regions. We also identified five types of hormone-responsive regulatory elements. Thirty-eight ABRE motifs associated with abscisic acid responsiveness were particularly abundant, with *GhGeBP12* containing the most motifs. Salicylic acid-responsive TCA-elements were present in *GhGeBP1*, *GhGeBP2*, *GhGeBP7*, *GhGeBP9*, *GhGeBP15*,

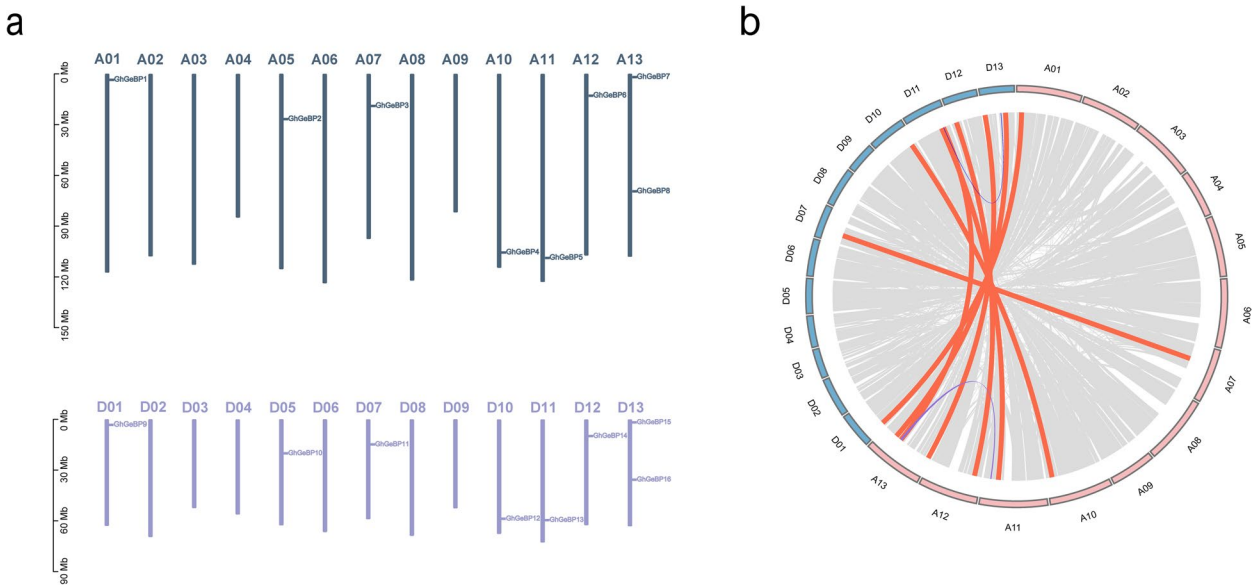

Fig. 2 *GhGeBP* genes chromosomal localization and duplication relationships. **a** Localization of *GhGeBP* on the chromosomes of upland cotton; **b** Gene duplication relationship among *GhGeBP* genes. The outer circle is color-coded to denote the different subgenomes: pink means the At subgenome and blue means the Dt subgenome. The line graph within the circle represents gene density across the chromosomes. Background shading indicates various types of gene duplication relationships: the gray background signifies all gene duplications within the genome, the red background shows duplications between chromosome groups, and the purple and blue backgrounds highlight duplications within the At and Dt subgenomes, respectively

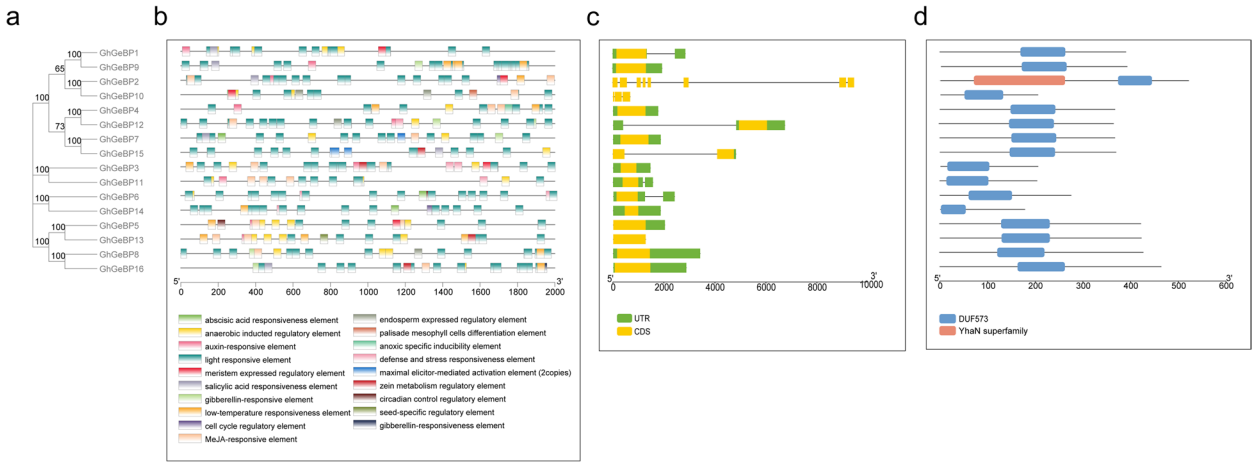

Fig. 3 *Cis*-acting elements and gene structures of the *GhGeBP* promoter were analyzed comparatively. **a** Clustering analysis of genes in line with the *GhGeBP*phylogenetic tree; **b** *Cis*-acting elements visualization and classification in the *GhGeBP* promoter region, where each color represents a specific functional category of *cis*-acting elements, enabling a visual differentiation of regulatory motifs associated with various biological processes; **c** Gene structure of *GhGeBP*, displaying the arrangement of exons, introns, and untranslated regions (UTRs); **d** The distribution of conserved domains on GhGeBP proteins

and *GhGeBP16*. The MeJA-responsive CGTCA motif was uniquely found in 11 *GhGeBP* genes (*GhGeBP2*, *GhGeBP3*, *GhGeBP4*, *GhGeBP5*, *GhGeBP7*, *GhGeBP8*, *GhGeBP10*, *GhGeBP11*, *GhGeBP12*, *GhGeBP13*, and *GhGeBP16*). Including GARE-motif, P-box, and TATC-box, elements that react to gibberellin were identified in *GhGeBP12*, *GhGeBP16*, *GhGeBP7*, *GhGeBP8*, and *GhGeBP9*. These results demonstrate that most *GhGeBP* genes might engage in multiple phytohormone signaling pathways. In addition, regarding stress-related responsive, TC-rich repeats (responsing defense and stress), GC-motif (anoxic inducibility), ARE (anaerobic induction), and LTR

(low-temperature responsiveness) were more prevalent. Endosperm-expressed elements, meristem-expressed elements, and palisade mesophyll cell differentiation elements, for plant organogenesis and development, were recognized in *GhGeBP10* promoter regions. Circadian control regulatory elements were detected in several genes, including *GhGeBP3*, *GhGeBP5*, and *GhGeBP6*. These results suggest that *GhGeBP* genes may be of importance in hormone signal transduction, cotton development, and response to abiotic stress. Table S5 contains detailed information on the *cis*-elements.

Ground on genomic sequence analysis and gff3 file, *GhGeBP* genes structures and domain distribution were conducted (Fig. 3c, d). Except for GhGeBP2, which contains an additional YhaN domain, all other proteins have only a single DUF573 domain. The exon number varied between 1 and 8, and *GhGeBP* genes mostly displayed either intron-less structures or contained only a few introns. Of the 16 *GhGeBP* genes, 56.25% (*GhGeBP3*, *GhGeBP4*, *GhGeBP5*, *GhGeBP7*, *GhGeBP8*, *GhGeBP9*, *GhGeBP13*, *GhGeBP14*, and *GhGeBP16*) contained only one exon and lacked introns, four were two-exon genes (*GhGeBP1*, *GhGeBP6*, *GhGeBP11*, and *GhGeBP15*), and three had three or more exons (*GhGeBP2*, *GhGeBP10*, and *GhGeBP12*). *GhGeBP2* had the highest number of exons (8) and introns (7). Notably, *GhGeBP* genes within the same phylogenetic cluster displayed similar gene structures. Furthermore, the 3D structure prediction of GhGeBP proteins certified the result above (Fig. S1).

Analysis of the tissue-specific pattern of *GhGeBPs* expression

To fully understand how the *GhGeBP* genes function, we analyzed their spatial and temporal expression profiles utilizing RNA-seq data in kinds of tissues (root, leaf, stem, sepal, torus, anther, bract, filament, petal, pistil, ovule, and fiber) at every five days between 10 and 25 days post-anthesis (DPA) under normal growth conditions (Fig. 4a). The results showed that 8 *GhGeBP* genes (*GhGeBP1*, *GhGeBP3*, *GhGeBP6*, *GhGeBP8*, *GhGeBP9*, *GhGeBP11*, *GhGeBP14*, and *GhGeBP16*) exhibited high expression levels relatively across all tissues. Contrarily, *GhGeBP12* and *GhGeBP15* displayed weak or no expression in any of the tissues analyzed. Interestingly, *GhGeBP5* and *GhGeBP7* were specifically expressed during ovule development, while *GhGeBP1* and *GhGeBP9* were strongly expressed during fiber development. Notably, *GhGeBP1* exhibited the highest expression level in fiber development (TPM ≥ 23), followed by *GhGeBP9* (TPM ≥ 43). Overall, throughout the reproductive cycle of upland cotton, most members of the *GhGeBP* gene family were expressed., suggesting their involvement in regulating the cotton to grow and develop.

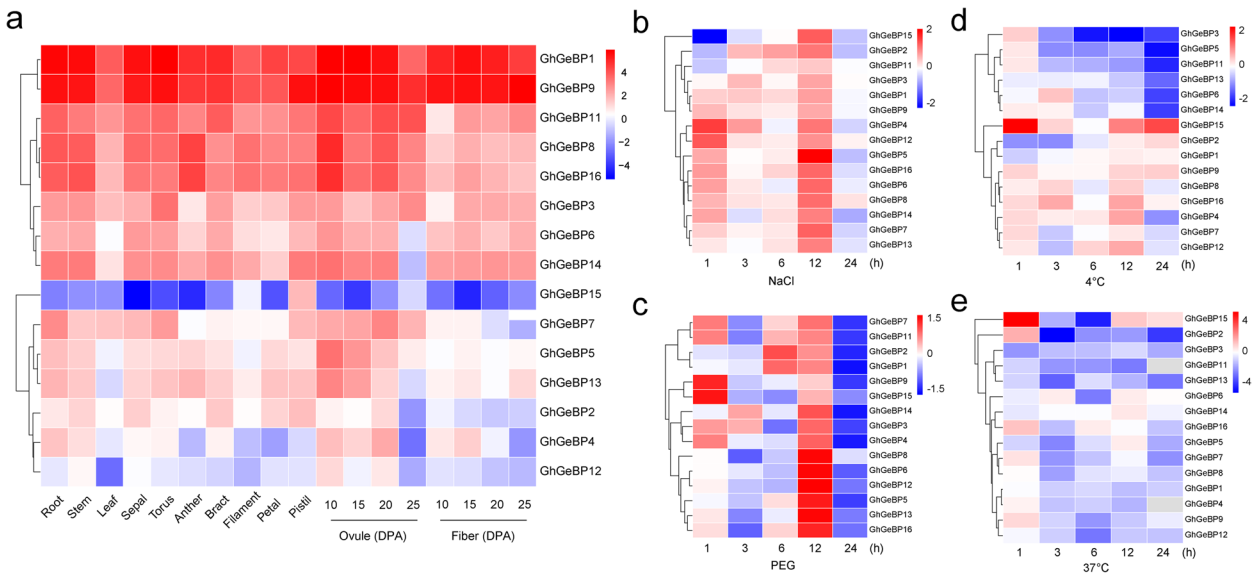

Fig. 4 Heatmap of *GhGeBP* genes' expression in diverse tissues and under various conditions of stress. **a** Heatmap of *GhGeBP* expression in diverse tissues, with color gradation from blue to red indicating an increase in expression levels. **b** Heatmap of the time-based expression pattern of *GhGeBP* genes at 1, 3, 6, 12, and 24 h after treatment under drought stress induced by 200 g/L PEG; **c** Heatmap of *GhGeBP* expression at 1, 3, 6, 12, and 24 h under salt (0.4 M NaCl) stress conditions; **d** Heatmap of *GhGeBP* expression at 1, 3, 6, 12, and 24 h under cold (4 °C) stress conditions; **e** Heatmap of *GhGeBP* genes expression at 1, 3, 6, 12, and 24 h under heat (37 °C) stress conditions

Analysis of the pattern of *GhGeBP* expression under the influence of abiotic stress

To investigate whether *GhGeBP* genes participate in abiotic stress responses, their expression patterns were analyzed under drought (200 g/L PEG), salt (0.4 M NaCl), cold (4 °C), and heat (37 °C) stress conditions (Fig. 4b, c, d, e). Collating the four stress conditions, more *GhGeBP* genes exhibited changed expression in answer to dryness and salt stresses than to cold and heat stresses. Most *GhGeBP* genes were up-regulated under salt stress, with the highest differential expression observed at 12 h, after which expression levels significantly decreased at 24 h. In contrast, *GhGeBP4* and *GhGeBP12* were consistently downregulated at 3, 6, 12, and 24 h, demonstrating that these two genes might make negative affect in answer to salt stress. Similarly, the pattern of *GhGeBP* expression under dryness stress (simulated with PEG) closely resembled those observed under salt stress. Notably, *GhGeBP9* and *GhGeBP15* were significantly downregulated at 3, 6, 12, and 24 h under PEG-induced drought conditions. Interestingly, *GhGeBP15* exhibited relatively low expression in all tissues but showed increased expression and significant upregulation under all four stresses which suggests that *GhGeBP15* may be important to enhance resistance to multiple nonliving stresses.

GWAS of characteristics of upland cotton that correlate with early maturity and fiber quality

In a GWAS of 355 cotton germplasm accessions, all 67 SNPs dispersed across the 16 *GhGeBP* genes with a minor allele frequency (MAF) > 0.05 and a resequencing data miss rate < 20% were used. 23 Of these SNPs were located in the At sub-genome and 43 in the Dt subgenome. Most SNPs were found in intergenic upstream regions (52.23%). Using a linear mixed model (LMM), total 16 and 6 important SNPs were identified for seven early maturity characteristics (FT, FBP, WGP, YPBF, NFFB, HNFFB, and PH) and five fiber quality characteristics (FU, FM, FL, FE and FS), respectively, across five *GhGeBP* genes (*GhGeBP4*, *GhGeBP9*, *GhGeBP11*, and *GhGeBP15*) (Table S6). Only *GhGeBP4* was found in the At subgenome, whereas the remaining genes (*GhGeBP9*, *GhGeBP11*, and *GhGeBP15*) were situated in the Dt subgenome. *GhGeBP9*, exhibited a robust SNP cluster (15 significant SNPs) associated with NFFB and FT, with $-\log_{10}(P)$ values ranging from 13.05 to 13.98 and a phenotypic variation explained (PVE) of over 14% (14.54–15.50). Additionally, a stable SNP (rsD07_14834977) in the promoter region of *GhGeBP11* with identified through eight traits (FT, HNFFB, WGP, YPBF, NFFB, PH, FBP, and FE), with a PVE ranging from 2.45 to 4.89% and $-\log_{10}(P)$ values ranging from 2.52 to 4.59. Three SNPs (rsD13_1727983, rsD13_1727984, and rsD13_1728125) downstream of *GhGeBP15* were associated with FM, explaining 1.98–2.09% of the observed PVE. Interestingly, two SNPs (rsA10_105653025 and rsA10_105655897) in the promoter and 3'UTR regions of *GhGeBP4* were linked to FS and FM, explaining 1.99% and 2.05% of the PV, respectively. Thus, *GhGeBP4* and *GhGeBP9* could be considered major candidate genes to continue dissecting.

Candidate gene for fiber quality on A10

In the association analysis of fiber-related traits, *GhGeBP4* was significantly associated with the highest number of traits (FM and FS) (Fig. 5a). Four SNPs were identified within *GhGeBP4*, located in the 2 kb upstream region and the 3'UTR (Fig. 5b). According to publicly available transcriptome data, *GhGeBP4* showed peak expression at 15 DPA as the fiber developed, followed by a gradual decline (Fig. 5c). Haplotype analysis classified *GhGeBP4* into two haplotypes, with HAP2 showing higher FS, FL, FU, and FE (Fig. 5d). This suggests that HAP2 may represent a superior haplotype for fiber quality, providing a new foundation for breeding elite fiber accessions. Gain insight into the genetic basis of HAP2 and its geographical distribution, 50 domestic upland cotton varieties containing HAP2 were grouped into four regional clusters: Northwest Inland Cotton Region (NIR) (29 varieties), Northern Super-Early Maturing Region (NSER) (2 varieties), Yellow River Region (YRR) (9 varieties), and Yangtze River Region (YZRR) (10 varieties) (Fig. 5e). The NIR cluster accounted for over 50% of the germplasm, consistent with the superior fiber quality of upland cotton in Xinjiang. Based on the variety approval years, the 355 germplasm resources were categorized into four breeding periods: pre-1950 (10 varieties), 1950–1979 (32 varieties), 1980–1999 (65 varieties), and 2000 to the present (175 varieties). An additional 73 varieties had unknown approval years. Nucleotide diversity analysis of this region (A10: 105,652,245 – 105,658,012) across these four breeding periods showed the highest diversity in *GhGeBP4* before 1950, followed by a gradual decrease, reflecting breeding goals focused on improving fiber quality traits (Fig. 5f). Further analysis of BILs derived from a cross between CRI36 and Hai7124 revealed that during the fiber elongation phase, although generally decreasing, the expression of *GhGeBP4* was higher in LF Group (long fiber group) in comparison to SF Group cotton (short fiber group) (Fig. 5g). Additionally, expression analyses of *GhGeBP4* at 10 DPA and 20 DPA between the cultivated variety (Maxxa) and a semi-wild type (TX_10) showed that *GhGeBP4* expression was consistently higher in the cultivated upland cotton than in the semi-wild type (Fig. 5h). These findings suggest a central role of *GhGeBP4* in developing fibers, potentially contributing

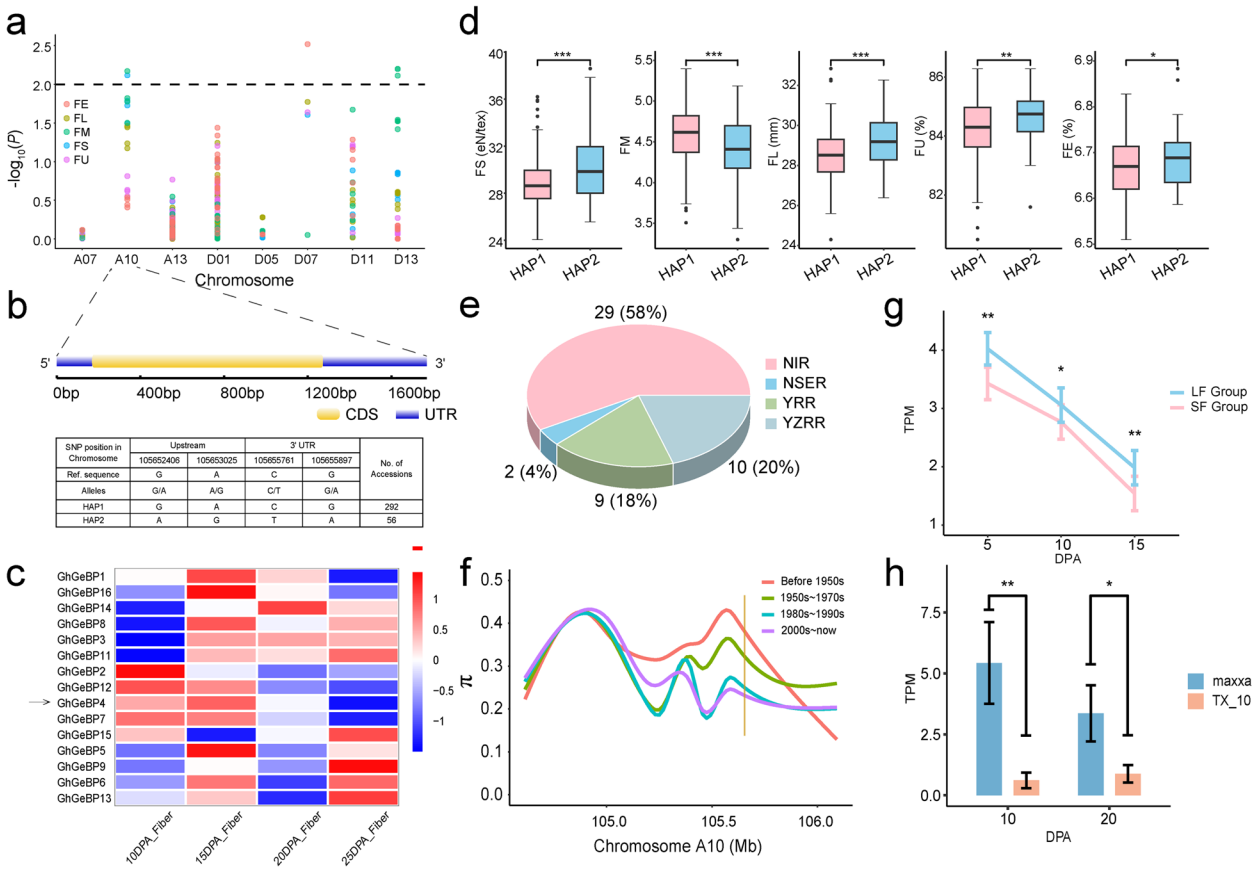

Fig. 5 Association analysis of characteristics of upland cotton that correlate with fiber quality. **a** Manhattan plot across five characteristics of upland cotton that correlate with fiber quality; **b** Gene structure and two haplotypes of *GhGeBP4*; **c** Heatmap of 15 *GhGeBP* family genes expression during the fiber development period; **d** Box plots comparing the phenotypic variation of the five different traits between the two haplotypes identified; **e** Pie charts of the distribution of the 50 HAP2-containing accessions in the four cotton ecological regions; **f** Nucleotide diversity during four distinct breeding periods, with a vertical line marking the location of *GhGeBP4*; **g** The expression of *GhGeBP4* on LF Group (long fiber group) and SF Group (short fiber group) cotton accessions at 5–15 DPA; **h** Significance analysis of *GhGeBP4* expression in cultivated (maxxa) and semi-wild (TX_10) species at 10–20 DPA. (** $P < 0.01$, * $P < 0.05$)

to the enhanced fiber qualities observed in cultivated varieties.

Candidate gene for early maturity on D01

In the association analysis of early maturity traits, SNPs within *GhGeBP9* showed significantly higher associations than other members of the *GhGeBP* gene family, particularly for NFFB (Fig. 6a). Most SNPs were in downstream regions, with no variants identified within the gene itself (Fig. 6b). Further haplotype analysis identified two haplotypes of *GhGeBP9*, HAP1 and HAP2. Varieties carrying the HAP2 haplotype exhibited lower NFFB, lower HNFFB, a shorter FT, a shorter FBP, and a shorter WGP (Fig. 6c). This suggests that HAP2 may represent a superior haplotype for early maturity traits, providing valuable insights for future research. The geographical distribution of these haplotypes showed that the distribution rate of the superior HAP2 haplotype was higher in NIR and NSER compared to YRR and YZRR (Fig. 6d). As a representative, a nucleotide diversity analysis of NIR was conducted. Nucleotide diversity analysis of chromosome D01 across three breeding periods showed that the nucleotide diversity of *GhGeBP9* was higher during the 1950–1970 s than in other periods, and its diversity has not changed significantly from the 1980s to the present, indicates that the compartment where *GhGeBP9* is located has been domesticated (Fig. 6e). Public transcriptome data showed that the *GhGeBP9* expression level in the standard upland cotton line TM-1 was slightly higher than in the early-maturing variety CRI50 at 0–5 DPS. After 5 DPS, the expression level of *GhGeBP9* in TM-1 gradually decreased and became lower than that in CRI50 (Fig. 6f). Moreover, comparing the expression levels of *GhGeBP9* in TM-1 and CRI50 at different developmental stages revealed considerably higher expression in the early-maturing variety (Fig. 6g). This indicates that

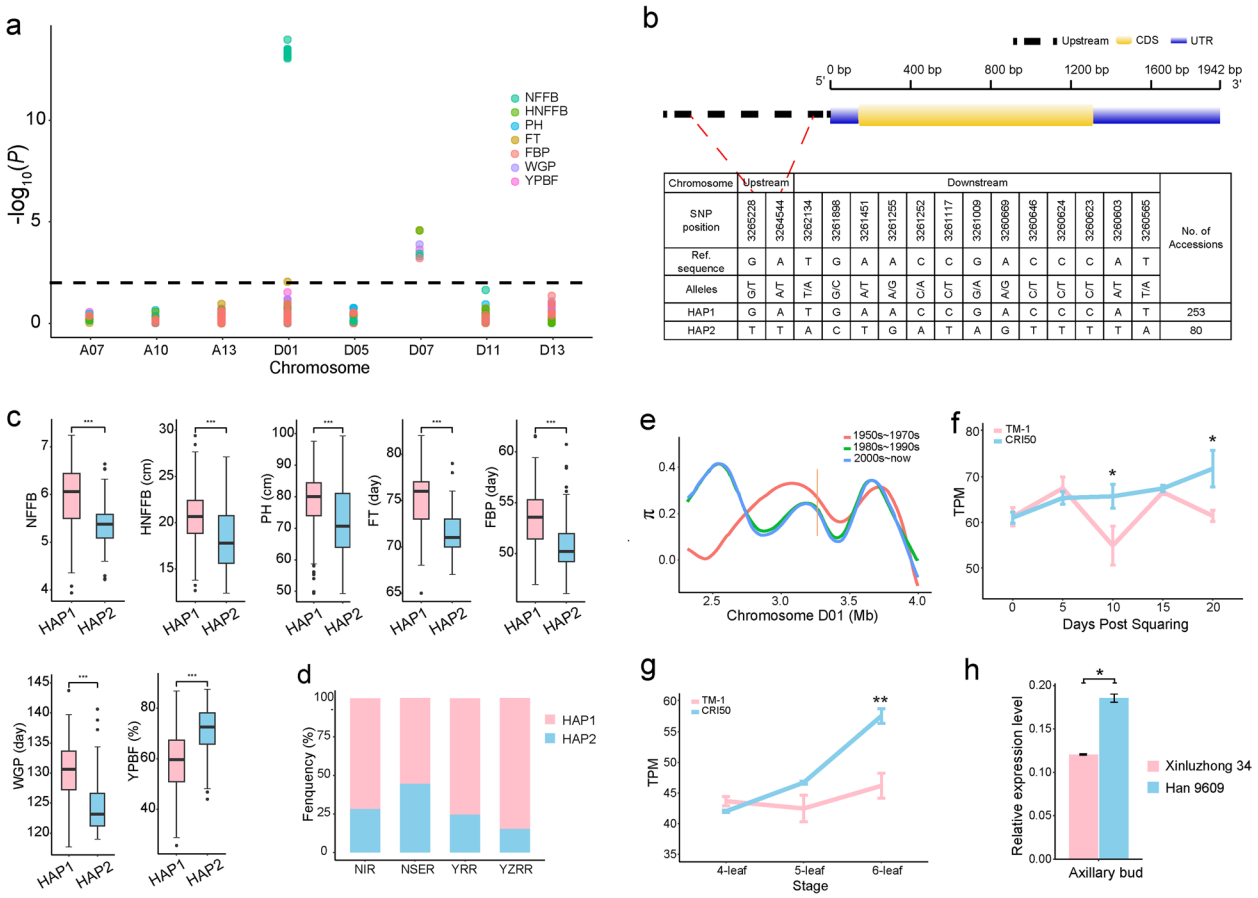

Fig. 6 Association analysis of characteristics of upland cotton that correlate with early maturity. **a** Manhattan plot of seven characteristics of upland cotton that correlate with early maturity; **b** Gene structure and two haplotypes of *GhGeBP9*; **c** Box plots of seven traits among different haplotypes; **d** Differences in haplotype distributions of *GhGeBP9* among the four cotton planting regions; **e** Nucleotide diversity of *GhGeBP9* on chromosome D01 over various breeding periods in NIR; **f** Expression pattern of *GhGeBP9* in two cotton varieties, TM-1 and CRI50, from 0 to 20 days post-squaring (DPS); **g** Comparative expression analysis of *GhGeBP9* in TM-1 and CRI50 across different leaf developmental stages; **h** Expression level analysis of *GhGeBP9* between 'Xinluzhong 34' which carried HAP1, and 'Han 9609' which carried HAP2. (** $P < 0.01$, * $P < 0.05$)

GhGeBP9 is central to influence the maturity of upland cotton. The qRT-PCR further validated higher expression levels in HAP2 accessions compared with HAP1, aligning with observed phenotypic differences (Fig. 6h).

Disscussion

Our extensive analysis of the *GhGeBP* gene family confirmed two key genes, *GhGeBP4* and *GhGeBP9*, associated with significant traits such as fiber quality and early maturity. Here, we discuss the roles of these genes in cotton breeding. *GhGeBP4* has been identified as a significant contributor to fiber quality traits and classified into two haplotypes. HAP2, one of which, exhibits superior fiber quality characteristics like higher FL, FS, and FE. This advantageous haplotype is prevalent in the Northwest Inland Cotton Region, aligning with the high-quality cotton produced in Xinjiang. The expression pattern of *GhGeBP4* at 15 DPA aligns with the stages of fiber elongation and secondary wall thickening, which are crucial for cotton fiber quality [50, 51]. Comparisons of expression data between long- and short-fiber cotton materials, as well as between wild cotton and cultivated varieties, corroborated this finding. The presence of *cis*-acting regulatory elements like light-responsive (G-Box, Box4) and hormone-responsive motifs (P-box, GARE-motif, ABRE, CGTCA) in the *GhGeBP4* promoter indicates its involvement in phytohormone signaling pathways. This is in line with the reported role of *GeBP* family genes in managing phytohormone pathways like gibberellins and cytokinins [9, 52]. These findings suggest that *GhGeBP4* may regulate fiber development through light and hormone signaling.

GhGeBP9 was identified as a significant contributor to early maturity traits in upland cotton. SNPs within *GhGeBP9* showed significant associations with the NFFB, FT. Haplotype analysis revealed two haplotypes,

HAP1 and HAP2, with HAP2 exhibiting superior early maturity traits. The geographical distribution of HAP2 aligns with the breeding goals in these regions, where early maturity is desirable [53]. *GhGeBP9* expression in CRI50 and TM-1 indicates that *GhGeBP9* is crucial to determine maturity timing. Identification of *GhGeBP4* and *GhGeBP9* as candidate genes for fiber quality and early maturity traits provides valuable targets for applying marker-assisted selection (MAS) in cotton breeding programs. *GhGeBP4* is a critical regulator of fiber development, while *GhGeBP9* influences early maturity traits. Both genes exhibit specific haplotypes associated with desirable phenotypes, offering potential for developing high-quality and early maturity cotton varieties. Further functional validation of these genes through CRISPR/Cas9 or RNA interference (RNAi) will clarify their precise roles in cotton physiology.

Gene family evolution is a crucial driver of plant genome diversity and adaptation. Gene family expansion, involving gene duplication and subsequent divergence, has led to the proliferation of transcription factor families such as *WRKY*, *ARF*, and *MYB* [54–57]. However, in some cases, gene families may also undergo contraction. A reduction in the number of *GhGeBP* genes was observed in our study, supported by evolutionary analysis using the ratio of Ka to Ks substitutions (Table S4). A ratio of Ka to Ks < 1 suggests purifying selection, indicating that most mutations are deleterious and that the gene is under strong functional constraint [58]. Despite the contraction of the *GhGeBP* gene family, key members like *GhGeBP4* and *GhGeBP9* still regulate important traits in upland cotton. These key genes suggest that even with a reduced number of members, the *GeBP* family retains its importance in cotton biology.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10983-y.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements
Not applicable.

Authors' contributions
LL, JW and RL wrote the original draft manuscript text, ZF, LL, and SY revised the main manuscript text, SZ, MY and YZ carried out data collation, JW, YX and FL completed data analysis, RL and NS performed data visualization. All authors reviewed the manuscript.

Funding
This research was sponsored by National Natural Science Foundation of China (32401297), Cotton Bio-breeding and Integrated Utilization Open Fund (CB2023A09) and Zhejiang A & F Student Research Training Program (2023KX007).

Data availability
The genomic resequencing data for GWAS analysis are available in the NCBI Sequence Read Archive under the accession PRJNA389777.Additional data are provided within the manuscript and 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.

Received: 13 July 2024 Accepted: 30 October 2024
Published online: 08 November 2024

References

1. Riechmann JL, Ratcliffe OJ. A genomic perspective on plant transcription factors. Curr Opin Plant Biol. 2000;3(5):423–34.
2. Yamaguchi-Shinozaki K, Shinozaki K. Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. Annu Rev Plant Biol. 2006;57:781–803.
3. Bartels D, Sunkar R. Drought and Salt Tolerance in plants. CRC Crit Rev Plant Sci. 2005;1:24.
4. Singh K, Foley RC, Oñate-Sánchez L. Transcription factors in plant defense and stress responses. Curr Opin Plant Biol. 2002;5(5):430–6.
5. Perazza D, Vachon G, Herzog M. Gibberellins promote trichome formation by up-regulating *GLABROUS1* in Arabidopsis. Plant Physiol. 1998;117(2):375–83.
6. Curaba J, Herzog M, Vachon G. *GeBP*, the first member of a new gene family in Arabidopsis, encodes a nuclear protein with DNA-binding activity and is regulated by *KNAT1*. Plant J. 2003;33(2):305–17.
7. Kazama D, Kurusu T, Mitsuda N, Ohme-Takagi M, Tada Y. Involvement of elevated proline accumulation in enhanced osmotic stress tolerance in Arabidopsis conferred by chimeric repressor gene silencing technology. Plant Signal Behav. 2014;9(3):e28211.
8. Khare D, Mitsuda N, Lee S, Song WY, Hwang D, Ohme-Takagi M, Martinoia E, Lee Y, Hwang JU. Root avoidance of toxic metals requires the GeBP-LIKE 4 transcription factor in *Arabidopsis thaliana*. New Phytol. 2017;213(3):1257–73.
9. Chevalier F, Perazza D, Laporte Fdr L, Hénanff Gl, Hornitschek P, Bonneville J-M, Herzog M, Vachon G. GeBP and GeBP-Like proteins are noncanonical leucine-Zipper transcription factors that regulate cytokinin response in Arabidopsis. Plant Physiol. 2008;146(3):1142–54.
10. Liu S, Liu Y, Liu C, Zhang F, Wei J, Li B. Genome-wide characterization and expression analysis of *GeBP* family genes in soybean. Plants. 2022;11(14):1848.
11. Umemoto N, Kakitani M, Iwamatsu A, Yoshikawa M, Yamaoka N, Ishida I. The structure and function of a soybean beta-glucan-elicitor-binding protein. Proc Natl Acad Sci USA. 1997;94(3):1029–34.
12. Gan Y, Liu C, Yu H, Broun P. Integration of cytokinin and gibberellin signalling by Arabidopsis transcription factors *GIS*, *ZFP8* and *GIS2* in the regulation of epidermal cell fate. Development. 2007;134:2073–81.
13. Jasinski S, Piazza P, Craft J, Hay A, Woolley L, Rieu I, Phillips A, Hedden P, Tsiantis M. *KNOX* action in Arabidopsis is mediated by coordinate regulation of cytokinin and gibberellin activities. Curr Biol. 2005;15(17):1560–5.
14. Noh B, Lee SH, Kim HJ, Yi G, Shin EA, Lee M, Jung KJ, Doyle MR, Amasino RM, Noh YS. Divergent roles of a pair of homologous jumonji/zinc-finger-class transcription factor proteins in the regulation of Arabidopsis flowering time. Plant Cell. 2004;16(10):2601–13.
15. Tsuda K, Tsuji T, Hirose S, Yamazaki K. Three Arabidopsis MBF1 homologs with distinct expression profiles play roles as transcriptional co-activators. Plant Cell Physiol. 2004;45(2):225–31.

16. Carles CC, Choffnes-Inada D, Reville K, Lertpiriyapong K, Fletcher JC. ULTRAPETALA1 encodes a SAND domain putative transcriptional regulator that controls shoot and floral meristem activity in Arabidopsis. Development. 2005;132(5):897–911.
17. Ray S, Dansana PK, Giri J, Deveshwar P, Arora R, Agarwal P, Khurana JP, Kapoor S, Tyagi AK. Modulation of transcription factor and metabolic pathway genes in response to water-deficit stress in rice. Funct Integr Genom. 2011;11:157–78.
18. Wang R, Wu X, Wang Z, Zhang X, Chen L, Duan Q, et al. Genome-wide identification and expression analysis of BrGeBP genes reveal their potential roles in Cold and Drought stress tolerance in Brassica rapa. Int J Mol Sci. 2023;24:13597.
19. Liu RX, Li HL, Qiao Z, Liu H-F, Zhao LL, Wang XF, et al. Genome-wide analysis of MdGeBP family and functional identification of MdGeBP3 in Malus domestica. Environ Exp Bot. 2023;208:105262.
20. Brininstool G, Kasili R, Simmons LA, Kirik V, Hülskamp M, Larkin JC. Constitutive Expressor of Pathogenesis-related Genes5affects cell wall biogenesis and trichome development. BMC Plant Biol. 2008;8(1):1.
21. Jing HC, Anderson L, Sturre MJG, Hille J, Dijkwel PP. *Arabidopsis CPR5* is a senescence-regulatory gene with pleiotropic functions as predicted by the evolutionary theory of senescence. J Exp Bot. 2007;58(14):3885–94.
22. Perazza D, Laporte F, Balagué C, Chevalier F, Remo S, Bourge M, Larkin J, Herzog M, Vachon G. GeBP/GPL transcription factors regulate a subset of *CPR5*-dependent processes. Plant Physiol. 2011;157(3):1232–42.
23. Wang M, Tu L, Yuan D, Zhu D, Shen C, Li J, Liu F, Pei L, Wang P, Zhao G, et al. Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. Nat Genet. 2019;51(2):224–9.
24. Hu Y, Chen J, Fang L, Zhang Z, Ma W, Niu Y, Ju L, Deng J, Zhao T, Lian J, et al. *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. Nat Genet. 2019;51(4):739–48.
25. Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, Ma Z, Shang H, Ma X, Wu J, et al. Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. Nat Biotechnol. 2015;33(5):524–30.
26. Li L, Hu Y, Wang Y, Zhao S, You Y, Liu R, Wang J, Yan M, Zhao F, Huang J, et al. Identification of novel candidate loci and genes for seed vigor-related traits in upland cotton (Gossypium hirsutum L.) via GWAS. Front Plant Sci. 2023;14:1254365.
27. Li L, Zhang C, Huang J, Liu Q, Wei H, Wang H, Liu G, Gu L, Yu S. Genomic analyses reveal the genetic basis of early maturity and identification of loci and candidate genes in upland cotton (*Gossypium hirsutum* L). Plant Biotechnol J. 2020;19(1):109–23.
28. Li H, Tang Z, Li J, Liu L, Zhou W. Bioinformatics analysis of the Arabidopsis GeBP transcription factor Gene Family. Mol Plant Breed. 2023:1–21. https://kns.cnki.net/kcms/detail/46.1068.S.20230915.1630.003.html.
29. Yu J, Jung S, Cheng CH, Ficklin SP, Lee T, Zheng P, Jones D, Percy RG, Main D. CottonGen: a genomics, genetics and breeding database for cotton research. Nucleic Acids Res. 2014;42(Database issue):D1229-1236.
30. Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, Madden TL, Matten WT, McGinnis SD, Merezhuk Y. BLAST: a more efficient report with usability improvements. Nucleic Acids Res. 2013;41(W1):W29-33.
31. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011;39(suppl2):W29-37.
32. Li B, Zhou G, Li Y, Chen X, Yang H, Li Y, Zhu M, Li L. Genome-wide identification of *R-SNARE* gene family in upland cotton and function analysis of *GhVAMP72l* response to drought stress. Front Plant Sci. 2023;14:1147932.
33. Katoh K, Misawa K, Kuma KI, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30(14):3059–66.
34. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74.
35. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. Ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8(1):28–36.
36. 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.
37. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genom Proteom Bioinform. 2010;8(1):77–80.
38. Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 2000;17(1):32–43.
39. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015;10(6):845–58.
40. Su J, Li L, Pang C, Wei H, Wang C, Song M, Wang H, Zhao S, Zhang C, Mao G. Two genomic regions associated with fiber quality traits in Chinese upland cotton under apparent breeding selection. Sci Rep. 2016;6(1):38496.
41. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44(7):821–4.
42. Wickham H. ggplot2. Wiley Interdisciplinary Reviews: Comput Stat. 2011;3(2):180–5.
43. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
44. Ma J, Jiang Y, Pei W, Wu M, Ma Q, Liu J, Song J, Jia B, Liu S, Wu J, et al. Expressed genes and their new alleles identification during fibre elongation reveal the genetic factors underlying improvements of fibre length in cotton. Plant Biotechnol J. 2022;20(10):1940–55.
45. Bao Y, Hu G, Grover CE, Conover J, Yuan D, Wendel JF. Unraveling *cis* and *trans* regulatory evolution during cotton domestication. Nat Commun. 2019;10(1):5399.
46. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417–9.
47. Huang G, Wu Z, Percy RG, Bai M, Li Y, Frelichowski JE, Hu J, Wang K, Yu JZ, Zhu Y. Genome sequence of *Gossypium herbaceum* and genome updates of *Gossypium arboreum* and *Gossypium hirsutum* provide insights into cotton A-genome evolution. Nat Genet. 2020;52(5):516–24.
48. Kolde R, Kolde MR. Package 'pheatmap'. R Package. 2015;1(7):790.
49. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. Methods (San Diego Calif). 2001;25(4):402–8.
50. Qin YM, Zhu YX. How cotton fibers elongate: a tale of linear cell-growth mode. Curr Opin Plant Biol. 2011;14(1):106–11.
51. Wen YZ, He P, Bai XH, Zhang HZ, Zhang YF, Yu JN. Strigolactones modulate cotton fiber elongation and secondary cell wall thickening. J Integr Agric. 2024;23:1850–63.
52. Huang J, Zhang Q, He Y, Liu W, Xu Y, Liu K, Xian F, Li J, Hu J. Genome-Wide Identification, expansion mechanism and expression profiling analysis of GLABROUS1 enhancer-binding protein (GeBP) Gene Family in Gramineae Crops. Int J Mol Sci. 2021;22:8758.
53. Zhao H, Chen Y, Liu J, Wang Z, Li F, Ge X. Recent advances and future perspectives in early-maturing cotton research. New Phytol. 2023;237(4):1100–14.
54. Dou L, Zhang X, Pang C, Song M, Wei H, Fan S, Yu S. Genome-wide analysis of the *WRKY* gene family in cotton. Mol Genet Genomics: MGG. 2014;289(6):1103–21.
55. Dai Y, Liu S, Zuo D, Wang Q, Lv L, Zhang Y, Cheng H, Yu JZ, Song G. Identification of *MYB* gene family and functional analysis of *GhMYB4* in cotton (*Gossypium* spp). Mol Genet Genomics: MGG. 2023;298(3):755–66.
56. Su J, Zhan N, Cheng X, Song S, Dong T, Ge X, Duan H. Genome-wide analysis of cotton *MYB* transcription factors and the functional validation of *GhMYB* in Response to Drought stress. Plant Cell Physiol. 2024;65(1):79–94.
57. Chao M, Dong J, Hu G, Li Y, Huang L, Zhang J, Tang J, Wang Q. Phylogeny, gene structures, and expression patterns of the auxin response factor (*GhARF2*) in upland cotton (*Gossypium hirsutum* L). Mol Biol Rep. 2023;50(2):1089–99.
58. Hurst LD. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet. 2002;18(9):486.

Publisher's Note