Plant Biotechnology Journal
Open Access

Plant Biotechnology Journal (2024) **22**, pp. 2756–2772 doi: 10.1111/pbi.14402

Regulation of a single inositol 1-phosphate synthase homeologue by HSFA6B contributes to fibre yield maintenance under drought conditions in upland cotton

Li'ang Yu[1], Anna C. Nelson Dittrich[1,†], Xiaodan Zhang[1,†], Jordan R. Brock[2,†], Venkatesh P. Thirumalaikumar[1,a], Giovanni Melandri[3], Aleksandra Skirycz[1,b], Patrick P. Edger[2,†], Kelly R. Thorp[4], Lori Hinze[5], Duke Pauli[3,6,*] and Andrew D.L. Nelson[1,*]

[1] *Boyce Thompson Institute, Cornell University, Ithaca, NY, USA*
[2] *Department of Horticulture, Michigan State University, East Lansing, MI, USA*
[3] *School of Plant Sciences, University of Arizona, Tucson, AZ, USA*
[4] *United States Department of Agriculture-Agricultural Research Service, Arid Land Agricultural Research Center, Maricopa, AZ, USA*
[5] *United States Department of Agriculture-Agricultural Research Service, Southern Plains Agricultural Research Center, College Station, TX, USA*
[6] *Agroecosystem Research in the Desert (ARID), University of Arizona, Tucson, AZ, USA*

Received 20 February 2024;
revised 15 May 2024;
accepted 21 May 2024.

**Correspondence* (Tel 979-220-0641; email an425@cornell.edu (A.D.L.N.); Tel 520-621-3656; email dukepauli@arizona.edu (D.P.))
[a] Present address: Purdue Proteomics Facility, Bindley biosciences, Purdue University, West Lafayette, IN, USA
[b] Present address: Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA
[†] These authors contributed equally.

Keywords: metabolomics, transcriptomics, phenomics, gene regulatory network, cotton fibre yield, drought response.

Summary

Drought stress substantially impacts crop physiology resulting in alteration of growth and productivity. Understanding the genetic and molecular crosstalk between stress responses and agronomically important traits such as fibre yield is particularly complicated in the allopolyploid species, upland cotton (*Gossypium hirsutum*), due to reduced sequence variability between A and D subgenomes. To better understand how drought stress impacts yield, the transcriptomes of 22 genetically and phenotypically diverse upland cotton accessions grown under well-watered and water-limited conditions in the Arizona low desert were sequenced. Gene co-expression analyses were performed, uncovering a group of stress response genes, in particular transcription factors GhDREB2A-A and GhHSFA6B-D, associated with improved yield under water-limited conditions in an ABA-independent manner. DNA affinity purification sequencing (DAP-seq), as well as public cistrome data from Arabidopsis, were used to identify targets of these two TFs. Among these targets were two lint yield-associated genes previously identified through genome-wide association studies (GWAS)-based approaches, *GhABP-D* and *GhIPS1-A*. Biochemical and phylogenetic approaches were used to determine that *GhIPS1-A* is positively regulated by GhHSFA6B-D, and that this regulatory mechanism is specific to *Gossypium* spp. containing the A (old world) genome. Finally, an SNP was identified within the GhHSFA6B-D binding site in *GhIPS1-A* that is positively associated with yield under water-limiting conditions. These data lay out a regulatory connection between abiotic stress and fibre yield in cotton that appears conserved in other systems such as Arabidopsis.

Introduction

Upland cotton (*Gossypium hirsutum* L.) is the world's top renewable textile fibre, supporting a multibillion-dollar industry with a global production of 120.2 million bales of cotton (~26 million metric tonnes). It is a major economically important crop for the U.S. and for Arizona, where upland cotton is planted on ~50 000 ha, mainly in the semi-arid environment of the low desert using surface irrigation to complement limited precipitation. Cotton productivity in semi-arid areas of the Southwestern U.S. is severely threatened by global climate change. Increasing climatic variability is responsible for hotter summers, with day and night temperatures far above the thermal optimum (30/22 °C) for the crop, and lower and erratic rainfall patterns which expose the crop to an increasing risk of drought (Alizadeh *et al*., 2020). Therefore, revealing the physio-genetics mechanisms that regulate cotton's response to arid conditions is of primary interest. Specifically, this information can be leveraged for the development of new elite cotton cultivars with improved adaptation to hotter and drier climatic conditions that are predicted in the near future.

In addition to being a critical fibre crop, cotton serves as an excellent model polyploid system for studying the impacts that interspecific hybridization has had on agronomic traits. The cotton species predominantly cultivated for fibre production, *G. hirsutum* (upland cotton) and *Gossypium barbadense* (Pima cotton), are new world allotetraploids believed to have formed ~1–2 million years ago from a transoceanic hybridization of an A genome diploid originating from Africa or Asia (e.g. *Gossypium arboreum*, tree cotton) and a D genome diploid from Central or South America (e.g. *Gossypium raimondii*; Chen *et al*., 2007; Wang *et al*., 2012). This unique combination of homeologous gene pairs in the allotetraploids resulted in superior fibre yield and quality over diploid progenitors that have since undergone additional selection in both *G. hirsutum* and *G. barbadense*. Despite high levels of sequence

© 2024 The Author(s). *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

conservation and collinearity between these two species and their diploid progenitors, an altered epigenetic landscape, as well as homeologue expression divergence, have contributed to *G. hirsutum*'s capacity to maintain lint yield under a wide range of environments (Chen *et al.*, 2020; Li *et al.*, 2021; Pan *et al.*, 2020; Peng *et al.*, 2022).

Like other crops, drought tolerance in cotton involves complex signalling pathways and transcriptional networks orchestrated by a number of transcription factors (TFs) and signalling proteins (Mahmood *et al.*, 2019; Shinozaki and Yamaguchi-Shinozaki, 2007; Takahashi *et al.*, 2020). These regulatory proteins are typically upstream of, or transcriptionally interconnected with, the genes necessary for coping with stress and maintaining crucial metabolic pathways. Examples of typical regulatory targets include genes encoding enzymes related to the production of protective metabolites, transporters, chaperones and lipid biosynthesis proteins (Gupta *et al.*, 2020; Malhotra and Sowdhamini, 2014; Singh and Laxmi, 2015). In cotton, multiple transcription factor families, such as *GhNACs*, *GhDREBs*, *GhERFs* and *GhWRKYs*, have been associated with the drought stress response (Chu *et al.*, 2015; Huang *et al.*, 2009, 2013; Ma *et al.*, 2017). Additionally, genome-wide association studies (GWAS)-based approaches have identified numerous candidate genes related to lint yield, fibre quality and other agronomically important traits (Fang *et al.*, 2014; Ma *et al.*, 2018; Sun *et al.*, 2021). Indeed, while germplasm exists with the ability to maintain fibre growth and quality under heat and drought conditions, little is known about how these traits arose in the cotton genome and the regulatory factors that connect them.

In this study, transcriptome sequencing was performed for 22 upland cotton accessions grown in the Arizona low desert and exposed to both well-watered and water-limited conditions. Phenotypic and metabolomic data were integrated into a co-expression network analysis to identify genes associated with improved yield under water-limited conditions. DAP sequencing was performed on two transcription factors, *GhDREB2A-A* and *GhHSFA6B-D*, shown to have the highest association with lint yield and stress response genes across the panel. Using transcriptomic, biochemical and phylogenomic approaches, *GhHSFA6B-D* binds to and positively regulates the lint yield-associated gene, *GhIPS1-A*, and the associated *GhHSFA6B-D* regulatory element is only present in the A subgenome and A genome diploid progenitors. We also identify a lint yield-associated single nucleotide polymorphism directly adjacent to this *GhIPS1-A* regulatory element that influences GhHSFA6B-D binding and appears to suggest additional selection at this locus during domestication.

Results and discussion

Genetic profiles of upland cotton panels in field experiment

To better understand the molecular mechanisms underlying cotton's performance under drought conditions, 22 diverse upland cotton accessions from the Gossypium Diversity Reference Set (Hinze *et al.*, 2015, 2016) were evaluated under water-limited (WL) and well-watered (WW) conditions in the field. These accessions were selected for reported variation in heat and drought tolerance as well as fibre qualities (Table S1). Leaf tissue from plants at the flowering and boll development stage was collected (two replicates each, ~100 DAS; Figure 1a). On average, ~27 million PE × 150 bp reads were generated for each replicate. Reads were then aligned using the RMTA pipeline (Peri *et al.*, 2020) to the *G. hirsutum* v2.0 reference genome (Chen *et al.*, 2020); (Table S1). Gene-level expression data were counted using *G. hirsutum* gene-level annotations as meta-features in FeatureCounts (Liao *et al.*, 2014; Table S1, parameter: -t gene, -g ID). The Pearson correlation coefficient (PCC) of normalized read counts revealed high correlation among replicates (average PCC = 0.904), except for accession 'Tipo Chaco' (PCC = 0.61, cutoff: 0.8) (Table S2). Thus, this accession was discarded from further analyses, resulting in 21 accessions for further analyses.

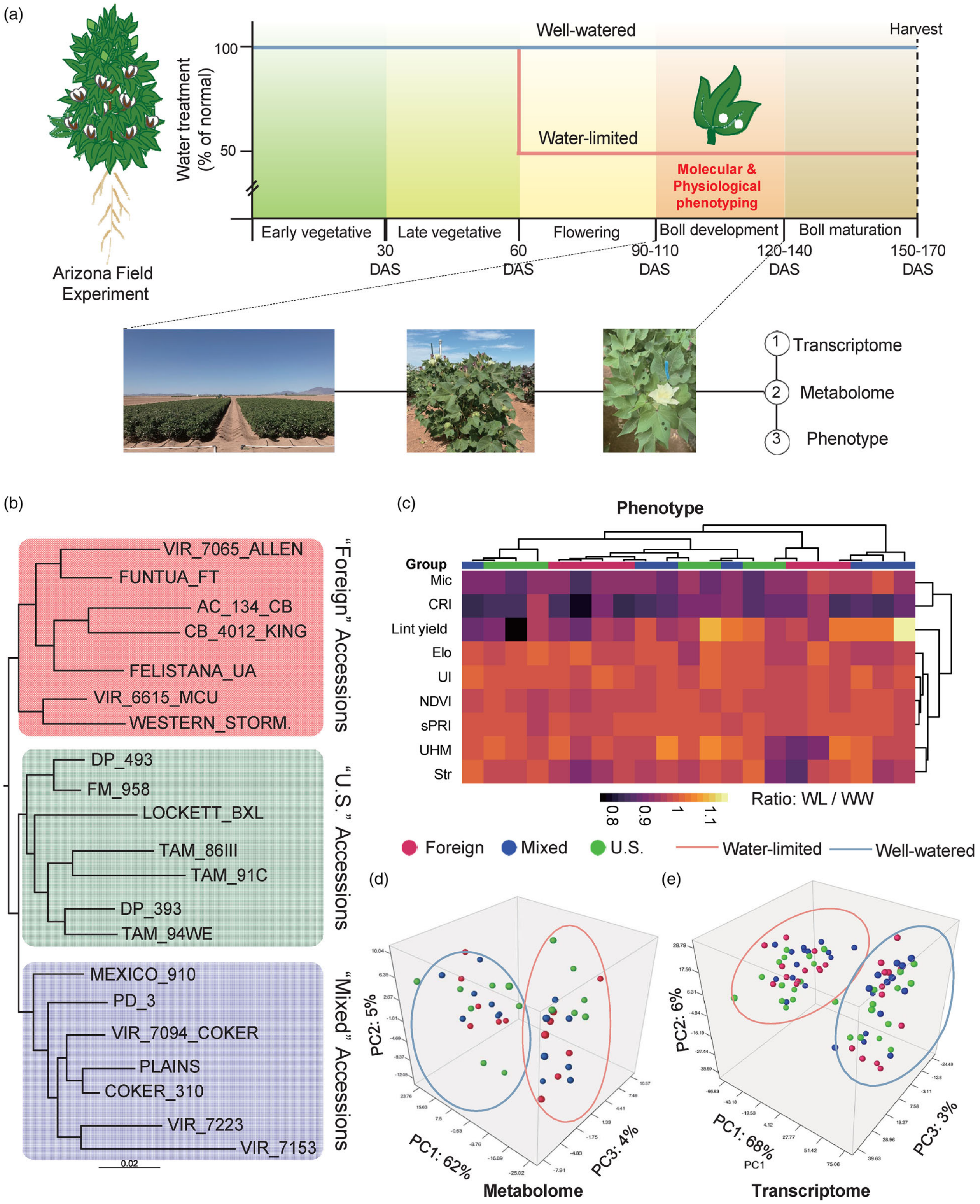
To determine the relatedness of each of these accessions, these RNA-seq data were used to perform variant calling relative to the reference genome (108 396 bi-allelic SNPs after filtering). These data were used to reconstruct a phylogeny of the 21 accessions, rooted by VIR_7153_D_10, which resulted in three major subgroups (Figure 1b). Based on information from USDA GRIN-Global, the first of these three subgroups is referred to as 'Foreign', as this group contains accessions developed primarily outside of the U.S. (Figure 1b and Table S1). The second subgroup consists of 'U.S.' accessions of breeding lines and commercial cultivars developed in the U.S. (Figure 1b and Table S1). The last group, 'Mixed', consists of commercial cultivars from the U.S. and improved breeding lines generally developed outside the U.S. (Figure 1b and Table S1). Thus, these 21 accessions serve as a genetically diverse framework to investigate the regulatory mechanisms controlling the physiological responses to drought in cotton.

Variation in physiological and molecular responses to drought in major subgroups

The phylogenetic groupings were used to determine the degree to which genotype and environment influenced physiological

Figure 1 Experiment information and genomic, metabolomic and transcriptomic profiling of the cotton panel used in this work. (a) Overview of the watering regime and timing of data collection for the experiment in the 2019 summer field season at Maricopa, AZ. Water levels were reduced to 50% of normal for the treatment plots at the early flowering stage. Metabolomic and transcriptomic data were collected at late flowering/early boll development, with phenotypic measurements taken throughout (See Materials and methods). (b) A phylogeny of the accessions used in this study based on filtered SNPs derived from the transcriptomic data. The three groups, 'Foreign', 'U.S.' and 'Mixed', reflect historical breeding information obtained from USDA GRIN-Global. (c) Overview of the phenotype profiles of 21 upland cotton accessions under water-limited (WL) and well-watered (WW) conditions: A cluster heatmap indicated the ratio of phenotypic data under WL condition relative to the WW condition (WL value / WW value, range: 0.8–1.2). Group colouring denotes accessions belonging to the three phylogenetic groups from 1b. (d) Three-dimensional PCA displaying the impact that treatment has on large-scale changes in normalized metabolite content between accessions. The different treatments have been denoted with light red (water-limited) and light blue (well-watered) ovals. (e) Three-dimensional PCA was used to display the top 10% (N = 3768) most variably expressed genes based on normalized read counts, with treatment groups denoted similarly to 1d.

14677652, 2024, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pbi.14402, Wiley Online Library on [27/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



responses to drought in the assembled panel. The effects of genotype (G: Foreign, U.S. and Mixed), environment (E: WL and WW) and G*E interaction on six fibre quality and agronomic traits and four vegetation indices were examined using two-way ANOVA (Table S3). Among these traits, three of the four vegetation indices were significantly affected by drought (Table S3; two-way ANOVA cut-off: $P < 0.05$, Figure S1). The treatment effect was particularly significant in the 'Foreign' subgroup (student *t*-test cutoff: $P < 0.05$). By contrast, long-term drought stress brought limited impacts to most fibre quality traits apart from the significantly reduced micronaire (two-way ANOVA: $P = 0.002$) observed in 'Mixed' and 'Foreign' subgroups (*t*-test: $P = 0.015$ and 0.047,

14677652, 2024, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pbi.14402, Wiley Online Library on [27/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

respectively, Figure S1, Table S3). In contrast to treatment effects, pronounced genotypic effects were observed between major subgroups ('Foreign, U.S. and Mixed') for five of the six fibre quality traits but not for the four vegetation indices (Table S3).

However, intragroup comparisons of responses to drought revealed stark differences. For instance, each subgroup contained one or more accessions that significantly outperformed close relatives when using a ratio of phenotypes from WL relative to WW condition as an indicator, particularly in fibre quality traits (Figure 1c). The lack of phylogenetic congruence in observed traits suggests that lint yield and drought-associated traits may have been selected differently in even closely related accessions. Importantly, the presence of outperforming accessions implies that this panel may be useful in identifying the factors associated with lint yield and drought stress.

We analysed metabolite profiles as molecular phenotypes for stress response to understand the effects of drought stress. Changes in the metabolome of the panel were measured by re-examining 451 metabolites previously analysed in Melandri *et al*. (2021) (27 GC–MS and 424 LC–MS/MS). A global profiling of the 451 metabolites revealed that the strongest effects were due to drought treatment (PC1, 62%; Figure 1d). The second and third principal components explained a further 5.0% and 3.4%, respectively, but the phylogenetic relationships (i.e. genotype) were not clearly separated along either of these axes. A large treatment effect was also observed in the metabolite profiles using a two-way ANOVA (Table S4). Drought treatment significantly impacted more than 95% of metabolites, but only 5% of the metabolites exhibited genotypic effects; around 25% of metabolites exhibited changes that could be associated with G*E effects ($P < 0.05$, Table S4). Among the 432 metabolites associated with significant treatment effects, 273 were up-regulated and 159 down-regulated by drought. Among these metabolites are the known glycine and proline (Fang *et al*. (2015); Table S4). Thus, these data demonstrate that alteration of specific metabolites associated with drought tolerance (osmoprotectants) is a common response to water-deficit stress conditions across our panel.

To test the basis for metabolic changes, we examined transcriptome profiles among these accessions in response to drought stress. Consistent with the observed metabolomic changes, 68% of the variation within the panel could be explained by the irrigation treatment (Figure 1e). In addition, neither PC 2 nor PC 3 could be attributed to genotypic differences. We then conducted pairwise comparisons of gene expression under the two conditions to identify differentially expressed genes (DEGs) for each accession (Table S5). We compared transcriptomes to determine if DEGs were shared between all intragroup accessions, or between all accessions within multiple subgroups in response to drought. Of the thousands of DEGs in each accession, there were few genes that were shared among all accessions within a subgroup, or between subgroups (Figure S2A, Table S5). The foreign accessions shared the least group-wide DEGs (up-regulated: 57 and down-regulated: 86), whereas the mixed accessions featured the most (up-regulated: 194 and down-regulated: 177) (Figure S2B). In contrast, there were 300 DEGs (194 upregulated and 106 downregulated) shared among all examined accessions (Table S6). Gene ontology (GO) enrichment of these shared up-regulated genes revealed an over-representation of the stress response and phosphorylation signal transduction genes while the shared down-regulated genes were involved in fatty acid biosynthetic processes and transport processes (Figure S3B). Thus, these shared DEGs may represent a common set of genes involved in drought stress response.

Upland cotton has a well-reported subgenome expression bias towards the D (new world) subgenome when comparing homeologous gene pairs across a wide range of tissues (Chen *et al*., 2020). Subgenome expression dominance is believed to be influenced by changes in the environment (Bird *et al*., 2018), thus, we next examined how subgenome expression dominance was impacted at a global scale (comparing all expressed genes in A and D subgenomes) across our panel in response to water-limiting conditions. After removal of genes with low expression (average TPM across samples < 1; 75 376 genes retained), we first compared expression dominance under well-watered and water-limited conditions separately. Under well-watered conditions, 10/21 accessions displayed a significant bias in mean gene expression towards the A subgenome (pair-wise *t*-test, $P < 0.05$; Figure S3), with the remainder not showing any bias. In contrast, under water-limiting conditions, the bias towards the A subgenome became more pronounced (17/21 accessions). Interestingly, three accessions with a bias under well-watered conditions lost that bias under water-limiting conditions: Western Stormproof, DP 393 and Mexico 910. Each of these three accessions has reported tolerance to hot and dry conditions, although they are not the only accessions with this trait in our panel (Table S1). Thus, this global analysis suggests that genes from the old world subgenome are predominantly expressed under hot (well-watered) and hot/dry (water-limited) conditions.

Using WGCNA to examine transcriptome–trait connections in response to drought

To determine a relationship between transcriptomic and phenotypic variability, we performed a weighted gene co-expression analysis (WGCNA) to uncover the association of gene networks with phenotypes. For the WGCNA, we incorporated the top 10% most variable genes across the 21 accessions under WL conditions (determined by median absolute deviation [MAD], $N = 4432$ genes). We obtained 22 modules that satisfied a scale-free topology ($R^2 = 0.86$; Figure S4A, Table S8) using a soft threshold (Beta = 7) for network construction (Figure S4B). These 22 modules displayed clear separation, with only 45 genes (<0.2%) unclassified (Figure S4C). Using PCC, these 22 well-clustered modules were then correlated with the trait dataset, which consisted of 10 phenotypes and 451 metabolites. Among these traits being tested, 5 traits (lint percentage, referred to here as lint yield, UI, sPRI, CRI and WI/NDVI) and 9 metabolites exhibited significant correlation to 18 modules (Figure 2, $P < 0.05$). In particular, lint yield displayed the highest positive correlation with the turquoise module ($r = 0.68$, $P = 7e\text{-}04$), whereas terpenoid and carbohydrate metabolites displayed a positive correlation with the salmon module ($P < 0.05$; Figure 2).

Identifying trait-associated functional modules using multiple data integration

Following module–trait correlation, we selected key module(s) for further functional analysis by incorporating the profile of GO enrichment, transcription factor (TF)-binding motif enrichment and gene expression variance (Figure 2a). Using Fisher test ($P < 0.01$) for GO enrichment, 14 of the 18 modules exhibited some degree of enrichment for genes involved in biological processes of interest. Notably, the lint yield-associated module contained a high degree of enriched stress response genes (*Q*-

14677652, 2024, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pbi.14402, Wiley Online Library on [27/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

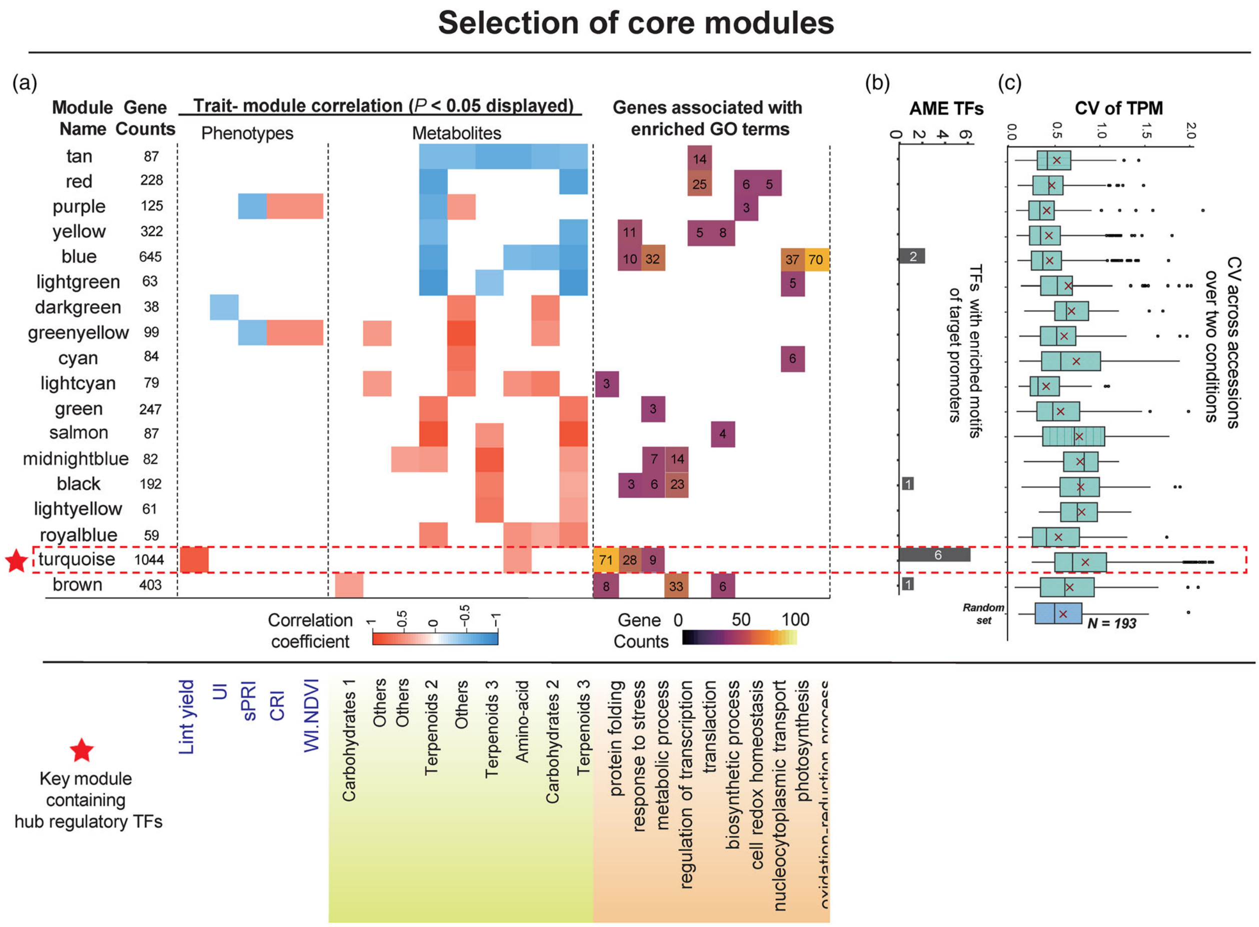


Figure 2 Trait and co-expression association in response to water-deficit conditions. (a) Modules of genes derived from a weighted gene co-expression analysis were correlated with phenotypic and metabolomic traits, with traits passing the significance threshold (*P* < 0.05) shown along the bottom. Modules are named according to colour, with the number of genes within each module shown (gene counts). Positive (red hues) and negative (blue hues) correlation values are shown for all significant trait–module correlations with scale bar below. GO term enrichment was performed on all genes within the module, with significant terms shown along the bottom. The number of genes associated with each GO term is shown within each box along with the 'purple' to 'yellow' scale displayed. (b) The number of TFs with enriched binding motifs within their respective module (TF and motifs are present in the same module) is shown. (c) Coefficient of variance (CV) of expression (TPM) was calculated for all genes within each module across all accessions and both conditions. CV was also calculated for a background set of genes based on an average module size (*N* = 193). The red dashed box outlines the 'turquoise' module, which was found to be correlated with both lint yield and stress response genes and was selected as a key module (highlighted with a red star) for further examination.

value: 1.6e^{-12}) as well as protein folding genes (*Q*-value: 4.6e^{-8}). Also, a large number of photosynthesis-related and redox-related genes (*Q*-value: 2.3e^{-10} and *Q*-value: 8.4e^{-4}, respectively) were identified within the blue module, negatively correlated with the abundance of multiple metabolites (Figure 2a). These data suggest that a co-expression network-based approach may uncover key genes integrating plant development and yield in response to water limitation.

To further investigate the upstream regulators of those genes associated with enriched GO terms, we took the 2-kb upstream promoter region sequences of all genes (WGCNA weight score > 0.1) in each of the 18 modules to perform motif enrichment analysis using analysis of motif enrichment (AME; McLeay and Bailey, 2010). To narrow down the list of possible enriched TFs/motifs, we focused on TF/motif pairs where the TF was also present in the same module. This approach uncovered four modules with enriched motifs corresponding to 10 different transcription factors, with the lint yield module (turquoise) containing the highest numbers of enriched TFs (*n* = 6, AME TFs, Figure 2b, Table S8). Under the hypothesis that significant trait-associated modules might display a more pronounced response to treatment, we assessed expression variation for genes within each module between accessions and conditions by calculating the coefficient of variance (CV). In support of our hypothesis, the turquoise module displayed the most significant (*P* = 0.0033) increase in CV relative to the background gene set (selected by random sampling: *N* = 193; Figure 2c). In sum, these data indicate that the genes within the 'turquoise' module (highlighted with a star in Figure 2a), and their associated TFs, may be critical for maintaining yield under water-limiting conditions.

DAP-seq of HSFA6B and DREB2A revealed the module-specific upregulation of their targets in response to drought

Among six TFs in the turquoise module that are associated with yield, four were AME TFs that are predicted to be associated with heat or drought stress based on homology with functionally

14677652, 2024, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pbi.14402, Wiley Online Library on [27/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

14677652, 2024, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pbi.14402, Wiley Online Library on [27/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

characterized TFs in *Arabidopsis* and cotton (Bian *et al*., 2020; Chen *et al*., 2017; Huang *et al*., 2016; Jacob *et al*., 2017; Kolmos *et al*., 2014; Nakashima *et al*., 2014; Weltmeier *et al*., 2009; Figure 3a). We further compared the gene from the enriched GO/KEGG terms (N = 119) in this module, as well as seven genes identified in previous GWAS, with lint yield for the presence of TF motifs (Fang *et al*., 2017; Su *et al*., 2016; Sun *et al*., 2021; Table S9). These four TFs, HSF7, HSF6, HSFA6B and DREB2A (Figure 3a, upper panel), are predicted to bind to a number of stress and heat response genes (Figure 3a, bottom panel). Notably, only HSFA6B and DREB2A are predicted to bind to the seven lint yield-associated genes in the module (Figure 3a, boxed, Table S10). This suggests that the cotton DREB2A and HSFA6B homologues are likely candidates for the observed association between water stress and fibre yield.

Given their association with fibre yield and drought stress, these two TFs likely regulate downstream stress-responsive genes. To test this, we performed DNA affinity purification sequencing (DAP-seq) using cotton DREB2A and HSFA6B and examined the DNA-binding ability of the homoeologues from both subgenomes (e.g. GhDREB2A-A and GhDREB2A-D). More than 40 million (DREB2A) and 37 million (HSFA6B) single-end reads (SE: 150 bp) per replicate enabled the identification of more than 20 000 peaks with high reproducibility for each sample after peak calling. For each TF, despite both being expressed in an *in vitro* wheat germ system (Figure S5), only a single homoeologue, namely GhHSFA6B-D, derived from the D subgenome (Gohir.D08G072600), and GhDREB2A-A, derived from the A subgenome (Gohir.A13G021700), could bind to DNA above background. Using irreproducibility discovery rate as a cutoff (IDR, $P < 0.05$), a total of 16 927 and 16 704 high confidence peaks were identified between replicates of GhDREB2A-A and GhHSFA6B-D (Figure 3b, Table S11). After filtering for GhHSFA6B-D and GhDREB2A-A peaks in the 5 kb upstream region of annotated coding genes (distal promoter), as well as peaks within the 5′ untranslated regions (UTRs) of coding genes (proximal promoter), a total of 5229 and 3178 genes were identified as likely regulatory targets of GhDREB2A-A and GhHSFA6B-D, respectively (Figure 3b, Tables S12 and S13). Genomic sequences associated with these peaks were further processed by motif analysis (MEME suite; Bailey *et al*., 2015) to identify binding sites and test the levels of conservation of consensus motifs. As expected, both target sequences of the two TFs revealed high-level conservation (E-value: 3.8^{e-855} across 3180 peaks and 5.9^{e-603} among 5218 peaks; Figure 3b), compared to core HSFA6B- and DREB2A-binding elements (HSEs and DREs) from other species (Nishizawa *et al*., 2006; Sakuma *et al*., 2006; Scharf *et al*., 2012). Interestingly, although our AME TF approach only predicted a GhHSFA6B-D-binding motif in the lint yield-associated gene *GhIPS1-A* (myo-inositol-phosphate synthase protein), we observed DAP-seq peaks for GhHSFA6B-D binding to both *GhIPS-A* and *GhABP-A* (Figure S6). In addition, we observed GhHSFA6B-D peaks in the promoter region of both *GhDREB2A-A* and *GhDREB2A-D* homoeologues (Figure S6) but did not observe reciprocal GhDREB2A-A peaks in the promoter of *GhHSFA6B-D*, suggesting that GhHSFA6B-D may act as a regulator of *GhDREB2A-A*. Together, these data suggest that GhHSFA6B-D and GhDREB2A-A may regulate expression of stress-responsive pathway genes under drought in cotton.

To determine if GhHSFA6B-D and GhDREB2A-A might have a specific impact under stress on genes found within the lint yield module, we profiled transcriptome changes between treatments. A comparison of average $\log_2$FC (WL relative to WW) for in-module targets of GhHSFA6B-D or GhDREB2A-A revealed a significant up-regulation relative to a similar number of their out-of-module targets ($P < 0.001$, Figure 3c, Table S14). The analysis of GO enrichment ($P < 0.01$) for these downstream target genes found that only the genes bound by GhHSFA6B-D in the module were enriched for stress response terms ($-\log_{10}$(Q-value) > 8), whereas both GhHSFA6B-D and GhDREB2A-A exhibited in-module specificity to chaperone/protein folding-related genes (Figure 3d). These data suggest that while both TFs display a high degree of regulatory connections between fibre yield and stress responses, GhHSFA6B-D may be a major regulator.

To better visualize the interactions between these two TFs and other genes within the lint yield module, including each other, we synthesized our data into a regulatory network using cytoscape (Figure 4a). To simplify this network, 866 genes in this module were reduced to a subset of 126 core genes by selecting genes with enriched GO terms, KEGG pathways, annotated as TFs or lint yield-associated genes, and those highly correlated with lint yield (trait-correlated genes; Figure 4a, see key). Trait-correlated genes are those with high module membership (MM), which is derived from the correlation between gene expression and the eigenvalue (first principal component) of the lint yield module, and gene significance (GS), a term reflecting the correlation between gene expression and lint yield (Figure 4b, orange circles). These target genes are those whose expression change across the 21 accessions is most likely to explain variation in lint yield in this panel. To further develop this network, we integrated the co-expression information derived from WGCNA, TF-target binding from our DAP-seq data and protein–protein interactions (PPIs) from the STRING database ((Szklarczyk *et al*., 2017); Figure 4b).

This network illustrates the high degree of connectivity between *GhHSFA6B-D* and the trait-correlated genes, both in terms of expression and direct connection (Figure 4b). Of the 24 trait-correlated genes, GhHSFA6B-D is predicted to bind 16 target genes based on DAP-seq analysis (Figure 4b and Table S10), whereas GhDREB2A-A is predicted to only bind two genes based on AME motif enrichment. As mentioned above, GhHSFA6B-D binds to both *GhDREB2A-A* and *GhDREB2A-D* homoeologues based on DAP-seq data. In addition, GhHSFA6B-D was found to bind to its own distal promoter region, suggesting it may act in an autoregulatory loop (Figure 4b and Table S10). Finally, despite AME predicting interactions between GhHSFA6B-D, GhDREB2A-A and the five lint yield-associated genes, only two DAP-seq-derived peaks between GhHSFA6B-D and *GhIPS1-A* and *GhABP-D* were observed. As the initial AME TF predictions were made based on DAP-seq data from *Arabidopsis* (O'Malley *et al*., 2016), it is not unexpected that TF binding preferences will have shifted slightly for these TFs in cotton, highlighting the importance of our TF binding validation.

DREB2A and HSFA6B affect the expression of both homologous *ABP* loci

The inferred gene regulatory network, developed from multiple lines of evidence, suggests a direct regulatory interaction among HSFA6B, DREB2A and two of the five lint yield-associated genes, *GhABP* and *GhIPS*. *GhABP-A* (Gohir.A12G006700) is an auxin-binding protein involved in cell elongation and cell division that was previously identified as a lint yield QTL (Zhu *et al*., 2020). TF-binding motif predictions based on the Arabidopsis Cistrome Database (O'Malley *et al*., 2016) identified DREB2A and HSFA6B

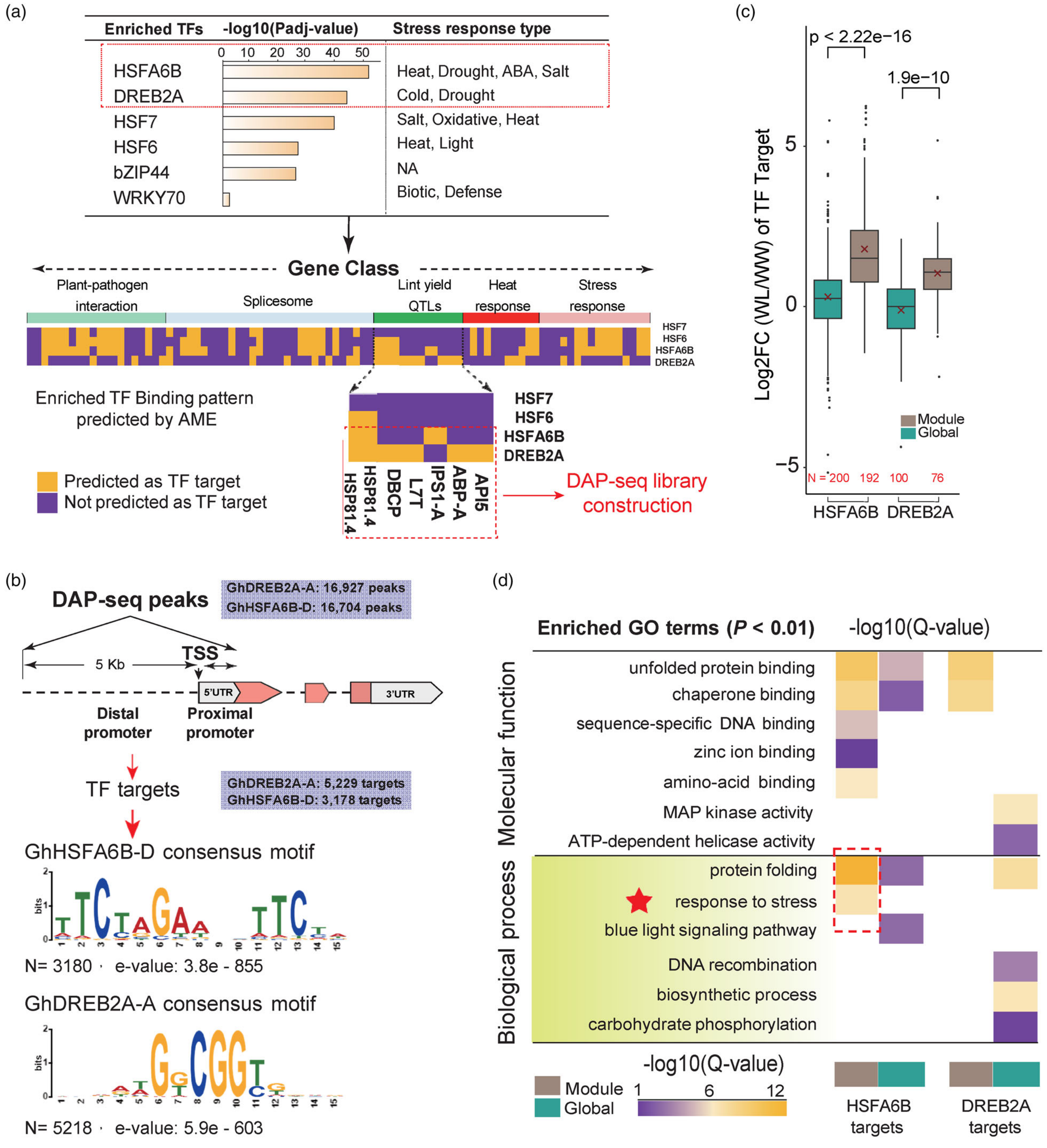


Figure 3 Identifying HSFA6B and DREB2A targets with DAP-seq. (a) Identity of the six key hub TFs within the lint yield-associated module (Top). For each TF, the $-\log_{10}$ (Adj-*P* value) level of motif enrichment among co-expressed genes within the module is shown, along with the reported stress response. For the four abiotic stress-associated TFs, their AME predicted binding to the genes with enriched GO terms within the module, as well as five GWAS-identified lint yield genes, is shown (bottom). As GhHSFA6B-D and GhDREB2A-A were both stress response regulators and predicted to bind to at least one of the five lint yield-associated genes, they were chosen for DAP-seq library creation (red dashed box). (b) Illustration of annotating peaks derived from GhHSFA6B-D and GhDREB2A-A DAP-seq data. The top panel highlights DAP-seq peaks within the 5-kb upstream regions (distal promoters) and 5′UTR regions (proximal promoters) of annotated genes. The bottom panel depicts the two consensus motifs identified from 3180 peaks (GhHSFA6B-D) and 5218 peaks (GhDREB2A-A) that passed stringency filters. (c) Comparison of the $\log_2$ fold change of transcript abundance (WL relative to WW) for GhDREB2A-A and GhHSFA6B-D DAP-seq targets between genes in the lint yield-associated module (192 GhHSFA6B-D targets and 76 GhDREB2A-A targets) and genes selected from transcriptome-wide through random sampling (*N* values shown below). Pairwise significance was performed using a paired Student's *t*-test. (d) Comparison of enriched GO terms among genes bound by GhHSFA6B-D and/or GhDREB2A-A within the lint yield module and targets from a transcriptome-wide group of expressed transcripts. The level of enrichment for each GO term is reflected by $-\log_{10}$ *Q*-value, with higher levels of enrichment corresponding to larger values. The 'star' logo highlights the 'stress response'-enriched GO term.

14677652, 2024, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pbi.14402, Wiley Online Library on [27/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

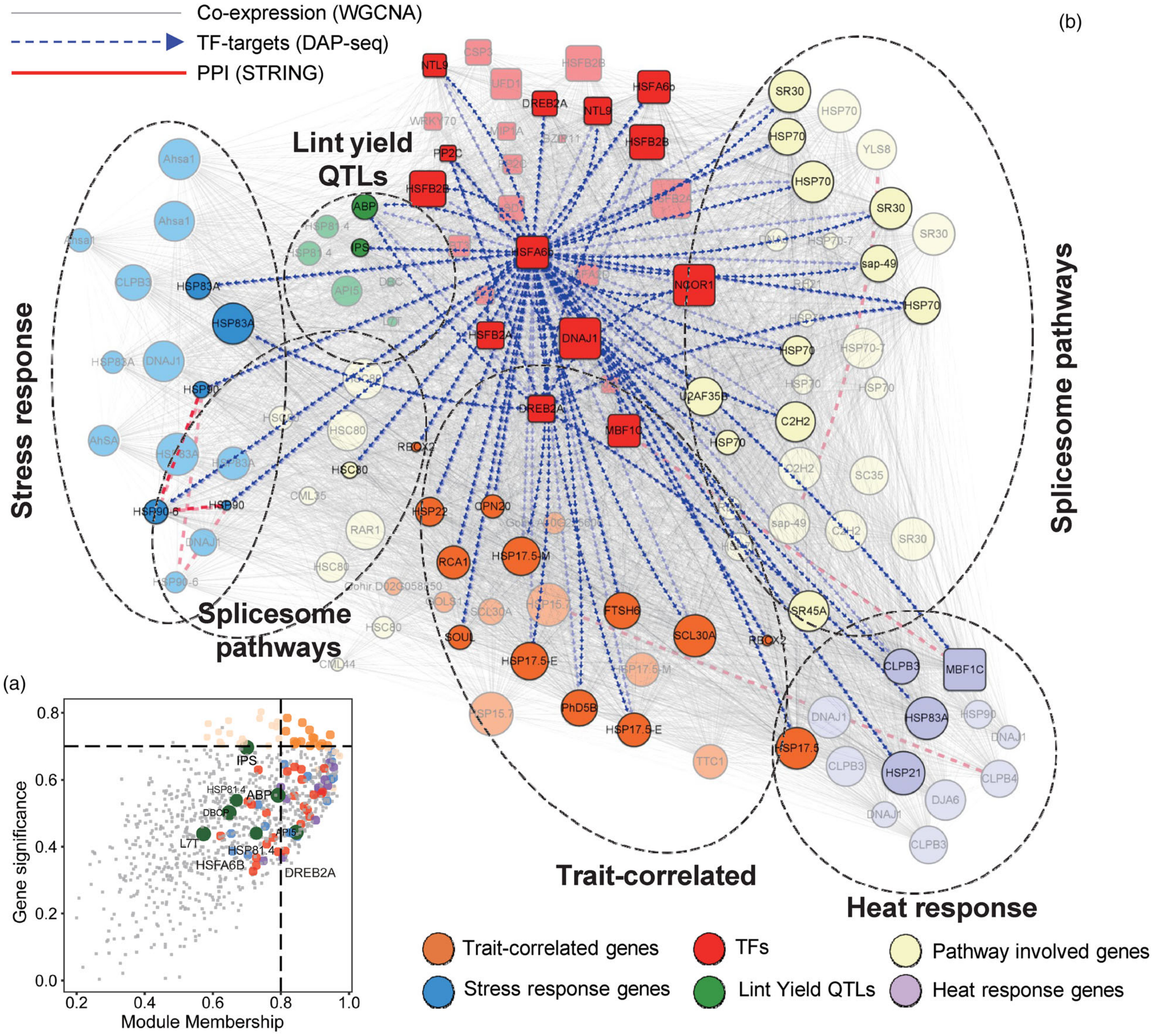


Figure 4 Integrated network display of a module of transcripts associated with lint yield. The module membership (MM) scores and gene significance (GS) for genes within the lint yield module were plotted with the linear model fitted line (a). The dashed black lines indicate the two cut-offs used to identify trait-related genes (orange dots, GS > 0.7, MM > 0.8). The integrated network derived from 133 of 866 selected genes within the lint yield-associated module was presented by layering multiple levels of information (b): gene categories are shown in the bottom line, as corresponded to the colour of highlighted dots in MM-GS plot, including genes associated with stress response (blue), TFs (red), heat response (light purple), spliceosome pathways, biotic response pathways and five lint yield QTLs (green), whereas connection types are shown in the top right. The solid 'grey' lines connected the co-expressed genes from transcriptional level, the dashed 'red' lines indicated protein–protein interactions derived from STRING database and the 'blue' dash-arrowed lines highlighted targets bound by DREB2A and HSFA6B, as supported by DAP-seq data.

motifs for both homeologous *GhABP* loci (*GhABP-D* and *GhABP-A*; yellow boxes, Figures S7 and S8A). *GhABP-A*, but not *GhABP-D* (Gohir.D12G006100; *GhABP-D*), was bound by both TFs based on our DAP-seq data (GhHSFA6B-D peak centre: −561 bp TSS and GhDREB2A-A peak centre: −1445 bp TSS, Figure S8A). *GhABP-D* was also absent from the lint yield module, and a comparison of expression versus lint yield across the 21 accessions revealed that *GhABP-D* was expressed at lower levels than *GhABP-A* (Figure S7). In addition, *GhABP-A* displayed a stronger correlation with the lint yield across the accessions. Despite the lower expression level of *GhABP-D* than *GhABP-A*, the abundance of both transcripts was positively correlated with both TFs (Figure S8C). A sequence comparison of the identified TF-binding regions between *GhABP-A* and *GhABP-D* revealed only minor changes near the GhDREB2A-A (two SNPs, Figure S8A, top) and within the HSFA6B motifs (four SNPs, Figure S8A, bottom). No changes were observed within the core dehydration-responsive elements (DREs) or heat-shock elements (HSEs) that these two TFs are known to bind to in other systems (Figure S8A, red boxes).

To determine if the observed SNPs are responsible for the altered transcript abundance and DAP-seq peaks arising from these two homologous loci, we performed an electrophoretic mobility shift assay (EMSA, Table S15), using ~50 bp probes that spanned the respective DRE and HSE motifs (Figure S8A, pink bars). Both recombinant *GhDREB2A-A* and *GhHSFA6B-D* bound to their respective labelled probes, causing a gel shift that was not

14677652, 2024, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pbi.14402, Wiley Online Library on [27/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

evident in the probe alone or probe + empty vector controls (Figure S8D,E). As expected, adding a large molar excess (200×) unlabelled probe to the reaction was sufficient to compete for protein binding. Interestingly, the probes corresponding to the *GhABP*-D homeologue were also able to compete for binding with GhDREB2A-A and GhHSFA6B-D (Figure S8D,E). These data suggest that the SNPs observed between the *GhABP*-A and *GhABP-D* distal promoter regions are insufficient to explain the specificity observed between these paralogues.

HSFA6B affects lint yield by modulating the expression of an inositol phosphate synthase (IPS)

The *IPS* gene, also known as MIPS in Arabidopsis, encodes for the myo-inositol-phosphate synthase protein (INO-1), which catalyses the rate-limiting step in the synthesis of myo-inositol-6-phosphate, a key source of phosphate in seed endosperm (Mitsuhashi *et al*., 2008). In addition, myo-inositol is a precursor of the osmoprotectants galactinol and raffinose and thus is critical in a number of abiotic and biotic stress responses (Vinson *et al*., 2020). In contrast to ABP, there are four IPS genes in upland cotton (Gohir.D03G043600 – *GhIPS1-D*, Gohir.A02G132300 – *GhIPS1-A*, Gohir.D11G224000 – *GhIPS2-D* and Gohir.A11G199700 – *GhIPS2-A*, Figure S9), two of which (*GhIPS1-D* and *GhIPS1-A*) show a positive correlation between RNA abundance and lint yield across our panel (Figure 5b). These two homeologous IPS genes fall in syntenic regions of the 'D' and 'A' subgenomes and are phylogenetically distinct from *GhIPS2-A* and *GhIPS2-D* (Figure 6a).

Despite a reported expression bias towards D subgenome homoeologues (Chen *et al*., 2020), only *GhIPS1-A* was predicted to contain a GhHSFA6B-D binding site based on DAP-seq data and the Arabidopsis Cistrome Database (Figure 5a, purple and yellow boxes, respectively). A pairwise comparison of the distal promoter regions of *GhIPS1-A* and *GhIPS1-D* revealed substantial polymorphisms between the two regions that appear to have disrupted the core HSE in this region in *GhIPS1-D*, as well as between *GhIPS1-D* and *GhIPS2-A*/*D* (Figure 5a, bottom; Figure S9A). An examination of *IPS1* distal promoter elements in *G. hirsutum*, *G. barbadensis*, *G. raimondii* (a D subgenome representative) and *G. arboreum* (old world cotton and an A subgenome representative) revealed conservation of the HSE within *GhIPS1-A* in *G. hirsutum*, *G. barbadensis* and *G. arboreum*, but not in *IPS1-D* loci for any of these species (Figure 6a). In agreement with the DAP-seq data, we observe a strong positive correlation between *GhHSFA6B-D* and *GhIPS1-A* transcript abundance, but not between *GhHSFA6B-D* and any of the other *GhIPS* loci (Figure 5c; Figure S8). Interestingly, an HSE and AtHSFA6B DAP-seq peak were observed upstream of the Arabidopsis IPS1 and IPS2 paralogues in Arabidopsis cistrome data (AT4G39800 and At2G22240, respectively; Figure 6a), suggesting this regulatory mechanism may be conserved between these two species.

An examination of polymorphisms within our RNA-seq data relative to the *TM-1* reference genome uncovered a lint yield-associated SNP directly adjacent to the predicted GhHSFA6B-A-binding motif (Figure 5a). While this nucleotide is a cytosine (C) at this position in the reference genome and *G. barbadensis* and *G. arboreum*, we observed two genotypes in our panel, either with a cytosine (C, $n = 8$) or a thymine (T, $n = 9$; Figure 5d). Interestingly, the 'C' genotypes, which were either the U.S. or 'mixed' accessions, showed increased lint yield under water-limited conditions relative to the 'T' genotypes (Figure 5d), suggesting this region might impact GhHSFA6B-D binding. To define the GhHSFA6B-D binding site more carefully in the *GhIPS1-A* promoter region, we designed a labelled probe centred on the core HSE (Figure 5a, pink box). A gel shift was observed when this probe was combined with *in vitro*-expressed GhHSFA6B-D protein (Figure 5e). As expected, this signal was abolished when a large excess of unlabelled probes was added. The addition of an unlabelled competitor, corresponding to the homologous *GhIPS1-A* promoter region with a disrupted HSE (Oligo 2, Figure 5e) was unable to abolish binding, even when adding a large molar excess (400× and 600×; Figure 5e). As the labelled probe contained the reference nucleotide (C) at the site of the observed lint yield-associated SNP, we next tested if altering this nucleotide, but leaving the rest of the HSE intact, had an impact on HSFA6B binding. Oligos with this site altered to an A, T or G nucleotide (SNP probes 1–3) could compete for GhHSFA6B-D binding when added in excess (200×; Figure 5e).

While the SNP oligos were competing with the reference oligo when added in excess due to its location outside of the HSE, it is not clear if the site of the SNP impacts GhHSFA6B-D binding. To precisely address this question, we performed a competition experiment using lower concentrations of two non-native SNP probes, SNP-A and SNP-G (SNP probes 1 and 3), that are not present in the genomes of our panel or a much larger sequenced panel comprised of 1024 accessions (Yuan *et al*., 2021; Table S16). Surprisingly, equimolar amounts of either non-native competitor SNP oligos were capable of competing with the reference probe to a higher degree than the unlabelled reference oligo (Figures 5b and 6c; Figure S10). These data suggest that this site, which contains the only observed SNP within the DAP-seq peak region of *GhIPS1-A* in our panel, has a strong influence on GhHSFA6B protein binding, *GhIPS1-A* expression and lint yield. Given the differentially accumulated frequency of SNPs at this site in extant accessions (wild, cultivated, improved and mutant) (Figure 6b), this site has likely been under different levels of selection in wild and cultivated accessions due to its connection to improved yield.

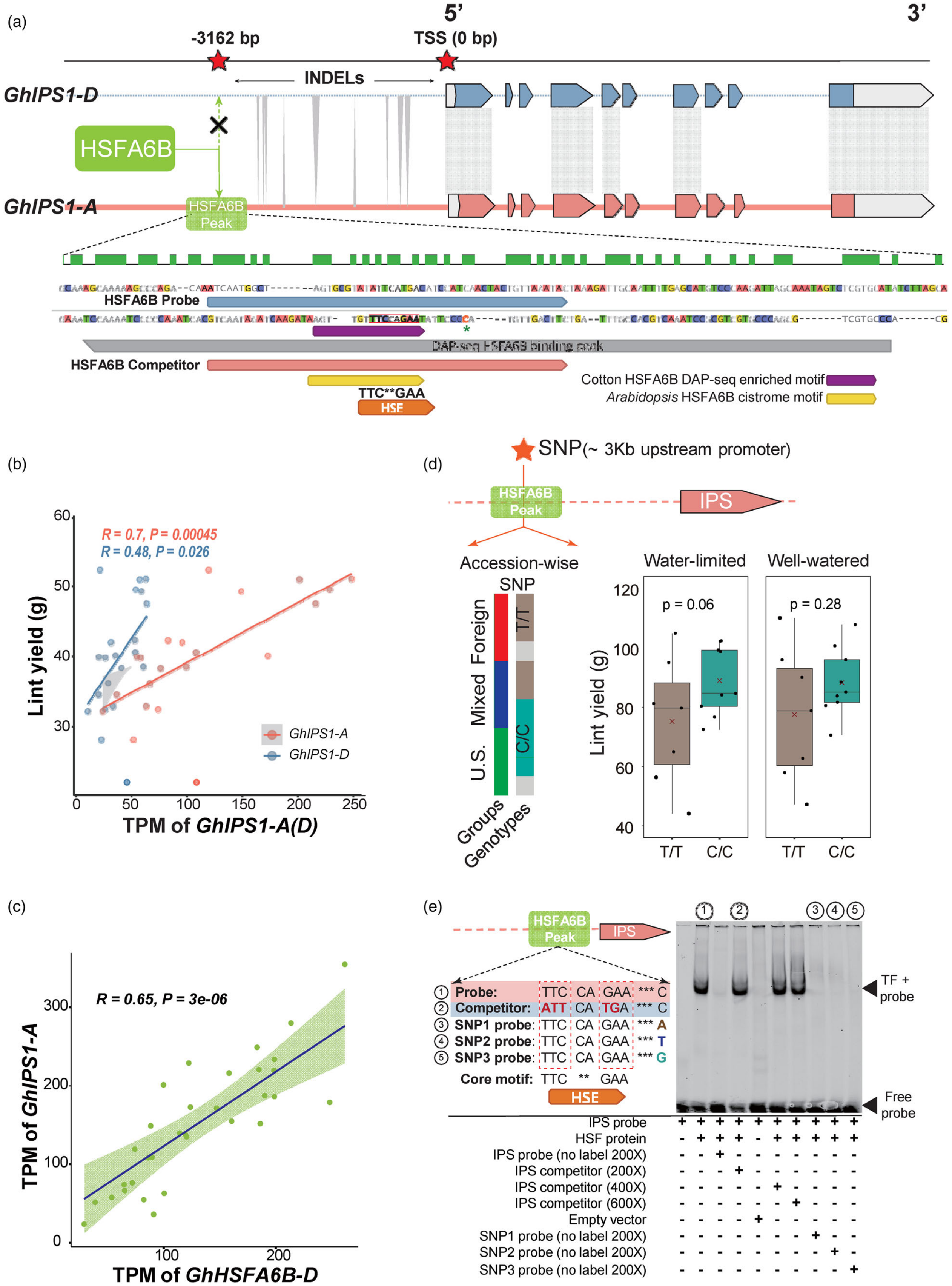
Conclusions

Understanding the regulatory crosstalk between agronomic traits of interest (e.g. yield) and heat and drought stress responses is critical for developing drought-tolerant cultivars with minimal impacts on yield (Alizadeh *et al*., 2020). Characterizing genotype-specific molecular responses under water-limiting conditions is challenging due to the genetic complexity underlying quantitative physiological traits (Tardieu *et al*., 2011; Welcker *et al*., 2011). In most crop species, this genetic complexity is influenced at the pre- and post-transcriptional level by *cis*-regulatory control, gene structural variation and post-translational modifications (Joshi *et al*., 2016; Tardieu *et al*., 2017). In cotton, this complexity is further compounded by a recent (~1.0–1.6 Ma) allopolyploidy and strong human selection within the last 8000 years (Chen *et al*., 2020). Despite the strong selection existing in current elite *G. hirsutum* cultivars relative to other domesticated crops (Chen *et al*., 2020; Ma *et al*., 2018; Su *et al*., 2016), variation in gene expression was identified across the panel for a cohort of transcripts that could be associated with improved yield under water-limiting conditions.

Despite the developmental stage at which we sampled for transcriptomics (primarily the cell elongation stage within the developing bolls; Ma *et al*., 2018), there was already a strong transcriptomic signal in our data connecting abiotic stress factors

14677652, 2024, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pbi.14402, Wiley Online Library on [27/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

14677652, 2024, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pbi.14402, Wiley Online Library on [27/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



© 2024 The Author(s). *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., **22**, 2756–2772

Figure 5 EMSA validation of interaction between *GhIPS1-A* and GhHSFA6B-D. (a) Top: Schematic representation of *GhIPS1-A* and its homologue, *GhIPS1-D* depicting the gene structure as well as the distal promoter region where the HSFA6B binding site was identified. Filled boxes represent annotated exons. Grey shading connecting the two genes represents sequence similarity, with the distal promoter region showing high structural and sequence variation. Bottom: An alignment for the DAP-seq peak for HSFA6B in the promoter region of *GhIPS1-A* and its corresponding region from *GhIPS1-D*. Also shown are the DAP-seq- and Cistrome-identified GhHSFA6B-D binding site (HSE, purple and yellow boxes, respectively), as well as the labelled probe and competitor oligos used for EMSA. A green asterisk denotes the site of the C:T SNP observed within our panel. (b) Pearson correlation among *GhIPS1-A* (red), *GhIPS1-D* (blue) expression (TPM) and lint yield. (c) Pearson correlation between TPM values of *GhIPS1-A* and *GhHSFA6B-D*. (d) Variation in lint yield between accessions associated with 'C/C' (blue) and 'T/T' (brown) genotypes of a SNP adjacent to the HSE element in the distal promoter of *GhIPS1-A* (~3.1 kb upstream of *GhIPS1-A* start). Accessions lacking sufficient coverage at this site are shown in grey. Accessions bearing the C/C or T/T genotypes were also divided based on their phylogenetic groupings (i.e. 'U.S.', 'Foreign' and 'Mixed') from Figure 1a. (e) Interaction between the GhHSFA6B-A protein and *GhIPS1-A* probe was examined by EMSA. The reaction components for each lane are listed below the gel image, including the IRD700-labelled probe, unlabelled probe (200×, same sequence as the labelled probe), competitor sequence (200×, 400× and 600×, containing the *GhIPS1-D* disrupted HSE), empty vector and the other competitor sequence (200×, containing one of the three SNP probes, A, T or G).

and lint yield. This group of lint yield and abiotic stress-associated genes were largely regulated by the well-known transcription factors *GhDREB2A-A* and *GhHSFA6B-D*. Importantly, based on DAP-seq, motif enrichment and co-expression, HSFA6B not only regulates *GhDREB2A-A* in this module but also regulates itself and two lint yield-associated genes, *GhABP* and *GhIPS1-A*. While *GhDREB2A-A* was previously shown to be regulated by *GhHSFA6B-D* in *Arabidopsis*, this regulation was dependent on ABA in general, and specifically the ABA response element TF AREB1 (Huang *et al*., 2016; Sakuma *et al*., 2006). In contrast, in cotton, this regulation does not appear to be ABA dependent, as none of the typical ABA-responsive elements, or ABA biosynthesis genes, were associated with this module, nor does HSFA6B contain a canonical AREB binding domain (ABRE) or share an expression profile with GhAREB1. However, this hypothesis requires further examination. This specialized regulatory mechanism appears to have emerged in cotton well before domestication. Indeed, the old world variant of IPS1 arising from the A subgenome (*GhIPS1-A*) is the only IPS to be targeted by *GhHSFA6B-D*. In fact, this HSFA6B – *IPS1* regulatory connection may reflect an evolutionarily conserved stress response mechanism, as it is also observed in Arabidopsis. Of importance to breeders, this regulatory mechanism appears to have undergone additional selection since speciation to maintain yield under water-limited conditions.

Methods and materials

Plant materials and experimental design

Plant growth conditions have been described in Melandri *et al*. (2021), where these same cotton accessions were examined for their metabolite profiles in response to drought stress over a 2-year field experiment. In brief, a panel of 22 upland cotton (*Gossypium hirsutum* L.) accessions (Table S1) were grown at the University of Arizona Maricopa Agricultural Center (MAC) in Maricopa, AZ, U.S., during the summer of 2019. The accessions were arranged in a randomized incomplete block design with half of the plots experiencing normal irrigation (well-watered, WW) and the other half treated with water-limited (WL) condition, starting when 50% of the plots were at first flower and consisting of approximately half of the WW irrigation amount. Leaf tissue was harvested ~50 days later at the boll development stage and used for metabolomics (details in Melandri *et al*. (2021)) and transcriptomics analyses. For RNA extraction, two biological replicates, comprised of 10 leaf discs (0.64 cm in diameter) of the upper-most expanded leaf derived from 5 randomly selected plants (2 leaf discs from each plant), in each plot were collected within a single day during a time window of 3 h (11:00–14:00). Leaf discs were stored in a 2 mL Eppendorf tube containing 1.5 mL of RNAlater (Fisher #AM7021, Waltham, MA) and stored on ice before being transferred from field to lab where they were then stored in a –80 °C freezer before further processing steps.

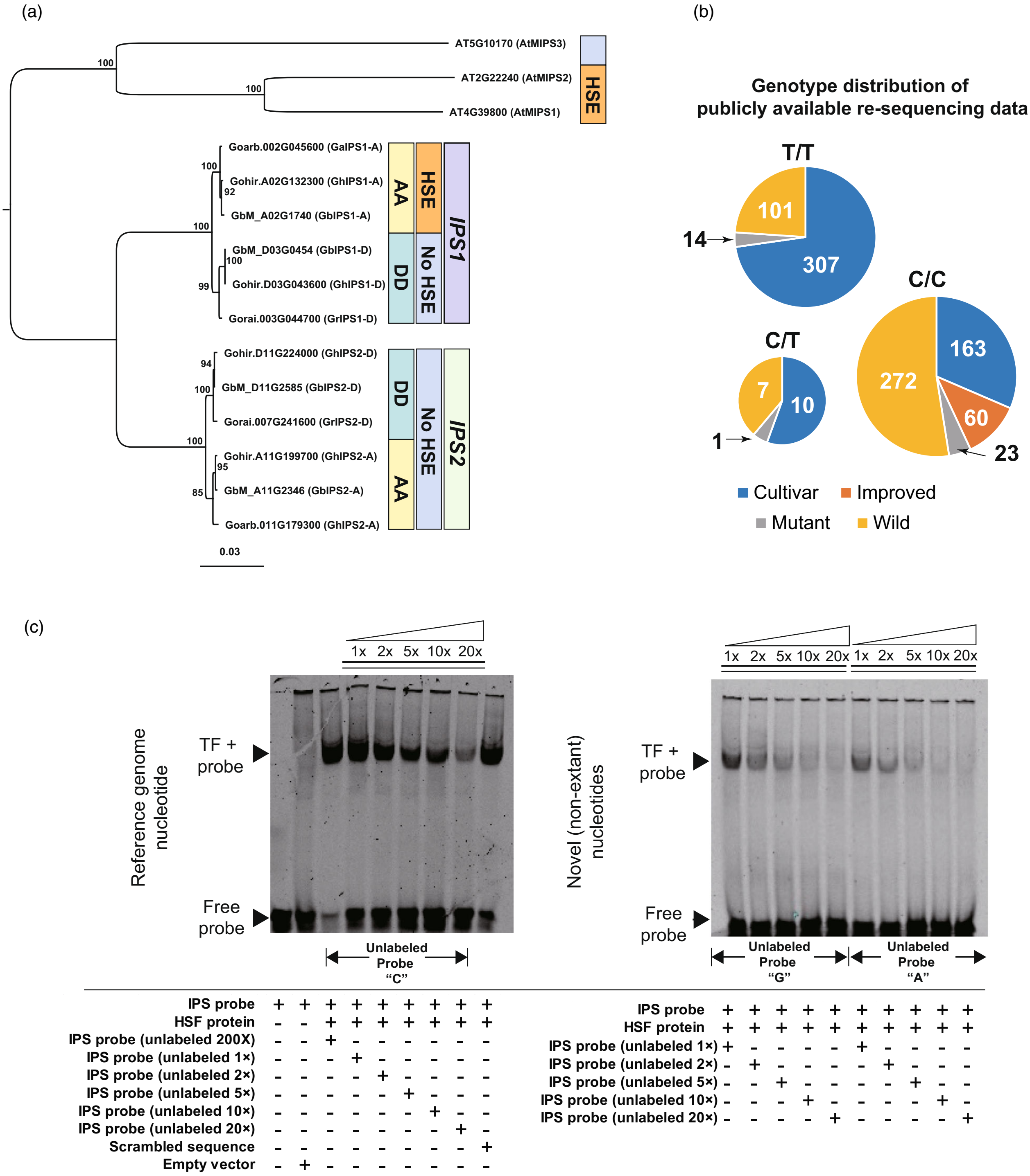
Metabolite and phenotypic data measurement

All details on the procedures used to collect/determine the phenotypic trait data and metabolite data used in this study can be found in Melandri *et al*. (2021). In brief, cotton fibre quantity and quality traits included lint yield (grams/plot), micronaire (Mic, units of air permeability), upper half mean length (UHM, inches), length uniformity (UI, per cent), strength (Str, grammes per tex) and elongation (Elo, per cent). Reflectance-based vegetation indices (VIs) included normalized difference vegetation index (NDVI), carotenoid reflectance index (CRI), scaled photochemical reflectance index (sPRI) and the ratio between the water index (WI) and NDVI (WI/NDVI). The levels of 451 metabolites were determined by untargeted gas chromatography–mass spectrometry (GC–MS, 27 metabolites) and liquid chromatography–mass spectrometry (LC–MS, 424 metabolites). Best linear unbiased estimators (BLUEs) of each phenotypic trait and metabolite value were generated for each cotton accession before being used for downstream statistical analyses.

RNA-seq library construction

Extraction of RNA from leaf disks was performed using methods from previous studies (Pang *et al*., 2011; Wu *et al*., 2002). Hot borate buffer was prepared containing 0.2 M sodium borate pH 9.0, 30 mM EGTA, 1% SDS and 1% sodium deoxycholate. Just before use, PVP-40, NP-40 and DTT were added to the hot borate buffer to a final concentration of 2%, 1% and 10 mM respectively. To extract RNA, a total of approximately 0.75 mL hot borate buffer per 50 mg tissue (~5 cotton leaf disks) was used. Buffer was heated to 80 °C, and 250 μL hot buffer was added to 5 leaf disks. Tissue was ground in the hot buffer in a mortar and pestle, then 15 μL of 20 mg/mL proteinase K (Roche #46950800: Indianapolis, IN/Sigma #3115887001: Burlington, MA) per sample was added and the tissue was ground again. A final 500 μL hot buffer was added before grinding a final time. Lysate was added to a Qiashredder column (Qiagen #79656: Germantown, MD) and centrifuged at 13 000 ***g*** for 1 min. Flow through was added to 0.5 volumes of 100% ethanol. The mixture was used as input for RNA cleanup using the RNeasy kit (Qiagen #74104) according to the manufacturer's instructions. Samples

14677652, 2024, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pbi.14402, Wiley Online Library on [27/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



IPS probe	+	+	+	+	+	+	+	+	+
HSF protein	-	-	+	+	+	+	+	+	+
IPS probe (unlabeled 200X)	-	-	+	-	-	-	-	-	-
IPS probe (unlabeled 1×)	-	-	-	+	-	-	-	-	-
IPS probe (unlabeled 2×)	-	-	-	-	+	-	-	-	-
IPS probe (unlabeled 5×)	-	-	-	-	-	+	-	-	-
IPS probe (unlabeled 10×)	-	-	-	-	-	-	+	-	-
IPS probe (unlabeled 20×)	-	-	-	-	-	-	-	+	-
Scrambled sequence	-	-	-	-	-	-	-	-	+
Empty vector	-	+	-	-	-	-	-	-	-

IPS probe	+	+	+	+	+	+	+	+	+	+
HSF protein	+	+	+	+	+	+	+	+	+	+
IPS probe (unlabeled 1×)	+	-	-	-	-	+	-	-	-	-
IPS probe (unlabeled 2×)	-	+	-	-	-	-	+	-	-	-
IPS probe (unlabeled 5×)	-	-	+	-	-	-	-	+	-	-
IPS probe (unlabeled 10×)	-	-	-	+	-	-	-	-	+	-
IPS probe (unlabeled 20×)	-	-	-	-	+	-	-	-	-	+

Figure 6 Identification of an HSE-adjacent SNP that impacts GhHSFA6B-D binding to *GhIPS1-A*. (a) A phylogenetic tree containing *GhIPS1* and *GhIPS2* homeologous genes from *G. hirsutum* (AADD), *G. barbadense* (AADD), *G. arboreum* (AA) and *G. raimondii* (DD) was constructed using *Arabidopsis thaliana* IPS homologues as an outgroup. The 'orange' and 'light blue' labels denote the presence of the HSE upstream of *GhIPS1-A* genes in all AA genome-containing species, but not in the distal promoters of *GhIPS1-D* genes found in DD genome species, nor in any of the *GhIPS2* paralogues in either AA or DD genomes. (b) Distribution of genotypes of the *GhIPS1-A* DAP-seq peak-associated SNP (C:T) in a published whole-genome sequencing-based panel (1024 accessions; Yuan *et al*., 2021). The distribution is summarized in pie charts reflecting the accession frequencies of the wild, cultivated and mutants and improved *G. hirsutum* accessions across the 'C/C' genotype (reference genome genotype), 'T/T' genotype and 'T/C' genotype. (c) The level of the GhHSFA6B-D to *GhIPS1-A* binding efficiency of different IPS-associated SNP variants was examined by competitions between the IDR700 dye-labelled reference oligos and non-labelled oligos with the reference nucleotide 'C', or the non-native SNP oligos 'G' and 'A' respectively. Left panel: Binding efficiency of the reference 'C' containing probe was tested by titrating in increasing amounts of the competitor probe (unlabelled reference 'C'). Right panel: The impact of this nucleotide position on GhHSFA6B-D binding was tested by competing the labelled reference probe with increasing amounts of two non-native probes ('G' or 'A'). Components in each reaction are shown below.

14677652, 2024, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pbi.14402, Wiley Online Library on [27/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

were eluted with RNase-free water. RNA-seq libraries were then generated from mRNA-enriched samples using the Amaryllis Nucleics Full Transcript Library Prep kit (YourSeq Duet, https://amaryllisnucleics.com/kits/duet-rnaseq-library-kit).

Transcriptome sequencing and data processing

Each of the 22 *G. hirsutum* accessions treated under both the WW or WL conditions was sequenced using Illumina HiSeqTM 2500 (San Diego, CA) paired-end libraries. Trimmomatic was used to trim adapters and low-quality reads (Bolger *et al*., 2014). Furthermore, reads were aligned to the *G. hirsutum* reference genome (Ghirsutum_527_v2.0, accession ID: VKGJ01000000), acquired from Phytozome (Chen *et al*., 2020), using RMTA v2.6.3 pipeline (Peri *et al*., 2020) with default parameters and further quantified reads mapped back to each locus using FeatureCounts (parameter: multimapping; reads: counted; and multi-overlapping reads: counted) and then transformed into length normalized transcripts per kilobase million (TPM) by custom R scripts (Liao *et al*., 2014) The raw feature counts for each transcript across 88 samples were normalized by DESeq2 (Love *et al*., 2014). To test the reproducibility of replicates, we used the Pearson correlation of normalized counts between each set of replicates. Replicates with $R^2 < 0.8$ and $P > 0.05$ between samples were removed, leading to only 21 accessions being further examined. Furthermore, pairwise comparison of TPM values under WL and WW conditions for each accession was performed to identify respective accession-specific DEGs (model: ~entry + treatment + entry: treatment). For each accession, significant DEGs under WL condition were classified using thresholds adjusted *P*-value < 0.01, $\text{Log}_2\text{FC} > 1$ (upregulated) or <1 (downregulated). We examined the average expression of all genes from A and D subgenomes with expression >1 TPM to determine whether there is a shift in global subgenome gene expression between WW and WL conditions. This analysis tests the hypothesis (Alger and Edger, 2020) that subgenomes may be adapted to different environments such that their expression dominance may be affected by spatiotemporal contexts.

Variants calling and phylogenetic analysis

Variant calling was conducted with the GATK4 pipeline using the haplotypecaller function (Brouard *et al*., 2019). The average mapping depth (DP) was calculated across 84 samples as screening cutoff which is equal to 4.12. Then, we filtered out variants with sites with DP < 3, depth by quality (DQ) < 2, genotype quality (GQ) < 20 and minor allele frequency < 0.025 by vcftools (Danecek *et al*., 2011). Phylogenetic relationships were inferred using the IQ-TREE pipeline with filtered SNPs (Nguyen *et al*., 2015). Briefly, variants calling files (VCFs) of all samples were transformed into phylip format by vcf2phylip tools, and a maximum-likelihood tree was constructed (parameter: -nt AUTO -m MFP, Ortiz, 2019) and plotted by Figtree (http://tree.bio.ed.ac.uk/software/figtree/). The 21 accessions were subdivided into different groups based on phylogenetic, geographic and historical breeding information obtained from USDA-GRIN (https://www.ars-grin.gov). A publicly available deep-sequenced (~20×) WGS dataset containing 17 of 21 accessions used for field experiments was obtained from the NCBI SRA (Table S1) to identify genome-wide variants. The reads were trimmed by trimmomatic (default settings, (Bolger *et al*., 2014)) and mapped to the reference genome using bwa-mem along with retention of unique mapped reads by picard (MarkDuplicates – remove = TRUE) for variants calling. The haplotypecaller calling was performed by GATK4 and filtered by VCFtools (--max-missing 0.9, --maf 0.05, minGQ 20, minQ 200 and minDP 5 (Danecek *et al*., 2011)). The filtered variants were annotated using VEP (default settings; McLaren *et al*., 2016) to classify variations to coding regions, promoter and inter-genic regions.

Principal component analysis (PCA)

To estimate the strength of treatment effects and the impact of phylogenetic relatedness on transcriptomic and metabolomic profiles, a principal component analysis (PCA) was performed using normalized read counts of 3,768 genes associated with the top 10% of transcriptomic variance or using plotPCA function of DESeq2 (Love *et al*., 2014), and normalized *Z*-score of 451 metabolites using 'factoextra' R package (https://cran.r-project.org/web/packages/factoextra). Contribution rates for each component were calculated using plotPCA. The first three components from transcriptome and metabolites were visualized in a three-dimensional PCA using Cubemaker (https://tools.altiusinstitute.org/cubemaker/).

Weighted gene co-expression network analysis

Gene co-expression network was constructed using the WGCNA R package to classify gene expression modules and explore module–trait relationships (Langfelder and Horvath, 2008). Genes with both a high median absolute deviation (MAD) score (top 10%) and genes with high expression (average TPM across all samples > 1) were retained for expression module classification. The pickSoftThreshold function was used to identify the optimal soft power threshold (=7) at which R^2 surpassed 0.85 and no further improvements in mean connectivity (module size) were observed, as performed previously (Zhu *et al*., 2018). Block-wise modules were constructed using the following parameters (power = 7, maxBlockSize = 5000, TOMType = 'unsigned', minModuleSize = 30, reassignThreshold = 0 and mergeCutHeight = 0.2). Trait–module relationships were derived using the 'modTraitCor' function in WGCNA. Modules that displayed a high trait correlation ($R^2 > 0.6$, $P < 0.05$) were selected for further analysis. To investigate modules with high module–trait membership, only co-expressed genes with strong connectivity (weight score > 0.1) were retained for downstream analysis.

Functional characterization of trait-correlated module

Gene functional annotations of filtered genes (weight score > 0.1) in lint yield-correlated module were obtained from the cotton functional genome database (CottonFGD) to perform enrichment of gene ontology (GO) and KEGG pathways (https://cottonfgd.org/; Zhu *et al*., 2017) using Fisher exact test ($P < 0.01$). Transcription factors (TFs), transcription regulators (TRs) and protein kinases within this module were classified using iTAK (Zheng *et al*., 2016). Reported protein–protein interactions (PPIs) among genes within the same module were screened using the STRING database (Szklarczyk *et al*., 2017; confidence level = 0.6, interaction source: database and experiments). To identify the upstream master TF(s) that may bind to genes within the lint yield module, the enrichment of consensus TF motifs was tested using the 2-kbp upstream region of the 866 co-expressed genes within the module. The enrichment test was performed by the analysis of motif enrichment (AME) pipeline (McLeay and Bailey, 2010) with the *Arabidopsis* DAP-seq profiles of O'Malley *et al*. (2016) as the consensus motif database (parameters: --scoring avg --method fisher --hit-lo-fraction 0.25 --e-value --kmer report-threshold 10.0, cutoff: TP values > 3, *P*-value < 0.001).

14677652, 2024, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pbi.14402, Wiley Online Library on [27/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

DNA affinity purification sequencing (DAP-seq) library construction

The four TFs were selected as hub genes to construct DAP-seq libraries, including two *GhDREB2A* (*Gohir.A13G021700* and Gohir.D13G022300) and two *GhHSFA6B* (*Gohir.D08G072600* and *Gohir.A08G064100*). DAP-seq assay was carried out as described by published protocol (Bartlett *et al*., 2017). The NEB next® DNA library prep master mix set for Illumina kit (NEB #E6040S) was implemented to prepare the DAP-seq gDNA library. The pIX-HALO vector (cat#G184A, Promega) was used to fuse the *GhDREB2A* and *GhHSFA6B* into the HaloTag. We further used the TNT SP6 high-yield wheat germ protein expression system (L3260, Promega) to express the four TFs-HaloTag fusion protein. Magens HaloTag beads (G7281, Promega) were used to purify the fusion protein. The fusion protein and 500 ng of library DNA were co-incubated in 40 μL PBS buffer for 1.5 h shaking in a cold room. The beads were washed with 200 μL PBS + NP40 (0.005%) for five times. The supernatant was discarded and an aliquot of 25 μL of elution buffer was added. Finally, beads were incubated at 98 °C for 10 min to elute DNA fragments. According to the fragment size of the library, the DAP-seq library concentration for a given read count was measured.

DAP-seq sequencing and peak analysis

DAP-seq libraries were sequenced by Illumina short reads platform (single end: 150 bp) with a total of expected 40 million reads for each protein to be sequenced. These reads were trimmed by TrimGalore (Krueger, 2015) and then mapped to the reference genome (*TM-1*: Ghirsutum_527_v2.0, Phytozome accession ID: VKGJ01000000) (Chen *et al*., 2020) using bowtie2 and sorted by sambamba. Before peak calling, only the unique mapped reads were retained by sambamba (parameter: -F '[XS] == null and not unmapped and not duplicate') to prevent false positives due to multiple mapped hits (Tarasov *et al*., 2015). Furthermore, the MAC2s tool was applied to call peaks (parameter: --keep-dupall -g 2.3-e9) and high-confidence peaks were captured using IDR (*P*adj < 0.05; Zhang *et al*., 2008). These high-quality peaks were annotated based on reference gene annotation using the ChIPseeker R package (Yu *et al*., 2015). Regulatory regions of a target gene were defined as the 5 kb upstream sequences before the transcription start sites (TSSs) and the downstream sequences bearing the longest 5′UTRs among respective isoforms. In particular, only DAP-seq peaks which fell into the regulatory regions were kept as TF-targeted genes. The sequences associated with DAP-seq peaks of TFs were harvested as query sequences to perform motif discovery by MEME suite (Bartlett *et al*., 2017) to identify conserved consensus motif sequences among peak sequences (parameter: -mod zoops -nmotifs 3 -minw 6 -maxw 50 -objfun classic -revcomp -markov_order 0).

Genomic analysis of TF binding regions

Paralogous genes of the selected lint module genes (*GhABP* and *GhIPS*) were identified using Blastn (Ye *et al*., 2006), which aligns the CDS sequences against the reference gene annotation (cut-off: *E*-value < 10–3 and identities > 90%). Those hits identified were further checked by functional annotation. Furthermore, the promoter and coding regions of these paralogous gene pairs were extracted and then aligned using multiple sequences alignment by Geneious (tool: MUSCLE, iteration: 500) to identify variants (SNPs and INDELs) between paralogous gene pairs, specifically the regions bearing the DAP-seq peak (Kearse *et al*., 2012). The potential binding sites of two TFs over the peak regions were defined by motif scan (FIMO) using *Arabidopsis* DAP-seq-derived consensus motif sequences (Motif ID: ERF48_col_a, HSFA6B_cal_a, HSFA6B_colamp_a) (Bailey *et al*., 2015).

In vitro protein expression and electrophoretic mobility shift assay (EMSA)

Protein expression was carried out using the *in vitro* transcription and translation system (TnT™ T7 Quick for PCR DNA, Promega: Madison, WI) and the EMSA reaction was carried out following manufacturer's instructions (odyssey). The vector containing the gene of interest (with flag tag) was PCR amplified to obtain the DNA template. The reaction mixture contained the DNA template, reaction buffer, amino acid mix and T7 RNA polymerase, following the manufacturer's instructions. The reaction was incubated at the recommended temperature for a specified duration, and the resulting protein product was subsequently analysed by SDS-PAGE to confirm expression.

Immunoblot analysis was performed by using standard wet transfer method. The proteins were blotted using nitrocellulose membrane (Sigma Aldrich: Burlington, MA). Rabbit monoclonal halo tag antibody was used in the (1:5000) dilution (v/v). Horseradish peroxidase-conjugated goat anti-rabbit antibodies were used as secondary antibodies. Blots were developed using supersignal chemiluminescent substrate following the manufacturer's instruction. The blots were imaged using the BioRAD gel imager following the linear curve of detection in the signal.

The interaction between the expressed protein and its putative DNA-binding partner was analysed using EMSA. A double-stranded DNA probe containing the target binding site was prepared by annealing complementary oligonucleotides labelled with a fluorophore. The binding reaction mixture consisted of the expressed protein, labelled DNA probe, binding buffer and appropriate cofactors. The reaction mixture was incubated at room temperature for 30 min to allow for protein–DNA complex formation. Subsequently, the samples were resolved on a non-denaturing polyacrylamide gel electrophoresis (PAGE) with a 6% acrylamide concentration. The gel was run at 120 volts for 3 h in 1× TBE. Following electrophoresis, the gel was visualized using LiCor Odyssey Gel Scanner to capture the mobility shifts indicating protein–DNA complex formation. For competition assays, unlabelled competitor DNA or unlabelled mutated competitor DNA was added to the binding reaction at increasing molar excesses.

Acknowledgements

The authors would like to acknowledge support from the NSF-IOS #2023310 (awarded to ADLN), NSF-IOS #2102120 (awarded to DP and ADLN), NSF-PRFB #2109178 (awarded to JRB), NSF-MCB #2226270 (awarded to AS), Cotton Incorporated Core Projects 18-384, 20-720 and 23-890 (to DP) and 22-639 (to ADLN), and the United States Department of Agriculture-Agricultural Research Service (USDA-ARS) (KT, CRIS #2020-13660-008-00-D; and LH, CRIS #3091-21000-041-00D). Mention of trade names or commercial products in this publication was solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. The USDA was an equal opportunity provider and employer. The authors would like to

14677652, 2024, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pbi.14402, Wiley Online Library on [27/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

thank other members of the Nelson, Skirycz and Pauli labs, as well as Dr. Jeffrey Chen (UT-Austin), for helpful discussion in the formulation of the analyses presented herein.

Conflict of interest

The authors declare no competing interests.

Author contributions

LY, ADLN and DP developed the project and performed RNA extractions and library prep. GM performed metabolomics analysis. KRT contributed to field data collection. XZ performed the DAP-seq library prep. LY, VPTK and XZ conducted the EMSA experiments. KRT and LH contributed to the execution of the field trial. LH contributed to the germplasm. LY, GM, DP and ADLN contributed to the writing of the manuscript. All authors contributed to the review and approval of the manuscript.

Data availability statement

The Illumina reads of transcriptome data have been deposited in the NCBI database under BioProject: PRJNA1118798. Analysis code is available here: https://github.com/Leon-Yu0320/Cotton_omics_studies.

References

Alger, E.I. and Edger, P.P. (2020) One subgenome to rule them all: underlying mechanisms of subgenome dominance. *Curr. Opin. Plant Biol.* **54**, 108–113.

Alizadeh, M.R., Adamowski, J., Nikoo, M.R., AghaKouchak, A., Dennison, P. and Sadegh, M. (2020) A century of observations reveals increasing likelihood of continental-scale compound dry-hot extremes. *Sci. Adv.* **6**, 1–12.

Bailey, T.L., Johnson, J., Grant, C.E. and Noble, W.S. (2015) The MEME suite. *Nucleic Acids Res.* **43**, W39–W49.

Bartlett, A., O'Malley, R.C., Huang, S.S.C., Galli, M., Nery, J.R., Gallavotti, A. and Ecker, J.R. (2017) Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nat. Protoc.* **12**, 1659–1672.

Bian, X., Li, W., Niu, C.F., Wei, W., Hu, Y., Han, J.Q., Lu, X. *et al.* (2020) A class B heat shock factor selected for during soybean domestication contributes to salt tolerance by promoting flavonoid biosynthesis. *New Phytol.* **225**, 268–283.

Bird, K.A., VanBuren, R., Puzey, J.R. and Edger, P.P. (2018) The causes and consequences of subgenome dominance in hybrids and recent polyploids. *New Phytol.* **220**, 87–93.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Brouard, J.S., Schenkel, F., Marete, A. and Bissonnette, N. (2019) The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments. *J. Anim. Sci. Biotechnol.* **10**, 1–6.

Chen, Z.J., Scheffler, B.E., Dennis, E., Triplett, B.A., Zhang, T., Guo, W., Chen, X. *et al.* (2007) Toward sequencing cotton (*Gossypium*) genomes. *Plant Physiol.* **145**, 1303–1310.

Chen, J., Nolan, T.M., Ye, H., Zhang, M., Tong, H., Xin, P., Chu, J. *et al.* (2017) Arabidopsis WRKY46, WRKY54, and WRKY70 transcription factors are involved in brassinosteroid-regulated plant growth and drought responses. *Plant Cell*, **29**, 1425–1439.

Chen, Z.J., Sreedasyam, A., Ando, A., Song, Q., de Santiago, L.M., Hulse-Kemp, A.M., Ding, M. *et al.* (2020) Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. *Nat. Genet.* **52**, 525–533.

Chu, X., Wang, C., Chen, X., Lu, W., Li, H., Wang, X., Hao, L. *et al.* (2015) The cotton WRKY gene GhWRKY41 positively regulates salt and drought stress tolerance in transgenic *Nicotiana benthamiana*. *PLoS One*, **10**, 1–21.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Fang, D.D., Jenkins, J.N., Deng, D.D., McCarty, J.C., Li, P. and Wu, J. (2014) Quantitative trait loci analysis of fiber quality traits using a random-mated recombinant inbred population in Upland cotton (*Gossypium hirsutum* L.). *BMC Genomics*, **15**, 397.

Fang, Y., Liao, K., Du, H., Xu, Y., Song, H., Li, X. and Xiong, L. (2015) A stress-responsive NAC transcription factor SNAC3 confers heat and drought tolerance through modulation of reactive oxygen species in rice. *J. Exp. Bot.* **66**, 6803–6817.

Fang, L., Wang, Q., Hu, Y., Jia, Y., Chen, J., Liu, B., Zhang, Z. *et al.* (2017) Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. *Nat. Genet.* **49**, 1089–1098.

Gupta, A., Rico-Medina, A. and Caño-Delgado, A.I. (2020) The physiology of plant responses to drought. *Science*, **368**, 266–269.

Hinze, L.L., Fang, D.D., Gore, M.A., Scheffler, B.E., Yu, J.Z., Frelichowski, J. and Percy, R.G. (2015) Molecular characterization of the *Gossypium* diversity reference set of the US national cotton germplasm collection. *Theor. Appl. Genet.* **128**, 313–327.

Hinze, L.L., Gazave, E., Gore, M.A., Fang, D.D., Scheffler, B.E., Yu, J.Z., Jones, D.C. *et al.* (2016) Genetic diversity of the two commercial tetraploid cotton species in the *Gossypium* diversity reference set. *J. Hered.* **107**, 274–286.

Huang, J.G., Yang, M., Liu, P., Yang, G.D., Wu, C.A. and Zheng, C.C. (2009) GhDREB1 enhances abiotic stress tolerance, delays GA-mediated development and represses cytokinin signalling in transgenic Arabidopsis. *Plant Cell Environ.* **32**, 1132–1145.

Huang, G.Q., Li, W., Zhou, W., Zhang, J.M., Di Li, D., Gong, S.Y. and Li, X.B. (2013) Seven cotton genes encoding putative NAC domain proteins are preferentially expressed in roots and in responses to abiotic stress during root development. *Plant Growth Regul.* **71**, 101–112.

Huang, Y., Niu, C., Yang, C. and Jinn, T. (2016) The heat stress factor HSFA6b connects ABA signaling and ABA-mediated heat responses. *Plant Physiol.* **172**, 1182–1199.

Jacob, P., Hirt, H. and Bendahmane, A. (2017) The heat-shock protein/chaperone network and multiple stress resistance. *Plant Biotechnol. J.* **15**, 405–414.

Joshi, R., Wani, S.H., Singh, B., Bohra, A., Dar, Z.A., Lone, A.A., Pareek, A. *et al.* (2016) Transcription factors and plants response to drought stress: current understanding and future directions. *Front. Plant Sci.* **7**, 1–15.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S. *et al.* (2012) Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.

Kolmos, E., Chow, B.Y., Pruneda-paz, J.L. and Kay, S.A. (2014) HsfB2b-mediated repression of PRR7 directs abiotic stress responses of the circadian clock. *Proc. Natl Acad. Sci. USA*, **111**, 16172–16177.

Krueger, F. (2015) *Trim Galore!: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data*. Babraham Institute.

Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Li, J., Yuan, D., Wang, P., Wang, Q., Sun, M., Liu, Z., Si, H. *et al.* (2021) Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. *Genome Biol.* **22**, 1–26.

Liao, Y., Smyth, G.K. and Shi, W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.

Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 1–21.

Ma, L., Hu, L., Fan, J., Amombo, E., Khaldun, A.B.M., Zheng, Y. and Chen, L. (2017) Cotton GhERF38 gene is involved in plant response to salt/drought and ABA. *Ecotoxicology*, **26**, 841–854.

Ma, Z., He, S., Wang, X., Sun, J., Zhang, Y., Zhang, G., Wu, L. *et al.* (2018) Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nat. Genet.* **50**, 803–813.

14677652, 2024, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pbi.14402, Wiley Online Library on [27/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Mahmood, T., Khalid, S., Abdullah, M., Ahmed, Z., Shah, M.K.N., Ghafoor, A. and Du, X. (2019) Insights into drought stress signaling in plants and the molecular genetic basis of cotton drought tolerance. *Cells*, **9**, 105.

Malhotra, S. and Sowdhamini, R. (2014) Interactions among plant transcription factors regulating expression of stress-responsive genes. *Bioinform. Biol. Insights*, **8**, 193–198.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P. *et al*. (2016) The ensembl variant effect predictor. *Genome Biol.* **17**, 1–14.

McLeay, R.C. and Bailey, T.L. (2010) Motif enrichment analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics*, **11**, 165.

Melandri, G., Thorp, K.R., Broeckling, C., Thompson, A.L., Hinze, L. and Pauli, D. (2021) Assessing drought and heat stress-induced changes in the cotton leaf metabolome and their relationship with hyperspectral reflectance. *Front. Plant Sci.* **12**, 1–19.

Mitsuhashi, N., Kondo, M., Nakaune, S., Ohnishi, M. and Hayashi, M. (2008) Localization of myo -inositol-1-phosphate synthase to the endosperm in developing seeds of Arabidopsis. *J. Exp. Bot.* **59**, 3069–3076.

Nakashima, K., Yamaguchi-Shinozaki, K. and Shinozaki, K. (2014) The transcriptional regulatory network in the drought response and its crosstalk in abiotic stress responses including drought, cold, and heat. *Front. Plant Sci.* **5**, 1–7.

Nguyen, L.T., Schmidt, H.A., Von Haeseler, A. and Minh, B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274.

Nishizawa, A., Yabuta, Y., Yoshida, E., Maruta, T., Yoshimura, K. and Shigeoka, S. (2006) Arabidopsis heat shock transcription factor A2 as a key regulator in response to several types of environmental stress. *Plant J.* **48**, 535–547.

O'Malley, R.C., Huang, S.S.C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M. *et al*. (2016) Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell*, **165**, 1280–1292.

Ortiz, E. (2019) *Convert a VCF matrix into several matrix formats for phylogenetic analysis. 2.0*.

Pan, Y., Meng, F. and Wang, X. (2020) Sequencing multiple cotton genomes reveals complex structures and lays foundation for breeding. *Front. Plant Sci.* **11**, 560096.

Pang, M., McD Stewart, J. and Zhang, J. (2011) A mini-scale hot borate method for the isolation of total RNA from a large number of cotton tissue samples. *Afr. J. Biotechnol.* **10**, 15430–15437.

Peng, R., Xu, Y., Tian, S., Unver, T., Liu, Z., Zhou, Z., Cai, X. *et al*. (2022) Evolutionary divergence of duplicated genomes in newly described allotetraploid cottons. *Proc. Natl Acad. Sci. USA*, **119**, e2208496119.

Peri, S., Roberts, S., Kreko, I.R., McHan, L.B., Naron, A., Ram, A., Murphy, R.L. *et al*. (2020) Read mapping and transcript assembly: a scalable and high-throughput workflow for the processing and analysis of ribonucleic acid sequencing data. *Front. Genet.* **10**, 1361.

Sakuma, Y., Maruyama, K., Qin, F., Osakabe, Y., Shinozaki, K. and Yamaguchi-Shinozaki, K. (2006) Dual function of an Arabidopsis transcription factor DREB2A in water-stress-responsive and heat-stress-responsive gene expression. *Proc. Natl Acad. Sci. USA*, **103**, 18822–18827.

Scharf, K., Berberich, T., Ebersberger, I. and Nover, L. (2012) The plant heat stress transcription factor (Hsf) family: structure, function and evolution. *Biochim. Biophys. Acta* **1819**, 104–119.

Shinozaki, K. and Yamaguchi-Shinozaki, K. (2007) Gene networks involved in drought stress response and tolerance. *J. Exp. Bot.* **58**, 221–227.

Singh, D. and Laxmi, A. (2015) Transcriptional regulation of drought response: a tortuous network of transcriptional factors. *Front. Plant Sci.* **6**, 1–11.

Su, J., Fan, S., Li, L., Wei, H., Wang, C. and Wang, H. (2016) Detection of favorable QTL alleles and candidate genes for lint percentage by GWAS in Chinese upland cotton. *Front. Plant Sci.* **7**, 1–11.

Sun, F., Chen, Q., Chen, Q., Jiang, M., Gao, W. and Qu, Y. (2021) Screening of key drought tolerance indices for cotton at the flowering and boll setting stage using the dimension reduction method. *Front. Plant Sci.* **12**, 1–10.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A. *et al*. (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362–D368.

Takahashi, F., Kuromori, T., Urano, K., Yamaguchi-Shinozaki, K. and Shinozaki, K. (2020) Drought stress responses and resistance in plants: from cellular responses to long-distance intercellular communication. *Front. Plant Sci.* **11**, 1–14.

Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J. and Prins, P. (2015) Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, **31**, 2032–2034.

Tardieu, F., Granier, C. and Muller, B. (2011) Water deficit and growth. Co-ordinating processes without an orchestrator? *Curr. Opin. Plant Biol.* **14**, 283–289.

Tardieu, F., Cabrera-Bosquet, L., Pridmore, T. and Bennett, M. (2017) Plant phenomics, from sensors to knowledge. *Curr. Biol.* **27**, R770–R783.

Vinson, C.C., Mota, A.P.Z., Porto, B.N., Oliveira, T.N., Sampaio, I., Lacerda, A.L., Danchin, E.G.J. *et al*. (2020) Characterization of raffinose metabolism genes uncovers a wild *Arachis galactinol* synthase conferring tolerance to abiotic stresses. *Sci. Rep.* **10**, 15258.

Wang, K., Wang, Z., Li, F., Ye, W., Wang, J., Song, G., Yue, Z. *et al*. (2012) The draft genome of a diploid cotton *Gossypium raimondii*. *Nat. Genet.* **44**, 1098–1103.

Welcker, C., Sadok, W., Dignat, G., Renault, M., Salvi, S., Charcosset, A. and Tardieu, F. (2011) A common genetic determinism for sensitivities to soil water deficit and evaporative demand: Meta-analysis of quantitative trait loci and introgression lines of maize. *Plant Physiol.* **157**, 718–729.

Weltmeier, F., Rahmani, F., Ehlert, A., Dietrich, K., Schütze, K., Wang, X., Chaban, C. *et al*. (2009) Expression patterns within the Arabidopsis C/S1 bZIP transcription factor network: availability of heterodimerization partners controls gene expression during stress response and development. *Plant Mol. Biol.* **69**, 107–119.

Wu, Y., Llewellyn, D.J. and Dennis, E.S. (2002) A quick and easy method for isolating good-quality RNA from cotton (*Gossypium hirsutum* L.) tissues. *Plant Mol. Biol. Report.* **20**, 213–218.

Ye, J., McGinnis, S. and Madden, T.L. (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res.* **34**, 6–9.

Yu, G., Wang, L.G. and He, Q.Y. (2015) ChIP seeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, **31**, 2382–2383.

Yuan, D., Grover, C.E., Hu, G., Pan, M., Miller, E.R., Conover, J.L., Hunt, S.P. *et al*. (2021) Parallel and intertwining threads of domestication in allopolyploid cotton. *Adv. Sci.* **8**, 1–17.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C. *et al*. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137.

Zheng, Y., Jiao, C., Sun, H., Rosli, H.G., Pombo, M.A., Zhang, P., Banf, M. *et al*. (2016) iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant*, **9**, 1667–1670.

Zhu, T., Liang, C., Meng, Z., Sun, G., Meng, Z., Guo, S. and Zhang, R. (2017) CottonFGD: an integrated functional genomics database for cotton. *BMC Plant Biol.* **17**, 1–9.

Zhu, G., Wang, S., Huang, Z., Zhang, S., Liao, Q., Zhang, C., Lin, T. *et al*. (2018) Rewiring of the fruit metabolome in tomato breeding. *Cell*, **172**, 249–261.e12.

Zhu, G., Gao, W., Song, X., Sun, F., Hou, S., Liu, N., Huang, Y. *et al*. (2020) Genome-wide association reveals genetic variation of lint yield components under salty field conditions in cotton (*Gossypium hirsutum* L.). *BMC Plant Biol.* **20**, 23.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 Comparison of six production-related traits and four vegetative indices from two watering conditions.

Figure S2 Transcriptome comparison of expression profiles in the three phylogenetic groups.

Figure S3 Comparison of global subgenome expression bias.

14677652, 2024, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pbi.14402, Wiley Online Library on [27/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Figure S4 Weighted co-expression network and trait–module membership.
Figure S5 Preparation of DAP-seq libraries.
Figure S6 EMSA validation of interaction between ABP and two TFs.
Figure S7 Distribution of GhDREB2A and GhHSFA6B DAP-seq peaks.
Figure S8 Correlation between TFs and target genes (TPM).
Figure S9 Sequence alignment and expression profile of four IPS in cotton.
Figure S10 Comparison of HSFA6B-IPS binding among three genotypes.
Data S1
Table S1 Summary of the 23 cotton accessions, sequencing and mapping based on *G. hirsutum* reference genome.
Table S2 Pearson correlation coefficient (PCC) of replicates for each sample.
Table S3 Two-way ANOVA test of 10 traits.
Table S4 Two-way ANOVA test of 451 metabolites.
Table S5 Summary of DEGs derived from pairwise comparisons of 21 accessions.
Table S6 List of commonly up-/down-regulated genes in cotton panel.
Table S7 The top 10% MAD score genes classified into modules.
Table S8 Summary of TFs with enriched motifs over lint yield-correlated module.
Table S9 Table S6 Summary of lint yield candidate genes identified from previous studies.
Table S10 Summary of core genes regulated by HSFA6B and DREB2A in cotton.
Table S11 Summary of DAP-seq data processing.
Table S12 Information of HSFA6B DAP-seq peak.
Table S13 Information of DREB2A DAP-seq peak.
Table S14 Expression profile of genes targeted by HSFA6B and DREB2A (Log2FC score).
Table S15 Probe and competitor sequences used in EMSA.
Table S16 Genotype summary of 1024 *G. hirsutum* accessions.

14677652, 2024, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pbi.14402, Wiley Online Library on [27/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License