Comparative multi-omic analyses reveal gene regulatory networks controlling physiological and metabolic responses to drought in upland cotton

Li'ang Yu [1], Anna C. Nelson Dittrich [1], Giovanni Melandri [2], Aleksandra Skirycz [1], Kelly R. Thorp[3], Lori Hinze[4], Andrew D.L. Nelson* [1], Duke Pauli* [2]

[1] The Boyce Thompson Institute, Cornell University, Ithaca, NY 14850, USA
[2] School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA
[3] United States Department of Agriculture-Agricultural Research Service, Arid Land Agricultural Research Center, Maricopa, AZ 85138, USA
[4] United States Department of Agriculture-Agricultural Research Service, Southern Plains Agricultural Research Center, College Station, TX 77845, USA

*Address correspondence to:
Dr. Andrew D. L. Nelson
The Boyce Thompson Institute
533 Tower Road, Ithaca, NY, 14850.
Phone (979-220-0641)
an425@cornell.edu

or

Dr. Duke Pauli
The University of Arizona
School of Plant Sciences
1140 E. South Campus Drive, Tucson, AZ, 85721
dukepauli@arizona.edu

Current Email addresses:
LY: ly76@cornell.edu, ACND: an426@cornell.edu, GM: giovannimelandri@gmail.com, AS: as4258@cornell.edu, KRT: kelly.thorp@usda.gov, LH: lori.hinze@usda.gov

Running title: Regulation of plant performance under drought stress in upland cotton.

Abstract

Drought stress substantially impacts crop physiology resulting in alteration of growth and productivity. Although many drought-responsive genes have been identified, and to a lesser extent functionally characterized, how these genes link drought stress to critical stress-responsive crop traits is still unclear. In this study, the transcriptomes of 22 genetically diverse upland cotton accessions grown under well-watered and water-limited conditions in the Arizona low desert were sequenced. Six cotton fiber-related traits, four reflectance-based vegetation indices, and the levels of 451 metabolites were collected to estimate the physiological response of the accessions to stress. After data processing, a SNP-based phylogeny clearly clustered 21 accessions to three subgroups (Foreign, U.S., and Mixed), but the analysis of transcriptome profiles and physiological responses to drought highlighted differences incongruent with the accession's phylogenetic relatedness. Co-expression network analysis revealed that genes highly correlated to fiber yield, reflectance-based vegetation indices, and carbohydrate levels were also correlated with drought-responsive transcription factors (TFs). Based on integrated information from TF-binding sites enrichment tests and published functional studies, DREB2A was identified as a potential hub regulator that cooperatively works with HSFA6B to target a network of genes associated with improved fiber yield. Additional TFs, including bZIP53 and four ERFs, were identified as stress response regulators linking carbon metabolism, in particular sucrose and trehalose-6-phosphate, with nutrient cycling under drought stress. Thus, despite the close phylogenetic relationships between these accessions, some stand out as having developed distinct regulatory networks connecting drought stress and production-related responses. Understanding and modulating these pathways represent high value targets for improving cotton production under reduced water conditions.

Key words: RNA-sequencing, metabolomics, comparative genomics, transcriptomics, WGCNA, gene regulation, cotton, lint yield, stress response, drought

Introduction

Upland cotton (*Gossypium hirsutum* L.) is the world's top renewable textile fiber, supporting a multibillion-dollar industry with a global production of 120.2 million bales of cotton (~26 million metric tons, USDA, 2022). It is a major economically important crop for the United States and for Arizona, where upland cotton is planted on ~50,000 ha (USDA 2021), 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 United States is severely threatened by global climate. This 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; Burke and Wanjura, 2010; Zhang et al., 2021). 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.

Central to developing stress resilient elite cotton cultivars is understanding how genetic variation impacts the cellular functions that are responsible for maintaining homeostasis. This is challenging as connecting genotype-to-phenotype necessarily means traversing multiple levels of the phenotype space, from intracellular all the way to gross morphology scales. Cellular metabolism represents an endophenotype that effectively connects gene expression and regulation to external phenotypic traits. Metabolite levels have been described as more directly linked to the plant phenotypic traits than gene transcripts and protein levels (Fernie and Stitt, 2012). This is due to the fact that the metabolome represents a summary of complex and interconnected cellular omics layers (e.g. transcriptome and proteome) of regulation which also incorporate the effect of the dynamic response to the changes in the external environment (Großkinsky et al., 2015; Melandri et al., 2022). With regard to mechanisms of plant adaptation to heat and drought stress conditions, there are several morphological, physiological, and molecular responses that interact to ultimately drive changes in plant metabolism (Zandalinas et al., 2018; Fahad et al., 2016). Abiotic stress conditions lead to a reprogramming of plant metabolism to balance the need for growth with that of survival; this dichotomy means that essential metabolic functions must be maintained while adaptation to new environmental conditions occurs in parallel (Obata and Fernie, 2012; Claeys and Inzé, 2013). For these reasons, only an integrative approach that links

physiological/biochemical responses with gene regulation can effectively disentangle the complexity hidden in plant stress biology.

Similar to other crops, drought tolerance in cotton involves complex signaling pathways and transcriptional networks orchestrated by a number of transcription factors (TFs) and signaling proteins (Mahmood et al., 2019; Takahashi et al., 2020; Shinozaki and Yamaguchi-Shinozaki, 2007). These regulatory proteins are typically upstream of, or transcriptionally interconnected with, the genes necessary for coping with the stress and maintaining crucial metabolic pathways. Examples of typical regulatory targets include genes encoding for enzymes related to production of protective metabolites, transporters, chaperones, and lipid biosynthesis proteins (Malhotra and Sowdhamini, 2014; Singh and Laxmi, 2015; Gupta et al., 2020). In cotton, multiple TFs, such as GhNACs, GhDREBs, GhERFs, and GhWRKYs, have been associated with the drought stress response (Huang et al., 2009; Chu et al., 2015; Huang et al., 2013; Ma et al., 2017). Additionally, numerous candidate genes related to lint yield, fiber quality, and other agronomically-important traits have been identified and described in the literature (Fang et al., 2014; Sun et al., 2021; Ma et al., 2018). Despite this increased knowledge, the intricate nature of drought responsive gene networks has not been fully explored in cotton, particularly the potential crosstalk connecting drought responses to stress-induced variation in phenotypic traits. Filling this knowledge gap would generate a deeper understanding of drought tolerance/resilience in upland cotton and simultaneously enable the engineering and breeding of stress tolerant cotton cultivars.

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. Transcriptional data were integrated with metabolic data (451 metabolite levels), fiber quantity and quality traits (6 traits), and reflectance-based vegetation indices (4 indices), which were collected in a previous study (Melandri et al., 2021), to characterize each accession's drought stress response. A comparative analysis of differentially expressed genes (DEGs), fiber traits, vegetation indices, and metabolites was performed between the 22 accessions to study the impact of genetic diversity on these features. Further, co-expression network analyses were performed to classify and identify genes strongly associated with physiological and molecular phenotypes. Finally, TF-binding site, protein-protein interaction, and published functional data were overlaid onto co-expression networks to identify key regulatory and metabolic genes that likely maintain plant homeostasis and yield during drought stress.

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 two-year field experiment. In brief, a panel of 22 upland cotton (*Gossypium hirsutum* L.) accessions (**Supplementary Table 1**) were grown at the University of Arizona Maricopa Agricultural Center (MAC) in Maricopa, AZ, United States 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 a water-limited (WL) conditions, 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 leaf discs (0.64 cm in diameter) of the upper-most expanded leaf of five randomly selected plants within each plot were collected using a leaf punch in a single day during a time window of 3 hours (11:00-14:00). Leaf discs were stored in a 2 mL Eppendorf tube containing 1.5 mL of RNAlater (Fisher #AM7021) 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 fiber quantity and quality traits included lint yield (grams/plot), micronaire (Mic, units of air permeability), upper half mean length (UHM, inches), length uniformity (UI, percent), strength (Str, grams per tex), and elongation (Elo, percent). 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 extractions and RNA-seq library preparation.

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, 1% sodium deoxycholate. Just before use, PVP-40 was added to 2%, NP-40 was added to 1%, and DTT was added to 10 mM final concentrations. 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 ul hot buffer added to 5 leaf disks. Tissue was ground in hot buffer, then 15 ul of 20 mg/ml Proteinase K (Roche #46950800/Sigma #3115887001) per sample was added and tissue was ground again. A final 500 ul hot buffer was added before grinding a final time. Lysate was added to a Qiashredder column (Qiagen #79656) and centrifuged at 13,000 x g for 1 minute. Flow through was added to 0.5 volumes of 100% ethanol. Mixture was used as input for RNA cleanup using the RNeasy kit (Qiagen #74104) according to the manufacturer's instructions. Samples 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

Two replicates each of the 22 *G. hirsutum* accessions treated with either WW or WL conditions were sequenced using Illumina HiSeqTM 2500 (San Diego, CA) paired-end libraries. We used Trimmomatic to trim adapters and low-quality reads (Bolger et al., 2014). Further, reads were aligned to the G. *hirsutum* reference genome (Ghirsutum_527_v2.0, acquired from Phytozome, (Goodstein et al., 2012) 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; 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 reproducibility of replicates, we used 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. Further, pair-wise 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, significantly DEGs under WL condition were classified using thresholds adjusted p-value < 0.01, Log_2FC > 1 (upregulated) or < 1 (downregulated).

Kinship and phylogenetic analyses

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) before downstream analysis. The kinship of 21 accessions was estimated by TASSEL with default settings (Bradbury et al., 2007) and visualized using the R pheatmap package (https://cran.r-project.org/web/packages/pheatmap). Phylogenetic relationships were inferred using the IQ-TREE pipeline with filtered SNPs (Nguyen et al., 2015). Briefly, VCFs of all samples were transformed into phylip format by vcf2phylip **(**Ortiz, E.M. 2019) tools and a maximum-likelihood tree was constructed using (parameter: -nt AUTO -m MFP). The final phylogenetic tree was visualized 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)

Principal component analysis of transcriptomes and metabolites

To estimate the strength of treatment effects and the impact of phylogenetic relatedness on transcriptomic and metabolomic profiles, we performed a principal component analysis (PCA) 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/).

Characterization of traits, metabolites, and gene expression patterns

The effects of irrigation treatment (WL and WW), phylogeny (Foreign, U.S., and Mixed

accessions) and combined interaction terms on 10 plants traits and 451 metabolites were estimated by two-way ANOVA Test AICcmodavg R package: model: Traits ~ genotype + environments + genotype*environments) (https://cran.r-project.org/web/packages/AICcmodavg). The 451 metabolites were divided into six main classes including terpenoids, amino acid & peptides, polar lipids, neutral lipids, and carbohydrates for further statistical analyses. A pairwise *t*-test was used to examine variation in plant traits between WL and WW conditions per group. The average numbers of DEGs across three phylogenetic groups, as well as group-wide DEGs numbers, were calculated in order to estimate the potential correlation between genetic and transcriptomic variation across the three groups. Group-wide DEGs were classified as genes consistently up (or down)-regulated across all accessions under the same group. Group-wide DEGs not found to be differentially expressed in other phylogenetic subgroups were classified as unique group-wide DEGs.

Profiling fold changes of genes involved in major metabolic pathways

To profile drought-induced transcriptomic changes in genes known to be associated with metabolism, cotton genes were assigned to different functional metabolic categories using Mercator annotation tools (MapMan 4.0; https://mapman.gabipd.org/web/guest/app/mercator; (Thimm et al., 2004; Usadel et al., 2009; Schwacke et al., 2019). The 67,414 (out of 75,376 annotated protein-coding genes in the v2.1 genome annotation) successfully assigned genes were further divided into functional categories based on their association with primary or secondary metabolism. Genes and metabolic pathways heavily impacted by WL conditions were identified by comparing the standard deviation of transcript abundance between treatments across the 21 accessions, and then only keeping transcripts with high variance (SD > 1) for visualization.

Correlating transcriptomic changes to phenotypic changes under drought

The relationship between changes in gene expression and physiological/metabolic traits was examined by performing a linear regression analysis. Changes in transcript abundance (Log_2 fold-change, denoted as Log_2FC) between conditions was used as the explanatory variable, and the WW/WL ratio of BLUEs which derived from Melandri et al., 2021 for each phenotypic measurement was used as the response variable. Recursive pairwise Pearson correlation analyses and linear model fitting was performed for each gene-phenotype combination using the cor () and

lm () base function in R. Genes that had a combination of significant ($p < 0.05$) expression-phenotype correlations, high regression coefficients ($R^2 > 0.5$), and high effects (top 10% of all calculated effects) were chosen for detailed exploration into how these genes might influence drought stress responses.

Identification of trait-related co-expression modules

Weighted gene coexpression network analysis was performed 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, 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 in Cytoscape (Shannon et al., 2003).

Selection of trait-correlated genes and hub genes

For each highly trait-correlated module, genes with both high gene-significance (GS; cutoff: |GS| > 0.6-0.8 depending on module, see **Supplemental Table 12, 13, and 14**), which denotes correlation between traits and gene expression, and pronounced module membership (MM; cutoff: |MM| > 0.7-0.8 depending on module, see **Supplemental Table 12, 13, and 14**), which describes the correlation between module eigenvalues and gene expression. Hub genes were classified as those with a connection strength (weight score) and number of edges (degree) within the top 10% in each module.

Functional analysis of genes from different modules

Gene functional annotations of filtered genes (weight score > 0.1) in each module were obtained

from the cotton functional genome database (CottonFGD) to perform enrichment of gene ontology and KEGG pathways (https://cottonfgd.org/; (Zhu et al., 2017). Enrichment was performed by a Fisher-exact test with a significance level = 0.001 in the ClusterProfiler R package (**Yu et al 2012**). Transcription factors (TFs), transcription regulators (TRs), and protein kinases from modules of interest were identified using annotation tools from the iTAK database (Zheng et al., 2016). Potential protein-protein interactions (PPIs) among genes within the same module were defined using the STRING database (Szklarczyk et al., 2017) with a confidence level of 0.5. To identify potential regulatory interactions between TFs and target genes, promoter regions, defined as 1 Kb region upstream of each co-expressed gene within a module, were used to test for motif enrichment using the Analysis of Motif Enrichment (AME) pipeline (McLeay and Bailey, 2010) with *Arabidopsis* DAP-seq information (Malley et al., 2016) as the database (parameters: --scoring avg --method fisher --hit-lo-fraction 0.25 –evalue –kmer report-threshold 10.0, cutoff: TP values > 3, P-value < 0.001). Each of these data types (PPI, TF-binding, and co-expression) were overlaid onto integrated networks for visualization and functional analysis.

Results

Phylogenetic relationship, diversity, and kinship of cotton accessions

To better understand the molecular mechanisms underlying cotton's performance under drought conditions, 22 diverse upland cotton accessions from the Gossypium Diversity Reference Set collection were evaluated under water limited (WL) and well-watered (WW) conditions (Hinze et al., 2016, 2015). Leaf tissue from late season cotton plants was collected, in replicate, from plants grown in an incomplete randomized block t design. Following RNA extraction and mRNA enrichment, RNA-sequencing was performed, generating on average ~27 million PE x 150 bp reads. Reads were then mapped using RMTA (Peri et al., 2020) with an 84.75% average mapping rate to the G. *hirsutum* v2.0 reference genome (Chen et al., 2020); **Supplementary Table 1**). Gene level expression data were generated using *G. hirsutum* gene annotations as meta features in FeatureCounts (Liao et al., 2014; **Supplementary Table 1**). Pearson correlation coefficient (PCC) of normalized read counts revealed high correlation among replicates (average PCC = 0.904), with the exception of accession "Tipo Chaco" (PCC = 0.61, cutoff: 0.8) (**Supplementary Table 2**). Thus, this accession was discarded from further analyses resulting in 21 accessions for further analyses.

To determine the relatedness of each of these accessions, we used these RNA-seq data to perform variant calling relative to the reference genome. After filtering, variant calling generated a total of 108,396 SNPs for kinship analysis and phylogenetic analysis. A kinship matrix clustered replicates of the 21 accessions with the exception of one COKER_310 replicate, which clustered with VIR_7094_COKER_310 (**Figure. S1**). Based on the kinship matrix, the VIR_7153_D_10 accession showed the most variability relative to the reference genome and to the other accessions (**Figure. S1**). Further, a reconstructed phylogeny of the 21 accessions, rooted by VIR_7153_D_10, formed three major subgroups. Based on information from USDA GRIN-Global, we refer to the first of these three subgroups as "Foreign", as this group contains accessions developed primarily outside of the U.S. (Salmon box; **Figure 1a** and **Supplemental Table 1**). The second subgroup consists of "U.S." accessions of breeding lines and commercial cultivars developed in the U.S. (Teal box; **Figure 1a** and **Supplemental Table 1**). The last group, "Mixed", consists of commercial cultivars from the U.S. and improved breeding lines generally developed outside the U.S. (Lavender box; **Figure 1a** and **Supplemental Table 1**). Thus, these 21 accessions form three sufficiently distinct clades that will serve as a comparative framework to investigate the regulatory

mechanisms controlling the physiological response to drought in cotton.

Variation in physiological and molecular responses to drought in major subgroups

These phylogenetic groupings were used to determine the degree to which genotype and environment had an effect on physiological 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 fiber quality and agronomic traits and four vegetative indices, using two-way ANOVA (**Table 1**). Among these traits, three of the four vegetative indices were significantly affected by drought (**Table 1**; two-way ANOVA cut-off: $p < 0.05$, **Figure. 1b and Figure. S2a**). The treatment effect is particularly significant in the "Foreign" subgroups (t-test cutoff: $p < 0.05$). By contrast, long-term drought stress brought limited impacts to most fiber quality traits with exception of the significantly reduced micronaire (two-way ANOVA: $p = 0.002$) observed in "Mixed" and "Foreign" subgroups (t-test: $p = 0.015$ and 0.047, **Figure. 1b, Figure. S2b, and Table 1**). In contrast to treatment effects, pronounced genotypic effects were observed between major subgroups ("Foreign, U.S. and Mixed) for five of the six fiber quality traits but not for the four vegetative indices (**Table 1**).

While significant genotype effects based on the subgroups inferred from our phylogeny were observed, intragroup comparisons of responses to drought revealed stark differences in performance. For instance, each subgroup contained one or more accessions that significantly outperformed close relatives, particularly in fiber quality traits (**Figures 1b and S2c**). The lack of phylogenetic congruence in observed traits suggests that lint yield and drought associated traits may have been selected for differently in even closely related accessions. Importantly, the presence of outperforming accessions implies that this panel may be useful in identifying regulatory crosstalk between lint yield and drought stress.

Next, changes in molecular phenotypes, as quantified by metabolite profiling, were assessed to understand the effects of drought stress. Changes in the metabolome of the panel were measured by re-examining changes in the levels of 451 metabolites obtained from paired leaf samples previously analyzed in Melandri et al., 2021 (27 GC-MS and 424 LC-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 importantly 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 (**Supplementary table 4**). Drought treatment significantly impacted more than 95% of metabolites, but only 5% of the metabolites exhibiting variations attributed to genotypic effects; around 25% of metabolites exhibited changes that could be associated with G*E effects ($p < 0.05$, **Table 2**). Among the 432 metabolites associated with significant treatment effect, 273 were up- and 159 down-regulated by drought. Among these metabolites are the known osmoprotectants glycine and proline (Fang et al., 2015); **Supplemental Table 4**). Even though 22 metabolites were associated with significant ($p < 0.05$) genotypic effects, these metabolites did not cluster as expected based on inferred relationships between the accessions (**Figure. 1e**). Thus, these data argue that alteration of specific metabolites associated with drought tolerance (osmoprotectants) is a common approach across our panel.

We examined how the transcriptome has changed in these accessions in response to drought stress. As expected based on metabolomic profiling, a PCA revealed that 68% of variation could be explained by the irrigation treatment (**Figure 2a**). In addition, neither PC 2 nor PC 3 could be attributed to differences in genotype. Pairwise comparisons of gene expression under the two conditions generated a range of differentially expressed genes (DEGs) for each accession **(Supplementary Table 5)**. The mixed accessions exhibited the highest average numbers as well as the highest intragroup variation of DEGs (both up and down-regulated, adjusted p-value < 0.05; **Figure. 2b**). In contrast, the seven foreign accessions displayed, on average, the smallest number of DEGs. Next, data were intersected to determine if there were DEGs shared between all intragroup accessions, or between all accessions within multiple subgroups, to identify genes associated with a shared drought response. Out of the thousands of differentially expressed genes in each accession, there were few genes that were shared among all accessions within a subgroup, or between subgroups (**Figures. 2c and 2d; Supplementary Table 5**). The foreign accessions shared the least group-wide DEGs (up-regulated: 57 and downregulated: 86) whereas the mixed accessions featured the most (up-regulated: 194 and down-regulated: 177) (**Figure. 2c**). In contrast, there were 300 DEGs (194 upregulated, 106 downregulated) shared among all examined accessions. GO enrichment of the up-regulated genes representing the shared response to drought stress revealed an over-representation of stress response and phosphorylation signal transduction process genes while the commonly down regulated genes were involved in fatty acid biosynthetic processes and transport processes (**Figure. 2d**). Thus, these consistently differentially expressed

genes may represent a core set of genes involved in drought stress in this panel.

Identifying candidate large effect genes associated with the drought stress response

Thus far, no strong relationship between metabolic or transcriptomic responses and phenotypes were observed that were congruent with the inferred phylogeny. We next asked if there were any transcriptional responses that could be linked to phenotypes or metabolite levels, irrespective of genotype. To identify linked responses, an all-way Pearson-correlation and regression analysis was performed of expression change (represented by $\log_2$FC of WW/WL conditions) with the WW/WL ratio of production related traits, vegetative indices, and metabolite levels. For gene-trait/metabolite pairs with significant correlations ($p < 0.05$) we plotted the change in expression vs trait for each accession. The slope of the best-fit line (blue line, **Figure 3a**) explaining these correlations was used as a proxy for "effect", whereas the R^2 for this line reflects how much of a particular effect could be explained by a specific gene (i.e., "contribution"; **Figure 3a**). Using this approach, 415,154 gene-trait pair correlations derived from 59,421 genes were generated (**Figure 3b**). The density distribution of the log2 transformed absolute values for effects revealed limited variation in distribution among vegetative indices, with the exception of CRI (**Figure 3c,** left). In contrast, production-related traits exhibited much more variation between traits (**Figure 3c,** center). Though the six classes of metabolites were featured by limited variation in their respective effects (**Figure 3c**, right), effects of gene-metabolites correlations were significantly higher compared with production traits and indices (**Figures 3b and 3c**, $p < 2.2e\text{-}16$), suggesting there are several genes with large effect on metabolic processes in cotton. Of these thousands of gene-trait pairs, we determined that substantially fewer could be considered to have a significant contribution ($R^2 > 0.5$; **Figure 3d**).

Consequently, we sampled the gene-phenotype correlations with both large effect (top 10% absolute values) and large contribution ($R^2 > 0.5$) for the three classes to identify high-effect trait-related genes. This filtering resulted in 16 genes associated with vegetative indices, 29 with production-related traits, and 2,340 with metabolite abundance. All 16 of the genes with a large effect and contribution to vegetative indices were associated with CRI (**Supplementary Table 7**). Within these 16 genes, of particular interest is the zinc-finger homeodomain containing transcription factor (TF), *ZHD8* (Gohir.D04G049000), which showed a strong positive relationship with CRI under drought conditions. The ZHD family of transcription factors has been

reported to regulate carotenoid biosynthesis in other systems (Shi et al., 2021). In addition, more than half of the production associated genes are related to lint yield (58%, **Supplementary Table 8**). Among these genes are *BIR2* (Gohir.A05G362600), which is associated with plant immunity in Arabidopsis), Gohir.D13G115251, a subunit of the 20S proteasome, and a fork-head domain containing TF, Gohir.A05G379500, that is orthologous to an *Arabidopsis* gene purportedly involved in fruit development. Lastly, among the 2,340 genes correlated with metabolite abundance (**Supplementary Table 9**), the two with the strongest association are Hexokinase 1 *HXK1* (Gohir.D13G201600), which is a glycolytic enzyme and glucose sensor (Dekkers et al., 2004) and acyl-coenzyme A oxidase (Gohir.A12G215400), which is involved in amino acid metabolism (Hayashi et al., 2002). Thus, while further studies are necessary, these large effect genes may be ideal candidates for engineering more efficient drought responses in cotton.

Understanding the transcriptional variability of genes involved in metabolism

Given the preponderance of genes whose expression significantly correlated with metabolite levels, we further characterized changes in metabolism associated transcripts with the metabolic pathway analysis tool, MapMan (Thimm et al., 2004; Usadel et al., 2009; Schwacke et al., 2019). A total of 62,976 out of 67,413 expressed genes were assigned to functional bins with MapMan categories. Of these, 11,418 genes were characterized into metabolism related categories. Among the 5,062 genes involved in primary and secondary metabolism, we identified significant differences in variation of gene expression across the 21 accessions (determined by per-gene standard deviation of $\log_2$FC; **Figure S3a**). High variation in response to drought suggests that these genes, and their associated metabolic pathways, are involved in the drought response in some, but not all, accessions. Genes involved in amino acid, cell wall, and lipids metabolism presented a pronounced standard deviation in gene expression (average SD > 1; **Figure S3a**). Further, we examined the expression pattern of genes with high standard deviation (SD > 1) within different categories of primary and secondary metabolism (**Figure S3b**). Despite displaying minimal variation between accessions, a large number of genes related to carbohydrate metabolism, including uracil-diphosphate glucose (UDP), trehalose, sucrose, and glycolysis pathways were upregulated (**Figure S3b**). In contrast, all starch synthesis genes were downregulated in response to drought (**Figure S3b**). In addition, the majority of amino-acid metabolism related genes were also down-regulated (**Figure. S3b**). For secondary metabolism,

more than half of genes involved in flavonoid and phenol-related pathways were upregulated in drought conditions (**Figure. S3c**). In sum, while the results above reveal a strong connection between metabolic and transcriptomic responses to drought stress, they did not uncover a single key pathway associated with a positive response (e.g., sustained yield).

Using WGCNA to examine transcriptome-phenome connections

The above correlative analyses revealed several genes that may be associated with physiological responses to drought stress. To determine if we could take advantage of correlations between transcriptomic and phenotypic variability to uncover gene networks regulating these phenotypes, we performed a weighted gene co-expression analysis (WGCNA). 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 = 4,552 genes). We obtained 22 modules that satisfied a scale-free topology ($R^2 = 0.86$; **Figure S4a**) 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 S4b and S4c).** From these 22 modules, three correlated with the phenotypic data. One was positively correlated with lint yield (Turquoise, $r = 0.68$, $p = 7e\text{-}04$; **Figure S4d**), one was negatively correlated with lint yield (Cyan, $r = -0.61$, $p = 0.005$, **Figure S4d**), and one positively correlated with both CRI and WI/NDVI (Green Yellow, $r = 0.58$ $p = 0.006$, $r = 0.53$, $p = 0.01$, **Figure S4d**). In addition, a total of nine modules were positively or negatively correlated with 20 metabolites (using cut-off: $r > 0.6$ or $r < -0.6$; **Table 3a and 3b**). These metabolites with high module-trait correlations were primarily classified as terpenoids, carbohydrates, and lipids (**Supplementary Table 10**). In sum, 11 modules of similarly expressed transcripts were identified that may provide insight as to how cotton maintains productivity under drought stress.

Functional analysis of trait-related gene modules

Lint-yield related turquoise module

To better understand the transcriptional networks associated with drought in our 21 cotton accessions, we first focused on the production-related modules. Despite clear indicators of stress, we observed a high degree of variability in yield under drought treatment (**Figures 1b and 1c**), suggesting mechanisms to maintain (or increase) yield under stress in our panel. Within the

Turquoise module, which was positively correlated with yield, a gene network of high confidence transcripts was developed by first filtering by weight score (cutoff > 0.1), leaving 866 out of 1,044 nodes (genes) connected by 134,205 edges. Trait-related genes were then identified using both the significance of a gene's association with the trait in question (lint yield), denoted as gene significance (GS), along with how well the gene's transcriptional profile matched the module (module membership. Using a GS > 0.7 and a MM > 0.8, 13 trait related genes were identified (**Figure 4a**). As expected based on their position within the module, a hierarchical heat map clearly clustered the expression profiles of these 13 genes into three groups (**Figure 4b**) with a significant positive correlation with lint yield (**Figure 4c**). Interestingly, more than half of these trait-related genes were associated with stress and/or heat responses, such as heat shock proteins (HSPs) and protein chaperones (CPNs; **Figure 4b**), indicating stress response genes are critical for maintaining yield.

To develop networks, hub genes were first identified within the 866 genes remaining within the module. A total of 311 genes were identified by keeping only those in the top 10% of connections (edges). A gene ontology (GO) enrichment analysis revealed a significant enrichment for stress response, protein folding, and translation-related processes in this group of genes. While the stress responsive hub genes were primarily HSPs or CPNs (**Supplementary Table 11**), these enriched GO terms included 24 highly connected TFs (Red squares, **Figure 4d**). In addition, this set of 311 genes contained a gene previously reported to be strongly associated with yield under non-drought conditions, Inositol-Phosphate Synthase (IPS; yellow hexagon, **Figure 4d**; Su et al., 2016). For visualization purposes, a child network was generated with approximately half of the genes, keeping only the most connected (n = 153, **Figure. 4d**). The full network of the 311 hub genes, with their enriched GO terms, can be found in **Supplementary Table 11**. In addition to the 13 trait-related genes, the lint yield candidate gene, IPS, and the 24 TFs mentioned above, the network also included 29 genes involved in RNA splicing (ko:03040), 20 genes involved in pathogen pathway (ko04626), 58 genes related to protein folding (GO:0006457), 29 genes to heat and stress response (GO:0006950, GO0009408), and 27 genes to metabolic process, protein process, and translation (GO:0019538, GO:0016485, GO:0006413).

To further explore the regulatory network within this module, TF-promoter binding data (DAP-seq) from *Arabidopsis* was used to infer promoter binding sites for each of the 24 TFs in the module with homologs in the *Arabidopsis* dataset (Malley et al., 2016). A search for

corresponding enriched TF-binding motifs within the promoter elements of the 153 genes within the network revealed 30 genes with the DREB2a dehydration response element (DRE; **Figure 4e**) enriched in their promoter regions. Aside from multiple stress-associated genes, the lint-related gene IPS also contained DREB2a binding sites within its promoter, providing the first direct regulatory connection between drought response and lint yield in our data. The lack of ABA-specific TFs in this network, as well as the generalized expression of the ABA-response element binding TFs in these accessions (**AREBs; Figure S5)** suggest this network, and DREB2A under these circumstances, are ABA-independent. The heat shock element (HSE) recognized by HSFA6B was also enriched within the promoters of 20 genes in this network (**Figure 4e**). Intriguingly, there is a significant positive correlation between normalized expression (TPM) and lint yield under drought for DREB2a, HSFA6b, and IPS ($p < 0.05$; **Figure 4f**). Thus, by overlaying TF-binding data onto the network, it is apparent that DREB2A and HSFA6b are both key regulators facilitating increased yield under drought stress.

Both DREB2a and HSFA6b have large edge numbers (418 and 526 respectively, **Supplementary Table 9**) as well as some potential regulatory overlap. Some genes, such as HSPs HSP18.3, HSP70, and HSP70-14 are likely regulated by both TFs as they contain both TF binding motifs in their promoters and are connected to both TFs in the network (**Figures 5a** and **4d**). The DRE-specific, and thereby likely DREB2a-regulated, genes include *DnaJ* protein, small and large HSP units, RCD1, and the aforementioned lint yield associated gene, IPS (**Figure. 5b)**. The list of HSFA6b regulated genes also includes HSPs, a different DnaJ1 chaperone protein, as well as three cotton genes without associated GO terms. A reciprocal best BLASTp (protein BLAST) determined that two of these three proteins were also mitochondrial heat shock proteins (denoted as *MtHSP* in **Figure 5c**), whereas the third was 87% identical to an *Arabidopsis* Calreticulin protein (*AtCRT1a*) associated with ER stress and calcium ion homeostasis (Christensen et al., 2008). In sum, a detailed analysis of the Turquoise module revealed a number of strong candidate genes for improving yield under drought stress (**Figure. 5**).

Modules associated with vegetative indices

We also examined the green-yellow module that is positively correlated with CRI and WI/NDVI (**Figure S6**). This module was much smaller than the Turquoise module, only containing 58 genes after filtering for MM and GS similarly as above. Within this module, there is a single TF,

NAC029, which features the highest edge numbers in the module and is connected to nine trait-related genes (**Figure S6**). NAC029 has been shown to confer enhanced drought tolerance in soybean and wheat (Wang et al., 2015; Huang et al., 2015). While no DAP-seq TF motif data were available for NAC029, it is connected in this network to a group of abiotic stress-responsive DNA demethylase family members and Calmodulin-like Ca^{2+} signaling proteins (GDSL and CML; Blue circles, **Figure S6**). Thus, this module may reflect a regulatory pathway associated with responding to water movement during drought stress.

Metabolite related modules

We also examined modules which displayed strong correlations with metabolites. There were two modules whose expression correlated with sugars, which are not only critical nutrient sources but also can play roles in plant stress responses (Rolland et al., 2006; Delorge et al., 2014). The Magenta module represents a glucose and photosynthesis-associated module of genes. This module is negatively correlated with glucose content (***r* = -0.61, Table 3b**). A functional enrichment analysis of the 75 genes within this module revealed that a majority of them were associated with photosynthesis, light harvesting, and carbon fixation (**Supplementary Table 13; blue circles, Figure S7a**). Within this module we identified four ethylene response TFs (ERFs) which are known to be involved in drought and heat responses. While there were no available TF binding motif data for these particular ERFs from *Arabidopsis* orthologs, they were connected, and clearly co-expressed, with photosynthesis and light harvesting genes in this network (Xie et al., 2019; **Figure S7a**). Protein-protein interaction (PPI) data were available for a large subset of the photosynthesis-related genes within this network, suggesting we were recovering entire complexes under similar regulation. Specifically, 17 genes involved in light reaction pathways exhibited differential expression among 21 accessions, indicating diversified levels of expression of light reaction upstream genes in response to drought treatment. (**Figure S7a**). Moreover, several genes in the Magenta module, including chlorophyll a/b binding proteins (CABs), Rubisco small subunit (RBCSs), and triose phosphate translocators (TPPTs) were negatively correlated to glucose content (**Figure. S7b**) and these genes were reported as sugar-repressible genes encoding photosynthesis components (Pego et al., 2000, **Figure S7C**). Thus, the observed ERFs may play a role in either increasing glucose utilization or shunting carbon metabolism towards other carbon intermediates during drought stress.

The largest and most connected of the sugar-associated modules was the Salmon module, which was positively correlated with levels of Norbornane-2-carboxylic acid (terpenoid), as well as negatively correlated with sucrose levels (**Table 3a and 3b**). The Salmon module consisted of 70 genes and 581 edges (**Supplementary Table 14**). The module membership of these 70 genes is highly correlated with both terpenoid and sucrose content (r =0.86 and r = -0.75). A total of 19 and 28 were identified as terpenoid and sucrose metabolite-related genes, respectively. Seventeen of these genes were strongly associated with both metabolites (|MM| > 0.8, |GS| > 0.7). K-means clustering of expression for these 17 genes were divided into three clusters (**Figure 6a and 6b**). As expected, these three expression clusters corresponded to antagonistic patterns of metabolite enrichment for sucrose and terpenoids. For instance, the most highly expressed genes corresponded to the highest levels of terpenoids and the lowest levels of sucrose. While trehalose-6-phosphate, a known osmolyte, was not directly measured, functional enrichment analysis revealed over-representation of trehalose biosynthesis genes (GO: 0005992) and enrichment in the KEGG pathway associated with starch and sucrose metabolism (Ko00500). Interestingly, three of the 17 genes associated with both traits were trehalose-6-phosphate synthases (TPS), which are integral to plant stress responses (Ponnu et al., 2011; Leyman et al., 2001). Finally, the list of trait-related genes also includes five stress responsive genes: two HVA22E and two OXS3 genes. HVA22E genes are part of a class of ER-associated, ABA/stress-induced proteins which negatively regulate plant cell death during abiotic stress (Guo and Ho, 2008) and OXS3s are oxidative stress-responsive genes which repress ABA signaling pathways (Xiao et al., 2021). Not only are these stress-responsive genes associated with other trait-related genes in this network, but they are co-expressed and highly correlated with TPS9 expression. Using a similar approach as above, we examined the sucrose/terpenoid network for TFs with connections to other genes that contained the appropriate TF motif in their promoter elements. Promoter regions of 36 genes were enriched with specific binding motifs of one transcription factor bZIP53 (blue arrows, **Figure 6c**). In particular, the bZIP53 (**Figure 6d**) motif was enriched in the promoter regions of one TPS9 and one OXS3 gene, and this bZIP53 TF was highly correlated with both content of both sucrose (negative correlation, $p < 0.05$) and terpenoids (positive correlation, $p < 0.05$**; Figure 6e**). Thus, we proposed that bZIP53 may be a key regulator in connecting oxidative stress responses with sucrose, trehalose-6-phosphate, and terpenoid metabolism in drought treated cotton.

Discussion:

Understanding the genetic interplay underlying plant performance under drought stress:

Understanding the molecular regulatory mechanisms connecting plant performance and drought response are crucial for developing drought-tolerant cultivars with minimal impacts on yield (Alizadeh et al., 2020; Sinclair, 2011). Here we took a comparative multi-omics approach to study the molecular and physiological tradeoffs that occur under reduced water availability in a diverse panel of field-grown cotton. Using this genetic diversity, we identified a cohort of genes strongly correlated with different drought-associated physiological responses. We also identified sets of co-expressed genes that were involved in fundamental biological processes, allowing for the inference of regulatory networks maintaining metabolic status quo during stress. Despite the reduced genetic variation existing in *G. hirsutum* (Chen et al., 2020) and within our panel, regulatory networks overlaid with public DAP-seq and PPI information identified target TFs whose expression profiles were associated with improved yield and increased metabolic function under reduced water conditions. Variation in expression of these TFs

Even among close relatives, cotton accessions display high variation in phenomic and transcriptomic responses to drought:

Characterizing genotype-specific molecular responses under water-limiting conditions is challenging due to the genetic complexity underlying quantitative physiological traits (Welcker et al., 2011; Tardieu et al., 2011, 2014). 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 (Tardieu et al., 2018; Joshi et al., 2016). In cotton, this complexity is further compounded by a recent (~1.0-1.6 Ma) allopolyploidy and strong human selection within the last 8,000 years (Chen et al., 2020). As genome sequence data are not available for all of these accessions, we inferred a phylogeny using transcriptome-based SNPs as a framework for further comparisons. Although this approach is incapable of identifying genetic differences outside of transcribed regions, three clades were resolved, largely based on geographic origin and extent of genetic improvement. Surprisingly, given other unified responses to drought stress among more phylogenetically distant relatives in the grasses (Chavez Montes et al., 2022), this transcriptome-based phylogeny was incongruent with patterns of metabolite or transcript abundance. Although there were clear treatment responses, very few genes were incorporated into a conserved, panel-

wide, response to drought stress. Disparate transcriptional responses were evident even among sister accessions, suggesting that there are ample genetic differences not directly captured in our RNA-sequencing data (e.g., promoter elements) driving this transcriptomic variation. The lack of a conserved response may also reflect the tropical (hot and humid) environment in which cotton evolved. In the absence of an evolutionarily conserved response, it appears that each of these accessions are responding to water-limited conditions through a number of different genetic mechanisms, some of which are beneficial to production-related traits.

Identifying the genetic mechanisms most closely associated with improved production is possible by inferring gene regulatory networks that utilize the transcriptomic variation within the panel. This transcriptomic variation, when associated with phenotypic data, can be used to identify individual genes with an outsized effect on traits of interest as well as cohorts of similarly expressed, trait-associated genes. Using similar approaches to those performed in maize, cacao, and rice (Lei et al., 2015; Do et al., 2013; Bae et al., 2008), we identified a number of genes that exhibited strong correlations with changes in plant performance traits and metabolite abundance. Among the list of large-effect genes were orthologs of several functionally studied genes including ZHD8, BIR2, HXK1, that might be ideal candidates in cotton breeding programs (Imkampe et al., 2017; Shi et al., 2021; Dekkers et al., 2004). Additionally, we identified several co-expression modules strongly correlated with lint yield, carbon flux, and vegetative indices. Enriched within these modules were genes involved in drought and heat stress responses, water deprivation response, and sucrose metabolism. These expression modules highlight the potential crosstalk between stress response genes and downstream physiological response.

Identification of key transcriptional regulators governing lint yield and carbon metabolism in cotton under drought:

The contrasting transcriptomic variation and coding sequence conservation in our panel suggests that transcriptional regulation by TFs may be key to influencing physiological responses to drought. An integration of TF-binding and PPI data, as well as functional characterization of orthologous TFs, identified factors governing potential crosstalk between lint yield and carbon flux in cotton in response to drought. Dehydration Responsive and Heat Shock Elements (DREs and HSEs) were enriched in the promoters of several lint-yield associated genes, including three HSPs (HSP18.3, HSP70, and HSP70-14) that were co-expressed with TFs DREB2A and HSFA6B.

This regulatory link has been reported previously in sunflower (Prieto-dapena et al., 2005; Almoguera et al., 2009), suggesting these two TFs may represent a commonly sampled genetic mechanism for regulating drought stress responses. DREB2A appears to directly regulate the expression of a number of drought and heat responsive genes that influence production and vegetative traits, and itself exhibits a significant correlation with lint yield. This observation is further supported from work in Arabidopsis, where AtDREB2A overexpression leads to enhanced stress tolerance by modulating the expression of a similar suite of heat shock factors (Sakuma et al., 2006). The direct regulatory connection between DREB2A and lint-yield associated gene IPS – and the strong positive correlation between both of these genes and yield, suggests that these two genes are prime targets for improving yield under water limiting conditions.

We also observed a potential regulatory mechanism governing carbon metabolism under water limiting conditions. Broadly, we identified two modules suggesting that drought stress results in decreased sucrose and glucose levels, and an increase in the expression of trehalose-6-phosphate (T6P) synthesis associated genes. T6P is not only an important storage carbohydrate, but has also been implicated in drought stress responses in a number of plants (Lin et al., 2019; MacIntyre et al., 2022; Delorge et al., 2014). Interestingly, the T6P synthases (TPSs) found to be regulated by the bZIP53 transcription factor, and inversely correlated with sucrose content, were the class II TPSs, which are themselves believed to have a regulatory rather than T6P synthesis role in plants. There are multiple reports of these regulatory TPSs being transcriptionally induced by stress and, at the genetic level, impacting stress responses in rice and Arabidopsis (Li et al., 2011; Miranda et al., 2007). T6P levels themselves need to be carefully regulated, as inappropriate levels in certain tissues may disrupt sink/source carbon flow and result in decreased biomass and altered leaf morphologies (Karim et al., 2007). Thus, bZIP53 in cotton may be pivotal in both regulating TP6 accumulation, but also in connecting this metabolic process to the nutrient-cycling processes governed by HVA22E and OS3. In sum, we propose that the GhbZIP53 ortholog is a key transcriptional regulator of central metabolism during drought stress in cotton, and that by modulating carbon assimilation in the leaf, this pathway can also influence lint yield.

Author Contributions and Acknowledgements:

LY, GM, ADLN, and DP developed the project. AND performed RNA extractions and library prep, GM performed metabolomics analysis. KRT contributed to field data collection. LY, GM,

and AS performed the analyses. 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 approve of the manuscript. The authors would like to acknowledge support from the NSF-IOS #2023310 (awarded to ADLN), NSF-IOS #2102120 (awarded to DP and ADLN), Cotton Incorporated Core Projects 18-384 and 20-720 (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 thank other members of the Nelson, Skirycz, and Pauli labs for helpful discussion in the formulation of the analyses presented herein.

Tables

Table 1 Effects of phylogeny, environment, and interaction terms to 10 traits among cotton accessions from two conditions (two-way ANOVA). NDVI = normalized difference vegetation index, CRI = carotenoid reflectance index, sPRI = scaled photochemical reflectance index, WI/NDVI = the ration between water index and NDVI.

Traits	Phylogeny (P)	Environment (E)	G*E
Lint yield	0.0286 *	0.06119	0.8886
Micronaire (Mic)	0.0335 *	0.0020 **	0.8549
Upper-half mean length (UHM)	0.00978 **	0.76844	0.81666
Length uniformity (UI)	0.0692	0.8343	0.9569
Strength (Str)	0.0155 *	0.1607	0.8218
Elongation (Elo)	0.308	0.769	0.985
NDVI	0.752	2.66e - 6 ***	0.298
sPRI	0.0677	0.1214	0.977
CRI	0.953	1.71e-11 ***	0.679
WI/NDVI	0.854	4.6e-05 ***	0.311

Table 2 Numbers of metabolites associated with significant effects of phylogeny, environment, and interaction terms (two-way ANOVA). NDVI = normalized difference vegetation index, CRI = carotenoid reflectance index, sPRI = scaled photochemical reflectance index, WI/NDVI = the ration between water index and NDVI.

Metabolites Class	Total counts	Phylogeny effects counts	Treatment effects counts	Interaction effects counts
Amino acids & peptides	14	2	14	1
Carbohydrates & conjugates	10	3	8	0
Neutral lipids	71	2	69	24
Polar lipids	144	8	140	32
Terpenoids	45	0	42	12
Other	167	7	159	42
Total	451	22	432	111

Table 3a The top 10 metabolites positively correlated to gene modules

Name	Main Class	Correlation	Module
Norbornane-2-carboxylic acid, 2-amino-	Terpenoids	0.86	salmon
Glutaric acid, 2-hydroxy-	Other	0.85	greenyellow
Docosanamide	Other	0.81	midnightblue
Citric acid	Other	0.78	greenyellow
Diethanolamine	Other	0.72	lightcyan
Sucrose	Carbohydrates	0.71	lightgreen
TG(18:0/20:0/20:4(5Z,8Z,11Z,14Z))	Neutral lipids	0.69	purple
N-gondoylethanolamine	Other	0.64	midnightblue
Pipericine	Other	0.64	midnightblue
Glucose, 1,6-anhydro-, beta-	Other	0.64	greenyellow

Table 3b The top 10 metabolites negatively correlated to gene modules

Name	Main Class	Correlation	Module
Norbornane-2-carboxylic acid	Terpenoids	-0.76	lightgreen
Sucrose	Carbohydrates conjugates	-0.75	salmon
(3S,3'S,5R,5'R,6R)	Terpenoids	-0.71	magenta
Glucose, 1,6-anhydro-, beta-	Other	-0.68	tan
2-Hexaprenyl-3-methyl-6-methoxy-1,4 benzoquinone	Other	-0.67	magenta
2,2'-Diketospirilloxanthin	Terpenoids	-0.66	green
Neoxanthin	Terpenoids	-0.65	magenta
Glucosylceramide (d18:1/24:1(15Z))	Polar lipids	-0.64	NA
DG	Neutral lipids	-0.64	greenyellow
Capsorubin	Terpenoids	-0.63	magenta
Xylobiose, D-	Carbohydrates conjugates	-0.62	magenta
Diadenosine heptaphosphate	Other	-0.61	magenta
Glucose	Carbohydrates conjugates	-0.61	magenta

Figure Legends

Figure. 1 Phylogeny, traits variations, and metabolite profiles of 21 distinct cotton accessions
A) A phylogeny of the accessions used in this study. The three groups, "Foreign", "U.S.", and "Mixed", reflect historical breeding information obtained from USDA GRIN-Global. B) Pairwise comparisons of two representative fiber quality traits (Lint yield and micronaire; Mic) from well watered (WW) and water limited (WL) conditions. C) Pairwise comparisons of two representative vegetative indices (carotenoid reflective index; CRI, and the ratio between the water index and the normalized difference vegetation index; NDVI), from two water conditions. Significance values for B) and C) were calculated using a parametric t-test in response to drought. D) Three-dimensional PCA of 451 metabolites derived from normalized metabolite content. Legend denotes accessions belonging to three phylogenetic groups and samples from WL and WW conditions. E) Z-score based heatmap of 22 metabolites belonging to five classes (see legend) with significant genotype effects (factor: Foreign, U.S., and Mixed, $p < 0.05$).

Figure. 2 Transcriptome comparison of expression profiles in the three phylogenetic groups
A) Three-dimensional PCA of the top 10% most variable genes (n = 3,768) derived from normalized read counts. B) The average numbers of differentially expressed genes (DEGs; down-regulated on left, up-regulated on the right) derived from per-accession pair-wise comparisons between the two conditions (WW and WL). An all way pairwise parametric t-test was performed to determine significance. C) Upset plot of genes that were differentially expressed in all representatives within a particular phylogenetic group. The vertical black bars stand for numbers of unique DEGs from each group, intersection between each two group, and intersection among three groups. D) Enriched gene ontology terms of commonly down-regulated (top) and up-regulated (bottom) genes among 21 accessions. The calculation is based on -log10 Q values (cutoff:3).

Figure. 3 Identifying genes making large contributions to phenotype
A) Illustrative scatter plot of the model used to identify genes making a large contribution to phenotype. A Pearson correlation between log2-transformed fold changes of gene expression and ratio of phenotypic measurements between the two conditions was performed for all phenotypic and metabolomic measurements. The X-axis depicts the explanatory variable, changes in gene

expression from two watering conditions, whereas the Y-axis depicts the responsive variable of phenotypic changes potentially regulated by gene expression. B) Boxplot comparison of log2-transformed absolute values of effects (Δy/Δx) for each significant ($p < 0.05$) pairwise linear regression among metabolites (M), production-traits (P), and vegetative indices (**V**). C) Density distribution of log2-transformed absolute values of effects (Δy/Δx) derived from pair-wise correlations of Log_2FC to phenotype ratio (WW/WL) represented by sub-categories (see legends) among metabolites (M), production-traits (P), and vegetative indices (V). D) Boxplot comparison of R^2 value of high-confidence ($p < 0.05$) Pearson correlations. The red dash line indicates the cut-off of correlations with high contribution rate ($R^2 > 0.5$), numbers of correlation pairs above the cut-off and their respective percent of total correlations are denoted alongside each boxplot.

Figure. 4 Network analysis of a module of genes associated with lint yield

A) The module membership (MM) scores and gene significance (GS) for genes contained within the lint-yield module were plotted with best fit determined by a linear model. The dashed red lines indicate the two cutoffs used to identify trait-related genes. B) Clustered heat map based on Z-score transformed TPM values across 21 accessions of trait-related genes selected from the network. Both row and column data were clustered by "correlation" distance clustering methods from pheatmap function in R. Red dashed line depicts the cutoff used to define clusters in C. C) Box and whiskers depicting lint yields for three different expression clusters of lint-yield associated genes. Significance values were derived from pair-wise t-tests. D) The integrated child network derived from 153 out of 866 genes within the lint-yield correlated turquoise module. Gene categories are shown in the bottom left, whereas connection types are shown in the top right. KEGG pathways are grouped with a red dashed oval. E) Motif logo of DREB2A and HSFA6B TFs, derived from Analysis of Motif Enrichment (AME), found within the promoter regions of genes within the lint yield module. Genes connected to DREB2A and HSF6AB, and containing the respective promoter binding motifs, are depicted as blue dashed lines in D. F) Pearson-correlation between lint yield and DREB2A, HSFA6B, and IPS gene expression (TPM) across 21 accessions.

Figure. 5 Proposed regulatory network of lint yield associated genes

Proposed regulatory network of DREB and HSF gene under water limited conditions. A) Depiction of three genes within the lint yield module that contain promoter binding motifs, and are connected in the module, to both DREB2A and HSFA6A. Log2 Fold Change values are shown for water limited conditions relative to control. B) Expression of genes likely to be specifically modulated by DREB2A, as they contain a DRE motif and are connected in the module to DREB2A. C) As in B), genes that look to be specifically modulated by HSFA6B. DRE = dehydration response element, HSE = heat shock element. Genes in B) and C) are divided based on major biological functions (shown in red on right).

Figure. 6 Network analysis of sucrose and trehalose-6-phosphate metabolism module

A) Cluster heat map based on Z-score transformed TPM values across 21 accessions of 17 trait-related genes that were selected from the parent network based on GS and MM as in Figure 5. Where possible, gene names based on ortology to Arabidopsis gene IDs are shown. Duplicated names suggest paralogs in cotton. B) Boxplot of sucrose (top) and terpenoid (bottom) content from three clusters of accessions denoted by the red dashed line in A). These clusters are loosely associated with high (1), medium (3), and low (2) gene expression in A). All-way pairwise t-tests were performed to determine significance between expression clusters. C) The integrated network of 70 genes from the sucrose/T6P/terpenoid module. The 17 trait-related genes depicted in A) are depicted by orange circles. Three TFs are denoted by red rectangles and four T6P synthase (TPS) genes are marked. D) Motif logo of bZIP53 TF derived from motif enrichment analysis of 1Kb promoter sequences of 70 genes in this module. E) Pearson-correlation between sucrose levels (left) and bZIP53 gene expression (TPM), as well as between terpenoid content (right) and bZIP53 gene expression (TPM) across the 21 accessions.

Supplementary Figures

Figure. S1 Kinship Matrix derived from 21 upland cotton accessions

The kinship matrix was generated by TASSLE and visualized by pheatmap function in R based on four reps per accession. The closely related samples were clustered together by K-means clustering.

Figure. S2 Phenotype variations across 11 traits from two water conditions

A) The 21 upland cotton accessions were split into Foreign, U.S., and Mixed groups for pairwise comparisons among six production-related traits from two water conditions. For each pair-wise comparison, a t-test was performed to effects of combined heat-drought stress. B) The 21 upland cotton accessions were split into t Foreign, U.S., and Mixed groups for pairwise comparisons among four vegetation indices from two water conditions. For each pair-wise comparison, a t-test was performed to estimate the effects of combined heat-drought stress.

Figure. S3 MapMan heatmap of biological processes and metabolisms of two accessions

A) Comparison of standard deviation of fold change of genes mapped to different primary and secondary metabolisms across 21 cotton accessions. Significant variations were estimated by Kruskal-Wallis test ($p < 0.05$). B) Heatmap of $\log_2$FC of genes with high SD (SD > 1) involved in primary metabolism along with certain type of metabolites been classified including carbohydrate, amino acid, cell wall, lipids, hormones, and nucleotides. C) Heatmap of $\log_2$FC of genes with high SD (SD > 1) involved in secondary metabolism along with certain type of metabolites been classified including phenol, phenlypropaniods, lignin, flavonoids, and alkaloids.

Figure. S4 Weighted Co-expression network and trait-module membership

A) The left panel represents the frequency of gene connection revealed by connectivity (k). The right panel shows the log-log plot of the connectivity, the correlation coefficient represents the scale of free topology. The 0.86 here indicated the high approximation of a scale-free of network of network derived from top 10% MAD score. B) The left panel is the relation between approximation of a scale-free of network and different soft threshold (power). The seven is the least numbers to ensure a scale free network. The right panel stands for the mean connectivity of network when different power score was applied. C) A heatmap plot depicting the topological

overlap matrix (TOM) supplemented by hierarchical clustering dendrograms and the module colors was created by TOMplot function from the WGCNA package. Each cell correspond to a single gene, the light color indicated the high topological overlap, and the darker orange and red colors represent the lower topological overlap. D) Module-trait relationship was estimated by correlation between gene expression from 21 accession and 10 traits. Modules were plotted from different rows with number of genes in each module been assigned marked on the left panel. Ten columns are corresponding to six plant production traits and four vegetation indices. Each cell contains the correlation coefficient (|r|) and the p-value between that traits and module eigengene values.

Figure. S5 The ABA-responsive transcription factors (AREBs) are likely not influencing lint yield in cotton under WL conditions

Hierarchical heat map of the Z-score normalized expression values for the seven annotated AREBs in the 21 cotton accessions under WL and WW conditions

Figure. S6 Network analysis of vegetative indices related module

The integrated gene network of 58 genes from light green module. The eight trait-related genes were marked in orange, with enriched GO terms (hydrolase activity and calcium binding) been marked. The co-expression and PPIs were connected by grey solid lines and red dash lines respectively. NAC022 co-expressed with GDSL gene which involved in stomatal opening controlling levels of the water stress.

Figure. S7 Network analysis of glucose metabolism module

A) The integrated gene network of 75 genes from magenta module. The 16 trait-related genes were marked in orange, with enriched GO terms (photosynthesis, light harvesting, and carbon fixation) been marked. The co-expression, DNA-protein interaction, and PPIs were connected by grey solid lines, the blue dashed arrows, and red dash lines respectively. B) Pearson-correlation between glucose content and CAB7, TPPT, and RBCS gene expression (TPM) across 21 accessions. C) Visualization of $\log_2$FC of genes involved in photosynthesis pathways, including glucose content correlated genes.

References

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.

Almoguera, C., Prieto-dapena, P., Díaz-martín, J., Espinosa, J.M., Carranco, R., and Jordano, J. (2009). The HaDREB2 transcription factor enhances basal thermotolerance and longevity of seeds through functional interaction with HaHSFA9. **12**: 1–12.

Bae, H., Kim, S.H., Kim, M.S., Sicher, R.C., Lary, D., Strem, M.D., Natarajan, S., and Bailey, B.A. (2008). The drought response of Theobroma cacao (cacao) and the regulation of genes involved in polyamine biosynthesis by drought and other stresses. Plant Physiol. Biochem. **46**: 174–188.

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

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., and Buckler, E.S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics **23**: 2633–2635.

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.

Burke, J. J., and Wanjura, D. F. (2010). "Plant responses to temperature extremes," in Physiology of Cotton, eds J. M. Stewart, D. M. Oosterhuis, J. J. Heitholt, and J. R. Mauney (Dordrecht: Springer), 123–128. doi: 10.1007/978-90-481-3195-2_12

Caturegli, L., Matteoli, S., Gaetani, M., Grossi, N., Magni, S., Minelli, A., Corsini, G., Remorini, D., and Volterrani, M. (2020). Effects of water stress on spectral reflectance of bermudagrass. Sci. Rep. **10**: 1–12.

Chávez Montes RA, Haber A, Pardo J, Powell RF, Divisetty UK, Silva AT, Hernández-Hernández T, Silveira V, Tang H, Lyons E, Herrera Estrella LR, VanBuren R, Oliver MJ. A comparative genomics examination of desiccation tolerance and sensitivity in two sister grass species. Proc Natl Acad Sci U S A. 2022 Feb 1;119(5):e2118886119. doi: 10.1073/pnas.2118886119.

Chen, Z.J. et al. (2020). Genomic diversifications of five Gossypium allopolyploid species and their impact on cotton improvement. Nat Genet **52**: 525–533.

Christensen, A., Svensson, K., Persson, S., Jung, J., Michalak, M., Widell, S., and Sommarin, M. (2008). Functional characterization of Arabidopsis calreticulin1a: A key alleviator of endoplasmic reticulum stress. Plant Cell Physiol. **49**: 912–924.

Chu, X., Wang, C., Chen, X., Lu, W., Li, H., Wang, X., Hao, L., and Guo, X. (2015). The cotton

WRKY gene GhWRKY41 positively regulates salt and drought stress tolerance in transgenic Nicotiana benthamiana. PLoS One **10**: 1–21.

Claeys, H. and Inzé, D. (2013). The Agony of Choice: How Plants Balance Growth and Survival under Water-Limiting Conditions. Plant Physiol. **162**: 1768–1779.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., and Durbin, R. (2011). The variant call format and VCFtools. Bioinformatics **27**: 2156–2158.

Dekkers, B.J.W., Schuurmans, J.A.M.J., and Smeekens, S.C.M. (2004). Glucose delays seed germination in Arabidopsis thaliana. Planta **218**: 579–588.

Delatte, T.L., Sedijani, P., Kondou, Y., Matsui, M., Jong, G.J. De, Somsen, G.W., Wiese-klinkenberg, A., Primavesi, L.F., Paul, M.J., and Schluepmann, H. (2011). Growth Arrest by Trehalose-6-Phosphate : An Astonishing Case of Primary Metabolite Control over Growth by Way of the SnRK1 Signaling Pathway 1 [C][W][OA]. **157**: 160–174.

Delorge I, Janiak M, Carpentier S, Van Dijck P. Fine tuning of trehalose biosynthesis and hydrolysis as novel tools for the generation of abiotic stress tolerant plants. Front Plant Sci. 2014 Apr 14;5:147. doi: 10.3389/fpls.2014.00147.

Dietrich, K., Weltmeier, F., Ehlert, A., Weiste, C., Stahl, M., Harter, K., and Dröge-Lasera, W. (2011). Heterodimers of the Arabidopsis transcription factors bZIP1 and bZIP53 reprogram amino acid metabolism during Low energy stress. Plant Cell **23**: 381–395.

Ding, L.N., Li, M., Wang, W.J., Cao, J., Wang, Z., Zhu, K.M., Yang, Y.H., Li, Y.L., and Tan, X.L. (2019). Advances in plant GDSL lipases: from sequences to functional mechanisms. Acta Physiol. Plant. **41**: 1–11.

Do, P.T., Degenkolbe, T., Erban, A., Heyer, A.G., Kopka, J., Köhl, K.I., Hincha, D.K., and Zuther, E. (2013). Dissecting Rice Polyamine Metabolism under Controlled Long-Term Drought Stress. PLoS One **8**.

Fahad, S., Hussain, S., Saud, S., Khan, F., Hassan, S., Amanullah, Nasim, W., Arif, M., Wang, F., and Huang, J. (2016). Exogenously Applied Plant Growth Regulators Affect Heat-Stressed Rice Pollens. J. Agron. Crop Sci. **202**: 139–150.

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**.

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.

Fernie, A.R., Stitt, M., On the Discordance of Metabolomics with Proteomics and Transcriptomics: Coping with Increasing Complexity in Logic, Chemistry, and Network Interactions Scientific Correspondence, Plant Physiology, Volume 158, Issue 3, March 2012, Pages 1139–1145, https://doi.org/10.1104/pp.112.193235

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D.S. (2012). Phytozome: A comparative platform for green plant genomics. Nucleic Acids Res. **40**: 1178–1186.

Guo, W.J. and Ho, T.H.D. (2008). An abscisic acid-induced protein, HVA22, inhibits gibberellin-mediated programmed cell death in cereal aleurone cells. Plant Physiol. **147**: 1710–1722.

Gupta, A., Rico-Medina, A., and Caño-Delgado, A.I. (2020). The physiology of plant responses to drought. Science (80-.). **368**: 266–269.

Hayashi, H., De Bellis, L., Hayashi, Y., Nito, K., Kato, A., Hayashi, M., Hara-Nishimura, I., and Nishimura, M. (2002). Molecular characterization of an Arabidopsis acyl-coenzyme a synthetase localized on glyoxysomal membranes. Plant Physiol. **130**: 2019–2026.

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., Frelichowski, J., and Percy, R.G. (2016). Genetic diversity of the two commercial tetraploid cotton species in the gossypium diversity reference set. J. Hered. **107**: 274–286.

Huang, G.Q., Li, W., Zhou, W., Zhang, J.M., Li, D. Di, 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, 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, Q., Wang, Y., Li, B., Chang, J., Chen, M., Li, K., Yang, G., and He, G. (2015). TaNAC29, a NAC transcription factor from wheat, enhances salt and drought tolerance in transgenic Arabidopsis. BMC Plant Biol. **15**: 1–15.

Imkampe, J. et al. (2017). The arabidopsis leucine-rich repeat receptor kinase BIR3 negatively regulates BAK1 receptor complex formation and stabilizes BAK1. Plant Cell **29**: 2285–2303.

Jaspers, P., Blomster, T., Brosché, M., Salojärvi, J., Ahlfors, R., Vainonen, J.P., Reddy, R.A., Immink, R., Angenent, G., Turck, F., Overmyer, K., and Kangasjärvi, J. (2009). Unequally redundant RCD1 and SRO1 mediate stress and developmental responses and interact with

transcription factors. Plant J. **60**: 268–279.

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

Karim S, Aronsson H, Ericson H, Pirhonen M, Leyman B, Welin B, Mäntylä E, Palva ET, Van Dijck P, Holmström KO. Improved drought tolerance without undesired side effects in transgenic plants producing trehalose. Plant Mol Biol. 2007 Jul;64(4):371-86. doi: 10.1007/s11103-007-9159-6.

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

Lei, L. et al. (2015). Ribosome profiling reveals dynamic translational landscape in maize seedlings under drought stress. Plant J. **84**: 1206–1208.

Leyman, B., Van Dijck, P., and Thevelein, J.M. (2001). An unexpected plethora of trehalose biosynthesis genes in Arabidopsis thaliana. Trends Plant Sci. **6**: 510–513.

Li HW, Zang BS, Deng XW, Wang XP. Overexpression of the trehalose-6-phosphate synthase gene OsTPS1 enhances abiotic stress tolerance in rice. Planta. 2011 Nov;234(5):1007-18. doi: 10.1007/s00425-011-1458-0.

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.

Lin Q, Yang J, Wang Q, Zhu H, Chen Z, Dao Y, Wang K. Overexpression of the trehalose-6-phosphate phosphatase family gene AtTPPF improves the drought tolerance of Arabidopsis thaliana. BMC Plant Biol. 2019 Sep 2;19(1):381. doi: 10.1186/s12870-019-1986-5.

Love, M.I., Anders, S., and Huber, W. (2014). Differential analysis of count data - the DESeq2 package.

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. 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.

MacIntyre AM, Meline V, Gorman Z, Augustine SP, Dye CJ, Hamilton CD, Iyer-Pascuzzi AS, Kolomiets MV, McCulloh KA, Allen C. Trehalose increases tomato drought tolerance, induces defenses, and increases resistance to bacterial wilt disease. PLoS One. 2022 Apr 27;17(4):e0266254. doi: 10.1371/journal.pone.0266254.

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**.

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

Malley, R.C.O., Carol, S., Song, L., Galli, M., Ecker, J.R., Malley, R.C.O., Huang, S.C., Song, L., Lewsey, M.G., Bartlett, A., and Nery, J.R. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape Resource Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell **165**: 1280–1292.

Miranda JA, Avonce N, Suárez R, Thevelein JM, Van Dijck P, Iturriaga G. A bifunctional TPS-TPP enzyme from yeast confers tolerance to multiple and extreme abiotic-stress conditions in transgenic Arabidopsis. Planta. 2007 Nov;226(6):1411-21. doi: 10.1007/s00425-007-0579-y.

McLeay, R.C. and Bailey, T.L. (2010). Motif Enrichment Analysis: A unified framework and an evaluation on ChIP data. BMC Bioinformatics **11**.

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.

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.

Obata, T. and Fernie, A.R. (2012). The use of metabolomics to dissect plant responses to abiotic stresses. Cell. Mol. Life Sci. **69**: 3225–3243.

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. African J. Biotechnol. **10**: 15430–15437.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods **14**: 417–419.

Paul Shannon, 1, Andrew Markiel, 1, Owen Ozier, 2 Nitin S. Baliga, 1 Jonathan T. Wang, 2 Daniel Ramage, 2, Nada Amin, 2, Benno Schwikowski, 1, 5 and Trey Ideker2, 3, 4, 5, 山本隆久, 豊田直平, 深瀬吉邦, and 大森敏行 (2003). Cytoscape: A Software Environment for Integrated Models. Genome Res. **13**: 426.

Pego, J. V., Kortstee, A.J., Huijser, C., and Smeekens, S.C.M. (2000). Photosynthesis, sugars and the regulation of gene expression. J. Exp. Bot. **51**: 407–416.

Peri, S., Roberts, S., Kreko, I.R., McHan, L.B., Naron, A., Ram, A., Murphy, R.L., Lyons, E., Gregory, B.D., Devisetty, U.K., and Nelson, A.D.L. (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**.

Ponnu, J., Wahl, V., and Schmid, M. (2011). Trehalose-6-phosphate : connecting plant metabolism and development. **2**: 1–6.

Prieto-dapena, P., Espinosa, M., and Jordano, J. (2005). Functional Interaction between Two Transcription Factors Involved in the Developmental Regulation of a Small Heat Stress Protein Gene Promoter 1 [w]. **139**: 1483–1494.

Rolland F, Baena-Gonzalez E, Sheen J. Sugar sensing and signaling in plants: conserved and novel mechanisms. Annu Rev Plant Biol. 2006;57:675-709. doi: 10.1146/annurev.arplant.57.032905.105441.

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. U. S. A. **103**: 18822–18827.

Schwacke, R., Ponce-Soto, G.Y., Krause, K., Bolger, A.M., Arsova, B., Hallab, A., Gruden, K., Stitt, M., Bolger, M.E., and Usadel, B. (2019). MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. Mol. Plant **12**: 879–892.

Shi, Y., Pang, X., Liu, W., Wang, R., Su, D., Gao, Y., Wu, M., Deng, W., Liu, Y., and Li, Z. (2021). SlZHD17 is involved in the control of chlorophyll and carotenoid metabolism in tomato fruit. Hortic. Res. **8**: 1–16.

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

Sinclair, T.R. (2011). Challenges in breeding for yield increase for drought. Trends Plant Sci. **16**: 289–293.

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. **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., Doncheva, N.T., Roth, A., Bork, P., Jensen, L.J., and Von Mering, C. (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.

Tang, J., Yang, X., Xiao, C., Li, J., Chen, Y., Li, R., Li, S., Shiyou, L., and Hu, H. (2020). GDSL lipase occluded stomatal pore 1 is required for wax biosynthesis and stomatal cuticular ledge formation. New Phytol. **228**: 1880–1896.

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., Parent, B., Caldeira, C.F., and Welcker, C. (2014). Genetic and physiological controls of growth under water deficit. Plant Physiol. **164**: 1628–1635.

Tardieu, F., Simonneau, T., and Muller, B. (2018). The Physiological Basis of Drought Tolerance in Crop Plants: A Scenario-Dependent Probabilistic Approach. Annu. Rev. Plant Biol. **69**: 733–759.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L.A., Rhee, S.Y., and Stitt, M. (2004). MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. **37**: 914–939.

Usadel, B., Poree, F., Nagel, A., Lohse, M., Czedik-Eysenberg, A., and Stitt, M. (2009). A guide to using MapMan to visualize and compare Omics data in plants: A case study in the crop species, Maize. Plant, Cell Environ. **32**: 1211–1229.

Wang, F. et al. (2015). GmWRKY27 interacts with GmMYB174 to reduce expression of GmNAC29 for stress tolerance in soybean plants. Plant J. **83**: 224–236.

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. 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.

Xiao, S., Jiang, L., Wang, C., and Ow, D.W. (2021). Arabidopsis OXS3 family proteins repress ABA signaling through interactions with AFP1 in the regulation of ABI4 expression. J. Exp. Bot. **72**: 5721–5734.

Xie, Z., Nolan, T.M., Jiang, H., and Yin, Y. (2019). AP2/ERF transcription factor regulatory networks in hormone and abiotic stress responses in Arabidopsis. Front. Plant Sci. **10**: 1–17.

Zandalinas, S.I., Mittler, R., Balfagón, D., Arbona, V., and Gómez-Cadenas, A. (2018). Plant adaptations to the combination of drought and high temperatures. Physiol. Plant. **162**: 2–12.

Zheng, Y. 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, G. et al. (2018). Rewiring of the Fruit Metabolome in Tomato Breeding. Cell **172**: 249-261.e12.

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.

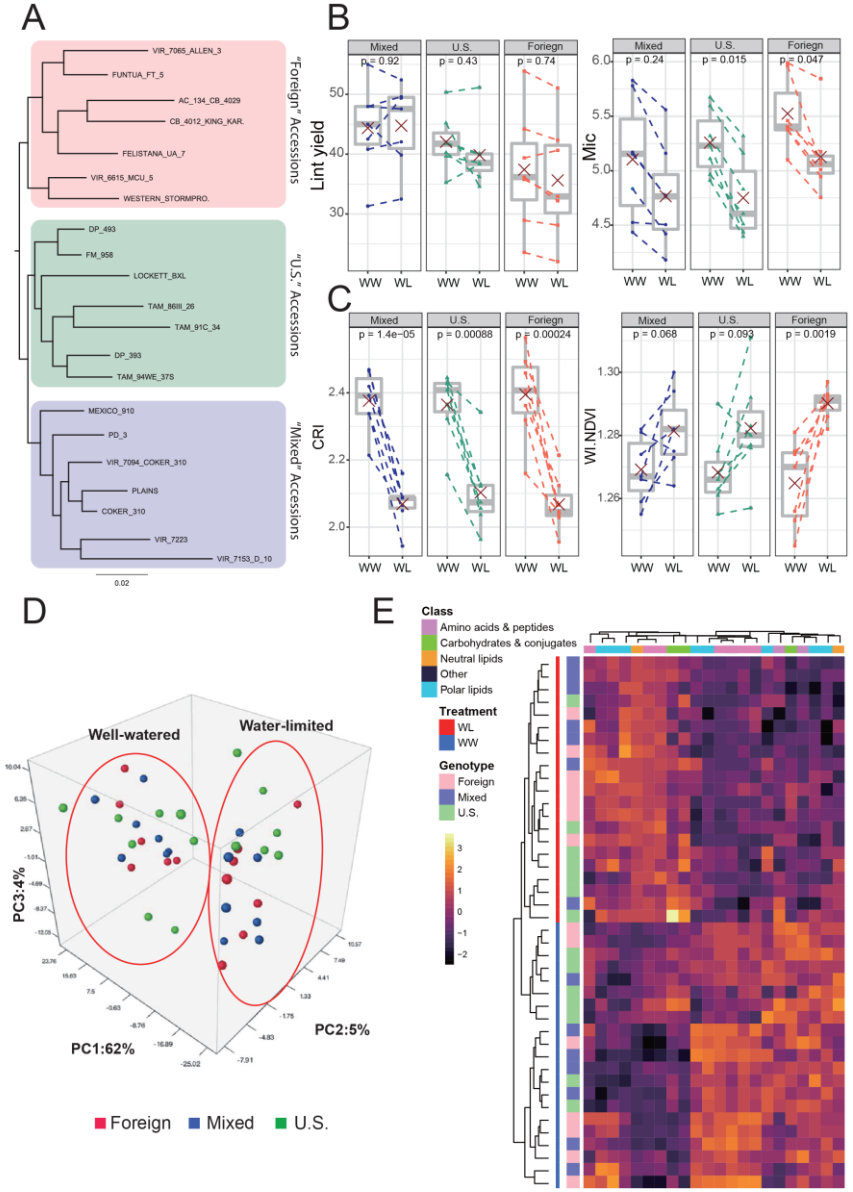

Figure. 1 Phylogeny, traits variations, and metabolite profiles of 21 distinct cotton accessions

A) A phylogeny of the accessions used in this study. The three groups, "Foreign", "U.S.", and "Mixed", reflect historical breeding information obtained from USDA GRIN-Global. B) Pairwise comparisons of two representative fiber quality traits (Lint yield and micronaire; Mic) from well watered (WW) and water limited (WL) conditions. C) Pairwise comparisons of two representative vegetative indices (carotenoid reflective index; CRI, and the ratio between the water index and the normalized difference vegetation index; NDVI), from two water conditions. Significance values for B) and C) were calculated using a parametric t-test in response to drought. D) Three-dimensional PCA of 451 metabolites derived from normalized metabolite content. Legend denotes accessions belonging to three phylogenetic groups and samples from WL and WW conditions. E) Z-score based heatmap of 22 metabolites belonging to five classes (see legend) with significant genotype effects (factor: Foreign, U.S., and Mixed, $p < 0.05$).

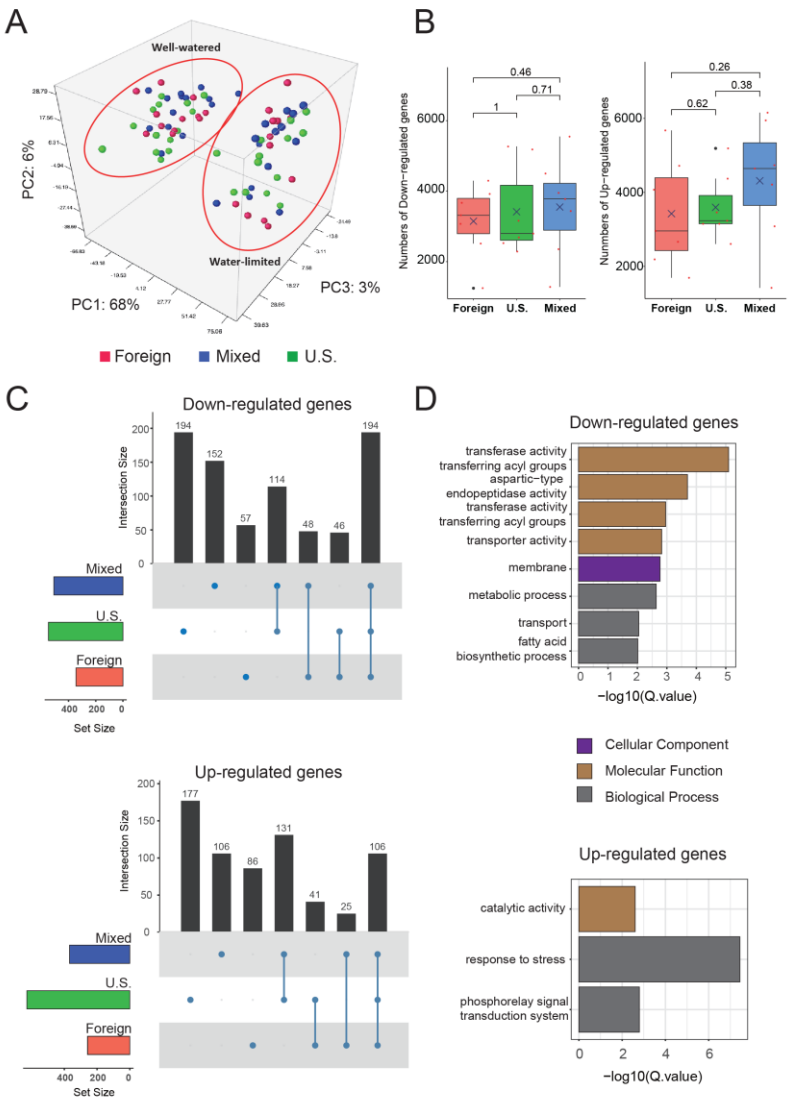

Figure. 2 Transcriptome comparison of expression profiles in the three phylogenetic groups

A) Three-dimensional PCA of the top 10% most variable genes (n = 3,768) derived from normalized read counts. B) The average numbers of differentially expressed genes (DEGs; down-regulated on left, up-regulated on the right) derived from per-accession pair-wise comparisons between the two conditions (WW and WL). An all way pairwise parametric t-test was performed to determine significance. C) Upset plot of genes that were differentially expressed in all representatives within a particular phylogenetic group. The vertical black bars stand for numbers of unique DEGs from each group, intersection between each two group, and intersection among three groups. D) Enriched gene ontology terms of commonly down-regulated (top) and up-regulated (bottom) genes among 21 accessions. The calculation is based on -log10 Q values (cutoff:3).

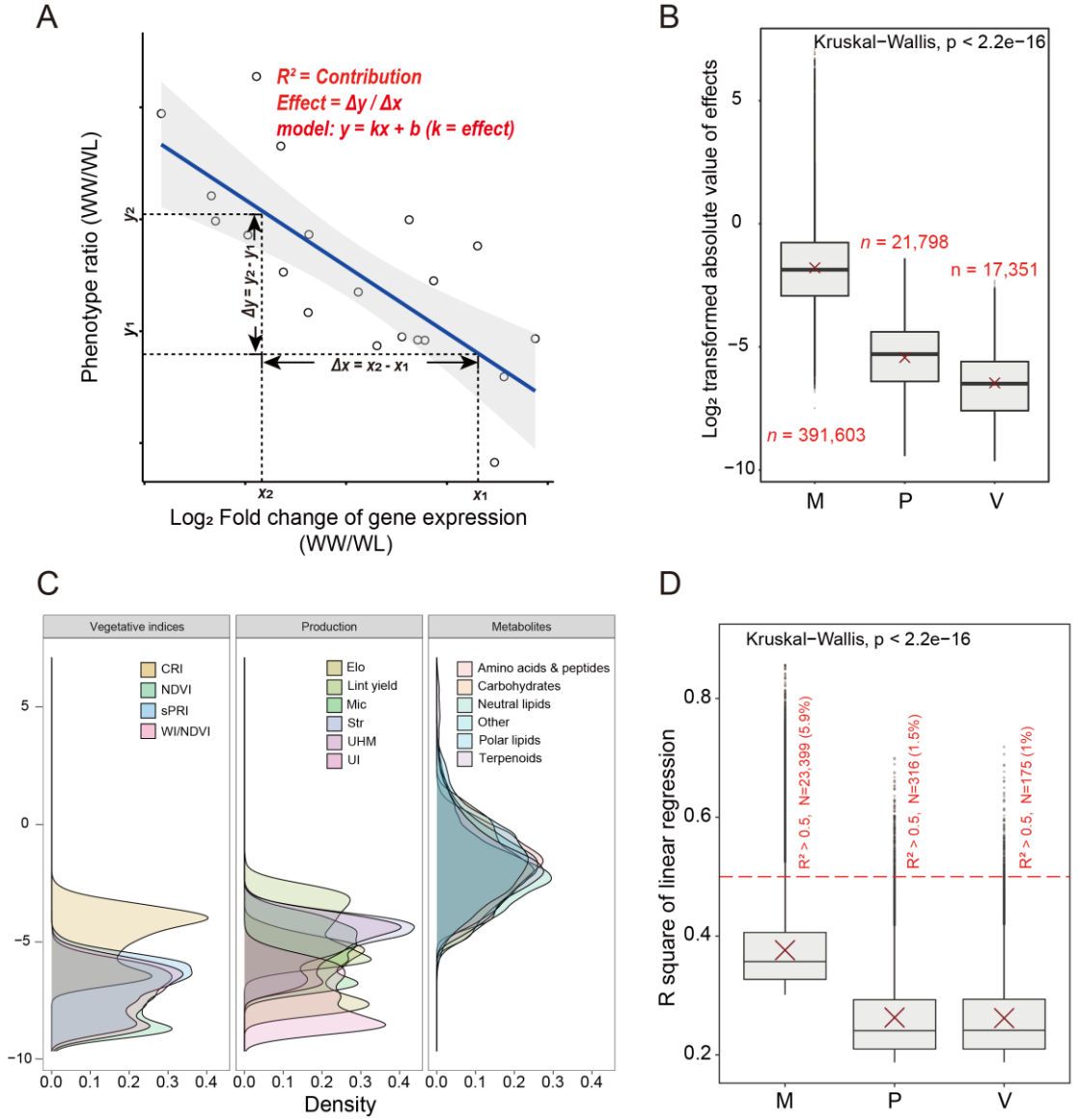

Figure. 3 Identifying large

A) Illustrative scatter plot of the model used to identify genes making a large contribution to phenotype. A Pearson correlation between log2-transformed fold changes of gene expression and ratio of phenotypic measurements between the two conditions was performed for all phenotypic and metabolomic measurements. The X-axis depicts the explanatory variable, changes in gene expression from two watering conditions, whereas the Y-axis depicts the responsive variable of phenotypic changes potentially regulated by gene expression. B) Boxplot comparison of log2-transformed absolute values of effects (Δy/Δx) for each significant ($p < 0.05$) pairwise linear regression among metabolites (M), production-traits (P), and vegetative indices (**V**). C) Density distribution of log2-transformed absolute values of effects (Δy/Δx) derived from pair-wise correlations of Log_2FC to phenotype ratio (WW/WL) represented by sub-categories (see legends) among metabolites (M), production-traits (P), and vegetative indices (**V**). D) Boxplot comparison of R^2 value of high-confidence ($p < 0.05$) Pearson correlations. The red dash line indicates the cut-off of correlations with high contribution rate ($R^2 > 0.5$), numbers of correlation pairs above the cut-off and their respective percent of total correlations are denoted alongside each boxplot.

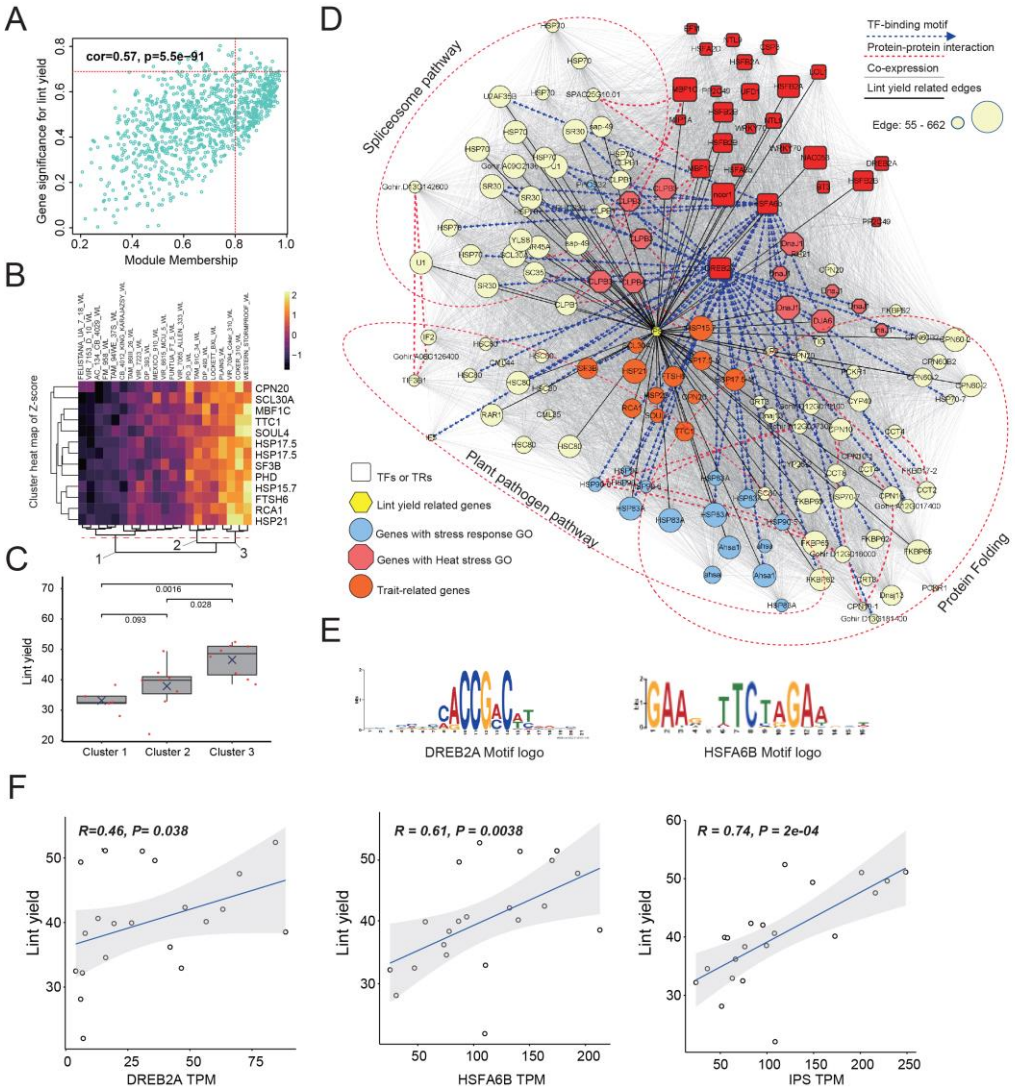

Figure. 4 Network analysis of a module of genes associated with lint yield

A) The module membership (MM) scores and gene significance (GS) for genes contained within the lint-yield module were plotted with best fit determined by a linear model. The dashed red lines indicate the two cutoffs used to identify trait-related genes. B) Clustered heat map based on Z-score transformed TPM values across 21 accessions of trait-related genes selected from the network. Both row and column data were clustered by “correlation” distance clustering methods from pheatmap function in R. Red dashed line depicts the cutoff used to define clusters in C. C) Box and whiskers depicting lint yields for three different expression clusters of lint-yield associated genes. Significance values were derived from pair-wise t-tests. D) The integrated child network derived from 153 out of 866 genes within the lint-yield correlated turquoise module. Gene categories are shown in the bottom left, whereas connection types are shown in the top right. KEGG pathways are grouped with a red dashed oval. E) Motif logo of DREB2A and HSFA6B TFs, derived from Analysis of Motif Enrichment (AME), found within the promoter regions of genes within the lint yield module. Genes connected to DREB2A and HSF6AB, and containing the respective promoter binding motifs, are depicted as blue dashed lines in D. F) Pearson-correlation between lint yield and DREB2A, HSFA6B, and IPS gene expression (TPM) across 21 accessions.

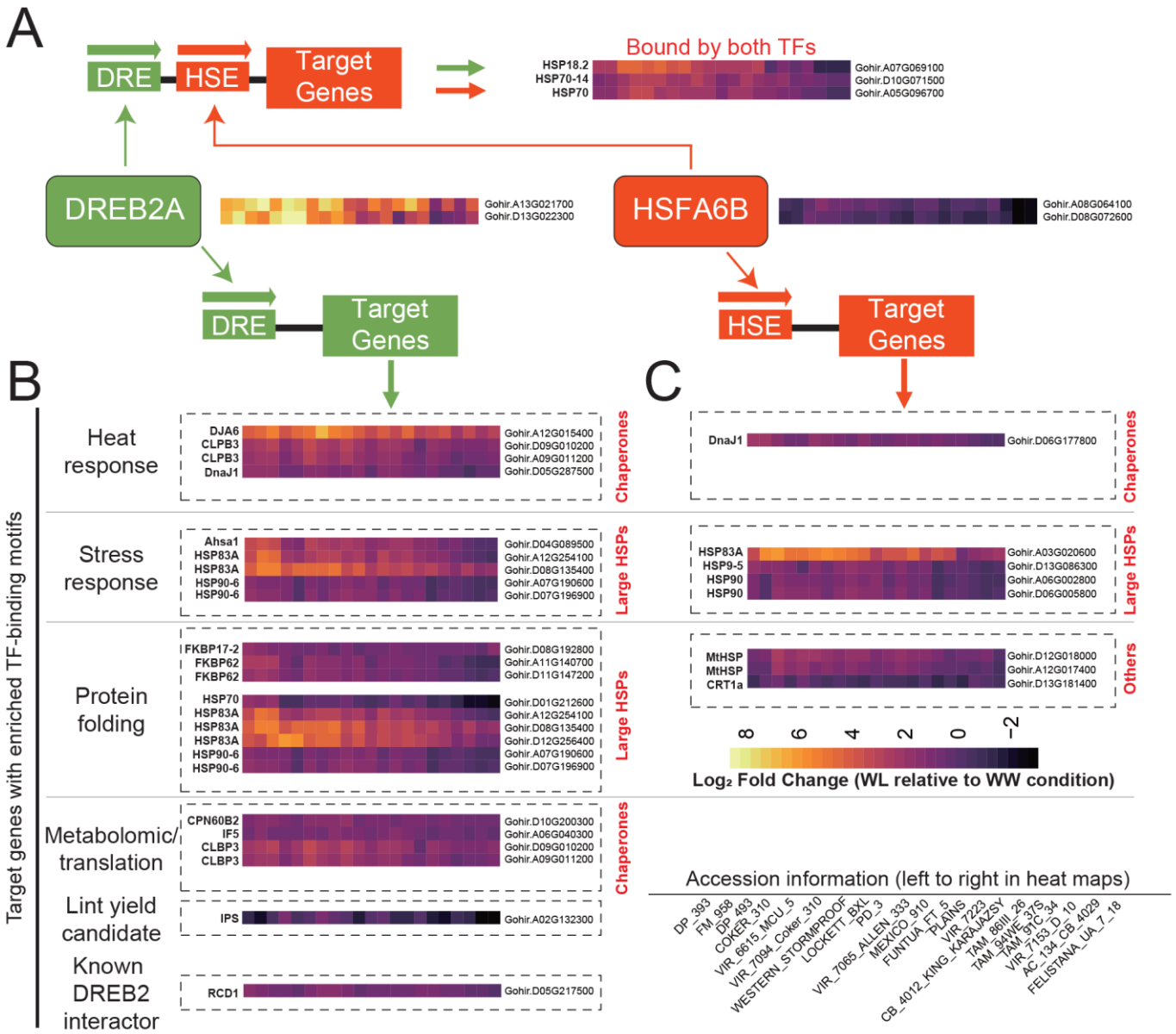

Figure. 5 Proposed regulatory network of lint yield associated genes

Proposed regulatory network of DREB and HSF gene under water limited conditions. A) Depiction of three genes within the lint yield module that contain promoter binding motifs, and are connected in the module, to both DREB2A and HSFA6A. Log2 Fold Change values are shown for water limited conditions relative to control. B) Expression of genes likely to be specifically modulated by DREB2A, as they contain a DRE motif and are connected in the module to DREB2A. C) As in B), genes that look to be specifically modulated by HSFA6B. DRE = dehydration response element, HSE = heat shock element. Genes in B) and C) are divided based on major biological functions (shown in red on right).

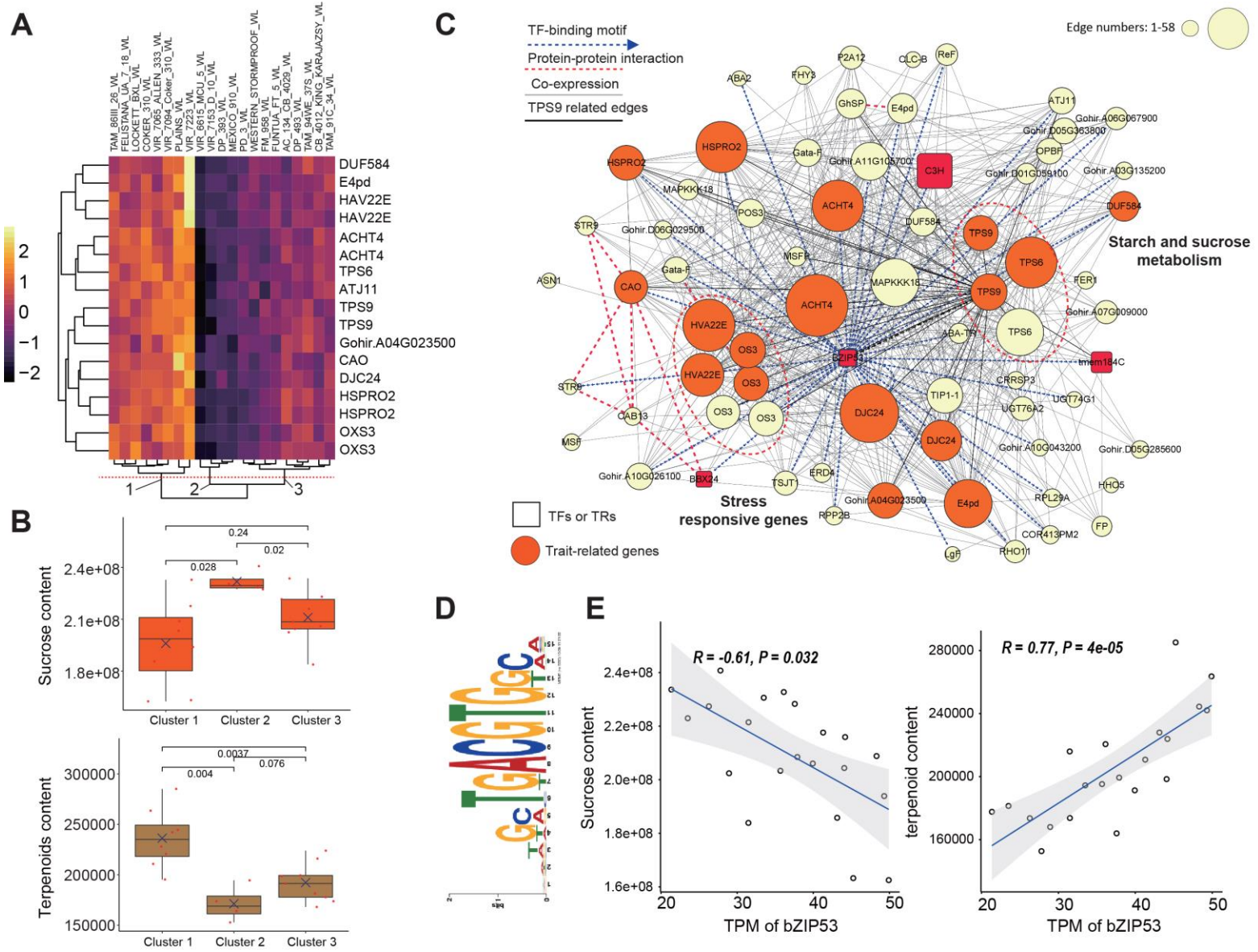

Figure. 6 Network analysis of sucrose and trehalose-6-phosphate metabolism module

A) Cluster heat map based on Z-score transformed TPM values across 21 accessions of 17 trait-related genes that were selected from the parent network based on GS and MM as in Figure 5. Where possible, gene names based on orthology to Arabidopsis gene IDs are shown. Duplicated names suggest paralogs in cotton. B) Boxplot of sucrose (top) and terpenoid (bottom) content from three clusters of accessions denoted by the red dashed line in A). These clusters are loosely associated with high (1), medium (3), and low (2) gene expression in A). All-way pairwise t-tests were performed to determine significance between expression clusters. C) The integrated network of 70 genes from the sucrose/T6P/terpenoid module. The 17 trait-related genes depicted in A) are depicted by orange circles. Three TFs are denoted by red rectangles and four T6P synthase (TPS) genes are marked. D) Motif logo of bZIP53 TF derived from motif enrichment analysis of 1Kb promoter sequences of 70 genes in this module. E) Pearson-correlation between sucrose levels (left) and bZIP53 gene expression (TPM), as well as between terpenoid content (right) and bZIP53 gene expression (TPM) across the 21 accessions.

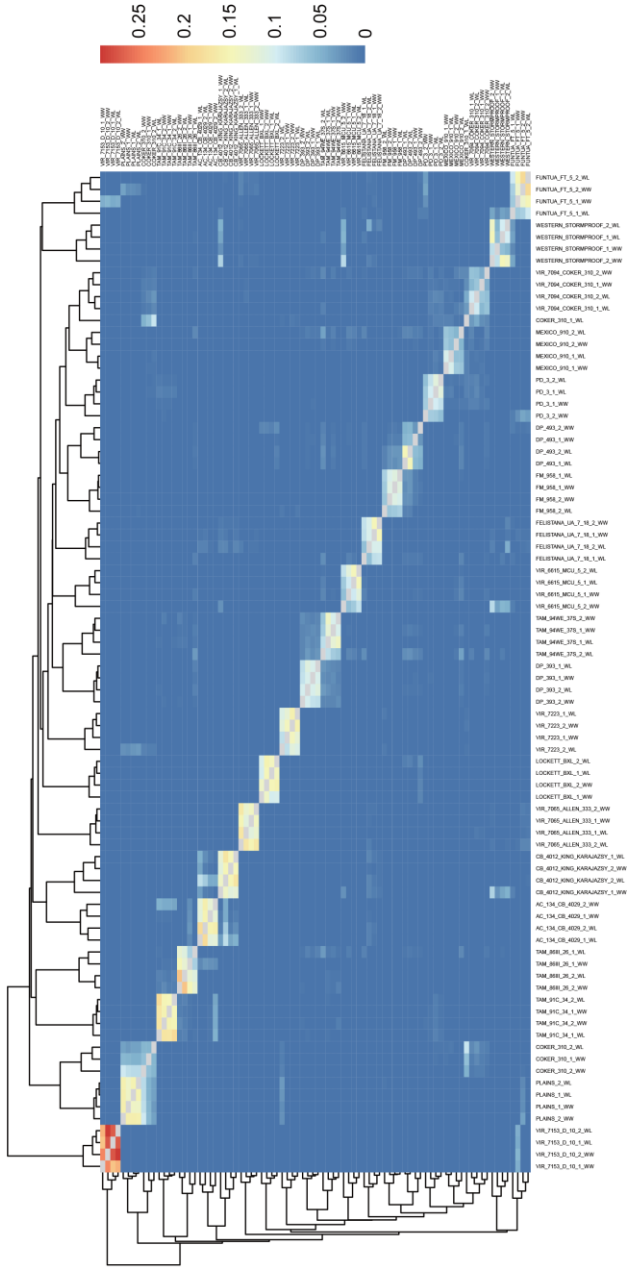

Figure. S1 Kinship Matrix derived from 21 upland cotton accessions

The kinship matrix was generated by TASSLE and visualized by pheatmap function in R based on four reps per accession. The closely related samples were clustered together by K-means clustering.

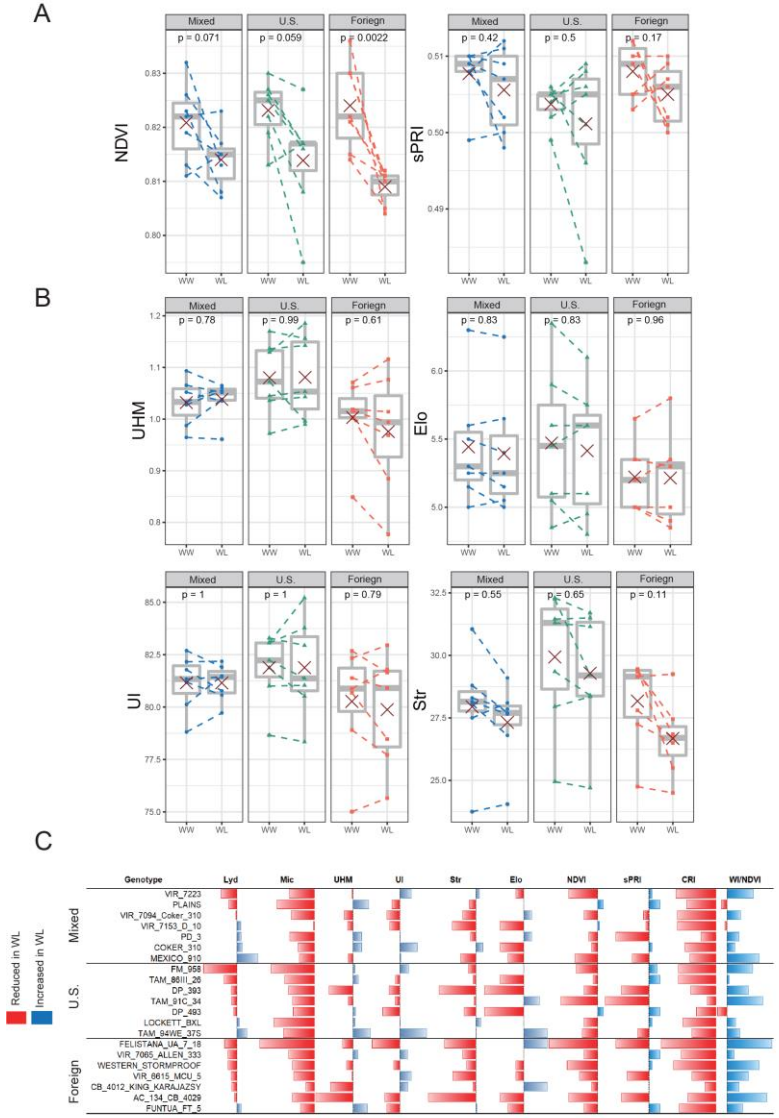

Figure. S2 Phenotype variations across 11 traits from two water conditions

A) The 21 upland cotton accessions were split into Foreign, U.S., and Mixed groups for pairwise comparisons among six production-related traits from two water conditions. For each pair-wise comparison, a t-test was performed to effects of combined heat-drought stress. B) The 21 upland cotton accessions were split into t Foreign, U.S., and Mixed groups for pairwise comparisons among four vegetation indices from two water conditions. For each pair-wise comparison, a t-test was performed to estimate the effects of combined heat-drought stress.

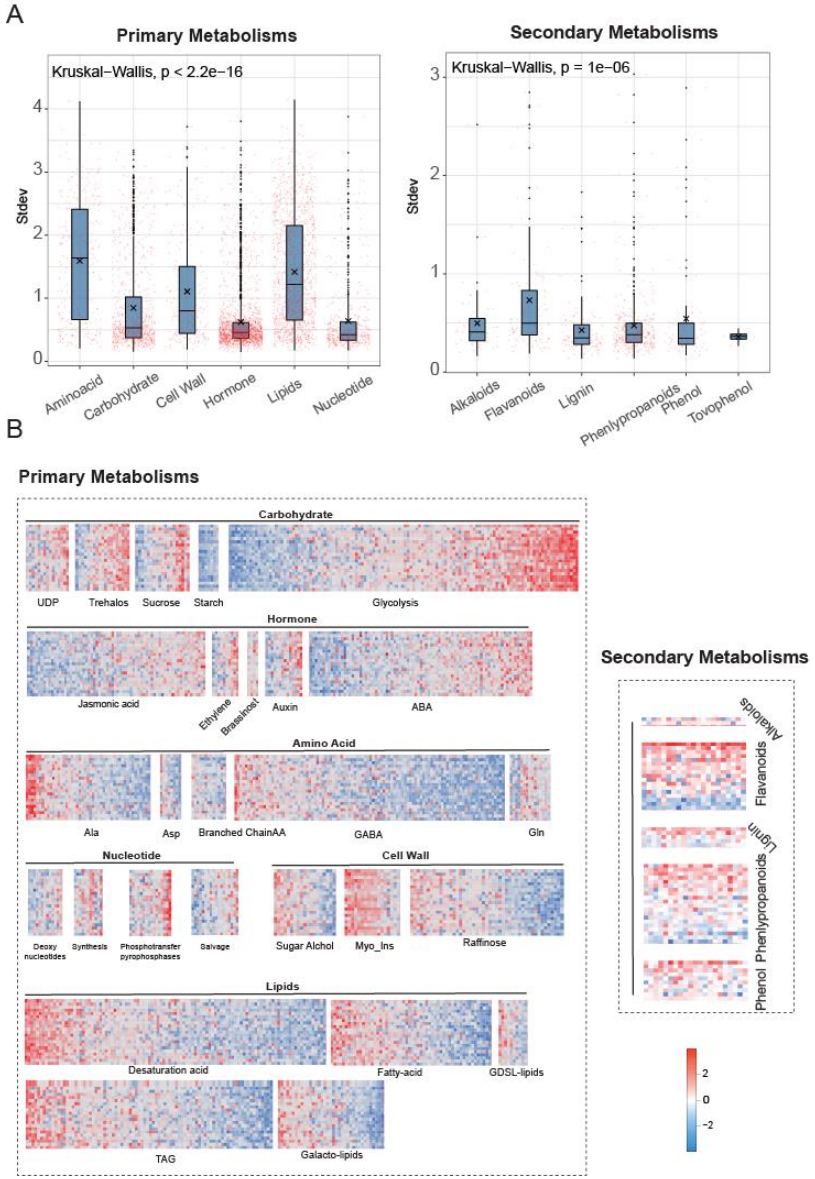

Figure. S3 MapMan heatmap of biological processes and metabolisms of two accessions

A) Comparison of standard deviation of fold change of genes mapped to different primary and secondary metabolisms across 21 cotton accessions. Significant variations were estimated by Kruskal-Wallis test ($p < 0.05$). B) Heatmap of $\log_2$FC of genes with high SD (SD > 1) involved in primary metabolism along with certain type of metabolites been classified including carbohydrate, amino acid, cell wall, lipids, hormones, and nucleotides. C) Heatmap of $\log_2$FC of genes with high SD (SD > 1) involved in secondary metabolism along with certain type of metabolites been classified including phenol, phenlypropaniods, lignin, flavonoids, and alkaloids.

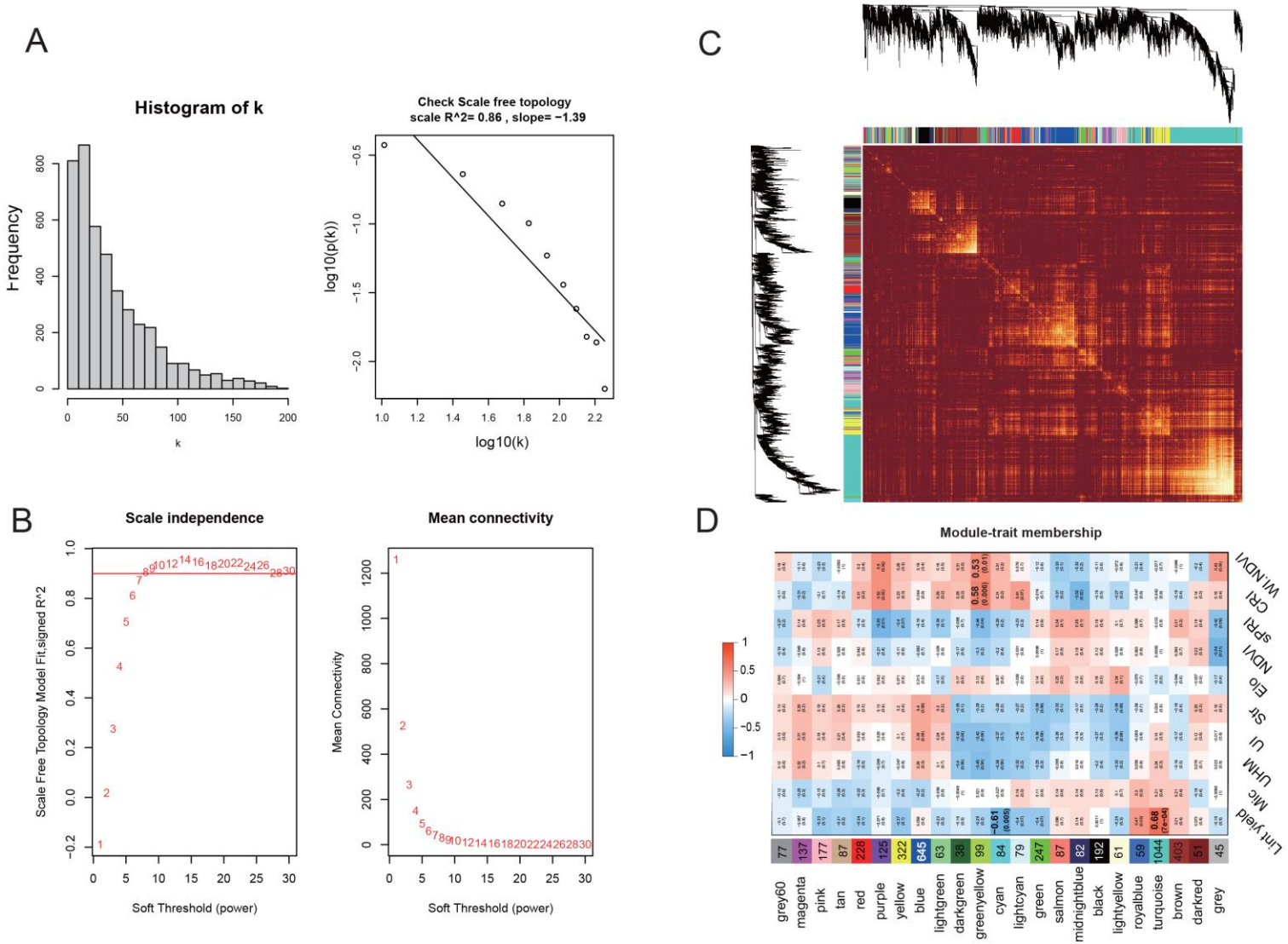

Figure. S4 Weighted Co-expression network and trait-module membership

A) The left panel represents the frequency of gene connection revealed by connectivity (k). The right panel shows the log-log plot of the connectivity, the correlation coefficient represents the scale of free topology. The 0.86 here indicated the high approximation of a scale-free of network of network derived from top 10% MAD score. B) The left panel is the relation between approximation of a scale-free of network and different soft threshold (power). The seven is the least numbers to ensure a scale free network. The right panel stands for the mean connectivity of network when different power score was applied. C) A heatmap plot depicting the topological overlap matrix (TOM) supplemented by hierarchical clustering dendrograms and the module colors was created by TOMplot function from the WGCNA package. Each cell correspond to a single gene, the light color indicated the high topological overlap, and the darker orange and red colors represent the lower topological overlap. D) Module-trait relationship was estimated by correlation between gene expression from 21 accession and 10 traits. Modules were plotted from different rows with number of genes in each module been assigned marked on the left panel. Ten columns are corresponding to six plant production traits and four vegetation indices. Each cell contains the correlation coefficient (|r|) and the p-value between that traits and module eigengene values.

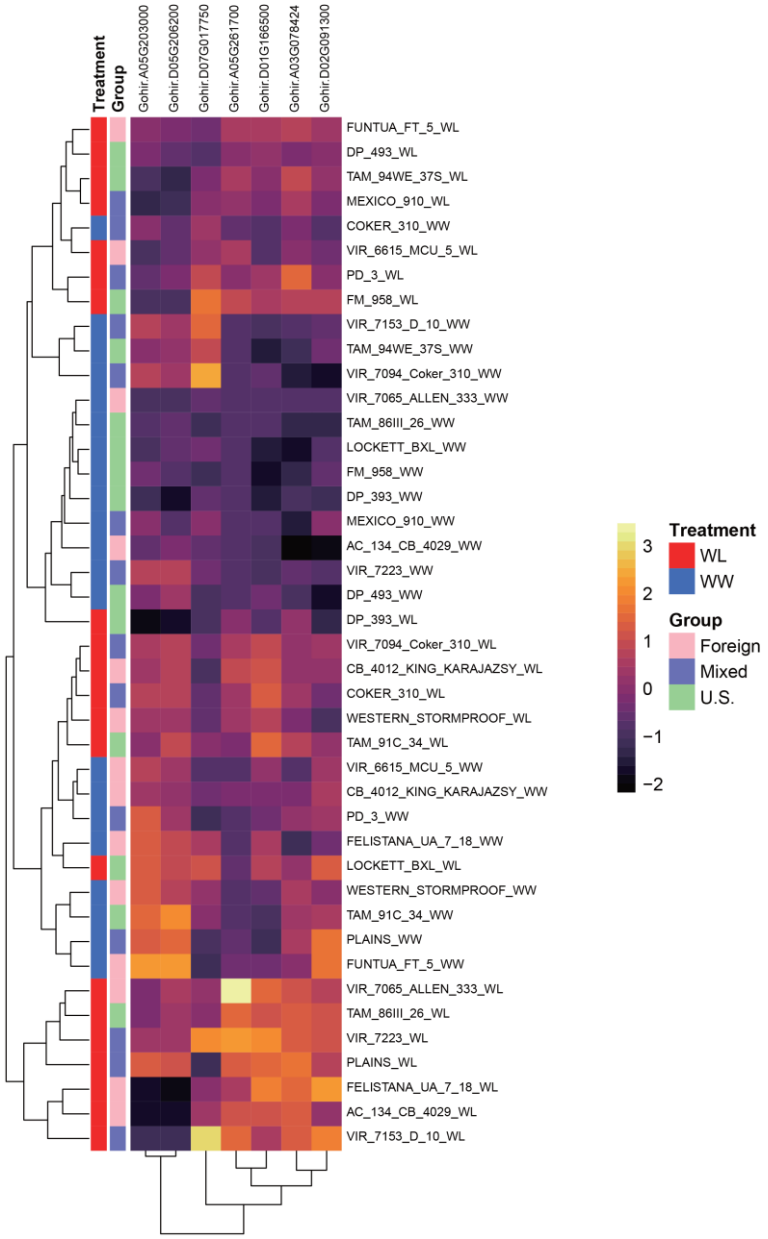

Figure. S5 The ABA-responsive transcription factors (AREBs) are likely not influencing lint yield in cotton under WL conditions

Hierarchical heat map of the Z-score normalized expression values for the seven annotated AREBs in the 21 cotton accessions under WL and WW conditions

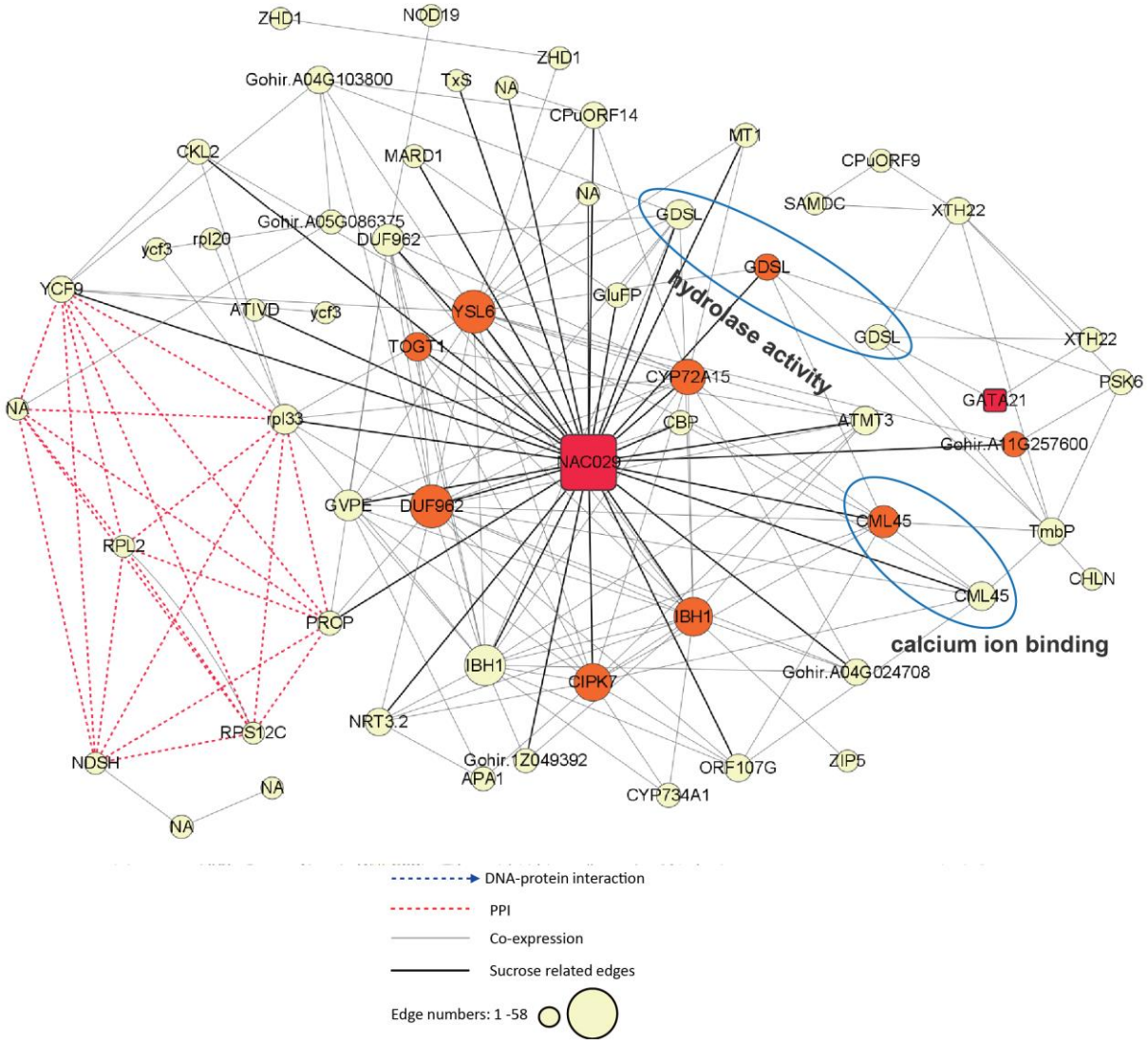

Figure. S6 Network analysis of vegetative indices related module

The integrated gene network of 58 genes from light green module. The eight trait-related genes were marked in orange, with enriched GO terms (hydrolase activity and calcium binding) been marked. The co-expression and PPIs were connected by grey solid lines and red dash lines respectively. NAC022 co-expressed with GDSL gene which involved in stomatal opening controlling levels of the water stress.

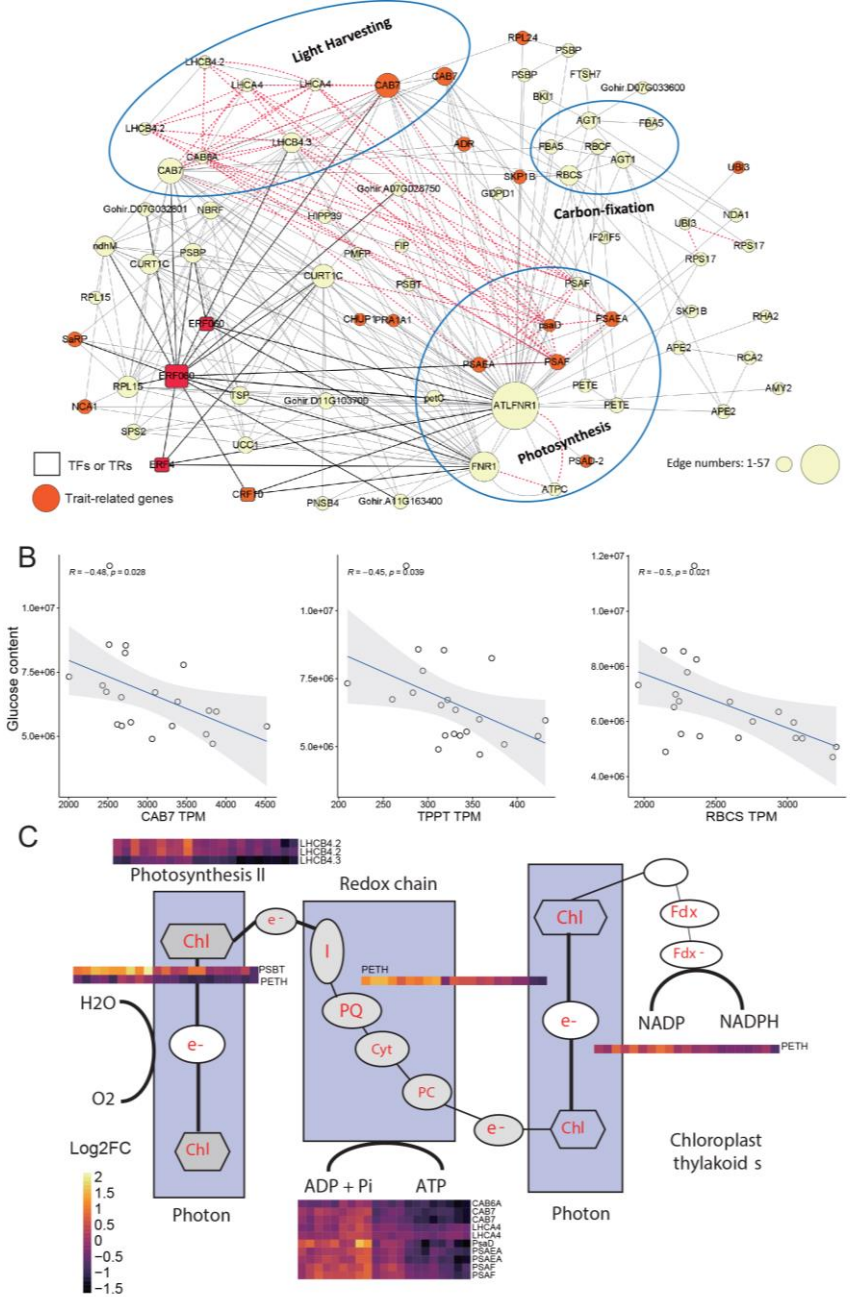

Figure. S7 Network analysis of glucose metabolism module

A) The integrated gene network of 75 genes from magenta module. The 16 trait-related genes were marked in orange, with enriched GO terms (photosynthesis, light harvesting, and carbon fixation) been marked. The co-expression, DNA-protein interaction, and PPIs were connected by grey solid lines, the blue dashed arrows, and red dash lines respectively. B) Pearson-correlation between glucose content and CAB7, TPPT, and RBCS gene expression (TPM) across 21 accessions. C) Visualization of log_2FC of genes involved in photosynthesis pathways, including glucose content correlated genes.