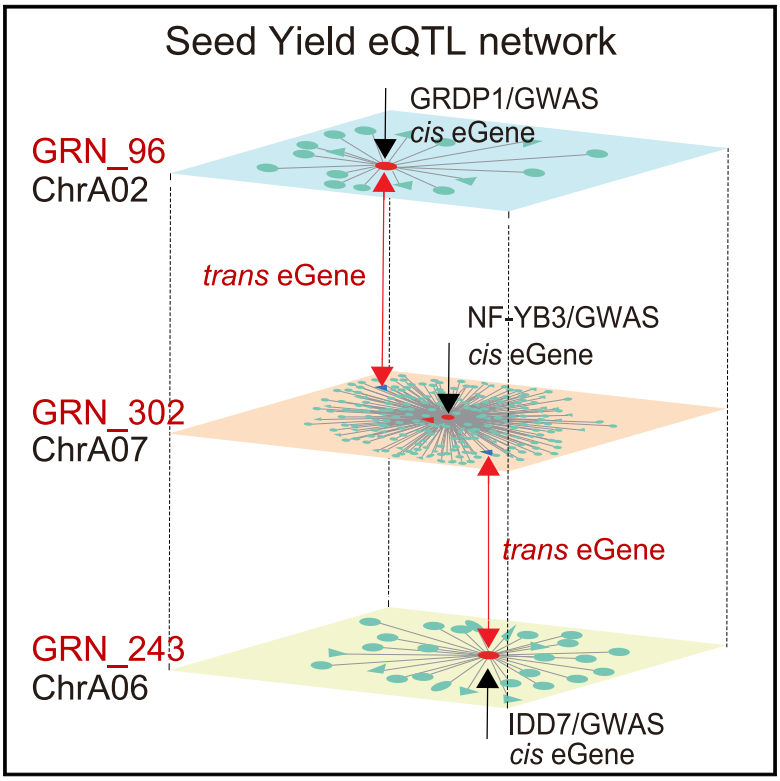
Cell Reports

Article

Integration of eQTL and machine learning to dissect causal genes with pleiotropic effects in genetic regulation networks of seed cotton yield

Graphical abstract



Highlights

- 12,207 eQTLs of 1 DPA ovule established with 558 transcriptome in cotton natural population
- eGenes enclosed in the cotton yield GRNs regulate the corresponding heritability as a group
- XGBoost is applied to predict seed cotton yield with gene importance ranking in GRNs
- Top-ranked eGenes *NF-YB3* and *GRDP1* in seed size regulation are validated

Authors

Ting Zhao, Hongyu Wu, Xutong Wang, ..., Lei Fang, Tianzhen Zhang, Xueying Guan

Correspondence

xueyingguan@zju.edu.cn

In brief

Zhao et al. construct a gene regulatory network (GRN) for cotton yield traits with 12,207 eQTLs in 1 DPA ovules, enclosing 735 genes in the functional GRN. The seed cotton yield is predicted using XGBoost and validated by manipulating the top-ranking eGenes recommended from machine learning analysis.

Zhao et al., 2023, Cell Reports *42*, 113111
September 26, 2023
https://doi.org/10.1016/j.celrep.2023.113111

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

Integration of eQTL and machine learning to dissect causal genes with pleiotropic effects in genetic regulation networks of seed cotton yield

Ting Zhao,[1,2] Hongyu Wu,[1] Xutong Wang,[3] Yongyan Zhao,[1,2] Luyao Wang,[2] Jiaying Pan,[1,2] Huan Mei,[1] Jin Han,[1] Siyuan Wang,[1] Kening Lu,[4] Menglin Li,[4] Mengtao Gao,[4] Zeyi Cao,[1] Hailin Zhang,[1] Ke Wan,[4] Jie Li,[4] Lei Fang,[1,2] Tianzhen Zhang,[1,2] and Xueying Guan[1,2,5,*]

[1]Zhejiang Provincial Key Laboratory of Crop Genetic Resources, The Advanced Seed Institute, Plant Precision Breeding Academy, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 300058, China

[2]Hainan Institute of Zhejiang University, Building 11, Yonyou Industrial Park, Yazhou Bay Science and Technology City, Yazhou District, Sanya 572025, China

[3]Hubei Hongshan Laboratory, Wuhan 430070, China

[4]State Key Laboratory of Crop Genetics and Germplasm Enhancement, Cotton Hybrid R & D Engineering Center (the Ministry of Education), College of Agriculture, Nanjing Agricultural University, Nanjing 210095, China

[5]Lead contact

*Correspondence: xueyingguan@zju.edu.cn

https://doi.org/10.1016/j.celrep.2023.113111

SUMMARY

The dissection of a gene regulatory network (GRN) that complements the genome-wide association study (GWAS) locus and the crosstalk underlying multiple agronomical traits remains a major challenge. In this study, we generate 558 transcriptional profiles of lint-bearing ovules at one day post-anthesis from a selective core cotton germplasm, from which 12,207 expression quantitative trait loci (eQTLs) are identified. Sixty-six known phenotypic GWAS loci are colocalized with 1,090 eQTLs, forming 38 functional GRNs associated predominantly with seed yield. Of the eGenes, 34 exhibit pleiotropic effects. Combining the eQTLs within the seed yield GRNs significantly increases the portion of narrow-sense heritability. The extreme gradient boosting (XGBoost) machine learning approach is applied to predict seed cotton yield phenotypes on the basis of gene expression. Top-ranking eGenes (*NF-YB3*, *FLA2*, and *GRDP1*) derived with pleiotropic effects on yield traits are validated, along with their potential roles by correlation analysis, domestication selection analysis, and transgenic plants.

INTRODUCTION

Genome-wide association studies (GWASs) are a common method for detecting associations between genetic variation and phenotype. GWASs can be traced back to the first decade of the 2000s,[1,2] or earlier. To date, tens of thousands of associated loci have been cataloged in major crops such as rice,[3] wheat,[4] maize,[5] and cotton.[6] However, although GWASs have been very successful in identifying loci associated with phenotypes, they still suffer from major limitations when it comes to pinning down causal genes, because of issues with population structure, the missing heritability of rare variations, and effects from *trans* regulation networks, among other issues.

Notably, although accumulated common genomic variants with small effect sizes can contribute to various traits, they may be filtered out from GWAS results by a stringent significance threshold.[7,8] In addition, GWASs focus on genomic variations in the form of common SNPs and neglect rare variations with minor allele frequency (MAF) values of less than 1%–5%.[9] More important, GWASs are limited in their ability to pinpoint causal variants and candidate genes because of the resolution of linkage disequilibrium (LD) in small population sample size. LD blocks can range in size from 30 kb in a common maize population[10] to ~100–200 kb in cultivated rice[3] and ~300–500 kb in cotton.[11,12] Consequently, the candidate genes associated within an LD block can range in number from a few to a few hundred. Theoretically, any gene within the LD block of a GWAS locus could not be excluded as having a causal effect on the corresponding trait. Additionally, the majority of GWAS loci are located in non-coding region and likely manifest their effects by regulating distant gene expression in *trans*.[13] Unfortunately, GWASs cannot accommodate the direct identification of genome-spanning gene regulatory networks (GRNs). Collectively, these limitations prevent further application of GWAS in navigating the critical step of hub gene selection for precision genome editing to improve crops.

One solution for overcoming the limitations of GWASs is to integrate relevant expression data in addition to genetic variations. Notably, gene expression changes are efficient at introducing phenotypic changes in crops.[14] Analysis of expression

quantitative trait loci (eQTLs) is a method of establishing connections between genetic variants and gene expression by identifying expression-associated SNPs (eSNPs) and their associated genes (eGenes). eGenes can locate either within the same region as an eSNP or in a distal region, and eQTL analysis can detect associated eGenes in *cis* and also in *trans*. Alternatively, multiple genes can be regulated by a single *trans*-eSNP, termed an eQTL hotspot, module, or GRN.[15] Therefore, each GRN is composed of eGenes both in *cis* and *trans*. The latest study based on the Genotype-Tissue Expression (GTEx) dataset (version 8) demonstrated that a median of 21% of GWAS loci from 87 tested complex traits colocalized with a *cis*-eQTL when aggregated across 49 tissue types.[16] Gene-gene interactions within a GRN are proposed to be components of the missing heritability for complex traits.[17] The enclosed gene number of GRNs, termed eQTL hotspots, ranges widely. Thus, although integrated eQTL and GWAS analysis can provide functional GRNs associated with phenotypes, prioritizing important genes in each GRN and their power to affect the phenotype is still a challenge.

Extreme gradient boosting (XGBoost) is a machine learning method, specifically a type of decision tree ensemble model for classification and regression modeling.[18,19] This tool is remarkable for its ability to process missing data efficiently and flexibly. It can also assemble weak prediction models, from which a reliable one can be built.[18,19] In a competition hosted by Kaggle.com, XGBoost was found to be the best algorithm for machine learning and prediction.[20] As it can evaluate the degree of feature importance, XGBoost can be used to prioritize genes according to their criticality, as has been reported in human populations.[21,22] Pioneering research has been conducted in plants, specifically mining N-responsive genes in *Arabidopsis thaliana* and maize.[23] However, the application of XGBoost to populations of crops or other plants is still in a preliminary stage.

Previous GWASs in cotton have revealed associated loci for multiple important agronomic traits including fiber production,[6,12] seed cotton yield,[24,25] fiber quality,[11,26] and abiotic stress tolerance.[27–29] Other recent studies have used eQTL hotspot analysis methods to dissect the genetic GRNs that regulate fiber quality traits and pollen sterility.[30,31] Nonetheless, although a large number of phenotype-associated loci have been reported in cotton populations, the causal genes and other important genes within functional GRNs are still largely unknown because of the lack of data mining methodology.

To characterize the genetic basis of cotton seed size and yield, identify causal genes, and elucidate the underlying GRNs associated with GWAS loci, we designed an integrative eQTL and GWAS analysis using transcriptomes from the China upland cotton population, CUCP1. The identified eGenes clustered into 38 functional GRNs on the basis of the colocalization of expression-associated lead SNPs (eSNPs) and phenotype-associated lead SNPs (pSNPs) within LD blocks. The joint additive effect of yield-related GRNs was validated on the basis of narrow heritability. Using XGBoost-derived feature importance ranking, the causal genes *NF-YB3*, *FLA2*, and *GRDP1* from GRN were validated as having functional impacts on seed development.

RESULTS

Study overview

Figure 1 illustrates the aim of this work, which is to construct GRNs and mine the genes that are important for seed size and fiber yield.

Data

A core germplasm was collected for the China upland cotton population, CUCP1, comprising a total of 279 *Gossypium hirsutum* accessions, including 34 wild/landrace accessions and 245 cultivated accessions. The collection of cultivated accessions is adapted from our previous GWAS catalog (Table S1).[12]

Cotton fiber differentiates from ovule epidermis at about −1 to 1 day post-anthesis (DPA). Approximately 25%–30% of the epidermal cell can differentiate into fiber cells, which largely determine the fiber yield on each seed.[32,33] To understand the expression variation associated with genetic variation in CUCP1 at the fiber-yield determining stage, we profiled the transcriptomes of 1 DPA lint-bearing ovules for all 279 accessions, with two biological replicates (Figure 1). The transcriptomes collectively provided 13.82 billion (mean 24.90 million per sample) paired-end reads, with an average unique mapping rate at 97.11% (cultivar 97.23%, wild 96.32%) to the TM-1 cotton reference genome[34] (Table S1). Previous reports indicate that most GWAS loci were mapped to non-coding regions, potentially pointing to non-coding variants.[13] Accordingly, to detect as many causal genes as possible, we also annotated long non-coding RNAs (lncRNAs) and quantified their transcription for eQTL mapping (see STAR Methods). A total of 37,108 protein-coding genes (PCGs) and 6,251 lncRNAs met the criteria for expressed genes (see STAR Methods), accounting for 50.99% of all annotated genes in the upland cotton TM-1 (version 2.0) reference genome[34]; These were used for further analysis. In parallel, whole-genome sequencing (WGS) of the accessions generated a total of 1,186,673 biallelic high-quality SNPs (MAF > 0.05 and missing ratio < 20%), which were used for eQTL mapping (Figure 1).[12]

Workflow

First, GWASs and eQTLs were integrated to obtain pSNPs and eSNPs, respectively. eGenes associated with the same eSNP were grouped as GRNs. pSNPs and eSNPs within the same LD block ($r^2 > 0.1$) were defined as having colocalization. Accordingly, a GRN also colocalized with a pSNP was considered as a functional GRN. Second, the eGenes in functional GRNs were used as the features for the XGBoost algorithm in predicting phenotype regression. The model's performance was evaluated using Pearson correlation coefficients (PCCs). Next, the eGenes were ranked according to the feature importance score exported from the model. Third, the top-ranked eGenes were selected for functional validation using heritability analysis, domestication sweep identification, and transgenic plants (Figure 1).

A map of eQTLs associated with fiber-bearing ovule development

The quality of the 558 (279 × 2) transcriptomes and their applicability to eQTL mapping were evaluated by calculating PCCs from

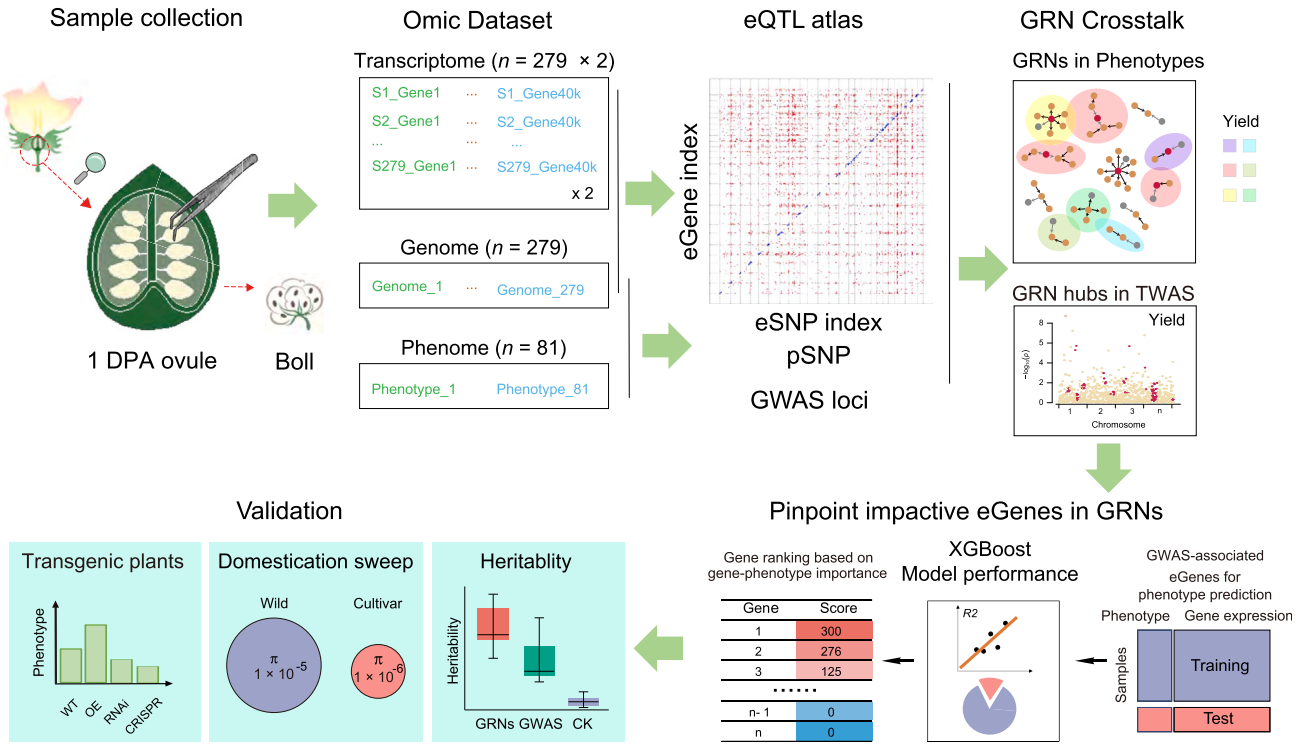


Figure 1. Graphic summary of datasets and analyses performed in the present study

The principal goal is to dissect the genetic networks underlying phenotypic correlations and mine important genes. This schematic chart represents our datasets and methods for network dissection and prioritization of important eGenes by integrating multiple omics (transcriptome, genome, and phenome).

the transcriptome profiles. The PCCs of the two biological replicates (mean r = 0.93) were found to be significantly higher than those of different accessions (mean r = 0.77, $p < 1 \times 10^{-16}$, Mann-Whitney test) (Figure 2A). Principal-component analysis (PCA) was also used to reveal the genetic similarities and differences of expression patterns (Figure 2B).

eQTL mapping was subsequently performed using Efficient Mixed Model Analysis Expedited (EMMAX) using the obtained SNPs and expression profiles. A total of 12,207 eQTLs were detected, involving 8,088 eSNPs and 6,449 eGenes (n = 5,197 PCGs, n = 1,252 lncRNAs), under a suggested threshold of $p < 2.18 \times 10^{-6}$ (Figure 2C; Table S2).[30] An average of 1 or 2 eQTLs were mapped for each eGene (Figure 2D), suggesting that the expression variation is under relatively simple genetic control. The mapped eQTLs were further classified as *cis* or *trans* according to relative eGene location using an empirical value (i.e., the SNP was within ±1 Mb of the transcription start site [TSS] or transcription termination site [TTS] of each gene),[35,36] yielding 3,444 *cis* eQTLs (involving 1,185 eGenes) and 8,763 *trans* eQTLs (involving 5,869 eGenes) (Figure 2E). For *cis*-eQTLs, the associated lead eSNPs were distributed predominantly in adjacent genes and enriched in proximity to TSSs or TTSs (Figures S1A and S1B). *Cis*-eQTLs showed higher association than *trans*-eQTLs did ($p < 2.2 \times 10^{-16}$, Mann-Whitney test) (Figure S1C). More than 77% of the eQTLs were shared between the two biological replicates (Figure S1D; Table S2). In addition, 21.42% of eQTLs showed variance between the wild and cultivated accessions (Figures S1E and S1F). The distribution patterns and frequencies of eQTLs reported here are consistent with previous reports in maize seedlings,[37] *Brassica napa* seeds,[38] rice shoots,[39] cotton 15 DPA fibers[30] and *Arabidopsis* shoots.[40] With regard to chromosomal location, the eQTLs identified here exhibited a significantly disproportionate distribution, forming 293 eQTL hotspots (Figure 2F) The most notable hotspots spanned 1,756 kb (from 88,974,035 to 90,730,903 bp) on chromosome ChrA07 and 7.23 kb (from 2,899,413 to 2,906,643 bp) on D08, which overlap with GWAS loci (Figure 2G).

Phenotypic relevance of GRNs derived from eQTLs

To systematically characterize the GRNs derived from eQTL analysis, eGenes either in *cis* or in *trans* associated the eSNPs within the same LD block ($r^2 > 0.1$) were grouped as one GRN (see STAR Methods). This yielded 1,014 GRNs, with the number of eGenes in each GRN ranging from 2 to 527, with an average of 13.

The pSNP dataset was adapted from the previous GWAS catalog and represents 187 GWAS loci.[12] The best linear unbiased prediction (BLUP) values for each trait were also calculated on the basis of phenotypic data from nine environments (Table S3). The association signals identified by BLUP and SI (seed index; the weight of 100 seeds) trait collected in 2018 (Dangtu) were consistent with those GWAS loci identified by the phenotype in 2007, 2008, and 2009, respectively (Figure S2A). With the above analysis, the pSNPs used in this study were reliable.

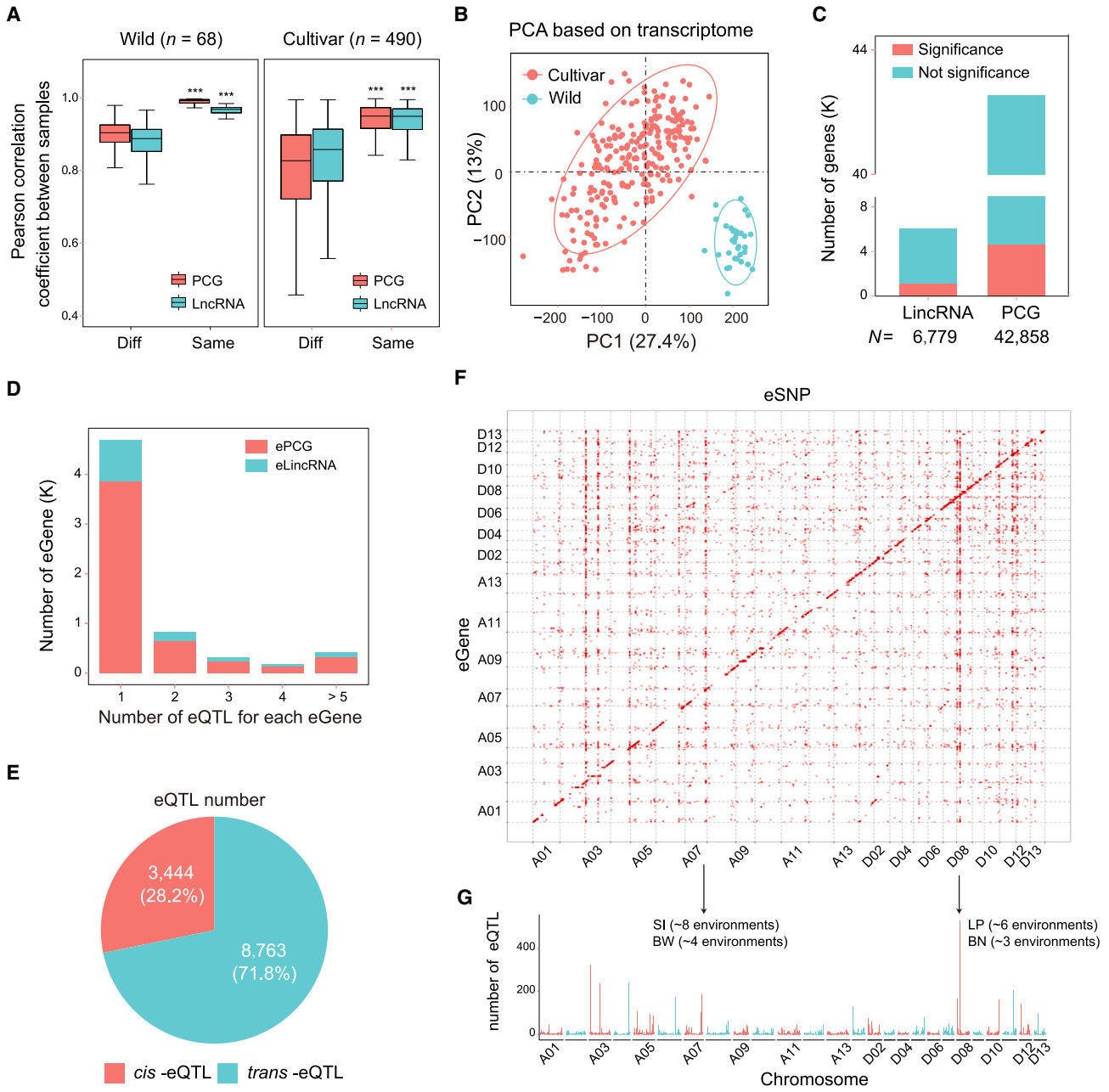


Figure 2. eQTL map for one day post-anthesis (DPA) lint-bearing ovules from 588 samples

(A) Pearson correlation coefficient (PCC) of samples on the basis of protein-coding gene (PCG) and lncRNA expression quantifications; correlations compare replicates of the same accession (Same) and randomly selected samples from different accessions (Diff). Quantifications were normalized to FPKM (fragments per kilobase of transcript per million mapped reads) before calculating the pairwise PCC. The boxplot shows the median and interquartile range (IQR). The end of the top line is the maximum or the third quartile (Q) + 1.5 × IQR. The end of the bottom line denotes either the minimum or the first Q − 1.5 × IQR. Dots are either more than the third Q + 1.5 × IQR or less than the first Q − 1.5 × IQR. Asterisks indicate significant differences by two-tailed Mann-Whitney test (***$p \leq 0.001$).

(B) Distinct separation of wild and cultivar groups was observed with principal-component analysis (PCA) of transcriptome profiles.

(C) Numbers of lncRNAs and PCGs associated with eQTLs.

(D) Number of eQTLs mapped for each eGene. The x axis represents the number of eQTLs mapped for each eGene, and the y axis represents the number of eGenes in each group (PCG and lncRNA).

(E) Pie chart showing the number and proportion of *cis*- and *trans*-eQTLs.

(F) Scatterplot of 8,088 high-confidence eSNP-expression associations, with expression of 6,449 eGenes (y axis) against 12,207 eQTLs. Each dot represents a detected eQTL.

(G) eQTL hotspot distribution across the genome, was determined in 1 Mb windows. The y axis indicates number of eGenes and is plotted against genetic location. Arrows indicate two eQTL hotspots co-located with cotton yield GWAS loci. SI, seed index; BW, boll weight; BN, boll number; LP, lint percentage.

To determine whether the identified GRNs tended to have functional consequences associated with phenotype, the pSNP and eSNP on the same LD were defined as being colocalized with each other. Accordingly, 38 GRNs (3.75% [38 of 1,014]) were colocalized with GWAS loci with the associated function (Figure 3A; Table S4). In total, this involved 1,090 eQTL and 701 non-redundant eGenes (Figure 3B; Table S5).

Among these 38 functional GRNs, 30 GRNs containing 657 eGenes were related to yield traits, while the other 8 GRNs containing 44 eGenes were related to fiber quality traits (Figure 3B; Table S5). Notably, it was notable that GRN_302 and GRN_808 were associated with yield phenotypes, together dominating (60.36%) the total eGenes in the functional GRNs (Tables S4 and S5). In detail, GRN_302 contained 230 eGenes, and its feature eSNP on chromosome A07 (A07:90680544) was colocalized with a GWAS locus represented by a pSNP associated with SI and BW (boll weight) (Figure 3B; Table S5). Meanwhile, GRN_808 contained 169 eGenes with the feature eSNP on chromosome ChrD08 (D08:2903486) was colocalized with a GWAS locus associated with LP (lint percentage) and BN (boll number) (Figure 3B; Table S5).

Previous studies have proposed that eGenes within the same GRN should exhibit similar expression patterns in unique cell types relevant to the phenotype of interest.[41] To confirm whether the eGenes in the same GRNs identified here were reliable with similar biological functions, their transcriptional activity was examined using previously published transcriptome profiles from 17 tissues in the upland cotton accession TM-1, which encompassed all developmental stages of seed and fiber.[42] For both GRN_302 and GRN_808, the eGenes exhibited a general trend of tissue-specific expression (Figure 3C). Specifically, over 60% of eGenes in GRN_302 were highly expressed in early ovule/seed development (about 0–5 DPA) (Figure 3C), while most eGenes in GRN_808 showed similar expression pattern in early ovule and fibers (Figure 3C). In addition, 13 eGenes within GRN_302 were found to be homologous to genes with reported roles in seed and embryo development, such as *FLA2*,[43] *EMB3147*, and *EMB2735* (Figure 3C). Within GRN gene-gene interactions were further examined using a protein-protein interaction (PPI) network database.[44] About 78.33% of the genes in GRN_302 and 79.50% of those in GRN_808 were supported as interactors by this PPI data. These proportions are higher than among randomly selected eGenes and genes in same LD (Figure 3D). The above analysis confirmed the eGenes in the GRNs revealed here are highly likely associated with seed development, with potential PPIs.

Crosstalk between functional GRNs involved in seed cotton yield

In the present study, each eGene was mapped with 1 or 2 eQTLs (Figure 2D), suggesting that eGenes can be associated with more than one genetic variation. Thus, pairwise comparisons were performed to identify eGenes shared by multiple functional GRNs. This yielded 18 instances of shared eGenes, indicative of crosstalk between different functional GRNs (Figure 4A). For example, the dominant network GRN_302 shared eGenes with GRN_96 and GRN_243 (Figure 4A); all three of these GRNs were associated with yield phenotypes at significance levels below the threshold of $p < 2.18 \times 10^{-6}$. Specifically, GRN_302 was associated with SI and BW, while GRN_96 and GRN_243 were associated with SI and LP (Figure 4B).

Regarding specific genes in the regulatory networks, *NF-YB3* (*GH_A07G2187*) and *FLA* (*GH_A07G2189*) are two *cis*-eGenes in GRN_302 (Figures 4C and S2B). A Manhattan plot showed that the genomic variants within the corresponding GWAS locus were significantly associated with the expression of *NF-YB3* and *FLA2* in all examined environments (Figures 4C, S2C, and S2D). *GRDP1* (*GH_A02G1719*) is a *trans*-eGene in GRN_302, but a *cis*-eGene in GRN_96 (Figure 4C). eQTL mapping of *GRDP1* revealed two regulatory variants located in two SI-GWAS loci on ChrA02 and ChrA07, which represent *cis*- and *trans*-regulation patterns in GRN_96 and GRN_302, respectively. This finding indicates an interaction between GRN_302 and GRN_96 in controlling SI (Figure 4C). Similarly, *IDD7* (*GH_A06G0949*) is a *trans*-eGene in GRN_302, and a *cis* -eGene in GRN_243 (Figure 4B). eQTL mapping of the *cis*-eGene *IDD7* likewise identified two regulatory variants located in two LP-GWAS loci in ChrA06 and ChrA07, which represent *cis*- and *trans*-regulation patterns in GRN_243 and GRN_302, respectively (Figure 4D).

Here we also adopted the transcriptome-wide and regulome-wide association studies (TWAS)[45] to validate the presumed causal role of the *NF-YB3*, *FLA2*, and *GRDP1*. In total, 297 expression-phenotype associations were found, involved with 83 PCGs, and 13 lncRNAs ($p < 9.2 \times 10^{-4}$) (Figure 4E; Table S6). Consistent with the integrative functional GRNs, *NF-YB3*, *FLA2*, and *GRDP1* all achieved significance in the TWAS on seed cotton yield traits (Figure 4E; Table S6).

Together with the shared eGenes from different GRNs, whether *cis* or *trans*, these GRNs might form an interactive connection that collectively constitutes a potential network regulating seed cotton yield (Figure 4F).

Prioritizing the important genes in GRNs using XGBoost

The power of each eGene in phenotype regulation is still unclear. Here we used the XGBoost algorithm to prioritize the most important eGenes. Taking the screened eGenes within major functional GRNs as features, XGBoost was used to construct regression models for the phenotype in each environment (Figure 5A). The mean PCC between the true values from the test data and the predicted values was high in SI and LP (Figure 5B). FL (fiber length) and FS (fiber strength) phenotypes were determined as controls, for which the obtained r values were lower than 0.1 (Figure 5B). This confirms the impacts of the identified GRNs on seed cotton yield.

Next, the feature importance for measuring the distinction in prediction was exported from the XGBoost model (Table S7). Among the important genes so identified, the potential pleiotropic genes *NF-YB3*, *FLA2*, *GRDP1*, and *IDD7* were ranked at the top (Figure 5C).

The interactive GRNs captured the missing heritability for seed size

Together with the shared eGenes, the GRNs form an interactive connection that comprises a potential regulatory network for seed cotton yield. In validating the effects of this extended network, a key question is whether the detected GRNs can

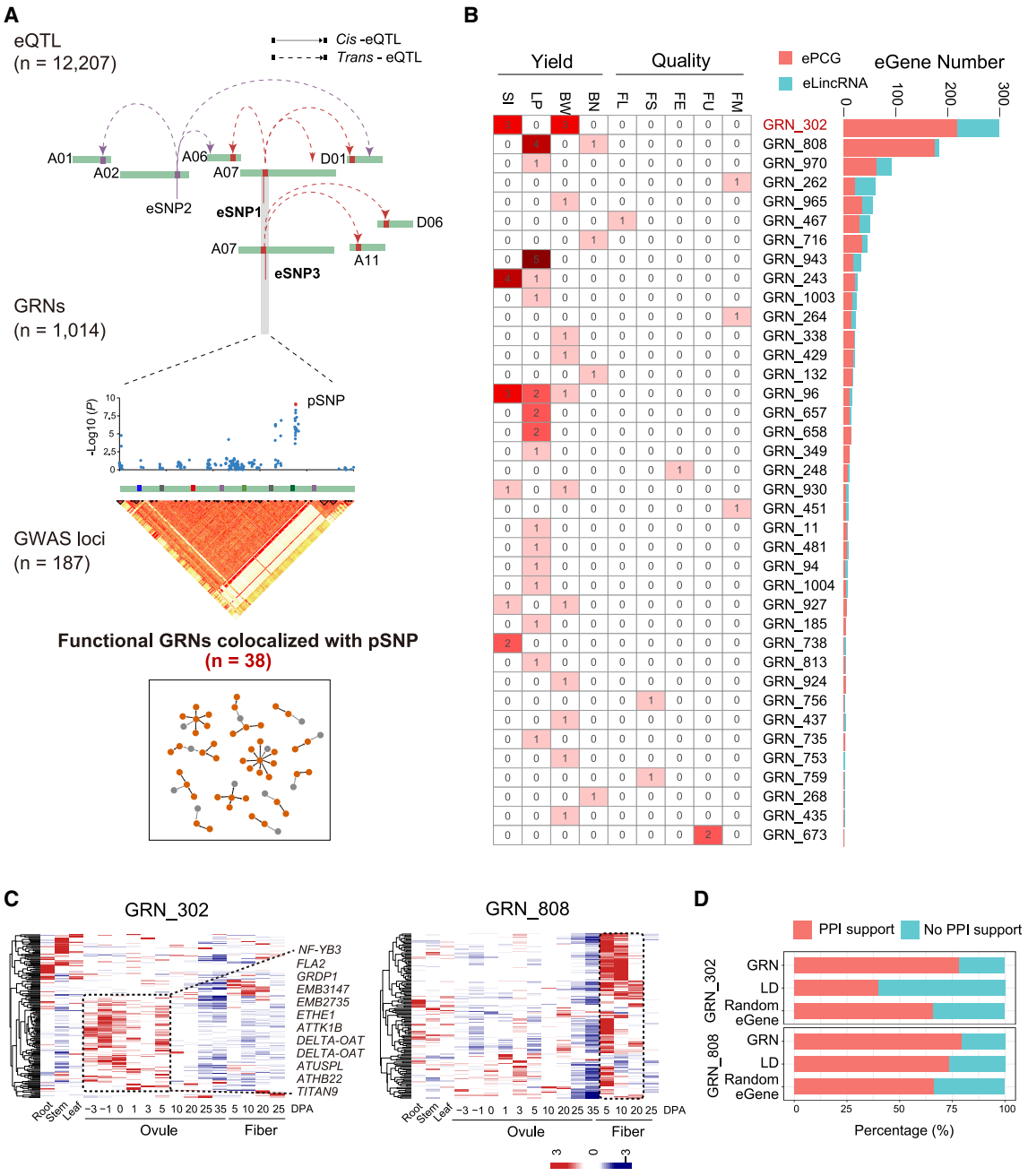


Figure 3. Gene regulatory networks (GRNs) with phenotype-associated feature SNPs

(A) Analytical workflow for functional GRN construction. Both GWAS and eQTL analysis were conducted to obtain phenotype-associated lead SNPs (pSNPs) and gene expression-associated lead SNPs (eSNPs), respectively. Those eSNPs/pSNPs within the same LD block ($r^2 > 0.1$) were merged into one lead SNP, and eGenes within an LD block were clustered into a GRN. GRNs having feature pSNPs were considered to be functional GRNs.

(B) Heatmap showing the 38 functional GRNs and their genotypic associations. Numbers in boxes indicate how many times the lead causal SNP satisfied the p value threshold for genome-wide significance in the GWAS. The bar plot in the right panel shows the number of PCGs (red) and lncRNAs (blue) in each GRN.

(C) Heatmap showing expression of eGenes in GRN_302 and GRN_808 across different tissues. Thirteen eGenes reported to play roles in seed and embryo development are highlighted.

(D) Bar plot showing the percentage of eGenes in GRN_302 and GRN_808 that are also identifiable as interactors in the STRING protein-protein interaction database (https://cn.string-db.org/). Randomly selected eGenes and genes in LD blocks were used as controls.

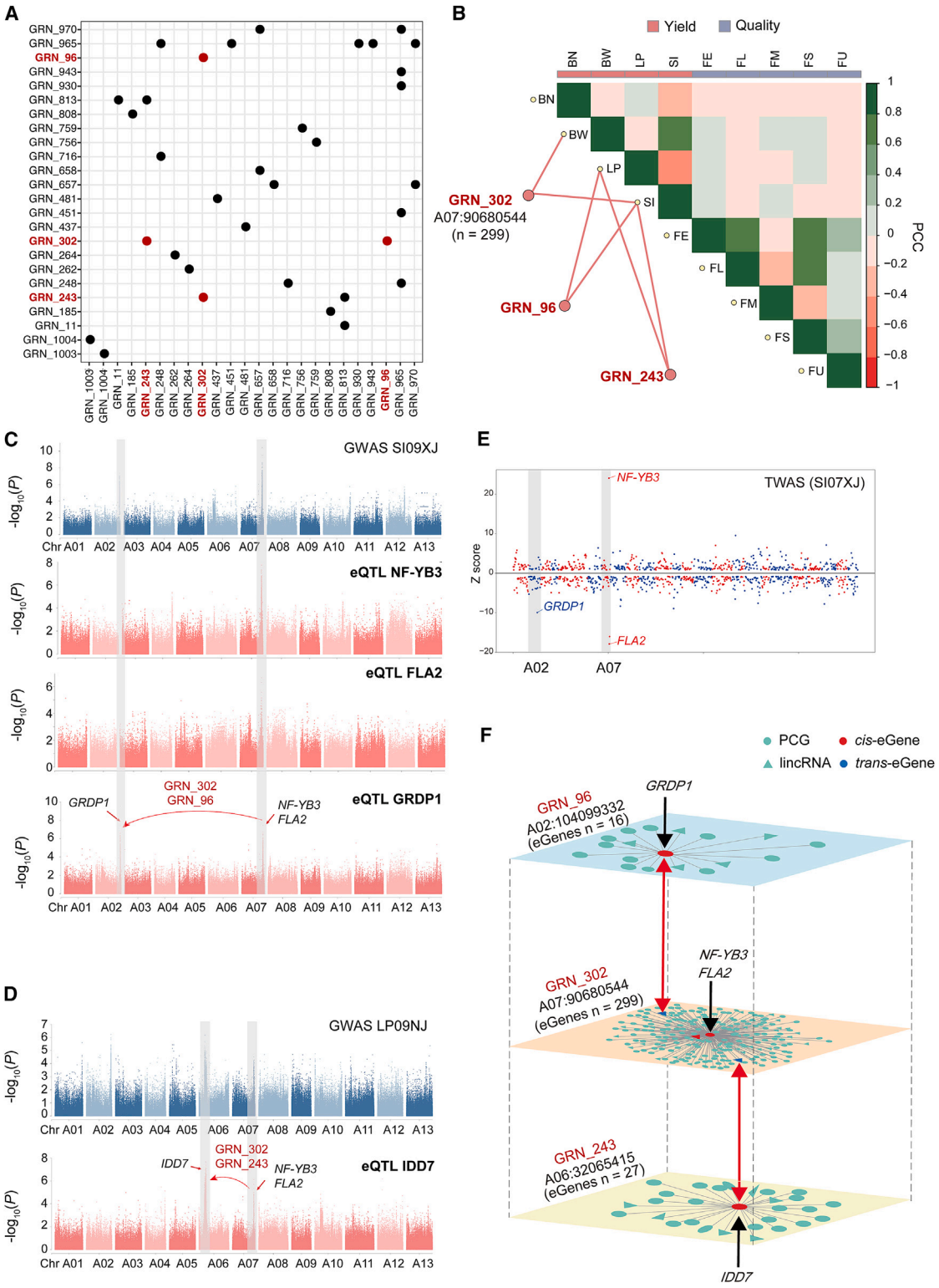


Figure 4. The joint additive effect of yield-related GRNs

(A) Overlap of eGenes across different functional GRNs. The GRNs that shared genes with GRN_302 are colored red.

(B) The relationship between representative GRNs (GRN_302, GRN_96, and GRN_243) and their associated phenotypes. The heatmap shows the Pearson correlation coefficients (PCCs) of different phenotypes.

(legend continued on next page)

provide quantitative power to explain the heritability, with increased portion. To quantify the relative genetic contribution of each GRN to phenotypic variation, narrow-sense heritability (h^2) can be calculated using the local variants.[46] Hereafter, h^2_{GRN} indicates the local variance explained by SNPs corresponding to *cis*- and *trans*-eGenes within a GRN. In the interests of comparison, we also calculated the variance explained by GWAS loci and randomly selected eSNPs, denoted $h^2_{GWAS\ +\ random}$, as well as the variance explained solely by randomly selected eSNPs, termed h^2_{random}.

The combined heritability of the effect of GRN_302 on SI was found to be significantly increased by about 3-fold (19.34%) compared with that of GWAS loci alone (7.00%), while the heritability of the same number of randomly selected eGenes is 1.84% (Figures 6A and S3A). The higher heritability of GRN_302 is highly likely because of the two *trans*-eGenes associated with the causal variants from GRN_96 and GRN_243 (Figure 4F). To test this hypothesis, the combined heritability estimated for GRN_302+GRN_96+GRN_243 were evaluated, the result showed that the combined heritability achieved an even higher level of significance compared with GWAS loci comprised of the causal variants of GRN_302, GRN_96, and GRN_243 ($h^2_{GRN302+GRN\_96+GRN243}$, 20.81%; h^2_{GWAS}, 8.86%; h^2_{random}, 1.50%) (Figures 6B and S3B), indicating that GRNs can capture phenotype-associated genes that have minor effects and are undetectable by GWAS. For the seed size associated phenotypes of SI, LP, BW, and BN, h^2_{GRN} was significantly higher than either $h^2_{GWAS,\ random}$ or h^2_{random} ($p = 2.2 \times 10^{-16}$, Mann-Whitney test) (Figures 6C and S3C). Meanwhile, the joint effects of GRNs as represented by h^2_{GRN} did not affect fiber quality traits (FS, FL, fiber uniformity [FU], and fiber micronaire [FM]), which is consistent with the GWAS results (Figures 6C and S3). This control confirmed that the GRNs reveal at 1 DPA ovule stage using eQTL network can predominantly explain the regulation on seed development and seed size.

NF-YB3, *FLA2*, and *GRDP1* are identified as the genes most likely responsible for the seed size phenotype

NF-YB3, *FLA2*, and *GRDP1* were top-ranked eGenes prioritized by machine learning and TWAS (Figure 5C). The feature eSNP on GRN_302 is located within a GWAS locus for which the LD block spans a 270.74 kb region on chromosome A07 and contains 11 annotated genes (Figure S3B). *NF-YB3* and *FLA2* are two *cis*-eGenes on this locus (Figures 4C and 7A). The expression of *NF-YB3* and *FLA2* was significantly correlated with SI and BW phenotypes (Figures 7B and S4). The homolog of *NF-YB3* in *Arabidopsis* is *AT4G14540*, which encodes a nuclear factor Y transcription factor, NF-YB3, that has been well-studied in embryogenesis and seed development.[47–54] The ectopic expression of cotton *NF-YB3* decreased the seed size in the transgenic *Arabidopsis* (Figure S5), which confirmed its direct impact on seed development. In addition, these data confirmed the eQTL analysis can efficiently navigate to seed regulation genes. The homolog of *FLA2* in *Arabidopsis* is *AT4G12730*, which encodes fasciclin-like arabinogalactan-protein 2 with reported function in seed development.[43,55] *FLA2* is actively expressed in the leaf, flower, and early ovules during seed development in both *Arabidopsis* and cotton (Figure S6). Statistical analysis further revealed that the expression of these two eGenes to be negatively correlated, suggesting that *NF-YB3* and *FLA2* may play antagonistic roles in coordinating seed development (Figures 7C and S4).

GRDP1 is a *trans*-eGene in GRN_302, but a *cis*-eQTL in GRN_96 (Figures 4C and 7A). The *GRDP1* homolog in *Arabidopsis* is *AtGRDP1* (*AT2G22660*), with reported data exclusive to the expression pattern in the embryo during seed development, seed germination, and ABA response.[56,57] Cotton cultivars of the AA haploid type, which predominate in CUCP1, exhibited relatively low *GRDP1* expression ($p < 10^{-16}$, Mann-Whitney test) (Figure 7C), and variation in *GRDP1* expression was found to be positively correlated with LP and negatively with SI (Figure 7C). In addition, *GRDP1* in GRN_96 is located in a selective sweep region associated with cotton domestication (Figure 7D) and exhibits fixed haplotypes in the cultivated population within CUCP1 (Figure 7D). The inclusion of *GRDP1* in a selective sweep region was also reported in a latest study using an upland cotton population with an even large sample size of 1,000 accessions.[58]

To confirm its function, we constructed transgenic cotton to alter the expression of the endogenous *GRDP1*. The T_2 generation of transgenic seeds from multiple independent lines of overexpressing *GRDP1* (*OE-GRDP1*), RNAi (*RNAi-GRDP1*), and antisense of *GRDP1* (*AS-GRDP1*) were obtained (Figure 7E). Real-time PCR examination confirmed that *GRDP1* expression were significantly higher in the *OE-GRDP1* lines than in the wild-type (WT), and significantly lower in the RNAi lines and antisense lines (Figure 7F). Furthermore, the SI was significantly larger in the *OE-GRDP1* lines than in the WT lines, while it was significantly lower in *AS-GRDP1* and *RNAi-GRDP1* lines (Figure 7G). The seed width and length showed a similar trend. On the contrary, LP was higher in *AS-GRDP1* and *RNAi-GRDP1* lines, while it was significantly higher in *OE-GRDP1* lines compared with WT (Figure 7G).

The above data confirmed that *GRDP1* has a direct effect on cotton seed size, and the natural variations in *GRDP1* structure and expression are of great potential in improving seed cotton yield in cultivated cotton populations. Moreover, the top-ranked eGenes in the integrative study are demonstrated to be the causal genes in the seed size associated GWAS loci.

(C) Manhattan plot for various traits and gene expression; from top to bottom, SI phenotype and expression of *FLA2*, *NF-YB3*, and *GRDP1*. The x axis is the SNP chromosomal location of SNP and the y axis is the strength of the association ($-\log_{10}$[p value]). The causal variations of GRN_302 and GRN_96 are highlighted.
(D) Manhattan plot for various traits and gene expression; from top to bottom, LP phenotype and expression of *IDD7*. The x axis is the SNP chromosomal location and the y axis is the strength of the association ($-\log_{10}$[p value]). The causal variations of GRN_302 and GRN_243 are highlighted.
(E) Manhattan plot of TWAS results for the SI phenotype. Each point represents a single *cis*-eGene. Genes whose expression is positively or negatively correlated with SI are plotted above or below the black bold line. The genomic positions of each eGene are plotted on the x axis. *NF-YB3*, *FLA2*, *GRDP1*, and *IDD7* are highlighted.
(F) Connections across GRN_302, GRN_243, and GRN96. Nodes represent eGenes (circles, PCGs; triangles, lncRNAs). The *cis*-eGenes in each GRN are highlighted in red. Lines represent *trans* regulation.

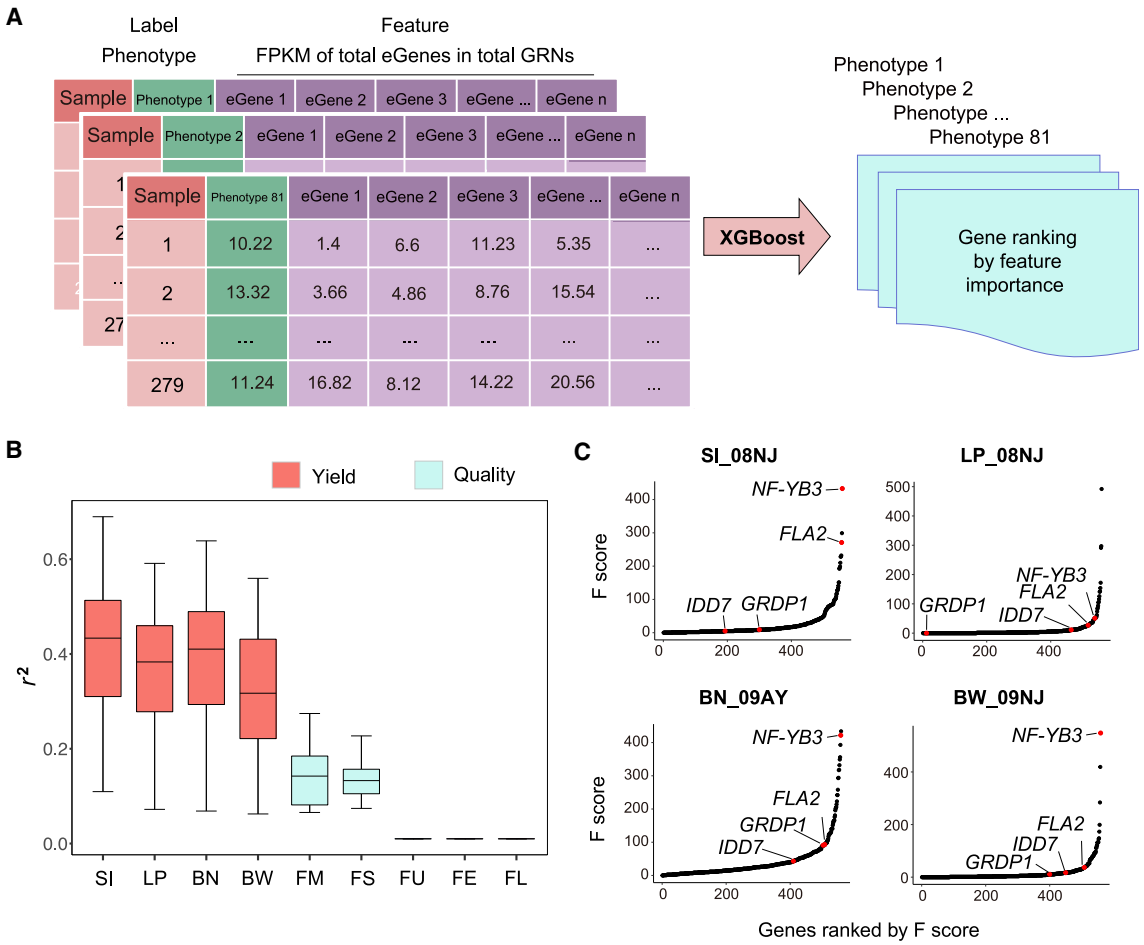


Figure 5. Prioritizing important genes in GRNs using XGBoost

(A) Machine learning workflow. The input data consisted of instances (samples) with labels (phenotypes) and values of features (eGenes). Instances were first split into training and testing sets. The training set was further split into a training subset (90%) and validation subset (10%) in a 5-fold cross-validation scheme. After tuning the model parameters, the optimal model was used to provide performance metrics on the basis of PCC r value between the predicted and actual values in each environment, predict labels in the testing set for model evaluation purposes, and obtain feature importance scores.

(B) Boxplot showing performance on the basis of PCC r value between the predicted and actual values in each phenotype.

(C) Coefficients are averaged from 100 iterations of model building.

(D) Feature importance (F score) of each eGene. The x axis represents genes ordered by F score within each phenotype and the y axis the F score value exported by XGBoost.

DISCUSSION

Mining potentially functional genes using an eQTL map

In this study we constructed an eQTL map of the 1 DPA ovule in cotton, comprising 12,207 eQTLs. This map serves as a bridge between phenotype-associated variation and gene expression. In total, 66 of 187 reported phenotypic GWAS loci were colocalized with *cis*- and *trans*-eQTLs.

Most studies prioritizing genes for complex traits have considered only *cis*-eQTL effects, despite the fact that *trans*-eQTLs account for a substantial portion of the eQTL regulation network. The effect of each *trans*-eQTL is in general as weak because of individual variation[35] and its alternative tissue-specific expression pattern rather than *cis*-eQTLs.[59] In this study, we estimated heritability on the basis of the variants of both *cis*- and *trans*-eGenes in GRN and successfully captured the missing heritability of seed size related traits (Figure 6), which emphasizes the importance of *trans*-eGenes as a group.

Crop domestication often involves genomic sweeps that remove rare variants from cultivated populations.[11] Thus, rare variants associated with phenotypes are difficult to detect by GWAS in a cultivated population.[60] However, if any critical genetic variants are present in an upstream regulatory module, the associated variations in important functional genes should be noticeable accordingly. The effects of these genetic variants on downstream gene expression in a GRN can also be detected as *trans*-eQTLs. The presented results demonstrate an integrative analysis using eQTL and GWAS can be used to construct a GRN can retrieve part of the "lost heritability" by mining a comprehensive resource that points to important genes.

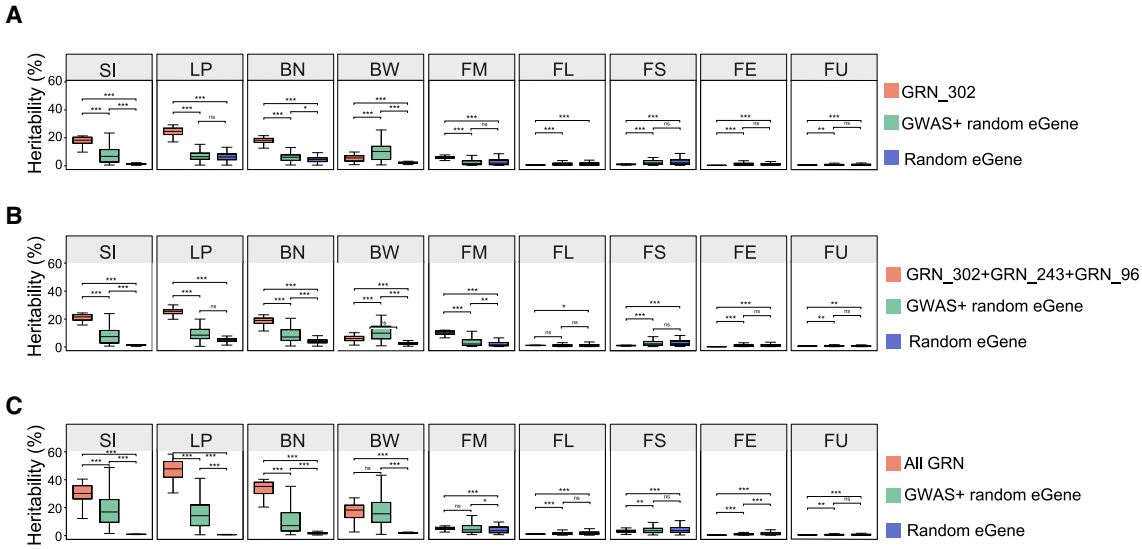


Figure 6. GRNs explained a larger fraction of heritability than the functional GWAS loci

(A–C) Boxplots show the estimated heritability (h^2) across different phenotypes explained by different sets of eSNPs shown in the corresponding color code. Boxes show the medians and IQRs. The end of the top line is the maximum or the third quartile (Q) + 1.5 × IQR. The end of the bottom line denotes either the minimum or the first Q − 1.5 × IQR. Dots indicate values outside those bounds. Asterisks indicate significant differences by two-tailed Mann-Whitney test (* $p \leq 0.05$, ** $p \leq 0.01$, and *** $p \leq 0.001$).

Crosstalk over GRNs reveals the pleiotropy of GWAS loci

Most agronomic traits are controlled by multiple quantitative loci, and many phenotypes tend to be integrated or controlled by pleiotropic genes.[61] Interaction between loci (i.e., epistasis) also increases the complexity of the genetic basis of a phenotype. Thus, the dissection of a single gene or a gene related to one specific trait is insufficient for molecular breeding.

In the present study, we discovered that different genetic variants associated with related but different phenotypes can influence the expression of the same genes via *cis*- or *trans*-regulatory mechanisms. To identify the eGenes that are mutually influenced by different loci, independent functional GRNs were compared in a pairwise manner, which revealed *cis*-eGenes to be shared across GRNs as *trans*-eGenes. The crosstalk between GRNs via *trans*-regulation by eGenes provides insight into the pleiotropy of GWAS loci. Using these pleiotropic genes from GWAS loci can aid in engineering multiple desirable phenotypes through gene and genome editing technologies.

Prioritizing important genes using machine learning

Machine learning methodologies show great promise for analyzing complex biological data. However, working with a large number of predictors (p) and a small number of samples (n) poses a major challenge.[62] This is particularly true for biological data because the large number of genomic variations and genes with expression data that must be examined can dramatically increase computational costs, especially since they are usually much more numerous than the sample.[63,64] This study demonstrates that using genes colocalized in GWAS loci and eQTL GRNs can effectively reduce the dimensionality of biological data for machine learning. Specifically, it narrows the focus from 49,637 expressed genes down to 701 trait-associated genes. By using XGBoost training, the top-ranking 661 genes in cotton yield GRNs were identified. This GRN navigation strategy successfully identified the top-ranking genes in seed size regulation.

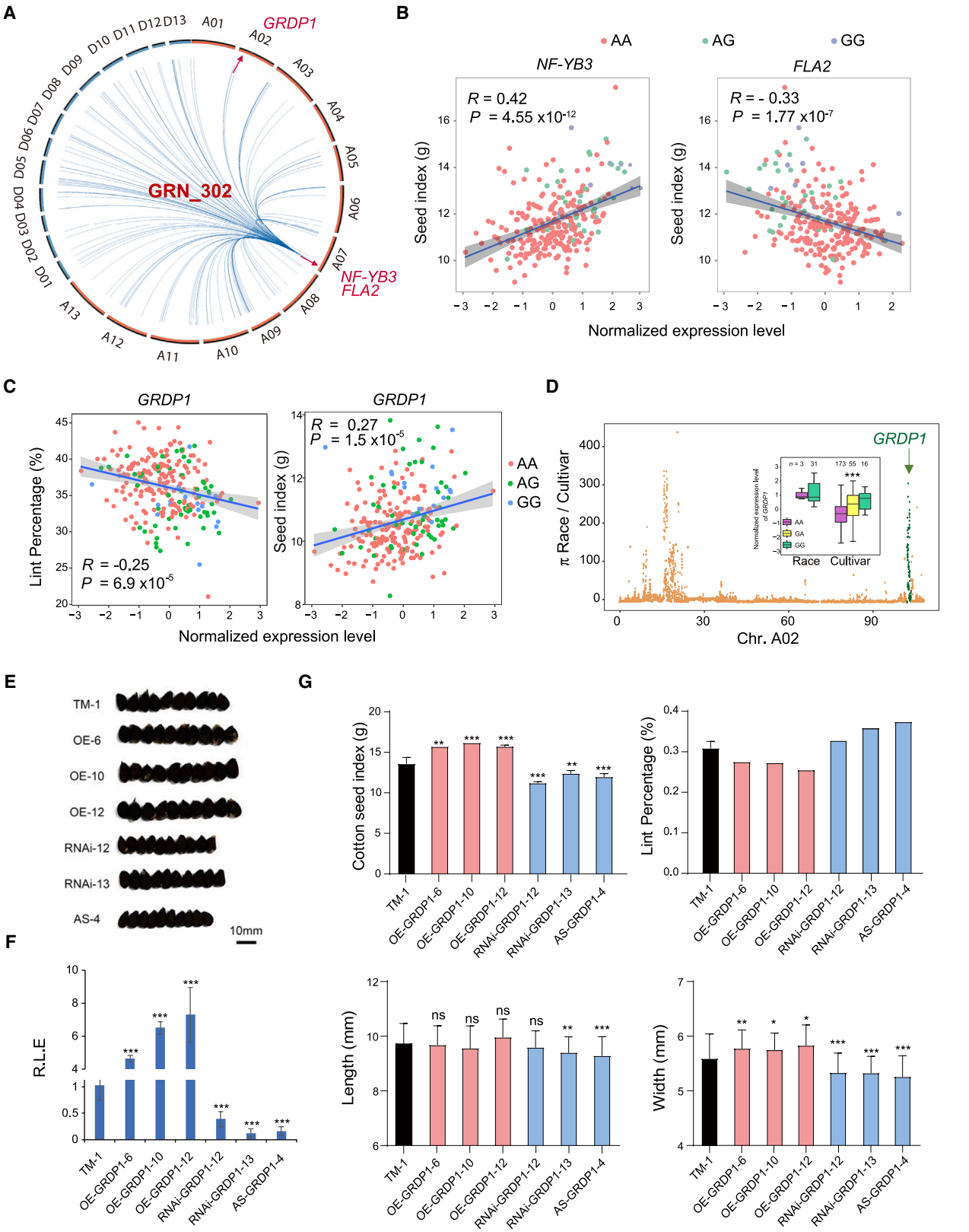
In addition to eQTLs, a wide spectrum of advanced molecular biotechnologies can be used to construct GRNs and reduce the dimensionality of biometrics for machine learning. Typically, GRNs can be characterized on the basis of physical interactions among molecules, or genetic regulatory relationships.[65] Among the state-of-the-art methods for GRN construction, one option is to use the Hi-C technology to establish the three-dimensional (3D) genome, which includes extensive DNA-DNA and DNA-RNA interactions.[66] Single-cell multi-omics can also be applied to uncover GRNs.[67] In future study, GRNs identified through different platforms and technologies can be further used to prioritize important genes with machine learning or deep learning algorithms.

Limitations of the study

Because of the population-wide working load in this study, only one developmental stage of ovule was selected. As a result, eQTLs that are specifically expressed in other tissues were not examined. This study used poly-A selected transcriptome sequencing technology, so any lncRNA without poly A was not considered. Another limitation is that the phenotype and RNA samples were collected in different years. According to our experience and open discussion, the phenotype remained stable across multiple environments.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:



(legend on next page)

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2023.113111.

ACKNOWLEDGMENTS

This work was financially supported in part by grants from the National Key Research and Development Program (2022YFF1001400), the National Natural Science Foundation of China (NSFC; 32000379), the Hainan Provincial Natural Science Foundation of China (323CXTD385 and 320LH002), Fundamental Research Funds for the Central Universities (226-2022-00153) and JCIC-MCP, and High-Performance Computing Platform of YZBSTCACC, National Key Laboratory of Cotton Bio-breeding and Integrated Utilization Open Fund (CB2023A01). We thank the Chinese national medium-term cotton gene bank at the Institute of Cotton Research (ICR) of the Chinese Academy of Agricultural Sciences (CAAS) and National Wild Cotton Nursery (Sanya, China) for kindly sharing the cotton accessions.

AUTHOR CONTRIBUTIONS

X.G. conceptualized the project. H.W., X.W., Y.Z., L.W., J.P., H.M., J.H., S.W., K.L., M.L., M.G., Z.C., H.Z., K.W., J.L. and L.F. conducted the experiments. T.Z. and X.W. performed the bioinformatics analysis. T. Zhao, T. Zhang, and X.G. prepared the manuscript. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: February 27, 2023
Revised: June 19, 2023
Accepted: August 24, 2023
Published: September 6, 2023

REFERENCES

1. Buckler, E.S., Holland, J.B., Bradbury, P.J., Acharya, C.B., Brown, P.J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J.C., et al. (2009). The genetic architecture of maize flowering time. Science *325*, 714–718. https://doi.org/10.1126/science.1174276.
2. Aranzana, M.J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., Lister, C., Molitor, J., Shindo, C., Tang, C., et al. (2005). Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. PLoS Genet. *1*, e60. https://doi.org/10.1371/journal.pgen.0010060.
3. Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., Zhu, C., Lu, T., Zhang, Z., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. *42*, 961–967. https://doi.org/10.1038/ng.695.
4. Juliana, P., Poland, J., Huerta-Espino, J., Shrestha, S., Crossa, J., Crespo-Herrera, L., Toledo, F.H., Govindan, V., Mondal, S., Kumar, U., et al. (2019). Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics. Nat. Genet. *51*, 1530–1539. https://doi.org/10.1038/s41588-019-0496-6.
5. Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., Han, Y., Chai, Y., Guo, T., Yang, N., et al. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat. Genet. *45*, 43–50. https://doi.org/10.1038/ng.2484.
6. He, S., Sun, G., Geng, X., Gong, W., Dai, P., Jia, Y., Shi, W., Pan, Z., Wang, J., Wang, L., et al. (2021). The genomic basis of geographic differentiation and fiber improvement in cultivated cotton. Nat. Genet. *53*, 916–924. https://doi.org/10.1038/s41588-021-00844-9.

Figure 7. Validation of *GRDP1* as a candidate gene of seed size regulation

(A) Circos plot of autosomes indicating the association of eGenes with the locus underlying GRN_302 on chromosome A07 (SNP A07:89225810). Lines with arrows indicate regulation of the expression of downstream genes.

(B) Linear regression analysis of *FLA2* and *NF-YB3* expression and seed index phenotype; the scatterplot illustrates the high inter-tissue correlation. Each point represents one accession. Gene expression values were normalized by normal quantile transform. Points are colored on the basis of SNP (A07:89225810) genotype.

(C) Linear regression analysis of *GRDP1* expression and seed index. The scatterplot illustrates the high inter-tissue correlation. Each point represents one accession. Gene expression values were normalized by normal quantile transform.

(D) Domestication sweeps in wild/landrace and cultivar populations. The value of π (wild)/π (cultivar) is plotted against position on chromosome A02. SNPs close to *GRDP1* are colored green. Boxplot presenting differences of expression in each accession according to index SNP genotypes in the wild/landrace and cultivar populations. Boxes show the medians and interquartile ranges (IQRs). The end of the top line is the maximum or the third quartile (Q) + 1.5 × IQR. The end of the bottom line denotes either the minimum or the first Q − 1.5 × IQR. Dots are either more than the third Q + 1.5 × IQR or less than the first Q − 1.5 × IQR. Asterisks indicate significant differences by two-tailed Mann-Whitney test (***$p \leq 0.001$).

(E) Photo image shows the seed size difference of wild-type, *GRDP1-OE*, *GRDP1-RNAi*, and *GRDP1-AS* lines. Scale bar, 10 mm.

(F) The bar plot shows the relative expression levels of wild-type, *GRDP1-OE*, *GRDP1-RNAi*, and *GRDP1-AS* lines. Asterisks indicate significant differences by two-tailed Student's t test (*$p \leq 0.05$, **$p \leq 0.01$, and ***$p \leq 0.001$).

(G) The histogram shows the weight of seed index, lint percentage, and seed length and width from wild-type, *GRDP1-OE*, *GRDP1-RNAi*, and *GRDP1-AS* lines. Asterisks indicate significant differences by two-tailed Student's t test (*$p \leq 0.05$, **$p \leq 0.01$, and ***$p \leq 0.001$). Error bars represent ±SD (n = 6).

7. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569. https://doi.org/10.1038/ng.608.

8. Gallagher, M.D., and Chen-Plotkin, A.S. (2018). The Post-GWAS Era: From Association to Function. Am. J. Hum. Genet. *102*, 717–730. https://doi.org/10.1016/j.ajhg.2018.04.002.

9. Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M.G., et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. *43*, 519–525. https://doi.org/10.1038/ng.823.

10. Hufford, M.B., Xu, X., van Heerwaarden, J., Pyhäjärvi, T., Chia, J.M., Cartwright, R.A., Elshire, R.J., Glaubitz, J.C., Guill, K.E., Kaeppler, S.M., et al. (2012). Comparative population genomics of maize domestication and improvement. Nat. Genet. *44*, 808–811. https://doi.org/10.1038/ng.2309.

11. Wang, M., Tu, L., Lin, M., Lin, Z., Wang, P., Yang, Q., Ye, Z., Shen, C., Li, J., Zhang, L., et al. (2017). Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. Nat. Genet. *49*, 579–587. https://doi.org/10.1038/ng.3807.

12. Fang, L., Wang, Q., Hu, Y., Jia, Y., Chen, J., Liu, B., Zhang, Z., Guan, X., Chen, S., Zhou, B., et al. (2017). Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. Nat. Genet. *49*, 1089–1098. https://doi.org/10.1038/ng.3887.

13. Ward, L.D., and Kellis, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. Nat. Biotechnol. *30*, 1095–1106. https://doi.org/10.1038/nbt.2422.

14. Lin, H.Y., Liu, Q., Li, X., Yang, J., Liu, S., Huang, Y., Scanlon, M.J., Nettleton, D., and Schnable, P.S. (2017). Substantial contribution of genetic variation in the expression of transcription factors to phenotypic variation revealed by eRD-GWAS. Genome Biol. *18*, 192. https://doi.org/10.1186/s13059-017-1328-6.

15. Das Gupta, M., and Tsiantis, M. (2018). Gene networks and the evolution of plant morphology. Curr. Opin. Plant Biol. *45*, 82–87. https://doi.org/10.1016/j.pbi.2018.05.011.

16. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330. https://doi.org/10.1126/science.aaz1776.

17. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753. https://doi.org/10.1038/nature08494.

18. Yuan, K.C., Tsai, L.W., Lee, K.H., Cheng, Y.W., Hsu, S.C., Lo, Y.S., and Chen, R.J. (2020). The development an artificial intelligence algorithm for early sepsis diagnosis in the intensive care unit. Int. J. Med. Inf. *141*, 104176. https://doi.org/10.1016/j.ijmedinf.2020.104176.

19. Chen, T., and Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining-KDD, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

20. Hou, N., Li, M., He, L., Xie, B., Wang, L., Zhang, R., Yu, Y., Sun, X., Pan, Z., and Wang, K. (2020). Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. J. Transl. Med. *18*, 462. https://doi.org/10.1186/s12967-020-02620-5.

21. Li, Q., Yang, H., Wang, P., Liu, X., Lv, K., and Ye, M. (2022). XGBoost-based and tumor-immune characterized gene signature for the prediction of metastatic status in breast cancer. J. Transl. Med. *20*, 177. https://doi.org/10.1186/s12967-022-03369-9.

22. Pezoulas, V.C., Papaloukas, C., Veyssiere, M., Goules, A., Tzioufas, A.G., Soumelis, V., and Fotiadis, D.I. (2021). A computational workflow for the detection of candidate diagnostic biomarkers of Kawasaki disease using time-series gene expression data. Comput. Struct. Biotechnol. J. *19*, 3058–3068. https://doi.org/10.1016/j.csbj.2021.05.036.

23. Cheng, C.Y., Li, Y., Varala, K., Bubert, J., Huang, J., Kim, G.J., Halim, J., Arp, J., Shih, H.J.S., Levinson, G., et al. (2021). Evolutionarily informed machine learning enhances the power of predictive gene-to-phenotype relationships. Nat. Commun. *12*, 5627. https://doi.org/10.1038/s41467-021-25893-w.

24. Su, J., Wang, C., Ma, Q., Zhang, A., Shi, C., Liu, J., Zhang, X., Yang, D., and Ma, X. (2020). An RTM-GWAS procedure reveals the QTL alleles and candidate genes for three yield-related traits in upland cotton. BMC Plant Biol. *20*, 416. https://doi.org/10.1186/s12870-020-02613-y.

25. Ma, Z., He, S., Wang, X., Sun, J., Zhang, Y., Zhang, G., Wu, L., Li, Z., Liu, Z., Sun, G., 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. https://doi.org/10.1038/s41588-018-0119-7.

26. Liu, W., Song, C., Ren, Z., Zhang, Z., Pei, X., Liu, Y., He, K., Zhang, F., Zhao, J., Zhang, J., et al. (2020). Genome-wide association study reveals the genetic basis of fiber quality traits in upland cotton (*Gossypium hirsutum* L.). BMC Plant Biol. *20*, 395. https://doi.org/10.1186/s12870-020-02611-0.

27. Li, B., Chen, L., Sun, W., Wu, D., Wang, M., Yu, Y., Chen, G., Yang, W., Lin, Z., Zhang, X., et al. (2020). Phenomics-based GWAS analysis reveals the genetic architecture for drought resistance in cotton. Plant Biotechnol. J. *18*, 2533–2544. https://doi.org/10.1111/pbi.13431.

28. Abdelraheem, A., Thyssen, G.N., Fang, D.D., Jenkins, J.N., McCarty, J.C., Wedegaertner, T., and Zhang, J. (2021). GWAS reveals consistent QTL for drought and salt tolerance in a MAGIC population of 550 lines derived from intermating of 11 Upland cotton (*Gossypium hirsutum*) parents. Mol. Genet. Genom. *296*, 119–129. https://doi.org/10.1007/s00438-020-01733-2.

29. Sun, H., Meng, M., Yan, Z., Lin, Z., Nie, X., and Yang, X. (2019). Genome-wide association mapping of stress-tolerance traits in cotton. Crop J. *7*, 77–88. https://doi.org/10.1016/j.cj.2018.11.002.

30. Li, Z., Wang, P., You, C., Yu, J., Zhang, X., Yan, F., Ye, Z., Shen, C., Li, B., Guo, K., et al. (2020). Combined GWAS and eQTL analysis uncovers a genetic regulatory network orchestrating the initiation of secondary cell wall development in cotton. New Phytol. *226*, 1738–1752. https://doi.org/10.1111/nph.16468.

31. Ma, Y., Min, L., Wang, J., Li, Y., Wu, Y., Hu, Q., Ding, Y., Wang, M., Liang, Y., Gong, Z., et al. (2021). A combination of genome-wide and transcriptome-wide association studies reveals genetic elements leading to male sterility during high temperature stress in cotton. New Phytol. *231*, 165–181. https://doi.org/10.1111/nph.17325.

32. Stewart, J.M. (1975). Fiber Initiation on the Cotton Ovule (*Gossypium Hirsutum*). Am. J. Bot. *62*, 723–730. https://doi.org/10.1002/j.1537-2197.1975.tb14105.x.

33. Kim, H.J., and Triplett, B.A. (2001). Cotton Fiber Growth in Planta and in Vitro. Models for Plant Cell Elongation and Cell Wall Biogenesis. Plant Physiol. *127*, 1361–1366. https://doi.org/10.1104/pp.010724.

34. Hu, Y., Chen, J., Fang, L., Zhang, Z., Ma, W., Niu, Y., Ju, L., Deng, J., Zhao, T., Lian, J., et al. (2019). *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. Nat. Genet. *51*, 739–748. https://doi.org/10.1038/s41588-019-0371-5.

35. Võsa, U., Claringbould, A., Westra, H.J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., et al. (2021). Large-scale *cis*- and *trans*-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat. Genet. *53*, 1300–1310. https://doi.org/10.1038/s41588-021-00913-z.

36. He, F., Wang, W., Rutter, W.B., Jordan, K.W., Ren, J., Taagen, E., DeWitt, N., Sehgal, D., Sukumaran, S., Dreisigacker, S., et al. (2022). Genomic variants affecting homoeologous gene expression dosage contribute to agronomic trait variation in allopolyploid wheat. Nat. Commun. *13*, 826. https://doi.org/10.1038/s41467-022-28453-y.

37. Wang, X., Chen, Q., Wu, Y., Lemmon, Z.H., Xu, G., Huang, C., Liang, Y., Xu, D., Li, D., Doebley, J.F., and Tian, F. (2018). Genome-wide Analysis

of Transcriptional Variability in a Large Maize-Teosinte Population. Mol. Plant *11*, 443–459. https://doi.org/10.1016/j.molp.2017.12.011.

38. Tang, S., Zhao, H., Lu, S., Yu, L., Zhang, G., Zhang, Y., Yang, Q.Y., Zhou, Y., Wang, X., Ma, W., et al. (2021). Genome- and transcriptome-wide association studies provide insights into the genetic basis of natural variation of seed oil content in Brassica napus. Mol. Plant *14*, 470–487. https://doi.org/10.1016/j.molp.2020.12.003.
39. Wang, J., Yu, H., Xie, W., Xing, Y., Yu, S., Xu, C., Li, X., Xiao, J., and Zhang, Q. (2010). A global analysis of QTLs for expression variations in rice shoots at the early seedling stage. Plant J. *63*, 1063–1074. https://doi.org/10.1111/j.1365-313X.2010.04303.x.
40. DeCook, R., Lall, S., Nettleton, D., and Howell, S.H. (2006). Genetic regulation of gene expression during shoot development in *Arabidopsis*. Genetics *172*, 1155–1164. https://doi.org/10.1534/genetics.105.042275.
41. Ongen, H., Brown, A.A., Delaneau, O., Panousis, N.I., Nica, A.C., and GTEx Consortium; and Dermitzakis, E.T. (2017). Estimating the causal tissues for complex traits and diseases. Nat. Genet. *49*, 1676–1683. https://doi.org/10.1038/ng.3981.
42. Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., Zhang, J., Saski, C.A., Scheffler, B.E., Stelly, D.M., et al. (2015). Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. Nat. Biotechnol. *33*, 531–537. https://doi.org/10.1038/nbt.3207.
43. Costa, M., Pereira, A.M., Pinto, S.C., Silva, J., Pereira, L.G., and Coimbra, S. (2019). In silico and expression analyses of fasciclin-like arabinogalactan proteins reveal functional conservation during embryo and seed development. Plant Reprod. *32*, 353–370. https://doi.org/10.1007/s00497-019-00376-7.
44. Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. *45*, D362–D368. https://doi.org/10.1093/nar/gkw937.
45. Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A.N., Knowles, D.A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., et al. (2019). Opportunities and challenges for transcriptome-wide association studies. Nat. Genet. *51*, 592–599. https://doi.org/10.1038/s41588-019-0385-z.
46. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011.
47. Mantovani, R. (1999). The molecular biology of the CCAAT-binding factor NF-Y. Gene *239*, 15–27. https://doi.org/10.1016/s0378-1119(99)00368-6.
48. Petroni, K., Kumimoto, R.W., Gnesutta, N., Calvenzani, V., Fornari, M., Tonelli, C., Holt, B.F., 3rd, and Mantovani, R. (2012). The promiscuous life of plant NUCLEAR FACTOR Y transcription factors. Plant Cell *24*, 4777–4792. https://doi.org/10.1105/tpc.112.105734.
49. Kwong, R.W., Bui, A.Q., Lee, H., Kwong, L.W., Fischer, R.L., Goldberg, R.B., and Harada, J.J. (2003). LEAFY COTYLEDON1-LIKE defines a class of regulators essential for embryo development. Plant Cell *15*, 5–18. https://doi.org/10.1105/tpc.006973.
50. Niu, B., Zhang, Z., Zhang, J., Zhou, Y., and Chen, C. (2021). The rice LEC1-like transcription factor OsNF-YB9 interacts with SPK, an endosperm-specific sucrose synthase protein kinase, and functions in seed development. Plant J. *106*, 1233–1246. https://doi.org/10.1111/tpj.15230.
51. Bai, A.N., Lu, X.D., Li, D.Q., Liu, J.X., and Liu, C.M. (2016). NF-YB1-regulated expression of sucrose transporters in aleurone facilitates sugar loading to rice endosperm. Cell Res. *26*, 384–388. https://doi.org/10.1038/cr.2015.116.
52. Bello, B.K., Hou, Y., Zhao, J., Jiao, G., Wu, Y., Li, Z., Wang, Y., Tong, X., Wang, W., Yuan, W., et al. (2019). NF-YB1-YC12-bHLH144 complex directly activates Wx to regulate grain quality in rice (Oryza sativa L.). Plant Biotechnol. J. *17*, 1222–1235. https://doi.org/10.1111/pbi.13048.
53. Pelletier, J.M., Kwong, R.W., Park, S., Le, B.H., Baden, R., Cagliari, A., Hashimoto, M., Munoz, M.D., Fischer, R.L., Goldberg, R.B., and Harada, J.J. (2017). LEC1 sequentially regulates the transcription of genes involved in diverse developmental processes during seed development. Proc. Natl. Acad. Sci. USA *114*, E6710–E6719. https://doi.org/10.1073/pnas.1707957114.
54. Feng, T., Wang, L., Li, L., Liu, Y., Chong, K., Theißen, G., and Meng, Z. (2022). OsMADS14 and NF-YB1 cooperate in the direct activation of OsAGPL2 and Waxy during starch synthesis in rice endosperm. New Phytol. *234*, 77–92. https://doi.org/10.1111/nph.17990.
55. Cagnola, J.I., Dumont de Chassart, G.J., Ibarra, S.E., Chimenti, C., Ricardi, M.M., Delzer, B., Ghiglione, H., Zhu, T., Otegui, M.E., Estevez, J.M., and Casal, J.J. (2018). Reduced expression of selected FASCICLIN-LIKE ARABINOGALACTAN PROTEIN genes associates with the abortion of kernels in field crops of Zea mays (maize) and of Arabidopsis seeds. Plant Cell Environ. *41*, 661–674. https://doi.org/10.1111/pce.13136.
56. Rodríguez-Hernández, A.A., Muro-Medina, C.V., Ramírez-Alonso, J.I., and Jiménez-Bremont, J.F. (2017). Modification of AtGRDP1 gene expression affects silique and seed development in Arabidopsis thaliana. Biochem. Biophys. Res. Commun. *486*, 252–256. https://doi.org/10.1016/j.bbrc.2017.03.015.
57. Rodríguez-Hernández, A.A., Ortega-Amaro, M.A., Delgado-Sánchez, P., Salinas, J., and Jiménez-Bremont, J.F. (2014). AtGRDP1 Gene Encoding a Glycine-Rich Domain Protein Is Involved in Germination and Responds to ABA Signalling. Plant Mol. Biol. Rep. *32*, 1187–1202. https://doi.org/10.1007/s11105-014-0714-4.
58. Yuan, D., Grover, C.E., Hu, G., Pan, M., Miller, E.R., Conover, J.L., Hunt, S.P., Udall, J.A., and Wendel, J.F. (2021). Parallel and Intertwining Threads of Domestication in Allopolyploid Cotton. Adv. Sci. *8*, 2003634. https://doi.org/10.1002/advs.202003634.
59. Consortium, G.T., Laboratory, D.A., Coordinating Center -Analysis Working, G., Statistical Methods groups-Analysis Working, G., Enhancing, G.g., Fund, N.I.H.C., et al.; Nih/Nci; Nih/Nhgri; Nih/Nimh; Nih/Nida (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213. https://doi.org/10.1038/nature24277.
60. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nat. Rev. Genet. *20*, 467–484. https://doi.org/10.1038/s41576-019-0127-1.
61. Wang, Z., Liao, B.Y., and Zhang, J. (2010). Genomic patterns of pleiotropy and the evolution of complexity. Proc. Natl. Acad. Sci. USA *107*, 18034–18039. https://doi.org/10.1073/pnas.1004666107.
62. Ma, C., Zhang, H.H., and Wang, X. (2014). Machine learning for Big Data analytics in plants. Trends Plant Sci. *19*, 798–808. https://doi.org/10.1016/j.tplants.2014.08.004.
63. Chen, Y., Wang, L., Li, L., Zhang, H., and Yuan, Z. (2016). Informative gene selection and the direct classification of tumors based on relative simplicity. BMC Bioinf. *17*, 44. https://doi.org/10.1186/s12859-016-0893-0.
64. Altman, N., and Krzywinski, M. (2018). The curse(s) of dimensionality. Nat. Methods *15*, 399–400. https://doi.org/10.1038/s41592-018-0019-x.
65. Wu, L., Han, L., Li, Q., Wang, G., Zhang, H., and Li, L. (2021). Using Interactome Big Data to Crack Genetic Mysteries and Enhance Future Crop Breeding. Mol. Plant *14*, 77–94. https://doi.org/10.1016/j.molp.2020.12.012.
66. Ouyang, W., Xiong, D., Li, G., and Li, X. (2020). Unraveling the 3D Genome Architecture in Plants: Present and Future. Mol. Plant *13*, 1676–1693. https://doi.org/10.1016/j.molp.2020.10.002.
67. Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nat. Methods *14*, 1083–1086. https://doi.org/10.1038/nmeth.4463.

68. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

69. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

70. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.

71. 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., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

72. Pertea, M., Kim, D., Pertea, G.M., Leek, J.T., and Salzberg, S.L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat. Protoc. *11*, 1650–1667. https://doi.org/10.1038/nprot.2016.095.

73. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578. https://doi.org/10.1038/nprot.2012.016.

74. Kang, Y.J., Yang, D.C., Kong, L., Hou, M., Meng, Y.Q., Wei, L., and Gao, G. (2017). CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. *45*, W12–W16. https://doi.org/10.1093/nar/gkx428.

75. Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. *44*, D279–D285. https://doi.org/10.1093/nar/gkv1344.

76. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. *42*, 348–354. https://doi.org/10.1038/ng.548.

77. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.

78. Falcon, S., and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. Bioinformatics *23*, 257–258. https://doi.org/10.1093/bioinformatics/btl567.

79. Chen, T.a.G., C. (2016). XGBoost: a scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco (Association for Computing Machinery), pp. 785–794. https://doi.org/10.1145/2939672.2939785.

80. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., de Geus, E.J.C., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252. https://doi.org/10.1038/ng.3506.

81. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504. https://doi.org/10.1101/gr.1239303.

82. Silva, I.T., Rosales, R.A., Holanda, A.J., Nussenzweig, M.C., and Jankovic, M. (2014). Identification of chromosomal translocation hotspots via scan statistics. Bioinformatics *30*, 2551–2558. https://doi.org/10.1093/bioinformatics/btu351.

83. Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Usinglme4. J. Stat. Software *67*. https://doi.org/10.18637/jss.v067.i01.

84. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

85. Browning, B.L., and Browning, S.R. (2016). Genotype Imputation with Millions of Reference Samples. Am. J. Hum. Genet. *98*, 116–126. https://doi.org/10.1016/j.ajhg.2015.11.020.

86. Battle, A., Khan, Z., Wang, S.H., Mitrano, A., Ford, M.J., Pritchard, J.K., and Gilad, Y. (2015). Genomic variation. Impact of regulatory variation from RNA to protein. Science *347*, 664–667. https://doi.org/10.1126/science.1260793.

87. Li, M.X., Yeung, J.M.Y., Cherny, S.S., and Sham, P.C. (2012). Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. Hum. Genet. *131*, 747–756. https://doi.org/10.1007/s00439-011-1118-2.

88. von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., and Bork, P. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. *33*, D433–D437. https://doi.org/10.1093/nar/gki005.

89. Fang, L., Zhao, T., Hu, Y., Si, Z., Zhu, X., Han, Z., Liu, G., Wang, S., Ju, L., Guo, M., et al. (2021). Divergent improvement of two cultivated allotetraploid cotton species. Plant Biotechnol. J. *19*, 1325–1336. https://doi.org/10.1111/pbi.13547.

90. Ge, X., Xu, J., Yang, Z., Yang, X., Wang, Y., Chen, Y., Wang, P., and Li, F. (2023). Efficient genotype-independent cotton genetic transformation and genome editing. J. Integr. Plant Biol. *65*, 907–917. https://doi.org/10.1111/jipb.13427.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-DDDDK tag (Binds to FLAG tag sequence)	Abcam	Cat#ab213519
HRP-labeled Goat Anti-Rat IgG(H + L)	Beyotime	Cat#A0192
Bacterial and virus strains		
DH5α Chemically Competent Cell	WEIDI	Cat#DL1001
LBA4404 Chemically Competent Cell	WEIDI	Cat#AC1030
GV3101 Chemically Competent Cell	WEIDI	Cat#AC1001
Biological samples		
Biological samples used in this study, see Table S1	This study	N/A
Chemicals, peptides, and recombinant proteins		
Trizol™	Invitrogen	Cat# 15596018
DNase I	Promega	Cat# Z3585
M-MLV reverse transcriptase	Promega	Cat# M1701
*Xba*I	NEW ENGLAND Biolab	Cat#R0145
*Bam*HI	NEW ENGLAND Biolab	Cat#R0136
Critical commercial assays		
RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system	Agilent	Cat#5067-1511
NEBNext® Ultra™ II RNA Library Prep Kit	NEW ENGLAND Biolab	Cat#E7420S
shoot apical meristem (SAM) cells-mediated transformation system (SAMT)	WIMI Biotechnology Co., Ltd	NA
Deposited data		
RNA sequence data for 279 accessions of 1-DPA cotton ovule.	This study	NCBI Bio Project: PRJNA730082
Whole genome resequencing data for 279 accessions of cotton leaf	Fang et al.[12]	NCBI Bio Project: PRJNA375965
Upland cotton, *G. hirsutum*, ac. TM-1 Reference Genome	Hu et al.[34]	http://cotton.zju.edu.cn/source/TM-1_V2.1.fa.gz
Phenotypic data for fiber traits used in association analysis.	Fang et al.[12]	https://mascotton.njau.edu.cn/info/1058/1132.htm or https://github.com/epi-cotton/eQTL_XGBoost/
Experimental models: Organisms/strains		
G. hirsutum ac. TM-1	College of Agriculture and Biotechnology, Zhejiang University	N/A
Arabidopsis thaliana, Col-0	ABRC (www.arabidopsis.org)	N/A
Oligonucleotides		
Primers used in this study	This paper	Table S8
Recombinant DNA		
pWMV062-AADA-OE-GRDP1	This paper	N/A
pWMV062-AADA-AS-GRDP1	This paper	N/A
pBI121_NF-YB3	This paper	N/A
Software and algorithms		
Fastp (v 0.12.2)	Chan et al.[68]	https://github.com/OpenGene/fastp; RRID:SCR_016962
SAMtools (v 1.16)	Li et al.[69]	https://samtools.sourceforge.net/mpileup.shtml; RRID:SCR_005227

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
GATK (v 3.7)	McKenna et al.[70]	https://software.broadinstitute.org/gatk/; RRID:SCR_001876
Picard (v 1.124)	Broad Institute et al.	http://broadinstitute.github.io/picard/; RRID:SCR_006525
VCFtools (v 0.1.13)	Danecek et al.[71]	https://vcftools.github.io/index.html; RRID:SCR_001235
Hisat2 (v 2.1.0)	Pertea et al.[72]	http://ccb.jhu.edu/software/hisat2/index.shtml; RRID:SCR_015530
StringTie (v 2.0)	Pertea et al.[72]	https://ccb.jhu.edu/software/stringtie/; RRID:SCR_016323
Cufflinks (v 2.2.1)	Trapnell et al.[73]	http://cole-trapnell-lab.github.io/cufflinks/cuffmerge/; RRID:SCR_014597
Coding Potential Calculator2 (v 0.1)	Kang et al.[74]	http://cpc2.cbi.pku.edu.cn; RRID:SCR_002764
Pfam	Finn et al.[75]	http://pfam-legacy.xfam.org/; RRID:SCR_004726
GCTA (v 1.92.1)	Yang et al.[46]	https://yanglab.westlake.edu.cn/software/gcta/
EMMAX (beta-07Mar2010)	Kang et al.[76]	https://genome.sph.umich.edu/wiki/EMMAX
Genetic type 1 Error Calculator (v 1.0)	Li et al.[77]	http://pmglab.top/gec/
GOstats (v 2.50.0)	Falcon et al.[78]	http://gostat.wehi.edu.au; RRID:SCR_008535
XGBoost (v 1.7.5)	Chen et al.[79]	https://xgboost.readthedocs.io/en/stable/; RRID:SCR_021361
FUSION	Gusev et al.[80]	http://gusevlab.org/projects/fusion/
Cytoscape (v 3.4.0)	Shannon et al.[81]	https://cytoscape.org; RRID:SCR_003032
Hot_scan	Silva et al.[82]	https://github.com/itojal/hot_scan; RRID:SCR_002840
STRING	Szklarczyk et al.[45] von Mering et al.[78]	https://cn.string-db.org/; RRID:SCR_005223
Plink (v 1.9)	Purcell et al.[77]	https://www.cog-genomics.org/plink/; RRID:SCR_001757
Pheatmap package	Raivo Kolde	https://cran.r-project.org/web/packages/pheatmap/index.html; RRID:SCR_016418
Original code of eQTL mapping	This study	https://github.com/epi-cotton/eQTL_XGBoost

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Xueying Guan (xueyingguan@zju.edu.cn).

Materials availability

All unique/stable materials and reagents generated in this study are available from the lead contact with a completed Materials Transfer Agreement.

Data and code availability

- mRNA sequencing data are deposited under the NCBI Bio Project: PRJNA730082.
- All original code has been deposited was displayed in https://github.com/epi-cotton/eQTL_XGBoost.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

We used 279 *Gossypium hirsutum* (Upland cotton) accessions collected from the Chinese national medium-term cotton gene bank at the Institute of Cotton Research (ICR) of the Chinese Academy of Agricultural Sciences (CAAS) and National Wild Cotton Nursery, Sanya, China and. Plants of the 279 accessions were grown in a farm environment during the April-October, 2018 at Dangtu, Anhui, China. Two independent biological samples of each accession were grown in different experimental fields.

Arabidopsis thaliana (Col-0) were grown on soil or petri dishes at 23°C under long-day photoperiod (16/8 h light/dark).

METHOD DETALIS

Plant material and growth conditions

A total of 279 accessions were collected from the Institute of Cotton Research at CAAS, including 34 wild/landrace *Gossypium hirsutum* (*Gh*) accessions, such as *G. palmeri*, *G. punctatum*, *G. morrilli*, *G. yucatanense*, *G. richmondi*, *G. marie-galante*, and *G. latifolium*, as well as 245 core germplasm samples (Table S1). The core germplasm accessions were previously genotyped by our laboratory,[12] while the whole-genome sequencing of 34 wild accessions was newly conducted in this study (Table S2). Plants of the 279 accessions were grown in a farm environment during the summer of 2018 in Dangtu, Anhui, China. Two independent biological samples were taken from each accession and grown in different experimental fields. For ovule collection, 16–18 plants were grown for each accession; the 1-DPA ovules collected were then bulked for total RNA extraction and sequencing. Leaves from the 34 wild/landrace *Gh* accessions were collected for DNA extraction, sequencing, and genotyping.

Phenotypic data for nine complex traits (seed index [SI], boll weight [BW], boll number [BN], lint percentage [LP], fiber elongation [FE], fiber micronaire [FM], fiber length [FL], and fiber strength [FS]) were collected over three years (2007, 2008, and 2009) from nine environments: three farms each in Anyang (AY) in the Yellow River cotton-growing area, Nanjing (NJ) in the Yangtze River cotton-growing area, and Korla in Xinjiang (XJ), the northwestern cotton-growing area.[12] The best linear unbiased prediction (BLUP) values (Bates et al., 2015) were estimated for different phenotypes using the R package lme4. These values segregated the genetic and environmental effects that influence the phenotypes.[83]

Sample preparation

Genomic DNA of 34 wild/landrace *Gh* accessions was extracted from young leaves using the CTAB method. For RNA profiling, 1-DPA ovules were harvested from 12:00 to 1:00 p.m. The aim was to collect samples in the shortest amount of time possible so as to minimize the effects of physiological changes. Harvested ovules were frozen with liquid nitrogen for RNA extraction. Total RNA was extracted using the Trizol method, following the the manufacturer's instructions, and RNA quality was verified with an Agilent 2100 Bioanalyzer. Transcriptome libraries were constructed using the standard Illumina RNA-seq protocol (Illumina, Inc., San Diego, CA, catalog no. RS-100-0801). RNA and DNA sequences were generated as 150 bp paired-end reads from libraries with inserts of 350 bp.

SNP identification and annotation

WGS data were quality controlled using fastp (V 0.12.2) with default parameters.[68] Genome and annotation files of TM-1 v2.1[34] were indexed using BWA index with the flag (-a bwtsw).[84] Reads were mapped to the reference genome using the BWA. SAM files were sorted, indexed, and converted to BAM files using SAMtools (V 1.16).[69] Only uniquely mapped non-duplicated reads were used for SNP calling according to the best practices pipeline of GATK (v3.7).[70] Duplicated reads in the resulting alignment BAM files were marked using Picard Tools (http://picard.sourceforge.net). SNPs were called based on a minimum phred-scaled confidence threshold of 20 (-stand_call_conf >20) using the GATK tool HaplotypeCaller and then filtered using the GATK tool VariantFiltration with the following requirements: Fisher strand value (FS) < 30.0 and quality by depth value (QD) > 2.0. For GWAS and eQTL analysis, SNPs having a high missingness rate (>20%) or low minor allele frequency (MAF <0.05) were removed using VCFtools (V 0.1.13) with the parameters (–remove-indels, –maf 0.05, –max-maf 0.95, –max-missing 0.8).[71] Missing genotypes were imputed using Beagle with the parameters (window = 50000, overlap = 5000, ibd = True).[85] This process identified 1.19 million autosomal SNPs, output in a variant call format (VCF) file.

LncRNA annotation

To examine the expression of non-coding sequences, we performed population-level transcript assembly of long non-coding RNAs. RNA-seq data were quality controlled using fastp (V 0.12.2) with default parameters.[68] An average of 24.34 million reads was obtained for each library. Clean RNA-seq reads (150 bp paired-end) were aligned to the *Gh* TM-1 v2.1 reference genome using Hisat2 (V 2.1.0) with parameter (–dta).[72] Mapped reads in each library were subsequently passed to StringTie (V 2.0) for transcript assembly[72] using annotated TM-1 transcripts[34] as a reference transcriptome; the transcripts so assembled were combined into a unified set using cuffmerge with parameter (-c 3).[73] Transcripts of less than 200 nt were discarded. Using Cuffcompare (V 2.2.1), transcripts were given a class code of "u," "x," or "i," respectively representing intergenic sequences, antisense sequences of known genes, and intronic sequences. The Coding Potential Calculator2 (CPC2) (V 0.1) was used to calculate the coding potential of transcripts of each given class ("u," "x," or "i") with default parameters.[74] All transcripts with CPC scores >0 were discarded. The remaining

transcripts were subjected to pfam_scan in order to exclude those containing known protein domains (cutoff <0.001).[75] The transcripts left after that step were considered candidate lncRNAs. To reduce isoform complexity, only the longest transcript of each locus was used for further analysis.

Expression profiling

Gene expression of the newly annotated transcripts, including lncRNAs, was quantified using StringTie (V 2.0).[72] Pearson's correlation coefficient was calculated for replicates using the cor () function in R. For comparison of transcriptomes across different tissues, raw RNA-seq were analyzed through our bioinformatics pipeline as described above.[42] Heatmaps of the expression of eGenes belonging to GRN_302 and GRN_808 were generated using the *pheatmap* package (https://cran.r-project.org/web/packages/pheatmap).

Genome-wide association analysis of eQTLs

The analysis was conducted on 279 individuals who had both genotype and gene expression data available. GWAS was performed for those accessions with a total of 1.19 million SNPs (MAF >5% and missing rate <20%). Population structure was calculated using GCTA (V 1.92.1) with the parameters (–make-grm –pca).[46] Only genes having FPKM >1 in more than 5% of accessions were defined as expressed for the purpose of eQTL mapping. The expression of each gene was normalized using QQ-normal in R as is commonly done in QTL studies.[86] Ultimately, a dataset comprising 42,858 PCGs and 6,779 lncRNAs was obtained and used to conduct downstream analyses. The first three genotyping principal components (PCs) and kinship matrix were employed as covariates to control false-positive associations. Genotype files were transposed using plink (V 1.9) with the parameters (–bfile –recode12 –output-missing-genotype0 –transpose –out).[77] Kinship matrices were obtained using the emmax-kin function of EMMAX with parameters (-v -d 10).[76] eQTL mapping was carried out using EMMAX with a mixed linear model and parameters (-v -d 10 -t -o -k -c).[76] The effective number of independent SNPs was calculated using the Genetic type 1 Error Calculator (GEC), and significant SNPs were identified using the threshold of $p < 2.18 \times 10^{-6}$ suggested by GEC.[87]

To reduce eQTL redundancy, we conducted linkage disequilibrium (LD) analysis for the associated SNPs. Lead SNPs within the LD block ($R^2 > 0.1$) for each trait were merged into one eQTL using plink (V 1.90) with parameters (-r2 -l -window 99,999).[77] The eQTLs were then further classified as *cis* or *trans* based on the distance between the marker SNP and the transcription start sites or transcription end sites of associated genes (threshold: 1 Mb).[35] Hotspots were identified using hot_scan with parameters (-m 5000, -s 0,05).[82] *Cis*- and *trans*-eGenes in GRN_302 were visualized using Cytoscape (version 3.4.0; www.cytoscape.org).[81]

Construction of GRNs

Linkage disequilibrium (LD) pruning was performed to provide a list of independent GWAS variants for downstream analyses. Pruning was carried out according to three linkage disequilibrium thresholds ($R^2 > 0.1$) using an in-house Perl script. To test whether the eGenes within a GRN have more connections and interactions among themselves, we downloaded PPI pairs from the STRING database (https://stringdb-static.org/download/protein.links.v11.5/3702.protein. links.v11.5.txt.gz),[44,88] which consisted of 16,029,730 PPI pairs.

Gene function enrichment analysis

To determine whether genes within a GRN share common functional features, we performed GO term enrichment analyses using a hypergeometric test in GOstats (V 2.50.0).[78] GO terms were retrieved from the annotation files of TM-1,[34] and categories that contained at least five genes were considered significantly enriched if having a false discovery rate-corrected $p < 0.05$.

GRN effect on heritability

Two GWAS catalogs previously published were employed to assess the impacts of SNPs in GRN contribute to phenotypic variability.[12,89] The analysis considered six complex agronomic traits: seed index (SI), boll weight (BW), boll number (BN), lint percentage (LP), fiber strength (FS), and fiber length (FL). The association of phenotypic variation with the GRN was evaluated using Genomic-Relationship-matrix Restricted Maximum Likelihood (GREML), performed in GCTA.[46] Three datasets were produced: (1) SNPs from *cis/trans* eGenes (n = 216) within GRN, (2) eSNPs in GRN_302 and randomly selected eGenes not in GRN, and (3) SNPs from randomly selected genes. The test and control groups in all three datasets used the same number of SNPs. The genetic relationship matrices for those datasets were built using GCTA (v 1.92.1) with the parameter (make-grm), then estimated the amount of phenotypic variation in FL, FS, FU, and LP that could be explained by each SNP set using GCTA with the parameter (mgrm).[46] We repeated this process 100 times, each time randomly sampling the set of SNPs.

Machine learning models for trait prediction

The predictive model for phenotype based on gene expression was constructed using an ensemble of gradient boosted trees (XGBoost).[79] The eGenes belonging to pSNP-eGene pairs were informative genes. For the SI trait, the 246 available individuals were initially partitioned into training and testing datasets consisting of 90% and 10% of the data, respectively. The testing samples were never used in training.

For prediction, we applied the XGBoost[79] module of python. The XGBoost classifier is a gradient boosting method. The goal function of the XGBoost algorithm model is $obj\ (\theta) = L(\theta) + \Omega(\theta)$, where $L(\theta)$ is the training loss function and $\Omega(\theta)$ is the complexity function of the tree. $L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i), l(y_i, \hat{y}_i)$ corresponds to the training loss function for each sample, where y_i represents the true value of the *i*th sample and $\hat{y}_i$ represents the estimated value of the *i* th sample. Then, $\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F$, where K represents the number of trees, F represents all possible DT, and f denotes a specific CART tree. $\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{i=1}^{T} w_i^2$, in which w_i is the score on the *i* th leaf node and T is the number of leaf nodes in the tree. By adjusting the parameters, the objective function was continuously optimized, and optimal results were ultimately obtained.[21] The grid search algorithm was used to optimize hyper-parameters in each iteration, which included max_depth, min_child_weight, gamma, subsample, col-sample_bytree, and learning_rate.

This process was repeated 100 times using different seeds to take into account the variation in the hyperparameter optimization procession. As a description of stability of the individual phenotype predictions, we computed the mean square error (MSE) and R^2 of the predictions in the test set. Finally, the importance of each gene was calculated.

Transcriptome-wide association (TWAS)

The TWAS was carried out using the functional summary-based imputation (FUSION) approach (http://gusevlab.org/projects/fusion/).[80] This method precomputes the functional weights of gene expression, and then integrated them with summary-level GWAS results to impute the association statistics between gene expression and phenotype. The FUSION approach only considers *cis*-eGenes, typically within 500 kb or 1 Mb; in this work, 1,085 *cis*-eGenes were included in the analysis.

Transgenic cotton and *Arabidopsis*

The transgenic cotton was transformed by WIMI Biotechnology Co., Ltd. using a shoot apical meristem (SAM) cells-mediated transformation system (SAMT).[90] To construct the over-expression and suppression vectors of *GRDP1*, total RNA was isolated from the cotton TM-1 using Trizol reagent (Invitrogen) according to the manufacturer's instructions. And was then treated with DNase I (Promega). First-strand cDNA was then synthesized using M-MLV reverse transcriptase (Promega). The Open read frames (ORFs) of *GRDP1* were amplified by regular PCR with added *Xba*I and *Bam*HI, and then inserted into the basic vector pWMV062-AADA controlled by the constitutive Cauliflower mosaicvirus (CaMV) 35S promoter. *OE-GRDP1* and *AS-GRDP1* constructs were introduced into *G. hirsutum* accession TM-1 via *Agrobacterium tumefaciens* (strain LBA4404) using SAMT.[90] The T_2 homozygous transgenic lines (confirmed by target gene PCR, target protein detection, and target gene real-time qPCR) were used for further analysis. The primers used for vector construction and PCR-based screening are provided in Table S8.

To generate transgenic over-expression lines of *Arabidopsis* plants, the coding region of *NF-YB3* were cloned into *Xba* I and *Bam*H I restriction sites of *pBI121* binary vectors, under the control of the CaMV 35S. The *pBI121_NF-YB3* plasmids were transformed into *Arabidopsis thaliana* Col-0 by *A. tumefaciens* (GV3101) using the floral dip method. Primers are listed in Table S8. The integration of the transgene into different transgenic lines was confirmed by PCR.

QUANTIFICATION AND STATISTICAL ANALYSIS

R software (v 4.3.1) was used for data analysis. Statistics are described in the corresponding section of method details and figure legends. For bar plot, data are shown as mean ± SD. To assess the statistical significance of a difference between two groups, two-tailed Student's *t* tests were used: *p ≤ 0.05, **p ≤ 0.01, ***p ≤ 0.001, ****p ≤ 0.0001, ns = non-significant. For non-parametric test, the Mann-Whitney U test was used.