Genome-wide Association Study Reveals Novel Quantitative Trait Loci and Candidate Genes of Lint Percentage in Upland Cotton Based on the CottonSNP80K Array

Yu Chen
State Key Laboratory of Cotton Breeding and Cultivation in Huang-huai-hai plain, Ministry of Agriculture and Rural Affairs of China, Cotton research Center of Shandong Academy of Agricultural Sciences
Yang Gao
State Key Laboratory of Crop Genetics and Germplasm Enhancement, College of Agriculture, Nanjing Agricultural University
Pengyun Chen
State Key Laboratory of Cotton Biology, Institure of Cotton Research of Chinese Academy of Agricultural Sciences
Juan Zhou
Key Laboratory of Cotton Breeding and Cultivation in Huang-Huai-Hai Plain, Ministry of Agriculture and Rural Affairs of China, Cotton Research Center of Shandong Academy of Agricultural Sciences
Chuanyun Zhang
Key Laboratory of Cotton Breeding and Cultivation in Huang-Huai-Hai Plain, Ministry of Agriculture and Rural Affairs of China, Cotton Research Center of Shandong Academy of Agricultural Sciences
Zhangqiang Song
Key Laboratory of Cotton Breeding and Cultivation in Huang-Huai-Hai Plain, Ministry of Agriculture and Rural Affairs of China, Cotton Research Center of Shandong Academy of Agricultural Sciences
Xuehan Huo
Key Laboratory of Cotton Breeding and Cultivation in Huang-Huai-Hai Plain, Ministry of Agriculture and Rural Affairs of China, Cotton Research Center of Shandong Academy of Agricultural Sciences
Zhaohai Du
Key Laboratory of Cotton Breeding and Cultivation in Huang-Huai-Hai Plain, Ministry of Agriculture and Rural Affairs of China, Cotton Research Center of Shandong Academy of Agricultural Sciences
Juwu Gong
State Key Laboratory of Cotton Biology, Institute of Cotton Research of Chinese Academy of Agricultural Sciences
Chengjie Zhao
Key Laboratory of Cotton Breeding and Cultivation in Huang-Huai-Hai Plain, Ministry of Agriculture and Rural Affairs of China, Cotton Research Center of Shandong Academy of Agricultural Sciences
Shengli Wang
Key Laboratory of Cotton Breeding and Cultivation in Huang-Huai-Hai Plain, Ministry of Agricultural and Rural Affairs of China, Cotton Research Center of Shandong Academy of Agricultural Sciences
Jingxia Zhang
Key Laboratory of Cotton Breeding and Cultivation in Huang-Huai-Hai plain, Ministry of Agriculture and Rural Affairs of China, Cotton Research Center of Shandong Academy of Agricultural Sciences
Furong Wang
Key Laboratory of Cotton Breeding and Cultivation in Huang-Huai-Hai Plain, Ministry of Agriculture and Rural Affairs of China, Cotton Research Center of Shandong Academy of Agricultural Sciences
Jun Zhang (mhzxzhangjun@shandong.cn)
Key Laboratory of Cotton Breeding and Cultivation in Huang-huai-hai Plain, Ministry of Agriculture and Rural Affairs of China, Cotton Research Center of Shandong Academy of Agricultural Sciences https://orcid.org/0000-0001-6542-3686

Research Article

Keywords: Reveals Novel Quantitative, Genes of Lint Percentage, haplotypes, qRT-PCR analysis

Posted Date: June 30th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-648403/v1

Abstract

Cotton (*Gossypium* spp.) is an important natural textile fiber and oilseed crop widely cultivated in the world. Lint percentage (LP, %) is one of the important yield factor, thus increasing lint percentage is a core goal of cotton breeding improvement. However, the underlying genetic and molecular mechanisms that control lint percentage in upland cotton remain largely unknown. Here, we performed a Genome-wide association study (GWAS) for LP based on phenotypic tests of 254 upland cotton accessions in four environments and BLUPs using the high-density CottonSNP80K array. A total of 41,413 high-quality single-nucleotide polymorphisms (SNPs) were screened and 34 SNPs within 22 QTLs were identified as significantly associated with lint percentage trait in different environments. In total, 175 candidate genes were identified from two major genomic loci (GR1 and GR2) of upland cotton and 50 hub genes were identified through GO enrichment and WGCNA analysis. Furthermore, two candidate/causal genes, *Gh_D01G0162* and *Gh_D07G0463*, which pleiotropically increased lint percentage were identified and further verified its function through LD blocks, haplotypes and qRT-PCR analysis. Co-expression network analysis showed that the candidate/causal and hub gene, *Gh_D07G0463*, was significantly related to another candidate gene, *Gh_D01G0162*, and the simultaneous pyramid of the two genes lays the foundation for a more efficient increase in cotton production. Our study provides crucial insights into the genetic and molecular mechanisms underlying variations of yield traits and serves as an important foundation for lint percentage improvement via marker-assisted breeding.

Key Message

A total of 34 SNPs within 22 QTLs associated with lint percentage were identified by a GWAS. Two candidate genes underlying this trait were detected based on significant SNPs as well.

Introduction

Cotton (*Gossypium* spp.) is the most important renewable natural textile fiber crop and is also a major source of oilseed worldwide (Hulse-Kemp et al. 2015). The allotetraploid upland cotton (*G. hirsutum* L., AADD, 2n=4x=52) is the most widely cultivated species and accounts for approximately 95% of the world's cotton lint production (Chen et al. 2007). Although fiber quality traits, which are of upmost importance to cotton breeding programs, have dominated these studies in cotton, increasing and improving the yield of cotton remains a major objective in cotton breeding for China and major cotton production countries in the world (Constable et al. 2015; Su et al. 2016; Sun et al. 2018). Lint yield-related traits, such as boll number (BN), boll weight (BW), lint percentage (LP), seed index (SI) and lint index (LI), are typically quantitative traits and controlled by multi-genes (Si et al. 2017; Song et al. 2019; Sun et al. 2018), and among which lint percentage (LP, %) is an important factor for determining cotton lint yield (Culp and Harrell 1975). However, little is known about the genetic and molecular mechanisms underlying variations for lint percentage in upland cotton despite having some related studies. Therefore, dissecting genetic variation and identifying candidate genes significantly associated with lint percentage is essential.

Quantitative trait locus (QTL) mapping has been widely used to dissect the genetic variation for complex traits of cotton, such as fiber quality (Wang et al. 2013; Fang et al. 2017c), yield-related traits (Abdurakhmonov et al. 2007; Rong et al. 2007; Shi et al. 2015), seed traits (An et al. 2009), drought and salt tolerance (Abdelraheem et al. 2018) and disease resistance (Gutierrez et al. 2011). Over the past three decades, several QTLs for lint percentage, which were distributed on different chromosomes, have been detected through bi-parental linkage mapping in upland cotton (Zhang et al. 2005; Shen et al. 2006; Abdurakhmonov et al. 2007; Wang et al. 2011; Said et al. 2013; Wang et al. 2013; Liu et al. 2015; Shang et al. 2016). However, most of these QTLs are not directly applicable to breeding due to low marker density and poor genetic diversity resulting in very large genetic regions and often unstable across populations (Su et al. 2016; Li et al. 2018b) and the molecular mechanisms underlying most of these QTLs remains largely unknown. Dissection the allelic variations and molecular mechanism for LP will facilitate the lint percentage trait improvement by combining traditional breeding methods and maker assisted selection (MAS) simultaneously.

Genome-wide association study (GWAS) based on linkage disequilibrium (LD) can effectively associate genotypes with phenotypes in natural populations (Sun et al. 2017; Jiang et al. 2018). And genome-wide association studies have the advantages of high resolution, cost efficiency, and non-essential pedigrees for detecting important QTLs or genes associated with complex traits compared with linkage mapping. Consequently, GWASs have been widely used and applied in many crops and fruiter for various traits such as *Arabidopsis* (Atwell et al. 2010), rice (Huang et al. 2011; Yang et al. 2018; Liu et al. 2019), maize (Kump et al. 2011; Yao et al. 2020), wheat (Wang et al. 2021; Yang et al. 2021), soybean (Fang et al. 2017a; Wen et al. 2018) and pear (Zhang et al. 2021), etc. With the completion of cotton genome sequencing (Paterson et al. 2012; Li et al. 2015; Zhang et al. 2015; Hu et al. 2019; Huang et al. 2020) and the establishment of a high-throughput genotyping platform based on the high-throughput array (Hulse-Kemp et al. 2015; Cai et al. 2017), a large number of single nucleotide polymorphism (SNP) markers have been developed, which greatly promoted the application genome-wide association analyses in cotton. Recently, GWAS research have mainly focused on fiber quality (Gapare et al. 2017; Sun et al. 2017; Dong et al. 2018; Li et al. 2018a; Tan et al. 2018; Yuan et al. 2019b) and cotton yield components (Su et al. 2016; Sun et al. 2018; Song et al. 2019; Xing et al. 2019; Zhu et al. 2020), disease tolerance (Li et al. 2017), reniform nematode resistance (Li et al. 2018c), salt tolerance (Yuan et al. 2019a), drought stress (Hou et al. 2018; Li et al. 2020) and other agronomic traits in cotton (Li et al. 2018b; Yuan et al. 2018; Fu et al. 2019). However, there are few reports for revealing loci and candidate genes of the lint percentage (LP, %) by GWAS and WGCNA analysis combining strategy in cotton.

To be better understanding the allelic variations in the cotton genome at a natural population level and identifying candidate genes significantly associated with lint percentage, we performed a GWAS for LP using 254 upland cotton accessions and the high-density CottonSNP80K array based on phenotypic tests in four environments and best linear unbiased prediction (BLUPs), which was the array developed by Nanjing Agricultural University (Cai et al. 2017). We identified 34 SNPs and two causal candidate genes significantly associated with lint percentage, and the candidate genes were further verified through RNA-seq and WGCNA analysis. These results would help us to better understand the genetic mechanism of the lint percentage variations of yield traits and enhance the foundation for genetic improvement in lint yield through marker-assisted breeding in cotton.

Materials And Methods

Plant materials and field experiments

A total of 254 upland cotton accessions were selected for GWAS in this study, comprising 214 accessions that originated from Chinese Yellow River region (YRR), eight that originated from Chinese Northwestern Inland Region (NIR), 17 from Chinese Yangtze River Region (YtRR), six from the northern specific early maturation region (NSEMR) in China and nine introduced from abroad (Table S1). All 254 accessions were planted at Linqing (36° 48′ N, 115° 41′ E), Shandong province of China and Anyang (36º 05′ N, 114º 29′ E), Henan province of China for 2016 and 2017, denoted 16LQ, 16AY, 17LQ and 17AY, respectively.

In each experimental environment, all accessions were planted in a single-row plot (5.0 m long and 0.76 m between rows). The field experiments used a randomized complete block design with three replications in each environment. The field management conformed to local practices.

Phenotypic evaluation and statistical analysis

Twenty naturally fully open bolls from the central and inner of each plant were randomly collected from each plot at the cotton plant maturity stage, and ginned. Lint percentage (LP, %) was evaluated by conventional methods of cotton breeding, which is fraction of lint weight to seed cotton weight.

To reduce environmental errors, the best linear unbiased predictors (BLUPs) for the lint percentage trait were estimated using the R software lme4 package (Bates and Maechler 2007). The BLUP values and single environments were used for the GWAS. Meanwhile, the broad-sense heritability of lint percentage was also calculated using the R software lme4 package. And the correlation coefficients for the lint percentage between environments were calculated and the analysis of variance (ANOVA) was conducted using R software. Statistical analysis of phenotypic data was conducted using SPSS 22.0 software. The frequency distribution of each trait was calculated using R package (R Core Team, Vienna, Austria).

SNP Genotyping

Genomic DNA of the 254 cotton accessions was extracted from young leaf tissue using a modified CTAB method (Paterson et al. 1993). The DNA quantity and quality were measured with Nano Drop 2000 and agarose gel electrophoresis. A CottonSNP80K array containing 77,774 SNPs (Cai et al. 2017) was used to determine the genotype of the 254 accessions. All SNP genotype data were treated with raw data normalization, clustering and genotype calling by Illumina Genome StudioGenotyping Module (Illumina). And the SNPs with a missing rate ≥ 0.30 and a minor allele frequency (MAF) < 0.05 were excluded to avoid problems of spurious LD and false positive associations. Finally, a final set of 41,413 high-quality SNPs were used for GWAS analysis.

Population Structure, Kinship (K), and LD Analyses

The population genetic structure of the 254 cotton accessions was estimated using a Bayesian model-based method in STRUCTURE 2.3.4 (Evanno et al. 2005). The number of population clusters was predefined as K=1–9, using an admixture model with twenty independent runs of 100,000 burn in length and 100,000 MCMC (Markov chain Monte Carlo) replication number. The optimal K value was determined by the logarithmic probability of LnP (K) and ΔK based on the rate of change of LnP (K) between successive K. A phylogenetic tree was constructed among the 254 cotton accessions were calculated with the distance matrix method in Phylip software (version 3.69) (Felsenstein 1989). PCA using EIGENSTRAT software and testing the principal component vector according to the Tracey-Widom method (Price et al. 2006). The correlation coefficient (r^2) of alleles was calculated to measure linkage disequilibrium (LD) in each group level using Haploview software (Barrett et al. 2005). The LD decay rate was also measured as the Chromosomal distance at which the average pairwise correlation coefficient dropped to half of its maximum value. The LD decay map was drawn using the R program.

Genome-wide Association Analysis and Identification of Candidate Genes

The Genome-wide Association Analysis was performed using the R/rMVP package with the FarmCPU model (Liu et al. 2016), which has the advantages of mixed linear model and stepwise regression, the significance threshold was selected through the number of makers (P values = -log10 (1/the number of total selected SNPs)). Moreover, python scripts were used to gene extraction of significant locus through the genome GTF files.

Candidate genes were confirmed based on gene annotations in the *G. hirsutum* acc.TM-1 genome (Zhang et al. 2015) within plus and minus 500 kb regions of significant SNPs. Except the genes whose FPKM value is equal to 0, the rest of all the candidate genes were subjected to Gene Ontology (GO) enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis.

Transcriptome sequencing, WGCNA and quantitative real-time PCR (qRT-PCR) analysis

The raw RNA-seq data of two materials with different lint percentage (LMY22 with higher LP, LY343 with lower LP) tissues (ovule and fiber developmental periods) were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (accession no. PRJNA546484). For transcriptome sequencing analysis, the expression levels of the materials were analyzed using TopHat and Cufflinks software (Trapnell et al. 2012). Normalized fragments per kilobase per million mapped read (FPKM) values are used to indicate the abundance of gene expression in each material.

R/WGCNA package (version 1.69) was used to construct the weighted gene co-expression network (Langfelder and Horvath 2008), and the pick Soft Threashold set was used to calculate the weight value. The Cytoscape/CytoHubba software (V_3.7.2) was visualized the gene network. Besides, R/mufzz and R/pheatmap packages were used for gene cluster analysis and heatmap construct, respectively. Final, gene ontology using the TBtools/GO enrichment to analyze and draw produced via the R/ggplot2 packages.

Total RNA was extracted from *G. hirsutum* acc. LMY22 and LY343 tissues, including ovules at 0 and fibers at 5, 10, 15, 20 and 25 DPA, and reverse transcription was performed using PrimeScript RT reagent Kit (TaKaRa, Kusatsu, Japan). The qRT-PCR was performed on a Light Cycler 480 II (Roche, Basel,

Switzerland) using SYBG Premix Ex Taq II (TaKaRa, Kusatsu, Japan). Expression levels of two causal genes (*Gh_D01G0162* and *Gh_D07G0463*) and four genes (*Gh_A13G0389*, *Gh_D01G0200*, *Gh_D07G0449* and *Gh_D07G0457*) have an interaction network relationship with the hub gene *Gh_D07G0463* were calculated according to $2^{-\Delta\Delta CT}$ method (Chen et al. 2018). The primer sequences of six genes are listed in Table S7.

Results

Phenotypic variation for lint percentage trait of the 254 upland cotton accessions

Phenotypic values for the lint percentage (LP, %) of 254 upland cotton accessions, collected from four environments and BLUPs in 2016 and 2017 (Fig. S1), were used for variation analysis. Continuous and extensive phenotypic variations were observed for LP trait among the 254 upland cotton accessions in each environment. LP values ranged from 22.90 to 48.32%, with a mean value of 40.69% across the four environments. The coefficient of variation (CV) ranged from 8.05 to 9.16%, and exhibited an approximately normal distribution pattern in four environments based on the skewness and kurtosis values (Fig. S1; Table 1). One-way analysis of variance (ANOVA) revealed that the genotype (G), environment (E), and the interaction of genotype and environment (G×E) all significantly affected the lint percentage trait. Besides, the broad-sense heritability (h_B^2) of LP was 86.65% (Table 1), indicating that the LP trait is less influenced by environment and highly stable inherited, and mainly controlled by genetic effects.

Genetic variation based on SNPs

The genotypes of 254 accessions were examined using Illumina GenomeStudio software. After removing low-quality SNP loci (minor allele frequency < 0.05 and call rate < 85%), a final set of 41413 high-quality SNPs (41413/77774, 53.25%) were used for subsequent screening polymorphic loci, the population structure (Q), relative kinship (K), and GWAS analysis (Table 2). The 41413 SNPs were unevenly distributed on 26 chromosomes of allotetraploid cotton genome, with 21962 and 19451 SNPs in the A and D subgenomes, respectively (Fig. 1; Table 2).

The SNP density of each chromosome was ranged from 23.68 kb/SNP to 79.70 kb/SNP (Chr.16/Chr02), with an average SNP density of 49.44 kb/SNP. Chr.06 (At06) is the longest (103,170.44 kb), with an average SNP density of 56.94kb/SNP; Chr.17 (Dt03) is the shortest (46,690.66 kb), with an average SNP density of 53.61kb/SNP. In addition, the polymorphism information content (PIC) values varied from 0.199 (Chr.19) to 0.276 (Chr.24), with a mean value of 0.237. Thus, the average gene diversity of the whole genome was 0.267, varying from 0.214 (Chr.19) to 0.317 (Chr.13) (Table 2).

Population structure and LD decay estimation

To estimate the number of subpopulations in the 254 upland cotton accessions, we performed a population structure analysis using all 41413 SNPs. The LnP(K) values continuously increased with K from 1 to 9, and Evanno's delta K value reached a sharp spike when K= 2 (Fig. 2b), which suggested that the 254 upland cotton accessions could be divided into two major subgroups (Fig. 2a-c). Furthermore, phylogenetic tree and principal component analysis (PCA) showed that two subgroups for the 254 upland cotton accessions were similar to the STRUCTURE analysis despite some accessions overlapping in the two subgroups (Fig. 2d, e). Each subgroup was composed of accessions from different ecological zones included the YtRR, YRR, NIR, NSEMR and abroad, and that was unrelated to geographical distribution (Table S1). Moreover, the LD decay of our population was approximately 520kb, when the $r^2 = 0.34$ at half of its maximum value (Fig. S2). Considering the LD decay distances and comparing with previous studies of the different natural population in cotton (Dong et al. 2018; Li et al. 2018a; Song et al. 2019; Su et al. 2016; Sun et al. 2017), we finally assumed that approximately 500 kb as the region of SNP-associated candidate genes for lint percentage trait in our research.

Genome-wide association study for the lint percentage trait

Based on 41413 high-quality SNPs and phenotypic values from four environments and the BLUPs, GWAS was performed to identify the associated loci in 254 upland cotton accessions. We identified 34 SNPs that were significantly associated with lint percentage trait (Table 3). These SNP loci were distributed on 15 chromosomes: At06, At07, At08, At11, At12, At13, Dt01, Dt02, Dt04, Dt05, Dt06, Dt07, Dt08, Dt11 and Dt12 (Fig. 3a, S3; Table 3). The phenotypic variation explained by these SNPs ranged from 5.91% to 14.48%, with an average of 8.91% (Table 3). Three significant SNPs (TM47821, TM59286 and TM63365), which were distributed on Dt01, Dt06 and Dt07, respectively, were simultaneously detected in four environments and BLUPs. Two significant SNPs (TM43813 and TM47822), which were distributed on At13 and Dt01, respectively, were simultaneously detected in one environments and BLUPs. Seven significant SNPs (TM28552, TM43079, TM47579, TM47823, TM63366, TM63368 and TM69118) were consistently detected in two environments. Two SNPs (TM50517 and TM58487) were only detected in BLUPs, and the rest 20 SNPs were only detected in one environment (Table 3). For instance, the SNP locus TM63365 on Chr.D07, simultaneously detected between the four environments and BLUPs, had the highest $-\log_{10}$(P) value (12.75) and explained the largest phenotypic variation (14.48%) in 17LQ. For the SNP locus TM28552 on Chr.A08, detected among 16AY, 16LQ and BLUPs, had the highest $-\log_{10}$(P) value (11.66) and phenotypic contribution rate (13.56%) in BLUPs (Table 3). It is generally believed that SNPs are considered a reliable and important site when it is detected in more than two environments, which can be used for further analysis.

Based on previous research results and combined with ours research, the ±500kb regions of significant SNPs could be defined as QTLs, and QTLs with overlapping regions can be regarded as the same locus. In this study, a total of 22 QTLs were detected, and similar to significant SNP loci, these QTLs were scattered across different chromosomes (Table S2). Most of these QTLs contained only one significant SNP except for three QTLs such as *qLP-Dt01-1*, *qLP-Dt07-1* and *qLP-Dt08* with three significant SNPs and the other six QTLs including *qLP-At11*, *qLP-At12-2*, *qLP-At13-2*, *qLP-Dt02*, *qLP-Dt04* and *qLP-Dt05-2* with two significant SNPs. Moreover, one QTL, *qLP-Dt01-1*, were co-localized with *qLP-C15-1*, which was previously reported (Table S2), eight of these QTLs (*qLP-At07*, *qLP-At12-1*, *qLP-At12-2*, *qLP-At13-3*, *qLP-Dt01-2*, *qLP-Dt07-2*, *qLP-Dt07-3*, and *qLP-Dt11*) were adjacent to DC40182, *qLP-C12-1*, *qLP-C12-2*, NAU3017, *qLP-D1-2*, *qLP-16-2*, *qLP-16-1*, and *qLP-C21-1*, respectively. The remaining 13 QTLs did not correspond to any reported QTLs, and these QTLs may be some novel QTLs identified via GWAS analysis, which need to be further verified through different genetic populations.

Identification and analysis of favorable SNP alleles and candidate genes for lint percentage

A total of 1388 candidate genes associated with 34 significantly related SNPs were identified through Genome-wide association study (Table S4), and were divided into 22 QTNs based on the SNP location information (Table S2). There were three non-synonymous SNP variations associated with LP in the peak, and only two SNPs, TM47821 (Dt01) and TM63365 (Dt07), were simultaneously detected in four environments and BLUPs (Table S3). Consequently, we focused on the peak in Dt01 and Dt07 for the subsequent analysis. To narrow the range of candidate genes associated with LP, we conducted the local LD analysis of the peak SNPs and examined non-synonymous SNPs variation in the GWAS. Finally, two major genomic region (GR1 and GR2) including both three loci with 96 and 79 genes were identified associated with LP on chromosomes Dt01 and Dt07, respectively (Table S5).

The first major genome region of 1.14Mb (GR1, between 1,200,000 and 1,260,000bp) on chromosome Dt01 including three significant SNP loci (TM47821, TM47822, and TM47823), of which, TM47821 was identified in the four different environments and the BLUPs (Fig. 4a, S3). The LD block analysis showed that the candidate SNP locus TM47821, which was marked by the red rectangle did not fall into any LD block and was located between the Block4 and Block5 (Fig. 4b). Interestingly, a non-synonymous SNP mutation (A + 264 G + 264) in exon 1 of *GH3.5* (*Gh_D01G0162*: exon 1: c.A264G: p.P89V) significantly associated with LP ($-\log_{10}P > 4.61$), which is a homologue of the auxin-responsive GH3 family protein in *Arabidopsis thaliana* (Fig. 4a, c; Table S3). And there were two haplotypes with distinct phenotypes in 254 upland cotton accessions: haplotype AA allele had significantly higher lint percentage values than those with the GG allele ($p < 0.01$) (Fig. 4d). Further, we randomly selected two materials with different lint percentage (LMY22 with higher LP and LY343 with lower LP) to verify the causal gene *Gh_D01G0162*. The qRT-PCR results indicated that *Gh_D01G0162* was significantly highly expressed in fiber of LMY22 at 0- and 5- DPA (Fig. 4e), and the number of fiber protrusions is much more in LMY22 (Fig. S4). The expression level of *Gh_D01G0162* in LMY22 gradually decreased while in LY343 gradually increased with the fiber development (Fig. 4e). These results suggested that candidate/causal gene *Gh_D01G0162* maybe participate in the early fiber development to affect the number of fiber protrusions that determined the lint percentage variation.

Another notable hotspot region was about 0.06Mb (GR2, between 4,904,213 and 4, 967, 256 bp) on chromosome Dt07, including three significant SNP loci (TM63365, TM63366, and TM63368), which the significance threshold was above the horizontal red lines (Fig. 5a), and SNP-TM63365, the green dot in red dotted region, was significantly associated with LP ($-\log_{10} P > 4.61$) in four environments and the BLUPs (Fig. 5a, S3; Table S3). The LD block analysis showed that the candidate SNP locus TM63365, fell into the block2 (Fig. 5b), in which there are seven closely linked SNPs (TM63358, TM63360, TM63361, TM63362, TM63363, TM63364 and TM63365) (Fig. 5b; Table S6) and eleven candidate genes were located according to the data of TM-1 genome sequencing (Table S6). Interestingly, a non-synonymous SNP variation (G/A) at 6360bp in exon 10 of *Gh_D07G0463* resulted in an amino acid change from glutamate to lysine , which is a homologue of the NADPH/respiratory burst oxidase protein D (RBOHD) in *Arabidopsis* (Fig. 5c; Table S5). According to the haplotype analysis, the accessions carrying the AA allele in 254 cotton accessions exhibited a significantly increased LP trait compared to the GG allele (Fig. 5d). Moreover, the expression of *Gh_D07G0463* was higher during the early fiber development (0 and 5 DPA) than other stages and gradually decreased except for the late fiber development (25DPA) in LMY22 and LY343 (Fig. 5e). All results revealed that the candidate/causal gene *Gh_D07G0463* may plays an important role in the early fiber development that affects LP.

Co-expression network analysis and hub gene identification for significant SNP loci

We found 1388 candidate genes related to the lint percentage via GWAS analysis (Table S4), then further validated these candidate genes using transcriptome sequencing data which were from our laboratory (PRJNA546484) (Wang et al. 2020). After filtering the genes whose FKPM is always equal to 0, and 1291 genes were left for subsequent WGCNA analysis. Six co-expression modules were gained according to the cluster of the gene and they were divided into three categories: biological process, cell component and molecular function by GO enrichment analyses (Fig. S5a-c). We chose the turquoise module with a larger number of genes for subsequent WGCNA analysis (Fig. S5a), and gained 50 hub genes. Transcriptome profiling showed the relative expression level of these hub genes were higher in the initial fiber development stage than the later (Fig. 6a; Table S8), and GO enrichment analyses indicated there are only two types of biological processes and molecular functions (Fig. 6b). Co-expression network analysis of hub gene *Gh_D07G0463* was visualized with Cytoscape software. Among them, 49 genes have an interaction network relationship with the hub gene *Gh_D07G0463* (Fig. 6c; Table S9). To further verify the relative function of the hub gene *Gh_D07G0463*, we randomly selected four from 49 genes for qRT-PCR analysis, the results showed that the relative expression levels of four genes were significantly higher in LMY22 (higher LP) than LY343 (lower LP) in the early stage of fiber development which mainly includes 0 and 5 DPA, consistent with the transcriptome data (Fig. 6d; Table S8). Another causal candidate gene *Gh_D01G0162* is also located in the turquoise module, but not a hub gene. We also performed a co-expression network analysis and showed that 25 genes interact with the causal candidate gene *Gh_D01G0162*, including the hub gene *Gh_D07G0463* (Fig. S6), suggesting that the two causal candidate genes *Gh_D01G0162* and *Gh_D07G0463* maybe play important roles in the early fiber development that affects the change of lint percentage.

Analysis of favorable SNP alleles

To further identify the cumulative effect of the favorable SNPs on lint percentage. In this study, we investigated the allelic variation of two SNPs loci on chromosome Dt01 (TM47821, A/G) and Dt07 (TM63365, A/G), which were significantly associated with lint percentage. Based on the SNP alleles of the two loci (TM47821, A/G and TM63365, A/G), the 254 upland cotton accessions were classified into three haplotypes (AA-AA, AA-GG, and GG-GG). The number of haplotypes AA-AA, AA-GG, and GG-GG are 109, 81 and 64 accessions, and the mean LP are 42.87%, 40.35%, and 37.39%, respectively, showing that pyramiding the favorable alleles could increase lint percentage (Fig. 7). These results suggested that increasing the frequency of elite alleles would significantly improve the lint percentage, thus enhancing the cotton yield.

Discussion

The CottonSNP80K array was efficient and valuable for GWAS in upland cotton

GWAS based on large-scale resequencing and high-density SNP arrays provides a powerful platform for the rapid identification of genetic variants and candidate genes associated with variations of agronomic trait that can be directly applied to crop improvement (Huang et al. 2011; Hulse-Kemp et al. 2015; Cai et al. 2017). However, it is especially critical to selected materials that should include a high degree of genetic diversity for GWAS (Li et al. 2018b). In this study, a total of 254 upland cotton (*G. hirsutum*) accessions were selected and formed into a natural panel for LP loci detection, which originating from different ecological cotton-growing areas in China and abroad (Table S1). Although most of the materials were from the Yellow River region (YRR) in china, it is consistent with the phenotypes accurately characterized in Anyang and Linqing, and also to meet the breeding aims for the production needs of the Yellow River region in China. At the same time, the range of pedigrees was very rich because of the wider range of genetic diversity among materials, which better to meets the needs of GWAS. Furthermore, the broad-sense heritability of LP was 86.65% in this study, which is similar to previously reported (Sun et al. 2018; Song et al. 2019; Xing et al. 2019). And it shows that the trait of LP is relatively stable and less affected by the environment factors, thus those markers significantly associated with LP from the GWAS should be useful for cotton molecular breeding.

Moreover, GWAS also has the advantage of a high resolution. High-throughput molecular markers and wide distribution on whole genome were efficient markers for trait-genes association of the GWAS (Cai et al. 2017; Huang et al. 2017; Xing et al. 2019). In our study, the 41,413 polymorphous SNP markers were filtered out from the 77,774 SNPs, accounting for 53.25% of the SNPs at a molecular level (Table 2). And the average density of polymorphic SNPs was approximately one SNP per 49.44 kb (Table 2), and this marker density is similar to that reported by Dong et al. (2018), Li et al. (2018a, b) and Yuan et al. (2018), in which the SNP markers from the CottonSNP80K array, but significantly better than that reported by Huang et al. (2017), Sun et al. (2018) and Song et al. (2019), in which the SNP markers were from the CottonSNP63K array. And it may be mainly due to differences in the selection of the reference genome. The average value of polymorphism information content (PIC) was 0.237, less than the value of Huang et al. (2017) (0.332), Sun et al. (2018) (0.285) and close to that of Yuan et al. (2018) (0.267), Song et al. (2019) (0.250). The LD decay distance of our population was approximate 520kb (Fig. S2), yet it was higher than the distance of Li et al. (2018b) (400 kb) but lower than the result of Sun et al. (2017) (820 kb), and is similar to the distance reported by Dong et al. (2018) (500 kb).Those inconsistent results with previous reports may be mainly due to differences in population structure, sizes and SNP marker filtering criteria and so on. In this study, the 254 accessions were divided into two subpopulations based on the population structure, Neighbor-joining phylogenetic tree and Principal component analysis (Fig. 2c-e), this result was also consistent with previous reports in cotton (Li et al. 2017; Sun et al. 2017; Yuan et al. 2018; Song et al. 2019;). However, there are also inconsistencies with reports by Huang et al. (2017) (k=3), Li et al. (2018a) (k=6) and Li et al. (2018b) (k=7).

Identification of novel stable QTL and elite-allele loci for marker-assisted breeding for lint percentage in cotton

Lint percentage is an important factor for determining cotton lint yield (Culp and and Harrell 1975) and also a typically quantitative trait, which is controlled by multi-genes (Si et al. 2017; Sun et al. 2018). Over the past three decades, several QTLs for LP have been detected based on linkage and association mapping in cotton (Liu et al. 2015; Said et al. 2013; Shang et al. 2016), and some of them were also identified by GWAS (Su et al. 2016; Huang et al. 2017; Sun et al. 2018; Song et al. 2019; Xing et al. 2019; Zhu et al. 2020). In our study, a total of 34 SNPs were identified to be significantly associated with LP (Table 3), corresponding with 22 QTLs (as defined in this study) (Table S2). Among them, one overlapped with previously reported QTLs for LP, eight were adjacent, and thirteen QTLs were novel. For instance, one QTL, *qLP-Dt01-1*, were co-localized with *qLP-C15-1* which was previously reported by Wang et al. (2013) (Table S2), eight QTLs, *qLP-At07*, *qLP-At12-1*, *qLP-At12-2*, *qLP-At13-3*, *qLP-Dt01-2*, *qLP-Dt07-2*, *qLP-Dt07-3*, and *qLP-Dt11*, which were adjacent to the markers /QTLs DC40182 (Zhang et al. 2016), *qLP-C12-1* (Shi et al. 2015), *qLP-C12-2* (Shi et al. 2015), NAU3017 (Zhang et al. 2016), *qLP-D1-2* (Si et al. 2017), *qLP-16-2* (Wang et al. 2011), *qLP-16-1* (Wang et al. 2011), and *qLP-C21-1*(Shi et al. 2015), respectively (Table S2). The other novel 13 QTLs with no reports were detected in different environments in this study, so they were also stable and reliable for further research.

As we know, elite-allele loci are valuable resources for crop breeding programs, and the cumulative effect of favorable SNP is an efficient way to improve target traits in crop plants (Su et al. 2016). In this study, we found two SNPs, TM47821 and TM63365, which significantly associated with lint percentage and had a positive effect on LP. Interestingly and importantly, the haplotype of elite-allele loci of the two significantly associated SNPs are AA-AA, and the accessions carrying AA alleles at TM47821 and TM63365 had higher LP than those harboring the GG allele (Fig.4d, 5d). Therefore, these stably inherited QTLs (unanimously identified in previous report and this study) which were repeatedly identified across different genetic populations and environments and the elite-allele loci (TM47821 and TM63365) with a positive effect on lint percentage, may display a great potential of marker-assisted breeding for lint percentage in cotton, but further work will be necessary to identify and verify.

Potential candidate genes and the underlying genetic and molecular mechanisms that control the lint percentage

Some candidate genes associated with LP were identified via GWAS using different association populations in cotton, such as, *Gh_A02G1268* (Su et al. 2016), *Gh_D08G2376* (Huang et al. 2017), *Gh_D03G1064*, *Gh_D03G1065*, *Gh_D03G1067* and *Gh_D03G1069* (Fang et al. 2017b), *Gh_D03G1064* and *Gh_D12G2354* (Sun et al. 2018), *Gh_D05G0313* and *Gh_D05G1124* (Song et al. 2019), *Gh_A10G0378* (Xing et al. 2019). In this study, we detected two candidate/causal genes that maybe participating in the early fiber development affecting lint percentage in upland cotton. A total of 1388 candidate genes for the 34 significantly related SNPs were found (Table S4), and those SNPs are divided into 22 QTNs based on the SNP location information (Table S2). According to the results of the type of significantly associated SNP and the phenotype tested in multi-environment, we particularly focused on the peak in Dt01 and Dt07 as the target chromosome region. Moreover, we conducted the local LD analysis of the peak SNPs and non-synonymous SNPs identified in the GWAS. Finally, two major genomic regions (GR1 and GR2) were identified associated with LP on chromosomes Dt01 and Dt07, respectively, both of which contain three loci with total of 175 genes (Table S5).

The number of fiber protrusions on ovule is one of the important factors affecting lint percentage, and is determined at fiber initiation. It was reported that genes controlling the lint percentage highly expressed during earlier stages in fiber development, especially the initiation and elongation phases (Zhang et al. 2011; Haigler et al. 2012; Su et al. 2016). We particularly focused on two genes, *Gh_D01G0162* and *Gh_D07G0463*, as their exon regions harbored polymorphic SNPs with non-synonymous mutation. Moreover, we randomly selected two accessions with significantly different lint percentage (LMY22 with higher LP,

LY343 with lower LP) from 254 natural populations to further verify the expression of the two genes. The qRT-PCR results indicated that *Gh_D01G0162* was significantly highly expressed in fiber of LMY22 at 0- and 5- DPA (Fig. 4e), and the number of fiber protrusions is much more in LMY22 (Fig. S4). According to the annotation of the cotton genome, *Gh_D01G0162* is a homologue of the auxin-responsive GH3 family protein in *Arabidopsis*. The GH3 proteins are reported to be involved in various developmental processes and environmental responses in plants (Jeong et al. 2021), especially related to hormones, such as auxin (Mellor et al. 2016) and JA (Staswick et al. 2002). Auxin promotes fiber development during early stages of fiber initiation, including 0- and 5- DPA (Lee et al. 2007). Similar to *Gh_D01G0162*, the expression of *Gh_D07G0463*, which is a homologue of the NADPH/respiratory burst oxidase protein D (RBOHD) in *Arabidopsis*, was also significantly highly expressed in fiber of LMY22 during the early fiber development (0 and 5 DPA) (Fig. 5e). And the *AtRBOHD* was involved in root growth (Foreman et al. 2003), and partly produced ROS that play roles during root-hair elongation (Gapper and Dolan 2006) and ROS accumulation (CM-H2DCF imaging) during root-hair elongation (Foreman et al. 2003). Cotton fibers are classified as seed trichomes, which share many similarities with leaf trichomes and root-hair in *Arabidopsis* (Lee et al. 2007). The results suggested that the causal gene *Gh_D01G0162* and *Gh_D07G0463* may plays a key role in the early stage of fiber development and regulate the variation of lint percentage related to the number of fiber protrusions, and the simultaneous pyramid of the two genes lays the foundation for a more efficient increase in cotton production. And these candidate genes were also validated by using transcriptome sequencing data which were from our laboratory (PRJNA546484) (Wang et al. 2020). All genes were performed by GO enrichment and WGCNA analysis, and 50 hub genes were identified, including *Gh_D07G0463*, one of the causal genes obtained via GWAS analysis. Transcriptome data showed that all of the hub genes were more highly expressed in the initial fiber development stages at 0 and 5 DPA than the later (Table S8; Fig. 6a). The qRT-PCR analysis revealed that four of hub genes, which were randomly selected from interaction network, were highly expressed at the early stage of fiber development at 0 and 5 DPA ovule and fiber development stages (Fig. 6d), consistent with the co-expression network analysis (Fig. 6a). We also performed a co-expression network analysis on another causal candidate gene *Gh_D01G0162* which is also located in the turquoise module, but not a hub gene. The results showed that 25 genes interact with the causal candidate gene *Gh_D01G0162*, including the hub gene *Gh_D07G0463* (Fig. S5), suggesting *Gh_D07G0463* may act in a common pathway with *Gh_D01G0162* to control the number of fiber protrusions in the early stage of fiber development, thus affect the lint percentage. Of course, this needs further experiments to verify.

In summary, the detected SNPs□QTLs and candidate/causal genes will help to better understand the genetic and molecular mechanisms underlying variations of lint percentage, and have potential use in lint percentage improvement via marker-assisted breeding in cotton.

Declarations

Acknowledgements

This work was supported by the National Natural Science Foundation of China (32072116, 31671742 and 31601345), Seed-Industrialized Development Program in Shandong Province (2020LZGC002), China Agriculture Research System of MOF and MARA (CARS-15-05), the Taishan Scholar Project of Shandong Province (ts201511070), State Key Laboratory of Cotton Biology (CB2019A18) and the Innovation Project in Shandong Academy of Agricultural Sciences (CXGC2016A01).

AUTHOR CONTRIBUTIONS

JZ and FRW designed the experiments. YC, YG, JZ, CYZ, JWG, CJZ, ZHD and SLW performed field trials, phenotypic evaluation and data collection. YC, YG, XHH, ZQS, CJZ and JXZ contributed to the preparation of cotton DNA samples. YC, PYC and CJZ analyzed the data. ZQS performed the preparation of cotton RNA samples and qRT-PCR. YC drafted the manuscript. JZ and FRW revised the manuscript. All authors read and approved the final manuscript.

Conflict of interest

The authors declare that there are no conflicts of interest in the reported research.

Ethical standards

The authors note that this research is performed and reported in accordance with ethical standards of the scientific conduct.

References

Abdelraheem A, Fang DD, Zhang J (2018) Quantitative trait locus mapping of drought and salt tolerance in an introgressed recombinant inbred line population of Upland cotton under the greenhouse and field conditions. Euphytica 214

Abdurakhmonov IY, Buriev ZT, Saha S, Pepper AE, Musaev JA, Almatov A, Shermatov SE, Kushanov FN, Mavlonov GT, Reddy UK, Yu JZ, Jenkins JN, Kohel RJ, Abdukarimov A (2007) Microsatellite markers associated with lint percentage trait in cotton, *Gossypium hirsutum*. Euphytica 156:141-156

An C, Jenkins JN, Wu J, Guo Y, McCarty JC (2009) Use of fiber and fuzz mutants to detect QTL for yield components, seed, and fiber traits of upland cotton. Euphytica 172:21-34

Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Al E (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* in bred lines. Nature 465:627-631

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263-265

Bates BD, Maechler M (2007) lme4: Linear mixed-effects models using S4 classes

Cai C, Zhu G, Zhang T, Guo W (2017) High-density 80 K SNP array is a powerful tool for genotyping *G. hirsutum* accessions and genome analysis. BMC Genomics 18:654

Chen Y, Liu G, Ma H, Song Z, Zhang C, Zhang J, Zhang J, Wang F, Zhang J (2018) Identification of introgressed alleles conferring high fiber quality derived from *Gossypium barbadense* L. in secondary mapping populations of *G. hirsutum* L. Front Plant Sci 9:1023

Chen ZJ, Scheffler BE, Dennis E, Triplett BA, Zhang T, Guo W, Chen X, Stelly DM, Rabinowicz PD, Town CD, Arioli T, Brubaker C, Cantrell RG, Lacape JM, Ulloa M, Chee P, Gingle AR, Haigler CH, Percy R, Saha S, Wilkins T, Wright RJ, Van Deynze A, Zhu Y, Yu S, Abdurakhmonov I, Katageri I, Kumar PA, Mehboob Ur R, Zafar Y, Yu JZ, Kohel RJ, Wendel JF, Paterson AH (2007) Toward sequencing cotton (*Gossypium*) genomes. Plant Physiol 145:1303-1310

Constable G, Llewellyn D, Walford SA, Clement JD (2015) Cotton breeding for fiber quality improvement. In: Cruz VMV, Dierig DA (eds) Industrial Crops: Breeding for bioenergy and bioproducts. Springer New York, pp 191-232

Culp TW, and Harrell DC (1975) Influence of lint percentage, boll size, and seed size on lint yield of upland cotton with high fiber strength. Crop Sci 15:741-746

Dong C, Wang J, Yu Y, Ju L, Zhou X, Ma X, Mei G, Han Z, Si Z, Li B, Chen H, Zhang T (2018) Identifying functional genes influencing *Gossypium hirsutum* fiber quality. Front Plant Sci 9:1968

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611-2620

Fang C, Ma Y, Wu S, Liu Z, Wang Z, Yang R, Hu G, Zhou Z, Yu H, Zhang M, Pan Y, Zhou G, Ren H, Du W, Yan H, Wang Y, Han D, Shen Y, Liu S, Liu T, Zhang J, Qin H, Yuan J, Yuan X, Kong F, Liu B, Li J, Zhang Z, Wang G, Zhu B, Tian Z (2017a) Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. Genome Biol 18:161

Fang L, Wang Q, Hu Y, Jia Y, Chen J, Liu B, Zhang Z, Guan X, Chen S, Zhou B, Mei G, Sun J, Pan Z, He S, Xiao S, Shi W, Gong W, Liu J, Ma J, Cai C, Zhu X, Guo W, Du X, Zhang T (2017b) Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. Nat Genet 49:1089-1098

Fang X, Liu X, Wang X, Wang W, Liu D, Zhang J, Liu D, Teng Z, Tan Z, Liu F, Zhang F, Jiang M, Jia X, Zhong J, Yang J, Zhang Z (2017c) Fine-mapping *qFS07.1* controlling fiber strength in upland cotton (*Gossypium hirsutum* L.). Theor Appl Genet 130:795-806

Felsenstein J (1989) PHYLIP-phylogeny inference package (Version 3.2). Cladistics-the International Journal of the Willi Hennig Society 5:164-166

Foreman J, Demidchik V, Bothwell J, Mylona P, Miedema H, Torres MA, P. L, . CS, . BC, JDG. J (2003) Reactive oxygen species produced by NADPH oxidase regulate plant cell growth. Nature 422:442-446

Fu Y, Dong C, Wang J, Wang Y, Li C (2019) Genome-wide association study reveals the genetic control underlying node of the first fruiting branch and its height in upland cotton (*Gossypium hirsutum* L.). Euphytica 215

Gapare W, Conaty W, Zhu Q-H, Liu S, Stiller W, Llewellyn D, Wilson I (2017) Genome-wide association study of yield components and fibre quality traits in a cotton germplasm diversity panel. Euphytica 213

Gapper C, Dolan L (2006) Control of plant development by reactive oxygen species. Plant Physiol 141:341-345

Gutierrez OA, Robinson AF, Jenkins JN, McCarty JC, Wubben MJ, Callahan FE, Nichols RL (2011) Identification of QTL regions and SSR markers associated with resistance to reniform nematode in *Gossypium barbadense* L. accession GB713. Theor Appl Genet 122:271-280

Haigler CH, Betancur L, Stiff MR, Tuttle JR (2012) Cotton fiber: a powerful single-cell model for cell wall and cellulose research. Front Plant Sci 3:104

Hou S, Zhu G, Li Y, Li W, Fu J, Niu E, Li L, Zhang D, Guo W (2018) Genome-wide association studies reveal genetic variation and candidate genes of drought stress related traits in cotton (*Gossypium hirsutum* L.). Front Plant Sci 9:1276

Hu Y, Chen J, Fang L, Zhang Z, Ma W, Niu Y, Ju L, Deng J, Zhao T, Lian J, Baruch K, Fang D, Liu X, Ruan YL, Rahman MU, Han J, Wang K, Wang Q, Wu H, Mei G, Zang Y, Han Z, Xu C, Shen W, Yang D, Si Z, Dai F, Zou L, Huang F, Bai Y, Zhang Y, Brodt A, Ben-Hamo H, Zhu X, Zhou B, Guan X, Zhu S, Chen X, Zhang T (2019) *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. Nat Genet 51:739-748

Huang C, Nie X, Shen C, You C, Li W, Zhao W, Zhang X, Lin Z (2017) Population structure and genetic basis of the agronomic traits of upland cotton in China revealed by a genome-wide association study using high-density SNPs. Plant Biotechnol J 15:1374-1386

Huang G, Wu Z, Percy RG, Bai M, Li Y, Frelichowski JE, Hu J, Wang K, Yu JZ, Zhu Y (2020) Genome sequence of *Gossypium herbaceum* and genome updates of *Gossypium arboreum* and *Gossypium hirsutum* provide insights into cotton A-genome evolution. Nat Genet 52:516-524

Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q, Li W, Guo Y, Deng L, Zhu C, Fan D, Lu Y, Weng Q, Liu K, Zhou T, Jing Y, Si L, Dong G, Huang T, Lu T, Feng Q, Qian Q, Li J, Han B (2011) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat Genet 44:32-39

Hulse-Kemp AM, Lemm J, Plieske J, Ashrafi H, Buyyarapu R, Fang DD, Frelichowski J, Giband M, Hague S, Hinze LL, Kochan KJ, Riggs PK, Scheffler JA, Udall JA, Ulloa M, Wang SS, Zhu QH, Bag SK, Bhardwaj A, Burke JJ, Byers RL, Claverie M, Gore MA, Harker DB, Islam MS, Jenkins JN, Jones DC, Lacape JM, Llewellyn DJ, Percy RG, Pepper AE, Poland JA, Mohan Rai K, Sawant SV, Singh SK, Spriggs A, Taylor JM, Wang F, Yourstone SM, Zheng X, Lawley CT, Ganal MW, Van Deynze A, Wilson IW, Stelly DM (2015) Development of a 63K SNP array for cotton and high-density mapping of intraspecific and interspecific populations of *Gossypium* spp. G3-Genes Genom Genet 5:1187-1209

Jeong J, Park S, Im JH, Yi H (2021) Genome-wide identification of GH3 genes in *Brassica oleracea* and identification of a promoter region for anther-specific expression of a GH3 gene. BMC Genomics 22:22

Jiang H, Ma B, qian Q, Gao Z (2018) The application of genome-wide association study (GWAS) in crop agronomic traits. Journal of Agricultural Biotechnology 26:1244-1257

Kump KL, Bradbury PJ, Wisser RJ, Buckler ES, Belcher AR, Oropeza-Rosas MA, Zwonitzer JC, Kresovich S, Mcmullen MD, Ware D (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nature Genet 43:163-168

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

Lee JJ, Woodward AW, Chen ZJ (2007) Gene expression changes and early events in cotton fibre development. Ann Bot 100:1391-1401

Li B, Chen L, Sun W, Wu D, Wang M, Yu Y, Chen G, Yang W, Lin Z, Zhang X, Duan L, Yang X (2020) Phenomics-based GWAS analysis reveals the genetic architecture for drought resistance in cotton. Plant Biotechnol J 18:2533-2544

Li C, Fu Y, Sun R, Wang Y, Wang Q (2018a) Single-locus and multi-locus genome-wide association studies in the genetic dissection of fiber quality traits in upland cotton (*Gossypium hirsutum* L.). Front Plant Sci 9:1083

Li C, Wang Y, Ai N, Li Y, Song J (2018b) A genome-wide association study of early-maturation traits in upland cotton based on the CottonSNP80K array. J Integr Plant Biol 60:970-985

Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, Ma Z, Shang H, Ma X, Wu J, Liang X, Huang G, Percy RG, Liu K, Yang W, Chen W, Du X, Shi C, Yuan Y, Ye W, Liu X, Zhang X, Liu W, Wei H, Wei S, Huang G, Zhang X, Zhu S, Zhang H, Sun F, Wang X, Liang J, Wang J, He Q, Huang L, Wang J, Cui J, Song G, Wang K, Xu X, Yu JZ, Zhu Y, Yu S (2015) Genome sequence of cultivated upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. Nat Biotechnol 33:524-530

Li R, Erpelding JE, Stetina SR (2018c) Genome-wide association study of *Gossypium arboreum* resistance to reniform nematode. BMC Genet 19:52

Li T, Ma X, Li N, Zhou L, Liu Z, Han H, Gui Y, Bao Y, Chen J, Dai X (2017) Genome-wide association study discovered candidate genes of *Verticillium wilt* resistance in upland cotton (*Gossypium hirsutum* L.). Plant Biotechnol J 15:1520-1532

Liu D, Liu F, Shan X, Zhang J, Tang S, Fang X, Liu X, Wang W, Tan Z, Teng Z, Zhang Z, Liu D (2015) Construction of a high-density genetic map and lint percentage and cottonseed nutrient trait QTL identification in upland cotton (*Gossypium hirsutum* L.). Mol Genet Genomics 290:1683-1700

Liu MH, Kang H, Xu Y, Peng Y, Wang D, Gao L, Wang X, Ning Y, Wu J, Liu W, Li C, Liu B, Wang GL (2019) Genome-wide association study identifies an NLR gene that confers partial resistance to *Magnaporthe oryzae* in rice. Plant Biotechnol J 18:1376-1383

Liu X, Huang M, Fan B, Buckler ES, Zhang Z (2016) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genet 12:e1005767

Mellor N, Band LR, Pencik A, Novak O, Rashed A, Holman T, Wilson MH, Voss U, Bishopp A, King JR, Ljung K, Bennett MJ, Owen MR (2016) Dynamic regulation of auxin oxidase and conjugating enzymes AtDAO1 and GH3 modulates auxin homeostasis. Proc Natl Acad Sci U S A 113:11022-11027

Paterson AH, Brubaker CL, Wendel JF (1993) A rapid method for extraction of cotton (*Gossypium* spp.) genomic DNA suitable for RFLP or PCR analysis. Plant Molecular Biology Reporter 11:122-127

Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J, Yoo MJ, Byers R, Chen W, Doron-Faigenboim A, Duke MV, Gong L, Grimwood J, Grover C, Grupp K, Hu G, Lee TH, Li J, Lin L, Liu T, Marler BS, Page JT, Roberts AW, Romanel E, Sanders WS, Szadkowski E, Tan X, Tang H, Xu C, Wang J, Wang Z, Zhang D, Zhang L, Ashrafi H, Bedon F, Bowers JE, Brubaker CL, Chee PW, Das S, Gingle AR, Haigler CH, Harker D, Hoffmann LV, Hovav R, Jones DC, Lemke C, Mansoor S, ur Rahman M, Rainville LN, Rambani A, Reddy UK, Rong JK, Saranga Y, Scheffler BE, Scheffler JA, Stelly DM, Triplett BA, Van Deynze A, Vaslin MF, Waghmare VN, Walford SA, Wright RJ, Zaki EA, Zhang T, Dennis ES, Mayer KF, Peterson DG, Rokhsar DS, Wang X, Schmutz J (2012) Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature 492:423-427

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genet 38

Rong J, Feltus FA, Waghmare VN, Pierce GJ, Chee PW, Draye X, Saranga Y, Wright RJ, Wilkins TA, May OL, Smith CW, Gannaway JR, Wendel JF, Paterson AH (2007) Meta-analysis of polyploid cotton QTL shows unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fiber development. Genetics 176:2577-2588

Said JI, Lin Z, Zhang X, Song M, Zhang J (2013) A comprehensive meta QTL analysis for fiber quality, yield, yield related and morphological traits, drought tolerance, and disease resistance in tetraploid cotton. BMC Genomics 14:1-22

Shang L, Liang Q, Wang Y, Zhao Y, Wang K, Hua J (2016) Epistasis together with partial dominance, over-dominance and QTL by environment interactions contribute to yield heterosis in upland cotton. Theor Appl Genet 129:1429-1446

Shen X, Guo W, Lu Q, Zhu X, Yuan Y, Zhang T (2006) Genetic mapping of quantitative trait loci for fiber quality and yield trait by RIL approach in Upland cotton. Euphytica 155:371-380

Shi Y, Li W, Li A, Ge R, Zhang B, Li J, Liu G, Li J, Liu A, Shang H (2015) Constructing a high-density linkage map for *Gossypium hirsutum* × *Gossypium barbadense* and identifying QTLs for lint percentage. Journal of integrative plant biology 57:450-467

Si Z, Chen H, Zhu X, Cao Z, Zhang T (2017) Genetic dissection of lint yield and fiber quality traits of *G. hirsutum* in *G. barbadense* background. Molecular Breeding 37

Song C, Li W, Pei X, Liu Y, Ren Z, He K, Zhang F, Sun K, Zhou X, Ma X, Yang D (2019) Dissection of the genetic variation and candidate genes of lint percentage by a genome-wide association study in upland cotton. Theor Appl Genet 132:1991-2002

Staswick PE, Tiryaki I, Rowe ML (2002) Jasmonate response locus JAR1 and several related Arabidopsis genes encode enzymes of the firefly luciferase superfamily that show activity on jasmonic, salicylic, and indole-3-acetic acids in an assay for adenylation. Plant Cell 14:1405-1415

Su J, Fan S, Li L, Wei H, Wang C, Wang H, Song M, Zhang C, Gu L, Zhao S, Mao G, Wang C, Pang C, Yu S (2016) Detection of favorable QTL alleles and candidate genes for lint percentage by GWAS in chinese Upland cotton. Front Plant Sci 7:1576

Sun Z, Wang X, Liu Z, Gu Q, Zhang Y, Li Z, Ke H, Yang J, Wu J, Wu L, Zhang G, Zhang C, Ma Z (2017) Genome-wide association study discovered genetic variation and candidate genes of fibre quality traits in *Gossypium hirsutum* L. Plant Biotechnol J 15:982-996

Sun Z, Wang X, Liu Z, Gu Q, Zhang Y, Li Z, Ke H, Yang J, Wu J, Wu L, Zhang G, Zhang C, Ma Z (2018) A genome-wide association study uncovers novel genomic regions and candidate genes of yield-related traits in upland cotton. Theor Appl Genet 131:2413-2425

Tan Z, Zhang Z, Sun X, Li Q, Sun Y, Yang P, Wang W, Liu X, Chen C, Liu D, Teng Z, Guo K, Zhang J, Liu D, Zhang Z (2018) Genetic map construction and fiber quality QTL mapping using the CottonSNP80K array in Upland cotton. Front Plant Sci 9:225

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7:562-578

Wang F, Gong Y, Zhang C, Liu G, Wang L, Xu Z, Zhang J (2011) Genetic effects of introgression genomic components from Sea Island cotton (*Gossypium barbadense* L.) on fiber related traits in upland cotton (*G. hirsutum* L.). Euphytica 181:41-53

Wang F, Xu Z, Sun R, Gong Y, Liu G, Zhang J, Wang L, Zhang C, Fan S, Zhang J (2013) Genetic dissection of the introgressive genomic components from *Gossypium barbadense* L. that contribute to improved fiber quality in *Gossypium hirsutum* L. Molecular Breeding 32:547-562

Wang F, Zhang J, Chen Y, Zhang C, Gong J, Song Z, Zhou J, Wang J, Zhao C, Jiao M, Liu A, Du Z, Yuan Y, Fan S, Zhang J (2020) Identification of candidate genes for key fibre-related QTLs and derivation of favourable alleles in *Gossypium hirsutum* recombinant inbred lines with *G. barbadense* introgressions. Plant Biotechnol J 18:707-720

Wang X, Guan P, Xin M, Wang Y, Chen X, Zhao A, Liu M, Li H, Zhang M, Lu L, Zhang J, Ni Z, Yao Y, Hu Z, Peng H, Sun Q (2021) Genome-wide association study identifies QTL for thousand grain weight in winter wheat under normal- and late-sown stressed environments. Theor Appl Genet 134:143-157

Wen Z, Tan R, Zhang S, Collins PJ, Yuan J, Du W, Gu C, Ou S, Song Q, An YC, Boyse JF, Chilvers MI, Wang D (2018) Integrating GWAS and gene expression data for functional characterization of resistance to white mould in soya bean. Plant Biotechnol J 16:1825-1835

Xing H, Yuan Y, Zhang H, Wang L, Mao L, Tao J, Wang X, Feng W, Wang H, Wang Q, Wei Z, Zhang G, Liu X, Li Z, Song X-L, Sun X-Z (2019) Multi-environments and multi-models association mapping identified candidate genes of lint percentage and seed index in *Gossypium hirsutum* L. Molecular Breeding 39

Yang M, Lu K, Zhao FJ, Xie W, Ramakrishna P, Wang G, Du Q, Liang L, Sun C, Zhao H, Zhang Z, Liu Z, Tian J, Huang XY, Wang W, Dong H, Hu J, Ming L, Xing Y, Wang G, Xiao J, Salt DE, Lian X (2018) Genome-wide association studies reveal the genetic basis of ionomic variation in rice. Plant Cell 30:2720-2740

Yang X, Zhong S, Zhang Q, Ren Y, Sun C, Chen F (2021) A loss-of-function of the dirigent gene *TaDIR-B1* improves resistance to Fusarium crown rot in wheat. Plant Biotechnol J 19

Yao L, Li Y, Ma C, Tong L, Du F, Xu M (2020) Combined genome-wide association study and transcriptome analysis reveal candidate genes for resistance to Fusarium ear rot in *maize*. J Integr Plant Biol 62:1535-1551

Yuan Y, Wang X, Wang L, Xing H, Wang Q, Saeed M, Tao J, Feng W, Zhang G, Song XL, Sun XZ (2018) Genome-wide association study identifies candidate genes related to seed oil composition and protein content in *Gossypium hirsutum* L. Front Plant Sci 9:1359

Yuan Y, Xing H, Zeng W, Xu J, Mao L, Wang L, Feng W, Tao J, Wang H, Zhang H, Wang Q, Zhang G, Song X, Sun XZ (2019a) Genome-wide association and differential expression analysis of salt tolerance in *Gossypium hirsutum* L at the germination stage. BMC Plant Biol 19:394

Yuan Y, Zhang H, Wang L, Xing H, Mao L, Tao J, Wang X, Feng W, Wang Q, Wang H, Wei Z, Zhang G, Song X-L, Sun X-Z (2019b) Candidate quantitative trait loci and genes for fiber quality in *Gossypium hirsutum* L. detected using single- and multi-locus association mapping. Ind Crop Prod 134:356-369

Zhang M, Zheng X, Song S, Zeng Q, Hou L, Li D, Zhao J, Wei Y, Li X, Luo M, Xiao Y, Luo X, Zhang J, Xiang C, Pei Y (2011) Spatiotemporal manipulation of auxin biosynthesis in cotton ovule epidermal cells enhances fiber yield and quality. Nat Biotechnol 29:453-458

Zhang MY, Xue C, Hu H, Li J, Xue Y, Wang R, Fan J, Zou C, Tao S, Qin M, Bai B, Li X, Gu C, Wu S, Chen X, Yang G, Liu Y, Sun M, Fei Z, Zhang S, Wu J (2021) Genome-wide association studies provide insights into the genetic determination of fruit traits of pear. Nat Commun 12:1144

Zhang S, Feng L, Xing L, Yang B, Zhou B (2016) New QTLs for lint percentage and boll weight mined in introgression lines from two feral landraces into *Gossypium hirsutum* acc TM-1. Plant Breeding 135:90-101

Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, Zhang J, Saski CA, Scheffler BE, Stelly DM, Hulse-Kemp AM, Wan Q, Liu B, Liu C, Wang S, Pan M, Wang Y, Wang D, Ye W, Chang L, Zhang W, Song Q, Kirkbride RC, Chen X, Dennis E, Llewellyn DJ, Peterson DG, Thaxton P, Jones DC, Wang Q, Xu X, Zhang H, Wu H, Zhou L, Mei G, Chen S, Tian Y, Xiang D, Li X, Ding J, Zuo Q, Tao L, Liu Y, Li J, Lin Y, Hui Y, Cao Z, Cai C, Zhu X, Jiang Z, Zhou B, Guo W, Li R, Chen ZJ (2015) Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. Nat Biotechnol 33:531-537

Zhang ZS, Xiao YH, Luo M, Li XB, Luo XY, Hou L, Li DM, Pei Y (2005) Construction of a genetic linkage map and QTL analysis of fiber-related traits in upland cotton (*Gossypium hirsutum* L.). Euphytica 144:91-99

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

Tables

Table1. Statistical analysis for lint percentage across the four environments

E	Min	Max	Mean	SD	CV (%)	Skewness	Kurtosis	F value			
								G	E	G×E	h_B^2(%)
16AY	23.38	47.74	41.08	3.76	9.16	-0.003	1.527	9.68^{***}	348.26^{***}	1.52^{***}	86.65
16LQ	24.41	48.32	41.33	3.38	8.17	-0.002	2.677				
17AY	23.40	45.40	39.51	3.18	8.05	-0.002	2.702				
17LQ	22.90	47.97	40.84	3.46	8.48	-0.003	2.864				

16AY, 16LQ, 17AY, and 17LQ indicate the four environments: 2016Anyang, 2016Linqing, 2017Anyang, and 2017Linqing, respectively; SD, standard deviation; CV, coefficient of variance; G, genotype; E, environment; G×E, the interaction of genotype and environment; h_B^2, broad-sense heritability.

**Significant at P = 0.01 level, ***Significant at P = 0.001 level

Table2. Summary of the polymorphic SNPs mapped in 26 Chromosomes of *G. hirsutum*

Chr.		No. of SNPs	Chr. Size(Kb)	Density of SNP (kb/SNP)	PIC	Gene diversity
Chr01	At01	1866	99884.70	53.53	0.240	0.274
Chr02	At02	1047	83447.91	79.70	0.228	0.252
Chr03	At03	1531	100263.05	65.49	0.210	0.216
Chr04	At04	884	62913.77	71.17	0.228	0.256
Chr05	At05	2083	92047.02	44.19	0.238	0.276
Chr06	At06	1812	103170.44	56.94	0.221	0.229
Chr07	At07	1812	78251.02	43.18	0.261	0.289
Chr08	At08	1803	103626.34	57.47	0.240	0.276
Chr09	At09	1805	74999.93	41.55	0.236	0.266
Chr10	At10	1731	100866.60	58.27	0.238	0.273
Chr11	At11	1444	93316.19	64.62	0.256	0.292
Chr12	At12	1652	87484.87	52.96	0.264	0.310
Chr13	At13	2492	79961.12	32.09	0.267	0.317
Chr14	Dt02	2036	67284.55	33.05	0.228	0.261
Chr15	Dt01	1606	61456.01	38.27	0.212	0.231
Chr16	Dt07	2336	55312.61	23.68	0.260	0.309
Chr17	Dt03	871	46690.66	53.61	0.236	0.272
Chr18	Dt13	1295	60534.30	46.74	0.247	0.290
Chr19	Dt05	1199	61933.05	51.65	0.199	0.214
Chr20	Dt10	1316	63374.67	48.16	0.251	0.288
Chr21	Dt11	1136	66087.77	58.18	0.209	0.227
Chr22	Dt04	801	51454.13	64.24	0.210	0.228
Chr23	Dt09	2063	50995.44	24.72	0.222	0.253
Chr24	Dt08	1989	65894.14	33.13	0.276	0.276
Chr25	Dt06	1575	64294.64	40.82	0.240	0.279
Chr26	Dt12	1228	59109.84	48.14	0.242	0.277

Chr., Chromosome; PIC, polymorphism information content

Table3. Genome-wide association loci of lint percentage (LP, %) in four environments and BLUPs.

SNP	Chr.	Position	Associated loci repeated times	Allele	16AY		16LQ		17AY		17LQ		BLUP	
					$-\log_{10}(P)$	R^2(%)	$-\log_{10}(P)$	R^2(%)	$-\log_{10}(P)$	R^2(%)	$-\log_{10}(P)$	R^2(%)	$-\log_{10}(P)$	R^2(%
TM13779	At06	5754594	1	[G/A]	9.76	11.40	-	-	-	-	-	-	-	-
TM21409	At07	73306766	1	[G/A]	11.52	13.42	-	-	-	-	-	-	-	-
TM28552	At08	72244157	2	[G/C]	5.39	6.98	12.45	14.18					11.66	13.5
TM39693	At11	89740275	1	[A/G]	-	-	-	-	-	-	5.84	7.42	-	-
TM39697	At11	89759992	1	[G/A]	-	-	7.57	9.62	-	-			4.89	6.54
TM42474	At12	72840369	1	[A/G]	-	-	6.68	8.17	-	-			-	-
TM43051	At12	85479081	1	[A/G]	-	-			-	-	6.38	8.19	-	-
TM43079	At12	85977716	2	[A/G]	-	-	6.08	7.72	5.23	6.76			-	-
TM43410	At13	4974703	1	[G/A]	-	-			6.33	7.96			-	-
TM43803	At13	14514908	1	[G/A]	-	-	7.98	9.91	-	-			-	-
TM43813	At13	14735925	1	[G/A]	-	-			-	-	10.27	11.92	9.48	13.1
TM47579	At13	75750702	2	[G/A]	-	-	6.08	7.72	-	-	6.62	7.99	5.37	7.05
TM47821	Dt01	1213051	4	[A/G]	4.89	6.72	5.08	6.21	5.88	7.17	5.51	7.07	5.00	7.03
TM47822	Dt01	1243639	1	[A/C]			5.39	6.98					6.62	7.99
TM47823	Dt01	1252464	2	[C/A]			5.23	6.76	6.10	7.82			5.83	7.07
TM48401	Dt01	13960072	1	[C/A]	-	-			4.70	6.33			-	-
TM50517	Dt02	3832229		[A/C]		-	-	-	-	-	-	-	4.70	6.35
TM50525	Dt02	4070554	1	[C/A]	-	-			5.21	6.43				
TM55676	Dt04	5890658	1	[G/A]	4.70	6.33	-	-	-	-	-	-	-	-
TM55684	Dt04	5965855	1	[C/A]	4.63	5.91	-	-	-	-	-	-	-	-
TM57333	Dt05	16633733	1	[G/A]	-	-	-	-	-	-	7.95	9.89	-	-
TM58487	Dt05	53107855		[A/C]	-	-	-	-	-	-			7.99	9.91
TM58495	Dt05	53168925	1	[A/T]	-	-	-	-	5.83	7.07	-	-		
TM59286	Dt06	7093234	4	[G/A]	5.08	6.24	8.60	10.55	7.40	9.15	7.47	9.34	7.10	8.79
TM63365	Dt07	4952057	4	[G/A]	4.80	6.82	6.21	8.31	6.99	8.73	12.75	14.48	7.90	9.85
TM63366	Dt07	4955474	2	[G/A]	5.37	7.02			5.84	7.73			5.99	7.32
TM63368	Dt07	4967256	2	[T/A]					6.21	7.83	6.68	8.17	5.22	6.89
TM63804	Dt07	13495950	1	[G/A]	5.96	7.25	-	-	-	-	-	-	-	-
TM64948	Dt07	31436518	1	[G/A]	9.38	11.22	-	-	-	-	-	-	-	-
TM69118	Dt08	50404556	2	[A/C]	8.53	10.01	-	-	8.59	10.07	-	-	9.36	11.3
TM69121	Dt08	50420004	1	[A/G]	-	-	6.60	8.44						
TM69157	Dt08	51639387	1	[G/A]	-	-	-	-	-	-	6.68	8.17	-	-
TM75638	Dt11	8991496	1	[A/C]	5.62	7.32	-	-	-	-	-	-	-	-
TM79184	Dt12	45493633	1	[G/C]	8.05	9.61	-	-	-	-	-	-	-	-

Chr., Chromosome; AY, Anyang; LQ, Linqing. 16 and 17 represent 2016 and 2017 years, respectively. -, no significant SNP was identified.

Figures

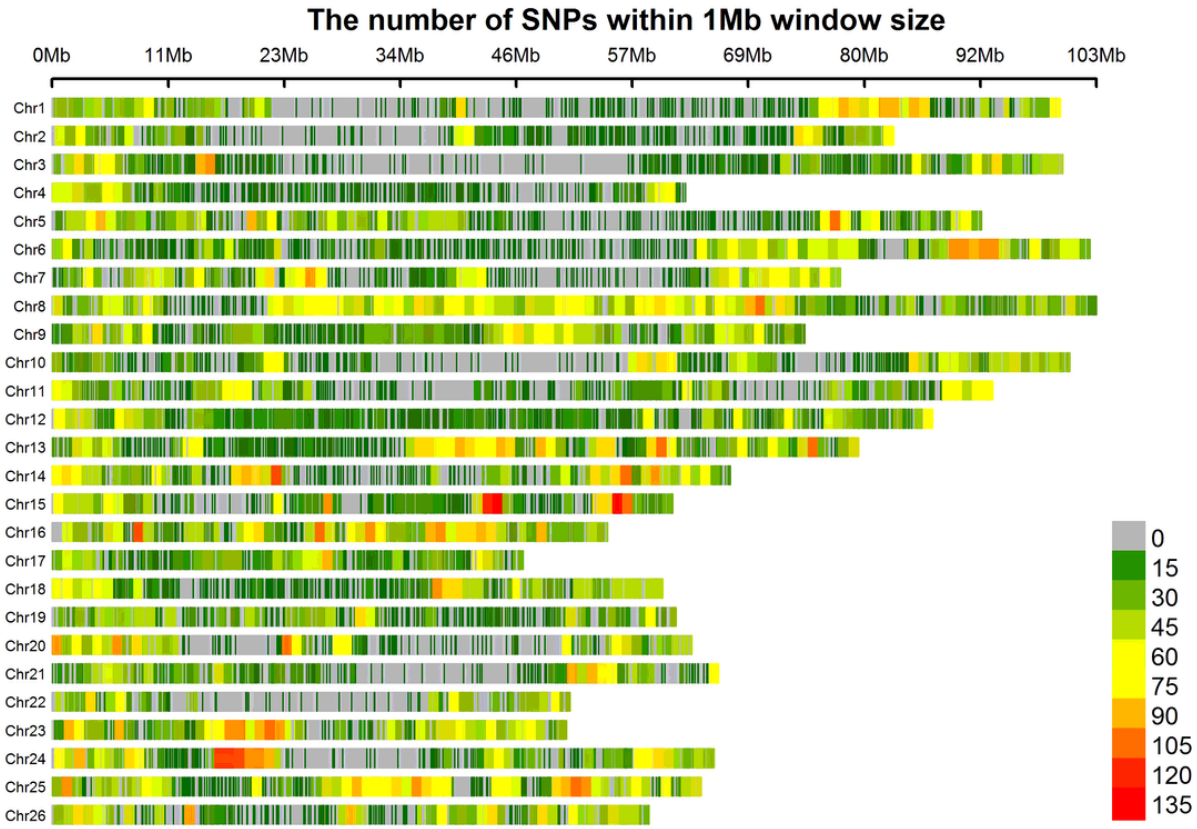

Figure 1

Distribution of 41413 polymorphic SNPs on the 26 chromosomes in upland cotton (G. hirsutun). The horizontal axis shows chromosome length (Mb); the different colors depict SNP density (the number of SNPs per window).

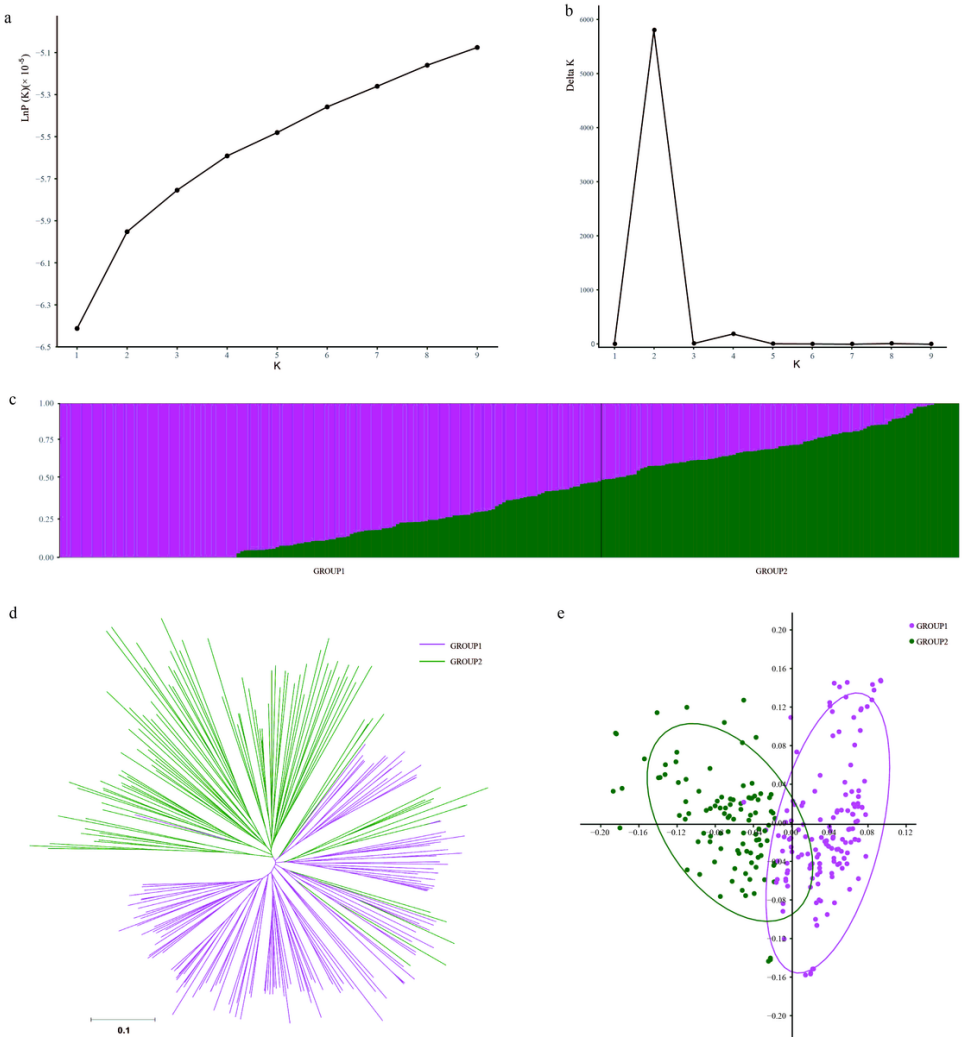

Figure 2

The results of population structure, principal component and phylogenetic analyses of 254 upland cotton (G. hirsutum) accessions. (a) Mean LnP (K) values of possible clusters (K) from 1 to 9. (b) Evanno's delta K values based on the rate of change of LnP (K) from 1 to 9. (c) Population structure based on STRUCTURE analysis at K = 2. (d) Neighbor-joining phylogenetic tree of 254 upland cotton accessions based on Nei's genetic distances. (e) Principal component analysis of 254 upland cotton accessions. Group 1 and Group 2 are represented by RGB (255, 31, 255) and RGB (13, 102, 13), respectively.

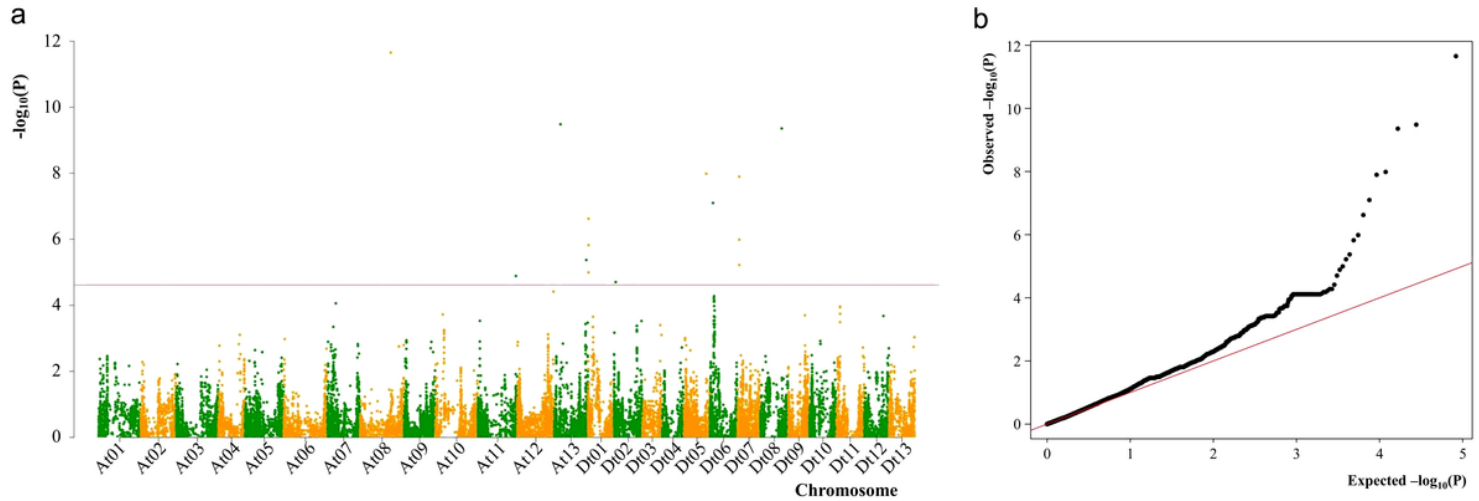

Figure 3

The results of a Genome-wide association study for LP in the BLUPs using the FarmCPU model. (a) Manhattan plot of the BLUPs across the four environments of lint percentage. The red horizontal line represents the significance threshold of −log10 (1/41413) = 4.61. (b) A quantile-quantile (QQ) plot of

the BLUPs for lint percentage.

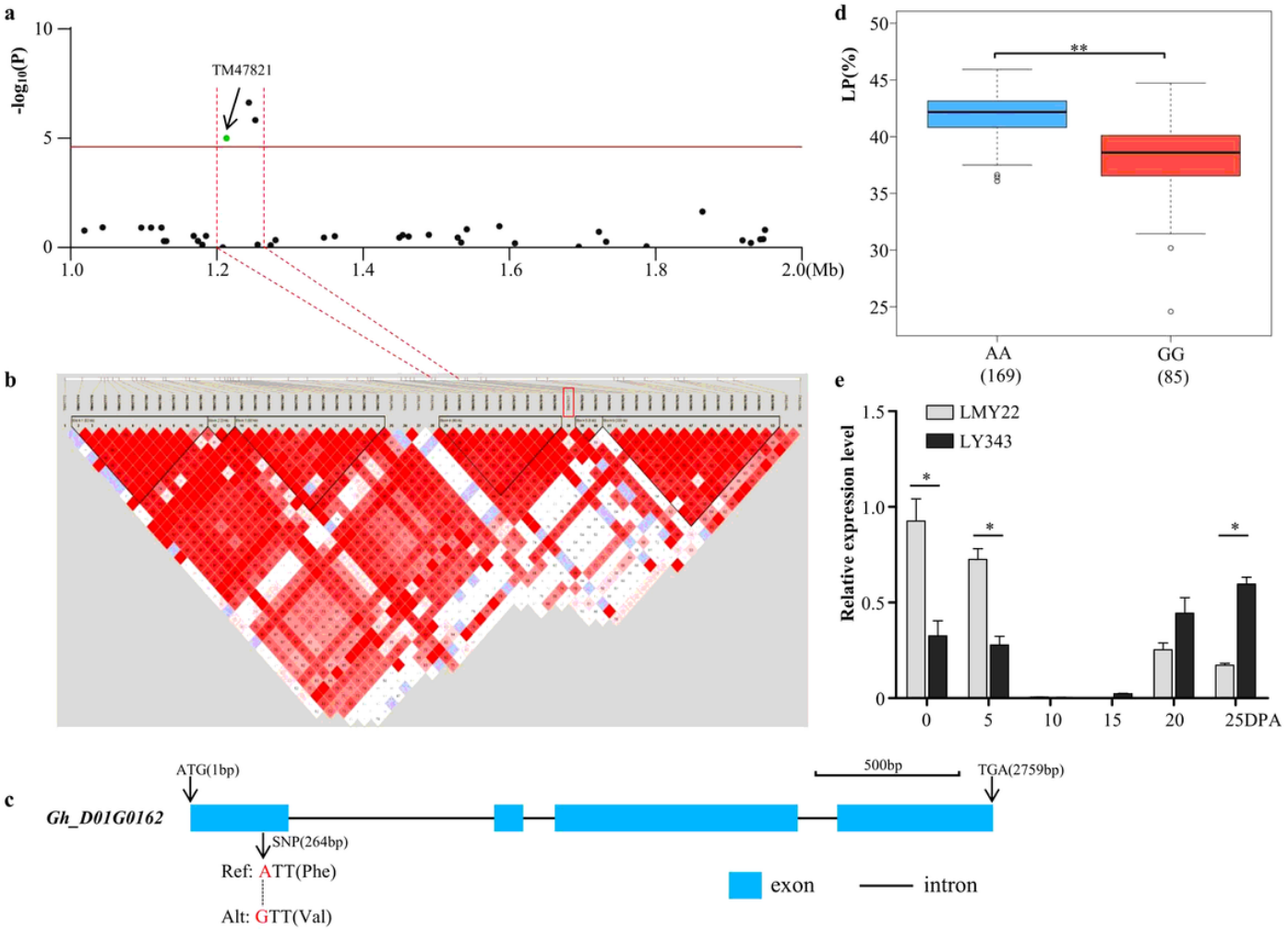

Figure 4

The GWAS results for the lint percentage and identification of the causal gene Gh_D01G0162 on chromosome Dt01 of G. hirsutum. (a) Manhattan plot for lint percentage on chromosome Dt01 which the region of the chromosome is from 1.20Mb to 1.26Mb. The horizontal red lines indicate a significance threshold of −log10 (1/41413) = 4.61. The green dot in the red dotted region represents the associated locus which was contained on the first exon of the gene Gh_D01G0162. (b) LD block analysis of SNPs in this region that includes plus and minus 500 kb regions associated with SNP (TM47821) positions. (c) Exon-intron structure of Gh_D01G0162 and DNA polymorphism in that gene. Blue rectangles and black lines indicate exons and introns, respectively. Ref and Alt stand for reference and alternate, respectively. (d) Boxplots for LP based on the allele of SNP (TM47821), box edges represent the 0.25 quantile and 0.75 quantile with the median values shown by bold lines. ** indicate significantly differential expression at 0.01 level. (e) Expression analysis of candidate gene Gh_D01G0162 associated with lint percentage by qRT-PCR. * indicate significantly differential expression at 0.05 level.

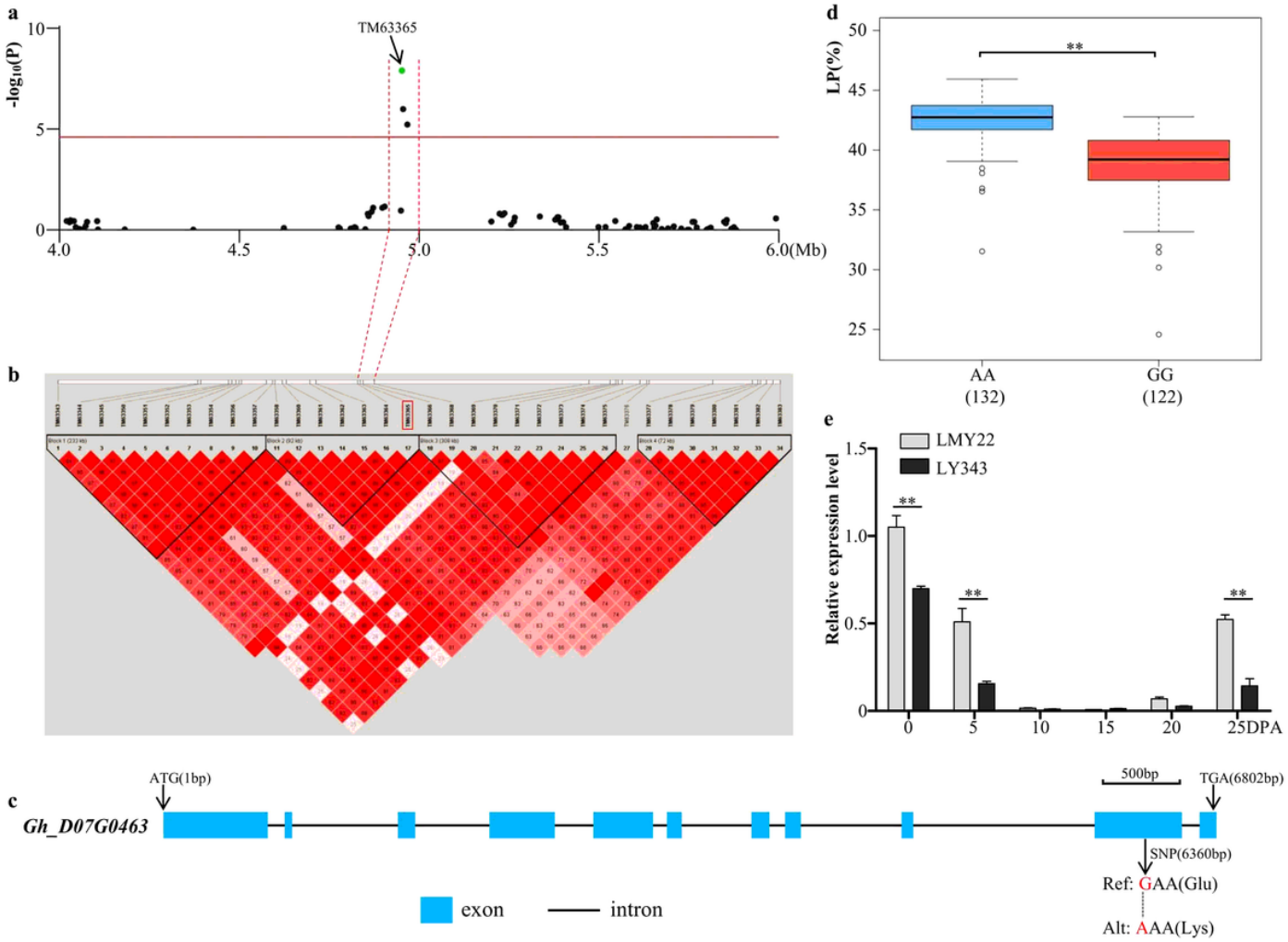

Figure 5

The GWAS results for the lint percentage and identification of the causal gene Gh_D07G0463 on chromosome Dt07 of G. hirsutum. (a) Manhattan plot for lint percentage on chromosome Dt07 which the region of the chromosome is from 4.9Mb to 5.0Mb. The horizontal red lines indicate a significance threshold of –log10 (1/41413) = 4.61. The green dot in the red dotted region represents the associated locus which was contained on the tenth exon of the gene Gh_D07G0463. (b) LD block analysis of SNPs in this region that includes plus and minus 500 kb regions associated with SNP(TM63365) positions. (c) Exon-intron structure of Gh_D07G0463 and DNA polymorphism in that gene. Blue rectangles and black lines indicate exons and introns, respectively. Ref and Alt stand for reference and alternate, respectively. (d) Boxplots for LP based on the allele of SNP (TM63365). ** indicate significantly differential expression at 0.01 level. (e) Expression analysis of causal gene Gh_D07G0463 associated with lint percentage by qRT-PCR. ** indicate significantly differential expression at 0.01 level.

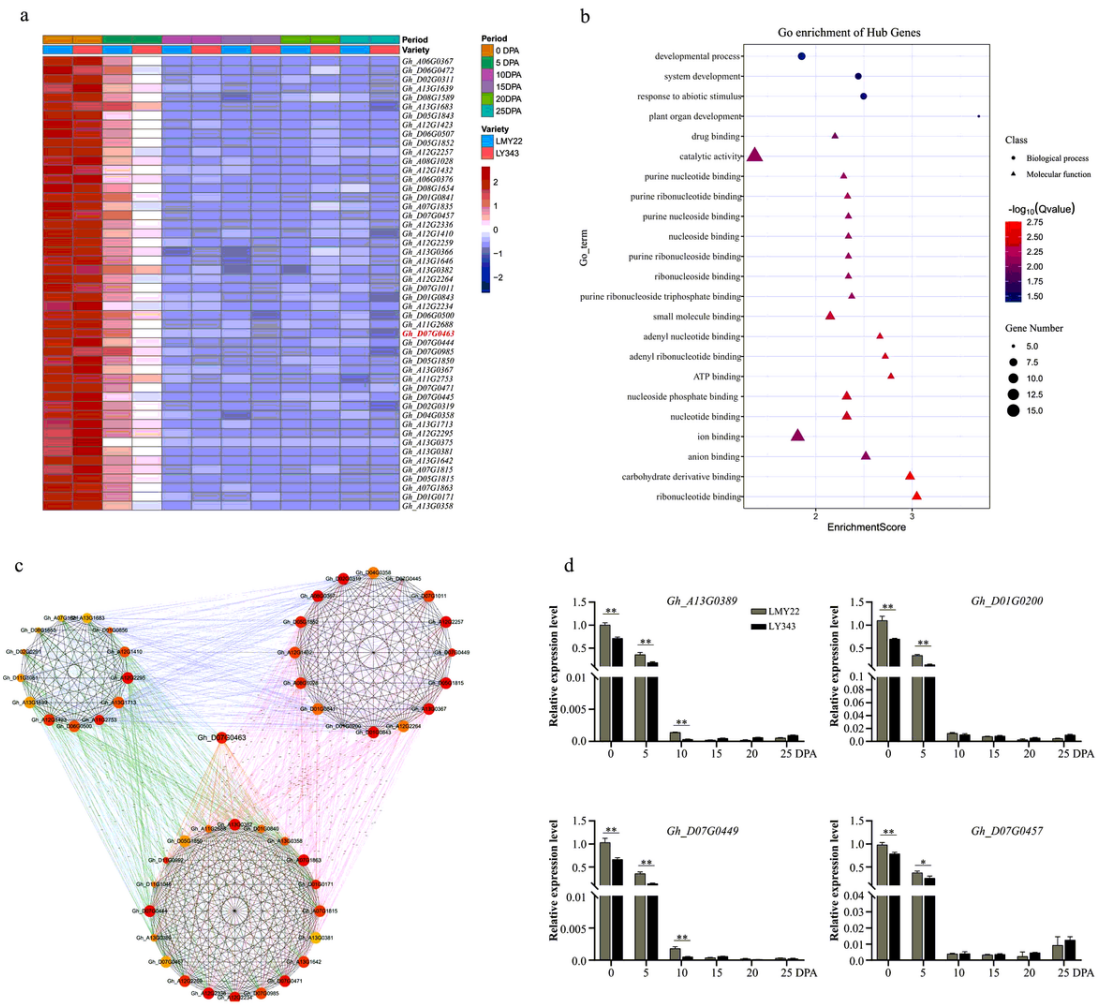

Figure 6

The results of the heat map, GO enrichment analyses, and gene co-expression networks for the hub gene. (a) Heat map of the 50 hub genes. (b) GO enrichment analyses of the 50 hub genes. (c) Co-expression network diagram of the hub gene Gh_D0G0463. (d) Expression analysis of gene Gh_A13G0389, Gh_D01G0200, Gh_D07G0449, Gh_D07G0457 by qRT-PCR. * and ** indicate significantly differential expression at 0.05 and 0.01 level, respectively.

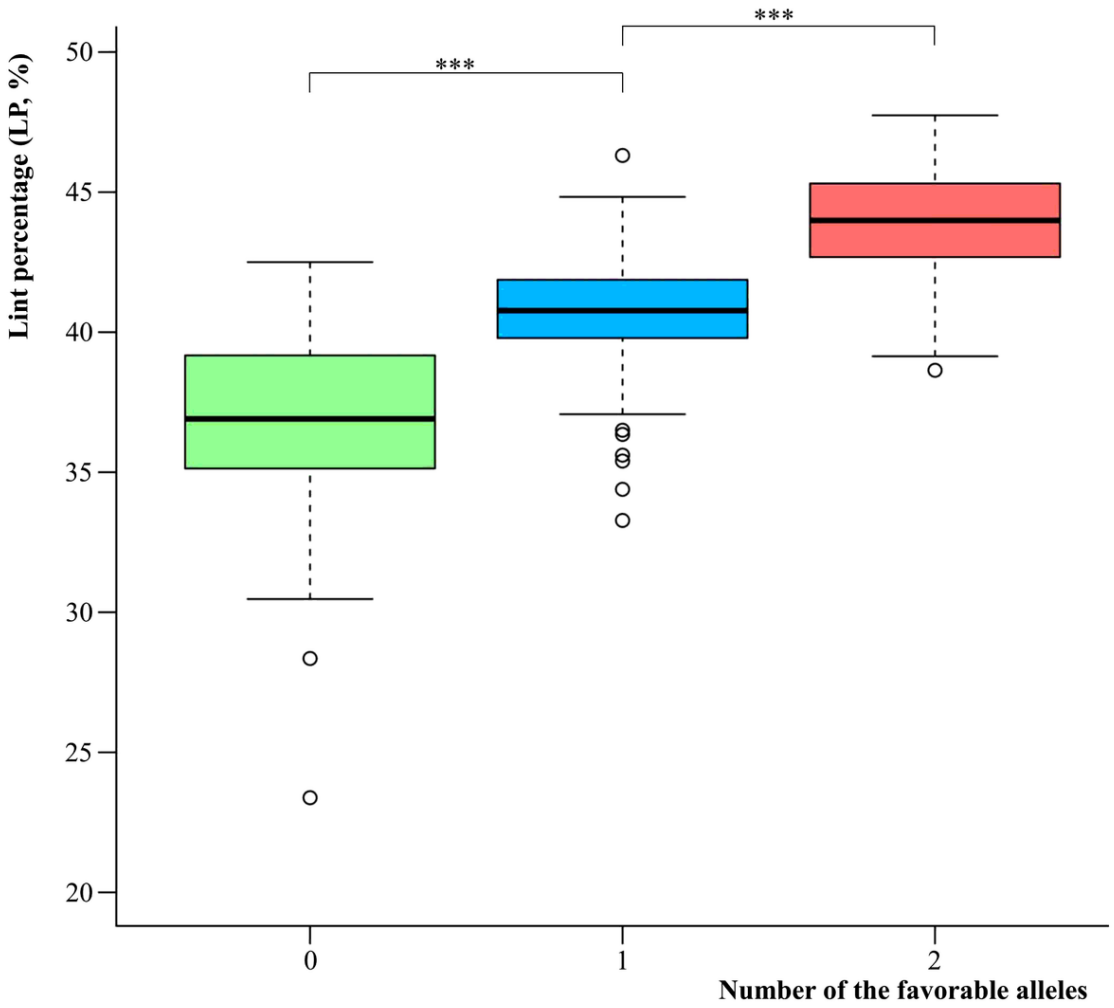

Figure 7

Box plot for lint percentage plotted as different numbers of favorable alleles. The X-axis represents the number of favorable alleles and the Y-axis represents the mean value of the lint percentage. The significance of differences was analyzed by a two-sided Wilcoxon test and *** indicate significantly differential expression at 0.001 level.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable19.xlsx
- Supplementarymaterial.docx