Theoretical and Applied Genetics (2019) 132:1991–2002
https://doi.org/10.1007/s00122-019-03333-0

ORIGINAL ARTICLE

Dissection of the genetic variation and candidate genes of lint percentage by a genome-wide association study in upland cotton

Chengxiang Song[1,2] · Wei Li[1] · Xiaoyu Pei[1] · Yangai Liu[1] · Zhongying Ren[1] · Kunlun He[1] · Fei Zhang[1] · Kuan Sun[1] · Xiaojian Zhou[1] · Xiongfeng Ma[1] · Daigang Yang[1]

Received: 19 September 2018 / Accepted: 20 March 2019 / Published online: 13 April 2019

Abstract

Key message **A genome-wide associated study identified six novel QTLs for lint percentage. Two candidate genes underlying this trait were also detected.**

Abstract Increasing lint percentage (LP) is a core goal of cotton breeding. To better understand the genetic basis of LP, a genome-wide association study (GWAS) was conducted using 276 upland cotton accessions planted in multiple environments and genotyped with a CottonSNP63K array. After filtering, 10,660 high-quality single-nucleotide polymorphisms (SNPs) were retained. Population structure, principal component and neighbor-joining phylogenetic tree analyses divided the accessions into two subpopulations. These results along with linkage disequilibrium decay indicated accessions were not highly structured and exhibited weak relatedness. GWAS uncovered 23 polymorphic SNPs and 15 QTLs significantly associated with LP, with six new QTLs identified. Two candidate genes, *Gh_D05G0313* and *Gh_D05G1124*, both contained one significant SNP, highly expressed during ovule and fiber development stages, implying that the two genes may act as the most promising regulators of LP. Furthermore, the phenotypic value of LP was found to be positively correlated with the number of favorable SNP alleles. These favorable alleles for LP identified in the study may be useful for improving lint yield.

Introduction

Cotton is a major source of natural textile fiber and a significant cash crop worldwide (Chen et al. 2007). Upland cotton (*Gossypium hirsutum* L.) occupies approximately 95% of global cotton production (Zhang et al. 2008). Lint yield, an important measure of cotton yield, depends on boll number (BN), lint percentage (LP), boll weight (BW) and other factors (Qin et al. 2015). Many studies have uncovered a significant positive correlation between LP and cotton yield, and LP is an important trait index for the breeding of high-yielding cotton (Immenkamp 2006). However, the genetic basis of LP is not fully understood. Identifying genetic variation in LP and the genes underlying this trait is therefore essential.

Most traits in plants are complex quantitative traits controlled by the small effects of multiple genes (Huang et al. 2010). Identification of genes underlying the target trait is therefore difficult. Analysis of quantitative trait loci (QTL) and genome-wide association studies (GWASs) are currently the most commonly used research methods to determine the genetic variation of a complex trait (Huang et al. 2018; Mitchell-Olds 2010). In the past few decades, QTL mapping has been widely used to dissect the genetic basis for cotton complex traits (Jamshed et al. 2016; Liu et al. 2017; Reinisch et al. 1994; Rong et al. 2004). QTL mapping of related traits in cotton has yielded fruitful results, with a total of 4892 QTLs for yield, fiber quality, stress resistance and seed traits currently identified. Among

Communicated by Alan H. Schulman.

Chengxiang Song and Wei Li contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00122-019-03333-0) contains supplementary material, which is available to authorized users.

✉ Xiongfeng Ma
maxiongfeng@caas.cn

✉ Daigang Yang
yangdaigang@caas.cn

[1] State Key Laboratory of Cotton Biology, Institute of Cotton Research of Chinese Academy of Agricultural Sciences, Anyang 455000, China

[2] College of Agriculture, Yangtze University, Jingzhou 434025, China

them, 327 LP QTLs distributed on different chromosomes have been detected (Said et al. 2015a, b). Because of the time-consuming nature of mapping-group construction and the low mapping accuracy of linkage analysis, fine mapping of QTLs for LP and map-based cloning of key genes is difficult to achieve (Cavanagh et al. 2008; Nie et al. 2016). GWAS is a more convenient and effective tool for discovering QTLs and candidate genes related to major traits in plants (Saidou et al. 2014; Zhu et al. 2008). Because of higher resolution, greater efficiency and suitability for use with large natural populations, GWAS has been widely applied to detect relationships between genetic loci and complex phenotypes in crops such as rice (Dong et al. 2018; Huang et al. 2010; Zheng et al. 2018), maize (Li et al. 2013; Tian et al. 2011; Zhao et al. 2018), rapeseed (Wang et al. 2018a; Wei et al. 2016) and soybean (Wang et al. 2018b; Wen et al. 2018; Zhou et al. 2015). However, conducting a genome-wide association analysis in cotton is relatively lagging, because of the complex genome of this species.

The completion of cotton genome sequence (Li et al. 2014; Paterson et al. 2012; Zhang et al. 2015) and the rapid evolution of gene array and high-throughput sequencing technologies (Cai et al. 2017; Hulse-Kemp et al. 2015) have led to the discovery of a large number of single-nucleotide polymorphism (SNP) markers and greatly promoted the use of genome-wide association analyses in cotton. Using a GWAS strategy with high-density SNP markers, researchers have recently detected many genetic loci associated with cotton yield components, fiber quality and disease resistance (Fang et al. 2017; Li et al. 2017; Ma et al. 2018; Wang et al. 2017). Similarly, GWAS has been used to investigate the LP trait. The 355 upland cotton accessions were genotyped by specific-locus amplified fragment sequencing (SLAF-seq), and combination with multiple environmental phenotypes in a GWAS, a gene, *Gh_A02G1268*, that may determine LP, was revealed (Su et al. 2016). The population structure and linkage disequilibrium (LD) of 503 upland cotton accessions were dissected using a CottonSNP63K array (Hulse-Kemp et al. 2015), and one candidate gene for LP, *Gh_D08G2376*, was detected (Huang et al. 2017).

In the present study, a population comprising of 276 upland cotton accessions was genotyped using a CottonSNP63K array and analyzed for structure, kinship and LD. Phenotype data were collected from seven environments and used for GWAS to determine the relationship between genetic loci and LP. The main objectives of this research were to: (1) determine the genetic structure and linkage disequilibrium level of this population, (2) identify loci associated with LP and (3) explore the candidate genes that control LP. These results should serve as useful information for the improvement breeding of LP in cotton.

Materials and methods

Plant materials and field experiments

A diverse collection of 276 upland cotton accessions was used for an association study (Table S1). These accessions were classified into five groups according to their origin: YRR (Yellow River region of China), YtRR (Yangtze River region of China), NW (Northwest China), NSEMR (Northern special early maturing region of China) and other countries of the world. All 276 accessions were grown in Anyang (Henan Province, China), Jingzhou (Hubei Province, China) and Jiujiang (Jiangxi Province, China) in 2016 and in Anyang, Jingzhou, Huanggang (Hubei Province, China) and Anqing (Anhui Province, China) in 2017 and designated as 16AY, 16JZ, 16JJ, 17AY, 17JZ, 17HG and 17AQ, respectively. In each experimental environment, all accessions were planted in a single-row plot (6.0 m long and 0.8 m between rows) with two replications (20–25 plants per replication). All field experiments were arranged in a complete randomized block design. The field management followed the local agricultural practices throughout the growing period.

Phenotypic evaluation and statistical analysis

During the open-boll bloom period, 25 naturally open bolls were randomly harvested from the middle of each plot. The lint fiber was ginned by roller gin, and LP was calculated based on fraction of lint weight to seed-cotton weight (Abdurakhmonov et al. 2007). Statistical analysis, calculation of Pearson linear correlation coefficients of LP between different environments and an analysis of variance (ANOVA) were conducted using R software (Team 2014). In addition, the broad-sense heritability (H^2) of LP was computed as $H^2 = \sigma_G^2/\left(\sigma_G^2 + \sigma_{GE}^2/n + \sigma_e^2/nr\right)$, where σ_G^2 is the genetic variance, σ_{GE}^2 is the genotype–environment interaction ($G \times E$) variance, σ_e^2 is the error variance, n represents the number of environments and r represents the number of replications. $\sigma_{G,}^2 \sigma_{GE}^2$ and σ_e^2 were estimated using the lmer function in the lme4 package of R. The best linear unbiased prediction (BLUP) of LP for each line across multiple environments was calculated using lme4 package as well (Bates et al. 2015).

SNP genotyping

Total DNA was extracted from young leaf tissues of each accession using a modified CTAB method (Zhang and Stewart 2000). A CottonSNP63K array (Hulse-Kemp et al. 2015), which contained 63,058 SNPs, was used to determine the

genotype of each mapping accession as the previous reports (Huang et al. 2017; Sun et al. 2017b). The genotyping was performed on an Illumina Infinium platform following the Illumina protocols. The SNP data were clustered and genotyped using Illumina GenomeStudio v2011.1. The SNP data were further screened according to the following criteria: SNP call rate > 0.85 and minor allele frequency > 0.05. In addition, according to the reported method (Sun et al. 2017b), the probe sequences of the SNP array were assigned to the *G. hirsutum* TM-1 reference genome (Zhang et al. 2015), and SNPs with the unique physical positions were retained for further analysis.

Population structure assessment and GWAS

The population genetic structure of the 276 accessions was analyzed using a Bayesian model-based method in STRUCTURE 2.3.4 (Evanno et al. 2005). The number of population clusters was predefined as $K = 1$–10, with five independent runs for each K. For each run, we performed 100,000 Markov chain Monte Carlo iterations after a burn-in period of 100,000 iterations. STRUCTURE HARVESTER (Earl and Vonholdt 2012), a free web-based program, was used to calculate the natural logarithm of the probability of the data (Ln P[K]) and the ad hoc statistic ΔK. The optimal K was chosen based on ΔK (Mezmouk et al. 2011). Finally, the Q matrix was obtained from CLUMPP software (Jakobsson and Rosenberg 2007) by integrating the results of the five repeated runs. In addition, principal component analysis (PCA) and calculation of a relative kinship matrix were performed using the GAPIT package (Lipka et al. 2012), with the first three principal components constituting the PCA matrix and the kinship matrix constructed according to the described method (VanRaden 2008). PowerMarker v3.25 (Liu and Muse 2005) was used to estimate the polymorphism information content (PIC) of the SNP markers, gene diversity and genetic distances among the 276 accessions. A neighbor-joining phylogenetic tree based on Nei's genetic distances (Nei 1972) was generated using MEGA 6.0 (Tamura et al. 2013). The LD parameter r^2 between pairs of SNPs was calculated with the $-r^2$ command in PLINK software (Purcell et al. 2007) based on a window size of 1000 following the reported method (Wang et al. 2017).

The association study between phenotype and genotype was performed using the GAPIT package in R under the mixed linear model (MLM) (Yu et al. 2006). The PCA matrix and kinship matrix were used as the fixed and random effects, respectively. The significance threshold for trait–marker associations was calculated according to the number of markers ($p = 1/n$, where n is the total number of SNPs used). By combining the GWAS results in different environments, an adjusted suggested genome-wide significance threshold of $p = 1.0 \times 10^{-3}$ was chosen in this study. Manhattan plots were generated using the R package qqman (Turner 2014). Heatmaps of LD on both sides of peak SNPs were produced using Haploview 4.2 (Barrett et al. 2005).

RNA-seq and quantitative real-time PCR (qRT-PCR) analysis

The raw RNA-seq data of *G. hirsutum* TM-1 tissues (root, stem, leaf, ovule and fiber developmental periods) were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (accession no. PRJNA248163). Expression analysis of the RNA-seq data was carried out using TopHat and Cufflinks software (Trapnell et al. 2012), with normalized fragments per kilobase per million mapped read (FPKM) values used as the gene expression levels.

Total RNA was extracted from *G. hirsutum* TM-1 tissues, including ovules at 0, 10, 20 and 30 days post-anthesis (DPA) and fibers at 10, 20 and 30 DPA, using TRIzol reagent (Tiangen, Beijing, China) and then reverse-transcribed using a PrimeScript RT Reagent Kit with gDNA Eraser (Takara, Tokyo, Japan). qRT-PCR amplifications were performed using SYBR Premix Ex *Taq* (2×) (Takara) on a LightCycler 480 96-well system (Roche, Mannheim, Germany). The *G. hirsutum histone3* gene was used as an internal reference. Expression levels of target genes were calculated using the comparative Ct method (Schmittgen and Livak 2008). Gene-specific primers are listed in Table S6.

Results

Analysis of LP phenotypic variation

We evaluated LP of 276 accessions in seven environments during 2016 and 2017 (Table 1). Extensive phenotypic variation was observed in each individual environment. LP values ranged from 10.49 to 49.62%, with a mean value of 37.60% across the seven environments. The coefficient of variation (CV) ranged from 7.68 to 11.20%. As indicated by skewness and kurtosis values, the LP trait exhibited an approximately normal distribution pattern in all environments (Table 1, Figure S1). In addition, the ANOVA revealed significant differences ($p < 0.001$) in the effects of genotype (G), environment (E) and the interaction of genotype and environment (G×E) (Table S2). The broad-sense heritability (H^2) of LP was 90.7% (Table S2), and a correlation analysis across different environments uncovered significant positive correlations among LP phenotypes in different environments (Figure S1). These results demonstrated that the LP trait is highly stable and mainly controlled by genetics.

Table 1 Phenotypic data statistics of lint percentage observed in seven environments

Environment	Min	Max	Mean	SD	CV (%)	Skewness	Kurtosis
16AY	20.30	49.62	36.85	3.35	9.09	− 0.625	2.446
16JZ	12.89	46.51	36.24	4.06	11.20	− 1.442	4.496
16JJ	14.52	43.96	35.97	3.79	10.55	− 1.213	3.782
17AY	10.49	46.20	38.86	3.71	9.54	− 2.053	11.821
17JZ	22.33	46.51	37.90	3.93	10.37	− 0.708	0.969
17HG	15.39	44.29	37.86	3.31	8.76	− 1.614	7.255
17AQ	28.83	45.26	39.53	3.04	7.68	− 1.085	1.480
BLUP	22.37	43.28	37.51	2.92	7.78	− 1.200	2.760

16AY, 16JZ and 16JJ represent the environment of Anyang, Jingzhou and Jiujiang in 2016; 17AY, 17JZ, 17HG and 17AQ represent the environment of Anyang, Jingzhou, Huanggang and Anqing in 2017; BLUP represents the best linear unbiased prediction across seven environments of lint percentage, SD and CV represent standard deviation and coefficient of variation, respectively

Analysis of genetic diversity based on SNPs

From the 63,058 SNPs, used to genotype the 276 tested accessions, a total of 10,660 high-quality SNPs meeting the filtering criteria were used for the subsequent analysis (Fig. 1, Table 2). These SNPs were unevenly distributed across the 26 chromosomes, with more SNPs found on the Dt subgenome (6480) than on the At subgenome (4180). The SNP density of chromosomes ranged from 86.43 kb/SNP (Dt07) to 731.71 kb/SNP (At06), with an average marker density of 237.32 kb/SNP. In addition, the polymorphism information content (PIC) values varied from 0.200 (Dt06) to 0.294 (At13) among the 26 chromosomes, with a mean value of 0.250. The mean gene diversity value of all chromosomes was 0.31 and ranged from 0.24 (Dt06) to 0.37 (At01, At05 and At13) (Table 2).

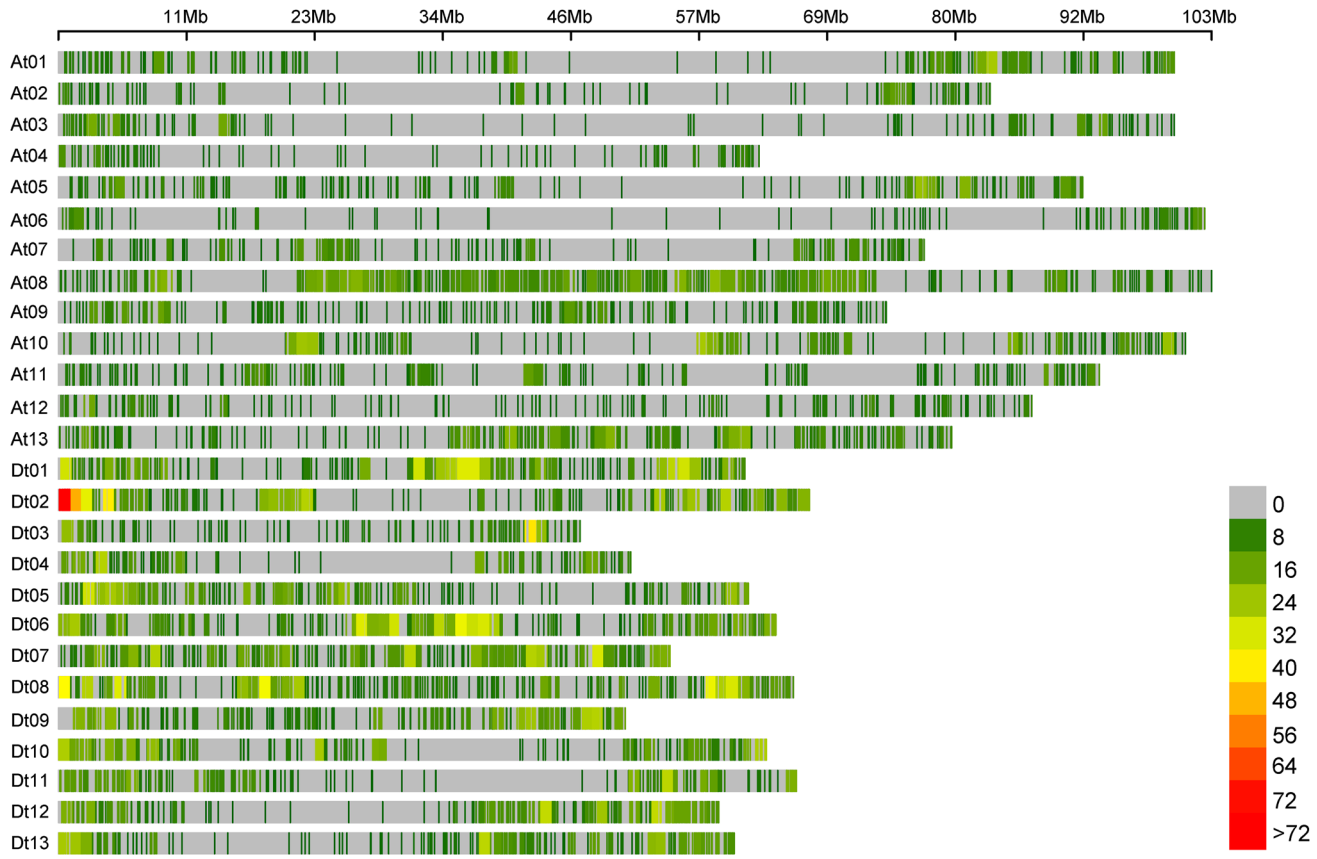

Fig. 1 Distribution of 10,660 polymorphic SNPs on the 26 chromosomes of an upland cotton association population. The horizontal axis indicates chromosome lengths, and the color legend depicts SNP density (the number of SNPs within a 1-Mb window)

Table 2 The summary of SNPs, PIC and gene diversity in 26 chromosomes of upland cotton

Chr	Chr length (kb)	SNPs	SNP density (kb/SNP)	PIC	Gene diversity
At01	99,884.700	351	284.57	0.293	0.37
At02	83,447.906	166	502.70	0.259	0.32
At03	100,263.045	225	445.61	0.280	0.35
At04	62,913.772	132	476.62	0.272	0.34
At05	92,047.023	324	284.10	0.292	0.37
At06	103,170.444	141	731.71	0.239	0.29
At07	78,251.018	275	284.55	0.250	0.31
At08	103,626.341	846	122.49	0.240	0.29
At09	74,999.931	303	247.52	0.271	0.34
At10	100,866.604	405	249.05	0.222	0.26
At11	93,316.192	294	317.40	0.213	0.26
At12	87,484.866	245	357.08	0.261	0.32
At13	79,961.121	473	169.05	0.294	0.37
Dt01	61,456.009	638	96.33	0.242	0.30
Dt02	67,284.553	694	96.95	0.289	0.36
Dt03	46,690.656	235	198.68	0.209	0.25
Dt04	51,454.130	259	198.66	0.266	0.33
Dt05	61,933.047	495	125.12	0.260	0.32
Dt06	64,294.643	697	92.24	0.200	0.24
Dt07	55,312.611	640	86.43	0.246	0.30
Dt08	65,894.135	677	97.33	0.273	0.34
Dt09	50,995.436	458	111.34	0.232	0.28
Dt10	63,374.666	462	137.17	0.264	0.33
Dt11	66,087.774	379	174.37	0.235	0.29
Dt12	59,109.837	436	135.57	0.252	0.31
Dt13	60,534.298	410	147.64	0.234	0.29

Chr Chromosome

Population structure and kinship analyses and LD decay estimation

STRUCTURE analysis indicated that values of Ln P(K) increased continuously as K was increased from 1 to 10, and there was no obvious inflexion point (Fig. 2a). However, the ΔK reached its maximum value when $K=2$ (Fig. 2b), indicating that the population could be separated into two subgroups (Fig. 2c). PCA gave a result similar to the STRUCTURE analysis, and some accessions were admixed between the two groups (Fig. 2d). The association population was divided into two clades in a neighbor-joining phylogenetic tree based on Nei's genetic distances (Fig. 2e). This classification was also supported by a kinship plot (Figure S2).

Most of the kinship coefficients (88.71%) were less than 0.2, with 58.74% equal to 0. Only 2.37% of kinship values were larger than 0.5 (Figure S3). These results indicated that weak relatedness was present in the accessions. Moreover, the LD decay, which corresponded to the distance at which r^2 was half of its maximum value, was approximately 530 kb (Fig. 3).

These results indicated that the accessions were not highly structured and exhibited weak relatedness and moderate LD decay. The association population was thus suitable for association mapping.

GWAS of the LP trait

A total of 23 SNP loci randomly distributed on 13 chromosomes were identified as significantly associated with the LP (Fig. 4a, Figure S4 and Table S3). The quantile–quantile (Q–Q) plot indicated that the MLM model can be used to identify association signal (Fig. 4b). Among these loci, seven were located on chromosomes Dt05, four on Dt10 and two on Dt13. The remaining 10 loci were positioned on chromosomes At01, At03, At05, At07, At10, Dt01, Dt02, Dt04, Dt09 and Dt11 (Figure S4 and Table S3). The phenotypic variation explained by these SNPs ranged from 4.20 to 10.23%, with an average of 5.68% (Table S3). Eleven significant SNPs were consistently detected in at least two environments. Four SNPs (i56741Gb, i61131Gt, i08888Gh and i00252Gh) were simultaneously detected in five environments and were distributed on chromosomes At03 and Dt05. Moreover, ten of these SNPs were also identified in BLUP. For example, the SNP locus i56741Gb on chromosome At03 had the highest $-\log_{10}(P)$ value (5.10) and explained the largest amount of phenotypic variation (10.23%) in 17JZ, and the $-\log_{10}(P)$ value and phenotypic variation explained in BLUP were 4.03 and 6.02%, respectively. For SNP loci on chromosome Dt05, i00252Gb recorded the highest $-\log_{10}(P)$ value (5.06) and phenotypic contribution rate (8.05%) and also possessed the highest value in BLUP (Table S3). Thus, these SNPs, which were detected in more than two environments and BLUP at the same time, were used for further analysis.

According to previous studies (Su et al. 2018; Sun et al. 2017b), the 200-kb upstream and downstream regions of significant SNPs could be defined as QTLs and considering QTLs with overlapping regions to be the same locus. Following the definition of QTL, 15 QTLs were detected in total (Table S4). Similar to significant SNP loci, these QTLs were scattered across different chromosomes. Most of these QTLs contained only one significant SNP, and the exceptions were *qLP-Dt05-1* (five significant SNPs), *qLP-Dt05-2* (two significant SNPs), *qLP-Dt10-2* (three significant SNPs) and *qLP-Dt13* (two significant SNPs). Moreover, nine QTLs were co-localized with 11 previously reported QTLs (Table S4). Six of these co-localized QTLs shared overlapping regions with known QTLs (*qLp-A-1*, *qLP-Chr10-1*, *qLP-Chr14-1*, *qLP-Chr21-2*, TMB0206 and MGHES46), and the remaining QTLs were adjacent to *qGhLP-c5*, JESPR220, NAU3269, *qLP-19* or *qLP-D10_16*.

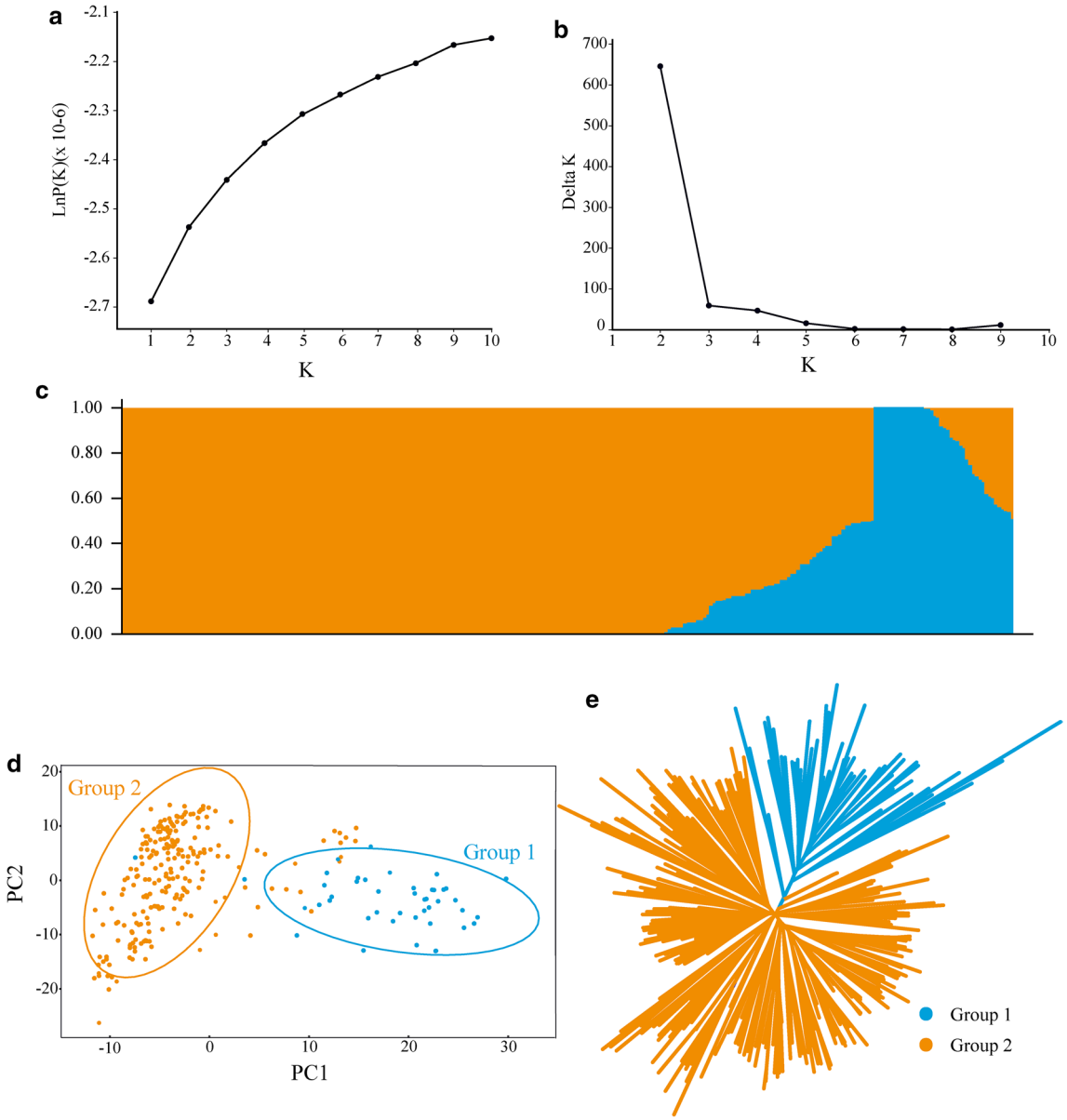

Fig. 2 The results of population structure, principal component and phylogenetic analyses of 276 upland cotton accessions. **a** Plot of mean Ln P(*K*) versus *K* for *K* = 1 to 10. **b** Plot of Δ*K* versus *K* for *K* = 1 to 10. **c** Population structure based on a STRUCTURE analysis at *K* = 2. The *y*-axis quantifies cluster membership, and the *x*-axis represents the different accessions. **d** Principal component plot of the test population. **e** Neighbor-joining phylogenetic tree based on Nei's genetic distances. Group 1 and Group 2 are represented by blue and orange, respectively (color figure online)

Candidate genes underlying associated loci

In total, 434 candidate genes were identified in the QTL regions (Table S5). Analysis of the TM-1 RNA-seq data revealed that 263 of these genes were specifically highly expressed in different organs, including roots, stems, leaves, ovules (−3, −1, 0, 1, 3, 5, 10, 20, 25 and 35 DPA), and fibers (5, 10, 20 and 25 DPA) (Figure S5). Some of these specifically expressed genes, such as *GhUPL7*, *GhTUB5* and *GhCK1*, have been previously determined to be involved in cotton fiber development (Table S5). Moreover, to narrow the range of candidate genes associated with LP, we conducted the local LD analysis of the peak SNPs and non-synonymous SNPs identified in the GWAS. Finally, we identified two genomic loci associated with LP.

The most significant SNP (i00252Gh) on Dt05 was selected the promising variant site, as i00252Gh was identified in five environments and exhibited the lowest *p* value (Fig. 5a and Figure S4). The candidate region was estimated to be 9.41–9.81 Mb (Fig. 5a, b). An LD block

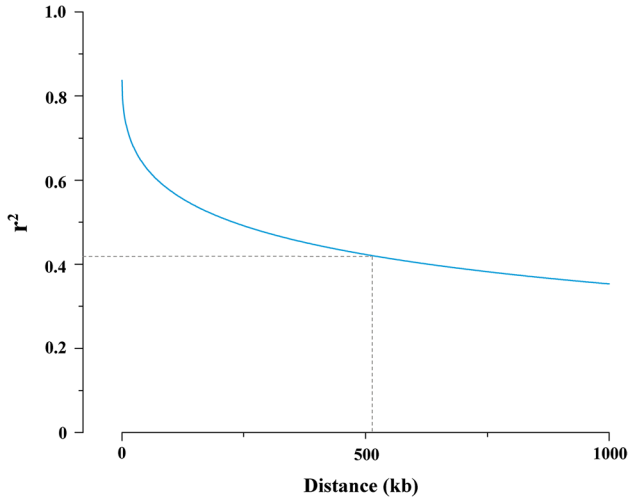

Fig. 3 Genome-wide average LD decay estimates of the association population. The black dashed line indicates the position where r^2 is at half of its maximum value

analysis indicated that the candidate SNP locus i00252Gh did not fall into any LD block (Fig. 5b). Interestingly, the peak SNP (i00252Gh) was located in the 10th exon region of *Gh_D05G1124*, a gene of unknown function homologous to a gene encoding a protein phosphatase 2C family protein in Arabidopsis. In addition, i00252Gh was a non-synonymous SNP (A/G) (Table S3) responsible for an aspartic acid to glycine amino acid substitution (Fig. 5c). The GG haplotype was found to have positive phenotypic effects on LP, as accessions carrying the GG allele had significantly higher LP values than those with the AA allele ($p < 0.001$) (Fig. 5d). Moreover, RNA-seq data for *Gh_D05G1124* from 17 different upland cotton tissues revealed that *Gh_D05G1124* was highly expressed during ovule and fiber development (Figure S5). qRT-PCR analysis indicated that the expression of this gene gradually increased during ovule and fiber development, with peak levels observed at 30 DPA in ovules as well as in fibers (Fig. 5e). These results suggest that *Gh_D05G1124* participates in ovule and fiber development and is a causative gene for LP in upland cotton.

There is a another notable hotspot region at the interval of 2.61–2.76 Mb on chromosome Dt05, where a novel non-synonymous SNP (i08888Gh) resulted in an amino acid change from asparagine to serine in the coding sequence (CDS) of gene *Gh_D05G0313* (Figure S6a–c). Accessions with the GG allele had significantly higher LP values than those harboring the AA allele ($p < 0.001$; Figure S6d). Furthermore, qRT-PCR analysis indicated that *Gh_D05G0313* was relatively high expressed in 20 and 30 DPA ovules and 30 DPA fibers (Figure S6e). The ortholog of *Gh_D05G0313* in Arabidopsis, *AtLUT2*, plays an important role in photosynthesis, an important process in plant organs, including developing cotton ovules and fibers.

Analysis of favorable SNP alleles

To identify the cumulative effect of favorable SNPs on LP, we selected the two significant SNPs i00252Gh and i08888Gh, which were found to have a positive effect on LP phenotypic performance. The 276 accessions were classified into three types (AA–AA, AG–AG/AG–AA/AG–GG and GG–GG) based on the SNP alleles of the two loci. A total of 134 accessions were genotyped as AA–AA, 126 accessions were heterozygous, and only 16 possessed the GG–GG genotype. The average LP values of the three genotype groups were 36.58%, 38.24% and 39.48%, respectively, showing that the more favorable alleles were pyramided in varieties, with the larger average LP values increasing (Fig. 6). These results suggest that LP is positively correlated with the number of favorable alleles and these favorable alleles displayed pyramiding effects on LP.

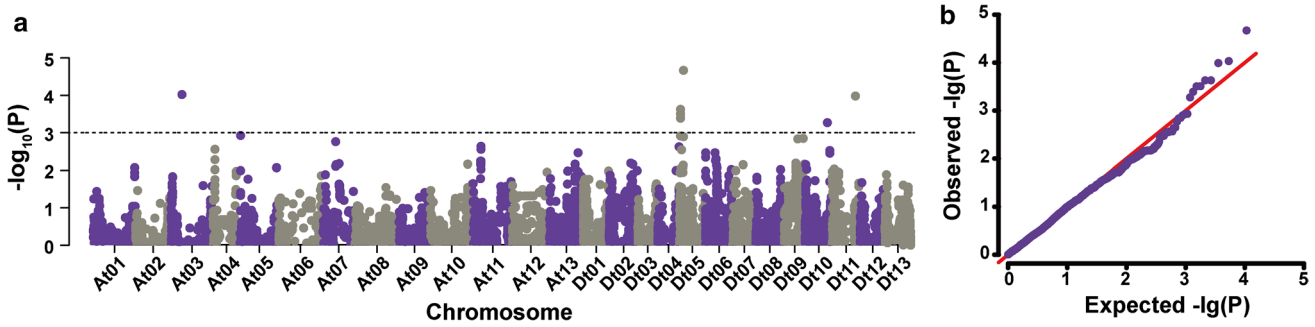

Fig. 4 Genome-wide association study (GWAS) for lint percentage (LP). **a** Manhattan plot of the best linear unbiased prediction (BLUP) across seven environments. The black dashed line represents the significance threshold. **b** A quantile–quantile (Q–Q) plot of the BLUP for LP

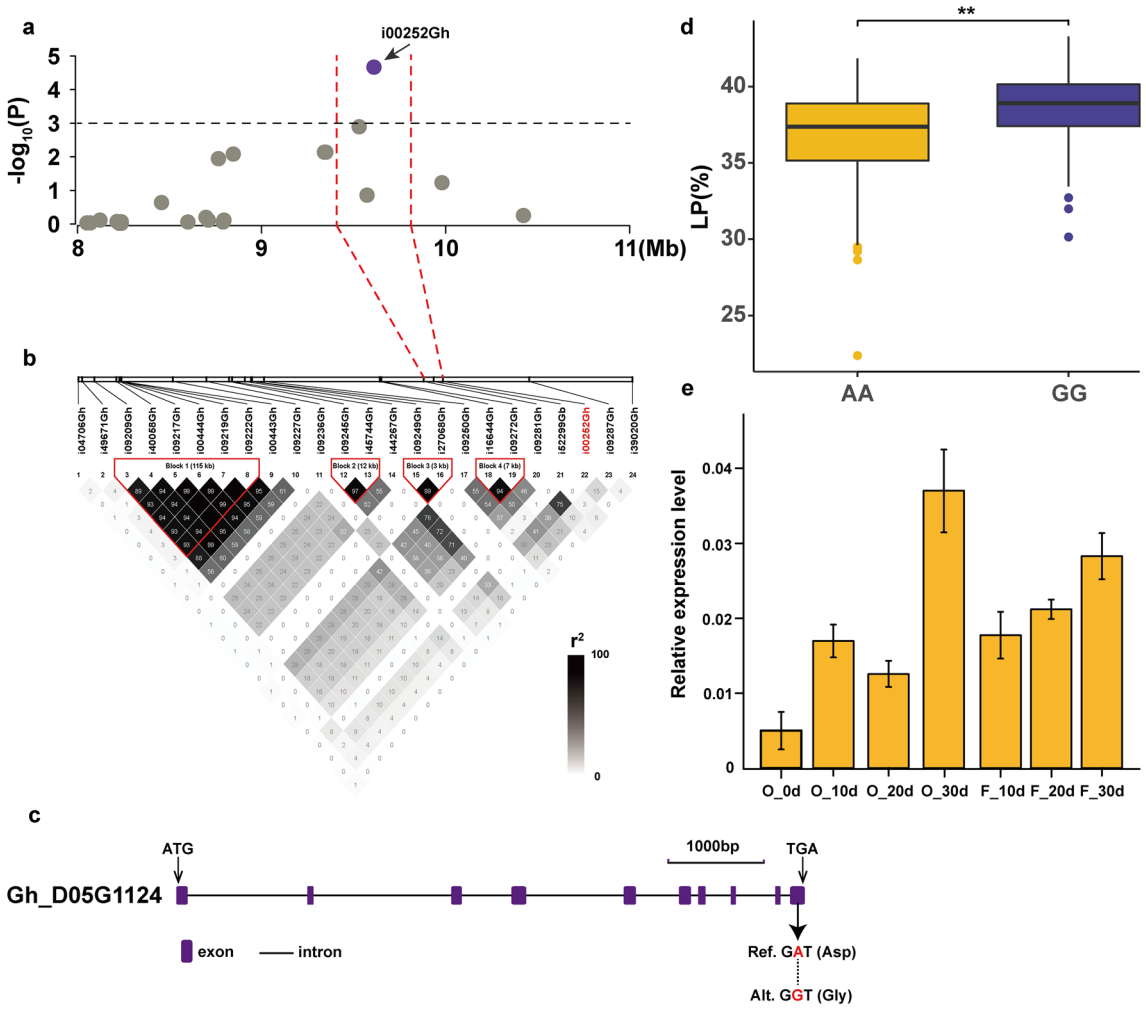

Fig. 5 GWAS results for lint percentage and identification of the causal gene for the peak on chromosome Dt05. **a** Local Manhattan plot for the candidate region on Dt05. The purple dot represents the peak SNP i00252Gh. Red dotted lines indicate the candidate region. **b** LD block analysis of SNPs in this region. The degree of linkage is represented by the coefficient of r^2. **c** Gene structure of *Gh_D05G1124* and a non-synonymous SNP within it. Purple rectangles and black lines indicate exons and introns, respectively. Ref and Alt stand for reference and alternate, respectively. **d** Box plots for LP based on the allele of SNP i00252Gh. The significance of differences was analyzed by a two-sided Wilcoxon test. **e** Tissue-specific expression profiles of *Gh_D05G1124*. Expression of *Gh_D05G1124* was investigated in ovule (0, 10, 20 and 30 DPA) and fiber (10, 20 and 30 DPA) developmental stages by qRT-PCR. *GhHis3* was used as an internal control. Error bars indicate the standard deviation of three technical replicates (color figure online)

Discussion

For GWAS, the wider range of genetic diversity among materials is especially critical (Li et al. 2018). In the present study, the 276 accessions originated from the five main cotton regions in China and other foreign countries, with more abundant genetic variation among materials. Moreover, the LP trait for the association panel was evaluated in seven environments during 2016 and 2017. The LP trait showed abundant phenotypic variation in each single environment, and multienvironment survey phenotypic data strategy would be enhanced the reliability of association mapping. In addition, the broad-sense heritability of LP was 90.7%, which is similar to previously reported values (Huang et al. 2017; Wang et al. 2015). This showed that the stability of LP was high, and the marker associated with LP can be stably detected and those markers should be useful for cotton breeding to adapt to different environments (Su et al. 2016).

Moreover, the high marker density is beneficial for the discovery of more elite loci and promising genes (Wang et al. 2018a). In our study, the average genome-wide density of polymorphic SNPs was one SNP per 273.32 kb. This marker density is similar to levels reported by Sun et al. (2017b) and Huang et al. (2017). The LD decay distance in the current study, 530 kb, was higher than the distance reported in cotton by Li et al. (2018) (400 kb) but lower

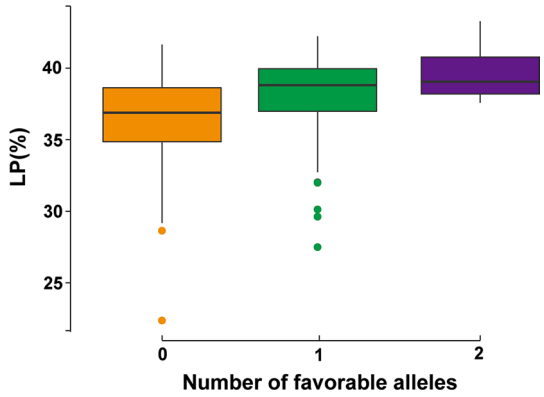

Fig. 6 Box plot of lint percentage versus the number of favorable alleles. The *x*-axis indicates LP, and the *y*-axis indicates the number of favorable SNP alleles

than the result of Sun et al. (2017b) (820 kb). The average PIC value of the markers was 0.250, less than the value of 0.332 obtained by Huang et al. (2017) and close to 0.285 reported by Sun et al. (2017b). These conflicting results may be mainly due to differences in population sizes and SNP-marker filtering criteria, as a similar phenomenon has been observed in soybean (Wen et al. 2018). Furthermore, population structure and relative kinship among individuals are the two important factors in controlling false positives (Lu et al. 2015). In this study, the 276 accessions were divided into two subpopulations by comprehensive analysis, which were unrelated to geographic origin. The lack of any geographic correlation may be due to extensive exchange and penetration of germplasm from different geographic origins during the process of cotton breeding. Overall, the association population was not highly structured and the LD level was moderate.

LP is a typical complex quantitative trait, which is controlled by multigene (Sun et al. 2018). In cotton, more than 327 QTLs for LP have been detected based on linkage and association mapping (Said et al. 2015a, b). Some of them were also identified by GWAS, especially the stably inherited QTLs (Huang et al. 2017; Su et al. 2016; Sun et al. 2018). In the present study, a total of 23 SNPs were found to be significantly associated with LP, half were identified in more than two environments and BLUP. The high proportion of significant SNPs identified in multiple environments reflects their high heritability. In addition, a total of 15 QTLs (as defined in this study) were detected. Among them, six were novel, while six overlapped with confidence regions of previously reported QTLs or GWAS signals for LP and three were near these regions. For instance, *qLP-At03*, *qLP-Dt02* and *qLP-Dt04* identified in this study overlapped with the confidence intervals of *qLP-A-1* (Wang et al. 2013), *qLP-Chr14-1* (Li et al. 2016) and TMB0206 (Abdurakhmonov et al. 2007). These results confirm the reliability of the LP-related associations determined in the present study. In addition, these stably inherited QTLs, which were repeatedly identified across different genetic backgrounds, populations and environments, may display a great potential of marker-assisted breeding for LP in cotton.

In cotton, several genes associated with LP, such as *Gh_A02G1268* (Su et al. 2016), *Gh_D08G2376* (Huang et al. 2017), *AIL6* and *EIL* (Fang et al. 2017), *Gh_D03G1064* and *Gh_D12G2354* (Sun et al. 2018) and *Gh_D02G0025* (Ma et al. 2018), have been detected via GWAS using different association populations. In the current study, 434 genes were found in the confidence intervals of identified QTLs. Among them, 263 genes were highly expressed in various organs including ovule and fiber developmental stages. We particularly focused on two of these genes, *Gh_D05G1124* and *Gh_D05G0313*, because their exon regions harbored polymorphic SNPs that were responsible for protein-coding differences. Moreover, qRT-PCR analysis revealed that both genes were highly expressed at the ovule and fiber development stages. The closest homologs of *Gh_D05G1124* and *Gh_D05G0313* in Arabidopsis are, respectively, *PP2C* (Protein phosphatase 2C family protein) and *AtLUT2*; those homologs are involved in protein phosphorylation and photosynthesis, two processes related to fiber development. Our results thus point to *Gh_D05G1124* and *Gh_D05G0313* as candidate genes for LP.

Elite-allele loci are valuable resources for crop breeding programs, and the accumulation of superior alleles is an efficient way to improve target traits in crop plants (Su et al. 2016). In wheat, the nine superior alleles contributing to a high thousand-kernel weight were uncovered in multiple environments in the cultivar Pindong34, and proper pyramiding of superior alleles was beneficial to increase wheat yield (Sun et al. 2017a). In rapeseed, the aggregation of superior alleles significantly associated with earliness resulted in earlier flowering or maturity (Zhou et al. 2018). In cotton, three favorable SNP alleles were selected to identify the effects of allelic variation on *Verticillium* wilt resistance in upland cotton, and it was found that the resistance of accessions was increased by pyramiding favorable SNP alleles (Li et al. 2017). In the present study, we similarly found two SNPs significantly associated with LP, i00252Gh and i08888Gh, that had a positive effect on LP. Accessions carrying GG alleles at i00252Gh and i08888Gh had higher LPs than those harboring the AA allele. The phenotypic value of LP increased continuously with the number of favorable alleles. This result indicates that those favorable alleles can be pyramided in a target line by marker-assisted selection. Out of the 276 upland cotton accessions, however, only 16 contained these favorable alleles. This scarcity indicates that these elite loci are not presently well utilized. The future application of favorable alleles thus has great potential in cotton breeding programs.

Author contribution statement DY, XM and WL conceived and designed the research. CS, ZR, KS and XZ performed the experiments. XP, YL, KH and FZ prepared the materials. CS and WL analyzed the data and wrote the paper. DY and XM revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements This work was supported by the National Key R&D Program for Crop Breeding (2016YFD0100306), the Key Project of Science and Technology of Henan Province of China (182102110306), and the Natural Science Foundation of Henan Province of China (152300410010).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

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

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

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixedeffects models using lme4. J Stat Softw 67(1):1–48

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

Cavanagh C, Morell M, Mackay I, Powell W (2008) From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Curr Opin Plant Biol 11:215–221

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

Dong H, Zhao H, Li S, Han Z, Hu G, Liu C, Yang G, Wang G, Xie W, Xing Y (2018) Genome-wide association studies reveal that members of bHLH subfamily 16 share a conserved function in regulating flag leaf angle in rice (*Oryza sativa*). PLoS Genet 14:e1007323

Earl DA, Vonholdt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 4:359–361

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 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 (2017) Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. Nat Genet 49:1089–1098

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Li W, Lin Z, Buckler ES, Qian Q, Zhang QF, Li J, Han B (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42:961–967

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 J, Li J, Zhou J, Wang L, Yang S, Hurst LD, Li WH, Tian D (2018) Identifying a large number of high-yield genes in rice by pedigree analysis, whole-genome sequencing, and CRISPR-Cas9 gene knockout. Proc Natl Acad Sci USA 115:E7559–e7567

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 (Bethesda) 5:1187–1209

Immenkamp M (2006) Correlation and path coefficient analysis for earliness and yield traits in cotton (*G. hirsutum* L.). Asian J Plant Sci 5:27–36

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801–1806

Jamshed M, Jia F, Gong J, Palanga KK, Shi Y, Li J, Shang H, Liu A, Chen T, Zhang Z, Cai J, Ge Q, Liu Z, Lu Q, Deng X, Tan Y, Or Rashid H, Sarfraz Z, Hassan M, Gong W, Yuan Y (2016) Identification of stable quantitative trait loci (QTLs) for fiber quality traits across multiple environments in *Gossypium hirsutum* recombinant inbred line population. BMC Genom 17:197

Li H, Peng Z, Yang X, Wang W, Fu J, Wang J, Han Y, Chai Y, Guo T, Yang N, Liu J, Warburton ML, Cheng Y, Hao X, Zhang P, Zhao J, Liu Y, Wang G, Li J, Yan J (2013) Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat Genet 45:43–50

Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, Li Q, Ma Z, Lu C, Zou C, Chen W, Liang X, Shang H, Liu W, Shi C, Xiao G, Gou C, Ye W, Xu X, Zhang X, Wei H, Li Z, Zhang G, Wang J, Liu K, Kohel RJ, Percy RG, Yu JZ, Zhu YX, Wang J, Yu S (2014) Genome sequence of the cultivated cotton *Gossypium arboreum*. Nat Genet 46:567–572

Li C, Dong Y, Zhao T, Li L, Li C, Yu E, Mei L, Daud MK, He Q, Chen J, Zhu S (2016) Genome-wide SNP linkage mapping and QTL analysis for fiber quality and yield traits in the upland cotton recombinant inbred lines population. Front Plant Sci 7:1356

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

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

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28:2397–2399

Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128–2129

Liu X, Teng Z, Wang J, Wu T, Zhang Z, Deng X, Fang X, Tan Z, Ali I, Liu D, Zhang J, Liu D, Liu F, Zhang Z (2017) Enriching an intraspecific genetic map and identifying QTL for fiber quality and yield component traits across multiple environments in upland cotton (*Gossypium hirsutum* L.). Mol Genet Genom 292:1281–1306

Lu Q, Zhang M, Niu X, Wang S, Xu Q, Feng Y, Wang C, Deng H, Yuan X, Yu H, Wang Y, Wei X (2015) Genetic variation and association mapping for 12 agronomic traits in indica rice. BMC Genom 16:1067

Ma Z, He S, Wang X, Sun J, Zhang Y, Zhang G, Wu L, Li Z, Liu Z, Sun G, Yan Y, Jia Y, Yang J, Pan Z, Gu Q, Li X, Sun Z, Dai P, Liu Z, Gong W, Wu J, Wang M, Liu H, Feng K, Ke H, Wang J, Lan H, Wang G, Peng J, Wang N, Wang L, Pang B, Peng Z, Li R, Tian S, Du X (2018) Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. Nat Genet 50:803–813

Mezmouk S, Dubreuil P, Bosio M, Decousset L, Charcosset A, Praud S, Mangin B (2011) Effect of population structure corrections on the results of association mapping tests in complex maize diversity panels. Theor Appl Genet 122:1149–1160

Mitchell-Olds T (2010) Complex-trait analysis in plants. Genome Biol 11:1–3

Nei M (1972) Genetic distance between populations. Am Nat 106:283–292

Nie X, Huang C, You C, Li W, Zhao W, Shen C, Zhang B, Wang H, Yan Z, Dai B, Wang M, Zhang X, Lin Z (2016) Genome-wide SSR-based association mapping for fiber quality in nation-wide upland cotton inbreed cultivars in China. BMC Genom 17:352

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

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575

Qin H, Chen M, Yi X, Bie S, Zhang C, Zhang Y, Lan J, Meng Y, Yuan Y, Jiao C (2015) Identification of associated SSR markers for yield component and fiber quality traits based on frame map and upland cotton collections. PLoS ONE 10:e0118073

Reinisch AJ, Dong JM, Brubaker CL, Stelly DM, Wendel JF, Paterson AH (1994) A detailed RFLP map of cotton, *Gossypium hirsutum* × *Gossypium barbadense*: chromosome organization and evolution in a disomic polyploid genome. Genetics 138:829–847

Rong J, Abbey C, Bowers JE, Brubaker CL, Chang C, Chee PW, Delmonte TA, Ding X, Garza JJ, Marler BS, Park CH, Pierce GJ, Rainey KM, Rastogi VK, Schulze SR, Trolinder NL, Wendel JF, Wilkins TA, Williams-Coplin TD, Wing RA, Wright RJ, Zhao X, Zhu L, Paterson AH (2004) A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (*Gossypium*). Genetics 166:389–417

Said JI, Knapka JA, Song M, Zhang J (2015a) Cotton QTLdb: a cotton QTL database for QTL analysis, visualization, and comparison between *Gossypium hirsutum* and *G. hirsutum* × *G. barbadense* populations. Mol Genet Genom 290:1615–1625

Said JI, Song M, Wang H, Lin Z, Zhang X, Fang DD, Zhang J (2015b) A comparative meta-analysis of QTL between intraspecific *Gossypium hirsutum* and interspecific *G. hirsutum* × *G. barbadense* populations. Mol Genet Genom 290:1003–1025

Saidou AA, Thuillet AC, Couderc M, Mariac C, Vigouroux Y (2014) Association studies including genotype by environment interactions: prospects and limits. BMC Genet 15:3

Schmittgen TD, Livak KJ (2008) Analyzing real-time PCR data by the comparative C(T) method. Nat Protoc 3:1101–1108

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

Su J, Li L, Zhang C, Wang C, Gu L, Wang H, Wei H, Liu Q, Huang L, Yu S (2018) Genome-wide association study identified genetic variations and candidate genes for plant architecture component traits in Chinese upland cotton. Theor Appl Genet 131:1299–1314

Sun C, Zhang F, Yan X, Zhang X, Dong Z, Cui D, Chen F (2017a) Genome-wide association study for 13 agronomic traits reveals distribution of superior alleles in bread wheat from the Yellow and Huai Valley of China. Plant Biotechnol J 15:953–969

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 (2017b) 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

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–2729

Team RDC (2014) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Computing 14:12–21

Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, Flint-Garcia S, Rocheford TR, McMullen MD, Holland JB, Buckler ES (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat Genet 43:159–162

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

Turner SD (2014) qqman: an R package for visualizing GWAS results using Q–Q and manhattan plots. Biorxiv. https://doi.org/10.1101/005165

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423

Wang C, Zhang T, Guo W (2013) The mutant gene negatively affects many aspects of fiber quality traits and lint percentage in cotton. Crop Sci 53:27–37

Wang H, Huang C, Guo H, Li X, Zhao W, Dai B, Yan Z, Lin Z (2015) QTL mapping for fiber and yield traits in upland cotton under multiple environments. PLoS ONE 10:e0130742

Wang M, Tu L, Lin M, Lin Z, Wang P, Yang Q, Ye Z, Shen C, Li J, Zhang L, Zhou X, Nie X, Li Z, Guo K, Ma Y, Huang C, Jin S, Zhu L, Yang X, Min L, Yuan D, Zhang Q, Lindsey K, Zhang X (2017) Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. Nat Genet 49:579–587

Wang B, Wu Z, Li Z, Zhang Q, Hu J, Xiao Y, Cai D, Wu J, King GJ, Li H, Liu K (2018a) Dissection of the genetic architecture of

three seed-quality traits and consequences for breeding in *Brassica napus*. Plant Biotechnol J 16:1336–1348

Wang YY, Li YQ, Wu HY, Hu B, Zheng JJ, Zhai H, Lv SX, Liu XL, Chen X, Qiu HM, Yang J, Zong CM, Han DZ, Wen ZX, Wang DC, Xia ZJ (2018b) Genotyping of soybean cultivars with medium-density array reveals the population structure and QTNs underlying maturity and seed traits. Front Plant Sci 9:610

Wei L, Jian H, Lu K, Filardo F, Yin N, Liu L, Qu C, Li W, Du H, Li J (2016) Genome-wide association analysis and differential expression analysis of resistance to Sclerotinia stem rot in *Brassica napus*. Plant Biotechnol J 14:1368–1380

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

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Zhang JF, Stewart JM (2000) Economical and rapid method for extracting cotton genomic DNA. J Cotton Sci 4:193–201

Zhang HB, Li Y, Wang B, Chee PW (2008) Recent advances in cotton genomics. Int J Plant Genom 2008:742304

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

Zhao Z, Zhang H, Fu Z, Chen H, Lin Y, Yan P, Li W, Xie H, Guo Z, Zhang X, Tang J (2018) Genetic-based dissection of arsenic accumulation in maize using a genome-wide association analysis method. Plant Biotechnol J 16:1085–1093

Zheng XM, Gong T, Ou HL, Xue D, Qiao W, Wang J, Liu S, Yang Q, Olsen KM (2018) Genome-wide association study of rice grain width variation. Genome 61:233–240

Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, Yu Y, Shu L, Zhao Y, Ma Y, Fang C, Shen Y, Liu T, Li C, Li Q, Wu M, Wang M, Wu Y, Dong Y, Wan W, Wang X, Ding Z, Gao Y, Xiang H, Zhu B, Lee SH, Wang W, Tian Z (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol 33:408–414

Zhou Q, Han D, Mason AS, Zhou C, Zheng W, Li Y, Wu C, Fu D, Huang Y (2018) Earliness traits in rapeseed (*Brassica napus*): SNP loci and candidate genes identified by genome-wide association analysis. DNA Res 25:229–244

Zhu C, Gore M, Buckler ES, Yu J (2008) Status and prospects of association mapping in plants. Plant Genome 1:5–20