Rapid identification of a candidate gene related to fiber strength using a superior chromosome segment substitution line from *Gossypium hirsutum* × *Gossypium barbadense* via bulked segregant RNA-sequencing

CURRENT STATUS: UNDER REVIEW

BMC Genetics  BMC Series

Qi Zhang
Cotton Research Institute

Pengtao Li
Anyang Institute of Technology
ORCiD: 0000-0002-3403-8686

Aiying Liu
Cotton Research Institute

Quanwei Lu
Anyang Institute of Technology

Juwu Gong
Cotton Research Institute

Qun Ge
Cotton Research Institute

Junwen Li
Cotton Research Institute

Wankui Gong
Cotton Research Institute

Haihong Shang
Cotton Research Institute

Yuzhen Shi
Cotton Research Institute
Abstract

*Gossypium* is the most widely cultivated commercial crop producing natural fiber around the world, and fiber strength principally determined during the secondary wall thickening period is a critical trait for fiber quality. Based on the developed BC 5 F 3:5 CSSLs (chromosome segment substitution lines) from *G. hisutum* CCRI36 × *G. barbadense* Hai1, the superior MBI9915 was chosen to construct the secondary segregated population BC 7 F 2 with its recurrent parent CCRI36, which was subjected to Bulk segregant RNA-sequencing (BSR-seq) for rapid identification of candidate genes related to fiber strength. Four fiber-transcriptome libraries were separately constructed and sequenced, including two parents (CCRI36 and MBI9915) and two extreme pools at 20 DPA (days post anathesis). Through multiple comparisons, 3742 DEGs (differentially expressed genes) and 3252 DEGs were separately identified between two parents and between two extreme pools, while 536 DEGs were overlapped between parent and extreme pool groups. A total of 831 high-probability SNPs (single nucleotide polymorphism) were identified relevant to fiber strength between two extreme pools through allelic-polymorphism comparison in mRNA sequences, and 18 correlated regions with 1981 annotation genes were finally screened by linkage analysis with SNP-index method, of which including only 12 common genes differentially expressed both between two parents and two pools. Interesting, there was one correlated region consistent with the previous study with the same parents on chromosome A07 with 13-14 Mb, and one common DEG (Gh_A07G0837) in the candidate region was identified in both parents and extreme pools, which has been reported to be involved in fiber strength development through regulating reactive oxygen species (ROS) activity. The
reliability of BSR-seq results was validated by the quantitative real-time PCR (qRT-PCR) experiments on 5 DEGs at 20 DPA. This study focuses on bulked segregant analysis of the extreme pools from segregation population developed by superior CSSL and its recurrent parent, indicating that BSR-seq can be efficiently applied on rapid identification for candidate genes related to the significant quantitative traits, which provides valuable contributions for comprehension of fiber strength formation in cotton.

Introduction

Cotton is one of the most important commercial crops all over the world, providing the major natural fibers for the textile industry. There are only four existing cultivated species of *Gossypium* worldwide, including diploid *G. herbaceum* and *G. arboretum* (2n = 2 × = 26), and tetraploid *G. hirsutum* and *G. barbadense* (2n = 4 × = 52) (Wendel and Albert, 1992; Grover et al., 2008). With the development of modern industry and people's living standard, higher requirements has been put forward for modern textile mills to achieve better spinning performance, which mainly rely on more efficient spinning technologies, such as the high-speed and automated operations (Bradow and Davidonis, 2000). Therefore, how to obtain the high-quality fiber has been the vital goal in cotton breeding at the present stage, meanwhile remaining a challenging task for cotton researchers because of genetic-basis complexities of fiber formation.

As the most widely cultivated cotton species, *G. hirsutum* and *G. barbadense* can contribute more than 95% fiber production per year, of which the former harbors the characteristics of high yield and moderate fiber performance, while the latter has the advantages of the premium fiber quality and high resistance of *Verticillium*
wilt, yet subjects to the disadvantage of low yield (Zhang et al., 2014; Shi et al., 2015). Chromosome segment substitution lines (CSSLs) arises to offer new thoughts in adequate utilization and integration of excellent traits derived from Upland cotton and Sea Island cotton, which also laid the theoretical foundation of developing new cotton varieties with high yield, fine quality, and disease resistance. To start with a hybridization between the recipient and donor parents, the CSSLs are subjected to multi-generation backcrosses with the recurrent parent and subsequently multi-generation self-crosses, combining marker assisted selection (MAS) to ultimately obtain the Near-isogenic lines (NILs). In consideration of reducing the interference of complex genetic background in population, the CSSLs are the ideal materials appropriate for dividing the complex and variable quantitative traits or phenotypes into a single inheritance factor, which have been widely applied to perform quantitative traits loci (QTL) mapping in tomato (Lycopersicon esculentum) (Eshed and Zamir, 1994), rice (Oryza sativa) (Bian et al., 2010), wheat (Triticum aestivum) (Liu et al., 2006), maize (Zea mays) (Lu et al., 2012), soybean (Glycine max) (Wang et al., 2013b), and cotton (Stelly et al., 2005; Luan et al., 2009; Song et al., 2017; Li et al., 2019).

Cotton fiber derived from ovule epidermal cells is a single-celled seed trichome. Although all the ovule epidermal cells in cotton have the potential to differentiate into fibers, only 15–25% of them can eventually develop into mature fibers (Lee et al., 2007). The development of cotton fiber can be divided into five overlapping stages: initiation (− 3 to + 3 DPA), elongation (3 to 23 DPA), transition, secondary wall synthesis (16 to 40 DPA), and maturation (40 to 50 DPA) (Haigler et al., 2012), which collectively affect the quality traits of fiber quality, including fiber length, strength, micronaire, elongation, and uniformity (Schubert et al., 1973). With the
rapid development of high-through sequencing technology and complete releasment of Gossypium genomes (Paterson et al., 2012; Li et al., 2014; Li et al., 2015; Wang et al., 2015; Zhang et al., 2015; Du et al., 2018; Wang et al., 2019), more efficient strategies have been applied to identify the genes responsible for fiber length and strength in cotton. Using re-sequencing technology and genome-wide association study (GWAS) strategy on a core collection of upland cotton, some significant candidate genes were identified, including Gh_A10G1256 associated with fiber length, and Gh_A07G1767, Gh_A07G1768, and Gh_A07G1769 related to fiber-strength (Ma et al., 2018). Through fine mapping of stable QTL and quantitative real-time PCR (qRT-PCR) verification, a LRR-RLK (leucine-rich repeat protein kinase) family protein (Gh_A07G1749) was identified as a candidate gene for fiber-strength QTL qFS07.1 (Fang et al., 2017). Transcriptome analysis on three CSSLs derived from CCRI36 (G. hirsutum) and Hai1 (G. barbadense) during the secondary wall thickening period indicated three differentially expressed genes (DEGs), namely as XLOC_036333 (MNS1), XLOC_029945 (FLA8), and XLOC_075372 (snakin-1), might play the significant roles in regulating fiber strength (Lu et al., 2017).

With the combination of bulked segregant analysis (BAS) method and RNA-seq technology, BSR (bulked segregant RNA-seq) can not only identify the DEGs by comparing the two constructed extreme pools, but also screen the SNPs co-isolated with mutant genes by linkage analysis (Liu et al., 2012). Due to saving plenty of manpower, material resources, and shortening the breeding cycle, BSR-seq has attracted more and more attentions in recent years, which has been an efficient strategy for rapid QTL mapping and efficient identification of candidate genes in connection with key traits in plants (Liu et al., 2012; Trick et al., 2012; Wang et al., 2013a; Yates et al., 2014). Through performing BSA-seq on waterlogging sensitive
and resistant pools in maize, a candidate gene (GRMZM2G055704) in the key QTL located in umc1619-umc1948onchromosome1 was identified to response waterlogging (Du et al., 2017). In order to screen the genes responsive to early and late flowering, BSR-seq association analysis on the extreme pools in tree peony obtained 291 unigenes, of which 7 DEGs (c42942.graph_c0, c58332.graph_c0, c58361.graph_c0, c57417.graph_c0, c46352.graph_c0, c53143.graph_c0, and c58526.graph_c0) were confirmed to relate with early and late flowering (Hou et al., 2018).

In the present study, the superior CSSL MBI9915 with high yield and excellent fiber quality was chosen to develop the secondary segregation population BC7F2 with CCRI36, which was derived from BC5F3:5 CSSL populations constructed by G. barbadense Hai1 as the donor parent and G. hirsutum CCRI36 as the recipient and recurrent parent (Li et al., 2017). Due to concentration on the significant traits of fiber strength, we performed BSA-seq analysis on the fiber samples collected from two parents (MBI9915 and CCRI36) and two extreme pools at 20 DPA, respectively. Plenty of DEGs were identified with comparisons between different samples, which subsequently underwent GO functional enrichment and KEGG pathway analysis. Through SNP-index association analysis method and qRT-PCR technology, candidate genes were finally screened to control fiber strength. These results provide a rapid and efficient method to identify the candidate genes by BSR-seq, which will facilitate the revelation of molecular mechanism of fiber formation in cotton.

Material and Methods

Plant materials

Two parents, the superior CSSL MBI9915 and its recurrent parent CCRI36, as well as
their segregation population BC$_7$F$_2$ were planted at the Anyang experimental farm of Institute of Cotton Research of Chinese Academy of Agricultural Sciences (Anyang, Henan Province) in 2016, where BC$_7$F$_2$:3 populations were planted through plant-to-row method in 2017. The high-yield and wide-adaptability Upland cotton CCRI36 as the recipient parent, and good-quality and disease-resistance Sea Island cotton Hai1 as the donor parent were chosen to conduct the hybridization in 2003, of which the offspring F$_1$ was subsequently subjected to five generations of backcrosses with the recurrent parent CCRI36 and three generations of self-crossings, eventually obtaining the developed BC$_5$F$_3$:5 CSSLs with plant-to-row and plant-to-family method, respectively (Zhang et al., 2012; Li et al., 2017). Based on the multi-year and multi-environment trait surveys, MBI9915 harboring the high-yield and superior-quality fiber was chosen for secondary population construction, and then underwent two-generation backcrossing with CCRI36 and one-generation self-crossing to develop BC$_7$F$_2$ and BC$_7$F$_2$:3 populations, which were planted with row length of 8 m, row spacing of 0.8 m, and plant spacing of 0.25 m.

Phenotypic Analysis

Naturally opened bolls from the BC$_7$F$_2$ population were harvested per plant in 2016, while 30 naturally opened bolls from BC$_7$F$_2$:3 population per row were harvested in 2017, of which the fiber quality traits (including fiber length, strength, micronaire, elongation, and uniformity) were tested with HFT9000 using HVICC international calibration cotton samples in the Cotton Quality Supervision and Testing Center of the Ministry of Agriculture of China.

RNA Isolation

Regarding the day tagging flowers as 0 DPA, 1 developing bolls of fiber samples
were taken from MBI9915 and CCRI36 at 20 DPA were immediately immersed in ice. Comprehensive consideration on the fiber performances of both BC$_7$F$_2$ and BC$_7$F$_{2:3}$, 16 ones with high fiber strength and 16 ones with low fiber strength were separately selected from BC$_7$F$_{2:3}$ lines at 20 DPA to construct high and low extreme pools. After being dissected from the developing bolls, fiber sampled were quickly frozen into the liquid nitrogen, and subsequently stored at -80°C. RNA extractions were conducted on each plant, and the obtained RNA samples were equally mixed to construct the extreme pools. RNA prep Pure Plant Kit (Tiangen, Beijing, China) was used to extract high-quality RNA samples. The quality and quantity of the RNA were verified using 1% agarose gel and spectrophotometer (Nanodrop 2000, Thermo Scientific, USA).

Library Construction and Transcriptome Sequencing

First, high-quality mRNA was enriched from total RNA using Oligo (dT) magnetic beads, and was randomly cleaved into short fragments by fragmentation buffer, then double-chain cDNA was synthesized with mRNA as a template. After purification, the cDNA segments were utilized to construct the cDNA libraries through end-repair, adding the tail of the ploy A, and selecting the size of the fragments. At last, the cDNA libraries were obtained by PCR amplification, and a total of 4 RNA libraries were sequenced using Illumina HiSeq™ 4000 sequence platform (Biomarker Technologies Corporation, Beijing, China).

By high-throughput sequencing, a large amount of raw data was obtained, and stored in FASTQ format file. The raw data was filtered to obtain clean reads. Clean Reads were aligned to the reference genome of TM-1 (Zhang et al., 2015), and the position information of reference genome or gene and the sequence characteristic information of the sequencing samples were obtained through STAR software (Dobin
et al., 2013).

Analysis of Differentially Expressed Genes (DEGs)

EBSeq (Leng et al., 2013) software was used to analyze differentially expressed genes, FPKM (fragments per kb per million of the mapped reads) (Mortazavi et al., 2008) parameter as an index for reviewing transcripts or gene expression levels. In the detection of differentially expressed genes, Fold Change (FC) ≥ 2 and False Discovery Rate (FDR) ≤ 0.01 are identified as the DEGs screening standard, of which the former represents the ratio of expression between two parents and pools at 20 DPA, and the latter is obtained by correcting the p-value of difference significance. The accepted Benjamini-Hochberg method was used to correct the p-value obtained from the original hypothesis test. In this experiment, p-value ≤ 0.05 was selected as the criterion for screening differentially expressed genes. To understand the functional classification of all the obtained DEGs, BLAST (Apweiler et al., 2004) software was used to annotate the coding genes in Gene ontology (GO) (Ashburner et al., 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases (Kanehisa et al., 2007).

BSR-Associate analysis

Single nucleotide polymorphism (SNP) callings were processed with GATK (McKenna et al., 2010) software toolkit (http://www.broadinstitute.org/gatk/guide/best-practices.php) with strict filtering of SNPs: SNP cluster filtering, near Indel SNP filtering and neighboring InDel filtering (Joke et al., 2012). Before doing correlation analysis, SNPs were filtered again. The criteria were as follows: firstly, SNP loci with multiple genotypes were filtered out, and then SNP loci with read support less than 4 were filtered out, Re-filter SNP loci with identical genotypes between mixing pools and SNP loci with recessive mixing pool genes not from recessive parents. Finally,
high-quality SNP loci were obtained for SNP-index analysis. SNP-index method was for marker association analysis based on genotype frequency differences between high fiber strength pool and low fiber strength pool (Altschul et al., 1997; Trapnell et al., 2010), which was mainly used to find significant differences in genotype frequencies between high fiber strength pool and low fiber strength pool using $\Delta$(SNP-index). $\Delta$(SNP-index) was calculated by subtracting the SNP-index of high fiber strength pool from low fiber strength pool.

Quantitative Real-Time PCR (qRT-PCR)

Five genes on chromosome 7 were chosen to conduct quantitative real-time PCR for verifying BSR-seq results. Two parents (CCRI36 and MBI9915) at 20 DPA with three technical replicates were used for qRT-PCR. The cDNAs synthesized by using a HiScript® II Q RT SuperMix for qPCR (+ gDNA wiper) kit (Vazyme Biotech, Nanjing, China). ChamQTM Universal SYBR® qPCR Master Mix (Vazyme Biotech, Nanjing, China) kit for real-time PCR was used for qRT-PCR amplification, which were performed on the ABI 7500 fast Real-Time PCR System (Applied Biosystems, USA), following the protocol of the kit. $\beta$-Actin gene (Primer sequence; F: 5′-ATCCTCCGTCTTTGACCTTG-3′, and R: 5′-TGTCGTCAGGCAACTCAT-3′) was used as an internal control to normalize the genes expression. The qRT-PCR Master mix was prepared as 20 µL, the process was employed according to the following steps: one cycle of 95 °C for 30 s; 40 cycle of 95 °C for 10 s, 60 °C for 30 s; one cycle of 95 °C 15 s, 60 °C for 60 s, 95 °C for 15 s. $2^{-\Delta \Delta C_t}$ method was followed for obtaining the relative gene expression (Livak and Schmittgen, 2001).

Gene clone, sequence and phylogenetic analysis

RNAs of CSSL MBI9915 and G. barbadense Hai1 were extracted according to the RNAIsolation method. NCBI Primer-blast Webpage was used to design the candidate
gene’s primers (F: ATGTTTCATTTCTTCTTGTC, R: TCATAAAACTAAAGGCTCCA).

TransScript All-in-one First-Strand cDNA Synthesis SuperMix for PCR kit (Transgen Biotech, Beijing, China) was used to Synthesis cDNA and TransStartFastPfu Fly DNA Polymerase kit (Transgen Biotech, Beijing, China) was used for amplifying the coding sequence (CDS) of the candidate gene. PCR conditions were one cycle of 95 °C for 2 min; 40 cycles of 95 °C for 20 s, 60 °C for 20 s, 72 °C for 40 s; one cycle of 72 °C for 5 min. The product was cloned into pEASY-Blunt Zero vector (Transgen Biotech, Beijing, China) for sequencing.

Alignment of Gh_A07G0837 protein sequences in CCRI36 and Hai1 was performed by DNAMAN program. Alignment of Gh_A07G0837 protein sequences among four cultivated species of Gossypium were performed by NCBI protein-blast. The phylogenetic tree of proteins in four cultivated species of Gossypium, CCRI36 and Hai1 was constructed with the MEGA program (version 7.0) by the neighbor joining (NJ) method with 1000 bootstraps (Kumar et al., 2016).

Results

Phenotypic data

On behalf of taking extreme-pool samples with more accurate phenotypic traits, the chosen individuals from BC7F2 population in 2016 were re-planted with plant-to-row to obtain BC7F2:3 populations in 2017, of which 30 naturally opened bolls were subjected to fiber-quality testing (Table 1). The average fiber-strength value of 16 individuals from high extreme pools was 39.6 in 2016, while the corresponding value of 16 individuals from low extreme pools in 2016 was 30.0; the mean fiber-strength values of 16 lines from high and low extreme pools in 2017 were 36.9 and 30.6, respectively. Referring to the fiber-strength comparisons between 2016 and
2017 materials, both higher consistency and significant difference were identified, which directly verified the reliability of sampling the individuals in 2016 for BSR-seq analysis.

| Line number in high fiber strength pool | FS(cN/tex) in 2016 | FS(cN/tex) in 2017 | Line number in low fiber strength pool | FS(cN/tex) in 2016 | FS(cN/tex) in 2017 |
|-----------------------------------------|--------------------|--------------------|----------------------------------------|--------------------|--------------------|
| 4379                                    | 40.0               | 34.7               | 4517                                   | 34.2               | 31.4               |
| 4416                                    | 38.7               | 37.2               | 4513                                   | 33.8               | 33.5               |
| 4401                                    | 39.5               | 38.4               | 4447                                   | 29.1               | 29.7               |
| 4402                                    | 38.8               | 35.3               | 4449                                   | 29.5               | 29.1               |
| 4435                                    | 41.1               | 40.2               | 4413                                   | 31.3               | 30.2               |
| 4437                                    | 39.6               | 40.9               | 4448                                   | 30.7               | 31.9               |
| 4405                                    | 38.5               | 35.7               | 4556                                   | 28.6               | 31.1               |
| 4404                                    | 38.5               | 33.1               | 4676                                   | 28.2               | 28.8               |
| 4429                                    | 40.7               | 38.0               | 4677                                   | 29.2               | 30.2               |
| 4381                                    | 38.7               | 35.5               | 4493                                   | 29.8               | 30.7               |
| 4380                                    | 39.4               | 34.2               | 4541                                   | 29.3               | 31.4               |
| 4434                                    | 41.9               | 40.7               | 4674                                   | 29.5               | 30.3               |
| 4436                                    | 39.8               | 35.6               | 4675                                   | 29.7               | 28.8               |
| 4430                                    | 39.3               | 40.5               | 4164                                   | 27.0               | 27.6               |
| 4378                                    | 40.1               | 33.5               | 4544                                   | 29.3               | 30.6               |
| 4403                                    | 38.7               | 37.6               | 4415                                   | 30.7               | 33.7               |
| Average                                  | 39.6               | 36.9               | Average                                | 30.0               | 30.6               |

Sequence Assembling and Analysis

For rapid identification of candidate genes related to fiber strength, BSR-seq was separately conducted on CCRI36, MBI99915, high and low extreme pools, and the obtained raw reads were then subjected to filtering low-quality reads, totally generating 186.622 million (approximately 27.9G data) clean reads from two parents and 369.247 million (approximately 55.2G data) clean reads from two extreme pools (Table 2). Subsequently, all the clean data mapped to the reference genome (Zhang et al., 2015), of which the alignment percentage of mapped reads was over 83.97%, while no less than 68.69% were uniquely mapped reads.

Meanwhile, the average percentages of Q30 and GC content were 93.83% and 44.56%, respectively, which laterally proved the reliability of our BSR-seq results.
Table 2  
Summary of the BSR-Seq outcomes of 4 samples.

| Samples     | Total Reads | Clean_Reads | Clean_Base   | Mapped Reads       | Uniq Mapped Reads | Q30(%) | GC(%) |
|-------------|-------------|-------------|--------------|--------------------|-------------------|--------|-------|
| CCRI36      | 91461204    | 45730602    | 13.67G       | 76814674(83.99%)   | 62824773(68.69%)  | 93.76  | 45.22 |
| MBI9915     | 95161446    | 47580723    | 14.23G       | 80622188(84.72%)   | 71780633(75.43%)  | 93.86  | 44.10 |
| High pool   | 183129156   | 91564578    | 27.38G       | 155155932(84.72%)  | 130288352(71.15%) | 93.79  | 44.77 |
| Low pool    | 186118560   | 93059280    | 27.82G       | 156286104(83.97%)  | 139208826(74.80%) | 93.93  | 44.16 |

Analysis of Differentially Expressed Genes (DEGs)

Based on the reference genome TM-1, a large number of differentially expressed genes (Fold change ≥ 2 and FDR < 0.01) were significantly identified by EBSeq (Leng et al., 2013), to be special 3742 DEGs (793 up-regulated and 2949 down-regulated ones) between the two parents and 3252 DEGs (2913 up-regulated and 339 down-regulated ones) between the extreme pools at 20 DPA, with 536 genes in common (Fig. 1A). Obviously, there were more DEGs in CCRI45-vs-MBI9915 group (Fig. 1B) than those in High pool-vs-Low pool group (Fig. 1C), of which the former group identified over 8 times as many down-regulated DEGs, while the latter group identified more than 4 times as many up-regulated DEGs, implying the dramatic variations between the two group. Having performing the hierarchical clustering analysis on DEGs from two pairwise-comparison groups, most genes showed diverse expression levels either in two parent group (Fig. 2A) or in two extreme pools (Fig. 2B), while similar phenomenon occurred on the slight DEGs with high-ratio expression qualities between two groups, which might be closely related to the mechanism regulation of fiber strength.

GO enrichment analysis was subsequently conducted on the whole DEGs from two parent group, two extreme pools, and common DEGs (Fig. 3), respectively, aiming at identifying the putative functional genes or significant signal pathways during fiber
development. Top GO terms were separately enriched into three main categories, and as for the DEGs from both two parent and two extreme pool groups or the common DEGs, cell part and cell were the predominant subcategories in cellular component, catalytic activity, and binding were the top subcategories in molecular functional category, and metabolic process, cellular process, and single-organism process were the most enriched subcategories in biological process.

Similarly, the whole DEGs were subjected to KEGG pathway analysis and were mapped onto five top-level subcategories of cellular processing, environmental information processing, genetic information processing, metabolism and organismal systems. At 20 DPA, DEGs in two parent groups (Fig. 4A) were mostly annotated into plant hormone signal transduction (ko04075), carbon metabolism (ko01200), while plant hormone signal transduction (ko04075) and starch and sucrose metabolism (ko00500) were the top two pathway subcategories for DEGs in two pool groups (Fig. 4B). The common DEGs both in two groups were mostly enriched to plant hormone signal transduction (ko04075) pathway subcategories (Fig. 4C).

qRT-PCR Verification of BSR-seq results

To confirm the reliability of BSR-seq results, qRT-PCR was performed on the five DEGs from two parents (CCRI36 and MBI9915) at 20 DPA, namely as Gh_A07G0831, Gh_A07G0837, Gh_A07G0838, Gh_A07G0858 and Gh_A07G0973, of which the primer sequences are enlisted in Table 3. According to the BSR-seq data, all the five DGEs showed significantly lower expression levels in MBI9915 than those in CCRI36, which presented the similar difference-pattern between two parents in qRT-PCR results. Only Gh_A07G0831 showed non-significant difference on the expression levels between two parents (Fig. 5A), while Gh_A07G0858 (Fig. 5D) and the other three DEGs (Gh_A07G0837, Gh_A07G0838 and Gh_A07G0973) (Fig. 5B, C, and D) had
significantly and extremely significantly higher expression qualities in CCRI36 compared with those in MBI9915, respectively. The qRT-PCR data were consistent with the BSR-seq results, fully demonstrating the reliability of this study.

**Table 3**

Primers are used for qRT-PCR verification.

| gene ID     | Forward primer | Reverse primer               |
|-------------|----------------|------------------------------|
| Gh A07G0831 | TCCTCAACGTGCTCAACGGTTT | CAACAAGGAAATGGCGCCCAAG      |
| Gh A07G0837 | CAGGCGAGTTACTGAGTACGCC | CATCTGCTCCGAAGCTCTTGAC      |
| Gh A07G0838 | CACCACACCACCATCACCACAC | TGCCAGAATGGTGGGTTTCGTT      |
| Gh A07G0858 | GAGGGAGCTGATGAGAGAAAG | AGGCTTGGTGCCCAAATGACAT      |
| Gh A07G0973 | CCGGAAAACACTGTGCGTGACT | ATCCCAACGCTTCAGCTTCTC       |

Chromosome region confirmation and candidate gene identification

To map candidate genes related to fiber strength, BSR-seq analyses were conducted on two extreme pools at 20 DPA, and 16 samples with high fiber strength and 16 samples with low fiber strength were collected and equally mixed for RNA sequencing, respectively. After being trimmed, the clean reads were aligned to the TM-1 reference genome for SNP identifications between high and low fiber-strength pools, resulting in a total of 831 high-probability SNPs via bulked segregant analysis. All the identified SNPs were quantified with SNP-index, of which the information were synthetically compared between high and low fiber-strength pools to calculate and plot the \( \Delta \text{(SNP-index)} \) against the TM-1 reference genome positions (Fig. 6). In this study, associated chromosomal regions were identified with \( |\Delta \text{(SNP-index)}| > 0.6365 \) as the threshold value.

Totally, 18 associated chromosomal regions located in 15 chromosomes (8 ones in A sub-genome and 7 ones in D sub-genome) were identified, and there were two different chromosomal regions in chromosome A05, A07, and A09, respectively (Table 4 and Fig. 6). Referred to the TM-1 genome information, a total of 1981 genes were annotated in 18 associated regions, including 141 ones with non-synonymous mutation, which were generally regarded the potential candidate genes
involved in fiber strength formation. Meanwhile, GO enrichment and KEGG pathway analyses were conducted on the whole annotated genes, and the most genes were enriched to cell part and cell, catalytic activity and binding, and cellular process and single-organism process in cellular component, molecular function, and biological process based on GO categories (Fig. 7A), while ribosome (ko03010), carbon metabolism (ko01200), biosynthesis of amino acids (ko01230) and protein processing in endoplasmic reticulum (ko04141) were top 4 pathways with the greatest number of genes (Fig. 7B).

Table 4
Information on 18 linkage domains.

| Chromosome number | Start Site | End Site | Size   |
|-------------------|------------|----------|--------|
| D08               | 5910000    | 9900000  | 3.99001|
| A13               | 64290000   | 68280000 | 3.99001|
| A09               | 43030000   | 20800000 | 0.05001|
| A09               | 45510000   | 49500000 | 3.99001|
| D04               | 48990000   | 49360000 | 0.37001|
| A11               | 20140000   | 24600000 | 4.46001|
| A10               | 9800000    | 13560000 | 3.76001|
| A01               | 88790000   | 92780000 | 3.99001|
| D07               | 31960000   | 33890000 | 1.93001|
| D10               | 56610000   | 60600000 | 3.99001|
| A07               | 1160000    | 1400000  | 0.24001|
| A07               | 13700000   | 20440000 | 6.74001|
| D02               | 12820000   | 16810000 | 3.99001|
| D03               | 9740000    | 10220000 | 0.48001|
| D06               | 15760000   | 19750000 | 3.99001|
| A05               | 10460000   | 14140000 | 3.68001|
| A05               | 24180000   | 24520000 | 0.34001|
| A06               | 24030000   | 24070000 | 0.04001|
| Total             |            |          | 50.02021|

Among all the 1981 genes, 69 genes and 93 genes were differentially expressed between two parents and between two extreme pools, respectively, while only 12 common DEGs were identified between two groups. According to annotation information of function genes shown in Table 5, 2 genes might play the significant roles during fiber development, namely as Gh_A05G1089 and Gh_A07G0837, which were separately annotated as abscisic acid transport in biological process and Golgi apparatus in cell component. Abscisic acid has been reported as a signal of secondary wall thickening to make great contribution to fiber development, and
Golgi apparatus were similarly of importance for plant growth (Ramsey and Berlin, 1976; Gokani et al., 1998; Kurek et al., 2002). Interestingly, the region of chromosome 7 associated with the correlation analysis was also consistent with the previous QTL localization region, of which segregation populations were constructed by MBI9915 and CCRI36. A associated region with 6.74 Mb on chromosome A07 (chr7: 13.7–20.4 Mb) was overlapped (chrA07:14–16 Mb) (Song et al., 2017), therefore the only overlapped DEG Gh_A07G0837 was deemed as a candidate gene related to fiber strength.

Table 5
Functional annotation of 12 important genes.

| Gene ID    | Annotation                                                                 |
|------------|---------------------------------------------------------------------------|
| Gh_A01G1560| PB1 domain                                                                |
| Gh_D06G0865| Pectinesterase                                                            |
| Gh_A10G0626| Glycerophosphoryl diester phosphodiesterase family                        |
| Gh_D08G0638| Protein of unknown function                                               |
| Gh_A07G0837| Golgi apparatus                                                           |
| Gh_A11G1567| Regulator of chromosome condensation (RCC1) repeat                       |
| Gh_A05G1141| Transcription factor regulating root and shoot growth via Pin3           |
| Gh_A13G1277| EPSP synthase (3-phosphoshikimate 1-carboxyvinyltransferase)             |
| Gh_D10G2220| Trypsin and protease inhibitor                                            |
| Gh_A13G1263| Intracellular membrane-bounded organelle                                  |
| Gh_A05G1089| Abscisic acid transport                                                  |
| Gh_A05G1261| Chlorophyll A-B binding protein                                           |

Clone, sequence and phylogenetic analysis on Gh_A07G0837

For further confirming the contribution of the candidate gene during fiber development, gene clone was conducted on the Gh_A07G0837 in CCRI36 and Hai1, respectively. The coding sequence (CDS) length was 807 bp both in CCRI36 and Hai1, while 2 SNPs were identified to separately locate in the 30th base and the 486th base (Fig. 8A), of which the former resulted in non-synonymous mutation, specifically encoding the aspartic acid in CCRI36 but the glutamic acid in Hai1, while the latter caused synonymous mutation. The CDS length of the candidate gene was only 627 bp in TM-1 reference genome, and the short 181 bp might be the second
intron sequence based on annotation information, indicating the probable occurrence of alternative splicing (AS) event in CCRI36 and Hai1. Subsequently, the phylogenetic relationships of Gh_A07G0837 protein sequences were analyzed among the CCRI36, Hai1 and other cotton species, and BLAST results showed the deduced amino acid sequence of Gh_A07G0837 in CCRI36 were highly similar to G. barbadense (99% identity and 95% similarity), while 81% identity and 97%, 95% and 95% similarity to G. hirsutum, G. raimondii and G. arboreatum. However, for the deduced amino acid sequence of Gh_A07G0837 in Hai1, G. barbadense (98% identity and 95% similarity) also had high similarity, G. hirsutum, G. raimondii and G. arboreatum was 80% identity and 97%, 95%, and 95% similarities (Fig. 8B).

Discussion:
Cotton (Gossypium L.), one of the most important cash crops, produces the vast majority of natural fiber, which is widely cultivated all over the world. With the development of modern textile industry and the improvement of human living standard, high yield has not been the only goal in the cotton breeding, while superior fiber quality has similarly attracted more and more attentions. However, either Upland cotton or Sea Island cotton, as the two most widely cultivated varieties, could not simultaneously meet the above-mentioned demands, of which the former subjects to the ordinary fiber performance and narrow genotypic milieu, and the latter has the disadvantages of poor yield and low adaptability (Ulloa et al., 2004; Zhang et al., 2009). An effective and economical method has been put forward to develop chromosome segment substitution lines (CSSLs) in cotton production, utilizing the conventional breeding (hybridization, backcross and self-crossing) and maker-assist selection (MAS) technologies to cultivate novel cotton
varieties harboring high yield and superior fiber quality. As to fiber strength, one of the critical traits determining cotton quality, its formation principally depends on the secondary wall thickening period, and in-depth studies on candidate gene identification will greatly contribute to the molecular-mechanism revelation of fiber development.

Based on the multiple-year and multiple-environment trait surveys of developed CSSLs BC$_5$F$_{3:5}$ constructed by G. hirsutum CCRI36 and G. barbadense Hai1, superior MBI9915 was chosen to develop the segregation population F$_2$ and F$_{2:3}$ together with its receptor and recurrent parent CCRI36, and 41 polymorphic SSR markers screened from the high-density linkage map of CCRI36 × Hai1 (Shi et al., 2015) were utilized to conduct QTL mapping, resulting in 24 QTLs related with cotton yield and fiber quality (Song et al., 2017). Meanwhile, transcriptome sequencing was performed on superior MBI9915 and ordinary CCRI36 during fiber development (at 10, 15, 20, 25, and 28 DPA), and 1801 identified DEGs were enriched into cell wall organization and response to oxidative stress and auxin pathways (Li et al., 2017).

The above-mentioned studies have lay the solid foundations for candidate gene identification related to fiber strength.

Combining bulked segregant analysis (BAS) method and RNA-seq technologies, BSR-seq could analyze the DEGs located in associated chromosomal regions by comparing two extreme pools, finally rapid identification of the candidate genes relevant to key quantitative traits. In this study, the CSSL MBI9915 and its recurrent parent CCRI36 were used to develop BC$_7$F$_2$ and BC$_7$F$_{2:3}$ populations, which were subjected to fiber quality tests. From BC$_7$F$_2$ population, 16 samples with high fiber strength and 16 samples with low fiber strength were collected at 20 DPA to
construct two extreme pool libraries, and then were separately sequenced together with transcriptome libraries of their parents (MBI9915 and CCRI36). Totally, 186.622 and 369.247 million raw reads were obtained from two parents and two extreme pools, respectively, and 93.311 and 184.623 million clean data were separately obtained from two parents and two extreme pools after low-quality read filtration, of which the mean Q30 was 93.83%, and GC content was 44.56%. Multiple comparisons between samples resulted in 3742 DEGs between two parents and 3252 DEGs between two extreme pools, of which the former included 793 up-regulated and 2949 down-regulated ones, while the latter included 2913 up-regulated and 339 down-regulated ones. Meanwhile, there were 536 common DEGs between parent group and extreme pool group, which might be closely related to the mechanism regulation of fiber strength. GO enrichment and KEGG pathway analyses were conducted on the whole DEGs, and cell part and cell were the predominant subcategories in cellular component, catalytic activity, and binding were the top subcategories in molecular functional category, and metabolic process, cellular process, and single-organism process were the most enriched subcategories in biological process, while plant hormone signal transduction (ko04075) was the most DEGs enriched pathway among two parents, two extreme pools, and common DEGs. Through SNP identifications between two extreme pools, 831 high-probability SNPs were screened via bulked segregant analysis, and 18 associated chromosome regions involved in 15 chromosomes were confirmed, including 8 ones in A sub-genome and 7 ones in D sub-genome. There were 1981 genes annotated in all the associated regions, of which 141 ones with non-synonymous mutation were the significant objects for further research related to fiber strength. Combined with the obtained DEGs from different samples, 69 and 93 DEGs were identified between two
parents and between two extreme pools, respectively, while only 12 common genes were differentially expressed between two groups. Two candidate genes were finally screened based on functional annotation of reference genome, namely as namely as Gh_A05G1089 (abscisic acid transport) and Gh_A07G0837 (Golgi apparatus), of which the former has been reported to make great contribution to fiber development as a signal of secondary wall thickening (Gokani et al., 1998; Kurek et al., 2002), while the latter was involved in fiber development (Ramsey and Berlin, 1976). What is interesting, the region of chromosome 7 associated with the correlation analysis was also consistent with the previous QTL localization region, and the associated region with 6.74 Mb on chromosome A07 (chr7: 13.7–20.4 Mb) was overlapped (chrA07:14-16 Mb) (Song et al., 2017), indicating the only overlapped DEGs Gh_A07G0837 was deemed as a candidate gene related to fiber strength. Subsequently, gene clone and sequence analyses were conducted on the candidate gene, and CCRI36 and Hai1 had two SNPs difference and one amino acid difference that induced the phenotypic difference between two parents. Using NCBI protein-blast, protein sequence in CCRI36 and Hai1 was highly similar with a hypothetical protein GOBAR AA28214 in G. barbadense (98% identity and 95% similarity). In Arabidopsis thaliana (AT4G13010.1), it was one of zinc-binding dehydrogenase family protein, named Chloroplast Envelope Quinone Oxidoreductase Homolog (CEQORH). CeQORH was associated with the inner membrane of the chloroplast envelope and imported into the organelle by an alternative import pathway (Stéphane et al., 2002; Claudia et al., 2013). Under oxidative stress conditions, CeQORH exhibited the activity of a NADPH-dependent to eliminate highly reactive a,b-unsaturated carbonyls (Curien et al., 2016). During the development of fiber strength, cells need to breathe to provide energy, and a series of redox
reactions occurred in cell respiration. Therefore, this candidate gene may play quite an important role in the development of cotton fiber strength.

Abbreviations

CSSLs
Chromosome segment substitution lines;
BSR-seq
Bulk segregant RNA-sequencing;
DPA
days post anathesis;
DEGs
differentially expressed genes;
SNP
single nucleotide polymorphism;
ROS
reactive oxygen species;
qRT-PCR
quantitative real-time PCR;
MAS
marker assisted selection;
NILs
Near-isogenic lines;
QTL
quantitative traits loci;
GWAS
genome-wide association study;
LRR-RLK
leucine-rich repeat protein kinase;
BSA
bulked segregant analysis;
FPKM
fragments per kb per million of the mapped reads;
FC
Fold Change;
FDR
False Discovery Rate;
GO
Gene ontology;
KEGG
Kyoto Encyclopedia of Genes and Genomes;
CDS
coding sequence;
AS
alternative splicing;
CEQORH
Chloroplast Envelope Quinone Oxidoreductase Homolog;
FS
fiber strength.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Availability of data and materials
The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

Funding
This study was supported by the National Natural Science Foundation of China (31621005 and 31801404) and Joint Funds of the National Natural Science Foundation (U1804103), the National Agricultural Science and Technology Innovation project for CAAS (CAAS-ASTIP-2016-ICR), and the Central Level of the Scientific Research Institutes for Basic R & D Special Fund Business (Y2017PT51).

**Author Contributions Statement**

YLY, YZS and QZ conceived and designed the study. QZ and PTL drafted the manuscript. AYL, QWL, JWG, and QG participated in segregation population construction and fiber sample collections. JWL, WKG, and HHS participated in BSR-seq and data analysis. All the authors read and approved the final manuscript.

**Acknowledgements**

Thanks to the helps of Shaoqi Li, Xianghui Xiao, Xiaoying Deng, Senmiao Fan, Jingtao Pan, Xiao Jiang, Zhibin Zhang, Liqiang Fan, Kang Lei, Ruixian Liu, Xianyan Zou, Zhen Zhang in this study.

**Contribution to the Field Statement**

Cotton is one of the most important fiber crops in the world. With the remarkable improvement of living standards and the rapid development of cotton textile industry, new requirements have been put forward for cotton production, not only for high cotton yield, but also for cotton fiber quality, requiring long, strong and fine cotton fiber. During the cotton fiber development, fiber strength plays an important role in determining the quality of fiber. The research on this aspect is of great significance to improve the strength of fiber, also plays an important role in improving the quality of fiber. It is a hot topic in cotton research. Using BSR-Seq method and introgressive population, we can quickly and effectively find genes related to fiber strength, and have certain predictability for selection. It was
important for us to improve fiber strength.

References

1. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 25, 3389-3402. doi: 10.1093/nar/25.17.3389

2. Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., and Magrane, M. (2004). UniProt: the universal protein knowledgebase. *Nucleic acids research* 32, D115-D119. doi: 10.1093/nar/gky092

3. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., and Eppig, J.T. (2000). Gene ontology: tool for the unification of biology. *Nature genetics* 25, doi: 10.1038/75556

4. Bian, J.M., Jiang, L., Liu, L.L., Wei, X.J., Xiao, Y.H., Zhang, L.J., Zhao, Z.G., Zhai, H.Q., and Wan, J.M. (2010). Construction of a new set of rice chromosome segment substitution lines and identification of grain weight and related traits QTLs. *Breeding Science* 60, 305-313. doi: 10.1270/jsbbs.60.305

5. Bradow, J.M., and Davidonis, G.H. (2000). Quantitation of fiber quality and the cotton production-processing interface: a physiologist’s perspective. *Cotton Sci* 4, 34-64.

6. Claudia, R., Christiane, R., John, G., Oscar, V., Diter, V.W., and Steffen, R. (2013). Three proteins mediate import of transit sequence-less precursors into the inner envelope of chloroplasts in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America* 110, 19962-
7. Curien, G., Giustini, C., Montillet, J.-L., Mas-Y-Mas, S., Cobessi, D., Ferrer, J.-L., Matringe, M., Grechkin, A., and Rolland, N. (2016). The chloroplast membrane associated ceQORH putative quinone oxidoreductase reduces long-chain, stress-related oxidized lipids. *Phytochemistry* **122**, 45-55. doi: 10.1016/j.phytochem.2015.11.015

8. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21. doi: 10.1093/bioinformatics/bts635

9. Du, H., Zhu, J., Su, H., Huang, M., Wang, H., Ding, S., Zhang, B., Luo, A., Wei, S., and Tian, X. (2017). Bulked segregant RNA-Seq reveals differential expression and SNPs of candidate genes associated with Waterlogging tolerance in maize. *Frontiers in plant science* **8**, doi: 10.3389/fpls.2017.01022

10. Du, X., Huang, G., He, S., Yang, Z., Sun, G., Ma, X., Li, N., Zhang, X., Sun, J., and Liu, M. (2018). Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. *Nature genetics* **50**, doi: 10.1038/s41588-018-0116-x

11. Eshed, Y., and Zamir, D. (1994). A genomic library of Lycopersiconpennellii in L. esculentum: a tool for fine mapping of genes. *Euphytica* **79**, 175-179. doi: 10.1007/BF00022516

12. Fang, X., Liu, X., Wang, X., Wang, W., Liu, D., Zhang, J., Liu, D., Teng, Z., Tan, Z., and Liu, F. (2017). Fine-mapping qFS07.1 controlling fiber strength in upland cotton (Gossypium hirsutum L.). *Theoretical and applied genetics* **130**, 795-806. doi: 10.1007/s00122-017-2852-1

13. Gokani, S., Kumar, R., and Thaker, V. (1998). Potential role of abscisic acid in
cotton fiber and ovule development. *Journal of plant growth regulation* 17, 1-5. doi: 10.1007/PL00007005

14. Grover, C.E., Yu, Y., Wing, R.A., Paterson, A.H., and Wendel, J.F. (2008). A phylogenetic analysis of indel dynamics in the cotton genus. *Molecular Biology and Evolution* 25, 1415-1428. doi: 10.1093/molbev/msn085

15. Haigler, C.H., Betancur, L., Stiff, M.R., and Tuttle, J.R. (2012). Cotton fiber: a powerful single-cell model for cell wall and cellulose research. *Frontiers in plant science* 3, doi: 10.3389/fpls.2012.00104

16. Hou, X., Guo, Q., Wei, W., Guo, L., Guo, D., and Zhang, L. (2018). Screening of genes related to early and late flowering in tree peony based on bulked segregant RNA sequencing and verification by quantitative real-time PCR. *Molecules* 23, doi: 10.3390/molecules23030689

17. Joke, R., Peter, D.R., Hui, Z., Anthony, L., Dominiek, S., John, C., Peter, V.L., Maarten, V.D.B., Kirsten, C., and Bernard, S. (2012). Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nature Biotechnology* 30, 61-68. doi: 10.1038/nbt.2053

18. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., and Tokimatsu, T. (2007). A Study on the “Hand Work” Education at the Kyoto City Kindergarten in the Meiji Era : The Case Study of Ryuchi Kindergarten. *Bulletin of the Faculty of Letters Aichi Prefectural University* 56, 53-72. doi:1093/nar/gkm882

19. Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular biology and evolution* 33, 1870-1874. doi: 10.1093/molbev/msw054

20. Kurek, I., Kawagoe, Y., Jacob-Wilk, D., Doblin, M., and Delmer, D. (2002).
Dimerization of cotton fiber cellulose synthase catalytic subunits occurs via oxidation of the zinc-binding domains. *Proceedings of the National Academy of Sciences* 99, 11109-11114. doi: 10.1073/pnas.162077099

21. Lee, J.J., Woodward, A.W., and Chen, Z.J. (2007). Gene expression changes and early events in cotton fibre development. *Annals of botany* 100, 1391-1401. doi: 10.1093/aob/mcm232

22. Leng, N., Dawson, J.A., Thomson, J.A., Ruotti, V., Rissman, A.I., Smits, B.M., Haag, J.D., Gould, M.N., Stewart, R.M., and Kendziorski, C. (2013). EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* 29, 1035-1043. doi: 10.1093/bioinformatics/btt087

23. Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R.J., Ma, Z., Shang, H., Ma, X., and Wu, J. (2015). Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. *Nature biotechnology* 33, doi: 10.1038/nbt.3208

24. Li, F., Fan, G., Wang, K., Sun, F., Yuan, Y., Song, G., Li, Q., Ma, Z., Lu, C., and Zou, C. (2014). Genome sequence of the cultivated cotton Gossypium arboreum. *Nature genetics* 46, doi: 10.1038/ng.2987

25. Li, P.-T., Wang, M., Lu, Q.-W., Ge, Q., Liu, A.-Y., Gong, J.-W., Shang, H.-H., Gong, W.-K., Li, J.-W., and Song, W.-W. (2017). Comparative transcriptome analysis of cotton fiber development of Upland cotton (Gossypium hirsutum) and Chromosome Segment Substitution Lines from G. hirsutum × G. barbadense. *BMC genomics* 18, doi: 10.1186/s12864-017-4077-8

26. Li, S.-Q., Liu, A.-Y., Kong, L.-L., Gong, J.-W., Li, J.-W., Gong, W.-K., Lu, Q.-W., Li, P.-T., Ge, Q., and Shang, H.-H. (2019). QTL mapping and genetic effect of chromosome segment substitution lines with excellent fiber quality from
Gossypium hirsutum x Gossypium barbadense. *Molecular Genetics and Genomics*, 1-14. doi: 10.1007/s00438-019-01566-8

27. Liu, S., Yeh, C.-T., Tang, H.M., Nettleton, D., and Schnable, P.S. (2012). Gene mapping via bulked segregant RNA-Seq (BSR-Seq). *PloS one* 7, doi: 10.1371/journal.pone.0036406

28. Liu, S., Zhou, R., Dong, Y., Li, P., and Jia, J. (2006). Development, utilization of introgression lines using a synthetic wheat as donor. *Theoretical and Applied Genetics* 112, 1360-1373. doi: 10.1007/s00122-006-0238-x

29. Livak, K.J., and Schmittgen, T.D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2− ΔΔCT method. *methods* 25, 402-408. doi: 10.1006/meth.2001.1262

30. Lu, M.-Y., Li, X.-H., Shang, A.-L., Wang, Y.-M., and Xi, Z.-Y. (2012). Characterization of a set of chromosome single-segment substitution lines derived from two sequenced elite maize inbred lines. *Maydica* doi: 10.3417/2009120

31. Lu, Q., Shi, Y., Xiao, X., Li, P., Gong, J., Gong, W., Liu, A., Shang, H., Li, J., and Ge, Q. (2017). Transcriptome analysis suggests that chromosome introgression fragments from sea island cotton (Gossypium barbadense) increase fiber strength in upland cotton (Gossypium hirsutum). *G3: Genes, Genomes, Genetics* 7, 3469-3479. doi: 10.1534/g3.117.300108

32. Luan, M., Guo, X., Zhang, Y., Yao, J., and Chen, W. (2009). QTL mapping for agronomic and fibre traits using two interspecific chromosome substitution lines of Upland cotton. *Plant breeding* 128, 671-679. doi: 10.1111/j.1439-0523.2009.01650.x

33. Ma, Z., He, S., Wang, X., Sun, J., Zhang, Y., Zhang, G., Wu, L., Li, Z., Liu, Z.,
and Sun, G. (2018). Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nature genetics* 50, doi: 10.1038/s41588-018-0119-7

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

35. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* 5, doi: 10.1038/nmeth.1226

36. Paterson, A.H., Wendel, J.F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., Llewellyn, D., Showmaker, K.C., Shu, S., and Udall, J. (2012). Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. *Nature* 492, doi: 10.1038/nature11798

37. Ramsey, J., and Berlin, J. (1976). Ultrastructure of early stages of cotton fiber differentiation. *Botanical Gazette* 137, 11-19. doi: 10.2307/2473754

38. *Gazette* 137, 11-19. doi: 10.2307/2473754

39. Schubert, A., Benedict, C., Berlin, J., and Kohel, R. (1973). Cotton fiber development-kinetics of cell elongation and secondary wall thickening. *Crop Science* 13, 704-709. doi: 10.2135/cropsci1973.0011183x001300060035x

40. Shi, Y., Li, W., Li, A., Ge, R., Zhang, B., Li, J., Liu, G., Li, J., Liu, A., and 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. doi: 10.1111/jipb.12288

41. Song, W., Wang, M., Su, W., Lu, Q., Xiao, X., Cai, J., Zhang, Z., Li, S., Li, P., and
Gong, J. (2017). Genetic and phenotypic effects of chromosome segments introgressed from Gossypium barbadense into Gossypium hirsutum. *PloS one* 12, doi: 10.1371/journal.pone.0184882

42. Stéphane, M., Daniel, S., Myriam, F., Didier, G., Jér?Me, G., Jacques, J., and Norbert, R. (2002). Non-canonical transit peptide for import into the chloroplast. *Journal of Biological Chemistry* 277, 47770-47778. doi: 10.1074/jbc.M207477200

43. Stelly, D., Saha, S., Raska, D., Jenkins, J., Mccarty, J., and Gutierrez, O. (2005). Registration of 17 upland (*Gossypium hirsutum*) cotton germplasm lines disomic for different *barbadense* chromosome or arm substitutions. *Crop Science* 45, 2663-2665. doi:10.2135/cropsci2004.0642

44. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* 28, doi: 10.1038/nbt.1621

45. Trick, M., Adamski, N.M., Mugford, S.G., Jiang, C.-C., Febrer, M., and Uauy, C. (2012). Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploid wheat. *BMC plant biology* 12, doi: 10.1186/1471-2229-12-14

46. Ulloa, M., Saha, S., Jenkins, J., Meredith Jr, W., Mccarty Jr, J., and Stelly, D. (2004). Chromosomal assignment of RFLP linkage groups harboring important QTLs on an intraspecific cotton (*Gossypium hirsutum*) joinmap. *Journal of Heredity* 96, 132-144. doi: 10.1093/jhered/esi020

47. Wang, M., Tu, L., Yuan, D., Zhu, D., Shen, C., Li, J., Liu, F., Pei, L., Wang, P.,
and Zhao, G. (2019). Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nature genetics* 51, doi: 10.1038/s41588-018-0282-x

48. Wang, R., Sun, L., Bao, L., Zhang, J., Jiang, Y., Yao, J., Song, L., Feng, J., Liu, S., and Liu, Z. (2013a). Bulk segregant RNA-seq reveals expression and positional candidate genes and allele-specific expression for disease resistance against enteric septicemia of catfish. *BMC genomics* 14, doi: 10.1186/1471-2164-14-929

49. Wang, S., Chen, J., Zhang, W., Hu, Y., Chang, L., Fang, L., Wang, Q., Lv, F., Wu, H., and Si, Z. (2015). Sequence-based ultra-dense genetic and physical maps reveal structural variations of allopolyploid cotton genomes. *Genome biology* 16, doi: 10.1186/s13059-015-0678-1

50. Wang, W., He, Q., Yang, H., Xiang, S., Zhao, T., and Gai, J. (2013b). Development of a chromosome segment substitution line population with wild soybean (*Glycine sojaSieb. et Zucc.*) as donor parent. *Euphytica* 189, 293-307. doi: 10.1007/s10681-012-0817-7

51. Wendel, J.F., and Albert, V.A. (1992). Phylogenetics of the cotton genus (Gossypium): character-state weighted parsimony analysis of chloroplast-DNA restriction site data and its systematic and biogeographic implications. *Systematic Botany*, 115-143. doi: 10.2307/2419069

52. Yates, S.A., Swain, M.T., Hegarty, M.J., Chernukin, I., Lowe, M., Allison, G.G., Ruttink, T., Abberton, M.T., Jenkins, G., and Skøt, L. (2014). De novo assembly of red clover transcriptome based on RNA-Seq data provides insight into drought response, gene discovery and marker identification. *BMC genomics* 15, doi: 10.1186/1471-2164-15-453
53. Zhang, J.-F., Dan, Y.-Z., Gu, Y.-J., Zhang, B.-C., Li, J.-W., Gong, J.-W., Liu, A.-Y.,
Shang, H.-G., Gong, M.-K., and Yuan, Y.-L. (2012). Evaluation of yield and fiber
quality traits of chromosome segment substitution lines population (BC5F3 and
BC5F34) in cotton. *Journal of Plant Genetic Resources* 13, 773-781. doi:
http://dx.doi.org/

54. Zhang, J., Percy, R.G., and Mccarty, J.C. (2014). Introgression genetics and
breeding between Upland and Pima cotton: a review. *Euphytica* 198, 1-12. doi:
10.1007/s10681-014-1094-4

55. Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., Zhang, J., Saski, C.A.,
Scheffler, B.E., and Stelly, D.M. (2015). Sequencing of allotetraploid cotton
(*Gossypium hirsutum* acc. TM-1) provides a resource for fiber improvement.
*Nature biotechnology* 33, 531. doi: 10.1038/nbt.3207

56. Zhang, Z.-S., Hu, M.-C., Zhang, J., Liu, D.-J., Zheng, J., Zhang, K., Wang, W.,
and Wan, Q. (2009). Construction of a comprehensive PCR-based marker
linkage map and QTL mapping for fiber quality traits in upland cotton
(*Gossypium hirsutum*). *Molecular Breeding* 24, 49-61. doi: 10.1007/s11032-
009-9271-1

**Figures**

![Figure 1](image)

**Figure 1**

Based on the reference genome TM-1, a large number of differentially expressed
Having performing the hierarchical clustering analysis on DEGs from two pairwise
GO enrichment analysis was subsequently conducted on the whole DEGs from twc
At 20 DPA, DEGs in two parent groups (Figure 4A) were mostly annotated into plant hormone signal transduction pathways.

Only Gh_A07G0831 showed non-significant difference on the expression levels between the two parents.

All the identified SNPs were quantified with SNP-index, of which the information was synthetically compared between high and low fiber-strength pools to calculate and plot the Δ(SNP-index) against the TM-1 reference genome positions.

GO enrichment and KEGG pathway analyses were conducted on the whole annotated genes, and the most genes were found to be involved in protein processing in endoplasmic reticulum (ko04141) as the top 4 pathways with the greatest number of genes (Figure 7B).

The coding sequence (CDS) length was 807 bp both in CCRI36 and Hai1, while 2 SNPs were identified to separate them from G. hirsutum, G. arimondii, and G. arboreum with 80% identity and 97%, 95%, and 95% similarities (Figure 8B).