Genomic and GWAS analyses demonstrate phylogenetic relationships of *Gossypium barbadense* in China and selection for fibre length, lint percentage and *Fusarium wilt* resistance

Nan Zhao1,†, Weiran Wang2,†, Corrinne E. Grover3,†, Kaiyun Jiang1, Zhanxia Pan4, Baosheng Guo5, Jiahui Zhu2, Ying Su1, Meng Wang2, Hushuai Nie1, Li Xiao2, Anhui Guo1, Jing Yang2, Cheng Cheng1, Xinmin Ning2, Bin Li1, Haijiang Xu2, Daniel Adjibolosoo1, Alifu Aierxi2, Pengbo Li4, Junyi Geng5, Jonathan F. Wendel3,* and Jie Kong2,* and Jinping Hua1,*

1Joint Laboratory for International Cooperation in Crop Molecular Breeding, Ministry of Education/College of Agronomy and Biotechnology, China Agricultural University, Beijing, China
2Institute of Economic Crops, Xinjiang Academy of Agricultural Sciences, Xinjiang, China
3Department of Ecology, Evolution and Organismal Biology, Iowa State University, Ames, IA, USA
4Institute of Cotton Research, Shanxi Agricultural University, Shanxi, China
5Cotton Research Institute, Hebei Academy of Agriculture and Forestry Sciences, Hebei, China

Received 1 July 2021; revised 22 October 2021; accepted 27 October 2021.
*Correspondence (Tel +86-10-62734748; fax +86-10-62734748; email jeping_hua@cau.edu.cn (J.H.); Tel +86-991-4504151; fax +86-991-4504151; email kongjie.258@163.com (J.K.); Tel +1-515-294-7172; fax +1-515-294-1337; email jfw@iastate.edu (J.F.W.))
†These authors are contributed equally.

Summary

Sea Island cotton (*Gossypium barbadense*) is the source of the world’s finest fibre quality cotton, yet relatively little is understood about genetic variations among diverse germplasms, genes underlying important traits and the effects of pedigree selection. Here, we resequenced 336 *G. barbadense* accessions and identified 16 million SNPs. Phylogenetic and population structure analyses revealed two major gene pools and a third admixed subgroup derived from geographical dissemination and interbreeding. We conducted a genome-wide association study (GWAS) of 15 traits including fibre quality, yield, disease resistance, maturity and plant architecture. The highest number of associated loci was for fibre quality, followed by disease resistance and yield. Using gene expression analyses and VIGS transgenic experiments, we confirmed the roles of five candidate genes regulating four key traits, that is disease resistance, fibre length, fibre strength and lint percentage. Geographical and temporal considerations demonstrated selection for the superior fibre quality (fibre length and fibre strength), and high lint percentage in improving *G. barbadense* in China. Pedigree selection breeding increased *Fusarium wilt* disease resistance and separately improved fibre quality and yield. Our work provides a foundation for understanding genomic variation and selective breeding of Sea Island cotton.

Keywords: *Gossypium barbadense*, GWAS, fibre quality, yield, *Fusarium wilt* disease resistance.

Introduction

Cotton (*Gossypium* spp.) production accounts for a majority of natural textile fibres produced worldwide (Zhang et al., 2014). While cotton has been domesticated independently four different times on two different continents, it is the two cultivated polyploid species (i.e. *G. hirsutum*, AD1, and *G. barbadense*, AD2) (Grover et al., 2020; Wendel and Grover, 2015) from Central and Northern South America that predominate in modern cotton commerce. These species are derived from a single allopolyplidization event approximately 1.5 million years ago that subsequently radiated into the seven known polyploid species (Wang et al., 2018). One of the polyploid species derived from this event, that is *G. barbadense*, is well known for its excellent fibre quality (Wang et al., 2019), particularly its superior extra-long fibres (Yu et al., 2013). Increasing demand for high-quality textiles has generated interest in understanding the genetics controlling fibre-related traits, particularly in Sea Island cotton, with the ultimate goal of genome-assisted breeding.

Both *G. hirsutum* and *G. barbadense* are allopolyploids derived from the union of two diploid genomes, A and D. The rapid development and application of genome sequencing technology to *Gossypium* have generated numerous insights into cotton genomics. The Peruvian diploid *G. raimondii* (D0) was the first cotton genome to be sequenced (Paterson et al., 2012; Wang et al., 2012), followed by genome assemblies of some (Udall et al., 2019) and resequencing of all 13 D-genome species (Grover et al., 2019). Similarly, genome assemblies and resequencing data sets have been published for the A-genome diploids, *G. arboreum* (A0) (Du et al., 2018; Huang et al., 2020; Li et al., 2014) and *G. herbaceum* (A1) (Huang et al., 2020). Genomic resources are also available for the allopolyploids, including nine genome assemblies of *Gossypium hirsutum* (AD1) genome (Chen et al., 2020; Hu et al., 2019; Huang et al., 2020; Li et al., 2015; Wang et al., 2019; Yang et al., 2019; Zhang et al., 2015) and four of *G. barbadense* (AD2) (Chen et al., 2020; Hu et al., 2019; Wang et al., 2019; Yuan et al., 2015), as well as thousands of resequenced accessions from both species
Upland cotton (Gossypium hirsutum) has been a focus of agronomic trait selection in crop plants (Huang and Han, 2017). Association studies (GWAS) have identified disease resistance-related pathways, including glutathione metabolism, glycolysis, plant hormone signal transduction, anthocyanin biosynthesis, and butanoate metabolism, using transcriptome sequencing of four different Sea Island cotton RILs with susceptible, highly susceptible, highly resistant, or super highly resistant phenotypes. Su et al. (2020) screened an E3 ubiquitin-protein ligase gene, GB_A03D0G33, which was correlated with fibre length and strength, via GWAS analyses using 6309 SNPs from 279 Sea Island cotton accessions. Yu et al. (2021) identified three fibre strength candidate genes, H0D1 (GB_D17G3437), encoding casein kinase I isoform delta-like protein, WDL2 (GB_D17G3460), encoding WDLI/WDVD-LIKE 1 protein, and TUBA1 (GB_D11G3471), encoding tubulin alpha-1 chain protein, and one lint percentage candidate gene, HERK1 (GB_A07G1034), encoding receptor-like protein kinase regulated by brassinosteroids and required for cell elongation) by GWAS of 240 Sea Island cotton accessions. Lu et al. (2017) identified three fibre strength-related genes, XLOC_025331, encoding mannosyl-oligosaccharide-a-mannosidase), XLOC_0259945 (FL48), and XLOC_025327 (snakin-1), using transcriptome analyses on three chromosome segments substitution lines (CSSLs) derived from CCRi45 (G. hirsutum, recurrent parent) x Hai1 (G. barbadense). Four years later, Li et al. (2021) revealed another eight fibre-related genes that they separately encoded O-fucosyltransferase family protein (GB_A02G0240), glutamine synthetase 2 (GB_A02G027), Ankyrin repeat family protein (GB_A02G0264), beta-6 tubulin (GB_D03G1742), WRKY DNA-binding protein 2 (GB_D03G1655), quinolinate synthase (GB_D07G0623), nuclear factor Y subunit B13 (GB_D07G0631) and leucine-rich repeat transmembrane protein kinase (GB_D07G0797), using the same materials and similar methods.

Here, we gained insights into the genotype-to-phenotype associations in G. barbadense using GWAS and a broad diversity of Sea Island cotton accessions, including a focus on lineages from China. We resequenced and phenotyped 336 Sea Island cotton accessions grown across 6 years and 4 locations to provide information on both genomic and phenotypic variations. Using these data, we performed GWAS for 15 important agronomic traits, including fibre quality, yield, resistance to Fusarium wilt disease, maturity-related traits and plant architecture, confirming the roles of five candidate genes responsible for four key traits using RNA-seq and transgenics. Finally, we explored the genetic basis for the improvement of Sea Island cotton through pedigreed analysis, identifying elite genes involved in fibre length and lint percentage. Our results provide a foundation for Sea Island cotton improvement and molecular perspectives into cotton breeding.

Results

Genomic variation and population structure

We generated 9.4 Tb high-quality resequencing data involving 336 accessions derived from Asia (274), Africa (32), the Americas (28) and Europe (2) (Table S1 and Figure S1). Approximately 98.77% of reads covered 97.09% of the reference genome (Wang et al., 2019), with an average of 11.2-fold depth (Tables S1 and S2). We identified 16.0 million (M) single nucleotide polymorphisms (SNPs; Table 1 and Table S3) and 2.3 M insertion/deletion polymorphisms (InDels; Table S4). We found that the number of SNPs in At was approximately 1.8 times that found in Dt (Table S5 and Figure S2a), congruent with the twofold size difference between the At and Dt subgenomes (Li et al., 2014). SNP density was 1.8 SNPs/kb in At and 1.9 SNPs/kb in Dt.
in Dt (Table S5 and Figure S2a), like that in G. hirsutum (Ma et al., 2018). Diversity (θs) within the two subgenomes was similar, albeit slightly lower for Dt (5.3 × 10^{-5}) than At (6.2 × 10^{-5}) (Table S5 and Figure S2a), which agrees with a recent report (Yuan et al., 2021).

For structure analysis, the natural logarithms of probability data (LnPK) and the ad hoc statistic ΔK were calculated (Dong et al., 2019; Huang et al., 2017; Su et al., 2018). The LnPK(Δ) value increased continuously from K = 1 to 7 without an obvious inflection point (Figure S2b). However, the ΔK value showed a spike at K = 2 (Figure S2c). This suggested two major gene pools, consistent with the phylogenetic tree (Figure 1a), population structure analysis (Figure 1b and Table S6) and principal component analysis (PCA; Figure S2d). Given intraspecies introgression due to geographic distribution and breeding practice, some landraces and transitional accessions were integrated into a third mixed subgroup by increasing a new middle level of ancestry proportion at K = 2 (at first, when the ancestry proportion of one accession belonging to K1 was over 0.7, it was categorized as K2; otherwise pop2, and then, accessions with the ancestry proportion from 0.5 to 0.7 were assigned into the mixed subgroup; Figure 1c, Figure S2e–f and Table S7). Hereafter, these subgroups were designated as ‘Pop1’ (76), ‘mixed’ (91) and ‘Pop2’ (169 accessions; Table S6 and Figure S1). Pop1 primarily included recently selected cultivars from China’s northwest inland cotton region, with longer and stronger fibres (fibre length, FL included recently selected cultivars from China’s northwest inland ‘Pop2’ (169 accessions; Table S6 and Figure S1). Pop1 primarily these subgroups were designated as ‘Pop1’ (76), ‘mixed’ (91) and subpopulation; Figure 1c, Figure S2e

| SNP category                  | No. of SNPs |
|-------------------------------|-------------|
| Upstream                      | 335,871     |
| Exonic                        |             |
| Stop gain                     | 5057        |
| Stop loss                     | 514         |
| Synonymous                    | 86,767      |
| Nonsynonymous                 | 160,616     |
| Intronic                      | 524,253     |
| Splicing                      | 1666        |
| Downstream                    | 270,135     |
| Upstream/downstream           | 25,959      |
| Intergenic                    | 14,429,742  |
| Ts                             | 10,789,456  |
| Tv                             | 5,242,423   |
| Ts/Tv                         | 2           |
| Total                         | 16,031,879  |

The decay rate of linkage disequilibrium (LD), that is the pairwise correlation coefficient (r²) from the maximum value to the half-maximum, was 388 kb for all 336 accessions and was close among populations (i.e. 373, 342 and 342 kb for Pop1, mixed and Pop2 respectively; Figure 1e). These LD values were higher than that of Upland cotton reported by Wang (296 kb; Wang et al., 2017a), but lower than that of Fang (1000 kb; Fang et al., 2017b).

**Genome-wide association studies identification**

We measured 15 traits (Table 2), including fibre quality (five), yield components (five), disease resistance (one), maturity (two) and plant architecture (two; Table 2), for the 336 Sea Island cotton accessions grown across four locations over six years (Table S8). Diverse phenotypic variations were observed for all traits (Table 2 and Table S9). Two of the fibre quality traits important for the spinning industry, FL and FS, were positively correlated with six traits FU, BN, FBN, SBW, SI and PH, while also being negatively associated with the other seven traits (i.e. FM, FE, LP, DP, GP, FNB and FBT; Figure S3). Using 4.1 M high-quality SNPs, we performed GWAS for these 15 traits. These analyses revealed 6,241 unique SNPs, including 437 same SNPs among different traits (Figure S4–S18, Table S10 and S11). The number of significant SNP varied among traits, for those were selected and improved by emphasis in our population, and the numbers of significantly associated SNPs were relatively large. For example, DP had the greatest number of associated SNPs, followed by two fibre quality traits (i.e. FM and FS). For traits that the selection and improvement degree were relatively low, their numbers of significant SNPs were small, such as LP, FL, SI, FBT, SBW and so on. Generally, the effective candidate regions with significant GWAS signals were defined as the LD blocks surrounding the signal peak (Yano et al., 2016), although these were sometimes enlarged slightly when candidate genes could not be identified (Fang et al., 2017b). Based on the 388 kb LD decay distance and candidate gene analysis, we defined 500 kb (slightly larger than 388 kb) upstream and downstream of a significant SNP signal peak (i.e. totally 1-Mb interval; Fang et al., 2017b) as the candidate region size and found 18 696 unique genes, involving in 6183 common genes related to at least two traits (Table S10 and S12). The total number of associated genes was highest for the fibre quality category, followed by maturity. From these, we chose key genes related to four agronomically important traits for further functional verification.

**Fibre length**

On chromosome A05, we identified one nonsynonymous SNP, within the candidate region located at 16.28–16.30 Mb, significantly correlated with fibre length (Figure 2a,b). The gene containing this SNP, Gbar_A05G017500, encoded a predicted U-box domain-containing E3 ubiquitin ligase (PUB4), named after FIBER LENGTH2 (GbFL2). The phenotypically associated SNP (16286973) resulted in a T/G transversion, leading to leucine (L) or valine (V) (Figure 2c), which was associated with either longer (T) or shorter (G) fibre respectively (Figure 2d). While most of the early introduced varieties in Pop2 had the long-fibre haplotype (T; Figure 2e), the proportion of short-fibre haplotypes (G) raise up to near equivalence in the ‘mixed’ population (Figure 2e), perhaps due to linkage drag associated with selection on other traits. The long-fibre (T) haplotype gained prominence again in Pop1, comprising 83.33% of haplotypes for this locus (Figure 2e).
Expression of *GbFL2* gradually decreased during fibre development (from 0 DPA to 20 DPA) and was lower in long-fibre varieties (Figure 2f). We validated expression pattern of *GbFL2* using qRT-PCR in FL extreme accessions (Figure 2g), namely, a negative regulation pattern. VIGS transformation of *GbFL2* in high and low FL lines showed increased fibre length relative to the wild type (Figure 2h-i), supporting the role of *GbFL2* in fibre elongation.

*GbFL2* is derived from the At chromosomes (i.e. A05) of AD 2 (Figure 2j), and the change in haplotype frequency during breeding is suggestive of directional selection during domestication (Figure 2k and Table S14).

**Fibre strength**

On chromosome D11, we identified one nonsynonymous SNP significantly correlated with fibre strength in the candidate locus at 64.20–64.25 Mb (Figure S19a-b). The sole gene contained within this locus, *Gbar_D11G032670*, encoded a putative casein kinase 1-like protein (HD16), named after *FIBER STRENGTH1* (*GbFS1*). The two alleles (C/T) detected at this position (D11:64227153) encoded two different amino acids, threonine (T) and isoleucine (I; Figure S19c), corresponding to low (C haplotype) and high fibre strength (T haplotype; Figure S19d). The early introduced varieties from Pop2 had the high-strength fibre haplotype (T), and the change in allele frequency hinted at directional selection during breeding in China (vs mixed and Pop1; Figure S19e). Of the 159 Chinese accessions, 144 contained the T/G haplotypes for *GbFL2* and T/C for *GbFS1* (the remaining 15 were missing information/nucleotides or had unique mutations; Figure S20 and Table S13). Among those 144 accessions, 41 accessions had the long/high-strength haplotype combination (TT), 35 exhibited short/low strength (GC), 58 had long/low strength (TC) and 10 exhibited short/high strength (GT). This suggested that although fibre length and strength were often regarded as the simultaneous targets of selection, Sea Island cotton breeding in China might have favoured fibre length as a priority (99 versus 45 accessions; Figure S19e).

*GbFS1* was highly expressed at most fibre developmental stages (5-20 DPA; Figure S19f) in low strength accessions, implying a negative regulation pattern (Figure S19f-g).

*GbFS1* was derived from the Dt subgenome (i.e. chromosome D11) of AD2, having been inherited from the D-genome ancestor (represented by the D 5 genome, Figure S19h), which is notable in that D-genome species have short, non-spinnable fibres. Interestingly, Dt homeolog of *GbFS1* showed directional selection in AD2 relative to their AD1 counterpart, suggesting selection of this advantageous mutation in Sea Island cotton (Figure S19h–i and Table S14).

**Lint percentage**

On chromosome A05, we also identified a strong signal associated with lint percentage (Figure 3a). The gene closest to this region (13.00–13.20 Mb), *Gbar_A05G014160*, had one nonsynonymous SNP (A05:13046765), that is a C/G transversion, resulting in an amino acid difference, that is either alanine (A) or glycine (G), the latter of which was associated with a significant improvement in lint percentage (Figure 3b–d). We designated the locus containing this gene *LINT PERCENTAGE*.
asparagine (N) (Figure S21c) and a serine (S) in their coding sequences, which resulted in a lysine (K) and a putative wall-associated receptor kinase-like 14 (WAKL14). Both GbDP1 and GbDP2 encoded a putative ATP-dependent RNA helicase (GbLP1), which encoded a putative ATP-dependent RNA helicase (DEAH12) targeted to the chloroplast. This putative protein contained a RING-type zinc-finger domain, a characteristic of the E3 ubiquitin ligase RBR family. Unlike GbFL2 and GbFS1, the number of accessions with high-lint-percent (haplotype (G) had generally been decreasing gradually since the first introduction of Sea Island cotton into China; however, the ratio of G/C fluctuated, first increasing and then decreasing (Figure 3e) in later accessions, as the need for greater lint production was balanced with fibre quality. High expression of GbLP1 occurred at the whole fibre developmental stages, especially at 0 and 5 DPA (Figure 3f), consistent with qRT-PCR validation in LP extreme accessions (Figure 3g). VIGS transformation in LP extreme Sea Island cotton lines with matching G/C haplotypes showed decreased lint percentage (Figure 3h), confirming the role of GbLP1 in lint-percentage forming. Like GbFL2, GbLP1 showed signatures consistent with positive selection (i.e. A2, Figure 3i–j and Table S14), while its homeolog (here, in the D genome) exhibits patterns consistent with purifying selection (Figure 3i,j and Table S14). Notably, most accessions had alleles conferring long fibre but with low strength and low lint percentage, followed by accessions exhibiting high lint percentage but with short fibre and low strength, implying the strongest directional selection was on long fibre, followed by high lint-percentage (Figure S20b and Table S13).

Fusarium wilt resistance

Cotton fusarium wilt disease, caused by the fungus Fusarium oxysporum f. sp. vasinfectum (FOV), is one of the most significant diseases impacting yield in G. barbadense. Here, we revealed a strong association signal cluster on chromosome D03 related to FOV disease percentage (DP) (Figures S21a; Figure 4a). We screened two closely linked candidate genes from this cluster, that is Gbar_D03G001430 (henceforth GbDP1) and Gbar_D03G001910 (GbDP2), at 0.8–1.0 Mb and 1.5–1.6 Mb respectively (Figure S21b; Figure 4b). GbDP1 encoded a putative zinc-finger homeodomain retein 6 (ZHD6), and GbDP2 encoded a putative wall-associated receptor kinase-like 14 (WAKL14). Both GbDP1 and GbDP2 had a nonsynonymous A/C transversion in their coding sequences, which resulted in a lysine (K)–asparagine (N) (Figure S21c) and a serine (S)–arginine (R) shift (Figure 4c), respectively, whose close linkage generally resulted in two haplotypes (AA and CC). Accessions carrying the CC-haplotype showed significantly lower disease percentage than those with the AA-haplotype (Figures S21d; Figure 4d). Most early introduced varieties had the high-disease-percentage haplotype (AA) (Figure S21e; Figure 4e); however, there had been 20 recently selected cultivars with low-disease-percentage haplotype (CC) in Xinjiang (in pop1) (Figures S20e; Figure 4f). Among the 178 Chinese Sea Island containing GbDP1 and/or GbDP2, 164 (92%) contained AA (144) or CC (20) (Figure S20c and Table S13). Both GbDP1 and GbDP2 exhibited high expression in susceptible (S) lines after FOV inoculation (Figures S21f; Figure 4f), implying a negative regulation pattern. GbDP1/2-silencing in susceptible lines (S_pCLCtVA-DP1/S_pCLCtVA-DP2) conferred increased resistance to FOV infection relative to empty-vector-carrying (S_pCLCrVA) and wild-type cotton susceptible lines (S_WT) (Figure S21-h and Figure 4g,h). These results suggested that GbDP1 and GbDP2 were two potential targets for conferring FOV resistance in G. barbadense. GbDP1 orthologs in AD2, AD1 and D6 were nearly identical but differed from the D5 ortholog by two SNPs in a 423-bp conserved region (Figure S21i-j and Table S14). This might reflect introgression from D6 into the AD1-AD2 lineage or, more likely, autapomorphic changes in the D6 lineage after divergence from the (now extinct) D-genome donor parents of the allotetraploids. For GbDP2, the case was more complex, and there were four kinds of variations in a 1509-bp conserved region among the AD2, AD1, D5 and D6 homologues: (i) the homologues in AD2 had three specific SNPs compared to those in AD1, D5 and D6; (ii) the homologues in AD2 and AD1 had three common SNPs compared to those in D5 and D6; (iii) the homologues in AD2, AD1 and D5 had five common SNPs compared to those in D6; (iv) the homologues in AD2, AD1 and D8 had two common SNPs compared to those in D5. These data demonstrated differential SNP introgression from AD1, D5 and D6 into AD2, reflecting differential selection of GbDP2 homologues after allotetraploidy (Figure 4i,j and Table S14). Despite the high level of conservation between AD1-AD2 for these two genes, evolution selection analyses of the Sea Island and Upland cotton D-Homoeologs suggested different histories of selection (Figure 4i,j, Figure S21i, j and Table S14). A simultaneous consideration of all four traits,
that is FL, FS, LP and DP, suggested priority selection on long fibre and high lint percentage once again (Figure S20d and Table S13).

Genomic characterization of a pedigree
We selected an intact pedigree from our GWAS population to examine the origin of two elite cultivars, XH39 and XH60, which were derived from the same initial breeding pool (11 common parents), but were later subjected to selection for longer fibre (in XH39) and high lint percentage (in XH60) respectively (Figures 5a–f and 6a–f). Comparisons among the parents and the two elite lines traced 4.57% and 3.78% of the genome in XH39 and XH60 respectively; through the historical crosses leading through these lineages, most of the traceable transmissions were
found in the At genome (Figures 5a–d, 6a–d and Table S15). We then analysed the overlapping genes between our broader GWAS analysis and those regions whose genetic transmission was traceable across the pedigree. For X3H9, we uncovered 32 and 178 genes controlling fibre length and strength in these overlapping regions (Figure 5g and Table S16), four of which simultaneously impacted FL and F5 (Figure S22a and Table S16). Two of these four genes, \textit{GbarA04G013270} and \textit{GbarA04G013290}, had high expression at 5–25 DPA fibres with a positive-regulation pattern, especially at 10 and 15DPA fibres of long-fibre line (Figure 5h). These were specifically passed from parent 86430, the parental origin of this gene, that is accession JH1, a preponderance of fibres in the early stage (0, 5, 10, 15, 20, 25 DPA), detected by RNA-seq (FPKM value). Data are average values with standard deviation (n = 3 varieties with three technical repeats). Single (**), double (***), and triple (****) asterisks mark statistical significance levels of P < 0.05, 0.01 and 0.001 respectively. (g) qRT-PCR analysis of GBFL2 expression in wild-type (WT), transgenic lines with empty VIGS vector (pCLCrVA) and target gene GBFL2 (pCLCrVA-FL2) of long-fibre (L) accession XHS8 and short-fibre (S) accession Ashi. The gene expression level in the long-fibre accession wild type (L_WT) was set to 1. GBUBQ7 is an internal control. (h) Fibre length (mm) of WT, pCLCrVA, (pCLCrVA-FL2) of long-fibre (L) accession XHS8 and short-fibre (S) accession Ashi. (i) VIGS phenotypes of GBFL2. (j) The evolutionary origin of GBFL2 (\textit{GbarA05G017500}). We built unrooted trees using the maximum-likelihood method in MEGA7, based on complete CDS sequences. (k) Selection analysis on homologous CDS sequences of GBFL2. Homologous sequence in each cotton species is represented by its genome name on the left side of the circle. The difference value (Ka-Ks) of each group of homologous comparisons is indicated by coloured rectangles according to the colour bar in the upper left corner. While Ka/Ks is generally used as an indicator of selective pressure, the presence of ‘Ks’ of these genes. (Figures 5f and 6f, Figure S22c and Table S16). Additional parental history, XH39 and XH60 both contained six, nine and 12 homologous regions (Table S19). Notably, the fibre length candidate gene, \textit{Gbar_A05G017500} (GBFL2), had a homolog in a previously identified fibre-length QTL region (TM10723_TM10747, TM10754; Su et al., 2020), and the fibre strength candidate, \textit{GbarA07G013290} (in CLC1), was contained in a large SNP-cluster on chromosome D11 that was proximal to a previously identified fibre strength loci (loci24; Fang et al., 2021; Table S19). Additional evidence was that a HD16 ortholog \textit{(GB_D11G3437)} on chromosome D11 was found to be associated with fibre strength in 240 \textit{Gossypium barbadense} accessions (Yu et al., 2021); this gene has 100% identity with our GBFS1, with the
same functional annotation, that is casein kinase I (Table S19). Our analysis also yielded additional, previously unrecognized loci associated with fibre quality, yield, and other traits (Table S9–S12), most of which were associated with fibre quality (followed by disease resistance, yield and maturity), and many of which were located on the chromosomes that originated in the non-lint producing parent (i.e. D03, D05, D09, D10, D11). These results were consistent with the previous reports that more loci associated with fibre quality and yield were in the DT than in the At subgenome (Ma et al., 2018b).

GWAS analysis of genes potentially contributing to agronomic traits

Sea Island cotton (Gossypium barbadense) has superior fibre quality, but poor adaption to biotic (e.g. Fusarium wilt disease) and abiotic (e.g. drought and salinity) stresses, as well as low yield, which collectively limit its commercial importance. Solutions for circumventing this productivity bottleneck include improving adaptability and yield of Sea Island cotton or transferring key genes from G. barbadense to Upland cotton. These twin paths lend significance to the goal of mining important genes responsible for fibre quality, yield and disease resistance in Sea Island cotton.

The abundance of data generated here revealed several genes related to key fibre and disease resistance traits in Sea Island cotton, including the underlying allelic and expression variation associated with domestication. We found alleles associated with superior fibre quality (FL and FS) and resistance to Fusarium oxysporum (DP) that included two genes encoding RING-type E3 ubiquitin ligase (E3) protein that had high expression in the early stages of fibre development, and the expression of the promoter of its Arabidopsis homolog, At3g19950, was activated in trichomes (Ho et al., 2010). Recently, another GWAS analysis using Sea Island cotton identified an E3 ubiquitin-protein ligase gene Gb_A03G03355 associated with fibre quality measures (Fang et al., 2021). Together, these studies supported the potential role of GbFL2 as an E3 ubiquitin-protein ligase contributing to fibre quality. The role for the other one, that is the ATP-dependent RNA helicase gene GbLP1 (Gbar_A05G014160), in fibre production was less clear; however, salinity stress suppressed ATP-dependent RNA helicase expression, which also reduced lint percentage (Gong et al., 2017). Because higher expression of GbLP1 was associated with increased lint production, this implied an indirect association between ATP-dependent RNA helicase and cotton lint percentage.

For fibre strength, we identified the candidate gene GbFS1, which encoded a Casein kinase I, a multifunctional protein kinase with serine/threonine protein kinase active sites (Gross and Anderson, 1998). In rice, casein kinase I phosphorylated the DELLA protein SLR1, stabilized SLR1 and negatively regulated gibberellic acid (GA) signalling (Dai and Xue, 2010). In Upland cotton, exogenously applied gibberellic acid (GAs) improved fibre strength in natural-coloured cottons (Zhang et al., 2017). GA promoted secondary cell wall development in cotton fibre cells by regulating expression of sucrose synthase genes (Xiao et al., 2019). We inferred that GbFS1 might regulate fibre strength via the gibberellic acid signalling pathway in Sea Island cotton.

We also detected two genes, GbDP1 and GbDP2, both on the D03 chromosome that appeared to have the role in resistance to Fusarium wilt disease. Recently, another newly identified gene (Gb_D03G02099) on D03 of Upland cotton was shown to affect resistance to FOV (Liu et al., 2021). It may be that the disease resistance networks are diverse, because Gb_D03G02099 encoded a GLUTAMATE RECEPTOR-LIKE (GLR) protein, whereas GbDP1 encoded a Zinc-finger homeodomain protein 6 (ZHD6). Alternatively, these proteins operated on different aspects of the same
network. Zinc-finger homeodomain (ZF-HD) subfamily proteins played specific roles in pathogen signalling and plant defences by activating CaM4 gene expression in response to pathogens (Park et al., 2007). The other defence gene, GbDP2, encoded a wall-associated receptor kinase-like 14 (WAKL14). Many WAKL genes were correlated with plant resistance and immune responses. For instance, CaWAKL08, a pathogen-induced wall-associated receptor-kinase like kinase in sweet orange, conferred resistance to citrus bacterial canker via ROS control and JA signalling (Li et al., 2020). OsWAKL21.2 activated rice immune responses by its kinase activity and Arabidopsis immune responses by its guanylate cyclase activity (Malukani and Ranjan 2020). Although the precise function of our five candidate genes identified here remains unclear, we confirmed their influences on their respective phenotypes using VIGS transgenic experiments in Sea Island cotton here.

Cotton improvement by GWAS and pedigree analysis

Modern Sea Island cotton cultivars principally were derived from three gene pools: Egyptian type, American type and Middle-Asian type (Abdullaev et al., 2017). In Xinjiang, Sea Island cotton varieties were derived from five backbone parents of Central Asia, including 2H3, C6022, 87631F, 5230Φ and 91221H. JH1, a variant with the early maturity of 91221H, was the core germplasm source used to produce more than 50 varieties, including at least 10 main cultivars. Here, we extracted a pedigree composed of 19 varieties, including 91221H and JH1 that were involved in developing the cultivar XH39, which has early maturity, high resistance and superb fibre quality. Our analyses revealed a dN/dS ratio and blocks of low diversity consistent with strong directional selection. The total size and gene number in the low diversity regions were biased towards the Dt subgenomes, supporting the notion that selection for fibre improvement has been asymmetric across genomes (Ma et al., 2019; Wang et al., 2017a).

In Upland cotton, pedigree-based genome resequencing has been an effective tool for researchers to detect candidate genes related to important traits via genetic transmission analysis (Fang et al., 2017; Lu et al., 2019; Ma et al., 2019). In our pedigree, XH39 and its related accession XH60 exhibited divergent improvement trends, with the former focussed on fibre quality (FL and FS), while the latter targeted yield (LP and SI). This divergence in selection might predict different targeted genes, but we found some common genes underlying traits of the same category, for example FLFS, LP and SI. While the genes in some categories, for example GP and DP, likely had no influence on fibre quality or yield, it is possible that some could simultaneously regulated fibre quality (i.e. FS), or yield (i.e. LP, BN) and growth period.

In summary, here we provided a detailed depiction of the Chinese Sea Island cotton gene pool, describing diversity and phylogenetic and population structure. We generated a WGS dataset for the community and demonstrated its utility via a comprehensive GWAS analysis. We further described the inferences of candidate genes to facilitate molecular-marker selection and genetic improvement for great disease resistance, superb fibre quality and high yield of cotton. Pedigree analysis of XH39 and XH60 provided evidences for the basis of increased fibre quality and yield, as well as the improvement to Fusarium wilt disease resistance. Our research laid a foundation for understanding polymorphism in Chinese Sea Island cotton, as well as introgression from other sources and artificial selection. In addition, the key genes identified here for fibre quality, yield and disease resistance can be further explored, for example, to decipher their participations in regulatory networks and the genotype-to-phenotype connections. Elite Chinese Sea Island varieties with excellent haplotype combinations have great agronomic potential for cotton improvement. The present study put forward a significant step to the exploration, understanding and utilization of the broad gene pool.

Methods

Sampling

We collected 336 G. barbadense accessions (including 19 accesses in pedigree analysis) derived from major global cotton-growing countries from seed stocks maintained at China Agricultural University, Beijing. The original diversity was evaluated based on their geographical distribution and breeding history. The geographic origins of these accessions included the major cotton-growing countries, that is China (Northwest Inland Region, Yellow River Basin, Yangtze River Basin, Southwest and
South China), the United States, the former Soviet Union (Uzbekistan, Tajikistan, Turkmenistan and Azerbaijan), Viet Nam, Syria, Antigua, Argentina, West Indies, Peru, Albania, Egypt and Sudan (Figure S1 and Table S1). Additionally, we had seven pair of samples that were initially bred from the same varieties, but they were not exactly genetically identical, because they came from the different lines, so they were not sample duplication in terms of genomic composition (Su et al., 2020); therefore, we labelled the source as cultivars/lines in Table S1.

Planting and phenotyping

Phenotyping of 15 traits was performed across four locations over six years (not four locations × six years, the detailed is in the next paragraph). Three locations were comprised of Yacheng in Hainan (H) Province (Southern China), and Korla (K) and Awat (A) in Xinjiang (Northwest Inland; Table S8). All accessions were planted in an experimental field with an arrangement-order design, including two replicates. Each plot at the H-site contained one row 4 m in length, 11–13 plants per row, ~33 cm between plants within each row and 75 cm between rows. Plot specifications at K and A locations contained 18–20 plants per row 2 m in length, ~11 cm between plants within each row and 66 cm between rows. Cotton was sown in mid-to-late April and was harvested in mid-to-late October in the Xinjiang locations, whereas the cotton was sown in mid-to-late October and was harvested in mid-to-late April in Hainan.

We characterized 15 traits and obtained a total of 119 sets of phenotypes. Nine traits (FL, FS, FM, FU, GP, BW, PH, FNFB and PH) were recorded in nine location×years sets (Table S9). SI, DP and FBT were assessed in six, four and one environment respectively (Table S9). Twenty naturally opened bolls were hand-harvested to calculate the SBW (g) and gin the fibres. SI was obtained after counting and weighing 100 cotton seeds. Fibre samples were separately weighed to calculate LP.

DNA isolation and genome resequencing

The leaves from a single plant of each accession were sampled and used for DNA extraction. Total genomic DNA was extracted with a Plant DNA Mini Kit (Cat # DN1502, Aidlab Biotechnologies, Ltd.), and 350-bp whole-genome libraries were constructed for each accession by random DNA fragmentation (350 bp), terminal repair, PolyA tail addition, sequencing connector addition, purification, PCR amplification and other steps (TrueSeq Library Construction Kit, Illumina Scientific Co., Ltd., Beijing, China). Subsequently, we used the Illumina HiSeq PE150 platform to generate 9.78 Tb raw sequences with 150 bp read length.

Sequencing reads quality checking and filtering

To avoid reads with artificial bias (i.e. low-quality paired reads, which primarily result from base-calling duplicates and adaptor contamination), we removed the following types of reads: (i) reads with ≥10% unidentified nucleotides (N); (ii) reads with adaptor sequences; (iii) reads with >50% bases having Phred quality Q ≤ 5. Consequently, 9.42 Tb high-quality sequences were used in subsequent analyses (Table S1).

Sequencing reads alignment

The remaining high-quality reads were aligned to the genome of *G. barbadense* 3–79 (http://cotton.hzau.edu.cn/EN/download. php; Wang et al., 2019) with BWA software (version: 0.7.8) with the command ‘mem -t 4 -k 32 -M’. BAM alignment files were subsequently generated in SAMTOOLS v1.4 (Li et al., 2009), and duplications were removed with the command ‘samtools rmdup’. Additionally, we improved the alignment performance through (i) filtering the alignment reads with mismatches ≤5 and mapping quality = 0 and (ii) removing potential PCR duplications. If multiple read pairs had identical external coordinates, only the pairs with the highest mapping quality were retained.

Population SNP detection

After alignment, SNP calling on a population scale was performed with the Genome Analysis Toolkit (GATK, version v3.1) with the UnifiedGenotyper method (McKenna et al., 2010). To exclude SNP-calling errors caused by incorrect mapping, only high-quality SNPs (depth ≥ 4 (1/3 of the average depth), map quality ≥20, the missing ratio of samples within the population ≤ of 10% (3,487,043 SNPs) or of 20% (4 052 759 SNPs), and minor allele frequency (MAF) >0.05) were retained for subsequent analyses. SNPs with the missing ratio ≤ of 10% were used in PCA/phylogenetic tree/structure analyses, whereas...
Nan Zhao et al.
SNPs with a missing ratio ≤ of 20% were used in the rest of the analyses.

Annotation of genetic variants

SNP annotation was performed according to the *G. barbadense* reference genome using the package ANNOVAR v1.0.0 (Wang et al., 2010). Based on the genome annotation, SNPs were categorized in exonic regions (overlapping a coding exon), intronic regions (overlapping an intron), splicing sites (within 2 bp of a splicing junction), upstream or downstream regions (within a 1 kb region upstream from the transcription start site or downstream from the transcription stop site), intergenic regions, transitions (ts), transversions (tv) and ts/tv. SNPs in exonic regions were further grouped into stop gain and stop loss (SNPs causing the gain and loss of stop codons), synonymous SNPs (those that did not cause amino acid changes) and nonsynonymous SNPs (those caused amino acid changes).

Phylogenetic and population structure analyses

To clarify phylogenetic relationships from a genome-wide perspective, an individual-based NJ (neighbour-joining) tree was constructed using P distance in TreeBestv1.9.2 software (http://treesoft.sourceforge.net/treebest.shtml). Bootstrap values were derived from 1000 resampling. Population genetic structure was assessed using the software Admixture (1.23). The number of assumed genetic clusters *K* ranged from 2 to 7, with 10 000 iterations for each run. We also conducted PCA to evaluate the genetic structure in GCTA 1.24.2 (http://cnsgenomics.com/software/gcta/pca.html) software (Li and Durbin, 2010). The significance level of the eigenvector was determined with the Tracy-Widom test.

Population genetic analysis

Fixation statistics *(Fst)* and nucleotide diversity *(θ)* were calculated in VCFTools v0.1.15 (Danecek et al., 2011), with sliding windows of 10 kb.

Linkage-disequilibrium analysis

The software Plink v1.07 (Purcell et al., 2007) was used to calculate the LD coefficient (*r²*) between pairwise high-quality SNPs; the parameters were set as: ‘--1d-window-r2 0 --ld-window 99999 --ld-window-kb 1000’, and the results were used to estimate LD decay.

GWAS analysis

Totally, 4 052 759 SNPs (MAF > 0.05; Quality ≥ 20; GQ ≥ 5; missing rate ≤ 0.2; depth ≥ 4) were used in GWAS for the 15 different traits. To correct for the effect of accession imbalance based on the geographical distribution, association analysis was conducted with the genome-wide efficient mixed-model association (GEMMA 0.94.1, http://www.xzlab.org/software.html) software package (Zhou and Stephens, 2012). For mixed-linear model analysis, we used the following equation:

\[ y = Xα + Sβ + Kμ + e \]

where \( y \) represents phenotype; \( α \) and \( β \) are fixed effects representing marker effects and non-marker effects respectively; and \( μ \) represents unknown random effects. \( X, S \) and \( K \) are the incidence matrices for \( α, β \) and \( μ \), respectively, and \( e \) is a vector of random residual effects. Additionally, the top three PCs were used to build up the \( S \) matrix for population structure correction, and the matrix of simple matching coefficients was used to build up the \( K \) matrix. The analyses were performed in the GEMMA software package. The parameters were set as: ‘gemma -bfile file -kinship -lmm 1 -o outfile -mss 0.2 -maf 0.05 -c covariates (GCTA: PCA)’. The effect values of our genetic markers were tested by \( F \) tests and corrected for multiple testing using Bonferroni correction. Only the most obvious SNP peak in Manhattan plot was chosen as the candidate SNP. Meanwhile, to estimate the difference between observed and predicted values of quantitative traits, all Manhattan results were validated by Q-Q plots.

Estimating breeding value

BLUP (Poland et al., 2011) was used to calculate the breeding values with lme4 packages in R (version 3.5.3). The formula was as follows:

\[ Y = μ + Line + Loc + Year + (Rep in Loc × Year) + (Line × Loc) + (Line × Year) + ε \]

where \( Y, μ, \) Line and Loc represent phenotype, intercept, variety effects and environmental effects respectively. Rep indicates different repetitions, and \( ε \) represents random effects. Rep in Loc × Year shows the interaction between repetition in the same location and year. Line × Loc is used to display the interaction between variety and environment. Line × Year is used to display the interaction between variety and year.
Transcriptome sequencing

Five Sea Island cotton extreme accesses were planted in the field in 2019 (Table S21). Bolls were collected during the initiation stage (0 DPA), cell elongation stage (5, 10, 15 DPA) and secondary-wall synthesis stage (20, 25 DPA). Total RNA was extracted from the fibres of the boll samples with an EASYspin RNA Plant Mini Kit (Cat # RN9002, Aidlab Biotechnologies., Ltd). The qualified RNA was treated with DNase I (Takara Biomedical Technology Co., Ltd., Beijing, China) was used for constructing cDNA library. HiSeq sequencing, assembling, mapping (HisAT 2.0.4 (Kim et al., 2015), with default parameters), analysing gene expression (HTSeq v0.6.1, -m union; Anders et al., 2015), detecting SNP (GATK v3.5, QUAL < 30.0 & QD < 5.0; McKenna et al. (2010), identifying differentially expressed genes (DESeq 1.10.1, P.adj < 0.05; Anders and Huber, 2010), GO (GOSeq, Release2.12, Corrected P-value < 0.05; Young et al., 2010) and KEGG (KOBASE v2.0, Corrected P-value < 0.05; Mao et al., 2005) annotation according to the method in our laboratory (Shi et al., 2015).

Functional characterization of GbDP1/2 genes

Sea Island cotton (Gossypium barbadense) highly susceptible (S) accession 15-3464 and highly resistant (R) accession T10-280 were used for VIGS transformation of GbDP1. Sea Island cotton highly resistant (R) accession (XH42) and highly susceptible (S) accession Su7871 were used for VIGS transformation of GbDP2. For virus-induced gene silencing (VIGS), 516-bp and 502-bp fragments from GbDP1 and GbDP2 were separately cloned into the PacI and SpeI sites of the pCLCrV-VA vector (primers used in Table S20). To analyse expression and silencing efficiency of GbDP1/2, leaves were harvested at 25 days post-inoculation (dpi) with FOV in two sets of resistant and susceptible wild-type (RWT) and S WT accessions, their DP-transgenic accession (S pCLCrVA-DP1, S pCLCrVA-DP2, R pCLCrVA-DP1 and R pCLCrVA-DP2) and empty-vector transformants (S pCLCrVA and R pCLCrVA). Total RNA (~2 μg) was extracted and was then reverse-transcribed in a 20 μL reaction mixture with PrimeScript™ RT reagent Kit with qDNA Eraser (Perfect Real Time) (Cat # RR047A; Takara). 1 μL sample aliquots were used as templates for qRT-PCR analysis. Three technical replicates per sample and three biological-replicate samples were analysed for each experiment. UBQ7 was used as the internal control for qRT-PCR data analysis.

VIGS experiments of GbFL2, GbFS1 and GbLP1

Sea Island cotton long-fibre accession XH58 and short-fibre accession Ashi were used for VIGS transformation receptors GbFL2. Sea Island cotton high-fibre-strength accession XH37 and low-fibre-strength accession LuoSailNa were used for VIGS transformation receptors GbFS1 respectively. Sea Island cotton high-lint-percentage accession Giza81 and low-lint-percentage accession C352 were used for VIGS transformation receptors GbLP1. For the VIGS experiments, plants were grown on the field of Hainan for 1 growing season. Inserts to generate pCLCrVA-FL2/FS1/LP1, approximately 500-bp fragments, were amplified from G. barbadense cDNA. Primer pairs to generate pCLCrVA-FL2/FS1/LP1 vectors are shown in Table S20. PCR fragments were cloned into the pCLCrV plasmid using PacI and SpeI. Plasmids pCLCrVB, pCLCrVA, and their derivatives were transformed into A. tumefaciens GV3101 (Shanghai Weidi Biotechnology Co., Ltd., Shanghai, China) using its supplied method. Before transformation, Agrobacterium containing pCLCrVA or one of its derivatives were mixed with an equal volume of Agrobacterium containing pCLCrVB (Gu et al., 2014). Mixed Agrobacterium solutions were infiltrated into the abaxial side of cotyledons of 2-week-old cotton seedlings using a needleless syringe as described previously (Gao et al., 2013). For 4-week-old cotton seedlings, Agrobacterium solutions were infiltrated into the abaxial side of cotyledons and leaves together with injection into the apical growth point of stems using a bevelled needle. Two months after infiltration, RNA was extracted from cotton leaves to measure the expression of the target genes using qRT-PCR. There were 100 individual plants for each treatment. All harvested plants were used for phenotyping of LP, but for FL and FS, we chose 10 positive individuals determined by qRT-PCR for phenotyping. Statistics of significance were carried out using two-tailed t tests.

Hierarchical filtering strategy of associated SNPs genes, QTL, candidate/key genes

Significant SNPs were screened by the criterion (-log(P-value) > 6). QTLs were defined according to the position of significant SNPs and the size of LD interval. In order to avoid missing key candidate genes, we defined the total size of QTL as 1Mb according to the criterion in Fang et al. (2017), namely, 500 kb upstream and downstream of associated SNPs, slightly larger than the size of actual LD interval (388 kb). That is, one significant SNP corresponds to 1Mb QTL interval. All genes in those QTL intervals were regarded initially as associated genes. These were filtered as follows: First, only the top three highest peak SNPs and the significant SNPs that could be repeatedly detected in at least two environments were regarded as key SNPs, thus narrowing the candidate SNPs from dozens to about 10. Second, only the genes closest to these key SNPs, at the same time having large-effect variations related to phenotype change and expressed differentially in extreme accesses, were considered as candidate genes for further transgenic validation. Finally, only genes silenced by VIGS and related to the phenotypes in question were deemed to be key genes (Figure S23).

Genetic transmission analysis

To detect genetically transmitted regions in a pedigree, we calculated the SNP ratio between parental accessions and XH39/CH60. A window size of 200 SNPs, with a step size of 20 SNPs, was used to perform genomic scans (Fang et al., 2017; Jiao et al., 2012). A window with the same SNP ratio ≥99% was considered as an inheritable fragment in the pedigree (Ma et al., 2019). We used our Sea Island cotton fibre transcriptome data to validate the function of candidate genes.

Acknowledgements

We thank K. W. and F. L., Institute of Cotton Research, China Academy of Agricultural Sciences, for providing seeds of 86 accessions. We thank X. Y., CAU for guidance in GWAS analysis. We thank H.L., B.Z., M.Z., Z.Z. and Y.W. for their helps in functional validation. This work was supported by National Key R & D Program for Crop Breeding (2016YFD0100306) to J.H., Key Cultivation Program of Science and Technology Innovation of Xinjiang Academy of Agricultural Sciences (xjcppy-001) to J.K., China Agriculture Research System (CARS-15-09) to A.A., Project of Agricultural Science and Technology Innovation of Shanxi Academy of Agricultural Sciences (YCX2018D2YS06) to P.L. and...
Phylogenomics of sea island cotton in 707

Project of Modern Agricultural Science and Technology Innovation of Hebei Academy of Agriculture and Forestry Sciences (2019-3-7) to B.G.

Conflicts of interest
The authors have no conflicts of interest to declare.

Author contributions
J.H. and J.F.W. conceived and designed the research. J.K., W.W., A.A., N.Z. and J.H. prepared the population material. W.W., J.Z., M.W., L.X., J.Y., X.N., H.X., A.A. and J.K. performed field experiments and phenotyping. W.W. and N.Z. performed data integration. N.Z. performed sampling, sequencing, genomic-variant and GWAS analyses. K.I., H.N., A.G. and N.Z. performed transcriptome analyses. N.Z. and K.I. conducted gene expression analysis. J.H., N.Z., B.L., I.D., K.J. and W.W. took part in functional validation. C.C. selected the pedigree accessions and N.Z. performed pedigree analysis. Z.P., B.G., J.G., P.L., Y.S., C.C., H.N. and C.E.G. contributed to the project discussion. N.Z. and J.H. wrote the manuscript. C.E.G., J.H. and J.F.W. revised the manuscript.

Data availability statement
Raw genome resequencing reads were deposited in NCBI sequence read archive (SRA) under BioProject ID: PRJNA720817 (Table S1). VCF file with final SNPs used for GWAS analyses is available from figshare repository (https://figshare.com/articles/(Table S1). VCF file with final SNPs used for GWAS analyses is sequence read archive (SRA) under BioProject ID: PRJNA720817. C.E.G., J.H. and J.F.W. revised the manuscript. C.E.G., J.H. and J.F.W. revised the manuscript.

References
Abdullaev, A.A., Salakhutdinov, I.B., Egamberdiev, S.S., Khurshut, E.E., Rizaeva, S.M., Ulloa, M. and Abdurakhmonov, I.Y. (2017) Genetic diversity, linkage disequilibrium, and association mapping analyses of Gossypium barbadense L. germplasm. PLoS One, 12, e0188125. Abdurakhmonov, I.Y., Saha, S., Jenkins, J.N., Buriev, Z.T., Shermatov, S.E., Scheffer, B.E., Pepper, A.E. et al. (2009) Linkage disequilibrium based association mapping of fiber quality traits in G. hirsutum L. variety germplasm. Genetika, 136, 401–417. Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. Genome Biol. 11, R106. Anders, S., Pyl, P.T. and Huber, W. (2015) HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics; 31, 166–169. Cai, C., Zhu, G., Zhang, T. and Guo, W. (2017) High-density 80 K SNP array is a powerful tool for genotyping G. hirsutum accessions and genome analysis. BMC Genom. 18, 654. Chen, Z.I., Sreedasyam, A., Ando, A., Song, Q., De-Santiago, L.M., Hulse-Kemp, A.M., Ding, M. et al. (2020) Genomic diversifications of five Gossypium allotriploid species and their impact on cotton improvement. Nat. Genet. 52, 525–533. Dai, C. and Xue, H.W. (2010) Rice early flowering1, a CKI, phosphorylates DELLA protein SR1 to negatively regulate gibberellin signalling. EMBO J 29, 1916–1927. Danecék, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E. et al. and 1000 Genomes Project Analysis Group. (2011) The variant call format and VCFTools. Bioinformatics, 27, 2156–2158. Dong, C., Wang, J., Yu, Y.U., Ju, L., Zhou, X., Ma, X., Mei, G. et al. (2019) Identifying functional genes influencing Gossypium hirsutum fiber quality. Front Plant Sci. 9, 1968. Du, X., Huang, G., He, S., Yang, Z., Sun, G., Ma, X., Li, N. et al. (2018) Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. Nat. Genet. 50, 796–802. Fan, L., Wang, L., Wang, X., Zhang, H., Zhu, Y., Guo, J., Gao, W. et al. (2018) A high-density genetic map of extra-long staple cotton (Gossypium barbadense) constructed using genotyping-by-sequencing based single nucleotide polymorphic markers and identification of fiber traits-related QTL in a recombinant inbred line population. BMC Genom. 19, 489. Fang, L., Gong, H., Hu, Y., Liu, C., Zhou, B., Huang, T., Wang, Y. et al. (2017a) Genomic insights into divergence and dual domestication of cultivated allotetraploid cottons. Genome Biol. 18, 33. Fang, L., Wang, O., Hu, Y., Jia, Y., Chen, J., Liu, B., Zhang, Z. et al. (2017b) Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. Nat. Genet. 49, 1089–1098. Fang, L., Zhao, T., Hu, Y., Si, Z., Zhu, X., Han, Z., Liu, G. et al. (2021) Divergent improvement of two cultivated allotetraploid cotton species. Plant Biotechnol. J. 19, 1325–1336. Feng, H., Li, X., Chen, H., Deng, J., Zhang, C., Liu, J., Wang, T. et al. (2018) GhHR2B2, a ubiquitin ligase, is involved in cotton fiber development via the ubiquitin-26s proteasome pathway. J. Exp. Bot. 69, 5059–5075. Gao, W., Long, L., Zhu, L.F., Xu, L., Gao, W.H., Sun, L.Q., Liu, L.L. et al. (2013) Proteomic and virus-induced gene silencing (VIGS) analyses reveal that gossypol, brassinosteroids, and jasmonic acid contribute to the resistance of cotton to Verticillium dahliae. Mol Cell Proteomics, 12, 3690–3703. Gong, W., Xu, F., Sun, J., Peng, Z., He, S., Pan, Z. and Du, X. (2017) ITRAQ-based comparative proteomic analysis of seedling leaves of two Upland cotton genotypes differing in salt tolerance. Front Plant Sci. 8, 2113. Gross, S.D. and Anderson, R.A. (1998) Casen kinase I: spatial organization and positioning of a multifunctional protein kinase family. Cell Signal. 10, 699–711. Grover, C.E., Arick, M.A., Thrash, A., Conover, J.L., Sanders, W.S., Peterson, D.G., Frelasichowski, J.E. et al. (2019) Insights into the evolution of the new world diploid cottons (Gossypium, subgenus houzingenia) based on genome sequencing. Genome Biol. Evol. 11, 53–71. Grover, C.E., Pan, M., Yuan, D., Arick, M.A., Hu, G., Brase, L., Stelly, D.M. et al. (2020) The Gossypium longicalyx genome as a resource for cotton breeding and evolution. Genes Genom. Genet. 10, 1457–1467. Gu, Z., Huang, C., Li, F. and Zhou, X. (2014) A versatile system for functional analysis of genes and microRNAs in cotton. Plant Biotechnol. J. 12, 638–649. Ho, M.H., Saha, S., Jenkins, J.N. and Ma, D.P. (2010) Characterization and promoter analysis of a cotton RING-type ubiquitin ligase (E3) gene. Mol. Biotechnol. 46, 140–148. Hu, Y., Chen, J., Fang, L., Zhang, Z., Ma, W., Niu, Y., Ju, L. et al. (2019) Gossypium barbadense and Gossypium hirsutum genomes provide insights into the origin and evolution of allotetraploid cotton. Nat. Genet. 51, 739–748. Huang, C., Nie, X., Shen, C., You, C., Li, W., Zhao, W., Zhang, X. et al. (2017) Population structure and genetic basis of the agronomic traits of upland cotton in China revealed by a genome-wide association study using high-density SNPs. Plant Biotechnol. J. 15, 1374–1386. Huang, G., Wu, Z., Percy, R.G., Bai, M., Li, Y., Frelasichowski, J.E., Hu, J. et al. (2020) Genome sequence of Gossypium herbaceum and genome updates of Gossypium arboreum and Gossypium hirsutum provide insights into cotton A-genome evolution. Nat. Genet. 52, 516–524. Huang, X. and Han, B. (2014) Natural variations and genome-wide association studies in crop plants. Annu. Rev. Plant. Biol. 65, 531–551. Huang, X., Kurata, N., Wei, X., Wang, Z.-X., Wang, A., Zhao, Q., Zhao, Y. et al. (2012) A map of rice genome variation reveals the origin of cultivated rice. Nature, 490, 497–501. Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q.I., Zhao, Y., Li, C. et al. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. 42, 961–967. Islam, S., Thyssen, G.N., Jenkins, J.N., Zeng, L., Delhom, C.D., McCarty, J.C., Deng, D.D. et al. (2016) A MAGIC population-based genome-wide association study reveals functional association of GHRB1_A07 gene with superior fiber quality in cotton. BMC Genom. 17, 903.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G. et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25, 537–541.

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

Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R.J., Ma, Z. et al. (2015) Genome sequence of cultivated upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. Nat. Biotechnol. 33, 524.

Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics, 26, 585–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G. et al. and 1000 Genome Project Data Processing Subgroup. (2009) The sequence alignment/map format and SAMtools. Bioinformatics, 25, 5078–5079.

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., Han, Y. et al. (2013) Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat. Genet. 45, 43–50.

Li, P.T., Lu, Q.W., Xiao, X.H., Yang, R. and Duan, X.X. (2021) Dynamic expression analysis and introgressive gene identification of fiber length using chromosome segment substitution lines from G. hirsutum × G. barbadense. J. Exp. Bot. 90, 129–144.

Li, Q., Hu, A., Qi, J., Dou, W., Qin, X., Zou, X., Xu, L. et al. (2020) CSWAKL08, a pathogen-induced wall-associating receptor-like kinase in sweet orange, confers resistance to citrus bacterial canker via ROS control and JA signaling. Hortic. Res. 7, 42.

Liu, R., Gong, J., Xiao, X., Zhang, Z., Li, L., Li, A., Lu, Q. et al. (2018) GWAS analysis and QTL identification of fiber quality traits and yield components in Upland cotton using enriched high-density SNP markers. Front. Plant Sci. 9, 1067.

Liu, S., Zhang, X., Xiao, S., Ma, J., Shi, W., Qin, T., Xi, H. et al. (2021) A single-nucleotide mutation in a GLUTAMATE RECEPTOR-LIKE gene confers resistance to Fusarium Wilt in Gossypium hirsutum. Adv. Sci. 8, 2002723.

Lu, Q., Shi, Y., Xiao, X., Li, P., Gong, J., Gong, W., Liu, A. et al. (2017) Transcriptome analysis suggests that chromosome introgression fragments from Sea Island cotton (Gossypium barbadense) increase fiber strength in Upland cotton (Gossypium hirsutum). Genes Genom. Genet. 7, 3469–3479.

Lu, X., Fu, X., Wang, D., Wang, J., Chen, X., Hao, M., Wang, J. et al. (2019) Resequencing of cv CR1-12 family reveals haplotype block inheritance and recombination of agronomically important genes in artificial selection. Biotechnol. J. 17, 945–955.

Ma, J., Geng, Y., Pei, W., Wu, M., Li, X., Liu, G., Li, D. et al. (2018a) Genetic variation of dynamic fiber elongation and developmental quantitative trait locus mapping of fiber length in Upland cotton (Gossypium hirsutum L.). BMC Genom. 19, 882.

Ma, X., Wang, L., Li, W., Zhang, Y., Zhou, X., Liu, Y., Ren, Z. et al. (2019) Resequencing core accessions of a pedigree identifies derivation of genomic segments and key agronomic trait loci during cotton improvement. Plant Biotechnol. J. 17, 762–775.

Ma, Z., He, S., Wang, X., Sun, J., Zhang, Y., Zhang, G., Wu, L. et al. (2018b) Resequencing a core collection of Upland cotton identifies genomic variation and loci influencing fiber quality and yield. Nat. Genet. 50, 803–813.

Mace, E.S., Tai, S., Gilding, E.K., Li, Y., Prentis, P.J., Bian, L., Campbell, B.C. et al. (2013) Whole-genome sequencing reveals untapped genetic potential in Africa’s indigenous cereal crop sorghum. Nat. Commun. 4, 2320.

© 2021 The Authors. Plant Biotechnology Journal published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd.
Wang, Y., Wu, Y., Yu, B., Yin, Z. and Xia, Y. (2017b) EXTRA-LARGE G PROTEINS interact with E3 ligases PUB4 and PUB2 and function in cytokinin and developmental processes. Plant Physiol. 173, 1235–1246.

Wendel, J.F., Brubaker, C.L. and Seelanan, T. (2010) The origin and evolution of Gossypium. In Physiology of Cotton, (Stewart, J.M., Oosterhuis, D.M., Heitholt, J.J. and Mauney, J.R., eds), pp. 1–18. Netherlands, Europe: Springer.

Wendel, J.F. and Grover, C.E. (2015) Taxonomy and evolution of the cotton genus, Gossypium. In Cotton, (Fang, D.D. and Percy, R.G., eds), pp. 25–44. New Jersey, US: John Wiley & Sons, Ltd.

Westengen, O.T., Huaman, Z. and Heun, M. (2005) Genetic diversity and geographic pattern in early South American cotton domestication. Theor. Appl. Genet. 110, 392–402.

Xiao, G., Zhao, P. and Zhang, Y. (2019) A pivotal role of hormones in regulating cotton fiber development. Front. Plant Sci. 10, 87.

Yang, Z., Ge, X., Yang, Z., Qin, W., Sun, G., Wang, Z., Li, Z. et al. (2019) Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. Nat. Commun. 10, 2989.

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P.C., Hu, L., Yamasaki, M. et al. (2016) Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat. Genet. 48, 927–934.

Yao, Z., Chen, Q., Chen, D., Zhan, L., Zeng, K., Gu, A., Zhou, J. et al. (2019) The susceptibility of sea-island cotton recombinant inbred lines to Fusarium oxysporum f. sp. vasinfectum infection is characterized by altered expression of long noncoding RNAs. Sci. Rep. 9, 2894.

Young, M.D., Wakefield, M.J., Smyth, G.K. and Oshlack, A. (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 11, R1.

Yu, J., Hui, Y., Chen, J., Yu, H., Gao, X., Zhang, Z., Li, Q. et al. (2021) Whole-genome resequencing of 240 Gossypium barbadense accessions reveals genetic variation and genes associated with fiber strength and lint percentage. Theor. Appl. Genet. 134, 3249–3261.

Yu, J., Zhang, K., Li, S., Yu, S., Zhai, H., Wu, M., Li, X. et al. (2013) Mapping quantitative trait loci for lint yield and fiber quality across environments in a Gossypium hirsutum × Gossypium barbadense backcross inbred line population. Theor. Appl. Genet. 126, 275–287.

Yuan, D., Grover, C.E., Hu, G., Pan, M., Miller, E.R., Conover, J.L., Hunt, S.P. et al. (2021) Parallel and intertwining threads of domestication in allopolyploid cotton. Adv. Sci. 8, 2003634.

Yuan, D., Tang, Z., Wang, M., Gao, W., Tu, L., Jin, X., Chen, L. et al. (2015) The genome sequence of Sea-island cotton (Gossypium barbadense) provides insights into the allopolyploidization and development of superior spinnable fibres. Sci. Rep. 5, 17662.

Zhang, J., Fang, H., Zhou, H., Sanogo, S. and Ma, Z. (2014) Genetics, breeding, and marker-assisted selection for Verticillium wilt resistance in cotton. Crop Sci. 54, 1289–1303.

Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., Zhang, J. et al. (2015) Sequencing of allotetraploid cotton (Gossypium hirsutum L. acc. TM-1) provides a resource for fiber improvement. Nat. Biotechnol. 33, 531–537.

Zhang, X., Guo, Z., Chen, Y., Abdalaiha, E.H.M.A., Dong, Z., Chen, D. et al. (2017) Developmental and hormonal regulation of fiber quality in two natural-colored cotton cultivars. J. Integr. Agric. 16, 1720-1729.

Zhao, W., Wang, H., Chen, W. and Li, Y. (2014) Genetic structure, linkage disequilibrium and association mapping of Verticillium wilt resistance in elite cotton (Gossypium hirsutum L.) germplasm population. PLoS One, 9, e86308.

Zhou, X. and Stephens, M. (2012) Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44, 821–824.

Zhou, Z., Jiang, Y.U., Wang, Z., Gou, Z., Luo, J., Li, W., Yu, Y. et al. (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat. Biotechnol. 33, 408–414.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.
Table S14 Selective pressure analysis based on $K_a$ and $K_s$ value.
Table S15 Genetic constitutions of XH39 and XH60.
Table S16 The same genes related to key traits in XH39 and XH60 pedigree identified by both GWAS and IBD analysis.
Table S17 Comparison of genome sequencing data in common Sea Island cotton samples in our and other studies.
Table S18 Comparison of fibre quality and yield phenotype data in our and other studies.
Table S19 Loci and genes for fibre length and fibre strength which overlap with previous researches on $G.~barbadense$.
Table S20 Primers used in this study.
Table S21 Fibre samples of Sea Island cotton used for transcriptome sequencing and analysis.