the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2022) **110,** 764–780 doi: 10.1111/tpj.15702

Introgression from *Gossypium hirsutum* is a driver for population divergence and genetic diversity in *Gossypium barbadense*

Pengpeng Wang[1,†], Na Dong[2,†], Maojun Wang[3,†], Gaofei Sun[4,†], Yinhua Jia[1,†], Xiaoli Geng[1,†], Min Liu[5], Weipeng Wang[2], Zhaoe Pan[1], Qiuyue Yang[2], Hongge Li[1], Chunyan Wei[2], Liru Wang[1], Hongkun Zheng[5], Shoupu He[1,*], Xianlong Zhang[3,*], Qinglian Wang[2,*] and Xiongming Du[1,*]

[1]*Institute of Cotton Research, Chinese Academy of Agricultural Sciences/Zhengzhou Research Base, State Key Laboratory of Cotton Biology, Zhengzhou University, Zhengzhou 450001, China,*

[2]*Henan Key Laboratory of Molecular Ecology and Germplasm Innovation of Cotton and Wheat, Collaborative Innovation Center of Modern Biological Breeding in Henan Province, Henan Institute of Science and Technology, Xinxiang 453003, China,*

[3]*National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, Hubei 430070, China,*

[4]*Anyang Institute of Technology, Anyang 455000, China, and*

[5]*Biomarker Technologies Corporation, Beijing, China*

Received 6 December 2021; revised 22 January 2022; accepted 3 February 2022; published online 7 February 2022.

*For Correspondence e-mails (duxiongming@caas.cn) [X.D.]; (wangql985@163.com) [Q.W.]; (xlzhang@mail.hzau.edu.cn) [X.Z.]; (heshoupu@caas.cn) [S.H.]).

†These authors contributed equally to this work.

SUMMARY

During the domestication and improvement processes, interspecific introgression from *Gossypium hirsutum* has reorganized the genomic architecture of *Gossypium barbadense*; however, the introgression details and the trait-related genetic loci remain largely unknown. Here, we perform a genome-wide population analysis and genetically recategorize 365 *G. barbadense* accessions into four subgroups which is different from previous categorizations. A total of 315 introgression events from *G. hirsutum* to *G. barbadense*, which primarily contributed to population divergence and agronomic trait variation in *G. barbadense*, are identified. We find that 70% introgression from *G. hirsutum* have greatly increased the genetic diversity and divergence of *G. barbadense*. Some loci are identified with divergent haplotype selection for adaptation to the environment at high latitudes. Through genome-wide association study and genome linkage disequilibrium interval haplotyping analyses, two fiber micronaire-related haplotype blocks are detected, one of which (*FM2*) is introgressed from *G. hirsutum*. Seven distinguished traits related to growth period, plant architecture, and stronger vegetative growth habit are found to have pleiotropic effects controlled by a single gene in *G. barbadense*. Thus, this study provides new insights into the breeding history of *G. barbadense* and highlights introgression is a driver for improving cultivars in *G. barbadense*.

Keywords: whole genome resequencing, introgression, fiber micronaire, adaptation.

INTRODUCTION

Cotton is the most important natural fiber crop worldwide. To meet the demands of the modern textile industry, simultaneous genetic improvement of fiber quality and yield is one of the main challenges for cotton breeders. The allotetraploid cotton species *Gossypium barbadense* is valued for its superior fiber quality but is cultivated in only a limited area because of its low fiber yield and narrow adaptability (Shi et al., 2016; Lu et al., 2017).

Gossypium barbadense originated west of the Andes, spread to eastern South America and the Caribbean, and then expanded worldwide because of transportation by European colonists during the Age of Exploration (Stephens and Moseley, 1974; Percy and Wendel, 1990; Jack and Dillehay, 1996; Piperno and Pearsall, 1998). The modern, improved germplasm pools of *G. barbadense* include Egyptian and American Pima, both of which originated from the 'Sea Island' cotton developed in the Caribbean in the late eighteenth century (McGowan 1960; Fryxell 1965).

Because of their derivation from the same founder (Ware 1936; Kearney 1943; McGowan 1960; Smith et al., 1999), both of these gene pools had fairly low genetic diversity (Kerr 1960; Smith et al. 1999). In the history of *G. barbadense* cultivation, diverse genetic sources, including Sea Island, Tanguis (from Brazil), and early Egyptian cultivars (cv. Giza) have been employed to improve fiber quality and adapt to various environments (Kerr 1960; Feaster and Turcotte, 1962; Feaster et al. 1967; Young et al., 1976; Percy and Turcotte, 1998; Percy, 2002; Smith et al., 1999; Ulloa et al., 2006). Revealing the genetic relationships among these sources would help improve our understanding of the cultivation history of *G. barbadense*.

To improve the adaptability and fiber yield of *G. barbadense*, a long-term breeding practice of interspecific hybridization with *Gossypium hirsutum*, which is the major cultivated allotetraploid cotton species because of its wide adaptability and high production, has been applied (Percy and Wendel, 1990; Wang et al., 1995). Since the last century, many interspecific crosses between *G. hirsutum* and *G. barbadense* have been developed to identify quantitative trait loci (QTLs) for fiber quality and yield traits (Jiang et al., 1998; Kohel et al., 2001; Paterson et al., 2003; Draye et al., 2005; Lacape et al., 2005; He et al., 2007; Lacape et al., 2009; Yu et al., 2013; Said et al., 2015; Chen et al., 2018). Several bidirectionally introgressed segments (ISs) from *G. hirsutum* to *G. barbadense* have been identified (Page et al., 2016; Fang et al., 2017; Hu et al., 2019; Wang et al., 2019). A recent report revealed a yield-increasing locus (Gb_INT13) in *G. barbadense* cultivars that might have been derived from *G. hirsutum* (Nie et al., 2020). During the northward migration of *G. barbadense* cultivation (from South America to the Caribbean and then worldwide), hybridization with *G. hirsutum* is thought to have played an essential role in reshaping the adaptation and phenotypes of *G. barbadense* (Peebles, 1954; Feaster and Turcotte, 1962; Percy and Wendel, 1990; Smith et al., 1999). Therefore, it is necessary to investigate the introgression events that occurred between these two cultivated tetraploid cottons.

In this study, the genomes of *G. barbadense* landraces and their Egyptian, American Pima, Central Asian, and Chinese cultivars were sequenced to determine the spread and breeding history of modern *G. barbadense*. Based on genome-wide association study (GWAS) and genome linkage disequilibrium interval genotyping of *G. barbadense*, we deciphered the genetic basis of population differentiation and the genetic diversity of *G. barbadense* and illuminated the introgression events and haplotype selection affecting associated fiber quality and environmental adaptation traits in modern cultivated *G. barbadense*. These data will serve as a reference supporting breeding programs of *G. barbadense*.

RESULTS

Population properties of *G. barbadense* worldwide

In this study, a total of 365 diverse *G. barbadense* accessions collected worldwide were used for resequencing, and five *G. hirsutum* accessions were introduced as the outgroup. A total of approximately 3.9 Tb Illumina sequence data were generated, with an average depth of 15.6×, and 99.68% of reads were mapped against the latest PacBio-assembled reference genome of *G. barbadense* (cv. 3-79) for each accession (Table S1). Mapping was used to identify genomic variants (Kang et al., 2010) within the *G. barbadense* population ($n = 365$), and a total of 3 729 095 high-quality single nucleotide polymorphisms (SNPs) and 2 472 396 indels were identified (Table S2). Of these SNPs, 18.13 and 1.04% were identified in the genic regions and the protein-coding exons (Table S2).

To explore the genetic relationships among the accessions, we constructed a phylogenetic tree and performed a population analysis with a total of 252 609 SNPs ($n = 370$, including five *G. hirsutum* accessions) with a missing rate of <20%, a minor allele frequency (MAF) of >0.05, and an r^2 (squared correlation of adjacent SNPs) of <0.2. When using *G. hirsutum* as the outgroup, all the *G. barbadense* accessions could be categorized into four groups, which were designated G1 ($n = 22$), G2 ($n = 174$), G3 ($n = 106$), and G4 ($n = 63$) (Figure 1a; Table S1). G1 included the primitive accessions; two accessions (CNH-64-85 and Line-Dar) were collected from Peru, where *G. barbadense* originated (McGowan 1960; Fryxell 1965; Stephens and Moseley, 1974; Percy and Wendel, 1990; Jack and Dillehay, 1996; Piperno and Pearsall, 1998). Most of the accessions in this branch were perennial and collected from the southwestern regions of China ($n = 20$) (Figure S1a). According to historical documents, the landraces of *G. barbadense* in G1 might have been directly introduced into China from South America in the seventeenth century during the Age of Exploration by sailing (Figure 1b). Due to its isolation in the mountainous areas of southwestern China, G1 retained its primitive landrace genotype. Most of the obsolete cultivars (developed during the nineteenth and early twentieth centuries) from the United States (cv. Pima series), Egypt (cv. Giza series), and Central Asia (Figure 1a; Figure S1b) were classified into the G2 group. Notably, this result was not consistent with a previous original geographic classification, which categorized *G. barbadense* accessions as Pima type, Egypt type, and Central Asia type according to their original regions of cultivation (Percy 2009). G3 mainly consisted of improved cultivars collected from the Yangtze River region of China, where cotton was first introduced from Egypt in the twentieth century (Figure 1b; Figure S1b). Finally, G4 contained all the modern cultivars grown in the northwestern region of China (Xinjiang

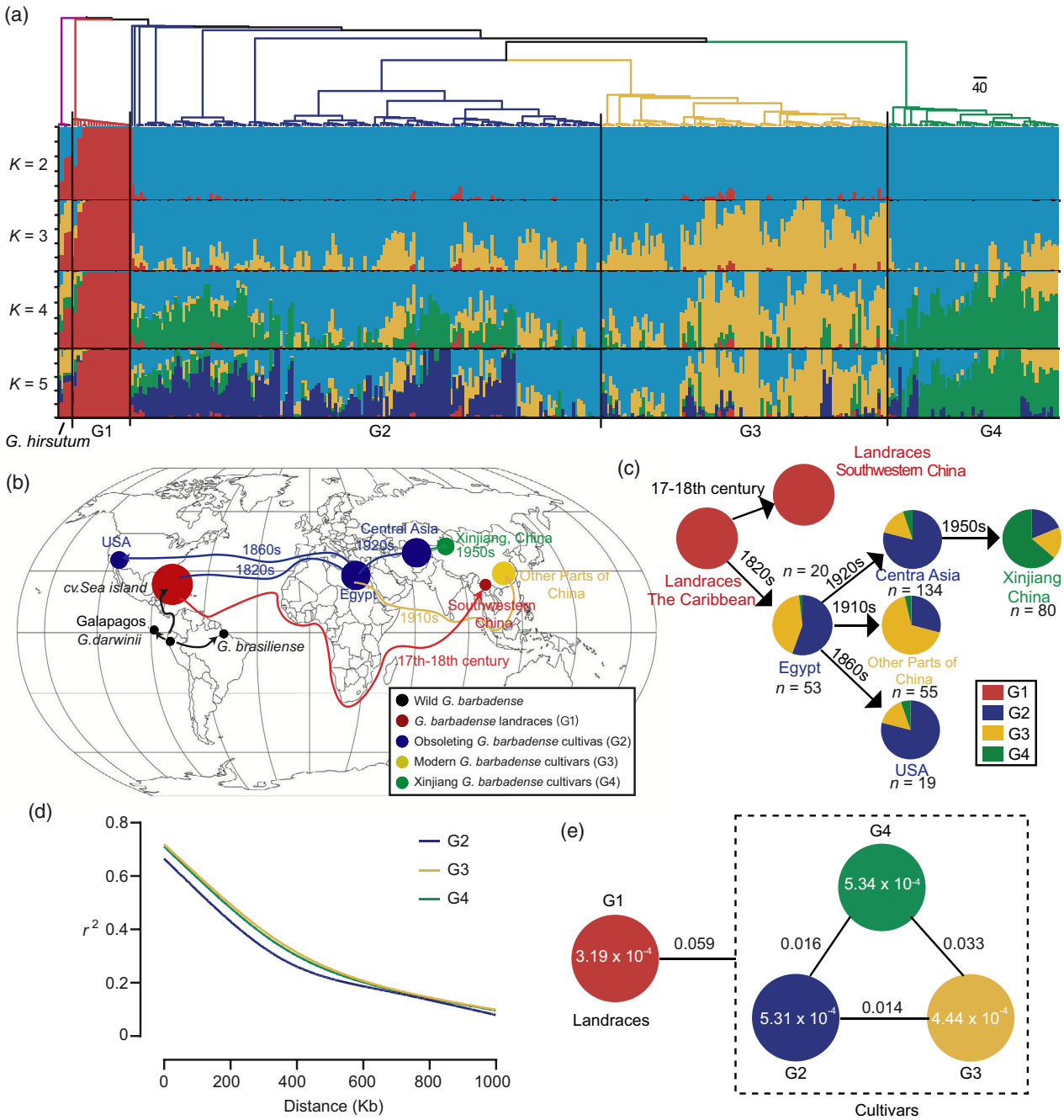


Figure 1. Genetic diversity and introduction history of *Gossypium barbadense*. (a) The upper panel shows a neighbor-joining tree of 365 *G. barbadense* and five *Gossypium hirsutum* accessions constructed using 252 609 SNPs. Branch colors indicate the four groups of *G. barbadense*. The lower panel shows the population structure based on different numbers of clusters (K = 2 to 5); all accessions (y-axis) are arranged in the same order as in the phylogenetic tree. (b) The left panel shows the dispersal route of *G. barbadense* among the major cultivation regions according to the literature. Dispersal routes of wild and semiwild *G. barbadense* (black), landraces (G1) (red), obsolete cultivars (G2) (blue), improved cultivars (G3) (orange), and modern cultivars (G4) (green) are shown by colored arrows. (c) The right panel indicates the genomic components of the four groups in the major cultivation regions. (d) Genome-wide averages of linkage disequilibrium (LD) decay in three cultivar groups. (e) Genetic diversity and population differentiation across four groups. Values in the circles represent the nucleotide diversity (π), and the values between the groups indicate population differentiation (F_{ST}). The cultivars shown include all the accessions in groups G2, G3, and G4.

Province), and it might have been introduced from the former Soviet Union after the 1950s and improved thereafter (Figure 1b, Figure S1b).

We compared traits related to fiber yield, fiber quality, maturity period, and morphology among the three groups (Figure S1c, Tables S3–S5). We found no significant

differences in most traits between groups G2 and G3. However, compared with G2 and G3, we found that G4 differed in terms of all aspects except fiber micronaire (FM), having excellent fiber length (FL) and fiber strength (FS), a greater leaf hair number (LHN), and a shorter growth stage (GS) (Figure S1c, Tables S3–S5), implying that modern cultivars (G4) have integrated more favorable traits than older cultivars. The high similarities of the genomic backgrounds (Figure 1a,b) and phenotypes (Figure S1d) of the accessions from the three early cultivation regions (United States/Pima, Egypt, and Central Asia) demonstrated that *G. barbadense* germplasm should be categorized into four groups according to genotype, consistent with its introduction and breeding history.

The squared correlation of adjacent SNPs (r^2) was calculated to investigate the extent of linkage disequilibrium (LD) within each 1000-bp window. The LD decay distance for group G2 for all SNPs was approximately 280 kb when the value of r^2 was set at 0.35 (when the LD value dropped to half), while it was slightly longer in G3 and G4 (approximately 340 kb) (Figure 1d). The LD decay of G1 was not calculated because of its small population size and low genetic diversity. We concluded that, as an often cross-pollinated crop, cultivated *G. barbadense* has an LD decay distance similar to that of cultivated *G. hirsutum* (296 kb) (Wang et al., 2017) and other cultivated *G. barbadense* (388 kb) (Zhao et al., 2021), but higher than that of cross-pollinated crops such as maize (*Zea mays*) (approximately 22-30 kb) (Hufford et al., 2012; Wang et al., 2020a) and rice (*Oryza sativa*) (approximately 123 kb in *indica* and approximately 167 kb in *japonica*) (Huang et al., 2010).

To explore the population divergence within *G. barbadense*, we estimated the genetic differentiation of the four groups using the pairwise fixation statistic (F_{ST}) (Figure 1e). We found that the F_{ST} value between the landrace group (G1) and cultivar groups (combined G2, G3, and G4) was much higher (0.059) than that within each cultivar group (average 0.021) (Figure 1e). Among the cultivar groups, the F_{ST} values of G2 versus G3 (0.014) and G2 versus G4 (0.016) were similar, while the F_{ST} value of G3 versus G4 was greater (0.033) (Figure 1e). These results indicate that both G3 and G4 might have originated from several primitive germplasm sources in G2 and diverged into two different genotypes due to different local ecological environments and breeding goals. Furthermore, the F_{ST} analysis identified 33 (G2 versus G3), 51 (G2 versus G4), and 39 (G3 versus G4) genomic regions with significant genetic divergence (top 5% of F_{ST} values) covering 4733, 6015, and 5108 genes, respectively (Tables S6–S8). These genes are essential for studying changes in the adaptability and trait improvement of *G. barbadense* in the future.

Combining resequencing and GWAS analysis, five loci associated with FL, lint percentage (LP), and *Fusarium wilt* resistance have been identified recently (Zhao et al., 2021). Compared with this research, our study focused on the introgression event from *G. hirsutum* to *G. barbadense*. We also found some genetic and genomic bases for the critical agronomic traits which were different from those of *G. hirsutum*.

Introgression from *G. hirsutum* restructured the genomic architecture of *G. barbadense*

Two tetraploid cotton species, *G. barbadense* and *G. hirsutum*, originated from nearby areas in America and may have introgressed with each other, and the agronomic traits of *G. barbadense* could be introgressed by upland cotton, but so far there is no molecular evidence to support this. To investigate the landscape of genomic exchanges between *G. barbadense* and *G. hirsutum*, we performed a genome-wide introgression analysis to identify genomic segments introgressed from *G. hirsutum* to *G. barbadense* (ISs). A total of 9017 introgressed bins were identified as introgressed from *G. hirsutum*, including 2541 high-quality introgressed bins (MAF > 0.05). By LD block analysis of the 2541 high-quality introgressed bins, a total of 315 IS regions covering an approximately 164.4-Mb genomic region were identified, representing approximately 7.3% of the whole *G. barbadense* genome (Figure 2a; Figures S2 and S3; Tables S9 and S10). We observed that the cultivars in G3 harbored the most ISs (approximately 60.4 Mb), followed by G2 (approximately 44.4 Mb) and G4 (approximately 44.2 Mb). The *G. barbadense* landraces (G1) harbored much fewer ISs (approximately 7.8 Mb) than those of the cultivars in other groups (Figure 2b, Table S10). This result indicated that human breeding introgressed a lot more segments of *G. hirsutum* into the *G. barbadense* than natural introgression. Among all the chromosomes, the four largest IS regions (>9 Mb) were identified, including one previously identified region located from approximately 51.1 to 71.2 Mb (IS-A01-22) (Page et al. 2016) on chromosome A01 and three new regions on A01 (IS-A01-28), A06 (IS-A06-12), and A10 (IS-A10-8) (Figure 2a,c; Figures S2–S6), and all these four IS regions overlapped with the high-diversity and high-divergence regions identified by π and F_{ST} (Figure 2c–h). In total, approximately 70% of the introgression regions (approximately 122.5 Mb) were found overlapped with these high-genetic-diversity regions, with more than 95% introgression regions overlapping with the high-diversity regions in chromosomes A01, D03, and D09 (Figure 2c–h, Table S11). These data demonstrated that most introgression from *G. hirsutum* greatly increased the genetic diversity and divergence of *G. barbadense* (Figures S4–S6).

We also found some introgressions that were directly related to the agronomic variations. Through genotyping, 92 IS regions were associated with at least one trait variation (Table S9). Among them, 11, nine, and seven IS

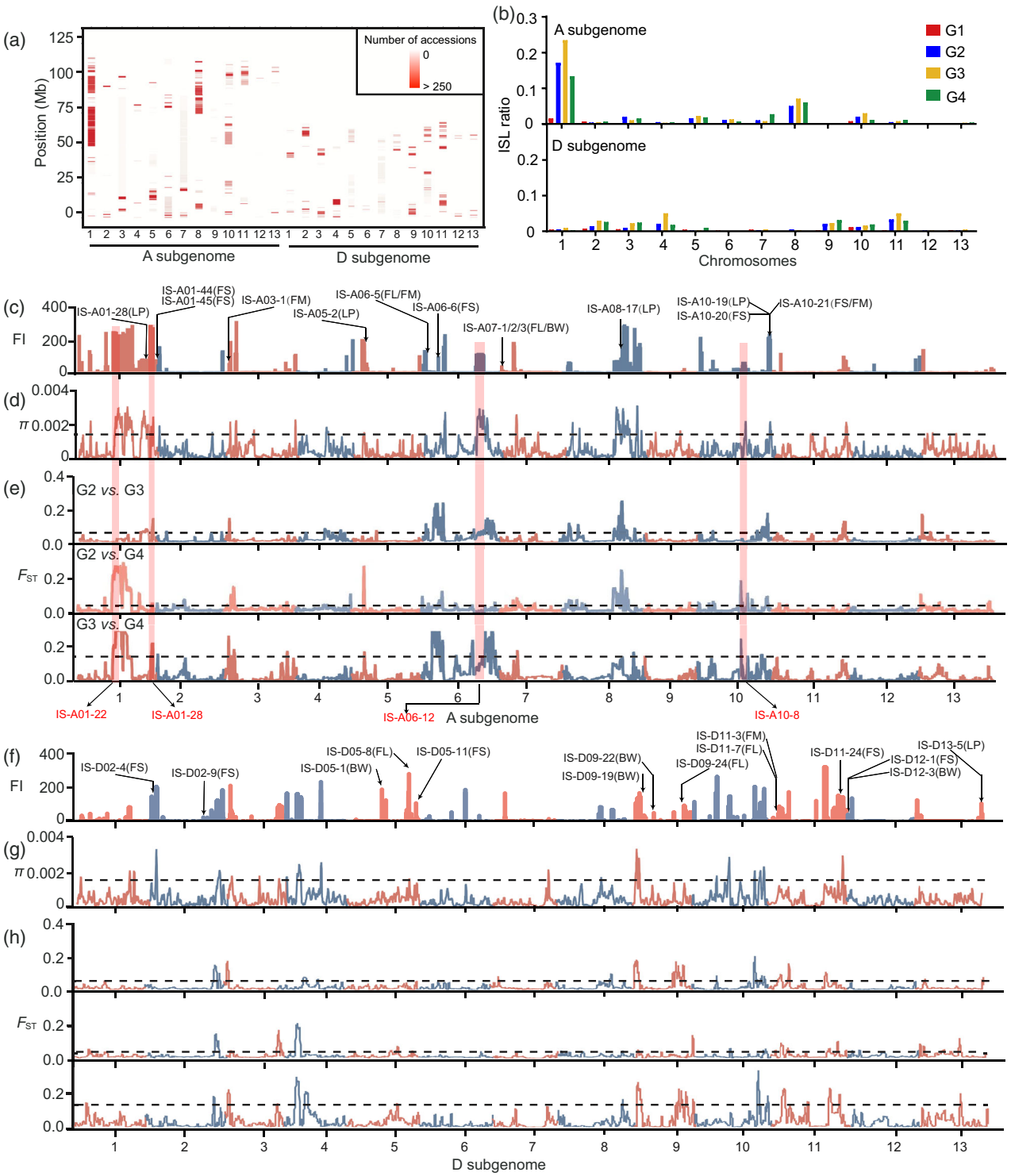


Figure 2. Introgression and population divergence in *G. barbadense*. (a) Heatmap of segment introgression from *G. hirsutum* to *G. barbadense*. The *y*- and *x*-axes show the positions of the introgressed segments (ISs) and chromosomes, respectively. Color intensity of the heatmap indicates the frequency of introgression (FI) at a given location, where FI = the number of accessions carrying ISs. (b) IS length (ISL) ratios of the four groups on each chromosome, where ISL ratio = the total length of the IS on a given chromosome/length of the chromosome. The *x*-axis shows the chromosomes. (c) Frequency of introgression (FI) in the A subgenome. The *y*-axis indicates the number of accessions. (d) Genetic diversity (π) of the A subgenome in 365 *G. barbadense* accessions. (e) Population differentiation (F_{ST}) between pairs of groups on the A subgenome. The *x*-axis of (c–e) indicates the 13 chromosomes on the A subgenome. (f–h) The same features as (c–e) for the D subgenome. The IS regions associated with related traits are shown above (c) and (f), and the related traits are shown in brackets. The major (length > 9 Mb) FI-π-F_{ST} overlapping regions are highlighted by transparent red boxes in red font.

regions were associated with FS, FL, and FM, respectively (Figure 2c,f).

Half of the 11 FS-associated IS regions had a negative effect (Table S12). Among them, IS-A10-20 had the largest impact, with an average decrease of 16% in fiber quality (Figure S7a,e). Four accessions (Figure S7b,f), GB0085, GB0387, GB0111, and GB0065, were chosen to identify the candidate genes associated with FS. By further analysis of all the introgressed gene expression patterns (Table S13), IS-A01-44 (Figure S7a) and IS-A10-20 (Figure S7e) were chosen to identify the candidate genes. *Gbar_A01G021870* was located in IS-A01-44, with higher expression levels in the accessions with ISs (GB0085, GB0387, GB0111) at the 15- and 20-days post-anthesis (DPA) fiber (Figure S7b–d). *Gbar_A10G022420* was located in IS-A10-20, with higher expression levels in the accession without ISs (GB0085), mainly at 15 DPA fiber (Figure S7g,h). *Gbar_A01G021870* encodes the bidirectional sugar transporter SWEET12 and *Gbar_A10G022420* encodes a macrophage erythroblast attacher. These results indicate that introgression events of IS-A01-44 and IS-A10-20 resulted in high expression of *Gbar_A01G021870* and *Gbar_A10G022420*, which affects FS in *G. barbadense*.

Among the nine FL- and seven FM-associated IS regions, seven and four IS regions had negative effects (Tables S14–S17), indicating that the introgression event from *G. hirsutum* to *G. barbadense* mainly led to shorter FL but better fiber fineness. IS-A07-2 was identified as a pleotropic IS region associated with FL and FM (Figure S8a,d). By gene expression pattern analysis (Figure S8c–g), *Gbar_A07G001990* and *Gbar_A07G002030* were identified as candidate genes, which encode arginine–tRNA ligase and leucine-rich receptor-like protein kinase, respectively.

Among the nine boll weight (BW)- and seven LP-associated IS regions, five and six IS regions had positive effects (Tables S18–S21). The higher expression of *Gbar_D12G002630* (Figure S8k) and *Gbar_A08G012800* (Figure S8o), which encode the two-component response regulator ARR18-like protein and natural resistance-associated macrophage 1, at the −1 and 5 DPA ovules in the accessions carrying the IS regions (Figure S8j,n) was involved in the regulation of BW and LP.

In conclusion, the introgression event from *G. hirsutum* was mainly favorable for the improvement of BW, LP, and FM, but negatively affected FS and FL in *G. barbadense*.

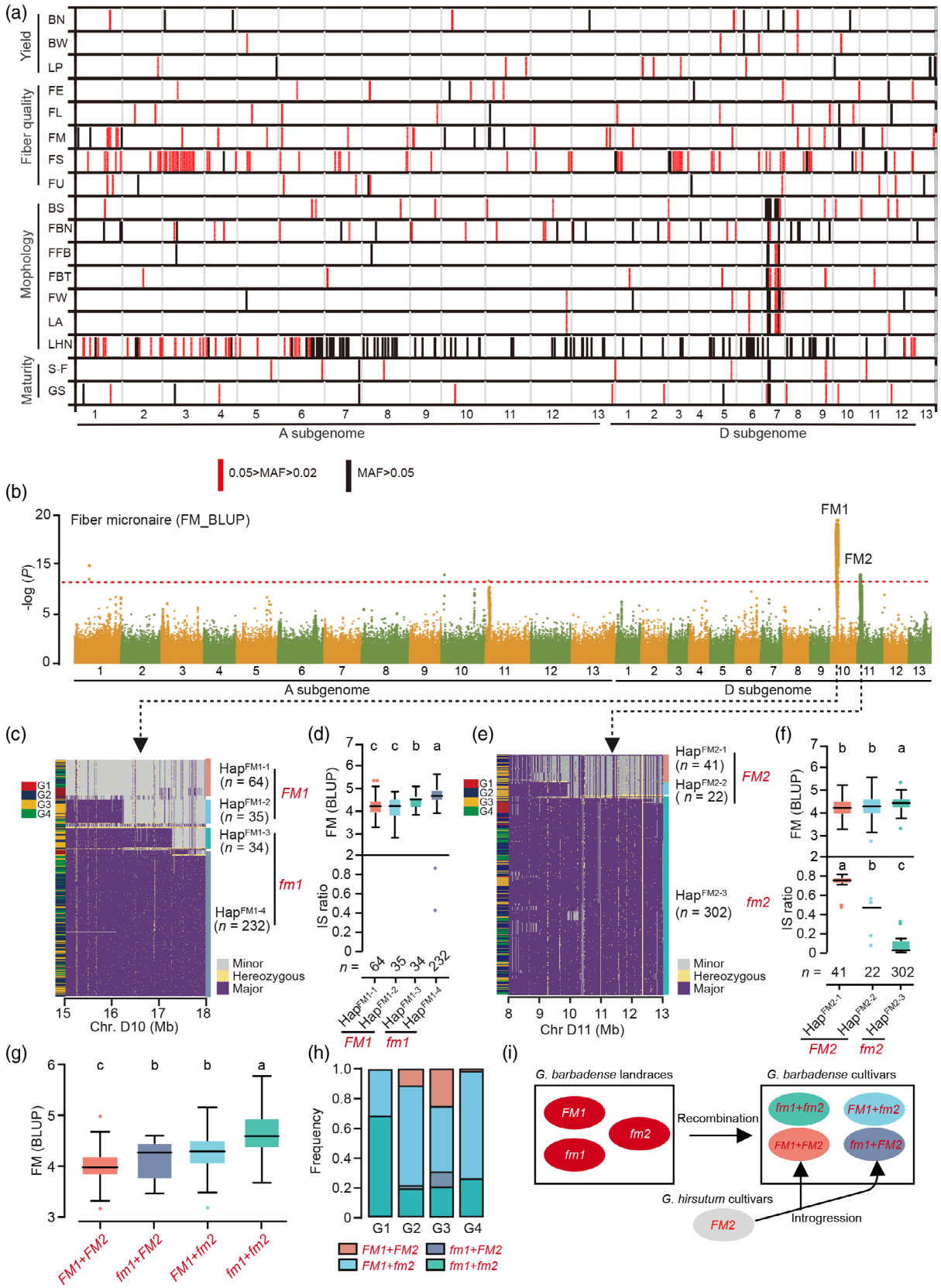
Whole genome analysis of important agronomic traits in *G. barbadense*

To dissect the genetic basis of agronomically important traits, we noted 17 fiber-, yield-, maturity-, and morphology-related traits of 326 cultivated *G. barbadense* accessions in four environments. The best linear unbiased prediction (BLUP) values (for multiple environmental traits) were used to perform GWAS with 3 797 297 SNPs (MAF > 0.05, missing rate < 20%). A total of 13 803 significant SNPs ($-\log(P) > 6.88$) were identified (Figure 3a; Table S22). Among the 17 traits examined, we detected the greatest number of significant SNPs for LHN (9813 SNPs) and FM (4362 SNPs), which were located on chromosomes A06 (9642 SNPs), D10 (2991 SNPs), and D11 (1289 SNPs) (Figure 3a; Table S22). A total of 6088 rare significant variations ($0.02 < \text{MAF} < 0.05$, $-\log(P) > 6.88$) were detected in our study (Figure 3a), distributed on nearly all chromosomes. This result indicated that many of the complex traits were associated with rare variations rather than common variations, which is consistent with previous studies (Mägi et al., 2012; Wagner, 2013; Sazonovs and Barrett, 2018; Luo et al., 2020). Our results also highlight the potential of GWAS to explore genes associated with rare variants. Moreover, many new loci associated with fiber quality, especially FS, have been identified. This rare mutation will provide a better understanding of the regulatory mechanisms of fiber quality.

Genetic source and potential improvement of FM in *G. barbadense* cultivars

FM is one of the most vital properties for spinning high-quality textiles (Bradow and Davidonis, 2000; Seagull et

Figure 3. Genomic distribution of all significant genome-wide association study (GWAS) signals ($-\log(P) > 6.88$) and genetic basis of fiber micronaire (FM). (a) Genomic distribution of significant signals for all investigated traits in *G. barbadense*. BN, number of bolls per plant; FBT, fruit branch type; FFB, first fruit branch; FL, fiber length; FM, fiber micronaire; FS, fiber strength; FU, fiber uniformity; GS, growth stage; LA, leaf area; LHN, leaf hair number; LP, lint percentage. Each significant GWAS signal is represented by a blue vertical line. (b) Manhattan plots of GWAS for FM (best linear unbiased prediction [BLUP]). Two major signals were labeled as FM1 (chromosome D10) and FM2 (chromosome D11). Horizontal red dashed lines indicate the significance threshold ($-\log(P) > 6.88$) of GWAS. (c) Genotype heatmap of locus FM1, ranging from approximately 15 Mb to approximately 18 Mb on chromosome D10 (*x*-axis), with all the accessions ordered according to local single nucleotide polymorphism (SNP) clustering (*y*-axis). Colored labels on the left indicates the four groups, consistent with Figure 1a. A total of four haplotypes were identified in this locus (Hap^{FM1-1} to Hap^{FM1-4}) and were further merged into two, representing a favorable haplotype (*FM1*) and an inferior haplotype (*fm1*) according to their phenotypes (Figure 3d). (d) Boxplots showing the FM values (upper panel) and ISL ratios (bottom panel) of the four haplotypes. (e) Genotype heatmap of locus FM2, ranging from approximately 8 to approximately 13 Mb on chromosome D11 (*x*-axis), with all the accessions ordered according to local SNP clustering (*y*-axis). Colored labels on the left indicate the four groups, consistent with Figure 1a. A total of three haplotypes were identified in this locus (Hap^{FM2-1} to Hap^{FM2-3}) and further merged into two, representing a favorable haplotype (*FM2*) and an inferior haplotype (*fm2*) according to their phenotypes (Figure 3f). (f) Boxplots showing the FM values (upper panel) and ISL ratios (bottom panel) of three haplotypes. (g) Boxplots showing the FM values for four haplotype combinations (*FM1* + *FM2*, *fm1* + *FM2*, *FM1* + *fm2*, and *fm1* + *fm2*). (h) Frequencies of the four haplotype combinations in the four groups. (i) Schematic diagram showing the proposed model of origin and recombination for FM haplotypes. For the boxplots, the center lines, box limits, and whiskers indicate the median, the upper and lower quartiles, and the 1.5× interquartile range, respectively. Dots indicate the outliers. Significant differences were tested by a two-sided Wilcoxon test.



al., 2000; Rodgers et al., 2017). Understanding the genetic basis of FM is essential for enhancing the fiber fineness of modern *G. barbadense* cultivars. Our GWAS identified two stable FM-related QTLs in four environments located at approximately 15 to 18 Mb on chromosome D10 (FM1) and approximately 8 to 13 Mb (FM2) on chromosome D11 (Figure 3b; Figures S9 and S10). Haplotype block analysis indicated that FM1 could be divided into four haplotype blocks designated $Hap^{FM1\text{-}1}$ to $Hap^{FM1\text{-}4}$ (Figure 3c). According to the genotyping data, four haplotype blocks were grouped into two types, named *FM1* (representing favorable FM haplotypes with lower FM values) and *fm1* (with unfavorable haplotypes) (Figure 3d). We found that both $Hap^{FM1\text{-}1}$ and $Hap^{FM1\text{-}4}$ were present in G1, which indicated that they were primitive haplotype blocks, and these haplotype blocks further artificially recombined to form $Hap^{FM1\text{-}2}$ and $Hap^{FM1\text{-}3}$ during the breeding process (Figure 3c). Moreover, introgression analysis showed that none of these haplotype blocks were derived from *G. hirsutum* (Figure 3d, bottom). These data suggest that the fiber fineness-related haplotypes ($Hap^{FM1\text{-}1}$ to $Hap^{FM1\text{-}4}$) at approximately 15 to 18 Mb on chromosome D10 originated from primitive *G. barbadense*.

To identify the candidate genes underlying the QTLs, we selected two accessions carrying contrasting FM1 haplotypes (Figure S11a) and investigated a total of 84 genes located in the FM1 genomic region at critical stages of fiber development (0, 10, and 25 DPA) (Table S23). According to gene annotations and expression levels, *Gbar_D10G011110*, which encodes a filament-like plant protein (Oda et al, 2015), was significantly highly expressed at 25 DPA (the stage of fiber cell secondary wall thickening and maturation, which is essential for FM), indicating it is a candidate gene underlying FM1 (Figure S11b).

For FM2 on chromosome D11 (Figure 3b,e), we found that all the accessions could be divided into three haplotypes ($Hap^{FM2\text{-}1}$ to $Hap^{FM2\text{-}3}$) (Figure 3e). $Hap^{FM2\text{-}1}$ and $Hap^{FM2\text{-}2}$ were identified as favorable haplotypes (*FM2*) with lower FM values (Figure 3f, top). Interestingly, the median IS ratio for *FM2* reached nearly 0.8 (Figure 3f, bottom), indicating that *FM2* might have been derived from *G. hirsutum*. Nearly all the accessions in G1 and G4 carried the normal haplotype block (*fm2*) (Figure 3e). This result demonstrated that a *G. hirsutum*-introgressed *haplotype block (FM2*) might enhance FM in a small subset of *G. barbadense* cultivars (most of the accessions were in G2 and G3) (Figure 3e). In the genomic region of *FM2*, a total of 411 genes were annotated (Table S24). Based on a comparison of the expression levels of these genes between two accessions carrying contrasting haplotype blocks (Figure S11c), *Gbar_D11G011390*, which encodes suppressor of gene silencing 3 (SGS3), was significantly highly expressed at 20 DPA in Xinhai 21 (*fm2*), and was predicted to be a candidate gene (Figure S11d).

To evaluate the combined effects of favorable haplotypes for FM traits in *G. barbadense*, we compared FM among accessions carrying multiple favorable allelic combinations (Figure 3g,h). We found that accessions carrying two favorable haplotypes (FM1 and FM2) showed better fineness than those carrying only one haplotype of FM1 or FM2 (Figure 3g), and G3 carried more FM-favorable haplotypes than other groups; however, G4 carried the least *FM2* (Figure 3h), which might explain why the FM trait of G4 was poor compared to those of other groups. In summary, we revealed that the excellent FM trait in modern *G. barbadense* originated from the recombination of FM haplotypes (*FM1*, *fm1*, and *fm2*) of *G. barbadense* landraces and the introgressed haplotype (*FM2*) from *G. hirsutum* (Figure 3i). This research also implies that breeders could improve FM by transferring the favorable haplotype of *FM2* into modern G4 cultivars to improve the quality *G. barbadense* varieties in the future.

Genetic basis of FS of *G. barbadense* on chromosome A03

FS is one of the most important fiber quality traits for *G. barbadense*. To understand the genetic basis of FS, rare mutations ($0.05 > MAF > 0.02$) and common ($MAF > 0.05$) variants were used together to identify FS-associated loci by GWAS. The results showed that 90 unique SNPs were associated with FS trait, and nearly all belonged to rare variations, except for three common SNPs (Table S25). Among the related rare mutation SNPs, the locus located from 5.60 to 6.50 Mb on chromosome A03 (Figure S12a) was considered as the most important. A high frequency of introgression (FI), nucleotide diversity (π), and population differentiation (F_{ST}) were identified in this region (Figure S12b–d), indicating that this region was related to the introgression event from *G. hirsutum*. Two haplotypes could be roughly categorized (Figure S12e,f), including the elite haplotype FS and the unfavorable haplotype named fs. Importantly, the SNP heatmaps indicated that the FS haplotype blocks originated from the primitive G1 group, and the fs haplotype was nearly the same as most of *G. hirsutum* (Figure S12e). This implies that these rare FS haplotypes already existed in the primitive *G. barbadense*. In addition, two candidate genes (Figure S12g,h) had higher expression levels at 25 DPA fiber in the accession carrying a favorable FS type (Xinhai 25). *Gbar_A03G004180*, which encodes ethylene-insensitive 3, acts as a positive regulator in the ethylene response pathway. The other gene, *Gbar_A03G004270*, encodes mitogen-activated protein kinase 3 (MAPK3), which also plays an important role in the ethylene response pathway. These two genes may work together to improve FS of the modern *G. barbadense* cultivars.

Identification of the haplotypes with LH and GS adapted for the high-latitude region

Gossypium barbadense cultivars are mainly grown in northwest China (mostly in Xinjiang Province), where they

are located in a high-latitude region that severely lacks precipitation (Figure 4g). To adapt to the local environment, the *G. barbadense* cultivars of Xinjiang have gained a series of unique characteristics, such as abundant leaf hairs and early maturity. LH is a beneficial trait responsible for increasing resistance to insects and pathogens, reflecting excess solar radiation and reducing water loss (Lee 1985; Wright et al., 1999; Lacape et al., 2005; Wan et al., 2014; Ding et al., 2015). In our study, the QTL genomic region highly associated with LH ranged from approximately 99.84 to approximately 107.49 Mb on chromosome A06 (Figure 4a). A previously cloned gene, *GbHD1*, that controls stem hair was recently uncovered in this region (103.25–103.26 Mb) (Ding et al., 2015) (Figure S13). A total of three haplotype blocks could be roughly categorized, including a hairy haplotype block ($Hap^{LH\text{-}1}$), a hairless haplotype block ($Hap^{LH\text{-}2}$), and an intermediate haplotype block ($Hap^{LH\text{-}3}$), which were mainly carried by the G4, G1, and G2/G3 groups, respectively. The hairy haplotype block might have originated from *G. barbadense* landraces (G1) (Figure 4b,c).

The maturity period is another adaptive trait in crops and is significantly influenced by the photoperiod (Fowler et al., 2001; Meyer and Purugganan, 2013; Song et al., 2013). GWAS signals for GS were found in the region ranging from approximately 14.91 Mb to 19.75 Mb on chromosome D07 (Figure 4d,e, Figure S14). The strongest signal was a non-synonymous mutation (SNP_15803388) located in the exon of a previously reported gene, *GbSP* (*Gbar_D07011870*), which was suggested to control the sympodial branch type and flowering time (Chen et al., 2015; Si et al., 2018). By analyzing the population genotype data, we found that the early maturity haplotype (*GS*) might have originated from an early Egyptian accession, cv. Ashmouni, developed in the 1860s (Smith et al. 1999). During the northward expansion of the cultivation region, this haplotype became predominant in G4 instead of the late-maturity haplotype (*gs*) (Figure 4e,f). Therefore, we concluded that GS was a domestication locus in *G. barbadense*, allowing it to adapt to long-day regions.

By examining the geographical distribution of major *G. barbadense* cultivation regions, we found that the selection of both LH and GS traits matched the changes in the planting regions of *G. barbadense*. In the last 200 years, due to decreasing precipitation and increasing daylength in new cultivation regions, such as Central Asia and Xinjiang in China, the enrichment of hairy and early maturity haplotypes in cultivars could be beneficial for rapid adaptation to water shortages and long-day environments (Figure 4g).

GWAS for plant architecture and vegetative growth-related traits of *G. barbadense* and their pleiotropic effects

In addition to high fiber quality and disease resistance, stronger vegetative growth habit is also an important characteristic trait for *G. barbadense*, which was distinguished from *G. hirsutum*. In this study, we found that all three maturity-related traits, including the first fruit branch (FFB), GS, and flowering stage (S-F), were positively correlated with each other, and the vegetative growth-related and plant architecture traits, including fruit branch number (FBN), fruit branch type (FBT), leaf area (LA), and fresh leaf weight (FW), were also positively correlated with each other (Table S26). Moreover, all maturity-related traits were significantly negatively correlated with vegetative growth-related traits, which is consistent with previous studies in other crops (Wu et al., 2014; Wang et al., 2018). GWAS results showed that the GS locus had a pleiotropic effect on the sympodial branch type (FBT), other maturity-related traits (FFB, S-F), and the vegetative growth-related traits (FBN, LA, and FW) (Figure 3a, Figure S15). All these traits, including plant type, GS, and leaf traits, were significantly correlated with one non-synonymous SNP (SNP_15 803 388) on *Gbar_D07G011870* (Figure 4d, Figure S15), which was previously cloned as a candidate gene for multiplex-type branching (Chen et al., 2015; Si et al., 2018). This result demonstrated that the *GbSP* gene (*Gbar_D07011870*), which might also affect plant architecture, similar to some *FT* genes in rice (Chardon and Damerval, 2005; Hedman et al., 2009; Huang et al., 2010), had pleiotropic effects on plant type, GS, and leaf traits. This pleiotropic effect on plant architecture and vegetative growth habits may also result in the limited population divergence and adaptation of *G. barbadense*.

DISCUSSION

Gossypium barbadense was previously categorized into three types, including Egypt, Pima, and Central Asia, according to source and geographical distribution. In this study, the whole population was categorized into four genetic groups based on deep resequencing of 365 *G. barbadense* accessions. We found no apparent genomic differences among obsolete cultivars from Egypt, Pima, and Central Asia (Figure S1b), suggesting that *G. barbadense* should be reclassified based on genomic data. The new four-group classification, including landraces (G1), obsolete cultivars (G2), improved cultivars (G3), and modern cultivars (G4), was more consistent with the breeding history of *G. barbadense* (from early to late). Notably, we also reported for the first time the genotype of a group of perennial *G. barbadense* landraces (G1) collected in the mountainous regions of southern China. Their genomes were similar to those of two primitive accessions collected from South America, implying a much earlier history of Chinese *G. barbadense* introduction (during the Age of Exploration) than previously documented (during 1920s–1950s) (Kong 2002; He 2004). Moreover, we found that the genotype of the G1 group differed from those of the other groups across the entire genome (Figure S1a), and thus, G1 could greatly expand the genetic diversity of *G. barbadense*.

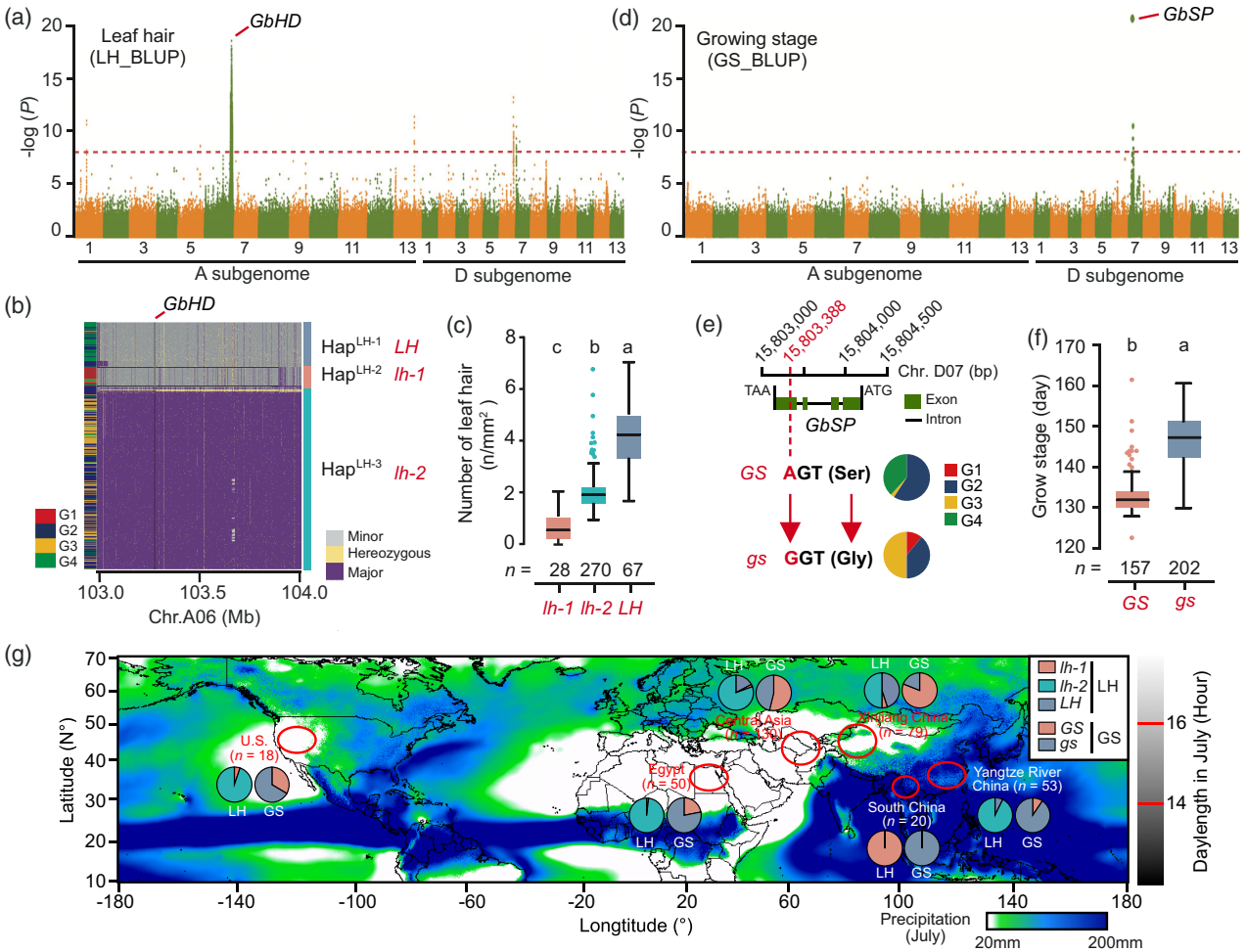


Figure 4. Genetic basis of two representative adaptive loci (leaf hair number and maturity period) in *G. barbadense*. (a) Manhattan plots of GWAS for leaf hair number (LH_BLUP). Horizontal red dashed lines indicate the significance threshold ($-\log_{10}(P) > 6.88$) of the GWAS. The position of the cloned gene *GbHD* is marked. (b) Genotype heatmap of locus LH, ranging from approximately 103 Mb to approximately 104 Mb on chromosome A06 (*x*-axis), with all the accessions ordered according to local SNP clustering (*y*-axis). Colored labels on the left indicates the four groups, consistent with Figure 1a. (c) A total of three haplotypes were identified in this locus (Hap^{LH-1} to Hap^{LH-3}) and the remaining were classified into three according to their phenotypes: a hairy haplotype (*LH*), an intermediate haplotype (*lh-2*), and a hairless haplotype (*lh-1*). (d) Manhattan plots of GWAS for growing stage (GS_BLUP). Horizontal red dashed lines indicate the significance threshold ($-\log_{10}(P) > 6.88$) of the GWAS. The position of the cloned gene *GbSP* is marked. (e) Gene model of *GbSP* and the causal mutation. The location of the strongest GWAS signal (D07:15803388, $-\log(P) = 22.43$) for GS (BLUP) is marked by a vertical red dashed line. Amino acid mutations and the frequencies of the two haplotypes (*GS* and *gs*) in the groups are shown. (f) Boxplot showing the growing stages of two haplotypes. The center lines, box limits, and whiskers indicate the median, the upper and lower quartiles, and the 1.5× interquartile range, respectively. Dots indicate the outliers. Significant differences were tested by a two-sided Wilcoxon test. (g) Geographic distribution and frequencies of haplotypes for LH and GS. Precipitation data are shown using the averages from 1979–2013 in July (Karger et al., 2017). Daylength (hours) in July is shown by a gradient from black to white on the right side of the graph.

Owing to the lack of any reproductive barrier between *G. hirsutum* and *G. barbadense*, abundant interspecific elite lines have been developed over the past decades. Genomic exchange has been found to occur between these two species on multiple occasions (Kohel et al., 2001; Zhang et al., 2002; Paterson et al., 2003; Lacape et al., 2005; He et al., 2007; Lacape et al., 2009; Fang et al., 2017). In a recent report, *G. barbadense* averaged 6.8 Mb of the total introgressed sequence per cultivar from *G. hirsutum* and only 5% of the genes under selection were located in introgressed regions (Yuan et al., 2021). Our study identified sequences with a total length of approximately 158.5 Mb in the *G. barbadense* genome (covering approximately 7.4% of the genome) that might have been derived from *G. hirsutum*. We also found that the current *G. barbadense* population was strongly affected by these introgressed fragments. Unexpectedly, we found that only a few of the investigated agronomic trait-related QTLs, except those for FM (D11) and FS (A03), overlapped with the ISs of *G. hirsutum*. This result was inconsistent with previous findings, which indicated that *G. hirsutum* contributed yield-related loci to *G. barbadense*. Therefore, an in-depth analysis of

the relationships between the genes within the ISs and improved traits in *G. barbadense* should be conducted.

Recently, many studies on introgression events between *G. hirsutum* and *G. barbadense* have been reported. By constructing recombinant inbred lines with *G. barbadense* introgressions, eight stable QTLs were identified, including qFL-A03-1, qFL-D07-1, and qFL-D13-1 (Wang et al., 2020b). As the reference genomes were different, no similar results were identified between this study and our study. Through sequencing and GWAS analysis of 229 *G. barbadense* accessions and 491 *G. hirsutum* accessions, many loci associated with agronomic traits have been identified (Fang et al., 2021). Among them, the pleotropic loci associated with FL, FS, and FU on chromosome A03 were close to the rare mutation SNPs associated with FS in our study (Figure S12).

Combining the whole genome resequencing and GWAS, several candidate genes associated with FS and LP were identified (Yuan et al., 2021). The loci associated with FS on chromosome D11 were close to our results (Table S26). However, in our study, the loci associated with FS on chromosome D11 could only be identified in the environment HN14. This FS-associated locus was also reported by GWAS of other *G. barbadense* populations (Zhao et al., 2021). One non-synonymous SNP significantly correlated with FS.

Previous studies showed that environmental change, especially as it relates to maturity, has a significant influence on FM (Verhalen et al., 1975; Bradow and Davidonis, 2000), causing difficulties in the precise mapping of FM-related QTLs in both cultivated *G. hirsutum* and *G. barbadense*. In the present study, we identified two stable FM-related haplotype blocks on chromosomes D10 (*FM1*) and D11 (*FM2*). Interestingly, the favorable haplotype *FM2* was derived from *G. hirsutum*, another cultivated tetraploid cotton with inferior fiber quality. Although most modern Xinjiang *G. barbadense* cultivars (G4) have well-integrated favorable traits, such as early maturity and excellent FL and FS, FM was one of their weakest traits (Figure S1c). The low frequency of *FM2* in modern cultivars (G4) might explain the poor FM of G4. In the future, incorporating haplotype *FM2* could fill this gap in modern *G. barbadense* cultivars. Meanwhile, we can also consider transforming the favorable fiber fineness genes located in the favorable haplotype *FM1* region in *G. barbadense* to *G. hirsutum* and thereby improving the fiber fineness of *G. hirsutum*.

The genetic diversity and environmental adaptability of *G. barbadense* were worse than those of *G. hirsutum*. Therefore, it is usually planted in a limited environment. In China, it is mainly planted in the Alar region of Xinjiang, Hainan, Yunnan, and other southern regions. This study found that the plant type traits (FBN and FBT), growth period-related traits (FFB, GS, and S-F), and vegetative growth-related traits (LA and FW) of *G. barbadense* were controlled by pleiotropic genes such as *GbSP*, which was previously also reported to function in nulliplex-type branching. It is also an important pleiotropic gene responsible for the decrease in genetic diversity and strong consistency in the plant growth habits of *G. barbadense*. These characteristics indicate that *G. barbadense* has strong vegetative growth habits and a longer growth period. It is not difficult to understand that a longer fiber growth period and higher nutrient accumulation result in better fiber quality. Therefore, this study preliminarily clarified the reasons behind the poor adaptability and good fiber quality of *G. barbadense* based on the development-related traits. In our study, we found that the pleiotropic gene *GbSP* plays an important role in the plant type traits, growth period, and vegetative growth of *G. barbadense*. We believe that *GbSP* improves the plant architecture and changes the GS from vegetative to reproductive growth, so if it is transferred to *G. hirsutum*, it will affect nutrient accumulation in the fibers. This may increase the fiber quality for *G. hirsutum*.

Collectively, our study sheds light on the breeding history of *G. barbadense* through genome-wide introgression and association analyses. The identification of favorable haplotypes could accelerate the production of high-quality *G. barbadense* cultivars and provide crucial data for the cotton research community.

METHODS

Sampling, DNA extraction, and resequencing

All 365 *G. barbadense* accessions used in this study were preserved in the China National Gene Bank, Institute of Cotton Research, Chinese Academy of Agriculture Sciences (Anyang, China), after collection from the main cotton cultivation regions worldwide, including China, South America, the United States, Egypt, and Central Asia (Table S1). For DNA extraction, we planted all the seeds in a greenhouse, collected the young leaves from a single seedling of each accession, and froze them at –80°C immediately. Genomic DNA was extracted following the CTAB method with some modifications (Paterson et al., 1993). The 500-bp DNA libraries were constructed according to the protocol provided by Illumina and sequenced on a HiSeq X Ten platform (Illumina, San Diego). A total of approximately 3898 Gb raw data were produced, with an average coverage depth of 15.6× for each accession.

Sequence alignment and variation calling

After obtaining raw reads from each accession, the following reads were removed: (i) reads aligned to adaptors; (ii) reads with ≥10% unidentified nucleotides; and (iii) reads with >50% bases having a Phred score of <10. The remaining high-quality reads were aligned against the reference genome of *G. barbadense* using the Burrows–Wheeler

Aligner program (ver. 0.7.12) with default settings (Wang et al., 2019; Li and Durbin, 2010). SAMtools (ver. 1.1) was used to generate consensus sequences for each accession and prepare input data for SNP calling using the Genome Analysis Toolkit (GATK, ver. 3.2–2), based on Bayesian estimation (Li et al., 2009; McKenna et al., 2010). To ensure the accuracy of SNP calling, some pre-treatments were performed, including marking duplicates with Picard software (ver. 1.119), local realignment, and base recalibration with GATK. The remaining SNPs called by GATK were used for SNP annotation using SnpEff software (Cingolani et al., 2012) (ver. 5.0), including SNP location and effect impact.

Population genetic analysis

A subset of 252 609 SNPs screened with missing rate < 20%, MAF > 0.05, and r^2 (squared correlation of adjacent SNPs) < 0.2 was used for phylogenetic and population structure analysis ($n = 370$). A neighbor-joining tree was built in FastTreeMP (Price et al., 2009) (ver. 2.19). The population structure was investigated with the software admixture_linux (ver. 1.23) (Alexander et al., 2009). The LD coefficient (r^2) between pairwise high-quality SNPs was calculated using Plink (ver. 1.07) (Purcell et al., 2007) with the parameters '--ld-window-r2 0 --ld-window 99 999 --ld-window-kb 1000'.

Genetic diversity analysis

The population fixation statistics (F_{ST}) among the subgroups and the nucleotide diversity (π) of each subgroup were calculated using VCFtools (ver. 0.1.12b) with 1-Mb sliding windows and a step size of 100 kb (Danecek et al., 2011). The regions with the top 5% of the F_{ST} and π values for each comparison were selected as the candidate high-divergence and high-diversity regions, respectively.

Identification of ISs

To identify the ISs derived from *G. hirsutum* in the *G. barbadense* genomes, we first used resequencing data from a total of 800 accessions, including 365 *G. barbadense*, six *Gossypium darwinii* (the wild relative of *G. barbadense*; Fryxell 1965), and 429 *G. hirsutum* (PRJNA399050) accessions, to call SNPs against the *G. barbadense* genome (Wang et al., 2019). A simplified workflow for identifying introgressed *G. barbadense* fragments in *G. hirsutum* has been reported (He et al., 2021). In this study, the workflow was used to identify introgressed *G. hirsutum* fragments in *G. barbadense* (Figure S16).

To confirm their genetic relationships, a phylogenetic tree of all accessions was constructed using FastTreeMP (Price et al., 2009) (ver. 2.19) (Figure S1a). Based on phylogenetic tree analysis, all 800 accessions were divided into two populations which well matched with two species, including the *G. hirsutum* population (*Gh*-POP, $n = 429$), the *G. barbadense* population (*Gb*-POP, $n = 365$), and six accessions of *G. darwinii*. First, 500 SNPs were considered as a bin. Second, the allele frequencies at each SNP site were calculated. For a single SNP (i.e., A/C), we calculated the allele frequency in *G. hirsutum* (F_C/F_A) (Figure S16a) as follows:

$$F_C = \frac{2 * N\_CC + N\_CA}{2 * n},\ FA = \frac{2 * N\_AA + N\_CA}{2 * n},$$

where $n = 429$, FC is the frequency of base C in *G. hirsutum*, FA is the frequency of base A in *G. hirsutum*, N_CC is the number of haplotype CC in *G. hirsutum*, N_CA is the number of haplotype CA in *G. hirsutum*, and N_AA is the number of haplotype AA in *G. hirsutum*.

In *G. barbadense*, fc/fa (Figure S16b) was calculated as follows:

$$fc = \frac{2 * N\_CC + N\_CA}{2 * n},\ fa = \frac{2 * N\_AA + N\_CA}{2 * n},$$

where $n = 365$, fc is the frequency of base C in *G. barbadense*, fa is the frequency of base A in *G. barbadense*, N_CC is the number of haplotype CC in *G. barbadense*, N_CA is the number of haplotype CA in *G. barbadense*, and N_AA is the number of haplotype AA in *G. barbadense*.

Third, the fragment similarity value for any given *G. barbadense* accession (P) was calculated based on 500 continuous SNP loci. The formula for the candidate accession (CC/AA/GT/. . ./GG) is as follows (Figure S16c):

$$P = \frac{2 * (FC - fc) + 2 * (FA - fa) + [(FG - fg) + (FT - ft)] + \ldots + 2 * (FG - fg)}{2 * n},$$

where $n = 500$; FA, FC, FG, and FT are the frequencies of different bases at SNP in *G. hirsutum*; and fa, fc, fg, and ft are the frequencies of different bases at SNP in *G. barbadense*. The bins with 500 SNPs and $P > 0.2$ were defined as ISs from *G. hirsutum* to *G. barbadense*.

After obtaining the ISs, the genotype of the IS bin was transferred to the normal type. Then, with the help of TASSEL, the rare bin with MAF < 0.05 was filtered. By LD block analysis, all the IS bins (MAF > 0.05) were clustered into 315 IS regions (Table S9). Each IS region was genotyped based on different agronomic traits (Table S4). The genetic effects of the main IS regions associated with FS (Table S12), FL (Table S14), FM (Table S16), BW (Table S18), and LP (Table S20) in different environments were calculated according to the following formula. For example, the effect of the IS region IS-A01-44 associated with FS (Table S12) = (FS − fs)/fs, where FS is the average FS of the accessions carrying the IS-A01-44 region in certain environments and fs is the average FS of the accessions with no introgression of IS-A01-44. A significance test was performed for all the calculations.

Phenotyping and statistical analysis

In this study, we investigated three yield-related traits, including boll number (BN), BW, and LP (%); five fiber

quality-related traits, including fiber elongation rate (FE, %), FL (mm), FM, FS (cN/tex), and fiber uniformity (FU, %); seven morphology-related traits, including boll setting (BS), FBN, FFB, FBT, FW, LA (cm^2), and LHN (count/mm^2); and two maturity-related traits, including flowering stage (S-F, day) and GS (day); (Table S3). All accessions were planted in three locations, including Akesu (41°15′N, 80°29′E, Xinjiang Province, China), Sanya (18°14′N, 109°31′E, Hainan Province, China), and Anyang (36°08′N, 114°48′E, Henan Province, China), with three replicates in each location. All the yield-, fiber quality-, and maturity-related traits were investigated in two locations for 3 years, and the morphology-related traits were investigated in Anyang in 2016 (Table S3). Field management, including watering, weed management, and fertilization, was performed according to the usual local practices in each location during the growing period. GS was calculated from the sowing day to the day that approximately 50% of the bolls had opened in one block. FBN, FBT, FFB, BN, FW, and LA were measured using approximately 30 randomly selected individuals for each block. The second, third, and fourth leaves of each plant were harvested to measure LA with an LI-3000C portable leaf area meter (LI-COR, Lincoln). LH was counted using an Olympus ix71 inverted microscope (Olympus, Tokyo). A total of 30 bolls were harvested from each block to measure BN, BW, and LP, and then the fiber was ginned for quality testing. All fiber quality-related traits were evaluated using an HFT900 instrument (Premier Evolvics Pvt. Ltd., Coimbatore) at the cotton quality testing center of the Ministry of Agriculture (Anyang, China). The BLUP value for each accession across the three environments was calculated using the R library 'lme4'. All data statistics, including Student's two-tailed *t*-test and one-way analysis of variance (ANOVA), were calculated by GraphPad Prism 8 (ver. 8.4.3).

GWAS

GWAS was conducted using Efficient Mixed-Model Association eXpedited (EMMAX) software (Kang et al., 2010) based on 3 797 297 SNPs (MAF > 0.05, missing rate < 20%) and 6 300 596 SNPs (MAF > 0.02, missing rate < 20%). The phenotypic contribution of each locus was calculated using ANOVA. In this study, we used a significance cutoff of $-\log_{10}(P) > 6.88$ and $-\log_{10}(P) > 7.10$ (MAF > 0.02) ($P = 0.5/n$, where n is the effective number of SNPs) to define the signal threshold for GWAS.

Haplotype block identification and haplotyping on chromosomes A03, A06, D10, and D11

The GWAS loci (haplotype blocks) were identified by the following steps: (i) The chromosome-wide LD blocks on chromosomes A03, A06, D10, and D11 were identified by TASSEL 5 (ver. 20200709) with an average r^2 of SNP pairs greater than 0.6. (ii) Any overlapping region between the LD blocks and the GWAS signal region ($-\log(P) > 6.88$) was identified as a GWAS locus. We then constructed a local phylogenetic tree based on the SNPs in each GWAS locus to generate a genotype heatmap. The haplotypes of each locus were roughly categorized according to the heatmap (Dai et al., 2020).

Transcriptome analysis

Four *G. barbadense* accessions (cv. Xinhai 25, cv. Maorad, cv. Hai7124, and cv. Xinhai21) with different FM values and haplotypes were used to identify the expression patterns of candidate genes located in the GWAS loci on chromosomes D10 and D11. Cotton ovules of cv. Xinhai 25, which had a favorable haplotype (*FM1*, FM_BLUP = 3.89; *FS*, FS_BLUP = 45.87), and cv. Maorad, which had an inferior haplotype (*fm1*, FM_BLUP = 4.66; FS_BLUP = 29.42), were sampled at 0, 10, and 25 DPA. Cotton ovules of cv. Hai7124, which had a favorable haplotype (*FM2*, FM_BLUP = 4.04), and cv. Xinhai21, which had an inferior haplotype (*fm2*, FM_BLUP = 4.64), were sampled at −1, 5, 10, 15, and 20 DPA.

Four *G. barbadense* accessions (cv. 3-79, cv. E24-3389, cv. Hai7124, and cv. Xinhai21) with different IS regions were used to identify candidate genes associated with fiber quality and yield. The four accessions were sampled from leaf, flower, −1, 5, 10, and 20 DPA ovules, and 10, 15, and 20 DPA fibers.

Total RNA was extracted following the instructions of the RNA Isolation Kit (TIANGEN Biotech, Beijing). The RNA sequencing libraries were prepared as described previously and sequenced on an Illumina HiSeq X Ten platform (Zhong et al., 2011). After removal of the adaptors and low-quality data, the clean reads were mapped against the *G. barbadense* genome (Wang et al., 2019) using HISAT2 (Kim et al., 2015) (ver. 2.2.0), and gene expression levels were calculated using StringTie (Pertea et al., 2015) (ver. 2.13).

Climate data

Climate data for the earth land surface area in July (full-blooming stage for cotton) from 1979 to 2013 were downloaded from CHELSA (Climatologies at High Resolution for the Earth's Land Surface Areas, www.chelsa-climate.org) according to the GPS coordinates of the original collection sites (Karger et al., 2017).

ACKNOWLEDGMENTS

The work was supported by the National Key Technology R&D Program, the Ministry of Science and Technology (2021YFF1000101-1, 2016YFD0100203, 2016YFD0101413, and 2016YFD0100306), Major Science and Technology Projects in Henan Province (161100510100), and the Agricultural Science and Technology Innovation Program of Chinese Academy of Agricultural Sciences.

AUTHOR CONTRIBUTIONS

PW, SH, XZ, QW, and XD conceived and designed the research. DN, XG, YJ, and XD collected materials. PW, YJ, ZP, HL, CW, and LW contributed to phenotyping. PW, ZP, ML, and HZ performed whole genome resequencing data

production. GS and SH performed GWAS, population diversity, and introgression analyses. PW, MW, XG, WW, and QY worked on data analysis. PW, MW, XG, SH, and XD worked on figure design and wrote the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The raw resequencing data and transcriptome sequencing reads generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA637990.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Phylogenetic tree, geographic distribution, and traits analysis of the accessions.

Figure S2. Single nucleotide polymorphism (SNP) heatmaps of the 800 accessions used to identify the introgression segments (ISs) from chromosomes A01 to A13.

Figure S3. SNP heatmaps of the 800 accessions used to identify the ISs from chromosomes D01 to D13.

Figure S4. Introgression from *Gossypium hirsutum* leads to the genetic divergence of *Gossypium barbadense* from chromosomes A01 to A08.

Figure S5. Introgression from *G. hirsutum* leads to the genetic divergence of *G. barbadense* from chromosomes A09 to D01.

Figure S6. Introgression from *G. hirsutum* leads to the genetic divergence of *G. barbadense* from chromosomes D02 to D13.

Figure S7. Genetic basis of the IS regions associated with fiber strength (FS) in *G. barbadense*.

Figure S8. Genetic basis of the IS regions associated with fiber length (FL), fiber micronaire (FM), boll weight (BW), and lint percentage (LP) in *G. barbadense*.

Figure S9. Manhattan plots of FM in four environments: best linear unbiased prediction (BLUP), HN2014, XJ2015, and XJ2016.

Figure S10. Linkage disequilibrium (LD) regions associated with FM on chromosomes D10 and D11.

Figure S11. Candidate genes associated with the FM trait on chromosomes D10 and D11.

Figure S12. Genetic basis of FS.

Figure S13. Strong signal regions associated with leaf hair number (LHN) on chromosome A06.

Figure S14. Manhattan plots of growth stage (GS) in four environments.

Figure S15. Pleotropic regions of plant architecture and vegetative growth habit traits on chromosome D07.

Figure S16. Identification of ISs from *G. hirsutum* to *G. barbadense*.

Table S1. Summary of all accessions sequenced in this study.

Table S2. Statistics of calling variations in all accessions.

Table S3. Description of all agronomical traits investigated in this study.

Table S4. Summary of all traits used in the genome-wide association study (GWAS).

Table S5. Statistics of all the phenotypic data.

Table S6. Genomic regions and genes that showed geographic divergence in the comparison of group G2 versus group G3 (F_{ST} > 0.0595).

Table S7. Genomic regions and genes that showed geographic divergence in the comparison of group G2 versus group G4 (F_{ST} > 0.0427).

Table S8. Genomic regions and genes that showed geographic divergence in the comparison of group G3 versus group G4 (F_{ST} > 0.1356).

Table S9. The number and rate of accessions carrying the introgression segments (ISs) in the four genotype groups in 315 introgression loci.

Table S10. Statistical analysis of the introgression region in each group.

Table S11. Comparison of the overlap region length of F_{ST} and ISs.

Table S12. Effects of the 11 ISs associated with fiber strength (FS) in different environments.

Table S13. Expression patterns of the genes located at the IS region associated with FS.

Table S14. Effect of the nine ISs associated with fiber length (FL) in different environments.

Table S15. Expression pattern of the genes located at the IS region associated with FL.

Table S16. Effect of the seven ISs associated with fiber micronaire (FM) in different environments.

Table S17. Expression patterns of the genes located at the IS region associated with FM.

Table S18. Effect of the nine ISs associated with boll weight (BW) in different environments.

Table S19. Expression patterns of the genes located at the IS region associated with BW.

Table S20. Effect of the seven ISs associated with lint percentage (LP) in different environments.

Table S21. Expression patterns of the genes located at the IS region associated with LP.

Table S22. Significant single nucleotide polymorphisms (SNPs) ($-\log(P) > 6.88$) identified in this study by GWAS.

Table S23. Expression profiles of the genes located at the region on chromosome D10 associated with FM.

Table S24. Expression profiles of the genes located at the region on chromosome D11 associated with FM.

Table S25. Significant SNPs ($-\log(P) > 6.88$) associated with FS identified in this study by GWAS.

Table S26. Pearson's correlation coefficients among fruit branch type (FBT), fruit branch number (FBN), first fruit branch (FFB), growth stage (GS), flowering stage (S-F), leaf area (LA), and fresh leaf weight (FW).

REFERENCES

Alexander, D., Novembre, J. & Lange, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655–1664. https://doi.org/10.1161/01.ATV.0000137190.63214.c5

Bradow, J.M. & Davidonis, G. (2000) Quantitation of fiber quality and the cotton production-processing interface: a physiologist's perspective. *Journal of Cotton Science*, **4**, 34–64.

Chardon, F. & Damerval, C. (2005) Phylogenomic analysis of the *PEBP* gene family in cereals. *Journal of Molecular Evolution*, **61**, 579–590. https://doi.org/10.1007/s00239-004-0179-4

Chen, W., Yao, J., Chu, L., Yuan, Z., Li, Y., Zhang, Y. *et al.* (2015) Genetic mapping of the nulliplex-branch gene (*gb_nb1*) in cotton using next-

generation sequencing. *TAG. Theoretical and Applied Genetics.*, **128**, 539–547. https://doi.org/10.1007/s00122-014-2452-2

Chen, Y.U., Liu, G., Ma, H., Song, Z., Zhang, C., Zhang, J. *et al.* (2018) Identification of introgressed alleles conferring high fiber quality derived from *Gossypium barbadense L.* in secondary mapping populations of *G. hirsutum L. Frontiers in Plant Science*, **9**, 1023. https://doi.org/10.3389/fpls.2018.01023

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L. *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80–92. https://doi.org/10.4161/fly.19695

Dai, P., Sun, G., Jia, Y., Pan, Z., Tian, Y., Peng, Z. *et al.* (2020) Extensive haplotypes are associated with population differentiation and environmental adaptability in Upland cotton (*Gossypium hirsutum*). *Theoretical and Applied Genetics*, **133**, 3273–3285. https://doi.org/10.1007/s00122-020-03668-z

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Ding, M., Ye, W., Lin, L., He, S., Du, X., Chen, A. *et al.* (2015) The hairless stem phenotype of cotton (*Gossypium barbadense*) is linked to a copia-like retrotransposon insertion in a homeodomain-leucine zipper gene (*HD1*). *Genetics*, **201**, 143–154. https://doi.org/10.1534/genetics.115.178236

Draye, X., Chee, P., Jiang, C., Decanini, L., Delmonte, T., Bredhauer, R. *et al.* (2005) Molecular dissection of interspecific variation between Gossypium hirsutum and Gossypium barbadense (cotton) by a backcross-self approach: II. Fiber Fifineness. *Theoretical and Applied Genetics*, **111**, 764–771. https://doi.org/10.1007/s00122-005-2061-1

Fang, L., Gong, H., Hu, Y., Liu, C., Zhou, B., Huang, T. *et al.* (2017) Genomic insights into divergence and dual domestication of cultivated allotetraploid cottons. *Genome Biology*, **18**, 33. https://doi.org/10.1186/s13059-017-1167-5

Fang, L., Zhao, T., Hu, Y., Si, Z., Zhu, X., Han, Z. *et al.* (2021) Divergent improvement of two cultivated allotetraploid cotton species. *Plant Biotechnology Journal*, **19**, 1325–1336. https://doi.org/10.1111/pbi.13547

Feaster, C.V. & Turcotte, E.L. (1962) Genetic basis for varietal improvement of Pima cottons. *USDA-ARS*, **34**, 31.

Feaster, C.V., Turcotte, E.L. & Young, E.F. (1967) Pima cotton varieties for low and high elevations. *USDA-ARS*, **34**, 90.

Fowler, D.B., Breton, G., Limin, A.E., Mahfoozi, S. & Sarhan, F. (2001) Photoperiod and temperature interactions regulate low-temperature-induced gene expression in barley. *Plant Physiology*, **127**, 1676–1681. https://doi.org/10.1104/pp.127.4.1676

Fryxell, P.A. (1965) Stages in the evolution of *Gossypium L. Advanced Frontiers Plant in Science*, **10**, 31–56.

He, D., Lin, Z., Zhang, X., Nie, Y. & Guo, X. (2007) QTL mapping for economic traits based on a dense genetic map of cotton with PCR-based markers using the interspecific cross of *Gossypium hirsutum* × *Gossypium barbadense*. *Euphytica*, **153**, 181–197. https://doi.org/10.1007/s10681-006-9254-9

Hedman, H., Kallman, T. & Lagercrantz, U. (2009) Early evolution of the *MFT-like* gene family in plants. *Plant Molecular Biology*, **70**, 359–369. https://doi.org/10.1007/s11103-009-9478-x

He, L. (2004) *Studies on pedigree relative and agronomy character evolution and its space-time distribution of economic traits in sea island cotton (G. barbadense L.)* in South Xinjjiang. Yangling: Northwest A&F University Press.

He, S., Sun, G., Geng, X., Gong, W., Dai, P., Jia, Y. *et al.* (2021) The genomic basis of geographic differentiation and fiber improvement in cultivated cotton. *Nature Genetics*, **53**, 916–924. https://doi.org/10.1038/s41588-021-00844-9

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q.I., Zhao, Y. *et al.* (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics*, **42**, 961–967. https://doi.org/10.1038/ng.695

Hufford, M.B., Xu, X., van Heerwaarden, J., Pyhäjärvi, T., Chia, J.-M., Cartwright, R.A. *et al.* (2012) Comparative population genomics of maize domestication and improvement. *Nature Genetics*, **44**, 808–811. https://doi.org/10.1038/ng.2309

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

Jack, R., Dillehay, T.D. & Ugent, D. (1996) Ancient cultigens or modern intrusions? evaluating plant remains in an Andean case study. *Journal of Archaeological Science*, **23**, 391–407. https://doi.org/10.1006/jasc.1996.0035

Jiang, C., Wright, R.J., El-Zik, K.M. & Paterson, A.H. (1998) Polyploid formation created unique avenues for response to selection in *Gossypium* (cotton). *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 4419–4424. https://doi.org/10.1073/pnas.95.8.4419

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B. *et al.* (2010) Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, **42**, 348–354. https://doi.org/10.1038/ng.548

Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W. *et al.* (2017) Climatologies at high resolution for the earth's land surface areas. *Scientific Data*, **4**, 170122. https://doi.org/10.1038/sdata.2017.122

Kearney, T.H. (1943) *Egyptian-type cottons: their origin and characteristics*. Report of Division of Cotton and Other Fiber Crops and Diseases. USDA Mimeo (unnumbered).

Kerr, T. (1960) The potentials of barbadense cottons. In *Proceedings of the 12th Ann*, pp. 57–60.

Kim, D., Langmead, B. & Salzberg, S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, **12**, 357–360. https://doi.org/10.1038/nmeth.3317

Kohel, R.J., Yu, J., Park, Y.H. & Lazo, G.R. (2001) Molecular mapping and characterization of traits controlling fiber quality in cotton. *Euphytica*, **121**, 163–172. https://doi.org/10.1023/A:1012263413418

Kong, Q. (2002) The analysis of comparative advantage of production situation for island cotton (*G. barbadense L.*) in China. *China Cotton*, **29**, 19–22.

Lacape, J.M., Nguyen, T.B. & Courtois, B. (2005) QTL analysis of cotton fiber quality using multiple × backcross generations. *Crop Science*, **45**, 123–140. https://doi.org/10.2135/cropsci2005.0123a

Lacape, J.M., Jacobs, J. & Arioli, T. (2009) A new interspecific, *Gossypium hirsutum* × *G. barbadense*, RIL population: towards a unified consensus linkage map of tetraploid cotton. *Theoretical and Applied Genetics.*, **119**, 281–292. https://doi.org/10.1007/s00122-009-1037-y

Lee, J.A. (1985) Revision of the genetics of the hairiness-smoothness system of *Gossypium*. *Journal of Heredity*, **76**, 123–126. https://doi.org/10.1007/BF01873581

Li, H. & Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T. & Durbin, R. (2009) The Sequence Alignment/Map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lu, Q., Shi, Y., Xiao, X., Li, P., Gong, J., Gong, W. *et al.* (2017) Transcriptome analysis suggests that chromosome introgression fragments from Sea Island cotton (*Gossypium barbadense*) increase fiber strength in Upland cotton (*Gossypium hirsutum*). *G3-Genes Genom Genet*, **7**, 3469–3479. https://doi.org/10.1534/g3.117.300108

Luo, L., Shen, J., Zhang, H., Chhibber, A., Mehrotra, D.V. & Tang, Z.Z. (2020) Multi-trait analysis of rare-variant association summary statistics using MTAR. *Nature Communications*, **5**, 2850.

Mägi, R., Asimit, J.L., Day-Williams, A.G., Zeggini, E. & Morris, A.P. (2012) Genome-wide association analysis of imputed rare variants: application to seven common complex diseases. *Genetic Epidemiology*, **36**, 785–796.

McGowan, J.C. (1960) *History of extra-long staple cottons*. Tucson: The University of Arizona Press.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303. https://doi.org/10.1101/gr.107524.110

Meyer, R.S. & Purugganan, M.D. (2013) Evolution of crop species: genetics of domestication and diversification. *Nature Reviews Genetics*, **14**, 840–852. https://doi.org/10.1038/nrg3605

Nie, X., Wen, T., Shao, P., Tang, B., Nuriman-guli, A., Yu, Y.U. *et al.* (2020) High-density genetic variation maps reveal the correlation between asymmetric interspecific introgressions and improvement of agronomic traits in Upland and Pima cotton varieties developed in Xinjiang, China. *The Plant Journal*, **103**, 677–689. https://doi.org/10.1111/tpj.14760

Oda, Y., Iida, Y., Nagashima, Y., Sugiyama, Y. & Fukuda, H. (2015) Novel coiled-coil proteins regulate exocyst association with cortical microtubules in xylem cells via the conserved oligomeric golgi-complex 2 protein. *Plant and Cell Physiology*, **56**, 277–286. https://doi.org/10.1093/pcp/pcu197

Page, J.T., Liechty, Z.S., Alexander, R.H., Kimberly, C., Hulse, A., Hamid, A. *et al.* (2016) DNA sequence evolution and rare homoeologous conversion in tetraploid cotton. *PLoS Genetics*, **12**, e1006012. https://doi.org/10.1371/journal.pgen.1006206

Paterson, A.H., Saranga, Y., Menz, M., Jiang, C.X. & Wright, R. (2003) QTL analysis of genotype × environment interactions affecting cotton fiber quality. *Theoretical and Applied Genetics*, **106**, 384–396. https://doi.org/10.1007/s00122-002-1025-y

Paterson, A.H., Brubaker, C.L. & Wendel, J.F. (1993) A rapid method for extraction of cotton (*Gossypium Spp.*) genomic DNA suitable for RFLP or PCR analysis. *Plant Molecular Biology Reporter*, **11**, 122–127. https://doi.org/10.1007/BF02670470

Peebles, R.H. (1954) Current status of American-Egyptian cotton breeding. In *Proceedings of the 6th Cotton Imp. Conference*, pp, 1–8.

Percy, R.G. & Wendel, J.F. (1990) Allozyme evidence for the origin and diversification of *Gossypium barbadense L. Theoretical and Applied Genetics*, **79**, 529–542. https://doi.org/10.1007/BF00226164

Percy, R.G. & Turcotte, E.L. (1998) Registration of extra-long staple cotton germplasm, 89590 and 8810. *Crop Science*, **38**, 1409. https://doi.org/10.1109/22.954814

Percy, R.G. (2002) Registration of five extra-long staple cotton germplasm lines possessing superior fiber length and strength. *Crop Science*, **42**, 988. https://doi.org/10.2135/cropsci2002.0988

Percy, R.G. (2009) The worldwide gene pool of *Gossypium barbadense L.* and Its Improvement. In *Genetics and genomics of cotton*. US: Springer, pp. 53–68. https://doi.org/10.1007/978-0-387-70810-2_3

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T. & Salzberg, S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, **33**, 290–295. https://doi.org/10.1038/nbt.3122

Piperno, D.R. & Pearsall, D.M. (1998) *The origins of agriculture in the lowland neotropics*. San Diego: Academic Press.

Price, M.N., Dehal, P.S. & Arkin, A.P. (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, **26**, 1641–1650. https://doi.org/10.1093/molbev/msp077

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D. *et al.* (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**, 559–575. https://doi.org/10.1086/519795

Rodgers, J., Zumba, J. & Fortier, C. (2017) Measurement comparison of cotton fiber micronaire and its components by portable near infrared spectroscopy instruments. *Textile Research Journal*, **87**, 57–69. https://doi.org/10.1177/0040517515622153

Said, J.I., Song, M., Wang, H., Lin, Z., Zhang, X., Fang, D.D. *et al.* (2015) A comparative meta-analysis of QTL between intraspecific *Gossypium hirsutum* and interspecific *G. hirsutum* × *G. barbadense* populations. *Molecular Genetics and Genomics*, **290**, 1003–1025. https://doi.org/10.1007/s00438-014-0963-9

Sazonovs, A. & Barrett, J.C. (2018) Rare-variant studies to complement genome-wide association studies. *Annual Review of Genomics and Human Genetics*, **19**, 97–112.

Seagull, R.W., Oliveri, V., Murphy, K., Binder, A. & Kothari, S. (2000) Cotton fiber growth and development 2. Changes in cell diameter and wall birefringence. *Journal of Cotton Science*, **4**, 97–104.

Shi, Y., Zhang, B., Liu, A. *et al.* (2016) Quantitative trait loci analysis of Verticillium wilt resistance in interspecific backcross populations of *Gossypium hirsutum* × *Gossypium barbadense*. *BMC Genomics*, **17**, 877. https://doi.org/10.1186/s12864-016-3128-x

Si, Z., Liu, H., Zhu, J., Chen, J., Wang, Q., Fang, L. *et al.* (2018) Mutation of SELF-PRUNING homologs in cotton promotes short-branching plant architecture. *Journal of Experimental Botany*, **69**, 2543–2553. https://doi.org/10.1093/jxb/ery093

Smith, C.W., Cantrell, R.G., Moser, H.S. & Oakley, S.R. (1999) History of cultivar development in the United States. In *Cotton*. New York: John Wiley & Sons, Inc. pp. 99–171.

Song, Y.H., Ito, S. & Imaizumi, T. (2013) Flowering time regulation: photoperiod- and temperature-sensing in leaves. *Trends in Plant Science*, **18**, 575–583. https://doi.org/10.1016/j.tplants.2013.05.003

Stephens, S.G. & Moseley, M.E. (1974) Early domesticated cottons from archaeological sites in central coastal Peru. *American Antiquity*, **39**, 109–122.

Ulloa, M., Hutmacher, R.B., Davis, R.M., Wright, S.D. & Marsh, B. (2006) Breeding for Fusarium Wilt race 4 resistance in cotton under field and greenhouse conditions. *Journal of Cotton Science*, **10**, 114–127.

Verhalen, L.M., Mamaghani, R., Morrison, W.C. & Mcnew, R.W. (1975) Effect of blooming date on boll retention and fiber properties in cotton. *Crop Science*, **15**, 47–52. https://doi.org/10.2135/cropsci1975.0011183X001500010014x

Wagner, M.J. (2013) Rare-variant genome-wide association studies: a new frontier in genetic analysis of complex traits. *Pharmaco Genomics*, **14**, 413–424.

Wang, B., Lin, Z., Li, X., Zhao, Y., Zhao, B., Wu, G. *et al.* (2020a) Genome-wide selection and genetic improvement during modern maize breeding. *Nature Genetics*, **52**, 565–571. https://doi.org/10.1038/s41588-020-0616-3

Wang, F., Lian, L., Liu, Y., Zhang, Y., Fang, R. & Liu, Q. (2018) RoTFL1c of Rosa multiflora has a dual-function in suppressing reproductive growth and promoting vegetative growth of Arabidopsis. *Science China Life Sciences*, **61**, 1599–1601.

Wang, F., Zhang, J., Yu, C., Zhang, C., Gong, J., Song, Z. *et al.* (2020b) Identification of candidate genes for key fibre-related QTLs and derivation of favourable alleles in *Gossypium hirsutum* recombinant inbred lines with *G. barbadense* introgressions. *Plant Biotechnology Journal*, **18**, 707–720. https://doi.org/10.1111/pbi.13237

Wang, G.L., Dong, J.M. & Paterson, A.H. (1995) The Distribution of *Gossypium hirsutum* chromatin in *G. barbadense* germ plasm: molecular analysis of introgressive plant-breeding. *Theoretical and Applied Genetics*, **91**, 1153–1161. https://doi.org/10.1007/BF00223934

Wang, M., Tu, L., Yuan, D., Zhu, D.E., Shen, C., Li, J. *et al.* (2019) Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nature Genetics*, **51**, 224–229. https://doi.org/10.1038/s41588-018-0282-x

Wang, M., Tu, L., Lin, M., Lin, Z., Wang, P., Yang, Q. *et al.* (2017) Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nature Genetics*, **49**, 579–587. https://doi.org/10.1038/ng.3807

Wan, Q., Zhang, H., Ye, W., Wu, H. & Zhang, T. (2014) Genome-wide transcriptome profiling revealed cotton fuzz fiber development having a similar molecular model as Arabidopsis trichome. *PLoS One*, **9**, e97313. https://doi.org/10.1371/journal.pone.0097313

Ware, J.O. (1936) Plant breeding and the cotton industry. In *USDA yearbook of agriculture*. Washington, DC: United States Government Printing Office, pp. 657–744.

Wright, R.J., Thaxton, P.M., El-Zik, K.H. & Paterson, A.H. (1999) Molecular mapping of genes affecting pubescence of cotton. *Journal of Heredity*, **90**, 215–219. https://doi.org/10.1093/jhered/90.1.215

Wu, R., Wang, T., McGie, T., Voogd, C., Allan, A.C., Hellens, R.P. *et al.* (2014) Overexpression of the kiwifruit *SVP3* gene affects reproductive development and suppresses anthocyanin biosynthesis in petals, but has no effect on vegetative growth, dormancy, or flowering time. *Journal of Experimental Botany*, **65**, 4985–4995.

Young, E.F., Feaster, C.V. & Turcotte, E.L. (1976) Registration of Pima S-3 cotton. *Crop Science*, **16**, https://doi.org/10.2135/cropsci1976.0011183X001600040050x

Yuan, D., Grover, C.E., Hu, G., Pan, M., Miller, E.R., Conover, J.L. *et al.* (2021) Parallel and Intertwining Threads of Domestication in Allopolyploid Cotton. *Advance Science*, **8**(10), 2003634.

Yu, J., Zhang, K., Li, S., Yu, S. & Zhang, J. (2013) Mapping quantitative trait loci for lint yield and fiber quality across environments in a *Gossypium hirsutum* × *Gossypium barbadense* backcross inbred line population. *Theoretical and Applied Genetics*, **126**, 275–287. https://doi.org/10.1007/s00122-012-1980-x

Zhang, J., Guo, W. & Zhang, T. (2002) Molecular linkage map of allotetraploid cotton (*Gossypium hirsutum L.* × *Gossypium barbadense L.*) with a haploid population. *Theoretical and Applied Genetics*, **105**, 1166–1174. https://doi.org/10.1007/s00122-002-1100-4

Zhao, N., Wang, W., Grover, C.E., Jiang, K., Pan, Z., Guo, B. *et al.* (2021) Genomic and GWAS analyses demonstrate phylogenomic

relationships of *Gossypium barbadense* in China and selection for fiber length, lint percentage, and Fusarium wilt resistance. *Plant Biotechnology of Journal*. Accepted Author Manuscript. https://doi.org/10.1111/pbi.13747

Zhong, S., Joung, J.G., Zheng, Y., Chen, Y., Liu, B., Shao, Y. *et al.* (2011) High-throughput illumina strand-specific RNA sequencing library preparation. *Cold Spring Harbor Protocols*, **8**, 940–949. https://doi.org/10.1101/pdb.prot5652