ARTICLES
https://doi.org/10.1038/s41588-021-00844-9
nature genetics

Check for updates

The genomic basis of geographic differentiation and fiber improvement in cultivated cotton

Shoupu He[1,10], Gaofei Sun[2,10], Xiaoli Geng[1,10], Wenfang Gong[1,10], Panhong Dai[1], Yinhua Jia[1], Weijun Shi[3], Zhaoe Pan[1], Junduo Wang[3], Liyuan Wang[1], Songhua Xiao[4], Baojun Chen[1], Shufang Cui[5], Chunyuan You[6], Zongming Xie[7], Feng Wang[8], Jie Sun[9], Guoyong Fu[1], Zhen Peng[1], Daowu Hu[1], Liru Wang[1], Baoyin Pang[1] and Xiongming Du[1]✉

Large-scale genomic surveys of crop germplasm are important for understanding the genetic architecture of favorable traits. The genomic basis of geographic differentiation and fiber improvement in cultivated cotton is poorly understood. Here, we analyzed 3,248 tetraploid cotton genomes and confirmed that the extensive chromosome inversions on chromosomes A06 and A08 underlies the geographic differentiation in cultivated *Gossypium hirsutum*. We further revealed that the haplotypic diversity originated from landraces, which might be essential for understanding adaptative evolution in cultivated cotton. Introgression and association analyses identified new fiber quality-related loci and demonstrated that the introgressed alleles from two diploid cottons had a large effect on fiber quality improvement. These loci provided the potential power to overcome the bottleneck in fiber quality improvement. Our study uncovered several critical genomic signatures generated by historical breeding effects in cotton and a wealth of data that enrich genomic resources for the research community.

Most modern cultivated cotton (~97%) is tetraploid *Gossypium hirsutum* ($(AD)_1$)[1], which was most likely domesticated on the Yucatan Peninsula of Mesoamerica[2,3]. Primitive *G. hirsutum* was grown in the tropics under high temperatures, plentiful rainfall and short days. Since this cotton was used for spinning, it spread widely with human activities. Long-term natural selection and artificial breeding effects under diverse environments have reshaped the ecological adaptability of the original *G. hirsutum*, allowing it to grow across a wide range of latitudes worldwide[4]. Although modern *G. hirsutum* cultivars have different areas suitable for growth, which were recognized by early breeders, the genomic basis of these distinguished ecotypes remains unknown. In both animals and plants, genomic divergence driven by chromosomal inversions within species is recognized as an essential signature of environmental adaptability[5,6]. However, previous *G. hirsutum* resequencing projects suggested that modern cultivated *G. hirsutum* was an admixed population with no geographical structure[7,8], in contrast to foxtail millet[9] and soybean[10]. Although the latest study revealed that the distinct haplotypes on two chromosomes (A06 and A08) might be responsible for adaptability in cultivated *G. hirsutum*[11], the relatively small population size ($n = 419$) and narrow genetic diversity (only cultivars) severely limited the exploration of haplotype evolution in entire *G. hirsutum* germplasm.

To meet the increasing textile quality demands of the spinning industry and coordinate mechanized harvesting, fiber quality improvement is the priority in current cotton breeding programs in China. Because the *G. hirsutum* genome is large (~2.3 gigabases (Gb)) and structurally complicated (allotetraploid), the mapping of fiber quality-related quantitative trait loci (QTLs) was a great challenge. Recently, most of the genomic studies in cotton were concentrated on genome assembling and comparison[12–14], resequencing projects only identified a few quality-related QTLs in *G. hirsutum* populations[7,15]; the origination and distribution of favorable loci of these QTLs were still unknown.

GenBank genomics has been suggested to act as a bridge connecting genetic diversity with breeding in crops[16]. Recently, large-scale resequencing projects (sample sizes >1,000) have revealed a complete set of novel variations in rice[17], unveiled the effects of introgression on shaping adaptability in wheat[18] and identified the essential genes responsible for important traits in melon[19]. GenBank conserved more than 7,000 *G. hirsutum* accessions collected worldwide over the past decades[20]. Previous studies covered <10% of the whole *G. hirsutum* collection, which limited the ability of QTL detection to explain the extensive variation in fiber properties in cotton germplasm.

In this study, we analyzed the genomic variations in 3,248 tetraploid cotton germplasms (termed 3K-TCG, which included 2,922 *G. hirsutum* accessions). On the basis of a large-scale investigation of genomic variation (Supplementary Figs. 1 and 2) and a multi-environmental genome-wide association study (GWAS) in a more diverse panel, we discuss the genomic basis of geographic differentiation, as well as findings of introgressed alleles that greatly enhance fiber quality.

Results

Genome mapping and variation detection. In this study, the resequencing data of 3,278 cotton genomes were mapped to the

[1]State Key Laboratory of Cotton Biology, Institute of Cotton Research of the Chinese Academy of Agricultural Sciences, Anyang, China. [2]School of Computer Science & Information Engineering, Anyang Institute of Technology, Anyang, China. [3]Research Institute of Economic Crops, Xinjiang Academy of Agricultural Sciences, Urumchi, China. [4]Institute of Industrial Crops, Jiangsu Academy of Agricultural Sciences, Nanjing, China. [5]Institute of Cotton, Hebei Academy of Agriculture and Forestry Sciences, Shijiazhuang, China. [6]Cotton Research Institute, Shihezi Academy of Agriculture Science, Shihezi, China. [7]Production & Construction Group Key Laboratory of Crop Germplasm Enhancement and Gene Resources Utilization, Biotechnology Research Institute of Xinjiang Academy of Agricultural and Reclamation Science, Shihezi, China. [8]College of Agronomy, Hunan Agricultural University, Changsha, China. [9]Key Laboratory of Oasis Eco-agriculture, College of Agriculture, Shihezi University, Shihezi, China. [10]These authors contributed equally: Shoupu He, Gaofei Sun, Xiaoli Geng, Wenfang Gong. ✉e-mail: duxiongming@caas.cn

PacBio-assembled reference genome (*G. hirsutum* 'Texas Marker 1')[12]. Among these accessions, 3,248 tetraploid cottons were screened (3K-TCG), including 2,922 *G. hirsutum* accessions, 309 *G. barbadense* accessions and 17 wild relatives (Supplementary Table 1). For the single nucleotide polymorphism (SNP) set of 3K-TCG, after filtering (minor allele frequency (MAF) > 0.05, missing rate <0.2), 6,711,614 SNPs and 937,526 InDels were identified. Among these variations, 9,482 SNPs and InDels were large-effect variations, which affected 8,088 genes (Supplementary Table 2).

Genetic diversity and population differentiation. When using the wild relative *G. mustelinum* as the outgroup species[21], on the basis of population structure analysis of genotype information and assisted with the phylogenetic analysis and principal component analysis (PCA), the 3K-TCG panel could be classified into eight subgroups (from G0 to G7) (Fig. 1a and Supplementary Fig. 3). In addition to wild species (G0) and *G. barbadense* (G7), all *G. hirsutum* accessions were categorized into six subgroups. Among them, G1 ($n = 292$) included most of the landraces collected from Central America (CAL), which was composed of seven previously recognized geographical races[22]. A clade of indigenous perennial accessions ($n = 43$) collected from islands of South China (historically named *G. purpurascens*) were also included in G1 (Supplementary Table 1), implying that *G. hirsutum* might have been introduced to China during the Age of Discovery, earlier than previous record (late of Qing Dynasty) (Supplementary Fig. 4). Consistent with our previous finding[23], all seven recognized landraces were genetically mixed (Fig. 1a), demonstrating that previous phenotype- or geography-based taxonomic classifications should be reconsidered. The G2 ($n = 76$) subgroup contained all accessions collected in South China (SCL) (Fig. 1a,b and Supplementary Fig. 4). These unidentified annual landraces are mainly grown in the remote mountains of South China and severely lack modern cultivation management, which exhibited significantly inferior economic traits compared with those of the improved cultivars (G3–G6) (Fig. 1c). The G3 ($n = 435$) subgroup was composed of most of the early-maturity accessions, nearly half of which ($n = 217$) are grown in Northwest China (NWC) and North China (NC) (Fig. 1a and Supplementary Table 1). The G4 ($n = 433$) subgroup included accessions from all three historical Chinese cotton planting areas. Most of the accessions in the G5 ($n = 429$) subgroup were cultivated in the Yangtze River region (YZR), located in southern China. The G6 subgroup contained the most accessions ($n = 1{,}258$), which were mainly from the Yellow River region (YER) of China and the United States. In this study, the whole-genomic genotypes of accessions ($n = 259$) collected from South China were genotyped for the first time, and were dispersed among all six subgroups (G1–G6), indicating a complicated introduction history of *G. hirsutum* in China (Supplementary Fig. 4).

Comparisons of average pairwise fixation statistic (F_{ST}) values for the subgroups demonstrated that the genetic divergence within improved subgroups (G3–G6) was low (0.019–0.067) (Fig. 1d). However, the average pairwise F_{ST} values were much higher (0.425–0.552) when using G1 than when using other subgroups, and those of G2 were intermediate (0.113–0.189) (Fig. 1d), indicating noteworthy genetic differentiation between landraces and improved cultivars. Windowed F_{ST} values across the whole genome further revealed broad genomic divergence between landrace subgroups (G1 or G2) and improved cultivars but the differentiation level of G2 was lower than that of G1 (Fig. 1e and Supplementary Fig. 5). Therefore, considering that G2 was not previously genotyped, we suggest that G2 is an undiscovered clade of *G. hirsutum* between recognized landraces (G1) and improved cultivars, which could be vital for studying *G. hirsutum* domestication. Interestingly, within improved cultivars (G3–G6), notable divergence was detected on chromosomes A06 and A08, which spanned broad genomic regions and specifically existed in the G3 (A06 genomic divergence) and G5 (A08 genomic divergence) subgroups, respectively (Fig. 1e). These highly divergent genomic regions might have driven the subgroup differentiation within improved cultivars.

Landrace-originated chromosomal inversions were related to the geographic differentiation of cultivated *G. hirsutum*. Like other crops cultivated in a broad range of areas worldwide, to adapt to the local environment and meet the breeding targets, the improved *G. hirsutum* was subject to strong natural and artificial selection. The dramatic genomic divergence on chromosomes A06 and A08 and the regular geographic origin of subgroups G3 (most accessions were from NC and NWC) and G5 (most accessions were from YZR) implied a possible relationship between genomic divergence and geographic differentiation within cotton cultivars (Fig. 1a,e). To unveil such a connection, we sorted the genotypes of 3K-TCG according to a phylogenetic tree constructed by using SNPs on chromosomes A06 and A08 only. Four major haplotypes were roughly identified on both chromosomes (Fig. 2a,b).

Pairwise F_{ST} comparisons within improved cultivar subgroups showed that the peak of the divergent regions ranged from ~77.5 to ~115.5 megabases (Mb) on chromosome A06 (Fig. 2a, top) and from ~22.5 to ~92.5 Mb on chromosome A08 (Fig. 2b, top). This result was consistent with previous reports using different populations and different variation identification technologies[11,23]. For chromosome A06, Hap-A06-3 and Hap-A06-4 were mainly carried by the G3 and G1 subgroups (Fig. 2a, bottom). For chromosome A08, Hap-A08-3 was specifically carried by the G5 subgroup (Fig. 2b, bottom). These three distinct haplotypes exclusively carried by the G3 and G5 subgroups distinguished them from the other subgroups. The genotypes of landraces (G1 and G2) enabled us to trace the origin of these haplotypes. We found that the landraces carried Hap-A06-3 and Hap-A06-4 on chromosome A06 while carrying Hap-A08-2, Hap-A08-3 and Hap-A08-4 on chromosome A08 (all these haplotypes were defined as primitive haplotypes) (Fig. 2c,d). This result indicated that the remaining haplotypes (Hap-A06-1, Hap-A06-2 and Hap-A08-1 were defined as derived haplotypes) might have resulted from recombination between primitive haplotypes during the domestication and selection process (Fig. 2c,d and Supplementary Fig. 6). Nearly all the Chinese registered cultivars (with known suitable cultivation regions) carrying Hap-A06-3/Hap-A06-4 (G3) and Hap-A08-3 (G5) were geographically distributed in the highest- and lowest-latitude regions, respectively (Fig. 2e and Supplementary Table 1), which exhibited a regular distribution pattern. To explore the cause of genomic divergence, we de novo assembled a genome carrying both Hap-A06-3 and Hap-A08-4 (ICR_XLZ 7) (Supplementary Table 3). According to a comparison with the reference genome (ICR_TM-1, carried Hap-A06-3 and Hap-A08-4)[12], the genomic synteny of the two chromosomes showed that the haplotype polymorphism was likely caused by several large-scale inversions (Extended Data Fig. 1).

To further understand the significance of genomic divergence in the cotton breeding history, we comprehensively analyzed the Chinese registered cultivars ($n = 851$) by integrating their A06–A08 haplotype combination, pedigree, suitable cultivation regions and release year. Among them, 14 A06–A08 haplotype combinations were classified. Type I to IV were exclusively carried by the G5 subgroup, which was cultivated in the YZR and showed the earliest registration time (the 1960s to 1980s) (Fig. 2f). In contrast, the G3 subgroup carried haplotype combinations from Type X to XIV and was mainly cultivated in the NWC, showing the latest registration time (from the 1990s to present) (Fig. 2f). In addition, the remaining accessions (carrying haplotype combinations from Type V to IX), which were registered at an intermediate time (from the 1980s to 1990s) but accounted for the majority, were mainly cultivated in the YER (Fig. 2f). Moreover, we tracked the sources of

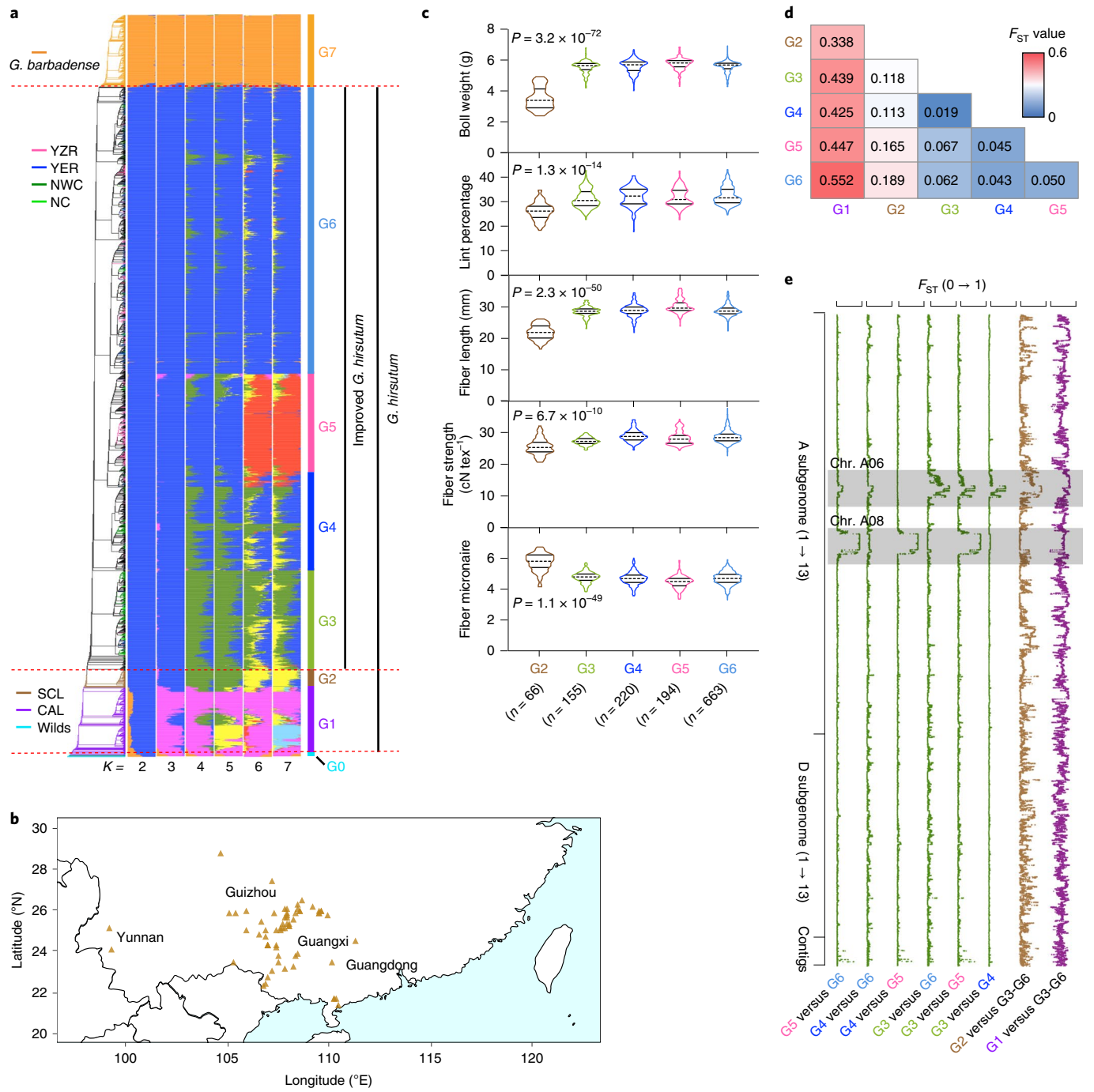


Fig. 1 | Population structure and divergence in tetraploid cotton. a, Phylogenetic tree and model-based clustering ($K = 2$–7) of 3,248 tetraploid cotton accessions (3K-TCG). Wild species *G. mustelinum* was used to root the tree (at the bottom). The entire population is separated into seven subgroups, as indicated by colored bands (right). Wilds, wild tetraploid cotton species; CAL, Central America landraces; SCL, South China landraces. All Chinese registered cultivars in the improved *G. hirsutum* population (from G3 to G6) are colored according to their suitable cultivation regions. YZR, Yangtze River region; YER, Yellow River region; NWC, Northwest China; NC, North China. **b**, Geographic distribution of all G2 accessions (SCL). The map was created by the 'maps'[42] package in R[43]. **c**, Comparisons of major agricultural traits among subgroups (G2–G6). In violin plots, the solid lines and dot lines indicate quartiles and medians, respectively. Significances were tested between G2 and subgroup with greatest (for fiber micronaire) or smallest (for boll weight, lint percentage, fiber length and fiber strength) medians, by two-tailed Student's *t*-test. **d**, Pairwise comparisons of fixation index (F_{ST}) values between subgroups in *G. hirsutum*. **e**, Comparisons of F_{ST} value across the entire genome for pairwise subgroups. The transparent gray bands highlight chromosomes (Chr.) A06 and A08. Subgroups G3–G6 indicate the combinations of all improved *G. hirsutum* from G3 to G6.

14 haplotype combinations, most of which could be found in the early United States or the Former Soviet Union introduced germplasm and six of them could be further tracked in landraces (G1 and G2) (Supplementary Fig. 7). Interestingly, Hap-A06-1 and Hap-A06-2 (mostly carried by G4 and G6) could not be found in landraces, demonstrating that they might have originated from an unknown source (Fig. 2c and Supplementary Fig. 6). On the basis of the historical planting area records for Chinese cotton cultivation

in the past 68 yr (Fig. 2g), we described the genomic landscape of chromosomes A06 and A08 for cotton germplasm introduction, breeding preferences and regional relocation in China (Fig. 2h). The alternative selection of haplotype combinations of Chinese cultivars in different periods reflected the genomic basis of geographic differentiation in Chinese cotton cultivars. Therefore, our study demonstrated that Hap-A06-3/Hap-A06-4 (G3) and Hap-A08-3 (G5) might be the essential genomic feature for cotton cultivars grown in different ecological environments (Fig. 2f–h).

Exotic introgressions for fiber quality improvement. Natural variations conserved in wild relatives, which can be used for crop improvement are abundant. Elite lines that have been created by successfully introgressing exotic alleles have played essential roles in many crop breeding processes[24]. Early *G. hirsutum* is thought to have experienced natural interspecific hybridization before domestication[25]. In the last century, attempts have been made to hybridize most wild relatives of the cotton genus with *G. hirsutum* and achieved remarkable breeding effects on fiber improvement[26]. However, the whole-genomic landscape of introgression and the functional introgressed fragments have rarely been reported. In this study, we investigated the introgression events in ~2,500 cultivated *G. hirsutum* germplasms by comparing their genomes with those of the 13 most likely exotic donors (Supplementary Table 1). Introgression events of high confidence (introgression index >0.2, accumulated introgressed fragment length >500 Mb) from three relatives were highlighted (Fig. 3a and Supplementary Table 4). For all 450 introgression lines (with known exotic donors), as expected due to the great efforts to hybridize *G. hirsutum* and *G. barbadense* ($(AD)_2$) in cotton breeding programs, *G. barbadense* had the most accumulated introgressed fragments, which were widely distributed on every chromosome (Fig. 3b). Introgressed fragments from another diploid cultivar species, *G. arboreum* (A_2), were primarily located on chromosome A09 (Fig. 3b and Supplementary Table 4). The fragments introgressed from wild D-genome diploid species were mostly located on chromosome D08 for *G. thurberi* (D_1) (Fig. 3b and Supplementary Table 4).

By integrating with GWAS results, a large-effect pleiotropic allele associated with both fiber length (*FL3*) and fiber strength (*FS2*) was identified (Fig. 4a) in this study. In some introgression lines (109 of 127), this locus was just overlapping with the *G. arboreum* introgressed region on chromosome A09 (named GaIR_A09, ranging from ~61.8 to 62.1 Mb; Fig. 3c and Supplementary Table 5). The accessions carrying GaIR_A09 ($n=81$) showed significantly better fiber length and strength than those without this fragment ($n=11$) (Fig. 3c). We further evaluated GaIR_A09 in the 3K-TCG population according to the cluster tree based on SNPs of GaIR_A09 and found that the clade of *G. hirsutum* introgression lines were tightly clustered with *G. arboreum* (Extended Data Fig. 2a). Therefore, GaIR_A09 was suggested to be a novel candidate locus introgressed from *G. arboreum* and responsible for fiber quality in modern cultivars; most accessions carrying *FL3*/*FS2* belonged to the interspecific hybridization programs such as 'Pee Dee' (United States) and 'Suyuan' (China) (Supplementary Table 5).

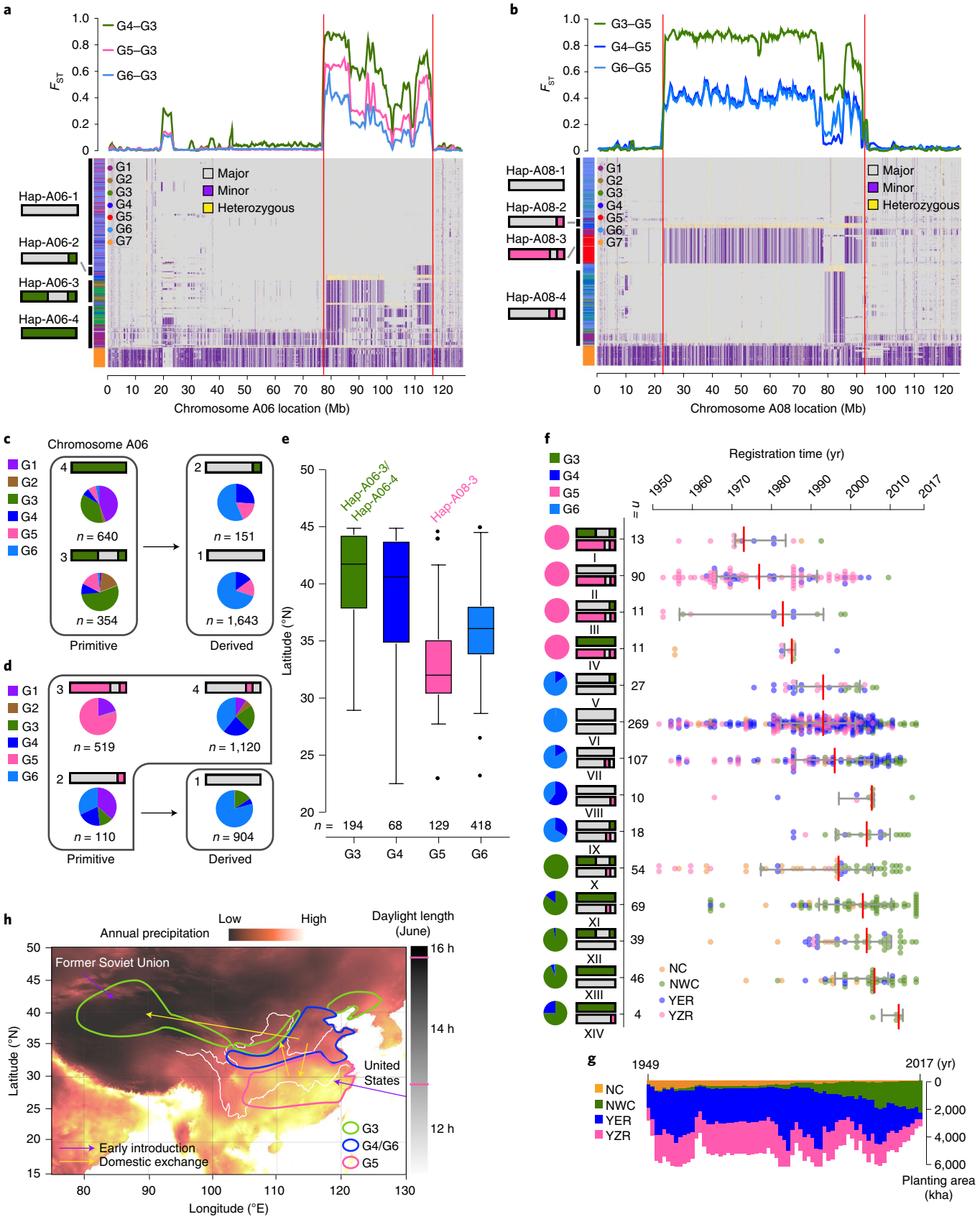
Another distinctive introgression event from *G. thurberi* occurred on chromosome D08 (Fig. 3a,b). The phenotypic comparison showed that the introgressed fragments had a substantial improvement on fiber strength (Fig. 3d). QTL mapping of a segregating population from a cross between an introgressed (with superior fiber quality) line and a nonintrogressed (with inferior fiber quality) line confirmed this locus (*FS3*) and further narrowed down the candidate genomic region ranging from ~7.8 to 60.4 Mb (ref. [27]) (Fig. 3d, Extended Data Fig. 2b and Supplementary Table 6). Nearly all the accessions carrying *FS3* belonged to the interspecific hybridization programs 'Suyuan' (Supplementary Table 6).

The developing thurberi–arboreum–hirsutum (TAH) triple-hybrid strains and their derived Pee Dee germplasm are considered remarkable breakthroughs in *G. hirsutum* germplasm enhancement programs. The Pee Dee program successfully broke the negative correlation between fiber yield and fiber strength, and >80 crucial germplasms from this program subsequently became the primary backbone parents of most cotton breeding programs worldwide[28]. Our work provides genomic evidence of interspecific hybridization efforts in the history of cotton breeding. By integrating with the documented background, we suggested that *FL3*/*FS2* probably originated from Pee Dee germplasm. Although the documentation mentioned that the Pee Dee germplasm should also contain *G. thurberi* introgressions, we did not detect *FS3* in any of the Pee Dee lines due to the limitation of sequenced Pee Dee lines in our study (Supplementary Table 4). Therefore, we suggested that *FL3*/*FS2* might come from the Pee Dee lines and further combined with *FS3* after introducing to China to develop the 'Suyuan' lines with superior fiber quality (Extended Data Fig. 2c). Confirmation of the precise genomic regions of these alien fragments would provide a foundation for identifying causal functional genes in the future.

The genetic architecture of cotton fiber quality. The properties of fiber quality determine its economic value, which is a concern for both the spinning industry and scientific research. In this study, on the basis of the latest assembled reference genome and a large, diverse germplasm population (1,245 samples), we investigated the loci that control major fiber properties by using multi-environmental phenotypic data. To efficiently confirm the haplotype, candidate genes and origin of each locus, we constructed a comprehensive GWAS analysis framework in this study (Supplementary Fig. 1). Three fiber length-related loci (the GWAS threshold value was $-\log P>7.35$), namely, *FL2*, *FL3* and *FL4*, were detected on chromosomes D11, A09 and A10, respectively (Fig. 4a). Among them, the strongest signal was *FL2* (D11: 24,509,312–24,800,082 base pairs (bp)) (Extended Data Fig. 3a,b). As a previously reported locus[15], *FL2* had the highest favorable allelic frequency (78.7%), with a fiber length-improving effect of 4.3%, indicating that it is common in the current germplasm (Fig. 4a and Extended Data Fig. 3e). According to the local SNP clustering analysis of 3K-TCG, we also found that *FL2* likely

Fig. 2 | Genomic divergence of chromosomes A06 and A08 impacts the geographic differentiation in improved *G. hirsutum*. **a**,**b**, Genomic divergence and haplotype classification of chromosomes A06 (**a**) and A08 (**b**). Colored lines represent the F_{ST} values for pairwise subgroups (top). Haplotype classification of two chromosomes in 3K-TCG (bottom). Sketches represent the major haplotypes of two chromosomes. **c**,**d**, Haplotype category of subgroups for chromosomes A06 (**c**) and A08 (**d**). Haplotypes observed in landraces (G1 or G2) and only in improved accessions are defined as primitive and derived haplotypes, respectively. **e**, Latitudinal distribution of Chinese *G. hirsutum* cultivars in G3–G6. Box limits and center lines indicate quartiles and medians, respectively. Whiskers denote 1.5× interquartile range and points show outliers. **f**, The relationship among registration time, geographic distribution and A06–A08 haplotype combination for Chinese *G. hirsutum* cultivars. Cultivars are represented by dots and colored according to their cultivated areas. Gray horizontal lines and red vertical lines indicate the interquartile range and median of registration time for cultivars carrying different haplotype combinations. **g**, Cotton sowing area in China from 1949 to 2017 (data were downloaded from the National Bureau of Statistics of China). **h**, Schematic of introduction, cultivation regions and domestic exchanges for Chinese *G. hirsutum* germplasm. Four major cotton cultivation regions in China are circled and colored according to the major subgroups they contained. Climate data (average annual precipitation from 1970 to 2000) were downloaded from www.worldclim.org (v.2.1)[44] and visualized by QGIS (v.3.16.3)[45].

originated from landraces (Extended Data Fig. 3f). *FL3* (A09) and *FL4* (A10) were detected as two novel loci with minor frequencies (6.7% and 15.7%) in the population (Fig. 4a and Extended Data Figs. 4 and 5). As an alien locus introgressed from the diploid cotton species *G. arboreum*, *FL3* exhibited the largest fiber length-improving effect of 15.6% (Extended Data Fig. 4). In contrast to *FL2* and *FL3*,



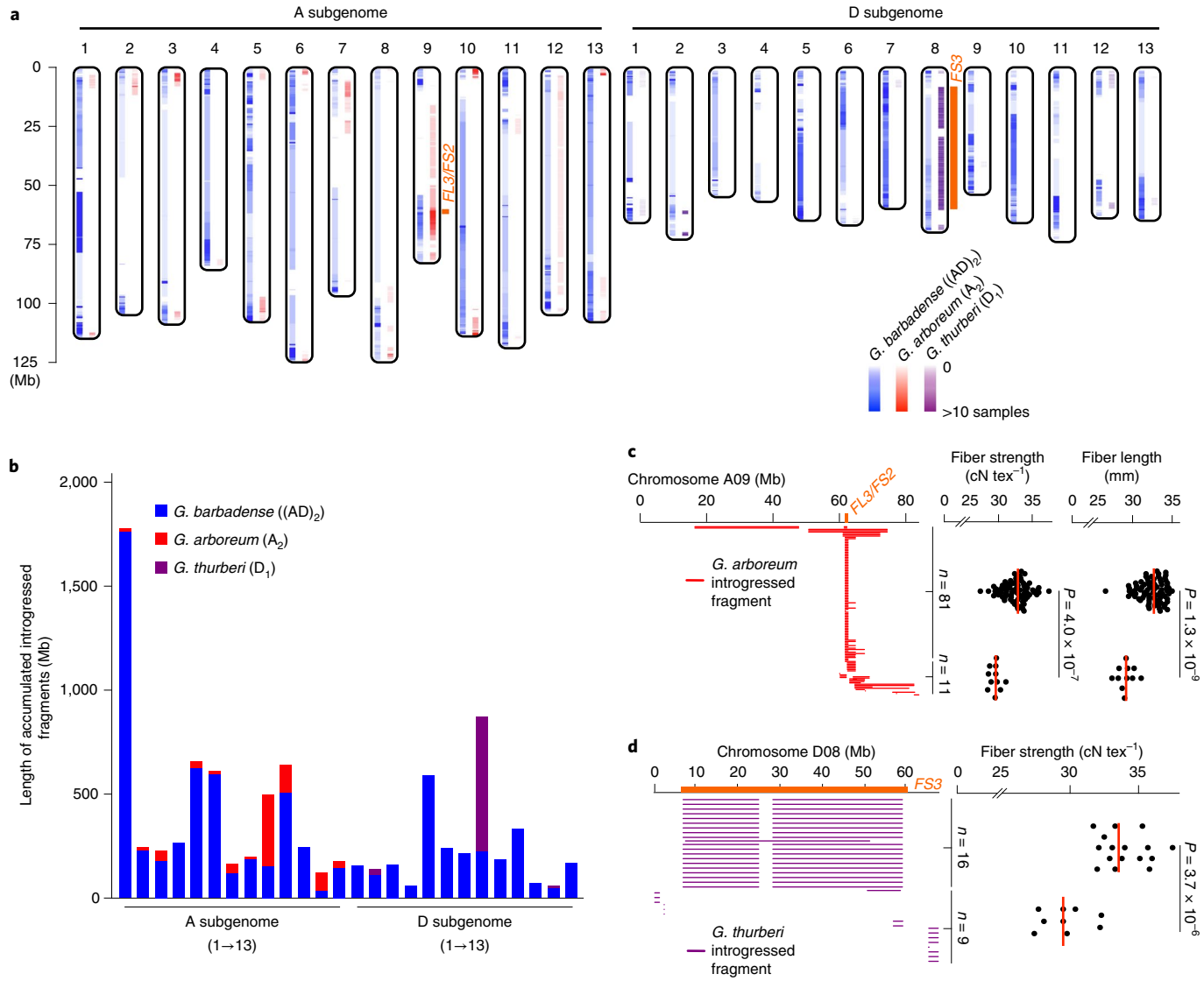


Fig. 3 | Interspecific introgressions in improved *G. hirsutum* and their effect on fiber quality improvement. a, Genome-wide landscape of introgressions in improved *G. hirsutum*. Colored heatmaps show the abundance of genomic regions existing introgressions from three exotic cotton species. Orange bands mark the regions overlapping with pleiotropic fiber-quality loci identified by the GWAS (chromosome A09) (Fig. 4a) and biparental QTL mapping (chromosome D08)[27]. **b**, Length of accumulated introgressed fragments from three cotton species in improved *G. hirsutum*. **c**, Schematic shows that all accessions ($n=127$) carried *G. arboreum* introgressed fragments with various lengths and locations on chromosome A09. Orange band highlights the specific overlapping region between introgression analysis and the GWAS signal (*FL3/FS2*) (Fig. 4a). Dot plots show the comparisons for fiber length and fiber strength between accessions that carry *FL3/FS2* ($n=81$) and those that do not carry *FL3/FS2* ($n=11$). **d**, Schematic shows that all accessions ($n=37$) carried introgressed *G. thurberi* fragments with various lengths and locations on chromosome D08. Orange band highlights the specific overlapping region between introgression analysis and biparental QTL mapping for fiber strength (*FS3*)[27]. Dot plots show the comparison for fiber strength between accessions that carry *FS3* ($n=16$) and those that do not carry *FS3* ($n=9$). In the dot plots of **c** and **d**, red vertical lines show the medians. All significances are tested by two-tailed Student's *t*-test.

FL4 was a SNP variant (A10: 111,991,164 bp) located in an exonic region and introduced a stop codon in the gene encoded NAC transcription factor 29 (*Gh_A10G233100*) (Extended Data Fig. 5). In addition to the above three loci, another previously identified locus of minor favorable allelic frequency[15], *FL5*, with a signal (–log *P* of peak SNP = 6.2) slightly lower than the threshold value, was located on chromosome A07 (88,391,473–88,557,465 bp) (Fig. 4a and Extended Data Fig. 6). Due to the significant positive correlation (Pearson $r=0.79$, $P<0.00001$) between fiber length and fiber strength, two major fiber strength loci, *FS1* and *FS2*, were co-located with *FL5* and *FL3*, with 5.4% and 14.2% fiber strength-improving effects, respectively (Extended Data Figs. 4 and 6). Both of them exhibited minor favorable allelic frequencies. Another alien locus introgressed from the wild diploid cotton species *G. thurberi* was not detected in the GWAS because of its low frequency in the population (Fig. 4a). Fiber elongation rate, a fiber-quality property representing the elongation ability of mature fiber cells, was negatively correlated with both fiber length (Pearson $r=-0.37$, $P<0.00001$) and fiber strength (Pearson $r=-0.50$, $P<0.00001$). Three alleles (favorable allelic frequency ranging from 35% to 63.7%) were detected for fiber elongation rate, located on chromosomes D04 (*FE1*), D01 (*FE2*) and A05 (*FE3*) (Fig. 4a). Two of them (*FE1* and *FE3*) were also detected in a previous study[29]. According to their distributions, we found that all three loci originated from

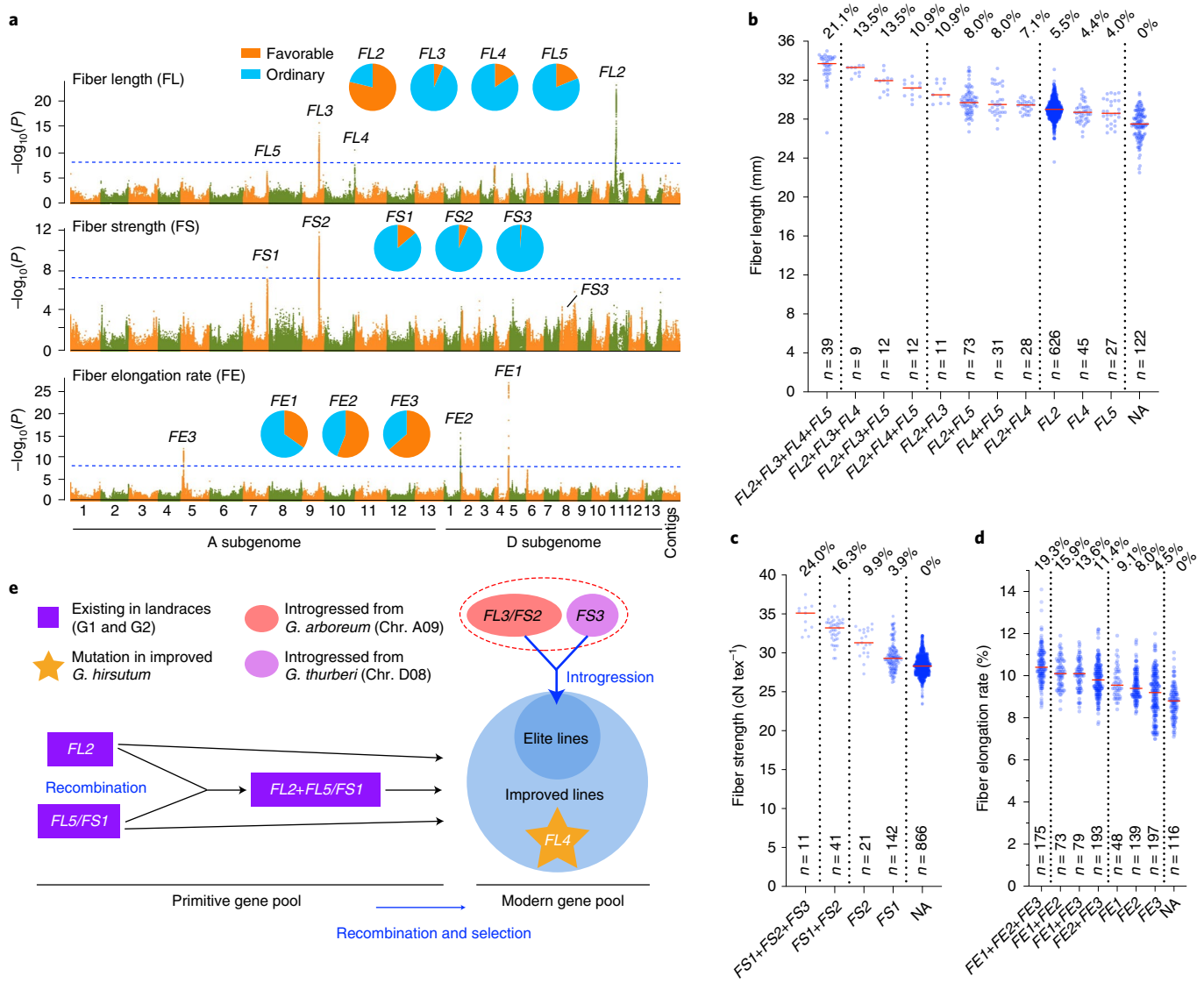


Fig. 4 | Genetic basis of fiber quality in *G. hirsutum*. **a**, Manhattan plots of GWAS for fiber length (FL), fiber strength (FS) and fiber elongation rate (FE). Pie charts represent allelic frequencies of each locus in the GWAS population. **b–d**, Dot plots show the effects of allelic combinations for FL (**b**), FS (**c**) and FE (**d**) in population. Blue dots represent the accessions categorized according to different allelic combinations. Red lines indicate the medians of each category. NA indicates accessions carrying no favorable alleles. **e**, Schematic illustrates the origin and recombination of favorable alleles related to fiber length and strength during domestication and breeding.

landraces (Extended Data Figs. 7–9). By integrating these findings with transcriptomes and quantitative PCR with reverse transcription (qRT–PCR) results, *Gh_D04G181300*, *Gh_D01G220400* and *Gh_A05G094100* were screened out as the possible candidate genes responsible for loci *FE1*, *FE2* and *FE3*, respectively (Supplementary Tables 7–12). Tubulin is the principal constituent of microtubules, which is essential for cotton fiber elongation[30]. In our study, *tubulin alpha 2* (*GhTUA2*) was specifically upregulated in the fiber elongation stage (from 10 to 15 days postanthesis (DPA)) and significantly repressed root elongation in *GhTUA2*-overexpressed *Arabidopsis* (Extended Data Fig. 7e,f), which implied a potential role in regulating cotton fiber elongation.

To evaluate the pyramiding effects of favorable alleles for each trait in the *G. hirsutum* germplasm, we compared fiber properties among accessions carrying multiple favorable allelic combinations (Supplementary Table 13). As expected, the accessions carrying more favorable alleles always exhibited better fiber length and fiber strength (Fig. 4b,c). For fiber length, most of the accessions carried only one favorable allele, *FL2*, which accounted for half of the GWAS population ($n=626$) (Fig. 4b). Surprisingly, for fiber strength, except for a small number of accessions that carried only the favorable allele *FS1* ($n=142$), most accessions ($n=866$) carried no favorable alleles (Fig. 4c). Among all identified alleles, two introgressed alleles, *FL3*/*FS2* and *FS3*, exhibited more significant enhancement effects on both fiber length (reaching a maximum of 21.1%) and fiber strength (reaching a maximum of 24.0%) when they were present in combination with another favorable allele (Fig. 4b,c). The positive result revealed by our study was that the potential for fiber quality improvement is still extremely high in the current cotton cultivar gene pool. Because the fiber elongation rate QTLs were negative factors, most of the elite accessions with the best fiber quality likely carried very few favorable alleles for fiber elongation rate (Extended Data Fig. 10). Therefore, reasonably combined favorable alleles of different traits should be carefully considered in the future breeding practice.

In this study, an abundance of genotypic data for wild relatives and landraces enabled us to trace the origin and recombination history of favorable alleles. For fiber quality-related loci, except the alien alleles (*FL3*/*FS2* and *FS3*) and the point mutation locus (*FL4*), all favorable alleles (*FL2*, *FL5*/*FS1*, *FE1*, *FE2* and *FE3*) could be found in landraces (primitive gene pool) (Fig. 4e). A favorable allelic combination, *FL2*+*FL5*/*FS1*, was also found to be naturally recombined in landraces (Fig. 4e). During the cotton domestication and breeding process, the existing favorable alleles in landraces were directly selected or recombined for cultivar development (improved lines). Together with interspecific hybridization efforts, various recombinants carrying more favorable allelic combinations were further developed into elite lines (Fig. 4e).

Discussion

Our large-scale sequencing data first revealed the genomic landscape of variations for the entire *G. hirsutum* germplasm collection. As the most notable signature in the genome, the extensive divergence on chromosomes A06 and A08 created the current population structure within cultivated *G. hirsutum*[11]. As reported in previous studies in both animals[31–36] and plants[37–39], population differentiation was mainly caused by large chromosomal inversions. This phenomenon is a typical evolutionary mechanism by which species rapidly adapt to different environments by repressing recombination to retain beneficial genotypes[5,40]. A recent report in wild sunflowers confirmed that the extreme-low-frequency recombination haplotype blocks (caused by chromosome inversions) were the critical genomic components to retained adaptive loci in ecotype populations[41]. Our findings strengthen the 'chromosomal inversion-population differentiation' theory in crops and specify the haplotypes related to geographic differentiation in cotton cultivars. To meet the challenges of Chinese cotton cultivation regions relocation (cotton planting areas have moved from the Yellow River region and Yangtze River region to Northwest China region), understanding the genomic basis of geographic differentiation is crucial for developing cultivars adapting new environment and promotes further targeted identification of causal adaptive genes in cotton.

In the GWAS population (n=1,245), in addition to a previously reported favorable allele (*FL2*)[15], only 16.4% of accessions carried two or more favorable alleles for fiber length (Fig. 4b). More than 80% of the accessions carried no favorable alleles for fiber strength (Fig. 4c). Our results uncovered the cause of the difficulty in current fiber quality improvement. In other words, due to the low frequency of most favorable alleles in modern cotton germplasm, the existing favorable alleles are far from completely used. Most importantly, benefiting from the large association population (n=1,245), we identified two large-effect alleles (*FL3*/*FS2* and *FS3*) introgressed from diploid cotton, which have the great potential to overcome the current bottleneck in fiber quality improvement in cotton breeding. Future identification of causal genes in these alien genomic fragments might reveal the molecular basis of the extensive interspecific hybridization performed in *Gossypium* in past decades. In summary, the extensive genotypic and phenotypic data provided by our study could be used as references for choosing minicore germplasm to construct pan-genome analysis and design molecular markers to clone functional genes or molecular breeding, thereby accelerating the process of new cultivar development.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00844-9.

Received: 19 November 2020; Accepted: 15 March 2021;
Published online: 15 April 2021

References

1. Lubbers, E. L. & Chee, P. W. in *Genetics and Genomics of Cotton* Part I (ed. Paterson, A. H.) 23–52 (Springer, 2009).
2. Brubaker, C. L. & Wendel, J. F. Reevaluating the origin of domesticated cotton (*Gossypium hirsutum*; Malvaceae) using nuclear restriction fragment length polymorphisms (RFLPs). *Am. J. Bot.* **81**, 1309–1326 (1994).
3. Yuan, D. et al. Parallel and intertwining threads of domestication in allopolyploid cotton. *Adv. Sci.* https://doi.org/10.1002/advs.202003634 (2021).
4. Wendel, J. F., Brubaker, C. L. & Seelanan, T. in *Physiology of Cotton* (eds Stewart, J. M. et al.) 1–18 (Springer, 2010).
5. Hoffmann, A. A. & Rieseberg, L. H. Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? *Annu. Rev. Ecol. Evol. Syst.* **39**, 21–42 (2008).
6. Wellenreuther, M. & Bernatchez, L. Eco-evolutionary genomics of chromosomal inversions. *Trends Ecol. Evol.* **33**, 427–440 (2018).
7. Fang, L. et al. Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. *Nat. Genet.* **49**, 1089–1098 (2017).
8. Wang, M. et al. Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nat. Genet.* **49**, 579–587 (2017).
9. Jia, G. et al. A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (*Setaria italica*). *Nat. Genet.* **45**, 957–961 (2013).
10. Zhou, Z. et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* **33**, 408–414 (2015).
11. Dai, P. et al. Extensive haplotypes are associated with population differentiation and environmental adaptability in upland cotton (*Gossypium hirsutum*). *Theor. Appl. Genet.* **133**, 3273–3285 (2020).
12. Yang, Z. et al. Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. *Nat. Commun.* **10**, 2989 (2019).
13. Huang, G. et al. 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 (2020).
14. Chen, Z. J. et al. Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. *Nat. Genet.* **52**, 525–533 (2020).
15. Ma, Z. et al. Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nat. Genet.* **50**, 803–813 (2018).
16. Mascher, M. et al. Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. *Nat. Genet.* **51**, 1076–1081 (2019).
17. Wang, W. et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43–49 (2018).
18. He, F. et al. Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. *Nat. Genet.* **51**, 896–904 (2019).
19. Zhao, G. et al. A comprehensive genome variation map of melon identifies multiple domestication events and loci influencing agronomic traits. *Nat. Genet.* **51**, 1607–1615 (2019).
20. Jia, Y., Sun, J. & Du X. in *World Cotton Germplasm Resources* (ed. Abdurakhmonov, I. Y.) 35–53 (IntechOpen, 2014).
21. Wendel, J. F., Rowley, R. & Stewart, J. M. Genetic diversity in and phylogenetic relationships of the Brazilian endemic cotton, *Gossypium mustelinum* (Malvaceae). *Plant Syst. Evol.* **192**, 49–59 (1994).
22. Hutchinson, J. B. Intra-specific differentiation in *Gossypium hirsutum*. *Heredity* **5**, 161–193 (1951).
23. He, S. et al. Introgression leads to genomic divergence and responsible for important traits in upland cotton. *Front. Plant Sci.* **11**, 929 (2020).
24. Zamir, D. Improving plant breeding with exotic genetic libraries. *Nat. Rev. Genet.* **2**, 983–989 (2001).
25. Wendel, J. F., Brubaker, C. L. & Percival, A. E. Genetic diversity in *Gossypium hirsutum* and the origin of upland cotton. *Am. J. Bot.* **79**, 1291–1310 (1992).
26. Beasley, J. O. The origin of American tetraploid *Gossypium* species. *Am. Nat.* **74**, 285–286 (1940).
27. Wang, L. et al. Alien genomic introgressions enhanced fiber strength in upland cotton (*Gossypium hirsutum* L.). *Ind. Crop. Prod.* **159**, 113028 (2021).
28. Campbell, B. T. et al. Genetic improvement of the Pee Dee cotton germplasm collection following seventy years of plant breeding. *Crop Sci.* **51**, 955–968 (2011).
29. Thyssen, G. N. et al. Whole genome sequencing of a MAGIC population identified genomic loci and candidate genes for major fiber quality traits in upland cotton (*Gossypium hirsutum* L.). *Theor. Appl. Genet.* **132**, 989–999 (2019).

30. Whittaker, D. J. & Triplett, B. A. Gene-specific changes in alpha-tubulin transcript accumulation in developing cotton fibers. *Plant Physiol.* **121**, 181–188 (1999).
31. Jones, F. C. et al. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55–61 (2012).
32. Cheng, C. et al. Ecological genomics of *Anopheles gambiae* along a latitudinal cline: a population-resequencing approach. *Genetics* **190**, 1417–1432 (2012).
33. Küpper, C. et al. A supergene determines highly divergent male reproductive morphs in the ruff. *Nat. Genet.* **48**, 79–83 (2016).
34. Wang, J. et al. A Y-like social chromosome causes alternative colony organization in fire ants. *Nature* **493**, 664–668 (2013).
35. Kirubakaran, T. G. et al. Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. *Mol. Ecol.* **25**, 2130–2143 (2016).
36. Berg, P. R. et al. Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated with large inversions. *Heredity* **119**, 418–428 (2017).
37. Lowry, D. B. & Willis, J. H. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol.* **8**, e1000500 (2010).
38. Fang, Z. et al. Megabase-scale inversion polymorphism in the wild ancestor of maize. *Genetics* **191**, 883–894 (2012).
39. Lee, C. R. et al. Young inversion with multiple linked QTLs under selection in a hybrid zone. *Nat. Ecol. Evol.* **1**, 119 (2017).
40. Rieseberg, L. H. Chromosomal rearrangements and speciation. *Trends Ecol. Evol.* **16**, 351–358 (2001).
41. Todesco, M. et al. Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature* **584**, 602–607 (2020).
42. Minka, T. P. & Deckmyn, A. maps: draw geographical maps. R package version 3.3.0 https://cran.r-project.org/web/packages/maps/ (2018).
43. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2019).
44. Fick, S. E. & Hijmans, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315 (2017).
45. QGIS Geographic Information System v.3.16.3 (Open Source Geospatial Foundation Project, 2020).

Methods

Sampling and genotyping. In this study, 3,278 accessions were analyzed. Among them, the 3K-TCG panel contained 3,248 samples, including 1,492 previously published genomes (PRJNA257154, PRJNA336461, PRJNA375965, PRJNA399050 and PRJNA414461) and 1,756 newly sequenced tetraploid cotton genomes (average 13.8× genome coverage) and the remaining accessions were other diploid cotton species (Supplementary Table 1). For sample collection, seeds of the cultivated accessions were sown in an incubator and young leaves were collected. Wild species were obtained from the National Wild Cotton Nursery (Sanya, China). Then, the genomic DNA of all samples was extracted by using the CTAB method[46]. The genomic DNA for each accession was used to construct 150-bp length paired-end sequencing libraries with an insert size of ~350-bp length. A total of ~61.64 terabases of new sequences was generated by the Illumina HiSeq 4000 sequencing platform (Novogene). The latest high-quality tetraploid cotton (*G. hirsutum* 'Texas Marker 1') genome[12] was used as the reference genome. Before mapping, we connected all the unassembled contigs into a pseudochromosome (named 'Contigs'). All reads of the 3,277 accessions were then aligned to the reference genome by BWA (v.0.7.12)[47] with default parameters (taking accession GB0300 as example, the input command was "bwa mem -t 8 -R "@RG\tID:GB0300\tLB:GB0300\tPL:ILLUMINA\tSM:GB0300" -M ref.fa GB0300_1.fq.gz GB0300_2.fq.gz >GB0300.sam") and all unaligned and low-quality (mapping quality <20) reads were discarded. Variants were called independently for each accession by GATK UnifiedGenotyper (v.3.8.0)[48] and then the raw population genotypes of the whole panel ($n=3,278$) were joined into a VCF file containing 65,696,715 SNPs (filter parameters, QD < 2.0| | MQ < 40.0| | FS > 60.0| | MQRankSum < −12.5| | ReadPosRankSum < −8.0) and 17,812,404 small InDels (filter parameters, QD < 2.0| | FS > 200.0| | InbreedingCoeff < −0.8) were identified in the raw SNP set. The effects of all variations were annotated by ANNOVAR[49].

Genetic diversity, population structure and divergence. A subset of 6,711,614 SNPs (3K-TCG, $n=3,248$) were filtered from the raw SNP set (MAF > 0.05, missing rate <0.2). The fixation statistic (F_{ST}) was used to estimate the level of genetic divergence by VCFtools (v.0.1.12b)[50]. First, the F_{ST} value at each SNP site was calculated for each pair of subgroups. Then, the average F_{ST} value was calculated in each 1-Mb length window with 100-kb length steps. For phylogenetic tree construction, PCA and population structure analysis, 66,969 SNPs was filtered by selecting one SNP from every other 100 SNPs (to reduce the computational burden and select the unlinked SNPs as possible). On the basis of the filtered 3K-TCG SNP set (one percentage), an approximate maximum-likelihood phylogenetic tree was constructed by FastTreeMP (v.2.19)[51] with default parameters. PCA was performed by the smartpca module of EIGENSOFT (v.6.1.4)[52]. Population structure was analyzed by ADMIXTURE[53] and the group membership for each accession was colored according to *K* values from 2 to 7 (Fig. 1a). The haplotype blocks were estimated by the strategy described in a previous study[11].

Sequencing and de novo assembly of the ICR_XLZ 7 genome. *G. hirsutum* 'Xinluzao 7' (ICR_XLZ 7) carrying the contrasting haplotype (Hap-A06-3 and Hap-A08-1) on chromosomes A06 and A08 was assembled to compare with that in the reference genome (ICR_TM-1, carrying Hap-A06-1 and Hap-A08-3) (Extended Data Fig. 1). The high-quality DNA was first sheared and concentrated by following the PacBio guidelines. Additionally, Hi-C experiments (~222.6 Gb Illumina reads, ~100× genome coverage) were also performed as previously described[54]. A 60-kb length library was constructed and sequenced on the basis of single-molecule real-time (SMRT) sequencing technology on the PacBio Sequel platform. A total of ~318 Gb of PacBio reads (N50 = 32,781 bp, ~138 × depth) were generated and assembled according to the Canu pipeline (v.1.8)[55]. The polished contigs (by Quiver) were further oriented and assembled by integrating Hi-C reads with a three-dimensional de novo assembly (3D DNA) pipeline[56]. The size of the assembled genome was ~2.3 Gb and the contig N50 length reached ~42.98 Mb (Supplementary Table 3).

Phenotyping. According to the pedigree background and phylogenetic tree, 1,260 accessions were initially screened to investigate fiber quality-related traits in various environments. Limited by the size of the population, it was difficult to cultivate the whole GWAS panel ($n=1,260$) in each environment. Therefore, the GWAS panel was equally divided into two subgroups ($n=630$) and grown in four environments (two locations for each environment) representing the major cotton cultivation regions in China for 2 yr (2017 and 2018). Shijiazhuang in Hebei Province (38.22° N, 114.32° E) and Anyang in Henan Province (36.07° N, 114.50° E) were selected to represent the YER. Yancheng in Jiangsu Province (33.34° N, 120.46° E) and Changsha in Hunan Province (28.38° N, 113.42° E) were selected to represent the YZR. Two adjacent fields in Shihezi in Xinjiang Province (44.40° N, 86.16° E and 44.41° N, 86.71° E) were selected to represent north Xinjiang region. Another two locations in Xinjiang, Kuche (41.82° N, 83.22° E) and Alaer (40.61° N, 81.33° E), were selected to represent the south Xinjiang region. Each accession was grown in the field with two random blocks and each block contained ~30 (YER and YZR) and ~60 (Xinjiang region) plants. All field management activities were strictly conducted according to local cultivation standards.

For fiber-quality testing, opened cotton bolls with uniform development condition were harvested in each block. After removing the seeds, the fiber quality was tested by a high-volume instrument (HVI9000) at the Urumqi Center of Supervision and Testing for the Quality of Cotton, Ministry of Agriculture, China. Three important fiber-quality properties—fiber length (mm), fiber strength (cN tex^{-1}) and fiber elongation rate (%)—were used to estimate best linear unbiased prediction values (two replicates for 2 yr) by using the R package lme4 (https://github.com/lme4/lme4)[57]. After discarding the accessions with low-quality data (missing or abnormal), the fiber-quality properties of 1,245 accessions were retained for further GWAS analysis.

GWAS. The genotypes of the GWAS panel ($n=1,245$) were filtered from the 3K-TCG SNP set with MAF > 0.05 and missing rate <0.2. A total of 1,122,352 high-quality SNPs were retained to perform a GWAS of the fiber-quality data in efficient mixed-model association expedited (EMMAX) software[58]. The significance threshold value of the GWAS was evaluated with the formula $P=0.05/n$ (where n is the total SNP number). The −log (*P* value) threshold for significance in the GWAS was ~7.35. To compare the effects of various favorable allele combinations, the median value of accessions carrying no favorable alleles (NA) was first set as the reference. The percentage increase in the median value (%) for other favorable allele combinations was calculated (Fig. 4b–d). To further identify potential candidate genes, we developed an analysis strategy by integrating multi-omics data (Extended Data Fig. 1).

Introgression analysis. The introgression analysis was based on the 3K-TCG SNP set. Here, a simplified workflow for identifying introgressed *G. barbadense* fragments is used as an example (Supplementary Fig. 8). First, we calculated the allele frequency at each SNP site (*P*) for both the *G. hirsutum* ($n=2,639$) and *G. barbadense* ($n=180$) populations. Then, the allele of any given *G. hirsutum* accession was compared with the alleles of the two populations site-by-site. For each accession, the similarity of every SNP site was calculated as follows: similarity value $= P_{Gbar} - P_{Ghir}$. The introgression index (introgressed from *G. barbadense* to *G. hirsutum*) was the average similarity value for every adjacent pair of the 500 SNP sites. All the genomic fragments with an introgression index >0.2 were defined as introgressed fragments (the adjacent fragments were connected). The introgressed fragments of the other 12 donor species were identified by using the same workflow.

Transcriptome. To identify the candidate genes in the regions with a GWAS signal, one superior-quality accession (GH1086) carrying all favorable alleles for fiber length (*FL2*, *FL3*, *FL4* and *FL5*) and strength (*FS1*, *FS2* and *FS3*) and an inferior-quality accession (GH1801) carrying the opposite alleles were planted in the experimental field in Anyang. Tissues representing the fiber cell initiation stage (ovules at −3, 0 and 3 DPA), fiber cell elongation stage (fiber at 5, 10 and 15 DPA) and fiber cell secondary wall synthesis stage (fiber at 20 and 25 DPA) were sampled in the field. For comparison, young root, stem and leaf tissues of both accessions were sampled in the greenhouse. Three replicates were prepared for each tissue. The total RNA of all samples was extracted using the RNAprep Pure Plant Kit (Tiangen). Messenger RNA sequencing was performed on the Illumina HiSeq 4000 sequencing platform (Novogene). Gene expression was calculated using RSEM software (v.1.3.1)[59].

Gene expression analysis. To assess the expression patterns of the candidate genes for fiber elongation rate, total RNA (~2 μg) was extracted from different tissues of selected samples (carrying opposite fiber elongation rate alleles) and reversely transcribed with EasyScript cDNA Synthesis SuperMix (TRANSGEN Biotech) in a 20-μl reaction mixture. Then 1 μl was used as template to perform qRT–PCR analysis. Three technical replicates per sample were analyzed and the average $2^{-\Delta\Delta C_t}$ values were used to determine the differences in gene expression by Student's *t*-test[60]. *Ubiquitin* (*Gh_A10G005800*) was used as an internal control in qRT–PCR results analysis. All primer sequences used in this study are listed in Supplementary Table 14.

***GhTUA2* overexpression vector construction and transformation.** The full-length complementary DNA sequence of *GhTUA2* was cloned and ligated into the two restriction sites (Xba1 and Sac1) of pBI121 vector under the control of CaMV 35S promoter. The primers were showed in Supplementary Table 14. Subsequently, the vector was transformed into *Arabidopsis* plants according to the floral dip method[61]. The transgenic seedlings were selected on MS agar medium containing 40 mg l^{-1} of kanamycin and 50 mg l^{-1} of cefotaxime, then transferred to soil and grown in a greenhouse before PCR confirmation. Homozygous transgenic *Arabidopsis* lines were obtained for further analysis. The wild-type *Arabidopsis* plants were used as the controls.

Statistical analyses. The two-tailed Student's *t*-test was performed in GraphPad Prism software (v.9.0).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All raw transcriptome data (PRJNA634606) and raw resequencing data (PRJNA605345) have been deposited at in the NCBI BioProject database. All supporting data (assembled genome sequence of *G. hirsutum* 'Xinluzao 7' (ICR_XLZ 7), genotype files for genetic diversity and population structure analysis and phenotype data for GWAS) are available in the cotton genomic variation database (CottonGVD) (http://120.78.174.209:30081/ftp).

Code availability

Introgression analysis pipeline can be accessed through https://github.com/sungaofei/3K-TCG.

References

46. Paterson, A. H., Brubaker, C. L. & Wendel, J. F. A rapid method for extraction of cotton (*Gossypium* spp.) genomic DNA suitable for RFLP or PCR analysis. *Plant Mol. Biol. Rep.* **11**, 122–127 (1993).
47. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
48. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
49. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
50. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
51. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
52. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
53. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
54. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
55. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
56. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
57. Bates, D., Mächler, M., Bolker, B. M. & Walker, S. C. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2014).
58. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
59. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
60. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the $_2{}^{-\Delta\Delta C_t}$ method. *Methods* **25**, 402–408 (2001).
61. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16**, 735–743 (1998).

Acknowledgements

This work was funded by the National Key Technology R&D Program, the Ministry of Science and Technology (grant nos. 2016YFD0100203 to X.D. and S.H. and 2016YFD0100306 to S.H.), the National Natural Science Foundation of China (grant nos. 31871677 to S.H. and 31671746 to X.D.), the Agricultural Science and Technology Innovation Program of the Chinese Academy of Agricultural Sciences and the National Crop Germplasm Resources Center (grant no. NICGR2019-12 to Y.J.). We thank the National Mid-term Gene Bank for Cotton at the Institute of Cotton Research, Chinese Academy of Agricultural Sciences, for providing the seeds; J. A. Udall of Southern Plains Agricultural Research Center, US Department of Agriculture for sharing the sequencing data in NCBI (PRJNA414461); K. Wang and F. Liu of the Institute of Cotton Research, Chinese Academy of Agricultural Sciences for providing the DNA samples of wild species and landraces; and J. Ma and X. Li (Research Institute of Economic Crops, Xinjiang Academy of Agricultural Sciences), Y. Li and C. Ye (Biotechnology Research Institute of Xinjiang Academy of Agricultural and Reclamation Sciences), Y. Qian and W. Jin (Institute of Cotton, Hebei Academy of Agriculture and Forestry Sciences), J. Liu and J. Zhao (Institute of Industrial Crops, Jiangsu Academy of Agricultural Sciences) and Z. Zhou (Hunan Agricultural University) for assisting in planting cottons and investigating phenotypes.

Author contributions

X.D. and S.H. conceived and designed the research. G.S., S.H., P.D. and Liyuan Wang performed the bioinformatics and data analysis. X.G., W.G., Y.J. and Z. Pan prepared the leaf tissues and extracted DNA samples. W.S., J.W., S.X., S.C., C.Y., Z.X., F.W., J.S., G.F., Liyuan Wang, Z. Peng, D.H., Liru Wang and B.P. participated in the phenotype data investigation. B.C. performed the qRT–PCR and overexpression experiment. S.H. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

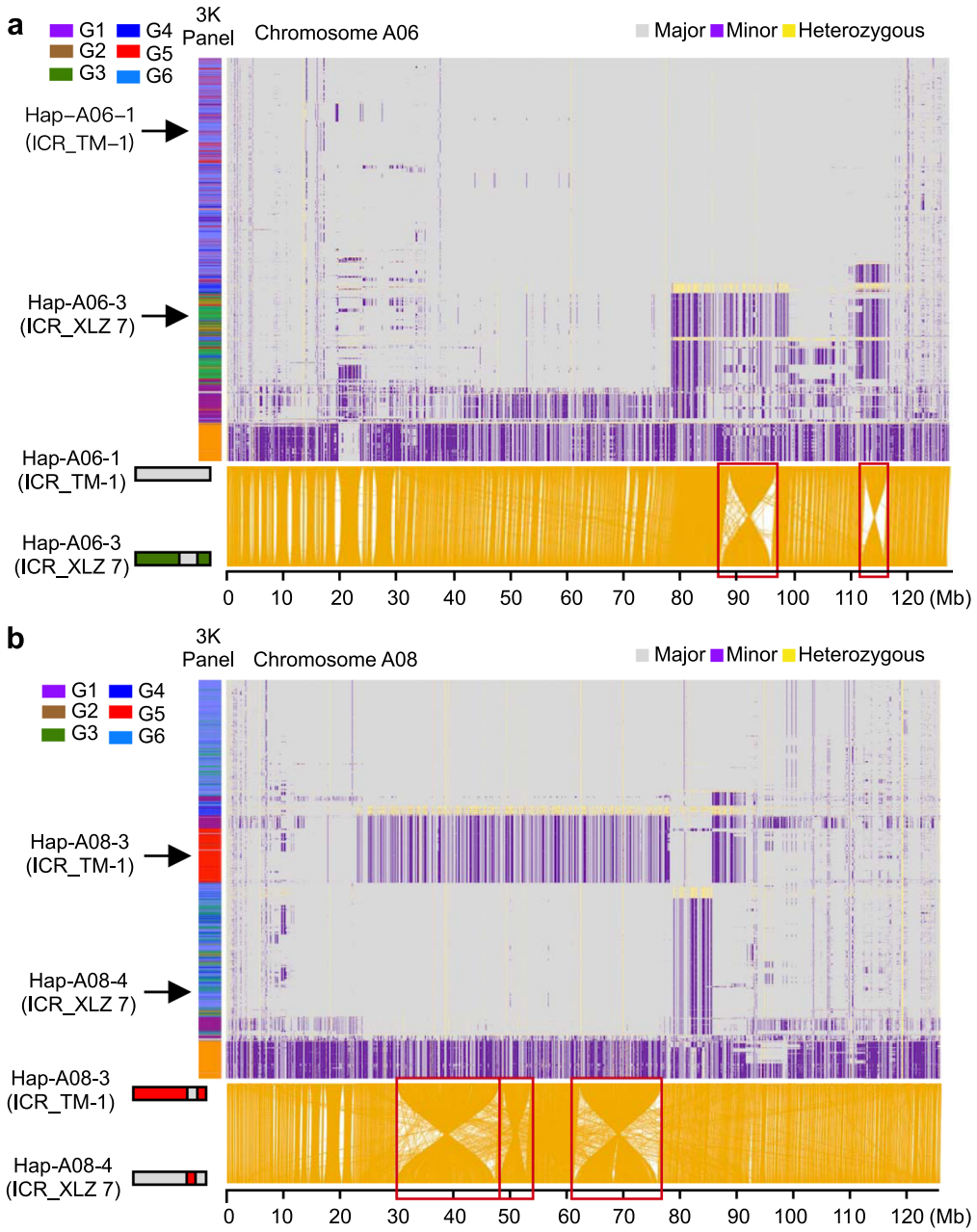
Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00844-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00844-9.

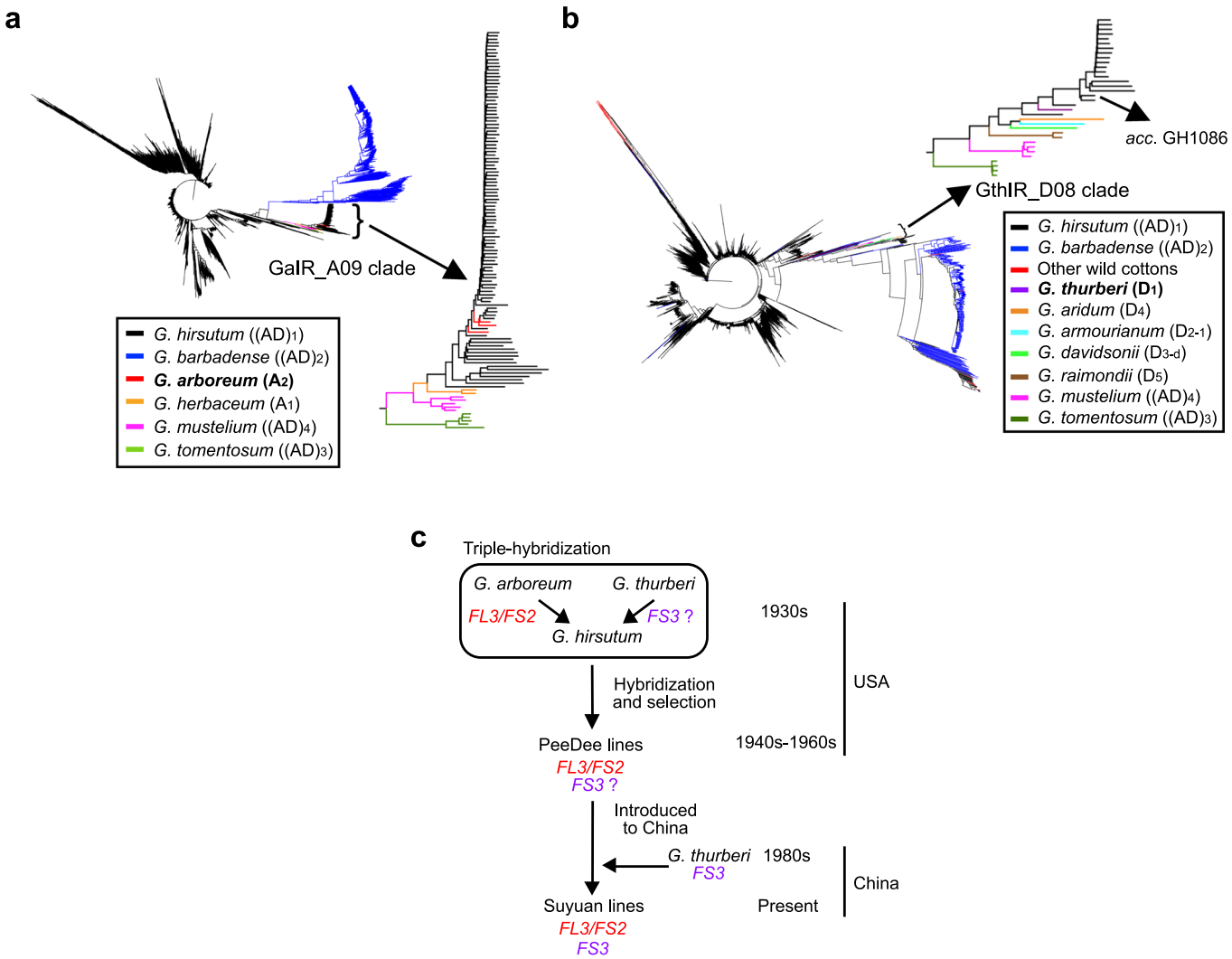
Correspondence and requests for materials should be addressed to X.D.

Peer review information *Nature Genetics* thanks Michael Bevan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

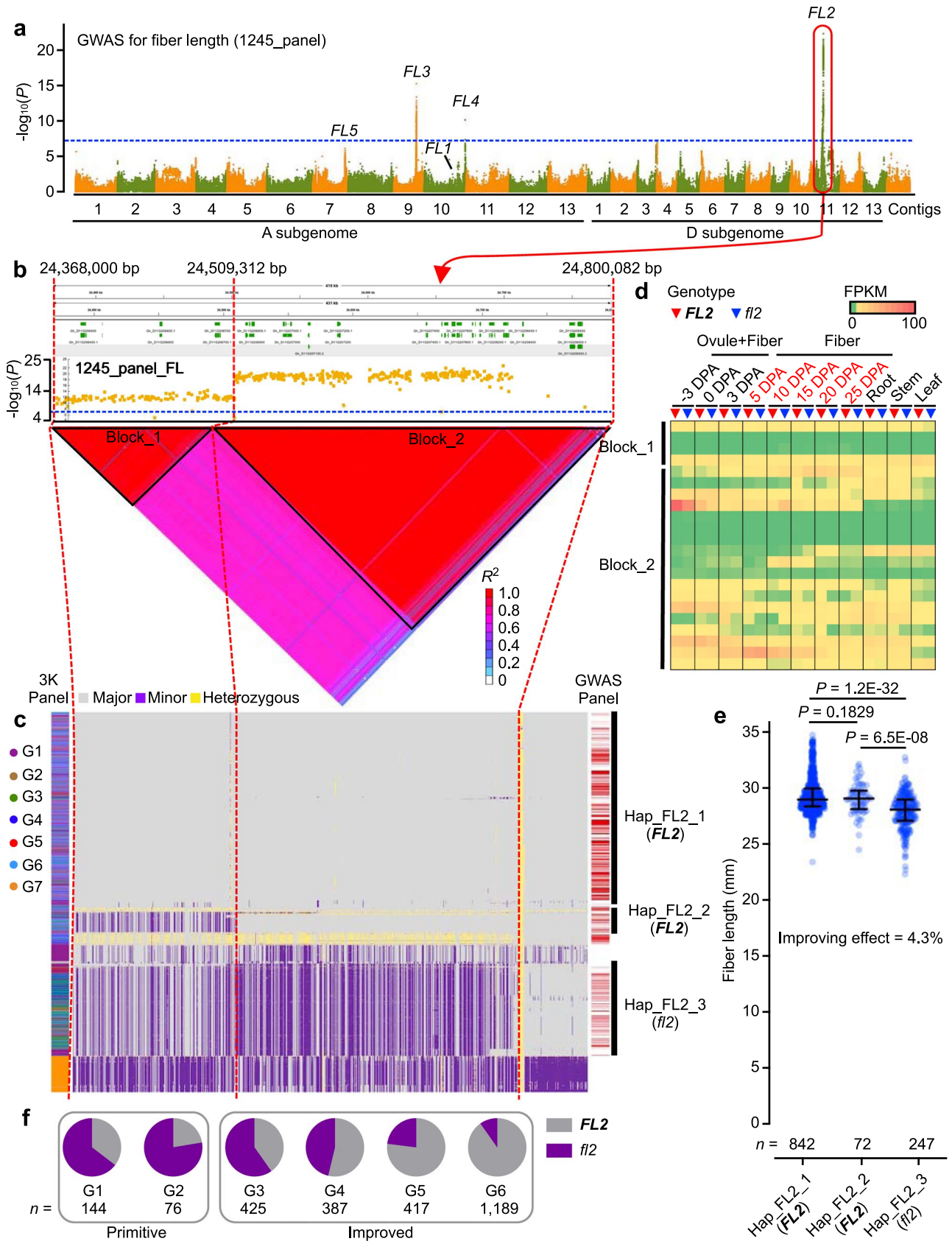
Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Extensive chromosomal inversions led to haplotype polymorphism on chromosomes A06 (a) and A08 (b) in *G. hirsutum*. For confirming chromosome inversions cause haplotype polymorphism on two chromosomes, we *de novo* assembled the genome of *G. hirsutum* 'Xinluzao 7' (ICR_XLZ 7) which carried the haplotype (Hap-A06-3 and Hap-A08-4) contrasting with the reference genome (ICR_TM-1, Hap-A06-1 and Hap-A08-3). Two and three major inversions are found on chromosomes A06 (**a**) and A08 (**b**), respectively (marked by red boxes).

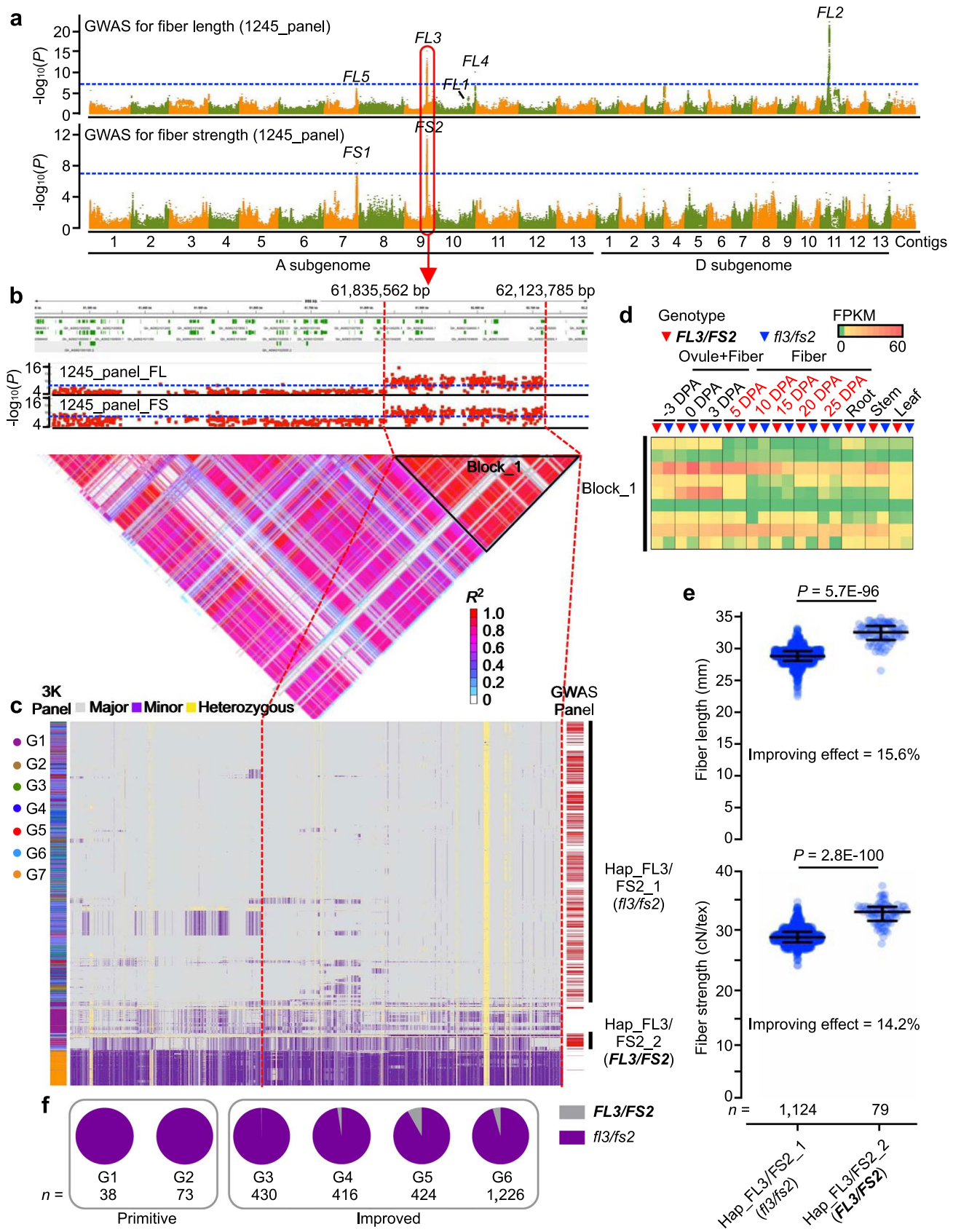


Extended Data Fig. 2 | Two fiber quality-related loci derived from introgressions of diploid cottons. a, Local clustering of 3,278 accessions based on the SNPs of *G. arboreum* introgressed region on chromosome A09 (GaIR_A09, ranged from ~61.8M to ~62.1 Mb) (*FL3/FS2*). A zoom-in view of the GaIR_A09 clade (right). *G. arboreum* (red branch) is clustered closely with *G. hirsutum* introgression lines. **b**, Local clustering of 3,278 accessions based on the SNPs of *G. thurberi* introgressed region on chromosome D08 (GthIR_D08, ranged from ~7.8Mb to ~60.4 Mb) (*FS3*). A zoom-in view of the GthIR_D08 clade (right). *G. thurberi* (purple branch) is clustered closely with all the introgression lines. **c**, The possible origination of *FL3/FS2* and *FS3* in Chinese elite cotton lines with superior fiber quality.



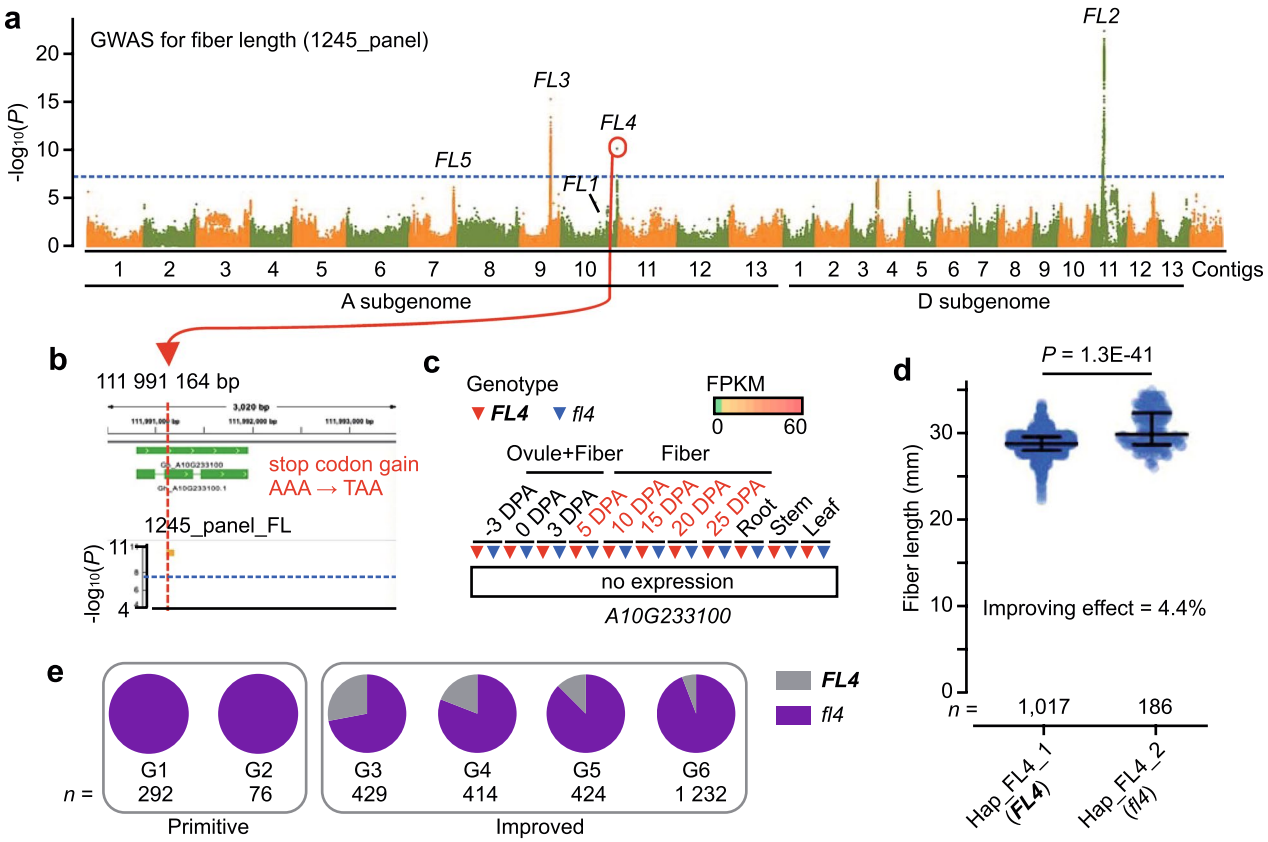
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | The genetic architecture of *FL2*. a, Manhattan plots of GWAS for fiber length in the GWAS panel. Red circle denotes the genomic location of *FL2* locus on chromosome D11. Blue dot line indicates the significant threshold of $-\log_{10}(P)$ value (7.35). **b**, Gene models (top), local Manhattan plot (middle) and local LD heatmap (bottom) in the *FL2* region. **c**, Haplotypes of *FL2* locus in the 3K-TCG panel. Accessions (vertical) are re-ordered according to the clustering based on regional SNPs (horizontal). The genotype of accessions is categorized into three haplotypes (Hap_FL2_1, Hap_FL2_2 and Hap_FL2_3). Colored lines (left) indicate the subgroup classification and the red lines (right) indicate the accessions selected for GWAS ($n = 1,245$). **d**, Gene expression profiles in the genomic region of *FL2*. Comparison of gene expression in various tissues between alternative haplotype (*FL2* and *fl2*). DPA, day postanthesis. **e**, Comparison of fiber length among different haplotypes of locus *FL2*. In scatter dot plot, horizontal lines and whiskers indicate the medians and interquartile ranges. Significances are tested by the two-tailed Student's *t*-test. **f**, Allelic frequency of locus *FL2* in *G. hirsutum* subgroups.

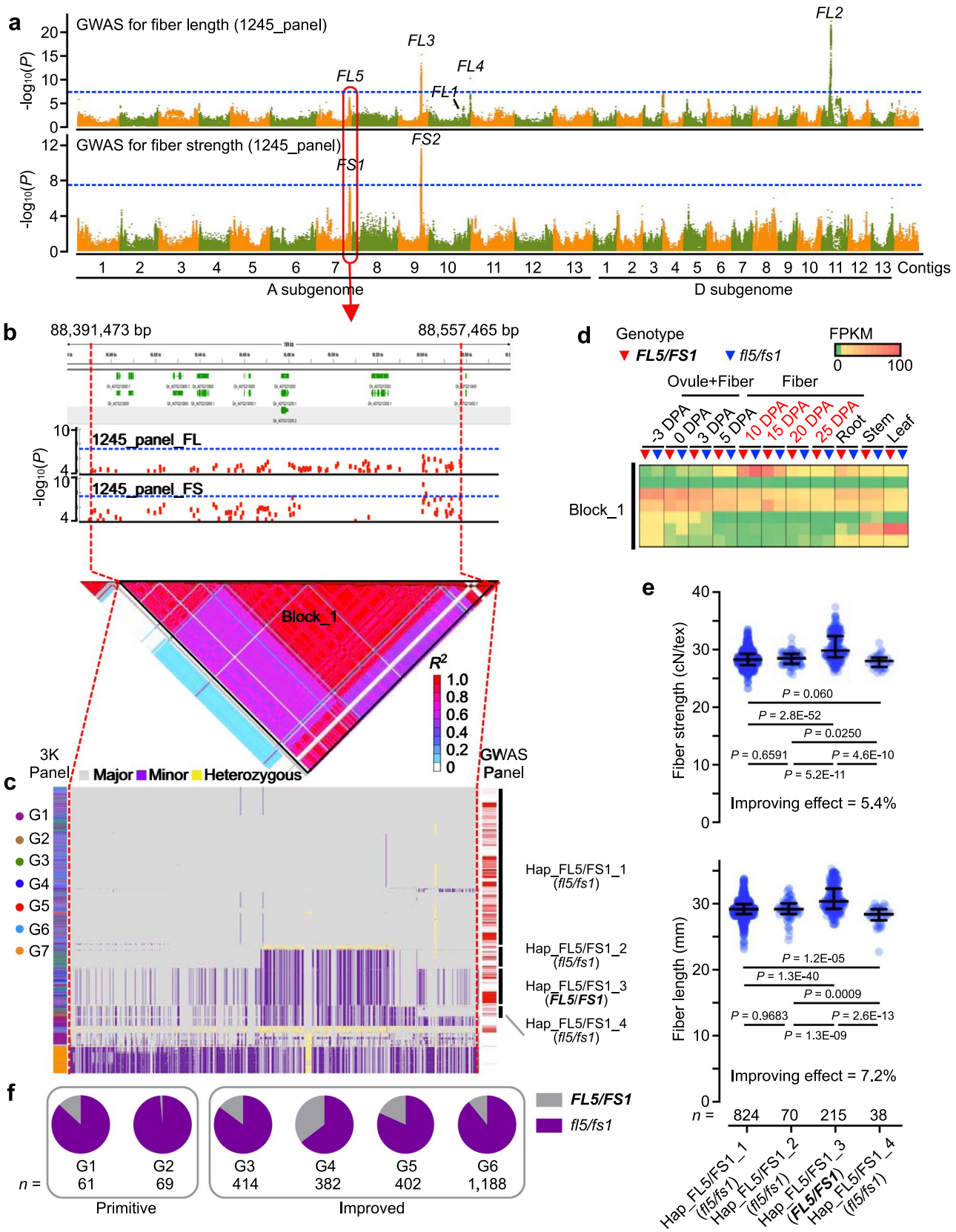


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | The genetic architecture of *FL3/FS2*. a, Manhattan plots of GWAS for fiber length (top) and fiber strength (bottom) in GWAS panel. Red circle denotes the genomic location of *FL3/FS2* locus on chromosome A09. Blue dot lines indicate the significant threshold of $-\log_{10}(P)$ value (7.35). **b**, Gene models (top), local Manhattan plots (middle) and LD heatmap (bottom) in the *FL3/FS2* region. **c**, Haplotypes of *FL3/FS2* locus in the 3K-TCG panel. Accessions (vertical) are re-ordered according to the clustering based on regional SNPs (horizontal). The genotype of accessions is categorized into two haplotypes (Hap_FL3/FS2_1 and Hap_FL3/FS2_2). Colored lines (left) indicate the subgroup classification, and the red lines (right) indicate the accessions selected for GWAS ($n = 1,245$). **d**, Gene expression profiles in the genomic region of *FL3/FS2*. Comparison of gene expression in various tissues between alternative haplotype (*FL3/FS2* and *fl3/fs2*). DPA, day postanthesis. **e**, Comparison of fiber length and fiber strength among different haplotypes of locus *FL3/FS2*. In scatter dot plot, horizontal lines and whiskers indicate the medians and interquartile ranges. Significances are tested by the two-tailed Student's *t*-test. **f**, Allelic frequency of locus *FL3/FS2* in *G. hirsutum* subgroups.

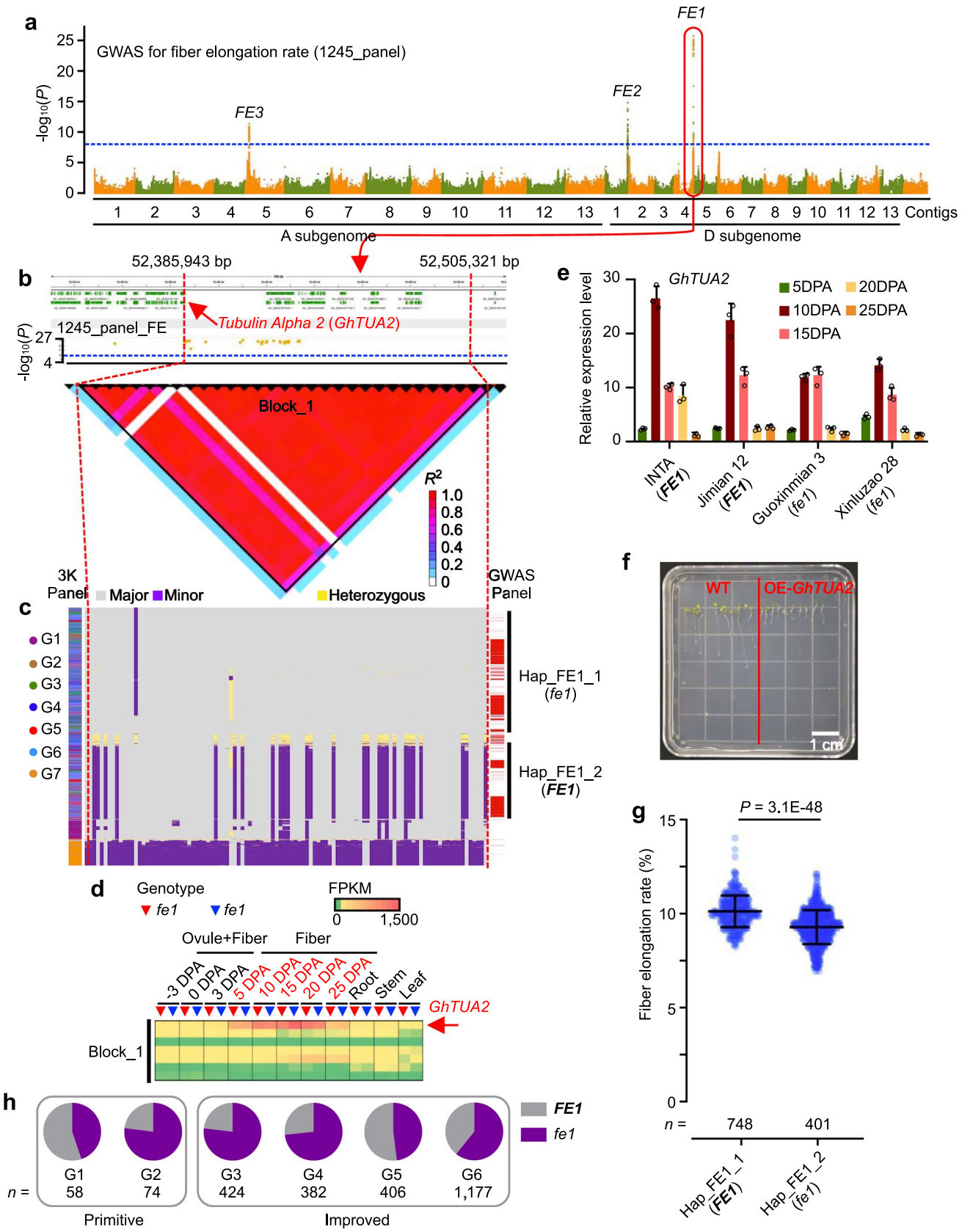


Extended Data Fig. 5 | The genetic architecture of *FL4*. a, Manhattan plots of GWAS for fiber length (top) and fiber strength (bottom) in GWAS panel. Red circle denotes the genomic location of *FL4* locus on chromosome A10. Blue dot lines indicate the significant threshold of -$\log_{10}(P)$ value (7.35). **b**, Gene model (top) and local Manhattan plots (bottom) in the *FL4* region. **c**, The expression of *Gh_A10G233100* in various tissues between alternative haplotype (*FL4* and *fl4*). DPA, day postanthesis. **d**, Comparison of fiber length among different haplotypes of locus *FL4*. In scatter dot plot, horizontal lines and whiskers indicate the medians and interquartile ranges. Significances are tested by the two-tailed Student's *t*-test. **e**, Allelic frequency of locus *FL4* in *G. hirsutum* subgroups.



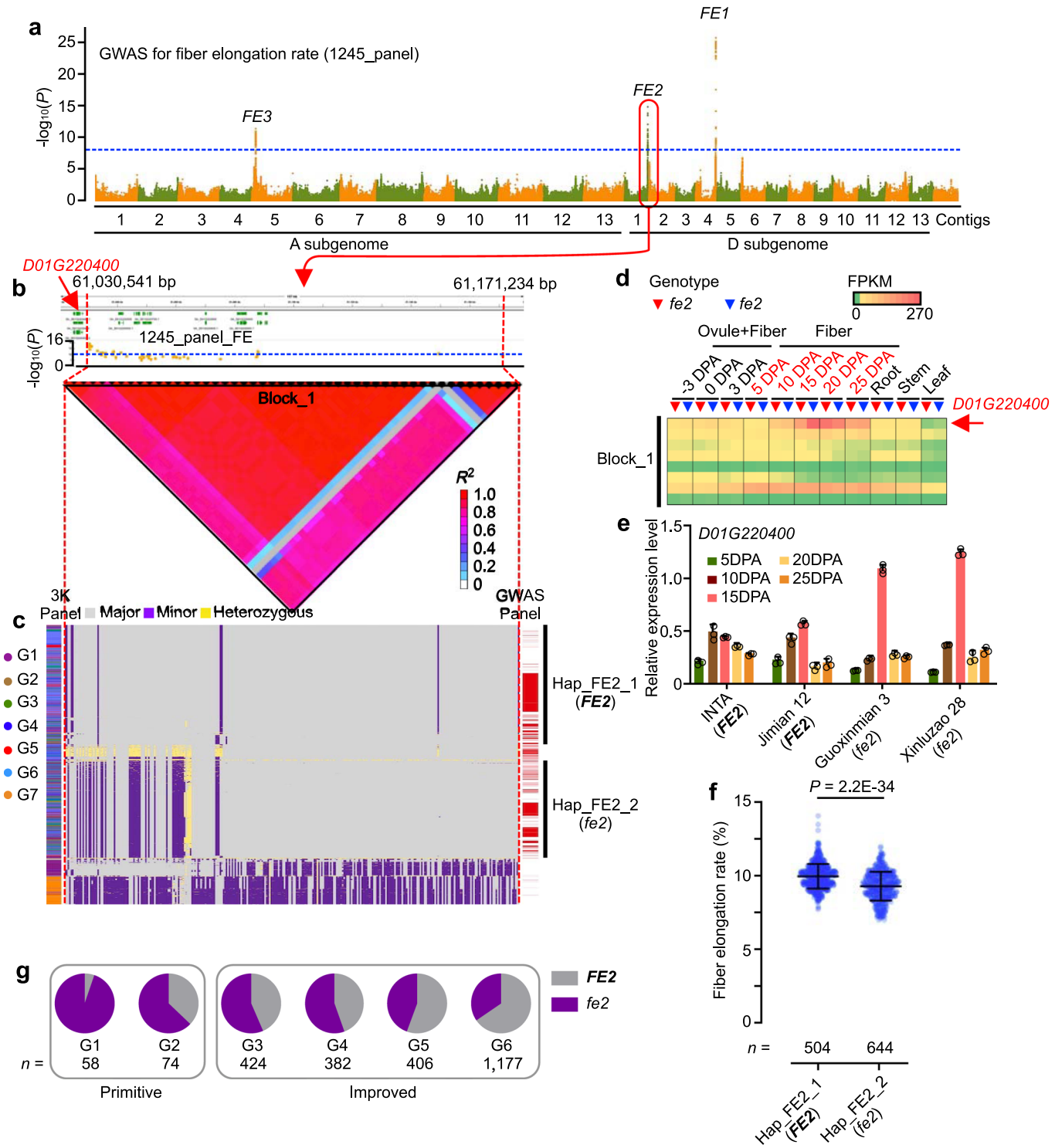
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | The genetic architecture of *FL5/FS1*. a, Manhattan plots of GWAS for fiber length (top) and fiber strength (bottom) in GWAS panel. Red circle denotes the genomic location of *FL5/FS1* locus on chromosome A07. Blue dot lines indicate the significant threshold of $-\log_{10}(P)$ value (7.35). **b**, Gene models (top), local Manhattan plots (middle), and LD heatmap (bottom) in the *FL5/FS1* region. **c**, Haplotypes of *FL5/FS1* locus in the 3K-TCG panel. Accessions (vertical) are re-ordered according to the clustering based on regional SNPs (horizontal). The genotype of accessions is categorized into four haplotypes (Hap_FL5/FS1_1, Hap_FL5/FS1_2, Hap_FL5/FS1_3 and Hap_FL5/FS1_4). Colored lines (left) indicate the subgroup classification, and the red lines (right) indicate the accessions selected for GWAS ($n = 1{,}245$). **d**, Gene expression profiles in the genomic region of *FL5/FS1*. Comparison of gene expression in various tissues between alternative haplotype (*FL5/FS1* and *fl5/fs1*). DPA, day postanthesis. **e**, Comparison of fiber length and fiber strength among different haplotypes of locus *FL5/FS1*. In scatter dot plot, horizontal lines and whiskers indicate the medians and interquartile ranges. Significances are tested by the two-tailed Student's *t*-test. **f**, Allelic frequency of locus *FL5/FS1* in *G. hirsutum* subgroups.

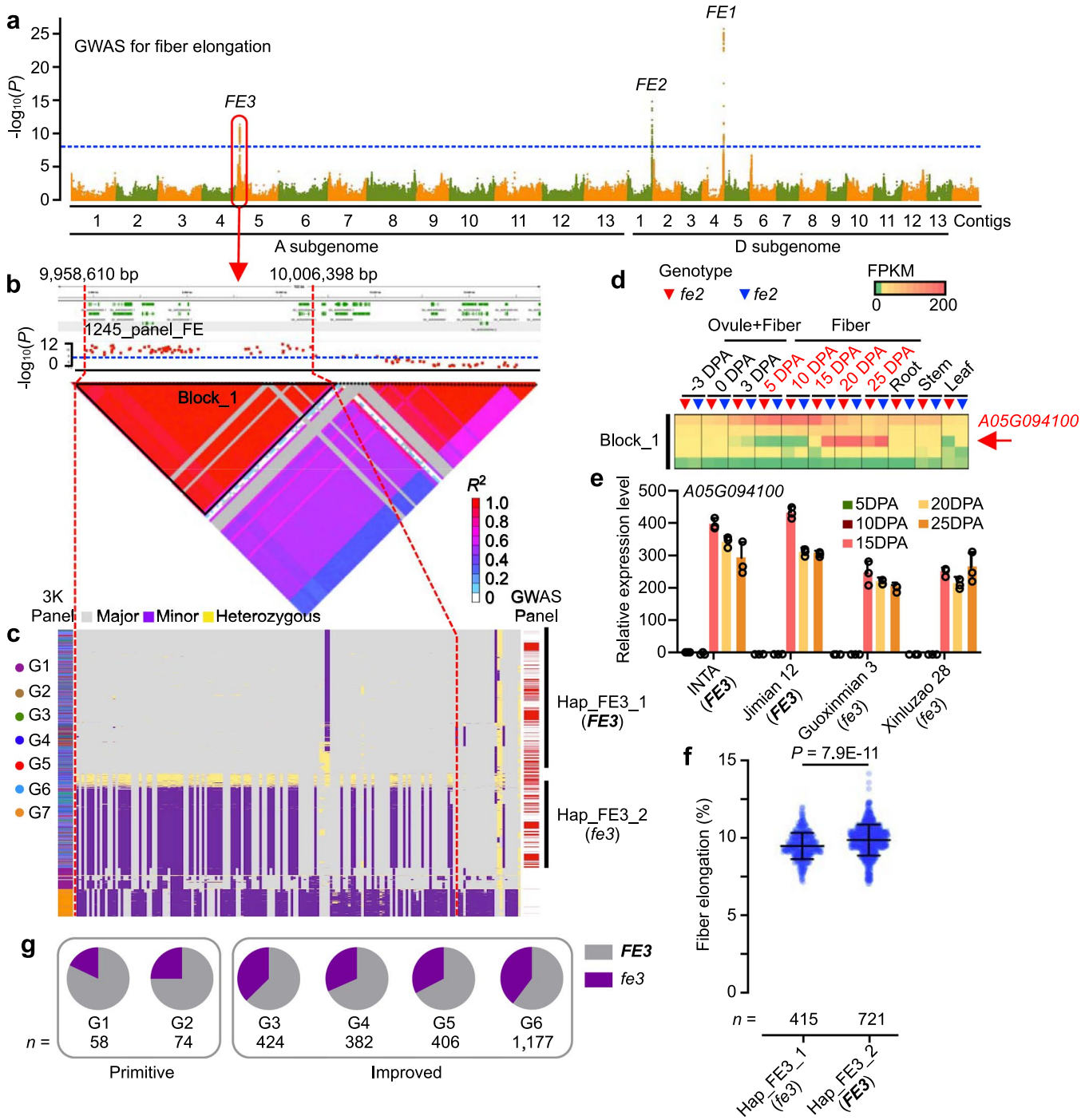


Extended Data Fig. 7 | See next page for caption.

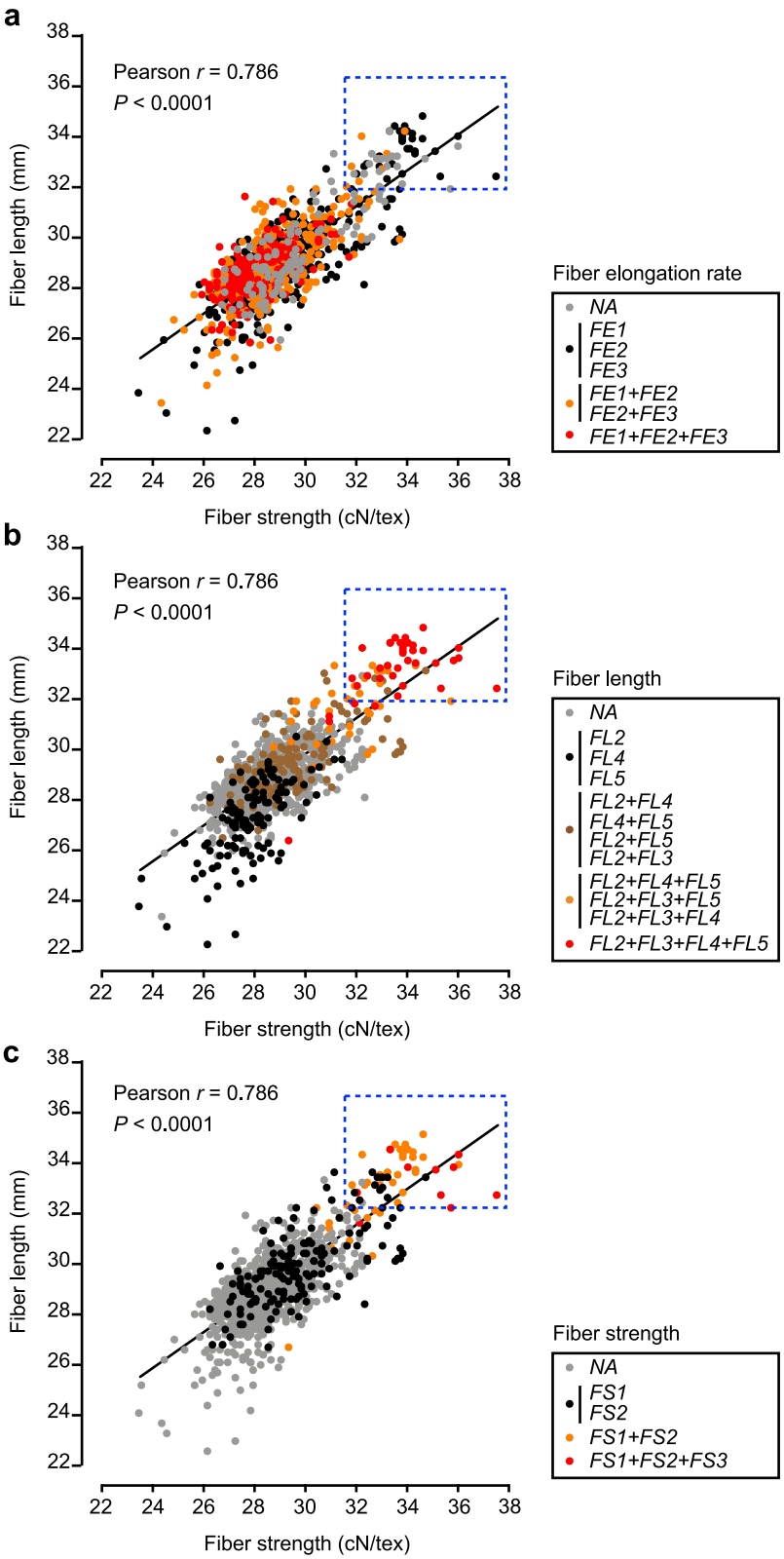
Extended Data Fig. 7 | The genetic architecture of *FE1*. **a**, Manhattan plots of GWAS for fiber elongation rate in GWAS panel. Red circle denotes the genomic location of *FE1* locus on chromosome D04. Blue dot lines indicate the significant threshold of $-\log_{10}(P)$ value (7.35). **b**, Gene models (top), local Manhattan plots (middle), and LD heatmap (bottom) in the *FE1* region. **c**, Haplotypes of *FE1* locus in the 3K-TCG panel. Accessions (vertical) are re-ordered according to the clustering based on regional SNPs (horizontal). The genotype of accessions is categorized into two haplotypes (Hap_FE1_1 and Hap_FE1_2). Colored lines (left) indicate the subgroup classification, and the red lines (right) indicate the accessions selected for GWAS ($n = 1,245$). **d**, Gene expression profiles in the genomic region of *FE1*. Comparison of gene expression in various tissues between alternative haplotype (*FE1* and *fe1*). DPA, day postanthesis. **e**, qRT–PCR analysis of *Gh_D04G181300* (*GhTUA2*) expression between accessions carrying alternative haplotype (mean ± s.d., $n = 3$ independent experiments). **f**, The root phenotype in *GhTUA2*-overexpressed *Arabidopsis*. **g**, Comparison of fiber elongation rate among different haplotypes of locus *FE1*. In scatter dot plot, horizontal lines and whiskers indicate the medians and interquartile ranges. Significances are tested by the two-tailed Student's *t*-test. **h**, Allelic frequency of locus *FE1* in *G. hirsutum* subgroups.



Extended Data Fig. 8 | The genetic architecture of *FE2*. **a**, Manhattan plots of GWAS for fiber elongation rate in GWAS panel. Red circle denotes the genomic location of *FE2* locus on chromosome D01. Blue dot lines indicate the significant threshold of -$\log_{10}(P)$ value (7.35). **b**, Gene models (top), local Manhattan plots (middle), and LD heatmap (bottom) in the *FE2* region. **c**, Haplotypes of *FE2* locus in the 3K-TCG panel. Accessions (vertical) are re-ordered according to the clustering based on regional SNPs (horizontal). The genotype of accessions is categorized into two haplotypes (Hap_FE2_1 and Hap_FE2_2). Colored lines (left) indicate the subgroup classification, and the red lines (right) indicate the accessions selected for GWAS ($n = 1{,}245$). **d**, Gene expression profiles in the genomic region of *FE2*. Comparison of gene expression in various tissues between alternative haplotype (*FE2* and *fe2*). DPA, day postanthesis. **e**, qRT-PCR analysis of *Gh_D01G220400* expression between accessions carrying alternative haplotype (mean ± s.d., $n = 3$ independent experiments). **f**, Comparison of fiber elongation rate among different haplotypes of locus *FE2*. In scatter dot plot, horizontal lines and whiskers indicate the medians and interquartile ranges. Significances are tested by the two-tailed Student's *t*-test. **g**, Allelic frequency of locus *FE1* in *G. hirsutum* subgroups.



Extended Data Fig. 9 | The genetic architecture of *FE3*. **a**, Manhattan plots of GWAS for fiber elongation rate in GWAS panel. Red circle denotes the genomic location of *FE3* locus on chromosome A05. Blue dot lines indicate the significant threshold of -$\log_{10}(P)$ value (7.35). **b**, Gene models (top), local Manhattan plots (middle), and LD heatmap (bottom) in the *FE3* region. **c**, Haplotypes of *FE3* locus in the 3K-TCG panel. Accessions (vertical) are re-ordered according to the clustering based on regional SNPs (horizontal). The genotype of accessions is categorized into two haplotypes (Hap_FE3_1 and Hap_FE3_2). Colored lines (left) indicate the subgroup classification, and the red lines (right) indicate the accessions selected for GWAS ($n = 1,245$). **d**, Gene expression profiles in the genomic region of *FE3*. Comparison of gene expression in various tissues between alternative haplotype (*FE3* and *fe3*). DPA, day postanthesis. **e**, qRT–PCR analysis of *Gh_A05G094100* expression between accessions carrying alternative haplotype (mean ± s.d., $n = 3$ independent experiments). **f**, Comparison of fiber elongation rate among different haplotypes of locus *FE3*. In scatter dot plot, horizontal lines and whiskers indicate the medians and interquartile ranges. Significances are tested by the two-tailed Student's *t*-test. **g**, Allelic frequency of locus *FE3* in *G. hirsutum* subgroups.



Extended Data Fig. 10 | Correlation of favorable allelic combinations for fiber elongation rate (a), fiber length (b), and fiber strength (c) in GWAS panel. Colored dots represent accessions carrying different allelic combinations. All accessions with superior fiber quality (fiber length > 32mm, fiber strength > 32cN/tex) are marked by blue rectangles.

nature research

Corresponding author(s): Xiongming Du

Last updated by author(s): Feb 5, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The climate data was downloaded from WorldClim (www.worldclim.org, ver.2.0). The new sequenced data (PRJNA605345 and PRJNA634606) was generated from Illumina HiSeq 4000 sequencing platform. The public re-sequencing data used was downloaded from NCBI (PRJNA414461, PRJNA399050, PRJNA375965,PRJNA336461,PRJNA257154). The reference genome and annotation files were download from www.cottonfgd.org (https://cottonfgd.org/about/download/assembly/genome.Ghir.CRI.fa.gz; https://cottonfgd.org/about/download/annotation/gene.Ghir.CRI.gff3.gz). For the phenotype data, the fiber quality data was measured by a high-volume instrument (HVI9000) and others were measured manually.
Data analysis	For genomic analysis, we used GATK (ver. 3.8.0), ANNOVAR (2014Nov12), VCFtools (ver. 0.1.12b), FastTreeMP (ver.2.19), EIGENSOFT (ver. 6.1.4), ADMIXTURE (ver. 1.23),Canu (ver.1.8), 3D DNA (ver. 180419), EMMAX (ver. 07Mar2010), RSEM (ver. 1.3.1). All statistical analysis was performed in Graphpad Prism (ver. 7.0). The BLUP values of phenotypic data was calculated using R package "lme4" (https://github.com/lme4/lme4) .

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All raw transcriptome data (PRJNA634606) and raw re-sequencing data (PRJNA605345) have been deposited at in the NCBI BioProject database. All supporting data (assembled genome sequence of G. hirsutum "ICR_XLZ_7", genotype files for genetic diversity and population structure analysis, and phenotype data for GWAS) is available in the cotton genomic variation database (CottonGVD) (http://47.112.109.227:30081).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size of planned GWAS panel was 1,260, after planning for two years in various environments, some accessions were failed to collected effective phenotype data. Finally, we kept 1,245 accessions for performing GWAS.
Data exclusions	The GWAS analysis in Fig. 4, we used 1,245 accessions, because the phenotype data of 1,245 accessions meet the requirement of BLUP calculation. Fiber micronaire data was excluded because it was not stable in different environments.
Replication	For qRT-PCR analysis, three independent experiments were performed; For RNA-seq data, three replicates for each samples were performed; For phenotype investigations, considering the environmental effect, we planted same genotype in two locations to represent Yellow River region, Yangtze River region, Northern Xinjiang Region and Southern Xinjiang Region, respectively. For each location, at two replications of each genotype were conducted. The detailed location information was described in Methods.
Randomization	The genotypes were randomly planted in each location.
Blinding	The experiment was conducted blindly. All genotypes were only labeled by numbers when planting, so the investigators did not know the exact accession names.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging