nature genetics

Asymmetric subgenome selection and *cis*-regulatory divergence during cotton domestication

Maojun Wang[1], Lili Tu[1], Min Lin[1,2], Zhongxu Lin[1], Pengcheng Wang[1], Qingyong Yang[1,2], Zhengxiu Ye[1], Chao Shen[1], Jianying Li[1], Lin Zhang[1], Xiaolin Zhou[1], Xinhui Nie[3], Zhonghua Li[1], Kai Guo[1], Yizan Ma[1], Cong Huang[1], Shuangxia Jin[1], Longfu Zhu[1], Xiyan Yang[4], Ling Min[4], Daojun Yuan[4], Qinghua Zhang[1], Keith Lindsey[5] & Xianlong Zhang[1]

Comparative population genomics offers an excellent opportunity for unraveling the genetic history of crop domestication. Upland cotton (*Gossypium hirsutum*) has long been an important economic crop, but a genome-wide and evolutionary understanding of the effects of human selection is lacking. Here, we describe a variation map for 352 wild and domesticated cotton accessions. We scanned 93 domestication sweeps occupying 74 Mb of the A subgenome and 104 Mb of the D subgenome, and identified 19 candidate loci for fiber-quality-related traits through a genome-wide association study. We provide evidence showing asymmetric subgenome domestication for directional selection of long fibers. Global analyses of DNase I–hypersensitive sites and 3D genome architecture, linking functional variants to gene transcription, demonstrate the effects of domestication on *cis*-regulatory divergence. This study provides new insights into the evolution of gene organization, regulation and adaptation in a major crop, and should serve as a rich resource for genome-based cotton improvement.

Early human domestication of wild plants was the first step in the development of modern crop varieties, and migration and differential directional selection over millennia has contributed to the adaptation of species in different environments for improved yield and quality traits[1]. In the current genomic era, high-throughput 'omics' technologies provide opportunities for detailed analyses of genetic change through domestication and for new, targeted and precise genome-based crop-breeding strategies[2,3].

Cotton is one of the most important economic crops in the world, both as a source of natural and renewable fiber for textiles, and as a source of seed oil and protein[4]. Allotetraploid Upland cotton arose from an intergenomic hybridization event approximately 1–2 Ma (ref. 5). Originally native to the Yucatan peninsula in Mesoamerica, it was first domesticated at least 4,000 to 5,000 years ago and was subsequently subjected to directional selection[6]. Modern varieties of cultivated cotton produce spinnable, fine white fibers, which are preferable to the sparser, coarse brown fibers of wild cotton. Previous molecular studies have shown that domestication has dramatically 'rewired' the transcriptome during fiber development[7,8]. What remains largely unknown, however, is the effect of human selection on the organization of the cotton genome and its gene-regulatory landscape. Using the recently published genome sequence of Texas Marker-1 (TM-1)[9,10] for comparison, we sought to address this question through a comprehensive population genomic analysis of multiple wild and cultivated cotton genotypes.

RESULTS

A genome-variation map for cotton

To construct an integrated variation map of Upland cotton, we collected a total of 352 diverse accessions for genomic sequence analysis[11], including 31 wild accessions and 321 cultivated accessions from around the world (**Fig. 1a** and **Supplementary Table 1**). A total of 6.1 Tb of sequence data were integrated, with an average depth of 6.9× (**Supplementary Table 1**). These data were mapped against the TM-1 genome to identify genomic variants[9]. We detected a total of 7,497,568 SNPs, 351,013 small deletion or insertion (indel) variants (shorter than 10 bp) and 93,786 structural variants (SVs) (**Table 1**, **Supplementary Fig. 1** and **Supplementary Tables 2–4**). The accuracy of the SNPs was estimated to be 98.2%, as determined by Sanger sequencing of 300 randomly selected SNPs in three individual accessions. In addition, we selected 50 representative accessions (10 wild and 40 cultivated cottons) from the 352 accessions for RNA sequencing (**Supplementary Table 5**) and generated 78,728 SNPs, of which more than 93.6% overlapped with SNPs from the resequencing data. This integrated variation data set provides a new resource for cotton genetics and breeding.

Cotton population properties and linkage disequilibrium

We explored the phylogenetic relationship among the 352 cotton accessions, by using whole-genome SNP analysis. These cottons were divided into three groups (**Fig. 1b** and **Supplementary Fig. 2**),

[1]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, China. [2]Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan, China. [3]Key Laboratory of Oasis Eco-agriculture of the Xinjiang Production and Construction Corps, College of Agronomy, Shihezi University, Shihezi, China. [4]College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China. [5]Department of Biosciences, Durham University, Durham, UK. Correspondence should be addressed to X.Z. (xlzhang@mail.hzau.edu.cn) or K.L. (keith.lindsey@durham.ac.uk).

Received 4 August 2016; accepted 10 February 2017; published online 6 March 2017; doi:10.1038/ng.3807

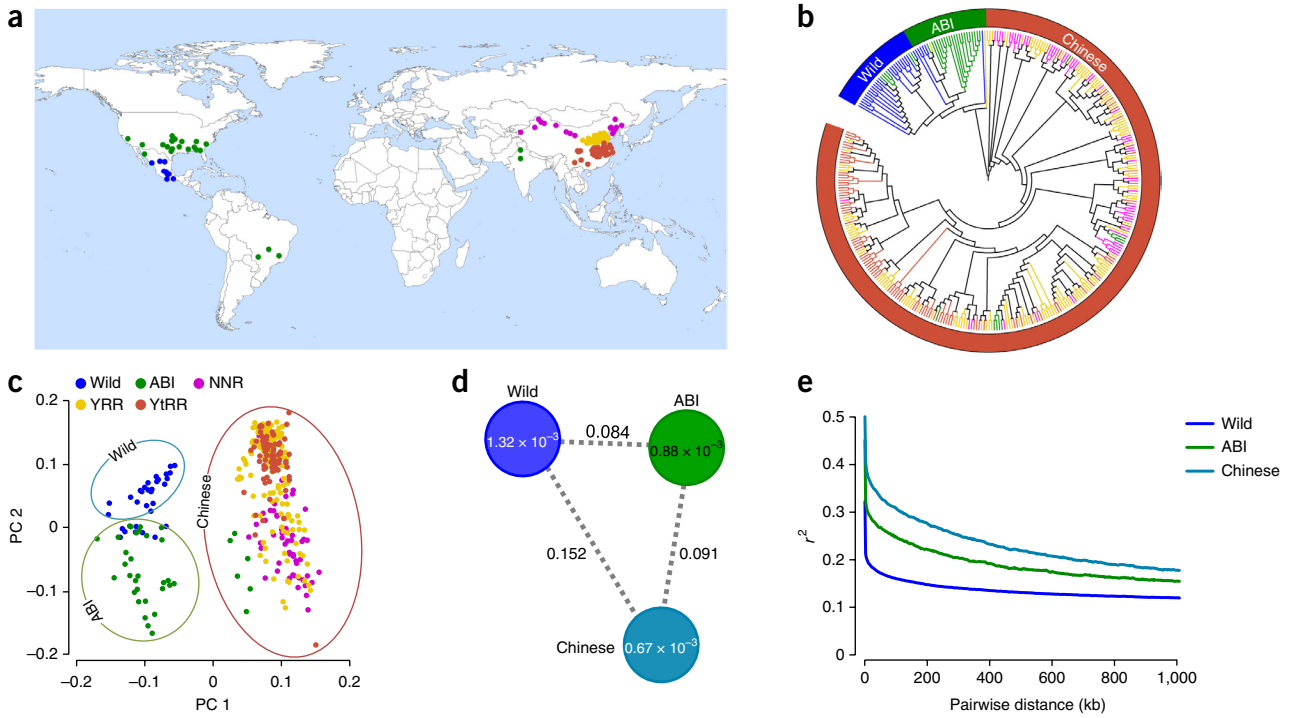

Figure 1 Geographic distribution and population diversity of Upland cotton accessions. (**a**) The geographic distribution of Upland cotton accessions. Each dot of a given color on the world map represents the geographic distribution of the corresponding cotton accession. (**b**) Neighbor-joining tree of all accessions constructed from whole-genome SNPs. The geographic distribution of each accession is represented by a tree branch with a color corresponding to that in **a**. The outer ring indicates groups emerging from the phylogenetic tree. Wild group, wild cottons; ABI group, cottons from America, Brazil and India; Chinese group, cottons from China. (**c**) PCA plots of the first two components for all accessions. The dot color scheme is as indicated in **a**. NNR, cottons from the NIR and the NSEMR; YRR, cottons from the YRR; and YtRR, cottons from the YtRR. PC, principal component. (**d**) Nucleotide diversity (π) and population divergence (F_{ST}) across the three groups. The value in each circle represents a measure of nucleotide diversity for this group, and the value on each line indicates population divergence between the two groups. (**e**) Decay of linkage disequilibrium (LD) in each group.

as supported by a principal component analysis (PCA; **Fig. 1c**). Wild cotton accessions clustered together (the wild group) except for a few accessions that clustered into the second group (the ABI group), which mainly comprised cottons from America, Brazil and India. The third group (the Chinese group) primarily consisted of cotton accessions in China, which were collected from the major Chinese cotton cultivation regions: the northwestern inland region (NIR), the northern specific early maturation region (NSEMR), the Yellow River region (YRR) and the Yangtze River region (YtRR)[12]. We observed that a few cotton accessions collected from North America clustered into the Chinese group, possibly because of the introduction of Upland cotton to China from America during the first 30 years of the twentieth century[13].

Crop species may experience population bottlenecks during domestication[14]. To examine this possibility in cotton, we measured the genetic diversity of each group by calculating π values. We found that the genetic diversity decreased from the wild cotton group ($\pi = 1.32 \times 10^{-3}$; the A subgenome (At, with lower-case 't' denoting tetraploid), 1.36×10^{-3}; the D subgenome (Dt), 1.25×10^{-3}) to the ABI group ($\pi = 0.88 \times 10^{-3}$; At, 0.96×10^{-3}; Dt, 0.66×10^{-3}) and to the Chinese group ($\pi = 0.67 \times 10^{-3}$; At, 0.72×10^{-3}; Dt, 0.56×10^{-3}) (**Fig. 1d** and **Supplementary Fig. 3**). These results indicated that a large amount of genetic diversity in both subgenomes has been lost during cotton domestication, especially for the Dt genome. Compared with other major crops, cotton exhibits low genetic diversity even within wild cotton accessions (**Supplementary Table 6**). To investigate population divergence, we calculated the population fixation statistics (F_{ST}) among groups (**Fig. 1d**). This analysis indicated large population divergence between the Chinese group and the wild group. Population divergence between the Chinese group and the ABI group was observed, thus suggesting that Upland cottons in China might have accumulated some genetic diversity from other sources after their introduction.

Linkage disequilibrium (LD; indicated by r^2) decreased with physical distance between SNPs in all cotton groups (**Fig. 1e**). The extent of LD for each group was measured as the chromosomal distance when LD decreased to half of its maximum value. In agreement with results for other crops, the extent of LD in cotton was lower in the wild group (84 kb; $r^2 = 0.16$) than in the cultivated groups. The LD decay occurred at 162 kb ($r^2 = 0.22$) in the ABI group and increased to 296 kb ($r^2 = 0.25$) in the Chinese group. The observed extent of LD in cultivated cotton groups was higher than that in cultivated maize (30 kb)[15], cultivated rice (123 kb in *Oryza indica*)[16] or cultivated soybean (133 kb)[17], but was lower than that of cultivated tomato (865.7 kb)[18]. For each group, the LD decay distance in the At was higher than that in the Dt (**Supplementary Fig. 4a**,**b**). For example, the extent of LD in the wild group was estimated to be 92 kb ($r^2 = 0.16$) in the At and 64 kb ($r^2 = 0.15$) in the Dt.

Selection signals during cotton domestication

Millennia of domestication have brought many morphological transformations to cotton, including an annualized growth cycle,

Table 1 Summary of the numbers of genomic variants in cotton populations

Category	Core set	Wild	ABI	Chinese
Sequence variants				
SNPs	7,497,568	5,603,940	4,528,637	4,632,445
Indels (<10 bp)	351,013	230,938	185,100	248,127
Structural variants (>10 bp)	93,786	76,821	60,201	59,663
Variants with effects on genes				
Nonsynonymous SNPs	86,633	67,914	55,179	63,270
SNPs that introduce stop codons	1,770	1,261	1,051	1,292
SNPs that disrupt stop codons	319	264	213	228
Frameshift indel	1,698	1,125	760	1,322
Nonframeshift indel	1,114	667	433	919
SVs that overlap with genes	12,511	11,876	10,963	11,193
SNPs in *cis*-regulatory elements				
Promoter DHSs	90,737	73,404	59,788	55,637
Enhancers	99,709	82,287	66,107	56,386

photoperiod insensitivity, loss of seed dormancy and superior spinnable white fiber[7,8]. To identify potential selective signals underlying these changes, we scanned genomic regions showing notable decreases in nucleotide diversity, by comparing cultivated accessions in the ABI and the Chinese groups with the wild group. In total, we identified 93 putative domestication sweeps supported by at least one likelihood method (XP–CLR) and π_w/π_c (w, wild; c, cultivated), occupying 178 Mb of the genome (74 Mb in the At and 104 Mb in the Dt) (**Fig. 2a**,**e** and **Supplementary Fig. 5**). These regions contain approximately 1,777 genes under selection, including 549 in the At and 1,228 in the Dt (**Supplementary Table 7**), thus suggesting that the Dt might be subject to stronger selection than the At.

To determine the genetic basis of cotton domestication, we overlapped selection sweeps with the locations of known quantitative trait locus (QTL) hotspots (containing at least four QTLs for the same trait within a 20-cM region)[19]. We found that 25 QTL hotspots overlapped with selection sweeps, and these QTL hotspots were associated with some major agronomic traits, including leaf hair and morphology, petal spots, cotton boll number and weight, resistance to *Verticillium* wilt and fiber quality (**Fig. 2a**,**e** and **Supplementary Table 8**). Of these QTL hotspots, 17 were associated with fiber-quality-related traits, including fiber length, fiber strength, micronaire value, fiber elongation rate and fiber uniformity. We investigated the nucleotide diversity of genes residing in the 25 QTL hotspots to identify putative loci with selection signals underlying these domestication-related traits. We identified 400 genes exhibiting low nucleotide diversity in cultivated cottons compared with wild cottons (π_w/π_c >4.8; **Supplementary Table 9**). We observed that 19 of 25 QTL hotspots containing 327 genes were located in the Dt.

Fiber-quality improvement has been one of the most important breeding goals during cotton domestication. To further identify candidate genes for fiber-quality-related traits, we performed a genome-wide association study (GWAS), using 267 cotton accessions and phenotypic data collected during 2012 and 2013 (**Supplementary Table 10**). Environmental effects were accounted for as described in our previous study[11]. We selected 2,020,834 high-quality SNPs with minor-allele frequency (MAF) >0.05 from the core set. The high-density SNP map was found to be superior to previous simple-sequence-repeat maps for GWAS[11]. A total of 19 association signals for fiber-quality-related traits, including 8 in the At and 11 in the Dt, were identified with $P < 4.9 \times 10^{-7}$ by using a compressed mixed linear model (MLM) (**Fig. 2b–d**,**f–i** and **Supplementary Table 11**). Among these associations, 16 signals were previously uncharacterized. Most candidate genes in the LD regions of GWAS signals were found to be highly expressed during cotton-fiber development (**Supplementary Table 12**). Three GWAS signals were identified as being under selection during domestication. Specifically, a GWAS signal associated with fiber strength was identified on chromosome A12 (**Fig. 2d**), where a *MYB DOMAIN PROTEIN* (*MYB*) gene and an *ACTIN DEPOLYMERIZING FACTOR* (*ADF*) gene reside. A GWAS signal associated with micronaire value was identified on chromosome D03 (**Fig. 2f**). This association was located near a *CINNAMYL ALCOHOL DEHYDROGENASE* (*CAD*) gene, which may have a role in the lignin pathway, which in turn affects the fiber micronaire value[20]. We also identified a GWAS signal associated with fiber elongation rate on chromosome D04 (**Fig. 2g**), where a gene encoding a gibberellin response protein is located.

Asymmetric subgenome domestication for long white fiber

The development of the trait for long white fiber in cultivated Upland cotton is the result of millennia of strong directional selection from its wild counterpart[21]. The observed change in fiber characteristics in cultivated Upland cotton is associated with changes in the expression patterns of genes affecting fibers[7,8,22]. However, the genetic basis of this developmental change remains largely unknown. To understand the relative contributions of the coexisting At and Dt genomes during domestication, we constructed ancestral pseudochromosomes to address this question at the subgenome level. We identified 15,456 homoeologous gene pairs, which we used to reconstruct an ancestral state for each of the 13 chromosomes in cotton diploids, similarly to a recent study in *Brassica*[23]. By comparing overlaps with domestication signals, we identified 620 homoeologous pairs that have been subjected to domestication selection in the At or Dt (192 in the At and 428 in the Dt) and only 34 homoeologous pairs with selection signals in both subgenomes (**Supplementary Fig. 6** and **Supplementary Table 13**). These results suggested that the coexisting subgenomes have been under asymmetric domestication selection (**Fig. 3a**).

Domestication selection probably increased fiber length through prolonging the elongation period of fiber development[21] (**Fig. 3b**). We identified a *FORMIN HOMOLOGY INTERACTING PROTEIN 1* (*FIP1*) gene, which is involved in actin-cytoskeleton organization[24,25], with a selection signal in the At but not in its Dt homoeolog (**Supplementary Fig. 6** and **Supplementary Table 14**). An altered regulation of the At *FIP1* in cultivated Upland cotton is predicted to be relevant to fiber elongation. Analysis of genes subjected to domestication selection in the Dt led us to identify 17 genes involved in stress-response pathways, such as reactive oxygen species (ROS) signaling (**Supplementary Fig. 6** and **Supplementary Table 14**).

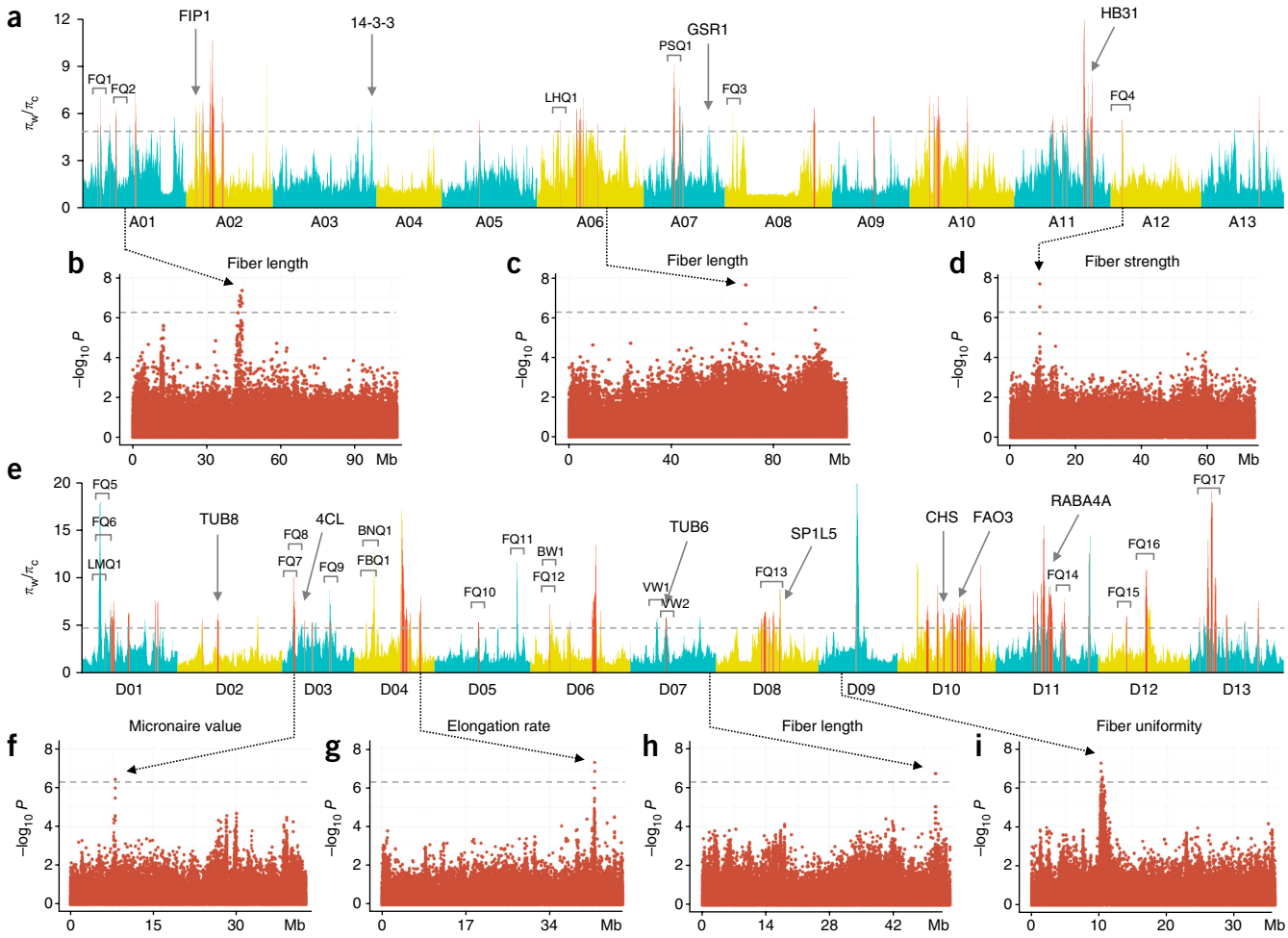

Figure 2 Genome-wide screening of domestication sweeps and GWAS on fiber-quality-related traits. (**a**,**e**) Selection signals in the A subgenome (At) (**a**) and the D subgenome (Dt) (**e**). The horizontal gray dashed lines show the genome-wide threshold for domestication sweeps identified from the ratio of nucleotide diversity between wild and cultivated cotton accessions (π_w/π_c >4.8). The results from the XP–CLR analytical tool are indicated by the red lines. The 25 QTL hotspots that overlap with domestication sweeps are shown in each chromosome. Genes with known function for fiber development under domestication selection are shown in corresponding chromosomes. These genes include *FIP1* (ref. 24), *14-3-3* (ref. 48), *GSR1* (ref. 49) and *HB31* (ref. 50) in the At, and *TUB6* (ref. 51), *TUB8* (ref. 51), *4CL*[52], *CHS*[52], *SP1L5* (ref. 53), *FAO3* (ref. 54) and *RABA4A*[55] in the Dt. The expression levels of these genes are shown in **Supplementary Figure 5**. (**b**–**d**) Significant GWAS associations for fiber length (**b**,**c**) and fiber strength (**d**) in the At. (**f**–**i**) Significant GWAS associations for micronaire value (**f**), fiber elongation rate (**g**), fiber length (**h**) and fiber uniformity (**i**) in the Dt. The horizontal gray dashed lines in **b**–**d** and **f**–**i** show the significance threshold of GWAS (1/*n* SNPs; 6.3). The other significant associations are presented in **Supplementary Table 11**.

High expression levels of these genes in wild cotton fibers may cause oxidative damage to developing fibers (**Supplementary Table 14**) and consequently may prevent rapid fiber elongation and promote developmental transition to secondary-cell-wall synthesis. Unexpectedly, we identified five homoeologous gene pairs involved in synthesis and deposition of secondary-cell-wall cellulose, which had selection signals only in the Dt (**Supplementary Table 14**). These genes, such as *TRICHOME BIREFRINGENCE-LIKE 43* (*TBL43*) and *COBRA-LIKE 4* (*COBL4*)[26,27], are highly expressed in wild cotton fibers at 20 days post anthesis (DPA). This finding may partially support the speculation that high concentrations of ROS in wild-cotton fiber development terminate fiber elongation in a manner associated with the developmental transition to secondary-cell-wall synthesis (**Fig. 3b**). This possibility was further supported by our genetic suppression of cytosolic *ASCORBATE PEROXIDASES* (*cAPXs*), in which an increased hydrogen peroxide content leads to the early initiation of secondary-cell-wall synthesis in fast-elongating fiber and gives rise to short fibers[28]. Therefore genetic evidence suggests that an asymmetric domestication selection between the At and the Dt subgenomes, which may modulate ROS levels, is associated with the development of the long-fiber trait in cultivated cotton (**Fig. 3b**).

Domestication has led to the transformation of cotton fiber from brown to white. To understand this phenomenon, we examined two homoeologous gene pairs subjected to domestication selection in only the Dt: *4-COUMARATE:COA LIGASE* (*4CL*) and *CHALCONE SYNTHASE* (*CHS*), which encode enzymes involved in the phenylpropanoid metabolic pathway[29] (**Fig. 3c** and **Supplementary Fig. 6**). For the *4CL* gene, we identified two nonsynonymous SNPs in the coding sequence and two SNPs residing in a promoter binding site for the Dof transcription factor (−369 bp to −378 bp; **Fig. 3c**). These SNPs displayed decreases in nucleotide diversity that occurred during domestication (**Fig. 3c**). We found that the two SNPs in the

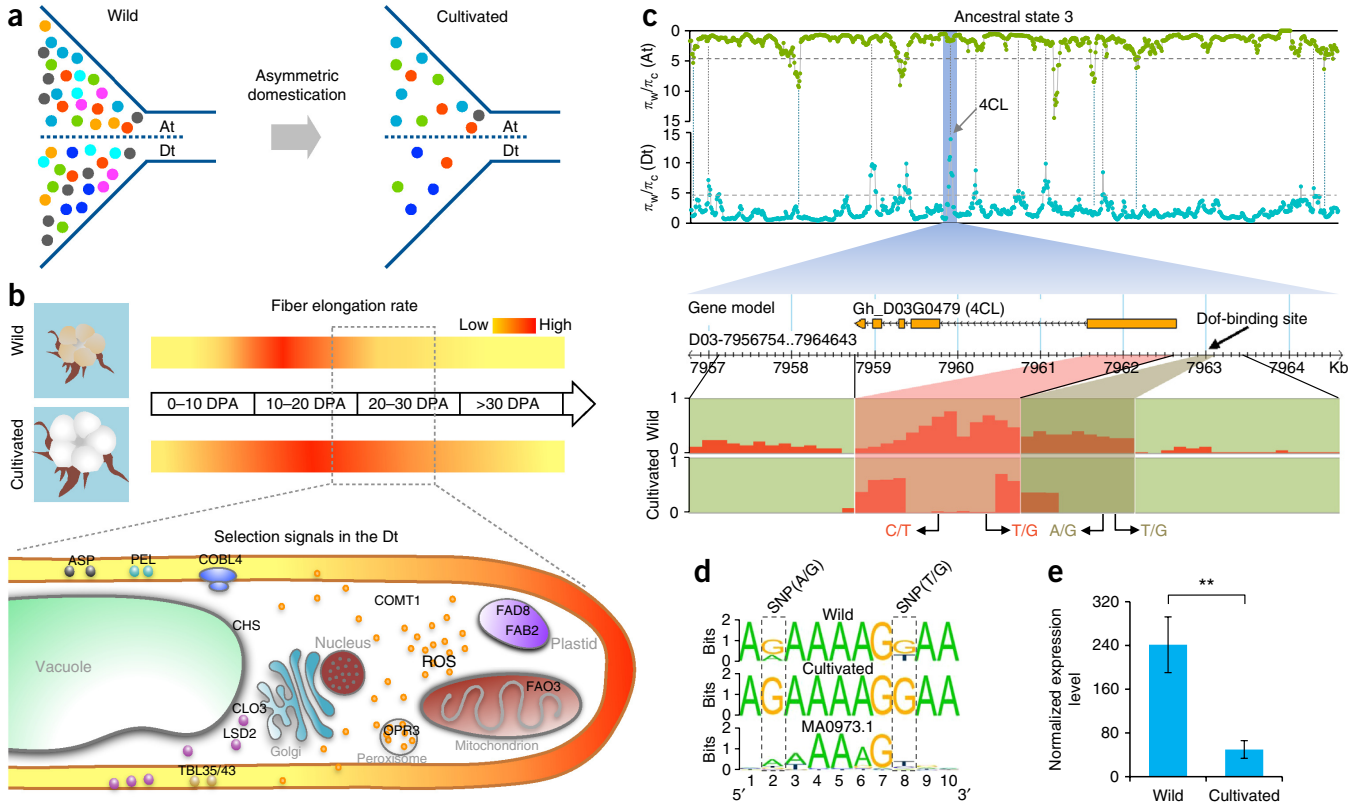

Figure 3 Asymmetric selection signals between the A subgenome (At) and the D subgenome (Dt). (**a**) Model of asymmetric domestication between the At and the Dt. (**b**) Effects of the Dt-specific selection signals on prolonged fiber elongation in cultivated cottons. Upper tract shows the morphological and developmental differences in fibers from wild and cultivated cottons. The heat map shows fiber elongation rate in wild and cultivated cotton. The dashed box shows a prolonged elongation period in cultivated cotton, with data from Applequist *et al.*[21]. The lower tract shows a model of a developing fiber. Genes with selection signals in the Dt are shown. These genes are downregulated in cultivated-cotton fiber development. Full descriptions of these genes are shown in **Supplementary Table 14**. (**c**) Selection signals in the *4-COUMARATE:COA LIGASE* (*4CL*) gene region. Upper tract shows asymmetric selection signals in ancestral state 3 in subgenomes. Vertical dashed lines show some homoeologous gene pairs with selection signals. Lower tract shows allele frequency of SNP variants in the *4CL* in wild and cultivated cotton groups. Nonsynonymous SNPs in the first exon are indicated in red. SNPs in the binding site for the Dof transcription factor are indicated in pale brown. (**d**) Sequence logos of the Dof-binding site in wild and cultivated cotton groups compared with that in *Arabidopsis* (JASPAR model MA0973.1). (**e**) Normalized expression levels of *4CL* at 10 DPA in wild and cultivated cottons, as determined by RNA-seq (two-sided *t*-test, $^{**}P < 0.01$). Error bars, s.d. of the normalized expression levels from different cotton accessions.

Dof-binding motif led to sequence variation departing from the canonical motif (**Fig. 3d**), thereby potentially affecting the transcription activity of *4CL*. This finding was supported by a significantly low expression level at 10 DPA in cultivated cottons (**Fig. 3e**). The enzyme CHS acts downstream of 4CL in this pathway, which catalyzes the first step of flavonoid synthesis, and its gene, *CHS*, has also been downregulated during domestication (**Supplementary Table 14**). Given the recognized functional role of flavonoids in brown-fiber pigmentation[29,30], selection signals at the *4CL* and *CHS* loci in the Dt may have driven the white-fiber trait characteristic of domestication.

Domestication effects on promoter *cis*-regulatory elements

Human selection of desirable agronomic traits not only affects the organization of functional genes but also may reshape the gene-regulatory landscape. In support of this idea, we found that many more variants were identified in intergenic compared with genic regions (**Table 1**). Specifically, intergenic noncoding variants can affect the activity of *cis*-regulatory elements (CREs)[31–33] and contribute to differential gene-expression patterns among populations (**Supplementary Fig. 7**). To investigate this possibility in cotton, we performed a global analysis of the effects of domestication on CREs in promoters.

We identified CREs in cotton with data from chromatin digestion with DNase I and subsequent sequencing (DNase–seq): active CREs can be detected because of their increased nuclease sensitivity, thereby reflecting an open chromatin conformation[34] (**Supplementary Fig. 8**). We identified a total of 188,360 DNase I–hypersensitive sites (DHSs) in cotton leaves and fibers, of which ~47% were common to both tissues (**Fig. 4a**). DHSs were preferentially identified in chromosomal arms, and approximately half were detected in promoter and intergenic regions (**Fig. 4b** and **Supplementary Fig. 9**). We found that DHSs were hypomethylated, in agreement with results from previous studies[34] (**Fig. 4c**). DHSs in promoter regions are commonly marked by high levels of active trimethylated histone H3 Lys4 (H3K4me3) and inactive trimethylated H3 Lys27 (H3K27me3), but exhibit low levels of active monomethylated H3 Lys4 (H3K4me1) and inactive dimethylated H3 Lys9 (H3K9me2) (**Fig. 4d**). Intergenic DHSs were also found to exhibit an enrichment in H3K4me3 and H3K27me3, but depletion of H3K9me2 and no enrichment of H3K4me1 (**Fig. 4e**). As predicted, the patterns of chromatin-modification marks in cotton were different between genic and TE regions (**Supplementary Fig. 10**). In addition, genes with promoter DHSs were generally expressed at higher levels in both tissues than in those without promoter DHSs

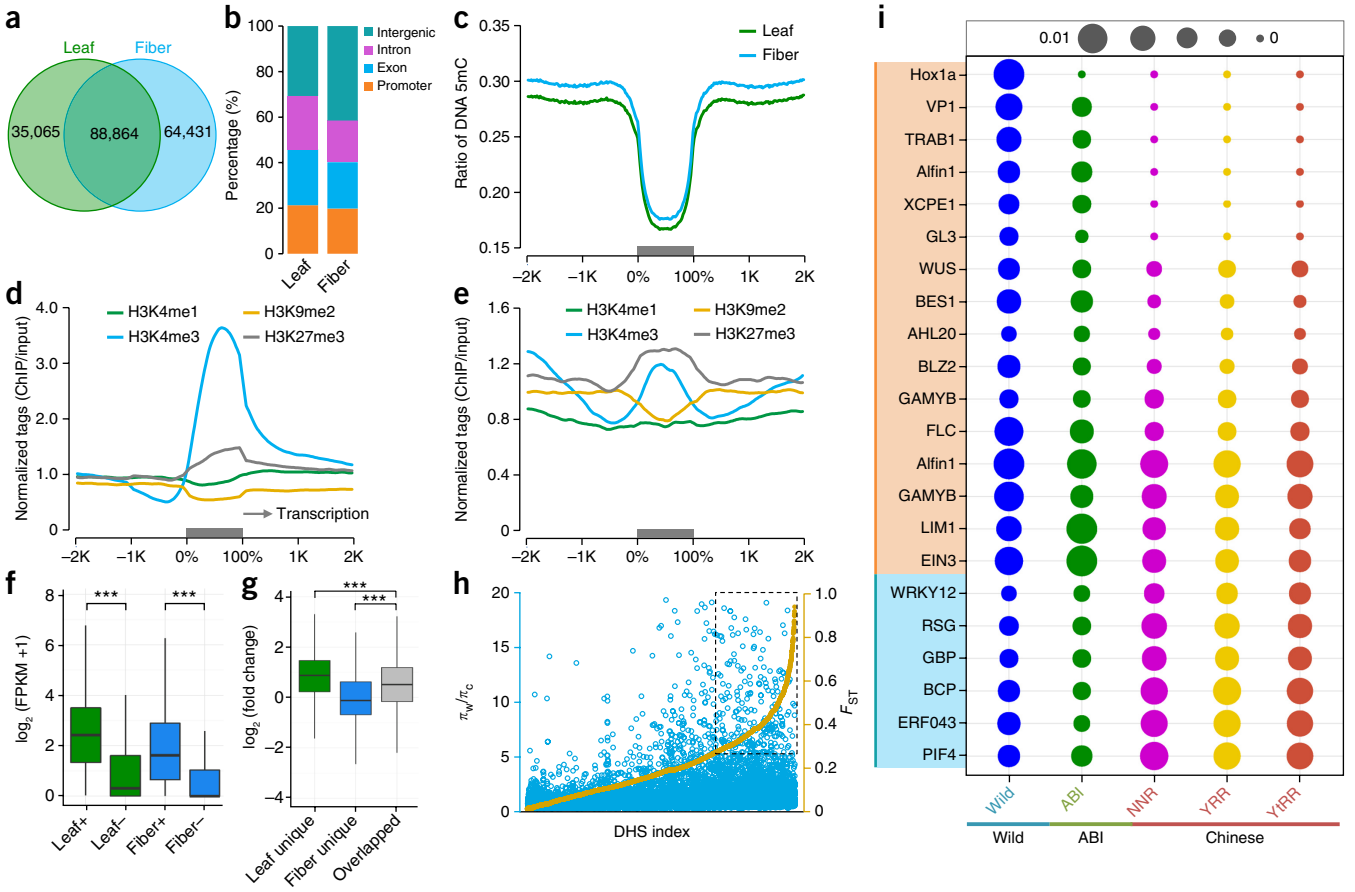

Figure 4 Characterization of cotton DNase I–hypersensitive sites (DHSs) and detection of selected DHSs during domestication. (**a**) Venn diagram showing the number of DHSs identified in cotton leaves and fibers at 10 DPA. (**b**) Genomic distribution of DHSs in genic and intergenic regions. (**c**) DNA-methylation levels (5-methylcytosine, 5mC) of DHSs in leaves and fibers. K denotes thousands. (**d**) Enrichment or depletion of chromatin-modification marks in promoter DHSs. ChIP, chromatin immunoprecipitation. (**e**) Enrichment or depletion of chromatin-modification marks in intergenic DHSs. For **c**–**e**, each DHS region was divided into 50 bins on average, and the flanking 2-kb regions were divided into 200 bins with equal length. (**f**) Comparisons of the expression levels among genes with promoter DHSs and those without promoter DHSs in leaf and fiber samples (two-sided Wilcoxon rank-sum test, $***P < 0.001$). In **f** and **g**, center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range. (**g**) Expression levels of tissue-specific promoter DHS-marked genes in leaf and fiber. For each group, the relative expression level was as the fold change of leaf versus fiber (two-sided Wilcoxon rank-sum test, $***P < 0.001$). (**h**) Detection of selected promoter DHSs during cotton domestication. All promoter DHSs were sorted on the basis of F_{ST}. The right *y* axis shows population divergence (F_{ST}) between wild and cultivated groups. Highly differentiated DHSs are indicated by the dotted box. (**i**) Nucleotide diversity of key transcription-factor-binding motifs identified from promoter DHSs in different cotton groups. For each motif, nucleotide diversity was scaled to the size of each respective circle.

(**Fig. 4f**), and tissue-specific promoter DHSs corresponded to higher levels of gene expression (**Fig. 4g**). These results demonstrated a close relationship between promoter DHS occurrence and relatively high transcriptional activity.

We examined genetic variants in promoter DHSs in our resequencing population and detected 90,737 SNPs in the 25,580 promoter DHSs (**Table 1**). Selection signals were detected for these promoter DHSs after domestication. A total of 738 DHSs (358 in the At and 380 in the Dt) were found to be under domestication selection ($\pi_w/\pi_c > 4.8$), of which 461 exhibited population divergence between cultivated and wild cotton accessions ($F_{ST} > 0.24$) (**Fig. 4h** and **Supplementary Table 15**). Of these DHSs with selection signals, we found that 281 DHS-related genes were differentially expressed. To investigate how variants in promoter DHSs might influence the expression of genes, we looked for associations between variants and transcription-factor-binding motifs. We discovered 178 motifs for 95 transcription factors in DHSs in cotton (**Supplementary Table 16**). We found that some well-known transcription-factor-binding motifs were under purifying selection in the cultivated groups, and some were under positive selection (**Fig. 4i** and **Supplementary Table 17**). For example, the TRAB1-binding motif, which relates to abscisic acid (ABA)-regulated transcription[35], was identified with a domestication-sweep signal. The GL3-binding motif, which participates in cotton-fiber initiation[36], was also under domestication selection. The PIF4-binding motif, which is important for high-temperature-mediated adaptation in plants[37], was identified as a positively selected motif. These results showed the effects of selection on *cis*-regulatory elements in promoter regions, which may be associated with the transcriptional regulation of genes contributing to desirable traits or adaptation.

Genome variation underlies distant regulatory divergence

Multiple genes can be considered to be organized into 'transcriptional factories' and transcribed in a high-order conformation[38]. A range of high-throughput methods, such as high-throughput chromosome conformation capture (Hi-C) and chromatin interaction analysis by paired-end tag sequencing (ChIA–PET), have been developed to

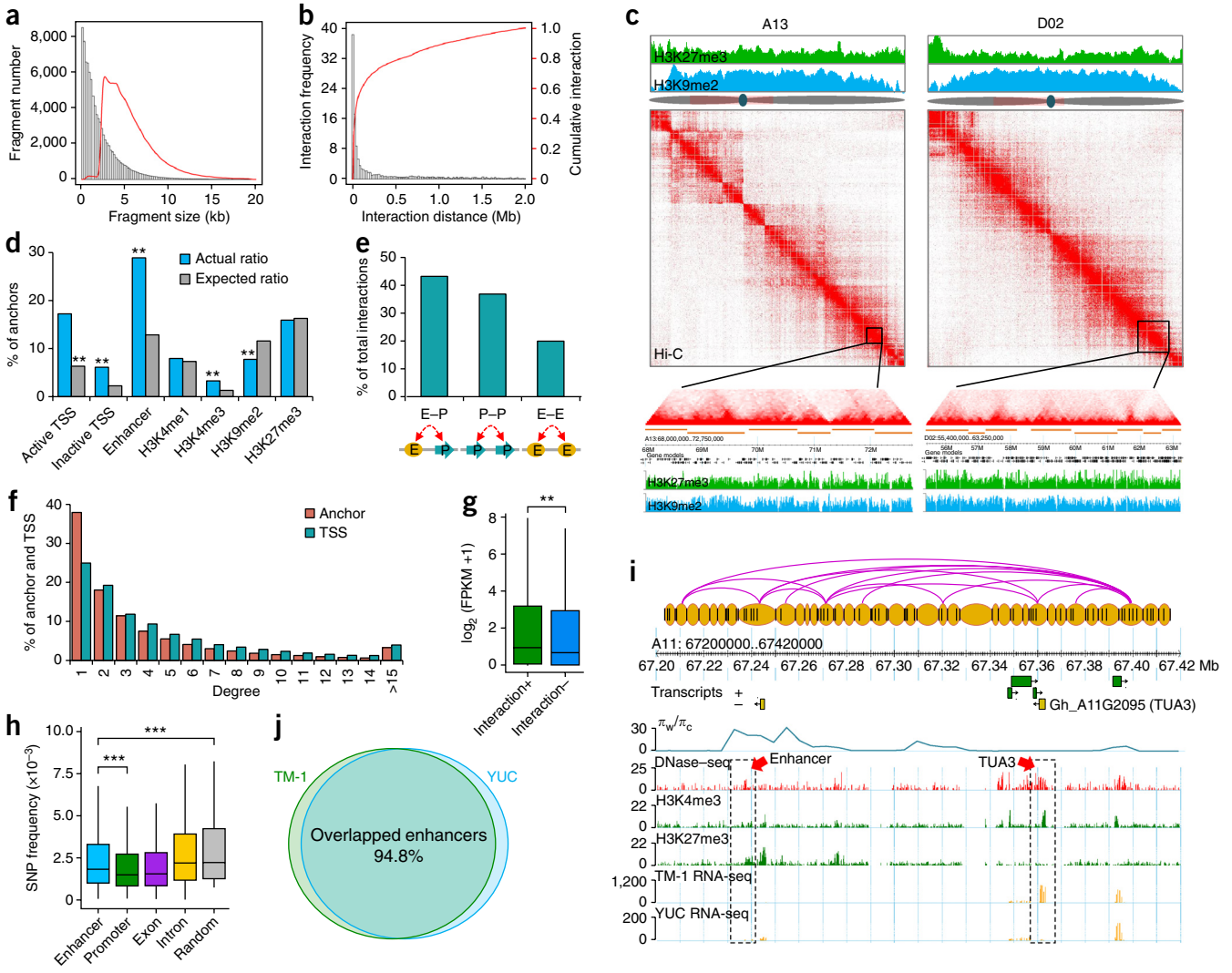

Figure 5 Characterization of the cotton chromatin interactome. (**a**) Size distribution of raw HindIII fragments (histogram) in the cotton genome, and anchors (red curve) used in this study. (**b**) Genomic distance between all interacting anchors. (**c**) Chromatin interaction in A13 and D02 chromosomes. Each heat map shows a normalized contact matrix, with strong contacts in red and weak contacts in white. Examples of TAD-like regions are shown below the heat maps. (**d**) Percentages of anchors involving *cis*-regulatory elements (CREs) and peaks of chromatin-modification marks. Actual enrichment ratios of CREs and ChIP peaks were compared with expected background values (two-sided Fisher's exact test, $^{**}P < 0.01$). TSS, transcription start site. (**e**) Percentage of promoter-centered interactions: enhancer–promoter (E–P), promoter–promoter (P–P) and enhancer–enhancer (E–E). (**f**) Degree distribution of anchors and promoters (TSS). (**g**) Expression analysis of genes with chromatin interaction and genes without chromatin interaction (two-sided Wilcoxon rank-sum test, $^{**}P < 0.01$). In **g** and **h**, center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range. (**h**) SNP frequencies in enhancer, promoter, exon and intron regions. SNP frequencies in these elements were compared with those in randomly selected genome regions (iteration number $n = 500$; two-sided Wilcoxon rank-sum test, $^{***}P < 0.001$). (**i**) One example of an enhancer under domestication selection. The upper tract shows chromatin interaction of anchors. Domestication selection is indicated by ratios of nucleotide diversity (π_w/π_c). The lower five tracts show enrichment of DNase–seq, ChIP–seq (H3K4me3 and H3K27me3) and RNA-seq in TM-1 and YUC accessions, respectively. The enhancer and gene regions are shown by dotted boxes and arrows. (**j**) Venn diagram showing the ratio of overlapped enhancers in TM-1 and YUC accessions.

elucidate 3D genome architecture in eukaryotic nuclei[39,40]. Several studies have shown that long-range chromatin interaction is an important mechanism for the regulation and coordination of gene transcription[41,42]. After we established a DHS landscape in cotton, our next aim was to characterize the effects of domestication on divergences in regulatory elements that are physically remote from, but functionally linked to, genes.

Hi-C analysis was carried out with the TM-1 accession to characterize global chromatin interactions. We generated 1.1 billion Hi-C paired-end reads, of which ~322 million were valid interaction reads (**Supplementary Table 18**). To exclude possible Hi-C bias, HindIII fragments of less than 2 kb were merged to obtain 305,682 chromosomal anchor regions (**Fig. 5a**). On the basis of a high-quality genome assembly of TM-1 (**Supplementary Fig. 11**), we used the Hi-C data to characterize the cotton chromatin interactome (**Supplementary Fig. 12**) and uncovered 737,377 midrange intrachromosomal interactions (20 kb–2 Mb). The number of interactions decreased rapidly with an increase in distance between sequences (**Fig. 5b**), but many topologically associated domain (TAD)-like regions were identified (**Fig. 5c**, **Supplementary Fig. 13** and **Supplementary Table 19**). We found that chromatin interactions were significantly enriched at promoters, distal DHSs such as enhancers and at regions marked by

the active chromatin mark H3K4me3, but were less frequent at regions marked by H3K9me2 (**Fig. 5d**).

Interactions involving promoters and distal DHSs, such as enhancers, were analyzed to construct a long-distance transcriptional-regulation map. We obtained 121,522 interactions, including 52,496 putative extragenic interactions (promoter to enhancer), 44,808 putative intergenic interactions between different genes (promoter–promoter interactions) and 24,218 putative enhancer–enhancer interactions (**Fig. 5e** and **Supplementary Table 20**). We found that only ~38% of putative enhancers and 25% of promoters were involved in a single interaction (**Fig. 5f**), thus indicating that transcription of most genes appears to be regulated by multiple long-range chromatin interactions. We observed that genes with relatively high levels of chromatin interaction exhibited higher expression levels than genes without interaction (**Fig. 5g**).

We next examined enhancer divergence. We identified a total of 99,709 SNPs in the 21,409 putative enhancers (**Table 1**). We found that enhancers exhibited a higher frequency of sequence variation than promoters or exons, and exhibited a lower frequency than introns (**Fig. 5h**), thus suggesting that enhancers have evolved rapidly. We then examined evidence for genomic selection of enhancers during cotton domestication. We identified 2,011 enhancers (496 in the At and 1,515 in the Dt) with selection signals associated with 1,651 gene promoters (**Supplementary Table 21**). One example indicated that an enhancer located 120 kb upstream of *TUBULIN ALPHA-3* (*TUA3*) has undergone strong selection, in agreement with the observed differentially high expression of *TUA3* in cultivated TM-1 compared with the wild YUC accession (**Fig. 5i**). DNase I digestion of chromatin on a representative wild cotton accession showed that more than 94% of enhancers were shared in wild and domesticated cottons (**Fig. 5j**), thus suggesting that domestication has had a limited effect on qualitative changes to enhancers.

DISCUSSION

Genome resequencing of 352 accessions of Upland cotton provided new insights into the genetic history of this important crop. By constructing a comprehensive variation map, we determined the genomic diversity and divergence of cotton. We found no obvious population divergence among geographic groups in China, probably because of frequent interbreeding by breeders within a short period after introduction. This finding is different from observations for cultivated rice and soybean, which were initially domesticated from wild forms in China several millennia ago[17,43]. Comparison of the wild and cultivated cottons has allowed for the identification of domestication sweeps. In this study, we primarily characterized key molecular signatures of selection responsible for spinnable long white fiber, of which some candidates were further identified by a GWAS analysis. We believe that these selection sweeps may enable future characterization of genes for other domestication-related agronomic traits. The variation map and selective sweeps should provide a valuable resource for future cotton improvement.

We identified the effects of domestication on *cis*-regulatory divergence through an integrated approach. We first presented a global analysis of DHSs by using DNase–seq, which has been demonstrated to be a highly efficient approach to map CREs in humans[44]. We provide evidence suggesting that directional selection through domestication has led to the divergence of CREs at promoters of at least some regulatory genes relevant to agronomic traits in cotton. Compared with promoters, distant CREs such as enhancers are less conserved among species but also are important for transcriptional regulation through long-range chromatin interactions[45]. With the DHS map, we provide a picture of 3D genome architecture, to link distant regulatory variants in enhancers to gene transcription. In contrast with isolated analyses of DHSs and 3D genome studies in *Arabidopsis*[46,47], this work is, to our knowledge, the first comprehensive functional interpretation of noncoding genetic variants in plants. Our approach to the characterization of functional variants should serve as a useful reference for other crops. Moreover, these data should facilitate future functional genomics studies for cotton and inform breeding strategies.

URLs. CottonGen TM-1 genome and annotation, https://www.cottongen.org/; iTOL browser, http://itol.embl.de/; HOMER software, http://homer.salk.edu/homer/; TRANSFAC database, http://www.gene-regulation.com/pub/databases.html; HiC-Pro software, https://github.com/nservant/HiC-Pro/.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank T. Zhang (Nanjing Agricultural University) for releasing resequencing data of wild cotton accessions. This work was supported by funding from the National Natural Science Foundation of China (31230056) to X.Z. and the National Natural Science Foundation of China (31201251) to D.Y.

AUTHOR CONTRIBUTIONS

X. Zhang, L.T. and M.W. conceived and designed the project. P.W., M.L., Q.Y., Z.Y., X. Zhou, M.W. and X.N. performed the experiments. M.W., P.W. and Q.Z. developed libraries and performed sequencing. M.W., C.S., J.L., L. Zhang, K.G., Y.M., Z. Li, C.H. and D.Y. analyzed the data. Z. Lin, L.T., S.J., L. Zhu, X.Y. and L.M. collected materials and managed sequencing. M.W. wrote the manuscript draft, which was revised by K.L. and X. Zhang.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Gross, B.L. & Olsen, K.M. Genetic perspectives on crop domestication. *Trends Plant Sci.* **15**, 529–537 (2010).
2. Varshney, R.K., Terauchi, R. & McCouch, S.R. Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. *PLoS Biol.* **12**, e1001883 (2014).
3. Crossa, J. *et al.* Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity (Edinb)* **112**, 48–60 (2014).
4. Chen, Z.J. *et al.* Toward sequencing cotton (*Gossypium*) genomes. *Plant Physiol.* **145**, 1303–1310 (2007).
5. Senchina, D.S. *et al.* Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. *Mol. Biol. Evol.* **20**, 633–643 (2003).
6. Stewart, J.M., Oosterhuis, D., Heitholt, J.J. & Mauney, J.R. *Physiology of Cotton* (Springer, 2010).
7. Rapp, R.A. *et al.* Gene expression in developing fibres of Upland cotton (*Gossypium hirsutum* L.) was massively altered by domestication. *BMC Biol.* **8**, 139 (2010).
8. Yoo, M.J. & Wendel, J.F. Comparative evolutionary and developmental dynamics of the cotton (*Gossypium hirsutum*) fiber transcriptome. *PLoS Genet.* **10**, e1004073 (2014).
9. Zhang, T. *et al.* Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat. Biotechnol.* **33**, 531–537 (2015).
10. Li, F. *et al.* Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nat. Biotechnol.* **33**, 524–530 (2015).
11. Nie, X. *et al.* Genome-wide SSR-based association mapping for fiber quality in nation-wide upland cotton inbreed cultivars in China. *BMC Genomics* **17**, 352 (2016).
12. Zhou, S.H. *Genogram of Cotton Varieties in China* (Sichuan Science and Technology Press, 2000).
13. Huang, Z.K. *Cotton Varieties and their Genealogy in China* (Chinese Agricultural Press, 2007).
14. Doebley, J.F., Gaut, B.S. & Smith, B.D. The molecular genetics of crop domestication. *Cell* **127**, 1309–1321 (2006).

15. Hufford, M.B. *et al.* Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44**, 808–811 (2012).
16. Huang, X. *et al.* Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961–967 (2010).
17. 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).
18. Lin, T. *et al.* Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet.* **46**, 1220–1226 (2014).
19. Said, J.I. *et al.* A comparative meta-analysis of QTL between intraspecific *Gossypium hirsutum* and interspecific *G. hirsutum* × *G. barbadense* populations. *Mol. Genet. Genomics* **290**, 1003–1025 (2015).
20. Han, L.B. *et al.* The dual functions of *WLIM1a* in cell elongation and secondary wall formation in developing cotton fibers. *Plant Cell* **25**, 4421–4438 (2013).
21. Applequist, W.L., Cronn, R. & Wendel, J.F. Comparative development of fiber in wild and cultivated cotton. *Evol. Dev.* **3**, 3–17 (2001).
22. Hovav, R. *et al.* The evolution of spinnable cotton fiber entailed prolonged development and a novel metabolism. *PLoS Genet.* **4**, e25 (2008).
23. Cheng, F. *et al.* Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in *Brassica rapa* and *Brassica oleracea*. *Nat. Genet.* **48**, 1218–1224 (2016).
24. Banno, H. & Chua, N.H. Characterization of the *Arabidopsis* formin-like protein AFH1 and its interacting protein. *Plant Cell Physiol.* **41**, 617–626 (2000).
25. Deeks, M.J., Hussey, P.J. & Davies, B. Formins: intermediates in signal-transduction cascades that affect cytoskeletal reorganization. *Trends Plant Sci.* **7**, 492–498 (2002).
26. Bischoff, V. *et al. TRICHOME BIREFRINGENCE* and its homolog *AT5G01360* encode plant-specific DUF231 proteins required for cellulose biosynthesis in *Arabidopsis*. *Plant Physiol.* **153**, 590–602 (2010).
27. Brown, D.M., Zeef, L.A., Ellis, J., Goodacre, R. & Turner, S.R. Identification of novel genes in *Arabidopsis* involved in secondary cell wall formation using expression profiling and reverse genetics. *Plant Cell* **17**, 2281–2295 (2005).
28. Guo, K. *et al.* Fibre elongation requires normal redox homeostasis modulated by cytosolic ascorbate peroxidase in cotton (*Gossypium hirsutum*). *J. Exp. Bot.* **67**, 3289–3301 (2016).
29. Feng, H. *et al.* Molecular analysis of proanthocyanidins related to pigmentation in brown cotton fibre (*Gossypium hirsutum* L.). *J. Exp. Bot.* **65**, 5759–5769 (2014).
30. Xiao, Y.H. *et al.* Transcriptome and biochemical analyses revealed a detailed proanthocyanidin biosynthesis pathway in brown cotton fiber. *PLoS One* **9**, e86344 (2014).
31. Maurano, M.T. *et al.* Large-scale identification of sequence variants influencing human transcription factor occupancy *in vivo*. *Nat. Genet.* **47**, 1393–1401 (2015).
32. Wittkopp, P.J. & Kalay, G. *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* **13**, 59–69 (2011).
33. Burgess, D.G., Xu, J. & Freeling, M. Advances in understanding *cis* regulation of the plant gene with an emphasis on comparative genomics. *Curr. Opin. Plant Biol.* **27**, 141–147 (2015).
34. Zhang, W. *et al.* High-resolution mapping of open chromatin in the rice genome. *Genome Res.* **22**, 151–162 (2012).
35. Hobo, T., Kowyama, Y. & Hattori, T. A bZIP factor, TRAB1, interacts with VP1 and mediates abscisic acid-induced transcription. *Proc. Natl. Acad. Sci. USA* **96**, 15348–15353 (1999).
36. Wang, S. *et al.* Control of plant trichome development by a cotton fiber MYB gene. *Plant Cell* **16**, 2323–2334 (2004).
37. Koini, M.A. *et al.* High temperature-mediated adaptations in plant architecture require the bHLH transcription factor *PIF4*. *Curr. Biol.* **19**, 408–413 (2009).
38. Cook, P.R. The organization of replication and transcription. *Science* **284**, 1790–1795 (1999).
39. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
40. Fullwood, M.J. *et al.* An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
41. Zhang, Y. *et al.* Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature* **504**, 306–310 (2013).
42. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).
43. Huang, X. *et al.* A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501 (2012).
44. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
45. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
46. Zhang, W., Zhang, T., Wu, Y. & Jiang, J. Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in *Arabidopsis*. *Plant Cell* **24**, 2719–2731 (2012).
47. Wang, C. *et al.* Genome-wide analysis of local chromatin packing in *Arabidopsis thaliana*. *Genome Res.* **25**, 246–256 (2015).
48. Zhou, Y. *et al.* Cotton (*Gossypium hirsutum*) 14-3-3 proteins participate in regulation of fibre initiation and elongation by modulating brassinosteroid signalling. *Plant Biotechnol. J.* **13**, 269–280 (2015).
49. Jakoby, M.J. *et al.* Transcriptional profiling of mature *Arabidopsis* trichomes reveals that *NOECK* encodes the MIXTA-like transcriptional regulator MYB106. *Plant Physiol.* **148**, 1583–1602 (2008).
50. Bueso, E. *et al. ARABIDOPSIS THALIANA HOMEOBOX25* uncovers a role for gibberellins in seed longevity. *Plant Physiol.* **164**, 999–1010 (2014).
51. He, X.C., Qin, Y.M., Xu, Y., Hu, C.Y. & Zhu, Y.X. Molecular cloning, expression profiling, and yeast complementation of 19 beta-tubulin cDNAs from developing cotton ovules. *J. Exp. Bot.* **59**, 2687–2695 (2008).
52. Tan, J. *et al.* A genetic and metabolic analysis revealed that cotton fiber cell development was retarded by flavonoid naringenin. *Plant Physiol.* **162**, 86–95 (2013).
53. Nakajima, K., Furutani, I., Tachimoto, H., Matsubara, H. & Hashimoto, T. *SPIRAL1* encodes a plant-specific microtubule-localized protein required for directional control of rapidly expanding *Arabidopsis* cells. *Plant Cell* **16**, 1178–1190 (2004).
54. Cheng, Q., Liu, H.T., Bombelli, P., Smith, A. & Slabas, A.R. Functional identification of AtFao3, a membrane bound long chain alcohol oxidase in *Arabidopsis thaliana*. *FEBS Lett.* **574**, 62–68 (2004).
55. Szumlanski, A.L. & Nielsen, E. The Rab GTPase RabA4d regulates pollen tube tip growth in *Arabidopsis thaliana*. *Plant Cell* **21**, 526–544 (2009).

ONLINE METHODS

Plant materials and resequencing. A total of 503 inbred cultivars of Upland cotton were collected as described in our previous study[11]. On the basis of the population structure analysis, a core germplasm set, including 282 accessions, was determined (**Supplementary Table 1**). Cotton plants were cultivated in a greenhouse in Wuhan, China. Young leaves were collected 4 weeks after planting and were immediately frozen in liquid nitrogen until use. Genomic DNA was extracted from leaves with the CTAB method[56]. For each accession, at least 5 μg DNA was used to construct a sequencing library with an Illumina TruSeq DNA Sample Prep Kit, according to the manufacturer's instructions. Paired-end sequencing (PE 150 bp) of each library was performed on an Illumina HiSeq 4000 system (**Supplementary Table 22**).

Mapping and variation calling. The allotetraploid cotton genome (*Gossypium hirsutum* L. acc. TM-1) and its annotation[9] were downloaded from CottonGen (URLs). Scaffolds with lengths less than 1,000 bp were excluded from further analysis. Paired-end resequencing reads were mapped to the TM-1 genome with BWA software with the default parameters. The PCR duplicates of sequencing reads for each accession were filtered with the Picard program, and uniquely mapping reads were retained in BAM format. Reads around indels from the BWA alignment were realigned with the IndelRealigner option in the Genome Analysis Toolkit (GATK)[57,58]. SNP and indel calling was performed with GATK and SAMtools software[59]. To obtain high-quality SNPs and indels, only variation detected by both software tools with a sequencing depth of at least 8 was retained for further analysis. SNPs with MAFs less than 1% were discarded, and indels with a maximum length of 10 bp were included. SNP annotation was carried out on the basis of that of the TM-1 genome, with snpEff software[60], and SNPs were categorized as being in intergenic regions, upstream (i.e., within a 2-kb region upstream of the transcription start site) and downstream (within a 2-kb region downstream of the transcription termination site) regions, in exons or introns. SNPs in coding sequences were further classified as synonymous SNPs or nonsynonymous SNPs. Indels in exons were classified according to whether they led to a frameshift effect.

Prediction of structural variation. Structural variations (SVs) were identified with three software tools: Breakdancer (version 1.3.6)[61], Delly (version 2)[62] and laSV (version 1.0.3)[63], which integrate most existing methods (read depth, read pair, split reads and *de novo* assembly of sequencing reads) for SV discovery. Breakdancer was run on all cotton accessions with the BWA alignment with the parameters (--q 20 --y 30). Delly, which uses paired-end mapping and a split-read method to discover SVs in the genome, was run separately for each sample with default settings. laSV, which first performs a reference-free *de novo* assembly of the sequencing reads and then compares the assembled contigs with the reference genome to identify SVs, was run separately for each sample, with parameters (--k 75 --l 150 --s 20). SVs (deletion, duplication, insertion and inversion) were retained if they were supported by at least two methods with a mapping depth of more than 10×. The breakpoint for each candidate SV was determined from the local assembly of sequencing reads with a de Bruijn algorithm.

Population-genetic analyses. To conduct the phylogenetic analysis, SNPs of all accessions were filtered with MAF = 0.05. These SNPs were used to construct a neighbor-joining tree with PHYLIP software[64] and were visualized with the online tool iTOL (URLs). PCA was performed with this SNP set with the smartpca program embedded in the EIGENSOFT package[65]. The population structure was analyzed with the Structure program, which infers the population structure by identifying different numbers of clusters (K)[66].

Linkage disequilibrium (LD) analysis. LD was calculated for each subpopulation with SNPs with MAF >0.05. To perform the LD calculation, plink software was applied with the parameters (--ld-window-r2 0 --ld-window 99999 --ld-window-kb 1000)[67]. LD decay was calculated on the basis of r^2 between two SNPs and averaged in 1-kb windows with a maximum distance of 1 Mb.

Identification of domestication sweeps. For domestication-sweep analysis, we combined cultivated cotton groups (ABI and Chinese groups) in a single group to exclude the potential effect of genetic drift. The genetic diversity in the wild group was compared with that in the cultivated group (π_w/π_c), because genomic regions in cultivated cottons should have a lower nucleotide diversity under domestication sweeps. Candidate domestication-sweep windows (100-kb windows sliding 20 kb) were identified with the top 5% of π_w/π_c values. We also used the XP–CLR method to scan for domestication-sweep regions (--w1 0.005 200 2000 1 --p0 0.95)[68]. To run XP–CLR, we assigned all SNPs to genetic positions on the basis of the published genetic map. Windows with the top 5% XP–CLR values were identified. Windows with a distance less than 50 kb were merged into a single nonoverlapping region. High-confidence domestication-sweep regions were identified by comparing XP–CLR analysis with the genetic diversity ratio (π_w/π_c).

To identify additional domestication effects, we calculated the population fixation statistics F_{ST} within 100-kb windows sliding 20 kb. Population-level F_{ST} was estimated as the average of all sliding windows. Windows with an empirical F_{ST} cutoff (top 5%) were regarded as highly differentiated regions. These regions were compared with the analysis of domestication sweeps. Genes with nonsynonymous SNPs in these regions were selected as being under selective pressure across groups.

Genome-wide association studies for fiber-quality-related traits. We used 2,020,834 high-quality SNPs (MAF >0.05) to perform GWAS for traits related to cotton fiber quality in 267 accessions. The traits included fiber length, fiber strength, micronaire value, fiber uniformity and fiber elongation rate. Association analyses were performed with TASSEL 5.0 with the compressed mixed linear model (P + G + Q + K)[69]. Kinship was derived from all these SNPs. The significant association threshold was set as $1/n$ (n, total SNP number). The significant association regions were manually verified from the aligned resequencing reads against the TM-1 genome with SAMtools[59].

Construction of ancestral chromosome state. To analyze selection signals at the subgenome level, we constructed the ancestral state for each of the 13 chromosomes in putative diploid ancestors. Homoeologous synteny blocks were identified in the 13 chromosome pairs between the At and the Dt subgenomes with MCScanX with default settings[70]. Syntenic gene pairs were identified in these syntenic blocks containing more than five aligned genes. A reciprocal blastp search was run with gene sequences from the At and Dt subgenomes. Gene pairs, which were identified in synteny blocks and also supported by blastp best hits between homologous chromosomes were retained as homoeologous genes. Genomic sequences consisting of gene regions and their flanking 2-kb sequences were ordered on the basis of the Dt subgenome and concatenated to construct the ancestral state.

RNA-seq and data analysis. Cotton leaves were sampled for gene expression analysis at the same developmental stage as for DNA resequencing. Total RNA was isolated as previously described[71]. A total of 2 μg RNA was used for library construction with an Illumina TruSeq RNA Kit, according to the manufacturer's instructions. RNA sequencing was performed on an Illumina HiSeq 3000 system (**Supplementary Table 22**). The clean reads were mapped to the TM-1 genome with TopHat (version 2.0.13)[72]. The expression level of each gene was determined with cufflinks (version 2.2.1) with a multiread and fragment bias correction method[73].

Bisulfite-treated DNA-sequencing data analysis. We downloaded bisulfite-treated DNA-sequencing data for leaf and fiber of TM-1 from the National Center for Biotechnology Information (NCBI) Sequence Read Archive collection (SRX710548–SRX710553). Trimmomatic software was applied to remove sequencing adapters and filter low-quality reads[74]. The clean reads for the two samples were mapped to the TM-1 genome with Bismark software (version 0.13.0; --N 1 --L 30)[75]. The multiple mapping and PCR-duplication reads were filtered to obtain a unique mapping BAM file. The Bismark methylation extractor program was run to extract potentially methylated cytosines. In this step, cytosines in CG, CHG and CHH contexts covered by at least three sequencing reads were retained and subjected to a two-sided binomial test (P-value cutoff of 1×10^{-5}).

DNase I digestion of chromatin. DNase I digestion of chromatin was conducted according to Zhang *et al.*[76] with some modifications. Briefly,

doi:10.1038/ng.3807

chromatin extraction was performed as described in our previous study[77]. For each sample, 100 g 10-DPA fiber and 1.5 g young leaves from TM-1 at the seedling stage were used for chromatin extraction. Extracted nuclei were washed once with 1× DNase I buffer and subjected to digestion with DNase I (Roche; lot no. 11781700). Nuclei were resuspended in 500 µL 1× DNase I buffer. A 20-µL aliquot was retained as an undigested control. Remaining nuclei were treated with 100 U DNase I and were incubated at 37 °C for 10 min. Immediately thereafter, both control and DNase I–digested nuclei of each sample were subjected to histone removal, DNA purification, RNase A treatment and fragment isolation. For each sample, DNase I digestion of chromatin was performed with at least two independent experiments.

DNase–seq and DHS identification. Purified DNA fragments of between 100 bp and 200 bp after DNase I digestion were isolated with a Pippin HT instrument (Sage Science). A total of 10 ng of the isolated fragments was used for library construction with an Illumina TruSeq Sample Prep Kit. Libraries were sequenced with an Illumina HiSeq 2000 system (**Supplementary Table 22**). After clipping of adapters and trimming of low-quality reads, clean reads were mapped to the TM-1 genome with Bowtie2 (version 2.2.4)[78]. The unique mapping data were processed to identify DHSs with the F-seq program with a 300-bp bandwidth[79]. MACS (version 1.4.2)[80], another peak-calling algorithm, was also run to identify DHSs, and randomly fragmented DNA-sequencing data were used as a control (*P*-value cutoff of 1×10^{-5}). Only peaks detected by both program tools were taken as candidate DHSs (**Supplementary Table 23**). Genome coverage of DNase–seq data in cotton was calculated with the coverageBed program embedded in the Bedtools package[81]. Chromosomal distribution of DHSs was analyzed in 1-Mb windows sliding 200 Kb.

Motif discovery. The promoter DHSs were screened for transcription-factor-binding motifs with the findMotifsGenome.pl program in HOMER software (URLs)[82], with the parameters '--size given --len 8,10,12 --chopify --mset plants'. In HOMER, motifs with cutoffs of $P < 0.01$ for known motifs and $P < 1 \times 10^{-12}$ for *de novo* motifs were retained. The 2-kb upstream sequences of genes were used for motif discovery with the Patch 1.0 program, which searches the TRANSFAC Public 6.0 database (URLs), with the following parameters: (i) the minimum length of sites was 8; (ii) the maximum number of mismatches was 1; (iii) the mismatch penalty was 100; and (iv) the lower score boundary was 87.5.

Chromatin immunoprecipitation (ChIP). Approximately 2 g of cotton leaves was cross-linked by vacuum infiltration with 1% formaldehyde for 35 min. Chromatin was extracted and fragmented to 200–500 bp by sonication. ChIP was performed as previously described[77]. Antibodies against H3K4me1 (Abcam; ab8895), H3K4me3 (Abcam; ab8580), H3K9me2 (Abcam; ab1220) and H3K27me3 (ABclonal; A2363) were cross-linked with Dynabeads protein A (Life Technologies; lot no. 165116310) and added to the sonicated samples for immunoprecipitation. All the ChIP experiments were carried out with samples from different plants in two independent experiments. Antibody validation is available on the manufacturers' websites.

ChIP–seq and data analysis. For each sample, a total of 10 ng ChIP DNA and input control DNA were used for library construction with an Illumina TruSeq Sample Prep Kit, according to the manufacturer's instructions. ChIP libraries were sequenced on an Illumina HiSeq 3000 system (**Supplementary Table 22**). The clean sequencing reads were mapped to the TM-1 genome with Bowtie2 (version 2.2.4)[78]. After removal of PCR duplication and multiple mapping reads, the unique mapping data were used to call histone-modification peaks with MACS software (version 2.1.0)[80]. The "--broad" parameter was on for calling H3K4me1, H3K9me2 and H3K27me3 peaks, and was off for calling H3K4me3 peaks (*P*-value cutoff of 1×10^{-5}). The Input DNA sequencing data were used as a control.

Hi-C experiments and sequencing. Cotton leaves were cross-linked in 20 ml of fresh ice-cold nuclei-isolation buffer and 1 ml of ~36% formaldehyde solution under vacuum for 40 min at room temperature. This reaction was quenched by the addition of 1 mL of 2 M glycine under vacuum infiltration for an additional 5 min. The clean samples were ground to powder in liquid nitrogen. The chromatin extraction was similar to that used in the DNase I–digestion experiment. The procedures were similar to those described previously[83]. Briefly, chromatin was digested for 16 h with 200 U (4 µl) HindIII restriction enzyme (Takara) at 37 °C. DNA ends were labeled with biotin and incubated at 37 °C for 45 min, and the enzyme was inactivated with 20% SDS solution. DNA ligation was performed by the addition of T4 DNA ligase (Fermentas) and incubation at 4 °C for 1 h followed by 22 °C for 4 h. After ligation, proteinase K was added to reverse cross-linking during incubation at 65 °C overnight. DNA fragments were purified and dissolved in 86 µL of water. Unligated ends were then removed. Purified DNA was fragmented to a size of 300–500 bp, and DNA ends were then repaired. DNA fragments labeled by biotin were finally separated on Streptavidin C1 beads (Life Technologies). Libraries were constructed with an Illumina TruSeq DNA Sample Prep Kit according to the manufacturer's instructions. TA cloning was performed to examine the quality of the Hi-C library. Hi-C libraries were sequenced on an Illumina HiSeq 3000 system (**Supplementary Table 22**). Hi-C was carried out with two independent experiments.

Hi-C data analysis. Raw Hi-C data were processed to filter low-quality reads and trim adapters with Trimmomatic (version 0.32)[74]. Clean reads were mapped to the TM-1 genome with a two-step approach embedded in the HiC-Pro software (version 2.7.1; URLs)[84]. After low-mapping-quality reads, multiple mapping reads and singletons were discarded, the unique mapping reads were retained in a single file. Read pairs that did not map close to a restriction site, or were not within the expected fragment size after shearing, were first filtered. Subsequent filtering analyses were performed to discard read pairs from invalid ligation products, including dangling-end and self-ligation products, and from PCR artifacts. The remaining valid read pairs were divided into intrachromosomal pairs and interchromosomal pairs. Contact maps were constructed with chromosome bins of equal sizes for 5 kb, 10 kb, 20 kb, 100 kb, 200 kb and 500 kb. The raw contact maps were then normalized with a sparse-based implementation of the iterative correction method in HiC-Pro.

Chromatin interactions (20 kb–2 Mb) were identified with a method of statistical confidence estimation, Fit-Hi-C[85]. To run Fit-Hi-C, fragments of less than 2 kb were merged to exclude possible Hi-C bias. Results from the second pass after an initial fit were used for further analysis. Fragments overlapping with intergenic DHSs or promoters were extracted to construct a regulatory interactome. Chromatin interactions with a false discovery rate (FDR) of 0.05 were retained and then compared with the genomic localization of intergenic DHSs and promoters to map promoter-centered interactions. TAD-like and boundary-like regions were identified with the TopDom method at 50-kb resolution[86]. TopDom was processed with a window size of 5.

Statistical analysis. To measure differential expression of genes in RNA-seq data, we used the DESeq package in R with the negative binomial distribution. Only genes with FDR <0.05 were retained.

Data availability. All the data sets generated during the current study are available in the NCBI Sequence Read Archive (SRA) under accession number SRP080913. The accession numbers are summarized in **Supplementary Table 22**. All the genomic variants can be downloaded from http://cotton.cropdb.org/cotton/download/data.php/.

56. 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. Report.* **11**, 122–127 (1993).
57. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
58. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
59. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
60. Cingolani, P. *et al.* 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 (2012).

61. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6**, 677–681 (2009).
62. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
63. Zhuang, J. & Weng, Z. Local sequence assembly reveals a high-resolution profile of somatic structural variations in 97 cancer genomes. *Nucleic Acids Res.* **43**, 8146–8156 (2015).
64. Felsenstein, J. PHYLIP-phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166 (1989).
65. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
66. Falush, D., Stephens, M. & Pritchard, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
67. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
68. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
69. Bradbury, P.J. *et al.* TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
70. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
71. Liu, D., Zhang, X., Tu, L., Zhu, L. & Guo, X. Isolation by suppression-subtractive hybridization of genes preferentially expressed during early and late fiber development stages in cotton. *Mol. Biol. (Mosk.)* **40**, 825–834 (2006).
72. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
73. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
74. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
75. Krueger, F. & Andrews, S.R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
76. Zhang, W. & Jiang, J. Genome-wide mapping of DNase I hypersensitive sites in plants. *Methods Mol. Biol.* **1284**, 71–89 (2015).
77. Wang, M. *et al.* Multi-omics maps of cotton fibre reveal epigenetic basis for staged single-cell differentiation. *Nucleic Acids Res.* **44**, 4067–4079 (2016).
78. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
79. Boyle, A.P., Guinney, J., Crawford, G.E. & Furey, T.S. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**, 2537–2538 (2008).
80. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X.S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).
81. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
82. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
83. Xie, T. *et al. De novo* plant genome assembly based on chromatin interactions: a case study of *Arabidopsis thaliana*. *Mol. Plant* **8**, 489–492 (2015).
84. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
85. Ay, F., Bailey, T.L. & Noble, W.S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.* **24**, 999–1011 (2014).
86. Shin, H. *et al.* TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.* **44**, e70 (2016).

doi:10.1038/ng.3807