

## Durham Research Online

---

### Deposited in DRO:

07 March 2017

### Version of attached file:

Accepted Version

### Peer-review status of attached file:

Peer-reviewed

### Citation for published item:

Wang, M. and Tu, L. and Lin, M. and Lin, Z. and Wang, P. and Yang, Q. and Ye, Z. and Shen, C. and Zhou, X. and Zhang, L. and Li, J. and Nie, X. and Li, Z. and Guo, K. and Ma, Y. and Jin, S. and Zhu, L. and Yang, X. and Min, L. and Zhang, Q. and Lindsey, K. and Zhang, X. (2017) 'Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication.', *Nature genetics.*, 49 (4). pp. 579-587.

### Further information on publisher's website:

<https://doi.org/10.1038/ng.3807>

### Publisher's copyright statement:

### Additional information:

## Use policy

---

The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a [link](#) is made to the metadata record in DRO
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Please consult the [full DRO policy](#) for further details.

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

3

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

9

10 <sup>1</sup>National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural  
11 University, Wuhan 430070, Hubei, China.

12 <sup>2</sup>Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics,  
13 Huazhong Agricultural University, Wuhan 430070, Hubei, China

14 <sup>3</sup>Key Laboratory of Oasis Eco-agriculture of the Xinjiang Production and  
15 Construction Corps, College of Agronomy, Shihezi University, Shihezi, Xinjiang,  
16 China.

17 <sup>4</sup>College of Plant Science and Technology, Huazhong Agricultural University,  
18 Wuhan 430070, Hubei, China

19 <sup>5</sup>Department of Biosciences, Durham University, Durham DH1 3LE, United  
20 Kingdom.

21

22 Correspondence should be addressed to X.Z. ([xlzhang@mail.hzau.edu.cn](mailto:xlzhang@mail.hzau.edu.cn)) or K.L.  
23 ([keith.lindsey@durham.ac.uk](mailto:keith.lindsey@durham.ac.uk))

24

25 Tel: +86-27-87280510

26 Fax: +86-27-87280196

27 **Comparative population genomics offers an excellent opportunity for**  
28 **unravelling the genetic history of crop domestication. Upland cotton (*Gossypium***  
29 ***hirsutum*) has long been an important economic crop, but a genome-wide and**  
30 **evolutionary understanding of the effects of human selection is largely**  
31 **unresolved. Here, we describe an integrated variation map for 352 wild and**  
32 **domesticated cotton accessions. This has allowed us to scan 93 domestication**  
33 **sweeps and identify 19 candidate loci for fiber quality-related traits by a**  
34 **genome-wide association study. We provide evidence to show asymmetric**  
35 **subgenome domestication for directional selection of long white fibers. Global**  
36 **analyses of DNase I-hypersensitive sites and 3-dimensional genome architecture,**  
37 **linking functional variants to gene transcription, reveal the effects of**  
38 **domestication on *cis*-regulatory divergence. This study provides new insights into**  
39 **the evolution of gene organization, regulation and adaptation in a major crop,**  
40 **and represents a rich resource for genome-based cotton improvement.**

41

42 Early human domestication of wild plants represented the first step in the  
43 development of modern crop varieties, and migration and differential directional  
44 selection over millennia has contributed to the adaptation of species in different  
45 environments for improved yield and quality traits<sup>1</sup>. In the current genomic era,  
46 high-throughput ‘omics’ technologies provide significant opportunities for a detailed  
47 analysis of genetic change through domestication and for new, targeted and precise  
48 genome-based crop breeding strategies<sup>2,3</sup>.

49 Cotton is one of the most important economic crops in the world, both as a  
50 source of natural and renewable fiber for textiles, and as a source of seed oil and  
51 protein<sup>4</sup>. Allotetraploid Upland cotton is formed from an inter-genomic hybridization  
52 event approximately 1–2 million years ago<sup>5</sup>. Originally native to the Yucatan  
53 peninsula in Mesoamerica, it was first domesticated at least 4,000 to 5,000 years ago,  
54 with subsequent directional selection<sup>6</sup>. Modern varieties of cultivated cotton produce  
55 spinnable fine white fiber, which is preferable to the sparser, coarse brown fiber of

56 wild cotton. Previous molecular studies have shown that domestication has  
57 dramatically rewired the transcriptome during fiber development<sup>7,8</sup>. What remains  
58 largely unknown, however, is the effect of human selection on the organization of the  
59 cotton genome and its gene regulatory landscape. Using as a comparator the recently  
60 published genome sequence of Texas Marker-1 (TM-1)<sup>9,10</sup>, we can address this  
61 question through a comprehensive population genome analysis of multiple wild and  
62 cultivated cotton genotypes.

63

## 64 **RESULTS**

### 65 **A genome variation map for cotton**

66 To construct an integrated variation map of Upland cotton, we collected a total of 352  
67 diverse accessions for genomic sequence analysis<sup>11</sup>. These included 31 wild  
68 accessions and 321 cultivated accessions from around the world (**Fig. 1a** and  
69 **Supplementary Table 1**). A total of 6.1 Tb of sequence data were integrated, with an  
70 average depth of 6.9× (**Supplementary Table 1**). These data were mapped against the  
71 TM-1 genome<sup>9</sup> to identify genomic variants. We detected a total of 7,497,568 SNPs,  
72 351,013 small indels (shorter than 10 bp) and 93,786 structural variants (SVs) (**Table**  
73 **1, Supplementary Fig. 1** and **Supplementary Tables 2-4**). The accuracy of SNPs  
74 was estimated to be 98.2%, determined by Sanger sequencing of 300 randomly  
75 selected SNPs in 3 individual accessions. In addition, we selected 50 representative  
76 accessions (10 wild and 40 cultivated cottons) from the 352 accessions for RNA  
77 sequencing (**Supplementary Table 5**), and generated 78,728 SNPs, of which more  
78 than 93.6% overlapped with SNPs from re-sequencing data. This integrated variation  
79 data set represents a new resource for cotton genetics and breeding.

80

### 81 **Cotton population properties and linkage disequilibrium**

82 We explored the phylogenetic relationship between the 352 cotton accessions using a  
83 whole-genome SNP analysis. These cottons can be divided into 3 groups (**Fig. 1b** and  
84 **Supplementary Fig. 2**), as supported by a principal component analysis (PCA; **Fig.**

85 **1c).** Wild cotton accessions cluster together (Group-I; the Wild group) except for a  
86 few accessions which cluster into a second group (Group-II; the ABI group), which  
87 mainly comprises cottons from America, Brazil and India. The third group (Group-III;  
88 the Chinese group) mostly consists of cotton cultivars in China, which were collected  
89 from the major Chinese cotton cultivation regions: the Northwestern Inland Region  
90 (NIR), the Northern Specific Early Maturation Region (NSEMR), the Yellow River  
91 Region (YRR) and the Yangtze River Region (YtRR)<sup>12</sup>. This group could be further  
92 classified into two subclades (Group-III-1 and Group-III-2; **Fig. 1b**), which exhibit  
93 different geographic distribution patterns. The subclade Group-III-1 is represented by  
94 cotton accessions from northern China (NIR and NSEMR), while Group-III-2  
95 includes the majority of accessions from southern China (YtRR). We observed that a  
96 few cotton accessions, which were collected from North America, clustered into  
97 Group-III, which might be due to the introduction of Upland cotton to China from  
98 America during the first thirty years of the 20<sup>th</sup> century<sup>13</sup>.

99 Crop species may experience population bottlenecks during domestication<sup>14</sup>. To  
100 examine this possibility in cotton, genetic diversity for each group was measured by  
101 calculating  $\pi$  values. We found that genetic diversity decreased from the Wild cotton  
102 group ( $\pi = 1.32 \times 10^{-3}$ ; the A-subgenome (At, the lower case t denotes tetraploid),  
103  $1.36 \times 10^{-3}$ ; the D-subgenome (Dt),  $1.25 \times 10^{-3}$ ) to the ABI group ( $\pi = 0.88 \times 10^{-3}$ ; At,  
104  $0.96 \times 10^{-3}$ ; Dt,  $0.66 \times 10^{-3}$ ) and to the Chinese group ( $\pi = 0.67 \times 10^{-3}$ ; At,  $0.72 \times 10^{-3}$ ;  
105 Dt,  $0.56 \times 10^{-3}$ ) (**Fig. 1d** and **Supplementary Fig. 3**). This shows that a large amount  
106 of genetic variation in both subgenomes has been lost during cotton domestication,  
107 especially for the Dt. Compared with other major crops, cotton possesses narrow  
108 genetic diversity even within wild cotton accessions (**Supplementary Table 6**). To  
109 investigate population divergence, we calculated the population fixation statistics ( $F_{ST}$ )  
110 among groups (**Fig. 1d**). This reveals large population divergence between the  
111 Chinese group and the Wild group. Population divergence between the Chinese group  
112 and the ABI group was observed, suggesting that Upland cottons in China have  
113 undergone population divergence after their introduction.

114 Linkage disequilibrium (LD; indicated by  $r^2$ ) was found to drop with physical  
115 distance between SNPs in all cotton groups (**Fig. 1e**). The LD extent for each group  
116 was measured as the chromosomal distance when LD dropped to half of its maximum

117 value. Consistent with other crops, the extent of LD in cotton is lower in the Wild  
118 group (84 kb;  $r^2 = 0.16$ ) than in the cultivated groups. The LD decay occurs at 162 kb  
119 ( $r^2 = 0.22$ ) in the ABI group and increases to 296 kb ( $r^2 = 0.25$ ) in the Chinese group.  
120 The observed LD extent in cultivated cotton groups is higher than is found in  
121 cultivated maize (30 kb)<sup>15</sup>, cultivated rice (123 kb in *Oryza indica*)<sup>16</sup> or cultivated  
122 soybean (133 kb)<sup>17</sup>, but lower than that of cultivated tomato (865.7 kb)<sup>18</sup>. For each  
123 group, LD decay distance in the At was found to be higher than that in the Dt  
124 (**Supplementary Fig. 4a,b**). For example, the LD extent of the Wild group was  
125 estimated to be 92 kb ( $r^2 = 0.16$ ) in the At and 64 kb ( $r^2 = 0.15$ ) in the Dt.

126

### 127 **Selection signals during cotton domestication**

128 Millennia of domestication has brought many morphological transformations to cotton,  
129 including an annualized growth cycle, photoperiod insensitivity, loss of seed  
130 dormancy, and superior spinnable white fiber<sup>7,8</sup>. To identify potential selective signals  
131 underlying these changes, we scanned genomic regions showing notable reductions in  
132 nucleotide diversity, by comparing cultivated accessions in the ABI and the Chinese  
133 groups with the Wild group. In total, we identified 93 putative domestication sweeps  
134 supported by at least one likelihood method (XP-CLR) and  $\pi_w/\pi_c$ , occupying 178 Mb  
135 of the genome (**Fig. 2a,e**). These regions harbored approximately 1,868 genes under  
136 selection, including 580 in the At and 1,288 in the Dt (**Supplementary Table 7**),  
137 suggesting that the Dt might be subject to stronger selection than the At.

138 To reveal the genetic basis of cotton domestication, we overlapped selection  
139 sweeps with the location of known QTL hotspots (containing at least four QTL for the  
140 same trait within a 20 cM region)<sup>19</sup>. We found that 25 QTL hotspots overlapped with  
141 selection sweeps, and these QTL hotspots were associated with some major  
142 agronomic traits, including leaf hair and morphology, petal spot, cotton boll number  
143 and weight, resistance to *Verticillium wilt* and fiber quality (**Fig. 2a,e** and  
144 **Supplementary Table 8**). Of these QTL hotspots, 17 of them were associated with  
145 fiber quality-related traits, including fiber length (FL), fiber strength (FS), micronaire  
146 value (MV), fiber elongation rate (FE) and fiber uniformity (FU). We investigated  
147 nucleotide diversity of genes residing in the 25 QTL hotspots to identify putative loci

148 with selection signals underlying these domestication-related traits. This led us to  
149 identify 400 genes exhibiting low nucleotide diversity in cultivated cottons when  
150 compared with wild cottons ( $\pi_w/\pi_c > 4.8$ ; **Supplementary Table 9**). Strikingly, 19 of  
151 25 QTL hotspots with 327 genes were located in the Dt.

152 Fiber quality improvement has been one of the most important breeding goals  
153 during cotton domestication. To further identify candidate genes for fiber  
154 quality-related traits, we performed a genome-wide association study (GWAS) using  
155 267 cotton accessions and phenotypic data collected during 2012 and 2013.  
156 Environmental effects were accounted for as described in our previous study<sup>11</sup>. We  
157 selected 2,020,834 high-quality SNPs with minor allele frequency (MAF > 0.05) from  
158 the core set. This high-density SNP map was found to be superior to previous  
159 SSR-maps for GWAS<sup>11</sup>. A total of 19 association signals for fiber quality-related  
160 traits, including 8 in the At and 11 in the Dt, were identified with  $P < 4.9 \times 10^{-7}$  using  
161 a compressed mixed linear model (MLM) (**Fig. 2b-d, f-i** and **Supplementary Table**  
162 **10**). Among these associations, 16 signals were previously uncharacterized. Most  
163 candidate genes in the LD regions of GWAS signals were found to be highly  
164 expressed during cotton fiber development (**Supplementary Table 11**). Three GWAS  
165 signals were identified as being under selection during domestication. Specifically, a  
166 GWAS signal associated with fiber strength was identified on chromosome A12 (**Fig.**  
167 **2d**), where a myb domain-encoding gene and an actin depolymerizing factor gene  
168 were found to reside. A GWAS signal associated with micronaire value was identified  
169 on chromosome D03 (**Fig. 2f**). This association was located near a cinnamyl alcohol  
170 dehydrogenase gene, which is a candidate for a role in the lignin pathway affecting  
171 fiber micronaire value<sup>20</sup>. We also identified a GWAS signal associated with fiber  
172 elongation rate on chromosome D04 (**Fig. 2g**), where a gibberellin response gene is  
173 located.

174

### 175 **Asymmetric subgenome domestication for long white fiber**

176 Most fiber characteristics in wild Upland cotton were probably inherited directly from  
177 its wild A-genome diploid ancestor post-allopolyploidization<sup>30</sup>, while fiber color is  
178 similar to that of its D-genome diploid ancestor. The development of the long white

179 fiber trait in cultivated Upland cotton is the result of millennia of strong directional  
180 selection from its wild counterpart. The observed change of fiber characteristics in  
181 cultivated Upland cottons is associated with changes in the expression patterns of  
182 fiber-related genes<sup>7,8,31</sup>. However, the genetic basis of this developmental change  
183 remains largely unknown. To understand the relative contributions of the co-existing  
184 At and Dt genomes during domestication, we constructed ancestral  
185 pseudochromosomes to address this question at the subgenome level. We identified  
186 15,456 homoeologous gene pairs, and used them to reconstruct an ancestral karyotype  
187 for each of the 13 chromosomes in cotton diploids, similar to a recent study in  
188 *Brassica*<sup>32</sup>. By comparing overlaps with domestication signals, we identified 620  
189 homoeologous pairs that have been subject to domestication selection in the At or Dt  
190 (192 in the At and 428 in the Dt), and only 34 homoeologous pairs with selection  
191 signals in both subgenomes (**Supplementary Fig. 6**). These results suggest that the  
192 co-existing subgenomes have been under asymmetric domestication selection (**Fig.**  
193 **3a**).

194 Domestication selection increased fiber length probably by effects on prolonging  
195 the elongation period of fiber development (**Fig. 3b**)<sup>30</sup>. We identified a formin  
196 homology interacting protein-coding gene (*FIP1*), which is involved in actin  
197 cytoskeleton organisation<sup>21,33</sup>, with a selection signal in the At but not in its Dt  
198 homoeolog (**Supplementary Fig. 6** and **Supplementary Table 12**). An altered  
199 regulation of the At *FIP1* in cultivated Upland cotton is predicted to be relevant to  
200 fiber elongation. Analysis of genes subjected to domestication selection in the Dt has  
201 led us to identify 17 genes involved in stress response pathways, such as reactive  
202 oxygen species (ROS) signaling (**Supplementary Fig. 6** and **Supplementary Table**  
203 **12**). High expression levels of these genes in wild cotton fibers may cause oxidative  
204 damage to developing fibers (**Supplementary Table 12**). Unexpectedly, we identified  
205 5 homoeologous gene pairs, involved in synthesis and deposition of secondary wall  
206 cellulose, with selection signals only in the Dt (**Supplementary Table 12**). These  
207 genes, such as *TRICHOME BIREFRINGENCE-LIKE 43 (TBL43)* and *COBRA-LIKE*  
208 *4 (COBL4)*<sup>34,35</sup>, were also highly expressed in wild cotton fibers at 20 days post  
209 anthesis (DPA). This is consistent with the view that high concentrations of ROS in  
210 wild cotton fiber development terminates fiber elongation, associated with the



211 developmental transition to secondary cell wall synthesis (**Fig. 3b**). This possibility is  
212 supported by our genetic suppression of cytosolic *ASCORBATE PEROXIDASES*  
213 (*cAPXs*), in which an increased content of hydrogen peroxide leads to the early  
214 initiation of secondary cell wall synthesis in fast elongating fiber and gives rise to  
215 short fibers<sup>36</sup>. Therefore genetic evidence suggests that an asymmetric domestication  
216 selection between the At and the Dt subgenomes, which might modulate ROS levels,  
217 is associated with the development of the long fiber trait in cultivated cotton (**Fig. 3b**).

218 Domestication has led to the transformation of cotton fiber from brown to white.  
219 To understand this phenomenon, we examined two homoeologous gene pairs only  
220 subjected to domestication selection in the Dt, *4-COUMARATE:COA LIGASE (4CL)*  
221 and *CHALCONE SYNTHASE (CHS)*, which encode enzymes involved in the  
222 phenylpropanoid metabolic pathway (**Fig. 3c** and **Supplementary Fig. 6**)<sup>37</sup>. For the  
223 *4CL* gene, we identified two nonsynonymous SNPs in the coding sequence and two  
224 SNPs residing in a Dof transcription factor binding site of the promoter (-369 bp to  
225 -378 bp; **Fig. 3c**). These SNPs display reductions in nucleotide diversity that occurred  
226 during domestication (**Fig. 3c**). Interestingly, we found that the two SNPs in the  
227 Dof-binding motif led to sequence variation departing from the canonical motif (**Fig.**  
228 **3d**), which might affect transcription activity of *4CL*, which is experimentally  
229 supported by a significantly low expression level at 10 DPA in cultivated cottons (**Fig.**  
230 **3e**). The enzyme CHS acts downstream of *4CL* in this pathway, catalyzing the first  
231 step of flavonoid synthesis, and its gene *CHS* has also been down-regulated during  
232 domestication (**Supplementary Table 12**). Given the recognized functional role of  
233 flavonoids in brown fiber pigmentation<sup>37,38</sup>, selection signals at the *4CL* and *CHS* loci  
234 in the Dt may have driven the white fiber trait characteristic of domestication.

235

### 236 **Effects of domestication on *cis*-regulatory elements in promoters**

237 Human selection of desirable agronomic traits not only affects the organization of  
238 functional genes, but may also reshape the gene regulatory landscape. In support of  
239 this idea, we found that many more variants were identified in intergenic compared  
240 with genic regions (**Table 1**). Specifically, intergenic non-coding variants can affect  
241 the activity of *cis*-regulatory elements (CREs)<sup>39-41</sup>, and can contribute to differential

242 gene expression patterns between populations (**Supplementary Fig. 7**). To investigate  
243 this in cotton, we performed a global analysis of the effects of domestication on CREs  
244 in promoters.

245 We identified CREs in cotton with data from chromatin digestion using DNase I  
246 followed by sequencing (DNase-seq): active CREs can be detected because of their  
247 increased nuclease sensitivity, reflecting an open chromatin conformation  
248 (**Supplementary Fig. 8**)<sup>42</sup>. We identified a total of 188,360 DNase I-hypersensitive  
249 sites (DHSs) in cotton leaves and fibers, of which ca. 47% are common to both tissues  
250 (**Fig. 4a**). DHSs were preferentially identified in chromosomal arms and  
251 approximately half were detected in promoter and intergenic regions (**Fig. 4b** and  
252 **Supplementary Fig. 9**). We found DHSs are hypo-methylated, consistent with  
253 previous studies<sup>42</sup> (**Fig. 4c**). DHSs in promoter regions are commonly marked by high  
254 levels of active H3K4me3 and inactive H3K27me3, with a depletion of active  
255 H3K4me1 and inactive H3K9me2 (**Fig. 4d**). Intergenic DHSs were also found to  
256 exhibit an enrichment of H3K4me3 and H3K27me3, but depletion of H3K9me2 and  
257 no enrichment of H3K4me1 (**Fig. 4e**). As predicted, the patterns of chromatin  
258 modification marks in cotton are different between genic and TE regions  
259 (**Supplementary Fig. 10**). In addition, genes with promoter DHSs are generally  
260 expressed at a higher level in both tissues than those without promoter DHSs (**Fig. 4f**),  
261 and tissue-specific promoter DHSs corresponded to higher levels of gene expression  
262 (**Fig. 4g**). These results reveal a close relationship between promoter DHS occurrence  
263 and relatively high transcriptional activity.

264 Genetic variants in promoter DHSs were examined in our resequencing  
265 population. We detected 90,737 SNPs in the 25,580 promoter DHSs (**Table 1**).  
266 Selection signals were detected for these promoter DHSs following domestication. A  
267 total of 738 DHSs (358 in the At and 380 in the Dt) are under domestication selection  
268 ( $\pi_w/\pi_c > 4.8$ ), of which 461 exhibit population divergence between cultivated and wild  
269 cotton accessions ( $F_{ST} > 0.24$ ) (**Fig. 4h**). Of these DHSs with selection signals, we  
270 found 281 DHS-related genes were differentially expressed. To investigate how  
271 variants in promoter DHSs might influence the expression of genes, we looked for  
272 associations between variants and transcription binding motifs. We discovered 178  
273 motifs for 95 transcription factors in DHSs (**Supplementary Table 13**). We found

274 that some well-known transcription binding motifs were under purifying selection in  
275 the cultivated groups, and some were under positive selection (**Fig. 4i** and  
276 **Supplementary Table 14**). For example, the TRAB1 binding motif, which relates to  
277 abscisic acid (ABA)-regulated transcription<sup>43</sup>, was identified with a domestication  
278 sweep signal. The GL3 binding motif, which participates in cotton fiber initiation<sup>44</sup>,  
279 was also under domestication selection. The PIF4 binding motif, which is important  
280 for high temperature-mediated adaptation in plants<sup>45</sup>, was identified as a positively  
281 selected motif. This reveals the effects of selection on *cis*-regulatory elements in  
282 promoter regions, which may be associated with the transcriptional regulation of  
283 genes contributing to desirable traits or adaptation.

284

### 285 **Genome variation underlies distant regulatory divergence**

286 Multiple genes can be considered to be organized into ‘transcriptional factories’ and  
287 transcribed in a high-order conformation<sup>46</sup>. A range of high-throughput methods, such  
288 as high-throughput chromosome conformation capture (Hi-C) and chromatin  
289 interaction analysis by paired-end tag sequencing (ChIA-PET), have been developed  
290 to understand 3D genome architecture in the eukaryotic nucleus<sup>47,48</sup>. Several studies  
291 have shown that long-range chromatin interaction is an important mechanism for the  
292 regulation and coordination of gene transcription<sup>49,50</sup>. Once we established a DHS  
293 landscape in cotton, the next aim was to characterize the effects of domestication on  
294 divergences in regulatory elements that are physically remote from, but functionally  
295 linked to, genes.

296 Hi-C analysis was carried out using the TM-1 accession to characterize global  
297 chromatin interactions. We generated 1.1 billion Hi-C paired-end reads, of which ca.  
298 322 million were valid interaction reads (**Supplementary Table 15**). To exclude  
299 possible Hi-C bias, *HindIII* fragments of less than 2 kb were merged to obtain  
300 305,682 chromosomal anchor regions (**Fig. 5a**). On the basis of a high-quality  
301 genome assembly of TM-1 (**Supplementary Fig. 11**), we used the Hi-C data to  
302 characterize the cotton chromatin interactome (**Supplementary Fig. 12**) and  
303 uncovered 737,377 mid-range intra-chromosomal interactions (20 kb–2 Mb). The  
304 number of interactions drops rapidly with an increase in distance between sequences

305 (Fig. 5b), but many topologically associated domain-like (TAD-like) regions were  
306 identified (Fig. 5c, Supplementary Fig. 13 and Supplementary Table 16). We  
307 found that chromatin interactions are significantly enriched at promoters, distal DHSs  
308 such as enhancers and at regions marked by the active chromatin mark H3K4me3, but  
309 are less frequent at regions marked by H3K9me2 (Fig. 5d).

310 Interactions involving promoters and distal DHSs, such as enhancers, were  
311 analyzed to construct a long-distance transcriptional regulation map. We obtained  
312 121,522 interactions, including 52,496 putative extragenic interactions (promoter to  
313 enhancer), 44,808 putative intergenic interactions between different genes  
314 (promoter-promoter interactions) and 24,218 putative enhancer-enhancer interactions  
315 (Fig. 5e and Supplementary Table 17). We found that only ca. 38% of putative  
316 enhancers and 25% of promoters are involved in a single interaction (Fig. 5f),  
317 indicating that transcription of most genes appears to be regulated by multiple  
318 long-range chromatin interactions. Interestingly, genes with relatively high levels of  
319 chromatin interaction exhibit higher expression levels than genes without interaction  
320 (Fig. 5g).

321 We next examined enhancer divergence. We identified a total of 99,709 SNPs in  
322 the 21,409 putative enhancers (Table 1). We found that enhancers exhibit a higher  
323 frequency of sequence variation than promoters or exons, and exhibit a lower  
324 frequency than introns (Fig. 5h). This suggests that enhancers have evolved rapidly.  
325 We then looked at evidence for genomic selection of enhancers during cotton  
326 domestication. We identified 2,011 enhancers (496 in the At and 1,515 in the Dt) with  
327 selection signals associated with 1,651 gene promoters (Supplementary Table 18).  
328 One example shows that an enhancer located 120 kb upstream of *TUBULIN ALPHA-3*  
329 (*TUA3*) has undergone strong selection, consistent with the observed differentially  
330 high expression of *TUA3* in cultivated TM-1 compared with the wild YUC accession  
331 (Fig. 5i). DNase I digestion of chromatin on a representative wild cotton accession  
332 revealed that more than 94% of enhancers are shared in wild and domesticated cottons  
333 (Fig. 5j), suggesting that domestication has had a limited effect on qualitative changes  
334 to enhancers.

335

## 336 **DISCUSSION**

337 Genome re-sequencing of 352 accessions of Upland cotton has provided new insights  
338 into the genetic history of this important crop. By constructing a comprehensive  
339 variation map, we have determined genomic diversity and divergence for cotton.  
340 Interestingly, we found no obvious population divergence between geographic groups  
341 in China, probably because of frequent migration of accessions for improvement  
342 breeding within a short period after introduction. This is different from observations  
343 for cultivated rice and soybean, which were initially domesticated from wild forms in  
344 China millennia ago<sup>17,51</sup>. Comparison of the wild and cultivated cottons has allowed  
345 the identification of domestication sweeps. In this study, we primarily characterized  
346 some key molecular signatures of selection responsible for spinnable fine white fiber,  
347 of which some candidates were further identified by a GWAS analysis. We believe  
348 that these selection sweeps could enable future characterization of genes for other  
349 domestication-related agronomic traits. The variation map and selective sweeps  
350 constitute a valuable resource for future cotton improvement.

351 We revealed the effects of domestication on *cis*-regulatory divergence through an  
352 integrated approach. We first present a global analysis of DHSs using DNase-seq,  
353 which was demonstrated to be a highly efficient approach to map CREs in human<sup>52</sup>.  
354 We provide evidence to suggest that directional selection through domestication has  
355 led to the divergence of CREs at promoters of at least some regulatory genes relevant  
356 to agronomic traits in cotton. Compared with promoters, distant CREs such as  
357 enhancers are less conserved among species but are also important for transcriptional  
358 regulation through long-range chromatin interactions<sup>53</sup>. With the DHS map, we  
359 provide a picture of 3D genome architecture, to link distant regulatory variants in  
360 enhancers to gene transcription. In contrast with isolated analyses of DHSs and 3D  
361 genome studies in *Arabidopsis*<sup>54,55</sup>, this represents the first comprehensive functional  
362 interpretation of non-coding genetic variants in plants. Our approach to the  
363 characterization of functional variants represents a useful reference for other crops.  
364 These data will facilitate future functional genomics studies for cotton and inform  
365 breeding strategies.

366

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

371

## 372 **METHODS**

373 Methods and any associated references are available in the online version of the  
374 paper.

375 **Accession codes.** The sequence data have been deposited in the NCBI Sequence Read  
376 Archive (SRA) under the BioProject accession PRJNA336461. All the genomic  
377 variants can be downloaded from <http://cotton.cropdb.org/cotton/download/data.php>.

378 *Note: Any Supplementary Information and Source Data files are available in the one*  
379 *version of the paper.*

380

## 381 **ACKNOWLEDGMENTS**

382 We thank T. Zhang (Nanjing Agricultural University) for releasing re-sequencing data  
383 of wild cotton accessions. This work was supported by funding from the National  
384 Natural Science Foundation of China (31230056 and 31201251).

385

## 386 **AUTHOR CONTRIBUTIONS**

387 X.Z., L.T. and M.W. conceived and designed the project. P.W., M.L., Q.Y., Z.Y.,  
388 X.Z., M.W. and X.N. performed the experiments. M.W., P.W. and Q.Z. developed  
389 libraries and performed sequencing. M.W., C.S., J.L., L.Z., K.G., Y.M., Z.L., C.H.  
390 and D.Y. analyzed the data. Z.L., L.T., S.J., L.Z., X.Y. and L.M. collected materials  
391 and managed sequencing. M.W. wrote the manuscript draft, and K.L. and X.Z.  
392 revised it.

393

394 **COMPETING FINANCIAL INTERESTS**

395 The authors declare no competing financial interests.

396 **Figure legends**

397 **Figure 1** Geographic distribution and population diversity of Upland cotton  
398 accessions. **(a)** The geographic distribution of Upland cotton accessions. Each dot of a  
399 given color on the world map represents the geographic distribution of the  
400 corresponding cotton accession. **(b)** Neighbour-joining tree of all accessions  
401 constructed from whole-genome SNPs. The geographic distribution of each accession  
402 is represented by a tree branch with a color corresponding to that in **Fig. 1a**. The outer  
403 ring indicates groups emerging from the phylogenetic tree. **(c)** PCA plots of the first  
404 two components for all accessions. The dot color scheme is as indicated in **Fig. 1a**.  
405 ABI represents cottons from America, Brazil and India; NNR represents cottons from  
406 the Northwestern Inland Region and the Northern Specific Early Maturation Region;  
407 YRR represents cottons from the Yellow River Region; and YtRR represents cottons  
408 from the Yangtze River Region. **(d)** Nucleotide diversity ( $\pi$ ) and population  
409 divergence ( $F_{ST}$ ) across the three groups. Value on each circle represents measure of  
410 nucleotide diversity for this group, and value on each line indicates population  
411 divergence between the two groups. **(e)** Decay of linkage disequilibrium (LD) in each  
412 group.

413

414 **Figure 2** Genome-wide screening of domestication sweeps and GWAS on fiber  
415 quality-related traits. **(a)** Selection signals in the A-subgenome (At) and **(e)** selection  
416 signals in the D-subgenome (Dt). The horizontal grey dashed lines show the  
417 genome-wide threshold for domestication sweeps identified from the ratio of  
418 nucleotide diversity between wild and cultivated cotton accessions ( $\pi_w/\pi_c > 4.8$ ). The  
419 results using the XP-CLR analytical tool are indicated by the red lines. The 25 QTL  
420 hotspots that overlap with domestication sweeps are shown in each chromosome.  
421 Genes with known function for fiber development under domestication selection are  
422 shown in corresponding chromosomes. These genes include *FIP1*<sup>21</sup>, *14-3-3*<sup>22</sup>, *GSRI*<sup>23</sup>,  
423 and *HB31*<sup>24</sup> in the At, and *TUB6*<sup>25</sup>, *TUB8*<sup>25</sup>, *4CL*<sup>26</sup>, *CHS*<sup>26</sup>, *SPIL5*<sup>27</sup>, *FAO3*<sup>28</sup> and  
424 *RABA4A*<sup>29</sup> in the Dt. The expression levels of these genes are shown in  
425 **Supplementary Fig. 5**. **(b–d)** Significant GWAS associations on fiber length **(b,c)**  
426 and fiber strength **(d)** in the At. **(f–i)** Significant GWAS associations on micronaire



427 value (**f**), fiber elongation rate (**g**), fiber length (**h**) and fiber uniformity (**i**) in the Dt.  
428 The horizontal grey dashed lines in **b–d** and **f–i** show the significance threshold of  
429 GWAS (1/n; 6.3). The other significant associations are presented in **Supplementary**  
430 **Table 10**.

431

432 **Figure 3** Asymmetric selection signals between the A-subgenome (At) and the  
433 D-subgenome (Dt). **(a)** A model of asymmetric domestication between the At and the  
434 Dt. The number of colored dots shows change of genetic diversity after domestication  
435 in each subgenome. **(b)** Effects of the Dt-specific selection signals on prolonged fiber  
436 elongation in cultivated cottons. Upper track shows the morphological and  
437 developmental differences of fibers between wild and cultivated cottons. The heatmap  
438 shows fiber elongation rate in wild/cultivated cotton. Dashed box shows a prolonged  
439 elongation period in cultivated cotton with data from Applequist *et al.* (2001)<sup>30</sup>.  
440 Lower track shows a model of developing fiber. Genes with selection signals in the Dt  
441 are shown. Compared with wild cotton, these genes are down-regulated in cultivated  
442 cotton fiber development, which could regulate reactive oxygen species (ROS) levels  
443 associated with prolonged fiber elongation. Full descriptions of these genes are shown  
444 in **Supplementary Table 12**. **(c)** Selection signals in the *4-coumarate:CoA Ligase*  
445 (*4CL*) gene region. Upper track shows asymmetric selection signals in ancestral  
446 karyotype 3 in the At and the Dt, which was reconstructed using homoeologous gene  
447 pairs. Vertical dashed lines show some homoeologous gene pairs with selection  
448 signals. Lower track shows allele frequency of SNP variants in the *4CL* in  
449 wild/cultivated cotton group. Nonsynonymous SNPs in the first exon are indicated in  
450 red. SNPs in the Dof transcription factor binding site are indicated in sky blue. **(d)**  
451 Sequence logos of the Dof-binding site in wild and cultivated cotton groups compared  
452 with that in *Arabidopsis* (JASPAR model: MA0973.1). **(e)** Normalized expression  
453 levels of *4CL* at 10 days post anthesis (DPA) in wild and cultivated cottons shown by  
454 RNA-seq (two-side *t*-test, \*\**P*-value < 0.01). Error bars, s.d. of the normalized  
455 expression levels from different cotton accessions.

456

457 **Figure 4** Characterization of cotton DNase I-hypersensitive sites (DHSs) and  
458 detection of selected DHSs during domestication. (a) Venn diagram showing the  
459 number of DHSs identified in cotton leaves and fibers at 10 days post anthesis (DPA).  
460 (b) Genomic distribution of DHSs in genic and intergenic regions. (c) DNA  
461 methylation levels of DHSs in cotton leaves and fibers. (d) Enrichment/depletion of  
462 chromatin modification marks in promoter DHSs. The grey arrow shows the  
463 transcription orientation of genes. (e) Enrichment/depletion of chromatin modification  
464 marks in intergenic DHSs. For c–e, each DHS region was divided into 50 bins on  
465 average, and the flanking 2 kb regions were divided into 200 bins with an equal length.  
466 For d–e, the ChIP-seq tags were normalized by Input DNA sequencing data. (f)  
467 Comparisons of the expression levels between genes with promoter DHSs and those  
468 without promoter DHSs in leaf and fiber samples (Wilcoxon rank sum test,  
469 \*\*\* $P$ -value < 0.001). (g) Comparisons of the expression levels of tissue-specific  
470 promoter DHS marked genes with those of overlapping promoter DHS marked genes  
471 between leaf and fiber. For each group, the relative expression level was calculated by  
472 fold-change of leaf versus fiber. The pattern of expression fold-change for  
473 tissue-specific DHS marked genes was compared with that of overlapping promoter  
474 DHS marked genes (\*\*\* $P$ -value < 0.001). (h) Detection of selected promoter DHSs  
475 during cotton domestication. All promoter DHSs were sorted by  $F_{ST}$ . The x axis  
476 shows the order of DHSs in this study. The left y axis shows ratio of nucleotide  
477 diversity for promoter DHSs between wild and cultivated cotton accessions ( $\pi_w/\pi_c$ ).  
478 The right y axis shows population divergence ( $F_{ST}$ ) between wild and cultivated  
479 populations. Highly differentiated DHSs are indicated by the shaded background. (i)  
480 Nucleotide diversity of key transcription factor binding motifs that were identified  
481 from promoter DHSs in different cotton groups. For each motif, nucleotide diversity  
482 was scaled to the size of each respective circle. Motifs with decreased diversity during  
483 domestication are represented by the orange bar and increased diversity by the green  
484 bar. Abbreviations representing cottons from different cultivation regions in China  
485 were the same as those in **Fig. 1c**.

486

487 **Figure 5** Characterization of cotton chromatin interactome and identification of  
488 promoter-centered interactions. (a) Size distribution of raw *HindIII* fragments

489 (histogram) in the cotton genome, and anchors (red curve) used in this study. **(b)**  
490 Genomic distances between all interacting anchors. The histogram shows frequency  
491 distribution of distances between anchors, and the red curve shows the cumulative  
492 proportion of interactions. **(c)** Chromatin interaction in A13 and D02 chromosomes.  
493 The repressive modification marks (H3K27me3 and H3K9me2) are shown for each  
494 chromosome. Each heatmap shows a normalized contact matrix, with strong contacts  
495 in red and weak contacts in white. Examples of topologically associated domain-like  
496 (TAD-like) regions are shown below the heatmaps. **(d)** Percentages of anchors  
497 involving *cis*-regulatory elements (CREs) and peaks of chromatin modification marks.  
498 Actual enrichment ratios of CREs and ChIP peaks were compared with expected  
499 background values (Fisher exact test, \*\**P*-value < 0.01). **(e)** Percentage of  
500 promoter-centered interactions for each type: enhancer-promoter (E-P),  
501 promoter-promoter (P-P) and enhancer-enhancer (E-E). **(f)** Degree distribution of  
502 anchor and promoter (TSS). The x axis represents degree distribution and y axis  
503 represents the proportions of anchor and TSS in each degree. **(g)** Expression analyses  
504 of genes with chromatin interaction and genes without chromatin interaction  
505 (Wilcoxon rank sum test, \*\**P*-value < 0.01). **(h)** SNP frequencies in enhancer,  
506 promoter, exon and intron regions. SNP frequency in these elements was compared  
507 with that in randomly selected genome regions (500 iterations; \*\*\**P*-value < 0.001).  
508 **(i)** One example of an enhancer under domestication selection. The upper track shows  
509 chromatin interaction of anchors represented by pink lines. Domestication selection is  
510 indicated by ratios of nucleotide diversity ( $\pi_w/\pi_c$ ) in 20 kb windows sliding 5 kb. The  
511 lower five tracks show sequencing tags of DNase-seq, ChIP-seq (H3K4me3 and  
512 H3K27me3) and RNA-seq in TM-1 and YUC accessions, respectively. The enhancer  
513 and gene regions were shown by colored background and arrows. **(j)** Venn diagram  
514 showing the ratio of overlapped enhancers in TM-1 and YUC accessions.

515

516 **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 introducing 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
Non-frameshift indel	1,114	667	433	919
SVs that overlap with genes	12,511	11,876	10,963	11,193
SNPs in <i>cis</i> -regulatory elements				
Promoter DHSs	90,737	73,404	59,788	55,637
Enhancers	99,709	82,287	66,107	56,386

517

518

## 519 **ONLINE METHODS**

### 520 **Plant materials and re-sequencing**

521 A total of 503 inbred cultivars of Upland cotton were collected as described in our  
522 previous study<sup>11</sup>. Based on the population structure analysis, a core germplasm set,  
523 including 282 accessions was determined (**Supplementary Table 1**). Cotton plants  
524 were cultivated in the greenhouse in Wuhan, China. Young leaves were collected 4  
525 weeks after planting and immediately frozen in liquid nitrogen until use. Genomic  
526 DNA was extracted from leaves using the CTAB method<sup>56</sup>. For each accession, at  
527 least 5 µg DNA was used to construct a sequencing library using the Illumina TruSeq  
528 DNA Sample Prep Kit following the manufacturer's instructions. Paired-end  
529 sequencing (PE 150-bp reads) of each library was performed on the Illumina HiSeq X  
530 Ten system.

531

### 532 **Mapping and variation calling**

533 The allotetraploid cotton genome (*Gossypium hirsutum* L. acc. TM-1) and its  
534 annotation<sup>9</sup> were downloaded from the Internet (see URLs). Scaffolds with lengths  
535 less than 1000 bp were excluded from further analysis. Paired-end re-sequencing  
536 reads were mapped to the TM-1 genome using BWA software with the default  
537 parameters. The PCR duplicates of sequencing reads for each accession were filtered  
538 using the Picard program, and uniquely mapping reads were retained in the BAM  
539 format. Reads around indels from the BWA alignment were realigned using the  
540 IndelRealigner option in Genome Analysis Toolkit (GATK)<sup>57,58</sup>. SNP and indel  
541 calling was performed using GATK and SAMtools software<sup>59</sup>. To obtain high-quality  
542 SNPs and indels, only variation detected by both software tools with sequencing depth  
543 of at least 8 was retained for further analysis. SNPs with minor allele frequencies less  
544 than 1% were discarded, and indels with a maximum length of 10 bp were included.  
545 SNP annotation was carried out based on that of the TM-1 genome, using the snpEff  
546 software<sup>60</sup>, and SNPs were categorized as being in intergenic regions, upstream (i.e.  
547 within a 2 kb region upstream of the transcription start site) and downstream (within a

548 2 kb region downstream of the transcription termination site) regions, in exons or  
549 introns. SNPs in coding sequences were further classified as synonymous SNPs or  
550 nonsynonymous SNPs. Indels in exons were classified according to whether they lead  
551 to a frame-shift effect.

552

### 553 **Prediction of structural variation**

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

567

### 568 **Population-genetic analyses**

569 To conduct the phylogenetic analysis, SNPs of all accessions were filtered with minor  
570 allele frequency (MAF) 0.05. These SNPs were used to construct a neighbour-joining  
571 tree using PHYLIP software<sup>64</sup> and visualized using the online tool iTOL (see URLs).  
572 Principal component analysis (PCA) analysis was performed using this SNP set with  
573 the smartpca program embedded in the EIGENSOFT package<sup>65</sup>. Population structure  
574 was analyzed using the Structure program which infers the population structure by  
575 identifying different numbers of clusters (K)<sup>66</sup>.

576

577 **Linkage disequilibrium (LD) analysis**

578 LD was calculated for each sub-population using SNPs with minor allele frequency  
579 (MAF) greater than 0.05. To perform the LD calculation, plink software was applied  
580 with the parameters (-ld-window-r2 0 -ld-window 99999 -ld-window-kb 1000)<sup>67</sup>. LD  
581 decay was calculated based on  $r^2$  between two SNPs and averaged in 1 kb windows  
582 with a maximum distance of 1 Mb.

583

584 **Identification of domestication sweeps**

585 For domestication sweep analysis, we combined cultivated cotton groups (ABI and  
586 Chinese groups) into a single group to exclude the potential effect of genetic drift. The  
587 genetic diversity in the wild group was compared with that in the cultivated group  
588 ( $\pi_w/\pi_c$ ), because genomic regions in cultivated cottons should have a lower nucleotide  
589 diversity under domestication sweeps. Candidate domestication sweeps windows (100  
590 kb windows sliding 20 kb) were identified with the top 5% of  $\pi_w/\pi_c$  values. We also  
591 used the XP-CLR method to scan for domestication sweep regions (-w1 0.005 200  
592 2000 1 -p0 0.95)<sup>68</sup>. To run XP-CLR, all SNPs were assigned to genetic positions  
593 based on the published genetic map. Windows with the top 5% XP-CLR values were  
594 identified. Windows with distance less than 50 kb were merged into a single  
595 non-overlapping region. High-confidence domestication sweeps regions were  
596 identified by comparing XP-CLR analysis with genetic diversity ratio ( $\pi_w/\pi_c$ ).

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

603

604 **Genome-wide association studies for fiber quality-related traits**

605 We used 2,020,834 high-quality SNPs (MAF > 0.05) to perform GWAS on cotton  
606 fiber quality-related traits in 267 accessions. The traits include fiber length, fiber  
607 strength, micronaire value, fiber uniformity and fiber elongation rate. Association  
608 analyses were performed using TASSEL 5.0 with the compressed mixed linear model  
609 (P + G + Q + K)<sup>69</sup>. Kinship was derived from all these SNPs. The significant  
610 association threshold was set as 1/n (n, total SNP number). The significant association  
611 regions were manually checked from the aligned re-sequencing reads against the  
612 TM-1 genome using SAMtools<sup>59</sup>.

613

#### 614 **Construction of ancestral karyotypes**

615 To analyze selection signals at the subgenome level, we constructed the ancestral  
616 karyotype for each of the 13 chromosomes in putative diploid ancestors.  
617 Homoeologous synteny blocks were identified in the 13 chromosome pairs between  
618 the At and the Dt subgenomes using MCScanX with default settings<sup>70</sup>. Syntenic gene  
619 pairs were identified in these syntenic blocks containing more than five aligned genes.  
620 A reciprocal blastp was run using gene sequences from the At and Dt subgenomes.  
621 Gene pairs, which were identified in syntenic blocks and also supported by blastp best  
622 hits between homologous chromosomes were retained as homoeologous genes.  
623 Genomic sequences consisting of gene regions and their flanking 2 kb sequences were  
624 ordered based on the Dt subgenome and concatenated to construct ancestral  
625 karyotypes.

626

#### 627 **RNA-seq and data analysis**

628 Cotton leaves were sampled for gene expression analysis at the same developmental  
629 stage as for DNA re-sequencing. Total RNA was isolated as previously described<sup>71</sup>. A  
630 total of 2 µg RNA were used for library construction using the Illumina TruSeq RNA  
631 Kit (Illumina, San Diego, CA, USA) following the manufacturer's instructions. RNA  
632 sequencing was performed on the Illumina HiSeq 3000 system (paired-end 150-bp  
633 reads). The clean reads were mapped to the TM-1 genome using Tophat (version



634 2.0.13)<sup>72</sup>. The expression level of each gene was determined using Cufflinks (version  
635 2.2.1) with a multi-read and fragment bias correction method<sup>73</sup>.

636

### 637 **Bisulfite-treated DNA sequencing data analysis**

638 We downloaded bisulfite-treated DNA sequencing data for leaf and fiber of TM-1  
639 from the National Center for Biotechnology Information (NCBI) Sequence Read  
640 Archive collection (SRX710548-SRX710553). Trimmomatic software was applied to  
641 clip sequencing adapters and filter low-quality reads<sup>74</sup>. The clean reads for the two  
642 samples were mapped to the TM-1 genome using Bismark software (version 0.13.0;  
643 -N 1 -L 30)<sup>75</sup>. The multiple mapping and PCR duplication reads were filtered to  
644 obtain a unique mapping BAM file. The Bismark methylation extractor program was  
645 run to extract potentially methylated cytosines. In this step, cytosines in CG, CHG and  
646 CHH contexts covered by at least three sequencing reads were retained for a binomial  
647 test ( $P$ -value cutoff  $1e-5$ ).

648

### 649 **DNase I digestion of chromatin**

650 DNase I digestion of chromatin was conducted accordingly to Zhang *et al* (2015) with  
651 some modifications<sup>76</sup>. Briefly, chromatin extraction was performed as described in  
652 our previous study<sup>77</sup>. For each sample, 100 g 10 DPA fiber and 1.5 g young leaves at  
653 the seedling stage were used for chromatin extraction, respectively. Extracted nuclei  
654 were washed once with  $1\times$  DNase I buffer before DNase I (Roche; Lot#11781700)  
655 digestion. Nuclei were re-suspended with 500  $\mu$ L  $1\times$  DNase I buffer. A 20  $\mu$ L aliquot  
656 was retained as undigested control. Remaining nuclei were treated with 100 U DNase  
657 I and were incubated at 37°C for 10 min. Immediately, both control and DNase I  
658 digested nuclei of each sample were subjected to histone removal, DNA purification,  
659 RNase A treatment and fragment isolation. For each sample, this experiment was  
660 performed for at least two biological replicates.

661

### 662 **DNase-seq and DHS identification**

663 Purified DNA fragments of between 100 bp and 200 bp following DNase I digestion  
664 were isolated with a Pippin HT (Sage Science, Beverly, MA, USA). A total of 10 ng  
665 of the isolated fragments was used for library construction using the Illumina TruSeq  
666 Sample Prep Kit. Libraries were sequenced using the Illumina HiSeq 2000 system  
667 (paired-end 100-bp reads). After clipping adapters and trimming low-quality reads,  
668 clean reads were mapped to the TM-1 genome using Bowtie2 (version 2.2.4)<sup>78</sup>. The  
669 unique mapping data were processed to identify DNase I hypersensitive sites (DHSs).  
670 To identify DHSs, we ran the F-seq program with a 300-bp bandwidth<sup>79</sup>. MACS  
671 (version 1.4.2)<sup>80</sup>, another peak-calling algorithm, was also run to identify DHSs. To  
672 run MACS, randomly fragmented DNA sequencing data were used as control  
673 (*P*-value 1e-5). Only peaks detected by both program tools were taken as candidate  
674 DHSs (**Supplementary Table 19**). Genome coverage of DNase-seq data in cotton  
675 was calculated using the coverageBed program embedded in the Bedtools package<sup>81</sup>.  
676 Chromosomal distribution of DHSs was analyzed in 1 Mb windows sliding 200 Kb.

677

#### 678 **Motif discovery**

679 The promoter DHSs were screened for transcription factor (TF) binding motifs using  
680 the findMotifsGenome.pl program in HOMER software (see URLs)<sup>82</sup>, with the  
681 parameters ‘-size given -len 8,10,12 -chopify -mset plants’. In HOMER, motifs with  
682 the *P*-value cutoffs of  $P < 0.01$  for known motifs and  $P < 1 \times 10^{-12}$  for *de novo* motifs  
683 were retained. The 2 kb upstream sequences of genes were used for motif discovery  
684 by the Patch 1.0 program, which searches the TRANSFAC Public 6.0 database (see  
685 URLs), with the following parameters: 1) the minimum length of sites was 8; 2) the  
686 maximum number of mismatches was 1; 3) the mismatch penalty was 100; 4) the  
687 lower score boundary was 87.5.

688

#### 689 **Chromatin immunoprecipitation (ChIP)**

690 Ca. 2 g of cotton leaves was cross-linked by vacuum infiltration with 1%  
691 formaldehyde for 35 min. Chromatin was extracted and fragmented to 200 to 500 bp  
692 by sonication. ChIP was performed as previously described<sup>77</sup>. Antibodies against

693 H3K4me1 (Abcam; ab8895), H3K4me3 (Abcam; ab8580), H3K9me2 (Abcam;  
694 ab1220) and H3K27me3 (ABclonal; A2363) were cross-linked with Dynabeads®  
695 protein A (Life Technologies; Lot#165116310) and respectively added to the  
696 sonicated samples for immunoprecipitation. All the ChIP experiments were carried  
697 out as two biological replicates.

698

### 699 **ChIP-Seq and data analysis**

700 For each sample, a total of 10 ng ChIP DNA and Input control DNA were used for  
701 library construction using the Illumina TruSeq Sample Prep Kit, according to the  
702 manufacturer's instructions. ChIP libraries were sequenced on the Illumina HiSeq  
703 3000 system (paired-end 150-bp reads). The clean sequencing reads were mapped to  
704 the TM-1 genome using Bowtie2 (version 2.2.4)<sup>78</sup>. After removing PCR duplication  
705 and multiple mapping reads, the unique mapping data were used to call histone  
706 modification peaks using MACS software (version 2.1.0)<sup>80</sup>. The "--broad" parameter  
707 was on for calling H3K4me1, H3K9me2 and H3K27me3 peaks, and was off for  
708 calling H3K4me3 peaks ( $P$ -value  $1e-5$ ). The Input DNA sequencing data was used as  
709 a control.

710

### 711 **Hi-C experiments and sequencing**

712 Cotton leaves were cross-linked in 20 ml of fresh ice-cold Nuclei Isolation Buffer and  
713 1 ml of ~36% formaldehyde solution under vacuum for 40 min at room temperature.  
714 This reaction was quenched by adding 1 mL of 2 M glycine under vacuum infiltration  
715 for additional 5 min. The clean samples were ground to powder in liquid nitrogen.  
716 Chromatin extraction was similar to that for the DNase I digestion experiment. The  
717 procedures were similar to those described previously<sup>83</sup>. Briefly, chromatin was  
718 digested for 16 h with 200 U (4  $\mu$ l) *Hind*III restriction enzyme (Takara) at 37°C. DNA  
719 ends were labelled with biotin, incubated at 37°C for 45 min, and enzyme was  
720 inactivated with 20% SDS solution. DNA ligation was performed by the addition of  
721 T4 DNA ligase (Fermentas) and incubated at 4°C for 1 h followed by 22°C for 4 h.  
722 After ligation, proteinase K was added to reverse cross-linking by incubation at 65°C

723 overnight. DNA fragments were purified and dissolved in 86  $\mu$ L of water. Un-ligated  
724 ends were then removed. Purified DNA was fragmented to a size of 300-500 bp  
725 followed by repair of DNA ends. DNA fragments labeled by biotin were finally  
726 separated on Streptavidin C1 beads (Life Technologies). Libraries were constructed  
727 using the Illumina TruSeq DNA Sample Prep Kit according to the manufacturer's  
728 instructions. TA cloning was performed to examine the quality of Hi-C library. Hi-C  
729 libraries were sequenced on the Illumina HiSeq 3000 system. The Hi-C experiment  
730 was carried out as two biological replicates.

731

### 732 **Hi-C data analysis**

733 Raw Hi-C data were processed to filter low-quality reads and trim adapters using  
734 Trimmomatic (version 0.32)<sup>74</sup>. Clean reads were mapped to the TM-1 genome using a  
735 two-step approach embedded in the HiC-Pro software (version 2.7.1; see URLs)<sup>84</sup>.  
736 After discarding low mapping quality reads, multiple mapping reads and singletons,  
737 the unique mapping reads were retained in a single file. Read pairs that did not map  
738 close to a restriction site, or were not within the expected fragment size following  
739 shearing, were first filtered. Subsequent filtering analyses were performed to discard  
740 read pairs from invalid ligation products, including dangling-end and self-ligation,  
741 and from PCR artifacts. The remaining valid read pairs were divided into  
742 intra-chromosomal pairs and inter-chromosomal pairs. Contact maps were constructed  
743 with chromosome bins of equal sizes for 5 kb, 10 kb, 20 kb, 100 kb, 200 kb and 500  
744 kb. The raw contact maps were then normalized using a sparse-based implementation  
745 of the iterative correction method in HiC-Pro.

746 Chromatin interactions (20 kb–2 Mb) were identified using a method of  
747 statistical confidence estimation, *Fit-Hi-C*<sup>85</sup>. To run *Fit-Hi-C*, fragments less than 2  
748 kb were merged to exclude possible Hi-C bias. Results from the second pass after an  
749 initial fit were used for further analysis. Fragments overlapping with intergenic DHSs  
750 or promoters were extracted to construct a regulatory interactome. Chromatin

751 interactions with a false discovery rate (FDR) of 0.05 were retained and then  
752 compared with genomic localization of intergenic DHSs and promoters to map  
753 promoter-centered interactions. Topologically associated domain-like (TAD-like) and  
754 boundary-like regions were identified using the TopDom method at a 50 kb  
755 resolution<sup>86</sup>. TopDom was processed with a window size of 5.

756 **References**

- 757 1. Gross, B.L. & Olsen, K.M. Genetic perspectives on crop domestication.  
758 *Trends Plant Sci.* **15**, 529–537 (2010).
- 759 2. Varshney, R.K., Terauchi, R. & McCouch, S.R. Harvesting the promising  
760 fruits of genomics: applying genome sequencing technologies to crop breeding.  
761 *PLoS Biol.* **12**, e1001883 (2014).
- 762 3. Crossa, J. *et al.* Genomic prediction in CIMMYT maize and wheat breeding  
763 programs. *Heredity (Edinb)* **112**, 48–60 (2014).
- 764 4. Chen, Z.J., Scheffler, B.E. & Dennis, E. Toward sequencing cotton  
765 (*Gossypium*) genomes. *Plant Physiol.* **145**, 1303–1310 (2007).
- 766 5. Senchina, D.S. *et al.* Rate variation among nuclear genes and the age of  
767 polyploidy in *Gossypium*. *Mol. Biol. Evol.* **20**, 633–643 (2003).
- 768 6. Stewart, J.M., Oosterhuis, D., Heitholt, J.J., Mauney, J.R. *Physiology of*  
769 *Cotton* (Springer Netherlands, Dordrecht, 2010).
- 770 7. Rapp, R.A. *et al.* Gene expression in developing fibers of Upland cotton  
771 (*Gossypium hirsutum* L.) was massively altered by domestication. *BMC Biol.*  
772 **8**, 139 (2010).
- 773 8. Yoo, M.J. & Wendel, J.F. Comparative evolutionary and developmental  
774 dynamics of the cotton (*Gossypium hirsutum*) fiber transcriptome. *PLoS Genet.*  
775 **10**, e1004073 (2014).
- 776 9. Zhang, T. *et al.* Sequencing of allotetraploid cotton (*Gossypium hirsutum* L.  
777 acc. TM-1) provides a resource for fiber improvement. *Nat. Biotechnol.* **33**,  
778 531–537 (2015).
- 779 10. Li, F. *et al.* Genome sequence of cultivated Upland cotton (*Gossypium*  
780 *hirsutum* TM-1) provides insights into genome evolution. *Nat. Biotechnol.* **33**,  
781 524–530 (2015).
- 782 11. Nie, X. *et al.* Genome-wide SSR-based association mapping for fiber quality  
783 in nation-wide upland cotton inbred cultivars in China. *BMC Genomics* **17**,  
784 352 (2016).
- 785 12. Zhou S.H. *Genogram of cotton varieties in China* (Sichuan Science and  
786 Technology Press, Chengdu, 2000).

- 787 13. Huang Z.K. *Cotton varieties and their genealogy in China* (Chinese  
788 Agricultural Press, Beijing, 2007).
- 789 14. Doebley, J.F., Gaut, B.S. & Smith, B.D. The molecular genetics of crop  
790 domestication. *Cell* **127**, 1309–1321 (2006).
- 791 15. Hufford, M.B. *et al.* Comparative population genomics of maize domestication  
792 and improvement. *Nat. Genet.* **44**, 808–811 (2012).
- 793 16. Huang, X. *et al.* Genome-wide association studies of 14 agronomic traits in  
794 rice landraces. *Nat. Genet.* **42**, 961–967 (2010).
- 795 17. Zhou, Z. *et al.* Resequencing 302 wild and cultivated accessions identifies  
796 genes related to domestication and improvement in soybean. *Nat. Biotechnol.*  
797 **33**, 408–414 (2015).
- 798 18. Lin, T. *et al.* Genomic analyses provide insights into the history of tomato  
799 breeding. *Nat. Genet.* **46**, 1220–1226 (2014).
- 800 19. Said, J.I. *et al.* A comparative meta-analysis of QTL between intraspecific  
801 *Gossypium hirsutum* and interspecific *G. hirsutum* × *G. barbadense*  
802 populations. *Mol. Genet. Genomics* **290**, 1003–1025 (2015).
- 803 20. Han, L.B. *et al.* The dual functions of *WLIM1a* in cell elongation and  
804 secondary wall formation in developing cotton fibers. *Plant Cell* **25**,  
805 4421–4438 (2013).
- 806 21. Banno, H. & Chua, N.H. Characterization of the *Arabidopsis* formin-like  
807 protein AFH1 and its interacting protein. *Plant Cell Physiol.* **41**, 617–626  
808 (2000).
- 809 22. Zhou, Y. *et al.* Cotton (*Gossypium hirsutum*) 14-3-3 proteins participate in  
810 regulation of fiber initiation and elongation by modulating brassinosteroid  
811 signalling. *Plant Biotechnol. J.* **13**, 269–280 (2015).
- 812 23. Jakoby, M.J. *et al.* Transcriptional profiling of mature *Arabidopsis* trichomes  
813 reveals that *NOECK* encodes the MIXTA-like transcriptional regulator  
814 MYB106. *Plant Physiol.* **148**, 1583–1602 (2008).
- 815 24. Bueso, E. *et al.* *ARABIDOPSIS THALIANA HOMEODOMAIN25* uncovers a role  
816 for gibberellins in seed longevity. *Plant Physiol.* **164**, 999–1010 (2014).
- 817 25. He, X.C. *et al.* Molecular cloning, expression profiling, and yeast  
818 complementation of 19 beta-tubulin cDNAs from developing cotton ovules. *J.*  
819 *Exp. Bot.* **59**, 2687–2695 (2008).

- 820 26. Tan, J. *et al.* A genetic and metabolic analysis revealed that cotton fiber cell  
821 development was retarded by flavonoid naringenin. *Plant Physiol.* **162**, 86–95  
822 (2013).
- 823 27. Nakajima, K. *et al.* *SPIRAL1* encodes a plant-specific microtubule-localized  
824 protein required for directional control of rapidly expanding *Arabidopsis* cells.  
825 *Plant Cell* **16**, 1178–1190 (2004).
- 826 28. Cheng, Q. *et al.* Functional identification of AtFao3, a membrane bound long  
827 chain alcohol oxidase in *Arabidopsis thaliana*. *Febs Letters* **574**, 62–68  
828 (2004).
- 829 29. Szumlanski, A.L. & Nielsen, E. The Rab GTPase RabA4d regulates pollen  
830 tube tip growth in *Arabidopsis thaliana*. *Plant Cell* **21**, 526–544 (2009).
- 831 30. Applequist, W.L., Cronn, R. & Wendel, J.F. Comparative development of  
832 fiber in wild and cultivated cotton. *Evol. Dev.* **3**, 3–17 (2001).
- 833 31. Hovav, R. *et al.* The evolution of spinnable cotton fiber entailed prolonged  
834 development and a novel metabolism. *PLoS Genet.* **4**, e25 (2008).
- 835 32. Cheng, F. *et al.* Subgenome parallel selection is associated with morphotype  
836 diversification and convergent crop domestication in *Brassica rapa* and  
837 *Brassica oleracea*. *Nat. Genet.* **48**, 1218–1224 (2016).
- 838 33. Deeks, M.J., Hussey, P.J. & Davies, B. Formins: intermediates in  
839 signal-transduction cascades that affect cytoskeletal reorganization. *Trends*  
840 *Plant Sci.* **7**, 492-498 (2002).
- 841 34. Bischoff, V. *et al.* *TRICHOME BIREFRINGENCE* and its homolog  
842 *AT5G01360* encode plant-specific DUF231 proteins required for cellulose  
843 biosynthesis in *Arabidopsis*. *Plant Physiol.* **153**, 590–602 (2010).
- 844 35. Brown, D.M. *et al.* Identification of novel genes in *Arabidopsis* involved in  
845 secondary cell wall formation using expression profiling and reverse genetics.  
846 *Plant Cell* **17**, 2281–2295 (2005).
- 847 36. Guo, K. *et al.* Fiber elongation requires normal redox homeostasis modulated  
848 by cytosolic ascorbate peroxidase in cotton (*Gossypium hirsutum*). *J. Exp. Bot.*  
849 **67**, 3289–3301 (2016).
- 850 37. Feng, H. *et al.* Molecular analysis of proanthocyanidins related to  
851 pigmentation in brown cotton fiber (*Gossypium hirsutum* L.). *J. Exp. Bot.* **65**,  
852 5759–5769 (2014).



- 853 38. Xiao, Y.H. *et al.* Transcriptome and biochemical analyses revealed a detailed  
854 proanthocyanidin biosynthesis pathway in brown cotton fiber. *PLoS One* **9**,  
855 e86344 (2014).
- 856 39. Maurano, M.T. *et al.* Large-scale identification of sequence variants  
857 influencing human transcription factor occupancy *in vivo*. *Nat. Genet.* **47**,  
858 1393–1401 (2015).
- 859 40. Wittkopp, P.J. & Kalay, G. *Cis*-regulatory elements: molecular mechanisms  
860 and evolutionary processes underlying divergence. *Nat. Rev. Genet.* **13**, 59–69  
861 (2012).
- 862 41. Burgess, D.G., Xu, J. & Freeling, M. Advances in understanding *cis* regulation  
863 of the plant gene with an emphasis on comparative genomics. *Curr. Opin.*  
864 *Plant Biol.* **27**, 141–147 (2015).
- 865 42. Zhang, W. *et al.* High-resolution mapping of open chromatin in the rice  
866 genome. *Genome Res.* **22**, 151–162 (2012).
- 867 43. Hobo, T., Kowyama, Y. & Hattori, T. A bZIP factor, TRAB1, interacts with  
868 VP1 and mediates abscisic acid-induced transcription. *Proc. Natl. Acad. Sci.*  
869 *USA* **96**, 15348–15353 (1999).
- 870 44. Wang, S. *et al.* Control of plant trichome development by a cotton fiber MYB  
871 gene. *Plant Cell* **16**, 2323–2334 (2004).
- 872 45. Koini, M.A. *et al.* High temperature-mediated adaptations in plant architecture  
873 require the bHLH transcription factor *PIF4*. *Curr. Biol.* **19**, 408–413 (2009).
- 874 46. Cook, P.R. The organization of replication and transcription. *Science* **284**,  
875 1790–1795 (1999).
- 876 47. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions  
877 reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
- 878 48. Fullwood, M.J. *et al.* An oestrogen-receptor- $\alpha$ -bound human chromatin  
879 interactome. *Nature* **462**, 58–64 (2009).
- 880 49. Zhang, Y.B. *et al.* Chromatin connectivity maps reveal dynamic  
881 promoter-enhancer long-range associations. *Nature* **504**, 306–310 (2013).
- 882 50. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a  
883 topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).
- 884 51. Huang, X. *et al.* A map of rice genome variation reveals the origin of  
885 cultivated rice. *Nature* **490**, 497–501 (2012).

- 886 52. Neph, S. *et al.* An expansive human regulatory lexicon encoded in  
887 transcription factor footprints. *Nature* **489**, 83–90 (2012).
- 888 53. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**,  
889 554–566 (2015).
- 890 54. Zhang, W., Zhang, T., Wu, Y. & Jiang, J. Genome-wide identification of  
891 regulatory DNA elements and protein-binding footprints using signatures of  
892 open chromatin in *Arabidopsis*. *Plant Cell* **24**, 2719–2731 (2012).
- 893 55. Wang, C. *et al.* Genome-wide analysis of local chromatin packing in  
894 *Arabidopsis thaliana*. *Genome Res.* **25**, 246–256 (2015).
- 895 56. Paterson, A.H., Brubaker, C.L. & Wendel, J.F. A rapid method for extraction  
896 of cotton (*Gossypium* spp.) genomic DNA suitable for RFLP or PCR analysis.  
897 *Plant Mol. Biol. Rep.* **11**, 122–127 (1993).
- 898 57. Li, H. & Durbin, R. Fast and accurate short read alignment with  
899 Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
- 900 58. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework  
901 for analyzing next-generation DNA sequencing data. *Genome Res.* **20**,  
902 1297–1303 (2010).
- 903 59. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools.  
904 *Bioinformatics* **25**, 2078–2079 (2009).
- 905 60. Cingolani, P. *et al.* A program for annotating and predicting the effects of  
906 single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila*  
907 *melanogaster* strain w(1118); iso-2; iso-3. *Fly* **6**, 80–92 (2012).
- 908 61. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of  
909 genomic structural variation. *Nat. Meth.* **6**, 677–681 (2009).
- 910 62. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end  
911 and split-read analysis. *Bioinformatics* **28**, I333–I339 (2012).
- 912 63. Zhuang, J. & Weng, Z. Local sequence assembly reveals a high-resolution  
913 profile of somatic structural variations in 97 cancer genomes. *Nucleic Acids*  
914 *Res.* **43**, 8146–8156 (2015).
- 915 64. Felsenstein, J. PHYLIP-phylogeny inference package (version 3.2). *Cladistics*  
916 **5**, 164–166 (1989).
- 917 65. Price, A.L. *et al.* Principal components analysis corrects for stratification in  
918 genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

- 919 66. Falush, D., Stephens, M. & Pritchard, J.K. Inference of population structure  
920 using multilocus genotype data: Linked loci and correlated allele frequencies.  
921 *Genetics* **164**, 1567–1587 (2003).
- 922 67. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and  
923 population-based linkage analyses. *Am. J. of Hum. Genet.* **81**, 559–575 (2007).
- 924 68. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for  
925 selective sweeps. *Genome Res.* **20**, 393–402 (2010).
- 926 69. Bradbury, P.J. *et al.* TASSEL: software for association mapping of complex  
927 traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
- 928 70. Wang, Y. *et al.* MCSScanX: a toolkit for detection and evolutionary analysis of  
929 gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
- 930 71. Liu, D., Zhang, X., Tu, L., Zhu, L. & Guo, X. Isolation by  
931 suppression-subtractive hybridization of genes preferentially expressed during  
932 early and late fiber development stages in cotton. *Mol. Biol.* **40**, 741–749  
933 (2006).
- 934 72. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions  
935 with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
- 936 73. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals  
937 unannotated transcripts and isoform switching during cell differentiation. *Nat.*  
938 *Biotechnol.* **28**, 511–515 (2010).
- 939 74. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for  
940 Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
- 941 75. Krueger, F. & Andrews, S.R. Bismark: a flexible aligner and methylation  
942 caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
- 943 76. Zhang, W. & Jiang, J. Genome-wide mapping of DNase I hypersensitive sites  
944 in plants. *Methods Mol. Biol.* **1284**, 71–89 (2015).
- 945 77. Wang, M. *et al.* Multi-omics maps of cotton fiber reveal epigenetic basis for  
946 staged single-cell differentiation. *Nucleic Acids Res.* **44**, 4067–4079 (2016).
- 947 78. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2.  
948 *Nat. Methods* **9**, 357–359 (2012).
- 949 79. Boyle, A.P., Guinney, J., Crawford, G.E. & Furey, T.S. F-Seq: a feature  
950 density estimator for high-throughput sequence tags. *Bioinformatics* **24**,  
951 2537–2538 (2008).

- 952 80. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X.S. Identifying ChIP-seq  
953 enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).
- 954 81. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for  
955 comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
- 956 82. Heinz, S. *et al.* Simple combinations of lineage-determining transcription  
957 factors prime *cis*-regulatory elements required for macrophage and B cell  
958 identities. *Mol. Cell* **38**, 576–589 (2010).
- 959 83. Xie, T. *et al.* *De novo* plant genome assembly based on chromatin interactions:  
960 a case study of *Arabidopsis thaliana*. *Mol. Plant* **8**, 489–492 (2015).
- 961 84. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data  
962 processing. *Genome Biol.* **16**, 259 (2015).
- 963 85. Ay, F., Bailey, T.L. & Noble, W.S. Statistical confidence estimation for Hi-C  
964 data reveals regulatory chromatin contacts. *Genome Res.* **24**, 999–1011 (2014).
- 965 86. Shin, H. *et al.* TopDom: an efficient and deterministic method for identifying  
966 topological domains in genomes. *Nucleic Acids Res.* **44**, e70 (2016).