LETTERS
https://doi.org/10.1038/s41588-018-0116-x
nature genetics

Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits

Xiongming Du[1,2,7], Gai Huang[3,4,7], Shoupu He[2,7], Zhaoen Yang[2,7], Gaofei Sun[1,7], Xiongfeng Ma[2,7], Nan Li[5,7], Xueyan Zhang[2], Junling Sun[2], Min Liu[6], Yinhua Jia[2], Zhaoe Pan[2], Wenfang Gong[2], Zhaohui Liu[3], Heqin Zhu[2], Lei Ma[2], Fuyan Liu[6], Daigang Yang[2], Fan Wang[6], Wei Fan[5], Qian Gong[2], Zhen Peng[2], Liru Wang[2], Xiaoyang Wang[2], Shuangjiao Xu[2], Haihong Shang[2], Cairui Lu[2], Hongkun Zheng[6], Sanwen Huang[5], Tao Lin[5]*, Yuxian Zhu[3,4]* and Fuguang Li[1,2]*

The ancestors of *Gossypium arboreum* and *Gossypium herbaceum* provided the A subgenome for the modern cultivated allotetraploid cotton. Here, we upgraded the *G. arboreum* genome assembly by integrating different technologies. We resequenced 243 *G. arboreum* and *G. herbaceum* accessions to generate a map of genome variations and found that they are equally diverged from *Gossypium raimondii*. Independent analysis suggested that Chinese *G. arboreum* originated in South China and was subsequently introduced to the Yangtze and Yellow River regions. Most accessions with domestication-related traits experienced geographic isolation. Genome-wide association study (GWAS) identified 98 significant peak associations for 11 agronomically important traits in *G. arboreum*. A nonsynonymous substitution (cysteine-to-arginine substitution) of GaKASIII seems to confer substantial fatty acid composition (C16:0 and C16:1) changes in cotton seeds. Resistance to fusarium wilt disease is associated with activation of *GaGSTF9* expression. Our work represents a major step toward understanding the evolution of the A genome of cotton.

Cotton is one of the world's most important commercial crops and is also a valuable resource for studying plant polyploidization[1]. *G. arboreum* was probably domesticated on Madagascar or in the Indus Valley (Mohenjo Daro), and was subsequently dispersed to Africa and other areas of Asia[2]. It was initially introduced to China more than 1,000 years ago as an ornamental plant[3,4]. Over the course of its adaptation to local agroecological environments, and under the influence of human selection, the Chinese *G. arboreum* population developed into a distinct geographical race referred to as 'sinense cotton'[4].

Although cotton breeders have constructed various genetic maps based on RFLP[5] and simple-sequence-repeat[6] markers, no causal genes responsible for the excellent agronomic and economic traits from *G. arboreum* or *G. herbaceum* have been identified. Likewise, efforts to introduce these important characteristics from diploids into tetraploids through intra- and interspecific hybridizations have not been productive[7–9]. The release of genome sequences for *G. raimondii*[10,11], *G. arboreum*[12], *Gossypium hirsutum*[13,14], and *Gossypium barbadense*[15,16] has provided the prerequisites for

Table 1 | Global statistical comparison of *G. arboreum* assembly between the updated genome and a previously published genome

Category	Previously published genome[12]					Updated genome				
	Numbers	N50 (kb)	Longest (Mb)	Size (Mb)	Percentage of assembly	Numbers	N50 (kb)	Longest (Mb)	Size (Mb)	Percentage of assembly
Contigs	40,381	72	0.8	1,561	NA	8,223	1,100	12.37	1,710	100
Scaffolds	7,914	665.8	5.9	1,694	100	NA	NA	NA	NA	NA
Anchored and oriented	3,740	790	5.9	1,532	90	3,720	730	12.37	1,573	92
Gene annotated	41,330	NA	NA	105	6.2	40,960	NA	NA	123	7.2
Repeat sequences	NA	NA	NA	1,160	68.5	NA	NA	NA	1,460	85.39

NA, not applicable.

[1]Research Base, Anyang Institute of Technology, State Key Laboratory of Cotton Biology, Anyang, China. [2]Institute of Cotton Research of the Chinese Academy of Agricultural Sciences, Anyang, China. [3]State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences, Peking University, Beijing, China. [4] Institute for Advanced Studies and College of Life Sciences, Wuhan University, Wuhan, China. [5]Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China. [6]Biomarker Technologies Corporation, Beijing, China. [7]These authors contributed equally: Xiongming Du, Gai Huang, Shoupu He, Zhaoen Yang, Gaofei Sun, Xiongfeng Ma, Nan Li. *e-mail: lintao_solab@126.com; zhuyx@whu.edu.cn; aylifug@163.com

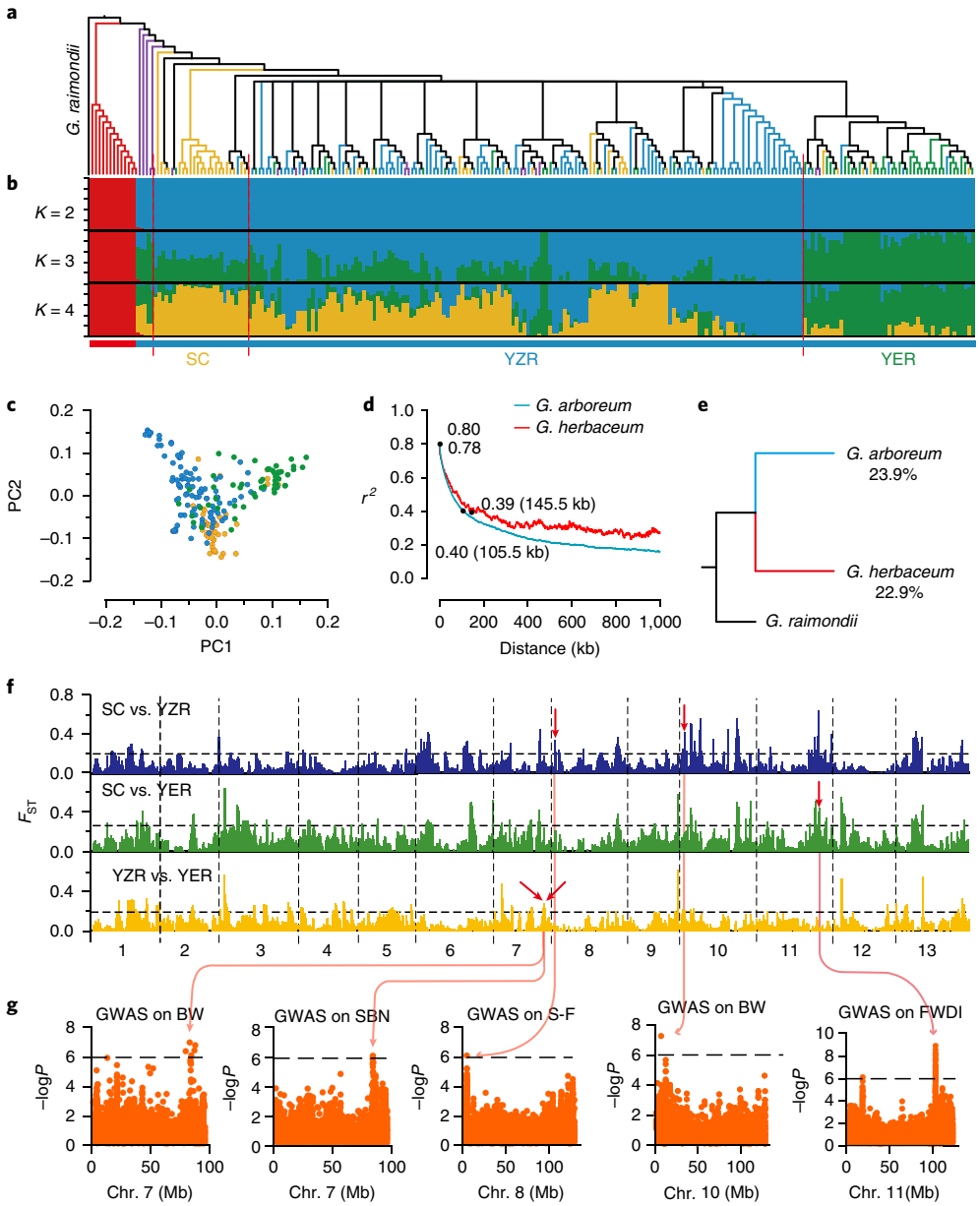

Fig. 1 | Genomic divergence and geographic-relationship analysis. a, Neighbor-joining tree of 243 diploid cotton accessions, based on whole-genome SNP studies. Branch colors indicate different geographical distributions of *G. arboreum*: South China (SC, orange); Yangtze River region (YZR, blue); Yellow River region (YER, green); non-Chinese and origin-unknown accessions (purple). *G. herbaceum* is in red. **b**, Population structure based on different numbers of clusters ($K = 2$–4). The *x* axis indicates the *G. herbaceum* group (red) and *G. arboreum* groups (blue), with all accessions arranged in the same order as in **a**. The left *y* axis quantifies genetic diversity in each accession, which is represented by a vertical color-coded column. **c**, Principal component (PC) analysis plots of the first two components for all Chinese *arboreum* accessions, using the same colors as in **a**. **d**, LD decay-distance analysis. **e**, Phylogenetic and ancestral-allele analysis of *Gossypium* species. **f**, Highly divergent genomic regions and overlapping GWAS signals. The vertical columns above the dashed lines indicate highly divergent regions; red arrows indicate highly divergent regions overlapped with GWAS signals shown in **g**. The GWAS signals were detected in any pairwise divergent regions between SC versus YZR, SC versus YER, or YZR versus YER. **g**, Local Manhattan plots obtained from GWAS signals of geographically selected traits. Dashed lines indicate the threshold for GWAS ($-\log P = 6$). BW, boll weight; SBN, sympodial branch node; S-F, flowering date; Chr., chromosome.

the study of population genetics, cultivation, and domestication. Genome-wide association studies have identified many candidate genes or quantitative trait loci (QTLs) in rice[17–19], maize[20,21], soybean[22], foxtail millet[23], cucumber[24], tomato[25], and upland cotton[26,27]. In this study, we reassembled a high-quality *G. arboreum* genome on the basis of PacBio long-reads and Hi-C technologies, and analyzed the population structure and genomic divergence trends of 243 diploid cotton accessions. We identified a number of candidate loci that may facilitate the genetic improvement of cotton lint production.

We generated 142.54 Gb of raw PacBio reads (approximately 77.6-fold genome coverage) by using SMRT sequencing technology and assembled these reads into 8,223 contigs, producing a 1,710 Mb *G. arboreum* genome with a contig N50 of 1,100 kb; the longest contig in the new assembly was 12.37 Mb (Table 1). We

Table 2 | A subset of trait-associated SNPs and candidate genes from GWAS analysis

Traits	Environment[a]	Chromosome	Position	SNP	MAF[c]	-log*P*	Gene	Annotation
Sympodial branch node	AY	8	127060603	A/G	0.31	6.09	*Ga08G2687*	Flowering locus T-like protein
C14:0	AY	2	9391129	C/T	0.11	7.78	*Ga02G0510*[d]	Lipase
C14:0	AY	6	2398472[b]	C/T	0.06	7.05	*Ga06G0267*[d]	Unknown protein
C16:0/C16:1	AY	11	122156027[b]	T/C	0.46	11.50/17.47	*Ga11G3851*[d]	3-Oxoacyl-[acyl-carrier-protein ACP] synthase III
Lint percentage	AY	5	16113663	A/G	0.26	6.38	*Ga05G1773*	ARM repeat superfamily protein
Lint percentage	AKS	13	4972998	T/C	0.44	6.30	*Ga13G0423*[d]	Leucine-rich repeat protein kinase family protein
Boll weight	AY	3	133,798,238[b]	G/A	0.11	8.61	*Ga03G2524*	Alpha-helical ferredoxin
Boll weight	SY	13	2,309,774[b]	C/T	0.08	6.70	*Ga13G0227*	ACT domain-containing protein
Boll weight	AKS	8	117,950,057[b]	G/A	0.21	6.18	*Ga08G2045*[d]	Unknown protein
Fusarium wilt disease index	Indoor	11	103,191,949	T/C	0.08	8.96	*Ga11G2353*[d]	Glutathione *S*-transferase F9
Seed fuzz	AY	8	862,509[b]	C/T	0.15	18.94	*Ga08G0117*	Casparian-strip membrane protein

[a]AY, Anyang (Henan); AKS, Akesu (Xinjiang), SY, Sanya (Hainan). [b]SNP located in exon or intron.[c]MAF, minor allele frequency. [d]Genes with tissue- or stage-specific expression patterns.

also generated ~125 million valid Hi-C interacting unique pairs with a coverage number >20 (Supplementary Tables 1 and 2). We anchored and oriented 1,573 Mb of the assembly onto 13 pseudo-chromosomes with the aid of Hi-C sequence data by using base-calling corrections. This genome, compared with the previously published genome[12], was found to have a substantially lower number of incongruities outside of the expected diagonal when the Hi-C data were mapped against the updated genome (Supplementary Fig. 1a,b). Moreover, this updated *G. arboreum* genome shares substantially longer syntenic blocks in the corresponding chromosomes of the At subgenome (potentially the closest sequenced species) (Supplementary Fig. 1c,d and Supplementary Table 3). 85.39% of the updated genome is composed of repeat sequences (Supplementary Table 4). We produced a new set of 40,960 consensus protein-coding-gene models by integrating currently available methods (Supplementary Tables 5 and 6).

A total of 230 *G. arboreum* (A_2) and 13 *G. herbaceum* (A_1) lines were collected from South China (SC), the Yangtze River region (YZR), and the Yellow River region (YER) (Supplementary Fig. 2) and were resequenced (Supplementary Table 7). These regions represent most of the phenotypic and geographical diversity known for diploid cottons in China. Approximately 18.30 billion 125-bp paired-end reads—approximately 2.29 Tb of raw sequence—were generated on the Illumina HiSeq 2500 platform, with an average coverage depth of ~6.0× for each accession. The updated genome was used as the reference genome for SNP identification. On average, 99.68% of the reads for each accession were successfully aligned (Supplementary Table 7). We identified 17,883,108 high-quality SNPs and 2,470,515 indels (ranging from 1 to 190 bp in length), an average of 10.5 SNPs and 1.4 indels per kilobase. A total of 242,449 SNPs (1.36%) and 16,816 (0.68%) indels were located in coding regions of 36,205 *G. arboreum* genes. A total of 128,512 (0.72%) nonsynonymous SNPs were identified in 31,549 genes, and 11,372 (0.46%) frame-shifted indels were identified in 8,117 genes; 25,117 variants showed potentially large effects, including SNPs causing premature stop codons or longer-than-usual transcripts, and indels resulting in frame shifts, the introduction of stop codons, or other disruptions of protein-coding capacity (Supplementary Tables 8 and 9).

A subset of 72,419 SNPs was screened in greater detail to construct a neighbor-joining tree by using the *G. raimondii* genome[11] as the outgroup. *G. herbaceum* and *G. arboreum* were clustered in two independent clades after branching from *G. raimondii* (Fig. 1a,b and Supplementary Fig. 3). The *G. arboreum* clade could be divided into SC, YZR, and YER groups that exhibited strong geographical distribution patterns, a result further supported by principal component analysis (Fig. 1a–c). These two species were independently domesticated from different wild progenitors[28].

Compared with the YZR and YER group accessions, the SC group accessions had relatively poor agronomic traits (Supplementary Fig. 4). Additionally, the SC group had higher nucleotide diversity ($\pi = 0.211 \times 10^{-3}$) than the YZR ($\pi = 0.197 \times 10^{-3}$) and YER ($\pi = 0.199 \times 10^{-3}$) groups. This result indicated that *G. arboreum* was initially cultivated in South China and extended further to the Yangtze and Yellow River regions, in agreement with findings from a previous report[7] based on molecular diversity using simple sequence repeats[29]. Linkage disequilibrium (LD) analysis indicated that the physical distance between SNPs (reported as half of its maximum value) occurred at ~105.5 kb ($r^2 = 0.40$) for *G. arboreum* and at ~145.5 kb ($r^2 = 0.39$) for *G. herbaceum* (Fig. 1d). These values are comparable to those for soybean (~83 kb)[22] and rice landraces (~123 kb in indica, ~167 kb in japonica)[17], but much higher than those of cultivated maize (22–30 kb)[20]. Approximately 23.9% or 22.9% of the *G. arboreum* or *G. herbaceum* alleles, respectively, were aligned to the *G. raimondii* genome (Fig. 1e), thus indicating that *G. arboreum* and *G. herbaceum* are equally diverged from *G. raimondii*.

Artificial selection plays an important role during crop domestication and migration[30]. Model-based clustering showed that the YER group was significantly different from the SC and YZR groups (Fig. 1b; $K = 4$). Pairwise fixation statistic (F_{ST}) analysis (SC versus YZR; SC versus YER; and YZR versus YER) identified 59, 53, and 51 genomic regions with significant genetic divergence (top 5% of F_{ST} values) covering 3,162, 2,879, and 3,308 genes, respectively (Fig. 1f and Supplementary Tables 10–12). A total of 21 divergent genomic regions (~43.5 Mb containing 915 genes) between the SC and YZR groups were conserved between the SC and YER groups (Fig. 1f and Supplementary Table 13).

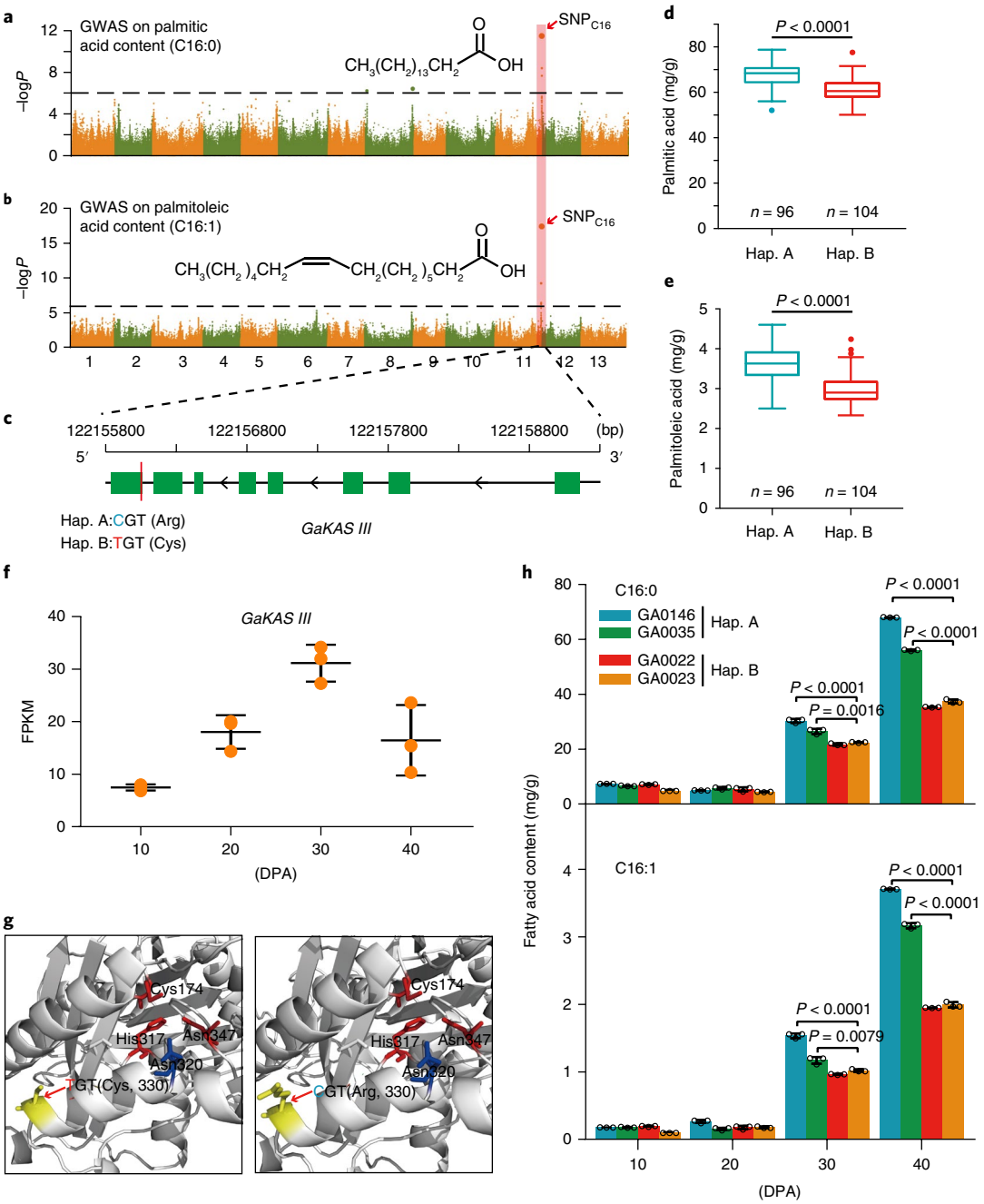

Fig. 2 | *GaKASIII* regulates cotton seed oil content. a,**b**, Manhattan plots for GWAS of the palmitic acid (C16:0) (**a**) and palmitoleic acid (C16:1) (**b**) content in cotton seeds. Dashed lines indicates the threshold for GWAS. The overlapping GWAS signals detected for both traits are highlighted by a shaded red column, and the strongest associated SNP_{C16} (chr. 11: 122156027 bp) is marked by a red arrow. **c**, Zoomed-in view of the strongest associated SNP_{C16} containing the *GaKASIII* (*Ga11G3851*) gene. Exons and introns are represented by boxes and lines, respectively. The position of the causal SNP is marked by a red line. Hap., haplotype. **d**,**e**, Comparisons of C16:0 (**d**) and C16:1 (**e**) content (mg/g) between haplotypes A and B in the GWAS population. In box plots, center line indicates median; box limits indicate upper and lower quartiles; whiskers denote 1.5× interquartile range; points shows outliers. **f**, *GaKASIII* expression during cotton ovule development. Data are presented as mean ± s.d. ($n = 3$ independent RNA-seq experiments). FPKM, fragments per kilobase per million reads. **g**, Protein-structure modeling of GaKASIII. Cys174, His317, and Asn347 (red) are the active sites. Asn320 (blue) functions as the CoA-binding site. The site for the p.Cys330/p.Arg330 substitution is marked in yellow. **h**, C16:0 and C16:1 fatty acid content analysis among haplotype A and B cotton accessions during ovule development. Data are presented as mean ± s.d. ($n = 3$ independent measurements). *P* values in this and all other figures were derived with two-tailed Student's *t* tests.

Manhattan plots and quantile–quantile plots for all 11 important traits from varied environments are shown in Supplementary Tables 14 and 15 and Supplementary Figs. 5–13. Among the 98 significant association signals (defined by $-\log P > 6$, including SNPs located both in genic and intergenic regions between two adjacent genes), 25 came from genic regions (exonic or intronic regions), including eight for morphological traits, six for yield, and three for seed oil traits. The remaining 73 signals came from noncoding regions (Supplementary Table 16). Major GWAS signals for agronomic traits that showed geographic differences in

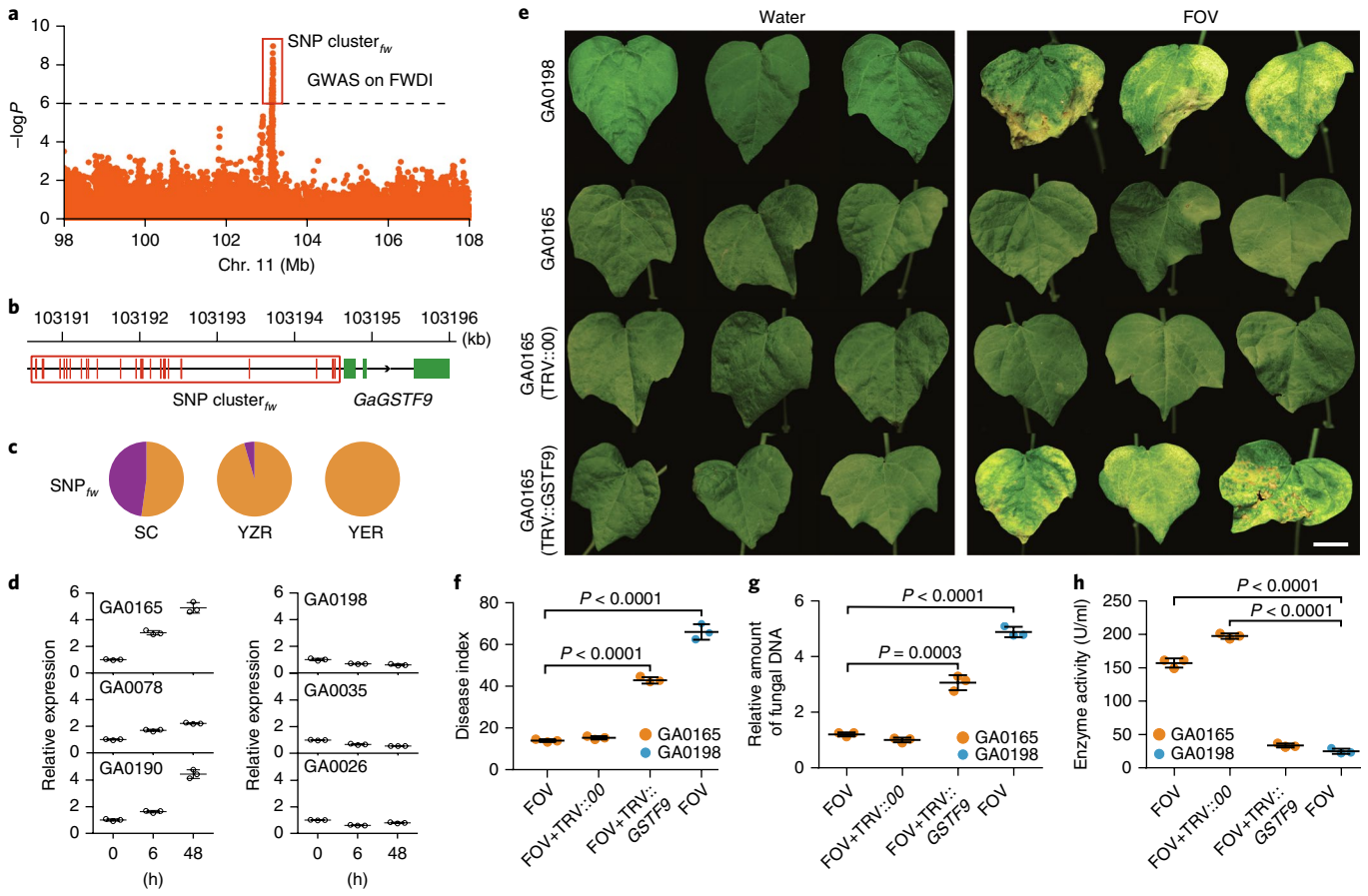

Fig. 3 | A genetic locus that underwent geographical isolation confers resistance to fusarium wilt disease. a, GWAS examining FWDI in a *G. arboreum* population. The strongest associated SNPs (SNP cluster$_{fw}$) are marked by red boxes with a black dashed line indicating the threshold for GWAS signals. **b**, Gene structure of *GaGSTF9* and its nearby strongly associated SNPs for FWDI (−log*P* > 6, red vertical lines). The corresponding SNP cluster$_{fw}$ is marked by a red box. **c**, The distribution of allele frequencies and geographical distribution of strongest SNP cluster$_{fw}$ (−log*P* = 8.96) in the GWAS population. The disease-susceptible and disease-tolerant alleles are shown in purple and orange, respectively. **d**, qRT-PCR analysis of *GaGSTF9* expression in roots after fusarium inoculation in highly tolerant (GA0165, GA0078 and GA0190) and highly susceptible (GA0198, GA0035, and GA0026) accessions. Gene expression at 0 h was set to 1. Data are presented as mean ± s.d. ($n = 3$ technical replicates). **e**, Disease symptoms of GA00198, GA0165, TRV::00, and TRV::GSTF9 plants after inoculation with water or FOV. The third true leaf was photographed. Scale bars, 1 cm. **f**, Disease index of GA00198, GA0165, TRV::00, and TRV:GSTF9 plants at 35 d post inoculation (dpi) with FOV. **g**, Relative content of FOV DNA in GA00198, GA0165, TRV::00, and TRV:GSTF9 plants at 35 dpi. **h**, GST enzyme activity measurement for GA0198, GA0165, TRV::00, and TRV::GSTF9 plants at 24 h post inoculation. Data in **f**–**h** are shown as mean ± s.d. ($n = 3$ independent experiments).

characteristics, such as sympodial branch node, flowering date, boll weight, and disease resistance, were found in conserved genomic regions (Fig. 1g, Table 2 and Supplementary Table 17). We thus conclude that maturity, yield, and disease-resistance traits have been under strong human and/or geographical selection.

Cotton is the world's sixth largest source of plant oil[31]. A significant SNP was detected in the eighth exon of the *GaKASIII* locus (*Ga11G3851*) on chromosome 11, which encodes 3-oxoacyl-[acyl-carrier-protein ACP] synthase III (Fig. 2a–c). *KASIII* encodes a key enzyme known to initiate fatty acid chain elongation from C2 to C4 and may ultimately determine the seed content of both palmitic acid (C16:0) and palmitoleic acid (C16:1)[32]. A polymorphism in *GaKASIII* results in a cysteine-to-arginine substitution in the conserved ACP_synthase_III_C domain (Fig. 2c). Haplotype B (TGT, cysteine) was mainly found in low-oil-content accessions, whereas haplotype A (CGT, arginine) was found in high-oil-content accessions (Fig. 2d,e). *GaKASIII* was expressed at the highest level at 30 d post anthesis (DPA) (Fig. 2f), which is a critical stage for seed oil accumulation[33]. Both C16:0 and C16:1 content accumulated at a significantly faster rate after 30 DPA in haplotype A accessions (Fig. 2h). Protein modeling with Phyre2 (ref. [34]) at >90% accuracy showed that this cysteine/arginine residue is located at an α-helix close to the enzyme active site and the CoA-binding site (Fig. 2g).

Fusarium wilt disease, caused by *Fusarium oxysporum* f. sp. *vasinfectum* (FOV), is one of the most severe threats to cotton production[35]. We performed GWAS for FOV resistance, as measured by the fusarium wilt disease index (FWDI), and found a strong association signal on chromosome 11 with a −log*P* value of 8.96 (Fig. 3a). Further analysis identified that this SNP cluster was localized in an upstream region of *Ga11G2353* (Fig. 3b), an ortholog of the *Arabidopsis GSTF9*, which encodes the Phi class of glutathione *S*-transferases involved in plant responses to biotic and abiotic stresses[36]. Accessions carrying the disease-susceptible allele 'T' were primarily found in the SC group, and all YER group members carried the disease-tolerant allele 'C' (Fig. 3c). *GaGSTF9* was upregulated only in tolerant lines after FOV inoculation of *G. arboreum* seedlings (Fig. 3d). *GaGSTF9*-silenced cotton lines (TRV::GSTF9, the virus-induced gene-silencing vector carrying the *GSTF9* gene) were found to be significantly more sensitive to FOV inoculation compared with empty-vector-carrying cotton lines (TRV::00)

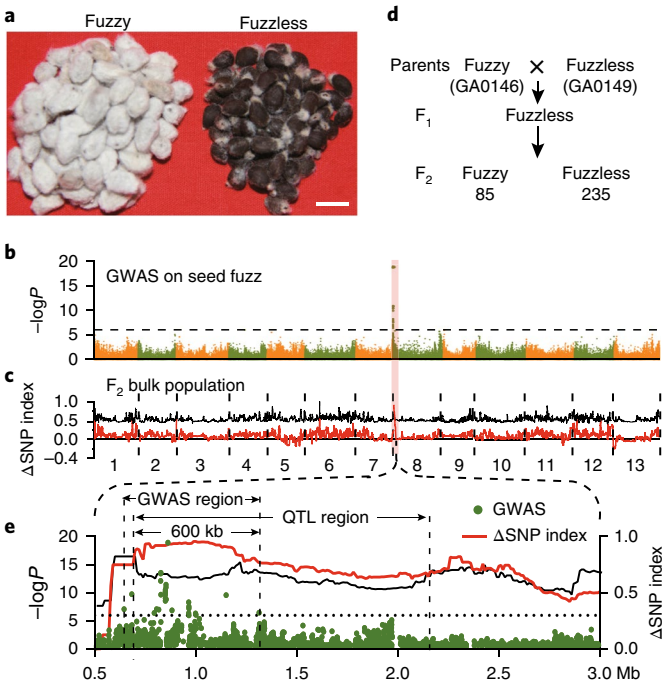

Fig. 4 | Both GWAS and QTL analysis identified the same region in the *G. arboreum* genome as being potentially important for seed fuzz development. a, The phenotypes of fuzzy (left) and fuzzless (right) *G. arboreum* seeds. Scale bars, 1 cm. **b**, Manhattan plot of GWAS examining seed fuzz in 215 *G. arboreum* accessions. **c**, QTL analysis of the fuzzless phenotype in the F_2 segregating population. The ΔSNP index (in which the SNP index of the fuzzless bulk population is subtracted from that of the fuzzy bulk population) and its 99% confidence interval are shown as red and black lines, respectively. The overlapping region in common between the GWAS and QTL analysis results after QTL-seq is highlighted for clarity. **d**, The F_2 population examined here was obtained by crossing a fuzzy accession (GA0146) and a fuzzless accession (GA0149). Crossing GA0146 with GA0149 resulted in a 1:3 (85:235, chi-square test, $P = 0.52$) segregating ratio for fuzzy versus fuzzless phenotypes. **e**, Local view around the overlapping region containing GWAS signals above the threshold (−log$P$ > 6). The left y axis indicates −log$P$ values for the GWAS results with its corresponding threshold value (−log$P$ > 6) shown by the dotted line. For the right y axis, the ΔSNP index and its corresponding 99% confidence interval are shown by red and black solid lines, respectively.

(Fig. 3e,f). Furthermore, the amount of fungal DNA was significantly higher, and the GST catalytic activity was significantly lower, in TRV::GSTF9 than in TRV::00 plants (Fig. 3g,h), thus suggesting that *GaGSTF9* may be a target for FOV resistance in *G. arboreum*.

Cotton fuzz comprises short fibers that cover the seed surfaces. We selected 158 fuzzy and 57 fuzzless accessions from *G. arboreum* accessions in a GWAS analysis that identified a strong association signal on chromosome 8 (~0.6 to ~1.3 Mb) (Fig. 4a,b). The ΔSNP index above 99% confidence intervals (QTL region) from a QTL analysis was also located on chromosome 8 (~0.70 to ~2.15 Mb) with a maximum of 0.959 (Fig. 4c). Analysis of an F_2 population obtained by crossing a fuzzy line (GA0146) with a fuzzless line (GA0149) identified a 1:3 segregation ratio for fuzzy and fuzzless phenotypes (Fig. 4d), thus indicating that a single locus controls fuzz initiation. When we zoomed in on the overlapping region obtained from QTL and GWAS analysis, we found that this approximately 600-kb region contains ten putative protein-encoding genes (Fig. 4e and Supplementary Fig. 14). Four genes encoding Casparian-strip membrane proteins[37] were found under/near the strongest signal in this region (−log$P$ = 18.95) (Supplementary Fig. 14a–d). A signal was located upstream of a putative B-type cyclin that has been reported to be involved in trichome or fiber development[38–40] (Supplementary Fig. 14f).

G. arboreum has an important role in the history of Chinese cotton cultivation[4]. The present study shows that the Chinese *G. arboreum* population exhibits distinct geographic patterns that are consistent with its introduction from SC to the YZR and the YER. Several phenotypes such as yield and disease-resistance traits changed substantially during the migration of cotton from SC to the YZR and further to the YER, thus suggesting positive inputs from local environments as well as human selection. The geographically selected genomic regions and overlapped QTLs detected in this study via pairwise comparisons of different germplasm groups represent an important high-resolution genetic resource that should greatly facilitate the improvement of complex cotton traits. Additionally, we identified a gene (*GaKASIII*) that may control fatty acid chain elongation and oil content, and we found that two typical promoter haplotypes of *GaGSTF9* are related to FOV resistance. Moreover, combined GWAS and QTL-seq identified a possible functional roles for Casparian-strip membrane proteins during fuzz cell development. Our study indicates that geographic isolation has affected the genetic basis of SC, YZR, and YER populations, and has also influenced the development and distribution of disease resistance and yield traits of *G. arboreum* in China.

URLs

RepeatMasker, http://www.repeatmasker.org/; PASA, http://pasapipeline.github.io/; PopGen package, https://metacpan.org/pod/Bio::PopGen::IO/.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0116-x.

Received: 25 October 2016; Accepted: 15 March 2018;
Published online: 7 May 2018

References

1. Wendel, J. F., Flagel, L. E. & Adams, K. L. Jeans, genes, and genomes: cotton as a model for studying polyploidy. in *Polyploidy and Genome Evolution* (eds. Soltis, P. S. & Soltis, D. E.) 181–207 (Springer, Berlin and Heidelberg, 2012).
2. Wendel, J. F., Brubaker, C. L. & Seelanan, T. The origin and evolution of *Gossypium*. in *Physiology of Cotton* (eds. Stewart, J. M. et al.) 1–18 (Springer Netherlands, Houten, the Netherlands, 2010).
3. Watt, G. *The Wild and Cultivated Cotton Plants of the World* (Longmans, London, 1907).
4. Institute of Cotton Research, CAAS & Institute of Industrial Crops, JAAS. *The Chinese Asiatic Cottons* (ChinaAgriculture Press, Beijing, 1989).
5. Desai, A., Chee, P. W., Rong, J., May, O. L. & Paterson, A. H. Chromosome structural changes in diploid and tetraploid A genomes of *Gossypium*. *Genome* **49**, 336–345 (2006).
6. Ma, X. X., Zhou, B. L., Lü, Y. H., Guo, W. Z. & Zhang, T. Z. Simple sequence repeat genetic linkage maps of A-genome diploid cotton (*Gossypium arboreum*). *J. Integr. Plant Biol.* **50**, 491–502 (2008).
7. Stanton, M. A., Stewart, J. M., Percival, A. E. & Wendel, J. F. Morphological diversity and relationships in the A-genome cottons, *Gossypium arboreum* and *G. herbaceum*. *Crop Sci.* **34**, 519–527 (1994).
8. Chen, Y. et al. A new synthetic amphiploid (AADDAA) between *Gossypium hirsutum* and *G. arboreum* lays the foundation for transferring resistances to *Verticillium* and drought. *PLoS One* **10**, e0128981 (2015).
9. Kulkarni, V. N., Khadi, B. M., Maralappanavar, M. S., Deshapande, L. A. & Narayanan, S. S. The worldwide gene pools of *Gossypium arboreum* L. and *G. herbaceum* L. and their improvement. in *Genetics and Genomics of Cotton* (ed. Paterson, A. H.) 69–97 (Springer, New York, 2009).
10. Wang, K. et al. The draft genome of a diploid cotton *Gossypium raimondii*. *Nat. Genet.* **44**, 1098–1103 (2012).

11. Paterson, A. H. et al. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* **492**, 423–427 (2012).
12. Li, F. et al. Genome sequence of the cultivated cotton *Gossypium arboreum*. *Nat. Genet.* **46**, 567–572 (2014).
13. 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).
14. 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).
15. Liu, X. et al. *Gossypium barbadense* genome sequence provides insight into the evolution of extra-long staple fiber and specialized metabolites. *Sci. Rep.* **5**, 14139 (2015).
16. Yuan, D. et al. The genome sequence of Sea-Island cotton (*Gossypium barbadense*) provides insights into the allopolyploidization and development of superior spinnable fibres. *Sci. Rep.* **5**, 17662 (2015).
17. Huang, X. et al. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961–967 (2010).
18. Huang, X. et al. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.* **44**, 32–39 (2011).
19. Huang, X. et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501 (2012).
20. Hufford, M. B. et al. Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44**, 808–811 (2012).
21. Chia, J. M. et al. Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.* **44**, 803–807 (2012).
22. 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).
23. Jia, G. et al. A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (*Setaria italica*). *Nat. Genet.* **45**, 957–961 (2013).
24. Qi, J. et al. A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nat. Genet.* **45**, 1510–1515 (2013).
25. Lin, T. et al. Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet.* **46**, 1220–1226 (2014).
26. Wang, M. et al. Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nat. Genet.* **49**, 579–587 (2017).
27. Fang, L. et al. Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. *Nat. Genet.* **49**, 1089–1098 (2017).
28. Wendel, J. F., Olson, P. D. & Stewart, J. M. Genetic diversity, introgression, and independent domestication of old world cultivated cottons. *Am. J. Bot.* **76**, 1795–1806 (1989).
29. Guo, W., Zhou, B. L., Yang, L. M., Wang, W. & Zhang, T. Z. Genetic diversity of landraces in *Gossypium arboreum* L. race sinense assessed with simple sequence repeat markers. *J. Integr. Plant Biol.* **48**, 1008–1017 (2006).
30. Olsen, K. M. & Wendel, J. F. A bountiful harvest: genomic insights into crop domestication phenotypes. *Annu. Rev. Plant Biol.* **64**, 47–70 (2013).
31. Liu, Q., Singh, S. P. & Green, A. G. High-stearic and high-oleic cottonseed oils produced by hairpin RNA-mediated post-transcriptional gene silencing. *Plant Physiol.* **129**, 1732–1743 (2002).
32. Yu, N., Xiao, W. F., Zhu, J., Chen, X. Y. & Peng, C. C. The *Jatropha curcas KASIII* gene alters fatty acid composition of seeds in *Arabidopsis thaliana*. *Biol. Plant.* **59**, 773–782 (2015).
33. Turley, R. B. & Chapman, K. D. Ontogeny of cotton seeds: gametogenesis, embryogenesis, germination, and seedling growth. in *Cotton Physiology* (eds. Stewart, J. M. et al.) 332–341 (Springer Netherlands, Houten, the Netherlands, 2010).
34. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
35. Oerke, E. C. Crop losses to pests. *J. Agric. Sci.* **144**, 31–43 (2005).
36. Edwards, R., Dixon, D. P. & Walbot, V. Plant glutathione S-transferases: enzymes with multiple functions in sickness and in health. *Trends Plant Sci.* **5**, 193–198 (2000).
37. Roppolo, D. et al. A novel protein family mediates Casparian strip formation in the endodermis. *Nature* **473**, 380–383 (2011).
38. Roppolo, D. et al. Functional and evolutionary analysis of the CASPARIAN STRIP MEMBRANE DOMAIN PROTEIN family. *Plant Physiol.* **165**, 1709–1722 (2014).
39. Schnittger, A., Schöbinger, U., Stierhof, Y. D. & Hülskamp, M. Ectopic B-type cyclin expression induces mitotic cycles in endoreduplicating *Arabidopsis* trichomes. *Curr. Biol.* **12**, 415–420 (2002).
40. Yang, C. et al. A regulatory gene induces trichome formation and embryo lethality in tomato. *Proc. Natl Acad. Sci. USA* **108**, 11836–11841 (2011).

Acknowledgements

This work was supported by funding from the National Natural Science Foundation of China (grants 31621005 to F. Li and 90717009 to Y.Z.), the National Key Technology R&D Program, the Ministry of Science and Technology (2016YFD0100203 to X.D. and 2016YFD0100306 to S. He), the National Science and Technology Support Program, the Ministry of Agriculture (2013BAD01B03 to X.D.), the Agricultural Science and Technology Innovation Program of the Chinese Academy of Agricultural Sciences (CAAS-ASTIP-IVFCAAS to S. Huang), and the leading talents of Guangdong Province Program (00201515 to S. Huang).

Author contributions

F. Li, Y.Z., X.D., and T.L. conceived and designed the research. F. Li and S. Huang managed the project. T.L., N.L., M.L., F. Liu, F.W., H. Zheng., and G.S. performed the genome sequencing, assembly, and bioinformatics. X.D., S. He, J.S., Z.Y., X.M., X.Z., Y.J., Z. Pan., W.G., Z.L., H. Zhu., L.M., D.Y., Q.G., Z. Peng., L.W., S.X., and X.W. prepared the samples, performed phenotyping, and contributed to data analysis. Y.Z. designed the molecular experiments, and Z.Y. and G.H. performed the molecular experiments and led interpretation of the molecular-data analysis. S. He, Z.Y., and G.H. prepared the figures and tables. Y.Z., S. He, G.H., Z.Y., T.L., S. Huang, H.S., C.L., and W.F. wrote and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0116-x.

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

Correspondence and requests for materials should be addressed to Y.Z., F.L. or T.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Genome sequencing and assembly. The same cultivated diploid cotton *G. arboreum* (cultivar Shixiya1, SXY1)[12] was used for sequencing and assembly. A total of ~142 Gb of raw data was obtained from 125 SMRT cells on a PacBio RSII instrument. Hi-C experiments were performed as previously reported[41]. De novo assembly of the PacBio reads was carried out with two assemblers: the Canu pipeline[42] and Falcon[43], with different parameters to achieve a higher consistency and longer continuity. We used Quiver to polish base-calling of contigs. The PacBio contigs were further clustered and extended into pseudochromosomes by using Hi-C data. The gaps in the pseudochromosomes were filled in Pbjelly, and a second round of polishing was performed in Quiver. llumina reads were used to correct base-calling. Syntenic-block analysis was performed as described previously[12].

Transposable element (TEs) annotation. Both homolog-based and de novo strategies were applied to identify repetitive sequences in the *G. arboreum* genome. De novo prediction software, including RepeatScout[44], LTR-FINDER[45], MITE[46], and PILER[47], was used to identify repeats within the genome. These results were then combined and merged in Repbase to form the *G. arboreum* repetitive sequence database, which was further classified into various categories with REPET[48]. The resulting repetitive sequences in the genome were identified by homolog searching in that database through RepeatMasker (see URLs).

Gene-model prediction, evaluation and annotation. We combined three different gene-model prediction methods in the present study. In the homolog-based gene-prediction model, we used geMoMa[49] to predict gene structures with homologous proteins obtained from NCBI. For de novo prediction, we used Augustus[50] with parameters trained by unigenes, by using transcriptome data obtained from pooled cotton tissues. In the transcriptome-based prediction, unigenes were first aligned to the genome assembly and were then filtered with PASA (see URLs). All predicted gene structures were integrated into a consensus set with EVidenceModeler (EVM)[51]. Genes were then annotated according to homologous alignments with BLAST[52] (E value $\leq 1\times10^{-5}$) against several databases including the nr[53] and nt databases of NCBI, Swiss-Prot, and TrEMBL. We further used InterProScan (v4.3)[54] to predict domain information and gene ontologies (GO terms)[55]. KAAS was used for KEGG pathway annotation[56].

Sampling and sequencing. A total of 243 cotton accessions, including 230 *G. arboreum* and 13 *G. herbaceum* accessions (Supplementary Table 7), were selected from the Chinese National Germplasm Mid-term Genebank (Anyang, China). Plants were grown in the greenhouses at the Institute of Cotton Research of the Chinese Academy of Agricultural Science (ICR, CAAS). Fresh young leaves were collected from single individuals of each accession and were immediately frozen in liquid nitrogen. Genomic DNA was extracted with a previously reported workflow[57]. At least 5 μg of genomic DNA for each accession was used to build paired-end-sequencing libraries with insert sizes of approximately 500 bp, according to vendor-provided instructions (Illumina). An average 6× coverage of the assembled genome, with 125-bp paired-end reads for each accession, was generated with the Illumina HiSeq 2500 platform. For the QTL-seq analysis[58], we sequenced two parent lines (GA0146 and GA0149) at 20× depth and two bulk populations (selected from an F_2 population and containing 20 progenies each for fuzzy and fuzzless phenotypes) at 30× depth.

SNP index and ΔSNP index. The SNP index was calculated for both fuzzy and fuzzless bulk samples expressing the proportion of reads containing SNPs that were identical to those in the fuzzy parent (GA0146). The ΔSNP index was calculated as (SNP index of fuzzy bulk) – (SNP index of fuzzless bulk). The average ΔSNP index was calculated with a 100-kb sliding window with a step size of 10 kb and was used to plot the ΔSNP index distributions in Fig. 4e. Statistical 99% confidence intervals of the ΔSNP index were calculated under the null hypothesis (no QTL)[58].

Sequence alignment, variation calling, and annotation. All the sequence reads for each accession were mapped to the newly updated genome (all unanchored contigs were connected by 1 kb 'N' sequence like contig1 + 1 kb length 'N' + contig2 + 1 kb length 'N' + contig3 + … and defined as chromosome 14) in the Burrows–Wheeler Aligner program (BWA, ver. 0.7.10)[59] with default parameters. We sorted the alignments according to mapping coordinates in Picard (ver. 1.118). After removing the reads with low mapping quality (MQ < 20), both paired-end and single-end mapped reads were used for SNP detection throughout the entire collection of cotton accessions in the GATK toolkit (ver. 3.2-2)[60]. Mapped reads were filtered by removal of PCR duplicates. First, the MarkDuplicates module was used to mark the duplication alignment; SNPs and indels identified by the HaplotypeCaller module were then used to perform base-quality recalibration with the BaseRecalibrator and IndelRealigner modules, respectively. Second, the genomic variants, in GVCF format for each accession, were identified with the HaplotypeCaller module and the GVCF model. Finally, after all of the GVCF files were merged, a raw population genotype file with the SNPs and indels was created in the HaplotypeCaller module and was filtered with the following parameters: 'QD < 2.0 || MQ < 40.0 || FS > 60.0 || MQRankSum < −12.5 || ReadPosRankSum < −8.0 −clusterSize 3 −clusterWindowSize 10' and 'QD < 2.0 || FS > 200.0 || ReadPosRankSum < −20.0'. The identified SNPs and indels were further annotated with ANNOVAR tool software[61] and were divided into groupings of variations occurring in intergenic regions, coding sequences, and introns, on the basis of newly updated *G. arboreum* genome annotation information.

Phylogenetic analysis and population-structure study. A subset of 72,419 SNPs (SNP quality > 2,000, minor allele frequency (MAF) > 0.05, and missing data < 20%) in the 243 cotton accessions from the entire SNP dataset was screened to build a neighbor-joining tree in PHYLIP (version 3.695)[62] with 100 bootstrap replicates. STRUCTURE software (version 2.3.1)[63] was used to infer the cotton population structure. The program was run on the subset of SNPs to estimate the group membership of each accession by using 10,000 iterations with K values from 2 to 4.

Linkage disequilibrium analysis. Haploview 4.20 (ref. [64]) software was used to calculate LD values for the *G. arboreum* and *G. herbaceum* accessions on the basis of SNPs (MAF > 0.05). The detailed parameters were as follows: -n -pedfile -info -log -minMAF 0.05 -hwcutoff 0 -dprime -memory 2096. LD decay was measured on the basis of the r^2 value and the corresponding distance between two given SNPs.

Population genetics analysis. Nucleotide diversity (π) analysis was applied to estimate the degree of variability within each group (SC, YZR, and YER), and the fixation statistic F_{ST} was applied to explain population differentiation on the basis of the variance of allele frequencies between two different groups. Both π and F_{ST} were calculated in the PopGen package (see URLs) of BioPerl (ver. 1.6.923). After filtering of SNPs with quality < 2,000, π values for the SC, YZR, and YER groups were calculated individually. F_{ST} values were initially calculated for each SNP through a variance component approach, and then the average F_{ST} of all SNPs in each 100-kb window was used as the value at the whole-genome level across different groups. Sliding windows with the top 5% of F_{ST} values for each comparison were selected as candidate highly divergent regions for further analysis. Adjacent windows were merged into a single divergent region.

Phenotyping. For phenotypic evaluations, we selected 215 of 230 resequenced accessions that displayed reliable phenotypes and planted them in Anyang (Henan province), Sanya (Hainan province), and Akesu (Xinjiang province) in 2014. Several traits were investigated in only one or two locations, owing to resource limitations. For drought-tolerance evaluation, the seedlings were watered every 3 d for a total of three weeks, and then water was withheld from 3-week-old seedlings. When the drought-sensitive accessions exhibited severe leaf-wilt symptoms, all of the plants were rewatered. Four days after the rewatering, the numbers of surviving plants of each accession were counted. Three replicates were performed at each location.

The FWDI evaluation followed the Chinese technical specifications for evaluating cotton diseases and pests (GB/T 22101.4-2009). The *F. oxysporum* strain Ag149 was inoculated in soil. The disease-susceptible Jimian-11 line and the highly disease-tolerant Zhongzhimian-2 line were used as maxima to calculate the disease index. The molecular detection of *F. oxysporum* DNA (fungal DNA) in cotton leaves was performed according to a previously described method[65].

The fatty acid composition of the lipid content of cotton seeds and ovules was evaluated according to previously described procedures[66].

GWAS. A total of 1,425,003 high-quality SNPs (MAF > 0.05, missing rate < 20%) in 215 *G. arboreum* accessions were used to perform GWAS for 20 traits in efficient mixed-model association expedited (EMMAX) software[67], which was designed to handle large-dataset analysis[68]. GWAS for several traits were conducted in multiple locations with different ecological environments, including Anyang (N36.02, E114.50°, altitude, 63 m), Sanya (N18.35°, E109.33°, altitude, 11 m), and Akesu (N41.11°, E80.54, altitude, 1,107 m). Population stratification and hidden relatedness were modeled with a kinship (K) matrix in the emmax-kin-intel package of EMMAX. The genome-wide significance thresholds of all tested traits were evaluated with the formula $P = 0.05/n$ (where n is the effective number of independent SNPs)[69]. The P-value thresholds for significance in the *G. arboreum* population were approximately 1.0×10^{-6}.

Ancestral-allele and phylogenetic analysis among species. To identify orthologous alleles, we first set up a saturation curve using different lengths of flanking sequences with 100 randomly picked alleles. We observed a plateau with a false-positive rate of 13.5% when the sequence length reached 901 bp (including the particular SNP) (Supplementary Fig. 15). We initially extracted each SNP (~18 million SNP set) and its flanking sequences (450-bp length in both directions) from *G. arboreum* genome. We then used the BLAST[52] algorithm (E value $<1\times10^{-10}$) to identify orthologous sequences in the *G. raimondii* genome. Only the top hit from the BLAST results was retained. A total of 4,487,496 corresponding hits were found in the *G. raimondii* genome. Those SNPs in each accession that were identical to SNPs in *G. raimondii* were defined as ancestral alleles. The ancestral-allele percentage was the average value of all *G. arboreum* and *G. herbaceum* accessions, respectively. To assess the phylogenetic relationships among *G. raimondii*, *G. arboretum*, and *G. herbaceum*, we used SNPs of 68,830 sites

that were present in all three species to construct a phylogenetic tree in FastTree[70] with default parameters.

Functional characterization of *GaGSTF9*. To analyze the expression pattern of *GaGSTF9*, samples were harvested at various incubation time points. Total RNA (~2 μg) was extracted and was then reverse transcribed in a 20-μl reaction mixture with EasyScript cDNA Synthesis SuperMix (TRANSGEN Biotech). Then 1-μl sample aliquots were used as templates for qRT–PCR analysis. Three technical replicates per sample and three biological-replicate samples were analyzed for each experiment. *Histone3* (*LOC108467150*) was used as the internal control for qRT–PCR data analysis. For virus-induced gene silencing, a 398-bp fragment from *GaGSTF9* was cloned into the XbaI and SacI sites of the pTRV-RNA2 vector. Glutathione *S*-transferase activity was measured with a Glutathione *S*-transferase (GSH-ST) assay kit (Jiancheng). All primers used in this study are presented in Supplementary Table 18.

Transcriptome data. We used Tophat and Cufflinks for the RNA-seq expression analysis[71]. For Fig. 2f (seed oil content), we downloaded the data from NCBI (SRX204555–SRX204558).

Statistical analyses. Student's two-tailed *t* tests and one-way ANOVA test were performed in GraphPad Prism software.

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

Data availability. The updated *G. arboreum* genome assembly data are accessible through the NCBI under accession PRJNA382310. All raw sequencing data for the 243 accessions have been deposited at in the NCBI BioProject database under accession number PRJNA349094. Supporting data (updated genome, raw SNP sets, input files for structure and nucleotide diversity, list of orthologous loci and phenotype data) can be downloaded from ftp://bioinfo.ayit.edu.cn/downloads/.

References

41. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
42. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
43. Chin, C. S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
44. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21** (Suppl. 1), i351–i358 (2005).
45. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
46. Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38**, e199 (2010).
47. Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21** (Suppl. 1), i152–i158 (2005).
48. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
49. Keilwagen, J. et al. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* **44**, e89 (2016).
50. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
51. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
52. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
53. Marchler-Bauer, A. et al. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **39**, D225–D229 (2011).
54. Hunter, S. et al. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* **40**, D306–D312 (2012).
55. Dimmer, E. C. et al. The UniProt-GO Annotation database in 2011. *Nucleic Acids Res.* **40**, D565–D570 (2012).
56. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
57. Paterson, A. H., Brubaker, C. L. & Wendel, J. F. A rapid method for extraction of cotton (*Gossypium* spp.) genomic DNA suitable for RFLP or PCR analysis. *Plant Mol. Biol. Rep.* **11**, 122–127 (1993).
58. Takagi, H. et al. QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *Plant J.* **74**, 174–183 (2013).
59. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
60. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
61. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
62. Felsenstein, J. PHYLIP-phylogeny inference package (version 3.2). *Cladistics* **5**, 163–166 (1989).
63. 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).
64. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
65. Haegi, A. et al. A newly developed real-time PCR assay for detection and quantification of *Fusarium oxysporum* and its use in compatible and incompatible interactions with grafted melon genotypes. *Phytopathology* **103**, 802–810 (2013).
66. Dowd, M. K. et al. Fatty acid profiles of cottonseed genotypes from the national cotton variety trials. *J. Cotton Sci.* **14**, 64–73 (2010).
67. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
68. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).
69. Li, M. X., Yeung, J. M. Y., Cherny, S. S. & Sham, P. C. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.* **131**, 747–756 (2012).
70. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
71. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).

nature research

Corresponding author(s): Yuxian Zhu

Reporting Summary

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

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

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

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	We used open source software and codes for data collection
Data analysis	All software used in the study are publicly available from the Internet and the corresponding versions were described in detail on the section of Online Methods.

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

Data

Policy information about availability of data

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

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

The updated G. arboreum genome assembly data are accessible through the NCBI under accession PRJNA382310. All raw sequencing data for the 243 accessions

have been deposited at the NCBI BioProject under the accession number PRJNA349094. These supporting data (raw SNP sets, input files for structure and nucleotide diversity, list of orthologous loci and phenotype data) are available from the website (ftp://bioinfo.ayit.edu.cn/downloads/).

Field-specific reporting

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

☒ Life sciences ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences

Study design

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

Sample size	No Sample size calculations
Data exclusions	No data exclusions
Replication	All attempts at replication were successful
Randomization	There was no randomization procedures
Blinding	No blinding was performed

Materials & experimental systems

Policy information about availability of materials

n/a	Involved in the study
☒	☐ Unique materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Research animals
☒	☐ Human research participants

Method-specific reporting

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ Magnetic resonance imaging