Genome sequence of the cultivated cotton Gossypium arboreum

Li, Fuguang; Fan, Guangyi; Wang, Kunbo; Sun, Fengming; Yuan, Youlu; Song, Guoli; Li, Qin; Ma, Zhiying; Lu, Cairui; Zou, Changsong; Chen, Wenbin; Liang, Xinming; Shang, Haihong; Liu, Weiqing; Shi, Chengcheng; Xiao, Guanhui; Gou, Caiyun; Ye, Wuwei; Xu, Xun; Zhang, Xueyan; Wei, Hengling; Li, Zhifang; Zhang, Guiyin; Wang, Junyi; Liu, Kun; Kohel, Russell J.; Percy, Richard G.; Yu, John Z.; Zhu, Yu-Xian; Wang, Jun; Yu, Shuxun

Published in:
Nature Genetics

DOI:
10.1038/ng.2987

Publication date:
2014

Document version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Li, F., Fan, G., Wang, K., Sun, F., Yuan, Y., Song, G., Li, Q., Ma, Z., Lu, C., Zou, C., Chen, W., Liang, X., Shang, H., Liu, W., Shi, C., Xiao, G., Gou, C., Ye, W., Xu, X., ... Yu, S. (2014). Genome sequence of the cultivated cotton Gossypium arboreum. Nature Genetics, 46(6), 567-572. https://doi.org/10.1038/ng.2987
Genome sequence of the cultivated cotton Gossypium arboreum

Fuguang Li1,11, Guangyi Fan2,11, Kunbo Wang1,11, Fengming Sun2,11, Youlu Yuan1,11, Guoli Song1,11, Qin Li3,11, Zhiying Ma4,11, Cairui Lu1, Changsong Zou1, Wenbin Chen2, Xinming Liang2, Haihong Shang1, Weiqing Liu2, Chengcheng Shi2, Guanghui Xiao3, Caiyun Gou2, Wuwei Ye1, Xun Xu2, Xueyan Zhang1, Hengling Wei1, Zhifang Li1, Guiyin Zhang4, Junyi Wang2, Kun Liu1, Russell J Kohel5, Richard G Percy6, John Z Yu5, Yu-Xian Zhu3, Jun Wang2,6–10 & Shuxun Yu1

The complex allotetraploid nature of the cotton genome (AADD; 2n = 52) makes genetic, genomic and functional analyses extremely challenging. Here we sequenced and assembled the Gossypium arboreum (AA; 2n = 26) genome, a putative contributor of the A subgenome. A total of 193.6 Gb of clean sequence covering the genome by 112.6-fold was obtained by paired-end sequencing. We further anchored and oriented 90.4% of the assembly on 13 pseudochromosomes and found that 68.5% of the genome is occupied by repetitive DNA sequences. We predicted 41,330 protein-coding genes in G. arboreum. Two whole-genome duplications were shared by G. arboreum and Gossypium raimondii before speciation. Insertions of long terminal repeats in the last 1–2 million years through hybridization and subsequent polyploidization events between the A- and D-subgenome progenitors. All tetraploid cotton species came from interspecific hybridization between the A-genome species G. arboreum (A2) and the D-genome species G. raimondii (D5). The A-genome species are cultivated, whereas the D-genome species do not produce spinnable fiber. Although G. arboreum (A2) and G. raimondii (D5) are the putative donor species for the A and D chromosome groups, respectively, tetraploid cotton species differ greatly with respect to plant morphology as well as economic characteristics, including fiber production, oil content and disease resistance. Furthermore, G. arboreum (1,746 Mb/1C) has a genome size that is almost twice that of G. raimondii (885 Mb/1C).4

In the current study, assisted by a high-resolution genetic map, we anchored and oriented 90.4% of the G. arboreum assembled scaffolds on 13 pseudochromosomes. This assembly was compared with that of G. raimondii to understand possible paths for genome evolution and species divergence. A highly homozygous single-seed descendant, derived from 18 successive generations of self-fertilization, of the cultivated diploid cotton Shixiya1 (SXY1) was used for DNA sequencing. Assembly of the G. arboreum genome and comparative studies with the genome of G. raimondii4,5 may provide new insights into the process of divergence among polyploid species5–9. We believe that the G. arboreum genome offers a diploid reference for the analysis of cotton agronomic traits, such as fiber quality10–13 and disease resistance14,15.

RESULTS
Genome sequencing and assembly
We sequenced and assembled the G. arboreum genome using the whole-genome shotgun approach. In brief, a total of 371.5 Gb of raw paired-end Illumina reads was generated by sequencing genome shotgun libraries with different fragment lengths that ranged from 180 bp to 40 kb (Supplementary Table 1). After filtering out low-quality reads, 193.6 Gb of high-quality sequence was obtained and used for the de novo assembly process (Supplementary Table 2). We next used 33,454 BAC end sequences (16,727 pairs; approximately 1-fold physical coverage) to improve the assembly. Over 90% of our BAC clones contained inserts of

1State Key Laboratory of Cotton Biology, Institute of Cotton Research of the Chinese Academy of Agricultural Sciences, Anyang, China. 2BGI-Shenzhen, Shenzhen, China. 3State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences, Peking University, Beijing, China. 4Key Laboratory for Crop Germplasm Resources of Hebei, Agricultural University of Hebei, Baoding, China. 5Crop Germplasm Research Unit, Southern Plains Agricultural Research Center, US Department of Agriculture–Agricultural Research Service (USDA-ARS), College Station, Texas, USA. 6Department of Biotechnology, University of Copenhagen, Copenhagen, Denmark. 7King Abdulaziz University, Jeddah, Saudi Arabia. 8Macau University of Science and Technology, Macau, China. 9Department of Medicine, University of Hong Kong, Hong Kong. 10State Key Laboratory of Pharmaceutical Biotechnology, University of Hong Kong, Hong Kong. 11These authors contributed equally to this work. Correspondence should be addressed to Y.-X.Z. (zhuyx2@pku.edu.cn), S.Y. (yu@cricaas.com.cn) or Jun Wang (wangj@genomics.org.cn).

Received 17 June 2013; accepted 24 April 2014; published online 18 May 2014; doi:10.1038/ng.2987

© 2014 Nature America, Inc. All rights reserved.
Table 1 Global statistics of *G. arboreum* genome assembly and annotation

| Category                  | Number | N50 (kb) | Longest (Mb) | Size (Mb) | Percent of the assembly |
|---------------------------|--------|----------|--------------|-----------|-------------------------|
| Total contigs             | 40,381 | 72.0     | 0.8          | 1,561     | NA                      |
| Total scaffolds           | 7,914  | 665.8    | 5.9          | 1,694     | 100                     |
| Anchored and oriented     | 3,740  | 790      | 5.9          | 1,532     | 90.4                    |
| scaffolds                 |        |          |              |           |                         |
| Genes annotated           | 41,330 | 105      |              |           | 6.2                     |
| miRNAs                   | 431    | 0.05     |              |           | <0.01                   |
| rRNAs                    | 10,464 | 1.2      |              |           | 0.07                    |
| tRNAs                    | 2,289  | 0.2      |              |           | 0.01                    |
| snRNAs                   | 7,619  | 0.8      |              |           | 0.04                    |
| Repeat sequences          | NA     | 1,160    |              |           | 68.5                    |

NA, not applicable or not analyzed.

≥50,000 bp in length (Supplementary Fig. 1). In-depth analysis showed that all scaffolds were supported by multiple (>80) paired-end linkages (Supplementary Fig. 2). The number of scaffolds decreased in libraries with increasing insert size (>2,000 bp). Using K-mer distribution analysis, the genome size was estimated to be 1,724 Mb (Supplementary Fig. 3), which is consistent with a previous report.

Our final assembly, performed with SOAPdenovo, showed that the *G. arboreum* genome is 1,694 Mb in total length (Table 1). Ninety percent of the assembly fell into 2,724 superscaffolds that were >148 kb in length, with the largest scaffold being 5.9 Mb. N50 (the size above which 50% of the total length of the sequence assembly can be found) for the contigs and scaffolds was 72 kb and 666 kb, respectively. Data quality was assessed by comparison with Sanger sequencing–derived *G. arboreum* sequences and by mapping available ESTs to the assembly. Nineteen of the twenty completely sequenced *G. arboreum* BAC clones available from GenBank were recovered in our assembly, with >98% sequence identity (Supplementary Table 3). Of the 55,894 Sanger sequencing–derived *G. arboreum* ESTs (>200 bp in length) available from NCBI, 96.37% were detected in our assembly (Supplementary Table 4). Using a restriction site–associated DNA (RAD) linkage map that we constructed during the current study based on 24,569 codominant SNP markers, 1,532 Mb or 90.4% of the assembled genome was anchored and oriented on 13 pseudochromosomes (Fig. 1, Supplementary Fig. 4 and Supplementary Table 5), which contained 97.1% of the predicted gene models. Collinear relationships existed between the SNP markers and most of the assembled *G. arboreum* pseudochromosomes, which indicates that our assembly is of high quality (Supplementary Fig. 5). Collinearity was also observed between our *G. arboreum* assembly and the tetraploid genetic map reported previously (Supplementary Fig. 6 and Supplementary Table 6). The GC content of the *G. arboreum* genome was comparable to those for the genomes of *Arabidopsis thaliana* and *Theobroma cacao* and *Arabidopsis thaliana* (Fig. 1e).

Figure 1 Characterization of the *G. arboreum* cotton genome.
(a) Percent coverage of TEs in non-overlapping windows (window size = 500 kb). Outer tick marks show the calculated lengths of 13 *G. arboreum* pseudochromosomes. (b) Gene density estimated on the basis of the number of genes in non-overlapping 500-kb windows. (c) Transcription state. The transcript level for each gene was estimated by averaging values of reads per kilobase of mapped cDNA per million reads (RPKM) from different tissues in non-overlapping 500-kb windows. (d) Marker density represented by the number of SNPs in non-overlapping 500-kb windows. (e) GC content estimated on the basis of the percentage of G+C nucleotides in 500-kb non-overlapping windows.

Genome annotation

We performed genome annotation by combining results obtained from *ab initio* prediction, homology search and transcriptome alignment. As much as 68.5% of the *G. arboreum* genome was composed of various types of repeat sequences (Fig. 1a, Table 1 and Supplementary Table 7). We believe that this genome has the greatest amount of repeat-containing sequences among sequenced eudicots. Long terminal repeat (LTR) retrotransposons accounted for 95.12% of all repeat sequences. In comparison with the *G. raimondii* genome, the *G. arboreum* genome had noticeable proliferation of *Gorge* elements, whereas LTR *Copia* elements tended to accumulate in the *G. raimondii* genome (Supplementary Table 8), as has been suggested previously. LTR retrotransposons in *G. arboreum* appeared to have inserted randomly along each chromosome (Fig. 1a), in a pattern substantially different from the ones observed for their insertion sites in soybean and potato, in which LTRs are clustered near the centromere and are found less frequently near telomeres.

A total of 41,330 protein-coding genes were identified in the *G. arboreum* genome, with an average transcript size of 2,533 bp (as determined by GLEAN) and a mean of 4.6 exons per gene (Fig. 1b, Table 1 and Supplementary Table 9). The genome encoded 431 micro-RNAs (miRNAs), 10,464 rRNAs, 2,289 tRNAs and 7,619 small nuclear RNAs (snRNAs) (Table 1). Among the annotated genes, 85.64% encoded proteins that showed homology to proteins in the TrEMBL database, and 68.71% were identified in InterPro (Supplementary Table 10). Over 96% of predicted coding sequences were supported by transcriptome sequencing data (Fig. 1c and Supplementary Table 9), which indicated high accuracy of *G. arboreum* gene predictions from the genome sequence. Orthologous clustering of the *G. arboreum* proteome with 3 closely related plant genomes identified 11,699 gene families in common, with 739 gene families that were present specifically in *G. arboreum* (Supplementary Fig. 7).
Figure 2 Phylogenetic and evolutionary analysis of the Gossypium genomes. (a) Phylogenetic analysis indicated that G. arboresum and G. raimondii separated ~5 million years ago (mya). (b) Fourfold-degenerate analyses suggested that the Gossypium genomes might have undergone two WGD events and one speciation event. (c) Large numbers of syntenic blocks were found when comparing either the G. raimondii (Gr) or G. arboresum (Ga) genome with the genome of T. cacao (Tc). Numbered rectangles represent the chromosomes.

Genome evolution
The current descendants of the two diploid cotton species that hybridized to create all cultivated allopolyploid cotton species, G. arboresum (A2) and G. raimondii (D2), exhibit a twofold difference in genome size. Molecular phylogenetic analyses suggested a divergence time for G. arboresum and G. raimondii of 2–13 million years ago, with their common ancestor having diverged from T. cacao 18–58 million years ago (Fig. 2a).

We scanned the G. arboresum genome for syntenic gene blocks identified on the basis of protein-coding sequence similarity. An age distribution was calculated for all duplicate gene pairs on the basis of the number of transversions at fourfold-degenerate sites. Using 1,917 paralogous gene pairs of similar age and excluding tandem or local duplications, a large peak centered around 0.17 synonymous transversions per site and a second peak at about 0.54 synonymous transversions per site were observed in both the G. arboresum and G. raimondii genomes, indicating that a recent and an ancient whole-genome duplication (WGD) event occurred in cotton (Fig. 2b). Both of these WGD events in G. arboresum coincided with those in G. raimondii, which were estimated to have occurred at 13–20 and 115–146 million years ago. The ancient duplication event corresponds to the ancient hexaploidization event that is shared among the eudicots. Using a whole-genome alignment approach, we found that both the G. raimondii and G. arboresum genomes showed close collinear relationships with the genome of T. cacao. There were 209 and 295 collinear blocks covering 82% and 66% of the G. arboresum and G. raimondii genomes, respectively (Fig. 2c and Supplementary Table 11). When the G. raimondii assembly reported by Paterson et al.6 was used for the analysis, a slightly lower number of collinear blocks with higher genome coverage were identified (Supplementary Table 11). Approximately 50% of the T. cacao genome was aligned with two segments in each of the cotton genomes (Fig. 2c), undoubtedly suggesting the existence of a Gossypium-specific WGD event after speciation from the T. cacao lineage.

Orthology maps and retrotransposition analysis
Whole-genome alignment also identified 68,863 orthologous hits (anchors) on the 2 cotton genomes, including 33,229 genes for G. arboresum and 34,204 genes for G. raimondii. These anchors spanned 17.1% of the A genome and 35.0% of the D genome (Supplementary Table 12). When the G. raimondii assembly by Paterson et al.6 was used for the analysis, we observed a total of 63,810 anchors that spanned 23% of the A genome and 48% of the D genome, indicating an improvement in data quality for the later assembly in this respect (Supplementary Table 12). Using these anchors, we found that the 2 cotton genomes shared 780 syntenic blocks that covered 73% and 88% of the assembled chromosomes in G. arboresum and G. raimondii, respectively (Supplementary Table 13). Chromosomes 1, 4–6 and 9–13 were highly collinear in the two cotton species, whereas large-scale rearrangements were observed on chromosomes 2 and 3 of G. raimondii and deletions and insertions were observed on chromosomes 7 and 8 of G. arboresum (Fig. 3a,b).

Expansions of transposable element (TE) families have been reported in the genome of G. raimondii. In G. arboresum, two major clusters of retrotransposition activity were found at 0–0.5 and 3.5–4.5 million years ago, with two additional minor sets of retrotransposition activity present around 1.0 and 7.0–8.0 million years ago (Fig. 3c). As a result, the difference in LTR numbers between these two cotton species showed two separate peaks at ~0–0.5 and ~3.5 million years ago (Fig. 3c). Data analysis indicated that the G. arboresum genome tended to harbor more LTRs inserted during the last 0.5 million years, whereas extremely low LTR activity was observed in the G. raimondii genome over the same period of time. Thus, LTR activities substantially contributed to the twofold increase in the size of the G. arboresum genome, as previously suggested. We confirmed our assertion by analyzing microsyntenic blocks on chromosome 7 of both G. arboresum and G. raimondii that were 3.5 and 1.5 Mb in length, respectively (Fig. 3d and Supplementary Fig. 8). The gene loci in these syntenic blocks were highly collinear, with 191 loci in G. raimondii and 216 loci in G. arboresum. However, 4,098 TEs were identified in G. arboresum, whereas only 1,542 TEs were found in G. raimondii (Fig. 3d). Close to 58% of the TEs reported in Figure 3d were Gorge elements in G. arboresum, whereas these elements represented only ~21% of TEs in G. raimondii (Supplementary Table 14).

Analysis of potential Verticillium wilt resistance genes
Verticillium wilt is a widespread and destructive cotton disease caused by the soil-borne fungus pathogen V. dahliae. G. raimondii is nearly immune to the pathogen, whereas G. arboresum is easily susceptible. Genes related to disease resistance (R genes) have critical roles during various stages of disease development in plants. Thus far, most of the cloned R genes encode NBS domains. We compared the G. arboresum genome to the G. raimondii and T. cacao genome sequences to identify differences in the expression patterns of NBS domain–encoding genes. We identified 391, 280 and 302 such genes in the genomes of G. raimondii, G. arboresum and T. cacao, respectively.

ARTICLES
Ethylene is a key modulator of cotton fiber cell growth

Ethylene is an important signaling molecule that promotes cotton fiber elongation in *G. hirsutum*\(^{10,11,12}\). Products of the 1-aminocyclopropane-1-carboxylic acid oxidase (*ACO*) gene are involved in the last rate-limiting step in ethylene biosynthesis during cotton fiber development\(^{10,11}\). In *G. raimondii* ovules collected at 3 days post-anthesis (dpa), the levels of *ACO1* transcripts were ~1,000-fold higher and the levels of *ACO3* transcripts were ~500-fold higher than in *G. arboresum* (Fig. 5a). These two genes were chosen because they accounted for >90% of all *ACO* transcripts during early fiber development (Supplementary Table 16). *GrACO1* and *GrACO3* showed 98% to 99% sequence identity in coding regions, respectively, with their counterparts in the *G. arboresum* genome (data not shown). Dot plots of promoter regions showed that a deletion of ~130 bp beginning at ~470 bp relative to the transcription start site of *GaACO1* resulted in loss of a putative MYB-binding site (Fig. 5b). The pivotal role of MYB transcription factors in regulating cotton fiber development and

---

**Figure 4 Analysis of NBS-encoding genes and potential molecular targets for resistance to *Verticillium* wilt.** (a) Analysis of gene numbers in all NBS-encoding disease resistance gene families in *G. arboresum* (Ga), *G. raimondii* (Gr) and *T. cacao* (Tc). (b–d) Analysis of gene numbers in the TN subfamily (b), the TNL subfamily (c) and the N subfamily (d). Numbers in circles represent the number of family members in each genome, and gene numbers for the MRCA are shown in rectangles. Numbers with plus and minus signs indicate the numbers of duplicated and deleted genes, respectively. (e) Example of tandem repeat and gene deletion of several TN subfamily members in *G. arboresum, G. raimondii* and *T. cacao*. Solid lines link orthologous genes on different chromosomes, and dashed lines represent inferred orthologous relationships plus putative deletions. Sections of *G. arboresum* chromosome 3 (Ga03) and *G. raimondii* chromosome 7 (Gr07) are enlarged for clarity, whereas all other pseudochromosomes are shown according to the scale bar. (f) qRT-PCR analysis of induced gene expression in the TN and TNL subfamilies upon *V. dahliae* infection. hpi, hours post-infection.
secondary cell wall biosynthesis has previously been proposed\(^{33,34}\). Although GaACO3 and GrACO3 share one common MYB-binding site ~150 bp upstream of the transcriptional start site, two small segments of mismatch located around ~750 bp to ~800 bp led to loss of two additional MYB-binding sites in G. arboreum (Fig. 5c). Very high levels of ACO transcripts in G. raimondii ovules in conjunction with an ethylene burst\(^{5,10}\) might force an early fiber senescence phenotype, whereas the inactivation of ACO gene transcription in G. arboreum ovules might be responsible for the short-fiber phenotype in this species (Fig. 5d). When all ACO transcripts were quantified from different cotton species (Fig. 5e; see Supplementary Table 17 for primer sequences), we found that fiber-specific upregulation of ACO expression, as in the case of G. hirsutum, seems to be required for normal fiber growth. Inactivation or overproduction of ACO, as in G. arboreum and G. raimondii, respectively, might suppress fiber development. Further functional characterization of biosynthesis and signaling pathways is required to pin down such a fundamental role for ethylene in the regulation of the growth of cotton fiber cells.

DISCUSSION

Comparative genomic analysis showed that LTR insertions and expansions of LTR families contributed substantially to forming the double-sized G. arboreum genome relative to G. raimondii. The amount of sequence encompassing LTR-type retrotransposons increased from 348 Mb in G. raimondii\(^5\) to 1,145 Mb (or 67.6%) in G. arboreum, whereas the protein-coding capacities of these two species remained largely unchanged (Table 1 and Supplementary Table 7). Typically, data obtained from the whole-genome shotgun approach in plants with genome sizes exceeding a few hundred megabases are quite difficult to assemble satisfactorily because of their high TE content\(^{35–37}\). Here 90.4% of the 1,694-Mb G. arboreum genome, which is the largest of all sequenced eudicots\(^{19–24}\), was anchored and oriented. Problems related to high TE content were circumvented by using a high-density genetic map that involved 154 F\(_2\) RAD lines. A previous study showed an association between LTR number and ecogeography in barley, with more LTR copies present in the genomes of plants from sites with greater environmental stress\(^37\). All wild Gossypium species seem to be adapted to arid and stressful environments. The different number of LTRs in G. arboreum and G. raimondii does not support the ecogeography hypothesis. There must be other unknown mechanisms and causes for the dramatic increase in the number of LTRs in the domesticated species G. arboreum.

Qualitative differences in several NBS-encoding subfamilies and in ACO gene expression between the G. arboreum and G. raimondii genomes were observed. Our data thus identify prime targets in deciphering the molecular mechanisms that control resistance to Verticillium wilt and fiber cell development. We suggest that this G. arboreum genome will be an essential reference for the assembly of tetraploid cotton genomes and for evolutionary studies of Gossypium species. It also provides an essential tool for the identification, isolation and manipulation of important cotton genes conferring agronomic traits for molecular breeding and genetic improvement.

URLs. Cotton Genome Project (CGP), http://cgp.genomics.org.cn/; LASTZ, http://www.bx.psu.edu/miller_lab/; RepeatModeler, http://www.repeatmasker.org/RepeatModeler.html.

METHODS

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

Accession codes. The G. arboreum genome sequence and gene annotation information have been deposited at the Cotton Genome Project (CGP), available at http://cgp.genomics.org.cn/, and all raw sequencing
data are accessible through the NCBI Sequence Read Archive (SRA) under accession SRA150181.

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

ACKNOWLEDGMENTS
We thank the National Mid-term Gene Bank for Cotton at the Institute of Cotton Research, Chinese Academy of Agricultural Sciences for providing the original SX1 seeds. This work was supported by the National Science Foundation for Distinguished Young Scholars (grant 31125020), the National Natural Science Foundation of China (grant 90717009) and the 863 National High-Tech Research Development Program in China (grant 2013AA102601).

AUTHOR CONTRIBUTIONS
E.L., K.W., G.F., F.S., Y.Y., G.S., Jun Wang and S.Y. designed the analyses. G.F., F.S., X.X., W.C., X.L., C.G., W.L., C.S. and Junyi Wang performed sequencing, assembly and genome annotation. E.L., K.W., Y.Y., G.S., C.L., C.Z., H.S. and W.Y. managed the project. Q.L., C.L., C.Z. and Y.-X.Z. performed genome analysis and physical map integration. C.L., Q.L., C.Z., G.S., W.Y. and H.W. prepared DNA and RNA samples and performed PCR analysis. G.S., K.W., Y.Y., C.L., H.S., Z.L., K.L. and G.Z. provided the homozygous seeds. Y.-X.Z., G.Z. and J.Y.Y. conceived the project and revised the manuscript. Jun Wang, Z.M., R.J.K., R.G.P. and J.Y.Y. performed genome analysis and physical map integration. C.L., Q.L., C.Z. and Y.-X.Z. conceived the project and wrote the manuscript. Jun Wang, Z.M., R.J.K., R.G.P. and J.Y.Y. conceived the project and revised the manuscript. S.Y. conceived and directed the project.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/.
ONLINE METHODS

Strain selection. We used DNA samples prepared from a specific genetic line (cultivar Shixiy1, SXY1) of the cultivated diploid cotton G. arboreum for sequencing and assembly. This cotton line was bred to near homozygosity after 18 successive generations of self-fertilization. G. arboreum cv. Shixiy1 (SXY1) and G. raimondii D5-3 (CMD#10) plants were maintained in the National Wild Cotton Nursery in Sanya, Hainan Province, China.

DNA extraction, library construction and sequencing. Fresh young leaves were collected, immediately frozen in liquid nitrogen and stored at −80 °C until DNA extraction. We used the standard phenol-chloroform method for DNA extraction with RNase A and proteinase K treatment to prevent RNA and protein contamination. Genomic libraries were prepared following the manufacturer’s standard instructions and sequenced on the Illumina HiSeq 2000 platform. To construct paired-end libraries, DNA was fragmented by nebulization with compressed nitrogen gas, and DNA ends were blunted before adding an A base to each 3′ end. DNA adapters with a single T-base 3′-end overhang were ligated to the above products. Ligation products were purified on 0.5%, 1% or 2% agarose gels, which each targeted a specific range of insert sizes. We constructed G. arboreum genome sequencing libraries with insert sizes of 180 bp, 250 bp, 350 bp, 500 bp, 800 bp, 2 kb, 5 kb, 10 kb, 20 kb and 40 kb.

Genome assembly. All sequences were assembled with SOAPdenovo software.16,17 Unusable reads were filtered out before assembly, including (i) reads that contained sequences with >10% “N” bases, (ii) reads with low-quality data (i.e., S7) for 65% of bases for short insert sizes or 80% of bases for long insert sizes, (iii) reads that contained >10 bp of adaptor sequence, (iv) reads with >10 bp that overlapped between two ends of reads of short insert size and (v) reads with identical sequences at the two ends. The 193.6 Gb of clean data were assembled to contigs and scaffolds using the de Bruijn graph–based assembler of SOAPdenovo16,17 with the following four steps.

Constructing K-mer graphs. SOAPdenovo split reads with an insert size of <1,000 bases into K-mers to use in constructing the de Bruijn graphs. The parameters of this step in the cotton genome assembly were as follows: SOAPdenovo-127mer pregraph –s cotton_0308.lib –K 75 –R –p 24 –D –o GOsakoD>–pregraph.log.

Building contigs. SOAPdenovo built contigs using the simplified de Bruijn graphs. The parameters of this step in the cotton genome assembly were as follows: SOAPdenovo-127mer contig –g GOsakoD –R >contig.log.

Linking scaffolds. SOAPdenovo aligned paired-end reads on the contig sequences. We subsequently calculated the total number of paired-end relationships between each pair of contigs and analyzed the rates of conflicting and consistent paired ends. A connecting relationship was regarded as a credible relationship between two contigs when that relationship met the given conditions. We then linked the contigs into scaffolds with N bases in a step-by-step manner from the paired-end reads with the shortest insert size to the paired-end reads with the longest insert size. We determined the orientation of the scaffolds on the basis of the paired-end relationship between reads and prescaffolds while we constructed the scaffolds. A reliable link was built between 2 contigs when there were ≥2 paired-end reads for short inserts or ≥5 paired-end reads for long inserts. We started with the smallest libraries, using the RANK parameter of the configuration files to control the order of the libraries. The parameters of this step in the cotton genome assembly were as follows: SOAPdenovo-127mer map –s cotton_0308.lib –p 16 –g GOsakoD –k 39 –map.log; SOAPdenovo-127mer scaffold –g GOsakoD –F –p 16 >scaffold.log.

Filling gaps. We used GapCloser to fill in gaps in the scaffolds. This program makes use of reads for local assembly within the gaps by aligning the other ends of paired-end reads onto scaffolds. The parameters were as follows: GapCloser –a /GOsakoD/scafSeq –b /cotton_0330.lib –o /GOsakoD/scafSeq.FG –t 64.

Genetic map construction. Genetic linkage maps were constructed to develop the integrated genome map for anchoring the scaffolds, using 154 individuals from an F2 population of a cross between G. arboreum cultivars SXY1 and AnhuiFuyang1Zisedahua (AFZ). Young leaves of parents and individuals from the F2 population were collected for DNA extraction using the Plant Genomics DNA Kit (Qiagen). Genomic DNA from this mapping population was digested with the restriction endonuclease EcoRI and processed into RAD libraries. Every 20 cotton individuals were pooled into 1 sequencing library with nucleotide multiplex identifiers (4, 6 and 8 bp), and each sequencing library was barcoded. Approximately 1,600 Mb of 50-bp reads (10 Mb of read data for each progeny on average) was generated on the HiSeq 2000 next-generation sequencing platform. The SNP calling process was carried out using the SOAP2+SNPpipeline48. JoinMap4.0 (ref. 40) was used to conduct the linkage analysis. A χ2 test was used to determine whether the RAD-based SNP markers corresponded with the expected segregation ratio. Unusable markers (P > 0.01) were filtered out before a genetic map was constructed. ABH-style markers were used to build the main framework of 13 linkage groups, and a logarithm of odds (LOD) score of 7 was initially set as the linkage threshold for linkage group identification. AC-style markers were also included to expand the length of linkage groups by reducing the LOD value to 3. All high-quality markers were used to construct the consensus maps with the F2 population in JoinMap4.0. We calculated the recombination fractions between all pairs of SNP markers in a scaffold and chose the SNP marker that had the minimum recombination fraction in the sum. Scaffold order was determined by tag SNPs, and inner SNPs were then used to orient the scaffolds.

Annotation of transposable elements. We searched the genome for tandem repeats using the software program Tandem Repeats Finder41. Both homology-based and de novo approaches were used to find TEs. The homology-based approach involved applying commonly used databases of known repetitive sequences along with such programs as RepeatProteinMask and RepeatMasker42. To identify DNA-level TEs, we used Repbase43 along with a database of plant repeating sequences and our de novo TE library to find repeats with RepeatMasker. We used four software programs, LTR_FINDER44, PILER45, RepeatModeler and RepeatScout46, to generate the de novo repeat library. These programs predict repeat sequences in different ways. LTR_FINDER retrieves full-length LTR retrotransposons, which usually contain ~18 bp of sequence that is complementary to the 3′ tail of certain tRNAs. PILER searches for repeats in the genome by aligning the genome with itself. RepeatScout builds a consensus sequence on the basis of aligning the fit-preferred alignment score. RepeatModeler uses two ab initio repeat prediction programs (RECON and RepeatScout), which identify repeat element boundaries and family relationships among sequences. Intact LTRs were predicted using the LTR_STRUC47 Windows-based software, and solo LTRs were predicted on the basis of comparison with intact ancestor LTRs. Single-copy gene families were used to construct the phylogenetic tree, and syntenic blocks were identified by MCscan48.

Gene prediction. Homology-based gene prediction. The target locations of homologous proteins were obtained by aligning the protein sequences of A. thaliana, Carica papaya, Populus trichocarpa, G. raimondii, Populus trichocarpa, T. cacao and Vitis vinifera to the G. arboreum genome using TBLASTN with an E-value parameter of 1 × 10−5. Proteins encoded by TEs were filtered out before alignment. We next extracted target sequences from the genome, including a 2,000-bp extension at both ends of the alignment with the genome sequence. ABLAT with an E-value parameter of 1 × 10−5 was used to find coding regions that were >150 bp in length during the process.
Transcriptome alignment. We used TopHat\textsuperscript{53} to map raw reads of cDNA sequence to the G. arabum genome sequence to identify potential areas of exons in addition to donor or receptor sites with the following parameters: \(-p \text{ 4-max-intron-length } \text{20,000} \text{– } 1 \text{ -r-20-mate-std-20 dev}}\text{20}. Cufflinks\textsuperscript{53} software was used to assemble potential alignments to transcripts, with the following parameters: \(-r \text{20,000} \text{– } 4\). To obtain complete gene models, we used the fifth-order Markov model to predict ORFs.

Combination of all methods. Data derived from homology-based, \textit{de novo} and transcriptome alignment methods were integrated to generate a consensus gene set by GLEAN\textsuperscript{54} with default parameters.

Phylogenetic analysis. Gene Ontology (GO)\textsuperscript{55} terms for each gene were obtained from its corresponding InterPro entry. We used OrthoMCL\textsuperscript{56} to confirm the genes that were orthologous among four species (\textit{A. thaliana, G. arabum, G. raimondii} and \textit{T. cacao}). All-versus-all BLASTP (E value \(\leq 1 \times 10^{-5}\)) comparison of all protein sequences for each species was performed, and orthologous genes were clustered by OrthoMCL (inflation parameter: 1.5). We performed multiple alignments of protein sequences with MUSCLE\textsuperscript{57} for each single-copy gene family and converted the protein alignments to coding sequences using a Perl script. Phase-1 sites were extracted from each family and concatenated to one supergene for every species, and Mrbayes 3.1.2 (ref. 58) was used to construct the phylogenetic tree. Branch-specific Ka/Ks values were calculated using codeml in the PAML 59 software package with (ref. 58) was used to construct the phylogenetic tree. Branch-specific Ka/Ks values were calculated using codeml in the PAML 59 software package with respect to the G. arboretum expression level at 0 dpa. For qRT–PCR analysis, total RNA (~2 μg) was reverse transcribed in a 20-μl reaction mixture using SuperScript II reverse transcriptase (Invitrogen). After the reaction, 1-μl aliquots were used as a template for PCR amplification with gene-specific primers. As an internal control, the UBQ7 transcript was used to quantify the relative transcript level of each target gene in each tissue type. Three biological replicates were analyzed for all qRT-PCR experiments.

Synteny analysis and whole-genome alignment. A BLASTP search (with an E-value cutoff of \(1 \times 10^{-5}\)) was performed to identify paralogous genes. Syntenic blocks (with at least five genes per block) were identified by MCscan\textsuperscript{48}. For the alignment results between these, each aligned block represented the orthologous pair derived from the common ancestor, and the sequences that contained the genes were used to show the intergenome relationships with their length information. The fourfold-degenerate value of the blocks was calculated as revised by the HKY model, and whole-genome alignment was carried out by LASTZ (see URLs) between four species, and MRBAYES 3.2 (ref. 58) was used to generate the consensus gene set by GLEAN 54 with default parameters.

LTR analysis. Intact LTRs from the A genome and D genome were predicted using LTR_STRUC\textsuperscript{57} Windows-based software. We constructed ancestor subfamilies on the basis of the following two rules: (i) family members had to share the same polypurine tract sequences and primer-binding-site sequences and (ii) the E value had to be \(< 1 \times 10^{-10}\) among transposable gene sequences. For each family, we aligned all intact LTRs with MUSCLE\textsuperscript{57} and manually corrected the alignments using MEGAS5.5 (ref. 60). We then constructed ancestor sequences for this family using the cons program (contained in the EMBL software package\textsuperscript{61}). To find more solo LTRs, we used the ancestor sequences for different families as queries for LTR prediction based on the standard prediction pipeline, and we followed the 80-80-80 rule (identity, \(>0.8\); align rate, \(>80\%\); alignment length, \(>80 \text{ bp}\))\textsuperscript{62} to find which family the solo LTRs belonged to. All intact LTRs and solo LTRs were used to calculate the insert time with the fourfold-degenerate value of the blocks was calculated as revised by the HKY model, and whole-genome alignment was carried out by LASTZ (see URLs) between four species, and MRBAYES 3.2 (ref. 58) was used to generate the consensus gene set by GLEAN 54 with default parameters.

Analysis of disease resistance- and fiber development–related genes. The largest class of characterized disease resistance (R) genes encodes intraacellular proteins that contain a nucleotide-binding site (NBS) and C terminus with leucine-rich repeats (LRRs)\textsuperscript{63}. These genes are rather abundant in plant genomes and have an important role in mediating resistance to pathogens\textsuperscript{19,64}.