nature biotechnology

OPEN

Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution

Fuguang Li[1,9], Guangyi Fan[2,9], Cairui Lu[1,9], Guanghui Xiao[3,4,9], Changsong Zou[1,9], Russell J Kohel[5,9], Zhiying Ma[6,9], Haihong Shang[1,9], Xiongfeng Ma[1,9], Jianyong Wu[1,9], Xinming Liang[2,9], Gai Huang[3,4], Richard G Percy[5], Kun Liu[1], Weihua Yang[1], Wenbin Chen[2], Xiongming Du[1], Chengcheng Shi[2], Youlu Yuan[1], Wuwei Ye[1], Xin Liu[2], Xueyan Zhang[1], Weiqing Liu[2], Hengling Wei[1], Shoujun Wei[1], Guodong Huang[2], Xianlong Zhang[7], Shuijin Zhu[8], He Zhang[2], Fengming Sun[2], Xingfen Wang[6], Jie Liang[2], Jiahao Wang[2], Qiang He[2], Leihuan Huang[2], Jun Wang[2], Jinjie Cui[1], Guoli Song[1], Kunbo Wang[1], Xun Xu[2], John Z Yu[5], Yuxian Zhu[3,4] & Shuxun Yu[1]

Gossypium hirsutum **has proven difficult to sequence owing to its complex allotetraploid (A_tD_t) genome. Here we produce a draft genome using 181-fold paired-end sequences assisted by fivefold BAC-to-BAC sequences and a high-resolution genetic map. In our assembly 88.5% of the 2,173-Mb scaffolds, which cover 89.6%~96.7% of the A_tD_t genome, are anchored and oriented to 26 pseudochromosomes. Comparison of this *G. hirsutum* A_tD_t genome with the already sequenced diploid *Gossypium arboreum* (AA) and *Gossypium raimondii* (DD) genomes revealed conserved gene order. Repeated sequences account for 67.2% of the A_tD_t genome, and transposable elements (TEs) originating from D_t seem more active than from A_t. Reduction in the A_tD_t genome size occurred after allopolyploidization. The A or A_t genome may have undergone positive selection for fiber traits. Concerted evolution of different regulatory mechanisms for Cellulose synthase (*CesA*) and 1-Aminocyclopropane-1-carboxylic acid oxidase1 and 3 (*ACO1,3*) may be important for enhanced fiber production in *G. hirsutum*.**

The allotetraploid species *G. hirsutum* L. is both the world's most important fiber crop and a model polyploid crop[1,2]. The *Gossypium* genus originated from paleo-hexaploidy of a eudicot progenitor and subsequent diversification into eight diploid genome groups, including A–G and K[3,4]. The A-genome diploids of the genus, which are native to Africa, diverged from the eudicot progenitor together with the Mexican-derived D-genome diploids about 5~10 million years ago (MYA)[5–7]. These two species were reunited geographically around 1~2 MYA by the transoceanic dispersal[8] of an A-genome ancestor resembling *G. arboreum* (AA) to the New World. Subsequently, the A-genome progenitor hybridized with a New World D-genome ancestor resembling *G. raimondii* (DD), followed by chromosome doubling. The resulting dicotyledonous allotetraploid cotton dispersed from the Americas to the western Pacific and diverged into at least five well-established allotetraploid *Gossypium* species ($A_tA_tD_tD_t$; $2n = 4x = 52$, where 't' stands for tetraploid), including *G. hirsutum*, the Upland cotton that accounts for more than 90% of commercial cotton production worldwide[8].

High-quality assembly of allopolyploid plant genomes is a formidable task because the genomes are large and have highly homeologous subgenomes. A draft genome of the 17-gigabase (Gb) monocotyledonous allohexaploid crop *Triticum aestivum* ($2n = 6\times = 42$, AABBDD) was recently reported by an international consortium[9]. As a result of their concerted efforts, 61% of the scaffolds were anchored and oriented on 21 pseudochromosomes[9]. In another study, 63% (712.3 Mb) of the 1,130-Mb allotetraploid ($2n = 4\times = 19$, $A_nA_nC_nC_n$) *Brassica napus* genome was assembled. 18,278 of the *B. napus* 20,702 scaffolds were assigned to either the A_n or C_n subgenomes, which constituted 19 chromosomes in total[10].

Here we sequenced and assembled the allotetraploid genome of *G. hirsutum* using DNA prepared from the highly homozygous TM-1 pure line[11,12]. We compared the *G. hirsutum* assembly to the putative ancestral species, *G. raimondii*[6] and *G. arboreum*[5], to investigate subgenome evolution and gene function including genes related to fiber biology.

[1]State Key Laboratory of Cotton Biology, Institute of Cotton Research of the Chinese Academy of Agricultural Sciences, Anyang, China. [2]BGI-Shenzhen, Shenzhen, 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]Crop Germplasm Research Unit, Southern Plains Agricultural Research Center, US Department of Agriculture–Agricultural Research Service (USDA-ARS), College Station, Texas, USA. [6]Key Laboratory for Crop Germplasm Resources of Hebei, Agricultural University of Hebei, Baoding, China. [7]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, China. [8]Department of Agronomy, Zhejiang University, Hangzhou, China. [9]These authors contributed equally to this work. Correspondence should be addressed to Y.Z. (zhuyx@whu.edu.cn), S.Y. (Yu@cricaas.com.cn) or J.Z.Y. (John.Yu@ars.usda.gov).

Received 8 October 2014; accepted 15 March 2015; published online 20 April 2015; doi:10.1038/nbt.3208

RESULTS

Assembly of the *G. hirsutum* genome

The allotetraploid genome of Upland cotton *G. hirsutum* has been estimated, using various methods, as 2.25–2.43 Gb[13,14]. We generated a total of 445.7 Gb, or 181-fold haploid genome coverage, of raw paired-end Illumina reads by sequencing whole genome shotgun (WGS) libraries of homozygous cv. 'TM-1' with fragment lengths ranging from 250 bp to 40 kb (**Supplementary Table 1**). Owing to the existence of abundant repetitive sequences and homeologous chromosomes (**Supplementary Table 2**), we were unable to assemble this allotetraploid genome satisfactorily using only the WGS data. Supplemental use of a bacterial artificial chromosome (BAC-to-BAC) sequencing strategy substantially improved the assembly. A total of 100,187 BACs, that corresponded to about fivefold genome coverage, were sequenced and used in the final assembly (**Supplementary Fig. 1a**,**b**). Each BAC was assembled individually before genome assembly (**Supplementary Fig. 1c**). Genome assembly used sequenced BACs and paired-end data. A total of 2,173 Mb of the *G. hirsutum* genome sequence was assembled using SOAPdenovo[15,16], with the largest scaffold being 8.4 Mb (**Table 1**). This corresponds to 96.7% of the previous estimation of nuclear DNA content[13], or 89.6% according to a more recent report[14]. The N50 (the size above which 50% of the total length of the sequence assembly can be found) of the contigs and scaffolds was 80 kb and 764 kb, respectively, which was better than the assembly that used WGS data only (N50 of contigs and scaffolds was 20 kb and 107 kb, respectively; **Table 1** and **Supplementary Tables 2** and **3**).

The quality of the assembly was assessed by aligning scaffolds to BAC sequences obtained using Sanger technology, and also by mapping available expressed sequence tags (ESTs) to the A_tD_t genome. All 223 sequenced BACs were recovered with >96% sequence identity in our assembly (**Supplementary Table 4**). Of the 108,790 transcripts derived from transcriptome sequencing of *G. hirsutum*, 98.9% were detected in our assembly (**Supplementary Table 5**). Of the assembled genome, 1,923 Mb or 88.5% was anchored and oriented to 26 pseudochromosomes according to a high-resolution genetic map that we constructed based on 39,662 co-dominant single-nucleotide polymorphism (SNP) markers produced from the TM-1 × 3-79 recombinant inbred line (RIL) population of 167 individual lines (**Supplementary Tables 6** and **7**).

Genome duplications and tetraploidization

All present-day allopolyploid cottons are thought to have arisen from a single common lineage derived from the hybridization of A_2 and D_5 genome ancestors[3,17,18]. However, early polyploidization events in the evolution of the *Gossypium* genus occurred before the AADD lineage emerged. In comparisons of all sequenced *Gossypium* genomes with the seven ancestral chromosome groups reported in grape (*Vitis vinifera*)[19], we found that the paleopolyploidy was shared at orthologous positions from the ancestor to *G. raimondii*, *G. arboreum* and *G. hirsutum* (**Fig. 1a**). These results confirm previous reports of the paleohexaploid origin of eudicot species such as *Theobroma cacao*[20] and *Eucalyptus grandis*[21]. Clear-cut chromosomal DNA fragmentation patterns were observed in eudicot genomes that underwent zero (as in *V. vinifera*[19] and *T. cacao*[20]), one (as in all three cotton species) or two (as in *Arabidopsis thaliana*[22] and *Glycine max*[23]) whole genome duplications during their evolution. It should be noted that chromosome 16 of *V. vinifera* might not have been assembled correctly, as it had much lower than expected percentages of remnant ancestral DNA (**Fig. 1a**).

According to our calculations, the ancient hexaploidization event shared among the eudicots[24] and the more recent duplication event in the progenitor of *G. arboreum* (A genome) and *G. raimondii* (D genome) occurred around 115~146 and 13~20 MYA, respectively (**Supplementary Fig. 2**), which is similar to what was previously reported[5,6]. An additional peak around 1.5 MYA that corresponds to the predicted hybridization and subsequent polyploidization event[7] was observed only in *G. hirsutum* (A_tD_t genome) (**Supplementary Fig. 2**). We assigned the 26 pseudochromosomes of *G. hirsutum* to either the A_t or D_t subgenome based on syntenic relationships with its predicted diploid ancestors as well as the published genetic maps[12] (**Fig. 1b**,**c** and **Supplementary Fig. 3**). The anchored A_t subgenome (1,170 Mb) is larger than the D_t subgenome (753 Mb), consistent with the sizes of the anchored A genome[5] (1,530 Mb, 90.4% of 1,694 Mb) and D genome[6] (567 Mb, 73.2% of 775 Mb). Co-linearity analysis of the current assembly indicated that the *G. hirsutum* genome is of high quality (**Supplementary Fig. 4**).

Table 1 ***G. hirsutum*** **genome assembly and annotation**

Category	Number	N50 (kb)	Longest (kb)	Size (Mb)	Percent of the assembly
Total contigs	44,816	80	784	2,090	–
Total scaffolds	8,591	764	8,400	2,173	100.0
Anchored and oriented scaffolds	4,023	853	8,400	1,923	88.5
Genes annotated	76,943			220	9.5
miRNAs	602			0.07	<0.01
rRNAs	2,153			0.6	<0.01
tRNAs	2,050			0.2	0.01
snRNAs	8,325			0.9	0.04
Repeat sequences	–			1,471	67.2

Annotation and gene content

We annotated 76,943 gene models in the *G. hirsutum* genome by combining *ab initio* gene prediction, homolog protein data search, EST alignment and assembly of RNA-seq reads (**Supplementary Table 8**). In bread wheat, a total of 124,201 gene loci have been identified, and 101,040 have been identified in oilseed rape[9,10]. These and our data suggest that the coding capacities of a plant species are basically proportional to its polyploidism. About 84.5% of our gene models have homolog matches in databases, including KEGG and GO (**Supplementary Table 9**). Of the gene models observed in *G. hirsutum*, 93.76% or 72,142 were evenly distributed along chromosomes, with 35,056 in the A or A_t subgenomes and 37,086 in the D or D_t subgenomes. Also, the Upland cotton genome encodes 602 microRNAs (miRNAs), 2,153 rRNAs, 2,050 tRNAs and 8,325 small nuclear RNAs (snRNAs) (**Table 1** and **Supplementary Table 10**).

We scanned the *G. hirsutum* genome, using an all-versus-all blastp approach, for syntenic blocks and found that the D_t and A_t subgenomes had high collinear relationships with the genomes of *G. raimondii* and *G. arboreum*, respectively. There were 1,801 collinear blocks between *G. arboreum* and *G. hirsutum*, which covered 68.2% and 65.9% of the respective A genome or A_t subgenome (**Fig. 1b** and **Supplementary Table 11**). Similarly, we observed a total of 2,241 collinear blocks between *G. raimondii* and *G. hirsutum*, which covered 91.9% and 88.8% of the respective genome or subgenome (**Fig. 1c** and **Supplementary Table 11**).

Comparative studies of cotton genomes

Plant genomes have been massively invaded by transposable elements (TEs), many of which are located near host genes[25]. The presence of TEs can activate or repress genes under specific biotic or abiotic conditions, or even at different developmental stages[26]. As much as 66% of the *G. hirsutum* assembly is composed of TEs (**Supplementary Table 12**).

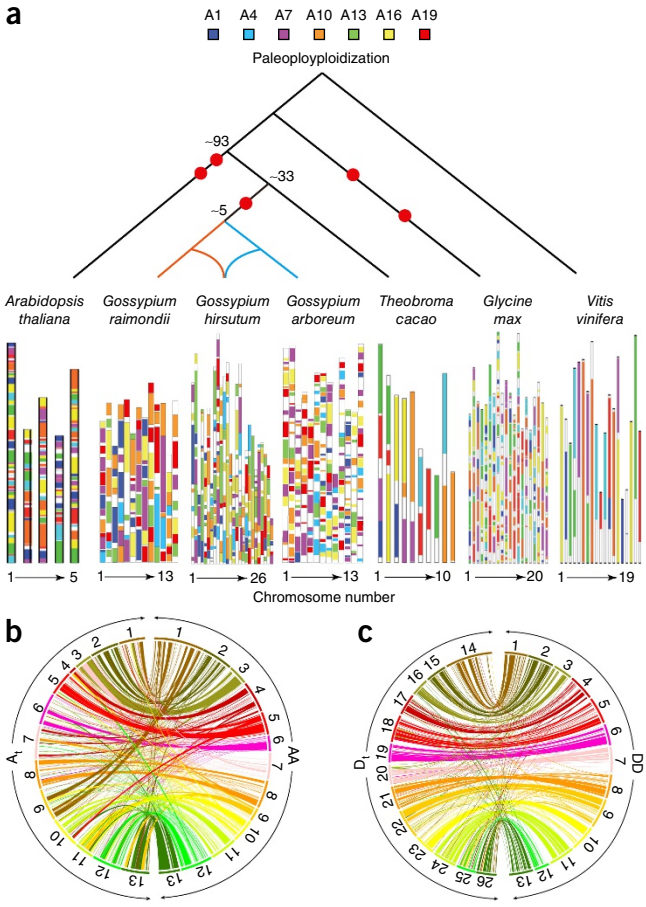

Figure 1 Evolution and syntenic analysis of the *G. hirsutum* genome. (**a**) *G. hirsutum* and six other genomes descended from common eudicot genome ancestors. Colored blocks within modern chromosomes of the species represent the chromatin origin from seven ancestral chromosomes. Numbers denote the predicted divergence times (MYA) and each red dot represents one whole genome duplication. (**b**) Syntenic blocks between the A_t subgenome in *G. hirsutum* and the diploid A genome in *G. arboreum* genome. (**c**) Syntenic blocks between the D_t subgenome in *G. hirsutum* and the diploid D genome in *G. raimondii*.

In common with *G. arboreum*[5] and *G. raimondii*[6], the A_t subgenome of *G. hirsutum* contains markedly higher amounts of *gypsy* than D_t, whereas the reverse is true for *copia*. However, both *copia* and *gypsy* were substantially more actively transcribed in the D_t subgenome (**Fig. 2a**). The estimated insertion time of long terminal repeat (LTR) retrotransposons, including both *copia* and *gypsy*, was determined using a spontaneous mutation rate[27] (**Fig. 2b**). *Copia* elements were remarkably more active than *gypsy* (Student's *t*-test, $P \leq 0.05$) in the recent 0~1 MYA time frame (**Fig. 2b**), with higher proportions of *copia* located near coding genes than *gypsy*-type (**Fig. 2c**, Student's *t*-test, $P \leq 0.05$). Our data indicate that the TEs of the D_t subgenome tend to be more active than that of the A_t subgenome after the tetraploidization[28].

The different abundances of retrotransposons and DNA transposons (62.81% and 2.52%, respectively, in our paper versus 52.29% and 1.08%, respectively, in Zhang *et al.*[28]) was probably the result of using different methods to detect TEs. We used the *de novo* method and the homolog search method based on Repbase, whereas Zhang *et al.*[28] used the MIPS database. As a result, only 0.43% of the TEs found are recognized as unclassified elements in our analysis, whereas 11.38% unclassified elements are reported by Zhang *et al.*[28]. Using our methodology, a further 9.43% of their unclassified elements can be recognized as retrotransposons, and an additional 1.25% of the unclassified elements can be recognized as DNA transposons. The MIPS database is constructed strictly according to the published cotton genomes whereas Repbase contains all repeat sequences from various plant genomes, further accounting for differences in findings.

Gene order was constructed by quartet alignments among the four sister chromosomes from the three *Gossypium* species. Using an iterative method based on the best match and the syntenic gene finder in MCScan[29], we identified 32,466 and 32,811 orthologous gene pairs within the A_t and D_t subgenomes, respectively, in *G. hirsutum* and their corresponding ancestral A and D diploid genomes (**Supplementary Table 13**). Of the 28,592 orthologous gene pairs identified between *G. arboreum* and *G. raimondii*, 22,876 were also conserved homeologous gene pairs in *G. hirsutum* (**Fig. 3a**). Of the orthologous gene pairs, 643 were lost from both *G. raimondii* and the D_t subgenome of *G. hirsutum*, and 478 were lost from both *G. arboreum* and the A_t subgenome (**Fig. 3a**), which indicates that the ancestral D genome suffered a higher frequency of genic sequence losses. We further identified gene losses in syntenic blocks in which at least two of the four *Gossypium* genomes have orthologous genes: 523 and 461 genes were absent from A_t and D_t, respectively, compared with 234 lost in the A genome and 390 lost in the D genome (**Fig. 3a**). These data suggest a higher rate of gene losses in the allotetraploid cotton than in both diploid species. Also, more genes were lost in the A_t subgenome than the D_t subgenome during the formation of *G. hirsutum*.

Homeologous exchanges (HEs) among cotton genomes

HEs, or transpositions, are characterized by the loss of a chromosomal region that is replaced by a duplicate copy of the corresponding homeologous region[10,30,31]. The synonymous substitution rates (*Ks* values) of orthologous DNA are used to evaluate evolutionary

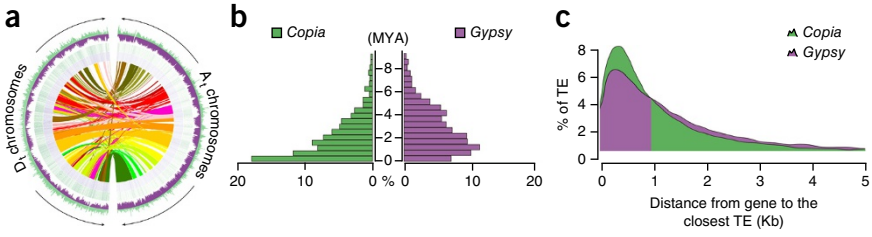

Figure 2 Characterization of *copia* and *gypsy* TEs in the *G. hirsutum* genome. (**a**) Statistics for these two types of TEs present in the D_t (left half) and A_t (right half) subgenomes. The outer circle shows the percent coverage of *copia* (green histogram) and *gypsy* (purple histogram) in nonoverlapping windows (window size = 500 kb). The following two inner circles indicate the *copia* and *gypsy* transcript levels, which were estimated by averaging values of reads ($\log_{10}$) from different tissues in nonoverlapping 500-kb windows. The links in the center indicate collinearity between A_t and D_t subgenomes. Only syntenic blocks of >1 Mb in length are shown. (**b**) Estimated insertion time for *copia* and *gypsy* LTR retrotransposons. (**c**) Distances from individual TEs to their nearest gene.

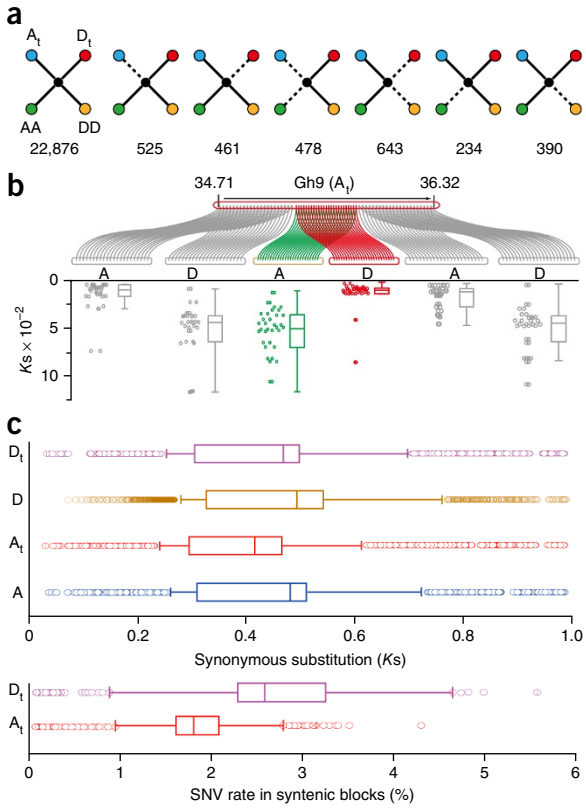

Figure 3 Evolution of gene models, DNA fragments and syntenic blocks among *G. hirsutum* (A_tD_t) and two diploid cotton genomes, *G. arboreum* (A) and *G. raimondii* (D). (**a**) Scenarios and statistics of gene conservation. Solid lines indicate currently observed genes, and dotted lines indicate lost genes. The numbers beneath each drawing represent the number of gene pairs found in the three different genomes that fit the specific model. From left to right, genes present in all four genomes, genes not observed in A_t, genes not observed in D_t, genes not observed in either A or A_t, genes not observed in either D or D_t, genes not observed in A, genes not observed in D. (**b**) HE of genomic segments between the A_t and D_t subgenomes in a region of *G. hirsutum* chromosome 9 (Gh9). The curves in the upper panel show homologous gene pairs between A_t and A or A_t and D. The lower panel shows the *Ks* value distribution for syntenic blocks, which indicates HE in the tetraploid cotton. The dot plots show distribution of *Ks* values and the boxplots display variations of *Ks* values between A and D genomes. Note that some of the dots (outliers) are not included in the boxplots due to their low probabilities. (**c**) Distribution of *Ks* values between four cotton genomes and *T. cacao* (upper panel) and single-nucleotide variation (SNV) rate (lower panel) among different cotton genomes.

distances among different plant species[32]. In the A-genome-derived chromosome 9 of *G. hirsutum*, a syntenic block in the middle of the chromosome showed a very low *Ks* value with a fragment from the D genome, whereas the two other blocks in this same region displayed low *Ks* values to DNA fragments from the A genome (**Fig. 3b**), suggesting that this A_t chromosome segment had undergone HE with a D_t homeologous region. Potential HEs were supported by multiple paired-end reads to ensure that they did not result from misassemblies (**Supplementary Fig. 5**). Putative HEs with junction distances ≤7 kb were selected for PCR verification using primers designed from both sides of the particular junction, and five out of seven potential HEs were successfully amplified in *G. hirsutum*, but not in *G. arboreum* or *G. raimondii* (**Supplementary Fig. 6**). Also, HE regions showed a considerably higher rate of multicopy genes than the genome as a whole, indicating that HEs may be associated with the evolution of high copy number genes (**Supplementary Fig. 7**). Using this approach across all *G. hirsutum* chromosomes, we identified 100 possible HEs, of which 54 were D_t to A_t (**Supplementary Table 14**), 46 were A_t to D_t (**Supplementary Table 15**). Further analysis of syntenic blocks involved in HEs indicated that 1,790 genes were transferred from D_t to A_t, whereas 1,530 genes were transferred from A_t to D_t. The average *Ks* values for collinearity-supported gene pairs were 0.463 for D_t versus *T. cacao* and 0.494 for D versus *T. cacao*, with 0.422 for A_t versus *T. cacao* and 0.476 for AA versus *T. cacao* (**Fig. 3c**, upper panel). The *Ks* values of both A_t and D_t subgenomes were substantially lower than those of the A and D diploid genomes, respectively. Similarly, analysis of intergenic collinear regions showed that the single-nucleotide variation rate for D_t versus D was greater than that for A_t versus A (**Fig. 3c**, lower panel). These results indicate that the D or D_t genome has higher mutation rates than A or A_t genome, which is consistent with a previous report for the *Adh* locus[33]. Further analysis revealed dN/dS ratios of 0.338 for D_t versus *T. cacao*, 0.270 for D versus *T. cacao*, 0.404 for A_t versus *T. cacao* and 0.314 for AA versus *T. cacao*. These data suggest that the A and A_t genomes are undergoing less purifying selection and greater positive selection than the D and D_t genomes; and that the genetic redundancy created by allotetraploidy may have allowed relaxed purifying selection in both the A_t and D_t subgenomes.

Deletions and fiber development

Ethylene is a key modulator of cotton fiber cell growth[34–36], and significant differences in ethylene or cell wall biosynthetic gene expression patterns during fiber development have been reported[34,37].

Early experiments showed 50- to 500-fold increases in *ACO* transcript levels in *G. hirsutum* and *G. raimondii* ovules, respectively, in comparisons with *G. arboreum*[5]. When grown in a semi *in vivo* ovule culture system, *G. arboreum* ovules released almost undetectable levels of gaseous ethylene over a 15-d culture period, whereas *G. raimondii* produced large amounts of this molecule (**Fig. 4a**). Although ovules of *G. hirsutum* did synthesize measurable levels of ethylene, the onset of this ethylene burst occurred at a much later stage with a lower peak value, as compared with *G. raimondii* (**Fig. 4a**). Gel shift assays were performed to investigate possible molecular mechanisms for regulating the expression of *ACO1* and *ACO3* that are rate-limiting during ethylene biosynthesis in cotton ovules[34]. Sequence alignment showed the deletion of a 128-bp fragment that resulted in the loss of a single MYB binding site in *ACO1* promoters from both the A genome and the A_t subgenome (**Fig. 4b**, upper panel). Significant mobility shift was observed only when fragment P6 from the D genome (either from D or D_t) was incubated with nuclear extracts prepared from 10 d post-anthesis (DPA) *G. hirsutum* ovules (**Fig. 4b**, middle and lower panels). When compared with the GaACO3 promoter, deletion of a large DNA fragment (123-bp) was observed in the GrACO3 promoter. Further scrutiny revealed several short sequence insertions (indels of 8 to 16 bp) in GrACO3 promoter that created two new and additional MYB-binding sites in *G. raimondii* (**Fig. 4c**, upper panel; and **Supplementary Fig. 8**). As a result, only fragment PVI from D_t or D genome showed substantial mobility shift whereas the same PVI fragment from A_t or A genome showed no obvious nuclear protein binding (**Fig. 4c**, middle and lower panels). A much weaker and similar binding of nuclear extracts on the PII fragment was observed for all four types of *ACO3* promoters (**Fig. 4c**, middle and lower panels). When the D genome–derived P6 and PVI fragments from the *ACO1* and *ACO3* promoter regions were individually incubated with nuclear extracts prepared from *G. raimondii* ovules

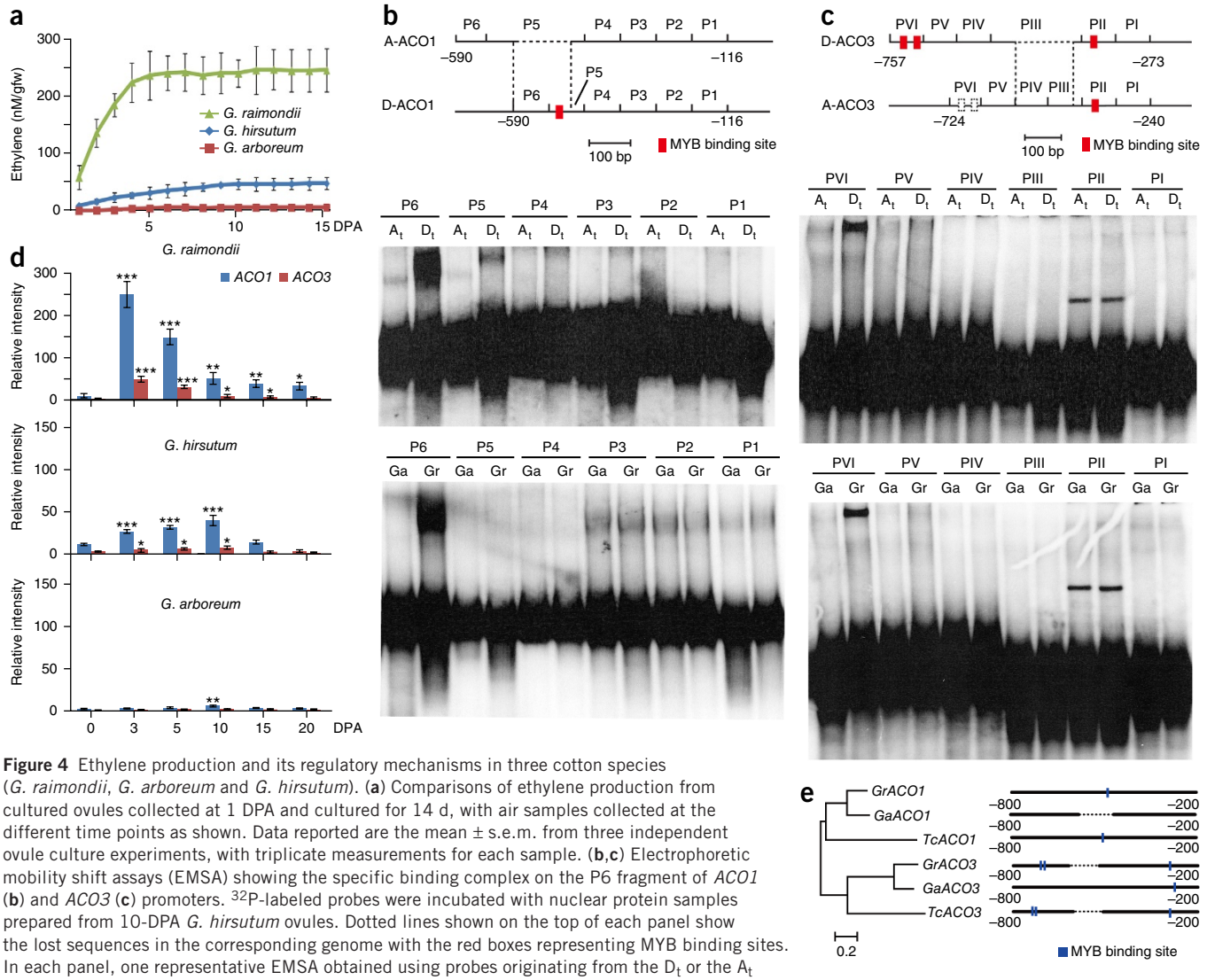

Figure 4 Ethylene production and its regulatory mechanisms in three cotton species (*G. raimondii*, *G. arboreum* and *G. hirsutum*). (**a**) Comparisons of ethylene production from cultured ovules collected at 1 DPA and cultured for 14 d, with air samples collected at the different time points as shown. Data reported are the mean ± s.e.m. from three independent ovule culture experiments, with triplicate measurements for each sample. (**b**,**c**) Electrophoretic mobility shift assays (EMSA) showing the specific binding complex on the P6 fragment of *ACO1* (**b**) and *ACO3* (**c**) promoters. ^{32}P-labeled probes were incubated with nuclear protein samples prepared from 10-DPA *G. hirsutum* ovules. Dotted lines shown on the top of each panel show the lost sequences in the corresponding genome with the red boxes representing MYB binding sites. In each panel, one representative EMSA obtained using probes originating from the D_t or the A_t subgenomes of *G. hirsutum* is shown in the middle, and data obtained using probes produced from *G. arboreum* (Ga) or *G. raimondii* (Gr) in the bottom. (**d**) Comparisons of the binding activity on P6 from *ACO1* and *ACO3* promoter regions of the three different cottons. Shown are data obtained from nucleoproteins prepared from 0-, 3-, 5-, 10-, 15- and 20-DPA ovules and incubated with P6 originating from *G. raimondii*, from the D_t copy of *G. hirsutum* and from *G. arboreum*. Error bars, mean ± s.e.m. from three independent EMSA experiments. (**e**) Phylogenetic and evolutionary analysis of *ACO1* and *ACO3* promoter regions from *G. raimondii*, *G. arboreum* and *T. cacao*. Scale bars, 100 bp. Statistical significance was determined using one-way analysis of variance software. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.

harvested at different developmental stages, we found that the nuclear protein binding activity reached its peak values around 3 DPA; these amounts were about four to five times the levels found in 10-DPA ovules (**Fig. 4d**, upper panel). In *G. hirsutum*, the highest binding activities to both *ACO1* and *ACO3* promoters appeared in 10-DPA ovules, but the peak intensities were less than 20% of those found in *G. raimondii* (**Fig. 4d**, middle panel). As expected, no obvious DNA binding activity was found using nuclear extracts from *G. arboreum* (**Fig. 4d**, lower panel).

Phylogenetic analysis revealed that cotton *ACO1* and *ACO3* are in the same clade with *TcACO1* (Thecc1EG030320t1) and *TcACO3* (Thecc1EG002179t1) of *T. cacao*, respectively (**Fig. 4e**). *GrACO1* and *GrACO3* displayed 43.8% and 43.3% identity, respectively, in promoter regions with their counterparts in *T. cacao*. Several important MYB binding sites observed in either *TcACOs* or *GrACOs* were not found in *GaACOs*, suggesting that *G. raimondii* evolved from the common eudicot ancestor before *G. arboreum*[3].

CesA upregulation and fiber elongation

Profound differences in fiber properties are found among the three cotton species. For example, *G. hirsutum* usually produces fibers >3 cm in lengths, whereas *G. arboreum* produces fibers of 1.3~1.5 cm long, and no spinnable fiber is produced by *G. raimondii* (**Fig. 5a**, left panel). *G. hirsutum* fiber cells undergo fast elongation until ~30 DPA, whereas those of *G. arboreum* stop growth around 20 DPA (**Fig. 5a**, right panel). As a load-bearing polymer of plant cell walls, cellulose microfibrils, which are produced by *cellulose synthase A* (*CesA*), have a key role in the formation of the secondary cell wall for cell growth[38,39]. The *CesA* genes are divided among six major clades (**Supplementary Figs. 9a** and **10**) and genome-wide analysis indicates that *G. hirsutum* contains 32 *CesA* and 64 cellulose synthase-like (*CSL*) genes, representing a notable increase over the number of *CesA* and *CSL* genes in *Arabidopsis*[22] and *T. cacao*[20] (**Supplementary Figs. 9b** and **11**).

Primary cell wall components produced by UGD, UGP and UER are also important for cotton fiber growth[37]. Transcriptome analyses

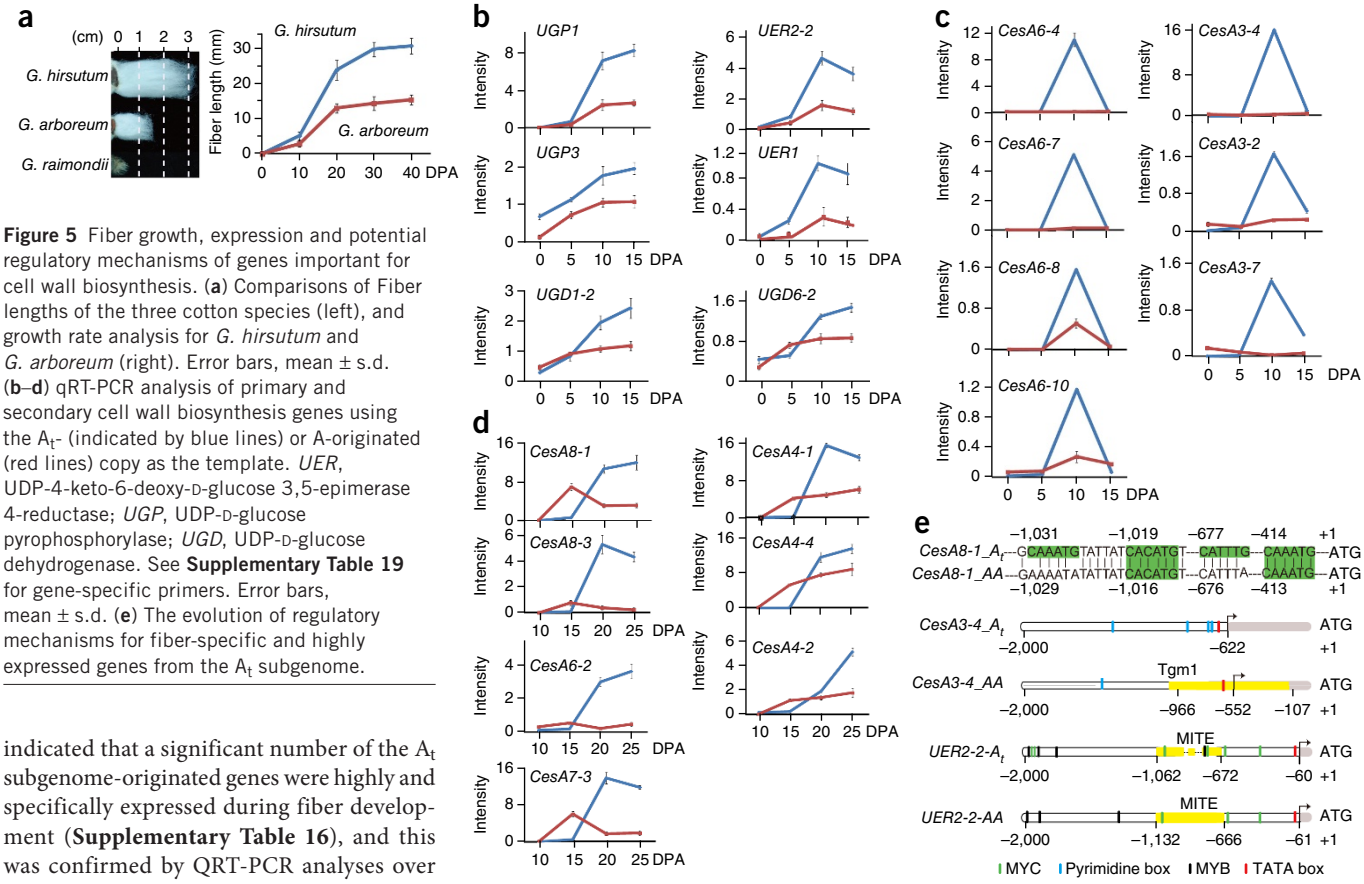

Figure 5 Fiber growth, expression and potential regulatory mechanisms of genes important for cell wall biosynthesis. (**a**) Comparisons of Fiber lengths of the three cotton species (left), and growth rate analysis for *G. hirsutum* and *G. arboreum* (right). Error bars, mean ± s.d. (**b**–**d**) qRT-PCR analysis of primary and secondary cell wall biosynthesis genes using the A_t- (indicated by blue lines) or A-originated (red lines) copy as the template. *UER*, UDP-4-keto-6-deoxy-D-glucose 3,5-epimerase 4-reductase; *UGP*, UDP-D-glucose pyrophosphorylase; *UGD*, UDP-D-glucose dehydrogenase. See **Supplementary Table 19** for gene-specific primers. Error bars, mean ± s.d. (**e**) The evolution of regulatory mechanisms for fiber-specific and highly expressed genes from the A_t subgenome.

indicated that a significant number of the A_t subgenome-originated genes were highly and specifically expressed during fiber development (**Supplementary Table 16**), and this was confirmed by QRT-PCR analyses over a longer growth period of time (**Fig. 5b**–**d**). In all cases, genes that originated from the A_t subgenome, but not their homeologs from the A subgenome, were expressed highly and specifically during either the primary (**Fig. 5b**,**c** and **Supplementary Table 17**) or the secondary (**Fig. 5d**) cell wall biosynthesis stages. Although a few *CSL* genes showed moderate fiber-specific upregulation, there was little difference between the A_t- and A-derived copies (**Supplementary Table 18**). This fiber-specific and high-level expression of cell wall biosynthesis genes was coupled with the evolution of more *cis*-elements through point mutations, as evidenced on the A_t*CesA*8-1 promoter (**Fig. 5e**, upper drawing). Transposition of a *Tgm1* (a subclass of TE) (**Fig. 5e**, middle drawing) and partial deletion of a MITE (**Fig. 5e**, lower drawing) from the ancestral A genome seems to have contributed to the considerably increased number of pyrimidine boxes or the conserved MYC- and MYB-binding sites present on A_t*CesA*3-4 and A_t*UER2-2* promoters.

DISCUSSION

Deciphering the allotetraploid genome of *G. hirsutum* provides unique insights into the earliest stages of postpolyploidy evolution and the relationships between *G. hirsutum* and its ancestral diploid species. Both *G. raimondii* and *G. arboreum* underwent cotton-specific whole genome duplication at ~16.6 MYA[5,6] after sharing the paleohexaploidization event common to all eudicots that occurred ~130.8 MYA[5,6,22]. We estimate that the modern allotetraploid *G. hirsutum* species resulted from hybridization of the two ancestral species about ~1.5 MYA, based on the calculations comparing A_t and D_t subgenomes, with their respective A or D diploid ancestral species (**Fig. 1a** and **Supplementary Fig. 2**). Up to 100 DNA segments have been replaced in one subgenome by sequences from another subgenome. Although most coding genes were conserved among all four genomes, close scrutiny showed that a large number of these genes were transferred from D_t to A_t, so that the current A_t subgenome may express more genes than the ancestral A genome. This segmental replacement apparently causes substantial differences in TE activities in the two subgenomes compared with the postulated ancestors and this seems to contribute to differences in fiber production and fiber quality of the allotetraploid species.

The assembly and analyses of the allotetraploid cotton genome show that the two subgenomes have different mutation rates because the D_t subgenome has higher single-nucleotide variation and *Ks* values than those of the A_t. Also, the allotetraploid genome enjoys lower purifying selection pressure than the ancestral diploid genomes, because both D_t and A_t subgenomes have higher dN/dS ratios than the D and A diploid genomes. Finally, the A_t subgenome may have undergone positive selection for fiber-related traits because the A_t showed a higher dN/dS ratio than the D_t. The significant changes in *Ks* rates among HEs and their flanking sequences suggest that HEs may be associated with "gene conversion," which involves homology-directed double-strand break repair between homeologous chromosomes or sister chromatids[40]. The *ACOs* and primary and secondary cell wall biosynthesis genes reported here might provide targets for engineering of improved fiber yield.

METHODS

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

Accession codes. The *G. hirsutum* genome sequence, including all contigs and scaffolds, has been deposited at NCBI BioProject under the accession number PRJNA259930 and also available at http://cgp.genomics.org.cn. Sequence data for *G. hirsutum* transcriptome

analyses are available in the NCBI Sequence Read Archive (SRA) under accession number SRA180756.

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

ACKNOWLEDGMENTS

We are grateful to Q. Li for his assistance with data analysis and with the early phase of manuscript preparation. This work was supported by the Major Program of Joint Funds (Sinkiang) of the National Natural Science Foundation of China (grant U1303282), the National Natural Science Foundation of China (grant 90717009) and the 863 National High-Tech Research Development Program in China (grant 2013AA102601), and United States Department of Agriculture, Agricultural Research Service (USDA-ARS project 3091-21000-038-01).

AUTHOR CONTRIBUTIONS

F.L., G.F., C.L., Z.M., R.J.K., X.X., J.Z.Y., Y.Z. and S.Y. designed the analyses. G.F., X.m.L., W.C., C.S., X. Liu, W.L., G.d.H., H.Z., J.L., J.W., Q.H., L.H., F.S., J.h.W. and X.X. performed sequencing, assembly and genome annotation. F.L., X.m.L., K.W., G.S., J.y.W., J.C., X.X., J.Z.Y. and S.Y. managed and coordinated the project. C.L., C.Z., H.S., G.X., J.Z.Y. and G.d.H. performed the genome analysis and physical map integration. F.L., C.L., C.Z., Z.M., H.S., X.M., K.W., G.S., J.y.W., J.C., K.L., W.Y., X.D., Y.Y., W.Ye, X.l.Z., H.W., S.Y., G.X., G.H., X.W., S.W., X.Z. and S.Z. prepared DNA/RNA samples and performed PCR analysis. Y.Z., G.X., H.S., C.Z., C.L. and G.H. performed transcriptome and lineage-specific gene functional analyses. R.J.K., R.G.P. and J.Z.Y conceived the project, provided the homozygous seeds and revised the manuscript. Y.Z., C.L., C.Z., G.X. and H.S. wrote the manuscript. S.Y., Y.Z. and F.L. conceived and directed the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Zhu, Y.-X. & Li, F.G. The *Gossypium raimondii* genome, a huge leap forward in cotton genomics. *J. Integr. Plant Biol.* **55**, 570–571 (2013).
2. Chen, Z.J. *et al.* Toward sequencing cotton (*Gossypium*) genomes. *Plant Physiol.* **145**, 1303–1310 (2007).
3. Wendel, J., Brubaker, C., Alvarez, I., Cronn, R. & Stewart, J.M. in *Genetics and Genomics of Cotton*, vol. 3 (ed. Paterson, A.H.) 3–22 (Springer, New York, 2009).
4. Paterson, A.H. *et al.* Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* **492**, 423–427 (2012).
5. Li, F.G. *et al.* Genome sequence of the cultivated cotton *Gossypium arboreum*. *Nat. Genet.* **46**, 567–572 (2014).
6. Wang, K. *et al.* The draft genome of a diploid cotton *Gossypium raimondii*. *Nat. Genet.* **44**, 1098–1103 (2012).
7. Wendel, J.F. & Albert, V.A. Phylogenetics of the cotton genus (*Gossypium*): character-state weighted parsimony analysis of chloroplast-DNA restriction site data and its systematic and biogeographic implications. *Syst. Bot.* **17**, 115–143 (1992).
8. Wendel, J.F. New world tetraploid cottons contain old-world cytoplasm. *Proc. Natl. Acad. Sci. USA* **86**, 4132–4136 (1989).
9. The International Wheat Genome Sequencing Consortium. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* **345**, 1251788 (2014).
10. Chalhoub, B. *et al.* Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* **345**, 950–953 (2014).
11. Kohel, R.J., Richmond, T.R. & Lewis, C.F. Texas Marker-1. Description of a genetic standard for *Gossypium hirsutum*. *Crop Sci.* **10**, 670–671 (1970).
12. Yu, J.Z. *et al.* A high-density simple sequence repeat and single nucleotide polymorphism genetic map of the tetraploid cotton genome. *G3* (*Bethesda*) **2**, 43–58 (2012).
13. Arumuganathan, K. & Earle, E.D. Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* **9**, 208–218 (1991).
14. Hendrix, B. & Stewart, J.M. Estimation of the nuclear DNA content of *Gossypium* species. *Ann. Bot.* **95**, 789–797 (2005).
15. Li, R. *et al.* The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**, 311–317 (2010).
16. Li, R. *et al.* *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
17. Wendel, J.F. & Cronn, R.C. in *Advances in Agronomy* (ed. Sparks, D.L.) 139–186 (Academic Press, 2003).
18. Zhang, H.B., Li, Y., Wang, B. & Chee, P.W. Recent advances in cotton genomics. *Int. J. Plant Genomics* **2008**, 742304 (2008).
19. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
20. Argout, X. *et al.* The genome of *Theobroma cacao*. *Nat. Genet.* **43**, 101–108 (2011).
21. Myburg, A.A. *et al.* The genome of *Eucalyptus grandis*. *Nature* **510**, 356–362 (2014).
22. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
23. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
24. Tang, H. *et al.* Unravelling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18**, 1944–1954 (2008).
25. Oliver, K.R., McComb, J.A. & Greene, W.K. Transposable elements: powerful contributors to angiosperm evolution and diversity genome. *Genome Biol. Evol.* **5**, 1886–1901 (2013).
26. Bennetzen, J.L. & Wang, H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol.* **65**, 505–530 (2014).
27. Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92–94 (2010).
28. Zhang, T. *et al.* Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat Biotechnol.* doi:10.1038/nbt.3207 (20 April 2015).
29. Tang, H. *et al.* Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
30. Udall, J.A., Quijada, P.A. & Osborn, T.C. Detection of chromosomal rearrangements derived from homeologous recombination in four mapping populations of *Brassica napus* L. *Genetics* **169**, 967–979 (2005).
31. Wright, R.J., Thaxton, P.M., El-Zik, K.M. & Paterson, A.H. D-subgenome bias of Xcm resistance genes in tetraploid *Gossypium* (cotton) suggests that polyploid formation has created novel venues for evolution. *Genetics* **149**, 1987–1996 (1998).
32. Yu, J. *et al.* The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol.* **3**, e38 (2005).
33. Small, R.L. & Wendel, J.F. Differential evolutionary dynamics of duplicated paralogous *Adh* loci in allotetraploid cotton (*Gossypium*). *Mol. Biol. Evol.* **19**, 597–607 (2002).
34. Shi, Y.H. *et al.* Transcriptome profiling, molecular biological, and physiological studies reveal a major role for ethylene in cotton fibre cell elongation. *Plant Cell* **18**, 651–664 (2006).
35. Qin, Y.M. *et al.* Saturated very-long-chain fatty acids promote cotton fibre and *Arabidopsis* cell elongation by activating ethylene biosynthesis. *Plant Cell* **19**, 3692–3704 (2007).
36. Qin, Y.M. & Zhu, Y.-X. How cotton fibres elongate: a tale of linear cell-growth mode. *Curr. Opin. Plant Biol.* **14**, 106–111 (2011).
37. Pang, C.Y. *et al.* Comparative proteomics indicates that biosynthesis of pectic precursors is important for cotton fibre and *Arabidopsis* root hair elongation. *Mol. Cell. Proteomics* **9**, 2019–2033 (2010).
38. Peng, L.C., Kawagoe, Y., Hogan, P. & Delmer, D. Sitosterol-β-glucoside as primer for cellulose synthesis in plants. *Science* **295**, 147–150 (2002).
39. McFarlane, H.E., Doring, A. & Persson, S. The cell biology of cellulose synthesis. *Annu. Rev. Plant Biol.* **65**, 69–94 (2014).
40. Rozen, S. *et al.* Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423**, 873–876 (2003).

ONLINE METHODS

Materials. We used the highly homozygous Upland cotton (*G. hirsutum*) genetic standard line TM-1 (ref. 11) for DNA sequencing. The homozygosity was reached through 51 successive generations of self-fertilization by single seed descent from one TM-1 seed. *G. hirsutum* cv. TM-1, *G. arboreum* cv. Shixiya1 and *G. raimondii* acc. D_5-3 (CMD#10) plants were maintained in the greenhouse for extraction of DNA or RNA. The tetraploid *G. barbadense* cv. 3-79 was crossed with TM-1 to develop an RIL population[12]. A total of 167 RIL individuals were successfully sequenced for the construction of genetic maps.

WGS sequencing. Genomic libraries were prepared following the manufacturer's standard instructions and sequenced on the Illumina HiSeq 2000 platform. To construct paired-end libraries, we fragmented DNA by nebulization with compressed nitrogen gas, and blunted DNA ends before adding an A base to each 3′-end. DNA adaptors 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, each of which targeted a specific range of insert sizes. We constructed *G. hirsutum* genome sequencing libraries with insert sizes of 250 bp, 350 bp, 500 bp, 800 bp, 2 kb, 5 kb, 10 kb, 20 kb and 40 kb. All libraries were sequenced on the Illumina sequencing platform, and paired-end reads from each library were obtained. To obtain high-quality reads for assembly, we first filtered out unusable reads, which included the following: (i) reads containing ≥10% "N" bases; (ii) reads with low-quality data (the ASCII value of base-64 ≤7) for 65% of the bases for short insert (<2 kb) or 80% of the bases for long insert (≥2 kb), (iii) reads containing >10-bp adaptor sequences, (iv) reads with >10 bp that overlapped between two ends of reads of short inserts and (v) reads with identical sequences at the two ends.

BAC pooling and sequencing. HindIII and BamHI were used to generate partially digested insert DNA for the construction of TM-1 BAC libraries. Insert sizes of these BAC libraries ranged between 80 kb and 180 kb on average. The Agilent Bravo Automated Liquid Handling Platform (Agilent) and Agilent BenchCel Microplate Handler (Agilent) were used for the construction of BAC libraries. The Adaptive Focused Acoustics (AFA) DNA fragmentation system (Covaris) was used to fractionate DNA samples of BACs. T4 DNA polymerase (Illumina) was used to convert overhangs resulting from fragmentation into blunt ends. Klenow fragment was used to add an A base to the 3′-end of each blunt DNA fragment for the ligation of index adapters having a single T base. DNA samples with different index adapters were pooled together, and DNA segments of 500 bp were selected for library construction. Subsequently, ten 96-well plates were pooled into a single lane for sequencing (i.e., 960 samples/lane), resulting in an average throughput of 100 million reads/lane. In total, 1,080 96-well samples, or 103,680 BAC clones, were obtained in which 100,187 BACs were sequenced successfully. For each BAC clone, a 500-bp paired-end library was constructed and sequenced with the HiSeq2000 sequencing platform. Each BAC clone was sequenced with >100-fold coverage on average.

Genome assembly. The allotetraploid genome of *G. hirsutum* was assembled through a BAC-to-BAC strategy combined with WGS sequencing. The primary steps were as follows.

1. We first assembled each BAC clone using SOAPdenovo[16] based on the BAC reads with expected read-length for each clone.
2. We also assembled all the WGS sequencing reads from the libraries with short insert sizes into contigs using SOAPdenovo[16].
3. All BAC assemblies and all the contigs assembled from the WGS sequencing were pooled into the overlap-layout-consensus (OLC) assembly, identical sequences were merged and redundant bases were filtered out from overlapping lengths using software contained within the Rabit package[41]. Then, we obtained the nonredundant super-contig sequences.
4. All the super-contigs obtained from the OLC assembly were linked to scaffolds using all the WGS reads from long inserts (2–40 kb) in a step-by-step fashion using SSPAC[42] with the following command: perl SSPACE_v1-1.pl -l lib.txt -s final3200.fa -k 3 -n 20 -x 0.
5. To fill the gaps (N regions) within the scaffold, all the WGS reads from libraries with short insert sizes were mapped to the scaffold, and the reads located in the gap regions were then locally assembled with the following GapCloser command: GapCloser -a scafSeq -b gap_all.lib -o scafSeq.FG1 -t 24.

Genetic map construction. Genetic linkage maps were constructed for anchoring the scaffolds to 26 chromosomes of the A_tD_t genome, using 167 individual RIL lines derived from an interspecifc cross of the *G. hirsutum* and *G. barbadense* genetic standards TM-1 and 3-79 (ref. 12). Genomic DNA from this mapping population was digested with the restriction endonuclease EcoRI and processed into restriction-site-associated DNA or 'RAD' libraries. Every 24 cotton individuals were pooled into one sequencing library with nucleotide multiplex identifiers (4, 5, 6, 7 and 8 bp), and each sequencing library was barcoded. Approximately 3,940 Mb of 90-bp reads (23 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 GATK pipeline. JoinMap4.0 (ref. 43) was used to conduct the linkage analysis. We used the method described in binmap[44] to combine the adjacent SNP markers into bins that can integrate markers with high quality and filter the mistakes among them. The primary steps were as follows: (i) using MPR (Maximum Parsimony of Recombination) to infer a parental genotype; (ii) using Bayesian inference to acquire reliable parents' genotype; (iii) using HMM (Hidden Markov Model) to correct the progeny genotype. We finally obtained 6,501 bins and utilized them in anchoring the scaffolds together with other SNP markers.

A logarithm of odds (LOD) score of 10 was initially set as the linkage threshold for linkage group identification. All high-quality markers were used to construct the consensus maps with the RIL population in JoinMap4.0 (ref. 43). 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. A scaffold order was determined by tag SNPs, and inner SNPs were then used to orient the scaffolds. A complete set of 26 pseudochromosomes of *G. hirsutum* was produced, with chromosomes 1–13 corresponding to the A_t subgenome and chromosomes 14–26 to the D_t subgenome.

Repeat content. We detected the repeat content of the *G. hirsutum* genome using two methods, the *de novo* method and the homolog search method based on Repbase[45]. Three *de novo* software packages, PILER-DF[46], RepeatModeler and LTR_FINDER[47], were used to predict TE sequences in the genome. Intact LTRs were predicted using LTR_FINDER, and the LTR insert time was calculated according to the rate of spontaneous mutations[27]. The homology-based approach involved applying commonly used databases of known repetitive sequences. We identified TEs at the DNA level with RepeatMasker[48] using Repbase[45] TE. We identified TEs at the protein level with RepeatProteinMask, which runs WuBlastX against the TE protein database. Then we combined the results from the DNA level and protein level and filtered out TE elements for which >80% of their domains were contained within the domain of a higher-scoring match. We then joined the overlapping TE elements that match the same TE in the library.

Tandem repeats in the genome assembly were identified using Tandem Repeat Finder[49] with the following parameters: 2 7 7 80 10 50 2000 -d -h. We also searched noninterspersed repeats in the genome using RepeatMasker[48] with its "-noint" parameter.

Gene prediction. We used *de novo* gene prediction, homology-based methods and RNA-seq data to predict gene structure and integrated all the results by GLEAN[50] to obtain a consensus gene set.

1. *De novo* gene prediction.
   Augustus[51] and GlimmerHMM[52] were used to obtain *de novo*–predicted gene structures from a version of the genome in which repeats were masked based on identified repeat sequences with lengths >500 bp, except for miniature inverted–repeat transposable elements (MITEs), which are usually found near genes or inside introns.
2. Homology-based prediction.
   Protein sequences of five sequenced genomes (*A. thaliana, G. raimondii, G. arboreum, C. papaya and T. cacao*) were aligned to the *G. hirsutum* genome using TBLASTN (E-value ≤ 1E-5). Then the homologous

genome sequences were aligned against the matching proteins using Genewise[53] to generate accurate spliced alignments. The expressed sequence tag sequences of *G. hirsutum* were aligned against the genome sequence using BLAT[54] (identity ≥ 0.95 and coverage ≥ 0.90) to generate spliced alignments. Then, Genewise[53] was applied to exploit the spliced alignments of expressed sequences to automatically model gene structure.

3. RNA-seq–based prediction.
   To assist in gene annotation, we first mapped RNA-seq reads of six transcriptomes from cotton leaf and fiber to the assembled genome using TopHat[55] to identify potential exons in addition to donor or receptor sites with the following parameters: –p 4-max-intron-length 20,000 –m 1 –r 20-mate-std-dev 20. Then we combined TopHat mapping results and applied Cufflinks[55] to predict transcript structures with the following parameters: –I 20,000 –p. To obtain complete gene models, we used the fifth-order Markov model to predict ORFs.
4. Integration of evidence for gene prediction.
   The GLEAN software[50] was used to integrate data derived from the three methods into a GLEAN-derived gene set. Short genes with coding sequences <150 bp and low-quality genes (gaps in >10% of the coding region) were filtered out.

Syntenic and *Ks* analysis. All-versus-all BLASTP (e-value < 1e-5) was used to detect orthologous and paralogous genes among *G. hirsutum, G. arboreum* and *G. raimondii*. Syntenic blocks (with at least five genes per block) were identified by MCScan[29] (MATCH_SCORE: 50, MATCH_SIZE: 5, GAP_SCORE: –3, E_VALUE: 1E-05). 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. Syntenic blocks were used to find homeologous exchanges (HEs) between the two subgenomes of tetraploid cotton by tracing the presence of A genome–derived haplotype blocks in the D_t genome, and D-derived blocks in the A_t genome. *Ks* values along the chromosome between tetraploid and two diploid genomes were calculated in nonoverlapping windows (window size = five genes). HE was defined when one fragment with a few sequent windows from A_t chromosome showed a lower *Ks* value (Student's *t*-test, $P \leq 0.05$) with a fragment from D genome than from A genome, whereas both flanking fragments displayed low *Ks* values with fragments from the A genome, and vice versa. *Ks* (the number of synonymous substitutions per synonymous site) and dN/dS ratios (the ratio of the number of nonsynonymous substitutions per nonsynonymous site to the number of synonymous substitutions per synonymous site) were calculated using yn00 in the PAML software[56] package with the branch model. To distinguish HEs and other DNA blocks with different mutation rates, the distribution of *Ks* values for the whole chromosome was compared with that of the particular HE. Only the DNA blocks with *Ks* values significantly deviating from the distribution of global *Ks* values (Student's *t*-test, $P \leq 0.05$) from both A and D genomes were acknowledged as candidate HEs. To avoid the recognition of possible misassembly as HEs, all the candidate HE regions with 50-kb flanking sequences were analyzed using the paired-end relation of reads from the libraries with insert sizes of 500 bp, 2 kb, 5 kb, 10 kb, 20 kb and 40 kb.

The 4DTv value of the blocks was calculated using a revised version of the HKY model. Gene family is defined here as a group of genes that are determined to descend by OrthoMCL[57] from the last common ancestor of the species under consideration. First, BLASTP was used to compare all the protein sequences with a database containing a protein data set of all species with e-value < 1e-5. Then clustering of genes was carried out with OrthoMCL[57] (inflation parameter: 1.5).

Analysis of ACOs and cell wall biosynthesis genes. *ACO*, *CesA*, *UER*, *UGP* and *UGD* gene families were identified using the HMMER 3.0 software package[58]. MEGA software[59] was used to construct the phylogenetic tree with the neighbor-joining method. Double-stranded DNA sequences corresponding to different fragments of the *ACO1* and *ACO3* promoters, respectively, were synthesized by Invitrogen and labeled with ^{32}P to generate the DNA probes. DNA-protein interactions were performed in 20-µl reactions that contained 1.5 µl 10× binding buffer (12.5 mM HEPES, 50% glycerol, 200 mM KCl, 2.5 mM EDTA, 5 mM DTT), 1 µg poly dI,dC (Sigma-Aldrich), 20 µM labeled DNA probe and 20 µg nuclear extract and were then incubated at room temperature for 1 h. The reactions were analyzed with 8% native PAGE in 0.5× Tris-borate-EDTA at 120 V. DNA binding activity was quantified using the Typhoon 9200 PhosphorImager (GE Healthcare).

To analyze ethylene gas production, 96 freshly collected 1-DPA wild-type ovules from each cotton species were cultured in 96-well plates in the dark at 30 °C for 14 d. Air samples (50 µl) were collected at each time point and injected into a 30-m HP-PLOT column (J&W) and were analyzed with a gas chromatograph (GC-14C; Shimadzu) equipped with a ñame-ionization detector. Ethylene production was determined by comparing with a standard ethylene sample run on the same GC using identical program. QRT-PCR experiments were performed as previously reported.

41. You, M. *et al.* A heterozygous moth genome provides insights into herbivory and detoxification. *Nat. Genet.* **45**, 220–225 (2013).
42. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. & Pirovano, W. Scaffolding preassembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
43. Stam, P. Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. *Plant J.* **3**, 739–744 (1993).
44. Xie, W. *et al.* Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. *Proc. Natl. Acad. Sci. USA* **107**, 10578–10583 (2010).
45. Jurka, J. Repbase Update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000).
46. Edgar, R.C. & Myers, E.W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21** (suppl. 1), i152–i158 (2005).
47. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
48. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics.* Chapter 4, Unit 4.10 (2009).
49. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
50. Elsik, C.G. *et al.* Creating a honey bee consensus gene set. *Genome Biol.* **8**, R13 (2007).
51. Stanke, M. *et al.* AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
52. Majoros, W., Pertea, M. & Salzberg, S. TigrScan and GlimmerHMM: Twoopen source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
53. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome Res.* **14**, 988–995 (2004).
54. Kent, W.J. BLAT: The BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
55. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
56. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).
57. Remm, M., Storm, C.E. & Sonnhammer, E.L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**, 1041–1052 (2001).
58. Finn, R.D., Clements, J. & Eddy, S.R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
59. Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).

doi:10.1038/nbt.3208