nature genetics

The draft genome of a diploid cotton *Gossypium raimondii*

Kunbo Wang[1,6], Zhiwen Wang[2,6], Fuguang Li[1,6], Wuwei Ye[1,6], Junyi Wang[2,6], Guoli Song[1,6], Zhen Yue[2], Lin Cong[2], Haihong Shang[1], Shilin Zhu[2], Changsong Zou[1], Qin Li[3], Youlu Yuan[1], Cairui Lu[1], Hengling Wei[1], Caiyun Gou[2], Zequn Zheng[2], Ye Yin[2], Xueyan Zhang[1], Kun Liu[1], Bo Wang[2], Chi Song[2], Nan Shi[2], Russell J Kohel[4], Richard G Percy[4], John Z Yu[4], Yu-Xian Zhu[3], Jun Wang[2,5] & Shuxun Yu[1]

We have sequenced and assembled a draft genome of *G. raimondii*, whose progenitor is the putative contributor of the D subgenome to the economically important fiber-producing cotton species *Gossypium hirsutum* and *Gossypium barbadense*. Over 73% of the assembled sequences were anchored on 13 *G. raimondii* chromosomes. The genome contains 40,976 protein-coding genes, with 92.2% of these further confirmed by transcriptome data. Evidence of the hexaploidization event shared by the eudicots as well as of a cotton-specific whole-genome duplication approximately 13–20 million years ago was observed. We identified 2,355 syntenic blocks in the *G. raimondii* genome, and we found that approximately 40% of the paralogous genes were present in more than 1 block, which suggests that this genome has undergone substantial chromosome rearrangement during its evolution. Cotton, and probably *Theobroma cacao*, are the only sequenced plant species that possess an authentic *CDN1* gene family for gossypol biosynthesis, as revealed by phylogenetic analysis.

Cotton is one of the most economically important crop plants worldwide. Its fiber, commonly known as cotton lint, is the principal natural source for the textile industry. Approximately 33 million ha (5% of the world's arable land) is used for cotton planting[1], with an annual global market value of textile mills of approximately $630.6 billion in 2011 (MarketPublishers; see URLs). Apart from its economic value, cotton is also an excellent model system for studying polyploidization, cell elongation and cell wall biosynthesis[2–5].

The *Gossypium* genus contains 5 tetraploid (AD_1 to AD_5, $2n = 4\times$) and over 45 diploid ($2n = 2\times$) species (where n is the number of chromosomes in the gamete of an individual), which are believed to have originated from a common ancestor approximately 5–10 million years ago[6]. Eight diploid subgenomes, designated as A to G and K, have been found across North America, Africa, Asia and Australia. The haploid genome size of diploid cottons ($2n = 2\times = 26$) varies from about 880 Mb (*G. raimondii* Ulbrich) in the D genome to 2,500 Mb in the K genome[7,8]. Diploid cotton species share a common chromosome number ($n = 13$), and high levels of synteny or colinearity are observed among them[9–12]. The tetraploid cotton species ($2n = 4\times = 52$), such as *G. hirsutum* L. and *Gossypium barbadense* L., are thought to have formed by an allopolyploidization event that occurred approximately 1–2 million years ago, which involved a D-genome species as the pollen-providing parent and an A-genome species as the maternal parent[13,14]. To gain insights into the cultivated polyploid genomes—how they have evolved and how their subgenomes interact—it is first necessary to have a basic knowledge of the structure of the component genomes. Therefore, we have created a draft sequence of the putative D-genome parent, *G. raimondii*, using DNA samples prepared from Cotton Microsatellite Database (CMD) 10 (refs. 15,16), a genetic standard originated from a single seed (accession D_5-3) in 2004 and brought to near homozygosity by six successive generations of self-fertilization. We believe that sequencing of the *G. raimondii* genome will not only provide a major source of candidate genes important for the genetic improvement of cotton quality and productivity, but it may also serve as a reference for the assembly of the tetraploid *G. hirsutum* genome.

RESULTS

Sequencing and assembly

A whole-genome shotgun strategy was used to sequence and assemble the *G. raimondii* genome. A total of 78.7 Gb of next-generation Illumina paired-end 50-bp, 100-bp and 150-bp reads was generated by sequencing genome shotgun libraries of different fragment lengths (170 bp, 250 bp, 500 bp, 800 bp, 2 kb, 5 kb, 10 kb, 20 kb and 40 kb) that covered 103.6-fold of the 775.2-Mb assembled *G. raimondii* genome (**Supplementary Table 1**). The resulting assembly appeared to cover a very large proportion of the euchromatin of the *G. raimondii* genome. The unassembled genomic regions are likely to contain heterochromatic satellites, large repetitive sequences or ribosomal RNA (rRNA) genes. Using a set of 1,369 molecular markers from a consensus genetic linkage map reported previously[17], 43.8% of the markers (599) were unambiguously located on the assembly, allowing us to anchor 73.2% of the assembled 567.2 Mb on the *G. raimondii* chromosomes (**Supplementary Fig. 1**).

[1]State Key Laboratory of Cotton Biology, Cotton Research Institute, 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]Crop Germplasm Research Unit, Southern Plains Agricultural Research Center, US Department of Agriculture–Agricultural Research Service (USDA-ARS), College Station, Texas, USA. [5]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [6]These authors contributed equally to this work. Correspondence should be addressed to S.Y. (yu@cricaas.com.cn), Jun Wang (wangj@genomics.org.cn) or Y.-X.Z. (zhuyx2@pku.edu.cn).

Received 11 January; accepted 5 July; published online 26 August 2012; doi:10.1038/ng.2371

Table 1 Global statistics for the *G. raimondii* genome assembly and annotation

Categories	Number	N50 (kb)	Longest (Mb)	Size (Mb)	Percent of the assembly
Total contigs	41,307	44.9	0.3	744.4	–
Total scaffolds	4,715	2,284	12.8	775.2	100
Anchored scaffolds	281		12.8	567.2	73.2
Anchored and oriented scaffolds	228		12.8	406.3	52.4
Genes annotated	40,976			115.7	14.9
miRNAs	348			0.04	<0.01
rRNAs	565			0.1	0.01
tRNAs	1,041			0.08	0.01
snRNAs	29			0.1	0.02
Transposable elements	148,740			441.4	57.0

The assembly, performed by SOAPdenovo[18,19], consisted of 41,307 contigs and 4,715 scaffolds and accounted for approximately 88.1% of the estimated *G. raimondii* genome[8] (**Table 1**). Over 73% of the assembly was in 281 chromosome-anchored scaffolds, with 228 of them both anchored and oriented (**Supplementary Fig. 1**). The N50 (the size above which 50% of the total length of the sequence assembly can be found) of contigs and scaffolds was 44.9 kb and 2,284 kb, respectively, with the largest scaffold measuring 12.8 Mb (**Supplementary Table 2**). As indicated by sequencing depth distribution analysis, 98.8% of the assembly was sequenced at 10× coverage (**Supplementary Fig. 2**). Of the 58,061 ESTs (>500 bp in length) reported in *G. raimondii*, 93.4% were identified in the assembly (**Supplementary Table 3**). Sequences of 24 of the 25 randomly selected, completely sequenced *G. raimondii* BAC clones downloaded from GenBank (AC243106–AC243130) were fully recovered from our assembly (**Supplementary Table 4**), supporting the view that the *G. raimondii* genome was assembled properly. Percentagewise, coding regions (exons), introns, DNA transposable elements, long terminal repeats (LTRs) and other repeat sequences made up 6.4%, 6.9%, 4.4%, 42.6% and 13.0% of the total genome content, respectively (**Fig. 1**). On most *G. raimondii* chromosomes, genes were more abundant in the subtelomeric regions (**Fig. 1**), as previously reported for *T. cacao*[20] and *Zea mays*[21]. Transposable elements were distributed largely in gene-poor regions (**Fig. 1**).

Gene content, annotation and analysis of major gene families

Genome annotation was performed by combining results obtained from *ab initio* prediction, homology search and EST alignment. We identified 40,976 protein-coding genes in the *G. raimondii* genome, with an average transcript size of 2,485 bp (GLEAN) and a mean of 4.5 exons per gene (**Table 1** and **Supplementary Table 5**). There were 348 micorRNAs (miRNAs), 565 rRNAs, 1,041 tRNAs and 1,082 small nuclear RNAs (snRNAs) in the *G. raimondii* genome (**Table 1** and **Supplementary Table 6**). Among the annotated genes, 83.69% encode proteins that show homology to proteins in the TrEMBL database, and 69.98% were identified in InterPro (**Supplementary Table 7**). As a result, 71.68% of the predicted genes were supported by at least two methods (**Supplementary Table 8**). Overall, 92.2% (37,780 of 40,976) of predicted coding sequences from the genome were supported by transcriptome sequencing data (**Supplementary Fig. 3**), which showed the high accuracy of *G. raimondii* gene predictions. Compared to the smaller *Arabidopsis thaliana* genome[22], the *G. raimondii* genome had a higher gene number, a similar exon number per gene and a lower mean gene density per 100 kb of genomic DNA sequence.

Comparative analysis of *G. raimondii* with *T. cacao*[20], *A. thaliana*[22] and *Z. mays*[23] showed that these four different plant species possess similar numbers of gene families, with a core set of 9,525 in common (**Supplementary Fig. 4**). Of the 16,113 *G. raimondii* gene families, all but 1,267 were conserved in at least 1 other plant genome (**Supplementary Fig. 4**). Analysis of species- and lineage-specific families identified potential inconsistencies between annotation projects but also reflected genuine biological differences in gene inventories.

Phylogeny, paleohexaploidization and whole-genome duplication

Although large-scale duplication events were predicted to have occurred during *Gossypium* evolution, the number and timing of these genome duplications are still being debated[24–26]. By examining 745 single-copy gene families from 9 sequenced plant genomes (**Supplementary Fig. 5**), we found that *G. raimondii* and *T. cacao* belong to a common subclade and probably diverged from a common ancestor approximately 33.7 million years ago (**Fig. 2a**). *Carica papaya* and *A. thaliana* belong to another subclade that diverged from the *G. raimondii*–*T. cacao* subclade approximately 82.3 million years ago (**Fig. 2a**).

Using substitution per synonymous site (Ks) values obtained from 3,195 paralogous gene pairs in the *G. raimondii* and *T. cacao* genomes, we observed 2 peaks at Ks values of 0.40–0.60 and 1.5–1.90 (**Fig. 2b**). The first peak appeared at approximately 16.6 (13.3–20.0) million years ago, corresponding to the whole-genome duplication event that was previously proposed in the *Gossypium* lineage[25,26]. The second peak appeared at approximately 130.8 (115.4–146.1) million years ago, corresponding to the paleohexaploidization event shared by the eudicots[27,28]. In *T. cacao*, a single peak value between 1.7–1.9 has been reported[20], which corresponds to the second peak observed in *G. raimondii* (**Fig. 2b**), indicating that the paleohexaploidization event

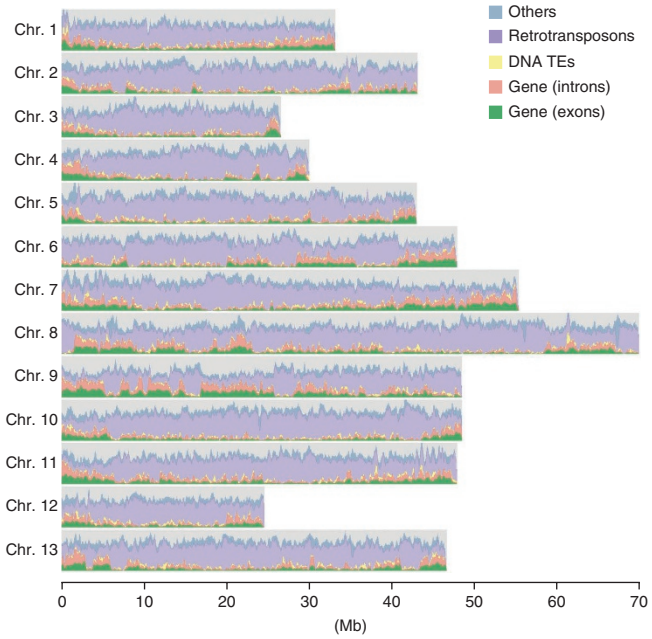

Figure 1 Genomic overview of the 13 assembled *G. raimondii* chromosomes. Major DNA components are categorized into exons, introns, DNA transposable elements (TEs), LTRs (retrotransposons) and other (repeat sequence other than DNA TEs and LTRs). Gray color indicates DNA elements not defined by the previous five terms. All categories were determined for 1.0-Mb windows with a 0.05-Mb shift.

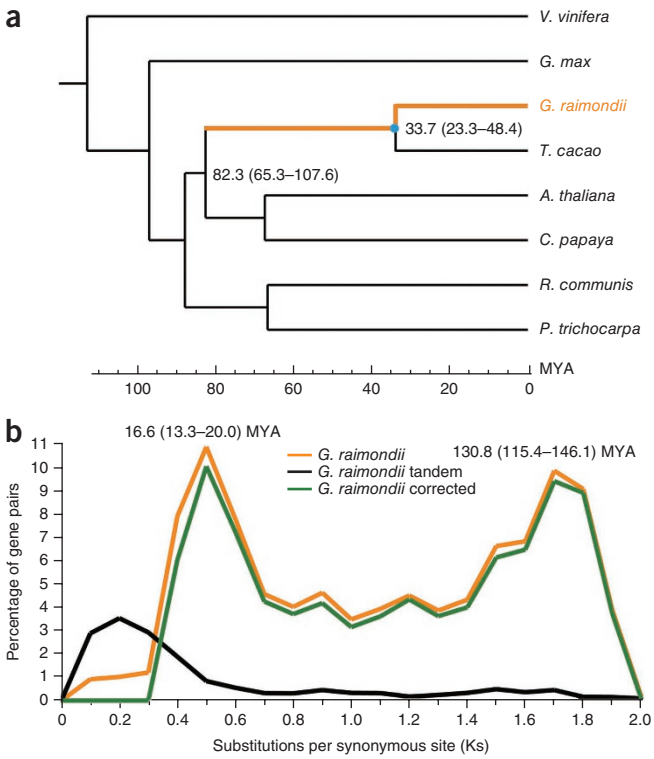

Figure 2 Genome evolution and duplication. (**a**) Phylogenetic analysis showed that *G. raimondii* and *T. cacao* separated approximately 33.7 million years ago (MYA). *O. sativa*, a monocot, was used as the outgroup. (**b**) Ks distributions of *G. raimondii*. Yellow line, Ks of all paralogous gene pairs; black line, Ks of tandem gene pairs only; green line, Ks of all except tandem gene pairs.

shared by the eudicots occurred between 115.4 and 146.1 million years ago in a common progenitor before speciation into the two present-day species 33.7 million years ago.

Comprehensive searches for evidence of whole-genome duplication were performed using an all-versus-all blastp approach comparing the *G. raimondii* and *T. cacao* genomes. Results indicated that the two genomes possess a moderate syntenic relationship, such that 463 collinear blocks (with ≥5 genes per block) covering 64.8% and 74.41% of the assembled *G. raimondii* and *T. cacao* genomes, respectively, are aligned (**Fig. 3a**, **Supplementary Fig. 6** and **Supplementary Table 9a**). Reciprocal best-BLAST-match analysis showed the existence of 133 duplicated and 43 triplicated regions in *G. raimondii* relative to *T. cacao* (**Fig. 3a**). There were 2,355 syntenic blocks among the 13 *G. raimondii* chromosomes. Among these blocks, 21.2% were found to involve only two chromosome regions, 33.7% spanned three chromosome regions and 16.2% traversed four chromosome regions (**Fig. 3b** and **Supplementary Fig. 7**). Chromosome 8 was highly fragmented, with 310 blocks that matched other chromosomes, probably as a result of multiple rounds of duplication, diploidization and chromosomal rearrangement in the genome (**Fig. 3b**). Thirty-nine triplicated chromosomal regions in the *G. raimondii* genome were observed (**Supplementary Table 9b**).

Expansion of transposable elements

Transposable elements are known to contribute substantially to changes in genome size, and they comprise approximately 57% (441 Mb in total length) of the *G. raimondii* genome (**Table 1** and **Supplementary Table 10**). In comparison, 24% of the *T. cacao*[20] genome and 14% of the *A. thaliana* genome are composed of transposable elements[22], suggesting that substantial transposable element proliferation in *G. raimondii* is partially accountable for the expansion of the *G. raimondii* genome. In-depth sequence analysis showed that the most widespread repetitive sequences in the *G. raimondii* genome were the *gypsy* and *copia*-like LTRs, which account for 33.83% and 11.10% of the genome, respectively (**Supplementary Table 10**). The growth rate of these LTR retrotransposons in *G. raimondii* and *T. cacao* tended to slow down after 0.5 and 0.7 million years ago, respectively (**Fig. 4a**). By contrast, the number of LTR retrotransposons has increased in *A. thaliana* since 1.5 million years ago (**Fig. 4a**).

Phylogenetic analysis supported the notion that a larger expansion of specific LTR retrotransposon clades had occurred in *G. raimondii* than in *T. cacao* and *A. thaliana* (**Fig. 4b**). An analysis of the repeat divergence rate distribution (percentage of substitutions in the corresponding region compared with consensus repeats in constructed libraries) independently confirmed the proliferation pattern for LTR retrotransposons in the *G. raimondii* genome (**Supplementary Fig. 8**). Coupled with higher transposable element content in its genome, *G. raimondii* was found to have a higher proportion of genes near (within 1 kb of) transposable elements than *T. cacao* and *A. thaliana* (**Fig. 4c**). By contrast, *T. cacao* maintained the greatest distance between its genes and transposable elements (**Fig. 4c**).

Simple sequence repeats (SSRs) in the *G. raimondii* genome

SSRs behave as polymorphic loci that provide a rich source of markers for cotton breeding as well as for genetic studies. A total of 15,503 di-, tri- and tetranucleotide SSRs, representing 34 distinctive motif families, were identified and annotated in the *G. raimondii* genome (**Supplementary Fig. 9**). We randomly selected 500 of them to study polymorphisms between the mapping parents *G. hirsutum* 'CCRI36'

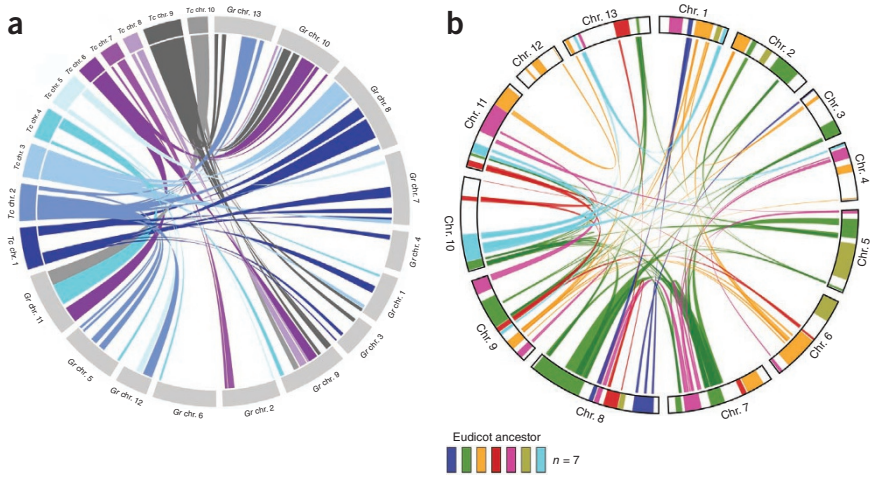

Figure 3 Comparison of syntenic blocks between the genomes of *T. cacao* and *G. raimondii* and reorganization of *G. raimondii* chromosomes. (**a**) Syntenic blocks between *T. cacao* and *G. raimondii*. *Tc*, chromosome of *T. cacao*; *Gr*, chromosome of *G. raimondii*. (**b**) Syntenic blocks among different *G. raimondii* chromosomes. The *G. raimondii* chromosomes are shown in the outer circle in mosaic form, with each color designating its origin from one of the seven ancient chromosomes. Only syntenic blocks longer than 700 kb are shown.

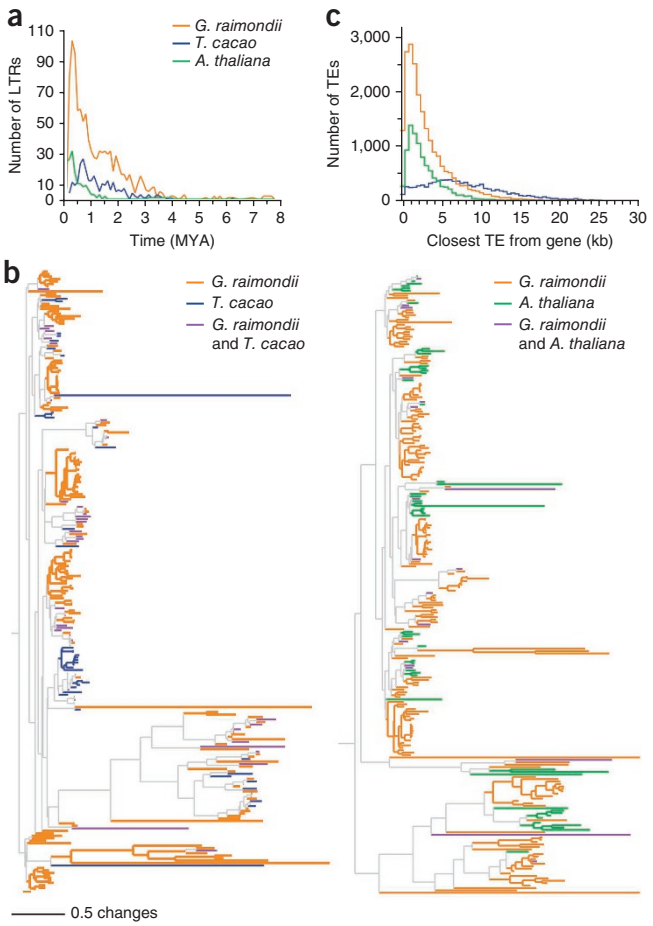

Figure 4 Comparisons of LTRs and transposable elements in the *G. raimondii*, *T. cacao* and *A. thaliana* genomes. (**a**) The distribution curve for the number and insertion time of LTRs in different plant genomes. (**b**) Phylogeny of LTR retrotransposons in the *G. raimondii*, *T. cacao* and *A. thaliana* genomes. (**c**) Distance distributions of nearest transposable elements (TEs) from each gene.

and *G. barbadense* 'Hai1' and found that 70 primer pairs, or 14%, showed polymorphisms. PCR amplification results for 15 of these primer pairs are shown in **Supplementary Figure 10**.

Analysis of genes involved in cotton fiber initiation and elongation

Qualitative transcript differences in key fiber development genes[2,3,29] were found between the non-fibered *G. raimondii* and the fibered *G. hirsutum* species, as revealed by transcriptome (RNA sequencing, RNA-seq) analysis using samples extracted from cotton ovules 3 days post-anthesis (DPA). Of the four sucrose synthase (Sus) genes identified in the genome, three (*SusB*, *Sus1* and *SusD*) were expressed at substantially higher levels in *G. hirsutum* than in *G. raimondii* (**Fig. 5a**). Several 3-ketoacyl-CoA synthase (KCS) genes, including *KCS2*, *KCS13* and *KCS6*, were only expressed in *G. hirsutum*, whereas intermediate levels of *KCS7* transcripts were observed in both *G. hirsutum* and *G. raimondii* (**Fig. 5b**), indicating that high-level expression of Sus and KCS family genes may indeed be required for fiber cell initiation and elongation. By contrast, extremely high amounts of transcripts encoding 1-aminocyclopropane-1-carboxylic acid oxidase (ACO) activities were recovered from *G. raimondii* at the 3-DPA stage (**Fig. 5c**), which is suggestive of a major role for the plant hormone ethylene during early fiber cell development.

Previous researchers have postulated that the cotton fiber is similar in form and origin to plant trichomes, hair-like epidermal cells that occur on various plant organs but are common to leaf and stem surfaces. As postulated, transcription factors that have important roles in *A. thaliana* trichome development may be related to factors involved in cotton fiber formation[4,30,31]. In *A. thaliana*, MYB and bHLH class transcription factors work in a complex in combination with TTG1 to specify a particular epidermal cell fate[30]. A total of 2,706 transcription factors, including 208 bHLH and 219 MYB class genes, were identified in the *G. raimondii* genome (**Supplementary Table 11**). A large number of MYB (**Fig. 5d**) and bHLH (**Fig. 5e**) genes were expressed predominantly in *G. hirsutum* ovules, with only remnant levels found in the ovules of *G. raimondii*, indicating that some of these genes may be required for early fiber development.

Gossypol biosynthesis genes

Cotton is known to produce a unique group of terpenoids that include desoxyhemigossypol, hemigossypol, gossypol, hemigossypolone and the heliocides. Cotton plants accumulate gossypol and related sesquiterpenoids in pigment glands as a defense against pathogens and herbivores. The majority of cotton sesquiterpenoids are derived from a common precursor, (+)-δ-cadinene, which is synthesized by

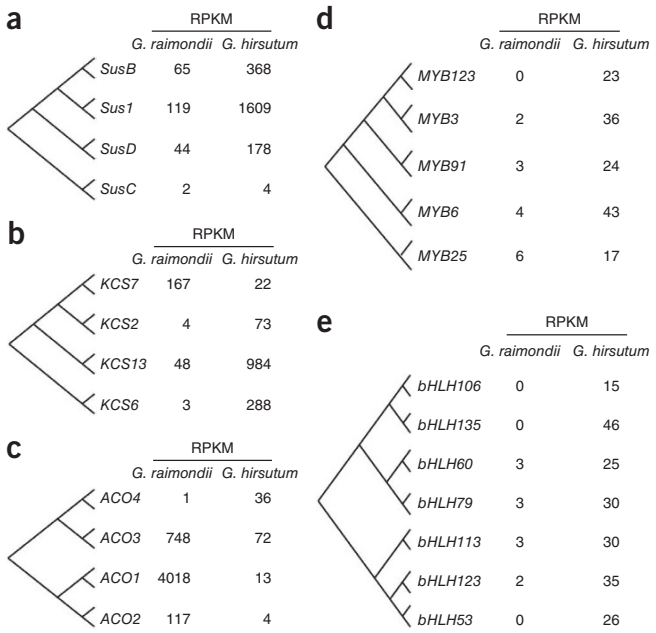

Figure 5 Topological trees and expression patterns of Sus, KCS, ACO, MYB and bHLH family genes in the transcriptome of *G. raimondii* and *G. hirsutum*. (**a**) Major sucrose synthase genes (Sus) were expressed at substantially higher levels in *G. hirsutum* ovules with developing fiber initials than in those of *G. raimondii*. (**b**) Substantially more 3-ketoacyl-CoA synthase (KCS) transcripts were found in *G. hirsutum* ovules. (**c**) Substantially more 1-aminocyclopropane-1-carboxylic acid oxidase (ACO) transcripts were found in *G. raimondii* ovules. (**d**) *G. hirsutum* preferentially expressed MYB transcription factors. (**e**) *G. hirsutum* preferentially expressed bHLH transcription factors. Shown in each panel are the topological tree (left) and comparison of expression levels (right) between the two cotton species. Expression levels were estimated by reads per kilobase of mapped cDNA per million reads (RPKM) values for each gene obtained by sequencing RNA samples from 3-DPA *G. raimondii* and *G. hirsutum* ovules.

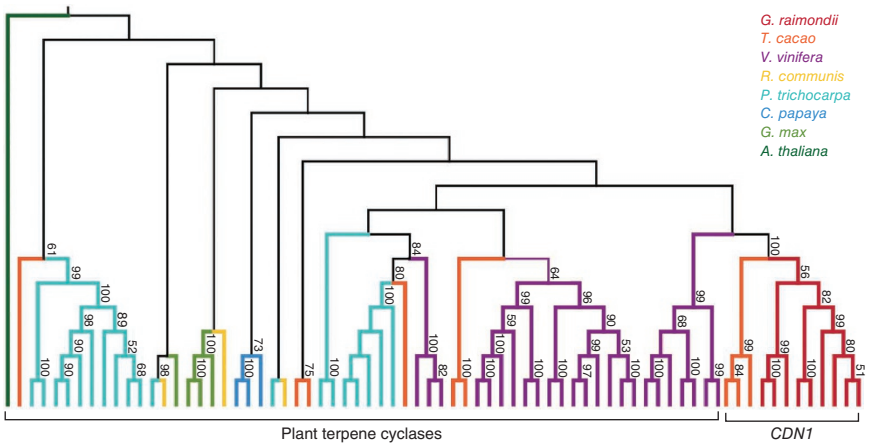

Figure 6 Phylogenetic analysis of the *CDN1* gene family in *G. max*, *P. trichocarpa*, *A. thaliana*, *C. papaya*, *V. vinifera*, *R. communis*, *T. cacao* and *G. raimondii*. The phylogenetic tree and multiple-sequence alignment were established using the neighbor-joining method with Mega 4 software[42]. Bootstrap numbers greater than 50 are shown on the branches.

(+)-δ-cadinene synthase (CDN) via cyclization of farnesyl diphosphate, in the first committed step in gossypol biosynthesis[32,33]. Previously, both CDN-A and CDN-C were reported to encode the proposed enzyme activity[34]. Phylogenetic analysis performed here using *G. raimondii* and eight other sequenced plant genomes, including *T. cacao*[20], *A. thaliana*[22], *Oryza sativa*[23], *C. papaya*[35], *Vitis vinifera*[36], *Populus trichocarpa*[37], *Glycine max*[38] and *Ricinus communis*[39], showed that, except for *O. sativa*, terpene cyclase gene families are common in various plant species (**Fig. 6** and **Supplementary Fig. 11**). However, *G. raimondii* and probably *T. cacao* were the only plant species that possess an authentic *CDN1* gene family with the proposed biochemical function (**Fig. 6** and **Supplementary Fig. 11**). It seemed that the ability to synthesize gossypol is related to both the paleohexaploidization and the whole-genome duplication events that were observed (**Fig. 2b**). No *CDN1* orthologs were found in *P. trichocarpa* or *C. papaya*, the most closely related subclade, suggesting that gossypol production evolved after the separation of these plant species. This conclusion was supported by a recent publication that indicated the key importance of two aspartate-rich Mg^{2+}-binding motifs, DDtYD and DDVAE, for gossypol biosynthesis[40]. All other plant terpene cyclase genes do not encode proteins with the DDVAE motif and thus cannot be recognized as CDN orthologs.

DISCUSSION

We have sequenced the genome of *G. raimondii* using a next-generation Illumina paired-end sequencing strategy, yielding an assembled sequence with 103.6-fold genome coverage. The draft sequence covered 88.1% of the estimated *G. raimondii* genome size. Compared with other sequenced plant genomes, *G. raimondii* showed substantially lower gene density with a high proportion of transposable elements despite being one of the smallest *Gossypium* genomes. One independent whole-genome duplication event occurred approximately 13.3 to 20.0 million years ago, and one paleohexaploidization event that is commonly found in eudicots was clearly observed in the *G. raimondii* genome. The dates of these events reported here agree with those proposed in previous studies[25,26]. *G. hirsutum*, an allotetraploid species, is believed to be the product of a hybridization of two parental diploid species with A and D genomes[41]. An average Ks value of 0.042 was previously reported for tetraploid formation on the basis of an analysis of 42 pairs of paralogous *G. hirsutum* genes[24].

Qualitative differences were found for genes encoding Sus, KCS and ACO activities by comparing the transcriptomes of fiber-bearing *G. hirsutum* and the non-fibered *G. raimondii*. These results indicate that Sus, KCS and ACO are necessary for cotton fiber development, as was proposed in previous individual studies[2,3,29]. Also, the MYB and bHLH transcription factors preferentially expressed in fiber reported herein may be used to elucidate the molecular mechanisms governing fiber initiation and early cell growth. Greater understanding of gossypol and related sesquiterpenoid biosynthesis genes may enable engineering of these genes for better defense against pathogens and herbivores in the cotton field. We suggest that sequencing of the *G. raimondii* genome is a major step toward fully deciphering and analyzing the genomes of the *Gossypium* family to improve cotton productivity and fiber quality.

URLs. Genome browser for *G. raimondii* at the Cotton Genome Project, http://cgp.genomics.org.cn/ *G. raimondii* genome sequencing data at NCBI BioProject, http://www.ncbi.nlm.nih.gov/bioproject/?term=%20PRJNA82769; MarketPublishers, http://marketpublishers.com/; CocoaGen DB, http://cocoagendb.cirad.fr/; *Arabidopsis* Information Resource, http://www.arabidopsis.org/; The Rice Annotation Project Database, http://rapdb.dna.affrc.go.jp/; The Hawaii Papaya Genome Project, http://asgpb.mhpcc.hawaii.edu/papaya/; genome assembly of *V. vinifera*, http://www.genoscope.cns.fr/spip/Vitis-vinifera-e.html; genome assembly of *G. max*, http://www.phytozome.net/soybean; Castor Bean Genome Database, http://castorbean.jcvi.org/; genome assembly of *P. trichocarpa*, http://www.phytozome.net/poplar; SOAPdenovo, http://soap.genomics.org.cn/; estclean, https://sourceforge.net/projects/estclean/; SSPACE, http://www.baseclear.com/landingpages/sspacev12/.

METHODS

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

Accession codes. *G. raimondii* genome sequencing data are available at NCBI BioProject under accession PRJNA82769. Sequencing data for *G. raimondii* and *G. hirsutum* transcriptome analyses are available in the NCBI Sequence Read Archive (SRA) under accessions SRA048621 and SRA048874.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank X.-Y. Chen for his valuable criticisms and suggestions to the manuscript. This work was supported by a grant from the China National Basic Research Program (grant 2010CB126000) and by the National Natural Science Foundation of China (grant 90717009).

AUTHOR CONTRIBUTIONS

K.W., F.L., G.S. and Z.W. designed the analyses. L.C., S.Z., B.W., Junyi Wang, Y. Yin, C.S. and N.S. performed sequencing, assembly and genome annotation. K.W., Z.W., F.L., Jun Wang, W.Y., C.G., Y. Yuan and Z.Y. managed the project. Y. Yuan, H.S., C.Z. and Q.L. performed the genome assembly and physical map integration. C.L., H.W., C.Z., H.S., K.L., X.Z. and Z.Z. prepared DNA and RNA samples. R.J.K., R.G.P. and J.Z.Y. conceived the project, provided the homozygous seeds and revised the manuscript. Y.-X.Z., H.S., C.Z. and

Q.L. performed transcriptome and linage-specific gene functional analyses. Y.-X.Z., H.S., C.Z., Q.L. and W.Y. wrote and revised the manuscript. S.Y. conceived and directed the project.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2371.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Jia, S. *et al. Transgenic Cotton*. Ch. 1, 8–17 (Science Press, Beijing and New York, 2006).
2. Ruan, Y.L., Llewellyn, D.J. & Furbank, R.T. Suppression of sucrose synthase gene expression represses cotton fiber cell initiation, elongation, and seed development. *Plant Cell* **15**, 952–964 (2003).
3. Shi, Y.H. *et al.* Transcriptome profiling, molecular biological, and physiological studies reveal a major role for ethylene in cotton fiber cell elongation. *Plant Cell* **18**, 651–664 (2006).
4. Wang, S. *et al.* Control of plant trichome development by a cotton fiber MYB gene. *Plant Cell* **16**, 2323–2334 (2004).
5. Qin, Y.M. & Zhu, Y.X. How cotton fibers elongate: a tale of linear cell-growth mode. *Curr. Opin. Plant Biol.* **14**, 106–111 (2011).
6. 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).
7. Hawkins, J.S. *et al.* Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Res.* **16**, 1252–1261 (2006).
8. Hendrix, B. & Stewart, J.M. Estimation of the nuclear DNA content of *Gossypium* species. *Ann. Bot.* **95**, 789–797 (2005).
9. Rong, J. *et al.* A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (*Gossypium*). *Genetics* **166**, 389–417 (2004).
10. Reinisch, A.J. *et al.* A detailed RFLP map of cotton, *G. hirsutum* × *G. barbadense*: chromosome organization and evolution in a disomic polyploidy genome. *Genetics* **138**, 829–847 (1994).
11. Brubaker, C.L., Paterson, A.H. & Wendel, J.F. Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. *Genome* **42**, 184–203 (1999).
12. Desai, A., Chee, P.W., Rong, J., May, O.L. & Paterson, A.H. Chromosome structural changes in diploid and tetraploid A genomes of *Gossypium*. *Genome* **49**, 336–345 (2006).
13. Sunilkumar, G., Campbell, L.A.M., Puckhaber, L., Stipanovic, R.D. & Rathore, K.S. Engineering cottonseed for use in human nutrition by tissue-specific reduction of toxic gossypol. *Proc. Natl. Acad. Sci. USA* **103**, 18054–18059 (2006).
14. Chen, Z.J. *et al.* Toward sequencing cotton (*Gossypium*) genomes. *Plant Physiol.* **145**, 1303–1310 (2007).
15. Yu, J.Z. A standard panel of *Gossypium* genotypes established for systematic characterization of cotton microsatellite markers. *Plant Breeding News* **148**, 1.07 (2004).
16. Blenda, A. *et al.* CMD: a cotton microsatellite database resource for *Gossypium* genomics. *BMC Genomics* **7**, 132–141 (2006).
17. Lin, L. *et al.* A draft physical map of a D-genome cotton species (*Gossypium raimondii*). *BMC Genomics* **11**, 395 (2010).
18. Li, R. *et al.* The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**, 311–317 (2010).
19. Li, R. *et al. De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
20. Argout, X. *et al.* The genome of *Theobroma cacao*. *Nat. Genet.* **43**, 101–108 (2011).
21. Schnable, P.S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
22. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
23. International Rice Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
24. Senchina, D.S. *et al.* Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. *Mol. Biol. Evol.* **20**, 633–643 (2003).
25. Blanc, G. & Wolfe, K.H. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**, 1667–1678 (2004).
26. Fawcett, J.A., Maere, S. & Van de Peer, Y. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc. Natl. Acad. Sci. USA* **106**, 5737–5742 (2009).
27. Tang, H. *et al.* Unraveling ancient hexaploidy through multiple-aligned angiosperm gene maps. *Genome Res.* **18**, 1944–1954 (2008).
28. Van de Peer, Y., Fawcett, J.A., Proost, S., Sterck, L. & Vandepoele, K. The flowering world: a tale of duplications. *Trends Plant Sci.* **14**, 680–688 (2009).
29. Qin, Y.M. *et al.* Saturated very-long-chain fatty acids promote cotton fiber and *Arabidopsis* cell elongation by activating ethylene biosynthesis. *Plant Cell* **19**, 3692–3704 (2007).
30. Larkin, J.C., Oppenheimer, D.G., Lloyd, A.M., Paparozzi, E.T. & Marks, M.D. Roles of the *GLABROUS1* and *TRANSPARENT TESTA GLABRA* genes in *Arabidopsis* trichome development. *Plant Cell* **6**, 1065–1076 (1994).
31. Walford, S.-A., Wu, Y., Llewellyn, D.J. & Dennis, E.S. GhMYB25-like: a key factor in early cotton fibre development. *Plant J.* **65**, 785–797 (2011).
32. Essenberg, M., Grover, P.B. Jr. & Cover, E.C. Accumulation of antibacterial sesquiterpenoids in bacterially inoculated *Gossypium* leaves and cotyledons. *Phytochemistry* **29**, 3107–3113 (1990).
33. Chen, X.Y., Chen, Y., Heinstein, P. & Davisson, V.J. Cloning, expression, and characterization of (+)-δ-cadinene synthase: a catalyst for cotton phytoalexin biosynthesis. *Arch. Biochem. Biophys.* **324**, 255–266 (1995).
34. Chen, X.Y., Wang, M., Chen, Y., Davisson, V.J. & Heinstein, P. Cloning and heterologous expression of a second (+)-δ-cadinene synthase from *Gossypium arboreum*. *J. Nat. Prod.* **59**, 944–951 (1996).
35. Ming, R. *et al.* The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452**, 991–996 (2008).
36. French-Italian Public Consortium for Grapevine Genome Characterization. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
37. Tuskan, G.A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596–1604 (2006).
38. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
39. Chan, A.P. *et al.* Draft genome sequence of the oilseed species *Ricinus communis*. *Nat. Biotechnol.* **28**, 951–956 (2010).
40. Gennadios, H.A. *et al.* Crystal structure of (+)-δ-cadinene synthase from *Gossypium arboretum* and evolutionary divergence of metal binding motifs for catalysis. *Biochemistry* **48**, 6175–6183 (2009).
41. Wendel, J.F. & Cronn, R.C. Polyploidy and the evolutionary history of cotton. *Adv. Agron.* **78**, 139–186 (2003).
42. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599 (2007).

ONLINE METHODS

Germplasm genetic resources. DNA samples of the D genome were obtained from CMD 10 (refs. 15,16), a genetic standard that originated from a single seed (accession D_5-3) in 2004 and was brought to near homozygosity by six successive generations of self-fertilization in the greenhouse. *G. raimondii* D_5-3 (CMD 10) was maintained in the nursery on the China National Wild Cotton Plantation in Sanya, and the *G. hirsutum* genetic standard, TM-1 (CMD 1), was grown under standard greenhouse conditions with the temperature maintained at 32 °C during the day time. Fresh young leaves were collected, immediately frozen in liquid nitrogen and stored at −80 °C until DNA extraction.

DNA extraction, library construction and sequencing. We used the standard phenol/chloroform method for DNA extraction, with RNase A and proteinase K treatment to prevent RNA and protein contamination. The extracted DNA was then precipitated with ethanol. Genomic libraries were prepared following the manufacturer's standard instructions and sequenced on the Illumina HiSeq 2000 platform. To construct the paired-end libraries, DNA was fragmented by nebulization with compressed nitrogen gas, the DNA ends were blunted and an A base was added to the 3′ ends. 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 targeted for each specific insert size and were purified from the gels (Qiagen Gel Extraction kit, 28704). We constructed *G. raimondii* genome sequencing libraries with insert sizes of 170 bp, 250 bp, 500 bp, 800 bp, 2 kb, 5 kb, 10 kb, 20 kb and 40 kb.

Genome assembly. The *G. raimondii* genome was assembled using SOAPdenovo with a *K*-mer of 41 and SSPACE software. We first assembled the reads with short insert size (<2 kb) to obtain long contigs. Then, the reads with long insert sizes (<40 kb) were aligned to the contigs to form scaffolds. Finally, we used the paired-end relationships of 40,000 library reads to construct super-scaffolds.

Chromosome anchoring. We aligned the marker sequences from the cotton consensus map[17] to the scaffolds using blastn (identities ≥ 95%; *e* value ≤ 1.0×10^{-6}; coverage ≥ 85%), and the best-scoring match was chosen in cases of multiple matches.

Genome synteny and whole-genome duplication analysis. We use blastp (identity ≥ 40%; *e* value ≤ 1.0×10^{-5}; match length of more than 100 amino acids) to detect paralogous genes in *G. raimondii* and *T. cacao*, and we applied OrthoMCL to detect gene families[43]. For each paralogous gene family, the Ks of each pair was calculated using the PAML package[44], and the median was selected to represent the Ks of the family.

RNA-seq analysis. Total RNA was isolated from 0-DPA ovules, 3-DPA ovules of *G. raimondii* D_5-3 (CMD 10) and *G. hirsutum* TM-1 (CMD 1) and from mature leaves of *G. raimondii* D_5-3 (CMD 10). Normalized pools were converted to full-length enriched cDNA using the SMART method and were sequenced using Illumina protocols. All reads were filtered to trim the adaptor sequences using estclean. Clean reads (with at least 20 nucleotides remaining after trimming) were then mapped to the *G. raimondii* gene models using CLC Genomics Workbench software 4 (CLC bio A/S Science), and matches were converted to RPKM to estimate gene expression levels.

43. Li, L., Stoeckert, C.J. Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
44. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

doi:10.1038/ng.2371