Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres

Andrew H. Paterson1, Jonathan F. Wendel2, Heidrun Gundlach3, Hui Guo1, Jerry Jenkins4,5, Dianchuan Jin6, Danny Llewellyn7, Kurtis C. Showmaker8, Shengqiang Shu4, Joshua Udall1, Mi-jeong Yoo7, Robert Byers8, Wei Chen9, Adi Doron-Faigenboim10, Mary V. Duke11, Lei Gong2, Jane Grimwood4,5, Corrinne Grover1, Kara Grupp1, Guanjing Hu2, Tae-ho Lee1, Jingping Li1, Lifeng Lin1, Tao Liu2, Barry S. Marler1, Justin T. Page1, Alison W. Roberts5,16, Elisson Romaneli1, William S. Sanders10, Emmanuel Szadkowski1, Xu Tan1, Haibao Tang1,14, Chunming Xu1,2, Jinpeng Wang1, Dong Zhang2, Lan Zhang2, Hamid Ashrafi1, Frank Bedon7, John E. Bowers1, Curt L. Brubaker7,17, Peng W. Chee18, Sayan Das1, Alan R. Gingle1, Candace H. Haigler19, Mary V. Duke11, Lei Gong2, Jane Grimwood4,5, Corrinne Grover1, Kara Grupp1, Guanjing Hu2, Tae-ho Lee1, Jingping Li1, Lifeng Lin1, Tao Liu2, Barry S. Marler1, Justin T. Page1, Alison W. Roberts5,16, Elisson Romaneli1, William S. Sanders10, Emmanuel Szadkowski1, Xu Tan1, Haibao Tang1,14, Chunming Xu1,2, Jinpeng Wang1, Dong Zhang2, Lan Zhang2, Hamid Ashrafi1, Frank Bedon7, John E. Bowers1, Curt L. Brubaker7,17, Peng W. Chee18, Sayan Das1, Alan R. Gingle1, Candace H. Haigler19, David Harker9, Lucia V. Hoffmann30, Ran Hovav4, Donald C. Jones33, Cornelia Lemke1, Shahid Mansoor22, Mehboob ur Rahman22, Lisa N. Rainville1, Aditi Rambani9, Umesh K. Reddy23, Jun-kang Rong1, Yehoshua Saranga24, brian E. Scheffler12, Jodi A. Scheffler5, David M. Stelly25, Barbara A. Triplett26, Allen Van Deynze1, Maite F. S. Valenzuela27, Vijay N. Waghmare28, Sally A. Walford1, Robert J. Wright29, Essam A. Zak1,30, Tianzhen Zhang31, Elizabeth S. Dennis2, Klaus F. X. Mayer1, Daniel G. Peterson2, Daniel S. Rokhsar4, Xiyan Wang18,26 & Jeremy Schmutz3,5

Polyplody often confers emergent properties, such as the higher fibre productivity and quality of tetraploid cottons than diploid cottons bred for the same environments1. Here we show that an abrupt five- to sixfold ploidy increase approximately 60 million years (Myr) ago, and allopolyploidy reuniting divergent Gossypium genomes approximately 1–2 Myr ago2, conferred about 30–36-fold duplication of ancestral angiosperm (flowering plant) genes in elite cottons (Gossypium hirsutum and Gossypium barbadense), genetic complexity equalled only by Brassica3 among sequenced angiosperms. Nascent fibre evolution, before allopolyploidy, is elucidated by comparison of spinnable-fibred Gossypium herbaceum and non-spinnable Gossypium longicalyx f genes to one another and the outgroup D genome of non-spinnable Gossypium raimondii. The sequence of a G. hirsutum A_D (in which ‘t’ indicates tetraploid) cultivar reveals many non-reciprocal DNA exchanges between subgenomes that may have contributed to phenotypic innovation and/or other emergent properties such as ecological adaptation by polyploids. Most DNA-level novelty in G. hirsutum recombines alleles from the D-genome progenitor native to its New World habitat and the Old World A-genome progenitor in which spinnable fibre evolved. Coordinated expression changes in proximal groups of functionally distinct genes, including a nuclear mitochondrial DNA block, may account for clusters of cotton-fibre quantitative trait loci affecting diverse traits. Opportunities abound for dissecting emergent properties of other polyploids, particularly angiosperms, by comparison to diploid progenitors and outgroups.

The Gossypium genus is ideal for investigating emergent consequences of polyploidy, A-genome diploids native to Africa and Mexican D-genome diploids diverged ~5–10 Myr ago4. They were reunited ~1-2 Myr ago by trans-oceanic dispersal of a maternal A-genome propagule resembling G. herbaceum to the New World5, hybridization with a native D-genome species resembling G. raimondii, and chromosome doubling (Fig. 1). The nascent A_D allopolyploid spread throughout the American tropics and subtropics, diverging into at least five species; two of these species (G. hirsutum and G. barbadense) were independently domesticated to spawn one of the world’s largest industries (textiles) and become a major oilseed.

New insight into Gossypium biology is offered by a genome sequence of G. raimondii Ublr. (chromosome number, 13) with ~8× longer scaffold N50 (18.8 versus 2.3 megabases (Mb)) compared with a draft1, and oriented to 98.3% (versus 52.4%) of the genome (Supplementary Table 1.3a). Across 13 pseudomolecules totalling 737.8 Mb, ~350 Mb (47%) of euchromatin span a gene-rich 2,059 centimorgan (cM), and ~390 Mb (53%) of heterochromatin span a repeat-rich 186 cM (Supplementary Discussion, sections 1.5 and 2.1). Despite having the least-repetitive DNA of the eight Gossypium genome types, G. raimondii is 61% transposable-element-derived (Supplementary Table 2.1). Long-terminal-repeat retrotransposons (LTRs) account for 53% of G. raimondii, but only 3% of LTR base pairs derive from 2,345 full-length elements. The 37,505 genes and 77,267 protein-coding transcripts annotated (Supplementary Table 2.3 and http://www.phytozome.com) comprise 44.9 Mb (6%) of the genome, largely in distal chromosomal regions (Supplementary Discussion, section 2.1).

Shortly after its divergence from an ancestor shared with Theobroma cacao at least 60 Myr ago,6 the cotton lineage experienced an abrupt five- to sixfold ploidy increase. Individual grape chromosome segments resembling ancestral eudicot genome structure, or corresponding chromosome segments, generally have five

1Plant Genome Mapping Laboratory, University of Georgia, Athens, Georgia 30602, USA. 2Department of Ecology, Evolution and Organismal Biology, Iowa State University, Ames, Iowa 50011, USA. 3MIPS/IBIS Institute for Bioinformatics and System Biology, German Research Center for Environmental Health (GmbH), 85764 Neuberger, Germany. 4Department of Energy Joint Genome Institute, Walnut Creek, California 94595, USA. 4HudsonAlpha Institute of Biotechnology, Huntsville, Alabama 35806, USA. 5Center for Genomics and Computational Biology, School of Life Sciences, and School of Sciences, Hebei United University, Tangshan, Hebei 063000, China. 6CSIRO Plant Industry, Canberra, ACT 2601, Australia. 7Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, Mississippi State, Mississippi 39762, USA. 8Plant and Wildlife Science Department, Brigham Young University, Provo, Utah 84602, USA. 9Department of Field Crops, Plant Science Institute, ARO, Bel-Dagan 50525, Israel. 10CSIRO Plant Industry, Canberra, ACT 2601, Australia. 11Jamie Whitten Delta States Research Center, USDA-ARS, Stoneville, Mississippi 38776, USA. 12Department of Biological Sciences, University of Rhode Island, Kingston, Rhode Island 02881, USA. 13Department of Genetics, Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, 21941-901, Brazil. 14Crag Venter Institute, Rockville, Maryland 20850, USA. 15Laboratory of Molecular Epigenetics of MOE, Unit of Plant Epigenetics, Institute of Genetics & Cytology, Northeast Normal University, Renmin Street, 5268 Changchun, China. 16Plant Reproductive Biology Extension Center, University of California, Davis, California 95616, USA. 17Bayer CropScience, Technologiepark 38, 9052 Gent, Belgium. 18Coastal Plain Experiment Station, University of Georgia, Tifton, Georgia 31793, USA. 19Departments of Crop Science and Plant Biology, North Carolina State University, Raleigh, North Carolina 27695, USA. 20Centro Nacional de Pesquisa em Algodão, EMBRAPA, Santo Antônio de Goiás, GO 73750-000, Brazil. 21Cotton Incorporated, Cary, North Carolina 27513, USA. 22National Institute for Biotechnology & Genetic Engineering, Faisalabad 38000, Pakistan. 23Department of Biology, West Virginia State University, Institute, West Virginia 25512, USA. 24Robert H. Smith Institute of Plant Sciences and Genetics in Agriculture, The Hebrew University of Jerusalem, Rehovot 76100, Israel. 25Department of Soil and Crop Science, Texas A&M University, College Station, Texas 77843, USA. 26Cotton Fiber Bioscience Research, USDA-ARS, New Orleans, Louisiana 70124, USA. 27Department of Microbiology, Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro 21941-971, Brazil. 28Central Institute for Cotton Research, Nagpur, 440010 Maharashtra, India. 29Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing, 210095 Jiangsu, China.
(infrequently six) best-matching *G. raimondii* regions and secondary matches resulting from pan-eudicot hexaploidy\(^7,8\) (Fig. 2 and Supplementary Table 3.1). Paralogous genes tracing to this five- to sixfold ploidy increase show a single peak of synonymous nucleotide-substitution (\(K_s\)) values, suggesting either one, or multiple closely spaced, event(s) (Supplementary Fig. 3.5). Pairwise cytological similarity among A-genome chromosomes\(^9\) suggests the most recent event was a duplication.

Paleopolyploidy may have accelerated cotton mutation rates: for 7,021 co-linearity-supported gene triplets, \(K_s\) rates and non-synonymous nucleotide-substitution (\(K_a\)) rates were, respectively, 19% and 15% larger for cotton–grape than cacao–grape comparisons (Supplementary Table 3.2). Adjusted for this acceleration (Supplementary Fig. 3.5), the cotton ploidy increase occurred about halfway between the pan-eudicot hexaploidy (<125 Myr ago\(^10\)) and the present, near the low end of an estimated range of 57–70 Myr ago\(^11\).

Paleopolyploidy increased the complexity of a Malvaceae-specific clade of Myb family transcription factors, perhaps contributing to the differentiation of epidermal cells into fibres rather than the mucilages of other Malvaceae. Among 204 R2R3, 8 R1R2R3 and 194 heterogeneous Myb transcription factors in *G. raimondii* (Supplementary Table 3.5), subgroup 9 has six members known only in Malvaceae (Fig. 3a), comprising a possible ‘fibre clade’ distinct from the *Arabidopsis thaliana* GL1-like subgroup 15 involved in trichome and root hair initiation and development\(^12\). Expressed predominantly in early fibre development, elite cultivated tetraploid cottons have higher expression of five (50%) of ten subgroup 9 genes compared with wild (undomesticated) tetraploids (Fig. 3a and Supplementary Table 5.3). Some subgroup 9 genes are also active in leaves, hypocotyls and cotyledons (Supplementary Fig. 3.8), consistent with specialization for different types of epidermal cell differentiation such as production of a ‘pulpy layer’ secreted from the teguments surrounding cacao seeds, and mucilages in other Malvaceae fruit (*Abelmoschus* (okra), *Cola* (kola)) and roots (*Althaea* (marshmallow)).

Cotton growers were early adopters of integrated pest management\(^13\) strategies to deploy intrinsic defences conferred by pest- and disease-resistance genes that evolved largely after the 5–6-fold ploidy increase.
NBS-encoding genes. The most NBS-rich (21%) region of T. cacao, on chromosome 7, corresponds to parts of G. raimondii chromosome triplets 2/10/13 and 7/9/4. In total, 27% and 25% of 294 mapped G. raimondii NBS genes are on these parts of chromosomes 7 and 9, often clustered in otherwise gene-poor surroundings (Supplementary Fig. 2.2). Most NBS clusters are species and chromosome specific (Fig. 3b and Supplementary Table 3.7), indicating rapid turnover and/or concerted evolution after cotton paleopolyploidy. In total, 230 (76.7%) NBS-encoding genes have experienced striking mutations (as detailed below) in the A genome since A–F divergence, reflecting an ongoing plant–pathogen ‘arms race’ (Supplementary Table 3.8).

Changes in gene expression during domestication have contributed to the deposition of >90% cellulose in cotton fibres, single-celled models for studying cell wall and cellulose biogenesis. G. raimondii has at least 15 cellulose synthase (CESA) sequences required for cellulose synthesis16 (Supplementary Table 3.3), with four single-gene Arabidopsis clades having three (CESA3, required in expanding primary walls) or two (CESA4, CESA7 and CESA8, each required in the thickening of secondary walls) clade members in G. raimondii16. G. raimondii has at least 35 cellulose-synthease-like (CSL) genes required for synthesis of cell wall matrix polysaccharides that surround cellulose microfibrils16 (Supplementary Table 3.4), including one family (CSL1) absent in Arabidopsis16. Elite tetraploids have higher expression than wild cottons in 6 (40%) of 15 CESA genes and 12 (34%) of 35 CSL genes (Supplementary Table 5.3).

A total of 364 G. raimondii microRNA precursors from 28 conserved and 181 novel families (Supplementary Table 3.12), are predicted regulators of 859 genes enriched for molecule binding factors, catalytic enzymes, transporters and transcription factors (Supplementary Fig. 3.11, 12). Four conserved and 35 novel mRNAs were specifically expressed in G. hirsutum fibres, respectively targeting 53 and 318 genes, most with homology to proteins involved in fibre development (Supplementary Table 3.14, 15). Among 183,690 short interfering RNAs (siRNAs) found, 33,348 (18.15%) were on chromosome 13 (Supplementary Fig. 3.12), a vast enrichment. Small RNA27–19 biogenesis proteins include 13 argonautes, 6 dicer-like (DCL) and 5 RNA-dependent RNA polymerase orthologues (Supplementary Table 3.16). G. raimondii seems to be the first eudicot with two DCL3 genes and two genes encoding RNA polymerase IVa (Supplementary Table 3.16), perhaps relating to control of its abundant retrotransposons.

From unremarkable hairs found on all Gossypium seeds, ‘spinnable’ fibres (fibres with a ribbon-like structure that allows for spinning into yarn) evolved in the A genome after divergence from the B, E and F genomes ~5–10 Myr ago (Fig. 1). To clarify the evolution of spinnable fibres, we sequenced the G. herbaceum A and G. longicalyx F genomes, which respectively differ from G. raimondii by 2,145,177 single-nucleotide variations (SNVs) and 477,309 indels, and 3,732,370 SNVs and 630,292 indels.

Specific genes are implicated in initial fibre evolution by both whole-genome and individual-nucleotide analyses. Across entire genomes, 36 G. herbaceum–G. raimondii and 11 G. herbaceum–G. longicalyx orthologue pairs show evidence of diversifying selection (|ω| > 1, P < 0.05) (Supplementary Table 4.1). A notable example, with G. herbaceum–G. raimondii 0.05, is Gorai.009G035800, a germin-like protein that is differentially expressed between normal and naked-seed cotton mutants during fibre expansion29 and between wild and elite G. barbadense at 10 days post-anthesis (DPA; Supplementary Table 5.3).

Among 114,202 SNVs in 29,015 G. herbaceum genes after G. herbaceum–G. longicalyx A–F divergence (using D as outgroup, so F is the same as D, and A differs from both), we identified striking mutations including 1,090 non-synonymous mutations in 959 genes comprising the most severe 1% of functional impacts inferred using a modified entropy function22; 3,525 frameshift mutations (3,021 genes), which respectively differ from G. raimondii by 1,077 (987) premature stops, 527 (513) splice-site mutations, 102 (102) initiation alterations and 95 (94) extended reading frames (Supplementary Table 4.2, 3). These striking mutations have an average genomic

---

**Figure 3 | Paleo-evolution of cotton gene families.** a. Myb subgroup 9 (ref. 12) originated from a gene on the progenitor of cacao chromosome 2 that formed two adjacent copies after Malvales–Brassicales divergence and then triplicated in cotton, with subsequent loss of one chromosome 8 and two chromosome 12 paralogues. One extant paralogue traces to pan-eudicot hexaploidy, Tc04 g009420, and reduplicated in cotton (Gorai.012G052500.1 and Gorai.011G122800.1) and Arabidopsis (At3g01140 and At5g15310). The other, Tc01 g036330, has reduplicated in cotton (Gorai.004G157600.1 and Gorai.001G169700.1). Asterisk indicates increased gene expression in elite versus wild tetraploids (Supplementary Table 5.3). b. The most NBS-rich region of T. cacao, on chromosome 7, corresponds to regions of G. raimondii chromosome triplets 2/10/13 and 7/9/4. Cacao chromosome 7 NBSs form a single branch, indicating lineage-specific expansion. G. raimondii chromosome 7 and 13 NBSs form distinct branches, indicating cluster/tandem duplication (gene numbers also reflect physical proximity of genes to one another).

A total of 300 (0.8%) G. raimondii genes encode nucleotide-binding site (NBS) domains (Supplementary Table 3.6), largely of coiled-coil (CC)-NBS and CC-NBS-leucine rich repeat subgroups (165, 55%). Like cereals44, after paleopolyploidy G. raimondii evolved clusters of new
distribution (Supplementary Fig. 2.2) but are over-represented in genes coding for cell-wall-associated, kinase or nucleotide-binding proteins (Supplementary Table 4.5).

Striking mutations in the A-genome lineage are enriched \((P = 2.6 \times 10^{-15}; \text{Supplementary Discussion, section 4.4})\) within fibre-related quantitative trait locus (QTL) hotspots in \(A_D\), tetraploid cottons\(^{22}\), suggesting that post-allopolyplody elaboration of fibre development\(^1\) involved recursive changes in \(A_t\) and new changes in \(D_t\) genes. Striking \(A\)-genome mutations have orthologues in 1,051 \(D_t\) and 951 \(A_t\) fibre QTL hotspots. Likewise, sequencing of \(G. hirsutum\) cultivar Acala Maxxa revealed 495 striking mutations in 391 genes, with 83 (21.2\%) in \(D_t\) fibre QTL hotspots and 73 (18.7\%) in \(A_t\) hotspots (Supplementary Table 4.6).

QTL hotspots affecting multiple fibre traits\(^{22}\) may reflect coordinated changes in expression of functionally diverse cotton genes. A total of 671 (1.79\%) genes with >100 reads per million reads were differentially expressed in fibres from wild versus domesticated \(G. hirsutum\) (mostly at 10 DPA) and/or \(G. barbadense\) (mostly at 20 DPA) (Supplementary Table 5.3). Among 48 genes upregulated in domesticated \(G. hirsutum\) at 10 DPA, 20 (42\%) are among 1,582 (4.2\%) genes within QTL hotspot \(D_t09.2\) (ref. 22) affecting length, uniformity, and short-fibre content, with 13 (27\%) out of 677 (1.8\%) genes in homoeologous hotspot \(A_t09\), affecting fibre elongation and fineness. Out of 45 genes downregulated in domesticated \(G. barbadense\) at 20 DPA, 16 (35.6\%) map to \(D_t09.2\), and 8 (17.7\%) to \(A_t09\). In 79\% of cultivated \(G. barbadense\), this \(A_t\) region (which was then thought to be on chromosome 5, and is now known to be on chromosome 9) has been unconsciously introgressed by plant breeders with \(G. hirsutum\) DNA, suggesting an important contribution to productivity of \(G. barbadense\) cultivars\(^{25}\).

A putative nuclear mitochondrial DNA (NUMT) sequence block\(^{24}\) has an intriguing relationship with fibre improvement. A \(G. raimondii\) chromosome 1 region includes many genes closely resembling mitochondrial homologues \((K_s \sim 0.22; \text{Supplementary Table 4.7a})\). NUMT genes experienced a coordinated change in expression associated with \(G. barbadense\) domestication. The 105 (0.2\%) genes upregulated at 10 DPA fibre of wild (versus elite) tetraploid \(G. barbadense\) (Supplementary Table 5.3) include 30 (37\%; \(P < 0.001\)) of the 81 NUMT genes, including 8 NADH dehydrogenase and 4 cytochrome-\(c\)-related genes. All are within the QTL hotspot \(D_t01\) that affects fibre fineness, length, and uniformity\(^{27}\), suggesting a fibre-specific change in electron transfer in \(G. barbadense\) domestication.

Emergent features of polyploids may be related to processes that render them no longer the sum of their progenitors and permit them to explore transgressive phenotypic innovations. Despite the \(A\)-genome origin of spinnable fibres, after 1–2 Myr of co-habitation in tetraploid nuclei most \(A_t\) and \(D_t\) homoeologues are now expressed in fibres at similar levels (Supplementary Table 5.4). Such convergence is not ubiquitous: gene families involved in the synthesis of seed oil show strong \(A\) bias in wild \(G. hirsutum\) and its sister \(G. tomentosum\), but strong \(D\) bias in an improved \(G. hirsutum\) (Supplementary Table 5.6).

Recruitment of \(D_t\) genome genes into tetraploid fibre development\(^1\) may have involved non-reciprocal DNA exchanges from \(A_t\) genes. In the ~40\% of Acala Maxxa \(A_t\) and \(D_t\) genes that differ in sequence from their diploid progenitors (Fig. 4), most mutations are convergent, with \(A_t\) genes converted to the \(D_t\) state at more than twice the rate (25\%) as the reciprocal (10.6\%). Known to occur between cereal paralogues diverged by 70 Myr\(^{24}\), non-reciprocal DNA exchanges are more abundant between cotton \(A_t\) and \(D_t\) genes separated by only ~5–10 Myr\(^1\). Such non-reciprocal exchanges explain prior observations including incongruent gene tree topology for 10\% (3 pairs) of \(G. hirsutum\) and \(D_t\) homoeologues in sequenced bacterial artificial chromosomes (BACs) (Supplementary Discussion, section 5.3); 13.2\% of tetraploid DNA markers that showed different subgenomic affiliations compared with the chromosomes to which they mapped, 9 of 13 being \(D_t\) biased \((A_t \rightarrow D_t)\)\(^{25}\); and expressed-sequence-tag-based evidence of phylogenetic incongruity for as many as 7\% of homeologous genes\(^{26}\).

Several factors may have favoured \(D_t\)-biased allele conversion in tetraploid cotton. The nascent polyploid may have gained fitness from \(D\)-genome alleles native to its New World habitat. Before fortifying its reproductive barriers, the nascent polyploid may have occasionally outcrossed to nearby \(D\)-genome diploids, increasing the likelihood of illegitimate recombination. Outcrossing may also have contributed to the origin of \(Gossypium gossypioides\), sister to \(G. raimondii\) and the only \(D\)-genome cotton containing many otherwise \(A\)-genome-specific repetitive DNAs\(^{27–29}\). \(D_t\)-biased allele conversion may have contributed to slightly greater protein-coding nucleotide diversity in the \(A_t\) compared with the \(D_t\) genome (Supplementary Table 5.7).

Whereas the \(G. raimondii\) reference sequence and \(G. hirsutum\) short-read sequences reveal much about tetraploid cotton genome structure and polyploid evolution, high-contiguity sequencing of polyploids may elucidate still-cryptic features. Tetraploid cotton sequencing appears feasible: among six pairs of \(A_t\) and \(D_t\) BAC clones, the most similar pair shows 99.1% shared \(D_t\) and 97.6% \(A_t\) content (Supplementary Table 6.1), sufficient divergence to de-convolute short-read sequence to the correct subgenome. Increased knowledge of molecular diversity is a foundation for integrating genomics with ecological and field-level knowledge of \(Gossypium\) species and their diverse adaptations to warm arid ecosystems on six continents.

**METHODS SUMMARY**

**Sequencing.** Reads were collected from Applied Biosystems 3730xl, Roche 454 XLR and Illumina Genome Analyzer IIx machines at the Joint Genome Institute
Assembly. A collection of 80,765,952 sequence reads used a modification of Arachne v.2007016, integrating linear (15x genome coverage) and paired (3.1x genome coverage) Roche 454 libraries corrected using 41.9 Gb Illumina sequence, with 1.54x paired-end Sanger sequences from two subclone, six fosmid and two BAC libraries. Cotton genetic and physical maps, and Vitis vinifera and T. cacao synteny were used to identify 51 joins across 64 scaffolds to form the 13 chromosomes (Supplementary Discussion, section 1). The remaining scaffolds were screened for contamination to produce a final assembly of 1,033 scaffolds (19.73M contigs) and 761.4 M. Sequences are in NCBI for G. raimondii (BioProject accession PRJNA171262), G. longicalyx (accession A1-97, SRA061243) and G. hirsutum (cultivar Calca Maxxa, SRS375727) genomes; G. hirsutum (SRA061240) and G. barbadense (SRA061309) fibre transcrip
tions; G. hirsutum (SRA061456) seed transcriptomes; and G. hirsutum microRNAs (SRA061415).

Annotation. PERTRAN software was used to construct transcript assemblies from ~1.1 billion pairs of G. raimondii paired-end Illumina RNA-seq reads, 250 million G. raimondii single end reads, and 150 million G. hirsutum single end reads. PASA™ was used to build transcript assemblies from 454 and Sanger resources (Supplementary Table 2.3). Loci were determined by transcript assembly and/or EXONERATE alignments of A. italiana, cacao, rice, soybean, grape and poplar peptides to repeat-soft-masked G. raimondii genome using RepeatMasker. Gene models were predicted by three homology-based predictors (Supplementary Discussion, section 2.2). Best-scoring gene predictions were improved by PASA, then filtered on the basis of peptide homology or expressed-sequence-tag evidence to remove Pfam transposable element domain models. ClustalW alignments of amino acid sequences (Fig. 3) were used to guide coding sequence alignments. Phylogenetic trees were constructed by bootstrap neighbour-joining with a Kimura 2-parameter model using ClustalW2, assessing internal nodes with 1,000 replicates.

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

Received 17 August; accepted 21 November 2012.

6. Carvalho, M. R., Herrera, F. A., Jaramillo, C. A., Wing, S. L. & Callejas, R. Paleocene and/or EXONERATE alignments of resources (Supplementary Table 2.3). Loci were determined by transcript assembly and/or ClustalW2, assessing internal nodes with 1,000 replicates. Discussion, section 2.2). Best-scoring gene predictions were improved by PASA, then filtered on the basis of peptide homology or expressed-sequence-tag evidence to remove Pfam transposable element domain models. ClustalW alignments of amino acid sequences (Fig. 3) were used to guide coding sequence alignments. Phylogenetic trees were constructed by bootstrap neighbour-joining with a Kimura 2-parameter model using ClustalW2, assessing internal nodes with 1,000 replicates.


Author Information Sequence data have been deposited in NCBI for G. raimondii (BioProject accession PRJNA171262); G. longicalyx (accession A1-97, SRA061660); G. herbaceum (accession A1-97, SRA061243) and G. hirsutum (cultivar Calca Maxxa, SRS375727) genomes; G. hirsutum (SRA061240) and G. barbadense (SRA061309) fibre transcriptions; G. hirsutum (SRA061456) seed transcriptomes; and G. hirsutum microRNAs (SRA061415). Reproductions and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.H.P. (paterson@plantbio.uga.edu) or J.S. (schmich@hudsonalpha.org).

This work is licensed under a Creative Commons Attribution-NonCommercial-Share Alike 3.0 Unported licence. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-sa/3.0

©2012 Macmillan Publishers Limited. All rights reserved.
METHODS

Sequencing. Reads were collected with standard protocols (http://www.jgi.doe.gov/sequencing/protocols/prots_production.html) on Applied Biosystems 3730xl, Roche 454 XLR and Illumina Genome Analyzer (GA)IIx machines at the US Department of Energy Joint Genome Institute. Linear 454 data included standard XLR (47 runs, 16.868Gb) and pre-release FLX+ data (5 runs, 3.262Gb). Eight paired 454, 3–4-kilobase (kb) average insert size and one paired 12-kb average insert size were sequenced on standard XLR (23 runs, 5.931Gb). One standard 400-base pair (bp) fragment library was sequenced at 2 × 150 (7 channels, 41.9Gb) on an Illumina GAIIx. One 2.5-kb average insert size (405,024 reads, 286.1Mb), one 6.5-kb average insert size library (374,125 reads, 263.0Mb), six fosmid libraries (1,222,643 reads, 702.1Mb) of 34–39-kb insert size, and two BAC libraries (107,520 reads, 77.5Mb) of 98-kb and 115-kb (73,728 reads, 48.8Mb) average insert size were sequenced on both ends for a total of 1,283,240 Sanger reads of 1.38Gb of high-quality bases. FLX+ data were collected at the Roche Service Center. BAC end sequence (BES) was collected using standard protocols at the HudsonAlpha Institute.

Genome assembly and construction of pseudomolecule chromosomes. Organellar reads were removed by screening against mitochondria, chloroplast and ribosomal DNA. Any Roche 454 linear read <200 bp was discarded. Roche 454 paired reads in which either was shorter than 500 bp were discarded. An additional de-duplication step was applied to the 454 paired libraries that identifies and retains only one copy of each PCR duplicate. All remaining 454 reads were compared against a full Illumina GA2x run and any insertion/deletions in the 454 reads were corrected to match the Illumina alignments. The sequence reads were assembled using our modified version of Arachne v.20071016 (ref. 31) with parameters maxcliq1 = 100, correct1_passes = 0 and BINGE_AND_PURGE = True. bcall = False maxcliq1 = 200 BINGE_AND_PURGE = True lap_ratio = 0.8 max_bad_look = 1000 (note Arachne error correction is on). This produced 1,263 scaffold sequences, with a scaffold L50 of 25.8Mb, 58 scaffolds larger than 100kb, and total genome size of 761.8Mb. Scaffolds were screened against bacterial proteins, organelle sequences and non-redundant GenBank and removed if found to be a contaminant. Additional scaffolds were removed if they: (1) consisted of >95% 24-nucleotide sequences that occurred four other times in scaffolds larger than 50 kb; (2) contained only unanchored RNA sequences; or (3) were <1kb in length.

The combination of BES/markers hybridized to fingerprint contigs32, 2,800 markers in a genetic map for the D genome in an A_{D}D_{plant}3 and 262 markers from the tetraploid genetic map31, along with Vitis vinifera and T. cacao synteny was used to identify breaks in the initial assembly. Markers were aligned to the assembly using BLAT35 (parameters: -t = dna -q = rna –extendThroughN). BES, physical map contigs, V. vinifera and T. cacao genes were aligned to the genome using BLAST36. Scaffolds were broken if they contained linkage group/syntenic discontinuity coincident with an area of low BAC/fosmid coverage. A total of 13 breaks were executed, and 64 of the broken scaffolds were oriented, ordered and joined using 51 joins to form the final assembly containing 13 pseudomolecule chromosomes. Each chromosome join is padded with 10,000 missing nucleotides. The final assembly contains 1,033 scaffolds (19,735 contigs) that cover 761.4 Mb of the genome with a contig L50 of 135.6kb and a scaffold L50 of 62.2Mb.

The assembly size is near the centre of genome-size estimates of 880 Mb from flow cytometry37, 630 Mb from Feulgen cytophotometry38, and 650 Mb39 and 770 Mb40 from re-naturation kinetics.

Completeness of the euchromatic portion of the genome assembly was assessed using 65,506 G. raimondii complementary DNA sequences obtained from GenBank, which were aligned to the assembly using BLAT37 (parameters: -t = dna -q = rna –extendThroughN). The aim of the completeness analysis was to obtain a measure of completeness of the assembly, rather than a comprehensive examination of gene space. cDNAs were aligned to the assembly using BLAT37 (parameters: -t = dna -q = rna –extendThroughN) and alignments that comprised ≥90% base-pair identity and ≥85% EST coverage were retained. The screened alignments indicate that 57,170 out of 63,506 (90.3%) cDNAs aligned to the assembly. The cDNAs that failed to align were primarily composed of stretches of polynucleotide sequences that failed to generate non-random alignments to any plant or other organism in the NCBI as of the release date.

Annotion. A total of 85,746 transcript assemblies were constructed from about 1.1 billion pairs of D5 paired-end Illumina RNA-seq reads, 55,294 transcript assemblies from 250 million D5 single-end Illumina RNA-seq reads and 62,526 transcript assemblies from 150 million G. hirsutum cotton single-end Illumina RNA-seq reads. All these transcript assemblies were constructed using PERITRAN software (in preparation). In total, 120,929 transcript assemblies were built using PASA32 from 56,638 D5 Sanger ESTs, 2.5 million D5 Roche 454 RNA-seq reads and all of the RNA-seq transcript assemblies. An additional 133,073 transcript assemblies were constructed using PASA from 296,214 G. hirsutum cotton Sanger ESTs and about 2.9 million G. hirsutum cotton 454 reads. The larger number of transcript assemblies from fewer G. hirsutum sequences is due to the fragmented nature of the assemblies. Loci were determined by transcript assembly alignments and/or EXONERATE alignments of peptides from A. thaliana, cacao, rice, soybean, grape and poplar peptides to repeat-soft-masked D5 genome using RepeatMasker (http://www.repetmasker.org) with up to 2,000-bp extensions on both ends, unless extending into another locus on the same strand. Gene models were predicted by homology-based predictors, FGENESH41, FGENESH_EST (similar to FGENESH + EST as splice site and intron input instead of peptide/translated open-reading frames) and GenomeScan42. The best scored predictions for each locus are selected using multiple positive factors including EST and peptide support, and one negative factor: overlap with repeats. The selected gene predictions were improved by PASA. Improvement includes adding untranslated regions, splicing correction, and adding alternative transcripts. PASA-improved gene model peptides were subject to peptide-homology analysis to above-mentioned proteomes in order to obtain Cacore and peptide coverage. Cacore is a peptide BLASTP score ratio maximum best hit BLASTP score and peptide coverage is highest percentage of peptide aligned to the best of homologues. PASA-improved transcripts were selected on the basis of Cacore, peptide coverage, EST coverage and its coding sequence (CDS) overlapping with repeats. The transcripts were selected if their Cacore was larger than or equal to 0.5 and peptide coverage larger than or equal to 0.5, or if it had EST coverage, but its CDS overlapping with repeats was less than 20%. For gene models whose CDS overlaps with repeats for more than 20%, its Cacore needed to be at least 0.9 and homology coverage at least 70% to be selected. The selected gene models were subject to Pfam analysis and gene models whose peptide was more than 30% in Pfam transposable element domains were removed. The final gene set had 37,505 protein-coding genes and 77,267 protein-coding transcripts.