Comparative genome analyses highlight transposon-mediated genome expansion shapes the evolutionary architecture of 3D genomic folding in cotton

Maojun Wang

Huazhong Agricultural University <https://orcid.org/0000-0002-4791-3742>

Jianying Li

National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, China

Pengcheng Wang

National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, China

Zhenping Liu

National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, China

Fang Liu

State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences

Guannan Zhao

National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, China

Zhongping Xu

Huazhong Agricultural University <https://orcid.org/0000-0003-2559-9091>

Kunbo Wang

State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences

Xianlong Zhang (✉ xlzhang@mail.hzau.edu.cn)

National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, China
<https://orcid.org/0000-0002-7703-524X>

Article

Keywords: Comparative genome analyses, highlight, transposon-mediated, genome expansion shapes, evolutionary architecture, 3D genomic folding in cotton

Posted Date: October 30th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-93594/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Molecular Biology and Evolution on May 11th, 2021. See the published version at <https://doi.org/10.1093/molbev/msab128>.

Abstract

Transposable element (TE) amplification has been recognized as a driving force mediating genome size expansion and evolution, but the effect on shaping of 3D genomic architecture remains largely unknown in plants. Here, we report three reference-grade cotton genome assemblies of *Gossypium rotundifolium* (K₂), *G. arboreum* (A₂) and *G. raimondii* (D₅) using Oxford Nanopore sequencing technology. Comparative genome analyses document the details of lineage-specific TE amplification contributing to three-fold change of genome size (K₂, 2.44 Gb; A₂, 1.62 Gb; D₅, 750.19 Mb), and indicate a relatively conserved gene content and synteny relationship among genomes. We found that approximately 17% of syntenic genes exhibit chromatin status switching of A/B compartment, and active TE amplification increases the proportion of A compartment in gene regions in K₂ and A₂ relative to D₅. We also found that only 42% of topologically associating domain (TAD) boundaries were conserved by comparing three genomes, and abundant TE amplification was linked to the organization of lineage-specific TADs. This study sheds light on the regulatory role of transposon-mediated genome expansion in the evolution of higher-order chromatin structure in plants.

Introduction

Transposable elements (TEs) are one kind of DNA sequences that can change the number of their copies or move their position in eukaryotic genome. TE amplification and elimination could affect phenotypic variation, gene transcription, genome evolution and population diversity^{1–4}. As the advance of three-dimensional (3D) genome mapping technologies, a recent study showed that TE could also influence 3D genome architecture, via affecting the organization of cell-specific topologically associating domains (TADs) in mammals⁵. Activation of long terminal repeat (LTR) retrotransposon fascinated the expansion of CCCTC-binding factor (CTCF; a well-known insulator protein for mediating TAD organization) binding sites among mammalian lineages, which promote the formation of TAD boundaries and then influence the transcription of adjacent genes⁶. In plants, such as *Arabidopsis*, maize, tomato and wheat, high-throughput chromosome conformation capture (Hi-C) maps have been used to uncover chromatin organization and detect genomic regulatory elements^{7–10}. In cotton, we established 3D genome architecture in diploids and allotetraploids, and found that the polyploidization process occurring approximately 1.5 million years ago (MYA) contributed to the status transition of A/B compartments and reorganization of TADs¹¹. However, the regulatory effect of remarkable genome size changes through differential TE accumulation on the evolution of higher-order chromatin organization was poorly uncovered in plants.

Cotton is a remarkable textile fiber crop, belonging to the *Gossypium* genus of malvaceae taxonomic family^{12,13}. The *Gossypium* genus contains more than 50 diploid species (A to G and K, $2n = 2 \times = 26$) and 7 tetraploid species (AD₁ to AD₇, $2n = 4 \times = 52$) from a genome allopolyploidization event. All the cotton species were originated from a common ancestor approximately 5–10 MYA¹⁴, and are widely geographically distributed in the world. The largest diploid genome, K genome with a similar size to

tetraploid cotton genome (AD), is estimated about 2600 Mb, representing three-fold of the smallest diploid genome (D)^{15,16}. Therefore, these characteristics show that cotton can serve as an excellent system for studying the evolutionary mechanism of genome polyploidization and genome size expansion. Recently, the assembly of multiple cotton genome sequences helped to uncover TE amplification in the *Gossypium* genus^{12,17-22}. It is found that the *G. arboreum* (A₂) and *G. raimondii* (D₅) genomes shared an early LTR insertion event at ~ 5.7 MYA, and the A₂ genome had an additional recent LTR amplification event at ~ 1 MYA after speciation²⁰. The recent LTR bursting in the A₂ genome was responsible for genome size expansion relative to D₅ genome.

To address the possible role of differential TE amplification in influencing 3D genome organization, we assembled three high-quality genomes for *G. rotundifolium* (K₂), *G. arboreum* (A₂) and *G. raimondii* (D₅) by integrating Oxford Nanopore sequencing technology, paired-end reads and Hi-C technologies. The assembly of reference-grade genomes allowed us to trace the evolutionary footprint of LTR retrotransposons contributing to genome expansion. We revealed the details of differential TE amplification in three genomes during the species divergence, and found that lineage-specific TE amplification was associated with A/B compartment switching and TAD reorganization. This study provides new insights into the TE-mediated chromatin structure changes and informs further evolutionary genomics research.

Results

Assembly of the *G. rotundifolium*, *G. arboreum* and *G. raimondii* genomes

In this study, we applied Oxford Nanopore sequencing technology to assemble *G. rotundifolium* (K₂), *G. arboreum* (A₂) and *G. raimondii* (D₅) genomes. *G. arboreum* and *G. raimondii* genomes have been *de novo* assembled previously using Illumina and PacBio reads^{20,23}, but both genomes have a number of sequence gaps and require an improvement in assembly contiguity. We generated a total of 304 Gb, 212 Gb, 125 Gb Nanopore sequencing data with a genome coverage 124×, 131×, 167 × for K₂, A₂ and D₅, respectively (**Supplementary Table 1**). We assembled 3,593, 1,173 and 366 contigs for *G. rotundifolium*, *G. arboreum* and *G. raimondii* with a contig length of 2.44 Gb, 1.62 Gb and 0.75 Gb, respectively (Table 1). These initial contigs were polished using Illumina paired-end reads with a genome coverage of 108×, 118×, 132 × for K₂, A₂ and D₅. The contig N50 is 5.33 Mb, 11.69 Mb and 17.04 Mb for K₂, A₂ and D₅, respectively. The maximum contig has a length of 32.72 Mb, 58.57 Mb and 43.74 Mb. After polishing contig using Illumina reads, we used high-through chromosome conformation capture (Hi-C) data to order and orient contigs, aimed at constructing pseudo-chromosomes of each species (Fig. 1a **and Supplementary Figs. 1–4**). In the Hi-C assisted assembly, 2,559, 485 and 201 contigs were placed on the 13 chromosomes of K₂, A₂ and D₅ genomes, occupying over 99% of genome length (Fig. 1b).

Table 1
Summary of genome assemblies and annotations of *G. rotundifolium*, *G. arboreum* and *G. raimondii*.

Genomic feature	<i>G. rotundifolium</i>	<i>G. arboreum</i>	<i>G. raimodii</i>
Total length of contigs	2,444,364,209	1,621,008,062	750,197,587
Total length of scaffolds	2,444,484,509	1,621,030,562	750,205,487
Total length of gaps	120,300	22,500	7,900
Percentage of anchoring, bp	99.28%	99.47%	99.57%
Percentage of anchoring and ordering, bp	93.16%	98.84%	99.01%
Number of contigs	3,593	1,173	366
Number of scaffolds	2,390	948	287
Contig N50, bp	5,326,689	11,691,474	17,043,680
Contig N90, bp	621,066	2,910,421	3,537,560
Scaffold N50, bp	177,839,665	129,592,444	57,716,579
Scaffold N90, bp	115,394,628	93,157,762	49,929,625
Maximum contig length, bp	32,728,186	58,575,076	43,739,617
Maximum scaffold length, bp	205,722,655	143,367,608	63,188,200
GC content	36.38%	35.16%	33.23%
Percentage of repeat sequences	80.92%	68.05%	57.04%
Number of genes	41,590	41,778	40,820

To verify the genome assembly completeness, we mapped the clean Illumina reads against each genome, and found that more than 97% of reads were aligned (**Supplementary Table 2**). More than 90% sequencing reads were perfectly mapped, suggesting the high sequence accuracy after base correction of Nanopore reads. We also performed Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis to estimate the assembly completeness in genic regions, which showed that 92.5%, 93.89% and 95.42% of BUSCO hits for K₂, A₂ and D₅ were found (**Supplementary Table 3**). Compared with recently published PacBio assemblies of A₂ and D₅ genomes^{18,20,23}, our assemblies have an improvement of 6.3 fold and 2.7 fold in contiguity, reducing gaps from 1.16 Mb to 22.5 Kb for A₂ and from 17.4 Kb to 7.9 Kb for D₅ (**Supplementary Tables 4–5 and Figs. 5–6**). These genomes should represent reference-grade genomes for the three diploid cotton species.

Genome annotation

We applied three approaches including *ab initio* prediction, homology searches and transcriptome-based analysis, to predict genes in the three genomes (Table 1). Totally, we predicted 41,590, 41,778 and 40,820 genes for K₂, A₂ and D₅ genomes, respectively. The genic regions have a similar length among the three genomes (Fig. 1c). About 27,014, 27,381 and 28,759 genes were transcribed in leaf tissue of K₂, A₂ and D₅ (Fragments Per Kilobase of transcript per Million fragments mapped > 0.1). A total of 41,184 (99.02%), 41,624 (99.63%) and 40,653 (99.59%) protein-coding genes of K₂, A₂ and D₅ genomes were functionally annotated based on InterPro, NR, Swiss-Prot and TAIR10 databases (**Supplementary Table 6**). In the three genomes, 20,782, 11,033 and 6,535 non-coding RNAs were predicted, including 132, 133 and 122 miRNAs (**Supplementary Table 7**). We also predict repeat sequences in the three genomes. The result shows that repeat related sequences have a length of 1,978 Mb, 1,103 Mb and 428 Mb for K₂, A₂ and D₅, occupying 80.92%, 68.05% and 57.04% of genomes, respectively (Fig. 1c). The long terminal repeat (LTR) retrotransposons exhibit higher proportions in the centromeric regions that were predicted using previous centromere-related long LTR regions (**Supplementary Table 8 and Fig. 7**). These data showed that these genomes have a little different length of genic regions, and the large different content of repeat-related sequences in three cotton genomes contributes to the three-fold genome size change.

TE amplification and genome expansion

To explore whether lineage-differential transposon amplification contributes to the genome size variation in these three species, we classified TEs into different categories. We found that 1,761 Mb (72.07%), 1,029 Mb (63.52%) and 366 Mb (48.85%) of transposable elements in K₂, A₂ and D₅ were class I LTRs, of which the vast majority were *Gypsy* (Fig. 2a). In different species, *Gypsy* exhibited a large different copy numbers among three species. Class II DNA transposon occupies 3.8%, 1.8%, and 3.1% of each genome. Compared with the TE number in D₅, three types of class I (*Gypsy*, *DIRS*, *LARD*) and one type of class II (*TIR*) have the most obvious change of copy number in these categories, occupying 80.04%, 62.71%, and 47.22% of all TEs in K₂, A₂ and D₅ genomes (**Supplementary Table 9**). It is observed that *Copia* did not show significant differences between K₂ and A₂ genomes. The amplification of TEs between K₂-A₂ and A₂-D₅ groups led to the larger size of intergenic sequences between genes (Fig. 2b; Mann-Whitney U test, $P < 2.2 \times 10^{-16}$), indicating the role of TEs in enlarging the proportion of genomic non-coding regions. The lineage-differential LTR retrotransposon amplification was responsible for the genome size variation of the three genomes.

To document the details of transposable element amplification, we analyzed full-length LTRs in the three genomes. We identified a total of 26,852, 21,590 and 3,911 full-length LTRs in K₂, A₂ and D₅ genomes respectively (**Supplementary Fig. 8**). These LTRs were clustered into families which could unravel the sequence similarity between different copies. It is found that 30% of LTRs in A₂ were clustered with a family size of > 20. As a comparison, only 12% of LTRs were clustered in K₂ and D₅ (Fig. 2c). This result indicates that LTRs in A₂ have higher sequence similarity than those in K₂ and D₅. To explore whether the sequence similarity is related to the time of LTR burst, we estimated the average age of LTR amplification in each species using the determined mutation rate per year. In K₂, the insertion time peak of LTR

retrotransposons was found at 4.5-5 MYA, while A_2 had a more recent amplification peak at 0.6-1 MYA (**Supplementary Fig. 9**). Furthermore, we found that *Gypsy*, *DIRS*, *LARD* and *TIR* have the largest insertion time in K_2 with a peak of 4.5-5 MYA (Fig. 2d). In agreement with previous estimates, these LTRs in D_5 had a peak of 3–4 MYA and 0.6-1 MYA in A_2 ²⁴. Of note is the observation that *LARD* have two amplification peaks (~ 1 MYA and ~ 4 MYA) in A_2 , of which the older peak was similar to the amplification time in K_2 and D_5 (Fig. 2d). The recent amplification of LTRs in A_2 may explain why LTRs have higher sequence similarity. In addition, these results suggest that K_2 genome has gained a large number of LTRs around 5 MYA compared with the A_2 genome. The phylogenetic tree also supports a huge *Gorge3*^{15,16} expansion of *Gypsy*-like retrotransposon in the clade III by comparing K_2 and A_2 with D_5 (Fig. 2e).

Comparative genomics and evolution

TE amplification has contributed to a three-fold genome size variation, but we don't know whether it has disrupted genomic synteny in gene regions. Here, we identified syntenic blocks between the K_2 genome and the A_2 or D_5 genome (Fig. 3a). We found that 84.11% and 88.78% of the K_2 genome shared genomic synteny in the K_2 - A_2 and K_2 - D_5 comparisons, respectively (**Supplementary Fig. 10**). In these regions, 26,579, 28,372 and 28,485 orthologous genes were included in K_2 , A_2 and D_5 genomes. Analysis of the syntenic block size showed that the K_2 genome has the largest length and D_5 has the smallest length, consistent with the three-fold change of genome size (Fig. 3b). Using the syntenic blocks, we identified a large rearrangement occurring in the Chr01-Chr02 chromosome between K_2 and A_2 genomes. This rearrangement did not occur in the K_2 - D_5 comparison. Since the rearrangement was also found between the *G. herbaceum* (A_1) and *G. arboreum* (A_2) genomes²⁰, it may suggest that this is a A_2 -specific event. Furthermore, comparison of the A_2 genome with other published diploid cotton genomes, including *G. thurberi* (D_1), *G. turneri* (D_{10}) and *G. longicalyx* (F_1), also revealed that this rearrangement only occurred in A_2 (**Supplementary Fig. 11**). We identified another large rearrangement in Chr13-Chr05 in the comparison of K_2 with A_2 and D_5 genomes (Fig. 3a). This rearrangement was supposed to be K_2 -specific by comparing with other diploid genomes (**Supplementary Fig. 12**).

We calculated synonymous substitution values (K_s) for syntenic gene pairs in each genome and between the genomes. We found that all the three species shared a common whole genome duplication (WGD) event occurring approximately 57–71 MYA (Fig. 3c). Analysis of orthologous genes showed that the three species might have undergone lineage divergence at the same period approximately 5.1–5.4 MYA (Fig. 3d), which was also shown between A_2 and D_5 previously²⁴. Further, we found that K_2 , A_2 and D_5 genomes were divergent from their relative species of *Gossypioidea kikii* approximately 8.5–10 MYA²⁵ (**Supplementary Fig. 13**). This suggests a speciation event between *Gossypium* and *Gossypioide* genus. It is estimated that 68%, 86%, 79% of LTRs (corresponding to ~ 5 MYA) emerged during the divergence of three species (**Supplementary Fig. 9a**).

Comparison of gene content in syntenic blocks among different species can reveal evolutionary genome organization. Since the three genomes are divergent from a common ancestor, we explored the extent of gene fractionation after speciation. In the syntenic blocks, 21,173 genes have consistent collinearity in the three genomes; 5,868 genes exhibit collinearity between D_5 and A_2 that are not found in K_2 , 2,736 genes exhibit collinearity between D_5 and K_2 that are not found in A_2 , and 3,972 genes exhibit collinearity between A_2 and K_2 that are not found in D_5 (Fig. 3e). To further analyze gene content at gene family level, we used OrthoMCL to identify gene clusters. The result shows that ~ 15% (unclustered genes and unique paralogs) of genes in each species were unique, these genes might have been under fast evolutionary process or represent lineage-specific genes (Fig. 3f). These results indicate that the three diploid genomes have a change of gene content during the lineage divergence.

Evolution of A/B compartment switching

TE amplification has been recognized as a driver for shaping higher-order chromatin structures in mammals, but we do not know whether it has a similar role in the organization of plant 3D genome. We have known that K_2 and A_2 genomes have gained widespread additional TE insertion relative to the D_5 genome and they shared conserved gene synteny occupying the vast majority of genomes. This provides an opportunity for exploring the effect of TE amplification on the organization of higher-order chromatin structures. We first analyzed the A and B compartments in the three genomes, which represent active and inactive chromatin status. We found that 44.1%, 47.3% and 46.6% of genomes could be categorized as A compartment in K_2 , A_2 and D_5 , respectively. The 55.1%, 52.1% and 53.0% of genomes were regarded as B compartments (Fig. 4a). A chromosome-level visualization showed that each D_5 chromosome tends to have two large A compartment on chromosome arms and one B compartment on the middle of each chromosome (**Supplementary Fig. 14**). However, K_2 and A_2 genomes have more status switching of A/B compartment in TE enriched regions. At the gene level, we noticed that 31,307 (75.2%) genes in K_2 and 31,331 (75.0%) genes in A_2 were located in the A compartment, and 4,431 (10.7%) and 4,518 (10.8%) genes were located in B compartment. In D_5 , 24,267 (59.5%) genes were in A compartment and 11,729 (28.7%) genes were in B compartment (Fig. 4b). These data show that more genes were found in A compartment and less genes in B compartment in the three genomes (Chi-square test, $P < 0.01$). Genes located in A compartment were enriched in some basic biological processes, such as auxin signaling, whereas genes in B compartment were involved in defense response, nucleotide integration and fatty acid metabolic processes (**Supplementary Fig. 15**). At the TE level, 59.2%, 59.1%, 61.9% of TEs in K_2 , A_2 and D_5 were located in B compartment (Fig. 4c). This data indicates that more genes and TEs in D_5 were located in B compartment than those in K_2 and A_2 .

To investigate the change of A/B compartment status among three species, we analyzed the chromatin status in syntenic gene regions. A comparison of homologous syntenic genes shows that 468 genes exhibited A-to-B transition in the comparison of K_2 with A_2 , 3,770 genes exhibited A-to-B transition in the comparison of K_2 with D_5 and 3,765 genes exhibited A-to-B transition in the comparison of A_2 with D_5 . Only 296, 73 and 67 genes exhibited B-to-A compartment switching in the three comparisons (Fig. 4d and

Supplementary Tables 10–12). About 17.4% (3,693/21,173) of syntenic genes in three genomes exhibited A/B compartment switching, 41 and 182 syntenic genes exhibited A-B-A and B-A-B switching in the comparison of K_2 - A_2 - D_5 (Fig. 4e and **Table 13**). To support this finding, genes in chromosome Chr06 were shown. In this chromosome, 32, 33 and 254 genes were located in B compartment in K_2 , A_2 and D_5 , respectively (Fig. 4f and **Supplementary Fig. 16**). To further characterize the biological role of homologous genes with chromatin status switching, we performed a functional enrichment analysis of the A-to-B and B-to-A orthologous genes in K_2 - A_2 and A_2 - D_5 comparisons (Figs. 4g, h and **Supplementary Fig. 17**). These results showed that A-to-B switching genes were enriched in the pathways of ion binding and transcription factor activity, while the B-to-A genes were intriguingly involved in fundamental activity such as ubiquitin transferase activity, pectate lyase activity, ATP binding (adjusted $P < 0.01$).

We then investigate the status of A/B compartment and their transcriptional activity. As expected, we found that genes and TEs in the A compartment display significantly higher expression levels than those in the B compartment (Fig. 4i). Further, in the comparison of K_2 with A_2 or D_5 , we found that the expression patterns of genes in K_2 with A-to-B and B-to-A status switching exhibited significantly higher and lower expression levels (Fig. 4j). This points to a relationship between the switching of chromatin status and the change of transcriptional activity. We also investigated TE activity in 5 Kb flanking regions for these switching genes. It is found that genes showing A-to-B compartment switching have more active TEs in K_2 genome and less in A_2 or D_5 genomes, and vice versa (Fig. 4k and **Supplementary Fig. 18**). This result suggested that active TEs might be involved in the switching of A to B compartment, which is linked to gene transcription.

Evolution of TAD organization

Topologically associating domains (TADs) represent megabase-sized local chromatin interaction domains of physical higher-order chromatin structures, which were separated by boundaries with an enrichment of specific DNA motifs²⁶. Our previous study has shown that cotton genome can be packaged into thousands of TADs in each genome, and polyploidization reshaped the organization of TADs in the comparison of diploids with tetraploid subgenomes¹¹. Using the new genome sequences, we identified TADs for three cotton species. The result showed that there are 2,541, 1,773 and 1,063 TADs ranging from 300 Kb to 3 Mb in K_2 , A_2 and D_5 genomes, occupying 2,187 Mb (96.7%), 1,524 Mb (95.6%) and 686 Mb (92.7%) of genomic length, respectively (**Supplementary Table 14**). The numbers of TADs in A_2 and D_5 were larger than our previous study¹¹, in which the tetraploid subgenomes were used as a reference to identify TADs since no high quality reference genome sequences were available at that time. The average sizes of TADs in K_2 and A_2 genomes are 860 Kb and 861 Kb, while D_5 genome has smaller TADs with an average size of 645 Kb (Fig. 5a). We characterized the gene composition of TAD boundaries that are responsible for TAD organization in K_2 , A_2 and D_5 genomes. The K_2 genome had the smallest gene number in TAD boundaries, while D_5 had the largest gene number (Fig. 5b). As expected, we found that genes in TAD boundaries tend to have significantly higher expression levels than those in TAD interior (Fig. 5c and **Supplementary Fig. 19**), consistent with our previous result¹¹.

The turnover of TAD boundaries indicates the reorganization of TADs in genomes. We compared TAD boundaries in syntenic blocks to explore TAD conservation and renewal in three genomes. We found that 406 TAD boundaries in K_2 were conserved in the comparison of three genomes, and K_2 , A_2 and D_5 genomes have 1,393, 580 and 131 lineage-specific boundaries, respectively (Fig. 5d). To support this finding, a syntenic region between K_2 (Chr08: 81.4–91.7 Mb) and D_5 (Chr08: 29.3–32.4 Mb) was shown. In this block, 5 boundaries were identified in D_5 and 11 boundaries were found in K_2 , of which 6 were specific in K_2 (Fig. 5e). In the comparison of A_2 and K_2 , a syntenic block was shown (Chr07: 70–79.5 Mb for K_2 and Chr07: 62.75–68.45 Mb for A_2). In this block, 7 boundaries were identified in A_2 and 10 boundaries were identified in K_2 , of which 3 were specific in K_2 (Fig. 5f). These showed that the comparison of syntenic blocks could help to identify lineage-specific TAD organization. A further analysis of the specific TAD boundaries showed that there were 69 sequence motifs in K_2 , 8 motifs in A_2 , 4 motifs in D_5 . We identified 13 motifs in conserved boundaries in three genomes (Fig. 5g and **Supplementary Table 15**). For example, K_2 genome has a *PABPC3* (poly(A) binding protein cytoplasmic 3) binding motif in lineage-specific boundaries, A_2 genome has an *AP2* (activating enhancer-binding protein 2) binding motif, and D_5 genome has a *CDF3* (cyclic dof factor 3) binding motif. The conserved boundaries in the three genomes are enriched in a bZIP (basic domain-leucine zipper) binding motif (Fig. 5h). Since gene transcription was found to have a role in the organization of TADs in mammals^{27,28}, we supposed that these transcriptional factor binding motifs might participate in the formation of TADs in each genome, similar to the finding that TCP transcriptional factor binding motif was enriched in TAD boundaries in rice¹⁰.

Effect of transposon amplification on TAD organization

To explore whether TE amplification led to changes of TAD organization in K_2 and A_2 genomes, we investigated TE content in TAD boundaries. It is found that the *Gypsy* LTR retrotransposons that were involved in 99% boundaries of K_2 , A_2 and D_5 , occupy the highest proportion of all TEs in boundaries (Fig. 6a). Of note is the finding that active TEs were enriched in TAD boundaries compared with the whole-genome level, and specific boundaries had a higher proportion of active TEs relative to conserved boundaries in K_2 and A_2 genomes (Figs. 6b, c). This result was coupled with the finding that more specific boundaries were located in A compartment than in B compartment (Fig. 6d). Over time, we classified the intact LTRs retrotransposons in each genome into ancient TEs and young TEs based on the median age of TE insertion within a species. We found that more young LTR retrotransposons were identified in lineage-specific TAD boundaries, and ancient TEs were more likely to exist in conserved boundaries (Fig. 6e, two-sided t-test, $P < 0.001$). In addition, young LTR retrotransposons were found to have higher expression levels than ancient LTR retrotransposons in the three genomes (Fig. 6f). In summary, these results indicated that the amplification of active young TEs in K_2 and A_2 genomes might contribute to the formation of lineage-specific TAD boundaries after divergence of the three species (Fig. 6g).

Discussion

In this study, we sequenced and assembled the first high-quality reference genome of *G. rotundifolium* (K₂), and updated the genome assemblies of *G. arboreum* (A₂) and *G. raimondii* (D₅). Compared with the four available published genome versions of A₂ and D₅^{12,18,20,23}, our assemblies have a considerable improvement in sequence contiguity (N50 reaching 11.69 Mb and 17.04 Mb). We document the details of the observed genome expansion of K₂ that was mainly caused by transposable element proliferation, such as *Gorge3* LTR retrotransposons^{15,16}. Compared with the smallest D₅ genome, the genome expansion of K₂ was deduced around 4.5-5 MYA, and the expansion of A₂ occurred around 0.6-1 MYA, consistent with a previous estimation²⁰. Despite the three-fold change of genome size, the three genomes shared a relatively high level of gene synteny with enlarged intergenic regions. This raises the possibility of TE expansion-reshaped regulatory relationship between non-coding regions and transcription of syntenic coding genes in K₂ and A₂ relative to D₅, on the basis of the recognized important role of non-coding intergenic sequence in transcriptional regulation²⁹. Our assembled K₂ genome, the known largest diploid species of *Gossypium* genus, will lay the foundation for further study of the effect of transposon amplification on genome size variation and the rewired transcriptional regulation.

Previous studies have shown that TE distribution or activity is involved in chromatin interaction in plants. It is found that the large genomes of maize and tomato have extensive chromatin loops, which is linked to A compartment⁸. In *Arabidopsis*, the KNOT engaged element (KEE) regions that represent heterochromatin islands of the 3D genome conformation show a preference for TE insertion, and are involved in the regulation of invasive DNA elements³⁰⁻³². Heat-induced transposon activation in *Arabidopsis* is associated with reduced chromosomal interactions in pericentromeric regions, which is involved in 3D genome reorganization³³. In rice, the density of TEs in H3K9me2-marked regions is higher than those in basal chromatin loop sites, suggesting that H3K9me2 binding sites with higher TE density might be involved in chromatin interactions³⁴. However, few study explored the relationship between TE amplification and 3D genome organization at the scale of genome evolution. In this study, we found that active TEs had a higher frequency in A compartment, and might have a role in the evolutionary switching of B to A compartment accompanied by genome size increase. Meanwhile, we linked active LTR retrotransposon expansion to the formation of lineage-specific TAD boundaries. It is interesting to explore the effect of TAD reorganization on gene transcription by genetic manipulation of these active TEs in boundaries in future study. Specifically, analysis of the transcriptional factor binding sites might help uncover the possible molecular mechanism underlying the formation of new TAD boundaries. In summary, we present some evidence for the evolutionary understanding of higher-order chromatin structure organization in *Gossypium* following activation of LTR retrotransposon amplification, and provide a topological basis for functional analysis of non-coding genomic sequences in complex genomes.

Methods

Cotton materials

Cotton plants of *Gossypium rotundifolium* (accession number K201), *Gossypium arboreum* (cultivar Shixiya-1) and *Gossypium raimondii* (accession number D502) are maintained perennially in National Wild Cotton Nursery and also cultivated in the greenhouse of Huazhong Agricultural University in Wuhan, China. Fresh young leaves were collected individually and immediately frozen in liquid nitrogen.

Library construction and nanopore sequencing

High quality genomic DNA from one single plant was extracted and inspected for purity, concentration and integrity using Nanodrop, Qubit and 0.35% agarose gel electrophoresis, respectively. Large DNA fragments (20–150 Kb) were collected using the BluePippin system. The DNA libraries were constructed using the SQK-LSK109 kit following the standard protocol of Oxford Nanopore Technologies (ONT). Briefly, DNA fragments were subject to optional fragmentation, end repair, ligation of sequencing adapters, tether attachment. The Qubit machine was used to quantify each DNA library. DNA sequencing was performed on the PromethION platform (R9.4.1; FLO-PRO002). The raw data in binary fast5 format from Nanopore sequencing were subject to base calling using the Guppy software in MinKNOW package. The processed reads were subject to removal of sequencing adapters and filtering of reads with low quality and short length (< 2000 bp) and then converted to fastq format for subsequent analysis. For each accession, we also constructed DNA libraries using the NEBNext® Ultra™ DNA Library Prep Kit for sequencing on the Illumina Novaseq 6000 platform (pair-end 150 bp).

Genome assembly and assessment

The Nanopore sequencing reads were corrected using Canu software (v1.3) with parameter corrected ErrorRate with 0.045³⁵. Then, clean reads were subject to *de novo* assembly using wtdbg software. The assembled contigs were calibrated using Racon software and polished with the Illumina sequencing reads using Pilon software (v1.22; parameters: –mindepth 10 –changes –fix bases) for three rounds of running³⁶. In total, we corrected 12,613,188, 6,004,300 and 27,230,681 SNPs, and 17,555,855, 9,185,630 and 31,044,977 InDels in A₂, D₅ and K₂, respectively. To assess the assembly quality, three analyses were performed: the Illumina reads were mapped to contigs using BWA³⁷ software (-mem) and the properly mapped reads were counted using the SAMTools (v0.1.19) software (-flagsstat)³⁸. The assemblies were used to search the CEGMA (v2.5) database, which contains 458 conserved core genes³⁹. The assemblies were aligned to the BUSCO database, which contains 1440 core genes⁴⁰.

Chromosome assembly using Hi-C

Hi-C data were used to construct chromosome-level assemblies for the three genomes. The Hi-C data of *G. arboreum* and *G. raimondii* were from our previous study¹¹, and the Hi-C data of *G. rotundifolium* were generated in this study with the same experimental method (HindIII digestion of chromatin). We performed a pre-assembly for error correction of contigs which required the splitting of contigs into segments of 50 kb on average. Hi-C data were mapped to these fragments and the unique mapping data were retained for the assembly using LACHESIS (v1.0) software⁴¹. Any two segments which showed inconsistent connection with information from the raw contigs were checked manually. The corrected

contigs were used to construct chromosome-level assemblies using LACHESIS with the parameters (CLUSTER_MIN_RE_SITES = 10, CLUSTER_MAX_LINK_DENSITY = 2, CLUSTER_NONINFORMATIVE_RATIO = 2, ORDER_MIN_N_RES_IN_TRUN = 219, ORDER_MIN_N_RES_IN_SHREDS = 216). To assess the assembly quality, the assemblies were split into 100-Kb bins serving as a reference for Hi-C data mapping using HiC-Pro software (v2.7.1)⁴². The placement and orientation errors showing obvious discrete chromatin interaction patterns were manually adjusted. The interaction matrices were shown with heatmaps at a 100-Kb resolution.

Transposon prediction

We used LTR_Finder (v1.07)⁴³ and RepeatScout⁴⁴ (v1.0.5) software with default parameters to construct a repetitive sequence library, representing structure-based prediction and *ab initio* prediction, respectively. The PASTECClassifier (v1.0) was used to classify the sequences in the library that were merged with the Repbase for the final repeat library⁴⁵. This library was used to predict repetitive sequences in each genome using RepeatMasker (-nolow -no_is -norna -engine wublast)⁴⁶.

LTR retrotransposon analysis

The LTR_Finder program with parameter settings (-C -M 0.8) was used to identify full-length LTR retrotransposons in each genome⁴³. The LTR sequences were extracted for the LTR family analysis using the CD-HIT program⁴⁷. For each full-length LTR retrotransposon, the LTR sequences were aligned using MUSCLE (v3.8.1551)⁴⁸ and the divergence distance between them was calculated with a Kimura two parameter (K2P) model using the distmat program embedded in the EMBOSS toolkit⁴⁹. The divergence time was estimated using the formula $T = K/2r$ (where K is the distance between two LTRs and r is the rate of nucleotide substitution per site per year, $r = 3.5 \times 10^{-9}$)¹⁴. The *Gossypium* retrotransposable *gypsy*-like element (*Gorge3*) sequences¹⁵ were aligned against the full-length LTRs from *G. rotundifolium*, *G. arboreum*, *G. raimondii* and *Gossypioides kirkii* (outgroup) using a reciprocal blastn (-e 1e-05) search. The number of full-length *Gorge3* sequences was 1,130 in K_2 , 963 in A_2 , 351 in D_5 and 59 in *Gossypioides kirkii*. MAFFT (v7.453)⁵⁰ was used for *Gorge3* 5' LTR domain with multiple sequence alignments in four species, and then phylogenetic tree was constructed using the IQ-TREE program⁵¹.

Gene prediction

To predict protein-coding genes, three different strategies were adopted, including *ab initio* prediction, homolog-based prediction and transcript-based prediction. Genscan⁵², Augustus (v2.4)⁵³, GlimmerHMM (v3.0.4)⁵⁴, SNAP (v2006-07-28)⁵⁵ were used for *ab initio* prediction. GeMoMa (v1.3.1)⁵⁶ was used for predicting genes based on homologous protein from other species (*Populus trichocarpa*, *Arabidopsis thaliana*, *Vitis vinifera*, *Theobroma cacao* and *Gossypium raimondii*). Hisat2 (v2.0.4)⁵⁷ and Stringtie (v1.2.3)⁵⁸ were used for reference-guided transcript assembly. PASA (v2.0.2)⁵⁹ was used to predict unigene sequences based on RNA-Seq data without reference-guided assembly. Finally, EVM (v1.1.1)⁶⁰ was used to integrate the prediction results obtained by the above three methods and PASA (v2.0.2)⁵⁹

was used to modify gene models. To identify pseudogenes, the GenBlastA (v1.0.4)⁶¹ program was used to scan each genome after masking predicted protein-coding sequences and then the immature stop codon and code shift mutations in the gene sequences were searched by GeneWise (v2.4.1)⁶². The functional annotation of genes was performed using InterProScan (v5.0)⁶³ with '-iplookup -goterms' parameter settings, NR (v20190625) with '-evaluate 1e-05 -best_hit_overhang 0.25 -max_target_seqs 5' and TAIR10 database. Gene ontology (GO) enrichment analysis was performed using a Fisher's exact test method.

Identification of centromeric regions

The previously identified centromeric regions were from published TM-1 reference genome, which were named as GhCR1-5'LTR, GhCR2-5'LTR, GhCR3-5'LTR and GhCR4-5'LTR^{22,64}. Four 5'LTR sequences were aligned to K₂, A₂ and D₅ genome sequences using MUMmer (v4.0)⁶⁵ with '-c 90 -l 40' parameter, followed by 'delta-filter - 1' to identify unique alignment regions. Then, we manually checked the alignment blocks to filter consensus alignment regions of three genomes for each chromosome. After filtering alignments, we selected 95% confidence interval for the median representing the centromeric region for each chromosome.

Comparative genomes and gene synteny analysis

The genomic sequences of *G. rotundifolium*, *G. arboreum* and *G. raimondii* were aligned using MUMmer (v4.0) with the following parameters: (1) nucmer -max match -c 90 -l 40, (2) delta-filter - 1. The syntenic blocks among the three genomes were constructed using the MCScanX⁶⁶ package with default settings. Each syntenic block has at least five homologous genes. The A₂ and D₅ reference genomes were compared with published genomes from CottonGen website (<https://www.cottongen.org/data/download>) by MUMmer (v4.0.0) and MCScanX. The Chr01-Chr02 large translocation of A₂-specific rearrangement and Chr13-Chr05 large translocation of K₂-specific rearrangement were confirmed by comparing with the published A₁, D₁, D₁₀ and F₁ genomes^{20,23,67}.

Analysis of A and B compartments

The Hi-C data of each species were aligned to the respective genome using HiC-Pro software. The valid interaction reads were used to construct the heatmaps of each chromosome at the resolution of 20 Kb, 50 Kb and 100 Kb. The raw contact maps were normalized using a sparse-based implementation of the iterative correction method embedded in HiC-Pro (v2.11.1)⁴². The A and B compartments were identified using the 50 Kb interaction matrix of each chromosome. The principal component analysis (PCA) method in HiTC (v1.0)⁶⁸ package was used to identify A and B compartments. The A compartment usually contains more genes and less transposable elements than does the B compartment. To analyze the A/B compartment status of homologous gene regions, genomic sequences of gene-body, upstream and downstream 2 Kb were extracted.

Analysis of topologically associating domains

The topologically associating domains of each species were identified using the HiTAD⁶⁹ software with default settings. In this analysis, the raw chromatin interaction matrix of each chromosome at the resolution of 50 Kb was constructed using HiC-Pro. Each matrix file was transformed into the cooler format using the toCooler tool of HiCPeaks (<https://github.com/XiaoTaoWang/HiCPeaks>). In each species, TADs with a size of 300 Kb-2 Mb were retained for further analysis. To identify conserved and lineage-specific TADs, we compared TAD boundaries located in syntenic blocks from the results of MCSScanX. Conserved boundaries were defined as those with a maximum boundary change of 3-resolution distance and sequence similarity supported by the MUMmer alignments between two genomes.

TAD boundary motif analysis

In each genome, the TAD boundary flanking 50 Kb were subjected to predict motifs with the findMotifsGenome.pl program in HOMER (v5.0)⁷⁰ software, with the parameters ‘-len 8,10,12 -size 200’. Then, motifs with cutoffs of $P \leq 0.01$ for known and $P \leq 1e-10$ for *de novo* prediction were selected. We used 1,000 uniformly distributed random genomic regions that did not overlap with TAD boundaries as a control set.

RNA-Seq and data analysis

For each species, total RNA from leaf was extracted using a Spectrum™ Plant Total RNA Kit (Sigma, STRN250). RNA libraries were constructed using the Illumina TruSeq RNA Library Preparation Kit (Illumina, San Diego, CA, USA) and sequenced on the Illumina HiSeq 4000 platform (pair-end 150 bp). The clean RNA sequencing data were mapped to each genome using hisat2 (v2.0.5)⁵⁷ software. The high-quality mapping reads were extracted using SAMTools (v0.1.19; -q 25)³⁸. After filtering PCR duplicates, the remaining reads were used to calculate the expression level of genes using Stringtie (v1.3.0)⁵⁸.

Data Availability

The Nanopore and Illumina sequencing data are available at the NCBI database (BioProject accession PRJNA646849). The genome sequence and annotation can be downloaded from the website <http://cotton.hzau.edu.cn/EN/download.php>.

Declarations

Competing interests

The authors declare no competing interests.

Author contributions

X.Z. and M.W. conceived and designed the project. K.W. and F.L. provided the materials. P.W. performed the Hi-C experiment. G.Z. extracted DNA and RNA samples. M.W. conducted PacBio and Illumina sequencing. M.W., J.L., Z.L., and Z.P. analyzed the sequencing data. M.W. and J.L. wrote the manuscript draft, and X.Z. and K.W. revised it. All authors read and approved the final manuscript.

Acknowledgements

This study was supported by National Transgenic Plant Research of China (2016ZX08005-001) to X.Z. and National Natural Science Foundation of China (31922069) to M.W.

References

1. Chen, J. *et al.* Tracking the origin of two genetic components associated with transposable element bursts in domesticated rice. *Nat. Commun* **10**, 641 (2019).
2. Niu, X.M. *et al.* Transposable elements drive rapid phenotypic variation in *Capsella rubella*. *Proc. Natl Acad. Sci. USA* **116**, 6908–6913 (2019).
3. Stein, J.C. *et al.* Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nat. Genet* **50**, 285–296 (2018).
4. Suh, A. Genome Size Evolution: Small Transposons with Large Consequences. *Curr Biol.* **29**, R241-R243 (2019).
5. Zhang, Y. *et al.* Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat. Genet* **51**, 1380–1388 (2019).
6. Schmidt, D. *et al.* Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335–48 (2012).
7. Concia, L. *et al.* Wheat chromatin architecture is organized in genome territories and transcription factories. *Genome Biol.* **21**, 104 (2020).
8. Dong, P.F. *et al.* 3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments. *Mol Plant* **10**, 1497–1509 (2017).
9. Dong, Q.L. *et al.* Genome-wide Hi-C analysis reveals extensive hierarchical chromatin interactions in rice. *Plant J.* **94**, 1141–1156 (2018).
10. Liu, C., Cheng, Y.J., Wang, J.W. & Weigel, D. Prominent topologically associated domains differentiate global chromatin packing in rice from *Arabidopsis*. *Nat. Plants* **3**, 742–748 (2017).
11. Wang, M. *et al.* Evolutionary dynamics of 3D genome architecture following polyploidization in cotton. *Nat. Plants* **4**, 90–97 (2018).
12. Paterson, A.H. *et al.* Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* **492**, 423–427 (2012).
13. Wendel, J.F. & Grover, C.E. Taxonomy and Evolution of the Cotton Genus, *Gossypium*. (2015).

14. Grover, C.E. *et al.* Re-evaluating the phylogeny of allopolyploid *Gossypium* L. *Mol Phylogenet Evol* **92**, 45–52 (2015).
15. Hawkins, J.S., Kim, H., Nason, J.D., Wing, R.A. & Wendel, J.F. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Res* **16**, 1252–61 (2006).
16. Hawkins, J.S., Proulx, S.R., Rapp, R.A. & Wendel, J.F. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. *Proc. Natl Acad. Sci. USA* **106**, 17811–6 (2009).
17. Chen, Z.J. *et al.* Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. *Nat. Genet* **52**, 525–533 (2020).
18. Du, X. *et al.* Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. *Nat. Genet* **50**, 796–802 (2018).
19. Hu, Y. *et al.* *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. *Nat. Genet* **51**, 739–748 (2019).
20. Huang, G. *et al.* Genome sequence of *Gossypium herbaceum* and genome updates of *Gossypium arboreum* and *Gossypium hirsutum* provide insights into cotton A-genome evolution. *Nat. Genet* **52**, 516–524 (2020).
21. Wang, K. *et al.* The draft genome of a diploid cotton *Gossypium raimondii*. *Nat. Genet* **44**, 1098–103 (2012).
22. Wang, M. *et al.* Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nat. Genet* **51**, 224–229 (2019).
23. Udall, J.A. *et al.* De Novo Genome Sequence Assemblies of *Gossypium raimondii* and *Gossypium turneri*. *G3 (Bethesda)* **9**, 3079–3085 (2019).
24. Naoumkina, M. *et al.* The Li2 mutation results in reduced subgenome expression bias in elongating fibers of allotetraploid cotton (*Gossypium hirsutum* L.). *PLoS One* **9**, e90830 (2014).
25. Udall, J.A. *et al.* The Genome Sequence of *Gossypioides kirkii* Illustrates a Descending Dysploidy in Plants. *Front Plant Sci* **10**, 1541 (2019).
26. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
27. Collombet, S. *et al.* Parental-to-embryo switch of chromosome organization in early embryogenesis. *Nature* **580**, 142–146 (2020).
28. Stadhouders, R., Fillion, G. J. & Graf, T. Transcription factors and 3D genome conformation in cell-fate decisions. *Nature* **569**, 345–354 (2019).
29. Gil, N. & Ulitsky, I. Regulation of gene expression by cis-acting long non-coding RNAs. *Nat. Rev. Genet* **21**, 102–117 (2020).
30. Grob, S., Schmid, M.W. & Grossniklaus, U. Hi-C analysis in *Arabidopsis* identifies the KNOT, a structure with similarities to the flamenco locus of *Drosophila*. *Mol Cell* **55**, 678–93 (2014).

31. Grob, S. & Grossniklaus, U. Invasive DNA elements modify the nuclear architecture of their insertion site by KNOT-linked silencing in *Arabidopsis thaliana*. *Genome Biol.* **20**, 120 (2019).
32. Grob, S. & Grossniklaus, U. Invasive DNA elements modify the nuclear architecture of their insertion site by KNOT-linked silencing in *Arabidopsis thaliana*. *Genome Biol.* **20**, 120 (2019).
33. Sun, L. *et al.* Heat stress-induced transposon activation correlates with 3D chromatin organization rearrangement in Arabidopsis. *Nat. Commun* **11**, 1886 (2020).
34. Zhao, L. *et al.* Chromatin loops associated with active genes and heterochromatin shape rice genome architecture for transcriptional regulation. *Nat. Commun* **10**, 3640 (2019).
35. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722–736 (2017).
36. Walker, B.J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
37. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).
38. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
39. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–7 (2007).
40. Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–2 (2015).
41. Burton, J.N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol* **31**, 1119–25 (2013).
42. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
43. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob DNA* **10**, 48 (2019).
44. Price, A.L., Jones, N.C. & Pevzner, P.A. De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1**, i351-8 (2005).
45. Bao, W., Kojima, K.K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11 (2015).
46. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics Chap. 4*, Unit 4 10 (2009).
47. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–2 (2012).
48. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–7 (2004).

49. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276–7 (2000).
50. Katoh, K. & Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
51. Nguyen, L.T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–74 (2015).
52. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 78–94 (1997).
53. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–7 (2005).
54. Majoros, W.H., Pertea, M. & Salzberg, S.L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–9 (2004).
55. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
56. Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S.O. & Grau, J. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics* **19**, 189 (2018).
57. Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–60 (2015).
58. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol* **33**, 290–5 (2015).
59. Haas, B.J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
60. Haas, B.J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
61. Lun, D.S., Sherrid, A., Weiner, B., Sherman, D.R. & Galagan, J.E. A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data. *Genome Biol.* **10**, R142 (2009).
62. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* **14**, 988–95 (2004).
63. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–40 (2014).
64. Wang, S. *et al.* Sequence-based ultra-dense genetic and physical maps reveal structural variations of allopolyploid cotton genomes. *Genome Biol.* **16**, 108 (2015).
65. Delcher, A.L., Phillippy, A., Carlton, J. & Salzberg, S.L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30**, 2478–83 (2002).
66. Tang, H. *et al.* Synteny and collinearity in plant genomes. *Science* **320**, 486–8 (2008).
67. Grover, C.E. *et al.* The *Gossypium longicalyx* Genome as a Resource for Cotton Breeding and Evolution. *G3 (Bethesda)* **10**, 1457–1467 (2020).

68. Servant, N. *et al.* HiTC: exploration of high-throughput 'C' experiments. *Bioinformatics* **28**, 2843–4 (2012).

69. Wang, X.T., Cui, W. & Peng, C. HiTAD: detecting the structural and functional hierarchies of topologically associating domains from chromatin interactions. *Nucleic Acids Res.* **45**(2017).

70. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576–89 (2010).

Figures

Figure 1

Genome assembly and feature description of *G. rotundifolium* (K2). a, Circos plot showing the chromosome-level features of *G. rotundifolium*. The tracks represent gene density (A), TE density (B), GC content (C), DNA methylation (D), A/B compartment (E), gene expression (F) and homologous gene pairs (G). In each track (A-D, F), the feature data is shown in 1 Mb windows sliding 200 Kb. b, Hi-C matrix of *G. rotundifolium*. In this heatmap, high chromatin contact frequency is shown with red color. c, Genomic components of *G. rotundifolium* (K2), *G. arboreum* (A2) and *G. raimondii* (D5). The data include genomic lengths of exon, intron, TE-related and other genomic regions.

Figure 2

Characterization of TE evolution in *G. rotundifolium* (K2), *G. arboreum* (A2) and *G. raimondii* (D5). a, Length of different TE classes. b, The distance of intergenic sequences between two adjacent genes in the three genomes. c, Clustering of full-length LTRs. d, Estimated insertion time for Gypsy, DIRS, LARD and TIR transposons. e, Polygenetic analysis of Gorge3 transposable elements. The phylogenetic tree

includes 2,503 Gorge3 sequences in four species, including 1,130 in K2, 963 in A2, 351 in D5 and 59 in Gossypioides kirkii.

Figure 3

Genomic synteny and evolution in *G. rotundifolium* (K₂), *G. arboreum* (A₂) and *G. raimondii* (D₅). a, Genome-wide syntenic blocks between *G. rotundifolium* (K₂), *G. arboreum* (A₂) and *G. raimondii* (D₅). The light grey lines indicate collinear blocks in three genomes. The dark grey lines indicate large rearrangements between K₂ versus A₂ and K₂ versus D₅. b, The length of syntenic blocks in *G.*

rotundifolium (K2), *G. arboreum* (A2) and *G. raimondii* (D5) genomes (Wilcoxon rank sum test, $P < 7.774 \times 10^{-6}$). c, Estimation of whole genome duplication time in *G. rotundifolium* (K2), *G. arboreum* (A2) and *G. raimondii* (D5) genomes. d, Estimation of species divergence among three genomes. e, Summary of conserved syntenic genes for three cotton genomes. The grey solid lines and grey dotted lines represent the conserved and lost genes in syntenic blocks, respectively. f, Summary of clustered genes and unique genes in *G. rotundifolium* (K2), *G. arboreum* (A2) and *G. raimondii* (D5) genomes based on the OrthoMCL analysis.

Figure 4

Characterization of A and B compartments in *G. rotundifolium* (K2), *G. arboreum* (A2) and *G. raimondii* (D5) genomes. a, The genomic length of A and B compartments in K2, A2 and D5. b, The number of A and B compartment covered genes in K2, A2 and D5 genome. c, The A and B compartments covered TEs. d, Percentage of conserved genes and A/B compartment switching genes in the K2-A2, K2-D5 and A2-D5 comparisons. e, The number of dynamic orthologous genes showing status switching of A/B chromatin compartments in three cotton genomes. Green boxes show genes in A compartment and yellow boxes

show genes in B compartment. f, Example of A/B compartment transition in Chr06 chromosome. Positive and negative values show A and B compartments, respectively. The chromosome length in each genome was normalized to an equal length. g-h, GO enrichment analysis of A-to-B and B-to-A switching genes in K2-A2 (g) and K2-D5 (h). The annotated genes in two genomes (K2 and A2; A2 and D5) were combined as references for enrichment analysis (Fisher's exact test, $P < 0.01$). The grey dashed line represent significance threshold. Top 10 GO terms of molecular function were shown. i, Gene and TE expression in the A (blue) and B (green) compartments. Two-sided Wilcoxon rank-sum test was used for significance analysis ($***P < 2.2 \times 10^{-16}$). j, The expression levels of switching genes in K2-A2 and K2-D5 groups. k, Active TE ratios in switching genes and their flanking 5 Kb regions.

Figure 5

Specific and conserved TADs in *G. rotundifolium* (K2), *G. arboreum* (A2) and *G. raimondii* (D5) genomes. a, TAD size in K2, A2 and D5 genomes ($***P < 2.2 \times 10^{-16}$). b, Number of genes in TAD boundaries (from -50 Kb to 50 Kb) ($***P < 2.2 \times 10^{-16}$). c, Gene expression in TAD boundaries and TAD interior of K2, A2 and D5 genomes ($***P < 2.2 \times 10^{-16}$). d, The number of specific and conserved TADs in K2, A2 and D5

genomes. K2 is used as the reference coordinate. e, TAD structures of collinear blocks on the Chr08 chromosome between K2 and D5. f, TAD structures of collinear blocks on the Chr07 chromosome between K2 and A2. The grey boxes indicate TADs. In each heatmap, strong interactions are highlighted in red color. The numbers of heatmap represent the conserved TADs in two genomes. The black arrows represent the specific TADs in K2 relative to D5 or A2. g, The number of transcriptional factor binding motifs in specific and conserved TAD boundaries. In this analysis, 380 conserved TAD boundaries were overlapped with promoters, and 1,126, 433, 120 lineage-specific TADs were overlapped with promoters in K2, A2 and D5 genomes, respectively. h, The most significantly enriched sequence motifs in specific and conserved TAD boundaries.

Figure 6

The effect of TE amplification on the reorganization of TAD boundary. a, Relative proportion of each TE category in TAD boundaries in K₂, A₂ and D₅ genomes. b, Percentage of active TEs and whole-genome TEs in TAD boundaries. c, Active TE coverage in specific and conserved TAD boundaries. The significance analysis was performed using a two-sided Wilcoxon rank-sum test ($***P < 2.2 \times 10^{-16}$). d, A/B compartment coverage in specific and conserved TAD boundaries ($***P < 2.2 \times 10^{-16}$). e, The ancient and young LTRs in specific and conserved TAD boundaries. Percentage of TAD boundaries in specific and

conserved TAD boundaries was calculated. LTRs were divided into ancient and young according to species divergence time in each cotton. f, Normalized expression of ancient and young TEs between lineage-specific and conserved TADs ($***P < 2.2 \times 10^{-16}$). g, LTR retrotransposon amplification induced organization of lineage-specific TAD during cotton evolution. The 'I' represents the proportion of specific (green) and conserved (blue) TADs. 'II' represents active (black arrow) lineage-specific LTRs in three genomes. 'III' represents the increase of TAD number. 'IV' represents the A/B compartment switching. Phylogenetic schematics showing the divergence between *G. raimondii* (D5) and *G. arboreum* (A2) at ~5.1 MYA, *G. raimondii* (D5) and *G. rotundifolium* (K2) at ~5.4 MYA. The divergence time between *Gossypium* and *Theobroma cacao* is estimated at ~60 MYA.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFigures119.pdf](#)
- [SupplementaryFigures119.pdf](#)
- [SupplementaryTables115.xlsx](#)
- [SupplementaryTables115.xlsx](#)
- [SupplementaryTables115.xlsx](#)
- [SupplementaryFigures119.pdf](#)