Comparative Genome Analyses Highlight Transposon-Mediated Genome Expansion and the Evolutionary Architecture of 3D Genomic Folding in Cotton

Maojun Wang [1,†], Jianying Li [1,†], Pengcheng Wang,[1] Fang Liu,[2] Zhenping Liu,[1] Guannan Zhao,[1] Zhongping Xu,[1] Liuling Pei,[1] Corrinne E. Grover [3], Jonathan F. Wendel [3,*], Kunbo Wang,[2,*] and Xianlong Zhang [1,*]

[1]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, Hubei, China

[2]State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang, Henan, China

[3]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, USA

[†]These authors contributed equally to this work.

***Corresponding authors**: E-mails: xlzhang@mail.hzau.edu.cn; jfw@iastate.edu; wangkunbo@caas.cn.

Associate editor: Jian Lu

Abstract

Transposable element (TE) amplification has been recognized as a driving force mediating genome size expansion and evolution, but the consequences for shaping 3D genomic architecture remains largely unknown in plants. Here, we report reference-grade genome assemblies for three species of cotton ranging 3-fold in genome size, namely *Gossypium rotundifolium* (K_2), *G. arboreum* (A_2), and *G. raimondii* (D_5), using Oxford Nanopore Technologies. Comparative genome analyses document the details of lineage-specific TE amplification contributing to the large genome size differences (K_2, 2.44 Gb; A_2, 1.62 Gb; D_5, 750.19 Mb) and indicate relatively conserved gene content and synteny relationships among genomes. We found that approximately 17% of syntenic genes exhibit chromatin status change between active ("A") and inactive ("B") compartments, and TE amplification was associated with the increase of the proportion of A compartment in gene regions (~7,000 genes) in K_2 and A_2 relative to D_5. Only 42% of topologically associating domain (TAD) boundaries were conserved among the three genomes. Our data implicate recent amplification of TEs following the formation of lineage-specific TAD boundaries. This study sheds light on the role of transposon-mediated genome expansion in the evolution of higher-order chromatin structure in plants.

Key words: **Gossypium, genome expansion, transposable element, chromatin compartment, TAD reorganization, 3D genome architecture.**

Article

Introduction

Transposable elements (TEs) play important roles in shaping the architecture of eukaryotic genomes. TE amplification and elimination affect phenotypic variation, gene transcription, genome evolution, and population diversity (Stein et al. 2018; Chen et al. 2019; Niu et al. 2019; Suh 2019). With the advance of 3D genome mapping technologies, recent studies have shown that TEs may also influence 3D genome architecture, in some cases affecting the organization of cell-specific topologically associating domains (TADs). In mammals (Zhang et al. 2019), activation of long terminal repeat (LTR) retrotransposons facilitated the expansion of CCCTC-binding factor (CTCF; a well-known insulator protein for mediating TAD organization) binding sites among mammalian lineages, which promoted the formation of TAD boundaries, thereby influencing the transcription of adjacent genes (Schmidt et al. 2012). In plants, such as *Arabidopsis*, rice, maize, tomato, and wheat, high-throughput chromosome conformation capture (Hi-C), chromatin interaction analysis by paired-end-tag sequencing (ChIA-PET), in situ Hi-C followed by chromatin immunoprecipitation (HiChIP) maps have been used to reveal chromatin organization and detect genomic regulatory elements (Dong et al. 2017, 2018; Liu et al. 2017; Li et al. 2019; Peng et al. 2019; Concia et al. 2020; Xu et al. 2020). In cotton, we established 3D genome architecture in diploids and allotetraploids and found that the polyploidization process occurring approximately 1.5 Ma contributed to the transition of A/B compartments and reorganization of TADs (Wang et al. 2018). However, the regulatory consequences of remarkable genome size changes through differential TE accumulation on the evolution of higher-order chromatin organization remain understudied in plants.

Cotton is a remarkable textile fiber crop, belonging to the genus *Gossypium* (Malvaceae) (Paterson et al. 2012; Wendel and Grover 2015). *Gossypium* contains more than 45 diploid species divided into eight monophyletic groups (designated A to G and K, $2n = 2\times = 26$) and a single allotetraploid clade

composed of 7 species (AD_1 to AD_7, $2n = 4\times = 52$). These species originated from a common ancestor approximately 5–10 Ma (Grover et al. 2015) and now are widely distributed throughout the tropics and subtropics. Notably, the largest diploid genomes belong to the K genome clade, whose large size (~2,600 Mb) is similar to the tetraploid cotton genomes (~ 2,400 Mb) and is approximately 3-fold larger than the smallest species (D genome; ~840 Mb) (Hawkins et al. 2006, 2009). The 3-fold genome size variation among the diploid *Gossypium* genus was mainly because of the copy number variation of the *Gossypium* retrotransposable *gypsy*-like element (*Gorge3*) (Hawkins et al. 2009). These characteristics make cotton an excellent system for studying the evolutionary mechanism and consequences associated with genome size expansion. Recently, multiple cotton genome sequences were used to uncover TE amplification in *Gossypium* (Paterson et al. 2012; Wang et al. 2012; Li et al. 2014; Grover et al. 2017; Du et al. 2018; Grover et al. 2019; Hu et al. 2019; Udall, Long, Hanson et al. 2019; Wang et al. 2019; Cai et al. 2020; Chen et al. 2020; Grover et al. 2020; Huang et al. 2020) for species that differ 2-fold in size. These species, *G. arboreum* (A_2) and *G. raimondii* (D_5), are thought to share an early LTR insertion event at ~5.7 Ma, with an A_2-specific amplification at ~0.5 Ma after speciation (Li et al. 2014); however, little is known about the composition of the timing of TE bursts in the largest cotton genomes (i.e., the K genomes) (Hawkins et al. 2006).

To address the role of differential TE amplification in influencing 3D genome organization, we assembled three high-quality genomes for *G. rotundifolium* (K_2), *G. arboreum* (A_2), and *G. raimondii* (D_5) by integrating Oxford Nanopore Technologies, paired-end reads, and Hi-C technologies. The assembly of reference-grade genomes allowed us to trace the evolutionary footprints of LTR retrotransposon contributions to genome expansion. We revealed the details of differential TE amplification in three genomes during 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 TE-mediated genome expansion accompanying chromatin structure change and informs further evolutionary genomics research.

Results

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

We applied Oxford Nanopore Technologies to assemble genomes of three cotton species, that is, *G. rotundifolium* (K_2), *G. arboreum* (A_2), and *G. raimondii* (D_5). Although *de novo* assemblies for *G. arboreum* and *G. raimondii* have previously been reported using Illumina and PacBio reads (Paterson et al. 2012; Li et al. 2014; Du et al. 2018; Udall, Long, Hanson et al. 2019; Huang et al. 2020), both genomes contain sequencing gaps and could benefit from improvements in assembly contiguity. For these assemblies, we generated a total of 304 Gb, 212 Gb, 125 Gb of Nanopore sequencing reads, which correspond to 124×, 131×, 167× genome coverage for K_2, A_2, and D_5, respectively (supplementary table 1, Supplementary Material online). Our initial assembly resulted in 3,593 contigs comprising 2.44 Gb in *G. rotundifolium* (N50 = 5.33 Mb); 1,173 contigs comprising 1.62 Gb in *G. arboreum* (N50 = 11.69 Mb); and 366 contigs comprising 0.75 Gb in *G. raimondii* (N50 = 17.04 Mb; table 1). These initial contigs were polished using Illumina paired-end reads with a genome coverage of 108×, 118×, 132× for K_2, A_2, and D_5, respectively. After polishing, we used high-throughput chromosome conformation capture (Hi-C) sequencing data to order and orient contigs, thereby constructing pseudo-chromosomes for each species (fig. 1*a* and supplementary figs. S1–S4, Supplementary Material online; supplementary table 2, Supplementary Material online). The Hi-C assisted assembly placed 2,559 *G. rotundifolium*, 485 *G. arboreum*, and 201 *G. raimondii* contigs on chromosomes for each species ($n = 13$), ultimately representing over 99% of assembled genome length (fig. 1*b* and table 1).

To verify genome assembly completeness, we mapped the clean Illumina reads against each genome and found that more than 97% of reads were aligned (supplementary table 3, Supplementary Material online). More than 90% sequencing reads were perfectly mapped, suggesting high sequence accuracy after base correction for the Nanopore reads. We also performed Benchmarking Universal Single-Copy

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

Genomic Feature	*G. rotundifolium*	*G. arboreum*	*G. raimondii*
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

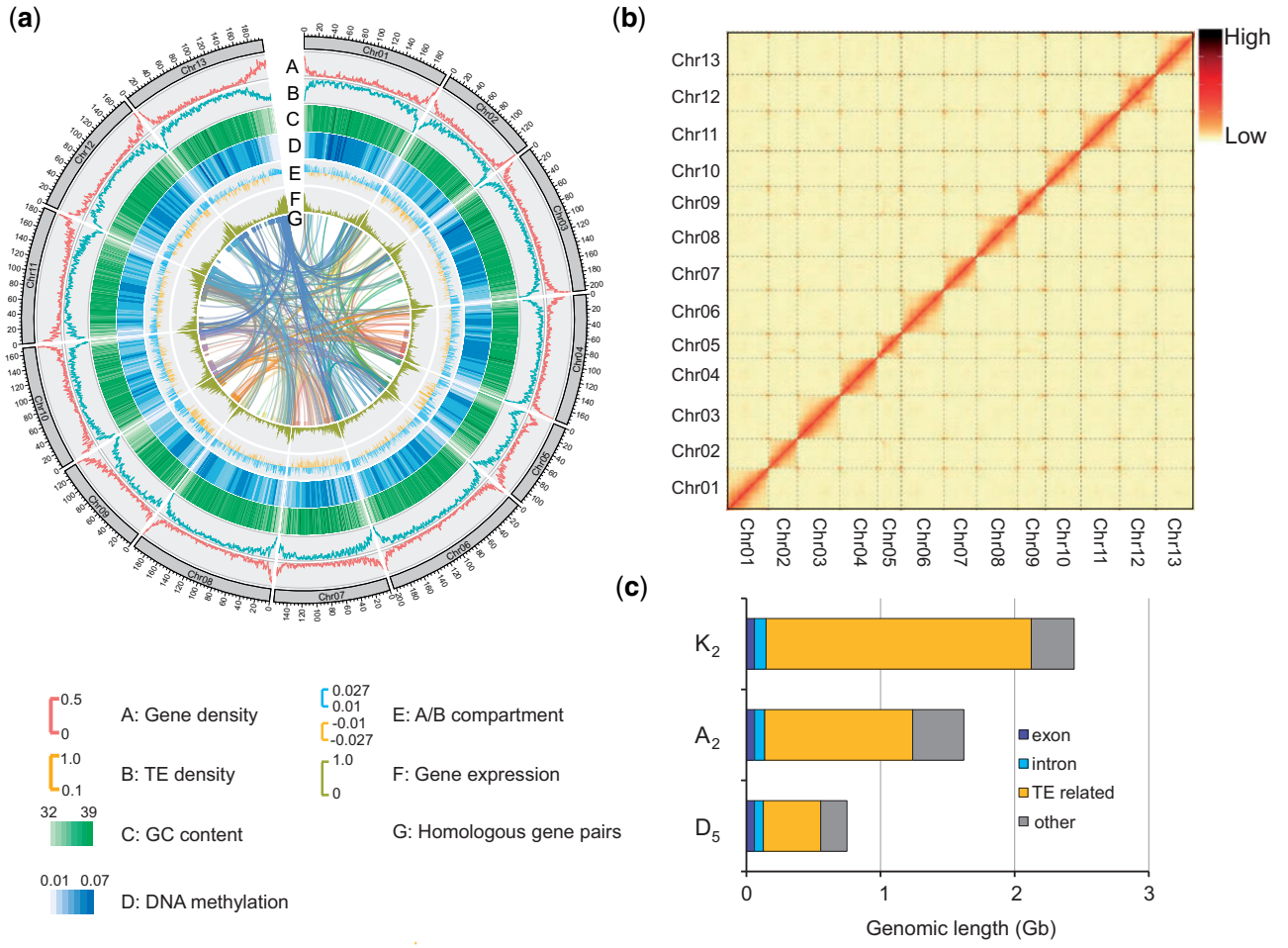

Fig. 1. Genome assembly and feature description of *G. rotundifolium* (K_2). (*a*) Circos plot showing chromosome-level features of *G. rotundifolium*. Tracks represent gene density (A), TE density (B), GC content (C), DNA methylation (D), A/B compartment (E), gene expression (F), and homologous gene pairs (G). In each track (A–D, F), feature data are 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* (K_2), *G. arboreum* (A_2) and *G. raimondii* (D_5). The data include genomic lengths of exon, intron, TE-related, and other genomic regions.

Orthologs (BUSCO) analysis to estimate the assembly completeness in genic regions, recovering 92.5%, 93.9%, and 95.4% of the 1,440 BUSCO analyzed for K_2, A_2, and D_5 (supplementary table 4, Supplementary Material online). Compared with recently published PacBio assemblies of A_2 and D_5 genomes (Paterson et al. 2012; Du et al. 2018; Huang et al. 2020), our assemblies exhibit improvements both in contiguity (6.3-fold and 2.7-fold, for *G. arboreum* and *G. raimondii*) and gap representation (reduced from 1.16 Mb to 22.5 Kb in A_2 and from 17.4 Kb to 7.9 Kb in D_5; supplementary tables 5 and 6, Supplementary Material online and supplementary figs. S5 and S6, Supplementary Material online). These metrics suggest that these genomes qualify as 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 and Supplementary Material online). In total, we predicted 41,590, 41,778, and 40,820 genes for K_2, A_2, and D_5 genomes, respectively, similar to that reported for these and other previously annotated diploid genomes (Wang et al. 2012; Li et al. 2014; Du et al. 2018; Huang et al. 2020) and comprising a similar total length among the three genomes (fig. 1*c*). Between 65 and 70% of annotated genes were transcribed in leaf tissue (27,014 expressed in K_2, 27,381 in A_2, and 28,759 in D_5) (supplementary table 7, Supplementary Material online). Noncoding RNA predictions number 20,782, 11,033, and 6,535 for *G. rotundifolium*, *G. arboreum*, and *G. raimondii*, respectively, and include 132, 133, and 122 miRNAs (supplementary table 8, Supplementary Material online). As in many plant species, repeats are abundant, occupying between 57 and 81% of each genome (K_2 = 1,978 Mb, A_2 = 1,103 Mb, and D_5 = 428 Mb) and scaling with genome size (fig. 1*c*). The long terminal repeat (LTR) retrotransposons in the chromosome centromeric regions were located using previous centromere-related long LTR sequences in *Gossypium hirsutum* (supplementary table 9, Supplementary Material online; and supplementary fig. 7, Supplementary Material online).

Congruent with previous surveys of these species (Hawkins et al. 2006, 2009), our data suggest that differential lineage-

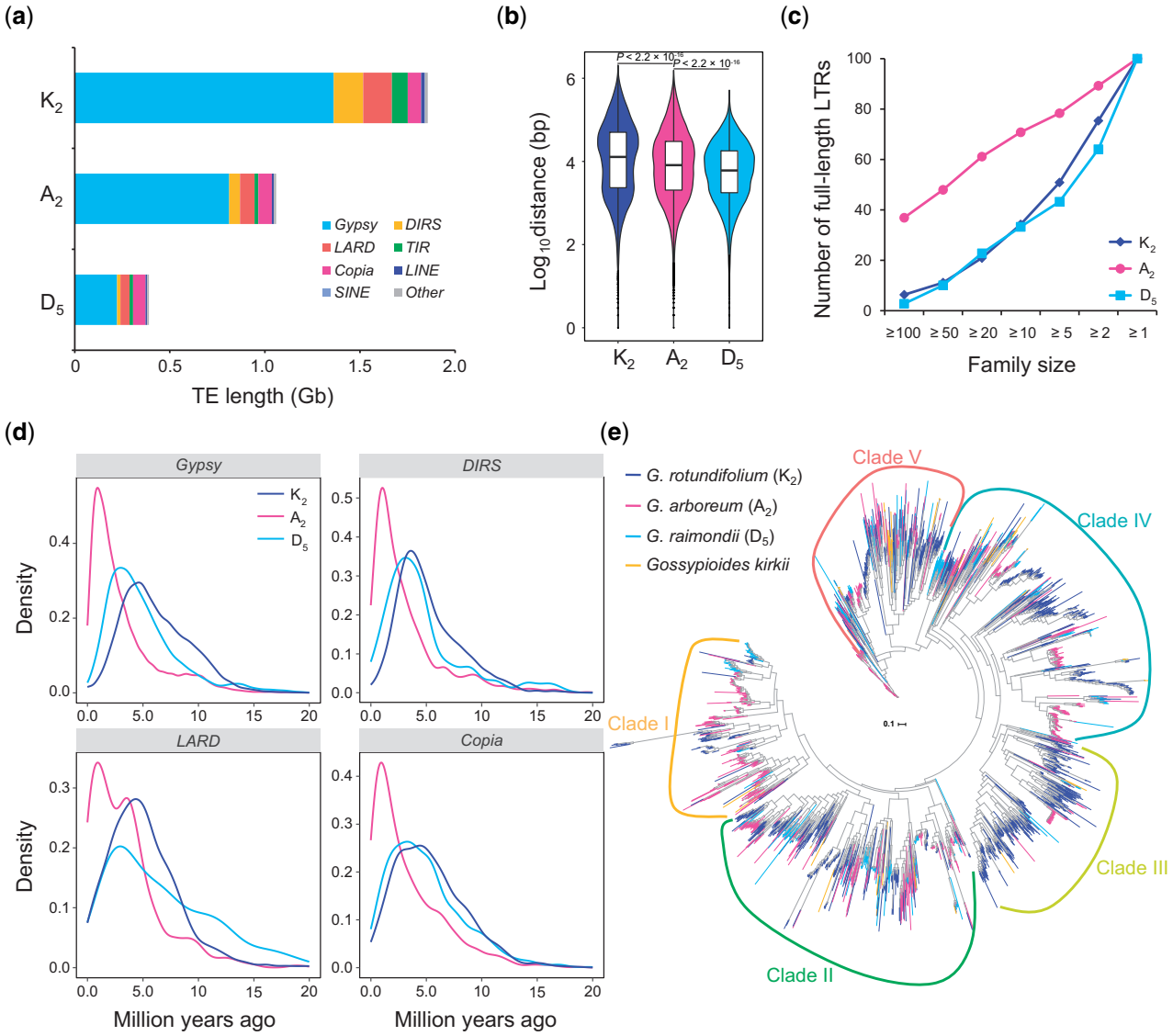

Fig. 2. Characterization of TE evolution in *G. rotundifolium* (K_2), *G. arboreum* (A_2) and *G. raimondii* (D_5). (*a*) Length of different TE classes. (*b*) The distance of intergenic sequences between two adjacent genes in the three genomes. (*c*) Clustering of full-length LTRs. (*d*) Estimated insertion time for *Gypsy*, *DIRS*, *LARD*, and *Copia* transposons. (*e*) Phylogenetic analysis of *Gorge3* transposable elements. The phylogenetic tree includes 2,503 *Gorge3* sequences in four species, including 1,130 in K_2, 963 in A_2, 351 in D_5, and 59 in *Gossypioides kirkii*.

specific retrotransposon amplification and/or retention largely contributes to the 3-fold genome size variation. Like many plants, a large portion of each genome was composed of class I LTR retrotransposons (72% in K_2, 64% in A_2, and 49% in D_5; fig. 2*a*), most of which were *Gypsy*. For both *G. rotundifolium* and *G. arboreum*, the difference in total *Gypsy* proportions among genomes was greater than the proportional difference in genome size. That is, while the *G. rotundifolium* and *G. arboreum* are approximately 3- and 2-fold larger in genome size than *G. raimondii*, the total length of *Gypsy* elements was 6- and 4-fold greater in *G. rotundifolium* and *G. arboreum*, respectively. Conversely, the total length of *Copia* elements was less than expected based on genome sizes alone and was more similar between the three species (i.e., 66–70 Mb in each; supplementary table 10, Supplementary Material online). After *Gypsy* elements, *DIRS* comprise the category that contributed most significantly to the *G. rotundifolium* genome (χ^2 test, $P < 2.2 \times 10^{-16}$), whereas *Copia* elements did not show significant differences among K_2, A_2, D_5 genomes (supplementary table 10, Supplementary Material online). As expected, the proportion of TEs per genome is associated with larger intergenic regions (fig. 2*b*; Mann–Whitney *U* test, $P < 2.2 \times 10^{-16}$), supporting a role for TEs in increasing the space between genes.

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 LTR retrotransposons in K_2, A_2, and D_5 genomes, respectively (supplementary fig. 8, Supplementary Material online). Clustering of these LTRs revealed that 30% of LTRs in A_2 belong to families of more than 20 elements, while only 12% of LTRs in K_2 and D_5 belong to these larger families

(fig. 2c). These results indicate that LTRs in A_2 have higher sequence similarity than those in K_2 and D_5, potentially indicating more recent amplification in A_2. To address this, we estimated the average age of LTR elements from each species using a previously determined mutation rate per year (Chen et al. 2020; Huang et al. 2020). In K_2, the insertion time peak of LTR retrotransposons was found at 4.5–5 Ma, while A_2 had a more recent amplification peak at 0.6–1 Ma (supplementary fig. 9, Supplementary Material online), with older proliferation events diagnosed for *Gypsy*, *DIRS*, *LARD*, and *Copia* elements (fig. 2d). In agreement with previous estimates, LTR elements in D_5 had a peak of 3–4 Ma and a much more recent peak (0.6–1 Ma) in A_2. Notably, *LARD* elements have two amplification peaks (~1 Ma and ~4 Ma) in A_2, the older of which was similar to the amplification time in K_2 and D_5 (fig. 2d). Thus, the second *LARD* amplification is likely lineage-specific for *G. arboreum* or the clade it represents. In addition, these results suggest that K_2 genome has gained a large number of LTRs around 5 Ma compared with the A_2 genome. The phylogenetic tree also supports a huge *Gorge3* (Hawkins et al. 2006; Hawkins et al. 2009, Supplementary Material online) expansion of *Gypsy*-like retrotransposon in clade III by comparing K_2 and A_2 with D_5 and *Gossypioides kirkii* (Udall, Long, Ramaraj et al. 2019) (fig. 2e).

Comparative Genomics and Evolution

While TE amplification has clearly contributed to the 2- to 3-fold genome size increase in K_2 (relative to A_2 and D_5), the effects on synteny in gene regions have not been characterized. Synteny between the K_2 genome and either the A_2 or D_5 genome (fig. 3a) is extensive, encompassing 84% and 89% of the K_2 genome (relative to A_2 and D_5, respectively; supplementary fig. 10, Supplementary Material online). In general, syntenic blocks in K_2 (average length ~11.3 Mb) were larger than in either A_2 (~6 Mb) or D_5 (~3.5 Mb), consistent with the 3-fold change of genome size (fig. 3b), although slightly fewer orthologous genes were detected in the K_2 syntenic regions (26,579) than in either A_2 or D_5 (28,372 and 28,485, respectively). Synteny revealed a large rearrangement between Chr01 and Chr02 in K_2 versus A_2 that was not present between K_2 and D_5. This A_2-specific rearrangement was also found between *G. arboreum* (A_2) and other published diploid cotton genomes (supplementary fig. 11, Supplementary Material online), including *G. thurberi* (D_1) (Grover et al. 2019), *G. turneri* (D_{10}) (Udall, Long, Hanson et al. 2019), and *G. longicalyx* (F_1) (Grover et al. 2020), and has been previously noted in comparison with the sister species, *G. herbaceum* (A_1) (Brubaker et al. 1999; Huang et al. 2020). We also identified a large rearrangement between Chr13 and Chr05 that appears to be specific to K_2; this rearrangement was detected relative to both the newly generated A_2 and D_5 genomes (fig. 3a), as well as relative to other published cotton genomes (supplementary fig. 12, Supplementary Material online).

We calculated synonymous substitution values (*Ks*) for syntenic gene pairs in each genome and between the genomes. We found that all the three species shared a common whole-genome duplication event occurring approximately 57–71 Ma (fig. 3c), as revealed previously (Paterson et al. 2012). Analysis of orthologous genes showed that the clades represented by these three species likely underwent temporally closely spaced divergence at the same period approximately 5.1–5.4 Ma (fig. 3d), congruent with other analyses (Senchina 2003; Wendel and Grover 2015). Further, we found that the K_2, A_2, and D_5 genomes diverged from their closest outgroup genus, *Gossypioides* (Udall, Long, Ramaraj et al. 2019), approximately 8.5–10 Ma, in agreement with other analyses (Grover et al. 2017) (supplementary fig. 13, Supplementary Material online).

Comparison of gene content in syntenic blocks among different species can reveal evolutionary changes in genome organization. Since the three genomes have diverged from a common ancestor, we explored the extent of syntenic gene loss and gain after speciation. A majority of the genes in syntenic blocks (i.e., 21,173 genes) are present and collinear among all three genomes. The comparison between K_2 and D_5/A_2 yielded the greatest number of missing genes (5,868 collinear in D_5 and A_2 and absent in K_2), whereas the comparison between A_2 versus D_5/K_2 yielded the fewest (2,736 genes collinear in D_5 and K_2, but absent in A_2). D_5 was intermediate, with 3,972 genes collinear between A_2 and K_2 that were not found in D_5 (fig. 3e). Gene family analysis using OrthoMCL resulted in ~15% of genes per species remaining as singletons (i.e., unclustered), which may represent genes and/or paralogs that are unique to specific lineages (represented here by these three species; fig. 3f). Together, these results hint at processes of gene birth and death that may have influenced gene content among these related species.

Evolution of A/B Compartment

TE amplification has been recognized as a driver for shaping higher-order chromatin structures in mammals (Diehl et al. 2020) and may play a similar role in the organization of plant 3D architecture. Similar to the A/B compartments of animal genomes, plant chromatin is partitioned into regions that are considered either "active" (A) or "inactive" (B), generally corresponding to euchromatin and heterochromatin, respectively. While the influence of the spatial organization on plant genomes has been surveyed for diverse species with varying genome size (Dong et al. 2017), little is understood about the evolution of A/B compartments among closely related species of variable genome size. Congruent with differences in genomes size, the K_2 genome had fewer active (~44% of the genome) and more inactive regions (55%), than did either A_2 or D_5 (~47% and 52%, and ~46% and 53%, respectively; fig. 4a). Chromosome-level visualization of A/B compartments shows a general blurring of A/B regions consistent with increasing genome size. That is, while the D_5 chromosomes tend to have a large and distinct A compartment on each chromosome arm which border a single, centralized B compartment, the larger K_2 and A_2 genomes exhibit more intercalation between A/B compartments, corresponding to the TE rich regions found in these larger genomes (fig. 4b). A majority of genes were found within A compartments for all genomes; however, interestingly, the smallest genome

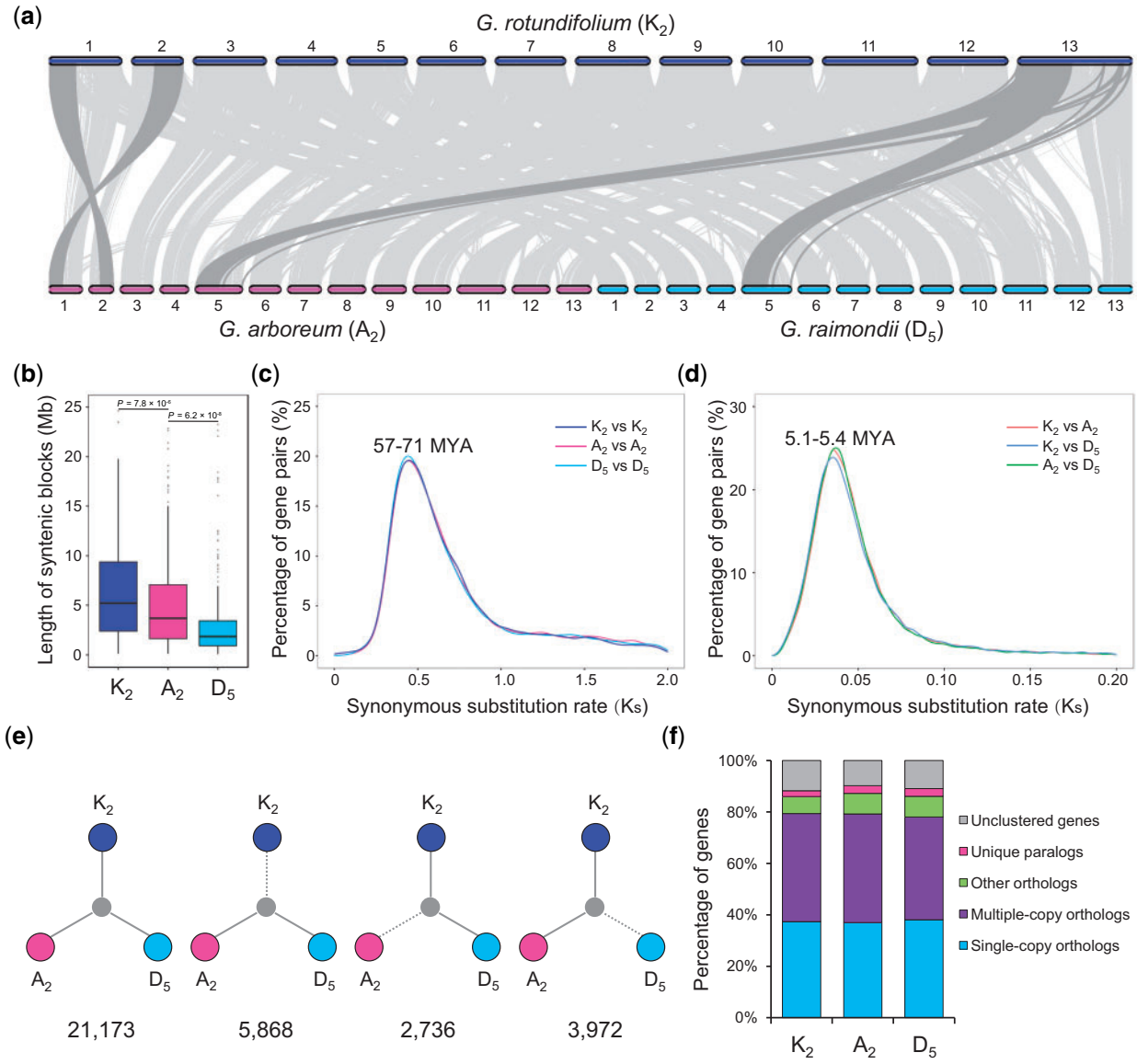

FIG. 3. Genomic synteny and evolution in *G. rotundifolium* (K_2), *G. arboreum* (A_2), and *G. raimondii* (D_5). (*a*) Genome-wide syntenic blocks between *G. rotundifolium* (K_2), *G. arboreum* (A_2), and *G. raimondii* (D_5). Light gray lines indicate collinear blocks in three genomes. Dark gray lines indicate large rearrangements between K_2 versus A_2 and K_2 versus D_5. (*b*) Length of syntenic blocks in *G. rotundifolium* (K_2), *G. arboreum* (A_2), and *G. raimondii* (D_5) genomes (Wilcoxon rank sum test, $P = 7.8 \times 10^{-6}$ for K_2 vs. A_2, $P = 6.2 \times 10^{-8}$ for K_2 vs. D_5). (*c*) Estimation of whole genome duplication time in *G. rotundifolium* (K_2), *G. arboreum* (A_2) and *G. raimondii* (D_5) genomes. (*d*) Estimation of species divergence among three genomes. (*e*) Summary of conserved syntenic genes for three cotton genomes. Gray solid lines and gray dotted lines represent conserved and lost genes in syntenic blocks, respectively. (*f*) Summary of clustered genes and unique genes in *G. rotundifolium* (K_2), *G. arboreum* (A_2), and *G. raimondii* (D_5) genomes based on the OrthoMCL analysis.

had approximately 7,000 fewer genes in the potentially euchromatic A regions than either of the larger genomes (24,267 genes in D_5 vs. 31,307 and 31,331 in K_2 and A_2, respectively) (supplementary table 11, Supplementary Material online). At the gene level, we noticed that 31,307 (75.2%) genes in K_2 and 31,331 (75.0%) genes in A_2 of chromosomes were located in the A compartment, and 4,431 (10.7%) and 4,518 (10.8%) genes were located in B compartment (fig. 4*c*). Approximately 7,000 of the A compartment genes were found in the B compartments of D_5, resulting in a significant difference in the number of genes assigned to each compartment for either of the larger genomes versus D_5 (χ^2 test, $P < 0.01$). Notably, the ratio of TEs in A compartment was also slightly greater for the larger genomes (χ^2 test, $P < 0.01$; fig. 4*d*), which might suggest that the pattern of A/B compartment changes exhibited by these larger genomes (K_2 and A_2) results in more diffuse boundaries and the inclusion of more genes and TEs in A (possibly euchromatic) regions.

To investigate the change of A/B compartment status among three genomes, we analyzed the chromatin status in syntenic gene regions. A comparison of homologous syntenic genes shows that 468 genes exhibited an A-to-B change in the comparison of K_2 with A_2, 3,770 genes exhibited an A-to-B change in the comparison of K_2 with D_5, and 3,765 genes

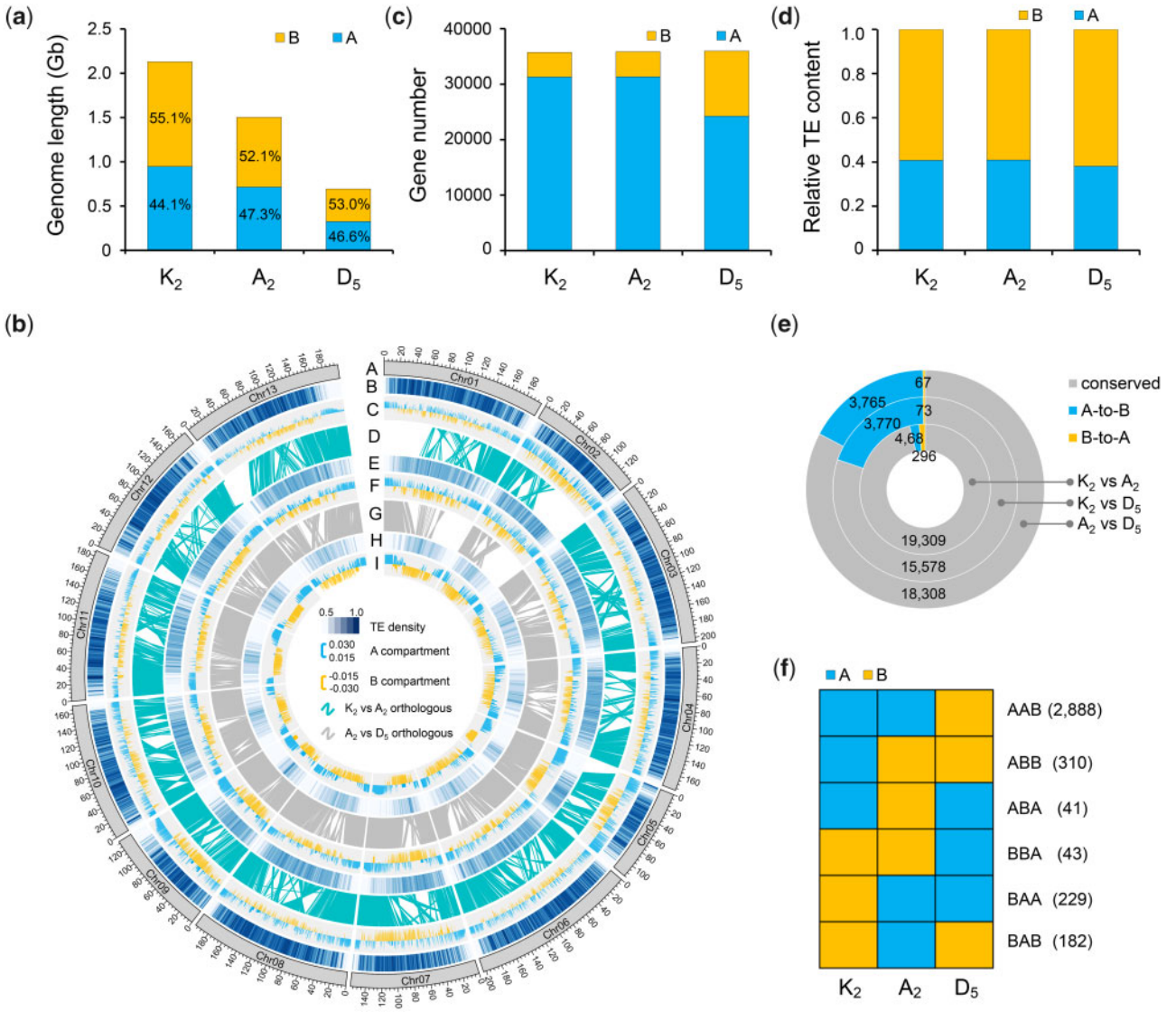

FIG. 4. Characterization of A and B compartments in *G. rotundifolium* (K_2), *G. arboreum* (A_2), and *G. raimondii* (D_5) genomes. (*a*) Genomic length of A and B compartments in K_2, A_2, and D_5. (*b*) Orthologous gene pairs and A/B compartments in *G. rotundifolium* (K_2), *G. arboreum* (A_2), and *G. raimondii* (D_5) genomes. Tracks represent chromosome length of the K_2 genome (A), TE density of the K_2 genome (B), A/B compartment regions in the K_2 genome (C), Orthologous gene pairs between K_2 and A_2 (D), TE density of the A_2 genome (E), A/B compartment regions in A_2 (F), orthologous gene pairs between A_2 and D_5 (G), TE density of the D_5 genome (H), A/B compartment regions in D_5 genome (I). In track A, the chromosome karyotype is shown based on K_2 genome in 1 Mb windows sliding 200 Kb. The chromosome length of D_5 and A_2 was normalized based on K_2 genome. (*c*) Number of genes in A and B compartments in K_2, A_2, and D_5 genomes. (*d*) Relative TE content in the A and B compartments. (*e*) Percentage of conserved genes and A/B compartment switching genes in the K_2-A_2, K_2-D_5, and A_2-D_5 comparisons. (*f*) Number of dynamic orthologous genes showing status switching of A/B chromatin compartments in three cotton genomes. Green and yellow boxes show genes in A and B compartments, respectively.

exhibited an A-to-B change in the comparison of A_2 with D_5. Only 296, 73, and 67 genes exhibited B-to-A compartment changes in the three comparisons (fig. 4*e* and supplementary tables 12–15, Supplementary Material online). Analysis of syntenic genes supports this, as more homologous genes are located in B compartments in D_5 relative to K_2 or A_2 than any other comparison (fig. 4*f* and supplementary fig. 14, Supplementary Material online). The functional enrichment analysis results suggest that A-to-B genes are enriched in the pathways of ion binding and transcription factor activity, whereas the B-to-A genes were intriguingly involved in fundamental activity, such as ubiquitin transferase activity, pectate lyase activity, and ATP binding (FDR < 0.05, supplementary fig. 15, Supplementary Material online).

As expected by the putatively euchromatic nature of A compartment, both genes and TEs in A compartment generally display significantly higher expression levels than those in B compartments, although the magnitude of the difference is reduced for D_5 gene expression (supplementary fig. 16, Supplementary Material online). The trend of higher expression in A compartment was reiterated when comparing both homologous (syntenic) genes and gene-associated TEs (within 5 Kb of syntenic genes), whereby expression in B compartments was generally lower than in A compartments

(supplementary fig. 17, Supplementary Material online). These results suggest that the delimitation between A and B regions in D_5 has measurable consequences for gene expression. This result suggested that expressed 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) are smaller (generally submegabase) domains located within A/B compartments that exhibit frequent within-domain interactions and less frequent interactions with loci located outside of the domain (Gibcus and Dekker 2013). First described in animals (Dixon et al. 2012; Nora et al. 2012; Sexton et al. 2012), TADs both contribute to physical higher-order chromatin structures, as well as influence gene expression (Gibcus and Dekker 2013). While plant genomes lack some TAD features that are canonical in animals (e.g., the insulator protein CTCF), TAD-like domain structures have been described for several plants (Wang et al. 2015; Dong et al. 2017; Liu et al. 2017; Dong et al. 2018; Concia et al. 2020), including cotton (Wang et al. 2018). This previous work in cotton found approximately 1,000 or more TADs whose organization changed over evolutionary time and across the boundaries of ploidy. Here, we identify TAD-like sequences for three cotton species, including one that has not previously been available for analysis (i.e., *G. rotundifolium*; K_2). In general, we find that the number of TAD regions increased with genome size from 1,063 TADs in D_5 to 2,541 in K_2 (supplementary table 16, Supplementary Material online). TADs ranged in size from 300 Kb to 3 Mb, averaging about 860 Kb in the larger K_2 and A_2 genomes, and about 25% lower (average = 645 Kb) in the smaller D_5 genome (fig. 5*a*). In total, we were able to predict TADs for over 90% of each genome. We note that the number of TADs in A_2 and D_5 was larger than previously reported (Wang et al. 2018); however, we attribute this to the use of the polyploid genome as the reference to identify TADs, due to an unavailability of high-quality reference genome sequences for those species at that time. 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. 5*b*). As expected, we found that genes in TAD boundaries tend to have significantly higher expression levels than those in the interior (fig. 5*c* and supplementary fig. 18, Supplementary Material online), consistent with our previous results (Wang et al. 2018).

Because the turnover of TAD boundaries can indicate structural reorganization, we compared TAD boundaries in syntenic blocks to explore TAD conservation and turnover in three genomes (fig. 5*d* and supplementary fig. 19, Supplementary Material online). We found that 406 TAD boundaries in K_2 were conserved among the three genomes, and that lineage-specific boundaries increased in number as the total number of TADs increased (1,393, 580, and 131 in K_2, A_2, and D_5, respectively; fig. 5*d* and supplementary table 17, Supplementary Material online). For example, in a syntenic region between K_2 (Chr08: 81.4–91.7 Mb) and D_5 (Chr08: 29.3–32.4 Mb) only about 45% of TAD boundaries (5) were conserved in D_5 (fig. 5*e*), and 70% (7) were conserved in the comparison of A_2 and K_2 (Chr07: 70–79.5 Mb for K_2 and Chr07: 62.75–68.45 Mb for A_2; fig. 5*f*). Motifs located at TAD boundaries have been associated with boundary delimitation and strength in animals (Ramirez et al. 2018; Stadhouders et al. 2019), but little is known about TAD boundary motifs in plants. Motif analysis at TAD boundaries reveal 69 specific sequence motifs that were specific to K_2, but only 8 and 4 specific motifs in A_2 and D_5, respectively. We identified 13 motifs in conserved boundaries in the three genomes (fig. 5*g* and supplementary table 18, Supplementary Material online). For example, the K_2 genome has a *PABPC3* (poly(A) binding protein cytoplasmic 3) binding motif in lineage-specific boundaries, A_2 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. 5*h*).

Effect of Transposon Amplification on TAD Organization

To explore whether TE gain in the K_2 and A_2 genomes and TE loss in D_5 genome led to changes in TAD organization, we investigated TE content in TAD boundaries. We found that 60%, 44%, and 26% of the genomic length of TAD boundaries were covered by *Gypsy* LTR retrotransposons in K_2, A_2, and D_5, respectively, occupying the highest proportion of all TE categories in TAD boundaries (fig. 6*a*). Of note is the finding that expressed TEs were enriched in TAD boundaries relative to the whole-genome (χ^2 test, $P < 2.2 \times 10^{-16}$) (fig. 6*b*). Specific TAD boundaries to a single species had a higher proportion of TEs relative to conserved boundaries in K_2 and A_2 genomes (fig. 6*c*). This result is consistent with the finding that more species-specific boundaries were located in A compartment than in B compartment (fig. 6*d*). In addition, we found that young LTR retrotransposons were more often associated with lineage-specific TAD boundaries, whereas ancient TEs were more likely to exist in conserved boundaries (fig. 6*e*, χ^2 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. 6*f*). In summary, these results indicated that recent amplification of expressed TEs in K_2 and A_2 genomes might contribute to the formation of lineage-specific TAD boundaries after the divergence of the three species (fig. 6*g*).

Discussion

TE Dynamics and Genome Evolution among Species That Vary 3-Fold in Genome Size

In this study, we sequenced and assembled the first high-quality reference genome of *G. rotundifolium* (K_2), and updated the genome assemblies of *G. arboreum* (A_2) and *G. raimondii* (D_5). Compared with the five available published genome versions of D_5 (Paterson et al. 2012; Udall, Long, Hanson et al. 2019; Chen et al. 2020) and A_2 (Du et al. 2018; Huang et al. 2020), our assemblies have a considerable

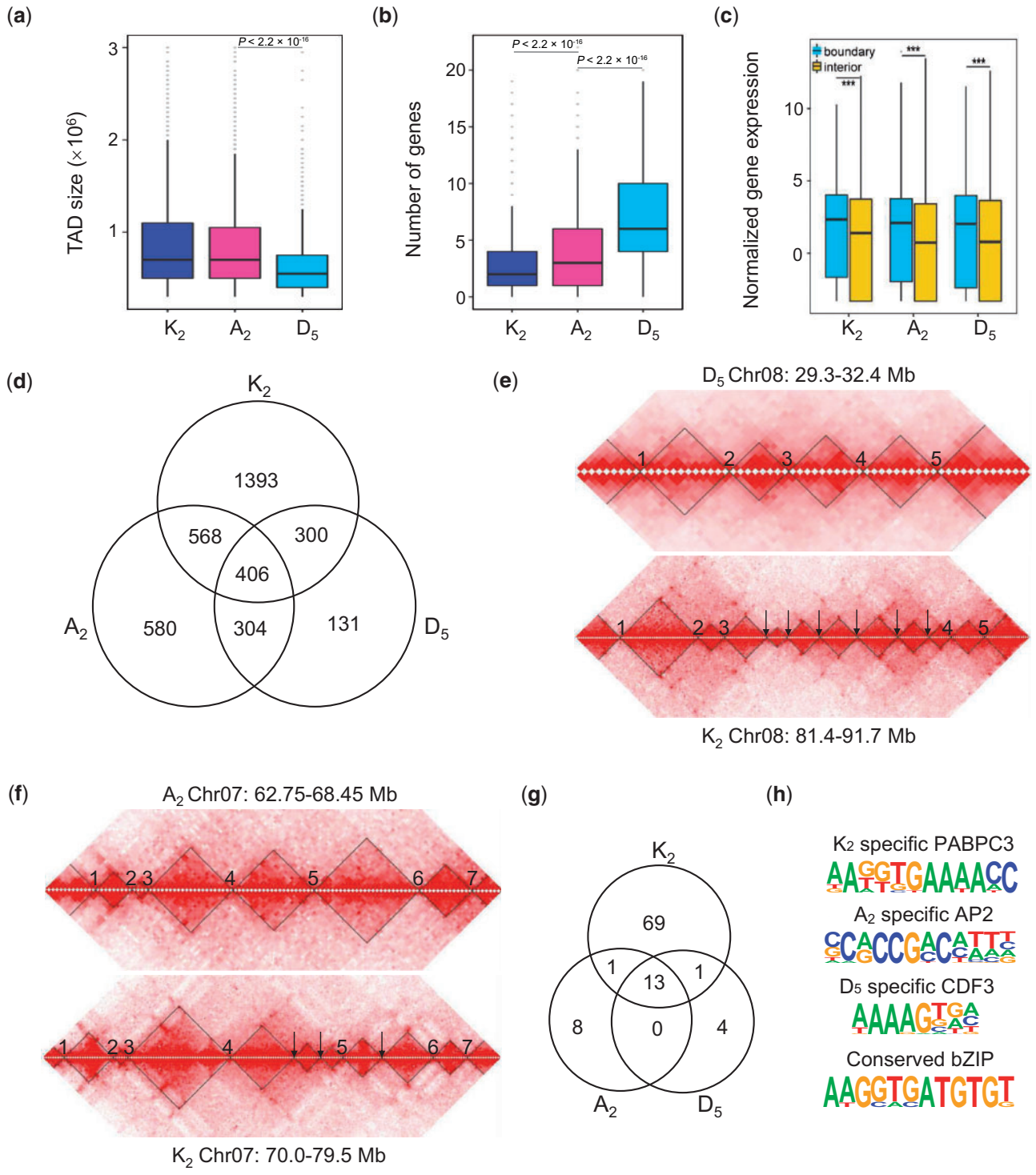

Fig. 5. Specific and conserved TADs in *G. rotundifolium* (K_2), *G. arboreum* (A_2), and *G. raimondii* (D_5) genomes. (*a*) TAD size in K_2, A_2 and D_5 genomes (Wilcoxon rank sum test, $^{***}P < 2.2 \times 10^{-16}$). (*b*) Number of genes in TAD boundaries (from −50 Kb to 50 Kb) (Wilcoxon rank sum test, $^{***}P < 2.2 \times 10^{-16}$). (*c*) Gene expression in TAD boundaries and TAD interior of K_2, A_2, and D_5 genomes (Wilcoxon rank sum test, $^{***}P < 2.2 \times 10^{-16}$). (*d*) Number of specific and conserved TADs in K_2, A_2, and D_5 genomes. K_2 is used as the reference coordinate. (*e*) TAD structures of collinear blocks on Chr08 between K_2 and D_5. (*f*) TAD structures of collinear blocks on Chr07 between K_2 and A_2. The gray boxes indicate TADs. In each heatmap, strong interactions are highlighted in red color. Numbers on heatmap represent the conserved TADs in two genomes. Black arrows represent specific TADs in K_2 relative to D_5 or A_2. (*g*) Venn diagram showing the number of transcription factor binding motifs in specific and conserved TAD boundaries. In this analysis, 380 conserved TAD boundaries overlapped with promoters, and 1,126, 433, and 120 lineage-specific TADs overlapped with promoters in the K_2, A_2, and D_5 genomes, respectively. (*h*) The most significantly enriched sequence motifs in specific and conserved TAD boundaries.

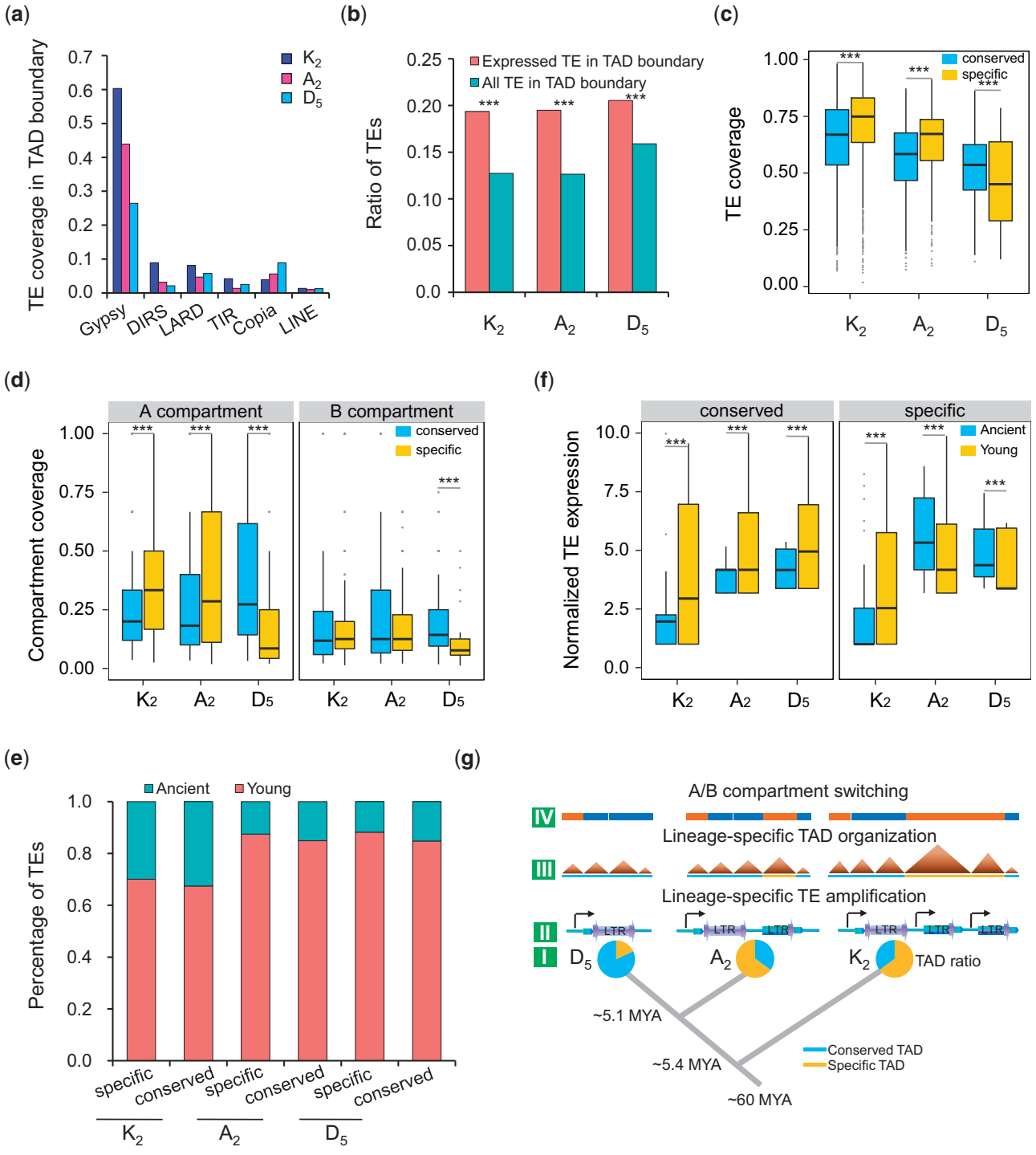

Fig. 6. The effect of TE amplification on the reorganization of TAD boundary. (*a*) TE coverage in TAD boundaries in the K_2, A_2, and D_5 genomes. (*b*) Ratio of expressed TEs and all TEs in TAD boundaries relative to the whole genome. (*c*) TE coverage in specific and conserved TAD boundaries. Significance 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 (Wilcoxon rank sum test, ***$P < 2.2 \times 10^{-16}$). (*e*) 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 divergence time. (*f*) Normalized expression of ancient and young TEs between lineage-specific and conserved TADs (Wilcoxon rank sum test, ***$P < 2.2 \times 10^{-16}$). (*g*) A conceptual model for 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 expressed (black arrow) lineage-specific LTRs in three genomes. "III" represents the increase of TAD number. "IV" represents A/B compartment switching. Phylogenetic schematic shows divergence between *G. raimondii* (D_5) and *G. arboreum* (A_2) at ~ 5.1 Ma, *G. raimondii* (D_5) and *G. rotundifolium* (K_2) at ~5.4 Ma. Divergence time between *Gossypium* and *Theobroma cacao* is estimated at ~60 Ma.

improvement in sequence contiguity (N50 reaching 11.69 Mb and 17.04 Mb). We present the first modern evaluation of the observed genome expansion in the lineage leading to K_2, which is a consequence of transposable element proliferation, especially LTR retrotransposon elements. The genome expansion of K_2 was placed around 4.5–5 Ma, and the expansion of A_2 more recently around 0.6–1 Ma, consistent with a previous estimation (Huang et al. 2020). The smallest genome (D_5) also exhibited a proliferation in TEs approximately 3–4 Ma, although this proliferation was likely offset by TE removal, as previously reported for *Gorge*3 LTR retrotransposons in this species (Hawkins et al. 2006, 2009). Despite the 3-fold change of genome size, the three genomes shared a relatively high level of gene synteny with enlarged intergenic regions. This raises the possibility that TE expansion influenced and/or reshaped regulatory relationships between noncoding regions and the transcription of syntenic coding genes in K_2 and A_2 relative to D_5, in particular, considering the recognized important role of noncoding intergenic sequence in transcriptional regulation (Gil and Ulitsky 2020). Our assembled K_2 genome, which has the largest genomes among diploid species of *Gossypium* genus, lays a foundation for further study of the effect of transposon amplification on genome size variation and the rewiring of transcriptional regulation concomitant with genome size expansion.

Consequences of TE Proliferation for Chromatin Architecture and Gene Expression

Previous studies have shown that TE distribution or activity is involved in chromatin interaction in plants. Maize and tomato, for example, exhibit extensive chromatin loops, which are linked to A compartments (Dong et al. 2017). In *Arabidopsis*, the KNOT engaged element 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 elements (Feng et al. 2014; Grob et al. 2014; Grob and Grossniklaus 2019). Heat-induced transposon activation in *Arabidopsis* is associated with reduced chromosomal interactions in pericentromeric regions, which is involved in 3D genome reorganization (Sun et al. 2020). In rice, the density of TEs in H3K9me2-marked regions is higher than in basal chromatin loop sites, suggesting that H3K9me2 binding sites with higher TE density might be involved in chromatin interactions (Zhao et al. 2019). However, relationships among TE amplification dynamics and 3D genome organization at the scale of genome evolution is largely unexplored. The three reference-grade genomes presented here provide an excellent opportunity to explore shifts in A/B boundaries during evolution using a known phylogenetic context and in the face of divergent and variable levels of TE accumulation. We found that expressed TEs had a higher frequency in A compartments and might have played a role in the evolutionary switching of B to A compartments as genome size increases. Given that some kinds of TEs tended to jump into active genic regions but did not enlarge genome size, further research is needed to explore the intricate relationship between the amplification of TEs and the expansion of active genomic regions. In addition, we linked expressed LTR retrotransposon expansion to the formation of lineage-specific TAD boundaries. The comparison of interspecies TAD showed that syntenic blocks could help to identify lineage-specific TAD organization. It will be of interest to discover whether the relationships we describe here among genome size, TE mobilization history, and higher-order chromatic dynamics are a general property of genome evolution in plants.

In addition to informing our understanding of the evolution of genome architecture, the effect of TAD reorganization on gene transcription is relevant to genetic manipulation for both functional genomics and crop improvement. Previous studies found that gene transcription had a role in the organization of TADs in mammals (Stadhouders et al. 2019; Collombet et al. 2020), raising a possibility that transcription factor binding motifs might participate in the formation of TADs, similar to the finding that a TCP transcriptional factor binding motif was enriched in TAD boundaries in rice (Liu et al. 2017). Of note is the observation that the enrichment of TCP binding sites in TAD boundaries of *Marchantia* was not required for TAD formation (Karaaslan et al. 2020). Specifically, analysis of transcription factor binding sites might help uncover possible molecular mechanisms underlying the formation of new TAD boundaries in plants, opening up prospects for future manipulation.

In summary, we present evidence for an 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 noncoding genomic sequences in complex genomes.

Materials and Methods

Cotton Materials

Plants of *G. rotundifolium* (accession number K201), *G. arboreum* (cultivar Shixiya-1) and *G. raimondii* (accession number D502) are maintained in the National Wild Cotton Nursery and are 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 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. 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, and 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; Biomarker Technologies). Nanopore data (binary fast5 format) was subjected to base calling using the Guppy software from the MinKNOW package. Processed

reads were subject to removal of sequencing adapters and filtering of reads with low quality and/or short length (<2,000 bp), and surviving reads were 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 (paired-end, 150 bp).

Hi-C Experiment and Library Construction

Fresh leaves (1 g) from *G. rotundifolium* were chopped with sharp blades, fixed with 1% formaldehyde solution, frozen in liquid nitrogen, and were used for nuclear extraction. Nuclei were digested with 30–50 U HindIII/DpnII for 15 h at 37°C. Digested chromatin was end-labeled with biotin-14-dCTP, and the DNA product was purified after blunt-end ligation. Then, the DNA was fragmented by ultrasound to a length of less than 500 bp. DNA fragments of 300–500 bp were captured by Streptavidin T1 magnetic beads. The library was prepared from the DNA isolated by the magnetic beads using the DNA library kit (Vazyme, #NDM607), and the obtained DNA library was sequenced (paired-end 150 bp reads) using the MGI2000 system.

Genome Assembly and Assessment

Nanopore sequencing reads were corrected via Canu (v1.3) with the parameter "correctedErrorRate = 0.045" (Koren et al. 2017). Clean reads were subsequently subject to *de novo* assembly using wtdbg (Ruan and Li 2019) (https://github.com/ruanjue/wtdbg). Assembled contigs were calibrated using Racon (Vaser et al. 2017) and then polished with the Illumina sequencing reads using Pilon (Walker et al. 2014) (v1.22; parameters: –mindepth 10 –changes –fix bases) for three iterations. In total, we corrected 12.6 million (M), 6.0 M and 27.2 M SNPs, and 17.6 M, 9.2 M, and 31.0 M InDels in the A_2, D_5 and K_2 assemblies, respectively. Assembly quality was assessed three ways. First, Illumina reads were mapped to the contigs using BWA (-mem) (Li and Durbin 2009), and the properly mapped reads were counted using SAMTools (v0.1.19; -flagstat) (Li et al. 2009). Second, the assemblies were evaluated for the 458 conserved core genes found in the CEGMA (v2.5) database (Parra et al. 2007). Finally, the assemblies also evaluated using the BUSCO embryophyta_odb9 data set, which contains 1,440 conserved eukaryotic genes (Simao et al. 2015).

Chromosome Assembly Using Hi-C

Hi-C data were used to construct chromosome-level assemblies for the three genomes. Hi-C data of *G. arboreum* and *G. raimondii* were previously published (Wang et al. 2018). Hi-C data of *G. rotundifolium* was newly generated here with two independent experiments (HindIII and DpnII for digestion of chromatin) (supplementary table 2, Supplementary Material online). Notably, up to 99.5% of A/B compartment regions and 96.4% of TAD boundaries overlapped in these two experiments (The method for A/B compartment and TAD analysis was described below), and the HindIII Hi-C data was used for further analysis. The resolution of Hi-C data sets was estimated as 20 Kb for *G. arboreum*, 10 Kb for *G. raimondii*, and 20 Kb for *G. rotundifolium* using the method described previously (Rao et al. 2014). We performed a preassembly for error correction of contigs, which required splitting the contigs into segments of 50 Kb (on average). Hi-C data were mapped to these fragments and unique mappings were retained for the assembly using LACHESIS (v1.0) (Burton et al. 2013). Any two segments that showed inconsistent connections with information from the raw contigs were checked manually. 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 assembly quality, each assembly was split into 100-Kb bins to serve as a reference for Hi-C data mapping using HiC-Pro (v2.7.1) (Servant et al. 2015). Obvious placement and orientation errors in chromatin interaction patterns were manually adjusted. The interaction matrices generated by HiC-Pro were displayed with heatmaps at a 100 Kb resolution.

Transposon Prediction

We used both LTR_Finder (v1.07) (Xu and Wang 2007) with "-C -M 0.8" and RepeatScout (v1.0.5) (Price et al. 2005) with default parameters to construct a repetitive sequence library, representing structure-based prediction and ab initio prediction, respectively. PASTEClassifier (v1.0) was used to classify sequences in the library with respect to repeat type, and these were subsequently merged with Repbase (version 19.06) for the final repeat library (Bao et al. 2015). This library was used to predict repetitive sequences in each genome using RepeatMasker (-nolow -no_is -norna -engine wublast) (Tarailo-Graovac and Chen 2009).

LTR Retrotransposon Analysis

LTR_Finder (Xu and Wang 2007) was used with parameter settings (-C -M 0.8) to identify full-length LTRs in each genome. Long-terminal repeat (LTR) sequences were clustered from each full-length LTR element using the CD-HIT program (Fu et al. 2012) with parameter "-d 0 -c 0.8 -aL 0.80 -T 0 -M 1500000" for LTR family analysis. For each full-length LTR retrotransposon, the 5′ LTR and 3′ LTR sequences were aligned using MUSCLE (v3.8.1551) (Edgar 2004) and the divergence distance between them was calculated with a Kimura two parameter (K2P) model using "distmat" from the EMBOSS toolkit (Rice et al. 2000). 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, $r = 3.5 \times 10^{-9}$) (Chen et al. 2020; Huang et al. 2020). According to the time of divergence (5 Ma) among the three *Gossypium* species, the burst time of full-length LTR retrotransposons were divided into ancient TE (≥5 Ma) and young TE (<5 Ma), depending on whether the burst was inferred to have occurred prior to or following divergence of these clades. The expression level of transposon was calculated based on the definition of Reads Per Kilobase per Million mapped reads (RPKM), and those with RPKM greater than 0.1 were considered as "expressed TE." *Gossypium* retrotransposable *Gypsy*-like element (*Gorge3*) sequences

(Hawkins et al. 2006) were aligned against the full-length LTR elements from *G. rotundifolium*, *G. arboreum*, *G. raimondii*, and *Gossypioides kirkii* (Udall, Long, Ramaraj et al. 2019) using a reciprocal blastn (-e 1e-05) search. MAFFT (v7.453) (Katoh and Standley 2013) was used for *Gorge*3 5′ LTR domain with multiple sequence alignments in four species, and then phylogenic tree was constructed using the IQ-TREE program (Nguyen et al. 2015).

Gene Prediction

To predict protein-coding genes, three different strategies were adopted, including ab initio prediction, homolog-based prediction, and transcript-based prediction. Genscan (Burge and Karlin 1997), Augustus (v2.4) (Stanke and Morgenstern 2005), GlimmerHMM (v3.0.4) (Majoros et al. 2004), SNAP (v2006-07-28) (Korf 2004) were used for ab initio prediction. GeMoMa (v1.3.1) (Keilwagen et al. 2018) was used for predicting genes based on homologous protein from other species (*Populus trichocarpa*, *Arabidopsis thaliana*, *Vitis vinifera*, *Theobroma cacao*, and *G. raimondii*). Hisat2 (v2.0.4) (Kim et al. 2015) and Stringtie (v1.2.3) (Pertea et al. 2015) were used for reference-guided transcript assembly. PASA (v2.0.2) (Haas et al. 2003) was used to predict unigene sequences based on RNA-Seq data without reference-guided assembly. Finally, EVM (v1.1.1) (Haas et al. 2008) was used to integrate the prediction results obtained by the above three methods, and PASA (v2.0.2) (Haas et al. 2003) was used to modify gene models. To identify pseudogenes, GenBlastA (v1.0.4) (She et al. 2009) was used to scan each genome after masking predicted protein-coding sequences and GeneWise (v2.4.1) (Birney et al. 2004) was used to identify premature stop codons and frameshift mutations relative to the intact reference proteins. The functional annotation of predicted genes was performed using 1) InterProScan (v5.0) (Jones et al. 2014) with "-iprlookup -goterms" parameter settings, 2) NR (v20190625) with "-evalue 1e-05 -best_hit_overhang 0.25 -max_target_seqs 5", and 3) The Arabidopsis Information Resource 10 (TAIR10) database (Lamesch et al. 2012). Gene Ontology (GO) enrichment analysis was performed using a Fisher's exact test method (Carbon et al. 2019). GO enrichment analysis was performed for genes showing A-to-B and B-to-A compartment status change, using different background gene sets (K2 and A2 genes were combined as a reference set and orthologous gene pairs showing A/B compartment status change were used as a test set; similarly, A2 and D5 genes were combined as another reference set).

Identification of Centromeric Regions

Previously identified centromeric regions from the published TM-1 reference genome, that is, GhCR1-5′LTR, GhCR2-5′LTR, GhCR3-5′LTR and GhCR4-5′LTR (Wang et al. 2015; Wang et al. 2019), were aligned to the K_2, A_2, and D_5 genome sequences using MUMmer (v4.0) (Delcher et al. 2002), with the parameters "-c 90 -l 40" followed by "delta-filter -1," to identify uniquely aligning regions. After manual filtering of alignments, the SPSS software (version 17.0) was used to calculate the 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 and 2) delta-filter -1. Syntenic blocks among the three genomes were constructed using MCScanX (Tang et al. 2008) with default settings and requiring a minimum of five homologous genes. The newly assembled A_2 and D_5 reference genomes were compared with published genomes (Paterson et al. 2012; Du et al. 2018; Udall, Long, Hanson et al. 2019; Huang et al. 2020) from CottonGen website (https://www.cottongen.org/data/download) by MUMmer (v4.0) and MCScanX. The Chr01-Chr02 large translocation of A_2-specific rearrangement and Chr13-Chr05 large translocation of K_2-specific rearrangement were confirmed by comparing with the published A_1 (Huang et al. 2020), D_1 (Grover et al. 2019), D_{10} (Udall, Long, Hanson et al. 2019; Udall, Long, Ramaraj et al. 2019) and F_1 (Grover et al. 2020) genomes. The single-copy gene families among three *Gossypium* genomes were extracted using an OrthoMCL analysis (Li et al. 2003).

Analysis of A and B Compartments

Hi-C interaction data can be used to partition the genome into two compartments, based on spatial organization of the chromatin and the relative paucity of interactions between compartments. Referred to as A/B compartments, these represent chromatin regions corresponding to open and closed chromatin, respectively. We evaluated each genome for the presence of A/B compartments, as described previously (Lieberman-Aiden et al. 2009). Briefly, Hi-C data for each species were aligned using HiC-Pro, as mentioned above. Valid interaction reads were used to construct heatmaps of each chromosome at resolutions of 20 Kb, 50 Kb, and 100 Kb. Raw contact maps were normalized using a sparse-based implementation of the iterative correction method embedded in HiC-Pro (v2.11.1) (Servant et al. 2015). The principal component analysis (PCA) method was used to identify A and B compartments by the HiTC (v1.0) package in R (Servant et al. 2012). Each chromosome was divided into consecutive 50 Kb bins for the construction of normalized interaction matrices as described in our previous study (Wang et al. 2018). Chromosomal bins with values of greater than zero were regarded as "A compartment," bins with values of less than zero were regarded as "B compartment." At the chromosome level, A compartment has a higher gene density and a lower transposon density than B compartment. To analyze the A/B compartment status of homologous gene regions among three *Gossypium* genomes, genomic sequences of gene body, upstream and downstream 2 Kb that were known to be important for gene transcriptional regulation, were extracted. In this analysis, we only considered the regions where the first principal component value changes from positive (A) to negative (B) or vice versa.

Analysis of Topologically Associating Domains

Topologically associating domains (TAD) are regions of highly selfinteracting chromatin that have distinct boundaries and which have been shown to align with coordinately related

gene clusters in some species. TAD regions for each species were identified using the HiTAD (Wang et al. 2017) software with default settings. In this analysis, the raw chromatin interaction matrix for each chromosome was constructed using HiC-Pro at a resolution of 50 Kb. 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 MCScanX. Conserved boundaries were defined as those with a maximum boundary change of 3-resolution distance (150 Kb) 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 used to predict motifs with the findMotifsGenome.pl program in HOMER (v5.0) (Heinz et al. 2010) software, with the parameters "-len 8,10,12 -size 200." Putative motifs were filtered with cutoffs of $P \leq 0.01$ for known and $P \leq 1e{-}10$ for *de novo* prediction. We used 1,000 uniformly distributed random genomic regions that did not overlap with TAD boundaries as a control set for nonboundary regions.

RNA-Seq and Data Analysis

For each species, leaf total RNA was extracted using the 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). After filtering of low-quality bases and sequence adapters, the clean RNA sequencing data were mapped to each genome using hisat2 (v2.0.4) (Kim et al. 2015) software. High-quality mapping reads were extracted using SAMTools (v0.1.19; -q 25) (Li et al. 2009). After filtering PCR duplicates using samtools (rmdup), the remaining reads were used to calculate the expression level of genes using Stringtie (v1.2.3) (Pertea et al. 2015).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

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. This study was also supported by the Fundamental Research Funds for the Central Universities (2662020ZKPY017). We also thank the high-performance computing center at National Key Laboratory of Crop Genetic Improvement in Huazhong Agricultural University.

Author Contributions

X.Z. and M.W. conceived and designed the project. K.W. and F.L. provided the materials. P.W. and L.P. 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.X. analyzed the sequencing data. M.W. and J.L. prepared the figures and wrote the manuscript draft, and X.Z., J.W., C.G., and K.W. revised it. All authors read and approved the final manuscript.

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 https://figshare.com/projects/Cotton_genomes/91826.

Code Availability

Bioinformatics in this study were performed with open-source software, which source data and codes for integrating figures have been deposited in a GitHub repository at https://github.com/HZAU-CottonLab/KAD_Genomes/.

References

Bao W, Kojima KK, Kohany O. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 6:11.

Birney E, Clamp M, Durbin R. 2004. GeneWise and Genomewise. *Genome Res*. 14(5):988–995.

Brubaker CL, Paterson AH, Wendel JF. 1999. Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. *Genome* 42(2):184–203.

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*. 268(1):78–94.

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol*. 31(12):1119–1125.

Cai Y, Cai X, Wang Q, Wang P, Zhang Y, Cai C, Xu Y, Wang K, Zhou Z, Wang C, et al. 2020. Genome sequencing of the Australian wild diploid species *Gossypium australe* highlights disease resistance and delayed gland morphogenesis. *Plant Biotechnol J*. 18(3):814–828.

Carbon S, Douglass E, Dunn N, Good B, Harris NL, Lewis SE, Mungall CJ, Basu S, Chisholm RL, Dodson RJ, et al. 2019. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res*. 47(D1):D330–D338.

Chen J, Lu L, Benjamin J, Diaz S, Hancock CN, Stajich JE, Wessler SR. 2019. Tracking the origin of two genetic components associated with transposable element bursts in domesticated rice. *Nat Commun*. 10(1):641.

Chen ZJ, Sreedasyam A, Ando A, Song Q, De Santiago LM, Hulse-Kemp AM, Ding M, Ye W, Kirkbride RC, Jenkins J, et al. 2020. Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. *Nat Genet*. 52(5):525–533.

Collombet S, Ranisavljevic N, Nagano T, Varnai C, Shisode T, Leung W, Piolot T, Galupa R, Borensztein M, Servant N, et al. 2020. Parental-to-embryo switch of chromosome organization in early embryogenesis. *Nature* 580(7801):142–146.

Concia L, Veluchamy A, Ramirez-Prado JS, Martin-Ramirez A, Huang Y, Perez M, Domenichini S, Rodriguez Granados NY, Kim S, Blein T, et al. 2020. Wheat chromatin architecture is organized in genome territories and transcription factories. *Genome Biol*. 21(1):104.

Delcher AL, Phillippy A, Carlton J, Salzberg SL. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res*. 30(11):2478–2483.

Diehl AG, Ouyang N, Boyle AP. 2020. Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. *Nat Commun*. 11(1):1796.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485(7398):376–380.

Dong PF, Tu XY, Chu PY, Lu PT, Zhu N, Grierson D, Du BJ, Li PH, Zhong SL. 2017. 3D chromatin architecture of large plant genomes determined by local A/B compartments. *Mol Plant*. 10(12):1497–1509.

Dong QL, Li N, Li XC, Yuan Z, Xie DJ, Wang XF, Li JN, Yu YA, Wang JB, Ding BX, et al. 2018. Genome-wide Hi-C analysis reveals extensive hierarchical chromatin interactions in rice. *Plant J*. 94(6):1141–1156.

Du X, Huang G, He S, Yang Z, Sun G, Ma X, Li N, Zhang X, Sun J, Liu M, et al. 2018. Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. *Nat Genet*. 50(6):796–802.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32(5):1792–1797.

Feng S, Cokus SJ, Schubert V, Zhai J, Pellegrini M, Jacobsen SE. 2014. Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in *Arabidopsis*. *Mol Cell*. 55(5):694–707.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28(23):3150–3152.

Gibcus JH, Dekker J. 2013. The hierarchy of the 3D genome. *Mol Cell*. 49(5):773–782.

Gil N, Ulitsky I. 2020. Regulation of gene expression by *cis*-acting long non-coding RNAs. *Nat Rev Genet*. 21(2):102–117.

Grob S, Grossniklaus U. 2019. Invasive DNA elements modify the nuclear architecture of their insertion site by KNOT-linked silencing in *Arabidopsis thaliana*. *Genome Biol*. 20(1):120.

Grob S, Schmid MW, Grossniklaus U. 2014. Hi-C analysis in *Arabidopsis* identifies the KNOT, a structure with similarities to the flamenco locus of *Drosophila*. *Mol Cell*. 55(5):678–693.

Grover CE, Arick MA, 2nd, Conover JL, Thrash A, Hu G, Sanders WS, Hsu CY, Naqvi RZ, Farooq M, Li X, et al. 2017. Comparative genomics of an unusual biogeographic disjunction in the cotton tribe (*Gossypieae*) yields insights into genome downsizing. *Genome Biol Evol*. 9(12):3328–3344.

Grover CE, Arick MA, 2nd, Thrash A, Conover JL, Sanders WS, Peterson DG, Frelichowski JE, Scheffler JA, Scheffler BE, Wendel JF. 2019. Insights into the evolution of the new world diploid cottons (*Gossypium*, Subgenus Houzingenia) based on genome sequencing. *Genome Biol Evol*. 11(1):53–71.

Grover CE, Gallagher JP, Jareczek JJ, Page JT, Udall JA, Gore MA, Wendel JF. 2015. Re-evaluating the phylogeny of allopolyploid *Gossypium* L. *Mol Phylogenet Evol*. 92:45–52.

Grover CE, Pan MQ, Yuan DJ, Arick MA, Hu GJ, Brase L, Stelly DM, Lu ZF, Schmitz RJ, Peterson DG, et al. 2020. The *Gossypium longicalyx* genome as a resource for cotton breeding and evolution. *G3 (Bethesda)*. 10(5):1457–1467.

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, et al. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*. 31(19):5654–5666.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol*. 9(1):R7.

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Res*. 16(10):1252–1261.

Hawkins JS, Proulx SR, Rapp RA, Wendel JF. 2009. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. *Proc Natl Acad Sci U S A*. 106(42):17811–17816.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol Cell*. 38(4):576–589.

Hu Y, Chen J, Fang L, Zhang Z, Ma W, Niu Y, Ju L, Deng J, Zhao T, Lian J, et al. 2019. *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. *Nat Genet*. 51(4):739–748.

Huang G, Wu Z, Percy RG, Bai M, Li Y, Frelichowski JE, Hu J, Wang K, Yu JZ, Zhu Y. 2020. Genome sequence of *Gossypium herbaceum* and genome updates of *Gossypium arboreum* and *Gossypium hirsutum* provide insights into cotton A-genome evolution. *Nat Genet*. 52(5):516–524.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30(9):1236–1240.

Karaaslan ES, Wang N, Faiß N, Liang Y, Montgomery SA, Laubinger S, Berendzen KW, Berger F, Breuninger H, Liu C. 2020. *Marchantia* TCP transcription factor activity correlates with three-dimensional chromatin structure. *Nat Plants*. 6(10):1250–1261.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30(4):772–780.

Keilwagen J, Hartung F, Paulini M, Twardziok SO, Grau J. 2018. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics*. 19(1):189.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. 12(4):357–360.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. 27(5):722–736.

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics*. 5:59.

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al. 2012. The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res*. 40(Database issue):D1202–D1210. (Database issue):

Li E, Liu H, Huang L, Zhang X, Dong X, Song W, Zhao H, Lai J. 2019. Long-range interactions between proximal and distal regulatory regions in maize. *Nat Commun*. 10(1):2633.

Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, Li Q, Ma Z, Lu C, Zou C, et al. 2014. Genome sequence of the cultivated cotton *Gossypium arboreum*. *Nat Genet*. 46(6):567–572.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.

Li H, 1000 Genome Project Data Processing Subgroup, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 13(9):2178–2189.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950):289–293.

Liu C, Cheng YJ, Wang JW, Weigel D. 2017. Prominent topologically associated domains differentiate global chromatin packing in rice from *Arabidopsis*. *Nat Plants*. 3(9):742–748.

Majoros WH, Pertea M, Salzberg SL. 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20(16):2878–2879.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 32(1):268–274.

Niu XM, Xu YC, Li ZW, Bian YT, Hou XH, Chen JF, Zou YP, Jiang J, Wu Q, Ge S, et al. 2019. Transposable elements drive rapid phenotypic variation in *Capsella rubella*. *Proc Natl Acad Sci U S A*. 116(14):6908–6913.

Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, et al. 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485(7398):381–385.

Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23(9):1061–1067.

Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J, et al. 2012. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* 492(7429):423–427.

Peng Y, Xiong D, Zhao L, Ouyang W, Wang S, Sun J, Zhang Q, Guan P, Xie L, Li W, et al. 2019. Chromatin interaction maps reveal genetic regulation for quantitative traits in maize. *Nat Commun*. 10(1):2632.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. 33(3):290–295.

Price AL, Jones NC, Pevzner PA. 2005. *De novo* identification of repeat families in large genomes. *Bioinformatics*. 21(Suppl 1):i351–358.

Ramirez F, Bhardwaj V, Arrigoni L, Lam KC, Gruning BA, Villaveces J, Habermann B, Akhtar A, Manke T. 2018. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun*. 9(1):189.

Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159(7):1665–1680.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*. 16(6):276–277.

Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. *Nat Methods*. 17(2):155–158.

Schmidt D, Schwalie PC, Wilson MD, Ballester B, Goncalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* 148(1-2):335–348.

Senchina DS, Alvarez I, Cronn RC, Liu B, Rong JK, Noyes RD, Paterson AH, Wing RA, Wilkins TA, Wendel JF. 2003. Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. *Mol Biol Evol*. 20(4):633–643.

Servant N, Lajoie BR, Nora EP, Giorgetti L, Chen CJ, Heard E, Dekker J, Barillot E. 2012. HiTC: exploration of high-throughput 'C' experiments. *Bioinformatics* 28(21):2843–2844.

Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J, Vert J-P, Heard E, Dekker J, Barillot E. 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol*. 16:259.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* 148(3):458–472.

She R, Chu JS, Wang K, Pei J, Chen N. 2009. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res*. 19(1):143–149.

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19):3210–3212.

Stadhouders R, Filion GJ, Graf T. 2019. Transcription factors and 3D genome conformation in cell-fate decisions. *Nature* 569(7756):345–354.

Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res*. 33(Web Server issue):W465–467.

Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang C, Chougule K, Gao D, Iwata A, Goicoechea JL, et al. 2018. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nat Genet*. 50(2):285–296.

Suh A. 2019. Genome size evolution: small transposons with large consequences. *Curr Biol*. 29(7):R241–R243.

Sun L, Jing Y, Liu X, Li Q, Xue Z, Cheng Z, Wang D, He H, Qian W. 2020. Heat stress-induced transposon activation correlates with 3D chromatin organization rearrangement in *Arabidopsis*. *Nat Commun*. 11(1):1886.

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008. Synteny and collinearity in plant genomes. *Science* 320(5875):486–488.

Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics*. Chapter 4:Unit 4:10.

Udall JA, Long E, Hanson C, Yuan D, Ramaraj T, Conover JL, Gong L, Arick MA, Grover CE, Peterson DG, et al. 2019a. *De novo* genome sequence assemblies of *Gossypium raimondii* and *Gossypium turneri*. *G3 (Bethesda)*). 9(10):3079–3085.

Udall JA, Long E, Ramaraj T, Conover JL, Yuan D, Grover CE, Gong L, Arick MA, 2nd Masonbrink RE, Peterson DG, et al. 2019b. The genome sequence of *Gossypioides kirkii* illustrates a descending dysploidy in plants. *Front Plant Sci*. 10: 1541.

Vaser R, Sovic I, Nagarajan N, Sikic M. 2017. Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Res*. 27(5):737–746.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9(11):e112963.

Wang C, Liu C, Roqueiro D, Grimm D, Schwab R, Becker C, Lanz C, Weigel D. 2015. Genome-wide analysis of local chromatin packing in *Arabidopsis thaliana*. *Genome Res*. 25(2):246–256.

Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S, et al. 2012. The draft genome of a diploid cotton *Gossypium raimondii*. *Nat Genet*. 44(10):1098–1103.

Wang M, Tu L, Yuan D, Zhu D, Shen C, Li J, Liu F, Pei L, Wang P, Zhao G, et al. 2019. Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nat Genet*. 51(2):224–229.

Wang M, Wang P, Lin M, Ye Z, Li G, Tu L, Shen C, Li J, Yang Q, Zhang X. 2018. Evolutionary dynamics of 3D genome architecture following polyploidization in cotton. *Nat Plants*. 4(2):90–97.

Wang S, Chen JD, Zhang WP, Hu Y, Chang LJ, Fang L, Wang Q, Lv FN, Wu HT, Si ZF, et al. 2015. Sequence-based ultra-dense genetic and physical maps reveal structural variations of allopolyploid cotton genomes. *Genome Biol*. 16(1):108.

Wang XT, Cui W, Peng C. 2017. HiTAD: detecting the structural and functional hierarchies of topologically associating domains from chromatin interactions. *Nucleic Acids Res*. 45(19):e163.

Wendel JF, Grover CE. 2015. Taxonomy and evolution of the cotton genus *Gossypium*. In: Fang DD, Percy RG, editors. Cotton. 2nd ed. Madison, WI: Agronomy Monograph. p. 25–44.

Xu G, Lyu J, Li Q, Liu H, Wang D, Zhang M, Springer NM, Ross-Ibarra J, Yang J. 2020. Evolutionary and functional genomics of DNA methylation in maize domestication and improvement. *Nat Commun*. 11(1):5539.

Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res*. 35(Web Server issue):W265–W268.

Zhang Y, Li T, Preissl S, Amaral ML, Grinstein JD, Farah EN, Destici E, Qiu Y, Hu R, Lee AY, et al. 2019. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat Genet*. 51(9):1380–1388.

Zhao L, Wang S, Cao Z, Ouyang W, Zhang Q, Xie L, Zheng R, Guo M, Ma M, Hu Z, et al. 2019. Chromatin loops associated with active genes and heterochromatin shape rice genome architecture for transcriptional regulation. *Nat Commun*. 10(1):3640.