Plant Communications
Research article

CellPress
Partner Journal

High-quality *Gossypium hirsutum* and *Gossypium barbadense* genome assemblies reveal the landscape and evolution of centromeres

Xing Chang, Xin He, Jianying Li, Zhenping Liu, Ruizhen Pi, Xuanxuan Luo, Ruipeng Wang, Xiubao Hu, Sifan Lu, Xianlong Zhang and Maojun Wang*

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

***Correspondence: Maojun Wang (mjwang@mail.hzau.edu.cn)**

https://doi.org/10.1016/j.xplc.2023.100722

ABSTRACT

Centromere positioning and organization are crucial for genome evolution; however, research on centromere biology is largely influenced by the quality of available genome assemblies. Here, we combined Oxford Nanopore and Pacific Biosciences technologies to *de novo* assemble two high-quality reference genomes for *Gossypium hirsutum* (TM-1) and *Gossypium barbadense* (3-79). Compared with previously published reference genomes, our assemblies show substantial improvements, with the contig N50 improved by 4.6-fold and 5.6-fold, respectively, and thus represent the most complete cotton genomes to date. These high-quality reference genomes enable us to characterize 14 and 5 complete centromeric regions for *G. hirsutum* and *G. barbadense*, respectively. Our data revealed that the centromeres of allotetraploid cotton are occupied by members of the centromeric repeat for maize (*CRM*) and *Tekay* long terminal repeat families, and the *CRM* family reshapes the centromere structure of the A_t subgenome after polyploidization. These two intertwined families have driven the convergent evolution of centromeres between the two subgenomes, ensuring centromere function and genome stability. In addition, the repositioning and high sequence divergence of centromeres between *G. hirsutum* and *G. barbadense* have contributed to speciation and centromere diversity. This study sheds light on centromere evolution in a significant crop and provides an alternative approach for exploring the evolution of polyploid plants.

Key words: genome assembly, centromere architecture, convergent evolution, polyploidization

Chang X., He X., Li J., Liu Z., Pi R., Luo X., Wang R., Hu X., Lu S., Zhang X., and Wang M. (2024). High-quality *Gossypium hirsutum* and *Gossypium barbadense* genome assemblies reveal the landscape and evolution of centromeres. Plant Comm. **5**, 100722.

INTRODUCTION

Polyploidization occurs widely in plants (Orr, 1990; Soltis et al., 2014; Wendel, 2015), increasing the potential for crop production and adaptation (Chalhoub et al., 2014; Appels et al., 2018; Wang et al., 2019; Zhuang et al., 2019). The allotetraploid cottons *Gossypium hirsutum* and *Gossypium barbadense* produce a major natural textile fiber that provides economic income for at least 100 countries worldwide, with an economic value of approximately US$500 billion per year (Chen et al., 2007). Around 1–2 million years ago, allotetraploid cotton was derived from a polyploidization event between two diploid ancestors with the A and D genomes (Wendel, 1989; Grover et al., 2015). The TM-1 genome of *G. hirsutum* and 3-79 genome of *G. barbadense* were the first cotton genomes to be published (Li et al., 2015; Yuan et al., 2015; Zhang et al., 2015), and several genome assemblies of *G. hirsutum* and *G. barbadense* have been released in recent years (Hu et al., 2019; Wang et al., 2019; Yang et al., 2019; Chen et al., 2020; Huang et al., 2020). However, these genomes are still fragmented and contain numerous gaps, which hampers the study of cotton centromere biology. With advances in sequencing technology, Oxford Nanopore Technologies (ONT) sequencing and Pacific Biosciences (PacBio) single-molecule real-time sequencing provide an opportunity to improve the quality of reference genomes for multiple species.

Centromeres are vital structures in the chromosomes of most eukaryotes and play an essential role in ensuring the correct

Published by the Plant Communications Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

segregation of chromosomes and the accurate inheritance of genetic material during meiosis and mitosis (Henikoff et al., 2001). Because the centromere and flanking regions contain abundant repetitive sequences, complete assembly of centromeric regions is challenging (Neumann et al., 2011; Miga, 2015). Recently, the Telomere-to-Telomere consortium announced a complete human reference genome after nearly 20 years of effort (Nurk et al., 2022); this assembly addressed the remaining 8% of sequences from the previous reference genome, mainly located in centromere and telomere regions, enabling the study of centromere organization, variation, and function (Altemose et al., 2022). In plants, complete centromeres in the model plants *Arabidopsis thaliana* and rice have provided interesting insights into their structure, satellite sequences, transposons, and methylation levels (Li et al., 2021; Naish et al., 2021; Song et al., 2021; Hou et al., 2022; Wang et al., 2022). These studies have advanced our understanding of the diversity and evolution of centromere sequences across species, thereby deepening our understanding of the influence of centromeres on genome evolution.

The active and functional centromere is defined by a conserved epigenetic mechanism, which is known as the centromere-specific histone H3 variant CENH3 in plants (Zhong et al., 2002; Sullivan and Karpen, 2004). In many animals and plants, the centromere region typically contains long arrays of centromere-specific satellite repeats (Melters et al., 2013). The length of these satellite sequences is generally between 150 and 200 bp; examples include the 171-bp α satellite DNA in human, the 180-bp monomer repeat in *Arabidopsis thaliana*, and the 156-bp satellite repeat in maize (Wu and Manuelidis, 1980; Willard, 1989; Maluszynska and Heslop-Harrison, 1991; Alfenito and Birchler, 1993; Ananiev et al., 1998). Centromeric satellite sequences can extend to several megabases or even the pericentromeric region (Melters et al., 2013; Altemose et al., 2022). However, satellite sequences are not a necessary component of centromeres. Several natural satellite-free centromeres have been discovered, including those of potato, *Equus asinus*, and zebra (Gong et al., 2012; Nergadze et al., 2018; Cappelletti et al., 2022), indicating that satellite sequences are not necessary for centromere function (Earnshaw and Migeon, 1985). Centromeres are also occupied by specific retrotransposons in some plants and animals. For example, the centromeric retrotransposon *Ty3-gypsy* is present and conserved within *Gramineae* (Miller et al., 1998). In maize, centromeric repeat for maize (*CRM*) elements are interspersed with satellite repeats (Ananiev et al., 1998). Further research on *CRM* found that this retroelement can regulate the loading of CENH3 protein through R-loops (Liu et al., 2021). In rice, the *CRR* element is located within the functional centromeric region (Cheng et al., 2002). The centromeric repeats *CRW* and *Quintas* were identified in wheat centromeres (Zhao et al., 2023). In *Brassica rapa*, *ALE* and *CRM* repeats invaded the centromeres (Zhang et al., 2023).

In cotton, several studies have used molecular and cytogenetic techniques to investigate the structure and function of centromeres. Luo et al. (2012) identified a *Gypsy*-like long terminal repeat (LTR) retrotransposon named centromere retroelement *Gossypium* (CRG) by fluorescence *in situ* hybridization. It was abundant in the centromere region of the *G. hirsutum* and D genomes and absent from the A genome. Han et al. (2016) also identified 10 centromeric repeats, 9 of which were absent from centromeres in the A genome, and 7 of which were similar to CRM. Four centromeric retrotransposons in *G. hirsutum* called GhCR1–4 were identified by sequencing a bacterial artificial chromosome (Zhang et al., 2014). Moreover, several researchers have attempted to find satellite repeat sequences in cotton, and a few cotton-specific centromere satellite sequences have been found (Melters et al., 2013; Han et al., 2016).

In this study, we incorporated Oxford Nanopore reads and PacBio reads for *de novo* assembly of high-quality genomes of *G. hirsutum* and *G. barbadense*. These higher-quality reference genomes provide an unprecedented opportunity to analyze the structure, sequence characteristics, epigenetic patterns, and evolution of tetraploid cotton centromeres. We examined structural changes in centromeres and convergent evolution of centromeres between tetraploid cotton subgenomes after tetraploidization. Comparative genomics revealed sequence divergence and evolutionary differences between *G. hirsutum* and *G. barbadense* centromeres.

RESULTS

Two high-quality genome assemblies for *G. hirsutum* and *G. barbadense*

We generated ∼345.3 Gb (∼160×) and ∼291.83 Gb (130×) ONT reads (Supplemental Tables 1 and 2) and ∼194.01 Gb (∼84×) and ∼210.98 Gb (∼91×) PacBio reads for *G. hirsutum* accession TM-1 and *G. barbadense* accession 3-79 (Wang et al., 2019). Using the two long-read sequencing technologies, we assembled genomes with contig N50s of 21.96 and 12.14 Mb, respectively. We then polished the contigs using ∼50× Illumina paired-end reads to correct SNPs and small insertions/deletions in the sequencing reads (Wang et al., 2019). The polished contigs were scaffolded into 26 chromosomes using high-throughput chromosome conformation capture (Hi-C) data (Huang et al., 2022; Pei et al., 2022), and 99.05% and 97.4% of the contig sequences were anchored to TM-1 and 3-79 (Supplemental Tables 3 and 4), respectively. The final chromosome-scale genomes were 2324 Mb (TM-1 HAU v2) and 2254 Mb (3-79 HAU v3) in length and included 556 and 1662 scaffolds, respectively (Table 1; Supplemental Table 3). The genome continuity of the current assemblies was significantly higher than that of previous high-quality assemblies for TM-1 and 3-79: the contig N50 increased from 4760 to 21 961 kb for TM-1 and from 2151 to 12 139 kb for 3-79 (Supplemental Table 5) (Wang et al., 2019; Yang et al., 2019).

To confirm the quality of the updated TM-1 HAU v2 and 3-79 HAU v3 genomes, we compared our assemblies with previously released genome versions, TM-1 ZJU v2.1 and TM-1 HAU v1.1 (Hu et al., 2019; Wang et al., 2019). We found highly consistent alignments of individual chromosomes and no obviously mis-oriented or mis-assembled regions (Figure 1; Supplemental Figures 1 and 2). The Hi-C heatmaps showed no signals of mis-orientations or other large structural errors (Supplemental Figure 3). Benchmarking Universal Single-Copy Orthologs (BUSCO) evaluation revealed 98.9% and 99.5% complete BUSCO genes from the embryophyta_odb10 dataset (Table 1), compared with 98.2% and 98.6% in the previous TM-1 HAU v1

Genomic feature	*G. hirsutum*	*G. barbadense*
Assembly		
Genome size (bp)	2 324 185 275	2 254 770 940
Scaffold number	582	1688
Scaffold N50 (bp)	108 550 191	93 110 895
Contig number	1418	2064
Contig N50 (bp)	21 961 441	12 139 909
Gap length (Mb)	0.8579	1.2415
Pseudochromosome size (bp)	2 302 180 022	2 169 217 327
Percentage anchored (%)	99.05	97.2
Complete BUSCOs (%)	98.9	99.5
Annotation		
TE percentage (%)	68.57	68.77
Gene number	77 782	76 754
Genes in pseudochromosomes	77 035	75 708

Table 1. Summary of *G. hirsutum* and *G. barbadense* genome assemblies.

and 3-79 HAU v2 assemblies (Wang et al., 2019) (Supplemental Table 5), indicating higher completeness of genic regions in the new assemblies. Because longer contigs can span previously ambiguous regions, we were able to correct many small inversions in the previous assemblies (Supplemental Figures 1, 2, and 4). Five corrected inversions in TM-1 HAU v2 and five in 3-79 HAU v3 were confirmed by Hi-C heatmaps and PCR amplifications (Supplemental Figure 5; Supplemental Table 6). Our TM-1 HAU v2 assembly also contained 25–93 Mb of new sequences compared with various previous versions (Supplemental Table 5) (Hu et al., 2019; Wang et al., 2019; Yang et al., 2019; Chen et al., 2020; Huang et al., 2020), mainly attributable to transposable elements (TEs) in intergenic and centromeric regions (Supplemental Figure 6; Supplemental Table 7). Our new assemblies thus offer notable improvements in genome completeness, mis-assembly correction, and centromeric regions.

We predicted 77 782 and 76 754 high-confidence protein-coding genes in the TM-1 HAU v2 and 3-79 HAU v3 genomes (Table 1; Supplemental Table 8), 77 035 and 75 708 of which were distributed on 26 chromosomes. A total of 97.3% of these genes were supported by RNA sequencing (RNA-Seq) reads, and more than 95% and 99% were supported by full-length transcripts. Gene Ontology (GO) annotation showed that 85.85% and 89.91% of the genes could be assigned GO terms (Supplemental Table 9). We predicted 6231 new genes in the TM-1 HAU v2 genome compared with the HAU v1.1 genome and ZJU v2.1 genome (Hu et al., 2019) and 3580 new genes in the 3-79 HAU v3 genome compared with the HAU v2 genome. Of these annotated genes, 5622 (89.8%) and 3572 (99.8%) were supported by RNA-Seq reads or full-length transcripts in the TM-1 HAU v2 and 3-79 HAU v3 genomes.

Structural variation is an important manifestation of individual genome variation and population differentiation. Comparative analysis revealed approximately 258 Mb of inversions and 56 Mb of translocations between the TM-1 HAU v2 and 3-79 HAU v3 genomes (Supplemental Tables 10 and 11). Of these, 172 Mb of inversions were located in the A_t subgenome, with the largest inversion (31.54 Mb) in the A06 chromosome of TM-1 (Wang et al., 2019), and 86 Mb of inversions were located in the D_t subgenome. We identified 48 973 presence segments (800–192 207 bp) with a total length of 103 Mb in the TM-1 HAU v2 genome (Supplemental Table 12) and 44 796 presence regions (562–14 167 bp) with a total length of 102 Mb in the 3-79 HAU v3 genome (Supplemental Table 13). In total, 4479 genes overlapped with these presence sequences in the TM-1 HAU v2 genome, and 3723 genes overlapped with them in the 3-79 HAU v3 genome (Supplemental Tables 14 and 15). The number of presence–absence variations (PAVs) in the A_t subgenome (30 789 in TM-1 HAU v2, 25 579 in 3-79 HAU v3) was higher than that in the D_t subgenome (14 863 in TM-1 HAU v2, 18 184 in 3-79 HAU v3), indicating that more genetic variation may have occurred in the A_t subgenome.

Centromere landscape and composition

Using CENH3 chromatin immunoprecipitation sequencing (ChIP-Seq) data (Han et al., 2016; Hu et al., 2019), we identified 25 potential centromeric regions for TM-1 HAU v2 (all except D08) and all centromeric regions for 3-79 HAU v3 (Figure 2A; Supplemental Figures 7 and 8); these ranged in length from 0.94 to 4 Mb and from 0.15 to 3.4 Mb (Supplemental Tables 16 and 17), respectively. The centromere position of Ghir_D08 was also not reported in the TM-1 ZJU v2.1 genome, and dispersed signals were observed in the TM-1 WHU D08 chromosome (Supplemental Figure 9A). CRG1 and CRG2 are centromere-specific retrotransposons that are enriched in the centromere regions of tetraploid and some diploid genomes (Luo et al., 2012). To obtain the accurate centromere position of the D08 chromosome, we aligned CRG1 and CRG2 to the TM-1 HAU v2 genome and found that CRGs were frequently present at approximately 26.07–28.76 Mb (Supplemental Figure 9B), consistent with the centromere feature visible on the Hi-C heatmap of the D08 chromosome. CRG1 and CRG2 were also

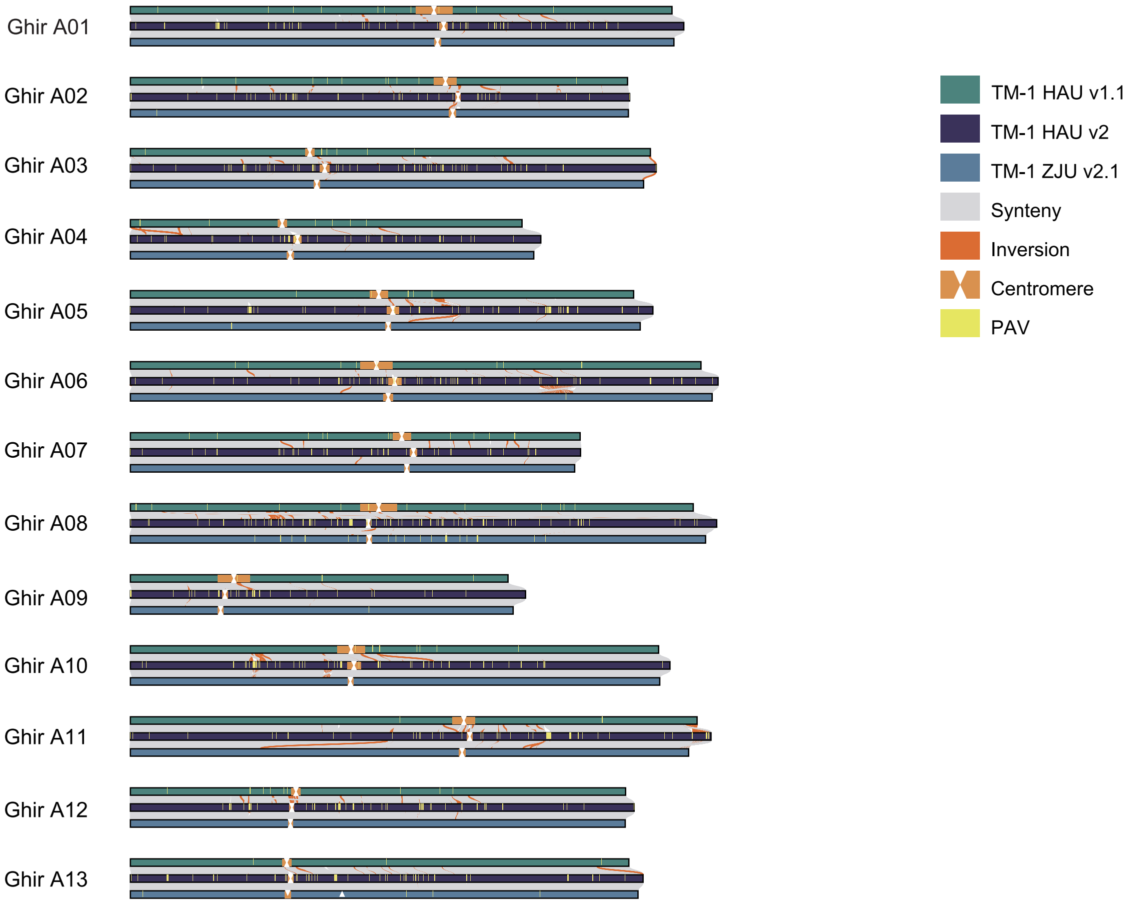

Figure 1. Genome collinearity.
Syntenic alignments among the TM-1 HAU v2, TM-1 HAU v1.1, and TM-1 ZJU v2.1 reference genomes. Syntenic regions are shown by gray lines. All presences (PAVs) are shown in yellow blocks. Centromeres are shown in khaki. The inversions between the TM-1 HAU v2 and TM-1 HAU v1.1 genomes and the TM-1 HAU v2 and TM-1 ZJU v2.1 genomes are shown in orange blocks.

markedly enriched in centromeric regions of other chromosomes (Supplemental Figure 9B). No gaps were detected in the centromeric region of chromosome D08, indicating that this centromere was also fully assembled.

Compared with the centromere positions reported in the TM-1 ZJU v2.1 and ZJU Hai7124 genomes, those in our new assemblies were improved in completeness and length. In total, 14 and 5 centromeres were gap free for the TM-1 HAU v2 and 3-79 HAU v3 genomes, and the remaining centromeres were substantially improved (Supplemental Tables 18 and 19), indicating that these centromeric regions are the most complete among currently released genomes. To assess genome completeness of the centromeres, we mapped Illumina, CLR, and Oxford Nanopore reads to the gap-free centromeres with 500-kb flanking regions and found that the ONT and CLR reads mapped evenly with no obvious breakpoints (Figure 2B; Supplemental Figures 10–13), again demonstrating the high quality of the genomes.

We observed that the DNA methylation percentage was highest in pericentromeric regions and decreased in centromeric regions, consistent with studies of maize, rice, and *Arabidopsis thaliana* (Zhang et al., 2008; Yan et al., 2010; Koo et al., 2011; Naish et al., 2021). CG methylation was higher than CHG methylation and CHH methylation in centromeric regions (Figure 2C; Supplemental Figures 14 and 15).

To better understand centromere composition, we analyzed that of TM-1 HAU v2 because of its higher genome quality. Sequence analysis revealed that 94.04% of the centromeric sequence of the TM-1 HAU v2 genome consisted of TEs, and no satellite repeats were observed, in contrast to human and *Arabidopsis*, which harbor abundant satellite repeats in their centromeres (Wu and Manuelidis, 1980; Willard, 1989; Maluszynska and Heslop-Harrison, 1991), suggesting that TEs have had a major influence on centromere formation. Unclassified LTRs (40.45%) and *Gypsy*-type retrotransposons (36.56%) were the major TE components of centromeric regions, and *Copia*-type retrotransposons were scarce (0.52%) in centromeres compared with the whole genome (4.4%). These results suggested that tetraploid cotton centromeres arose from retrotransposons and that *Gypsy*-type retrotransposons contributed more than *Copia* during centromere evolution (Figure 2D). In addition to LTR transposons, we found a large proportion of DNA transposons (16.44%) and grouped them into six superfamilies, the most abundant of which was the Mutator DNA transposon superfamily (13.78%) (Figure 2D). Notably, the proportion of Mutator in the centromeric region was greater than that in the non-centromeric

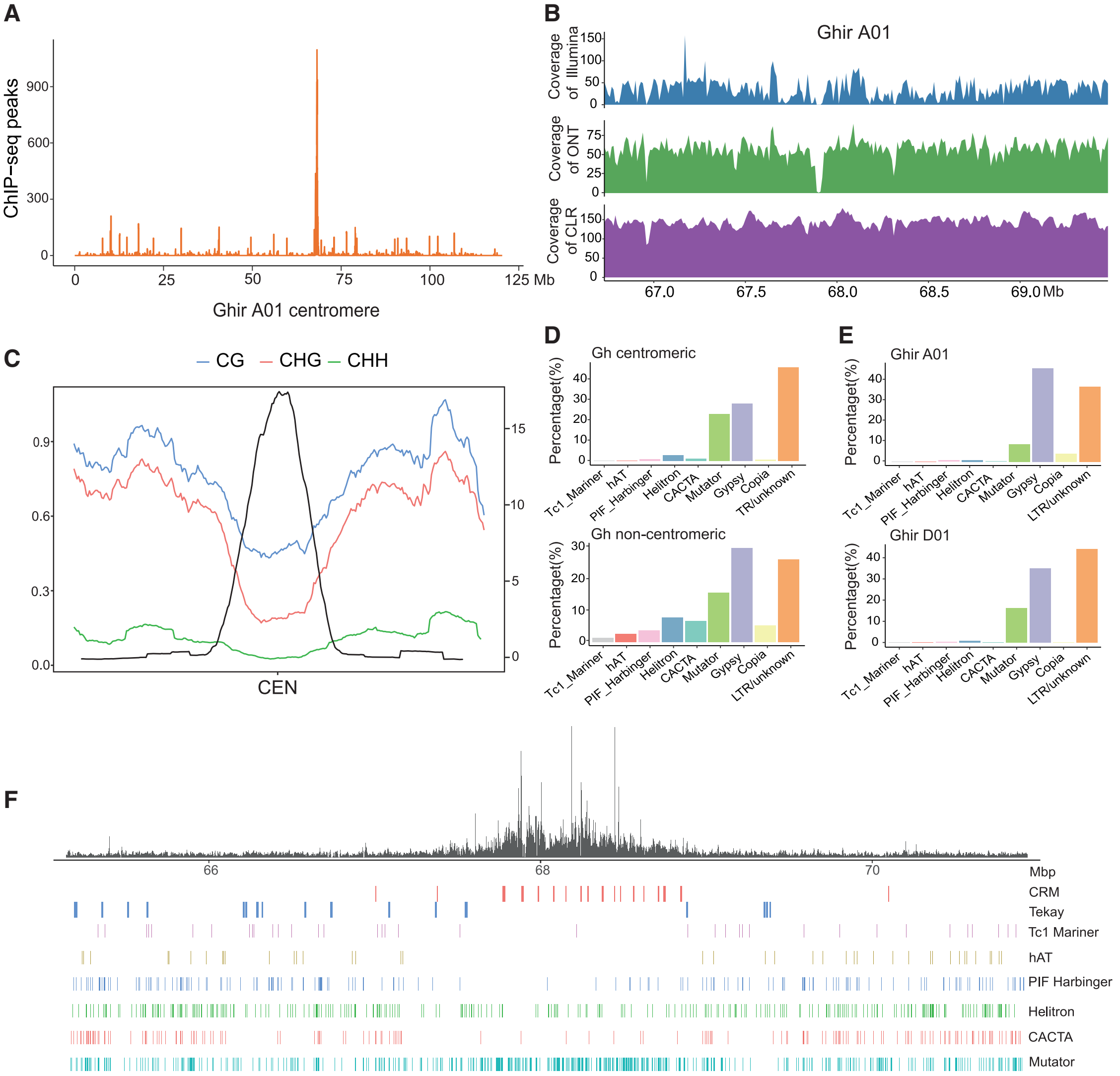

Figure 2. Characterization of centromeric regions in *G. hirsutum*.

(A) Whole-genome mapped CENH3 ChIP-Seq read peaks of chromosomes, with the A01 chromosome of *G. hirsutum* as an example. The x axis shows the positions on the chromosome. The y axis shows the normalized density of uniquely mapped reads at the centromere position. Sharp peaks indicate centromere regions.

(B) Illumina, CLR, and ONT read coverage of the complete *G. hirsutum* A01 centromere and 500-kb upstream and downstream regions.

(C) Methylation levels of centromeric regions (extended 2 Mb left and right). The ChIP-Seq peak region is the centromere position. The black curve indicates the CENH3 ChIP-Seq binding peak.

(D) Content of transposable elements (TEs) in centromeric and non-centromeric regions of *G. hirsutum*.

(E) Proportion of TEs in the A01 and D01 chromosomes of *G. hirsutum.*

(F) Distribution of DNA transposons and intact LTRs in the centromere (extended 2 Mb left and right) of the A01 chromosome of *G. hirsutum*.

region (5.94%) in the TM-1 HAU v2 genome. The content of Mutator and Helitron DNA transposons was significantly higher in centromeres of the D_t subgenome than in those of the A_t subgenome (Figure 2E; Supplemental Figure 16) (P = 6.103e−04 and P = 0.01635, Wilcoxon test).

To further explore the centromere landscape, we profiled the TE distribution of centromeres in the TM-1 HAU v2 genome. We classified the reverse transcriptase (RT) domains of 967 intact LTR retrotransposons into nine clades and found that centromeres were mainly invaded by the *CRM* (698) and *Tekay* (60) families, which were interspersed with DNA transposons (Figure 2F; Supplemental Tables 20–23). The *CRM* family was concentrated in the CENH3-enriched region within the centromere, whereas the *Tekay* family was rarely found in the CENH3-enriched region and more frequently found in the CENH3-depleted region (Figure 2F). This observation suggests that the *CRM* and *Tekay* LTRs within centromeric regions may play an important role in the establishment of centromere architecture and function.

We next predicted the insertion time of full-length LTRs in the centromeric region and found that LTRs in the centromeres were significantly younger than those in the whole genome (Supplemental Figure 17A). To avoid miscalculation caused by the number of centromeric LTRs, we randomly selected the same number of complete LTRs from non-centromeric regions to calculate the LTR insertion time (Supplemental Figure 17B). We found that the LTR insertion time of non-centromeric regions was similar to that of the whole genome, confirming that centromeric LTRs were younger and more active and indicating rapid evolution of the centromeres. To investigate the difference between sequences inside and outside the centromeres, we constructed a phylogenetic tree using complete LTR-RTs. Overwhelmingly, most of the non-centromeric LTRs clustered together, and centromeric LTRs clustered together, suggesting a difference between LTRs from centromeric and non-centromeric regions (Supplemental Figure 18A). In centromeric regions, no specific LTRs were found to be inserted into homoeologous chromosomes between the two subgenomes (Supplemental Figure 18B).

CRM and *Tekay* have shaped the architecture of cotton centromeres

To study the evolutionary history of centromeres in allotetraploid cotton, we identified 12 and 11 centromere positions in the ancestral diploids *G. raimondii* (D_5) and *G. arboreum* (A_2), respectively (Supplemental Tables 24 and 25; methods). In the A_2 genome (Wang et al., 2021), we identified 63.8% LTRs from the *Tekay* family, which were evenly distributed throughout the centromeric regions, and 16.7% LTRs from the *CRM* family. In the D_5 genome (Wang et al., 2021), the centromeres contained 86.8% *CRM* retrotransposon elements, which were distributed throughout the centromeric regions, and 1.6% LTRs from the *Tekay* family (Figure 3A and 3B). The classes of DNA transposons were the same as those in centromeres of tetraploid cotton.

To gain insight into the evolution of centromere composition, we investigated the content and distribution of *CRM and Tekay* in the A_2 and D_5 genomes. *CRM* content was increased in the A_t centromeric regions of the two tetraploid cottons relative to the A_2 genome (from 16.7% to 65.4% and 70.9%), whereas the *Tekay* content was decreased in the tetraploids (from 63.8% to 8% and 8%). However, the centromeres of the D_t subgenome showed the opposite trend: *CRM* content decreased (from 86.8% to 71.2% and 73.5%), and *Tekay* content increased (from 1.3% to 4% and 10%). We also observed that *Tekay* content was decreased in the non-centromeric A_t subgenome but increased in the non-centromeric D_t subgenome (Figure 3A). These findings suggest that the proportions of *Tekay* insertions in centromeric and non-centromeric regions were similar in the D_t subgenome and that *CRM* was more actively inserted into centromeric regions than non-centromeric regions in the A_t subgenome after polyploidization.

More interestingly, the distributions of *CRM* and *Tekay* were also altered dramatically. In contrast to the A_2 centromere, the CENH3-enriched region of the A_t subgenome was invaded and occupied by *CRM*, and *Tekay* was dramatically reduced and located more peripherally (Figure 3B). To further examine this change, we analyzed the CENH3-binding capability of these two LTRs. As expected, *CRM* and the surrounding sequences showed higher levels of binding than *Tekay* (Figure 3C), suggesting that CENH3 shows a preference for binding to *CRM*. These results demonstrated that *CRM* played a vital role in positioning and shaping the structure of the A_t subgenome centromeres in tetraploid cotton. The D_t subgenome centromeres also differed from the ancestral diploid D_5 centromeric regions. Although *CRM* still occupied the CENH3-enriched regions of the D_t subgenome, a few *Tekay* LTRs were inserted into the D_t subgenome centromere regions (Figure 3B). These findings suggest that centromere structures tended to be similar between subgenomes after polyploidization.

To further understand the evolutionary history of *CRM* and *Tekay*, we constructed a phylogenetic tree of centromeric LTRs colored by the A/D subgenomes and clustered them into five subclades. *CRM* formed a monophyletic group, indicating that it was derived from a common ancestor and underwent relatively recent amplification within the centromere (Figure 3D) and suggesting that *CRM* was derived from the D_t subgenome centromere. LTR identity comparisons also showed that *CRM* identities, on average, were significantly higher than those of *Tekay* in the centromeres (*CRM*: 98.86%; *Tekay*: 98.76%, $P < 2.2e-16$ in TM-1 HAU v2, Wilcoxon test), indicating that centromeric *CRMs* were younger than *Tekays*, and the *CRM* and *Tekay* retrotransposon elements were younger in centromeric than in non-centromeric regions (*CRM*: 98.05%, $P < 2.2e-16$; *Tekay*: 98.25%, $P = 0.0002513$, Wilcoxon test) (Supplemental Figure 18C), indicating the expansion of *CRM* in centromeres. These results suggested that the *CRM* from the D_t subgenome centromeres may have invaded the A_t subgenome centromeres after polyploidization and been amplified to reshape centromeric structure. We also observed that LTRs from the A_t subgenome and D_t subgenome were frequently intertwined (Figure 3D). This evidence suggests that *CRM* and *Tekay* interweaved across the subgenomes, driving the convergent evolution of centromere architecture in tetraploid cotton. These findings reveal changes in centromeres during the evolution of allotetraploid cotton and provide a new perspective for study of the polyploidization process.

Centromere repositioning between *G. hirsutum* and *G. barbadense*

Given that *G. hirsutum* and *G. barbadense* originated from the same hybridization event, the centromere positions of each homologous chromosome are expected to be the same in these two cotton species (Wendel, 1989; Grover et al., 2015). However, we observed six shifts in centromere physical positions between TM-1 and 3-79, which were caused by large pericentric inversions (Figure 4A; Supplemental Figure 19). This phenomenon has been described as centromere repositioning in some animals and plants (Montefalcone et al., 1999; Carbone et al., 2006; Han et al., 2009). Interestingly, although the A02 chromosome underwent a pericentric inversion event, the centromere did not form in the expected position (Supplemental Figure 19), suggesting that evolutionary forces mildly resist positional change. We also examined the effect of the centromeric shift on gene expression levels. Because of centromere repositioning, a total of 102 genes located near the centromere in TM-1 and their homologous genes in 3-79 were relocated distal to the centromere, and 351 genes underwent this change in 3-79; however, there were no significant changes in gene expression levels ($P = 0.1844$, $P = 0.5665$).

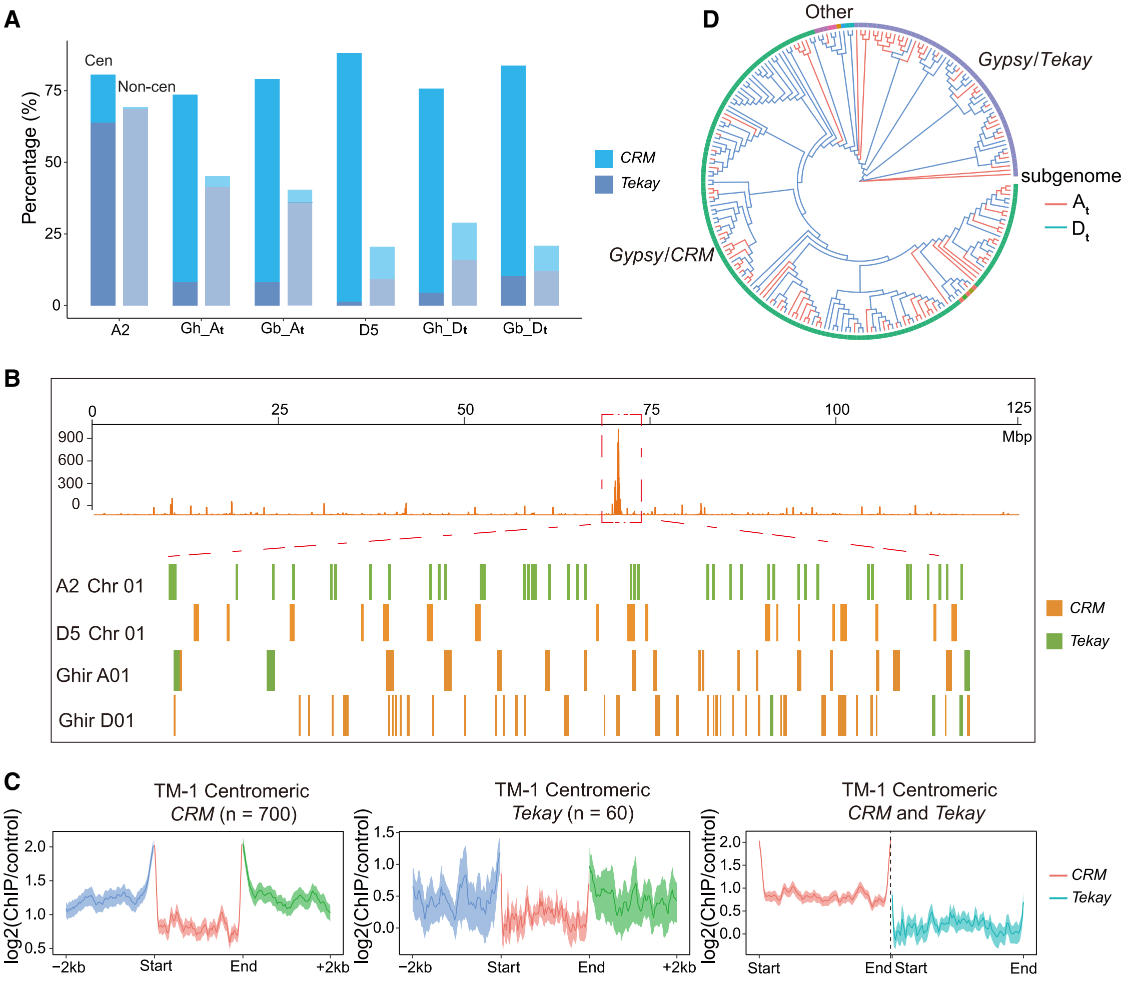

Figure 3. Differences in the distribution of centromeric *CRM* and *Tekay* retroelements in the progenitors *G. arboreum* (A_2) and *G. raimondii* (D_5) and the allotetraploid cotton species *G. hirsutum* $(AD)_1$ and *G. barbadense* $(AD)_2$.

(A) Percentage of *CRM* and *Tekay* elements in centromeric and non-centromeric regions of the *G. arboretum* and *G. raimondii* genomes and the subgenomes of allotetraploid cotton *G. hirsutum* and *G. barbadense*. The left columns with darker colors indicate centromeric regions, and the right columns with lighter colors indicate non-centromeric regions.

(B) Distribution of *CRM* and *Tekay* elements in diploid progenitors and allotetraploid cotton. The CENH3 ChIP-Seq signal on the centromere of the TM-1 Ghir_A01 chromosome is shown. The second to fifth tracks show the distribution of LTR elements in an enlarged view of the centromere region.

(C) ChIP-Seq signals around centromeric *CRM* and *Tekay* LTRs. Shaded ribbons indicate 95% confidence intervals of the mean values for windows.

(D) Phylogenetic tree of LTRs from centromeric regions, with branches colored by subgenomes in *G. hirsutum*. Different colored blocks on the outer circle represent different LTR families. Red lines indicate the A_t subgenome, and blue lines indicate the D_t subgenome.

The centromeric sequence similarity of chromosomes between TM-1 and 3-79 showed that centromeric sequence similarity was higher in the A_t subgenome than in the D_t subgenome, and centromeric similarity between the two subgenomes was higher in TM-1 than in 3-79 (Figure 4B), indicating that *G. hirsutum* underwent intense sequence variation during its formation and differentiation, perhaps due to substantial TE insertion into the centromere region. To further investigate the sequence variations that occurred within the centromeres during evolution, we mapped the CENH3 ChIP-Seq reads from *G. hirsutum* and *G. barbadense* to the genome of the *G. raimondii* (D_5) progenitor. The reads from *G. barbadense* showed partially significant peaks (10 out of 13 chromosomes) in the D_5 genome that coincided with the ancestral centromere position (Supplemental Figure 20). However, no obvious peaks were preserved in any ancestral centromere position of D_5 for *G. hirsutum* CENH3 ChIP-Seq reads (0 out of 13 chromosomes) (Figure 4C; Supplemental Figure 21).

To further clarify the divergence and conservation of cotton centromeres after the polyploidization event, we mapped the *G. raimondii* ChIP-Seq data to the *G. hirsutum* and *G. barbadense* genomes. Although binding peaks of CENH3 ChIP-Seq reads were observed on both genomes, the peak regions were not consistent with the actual centromere regions of the genomes. There were 10/13 CENH3 peaks consistent with *G. barbadense* centromere positions and 6/13 CENH3 peaks consistent with *G. hirsutum* centromere positions (Supplemental Figure 22). We also mapped the CENH3 ChIP-Seq reads from *G. hirsutum* and *G. barbadense* to the diploid A_2 genome (Supplemental Figures 23 and 24). The results were consistent with the mapping of CENH3 ChIP-Seq reads to the diploid D_5 genome. There were clear peaks when

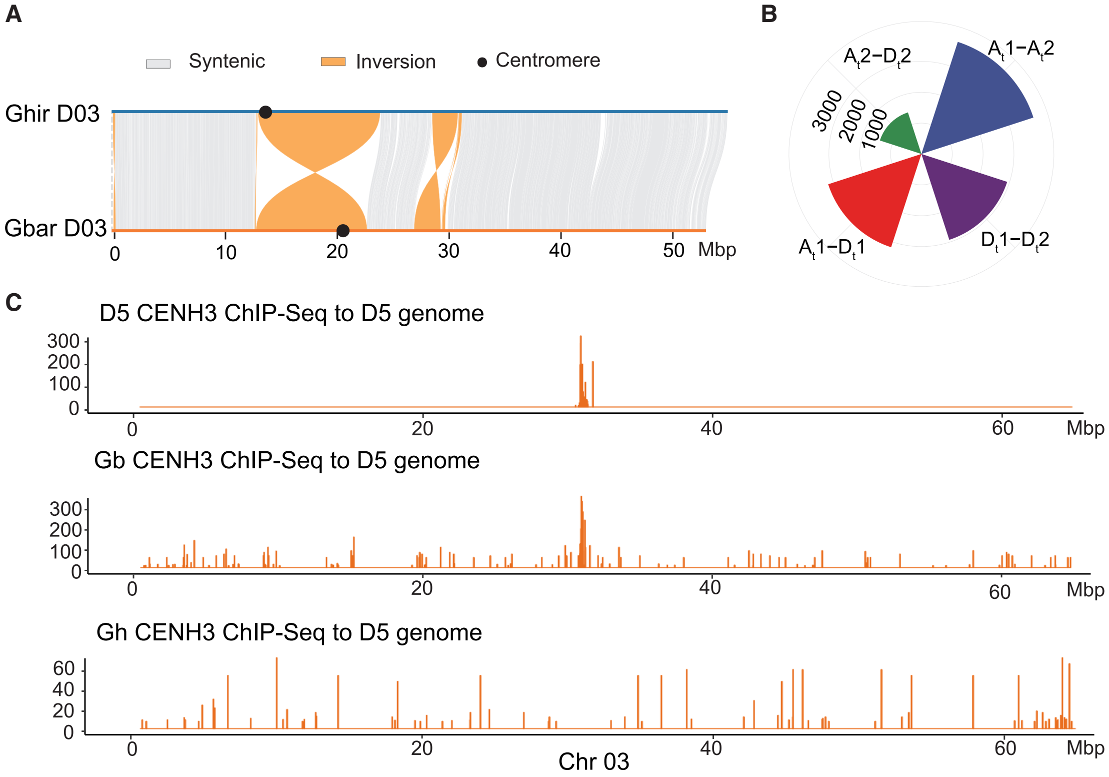

Figure 4. Centromeric sequence divergence between *G. hirsutum* and *G. barbadense*.

(A) A pericentric inversion causes a shift in centromere position. The black point represents the centromere position. The orange block represents the pericentric inversion.

(B) Centromeric sequence similarity between *G. hirsutum* and *G. barbadense*. The y axis shows the number of conserved blocks after filtering with the criteria block length > 2000 bp and identity > 95%.

(C) Genome-wide mapping of *G. hirsutum* CENH3 ChIP-Seq reads to the *G. raimondii* reference genome. The first track shows the distribution of *G. raimondii* CENH3 ChIP-Seq reads along chromosome (Chr) 03 of the *G. raimondii* reference genome, which indicates the real centromere position. The second track shows the density of *G. barbadense* CENH3 ChIP-Seq reads along Chr 03 of the *G. raimondii* reference genome. The third track shows the enrichment of *G. hirsutum* CENH3 ChIP-Seq reads along Chr 03 of the *G. raimondii* reference genome.

the CENH3 ChIP-Seq reads from *G. barbadense* were mapped to the A_2 genome, but no clear peaks were present at any ancestral centromere position of A_2 for the *G. hirsutum* CENH3 ChIP-Seq reads. These results indicate that the ancestral centromeric sequences were mainly retained in *G. barbadense* but that substantial variation occurred in the centromeres of *G. hirsutum*. These results highlight the evolutionary divergence of centromeric sequences during the evolution and differentiation of *G. hirsutum* and *G. barbadense*.

Expression and distribution of centromere genes

We detected 150 and 227 high-confidence genes in the centromeric regions of the TM-1 HAU v2 and 3-79 HAU v3 genomes (Supplemental Table 26), respectively. In total, 67 in TM-1 (44.6%) and 98 in 3-79 (43.2%) were expressed in at least one tissue (Fragments Per Kilobase of exon model per Million mapped fragments (FPKM) > 1), as supported by RNA-Seq data from 27 and 10 different tissues or developmental periods in TM-1 and 3-79, respectively. The expression levels of centromeric genes were significantly lower than those of non-centromeric genes ($P < 2.2e-16$). Surprisingly, TM-1 centromeric genes showed significantly lower expression levels than 3-79 centromeric genes ($P < 2.2e-16$; Figure 5A). The percentage of expressed genes was also unbalanced in centromeres from the two subgenomes. In TM-1, 31.58% of the genes in the A_t subgenome centromeres were expressed, compared with 50% of those in the D_t subgenome centromeres. A similar phenomenon was also observed for 3-79, with 30.33% of genes in the A_t subgenome centromeres and 58.2% of the genes in the D_t subgenome centromeres expressed. Interestingly, the distribution of expressed gene numbers was similar for TM-1 and 3-79 chromosomes, with the highest number of expressed genes located on chromosomes Ghir D12 and Gbar D12 (Supplemental Table 26). In addition, 24 homologous genes were found in the centromere, 21 of which were located in the D_t subgenome, and some genes had structural variations between TM-1 and 3-79. For instance, in a pair of homologous genes, Ghir_D04G16752 was not expressed in *G. hirsutum*, but Gbar_D04G15822 was expressed in *G. barbadense*.

One hundred fourteen centromeric genes in *G. hirsutum* had orthologous genes in *G. barbadense*, 15 of which were located in non-centromeric regions of *G. barbadense*. Likewise, 169 centromeric genes in *G. barbadense* had orthologous genes in *G. hirsutum*, 70 of which were located in non-centromeric regions of *G. hirsutum*, indicating that these genes may have been transferred from centromeric to non-centromeric regions during

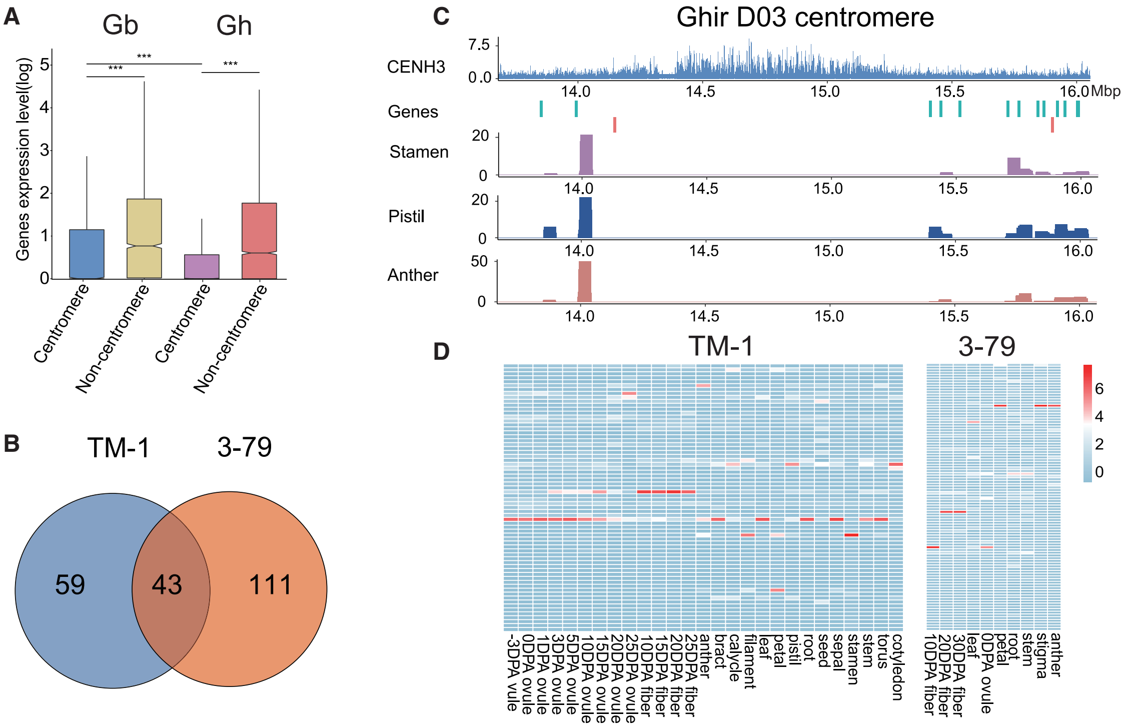

Figure 5. Distribution and expression levels of centromeric genes in CENH3 and H3 subdomains of *G. hirsutum*.

(A) Comparison of gene expression levels between centromere and non-centromere regions of TM-1 and 3-79. Asterisks indicate extremely high significance (***$P < 2.2e-16$).

(B) Venn diagram of gene families in the centromeric regions of TM-1 and 3-79.

(C) Gene distribution and expression levels in the centromeric region. The first track shows the CENH3 ChIP-Seq read density at the centromere of the D03 chromosome of *G. hirsutum*, with clear CENH3-enriched and CENH3-depleted subdomains. The second and third tracks show gene distribution in the centromere. Expressed genes (FPKM $\geq$ 1) are shown in blue, and non-expressed genes (FPKM < 1) are shown in red. The fourth to sixth tracks show gene expression levels in the stamen, pistil, and anther.

(D) Heatmaps of centromeric gene expression levels in different tissues/periods for TM-1 and 3-79. All genes were expressed (FPKM $\geq$ 1) in at least one tissue, and the gene expression level was normalized.

species divergence. Six genes were located in the *G. hirsutum* presence regions, and 27 genes were located in the *G. barbadense* presence regions. Further comparative analysis revealed 43 shared gene families in centromeres of TM-1 and 3-79 (Figure 5B). Collinearity analysis revealed that 54 of all centromere genes were syntenic. These findings highlight the opportunity provided by the improved genomes to explore the structure and function of genes in centromeric regions.

We observed an uneven distribution of CENH3 peaks within centromeres, with CENH3-enriched and CENH3-depleted subdomains, as previously reported for CENH3 distribution in rice and other species (Yan et al., 2008; Su et al., 2016). We found that the majority of putative centromeric genes were located in CENH3-depleted subdomains and were expressed in at least one tissue. For instance, in the centromeric region of the D03 chromosome of *G. hirsutum*, we observed obvious CENH3-enriched and CENH3-depleted subdomains, with no genes in the CENH3-enriched region, and 13 of the 15 genes in H3 subdomains were expressed (Figure 5C). These results suggested that the absence of genes and transcription might provide an environment for CENH3 binding.

To examine the overall expression levels of centromeric genes (FPKM > 1 genes), we constructed a heatmap of gene expression levels. Most expressed genes in centromeric regions were weakly expressed (1 < FPKM < 5) in most tissues. Only seven genes of TM-1 were highly expressed in some tissues and/or developmental stages (Figure 5D); these included Ghir_D13G72549 (FPKM > 26), which was highly expressed in the ovule, root, leaf, and stem. Only five genes were highly expressed in some tissues and/or developmental stages of 3-79; these included Gbar_D12G68262, which was highly expressed in fiber at 10 days post anthesis (DPA) (FPKM > 73), and Gbar_A07G37900, which was highly expressed in fiber at 20 DPA (FPKM > 130) and 30 DPA (FPKM > 130). We annotated the centromeric genes to better understand their functions and biological processes. Approximately 94% and 95% of the expressed centromeric genes (FPKM > 1) were assigned GO terms in TM-1 and 3-79, respectively. As expected, most centromeric genes were associated with the inheritance of genetic information, including processes such as cell-cycle phase transition, cytoskeletal motor activity, and chromatin organization or disassembly. However, centromeric genes were also involved in xyloglucan metabolic processes, plant-type secondary cell wall biogenesis, and other biological processes (Supplemental Tables 27 and 28). These results raise the possibility that centromeric genes participate in the inheritance of genetic information, reflecting the importance of higher-quality assembly of the centromere region.

DISCUSSION

Improved assembly quality for *G. hirsutum* and *G. barbadense*

We integrated multiple technologies, including ONT reads, PacBio CLR reads, Illumina paired-end reads, and Hi-C data, to *de novo* assemble the genomes of *G. hirsutum* (TM-1) and *G. barbadense* (3-79). The contig N50 was improved by 4.6-fold and 5.6-fold, respectively, compared with recently published genomes, indicating improvements in genome continuity (Hu et al., 2019; Wang et al., 2019; Yang et al., 2019; Chen et al., 2020; Huang et al., 2020). We also added 93 Mb of new DNA sequence compared with the TM-1 HAU v2 reference genome and filled hundreds of gaps with longer reads, providing the first assembly of gap-free centromeres in cotton. This will enable new studies of the sequence, structure, epigenetic modification, and evolution of centromeres. The updated reference genome also enabled us to annotate more genes that were absent in previous reference genomes, most of which were supported by BUSCO evaluation and RNA-Seq data or full-length transcripts. These resources will advance the study of genome structure and function in cotton. However, complete and accurate assembly of the centromere region still requires a combination of multiple sequencing technologies and extensive, time-consuming manual correction processes, as well as ingenious design ideas (Naish et al., 2021; Song et al., 2021; Nurk et al., 2022).

The role of *CRM* and *Tekay* in centromere evolution

Maize centromeres contain four kinds of *CRM*, *CRM1–CRM4* (Schnable et al., 2009), which have different functions in the centromere region (Sharma and Presting, 2008). *CRM1* and *CRM2* constitute the majority of the centromeric sequences. *CRM2* is found at deposition sites of CENH3 histone and can phase CENH3 nucleosomes by the R-loop (Gent et al., 2011). *CRM1* is the most active *CRM* element in the centromere and has been associated with expansion of centromere size during maize evolution (Sharma and Presting, 2008). Liu et al. (2020) revealed that circular RNA from *CRM1* was associated with CENH3 localization. Furthermore, R-loops derived from *CRM1* and *CRM2* play an important role in maintenance of centromere function and formation of pericentromeric heterochromatin (Liu et al., 2021). Han et al. (2016) found that seven cotton centromere-related retrotransposons also belonged to group A of the *CRM* clade, whose members can recognize the centromeric chromatin region.

In *G. hirsutum* (TM-1) and *G. barbadense* (3-79), centromeres were mainly occupied by members of the *CRM* and *Tekay* families. Invasion and amplification of *CRM* from the D_t subgenome in the CENH3-enriched domain shaped centromere architecture of the A_t subgenome after polyploidization, and *Tekay* was gradually marginalized. In previous studies, some researchers have also reported invasion of the A genome by CRG centromeric retrotransposon elements from the D genome after allopolyploid formation (Luo et al., 2012; Han et al., 2016). In this study, we found that CRG elements were structurally annotated as *CRM* LTR elements. This evidence further supports our conclusion. The concentration of *CRM* in CENH3-enriched regions may provide a structural foundation for recruitment of histone CENH3 and other centromere-associated proteins, whereas *Tekay* in the CENH3-depleted region may serve a different function, highlighting the structural discrepancy between CENH3-enriched and CENH3-depleted regions within centromeres. After polyploidization, centromere structure of the two subgenomes became more similar. We found strong intertwining of centromeric *CRM* and *Tekay* in the two subgenomes, suggesting that these two LTRs might drive the convergent evolution of centromeres to maintain centromere function for pairing of homologous chromosomes during meiosis and to ensure that allopolyploids adopt the proper conformation for association during meiosis to avoid unbalanced gamete formation. Previous experimental evidence also showed that centromeric sequences were not conserved across species (Balzano and Giunta, 2020), but the structure and function of centromeres were highly conserved. Melters et al. (2012) confirmed that convergent evolution of centromere structure (rather than sequence) is critical to centromere function. This finding provides an alternative strategy for exploring polyploidy in plants.

Effects of human breeding activities and the environment on evolution of centromere sequences

We found that the centromere sequences of upland cotton were more active than those of sea island cotton, a result that may reflect breeding activities and environmental adaptability. Upland cotton has a high fiber yield and strong environmental adaptability and is widely planted all over the world, whereas sea island cotton is planted only in some areas (Hu et al., 2019; Wang et al., 2019; Li et al., 2023). Upland cotton can survive and reproduce in a variety of environmental conditions, and this wide adaptability may lead to more variation in the centromere sequence to adapt to different environments (Hutchinson, 1951; Ulloa et al., 2007; Fang et al., 2017). In addition, the extensive selection of upland cotton varieties by breeders has caused a greater selection pressure, and selected genes linked to the centromere may promote variation in centromere sequence (Li et al., 2023). Studies of maize centromere sequences have shown that changes in centromere DNA repeats are driven largely by selection for centromere-linked genes. Post-domestication selection of centromere-associated genes favors selection that may affect key domestication or agricultural traits (Schneider et al., 2016). By contrast, sea island cotton has a relatively narrow growth range and is planted in a specific area, limited by specific ecological conditions (Ulloa et al., 2007; Ma et al., 2021). More importantly, sea island cotton varieties are less frequently selected by breeders, probably resulting in relatively less variation in centromeric sequences.

Functional implications of centromeric genes

The centromere is located in the heterochromatin region of the chromosome, which contains a large number of repeat sequences and is considered to be transcriptionally silent (Yan et al., 2008; Wu et al., 2011). However, active genes were recently found in centromeric regions of rice, *Arabidopsis*, potato, and wheat (Copenhaver et al., 1999; Nagaki et al., 2004; Wu et al., 2004; Gong et al., 2012; Su et al., 2019). In rice, for example, *RH2* controls setting rate, *OsGT61-1* regulates plant height, *OsTrx1* controls flowering time, and *OsSNAP32* is involved in response to biotic and abiotic stresses (Bao et al., 2008; Singh et al., 2010; Chern et al., 2012; Choi et al., 2014). However, few such studies have been reported in cotton. Here, we systematically analyzed the expression and function of centromeric genes in allotetraploid cotton. Some centromeric genes showed high expression levels in specific tissues and periods. Functional annotation revealed that most centromeric genes were associated with cell division, but some were also involved in other important metabolic pathways, such as the hemicellulose metabolic process and cellular response to stress. These centromeric sequences provide an unprecedented opportunity to understand the function and evolution of centromeric genes and to reveal the association between centromeric genes and important agronomic traits, thus facilitating the improvement of cotton germplasm.

METHODS

Plant materials and DNA extraction

TM-1 and 3-79 were planted in Wuhan, China. We collected fresh young leaves from each species and isolated high-quality DNA for Nanopore sequencing using the cetyltrimethylammonium bromide (CTAB) method (Paterson et al., 1993).

Library preparation and Oxford Nanopore sequencing

For construction of the Nanopore sequencing library, extracted genomic DNA was examined with the Nanodrop and Qubit systems and by 0.35% agarose gel electrophoresis. DNA fragments of 100 bp to 50 kb were collected by electrophoresis (Sage Sciences) and processed using the Ligation Sequencing 1D Kit (Catalog No. SQK-LSK109; ONT, Oxford, UK) according to the standard protocol. The ONT platform was used to construct and sequence the DNA libraries. In total, 345.3 Gb and 291.83 Gb of clean data were generated for *G. hirsutum* and *G. barbadense*, respectively. The N50 values of ONT reads for the two species were 34 668 and 28 839 bp, and the largest reads were 263 147 and 177 895 bp in length.

Genome assembly, correction, and chromosome anchoring

We assembled the TM-1 and 3-79 genomes with Canu (version 2.1.1) (Koren et al., 2017), which included correction, trimming, and assembly in three steps. We performed these steps manually. First, ONT reads and PacBio reads from both genomes were corrected and trimmed using Canu with default parameters (correctedErrorRate = 0.045 for PacBio reads; correctedErrorRate = 0.144 for Nanopore reads). Trimmed high-quality reads from ONT (~40×) and PacBio (~40×) were delivered as input to Canu using a mix of formats with default parameters. To improve base quality, we aligned Illumina paired-end reads (~50×) to contigs using BWA–MEM and polished them with Pilon (version 1.23) (–fix bases –mindepth 10 –minmq 30) (Li, 2013; Walker et al., 2014). High-quality paired-end Hi-C reads based on DpnII (Huang et al., 2022; Pei et al., 2022) for *G. hirsutum* TM-1 and *G. barbadense* 3-79 were mapped to the two contig-scale assemblies using Juicer (version 1.6) (Durand et al., 2016b). The original contigs were organized into chromosomes with the 3D-DNA pipeline (version 180 419) (-r 2 -i 15000 –build-gapped-map) (Dudchenko et al., 2017). Finally, we used Juicebox Assembly Tools (v1.11.08) to manually correct and refine the connections (Durand et al., 2016a).

Genome assessment

The new genome assembly was aligned to previously published genomes using Mummer v4.0.0 (-c 90 -l 1000), and the delta-filter module was used to filter the output delta file (-i 99 -l 15000) (Marçais et al., 2018). A dot plot of genomic alignment was generated with mummerplot. BUSCO (v4.1.4) analysis was performed using the embryophyta_odb10 database (1614 genes) with default parameters (Manni et al., 2021).

Repeat annotation

We combined homolog-based, structure-based, and *de novo* approaches to identify repeat sequences in TM-1 and 3-79. First, we identified genome-specific repetitive elements using RepeatModeler with default parameters to construct a custom TE library (Flynn et al., 2020). Second, we used RepeatMasker with default parameters, combined with the *de novo* TE library of TM-1 and Repbase (Smit et al.), to identify repeats and obtain a detailed annotation of the repeats present in the genome assembly. TEs were annotated using EDTA software, which integrates LTR_FINDER, LTRharvest, Generic Repeat Finder, TIR-Learner, MITE-Hunter, and HelitronScanner software (Ou et al., 2019). The same approach was used to annotate repeats in the 3-79 genome.

Gene annotation

Repeat regions were masked with RepeatMasker using the custom repeat database generated by RepeatModeler. Genes were predicted by a combination of *ab initio*- and homology-based methods as described previously (https://github.com/xingchang-web/NAM-genomes/tree/master/gene-prediction). For transcription evidence, RNA-Seq data from 27 tissues of TM-1 and 10 tissues of 3-79 were downloaded from NCBI (Supplemental Table 29). The assembled transcripts were generated using Trinity, StringTie, Cufflinks, and Strawberry, and the optimal transcripts for each locus were then selected from these assembled transcripts using Mikado (Haas et al., 2003; Trapnell et al., 2012; Liu and Dickerson, 2017; Kovaka et al., 2019). To obtain the assembled transcripts, we first mapped RNA-Seq reads to each genome using two rounds of STAR (–outSAMattributes All, –outSAMmapqUnique 10, –outFilterMismatchNmax 0). The splice junctions produced in the first round were used in the second round to improve the alignments. For selection of optimal transcripts, the high-confidence splice junctions and mapped RNA-Seq reads generated in the above steps were used to predict open reading frames of the assembled transcripts (Mapleson et al., 2018), which were generated by TransDecoder (B.J. Haas, https://github.com/TransDecoder/TransDecoder), and the full-length transcripts were searched against SwissProt.

For *ab initio* prediction, we used BRAKER, which combines GeneMark and AUGUSTUS software, with the masked genome (Brůna et al., 2021). Proteins were generated by the above evidence-based pipeline and from mapped RNA-Seq read bam files by STAR. BRAKER was run in two rounds with default options. The output file was in GFF3 format (–gff3). Finally, we used PASA software with default options to improve the models iteratively obtained from *ab initio*- and homology-based methods. Finally, TE-overlapping genes were filtered using TEsorter tools (version 1.3) (Zhang et al., 2022). Functional annotation of genes was performed using the eggNOG-Mapper online tool (Cantalapiedra et al., 2021).

Gene expression analysis

We chose 27 and 10 different tissues/developmental periods for TM-1 and 3-79 (Hu et al., 2019; Wang et al., 2019). HISAT2 was used to map the RNA-Seq data to the reference genome (version 2.1.0) (Kim et al., 2015), and SAMtools was used to remove duplicate reads and filter the results ($q > 20$) (Danecek et al., 2021). FPKM values were calculated with StringTie (version 2.1.4) (Kovaka et al., 2019).

Identification of structural variation

To detect the structural variation between TM-1 and 3-79, we compared the TM-1 and 3-79 genomes using Mummer (v 4.0) (Marçais et al., 2018) and identified synteny and structural rearrangements using SyRI with default parameters (version 1.5.4) (Goel et al., 2019). Presence/absence variations were identified with the scanPAV pipeline (Giordano et al., 2018), which divides the presence assembly into chunks of 1000 bp and maps them against the absence assembly. The resulting 1000-base chunks in the absence assembly were merged as absence PAVs.

Centromere identification

We detected the centromere positions of *G. hirsutum*, *G. barbadense*, and *G. raimondii* as described previously (Hu et al., 2019). First, we mapped

ChIP-Seq reads to each genome assembly with Bowtie 2 (Langmead and Salzberg, 2012). Uniquely mapped reads (1-bp mismatch allowed and mapping quality >30) were retained. Then, we used SICER v1.1 to identify the CENH3 domain (-w 200, -egf 0.98 -fdr 0.01 -g 600 -e 1000) (Zang et al., 2009). Islands with enrichment levels greater than 5 and false discovery rate threshold lower than 0.01 were retained and gaps of 1000 bp were allowed when defining the CENH3 domains. Locations of CENH3 ChIP-Seq peaks in terminal centromeric regions were determined by visual inspection. The centromere position of Ghir_D08 was identified by mapping the CRG1 and CRG2 sequences to the Ghir_D08 chromosome by BLASTN alignment (Luo et al., 2012). To characterize the centromeres of *G. arboreum* (A_2), we aligned the centromeric sequences of TM-1 and 3-79 to the A_2 genome using Mummer. According to the criteria of identity >90% and alignment length >2500 bp, we filtered the sequence and used a custom Python script to calculate the 95% confidence interval to determine the representative centromeric regions. To accurately delimit the boundaries of the centromeres as much as possible, we discarded centromere regions that were inconsistent with those reported in Wang et al. (2021), which were detected on the basis of centromere-related LTR retrotransposons. We used a custom Python script to identify the breakpoint in the centromere, which did not contain "Ns" and was considered to be a complete centromere. PacBio CLR reads and ONT reads were mapped to the genomes with Minimap2 (Li, 2018).

TE analysis of centromeres

Centromeric LTR sequences were extracted using BEDTools (version v2.27.0) with the EDTA annotations to further identify TE protein domains using TEsorter with the REXdb database (Quinlan and Hall, 2010; Zhang et al., 2022). According to the positions of the transposons, we visualized centromere structure using ggplot2.

Construction of LTR phylogenetic tree

The RT domains were extracted from the TEsorter results using the python script concatenate_domains.py. Mafft (version 7.453) was used to perform multiple sequence alignment. A phylogenetic tree was constructed from the resulting alignment using IQ-TREE (Zhang et al., 2022).

ChIP-Seq read binding capability of centromeric TEs

A Perl script (https://github.com/oushujun/EDTA/blob/master/util/split_overlap.pl) was used to split overlapping TEs across the whole genome. A custom Python script was used to extract centromeric *CRM* elements and their flanking 2-kb regions. These regions were divided into 80 windows using BEDTools. BWA–MEM was used to align the ChIP-Seq data to the reference genome, and the bam file was filtered and sorted using SAMtools. The ChIP-Seq read depth for each window was calculated from the bam file using Mosdepth (version 0.3.2) (Pedersen and Quinlan, 2018).

Centromeric DNA methylation analysis

The bisulfite-treated DNA sequencing data were downloaded from NCBI (Supplemental Table 29) (Song et al., 2017). Trimmomatic (version 0.32) was used to trim low-quality reads (https://github.com/usadellab/Trimmomatic). The clean data were mapped to the reference genome with Bismark (version 0.13.0) (-N 1, -L 30) (Krueger and Andrews, 2011). Potentially methylated cytosines were extracted from the uniquely mapped reads using the Bismark methylation extractor with the default parameters. We used the reported bisulfite non-conversion rate (0.04). We retained the methylated cytosines that were covered by more than three reads and further filtered using the binomial distribution (*P* value cutoff 1e−5) and Fisher's exact test (false discovery rate < 0.01).

DATA AND CODE AVAILABILITY

The genome assemblies and annotations of *G. hirsutum* and *G. barbadense* are available at https://doi.org/10.6084/m9.figshare.22682833. The raw ONT sequence reads of *G. hirsutum* and *G. barbadense* are available at the Genome Sequence Archive of the China National Genomics Data Center under accession number: CRA012599 (https://bigd.big.ac.cn/gsa/browse/CRA012599).

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Plant Communications Online*.

FUNDING

This study was supported by the National Natural Science Foundation of China (32170645) and the National Key Research and Development Program of China (2021YFF1000900).We thank the high-performance computing platform at National Key Laboratory of Crop Genetic Improvement in Huazhong Agricultural University.

AUTHOR CONTRIBUTIONS

M.W. and X.Z. conceived and designed the study. X.C., X. He., R.P., R.W., and S.L. analyzed the data. X.L. and X. Hu performed the experiment. J.L. and Z.L. contributed to project discussion. X.C. wrote the manuscript draft, and M.W. and X.Z. revised it. All authors reviewed the manuscript.

ACKNOWLEDGMENTS

We thank the high-performance computing platform at the National Key Laboratory of Crop Genetic Improvement at Huazhong Agricultural University. No conflict of interest is declared.

Received: April 24, 2023
Revised: June 16, 2023
Accepted: September 19, 2023
Published: September 22, 2023

REFERENCES

Alfenito, M.R., and Birchler, J.A. (1993). Molecular characterization of a maize B chromosome centric sequence. Genetics **135**:589–597. https://doi.org/10.1093/genetics/135.2.589.

Altemose, N., Logsdon, G.A., Bzikadze, A.V., Sidhwani, P., Langley, S.A., Caldas, G.V., Hoyt, S.J., Uralsky, L., Ryabov, F.D., Shew, C.J., et al. (2022). Complete genomic and epigenetic maps of human centromeres. Science **376**, eabl4178. https://doi.org/10.1126/science.abl4178.

Ananiev, E.V., Phillips, R.L., and Rines, H.W. (1998). Chromosome-specific molecular organization of maize (Zea mays L.) centromeric regions. Proc. Natl. Acad. Sci. USA **95**:13073–13078. https://doi.org/10.1073/pnas.95.22.13073.

Appels, R., International Wheat Genome Sequencing Consortium IWGSC, Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., Pozniak, C.J., Stein, N., Choulet, F., Distelfeld, A., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science **361**, eaar7191. https://doi.org/10.1126/science.aar7191.

Balzano, E., and Giunta, S. (2020). Centromeres under Pressure: Evolutionary Innovation in Conflict with Conserved Function. Genes **11**:912.

Bao, Y.-M., Wang, J.-F., Huang, J., and Zhang, H.-S. (2008). Molecular cloning and characterization of a novel SNAP25-type protein gene OsSNAP32 in rice (Oryza sativa L.). Mol. Biol. Rep. **35**:145–152. https://doi.org/10.1007/s11033-007-9064-8.

Brůna, T., Hoff, K.J., Lomsadze, A., Stanke, M., and Borodovsky, M. (2021). BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom. Bioinform. **3**:lqaa108. https://doi.org/10.1093/nargab/lqaa108.

Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional

Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol. Biol. Evol. **38**:5825–5829. https://doi.org/10.1093/molbev/msab293.

Cappelletti, E., Piras, F.M., Sola, L., Santagostino, M., Abdelgadir, W.A., Raimondi, E., Lescai, F., Nergadze, S.G., and Giulotto, E. (2022). Robertsonian Fusion and Centromere Repositioning Contributed to the Formation of Satellite-free Centromeres During the Evolution of Zebras. Mol. Biol. Evol. **39**:msac162. https://doi.org/10.1093/molbev/msac162.

Carbone, L., Nergadze, S.G., Magnani, E., Misceo, D., Francesca Cardone, M., Roberto, R., Bertoni, L., Attolini, C., Francesca Piras, M., de Jong, P., et al. (2006). Evolutionary movement of centromeres in horse, donkey, and zebra. Genomics **87**:777–782. https://doi.org/10.1016/j.ygeno.2005.11.012.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I.A.P., Tang, H., Wang, X., Chiquet, J., Belcram, H., Tong, C., Samans, B., et al. (2014). Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. Science **345**:950–953. https://doi.org/10.1126/science.1253435.

Chen, Z.J., Sreedasyam, A., Ando, A., Song, Q., De Santiago, L.M., Hulse-Kemp, A.M., Ding, M., Ye, W., Kirkbride, R.C., Jenkins, J., et al. (2020). Genomic diversifications of five Gossypium allopolyploid species and their impact on cotton improvement. Nat. Genet. **52**:525–533. https://doi.org/10.1038/s41588-020-0614-5.

Chen, Z.J., Scheffler, B.E., Dennis, E., Triplett, B.A., Zhang, T., Guo, W., Chen, X., Stelly, D.M., Rabinowicz, P.D., Town, C.D., et al. (2007). Toward Sequencing Cotton (Gossypium) Genomes. Plant Physiol. **145**:1303–1310. https://doi.org/10.1104/pp.107.107672.

Cheng, Z., Dong, F., Langdon, T., Ouyang, S., Buell, C.R., Gu, M., Blattner, F.R., and Jiang, J. (2002). Functional Rice Centromeres Are Marked by a Satellite Repeat and a Centromere-Specific Retrotransposon. Plant Cell **14**:1691–1704. https://doi.org/10.1105/tpc.003079.

Chern, M., Bai, W., Sze-To, W.H., Canlas, P.E., Bartley, L.E., and Ronald, P.C. (2012). A rice transient assay system identifies a novel domain in NRR required for interaction with NH1/OsNPR1 and inhibition of NH1-mediated transcriptional activation. Plant Methods **8**:6. https://doi.org/10.1186/1746-4811-8-6.

Choi, S.C., Lee, S., Kim, S.-R., Lee, Y.-S., Liu, C., Cao, X., and An, G. (2014). Trithorax Group Protein Oryza sativa Trithorax1 Controls Flowering Time in Rice via Interaction with Early heading date3. Plant Physiol. **164**:1326–1337. https://doi.org/10.1104/pp.113.228049.

Copenhaver, G.P., Nickel, K., Kuromori, T., Benito, M.-I., Kaul, S., Lin, X., Bevan, M., Murphy, G., Harris, B., Parnell, L.D., et al. (1999). Genetic Definition and Sequence Analysis of Arabidopsis Centromeres. Science **286**:2468–2474. https://doi.org/10.1126/science.286.5449.2468.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience **10**:giab008. https://doi.org/10.1093/gigascience/giab008.

Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., et al. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science **356**:92–95. https://doi.org/10.1126/science.aal3327.

Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S., and Aiden, E.L. (2016a). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Syst. **3**:99–101. https://doi.org/10.1016/j.cels.2015.07.012.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016b). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst. **3**:95–98. https://doi.org/10.1016/j.cels.2016.07.002.

Earnshaw, W.C., and Migeon, B.R. (1985). Three related centromere proteins are absent from the inactive centromere of a stable isodicentric chromosome. Chromosoma **92**:290–296. https://doi.org/10.1007/BF00329812.

Fang, L., Wang, Q., Hu, Y., Jia, Y., Chen, J., Liu, B., Zhang, Z., Guan, X., Chen, S., Zhou, B., et al. (2017). Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. Nat. Genet. **49**:1089–1098. https://doi.org/10.1038/ng.3887.

Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., and Smit, A.F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. USA **117**:9451–9457. https://doi.org/10.1073/pnas.1921046117.

Gent, J.I., Schneider, K.L., Topp, C.N., Rodriguez, C., Presting, G.G., and Dawe, R.K. (2011). Distinct influences of tandem repeats and retrotransposons on CENH3 nucleosome positioning. Epigenet. Chromatin **4**:3. https://doi.org/10.1186/1756-8935-4-3.

Giordano, F., Stammnitz, M.R., Murchison, E.P., and Ning, Z. (2018). scanPAV: a pipeline for extracting presence–absence variations in genome pairs. Bioinformatics **34**:3022–3024. https://doi.org/10.1093/bioinformatics/bty189.

Goel, M., Sun, H., Jiao, W.-B., and Schneeberger, K. (2019). SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. **20**:277. https://doi.org/10.1186/s13059-019-1911-0.

Gong, Z., Wu, Y., Koblízková, A., Torres, G.A., Wang, K., Iovene, M., Neumann, P., Zhang, W., Novák, P., Buell, C.R., et al. (2012). Repeatless and repeat-based centromeres in potato: implications for centromere evolution. Plant Cell **24**:3559–3574. https://doi.org/10.1105/tpc.112.100511.

Grover, C.E., Gallagher, J.P., Jareczek, J.J., Page, J.T., Udall, J.A., Gore, M.A., and Wendel, J.F. (2015). Re-evaluating the phylogeny of allopolyploid Gossypium L. Mol. Phylogenet. Evol. **92**:45–52. https://doi.org/10.1016/j.ympev.2015.05.023.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. **31**:5654–5666. https://doi.org/10.1093/nar/gkg770.

Han, J., Masonbrink, R.E., Shan, W., Song, F., Zhang, J., Yu, W., Wang, K., Wu, Y., Tang, H., Wendel, J.F., and Wang, K. (2016). Rapid proliferation and nucleolar organizer targeting centromeric retrotransposons in cotton. Plant J. **88**:992–1005. https://doi.org/10.1111/tpj.13309.

Han, Y., Zhang, Z., Liu, C., Liu, J., Huang, S., Jiang, J., and Jin, W. (2009). Centromere repositioning in cucurbit species: Implication of the genomic impact from centromere activation and inactivation. Proc. Natl. Acad. Sci. USA **106**:14937–14941. https://doi.org/10.1073/pnas.0904833106.

Henikoff, S., Ahmad, K., and Malik, H.S. (2001). The Centromere Paradox: Stable Inheritance with Rapidly Evolving DNA. Science **293**:1098–1102. https://doi.org/10.1126/science.1062939.

Hou, X., Wang, D., Cheng, Z., Wang, Y., and Jiao, Y. (2022). A near-complete assembly of an Arabidopsis thaliana genome. Mol. Plant **15**:1247–1250. https://doi.org/10.1016/j.molp.2022.05.014.

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**:739–748. https://doi.org/10.1038/s41588-019-0371-5.

Huang, G., Wu, Z., Percy, R.G., Bai, M., Li, Y., Frelichowski, J.E., Hu, J., Wang, K., Yu, J.Z., and 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**:516–524. https://doi.org/10.1038/s41588-020-0607-4.

Huang, X., Tian, X., Pei, L., Luo, X., Zhang, Y., Zhang, X., Zhang, X., Zhu, L., and Wang, M. (2022). Multi-omics mapping of chromatin interaction resolves the fine hierarchy of 3D genome in allotetraploid cotton. Plant Biotechnol. J. **20**:1639–1641. https://doi.org/10.1111/pbi.13877.

Hutchinson, J.B. (1951). Intra-specific differentiation in Gossypium hirsutum. Heredity **5**:161–193. https://doi.org/10.1038/hdy.1951.19.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods **12**:357–360. https://doi.org/10.1038/nmeth.3317.

Koo, D.H., Han, F., Birchler, J.A., and Jiang, J. (2011). Distinct DNA methylation patterns associated with active and inactive centromeres of the maize B chromosome. Genome Res. **21**:908–914. https://doi.org/10.1101/gr.116202.110.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. **27**:722–736.

Kovaka, S., Zimin, A.V., Pertea, G.M., Razaghi, R., Salzberg, S.L., and Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. **20**:278. https://doi.org/10.1186/s13059-019-1910-1.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics **27**:1571–1572. https://doi.org/10.1093/bioinformatics/btr167.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**:357–359. https://doi.org/10.1038/nmeth.1923.

Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R.J., Ma, Z., Shang, H., Ma, X., Wu, J., et al. (2015). Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. Nat. Biotechnol. **33**:524–530. https://doi.org/10.1038/nbt.3208.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at arXiv. https://doi.org/10.48550/arXiv.1303.3997.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **34**:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Li, K., Jiang, W., Hui, Y., Kong, M., Feng, L.-Y., Gao, L.-Z., Li, P., and Lu, S. (2021). Gapless indica rice genome reveals synergistic contributions of active transposable elements and segmental duplications to rice genome evolution. Mol. Plant **14**:1745–1756. https://doi.org/10.1016/j.molp.2021.06.017.

Li, Y., Si, Z., Wang, G., Shi, Z., Chen, J., Qi, G., Jin, S., Han, Z., Gao, W., Tian, Y., et al. (2023). Genomic insights into the genetic basis of cotton breeding in China. Mol. Plant **16**:662–677. https://doi.org/10.1016/j.molp.2023.01.012.

Liu, R., and Dickerson, J. (2017). Strawberry: Fast and accurate genome-guided transcript reconstruction and quantification from RNA-Seq. PLoS Comput. Biol. **13**, e1005851. https://doi.org/10.1371/journal.pcbi.1005851.

Liu, Y., Su, H., Zhang, J., Liu, Y., Feng, C., and Han, F. (2020). Back-spliced RNA from retrotransposon binds to centromere and regulates centromeric chromatin loops in maize. PLoS Biol. **18**, e3000582. https://doi.org/10.1371/journal.pbio.3000582.

Liu, Y., Liu, Q., Su, H., Liu, K., Xiao, X., Li, W., Sun, Q., Birchler, J.A., and Han, F. (2021). Genome-wide mapping reveals R-loops associated with centromeric repeats in maize. Genome Res. **31**:1409–1418. https://doi.org/10.1101/gr.275270.121.

Luo, S., Mach, J., Abramson, B., Ramirez, R., Schurr, R., Barone, P., Copenhaver, G., and Folkerts, O. (2012). The Cotton Centromere Contains a Ty3-gypsy-like LTR Retroelement. PLoS One **7**, e35261. https://doi.org/10.1371/journal.pone.0035261.

Ma, Z., Zhang, Y., Wu, L., Zhang, G., Sun, Z., Li, Z., Jiang, Y., Ke, H., Chen, B., Liu, Z., et al. (2021). High-quality genome assembly and resequencing of modern cotton cultivars provide resources for crop improvement. Nat. Genet. **53**:1385–1391. https://doi.org/10.1038/s41588-021-00910-2.

Maluszynska, J., and Heslop-Harrison, J.S. (1991). Localization of tandemly repeated DMA sequences in Arabidopsis thaliana. Plant J. **1**:159–166. https://doi.org/10.1111/j.1365-313X.1991.00159.x.

Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., and Zdobnov, E.M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol. Biol. Evol. **38**:4647–4654. https://doi.org/10.1093/molbev/msab199.

Mapleson, D., Venturini, L., Kaithakottil, G., and Swarbreck, D. (2018). Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. GigaScience **7**:giy131. https://doi.org/10.1093/gigascience/giy131.

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. PLoS Comput. Biol. **14**, e1005944. https://doi.org/10.1371/journal.pcbi.1005944.

Melters, D.P., Paliulis, L.V., Korf, I.F., and Chan, S.W.L. (2012). Holocentric chromosomes: convergent evolution, meiotic adaptations, and genomic analysis. Chromosome Res. **20**:579–593. https://doi.org/10.1007/s10577-012-9292-1.

Melters, D.P., Bradnam, K.R., Young, H.A., Telis, N., May, M.R., Ruby, J.G., Sebra, R., Peluso, P., Eid, J., Rank, D., et al. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. **14**:R10. https://doi.org/10.1186/gb-2013-14-1-r10.

Miga, K.H. (2015). Completing the human genome: the progress and challenge of satellite DNA assembly. Chromosome Res. **23**:421–426. https://doi.org/10.1007/s10577-015-9488-2.

Miller, J.T., Dong, F., Jackson, S.A., Song, J., and Jiang, J. (1998). Retrotransposon-Related DNA Sequences in the Centromeres of Grass Chromosomes. Genetics **150**:1615–1623. https://doi.org/10.1093/genetics/150.4.1615.

Montefalcone, G., Tempesta, S., Rocchi, M., and Archidiacono, N. (1999). Centromere repositioning. Genome Res. **9**:1184–1188. https://doi.org/10.1101/gr.9.12.1184.

Nagaki, K., Cheng, Z., Ouyang, S., Talbert, P.B., Kim, M., Jones, K.M., Henikoff, S., Buell, C.R., and Jiang, J. (2004). Sequencing of a rice centromere uncovers active genes. Nat. Genet. **36**:138–145. https://doi.org/10.1038/ng1289.

Naish, M., Alonge, M., Wlodzimierz, P., Tock, A.J., Abramson, B.W., Schmücker, A., Mandáková, T., Jamge, B., Lambing, C., Kuo, P., et al. (2021). The genetic and epigenetic landscape of the Arabidopsis centromeres. Science **374**, eabi7489. https://doi.org/10.1126/science.abi7489.

Nergadze, S.G., Piras, F.M., Gamba, R., Corbo, M., Cerutti, F., McCarter, J.G.W., Cappelletti, E., Gozzo, F., Harman, R.M., Antczak, D.F., et al. (2018). Birth, evolution, and transmission of satellite-free mammalian centromeric domains. Genome Res. **28**:789–799. https://doi.org/10.1101/gr.231159.117.

Neumann, P., Navrátilová, A., Koblížková, A., Kejnovský, E., Hřibová, E., Hobza, R., Widmer, A., Doležel, J., and Macas, J. (2011). Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mob. DNA **2**:4. https://doi.org/10.1186/1759-8753-2-4.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science **376**:44–53. https://doi.org/10.1126/science.abj6987.

Orr, H.A. (1990). "Why Polyploidy is Rarer in Animals Than in Plants" Revisited. Am. Nat. **136**:759–770. https://doi.org/10.1086/285130.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B., Elliott, T.A., Ware, D., Peterson, T., et al. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. **20**:275. https://doi.org/10.1186/s13059-019-1905-y.

Paterson, A.H., Brubaker, C.L., and Wendel, J.F. (1993). A rapid method for extraction of cotton (Gossypium spp.) genomic DNA suitable for RFLP or PCR analysis. Plant Mol. Biol. Rep. **11**:122–127. https://doi.org/10.1007/BF02670470.

Pedersen, B.S., and Quinlan, A.R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics **34**:867–868. https://doi.org/10.1093/bioinformatics/btx699.

Pei, L., Huang, X., Liu, Z., Tian, X., You, J., Li, J., Fang, D.D., Lindsey, K., Zhu, L., Zhang, X., and Wang, M. (2022). Dynamic 3D genome architecture of cotton fiber reveals subgenome-coordinated chromatin topology for 4-staged single-cell differentiation. Genome Biol. **23**:45. https://doi.org/10.1186/s13059-022-02616-y.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**:841–842. https://doi.org/10.1093/bioinformatics/btq033.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., et al. (2009). The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science **326**:1112–1115. https://doi.org/10.1126/science.1178534.

Schneider, K.L., Xie, Z., Wolfgruber, T.K., and Presting, G.G. (2016). Inbreeding drives maize centromere evolution. Proc. Natl. Acad. Sci. USA **113**:E987–E996. https://doi.org/10.1073/pnas.1522008113.

Sharma, A., and Presting, G.G. (2008). Centromeric retrotransposon lineages predate the maize/rice divergence and differ in abundance and activity. Mol. Genet. Genom. **279**:133–147. https://doi.org/10.1007/s00438-007-0302-5.

Singh, A., Singh, U., Mittal, D., and Grover, A. (2010). Transcript expression and regulatory characteristics of a rice glycosyltransferase OsGT61-1 gene. Plant Sci. **179**:114–122. https://doi.org/10.1016/j.plantsci.2010.03.005.

Smit, A., Hubley, R & Green, P. RepeatMasker Open-4.0.

Soltis, D.E., Visger, C.J., and Soltis, P.S. (2014). The polyploidy revolution then, and now: Stebbins revisited. Am. J. Bot. **101**:1057–1078. https://doi.org/10.3732/ajb.1400178.

Song, J.-M., Xie, W.-Z., Wang, S., Guo, Y.-X., Koo, D.-H., Kudrna, D., Gong, C., Huang, Y., Feng, J.-W., Zhang, W., et al. (2021). Two gap-free reference genomes and a global view of the centromere architecture in rice. Mol. Plant **14**:1757–1767. https://doi.org/10.1016/j.molp.2021.06.018.

Song, Q., Zhang, T., Stelly, D.M., and Chen, Z.J. (2017). Epigenomic and functional analyses reveal roles of epialleles in the loss of photoperiod sensitivity during domestication of allotetraploid cottons. Genome Biol. **18**:99. https://doi.org/10.1186/s13059-017-1229-8.

Su, H., Liu, Y., Liu, C., Shi, Q., Huang, Y., and Han, F. (2019). Centromere Satellite Repeats Have Undergone Rapid Changes in Polyploid Wheat Subgenomes. Plant Cell **31**:2035–2051. https://doi.org/10.1105/tpc.19.00133.

Su, H., Liu, Y., Liu, Y.-X., Lv, Z., Li, H., Xie, S., Gao, Z., Pang, J., Wang, X.-J., Lai, J., et al. (2016). Dynamic chromatin changes associated with de novo centromere formation in maize euchromatin. Plant J. **88**:854–866. https://doi.org/10.1111/tpj.13305.

Sullivan, B.A., and Karpen, G.H. (2004). Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. Nat. Struct. Mol. Biol. **11**:1076–1083. https://doi.org/10.1038/nsmb845.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. **7**:562–578. https://doi.org/10.1038/nprot.2012.016.

Ulloa, M., Brubaker, C., and Chee, P. (2007). Cotton. In Technical Crops, C. Kole (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 1–49. https://doi.org/10.1007/978-3-540-34538-1_1.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., and Earl, A.M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One **9**, e112963. https://doi.org/10.1371/journal.pone.0112963.

Wang, B., Yang, X., Jia, Y., Xu, Y., Jia, P., Dang, N., Wang, S., Xu, T., Zhao, X., Gao, S., et al. (2022). High-quality Arabidopsis thaliana Genome Assembly with Nanopore and HiFi Long. Dev. Reprod. Biol. **20**:4–13. https://doi.org/10.1016/j.gpb.2021.08.003.

Wang, M., Li, J., Wang, P., Liu, F., Liu, Z., Zhao, G., Xu, Z., Pei, L., Grover, C.E., Wendel, J.F., et al. (2021). Comparative Genome Analyses Highlight Transposon-Mediated Genome Expansion and the Evolutionary Architecture of 3D Genomic Folding in Cotton. Mol. Biol. Evol. **38**:3621–3636. https://doi.org/10.1093/molbev/msab128.

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**:224–229. https://doi.org/10.1038/s41588-018-0282-x.

Wendel, J.F. (1989). New World tetraploid cottons contain Old World cytoplasm. Proc. Natl. Acad. Sci. USA **86**:4132–4136. https://doi.org/10.1073/pnas.86.11.4132.

Wendel, J.F. (2015). The wondrous cycles of polyploidy in plants. Am. J. Bot. **102**:1753–1756. https://doi.org/10.3732/ajb.1500320.

Willard, H.F. (1989). The genomics of long tandem arrays of satellite DNA in the human genome. Genome **31**:737–744. https://doi.org/10.1139/g89-132.

Wu, J., Yamagata, H., Hayashi-Tsugane, M., Hijishita, S., Fujisawa, M., Shibata, M., Ito, Y., Nakamura, M., Sakaguchi, M., Yoshihara, R., et al. (2004). Composition and structure of the centromeric region of rice chromosome 8. Plant Cell **16**:967–976. https://doi.org/10.1105/tpc.019273.

Wu, J.C., and Manuelidis, L. (1980). Sequence definition and organization of a human repeated DNA. J. Mol. Biol. **142**:363–386. https://doi.org/10.1016/0022-2836(80)90277-6.

Wu, Y., Kikuchi, S., Yan, H., Zhang, W., Rosenbaum, H., Iniguez, A.L., and Jiang, J. (2011). Euchromatic Subdomains in Rice Centromeres Are Associated with Genes and Transcription. Plant Cell **23**:4054–4064. https://doi.org/10.1105/tpc.111.090043.

Yan, H., Talbert, P.B., Lee, H.-R., Jett, J., Henikoff, S., Chen, F., and Jiang, J. (2008). Intergenic Locations of Rice Centromeric Chromatin. PLoS Biol. **6**:e286. https://doi.org/10.1371/journal.pbio.0060286.

Yan, H., Kikuchi, S., Neumann, P., Zhang, W., Wu, Y., Chen, F., and Jiang, J. (2010). Genome-wide mapping of cytosine methylation revealed dynamic DNA methylation patterns associated with genes and centromeres in rice. Plant J. **63**:353–365. https://doi.org/10.1111/j.1365-313X.2010.04246.x.

Yang, Z., Ge, X., Yang, Z., Qin, W., Sun, G., Wang, Z., Li, Z., Liu, J., Wu, J., Wang, Y., et al. (2019). Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. Nat. Commun. **10**:2989. https://doi.org/10.1038/s41467-019-10820-x.

Yuan, D., Tang, Z., Wang, M., Gao, W., Tu, L., Jin, X., Chen, L., He, Y., Zhang, L., Zhu, L., et al. (2015). The genome sequence of Sea-Island cotton (Gossypium barbadense) provides insights into the allopolyploidization and development of superior spinnable fibres. Sci. Rep. **5**, 17662. https://doi.org/10.1038/srep17662.

Zang, C., Schones, D.E., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics **25**:1952–1958. https://doi.org/10.1093/bioinformatics/btp340.

Zhang, L., Liang, J., Chen, H., Zhang, Z., Wu, J., and Wang, X. (2023). A near-complete genome assembly of Brassica rapa provides new insights into the evolution of centromeres. Plant Biotechnol. J. **21**:1022–1032. https://doi.org/10.1111/pbi.14015.

Zhang, R.-G., Li, G.-Y., Wang, X.-L., Dainat, J., Wang, Z.-X., Ou, S., and Ma, Y. (2022). TEsorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes. Hortic. Res. **9**:uhac017. https://doi.org/10.1093/hr/uhac017.

Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., Zhang, J., Saski, C.A., Scheffler, B.E., Stelly, D.M., et al. (2015). Sequencing of allotetraploid cotton (Gossypium hirsutum L. acc. TM-1) provides a resource for fiber improvement. Nat. Biotechnol. **33**:531–537. https://doi.org/10.1038/nbt.3207.

Zhang, W., Lee, H.-R., Koo, D.-H., and Jiang, J. (2008). Epigenetic Modification of Centromeric Chromatin: Hypomethylation of DNA Sequences in the CENH3-Associated Chromatin in Arabidopsis thaliana and Maize. Plant Cell **20**:25–34. https://doi.org/10.1105/tpc.107.057083.

Zhang, W., Cao, Y., Wang, K., Zhao, T., Chen, J., Pan, M., Wang, Q., Feng, S., Guo, W., Zhou, B., and Zhang, T. (2014). Identification of centromeric regions on the linkage map of cotton using centromere-related repeats. Genomics **104**:587–593. https://doi.org/10.1016/j.ygeno.2014.09.002.

Zhao, J., Xie, Y., Kong, C., Lu, Z., Jia, H., Ma, Z., Zhang, Y., Cui, D., Ru, Z., Wang, Y., et al. (2023). Centromere repositioning and shifts in wheat evolution. Plant Commun. **4**, 100556. https://doi.org/10.1016/j.xplc.2023.100556.

Zhong, C.X., Marshall, J.B., Topp, C., Mroczek, R., Kato, A., Nagaki, K., Birchler, J.A., Jiang, J., and Dawe, R.K. (2002). Centromeric Retroelements and Satellites Interact with Maize Kinetochore Protein CENH3. Plant Cell **14**:2825–2836. https://doi.org/10.1105/tpc.006106.

Zhuang, W., Chen, H., Yang, M., Wang, J., Pandey, M.K., Zhang, C., Chang, W.-C., Zhang, L., Zhang, X., Tang, R., et al. (2019). The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. Nat. Genet. **51**:865–876. https://doi.org/10.1038/s41588-019-0402-2.