PNAS RESEARCH ARTICLE | CELL BIOLOGY

Pan-centromere reveals widespread centromere repositioning of soybean genomes

Yang Liu[a,1], Congyang Yi[a,b,1], Chaolan Fan[a,1], Qian Liu[a,b,1], Shulin Liu[a,1], Lisha Shen[a], Kaibiao Zhang[a,b], Yuhong Huang[a,b], Chang Liu[a,b], Yingxiang Wang[c], Zhixi Tian[a,2], and Fangpu Han[a,2]

Edited by David Baulcombe, University of Cambridge, Cambridge, United Kingdom; received June 16, 2023; accepted September 6, 2023

Centromere repositioning refers to a de novo centromere formation at another chromosomal position without sequence rearrangement. This phenomenon was frequently encountered in both mammalian and plant species and has been implicated in genome evolution and speciation. To understand the dynamic of centromeres on soybean genome, we performed the pan-centromere analysis using CENH3-ChIP-seq data from 27 soybean accessions, including 3 wild soybeans, 9 landraces, and 15 cultivars. Building upon the previous discovery of three centromere satellites in soybean, we have identified two additional centromere satellites that specifically associate with chromosome 1. These satellites reveal significant rearrangements in the centromere structures of chromosome 1 across different accessions, consequently impacting the localization of CENH3. By comparative analysis, we reported a high frequency of centromere repositioning on 14 out of 20 chromosomes. Most newly emerging centromeres formed in close proximity to the native centromeres and some newly emerging centromeres were apparently shared in distantly related accessions, suggesting their emergence is independent. Furthermore, we crossed two accessions with mismatched centromeres to investigate how centromere positions would be influenced in hybrid genetic backgrounds. We found that a significant proportion of centromeres in the S9 generation undergo changes in size and position compared to their parental counterparts. Centromeres preferred to locate at satellites to maintain a stable state, highlighting a significant role of centromere satellites in centromere organization. Taken together, these results revealed extensive centromere repositioning in soybean genome and highlighted how important centromere satellites are in constraining centromere positions and supporting centromere function.

centromere repositioning | soybean | centromere satellites | CENH3

The centromere is the chromosomal structure that ensures proper segregation of chromosomes during mitosis and meiosis (1). This is achieved by providing the assembly site for the kinetochore that connects centromeric chromatin to the spindle microtubules (2). Centromeres have distinct characteristics that make them unique chromosome loci. In most eukaryotes, they are defined epigenetically by the histone H3 variant, CENH3, which is both necessary and sufficient for centromere function (3). In addition, centromeres exhibit characteristic chromatin modification signatures and genetically encoded structural features, like non-B-form conformations, which may act to promote the faithful formation of a centromere on each chromosome (4–8).

In many species, centromere DNA consists of megabase-sized arrays of one or more satellite repeat units, but these sequences differ widely in sequence composition and length among species (9–11). For example, in tribe *Fabeae*, centromere repeat sizes range from 33 bp in *Vicia tetrasperma* to 2,979 bp in *Pisum fulvum*, and most centromeric satellite families were species-specific (12). Furthermore, in the *Equus asinus*, an extraordinarily high number of centromeres were found completely devoid of satellite DNA (13). The dramatic size and sequence variability of centromeres, combined with their essential and conserved feature in chromosome segregation, present a centromere paradox.

Centromere repositioning refers to a de novo centromere formation at another chromosomal position without sequence rearrangement (14). This phenomenon was used to explain the emergence of evolutionary-new centromeres (ENCs) (15). ENCs can appear during evolution at an ectopic location in a chromosomal region. It is hypothesized that the process is usually accompanied by the inactivation of the old centromere, whereas the newly formed centromere does not contain satellite repeats initially, but matured gradually through acquisition of repeats. The first case of ENC was discovered in non-human primate species (16). Since then, the ENCs were relatively frequently reported in fungi, mammals, and plants (15, 17–25). However, the mechanism underlying their occurrence has not been completely elucidated.

Significance

Centromeres are crucial for ensuring accurate chromosome segregation in eukaryotic organisms. We identified centromere satellites associated with chromosome 1 of soybean, revealing significant rearrangements and impacting the localization of CENH3. Comparative analysis shows frequent centromere repositioning across chromosomes, with new centromeres forming near native ones or independently emerging. Hybrid crosses demonstrate significant changes in centromere size and position, highlighting their dynamic nature. Crucially, our findings emphasize the role of centromere satellites in maintaining stable positions, underscoring their importance in centromere organization.

Author contributions: Y.L., Z.T., and F.H. designed research; Y.L., C.Y., C.F., Q.L., S.L., L.S., K.Z., Y.H., C.L., Y.W., and Z.T. performed research; Y.L. and C.Y. analyzed data; and Y.L. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] Y.L., C.Y., C.F., Q.L., and S.L. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: zxtian@genetics.ac.cn or fphan@genetics.ac.cn.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2310177120/-/DCSupplemental.

Published October 10, 2023.

Soybean is the most widely consumed legume source of protein for human nutrition. Cultivated soybean (*Glycine. max* [L.] Merr.) was domesticated in China 5,000 y ago from wild soybean (*G. soja* Sieb. & Zucc.) in a process that dramatically changed its morphology and organ size (26, 27). After domestication, ancient farmers developed many phenotypically diverse landraces through both artificial and natural selection. Subsequently, the success of genetic improvement led to the replacement of local landraces with cultivars that meet societal food demands related to population increase (28). Comprehensive resequencing analyses of wild soybeans, landraces, and cultivars have identified molecular footprints of domestication and genetic improvement (29–31). Soybean provides an excellent system to study how genomes were shaped through a complex interaction of demography and selection. At present, several studies reported a number of soybean genome information. The first referenced soybean genome, Williams 82, facilitated the advance in soybean functional genomics (32). Subsequently, the sequencing of the wild soybean (*G. soja* W05) genome explored the diversity of lost genes in cultivated soybean and promoted comparative genomic and evolutionary studies (33). Furthermore, the recent construction of the soybean pan-genome provided high-quality reference genome information, which identified numerous genetic variations that are responsible for important traits (34). Zhonghuang 13 (ZH13) is the most widely planted soybean cultivar in China. Compared with the pan-genome of soybean, the genome assembly of ZH13 is also high-quality with higher completeness and accuracy than the reference genome (W82) (35). The released pan-genome of soybean, as well as the high-quality reference of ZH13, have offered valuable genomic resources to enable surveys of centromere diversity among populations. Previous studies indicated the centromere regions of soybean chromosomes contain three kinds of tandem repetitive sequences, CentGm-1, CentGm-2, and CentGm-4 (36, 37). Centromere-specific retrotransposons have also been identified in soybeans (36). However, the scope of centromere structural and sequence diversity within and between populations has not been investigated in detail.

Here, we carried out a comparative genomics to explore soybean centromere diversity and dynamics in terms of both genetics (sequence composition) and epigenetics (CENH3 localization). We performed CENH3-ChIP-seq of 27 soybean accessions, including wild soybeans, landraces, and cultivars, and analyzed the variation in the size and sequence composition of centromeres. We also characterized the centromere positions by mapping the CENH3-ChIP-seq reads from each accession to the ZH13 genome. We asked how frequently ENCs formed in the genome evolution of soybean.

Results

Genome-Wide Mapping of Functional Centromeres in 27 Soybean Accessions. To accurately position soybean centromeres, we first cloned soybean *CENH3* (GmCENH3) and raised specific antibody generated from a peptide corresponding to the N-terminal region of the GmCENH3 sequence. The antibody can localize to all functional soybean centromeres (*SI Appendix*, Fig. S1). We performed chromatin immunoprecipitation sequencing (ChIP-seq) with anti-CENH3 antibody in a set of 27 accessions including 3 wild soybeans, 9 landraces, and 15 cultivars selected to represent the extensive phylogenetic relationships and geographic distributions (34). The immunoprecipitated DNA fragments were sequenced on the NovaSeq platform along with control DNA samples extracted from chromatin preparations prior to ChIP (input control). Fluorescence in situ hybridization (FISH) using the ChIPed DNA as probe showed bright signals in the centromeres, supporting the actual accumulation of centromeric DNA in our ChIPed DNA (*SI Appendix*, Fig. S2). In total, we obtained an average of 11 million reads (on 3× coverage) (*SI Appendix*, Table S1). After removing short reads with poor quality and PCR duplicates, the filtered reads were mapped to their respective reference genomes (*SI Appendix*, Fig. S3). Significant enrichment of CENH3 peaks was observed within the centromere regions across all 20 soybean chromosomes. The mapping of these peaks in the ZH13 accession is illustrated in Fig. 1 and *SI Appendix*, Fig. S3 provides the corresponding mapping for other accessions. The size of the core region of CENH3 binding in ZH13 centromeres varies from 0.5 to 2.7 Mb, with an average centromere size of ~1.3 Mb (Table 1), which was significantly smaller than that of the soybean centromeric satellite arrays (CentGm), which ranged from ~1 to 4.5 Mb (Fig. 1 and Table 1). These findings are consistent with the recent analyses in *Arabidopsis*, where AthCEN178 array sizes (~1.5 to 6.5 Mb) were also found to be larger than the regions of CENH3 enrichment (~1 to 2 Mb) (38, 39). This suggests that the size range of CENH3-enriched domains relative to the larger satellite arrays may represent a common feature in centromere organization among diverse plant species. Furthermore, we observed that several chromosomes exhibited multiple CENH3-enriched subdomains (Fig. 1, e.g., chromosome 5 and chromosome 14), potentially arising from assembly gaps in the centromeric sequences of ZH13 (Dataset S1).

The Structure and Epigenetics of Satellite Arrays Associated with CENH3 Nucleosomes. Previous studies indicated that the centromeres of cultivated soybean are composed of three distinct satellite repeats (CentGm-1, CentGm-2, and CentGm-4) (36, 37). However, it is still unclear whether other lines or wild soybeans have unique centromere satellites. To comprehensively identify centromere-specific repeats within the soybean genome, we initiated the process by employing RepeatExplorer to analyze and characterize repeat clusters. Among these, a total of 26 clusters exhibited a CENH3-ChIP/input ratio exceeding 2 (*SI Appendix*, Fig. S4*A*). Using Basic Local Alignment Search Tool (BLAST), we found seven clusters were specifically mapped to the centromeres (*SI Appendix*, Fig. S4*B*), among which five clusters were identified as tandemly arranged satellite repeats with a unit size of 91, 92, 273, 413, and 444 bp, respectively. We named the tandem repeats CentGm91, CentGm92, CentGm273, CentGm413, and CentGm444 according to their length. The alignment results indicated that the CentGm91, CentGm92, and CentGm413 were the previously identified CentGm-1, CentGm-2, and CentGm-4, respectively (*SI Appendix*, Fig. S5), while the CentGm273 and CentGm444 have not yet been described. Two additional clusters, CL26 and CL33, represented soybean centromeric-specific long terminal repeat retrotransposons (LTR-RTs). Through a comprehensive search in the SoyTEdb database (40), we confirmed CL26 as the known soybean centromere-enriched retrotransposon 12 (Gmr12), and CL33 aligns with Gmr17 in soybean (41). Both of these retrotransposons are categorized within the *Gypsy*-like families. (*SI Appendix*, Fig. S4*C*). We also employed the Tandem Repeat Annotation and Structural Hierarchy (TRASH) tool in its de novo mode to annotate tandem repeats (42). This strategy enabled us to thoroughly identify CentGm repeat arrays located at the anticipated centromere positions (*SI Appendix*, Fig. S4*E*). TRASH not only confirmed the five distinct satellite populations identified by RepeatExplorer but also revealed new variants of CentGm413. Different centromeres were predominantly characterized by a single repeat class, with CentGm91 being predominant in Cen1, Cen3, Cen4, Cen6, Cen7, Cen10, Cen11, and Cen17. Conversely, CentGm92 dominated in Cen2, Cen5, Cen8, Cen9, Cen12,

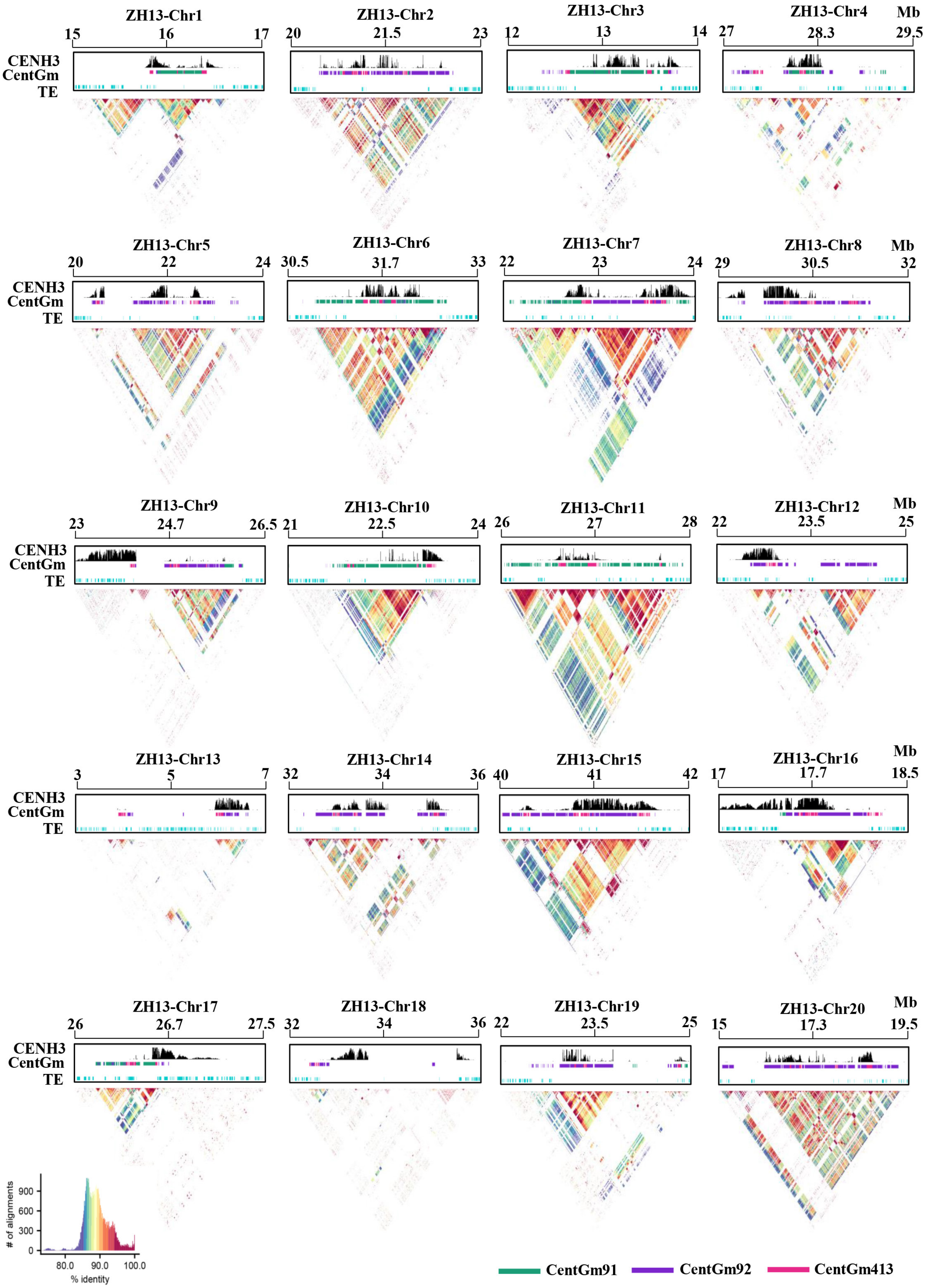

Fig. 1. Genome-wide mapping of CENH3-ChIP-Seq in ZH13. Characteristics of the 20 centromeres in ZH13 accession. The first track illustrates CENH3 enrichment [$\log_2$(ChIP/Input)] across soybean centromeres. The second track depicts the distribution of soybean centromeric satellites. The third track showcases the distribution of transposable elements (TEs). Heatmaps display the pairwise sequence identity of soybean centromeric satellites among all non-overlapping 2 kb regions on the centromere.

Table 1. Positions of functional CENH3-Binding regions in ZH13

Chr	Chr size (Mb)	Cen location (Mb)	Cen size (Mb)	CentGm arrays (Mb)
1	59	15.7 to 16.6	0.9	1.1
2	52	20.5 to 22.5	2.0	3.1
3	47	12.5 to 14.0	1.5	2.1
4	53	27.7 to 28.4	0.7	2.2
5	45	20.3 to 23.0	2.7	4.1
6	51	31.3 to 32.3	1.0	1.5
7	46	22.6 to 23.9	1.3	2.3
8	49	29.1 to 30.5	1.4	3.2
9	51	23.0 to 24.2	1.2	3.3
10	54	22.5 to 23.5	1.0	2.1
11	40	26.5 to 27.1	0.6	2.2
12	44	22.3 to 23.0	0.7	3.3
13	47	5.9 to 6.7	0.8	3.4
14	53	32.9 to 35.2	2.3	2.5
15	53	40.2 to 41.7	1.5	2.0
16	38	17.0 to 18.0	1.0	1.5
17	42	26.5 to 27.0	0.5	1.0
18	60	32.8 to 33.7	0.9	4.0
19	52	23.0 to 24.0	1.0	2.5
20	52	16.0 to 18.7	2.7	4.5

Cen, centromere. Chr, chromosome. CentGm, soybean centromeric satellites.

Cen14, Cen15, Cen16, Cen17, Cen18, Cen19, and Cen20. While CentGm413 and its variants were present in relatively fewer numbers, they exhibited a relatively uniform distribution across all chromosomes. Notably, CentGm273 and CentGm444 exhibited specific localization at Cen1 (*SI Appendix*, Fig. S4*E*). Unlike retrotransposon-rich centromere in many plants, we found that tandem repeat content is very high in soybean, suggesting satellite repeats are the major component of soybean centromeres (*SI Appendix*, Fig. S4*D*).

To ascertain the centromeric distribution, we designed specific oligo probes for each of the five computationally identified tandem repeats. FISH analyses showed that CentGm91 and CentGm92 marked distinct subsets of ZH13 centromeres, and CentGm413 were present at all functional centromeres of ZH13 (Fig. 2 *A* and *B*). In contrast, the FISH signals for CentGm273 and CentGm444 repeats were primarily localized in the centromeric regions of a pair of chromosomes (Fig. 2*C*). However, co-localization signals were not observed between CentGm273 and CentGm444, indicating that they are distributed at different locations within the Cen1 of ZH13 (Fig. 2 *C* and *D*), which is consistent with the genomic annotation data (Fig. 2*E*). Taken together, our results showed that the centromeres of soybean contain five types of satellites differed in their nucleotide sequence, monomer length (91 to 444 bp), and abundance, and also highlights the presence of two types of centromeric retrotransposons.

To investigate the epigenetic features of the ZH13 centromeres, we analyzed DNA methylation patterns in CG, CHG, and CHH contexts, along with a comparison of repeat structure and CENH3 location. Specifically focusing on the relatively abundant CentGm91, CentGm92, and CentGm413 satellites in soybean centromeres, we examined their binding patterns with CENH3 nucleosomes. Our findings revealed a highly phased pattern of CENH3 nucleosomes with CentGm91 and CentGm92, while CentGm413 exhibited a less pronounced pattern (*SI Appendix*, Fig. S6 *A–C*). Notably, for CENH3 nucleosomes associated with CentGm92, we observed increased methylation levels in CG, CHG, and CHH contexts in spacer regions at satellite edges (*SI Appendix*, Fig. S6*B*), resembling the chromatin state of CEN180 in *Arabidopsis* (38). Conversely, for CENH3 nucleosomes bound to CentGm91, increased methylation levels were observed in CG and CHG contexts, while CHH methylation levels were reduced (*SI Appendix*, Fig. S6*A*). These outcomes emphasize the varied epigenetic and chromatin attributes associated with distinct satellite sequences, along with the organization of CENH3 nucleosomes (*SI Appendix*, Fig. S6).

Furthermore, we assessed centromeric LTR-RTs with respect to their enrichment in CENH3 ChIP-seq data. We observed a noticeable reduction in CENH3 enrichment relative to the surrounding CentGm arrays within the centromeres. However, the levels of CENH3 enrichment remained higher than those observed in LTR-RTs located outside of the centromeric regions (*SI Appendix*, Fig. S6*D*).

Structural Features of the Cen1 in 27 Accessions. The distinct positioning of CentGm273 and CentGm444 on Cen1 has inspired us to further explore the distribution patterns of these five sequences, namely CentGm91, CentGm92, CentGm413, CentGm444, and CentGm273, across the Cen1 of 27 soybean accessions. By aligning these sequences to their respective reference genomes, we have observed consistent arrangement patterns and proportional distribution of these five sequences among the 27 accessions (Fig. 3 *A* and *C*). This stability in arrangement and proportional composition observed within the centromeres aligns with similar stability observed within the chromosome arms (*SI Appendix*, Fig. S7). Physical mapping showed CentGm273 mapped to the edge of the core Cen1, especially abundant in the short arm of chromosome 1, while CentGm444, together with CentGm91, CentGm92, and CentGm413 made up the core Cen1 in most of the accessions. This observation aligns with cytogenetic findings indicating the absence of co-localization between CentGm273 and CentGm444. In the majority of the accessions, the five sequences, namely CentGm273, CentGm444, CentGm91, CentGm92, and CentGm413, exhibit the aforementioned distribution pattern. However, in two other accessions, C1 and L7, the relative positions of CentGm273 and the other four sequences have undergone changes, most notably with CentGm273 shifting to the right of CentGm91 (Fig. 3*A*). Furthermore, the results of collinearity analysis demonstrate a large inversion in the centromeric regions of C1 and L7 when compared to the reference accession ZH13 (Fig. 3*B*), supporting the interchange of centromere satellite sequence positions in C1 and L7. To further investigate the impact of the centromeric inversion on CENH3 localization, we mapped CENH3-ChIP onto their respective reference genomes. Intriguingly, our analysis revealed distinct patterns in different accessions. In the case of C1 and L7, CENH3 predominantly bound to CentGm273, whereas in the remaining accessions, it showed a preference for the other four sequences (Fig. 3*A*). Notably, in the absence of the centromeric inversion observed in C11, the CENH3 nucleosomes extended towards the CentGm273 (Fig. 3*A*). These findings unequivocally demonstrate that the rearrangement of the centromeric region

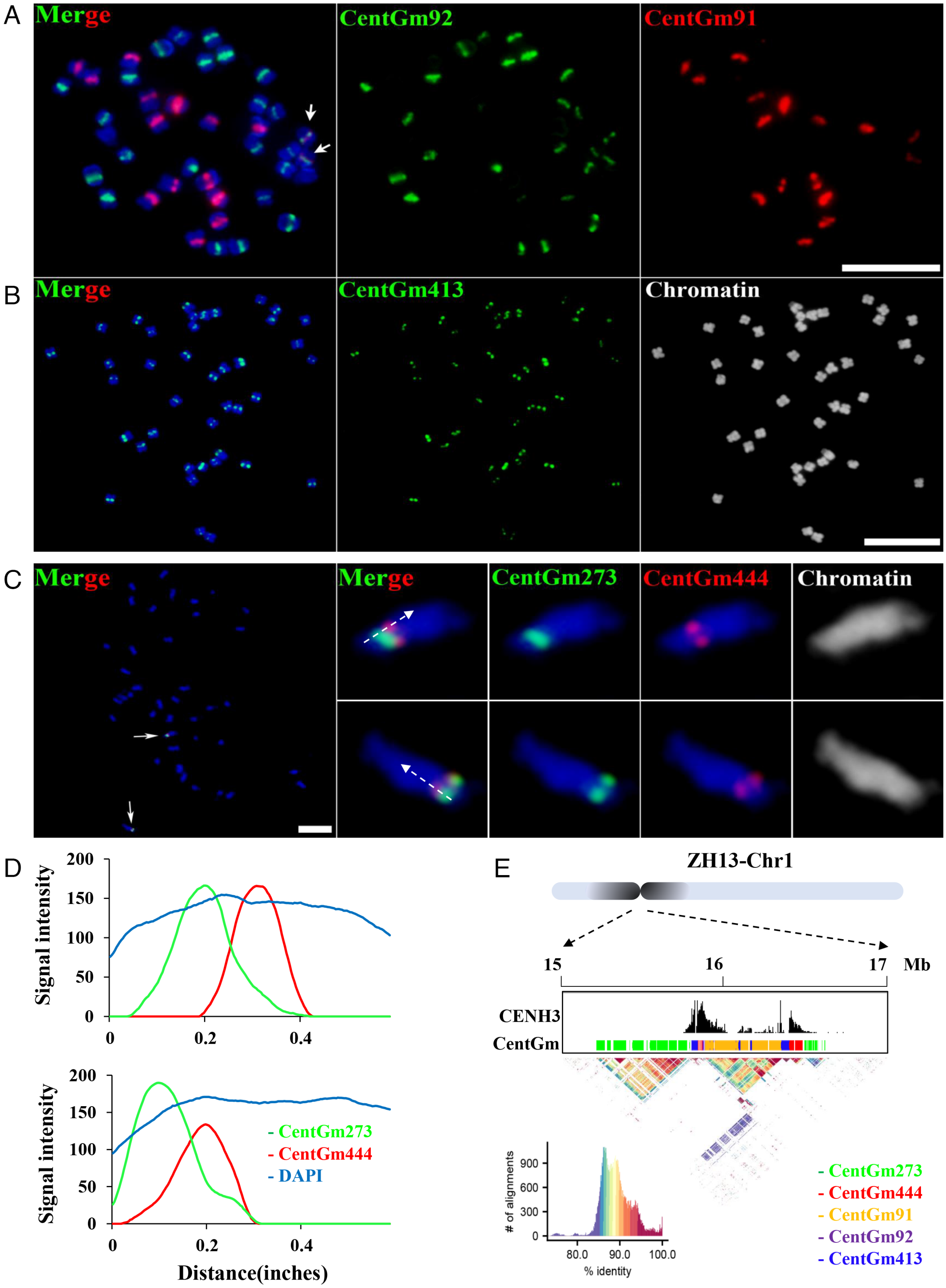

Fig. 2. Cytological analysis of the DNAs that are associated with the CENH3 nucleosomes in ZH13. (*A*) FISH signals of CentGm91 (red) and CentGm92 (green) in ZH13. (Bar, 10 µm.) (*B*) FISH signals of CentGm413 (green) in ZH13. (Bar, 10 µm.) (*C*) FISH signals of CentGm444 (red) and CentGm273 (green) in ZH13. The boxed *Insets* show high-magnification images of chromosome 1. The dashed line indicates the section of chromatin used for quantifying signal intensity. (Bar, 10 µm.) (*D*) Fluorescent profiles indicate that CentGm273 and CentGm444 do not co-localize. The horizontal axis represents the starting position of the dashed line in Fig. 2*C* (unit: inches), and the y-axis corresponds to signal intensity. (*E*) Annotation of CentGm273 and CentGm444 at distinct positions within the Cen1 of the ZH13 reference genome. The different layers demonstrate the CENH3 enrichment [$\log_2$(ChIP/Input)], the CenGm annotations, and pairwise satellite sequence similarity, respectively.

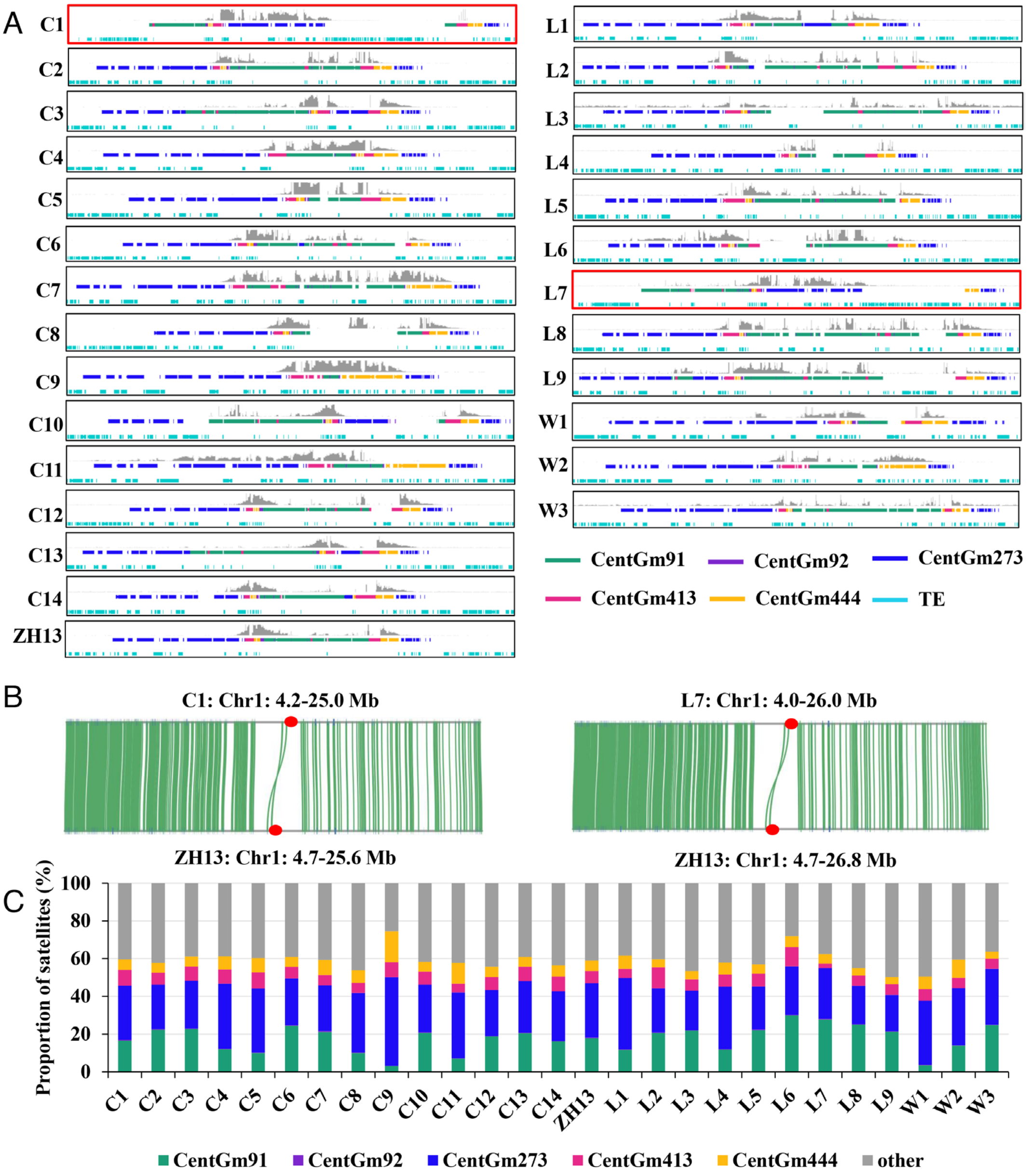

Fig. 3. Distribution characteristics of centromere satellites, TE, and CENH3-ChIP on chromosome 1 in soybean. (*A*) Structural features of the Cen1 in 27 accessions. The 27 outlines represent the structural variations of Cen1 in these 27 accessions. In each outline, the first track illustrates the enrichment of CENH3 [$\log_2$(ChIP/Input)] across Cen1. The second track depicts the distribution of soybean centromeric satellites, while the third track showcases the distribution of TE. The red outline corresponds to two accessions, C1 and L7, where the relative positions of CentGm273 and the other four sequences have undergone alterations, influencing the binding pattern and localization of CENH3. (*B*) Comparative analysis of syntenic sequences at the centromeric and pericentromeric regions between ZH13 and other accessions. Approximate 20 Mb region around the functional centromeric domains are compared. The red circles indicate centromeres. (*C*) The compositions of five centromere satellites (CentGm91, CentGm92, CentGm273, CentGm413, and CentGm444) were assessed across the Cen1 of 27 soybean accessions.

directly affects the binding pattern and localization of CENH3. This, in turn, indicates a potential disruption in the centromere structure and function, which could have broader implications for the stability and proper functioning of the affected chromosomes.

Centromere Repositioning Occurred Frequently in the Soybean Genome. In order to investigate the dynamic changes in centromeres for all 27 accessions across all 20 chromosomes, we utilized CENH3 ChIP-seq reads from the 27 soybean accessions and mapped them to the ZH13 reference genome. This allowed us to determine the positions of centromeres in the ZH13 reference genome for all 27 accessions. Different CENH3 ChIP-seq signatures were observed for all chromosomes (*SI Appendix*, Fig. S8). For example, the functional centromeres in all accessions were located at the region of 24 to 35 Mb in chromosome 8, but the location of the core CENH3 region varies among the lines (Fig. 4*A*). The three wild soybeans showed CENH3 enrichment at Cen8M (Middle; ZH13 RefGen_V2 coordinates 8:29.9 to 31.3 Mb), a region of ancestral CENH3 location defined by large amounts of CentGm92 and CentGm413 and small amounts of CentGm91 in the ZH13 reference genome (Fig. 4*E*). Another 10 accessions showed the enrichment of CENH3-ChIP-seq reads at the same position (Fig. 4*E*). In contrast, the functional centromeres of all other accessions were located either upstream or downstream of Cen8M. CEN8M1 (Coordinates 8:29.1 to 30.6 Mb), that partially overlapped left with Cen8M, was used by 10 accessions

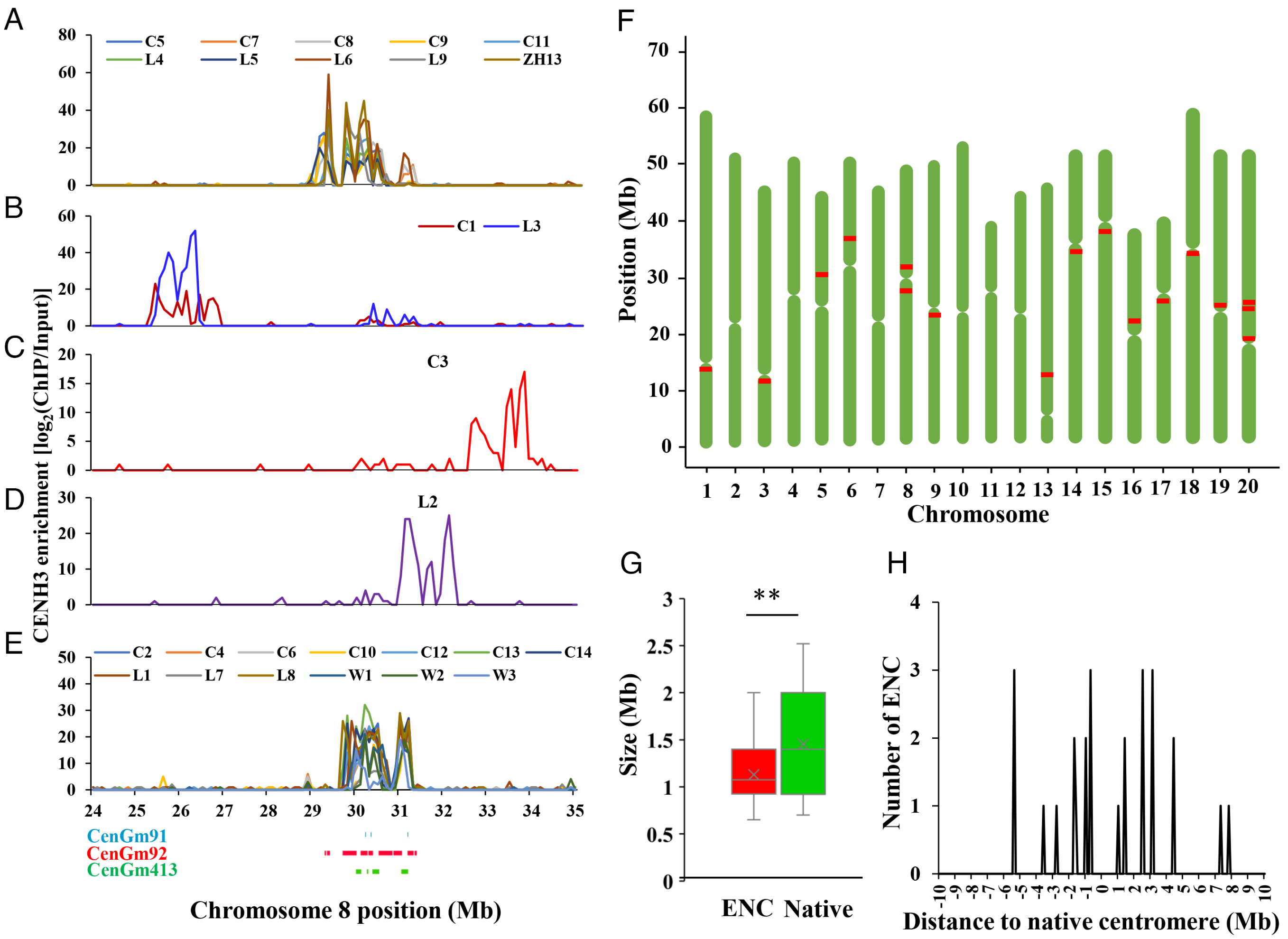

Fig. 4. Centromere repositioning occurred frequently in the evolution of soybean genome. (*A*–*E*) CENH3 ChIP-seq data from 27 accessions were consistently mapped to the ZH13 reference genome. Five distinct types of CENH3 ChIP-seq signatures were identified. The enrichment of CENH3 [$\log_2$(ChIP/Input)] was graphed within 100-kb windows along chromosome 8 of the ZH13 reference genome, with the *X* axis indicating the specific position on chromosome 8. (*F*) Locations of ENCs across chromosomes (red rectangle). (*G*) Box-plots showed the ENCs were significantly smaller in size than native centromeres. (*t* test; $^{**}P < 0.01$). (*H*) The distance of ENCs to native centromeres. The negative value on the *X* axis indicates the distance of ENC located to the short arm of a chromosome to the native centromere. The positive value on the *X* axis indicates the distance of ENC located to the long arm of a chromosome to the native centromere.

to load CENH3 (Fig. 4*A*), and CEN8M2 (Coordinates 8:31.1 to 32.3 Mb), that partially overlapped right with Cen8M, was used by L2 to load CENH3 (Fig. 4*D*). In the C1 and L3 accessions, the centromeres moved along the longer upper arm 5 Mb away from the ancestral position (Fig. 4*B*), whereas the C3 relocated the centromere along short arm 2 Mb away from the ancestral position (Fig. 4*C*). We also aligned CENH3 ChIP-seq data from C1, L3, and C3 to their respective genomes. Similarly, we observed that CENH3 localization did not coincide with CentGm on chromosome 8 across these three accessions (*SI Appendix*, Fig. S9), in line with the aforementioned observations.

These positional shifts of CENH3 represent the phenomenon of "centromere repositioning," that is, de novo centromere formation in a different position on the same chromosome. The centromere positions occupied by CentGm91, CentGm92, and CentGm413 were considered as the ancestral ones and CENH3 locations from other lines separated from, or abutted the ancestral CENH3 location represented the ENC. In total, we identified 55 ENCs in 14 of the 20 chromosomes (*SI Appendix*, Table S2). These ENCs range in size from 1.5 to 2.3 Mb, which are smaller compared to progenitor centromeres (Fig. 4*G*). We found that the distribution of the ENCs across the genome is nonrandom. They tend to form in close proximity to the progenitor centromere (Fig. 4 *F* and *H*), which was consistent with previous findings in *Candida albicans* and maize (43, 44). We also identified several chromosomes with recurrent ENCs in these accessions. For example, on chromosome 9, we detected 18 ENCs with nearly identical locations that abutted the progenitor centromere. On chromosome 16, we identified 5 ENCs with three different types of locations (*SI Appendix*, Table S2 and Fig. S8). The recurrence of ENCs suggested that certain genomic regions are favorable for participating in the formation of centromeres. We performed collinearity analysis of wild soybean and landraces, and cultivars with ZH13. We found the gene orders in centromeric and pericentromeric regions are highly conserved with good collinearities. No large structural rearrangements were observed except for the centromeric region of chromosome 19 in the C1 (*SI Appendix*, Fig. S10), highlighting that centromeres have not undergone major rearrangement during soybean domestication and improvement and the shifting of the centromere positions likely resulted from the centromere repositioning events.

In order to gain deeper insights into the sequence characteristics underlying the ENCs, we collected the 10 instances situated at a considerable distance from the native centromere. For each of these instances, we extracted the corresponding sequences and subjected them to annotation using the TRASH and External

Tandem Repeat Annotation (EDTA) tools. Our results reveal that the GC content in these 10 ENCs remains relatively stable and comparable to the native centromere, accounting for approximately 40%. Repetitive DNA constitutes approximately 76.8% of each ENC loci, with a range spanning from 57.8 to 90.4%. Notably, repetitive DNA constitutes a significant portion of each ENC locus, ranging from 57.8 to 90.4%, with an average of approximately 76.8%. Among the various repetitive sequences, LTR-RTs emerge as the prevailing element. However, this composition significantly differs from that of the native centromere, where satellites are the predominant component, accounting for nearly 90% (*SI Appendix*, Fig. S9*E*). This underscores the distinctive sequence characteristics and repetitive DNA content of these ENCs compared to the native centromeres.

Centromere Positions Were Not Rigidly Fixed after Subsequent Generations in Hybrid Genetic Backgrounds. Next, we investigated how centromere positions would be influenced in genetic crosses involving accessions with mismatched centromeres. Specifically, we examined the scenarios where one parent carried a repositioned centromere while the other parent retained the original centromere. Previous studies have shown that in F1 hybrids, both centromeres tend to maintain their positions (45). However, it remained a mystery as to how centromere positions could potentially change or remain stable after subsequent generations in hybrid genetic backgrounds. In comparison to the ZH13 accession, the C8 accession exhibited changes in centromere positions in chromosome 13 (with a distance of 1.8 Mb) and chromosome 15 (with a distance of 3 Mb). Similarly, the L4 accession also displayed centromere position changes in chromosome 13 (with a distance of 1.7 Mb) and chromosome 16 (with a distance of 5 Mb) (*SI Appendix*, Fig. S8). Additionally, besides these specific centromeres, we observed differences in the CENH3-ChIP patterns between the ZH13 and C8 accessions, as well as between ZH13 and L4 accessions in several other centromeres (*SI Appendix*, Fig. S8). We conducted individual crosses between C8 and ZH13, as well as L4 and ZH13, followed by self-pollination for nine generations. This breeding process yielded two distinct lines, namely ZH13-C8 and ZH13-L4 (*SI Appendix*, Fig. S11*A*). Subsequently, CENH3-ChIP-seq was conducted on three individuals from each line, and the resulting data were aligned to the ZH13 reference genome. This allowed us to analyze and compare the read distribution patterns and centromere positions between the offspring individuals and their respective parental accessions.

To account for the potential impact of divergent parent centromeres on offspring centromeres, we categorized the centromere alterations into two types based on the CENH3-ChIP pattern between the parental accessions. The first type, referred to as parents-same (PS), indicated the presence of identical patterns of positioning and spacing of the CENH3 nucleosomes observed between the parents. The second type, known as parents-difference (PD), denoted distinct patterns of positioning, and spacing of the CENH3 nucleosomes observed between the parents. Among the 30 identified PS-type centromeres, all these centromeres maintained their state in the ZH13-C8-S9 lines (*SI Appendix*, Table S3 and Fig. 5*A*, with Cen5 as an example). In contrast, among the 30 identical centromeres between ZH13 and L4, five centromeres (Cen8 and Cen9) displayed distinct CENH3-ChIP-seq patterns in the ZH13-L4-S9 lines compared to both ZH13 and L4 (*SI Appendix*, Table S3 and Fig. 5*A*, with Cen9 as an example). Specifically, the CENH3 read coverages in the first and third individuals of the ZH13-L4-S9 lines exhibited higher values on the right side of Cen8 and lower values on the left side compared to their respective parents (*SI Appendix*, Fig. S11*B*). Additionally, in the case of Cen9, all three individuals showed a shift in centromere position, where CENH3 displayed a propensity to bind to centromere tandem repeats on the right side and partially lost its ability to bind non-tandem repeats on the left side (Fig. 5*A*). These findings indicate that minor variations in the CENH3 binding pattern were detected between the S9 individuals and their parental centromeres in the PS type.

For the PD type, we conducted an analysis of CENH3-ChIP-seq and SNP frequencies to determine the origin of centromeres in the ZH13-C8-S9 and ZH13-L04-S9 lines. For instance, in the case of Cen18 in ZH13-C8-S9, individual 1 exhibited a CENH3 pattern resembling that of ZH13, while individuals 2 and 3 showed a CENH3 pattern similar to that of C8, which correlated with their divergence in SNP frequency from ZH13 (Fig. 5*B*). Regarding Cen13, which displayed significant centromere positional differences between ZH13 and C8, we found that all three individuals inherited the centromeres from ZH13, and the same was observed for Cen13 in the ZH13-L4 lines (*SI Appendix*, Fig. S11*B*). In addition to these stable centromeres inherited from parents, we found that a number of CENH3 profiles in S9 were not identical with parents (*SI Appendix*, Table S3). For example, in the case of Cen16, we observed substantial divergence in the CENH3 profiles of all three individuals from both ZH13-C8-S9 and ZH13-L4-S9, compared to their respective parents (Fig. 5*C*). By analyzing SNP frequencies, we determined that individual 1 and individual 2 in the ZH13-C8-S9 line inherited Cen16 from ZH13, while individual 3 in the ZH13-C8-S9 line had Cen16 contributed by C8 (*SI Appendix*, Fig. S11*C*). In contrast, all three individuals from the ZH13-L4-S9 line inherited Cen16 from L4 (*SI Appendix*, Fig. S11*D*). In total, we identified 11 (36.7%) centromeres from ZH13-C8-S9 and 16 (53.3%) centromeres from ZH13-L4-S9 that showed distinct CENH3 profiles compared to their parents (*SI Appendix*, Fig. S12 and Table S3). These differences in CENH3 profiles included variations in CENH3 sequence abundance, increases or decreases, and changes in CENH3 binding positions (*SI Appendix*, Fig. S12). Particularly, in cases where CENH3 binding position changes occurred in the S9 generation, the parental centromeres often did not bind to centromere satellites, while in S9, CENH3 tended to undergo positional shifts and associate with centromere satellites (*SI Appendix*, Fig. S12, Chr9 in ZH13-L4 and Chr16 in ZH13-L4). These findings suggest that the inheritance of centromere positions in subsequent generations is influenced by the divergence of parent centromeres. When the centromere positions and spacing patterns between parents are consistent (PS type), the majority of centromeres in the S9 generation remain unchanged. However, when there are differences in CENH3 profiles between parents (PD type), a significant proportion of centromeres in the S9 generation undergo changes in size and position compared to their parental counterparts. Additionally, the role of centromere satellites in centromere stability is highlighted by these observations.

Discussion

A typical centromere is composed of tandem DNAs and retrotransposons (46). However, the relative amounts of satellites and retrotransposon vary greatly between species. For example, wheat and oat centromeres are primarily composed of retrotransposons (47), while *Arabidopsis* and rice centromeres mainly contain satellites (48–50). Notably, thanks to the completion of the *Arabidopsis* Telomere-to-Telomere (T2T) genome, two recent outstanding studies have revealed that the centromeres in this model plant species are invaded by *ATHILA* retrotransposons. These retrotransposons have been found to disrupt the genetic and epigenetic organization within the centromeres (38, 39). CENH3 are highly adaptable to

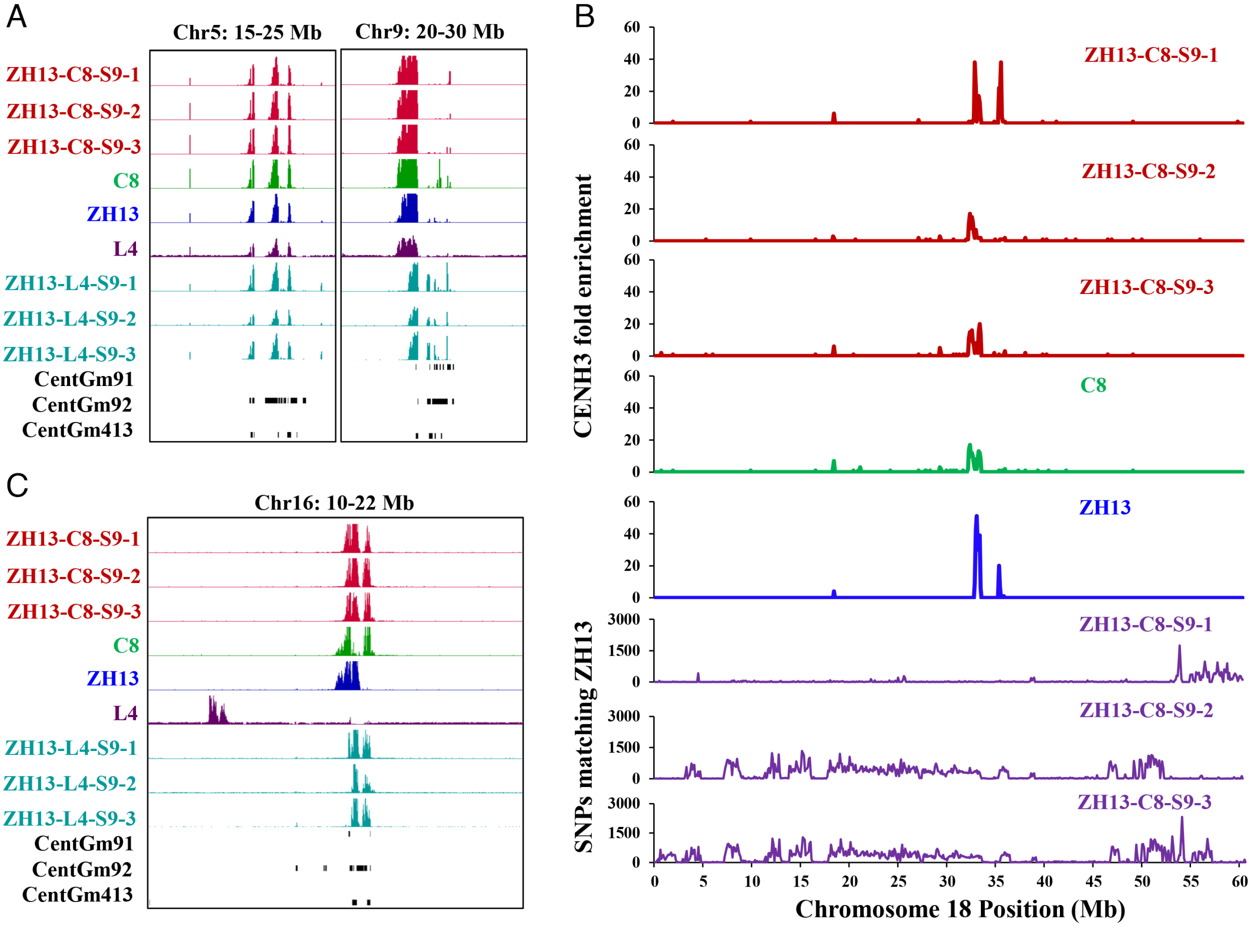

Fig. 5. Centromere positions were not rigidly fixed after subsequent generations in hybrid genetic backgrounds. (*A*) The CENH3 profiles in ZH13, C8, L4, ZH13-C8-S9, and ZH13-L4-S9 using the Integrative Genomic Viewer tool. The left picture shows the 10-Mb window of Chr5 and the right picture shows the 10 Mb window of Chr9. (*B*) Differences in CENH3 profiles correlate with genetic change. "CENH3 fold enrichment" is the normalized ratio of CENH3-ChIP-seq reads to total nucleosome control reads per 100-kb locus across the genome. "SNPs matching ZH13" is the number of SNPs in 100-kb bin that matched ZH13 for each individual. (*C*) The CENH3 profiles in ZH13, C8, L4, ZH13-C8-S9, and ZH13-L4-S9 using the Integrative Genomic Viewer tool. The picture shows the 12-Mb window of Chr16.

binding with various satellite sequences and lengths. Such adaptability enables them to function effectively in diverse genomes and under varying conditions, ensuring proper chromosome segregation and genetic stability. However, fully established centromeric satellite repeats often consist of monomers with lengths in the range of a single nucleosome, typically ranging from 150 to 180 bp. Examples include pAL1 in *Arabidopsis* (50), CentO in rice (48), CentC in maize (51), and Ss1 in *Saccharum spontaneum* (52, 53). The advantage of such satellite repeats with monomeric lengths lies in their adaptability for a single CENH3 nucleosome, facilitating the assembly of translationally phased CENH3 nucleosome arrays, which are essential for centromere establishment and maintenance (54). In contrast, newly emerged centromeric repeats vary widely in their monomeric lengths, ranging from a few hundred base pairs to several kilobases (55, 56). These "odd" repeats are often restricted to a few centromeres and have not spread to all centromeres (57–59). In our study, we identified CentGm91, CentGm92, and CentGm413 as the major components of soybean centromeres, and their lengths do not fall within the range of a single nucleosome. Moreover, the distinct distribution and intensities revealed by FISH suggested they might undergo a dynamic amplification and adaptation. Based on these observations, we propose that the frequent centromere repositioning in the soybean genome may be partly attributed to the presence of immature satellite structures that are not fully compatible with centromere function. Interestingly, we also identified two classes of satellite repeats specially located in Cen1 (Fig. 2*C*). These two satellites were not only found in landraces and cultivars, but also in wild soybeans. However, their distribution patterns with the other three satellites were distinct in different accessions. The most significant difference is that CentGm273 located to the right side of CentGm91 in C1 and L7 lines, where they bind CENH3 to form a functional centromere, whereas in the other lines including three wild soybeans, CentGm273 moved to the left side relative to CentGm91 and lost the ability to recruit CENH3, except the line C11 (Fig. 3*A*). These results suggested large-scale inversion in Cen1 occurred during soybean domestication and improvement and this inversion had obvious effect on centromere location (Fig. 3*B*). When the two satellites were formed remained unclear, however, analysis of the presence of CentGm273 and CentGm444 in common bean, the close relative to soybean, would provide more information about their origin.

The phenomenon of centromere repositioning, particularly in plant species, has been a subject of growing interest and investigation. While there are fewer reports on centromere repositioning in plants compared to animals, the existing studies have provided valuable insights into the dynamics and mechanisms of this process. For instance, the first observed ENC in plants was detected on truncated barley telosomes (60). Additionally, Han et al. utilized FISH mapping to observe different centromere positions between two pairs of chromosomes in cucumber and melon (18). In another inter-species comparison, Liao et al. identified a centromere repositioning event on chromosome 12 in *Oryza brachyantha*, moving the centromere approximately 400 kb away from its original location when compared to two Oryza (rice) genomes and the outgroup genome of *Leersia perrieri* (19). Maize (*Zea mays*) has been a notable model for studying de novo centromere formation and centromere inactivation. Wang et al. observed centromere repositioning and size expansion on one of the eight maize chromosomes after transfer to the oat background in oat-maize hybrid lines (61). Additionally, Liu et al. demonstrated the inactivation of a newly formed centromere on an engineered maize chromosome, with de novo centromeres forming elsewhere on that chromosome (62). Furthermore, studies by Zhao et al. revealed the propensity for de novo centromere formation within a latent region 2 Mb away from the original centromere on maize chromosome 3 (43). Schneider et al. conducted a large-scale study in domesticated maize lines and documented 57 independent centromere shifts associated with the decay of original centromeres (22). A notable example of ENCs in plants comes from the largest crucifer tribe, Arabideae, which comprises approximately 550 species in the mustard family, Brassicaceae. Interestingly, despite the virtual absence of chromosome number variation, the intra-tribal diversification within Arabideae exhibited a high frequency of ENCs (17). This example of ENCs adds to our understanding of centromere dynamics in plant genome evolution. In this study, our CENH3-ChIP approach allowed us to determine the precise positions of ENCs in the 27 analyzed lines. Some ENCs are apparently shared at different, well-defined clades, such as the ENC at chromosome 1 (ENC1) in lines of C1, C11, and L7, and ENC16 in lines of C7, L1, and L4. The same position of ENCs in distantly related accessions suggests that certain regions of the genome have a propensity to form centromeres (63, 64). In *C. albicans*, centromere proximal regions are the most preferred for assembly of induced new centromeres (44). Similarly, in chicken, 76% of the induced new centromeres on the Z chromosome form near the original centromere (65). In maize, at least three independent new centromere activation events occurred within the pericentromeric region of maize chromosome 3 (43). Our results showed new centromere formed at multiple positions in the soybean genome, but most formed in close proximity to the native centromere. Meiotic recombination is suppressed in centromeres of eukaryotic chromosomes (66). It is reasonable to predict that the distally formed new centromeres have the potential to form disastrous meiotic anaphase bridges if recombination occurs between mismatched centromeres in a heterozygote. Thus, a chromosome with a new centromere formed adjacent to the native centromere will be conducive to survive because crossovers will not likely occur between the native centromere and new centromere (43, 67).

Previous results revealed that in hybrids (three mules and one hinny), centromere sliding can occur in one generation. This positional movement is not large. In other words, the centromeric domain in the hybrid is always partially overlapping the domain of the parent (13). Here, we showed extensive centromere repositioning occurred after subsequent generations in hybrid genetic backgrounds. Similar to the slight movements of centromere in one generation of hybrids (three mules and one hinny) (13), we found the soybean centromeres do not jump to a completely new location in the ninth generation of hybrids. They remained partially overlapping the centromere domain of the parent. Furthermore, the movement of centromeric domains do not seem to be random. They are inclined to bind to centromere satellites, suggesting tandem repeats may provide favorable environment for CENH3. Though centromeres are believed to be determined by epigenetics, independent of the underlying DNA sequence, as revealed by the finding of new centromere and satellite-free centromere, it is clear that abundant satellites found in many eukaryotic centromeres also contribute to centromere stability (68, 69). Recent studies showed that centromeric DNA sequences from yeast, humans, and plants are enriched in non-B-form DNA, which were recognized as important players in centromere activity and stability (4–6, 8). While the mechanism underlying the role of centromere satellites in centromere function remains unknown, it is evident that various factors, including proper transcription levels, recombination, and non-B-form structures, collaboratively contribute to subtly influencing the positioning and stability of centromeres.

Methods

Plant Materials. All soybean seeds were planted in a mixture of soil (pindstrup, Denmark) and vermiculite at 25 °C with a photoperiod of 16-h light/8-h dark in a greenhouse at the Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing (at N 40.22°, E 116.23°). The leaves of 10-d-old seedlings were collected and frozen immediately in liquid nitrogen and then stored at −80 °C for subsequent analysis.

ChIP and ChIP-Seq. ChIP was conducted as previously described with some modifications (8). Briefly, 3 to 5 g young leaves was ground into a fine powder using liquid nitrogen. Nuclei were isolated in lysis buffer 1 (1 mM EDTA, 1 × Complete Mini EDTA-free protease inhibitor cocktail, 50 mM 4-(2-Hydroxyethyl)-1-piperazi neethanesulfonic acid (HEPES) pH 7.5, 150 mM NaCl, 1.0% TritonX-100, and 10% glycerol, 5 mM β-ME). Pelleted nuclei were washed with lysis buffer 2 (1 mM EDTA, 1 × Complete Mini EDTA-free protease inhibitor cocktail, 50 mM HEPES pH 7.5, 150 mM NaCl, 1.0% TritonX-100, 10% glycerol) and centrifuged at 7,000 g for 3 min at 4 °C twice. Pelleted nuclei were then washed with MNase buffer (10% sucrose, 50 mM Tris-HCl, pH 7.5, 1 mM $CaCl_2$, 1 mM $MgCl_2$, and 0.1 mM Phenylmethylsulfonyl fluoride) and centrifuged at 7,000 g for 3 min at 4 °C once. Pelleted chromatin was resuspended in MNase buffer and digested with MNase (New England BioLabs, 0.5 U) at 37 °C for 24 min to shear the chromatin into a size range suitable for sequencing. The reaction was quenched with 0.5 M EDTA and centrifuged at 12,000 rpm for 3 min at 4 °C. Input sample was taken from the supernatant and anti-CENH3 antibody (1:400 dilution) was added into the remainder, incubating at 4 °C overnight. Immunocomplexes were recovered by 50 μL protein A beads and were sequentially washed in washing buffer (50 mM Tris-HCl pH 7.5, 0.2 mM PMSF, 1 × Complete Mini EDTA-free protease inhibitor cocktail, 10 mM EDTA) containing 50, 100, and 150 mm NaCl. The bound immune complex was eluted with elution buffer (50 mM NaCl, 20 mM Tris-HCl, 5 mM EDTA, 2% SDS) at 65 °C for 40 min. Nucleic acids were resuspended in 15 uL of TE buffer (pH 8.0). The ChIP and input DNA samples were then used to construct the libraries according to the protocol provided by the NadPrep DNA library Preparation Kit for Illumina (Nanodigmbio, cat#1002101) and NadPrep UDI Adapter Kit Set C1 for Illumina (Nanodigmbio, cat#1003221). Libraries were sequenced on the Illumina NovaSeq platform at 150-bp paired-end reads.

ChIP-Seq Raw Data Processing and Alignment. The adapters and low-quality reads were removed using Trimmomatic (version 0.36) (70) with the parameters "ILLUMINACLIP: adapter. fa: 2:30:10 LEADING:20 TRAILING:20 MINLEN:36 SLIDINGWINDOW: 4:20." The quality-controlled reads were then aligned to the soybean ZH13 reference genome (35) using Burrows Wheeler Aligner BWA-MEM software (71) with default parameters and were further filtered by SAMtools (version 1.3.1) (72) for nonunique and duplicated reads. For visualization, aligned and filtered BAM files were converted to normalized coverage files (bigwig) with 5 bp bins using Deeptools (version 3.5.1) (73). The Integrative Genomics Viewer (IGV) (74) is used to view CENH3-ChIP-seq data. For the Figs. 3 *A–E* and 4*B*, MACS2 (75) was used to call CENH3-ChIP-seq peaks using input as a control with the settings: -f BAMPE -g 1.0e^9. Peak numbers were counted in each 100-kb bin along the chromosomes, and the plots were produced using Excel software.

Identifying the Centromeric Repeats and Full-Length LTR-RTs (flLTR-RTs). The TRASH pipeline was used to identify CentGm91, CentGm92, CentGm413, CentGm273, and CentGm444 satellites in soybean centromeres, and the StainedGlass package was used to generate sequence identity heat maps. For the analysis of RepeatExplorer, a total of 10 million randomly selected sequence reads from the input control were analyzed using Web-based Galaxy RepeatExplorer software (https://repeatexplorer-elixir.cerit-sc.

cz/galaxy/). To identify repeats associated with CENH3 nucleosomes, the CENH3-ChIP-seq and input reads were subjected to BLAST analysis against the cluster repeats with the parameters "-outfmt 6 -evalue 1e-8 -num_alignments 500." The flLTR-RTs were identified using LTRharvest (76), where LTR_finder (77) and LTRharvest (78) were employed to identify flLTR-RTs based on the principle of structure *ab initio*. The following parameter settings were used: "ltr_finder -D 25000 -d 3000 -L 2000 -l 100 -p 20 -C -M 0.85" and "ltrharvest -overlaps best -seed 30 -minlenltr 100 -maxlenltr 2000 -mindistltr 3000 -maxdistltr 25000 -similar 85 -mintsd 4 -maxtsd 20 -motif tgca -motifmis 1 -vic 60 -xdrop 5 -mat 2 -mis -2 -ins -3 -del -3," respectively. LTR_retriever was then used to integrate both candidates.

Immunolocalization and FISH. Immunolocalization and FISH were performed as described (79). For the immunostaining assay, an anti-CENH3 antibody was raised against peptides designed according to the previously identified soybean CENH3 protein sequences (36). For the FISH assay, mitotic chromosomes were prepared from root apical meristems. Oligonucleotide probes were either 5′-labeled with red fluor (5′ TEX 615) or green fluor (5′ 6-FAM) during their synthesis (BGI Tech). The oligos sequences are listed in *SI Appendix*, Table S4. Images were collected by confocal microscopy (Zeiss Cell Observer SD) and were processed with ZEN 2009 Light Edition (Zeiss, http://www.zeiss.com/) and Adobe Photoshop CS 6.0 software.

Synteny Analysis. We constructed the synteny map using orthologous genes as markers at each centromeric region. MCScanX (version 1.1) (80) was used to identify homologous blocks, requiring at least five collinear gene pairs within one block and fewer than 25 intervening genes.

Determining SNP Identity. The quality-controlled reads from ZH13-C8-S9 and ZH13-L4-S9 were aligned to the ZH13 reference genome (35) using BWA-MEM (71). Low mapping quality reads and duplicated reads were removed by SAMtools (version 1.3.1) (72) and Picard tools (http://broadinstitute.github.io/picard) (version 2.26.9). The SNPs were identified using GATK (The Genome Analysis Toolkit, version 3.8.1) (81) and bcftools (Tools for manipulating Variant Call Format and Binary Variant Call Format, version 1.15.1). The resulting sequences were filtered using GATK (parameters: -T VariantFiltration --filterExpression "QD < 2.0 || FS > 200.0 || SOR > 10.0 || MQRankSum < −12.5 || ReadPosRankSum < −8.0" --filterName "PASS"). We divided the genome into 100-kb bins and calculated the number of SNPs in each bin that matched ZH13 for each individual.

Data, Materials, and Software Availability. The anti-CENH3 ChIP-seq and input data are submitted to the National Genomics Data Center (NGDC) Genome Sequence Archive (GSA; https://bigd.big.ac.cn/gsa/) under accession number CRA009445 (82). All other data are included in the manuscript and/or supporting information.

ACKNOWLEDGMENTS. We thank James A. Birchler from University of Missouri, Gernot Presting from University of Hawaii, Jonathan Gent from University of Georgia, and Nathan D. Han from The Edison Family Center for critical reading of the article and helpful comments. We thank Lili Dong (Beijing Institute of Genomics, Chinese Academy of Sciences) for helping us to upload the sequencing data. This work was supported by the National Natural Science Foundation of China (31991212).

Author affiliations: [a]State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Innovation Academy for Seed Design, Chinese Academy of Sciences, Beijing 100101, China; [b]College of Advanced Agricultural Sciences, University of the Chinese Academy of Sciences, Beijing 100049, China; and [c]Guangdong Laboratory for Lingnan Modern Agriculture, College of Life Sciences, South China Agricultural University, Guangzhou 510642, China

1. L. Comai, S. Maheshwari, M. P. A. Marimuthu, Plant centromeres. *Curr. Opin. Plant Biol.* **36**, 158–167 (2017).
2. H.-G. Yu, R. K. Dawe, E. N. Hiatt, R. K. Dawe, The plant kinetochore. *Trends Plant Sci.* **5**, 543–547 (2000).
3. J. Zhou *et al.*, Centromeres: From chromosome biology to biotechnology applications and synthetic genomes in plants. *Plant Biotechnol. J.* **20**, 2051–2063 (2022).
4. Y. Liu *et al.*, Genome-wide mapping reveals R-loops associated with centromeric repeats in maize. *Genome Res.* **31**, 1409–1418 (2021).
5. S. Kasinathan, S. Henikoff, Non-B-Form DNA is enriched at centromeres. *Mol. Biol. Evol.* **35**, 949–962 (2018).
6. V. S. P. Patchigolla, B. G. Mellone, Enrichment of Non-B-Form DNA at D. melanogaster centromeres. *Genome Biol. Evol.* **14**, evac054 (2022).
7. Y. Liu, C. Wang, H. Su, J. A. Birchler, F. Han, Phosphorylation of histone H3 by Haspin regulates chromosome alignment and segregation during mitosis in maize. *J. Exp. Botany* **72**, 1046–1058 (2021).
8. Q. Liu *et al.*, Non-B-form DNA tends to form in centromeric regions and has undergone changes in polyploid oat subgenomes. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2211683120 (2023).
9. W. H. Shang *et al.*, Chickens possess centromeres with both extended tandem repeats and short non-tandem-repetitive sequences. *Genome Res.* **20**, 1219–1228 (2010).
10. O. K. Smith *et al.*, Identification and characterization of centromeric sequences in Xenopus laevis. *Genome Res.* **31**, 958–967 (2021).
11. D. P. Melters *et al.*, Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.* **14**, R10 (2013).
12. L. Ávila Robledillo *et al.*, Extraordinary sequence diversity and promiscuity of centromeric satellites in the Legume Tribe Fabeae. *Mol. Biol. Evol.* **37**, 2341–2356 (2020).
13. S. G. Nergadze *et al.*, Birth, evolution, and transmission of satellite-free mammalian centromeric domains. *Genome Res.* **28**, 789–799 (2018).
14. I. Schubert, What is behind "centromere repositioning"? *Chromosoma* **127**, 229–234 (2018).
15. M. Rocchi, N. Archidiacono, W. Schempp, O. Capozzi, R. Stanyon, Centromere repositioning in mammals. *Heredity* **108**, 59–67 (2012).
16. G. Montefalcone, S. Tempesta, M. Rocchi, N. Archidiacono, Centromere repositioning. *Genome Res.* **9**, 1184–1188 (1999).
17. T. Mandáková, P. Hloušková, M. A. Koch, M. A. Lysak, Genome evolution in Arabideae was marked by frequent centromere repositioning. *Plant Cell* **32**, 650–665 (2020).
18. Y. Han *et al.*, Centromere repositioning in cucurbit species: Implication of the genomic impact from centromere activation and inactivation. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 14937–14941 (2009).
19. Y. Liao *et al.*, Comparison of Oryza sativa and Oryza brachyantha genomes reveals selection-driven gene escape from the centromeric regions. *Plant Cell* **30**, 1729–1744 (2018).
20. M. Ventura *et al.*, Evolutionary formation of new centromeres in Macaque. *Science* **316**, 243–246 (2007).
21. G. Chiatante *et al.*, Centromere repositioning explains fundamental number variability in the New World monkey genus Saimiri. *Chromosoma* **126**, 519–529 (2017).
22. K. L. Schneider, Z. Xie, T. K. Wolfgruber, G. G. Presting, Inbreeding drives maize centromere evolution. *Proc. Natl. Acad. Sci. U.S. A.* **113**, E987–E996 (2016).
23. D. Tolomeo *et al.*, Epigenetic origin of evolutionary novel centromeres. *Sci. Rep.* **7**, 41980 (2017).
24. L. Carbone *et al.*, Evolutionary movement of centromeres in horse, donkey, and zebra. *Genomics* **87**, 777–782 (2006).
25. F. M. Piras *et al.*, Uncoupling of satellite DNA and centromeric function in the Genus Equus. *PLOS Genet.* **6**, e1000845 (2010).
26. R. F. Wilson, "Soybean: Market driven research needs" in *Genetics and Genomics of Soybean*, G. Stacey, Ed. (Springer New York, New York, NY, 2008), pp. 3–15.
27. J. Orf, B. W. Diers, H. R. Boerma, "Soybeans: Improvement, production, and uses", *Agron. Monogr. Genetic Improvement: Conventional and Molecular-Based Strategies*, H. R. Boerma, J. E. Specht, Eds. (American Society of Agronomy, 2004), **vol. 16**, pp. 417–450.
28. E. J. Sedivy, F. Wu, Y. Hanzawa, Soybean domestication: The origin, genetic architecture and molecular bases. *New Phytol.* **214**, 539–553 (2017).
29. H.-M. Lam *et al.*, Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* **42**, 1053–1059 (2010).
30. Z. Zhou *et al.*, Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* **33**, 408–414 (2015).
31. Y.-H. Li *et al.*, Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. *BMC Genomics* **14**, 579 (2013).
32. J. Schmutz *et al.*, Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
33. M. Xie *et al.*, A reference-grade wild soybean genome. *Nat. Commun.* **10**, 1216 (2019).
34. Y. Liu *et al.*, Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176.e113 (2020).
35. Y. Shen *et al.*, Update soybean Zhonghuang 13 genome to a golden reference. *Sci. China Life Sci.* **62**, 1257–1260 (2019).
36. A. L. Tek, K. Kashihara, M. Murata, K. Nagaki, Functional centromeres in soybean include two distinct tandem repeats and a retrotransposon. *Chromosome Res.* **18**, 337–347 (2010).
37. N. Gill *et al.*, Molecular and chromosomal evidence for allopolyploidy in soybean. *Plant Physiol.* **151**, 1167–1174 (2009).
38. M. Naish *et al.*, The genetic and epigenetic landscape of the Arabidopsis centromeres. *Science* **374**, eabi7489 (2021).
39. P. Wlodzimierz *et al.*, Cycles of satellite and transposon evolution in Arabidopsis centromeres. *Nature* **618**, 557–565 (2023).
40. J. Du *et al.*, SoyTEdb: A comprehensive database of transposable elements in the soybean genome. *BMC Genomics* **11**, 113 (2010).
41. J. Du *et al.*, Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: Insights from genome-wide analysis and multi-specific comparison. *Plant J. Cell Mol. Biol.* **63**, 584–598 (2010).
42. P. Wlodzimierz, M. Hong, I. R. Henderson, TRASH: Tandem repeat annotation and structural hierarchy. *Bioinformatics* **39** (2023).
43. H. Zhao *et al.*, Recurrent establishment of de novo centromeres in the pericentromeric region of maize chromosome 3. *Chromosome Res.* **25**, 299–311 (2017).
44. J. Thakur, K. Sanyal, Efficient neocentromere formation is suppressed by gene conversion to maintain centromere function at native physical chromosomal loci in Candida albicans. *Genome Res.* **23**, 638–652 (2013).
45. J. I. Gent, N. Wang, R. K. Dawe, Stable centromere positioning in diverse sequence contexts of complex and satellite centromeres of maize and wild relatives. *Genome Biol.* **18**, 121 (2017).
46. J. Jiang, J. A. Birchler, W. A. Parrott, R. Kelly Dawe, A molecular view of plant centromeres. *Trends. Plant Sci.* **8**, 570–575 (2003).

47. H. Su *et al.*, Centromere satellite repeats have undergone rapid changes in polyploid wheat subgenomes. *Plant Cell* **31**, 2035–2051 (2019).
48. Z. Cheng *et al.*, Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell* **14**, 1691–1704 (2002).
49. J. Maluszynska, J. S. Heslop-Harrison, Localization of tandemly repeated DMA sequences in Arabidopsis thaliana. *Plant J.* **1**, 159–166 (1991).
50. K. Nagaki *et al.*, Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of Arabidopsis thaliana centromeres. *Genetics* **163**, 1221–1225 (2003).
51. T. K. Wolfgruber *et al.*, Maize centromere structure and evolution: Sequence analysis of centromeres 2 and 5 reveals dynamic Loci shaped primarily by retrotransposons. *PLoS Genet.* **5**, e1000743 (2009).
52. W. Zhang *et al.*, Isolation and characterization of centromeric repetitive DNA sequences in Saccharum spontaneum. *Sci. Rep.* **7**, 41659 (2017).
53. Y. Huang *et al.*, The formation and evolution of centromeric satellite repeats in Saccharum species. *Plant J.* **106**, 616–629 (2021).
54. D. Hasson *et al.*, The octamer is the major form of CENP-A nucleosomes at human centromeres. *Nat. Struct. Mol. Biol.* **20**, 687–695 (2013).
55. Z. Gong *et al.*, Repeatless and repeat-based centromeres in potato: Implications for centromere evolution. *Plant Cell* **24**, 3559–3574 (2012).
56. H. Zhang *et al.*, Boom-Bust turnovers of megabase-sized centromeric DNA in solanum species: Rapid evolution of DNA sequences associated with centromeres. *Plant Cell* **26**, 1436–1447 (2014).
57. H. R. Lee *et al.*, Chromatin immunoprecipitation cloning reveals rapid evolutionary patterns of centromeric DNA in Oryza species. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 11793–11798 (2005).
58. P. Neumann *et al.*, Stretching the rules: Monocentric chromosomes with multiple centromere domains. *PLoS Genet.* **8**, e1002777 (2012).
59. A. Iwata *et al.*, Identification and characterization of functional centromeres of the common bean. *Plant J. Cell Mol. Biol.* **76**, 47–60 (2013).
60. S. Nasuda, S. Hudakova, I. Schubert, A. Houben, T. R. Endo, Stable barley chromosomes without centromeric repeats. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 9842–9847 (2005).
61. K. Wang, Y. Wu, W. Zhang, R. K. Dawe, J. Jiang, Maize centromeres expand and adopt a uniform size in the genetic background of oat. *Genome Res.* **24**, 107–116 (2014).
62. Y. Liu *et al.*, Sequential de novo centromere formation and inactivation on a chromosomal fragment in maize. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E1263–E1271 (2015).
63. M. Ventura *et al.*, Recurrent sites for new centromere seeding. *Genome Res.* **14**, 1696–1703 (2004).
64. O. Capozzi *et al.*, Evolutionary descent of a human chromosome 6 neocentromere: A jump back to 17 million years ago. *Genome Res.* **19**, 778–784 (2009).
65. W.-H. Shang *et al.*, Chromosome engineering allows the efficient isolation of vertebrate neocentromeres. *Dev. Cell* **24**, 635–648 (2013).
66. K. H. Choo, Why is the centromere so cold? *Genome Res.* **8**, 81–82 (1998).
67. J. C. Lamb, J. M. Meyer, J. A. Birchler, A hemicentric inversion in the maize line knobless Tama flint created two sites of centromeric elements and moved the kinetochore-forming region. *Chromosoma* **116**, 237–247 (2007).
68. P. B. Talbert, S. Henikoff, The genetics and epigenetics of satellite centromeres. *Genome Res.* **32**, 608–615 (2022).
69. J. I. Gent, K. Wang, J. Jiang, R. K. Dawe, Stable patterns of CENH3 occupancy through maize lineages containing genetically similar centromeres. *Genetics* **200**, 1105–1116 (2015).
70. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
71. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
72. H. Li *et al.*, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
73. F. Ramírez *et al.*, deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
74. H. Thorvaldsdóttir, J. T. Robinson, J. P. Mesirov, Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
75. Y. Zhang *et al.*, Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
76. D. Ellinghaus, S. Kurtz, U. Willhoeft, LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinform.* **9**, 18 (2008).
77. Z. Xu, H. Wang, LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic acids Res.* **35**, W265–W268 (2007).
78. S. Ou, N. Jiang, LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
79. Y. Liu *et al.*, Cohesion and centromere activity are required for phosphorylation of histone H3 in maize. *Plant J.* **92**, 1121–1131 (2017).
80. Y. Wang *et al.*, MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
81. A. McKenna *et al.*, The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
82. Y. Liu *et al.*, Pan-centromere reveals widespread centromere repositioning of soybean genomes. Genome Sequence Archive. https://ngdc.cncb.ac.cn/gsa/browse/CRA009445. Deposited 3 January 2023.