Genome Biology

RESEARCH Open Access

Genetic diversity and architectural dynamics of soybean centromeres

Yicheng Huang[1,6], Enlai Guan[1], Shipeng Song[2], Dal-Hoe Koo[3], Monica A. Schmidt[4], Handong Su[1,5], Chunli Chen[2] and Jianwei Zhang[1*]

*Correspondence:
jzhang@mail.hzau.edu.cn

[1] National Key Laboratory of Crop Genetic Improvement, Hubei Hongshan Laboratory, Shenzhen Institute of Nutrition and Health, Huazhong Agricultural University, Wuhan 430070, China
Full list of author information is available at the end of the article

Abstract

Background: Centromere function is fundamental and conserved across eukaryotes, despite highly divergent DNA sequences, even among closely related species. These regions often contain rapidly evolving repeats and retrotransposons, yet play a crucial role in chromosome segregation. Soybean, which harbors two distinct types of centromeric satellite repeats, is an ideal model for studying centromeric repeat organization and function.

Results: Here we generate the complete map of centromeric satellite repeats revealing the organizational patterns of different types of centromeric satellite repeats within centromeres. These maps are constructed using three recently available telomere-to-telomere soybean genomes. We find that certain centromeric satellite repeats exhibit chromosome-specific evolutionary trajectories and may serve distinct functional roles in centromere activity. We further analyze the potential relationship between centromere-specific histones H3 (CENH3) and centromeric satellite repeats, identifying consensus motifs associated with CENH3-binding sites. We also analyze the higher-order tandem repeats of the centromere and propose a hypothetical model of centromeric DNA replication.

Conclusions: We conclude that *CentGm*-1 and *CentGm*-4 evolve independently. The observation that completely identical *CentGm*-4 sequences consistently appear on the same chromosome across different soybean varieties indicates a stronger chromosome-specific preference for *CentGm*-4. We propose a model in which replication templates within the centromere region originate from multiple CENH3-nucleosome complexes bound to *CentGm* sequences. Both *CentGm*-1 and *CentGm*-4 contain similar motifs with the potential to bind CENH3 protein. The findings provide a new insight into the mechanisms behind centromere diversity and dynamics.

Background

The centromere serves as the foundation for kinetochore formation, which mediates accurate chromosome segregation by binding to spindle microtubules. However, a complete understanding of centromeres remains elusive due to their high repetitiveness, sequence similarity, and structural complexity. The absence of complete centromeric sequences in previous genome assemblies based on short-read sequencing technology resulted in the clustering and assembly of individual repeats into a single sequence, relegating centromere in the "grey side" of the genomes [1, 2]. Advances in third-generation sequencing technologies have significantly improved both read length and accuracy, greatly enhancing the ability to observe and assemble repetitive sequences, particularly those in the centromeric region [3–5]. Despite significant advances in genome assembly, the structural and functional organization of centromeres remains poorly understood.

Soybean is not only one of the fastest-growing global crops in recent decades, but it is emerging as a unique model for centromere research. Unlike most other known plant species, mainly monocots, which have a single centromeric repeat [6, 7], soybean has two major distinct types of centromeric satellite repeats (*CentGm*-1 and *CentGm*-4) in its functional centromeres [8]. Recently, telomere-to-telomere (T2T) genomes of different soybean cultivars (WM82, ZH13 and Jack) have been released to the community [9–14]. The growing availability of complete genomes, including centromeres located within highly repetitive heterochromatin regions, presents new opportunities to investigate centromere structure and function.

The centromere formation is notably diverse and exhibits rapidly evolving DNA compartments, yet it maintains a highly conserved mechanism for chromosome segregation, known as the centromere paradox [15]. Most eukaryotic centromeres contain numerous satellite repeat sequences organized into large tandem repeat arrays [15], forming chromosome-specific higher order repeat (HOR) structures [16–21]. The role of these satellite repeats in binding the centromere-specific histone H3 (CENH3) in nucleosomes (analogous to CENP-A in animals) remains unclear, despite their essential function in separating sister chromatids during meiosis [22]. Consensus sequences in centromeric regions from some species have been identified that are sufficient in binding to CENH3 proteins [23, 24]. In humans, there exists a 17 bp motif in α-satellite repeats known as the CEN-B box, which binds to the CENP-A protein in centromeres [25]. However, some centromeres lacking α-satellite DNA have also been identified in human patients, suggesting that these satellite repeats are neither necessary nor sufficient for centromere function [26–28]. Moreover, in some animals, such as horses [29, 30] and chickens [31], functional centromeres with or without tandem repeat sequences have been identified on different chromosomes.

Although satellite repeat sequences are not directly implicated in centromere functions, the widespread presence of tandem repeats in the centromeres of many plant and animal species remains poorly understood. This raises an intriguing question: do different combinations of tandem repeat sequences serve distinct functional roles, and do they contribute to maintaining centromere integrity amid rapid evolutionary changes? With the increasing availability of gap-free reference genomes and the unique presence of two significantly divergent types of satellite repeat sequences, soybean emerges as an ideal model for investigating centromere structure and function. In this study, we provide the complete centromeric satellite maps for three currently available complete soybean genomes, revealing centromere-specific chromosomal structures by exploring the organizational patterns of different satellite repeats within centromeres. We analyzed these higher-order tandem repeats and proposed a model outlining how centromeric repetitive regions might replicate, hypothesizing the mechanisms responsible for the accumulation of centromeric tandem repeat sequences.

Results

Two major types of centromeric satellite sequences in soybean

Soybean centromeres are composed of megabase-scale arrays of tandemly arranged satellite DNA (*CentGm*-1 and *CentGm*-4), which were previously identified by ChIP-seq [8] and are presumed to have an affinity for binding centromere-specific histone H3 (CENH3) proteins. Multiple genome assemblies are available for different soybean varieties; for consistency, we used the assemblies from Wang et al. for WM82 [9], Zhang et al. for ZH13 [12], and Huang et al. for Jack [14]. We developed a pipeline to de novo identify *CentGm* sequences in the WM82, ZH13 and Jack genomes, thereby enabling the detection of potential centromeric regions (Additional file 1: Fig. S1, Additional file 2: Table S1), which were found to be consistent with the ChIP-seq coverage regions (Additional file 1: Fig. S2). We subsequently identified two predominant types of tandem repeats sequences (TRSs) (~ 91 bp and ~ 410 bp) based on sequence similarity, which occur in the centromeric regions of all chromosomes and are classified as *CentGm*-1 and *CentGm*-4 sequences (Additional file 2: Table S2). A total of 684,595 *CentGm*-1 monomers (62,685,276 bp) and 8,417 *CentGm*-4 monomers (3,460,061 bp) were identified in Jack, 626,713 *CentGm*-1 monomers (57,197,800 bp) and 9,307 *CentGm*-4 monomers (3,839,892 bp) in ZH13, and 656,688 *CentGm*-1 monomers (60,147,825 bp) and 11,101 *CentGm*-4 monomers (4,582,198 bp) in WM82 (Additional file 1: Fig. S3, Additional file 2: Table S2).

By aligning similar *CentGm* monomers from all chromosomes, we determined the most prevalent base for each locus within these two predominant *CentGm* types and generated two representative *CentGm* sequences (Fig. 1a, Additional file 1: Figs. S4 and S5). As a result, a 93-bp representative *CentGm*-1 sequence and a 451-bp representative *CentGm*-4 sequence were identified in soybean cv Jack, a 93-bp *CentGm*-1 sequence and a 457-bp *CentGm*-4 sequence in ZH13, a 93-bp *CentGm*-1 sequence and a 454-bp *CentGm*-4 sequence in WM82 (Additional file 1: Figs. S6 and S7). We validated the centromeric repeat regions using fluorescence in situ hybridization (FISH), as these regions are prone to misassembly due to their high density of repetitive elements (Additional file 1: Fig. S8, Additional file 2: Table S3).

Through analyzing the distribution of the *CentGm*-1 and *CentGm*-4 monomers in the centromeric regions, we found that similar monomers tend to aggregate in adjacent positions (Fig. 1b) and can be obviously clustered into two groups (Fig. 1b, Additional file 1: Figs. S11–S30) based on their similarity. We thereby re-identified all *CentGm*-1 and *CentGm*-4 monomers in each chromosome of the WM82, ZH13 and Jack genome assemblies, aligning them to a common start point based on the representative *CentGm* sequences. The detailed results show that *CentGm*-1 and *CentGm*-4 satellite repeats

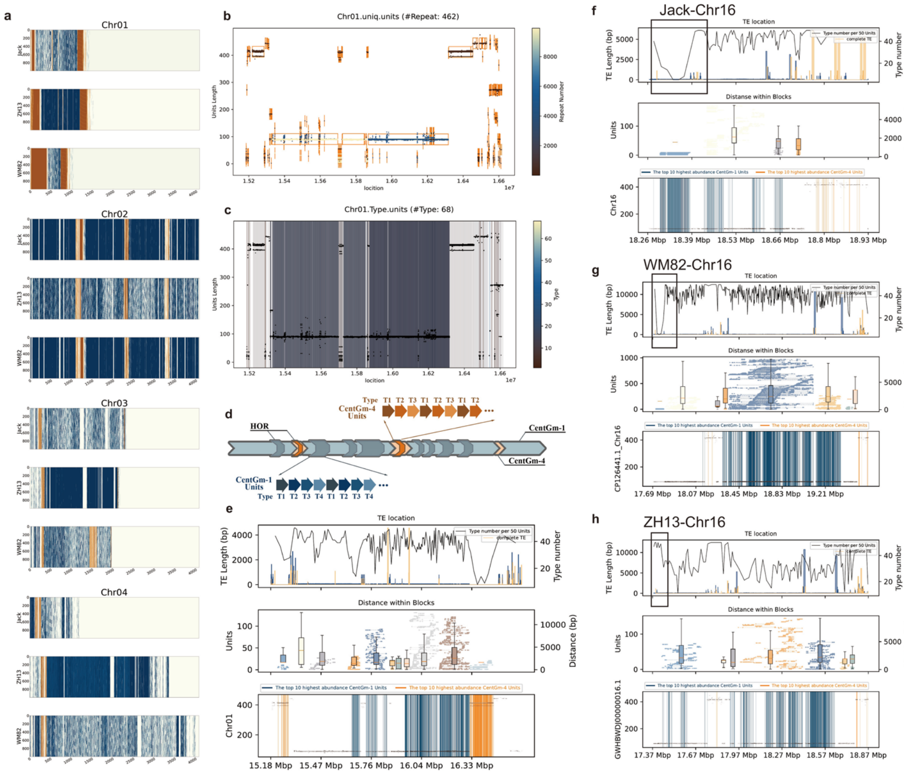

Fig. 1 Two major types of centromeric satellite sequences in soybean. **a** Distribution of *CentGm*-1 (blue) and *CentGm*-4 (orange) in Chr01/Chr02/Chr03/Chr04 of Jack, ZH13 and WM82. Color intensity indicates identity to representative *CentGm* sequences. Centromeric regions from different chromosomes were divided into 1-kbp segments and arranged left-to-right in a rectangle layout (Additional file 1: Fig. S9). Yellow indicates non-*CentGm* sequences, while orange and blue denote *CentGm*-4 and *CentGm*-1, respectively. **b** The distribution of repetitive sequence lengths and numbers in soybean Jack Chr01 centromere region, with orange boxes highlighting the consecutive identical repetitive sequences. The y-axis denotes the length of the repetitive sequences, and the x-axis indicates their positions on the centromere. Different colors correspond to the repeat counts of different repetitive sequences, and the orange boxes outline the entire regions occupied by the same type of satellite repeats. **c** The *CentGm* clustering results for potential Chr01 centromeric regions in soybean Jack reveal sequence repetitiveness, with identical monomers highlighted in the same color to indicate sequence similarity. **d** Diagram of the typical centromere structure of soybean. Similar *CentGm* satellite repeat sequences are arranged to form higher ordered repeats (HORs). **e** Distribution features of *CentGm* monomers and transposon elements (TEs) in Chr01 centromeric region in Jack. (Top) The location of complete TEs (orange line) and incomplete TEs (blue line), and the number of *CentGm* types per 50 *CentGm* monomers (black line); (Middle) Distances between the nearest identical *CentGm* monomers within the same block. The colored dots represent individual *CentGm* monomers, and the same monomers plotted in the same row connected by a colored line. Box plots represent distance distribution of the same *CentGm* monomers. (Bottom) The location of the top 10 high-frequency *CentGm* monomer subtypes (blue for HF10-*CentGm*-1, orange for HF10-*CentGm*-4). The Y-axis is the length of *CentGm* monomers. Black dots represent all *CentGm* monomers. **f** (**g**) and (**h**) Length variation of the *CentGm*-1 fragments on Chr16 of Jack, WM82, and ZH13, respectively, with a *CentGm*-1 fragment marked in a black box. Details are the same as in Fig. 1e

form tandem arrays through distinct combinatorial patterns but no tendency of decreasing conservation from central to peripheral sequences in soybean centromeric regions. Despite various lengths of satellite repeat arrays among different soybean varieties, most centromeres form multiple *CentGm*-enriched loci with high sequence similarity and maintain a relatively conserved alternating composition of “*CentGm*-1 clusters” and “*CentGm*-4 clusters” (Fig. 1d, Additional file 1: Figs. S9 and S10). For instance, Chr01 follows a "*CentGm*-4 cluster → *CentGm*-1 cluster → *CentGm*-4 cluster" pattern, whereas Chr02 exhibits an interwoven arrangement of four *CentGm*-1 clusters and three *CentGm*-4 clusters (Fig. 1a). However, exceptions were observed in specific chromosomal organizations. In Chr03, Jack and ZH13 completely lack the second *CentGm*-4 cluster found in WM82 (Fig. 1a). Additionally, Chr20 demonstrates variety-specific differences in cluster composition, with distinct alternating ratios of *CentGm*-1 clusters to *CentGm*-4 clusters among Jack (3:2), ZH13 (4:3), and WM82 (5:3) (Additional file 1: Fig. S10).

In addition to the *CentGm* clusters, the centromeric regions contain a high abundance of transposable elements (TEs), predominantly long terminal repeats (LTR) retrotransposons of the *Gypsy* type (Additional file 1: Figs. S11–S30, Additional file 2: Table S4). Furthermore, we identified that TEs preferentially insert at the boundaries between different *CentGm* clusters, as well as in regions with greater variation of *CentGm* repeat types (Fig. 1e, Additional file 1: Figs. S11–S30). Certain tandem *CentGm* monomers within some clusters exhibited high sequence identity with no complete or partial TE sequences detected (Additional file 1: Figs. S11–S30), though their length and distribution varied significantly among the three soybean varieties. In the proximal region of Chr16, between the first pair of TE insertion sites, there is a *CentGm*-1 cluster composed of simple, homogeneous repeats of a single type. In soybean Jack, this cluster (18,261,459–18,283,966 bp; 22,507 bp in total) (Fig. 1f) is the longest with a simple composition, whereas in WM82 it is shorter (17,693,749–17,708,206 bp; 14,457 bp in total) (Fig. 1g) and nearly absent in ZH13 (Fig. 1h). These homogeneous fragments predominantly occur in regions where *CentGm* cluster length varies among the varieties. TEs are frequently observed at the boundaries of the *CentGm* cluster, indicating they might play a role in either bridging or disrupting distinct arrays. The high sequence similarity within these fragments likely contributes to the observed discrepancies in the centromeric sequence lengths among various soybeans.

Independent evolution process of *CentGm*—1 and *CentGm*—4

We have observed *CentGm* arrays of varying lengths across different soybean varieties, raising the intriguing possibility that the length variations may contribute to the rapid evolution of centromeres driven by the tandem repeat accumulation. To explore this, we selected the top10 high-frequency *CentGm*-1 and *CentGm*-4 monomers (referred to as HF10-*CentGm*-1 and HF10-*CentGm*-4) on each chromosome (Additional file 1: Figs. S11–S30, Additional file 2: Tables S5–S10). These most frequent monomers formed abundant tandem repeat arrays (*CentGm*-1 copy number > 50 and *CentGm*-4 copy number > 10) at multiple sites within the centromeric regions and notably lacked TEs.

Phylogenetic analysis of HF10-*CentGm*-1 and HF10-*CentGm*-4 revealed that most high-frequency *CentGm* monomers from the same chromosome clustered within the same branch (Fig. 2). HF10-*CentGm*-1 formed two major branches, corresponding to 92-bp and 91-bp subtypes, which were previously reported and considered as distinct centromeric satellite repeats [32]. In contrast, HF10-*CentGm*-4 exhibited greater sequence diversity, with monomer lengths ranging from 313 to 441 bp. Notably, *CentGm* monomers located on the same chromosome but from different soybean varieties also tended to cluster together, indicating the *CentGm* types are conserved across varieties. Despite the physical proximity

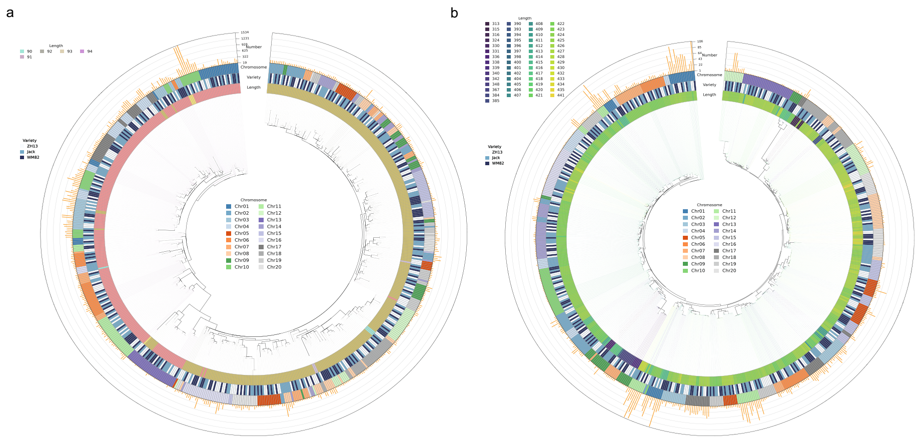

Fig. 2 Phylogenetic analysis of HF10-*CentGm*. From inner to outer, each track represents the length of *CentGm* monomers, soybean variety, chromosome, and copy number of each *CentGm* monomer. **a** HF10-*CentGm*-1, **b** HF10-*CentGm*-4. *CentGm*-1 and *CentGm*-4 undergone independent evolutionary trajectories

of *CentGm*-1 and *CentGm*-4 on the same chromosome, HF10-*CentGm*-1 and HF10-*CentGm*-4 showed distinct clustering patterns, suggesting these two repeat families have likely undergone independent evolutionary trajectories (Fig. 2).

Higher ordered repeats within a conservative center

In the three soybean genomes, the monomer subtypes, distribution, and abundance of HF10-*CentGm*-1 and HF10-*CentGm*-4 on corresponding chromosomes are not entirely consistent (Additional file 1: Figs. S31 and S32). These inconsistencies are particularly evident in regions where the *CentGm* array lengths vary among the varieties. For instance, HF10-*CentGm*-4 is concentrated in the distal portion of the centromeric region on Chr01 (Fig. 1e, Additional file 1: Fig. S11), however, the monomer subtypes of HF10-*CentGm*-4 in ZH13 and WM82 differ from those observed in Jack (Additional file 1: Fig. S32).

By dividing the centromere regions into blocks based on TE locations, we found that *CentGm* monomers – particularly those enriched in HF10-*CentGm* – are arranged in a mosaic-like pattern. Rather than being positioned adjacently within a block, these monomers are interwoven and form higher ordered repeats (HORs), where identical monomers are distributed in a non-contiguous yet orderly fashion (Fig. 1e, Additional file 1: Figs. S11–S30 and S33). Interestingly, similar repetitive monomers within the same block tend to be spaced at consistent intervals (Fig. 1e, Additional file 1: Figs. S11–S30), suggesting an underlying structural regularity. Given that the genome is unlikely to undergo such high-frequency recombination across short DNA segments, these patterns point toward the possibility of distinct replication mechanisms. Altogether, these findings suggest that highly frequent *CentGm* monomers may play a significant role in shaping variation in the length and composition of centromeric satellite repeats among soybean varieties.

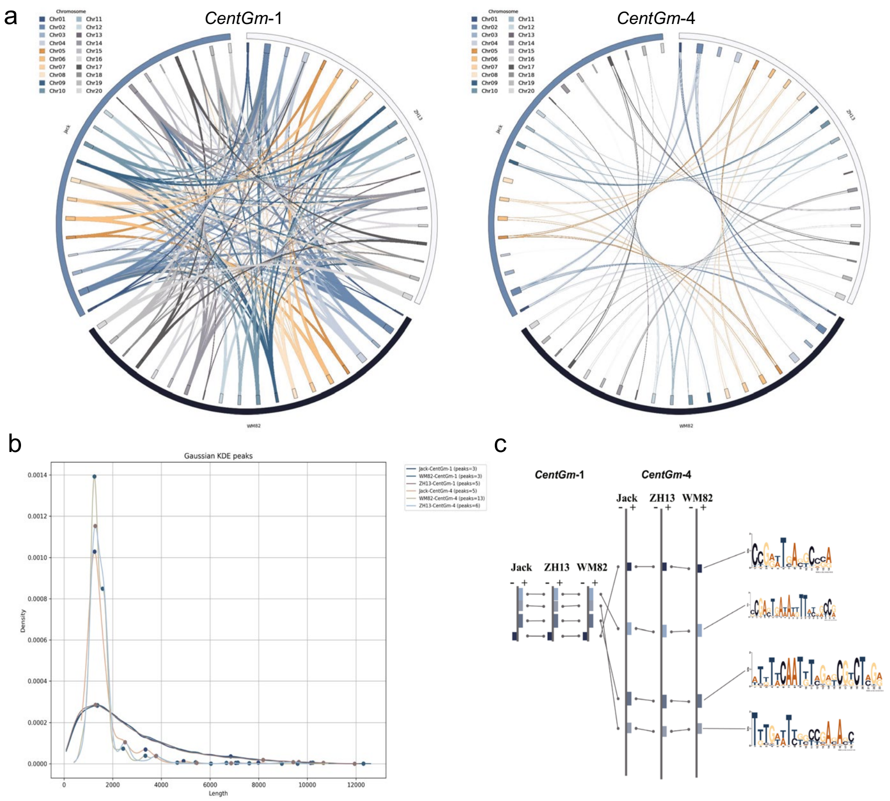

Fig. 3 Soybean *CentGm* monomers across different varieties. **a** Shared *CentGm*-1 and *CentGm*-4 monomers across all three soybean varieties: Jack, ZH13 and WM82. **b** Gaussian kernel density estimation (KDE) of distances between the same *CentGm*-1 or *CentGm*-4 monomers within centromeric regions. Each curve represents the smoothed distribution of distances between the same *CentGm* monomers. Colored markers indicate detected local peaks in the KDE curve, with peak counts labeled in the legend. **c** Four motifs shared in both *CentGm*-1 and *CentGm*-4 that are potentially functionally important for binding CENH3. "±" indicates the DNA strand orientation

Shared *CentGm* sequences occur in different soybean varieties

We analyzed all *CentGm* sequences to determine whether certain monomer types are shared across chromosomes (Additional file 1: Fig. S34). Our results revealed that some *CentGm*-1 monomers are recurrently present on multiple chromosomes, whereas *CentGm*-4 monomers tend to be restricted to a single chromosome. Notably, Chr10 contains *CentGm*-1 monomers that are frequently found across several other chromosomes (Additional file 1: Fig. S34). A similar pattern is also observed among different soybean varieties: *CentGm*-1 monomers are often distributed across multiple chromosomes, while *CentGm*-4 monomers remain more chromosome-specific and conserved (Fig. 3a). For example, the *CentGm*-4 monomer on Chr13 is shared between Jack and WM82 but is markedly different in ZH13 (Fig. 3a). These observations suggest a potential functional distinction between the two *CentGm* types, with *CentGm*-4 possibly playing a role in chromosome-specific centromere identity.

Furthermore, we investigated whether all *CentGm* monomers display similar arrangement patterns within the HOR regions formed by HF-*CentGm*. To this end, we calculated the distances between identical *CentGm* monomers within each region delineated by TEs. After removing outliers caused by occasional insertions of non-repetitive sequences between *CentGm* monomers, we found that the median distances of the same *CentGm*-1 or *CentGm*-4 monomers are relatively consistent across all chromosomes, ranging from 1,067 bp to 3,702 bp (Fig. 3b, Additional file 1: Fig. S35; Additional file 2: Tables S11-S13). These results are similar to the patterns observed in HF-*CentGm* regions and suggest that the broader *CentGm* arrays may have originated from a limited number of HF-*CentGm* monomer types, followed by successive rounds of mutation and recombination, contributing to the diversity of centromeric satellite repeat composition.

Statistical analysis revealed that the distances between identical *CentGm*-1 (*CentGm*-4) monomers did not differ significantly among the three soybean accessions (Jack, WM82, and ZH13), with FDR-adjusted p-values > 0.05 and Cohen's d values < 0.5, indicating stable length distributions within each *CentGm* type (Additional file 2: Tables S11–S12). In contrast, comparisons between *CentGm*-1 and *CentGm*-4 revealed highly significant differences ($p < 0.05$, FDR-adjusted p-values < 0.05), with Cohen's d ranging from 0.64 to 0.78 representing moderate to large effect sizes (Table 1). These findings align with our overall hypothesis that *CentGm*-1 and *CentGm*-4 originated from a limited number of HF-*CentGm* monomer types and evolved independently.

Functional soybean *CentGm* monomers are bound with CENH3

By reanalyzing the ChIP-seq data obtained from soybean ZH13 [32], we found that the majority of CENH3 proteins are enriched in regions containing both *CentGm*-1 and *CentGm*-4 sequences across all chromosomes (Additional file 1: Figs. S11–S30). Although CENH3 exhibits binding capability to both types of *CentGm* sequences (Additional file 1: Fig. S36), our analysis revealed a distinct preference in its binding affinity. CENH3 binds more frequently to *CentGm*-4 regions, despite the higher overall abundance of *CentGm*-1 (Additional file 1: Figs. S11–S30). This preferential binding suggests that *CentGm*-4 may serve as a more functionally relevant platform for centromere assembly and kinetochore formation, potentially playing a dominant role in centromere identity and specification.

The occurrence of two types of centromeric repeats in soybean offers an opportunity to explore the conserved sequence features within centromeric regions. To investigate the potential functional relationships between different *CentGm* types, we utilized the top five high-frequency *CentGm*-1 and *CentGm*-4 sequences (HF5-*CentGm*), along with two representative *CentGm* sequences, to search for conserved motifs potentially involved in CENH3 binding (Additional file 1: Fig. S37). This analysis identified four putative CENH3-binding motifs that are conserved in all three soybean varieties studied (Jack, ZH13 and WM82) (Fig. 3c, Additional file 2: Table S14). We propose that CENH3 recognizes and binds to these conserved motifs, enabling the coexistence and evolutionary retention of both *CentGm*—1 and *CentGm*-4 repeats in soybean centromeres.

Furthermore, we observed that the spatial arrangement of these motifs differs between the two *CentGm* types (Fig. 3c), suggesting distinct modes of CENH3 binding. Notably, these motifs represent the first putative CEN-B box sequences identified in any plant species. This discovery provides key insights into the functional architecture of soybean centromeric regions. Future studies on the positional relationship of these motifs

Table 1 Statistical comparison of distances between identical *CentGm*-1 and *CentGm*-4 across different soybean varieties

				Jack-*CentGm*-1			**Jack-*CentGm*-4**			**WM82-*CentGm*-1**			**WM82-*CentGm*-4**			**ZH13-*CentGm*-1**		
	Mean (bp)	**Std**	**Number**	**P value**	**Cohen's d**	**P value (FDR)**	**P value**	**Cohen's d**	**P value (FDR)**	**P value**	**Cohen's d**	**P value (FDR)**	**P value**	**Cohen's d**	**P value (FDR)**	**P value**	**Cohen's d**	**P value (FDR)**
Jack-*CentGm*-1	2,728	2,031	170,355															
Jack-*CentGm*-4	1,634	1,010	2,237	0.005	0.693	0.012												
WM82-*CentGm*-1	2,759	2,054	177,275	0.419	−0.042	0.477	0.003	0.696	0.010									
WM82-*CentGm*-4	1,554	799	3,826	0.002	0.764	0.010	0.446	0.096	0.477	0.003	0.747	0.010						
ZH13-*CentGm*-1	2,730	2,093	194,827	0.412	0.031	0.477	0.005	0.645	0.012	0.477	0.005	0.477	0.003	0.782	0.010			
ZH13-*CentGm*-4	1,614	914	2,695	0.006	0.684	0.012	0.376	0.036	0.477	0.003	0.734	0.010	0.274	−0.084	0.411	0.013	0.649	0.022

Statistical analysis revealed no significant differences in the distances between identical *CentGm*-1 or *CentGm*-4 monomers among the three soybean varieties, whereas comparisons between *CentGm*-1 and *CentGm*-4 monomers revealed highly significant differences across all varieties may provide a deeper understanding of how multiple *CentGm* sequences interact with CENH3 to form and maintain the higher-order centromeric structure.

Novel centromeric DNA replication model

We further discovered that high-frequency *CentGm* subtypes (HF10-*CentGm*) cluster into homogenized arrays of varying lengths and exhibit chromosome-specific evolutionary patterns (Fig. 2). In the yeast *Schizosaccharomyces pombe*, centromeric DNA has been shown to contain a high density of segments, including repetitive regions, that mediate autonomous replication [33]. In mammals, RNAP II actively transcribesα-satellite DNA at kinetochores of mitotic chromosomes, and its inhibition results in chromosome missegregation, such as increased lagging chromosomes [34]. Integrating these insights with our findings, we hypothesize that centromeres may possess a distinct replication mechanism involving a template-containing polymerase (Fig. 4), akin to telomerase [35]. This mechanism could explain the formation of HORs, where HF10-*CentGm* subtypes are concentrated in multiple adjacent, interlaced fragments (Additional file 1: Fig. S33).

Supporting this model, previous studies have identified numerous interstitial telomeric repeats (ITRs) of non-telomeric origin within the centromeric regions of *Solanum* species [36, 37]. These elements may be amplified via an extrachromosomal circular DNA (eccDNA) mechanism mediated by ITRs, potentially giving rise to new centromeric repeats – reinforcing our hypothesis of a template-based replication in centromere evolution. This process likely contributes to variable lengths of *CentGm* regions across chromosomes (Fig. 1e). Furthermore, TEs frequently appear at the junctions between *CentGm*-1 and *CentGm*-4 arrays, potentially introducing sequence variations, disrupting replication templates of HF-*CentGm*, and contributing to *CentGm* diversity among soybean varieties.

Based on the detailed centromere sequence data from the three soybean genomes, we propose a replication model centromeric repetitive regions (Fig. 4b). Previous studies have shown that replication forks originating from euchromatic regions often stall at centromeres due to CENH3-bound nucleosomes [38], implying that replication of centromere regions operates through a distinct mechanism. In our model, conserved sequence motifs identified in both *CentGm*-1 and *CentGm*-4 (Fig. 3c) are bound by CENH3 and serve as templates for a specialized polymerase to replicate the nucleosome-rich centromeric regions. We propose that within these regions, polymerases use templates associated with existing CENH3 complexes to extend HOR arrays, switching templates as needed when sequence changes are encountered. This model aligns with prior observations that centromere replication occurs rapidly and late in S phase [39].

(See figure on next page.)

Fig. 4 Hypothetical replication model of centromeric repetitive regions. **a** Schematic illustrating the hypothesis of how a centromeric sequence template might be obtained from a nucleosome. **b** Hypothetical process of rapid centromere replication involving multiple simultaneously active polymerases that change templates. **c** FISH of *CentGm*-1 and *CentGm*-4 repeats during the transition from interphase chromatin to prophase chromosomes of WM82. Bars = 5 µm. *CentGm*-4 signal labelled red. *CentGm*-1 signal labelled green. From the first to the third row, scattered fluorescence signals around centromeres gradually increase and then decrease, suggesting that extrachromosomal *CentGm* clusters might form progressively during late S phase

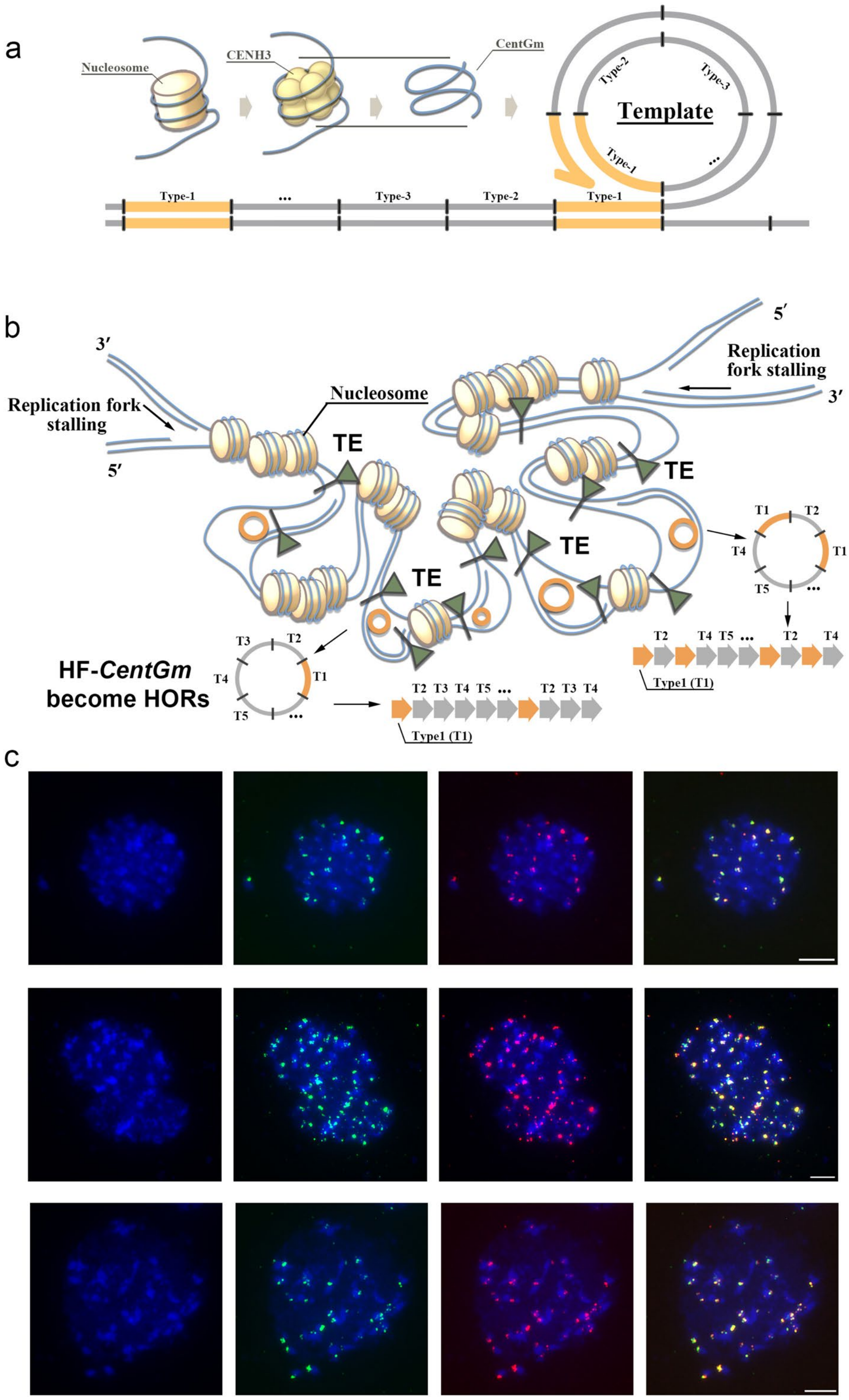

Fig. 4 (See legend on previous page.)

Supporting this, our FISH experiments on soybean interphase cells revealed numerous dispersed *CentGm* fluorescence signals, which likely coalesce into defined centromeric regions during the transition from interphase chromatin to prophase chromosomes (Fig. 4c). Moreover, the chromosome-specific nature of *CentGm*-4, its preferential association with CENH3 (Additional file 1: Fig. S36), and the broader chromosomal distribution of *CentGm*-1 suggest that these repeat types play distinct roles: *CentGm*-4 may contribute to chromosome-specific recognition, whereas *CentGm*-1 may facilitate efficient centromere replication.

Discussion

In this study, we analyzed the centromeric satellite repeats across three soybean varieties and identified two distinct types, *CentGm*-1 and *CentGm*-4, that both bind the centromeric histone CENH3 and co-exist across chromosomes (Fig. 1, Additional file 1: Figs. S11–S30). Despite variation in array length and composition, their organization is largely conserved on corresponding chromosomes, with *CentGm*-1 subtypes being broadly distributed, while *CentGm*-4 subtypes display more chromosome specificity. These observations suggest the two repeat types may serve distinct functional roles within the centromere.

The architectural dynamics of centromeres—characterized by diverse yet functionally conserved repeat sequences—may be driven by a specialized mechanism tailored to maintain centromere stability and replication efficiency. We propose that centromeric DNA replication is rapid and efficient, as the newly synthesized strands remain associated with CENH3-bound nucleosomes, thereby preserving centromere identity throughout cell division. In our model, replication initiates simultaneously at multiple locations, where polymerases utilize CENH3-bound sequence variants as templates to produce interlaced HOR arrays (Fig. 4b). These fragments are subsequently linked together, and TEs may introduce further variation or disrupt replication, contributing to architectural complexity. Alongside template-guided replication, conventional replication and sequence exchange likely co-occur, particularly at the edges of HOR arrays, where diverse *CentGm* subtypes tend to accumulate. These *CentGm* variants may perform complementary functions, supporting both centromere integrity and rapid replication. This model explains the extensive sequence variability of *CentGm* in centromeric regions, where all variants retain CENH3-binding capacity due to their origin from templates stabilized by CENH3-bound nucleosomes. Thus, we propose that centromere function – mediated by CENH3 binding to conserved motifs – preceded and guided the evolutionary accumulation of satellite repeats, establishing a dynamically self-reinforcing centromere architecture.

Conclusions

We have successfully deciphered the complex centromere regions in soybean genomes, uncovering intricate arrays of repetitive sequences interspersed with TEs. The presence of two major *CentGm* types likely reflects soybean's evolutionary history as a paleopolyploid species derived from two unknown ancestors. It is possible that both *CentGm* types were retained due to their roles in maintaining centromere stability. If so, similar patterns of multiple types of centromeric satellite repeats might be expected in other polyploid (particularly allopolyploid) genomes, such as *Fragaria vesca* (strawberry), which contains at least two or more satellite repeat types (Type I, II, III, IV-1, IV-2) with monomer lengths typically falling into two categories: ~ 146 bp and ~ 158 bp [40, 41]. Unlike soybeans, which maintain two completely distinct sets of repeats, the two strawberry types share > 80% identity and > 90% coverage, suggesting functional interchangeability, with some chromosomes containing only one type. In soybean, the co-existence of *CentGm*-1 and *CentGm*-4 enabled comparative analyses that led to the identification of the first putative CENH3-binding sites in a plant genome.

Our further investigation of centromeric repeat organization led to the development of a replication model in which template-containing polymerases simultaneously replicate multiple regions. Based on high-resolution centromere sequence architecture, this novel model offers new insights into centromere diversity and dynamics and provides a valuable framework for future experimental validation as more advanced technologies emerge.

Methods

De novo *discovery of CentGm sequences*

The methods and procedures for identifying *CentGm* sequences and generating statistical charts have been packaged into the unitFinder software. This tool can perform de novo identification by inputting a single chromosome sequence via the -seq parameter, or reference-based identification by supplying potential satellite repeat sequences via the -cen parameter. The detailed steps are as follows:

First, we located the potential centromere regions by finding the tandem repeats sequences [42] (Additional file 1: Fig. S1b, Additional file 2: Table S1). Tandem Repeats Finder was used to find tandem repeat sequences (TRSs) which contain *CentGm* sequences with parameters “2 7 7 80 10 50 500 -f -d -m”. We employed an iterative approach to ensure the detection of shorter tandem repeats within higher-order TRSs (Additional file 1: Fig. S2). Using unitFinder to identify potential satellites without reference, the repeated command is as follows: “python unitFinder.py -seq Chr*.fasta -name Chr*”.

After finding the smallest TRS in potential centromeric regions, we used nucmer with parameter “-c 10 -l 10” to align the TRS to each other, and the TRS with identity > = 50.0 and coverage ≥ 80% will be grouped as the same TRS. We grouped the TRSs based on sequence similarity, and those present across all chromosomes were considered as the potential *CentGm* sequences (Additional file 1: Fig. S2a, Additional file 2: Table S1). Results from each chromosome were merged and potential *CentGm* sequences identified using scripts typeMerge_type_raw.py and getChrTypeFromMerge.py, respectively.

Finally, we reidentified the potential centromere regions by aligning potential *CentGm* sequences to the genome to find TRSs within these regions (Additional file 1: Fig. S2b, Additional file 2: Table S1) and using unitFinder command as follows: “python unitFinder.py -seq Chr*.fasta -name Chr* -cen *.subtype.fa”. The TRSs shared in all chromosomes were considered as the *CentGm* sequences.

Generation of HF-*CentGm* sequences and Representative *CentGm*

Centromeres display a large number of *CentGm* sequences, so we only selected the TRS of high frequency in contiguous regions based on copy number. TRSs shared in all chromosomes were selected by using nucmer with the parameter '-c 5 -l 5' and grouped the TRS with identity ≥ 55.0 and coverage ≥ 70%. Any TRSs, appearing in all chromosomes, were considered as *CentGm* sequences. We reordered the TRS sequences within a group by aligning them with each other using nucmer with the parameter "-c 10 -l 10", so that they can have same start locus and same end locus. To generate the representative *CentGm* sequences, we aligned TRSs within each group used Clustal-Omega (1.2.4-foss-2016b) with the parameter "-v -force" and counted the most common base type for each locus. Two representative *CentGm* sequences were obtained in the three soybean varieties (Additional file 1: Figs. S4 and S5) and utilized to align all *CentGm* monomers from the three soybeans by running nucmer with the parameter "-c 10 -l 10", then the *CentGm* of the highest frequency on each chromosome were identified. The representative *CentGms* were generated using scripts complete-Gm1.unit.py, complete-Gm4.unit.py and unify.py. The script getType.py was used to align all *CentGm* monomers to a common start point based on the representative *CentGm* sequences.

The phylogenetic analysis of the 10 most abundant types was performed using MEGAX5 [43] with the minimum evolution method.

Statistical analysis of distances between identical *CentGm* monomers

Mann–Whitney U test was conducted to compare the distances between identical *CentGm* monomers across different soybean accessions. P-values were adjusted for multiple comparisons using the Benjamini–Hochberg procedure to control the false discovery rate (FDR). Additionally, Cohen's d was calculated to assess the practical significance of the observed effects, with values > 0.5 considered indicative of a meaningful effect in real applications. All statistical analyses were performed with the SciPy library in Python (version 1.11.2).

Motifs of *CentGm* sequences

MEME (version 5.0.5) program [44] was used to find motifs in *CentGm* sequences with parameters "-dna -oc. -nostatus -time 14,400 -mod oops -nmotifs 5 -minw 6 -maxw 50 -objfun classic -revcomp -markov_order 0".

Fluorescence in situ hybridization (FISH) and image analysis

Chromosomal preparation and FISH experiments were performed following the protocol described by Huang et al. [14]. Mitosis prophase chromosomes of Jack were prepared as described by Koo et al. [45]. For probe preparation, sequences of centromeric repeats (*CentGms*) were used to design the PCR primers (Additional file 2: Table S3).

The *CentGm*—4 and 92-bp centromeric probes were labeled with biotin-16-dUTP and digoxigenin-11-dUTP, respectively, using the PCR Kit (PCR DIG Probe DIG Synthesis Kit, 11,636,090,910; Biotin-16-dUTP, 11,093,070,910). The FISH reaction mixture, containing 20 µL of hybridization mixture (50 ng of each labeled probe), was added to the denatured chromosome slides and incubated for 18 h at 37 °C for hybridization. After post-hybridization washes, the digoxigenin- and biotin-labeled probes were detected using digoxigenin antibody (Anti-Digoxigenin-Fluorescein, Fab fragments, 39,516,300) and biotin antibody (Streptavidin-Cy3TM, S6402), respectively.

Chromosomes were counterstained with 4′,6-diamidino-2-phenylindole (DAPI) in Vectashield antifade solution (Vector Laboratories, Burlingame, CA). Slides were examined under the ZEISS Axio Imager M2 fluorescence microscope with the filters (365 EX/445 ± 50 EM for the DAPI blue fluorescence, 500 ± 20 EX/535 ± 30 EM for *CentGm*-1 centromeric probe FISH signals, and 545 ± 25 EX/605 ± 70 EM for *CentGm*-4 centromeric probe FISH signals). Images were captured with a Zeiss Axio Imager M2 microscope (Carl Zeiss Microscopy LLC, Thornwood, NY) using a cooled CCD camera CoolSNAP HQ2 (Photometrics, Tucson, AZ) and AxioVision 4.8 software. The final contrast of the images was processed using Adobe Photoshop 2024 software.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03924-9.

Additional file 1: Figs. S1–S37. Fig. S1. De novo identification of *CentGm* monomers in soybean. Fig. S2. Comparison between potential centromeric regions identified using tandem repeat and *CentGm* monomers. Fig. S3. Total length of *CentGm*-1 and *CentGm*-4 monomers within different chromosomes in three soybean varieties. Fig. S4. Representative *CentGm*-1 generated by aligning similar TRSs from all chromosomes in the three available soybean genomes. Fig. S5. Representative *CentGm*-4 generated by aligning similar TRSs from all chromosomes in the three available soybean genomes. Fig. S6. Alignment of representative *CentGm*-1 from Jack, ZH13 and WM82. Fig. S7. Representative *CentGm*-4 in Jack, ZH13 and WM82. Fig. S8. FISH mapping of *CentGm*-1 and *CentGm*-4 repeats on mitotic metaphase chromosome of soybean. Fig. S9. Visualization of *CentGm* cluster distribution on centromeres using a 2D linear-to-planar conversion. Fig. S10. Distribution of *CentGm*-1 clusters and *CentGm*-4 clusters in the centromeric regions of Jack, ZH13 and WM82. Figs. S11–S30. Soybean CentGm monomers in the centromeric region of Chr01–Chr20. Fig. S31. The distribution of HF10-*CentGm*-1 in three soybean genomes. Fig. S32. The distribution of HF10-*CentGm*-4 in three soybean genomes. Fig. S33. Mosaic-like arrangement of HF-*CentGm* monomer subtypes in Chr01 centromeres. Fig. S34. Shared *CentGm*-1 and *CentGm*-4 monomers in Jack, ZH13 and WM82. Fig. S35. Violin plot showing distances between identical *CentGm*-1 (top) or *CentGm*-4 (bottom) monomers within centromeric regions. Fig. S36. CENH3 ChIP-seq alignment across complete chromosomes of soybean ZH13. Fig. S37. Motifs shared among *CentGm*-1 and *CentGm*-4 monomers that possibly serve as functional binding sites for the centromere-specific histone CENH3.

Additional file 2: Tables S1–S14. Table S1. Location of predicted potential centromeric regions in three soybean varieties: Jack, ZH13 and WM82. Table S2. *CentGm*-1 and *CentGm*-4 monomers identified in centromeres of Jack, ZH13 and WM82. Table S3. PCR primers for centromeric repeat amplification used in the production of FISH probes. Table S4. The amount of various TEs within the centromeric regions of the three soybean varieties Jack, ZH13 and WM82. Table S5. The top 10 high-frequency *CentGm*-1 monomer subtypes (>80 bp) in Jack. Table S6. The top 10 high-frequency *CentGm*-4 monomer subtypes (>300 bp) in Jack. Table S7. The top 10 high-frequency *CentGm*-1 monomer subtypes (>80 bp) in ZH13. Table S8. The top 10 high-frequency *CentGm*-4 monomer subtypes (>300 bp) in ZH13. Table S9. The top 10 high-frequency *CentGm*-1 monomer subtypes (>80 bp) in WM82. Table S10. The top 10 high-frequency *CentGm*-4 monomer subtypes (>300 bp) in WM82. Table S11. Skewed distribution of distances between the same *CentGm*-1 and *CentGm*-4 monomers. Table S12. Distances between the same *CentGm*-1 monomers in the three soybean varieties Jack, ZH13 and WM82. Table S13. Distances between the same *CentGm*-4 monomers in the three soybean varieties Jack, ZH13 and WM82. Table S14. Four motifs shared between *CentGm*-1 and *CentGm*-4 in Jack, ZH13 and WM82.

Acknowledgements

We thank the National Key Laboratory of Crop Genetic Improvement and the Supercomputing Center at Huazhong Agricultural University (HZAU) for providing the bioinformatics computing resources in this study.

Peer review information

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions

Y.H. and J.Z. designed the research; H.Y. performed script programming; H.Y. and E.G. conducted data analysis; S.S. and D.-H.K. performed FISH experiments on WM82 and Jack, respectively; M.S., H.S. and C.C. contributed the reagents, materials, and valuable suggestions; Y.H. and J.Z wrote and edited the paper. All authors read and approved the final manuscript.

Funding

This research was supported by the Biological Breeding—National Science and Technology Major Project (2023ZD04073), the Major Project of Hubei Hongshan Laboratory (2022HSZD031) and HZAU Start-up Fund to J.Z.; HZAU Special Funds for Interdisciplinary Scientific Research (SZYJY2022011) to H.S; National Institute of Food and Agriculture award 2014–33522-22531 to M.S.

Data availability

All datasets described in this study are publicly available. The reference genome assemblies for soybean cultivars Jack, WM82, and ZH13 can be accessed through DDBJ/ENA/GenBank under accession numbers JAGXCU000000000 [14, 46] and GCA_030864155 [9, 47], and through the Genome Warehouse at the National Genomics Data Center (NGDC) under accession number GWHBWDJ00000000 [12, 48], respectively. The raw sequencing data for ZH13 are available under project PRJCA015269 [49] from the Genome Sequence Archive at the NGDC, China National Center for Bioinformation. ZH13 ChIP-seq data were obtained from NGDC under accession numbers CRR638211–CRR638216 [50]. The centromeric repeat identification tool 'unitFinder' was developed as an open-source Python program on the Linux platform and is available for download at Github [51] and Zenodo [52]. The software is released under a permissive MIT license.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]National Key Laboratory of Crop Genetic Improvement, Hubei Hongshan Laboratory, Shenzhen Institute of Nutrition and Health, Huazhong Agricultural University, Wuhan 430070, China. [2]National Key Laboratory for Germplasm Innovation and Utilization for Fruit and Vegetable Horticultural Crops, Huazhong Agricultural University, Wuhan 430070, China. [3]Wheat Genetics Resource Center, Department of Plant Pathology, Kansas State University, Manhattan, KS 66506, USA. [4]School of Plant Sciences, BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA. [5]Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory, Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518000, China. [6]Present Address: Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Shenzhen Key Laboratory of Agricultural Synthetic Biology, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518124, China.

Received: 9 June 2025 Accepted: 17 December 2025
Published online: 05 January 2026

References

1. Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, et al. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Brief Funct Genomics. 2012;11:25–37. https://doi.org/10.1093/bfgp/elr035.
2. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. 2010;95:315–27. https://doi.org/10.1016/j.ygeno.2010.03.001.
3. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS ONE. 2012;7:e47768. https://doi.org/10.1371/journal.pone.0047768.
4. Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M. Improved data analysis for the MinION nanopore sequencer. Nat Methods. 2015;12:351–6. https://doi.org/10.1038/nmeth.3290.
5. Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ. Centromere reference models for human chromosomes X and Y satellite arrays. Genome Res. 2014;24:697–707. https://doi.org/10.1101/gr.159624.113.
6. Cheng Z, Dong F, Langdon T, Ouyang S, Buell CR, Gu M, et al. Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. Plant Cell. 2002;14:1691–704. https://doi.org/10.1105/tpc.003079.
7. Wolfgruber TK, Sharma A, Schneider KL, Albert PS, Koo DH, Shi J, et al. Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons. PLoS Genet. 2009;5:e1000743. https://doi.org/10.1371/journal.pgen.1000743.

8. Tek AL, Kashihara K, Murata M, Nagaki K. Functional centromeres in soybean include two distinct tandem repeats and a retrotransposon. Chromosome Res. 2010;18:337–47. https://doi.org/10.1007/s10577-010-9119-x.
9. Wang L, Zhang M, Li M, Jiang X, Jiao W, Song Q. A telomere-to-telomere gap-free assembly of soybean genome. Mol Plant. 2023;16:1711–4. https://doi.org/10.1016/j.molp.2023.08.012.
10. Garg V, Khan AW, Fengler K, Llaca V, Yuan Y, Vuong TD, et al. Near-gapless genome assemblies of Williams 82 and Lee cultivars for accelerating global soybean research. Plant Genome. 2023;16:e20382. https://doi.org/10.1002/tpg2.20382.
11. Espina MJC, Lovell JT, Jenkins J, Shu S, Sreedasyam A, Jordan BD, et al. Assembly, comparative analysis, and utilization of a single haplotype reference genome for soybean. Plant J. 2024;120:1221–35. https://doi.org/10.1111/tpj.17026.
12. Zhang C, Xie L, Yu H, Wang J, Chen Q, Wang H. The T2T genome assembly of soybean cultivar ZH13 and its epigenetic landscapes. Mol Plant. 2023;16:1715–8. https://doi.org/10.1016/j.molp.2023.10.003.
13. Zhang A, Kong T, Sun B, Qiu S, Guo J, Ruan S, et al. A telomere-to-telomere genome assembly of Zhonghuang 13, a widely-grown soybean variety from the original center of *Glycine max*. Crop J. 2024;12:142–53. https://doi.org/10.1016/j.cj.2023.10.003.
14. Huang Y, Koo DH, Mao Y, Herman EM, Zhang J, Schmidt MA. A complete reference genome for the soybean cv. Jack. Plant Commun. 2024;5:100765. https://doi.org/10.1016/j.xplc.2023.100765.
15. Henikoff S, Ahmad K, Malik HS. The centromere paradox: stable inheritance with rapidly evolving DNA. Science. 2001;293:1098–102. https://doi.org/10.1126/science.1062939.
16. Manuelidis L. Repeating restriction fragments of human DNA. Nucleic Acids Res. 1976;3:3063–76. https://doi.org/10.1093/nar/3.11.3063.
17. Manuelidis L, Wu JC. Homology between human and simian repeated DNA. Nature. 1978;276:92–4. https://doi.org/10.1038/276092a0.
18. Wevrick R, Willard HF. Physical map of the centromeric region of human chromosome 7: relationship between two distinct alpha satellite arrays. Nucleic Acids Res. 1991;19:2295–301. https://doi.org/10.1093/nar/19.9.2295.
19. Alexandrov I, Kazakov A, Tumeneva I, Shepelev V, Yurov Y. Alpha-satellite DNA of primates: old and new families. Chromosoma. 2001;110:253–66. https://doi.org/10.1007/s004120100146.
20. Rudd MK, Wray GA, Willard HF. The evolutionary dynamics of alpha-satellite. Genome Res. 2006;16:88–96. https://doi.org/10.1101/gr.3810906.
21. Alkan C, Cardone MF, Catacchio CR, Antonacci F, O'Brien SJ, Ryder OA, et al. Genome-wide characterization of centromeric satellites from multiple mammalian genomes. Genome Res. 2011;21:137–45. https://doi.org/10.1101/gr.111278.110.
22. McKinley KL, Cheeseman IM. The molecular basis for centromere identity and function. Nat Rev Mol Cell Biol. 2016;17:16–29. https://doi.org/10.1038/nrm.2015.5.
23. Masumoto H, Masukata H, Muro Y, Nozaki N, Okazaki T. A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. J Cell Biol. 1989;109:1963–73. https://doi.org/10.1083/jcb.109.5.1963.
24. Kixmoeller K, Allu PK, Black BE. The centromere comes into focus: from CENP-A nucleosomes to kinetochore connections with the spindle. Open Biol. 2020;10:200051. https://doi.org/10.1098/rsob.200051.
25. Tanaka Y, Nureki O, Kurumizaka H, Fukai S, Kawaguchi S, Ikuta M, et al. Crystal structure of the CENP-B protein-DNA complex: the DNA-binding domains of CENP-B induce kinks in the CENP-B box DNA. EMBO J. 2001;20:6612–8. https://doi.org/10.1093/emboj/20.23.6612.
26. Voullaire LE, Slater HR, Petrovic V, Choo KH. A functional marker centromere with no detectable alpha-satellite, satellite III, or CENP-B protein: activation of a latent centromere? Am J Hum Genet. 1993;52:1153–63.
27. Tyler-Smith C, Gimelli G, Giglio S, Floridia G, Pandya A, Terzoli G, et al. Transmission of a fully functional human neocentromere through three generations. Am J Hum Genet. 1999;64:1440–4. https://doi.org/10.1086/302380.
28. Marshall OJ, Chueh AC, Wong LH, Choo KH. Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution. Am J Hum Genet. 2008;82:261–82. https://doi.org/10.1016/j.ajhg.2007.11.009.
29. Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science. 2009;326:865–7. https://doi.org/10.1126/science.1178158.
30. Piras FM, Nergadze SG, Magnani E, Bertoni L, Attolini C, Khoriauli L, et al. Uncoupling of satellite DNA and centromeric function in the genus *Equus*. PLoS Genet. 2010;6:e1000845. https://doi.org/10.1371/journal.pgen.1000845.
31. Shang WH, Hori T, Toyoda A, Kato J, Popendorf K, Sakakibara Y, et al. Chickens possess centromeres with both extended tandem repeats and short non-tandem-repetitive sequences. Genome Res. 2010;20:1219–28. https://doi.org/10.1101/gr.106245.110.
32. Liu Y, Yi C, Fan C, Liu Q, Liu S, Shen L, et al. Pan-centromere reveals widespread centromere repositioning of soybean genomes. Proc Natl Acad Sci U S A. 2023;120:e2310177120. https://doi.org/10.1073/pnas.2310177120.
33. Smith JG, Caddle MS, Bulboaca GH, Wohlgemuth JG, Baum M, Clarke L, et al. Replication of centromere II of *Schizosaccharomyces pombe*. Mol Cell Biol. 1995;15:5165–72. https://doi.org/10.1128/MCB.15.9.5165.
34. Chan FL, Marshall OJ, Saffery R, Kim BW, Earle E, Choo KH, et al. Active transcription and essential role of RNA polymerase II at the centromere during mitosis. Proc Natl Acad Sci U S A. 2012;109:1979–84. https://doi.org/10.1073/pnas.1108705109.
35. Higa M, Fujita M, Yoshida K. DNA replication origins and fork progression at mammalian telomeres. Genes. 2017. https://doi.org/10.3390/genes8040112.
36. He L, Liu J, Torres GA, Zhang H, Jiang J, Xie C. Interstitial telomeric repeats are enriched in the centromeres of chromosomes in *Solanum* species. Chromosom Res. 2013;21:5–13. https://doi.org/10.1007/s10577-012-9332-x.
37. Barchi L, Pietrella M, Venturini L, Minio A, Toppino L, Acquadro A, et al. A chromosome-anchored eggplant genome sequence reveals key events in Solanaceae evolution. Sci Rep. 2019;9:11769. https://doi.org/10.1038/s41598-019-47985-w.

38. Giunta S, Herve S, White RR, Wilhelm T, Dumont M, Scelfo A, et al. CENP-A chromatin prevents replication stress at centromeres to avoid structural aneuploidy. Proc Natl Acad Sci USA. 2021. https://doi.org/10.1073/pnas.2015634118.
39. Shelby RD, Monier K, Sullivan KF. Chromatin assembly at kinetochores is uncoupled from DNA replication. J Cell Biol. 2000;151:1113–8. https://doi.org/10.1083/jcb.151.5.1113.
40. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, et al. The genome of woodland strawberry (*Fragaria vesca*). Nat Genet. 2011;43:109–16. https://doi.org/10.1038/ng.740.
41. Jin X, Du H, Chen M, Zheng X, He Y, Zhu A. A fully phased octoploid strawberry genome reveals the evolutionary dynamism of centromeric satellites. Genome Biol. 2025;26:17. https://doi.org/10.1186/s13059-025-03482-0.
42. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573–80. https://doi.org/10.1093/nar/27.2.573.
43. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2018;35:1547–9. https://doi.org/10.1093/molbev/msy096.
44. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME suite: tools for motif discovery and searching. Nucleic Acids Res. 2009;37:W202-208. https://doi.org/10.1093/nar/gkp335.
45. Koo DH, Molin WT, Saski CA, Jiang J, Putta K, Jugulam M, et al. Extrachromosomal circular DNA-based amplification and transmission of herbicide resistance in crop weed *Amaranthus palmeri*. Proc Natl Acad Sci USA. 2018;115:3332–7. https://doi.org/10.1073/pnas.1719354115.
46. Huang Y, Koo DH, Mao Y, Herman EM, Zhang J, Schmidt MA. A complete reference genome for the soybean cv. Jack. Plant Commun. 2024. https://doi.org/10.1016/j.xplc.2023.100765.
47. Wang L, Zhang M, Li M, Jiang X, Jiao W, Song Q. NCBI GenBank. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_030864155.1/ (2023).
48. Zhang C, Xie L, Yu H, Wang J, Chen Q, Wang H. National Genomics Data Center. https://ngdc.cncb.ac.cn/search/all?&q=GWHBWDJ00000000 (2023).
49. Zhang C, Xie L, Yu H, Wang J, Chen Q, Wang H. https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA015269 (2023).
50. Liu Y, Yi C, Fan C, Liu Q, Liu S, Shen L, Zhang K, Huang Y, Liu C, Wang Y, et al. https://ngdc.cncb.ac.cn/gsa/browse/CRA009445 (2023).
51. Huang Y, Guan E, Song S, Koo DH, Schmidt MA, Su H, Chen C, Zhang J. Github. https://github.com/HuangYicheng-Bio/unitFinder (2025).
52. Huang Y, Guan E, Song S, Koo DH, Schmidt MA, Su H, et al. 2025. Zenodo. https://doi.org/10.5281/zenodo.17899110.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.