JIPB Journal of Integrative Plant Biology

Research Article
https://doi.org/10.1111/jipb.13607

A centromere map based on super pan-genome highlights the structure and function of rice centromeres[oo]

Yang Lv[1,2†], Congcong Liu[2†], Xiaoxia Li[2†], Yueying Wang[1†], Huiying He[2], Wenchuang He[2], Wu Chen[2], Longbo Yang[2], Xiaofan Dai[2], Xinglan Cao[2], Xiaoman Yu[2], Jiajia Liu[2], Bin Zhang[2], Hua Wei[2], Hong Zhang[2], Hongge Qian[2], Chuanlin Shi[2], Yue Leng[2], Xiangpei Liu[2], Mingliang Guo[2], Xianmeng Wang[2], Zhipeng Zhang[2], Tianyi Wang[2], Bintao Zhang[2], Qiang Xu[2], Yan Cui[2], Qianqian Zhang[2], Qiaoling Yuan[2], Noushin Jahan[1], Jie Ma[1], Xiaoming Zheng[3,4], Yongfeng Zhou[2], Qian Qian[1,2,3*], Longbiao Guo[1*] and Lianguang Shang[2,3*]

1. State Key Laboratory of Rice Biology, China National Rice Research Institute, Hangzhou 310006, China
2. Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China
3. Yazhouwan National Laboratory, No. 8 Huanjin Road, Yazhou District, Sanya City 572024, China
4. National Key Facility for Crop Gene Resources and Genetic Improvement, Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing 100081, China

†These authors contributed equally to this work.

*Correspondence: Qian Qian (qianqian188@hotmail.com); Longbiao Guo (guolongbiao@caas.cn); Lianguang Shang (shanglianguang@caas.cn, Dr. Shang is fully responsible for the distributions of all materials associated with this article)

Yang Lv

Lianguang Shang

ABSTRACT

Rice (*Oryza sativa*) is a significant crop worldwide with a genome shaped by various evolutionary factors. Rice centromeres are crucial for chromosome segregation, and contain some unreported genes. Due to the diverse and complex centromere region, a comprehensive understanding of rice centromere structure and function at the population level is needed. We constructed a high-quality centromere map based on the rice super pan-genome consisting of a 251-accession panel comprising both cultivated and wild species of Asian and African rice. We showed that rice centromeres have diverse satellite repeat CentO, which vary across chromosomes and subpopulations, reflecting their distinct evolutionary patterns. We also revealed that long terminal repeats (LTRs), especially young Gypsy-type LTRs, are abundant in the peripheral CentO-enriched regions and drive rice centromere expansion and evolution. Furthermore, high-quality genome assembly and complete telomere-to-telomere (T2T) reference genome enable us to obtain more centromeric genome information despite mapping and cloning of centromere genes being challenging. We investigated the association between structural variations and gene expression in the rice centromere. A centromere gene, *OsMAB*, which positively regulates rice tiller number, was further confirmed by expression quantitative trait loci, haplotype analysis and clustered regularly interspaced palindromic repeats (CRISPR)/CRISPR-associated protein 9 methods. By revealing the new insights into the evolutionary patterns and biological roles of rice centromeres, our finding will facilitate future research on centromere biology and crop improvement.

Keywords: centromere, super pan-genome, CentO satellite repeat, rice

Lv, Y., Liu, C., Li, X., Wang, Y., He, H., He, W., Chen, W., Yang, L., Dai, X., Cao, X., et al. (2024). A centromere map based on super pan-genome highlights the structure and function of rice centromeres. J. Integr. Plant Biol. **00:** 1–12.

INTRODUCTION

Centromeres are chromosomal structures which are key for chromosome segregation and genetic stability during cell division (Altemose et al., 2022; Zhou et al., 2022). They are highly diverse and complex and constitute part of the "dark matter" of the genome; thus, they pose significant challenges for plant genome research (Naish et al., 2021). Most plant centromeres consist of satellite repeats, transposons, and a few genes, and the high repetitive sequence content hinders the study of their structure, features, and function (Furuyama and Biggins, 2007). Despite these challenges, researchers have not only elucidated the structure and evolution of centromeres but also revealed their role in plant growth and adaptation (Jiang et al., 2003; Nagaki et al., 2004).

Rice (*Oryza sativa*) is a major staple crop and a model system for plant genomics and breeding (Shang et al., 2022). One of the most challenging aspects of rice genome research is characterizing the centromere, which is mainly composed of CentO satellite repeats and the centromere-specific retrotransposon of rice (CRR) (Dong et al., 1998; Ma and Jackson, 2006). These elements form tandem or interspersed arrays in the core centromere region and are essential for rice centromere function and formation (Yan et al., 2006; Jiang, 2013). It is challenging to map and clone centromere genes, and few studies have reported their cloning and function (Nagaki et al., 2004). Therefore, many aspects of rice centromere structure and function remain unknown (Cheng et al., 2002). Pan-genome and high-quality sequencing data obtained using high-throughput sequencing technology can provide better insight into rice centromere regions (Wlodzimierz et al., 2023). This approach can effectively identify and validate the function of genes and the variation in centromere sequence and structure between chromosomes and within species.

Several studies have examined the diversity of centromere regions within species at the population level. For example, a study of wheat 10+ genome accessions revealed the evolutionary diversity and functional conservation of wheat centromeres (Ma et al., 2023); a study of 27 soybean varieties revealed the formation mechanism of new centromeres in soybean genome evolution (Liu et al., 2023); and a study of 66 *Arabidopsis thaliana* accessions revealed significant intra- and inter-species diversity of centromere sequences and showed that satellite repeat homogenization can mediate a rapid cycle of transposon invasion and removal in centromere arrays (Wlodzimierz et al., 2023). However, there have been few population-scale studies based on high-quality assemblies in rice, making it difficult to know the true extent of genomic variation. Notably, studies of rice centromeres have primarily focused on *O. sativa* (*Os*) accessions, with *O. rufipogon* (*Or*), *O. glaberrima* (*Og*), and *O. barthii* (*Ob*) accessions remaining underexplored.

Although rice centromeres have been investigated for a long time, the identification and cloning of centromere genes remain challenging because of the low recombination rate of these regions, which hinders the application of conventional mapping methods (Nagaki et al., 2004). To date, only a few centromere genes have been cloned, such as *OsMRPL15* in CEN 8, and none of them have been shown to regulate agronomic traits (Xie et al., 2023). High-quality genome assembly and complete telomere-to-telomere (T2T) rice reference genome (Shang et al., 2023) enable us to overcome the limitations to obtain the complete genomic information. Specifically, based on super pan-genome (Shang et al., 2022), we constructed a high-quality rice centromere map to analyze the diversity of centromeric elements and the rapid evolution of centromeres. Moreover, we identified and functionally validated a centromere gene, *OsMAB*, that positively regulates rice tiller number.

RESULTS

Construction of a rice centromere map

Based on the super pan-genome (Shang et al., 2022), we obtained 251 mini-core germplasm accessions collected globally, including both cultivated and wild species of Asian rice and African rice (mainly eight *Ob*, 11 *Og*, 26 *Or*, and 202 *Os* accessions). We also obtained the published Nanopore long-read data and Illumina short-read data of these 251 accessions, the transcriptome data of 202 accessions and the high-quality chromosome-level genome sequences of 232 accessions. A high-accuracy and chromosome-level genome is as critical as a centromere map. To this end, we *de novo* assembled 19 chromosome-level African rice assemblies (Table S1) with Nanopore long reads and contig N50 of 15.7–26.3 Mb, and we polished them using NextPolish. The average BUSCO (Benchmarking Universal Single-Copy Orthologs) and mapping rate of Illumina short reads were 97.77% and 99.74%, respectively (Table S1), confirming the continuity and quality of the genome. To construct the pan centromere map, we aligned the CENH3 (centromere-specific histone H3-like protein) binding regions and additional 2 M sequences on both sides from the reference genomes of Nipponbare (Shang et al., 2023), MH63 and ZS97 (Song et al., 2021), and obtained the centromere spanning regions of 251 accessions, with an average length of 5.08 Mb (Table S2). To confirm the quality of the centromere genomes, we used high-quality ONT (Oxford Nanopore Technologies) reads to map each genome and obtained the average coverage depth of each centromere spanning region as 87.59 (Table S2). We also used Illumina short reads (251 accessions) to align the corresponding genomes and obtained the average base accuracy of more than 99.99% for the centromere spanning region (Table S2). These results indicate the high-quality base accuracy and integrity of the centromere regions in the genome sequences.

Using Tandem Repeats Finder (TRF) (Benson, 1999) and Extensive *de novo* TE (transposable element) Annotator (EDTA) (Ou et al., 2019), we identified CentO satellite repeats (Chang et al., 2023; Fu et al., 2023) and centromere-specific long terminal repeats (LTRs) from the assembled genomes

(Benson, 1999; Song et al., 2021) and delimited CentO-enriched regions (CoERs). We used continuous contig spanning CoERs as target samples for the centromere map to avoid gaps and ensure quality, and obtained 2,188 complete CoERs for constructing the rice centromere map (Figure 1A; Table S3), with mean coverage depth of 86.61 (visualization of partial samples in Figure S1) and base accuracy over 99.99% from Illumina reads. Significant differences in the centromere elements were observed between CoERs within the centromere map. We identified a total of 487.14 Mb of CoERs sequences, 2,056,749 CentO repeats and 88,226 LTRs in the centromere map and found that *O. sativa indica* (*Osi*) had the highest percentage of CoERs sequence length (60.30%), CentO copy (57.16%), and LTR number (65.83%) among all subpopulations (Figure S2; Table S4). We found that CoERs varied widely in length, CentO copy and LTR number, and in abundance across different chromosomes and subpopulations (Figures 1B–E, S3, and S4), therefore indicating the diversity and variation of rice centromere elements.

Diversity of centromere repeats in rice genome

We identified the CentO monomer and repeat copies across the centromere map (Figure 2A) to characterize the centromere sequences and found that the most common satellite repeat was composed of 155-bp CentO units, with great variation in abundance among different chromosomes and subpopulations (Figure S5). For example, 155 bp CentO dominates CEN8 in the *Ob*, *Og*, *Or*, *Osi* and *O. sativa japonica* (*Osj*) accessions; 165 bp CentO is the major component in the *Ob*, *Og*, *Or* and *Osj* accessions, while in the *Osi* subpopulation, the 154 bp CentO was most abundant; and CEN 4 varies in CentO repeat length between African and Asian rice. These results suggest there is substantial diversity of CentO repeat sequences among different chromosomes and subpopulations. Moreover, the similarity of CentO repeats differed among different centromeres, being higher within each chromosome (96.55%–98.84%) than between different chromosomes (91.59%–97.69%) (Figure 2B). Population-level analysis of CentO supported the model that

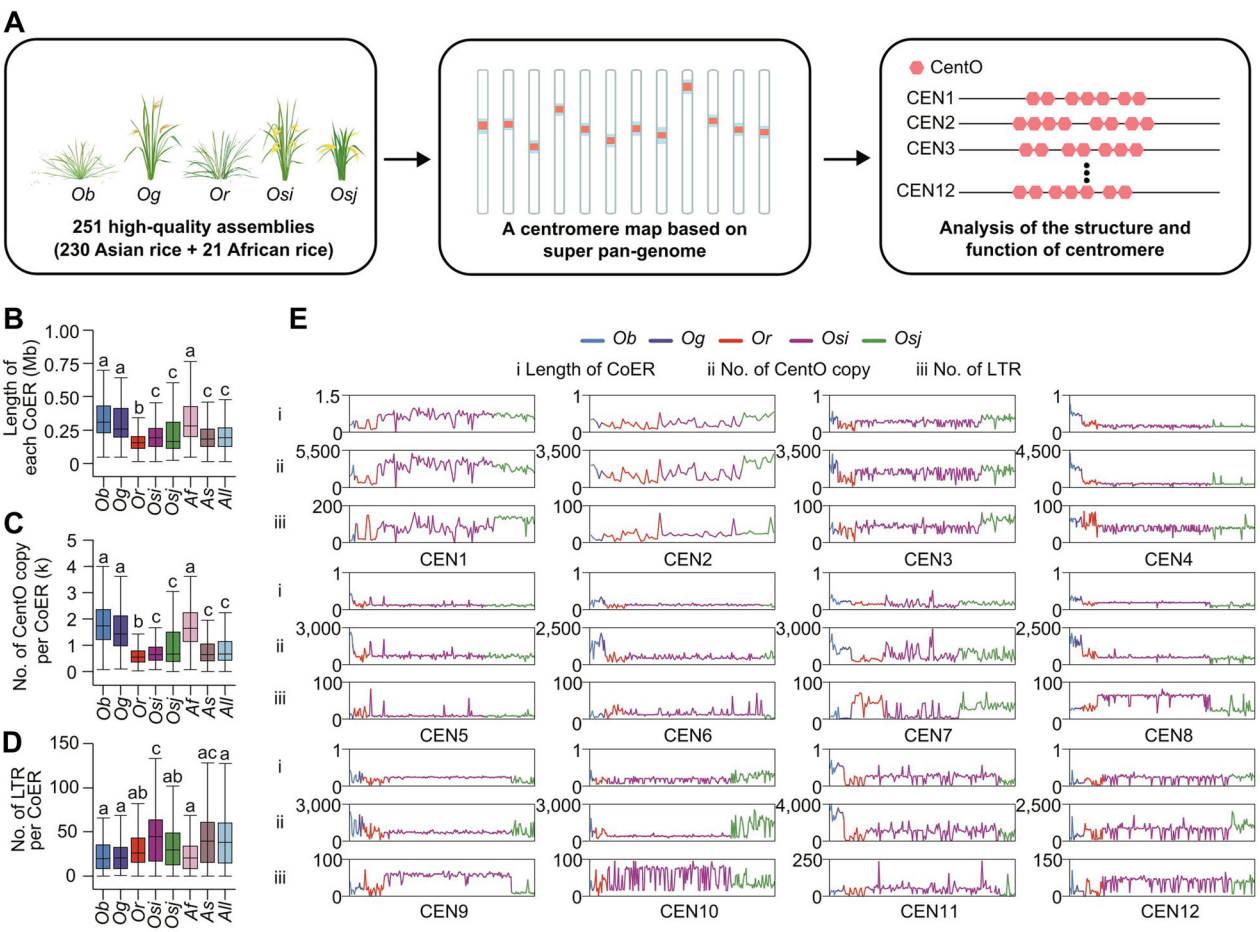

Figure 1. The centromere map of 251 rice mini-core germplasm accessions

(A) Schematic representation of the construction of the rice centromere map. The pink region on the chromosome indicates the CentO-enriched region (CoER), and the blue region on the chromosome indicates the peri-CoER. **(B–D)** Tukey's multiple comparisons test of the CoER length **(B)** and the number of CoER elements with CentO satellite **(C)** and long terminal repeats (LTRs) **(D)** across different subpopulations in the complete CoER dataset of the centromere map. Significance is represented in alphabetical notation ($P < 0.05$). **(E)** Landscape of CoER length (i), CentO repeat copy (ii) and LTR number (iii) in the complete CoERs across different chromosomes. *Ob*, *Og*, *Or*, *Osi*, *Osj*, As, Af and All refer to *Oryza barthii*, *O. glaberrima*, *O. rufipogon*, *O. sativa indica*, *O. sativa japonica*, African rice, Asian rice, and the 251 accessions, respectively.

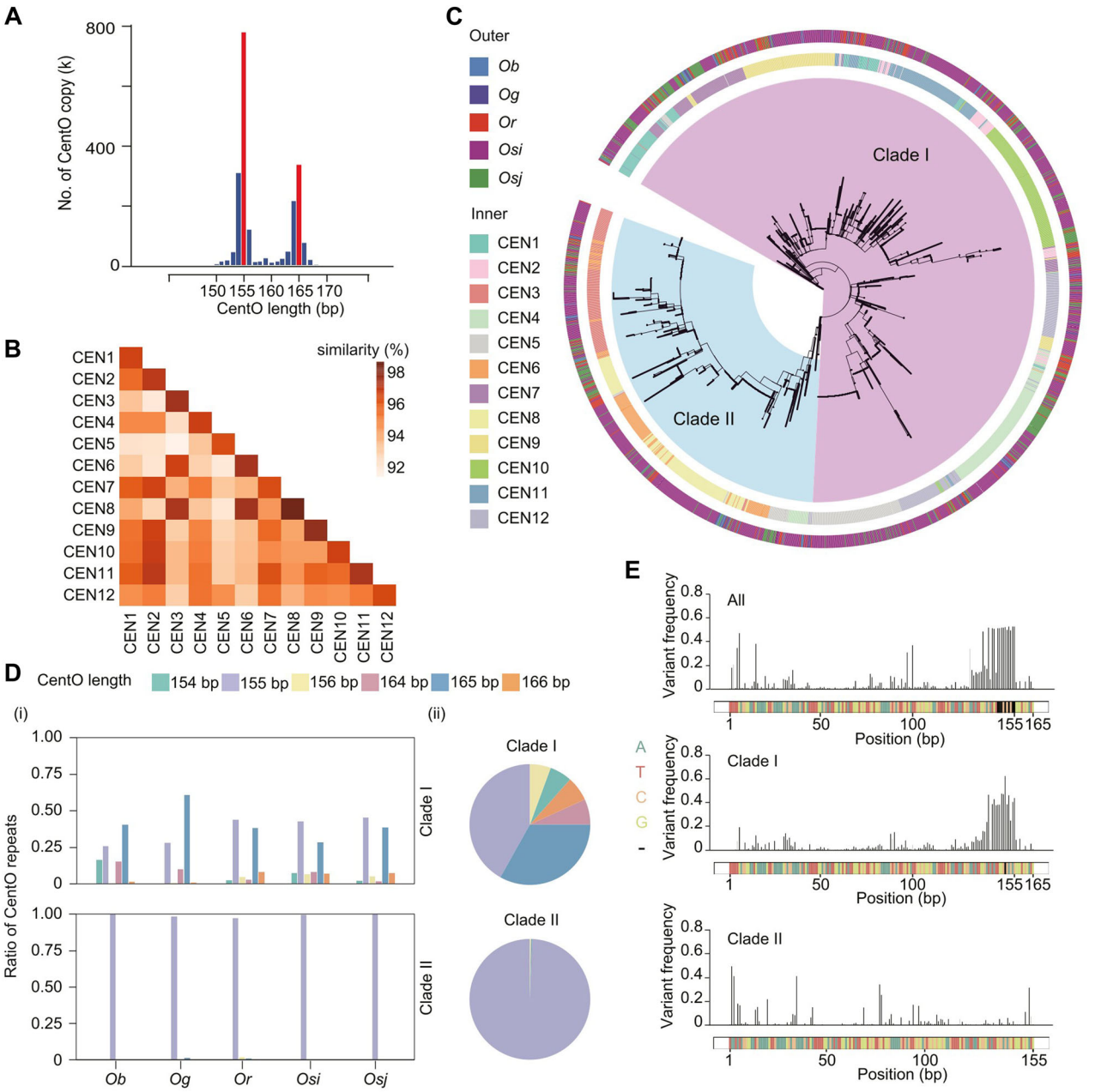

Figure 2. High diversity analysis of CentO repeats in the rice centromere map

(A) CentO repeat copy across all complete CentO-enriched regions (CoERs), with red bar highlighting the 155 and 165 bp CentO satellite repeats. **(B)** Heatmap showing the identity of CentO consensus sequences on each CoER between centromeres. With light red (lowest) to dark red (highest), based on the average similarity ratio (%) of the consensus sequences on each chromosome for each complete CoER. **(C)** Maximum likelihood phylogenetic tree of CentO consensus repeats from 2,188 complete CoERs. Tracks from outer to inner circles indicate: subpopulation distribution, centromere distribution of different chromosomes. **(D)** Comparison of CentO satellite repeats from two clades. (i) Two bar charts showing the distribution of different lengths of CentO repeats in each subpopulation within Clade I and Clade II. The *x*-axis shows the CentO monomer length, and the *y*-axis shows the ratio of CentO repeats with that subpopulation. (ii) Two pie charts showing the ratio of different lengths of CentO repeats within Clade I and Clade II. The CentO repeat lengths category is shown in the legend. Excluded were the CentO repeat lengths that are <10% of the total. **(E)** Frequency of CentO sequence variant sites across the consensus repeats from the CoERs. The color-coded DNA sequence is shown on left (dark green, A; red, T; orange, C; light green, G; -, gap).

the centromere repeats have undergone local homogenization (Cheng et al., 2002; Lamb et al., 2007; Wlodzimierz et al., 2023) and revealed distinct evolutionary patterns of CentO repeats between different chromosomes.

We applied phylogenetic and principal component analyses (PCA) (Figures 2C, S6) to CentO consensus sequence from each centromere to explored their evolution among different chromosomes and subpopulations. The phylogenetic analysis demonstrated that the CentO consensus sequence clustered by chromosome rather than by subpopulation, indicating limited sequence exchange and local homogenization between CentO repeats on different

chromosomes. We further classified CentO consensus sequence from each centromere into two clades (Clade I and Clade II) with distinct CentO repeat lengths and variation patterns (Figure 2D). This indicates that the centromeres in the two clades differ in evolutionary stages, possibly affected by different selective pressures or constraints on CentO repeats among different chromosomes.

To further explore the dynamic variation of the CentO repeat length during centromere evolution, we examined the subpopulation distribution and frequency of CentO variation along the consensus sequences from the two clades (Figure 2E). We found that the CentO consensus sequence in Clade II had multiple sites of variation, but were more conserved in length and sequence on CEN 3, 6, and 8 (Figure S5). This indicates that the CentO repeats on CEN 3, 6, and 8 may have a common origin from a CentO sequence. On the other hand, the CentO repeats in Clade I showed more variation and polymorphism, and the CentO lengths on CEN 2, 4, 5, 7, 10, 11, and 12 were relatively stable in African rice, but changed in Asian rice accessions (including *Or* and *Os*) (Figure S5). Moreover, CentO repeats in Clade I exhibited many polymorphisms, involving the 10 bp repeat at the terminal end of the CentO consensus repeat. We also performed a variation frequency analysis of the CentO consensus sequence in both clades and revealed numerous sites of variation, especially at the beginning (~30 bp) and end (~150 bp) of the CentO consensus repeat (Figure 2E). These sites and segment repeats accounted for the diversity of the CentO sequences. The significant differences in repeat length and sequence between CentO repeats reflect the expansion and variation of rice centromeres across different chromosomes and subpopulations.

Long terminal repeats facilitate centromere amplification and evolution

We annotated the repetitive sequences in each rice genome using EDTA (Ou et al., 2019) and found that LTR-type retrotransposons were overall significantly enriched in the centromere regions (Figures 3A, S7) and that there was a positive correlation between the number of LTRs and the length of the CoERs ($R = 0.67$, $P < 2.2e\text{-}16$) (Figure 3B). Moreover, the number of LTR in the CoERs varied widely across different subpopulations, ranging from one to 127. The coefficient of variation (0.03–1.86) and the *SD* (0.71–61.89) (Figure 3C) reflected the high diversity and divergence of centromere structure and function among different subpopulations.

The distribution of LTRs in various chromosomal regions in the rice genomes showed that CoERs had a significantly lower LTR density than peri-CoERs (Figure 3D, E), which was consistent across different chromosomes and subpopulations (Figure S8). We then estimated the age of centromeric LTR retrotransposons by analyzing the insertion times of intact LTRs and found that the LTR insertion time differed among different chromosomal regions. The CoERs contained much younger LTRs (inserted an average of 0.46 million years ago, MYA) than other chromosomal regions, while the peri-CoERs had relatively older LTRs (inserted an average of 0.82 MYA) (Figure 3F). Gypsy LTRs in the CoERs were more abundant (94.48%) and younger (0.44 MYA) compared with the peri-CoERs (88.18% and 0.81 MYA) and the whole genome (79.51% and 0.82 MYA) (Figures 3G, S9). Therefore, the removal of inserted transposons may maintain the functional structure of the centromere, implying that LTR-type retrotransposons, especially Gypsy LTRs, play an important role in the expansion and evolution of rice centromeres.

OsMAB regulates rice tiller number in CEN12

The mapping and cloning of centromeric genes is challenging, and the constraints of sequencing technologies and assembly strategies make it hard to study them using forward genetics methods. High-quality genome assembly and complete T2T reference genome enable us to overcome the limitations of sequence gaps in the centromeric regions and obtain the complete genomic information (Nagaki et al., 2004). We used the high-quality pan-structural variation (SV) map from our previous study, which was built with the complete T2T Nipponbare genome (Shang et al., 2023) and Nanopore long reads (Shang et al., 2022), to study the biological roles of genes in the centromeres. Most of the genes in centromeres have unknown functions (Nagaki et al., 2004; Wu et al., 2004), and gene expression varies among accessions (Figure S10). To investigate the correlation between SVs and gene expression levels, we performed expression QTL (eQTL) analysis using *Os* accessions from our pan-genome germplasm (Shang et al., 2022). We identified 15 significant cis-eQTL signals in centromere regions (Table S5). We then validated the effect of the candidate signals on gene expression by conducting haplotype analysis using gene expression and agronomic traits (Table S6). Among them, we found a 6,758 bp deletion in the promoter region of *AGI-S_Os12g018490* (*OsMAB*), which is a gene located 38.07 km away from the CoERs on CEN 12 that encodes a matrix attachment region binding protein with unknown function (Figure 4A). The accessions with the deletion variant had higher expression levels of *OsMAB* and more tillers than those without the deletion variant (Figure 4B, C). Furthermore, we determined that the deleted sequence was a DNA TE sequence, suggesting that the deletion of the TE in the centromeric region may affect rice tiller number by altering the expression of *OsMAB*.

To further confirm the biological function of *OsMAB*, we employed clustered regularly interspaced palindromic repeats (CRISPR)/CRISPR-associated protein 9 (Cas9) to knock out *OsMAB* in Nipponbare, obtaining two homozygous knockout (KO) lines: *KO-1* and *KO-2* (Figure 4D). Both lines had a premature stop codon due to a deletion of G or GT in the coding region (Figure 4E, F). These KO lines showed a significant decrease in tiller number compared with the wild type (WT) (Figure 4G). We next measured the expression of other tillering-related genes in both the WT and KO lines. The results revealed that *HTD1*, *HTD2*, *D10*, and *OsSPL14* were downregulated, while *MOC1* and *D53* were upregulated in the

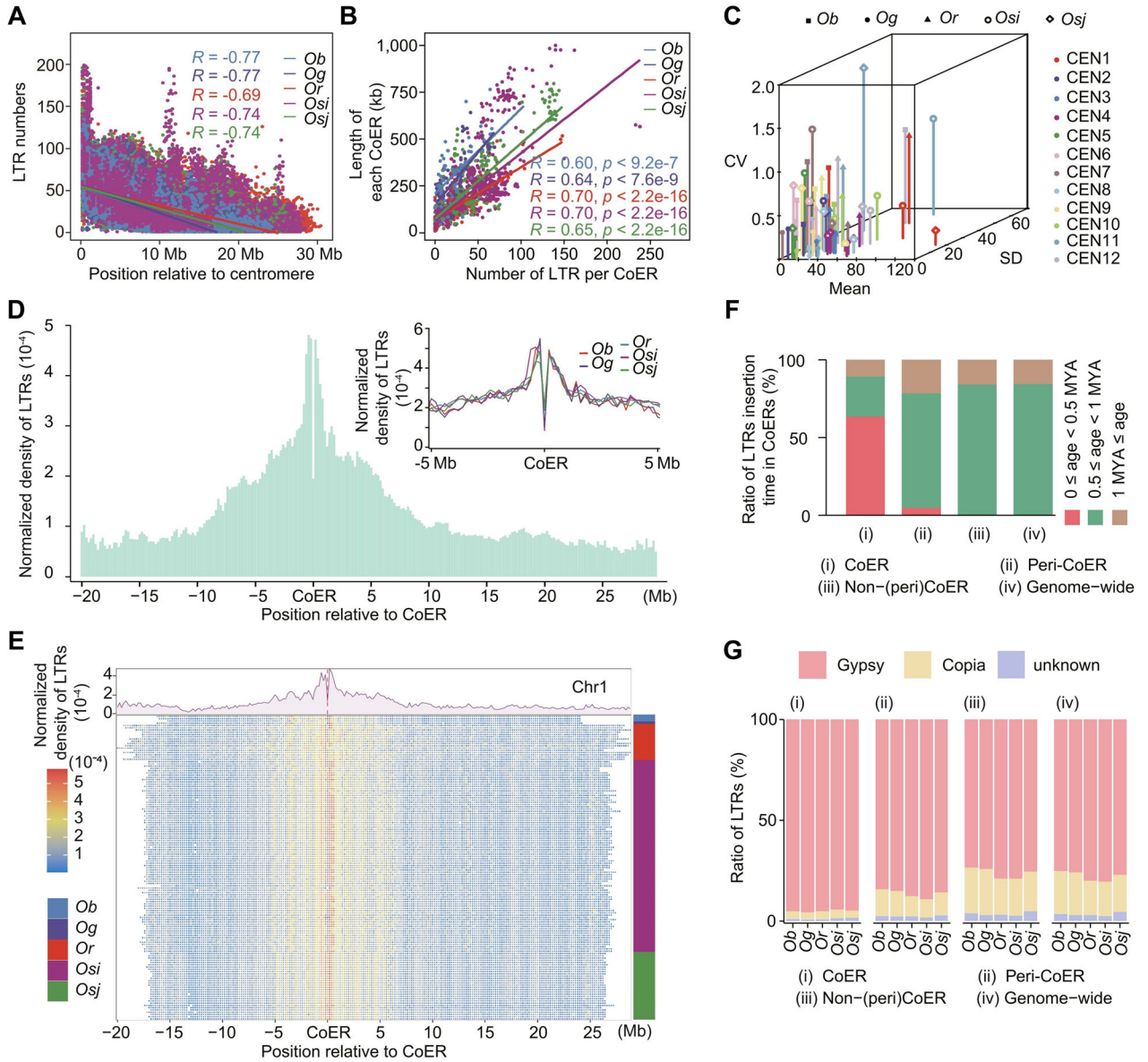

Figure 3. Characterization of long terminal repeats (LTRs) in the rice centromere map

(A, B) The linear regression curves show large correlation between the number of LTRs (*y*-axis) and the distance from the centromere (*x*-axis) (except for some accessions on chromosomes 9 and 11, *P*-values are < 2.2e-16 by Spearman correlation analysis) **(A)**, and the number of LTRs within CentO-enriched regions (CoERs) (*y*-axis) and the CoER length (*x*-axis) **(B)**. The linear regression curves for each subpopulation are shown as *R*-values and *P*-values in the legend. **(C)** Number variation of LTR elements in complete CoERs dataset across different subpopulations and chromosomes. The *x*, *y*, and *z* axes represent the mean, *SD*, and coefficient of variation (CV) of LTR numbers, respectively. **(D)** The genome-wide distribution of LTRs in complete CoERs across different subpopulations. The *x*-axis represents the distance from the CoER origin to the flanking regions (200 kb) on either side. The *y*-axis represents the LTR density, and was calculated by dividing the number of LTRs in the CoER length. The subplot that shows the LTR distribution patterns in the 5 Mb flanking regions of CoER (above the main plot). **(E)** The LTR distribution patterns with a representative example of the CoERs on chromosome 1. Upper panel, density plot of LTR numbers in CoER start sites and genomic regions on chromosome 1. Lower panel, heatmap of LTR number distribution in each CoER start site and genomic regions on chromosome 1. The red dashed line indicates the CoERs. The window size is 200 kb. The heatmap shows the LTR density between different genome regions on chromosome 1 for each CoER, ranging from blue (lowest) to red (highest). The color-coded representation indicates different subpopulations of accessions. **(F, G)** The percentage of LTRs with different insertion times **(F)** and different LTR families **(G)** in different genomic regions across complete CoERs. The regions are defined as the core centromere region (CoER) (i), the 1 Mb flanking region of the CoER on either side (peri-CoER) (ii), the rest of the genome (non-(peri)CoER) (iii), and genome-wide (iv). *Ob*, *Og*, *Or*, *Osi*, and *Osj* refer to *Oryza barthii*, *O. glaberrima*, *O. rufipogon*, *O. sativa indica*, and *O. sativa japonica*, respectively.

KO lines (Figure S11). Lastly, to determine the subcellular localization of OsMAB, we fused its full-length complementary DNA (cDNA) sequence to the N-terminus of green fluorescent protein (GFP) and expressed it in rice protoplasts. We observed that the green fluorescence signal from OsMAB-GFP and the red fluorescence signal from the GS2-RFP, a gene that localizes to the nucleus (Hu et al., 2015), colocalized, suggesting that OsMAB is a nuclear protein (Figure 4H). Together with the

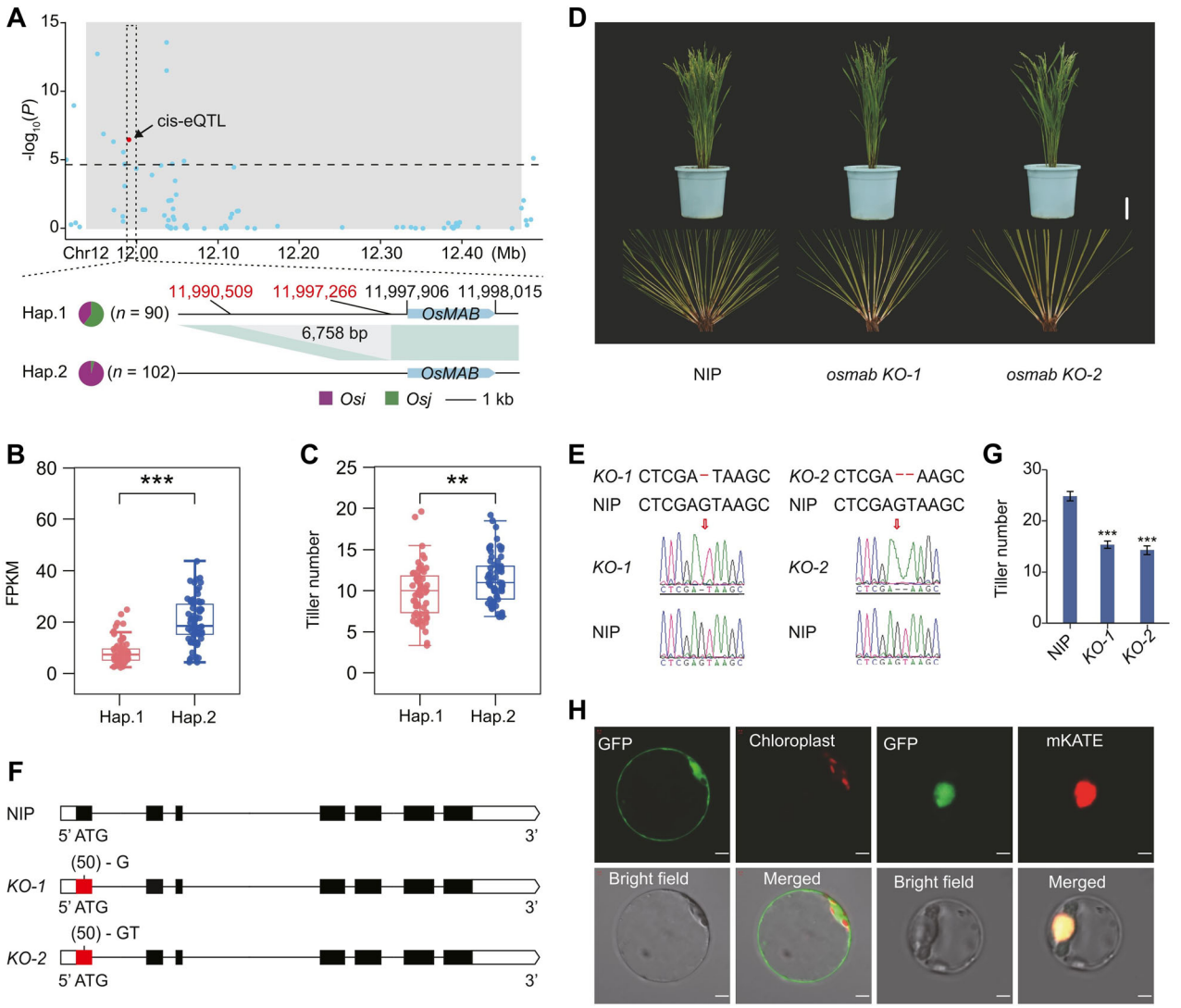

Figure 4. Identification of a centromeric gene *OsMAB* based on structural variation – expression quantitative trait loci (SV-eQTL) analysis

(A) Local Manhattan plot of *OsMAB* expression level and SV. The gray region in the Manhattan plot indicates centromere region, and the centromere position information was obtained from the CENH3 (centromere-specific histone H3-like protein) chromatin immunoprecipitation - sequencing of the telomere-to-telomere (T2T) Nipponbare genome. The black dashed line represents *P*-value threshold (*P*-value = 2.48e-5). **(B, C)** Fragments per kilobase of transcript per million mapped reads (FPKM) **(B)** of *OsMAB* and tiller numbers **(C)** of accessions with (Hap.2) or without (Hap.1) and the deletion in the promoter of *OsMAB*. $^{**}P < 0.01$, $^{***}P < 0.001$ by Wilcox test **(B, C)**. **(D)** Morphology of NIP and *osmab* mutant. Bar, 10 cm. **(E)** Sequencing comparison of mutation sites in the wild type and the *osmab* mutant. **(F)** Gene structure and variation sites within coding sequences (CDs) of the *OsMAB*. **(G)** Tiller numbers of NIP compared with *osmab* mutant ($n = 12$). $^{***}P < 0.001$ by Student's *t*-test. **(H)** Subcellular localization of *OsMAB* (bar, 5 μm).

haplotype analysis results, these results indicated that OsMAB positively regulates rice tiller number and acts in the nucleus.

DISCUSSION

High-quality genomes are essential for exploring the structure and function of centromeres, but the complexity of centromeric regions presents significant challenges for genome sequencing and assembly (Naish et al., 2021). Current research on rice centromeres is mainly limited to a single or few materials. The construction of pan-genomes enabled us to conduct population-level studies on rice centromeres (Ma et al., 2023; Wlodzimierz et al., 2023), and we used the super pan-genome to construct a rice centromere map, which covered 251 mini-core germplasm resources collected globally, including cultivated and wild species of Asian and African rice. The integrity of the core centromeres is essential for studying centromeric elements. Our centromere map consisting of 2,188 complete CoERs, provides us with a valuable strategy to investigate the structure, function, and evolution of centromeres at the population perspective.

Our centromere map enabled us to conduct a comprehensive analysis of centromeric elements. We observed

significant variation in centromere elements across different chromosomes and subpopulations, with differences in CoER length, CentO copy, and LTR number. The distribution of CentO sequences of varying lengths in the centromere map was uneven across different chromosomes and subpopulations, implying different evolutionary rates and patterns of CentO repeats. We also performed phylogenetic analysis and PCA of the CentO consensus repeat sequences in each CoER. We observed that the CentO consensus sequence clustered more strongly by chromosome than by subpopulation. We found that the CentO repeats on CEN 3, 6, and 8 might share a common origin, whereas those on other chromosomes showed more variation and polymorphisms. We found significant differences in CentO repeat length and sequence among different chromosomes and subpopulations, indicating the distinct expansion and variation of centromeres. Furthermore, we detected a significant enrichment of LTR-type retrotransposons in the peri-CoERs. By comparing the insertion times of LTRs in different regions, we noticed that young Gypsy-type LTRs had more frequent insertions into the core centromere region, implying that inserted LTRs promoted the expansion of centromeres and played a crucial role in centromere formation and evolution.

Rice centromeres are a fascinating research area, and genes in the centromere with unknown functions await exploration (Liao et al., 2018). However, centromere gene mapping and functional validation studies are scarce (Xie et al., 2023) because the genes in centromeres are few and difficult to map because of suppression of recombination and high heterogeneity. We used eQTL and haplotype analyses to identify a centromere gene, *AGIS_Os12g018490* (*OsMAB*), with a 6,758 bp deletion in the promoter region that regulates rice tiller number and which we validated by CRISPR/Cas9. The centromere map provides a powerful tool for further understanding the rice centromere that advances centromere biology research and crop improvement.

In summary, we constructed a high-quality centromere map that enabled us to revisit and deepen our understanding of the centromere structure, function, and evolution at the population level. Our results revealed the diversity of the centromeric elements CentO repeats and LTRs across different chromosomes and subpopulations, and their roles in centromere formation and evolution. We also identified a centromere gene that affects rice tiller number by the SV of the rice pan-genome. Our study demonstrated the construction of the rice centromere map to elucidate centromere structure and gene function. Future studies using these datasets may reveal the role of centromeres in rice adaptation and breeding.

MATERIALS AND METHODS

Materials and data collection

We downloaded the high-quality assemblies and TE annotation files of 232 accessions from the Genome Warehouse (https://bigd.big.ac.cn/gwh/) under BioProject PRJCA018476 and Figshare (https://figshare.com/articles/dataset/Annotations_of_250_high_quality_rice_genomes/23803956), respectively. The ONT long reads and Illumina short reads were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive under BioProjects PRJNA656318 and PRJNA692836, respectively. Leaf transcriptome data of 202 *Os* accessions from our previous study (Shang et al., 2022) were downloaded from the NCBI under BioProject PRJNA692672. Field phenotypes were investigated during planting in Shenzhen, Lingshui, and Hainan from 2019 to 2021.

De novo genome assembly and assessment

We *de novo* assembled 19 African rice genomes using Nanopore long reads with quality larger than seven using NextDenovo (Hu et al., 2023) (version 2.4; parameter "genome size = 370 Mb, read cutoff = 50,000, seed cutoff = 56,000, and seed depth = 45"), and raw contigs were polished three times with Nanopore long reads and Illumina short reads using NextPolish (version 1.4.1) (Hu et al., 2020) with default parameters. Illumina short reads were then aligned to each assembly with BWA (version 0.7.17-r1188) (Li and Durbin, 2009). The mutation sites in the assembly of each accession were identified with FreeBayes (version 1.3.1) (Garrison and Marth, 2012), and variants were removed (parameters "QUAL > 20 & DP > 10 & AO > 10"). The homozygous sites were replaced, and false duplications were artificially removed from each assembly using purge Haplotigs (version 1.0.3) (Roach et al., 2018) with parameters "low, middle, and high" read depth cutoff. The various contaminating DNA reads were aligned to the NCBI NR database (downloaded on 4 June 2021) using DIAMOND (version 0.9.24) (Buchfink et al., 2015) with default parameters. Contigs in which more than 50% of the sequences of protein-coding annotations aligned to non-Viridiplantae organisms were considered contaminants and filtered out. All super contigs from *Og* accessions and *Ob* accessions were respectively aligned to the NH265 and NH286 genomes using RaGOO RagTag (version 2.1.0) (Alonge et al., 2022) to construct pseudo-chromosomes, and an average of 99.7% of contigs were anchored to the chromosomes.

To evaluate the quality of each genome assembly, the completeness of assemblies was first evaluated through alignment to the complete T2T Nipponbare genome (Shang et al., 2023) using MUMmer (version 4.0.0, beta with parameters "--mum -t 10 -c 90 -l 40") (Marçais et al., 2018), and alignments were further manually modified. BUSCO (Simão et al., 2015) was used to evaluate each assembly, which presented an average of 97.8% of the 1,440 single-copy Embryophyta genes. We estimated mapping rates by mapping Illumina short reads to the final assembly.

Long terminal repeat annotation and insertion time analysis

LTR elements of the 19 African rice assemblies were predicted by EDTA (version 1.9.6) (Ou et al., 2019) (Table S7). The curated TE library (rice 6.9.5. liban) from the EDTA

package (version 1.9.6) was used to annotate TEs in the 19 whole-genome assemblies of rice with parameters "--overwrite 1 --sensitive 1 --anno 1." Insertion times of LTR retrotransposons were estimated using LTR_retriever (version 2.9.0) (Ou and Jiang, 2018), assuming the same mutation rate, and divergence was estimated between the left and right ends of the LTR sequence of each intact element.

Construction and assessment of centromere map

The CENH3 chromatin immunoprecipitation – sequencing (ChiP-seq) and 2 M flanking sequences of MH63 and ZS97 (Benson, 1999; Song et al., 2021) were aligned to the complete T2T Nipponbare genome (Shang et al., 2023) using MUMmer (version 4.0.0) (Marçais et al., 2018); we merged the alignment results using bedtools (merge -i -d 1000000) (Quinlan and Hall, 2010) and selected the homologous regions with the highest coverage, then defined the boundary start and end points of the centromere spanning regions as the union of the regions obtained from the two alignment results and the CENH3 ChiP-seq and 2 M flanking positions of complete T2T Nipponbare, ensuring complete coverage of the centromere regions. We aligned the centromere spanning region sequences to 251 genomes using mummer (nucmer --mum -l 1000) (Marçais et al., 2018) and filtered the highest-coverage homologous regions. We manually checked and corrected the results, identifying the centromere spanning regions of each chromosome. Using TRF (version 4.10.0, parameters "2 6 6 80 10 50 200 -f -d -m") (Benson, 1999; Chang et al., 2023; Fu et al., 2023), we annotated CentO repeats in 251 rice assemblies, and checked the alignments with pRCS2 sequence (Dong et al., 1998) using blastn (version 2.5.0+, default parameters) (Camacho et al., 2009). Based on the alignment results, the regions (50 kb as continuous window) with the most copies of continuous CentO repeats in each chromosome were classified as CoERs, which we manually verified. In addition, based on the previous study (Table S8), we carried out a one-way analysis of variance ($P = 0.9975$) to further examine the positions of CoERs (Cheng et al., 2002; Song et al., 2021; Shang et al., 2023). A chromosomal CoER was considered a complete CoER when the CoERs were entirely contained within a contig level sequence.

To evaluate the integrity and accuracy of each centromere spanning region and CoERs, we aligned the ONT long reads from 251 accessions (Shang et al., 2022) to each corresponding genome by using minimap2 (version 2.26-r1175, parameters: -ax map-ont) (Li, 2018). Using bedcov of samtools (version 1.5) (Danecek et al., 2021) to count the coverage depth of ONT reads (50 kb window, 10 kb step), and the average, *SD* and median for each centromeric sample were calculated on R4.3.0. Illumina reads from 251 accessions were aligned to their genome by BWA (version 0.7.17-r1188) (Li and Durbin, 2009). SAMtools (version 1.5, parameters "mpileup-ugf") was used to generate BCF (BIM Collaboration Format) files. BCFtools call was used for single nucleotide polymorphism (SNP) calling. Variants were filtered with parameters "QUAL > 20 & DP > 10 & AO > 10" and the homozygous sites were replaced by custom Perl script.

Analysis of satellite repeats and phylogenetic tree construction

Based on the 2,188 complete CoER sequences from the centromere map, CentO repeats of each CoERs sequence were aligned using the R package msa (version 1.30.1, parameters "type = DNA") (Bodenhofer et al., 2015). A consensus matrix was developed using the Biostrings R package (version 2.66.0). The most common nucleotides at each position, with a frequency of at least 50%, were extracted and used to generate a CentO consensus sequence using an R custom script. The CentO consensus sequences were aligned using MAFFT (version 7.520) (Katoh et al., 2002) with default parameters and manual correction, then we computed the frequency of different bases at each site to identify the variant sites in the CentO consensus sequences.

We determined the centromeric repeat similarity by comparing the CentO consensus from each CoER and calculated the percentage of identical nucleotides in the aligned regions using blastn (blastn -outfmt 6 -evalue 1e-5) (Camacho et al., 2009). We calculated the average similarity score for each chromosome comparison from all the pairwise alignments and used the R ggplot package to create a heatmap displaying the average similarity scores for each pair. The maximum likelihood phylogenetic tree based on CentO consensus sequences was built using IQ-TREE (version 2.2.2.3) (Nguyen et al., 2015) with parameter "-m MFP -T AUTO -bb 1000." iTOL (version 6.3.1, https://itol.embl.de/) was used for tree visualization. PCA was performed using the R package adegenet (version 2.1.10) (Jombart and Ahmed, 2011).

Structural variation identification

The high-quality ONT long reads of the 251 rice accessions were aligned to the complete T2T Nipponbare reference genome using minimap2 (version 2.24-r1122; parameters: -ax map-ont -p 0.0001). SVs were called using cuteSV (version 2.0.3, parameters: --max_cluster_bias_INS 100 --diff_ratio_merging_INS 0.3 --max_cluster_bias_DEL 100 --diff_ratio_merging_DEL 0.3 -l 50 -L 1000000 --genotype -S) (Jiang et al., 2020). SVs were removed if they: (1) were smaller than 50 bp; (2) were labeled "IMPRECISE" by cuteSV; or (3) had a read depth <30. Survivor (version, parameters: 1000 1 1 -1 -1 50) was used to merge SVs called separately in each accession.

Expression QTL analysis

Clean RNA-seq reads of leaves were mapped to the complete T2T Nipponbare genome using hisat2 (version 2.2.0) (Kim et al., 2019) with default parameters. Fragments per kilobase of transcript per million mapped reads (FPKM) were calculated using StringTie (version v2.2.1, with parameters -o -p -G -e -A) (Pertea et al., 2016). For eQTL analysis, expression data from 193 accessions were used. To obtain a high-quality normal distribution of expression values for each gene, genes with a mean FPKM value larger than 1 were used in the downstream analysis; 18,497 genes met this condition. The FPKM values of

each gene were normalized using the quantile-quantile normalization function in R (version 3.1.2). SVs with allele frequency ≥0.05 and maximum deletion rate ≤ 0.2 were removed by VCFtools (version 0.1.16) (Danecek et al., 2011). PCA was performed to infer population structure (plink2 -allow-extra-chr -pca 10). The first 20 factors in the R package PEER results and the first five principal components from PCA were used as covariates. The linear regression model of the MatrixEQTL package (version 2.2) (Shabalin, 2012) was used to detect associations between SV-gene pairs. Genome-wide error thresholds were calculated using the Benjamin-Hochberg method ($\alpha = 0.05$), and the corrected *P*-value, $P = 2.48e\text{-}05$, was calculated by the p.adjust software package (version 3.1.2) in R.

Generation of gene KOs

To design the *OsMAB* knockdown target site, we utilized the website developed by Liu Yaoguang's team (http://skl.scau.edu.cn/home/). We constructed the *OsMAB* knockdown vector using the multi-target CRISPR (monocotyledonous plants) method (Ma et al., 2015). The constructed *OsMAB* KO vector was transformed into Nipponbare plants by *Agrobacterium tumefaciens* EHA105-mediated transformation to obtain KO transgenic lines.

Green fluorescent protein

To further study the function of *OsMAB*, an *OsMAB*-GFP vector was constructed. The full-length cDNA of *OsMAB* in Nipponbare was ligated into the pUbi::GFP vector for subcellular localization analysis. The constructed pUbi-OsMAB::GFP vector was transformed into rice protoplasts, which were incubated at 28°C for 14–16 h in the dark and then photographed with a confocal microscope (Zeiss, LSM700) to view the localization of OsMAB in the cells.

Reverse transcription – polymerase chain reaction (RT-PCR) expression analysis

Total RNA was extracted from seedling leaves using a microRNA extraction kit (Axygen) or TRIzol reagent (Invitrogen). RNA was extracted from seedling leaves of *OsMAB*-positive KO plants grown in Lingshui, Hainan in March 2023. Genomic DNA (TaKaRa) in the extracted RNA was removed with DNase, and 1 μg of total RNA was reverse transcribed into cDNA with a reverse transcription kit (Toyobo). RT quantitative PCR analysis was performed using SYBR Green Master reagent and the rice *Actin1* gene as a control. The relative gene expression levels were calculated using a previous method (Livak and Schmittgen, 2001). All experimental samples were subjected to three biological repeats. All primers used for analysis are listed in Table S9.

Availability of data and materials

The 19 African rice genome assemblies present in our study are available from the Genome Warehouse database (https://bigd.big.ac.cn) under accession number PRJCA019158. Scripts for the primary analyses are available under the MIT license at https://github.com/liu2c/centromere.

ACKNOWLEDGEMENTS

We are very grateful to the high-performance computing centers of Agricultural Genomics Institute at Shenzhen, and high-performance computing centers of China National Rice Research Institute for their technical support. This research was supported by the National Natural Science Foundation of China (32188102, 32372148), Innovation Program of Chinese Academy of Agricultural Sciences, the Youth Innovation of Chinese Academy of Agricultural Sciences (Y20230C36), Guangdong Basic and Applied Basic Research Foundation (2023B1515020053), and the Youth Program of Guangdong Basic and Applied Research (2021A1515111123).

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

L.S., Q.Q., and L.G. conceived and supervised the study. X.L. and H.H. assembled and annotated the genomes. C.L., T.W., and Q.X. collected previously released assemblies of rice genomes. C.L. evaluated the quality of rice assemblies. Y.Lv, C.L., and X.L. constructed the rice centromere map and the complete CoERs dataset. H.Q., L.Y., and X.Y. analyzed the haplotypes. Y.W., Bin Z., X.W., Z.Z., and Y.Leng contributed to the experimental design. Y.W., X.C., Y.C., and Y.Lv designed and participated in the experiment. L.S., Y.Z., J.M., X. Z., W.H., W.C., H.W., X.L., C.S., and F.D. discussed the results. Y.Lv, X.L. wrote the manuscript, and H.Z., M.G., Bintao Z., N.J., Q.Z., and Q.Y. revised it. All authors discussed the results and commented on the manuscript.

Edited by: Zhizhong Gong, China Agricultural University, China

Received Dec. 19, 2023; **Accepted** Dec. 28, 2023; **Published** Dec. 30, 2023

OO: OnlineOpen

REFERENCES

Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., Wang, X., Lippman, Z.B., Schatz, M.C., and Soyk, S. (2022). Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol. **23:** 258.

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

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Res. **27:** 573–580.

Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C., and Hochreiter, S. (2015). msa: An R package for multiple sequence alignment. Bioinformatics **31:** 3997–3999.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods **12:** 59–60.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: Architecture and applications. BMC Bioinformatics **10:** 421.

Chang, X., He, X., Li, J., Liu, Z., Pi, R., Luo, X., Wang, R., Hu, X., Lu, S., Zhang, X., et al. (2023). High-quality Gossypium hirsutum and Gossypium barbadense genome assemblies reveal the landscape and evolution of centromeres. Plant Commun. https://doi.org/10.1016/j.xplc.2023.100722

Cheng, Z., Dong, F., Langdon, T., Ouyang, S., Buell, C.R., Gu, M., Blattner, F.R., and Jiang, J. (2002). Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. Plant Cell **14:** 1691–1704.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M. A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics **27:** 2156–2158.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. Gigascience **10:** giab008.

Dong, F., Miller, J.T., Jackson, S.A., Wang, G.L., Ronald, P.C., and Jiang, J. (1998). Rice (Oryza sativa) centromeric regions consist of complex DNA. Proc. Natl. Acad. Sci. U.S.A. **95:** 8135–8140.

Fu, A., Zheng, Y., Guo, J., Grierson, D., Zhao, X., Wen, C., Liu, Y., Li, J., Zhang, X., Yu, Y., et al. (2023). Telomere-to-telomere genome assembly of bitter melon (*Momordica charantia* L. var. *abbreviata* Ser.) reveals fruit development, composition and ripening genetic characteristics. Hortic. Res. **10:** uhac228.

Furuyama, S., and Biggins, S. (2007). Centromere identity is specified by a single centromeric nucleosome in budding yeast. Proc. Natl. Acad. Sci. U.S.A. **104:** 14706–14711.

Garrison, E.P., and Marth, G.T. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv*. https://doi.org/10.48550/arXiv.1207.3907

Hu, J., Fan, J., Sun, Z., and Liu, S. (2020). NextPolish: A fast and efficient genome polishing tool for long-read assembly. Bioinformatics **36:** 2253–2255.

Hu, J., Wang, Y., Fang, Y., Zeng, L., Xu, J., Yu, H., Shi, Z., Pan, J., Zhang, D., Kang, S., et al. (2015). A rare allele of GS2 enhances grain size and grain yield in rice. Mol. Plant **8:** 1455–1465.

Hu, J., Wang, Z., Sun, Z., Hu, B., Ayoola, A.O., Liang, F., Li, J., Sandoval, J.R., Cooper, D.N., Ye, K., et al. (2023). An efficient error correction and accurate assembly tool for noisy long reads. *BioRxiv*. https://doi.org/10.1101/2023.03.09.531669

Jiang, J. (2013). Rice centromeres. In *Plant Centromere Biology*, J. Jiang, and J.A. Birchler, eds, (Iowa: John Wiley & Sons, Inc.), pp. 15–24.

Jiang, J., Birchler, J.A., Parrott, W.A., Dawe, R.K. (2003). A molecular view of plant centromeres. Trends Plant Sci. **8:** 570–575.

Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., Liu, Y., Liu, B., and Wang, Y. (2020). Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. **21:** 189.

Jombart, T., and Ahmed, I. (2011). adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. Bioinformatics **27:** 3070–3071.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. **30:** 3059–3066.

Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. **37:** 907–915.

Lamb, J., Yu, W., Han, F., and Birchler, J. (2007). Plant chromosomes from end to end: Telomeres, heterochromatin and centromeres. Curr. Opin. Plant Biol. **10:** 116–122.

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics **34:** 3094–3100.

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

Liao, Y., Zhang, X., Li, B., Liu, T., Chen, J., Bai, Z., Wang, M., Shi, J., Walling, J.G., Wing, R.A., et al. (2018). Comparison of *Oryza sativa* and *Oryza brachyantha* genomes reveals selection-driven gene escape from the centromeric regions. Plant Cell **30:** 1729–1744.

Liu, Y., Yi, C., Fan, C., Liu, Q., Liu, S., Shen, L., Zhang, K., Huang, Y., Liu, C., Wang, Y., et al. (2023). Pan-centromere reveals widespread centromere repositioning of soybean genomes. Proc. Natl. Acad. Sci. U.S.A. **120:** e2310177120.

Livak, K.J., and Schmittgen, T.D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2–ΔΔCT Method. Methods **25:** 402–408.

Ma, H., Ding, W., Chen, Y., Zhou, J., Chen, W., Lan, C., Mao, H., Li, Q., Yan, W., and Su, H. (2023). Centromere plasticity with evolutionary conservation and divergence uncovered by wheat 10+ genomes. Mol. Biol. Evol. **40:** msad176.

Ma, J., and Jackson, S.A. (2006). Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. Genome Res. **16:** 251–259.

Ma, X., Zhang, Q., Zhu, Q., Liu, W., Chen, Y., Qiu, R., Wang, B., Yang, Z., Li, H., Lin, Y., et al. (2015). A robust CRISPR/Cas9 system for convenient, high-efficiency multiplex genome editing in monocot and dicot plants. Mol. Plant **8:** 1274–1284.

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. PLoS Comput. Biol. **14:** e1005944.

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

Naish, M., Alonge, M., Wlodzimierz, P., Tock, A.J., Abramson, B.W., Lambing, C., Kuo, P., Yelina, N., Hartwick, N., Colt, K., et al. (2021). The genetic and epigentic landscape of the Arabidopsis centromeres. Science **374:** eabi7489.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. **32:** 268–274.

Ou, S., and Jiang, N. (2018). LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. **176:** 1410–1422.

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

Pertea, M., Kim, D., Pertea, G.M., Leek, J.T., and Salzberg, S.L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat. Protoc. **11:** 1650–1667.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics **26:** 841–842.

Roach, M.J., Schmidt, S.A., and Borneman, A.R. (2018). Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics **19:** 460.

Shabalin, A.A. (2012). Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. Bioinformatics **28:** 1353–1358.

Shang, L., He, W., Wang, T., Yang, Y., Xu, Q., Zhao, X., Yang, L., Zhang, H., Li, X., Lv, Y., et al. (2023). A complete assembly of the rice Nipponbare reference genome. Mol. Plant **16:** 1232–1236.

Shang, L., Li, X., He, H., Yuan, Q., Song, Y., Wei, Z., Lin, H., Hu, M., Zhao, F., Zhang, C., et al. (2022). A super pan-genomic landscape of rice. Cell Res. **32:** 878–896.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31:** 3210–3212.

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

Wlodzimierz, P., Rabanal, F.A., Burns, R., Naish, M., Primetis, E., Scott, A., Mandáková, T., Gorringe, N., Tock, A.J., Holland, D., et al. (2023). Cycles of satellite and transposon evolution in Arabidopsis centromeres. Nature **618:** 557–565.

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

Xie, E., Chen, J., Wang, B., Shen, Y., Tang, D., Du, G., Li, Y., and Cheng, Z. (2023). The transcribed centromeric gene OsMRPL15 is essential for pollen development in rice. Plant Physiol. **192:** 1063–1079.

Yan, H., Ito, H., Nobuta, K., Ouyang, S., Jin, W., Tian, S., Lu, C., Venu, R. C., Wang, G.L., Green, P.J., et al. (2006). Genomic and genetic characterization of rice Cen3 reveals extensive transcription and evolutionary implications of a complex centromere. Plant Cell **18:** 2123–2133.

Zhou, J., Liu, Y., Guo, X., Birchler, J.A., Han, F., and Su, H. (2022). Centromeres: From chromosome biology to biotechnology applications and synthetic genomes in plants. Plant Biotechnol. J. **20:** 2051–2063.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article: http://onlinelibrary.wiley.com/doi/10.1111/jipb.13607/suppinfo

Figure S1. Assembly quality assessment using Nanopore long-read coverage depth for complete CentO-enriched regions (CoERs)
Figure S2. Comparison of the total CentO-enriched regions (CoERs) length, CentO repeats copy, and long terminal repeat (LTR) numbers across different subpopulations in the centromere map
Figure S3. CentO-enriched regions (CoERs) lengths across different subpopulations on different chromosomes in the centromere map
Figure S4. Number of CentO copies (A), and long terminal repeats (LTRs) (B) across different subpopulations on different chromosomes in the centromere map
Figure S5. The distribution and diversity of different CentO monomer lengths across different subpopulations on different chromosomes in the centromere map
Figure S6. Principal component analysis of CentO consensus sequence from each CentO-enriched regions (CoERs)
Figure S7. The linear regression of the number of long terminal repeats (LTRs) (*y*-axis) and the distance from the CentO-enriched regions (CoERs) (*x*-axis) across different subpopulations on each chromosome
Figure S8. The genome-wide distribution of long terminal repeats (LTRs) from complete CentO-enriched region (CoER) samples of different chromosomes
Figure S9. The ratio of long terminal repeats (LTRs) and Gypsy LTRs insertion times in different chromosomal regions across different subpopulations from the complete CentO-enriched regions (CoERs) dataset
Figure S10. Expression level of the centromeric genes in Osi and Osj accessions
Figure S11. Expression of tiller-related genes by reverse transcription quantitative polymerase chain reaction (RT-qPCR) in Knockout-1 (KO-1), KO-2, and wild type (WT)
Table S1. Summary of quality assessments from 19 African rice assemblies
Table S2. Location and evaluation of genome sequence quality for centromere spanning regions across 251 rice genomes
Table S3. Number of complete CentO-enriched regions (CoERs) on each chromosome
Table S4. Summary statistics of the rice complete CentO-enriched regions (CoERs) dataset in the centromere map
Table S5. Significant structural variation (SV) expression quantitative trait loci (eQTL) signaling sites were located in the centromeric region
Table S6. Haplotype analysis of the gene expression level and agronomic traits data for significant expression quantitative trait loci (eQTL) signals
Table S7. Transposable elements (TEs) sequence annotation of the 19 accessions
Table S8. Evaluation of strategies for CentO-enriched regions (CoERs) definition
Table S9. List of primers for analysis