# T2T Assembly of Oryza australiensis Reveals Centromere Repositioning Mechanisms and Identifies the EE Genome as Direct Donor of the Extinct DD Lineage

**Tao Zhang**

zhangtao@cib.ac.cn

Chengdu Institute of Biology, Chinese Academy of Sciences https://orcid.org/0000-0002-7897-0205

**Yu Bao**

**Hanli You**

Yangzhou University

**Shuo Liu**

Yangzhou University

**Guanqing Liu**

Yangzhou University

**Yuechao Wu**

Yangzhou University

**Qingqing Yang**

Yangzhou University

**Qi You**

Yangzhou University

**Peng Liu**

Yangzhou University

**Chuandeng Yi**

College of Agriculture, Yangzhou University

**Wenli Zhang**

Nanjing Agricultural University

**Zhukuan Cheng**

Institute of Genetics and Developmental Biology https://orcid.org/0000-0001-8428-8010

**Biological Sciences - Article**

**Keywords:** Genome assembly, O. australiensis, Centromere, Transposon, Evolutionary

**Posted Date:** September 17th, 2025

**DOI:** https://doi.org/10.21203/rs.3.rs-7517070/v1



**Additional Declarations:** There is **NO** Competing Interest.

---

**Version of Record:** A version of this preprint was published at Nature Communications on May 27th, 2026. See the published version at https://doi.org/10.1038/s41467-026-73769-8.

# T2T Assembly of *Oryza australiensis* Reveals Centromere Repositioning Mechanisms and Identifies the EE Genome as Direct Donor of the Extinct DD Lineage

Yu Bao[1†], Hanli You[1†], Shuo Liu[1], Guanqing Liu[1], Yuechao Wu[1,2], Qingqing Yang[1], Qi You[1], Peng Liu[3], Chuandeng Yi[1*], Wenli Zhang[4*], Zhukuan Cheng[1*], Tao Zhang[1,2,3*]

1. Jiangsu Key Laboratory of Crop Genomics and Molecular Breeding/Zhongshan Biological Breeding Laboratory/Jiangsu Co-Innovation Center for Modern Production Technology of Grain Crops, Agricultural College of Yangzhou University, Yangzhou 225009, China
2. Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu 610213, China
3. Joint International Research Laboratory of Agriculture and Agri-Product Safety/The Ministry of Education of China, Yangzhou University, Yangzhou 225009, China.
4. National Key Laboratory for Crop Genetics and Germplasm Enhancement and Utilization, Collaborative Innovation Center for Modern Crop Production Co-Sponsored By Province and Ministry (CIC-MCP), Nanjing Agricultural University, Nanjing, 210095, China

†These authors contributed equally to this work.
*Corresponding author(s)
E-mail: zhangtao@yzu.edu.cn (T.Z.); zkcheng@genetics.ac.cn (ZK.C.); wzhang25@njau.edu.cn (WL.Z.); cdyi@yzu.edu.cn (CD.Y.)

## Abstract

*Oryza australiensis* (EE genome), as the sole representative of the EE genome in the genus *Oryza*, holds significant research value due to its extreme environmental adaptability, unique grain pigmentation and starch properties, and making it a critical resource for deciphering polyploid evolution and mining stress-resistance genes for crop breeding. Here, a telomere-to-telomere (T2T) assembly of *O. australiensis* has been generated, yielding a 901 Mb chromosome-scale genome with 12 complete chromosomes (BUSCO:99.2%; LAI:22.67). This enabled comprehensive annotation of centromeres, telomeres, and 36,753 genes, revealing non-canonical telomeric repeats and three distinct centromere repositioning mechanisms in *Oryza*: inversion-driven shifts, segmental duplication-mediated neocentromeres, and transposon burst-induced

neocentromere formation to the EE genome. Crucially, transposon dynamics revealed synchronized amplification of *Angela* LTR retrotransposons in EE genome and the DD subgenome of allotetraploid *O. alta* (CCDD) prior to 1.7 million years ago (Mya). Phylogenetic and synteny analyses of conserved *Angela* LTR integrase domains, alongside a unique 275-bp repeat marker exclusive to EE and DD, demonstrate that the pre-1.7 Mya EE lineage served as the direct donor of the extinct DD genome. This resolves a major controversy in *Oryza* evolution and establishes transposon-centric approaches for reconstructing paleogenomic histories.

**Key word**: Genome assembly; *O. australiensis*; Centromere; Transposon; Evolutionary

## Introduction

The genus *Oryza* comprises 27 species, including 25 wild rice species and two independently domesticated cultivated species[1-8]. Based on the molecular markers and subsequent cytogenetic analysis, the genus *Oryza* is classified into 11 distinct genome types, including six diploids (AA, BB, CC, EE, FF, GG) and five allotetraploids (BBCC, CCDD, HHJJ, HHKK, KKLL)[9]. Among these, *O. australiensis*, a unique EE genome clade in the *Oryza*, originated from Australia and is mainly distributed in the tropical north of Australia. *O. australiensis* is the closest extant relative to the extinct species in the DD genome[8,10]. The CCDD genome allotetraploid rice originated from a single hybridization event, where the CC genome species (*O. officinalis* or *O. rhizomatis*) served as the maternal parent and the extinct species with the DD genome type served as the paternal donor[5,7,10,11]. Due to the extinction of the diploid DD genome, researchers commonly use the DD subgenome of CCDD to study the evolutionary relationship between EE and DD in *Oryza*[6,10,12]. In 2010, Ammiraju et al. studied the evolution of *Oryza* by cloning the *Adh1* fragment of *Oryza* genomes and revealed that EE genome and DD subgenomes (from *Oryza alta*) divergence at 7.08 Mya[6]. Since this divergence time significantly predates the formation of CCDD allotetraploids (~1.6 Mya) and no marked difference in evolutionary rates was observed between DD and EE lineages, Ammiraju et al. excluded EE genome as the direct donor of the DD subgenome.[5,6]. Recently, Long et al. assembled 13 wild rice genomes and revealed

through single-copy genes of the rice genus that the divergence time of EE genome and DD subgenomes at 4.44 Mya[13]. The substantial discrepancy in divergence times between studies is perplexing and likely stems from methodological limitations: single-copy genes are prone to biases caused by post-polyploidization evolutionary rate heterogeneity[14,15], and short sequence fragments (e.g., *Adh1*) inadequately capture deep divergence events. Consequently, evidence independent of single-copy genes, such as transposon dynamics and chromosomal structural evolution, is urgently needed to reassess the EE-DD relationship.

To date, the chromosome-level reference genome of *O. alta* has been released[16]. However, high-quality *O. australiensis* genome sequence and their annotations are still lacking, which seriously hinders our in-depth analysis of the evolutionary relationship between EE and DD. More than a decade ago, the genome of *O. australiensis* was estimated to be 965 Mb, which is more than twice the size of that of *O. sativa*[17]. The large size of the genome was caused by the expansion of long terminal repeats (LTR), with three retrotransposon elements (RIRE1, Kangourou, and Wallabi) accounting for about 60% of the genome[18]. Recently, Phillips et al. combined Oxford Nanopore long reads and Illumina short reads to complete a ~858 Mb genome assembly with 850 contigs[19]. Long et al. and Fornasiero, A. et al. completed an 881 Mb and 903.3 Mb *O. australiensis* genome using PacBio reads, respectively[8,13]. However, despite these advancements, chromosome level T2T genome assembly, centromere and telomere annotations, and in-depth investigations into genome architecture and repetitive sequence dynamics remain lacking. These critical gaps severely hinder the elucidation of structural evolution and donor relationships between the EE and DD genomes.

Here, we integrated Oxford Nanopore sequencing (ONT) PromethION long reads, Ultra-long reads and Illumina reads from *O. australiensis* to assemble a high-quality genome spanning approximately 901 Mb into 12 complete telomere-to-telomere chromosomes. Then, we have completed the systematic annotation of the *O. australiensis* genome, successfully deciphering key genomic elements including gene structures, repetitive sequences, telomeres, and centromeres. Our study revealed the presence of non-canonical telomeric repeat sequences in its telomere regions and

identified widespread centromere repositioning events across its chromosomes. In addition, in view of the long-standing controversy over the evolutionary relationship between EE genome and DD subgenomes in *Oryza*, we conducted comprehensive comparisons of transposon insertion timing and sequence characteristics between EE genome and DD subgenomes. Through these comparisons, we clarified the divergence time of EE and DD and determined that the pre-divergence EE genome served as the direct donor of DD subgenome in the context of CCDD. In general, this study completed the complete genome assembly and structural analysis of *O. australiensis*, clarified the evolutionary history of EE and DD, and resolved the long-debated evolutionary relationship between the EE and DD genomes of *Oryza*.

# Result

### Telomere-to-telomere genome assembly of *O. australiensis*

In this study, the *O. australiensis* genome was *de novo* assembled using 100× ONT PromethION long reads sequencing, 20× Ultra-long reads sequencing, and 90× Illumina short reads sequencing (Supplementary Table 1). For the raw datasets, the N50 length of the PromethION and Ultra-long reads was greater than 33 Kb and 100 Kb (Supplementary Figure 1), respectively. The PromethION and Ultra-long reads were initially assembled by Nextdenovo software into 16 contigs with a total length of 899.3 Mb. Then, the duplicated contigs, organelle contigs, and contaminants were removed. Further assembly step involved telomere patched and gap filling by examining the coverage relationship between ONT reads and the assembled contigs. Subsequently, Arang Rhie's method was employed to polish the assembled genome[20] (Supplementary Figure 2). The final assembly genome size was ~901 Mb and consisted of 12 telomere-to-telomere(T2T) chromosomes, with scaffolds N50 size of 76.6 Mb (Fig. 1a; Table 1).

The quality of the T2T genome was evaluated through mapping Illumina and ONT sequencing reads to the *O. australiensis* genome. The high reliability of the genome assembly was supported by high genome coverage rates (99.12% Illumina, 99.97% ONT) and mapping rates (98.62% Illumina, 99.79% ONT), as well as rigorous validation through Benchmarking Universal Single-Copy Orthologs (BUSCOs) values of 99.2%, LTR assembly index (LAI) values of 22.67, and consensus quality values

(QV) of 35.65 (Table1; Supplementary Table 1).

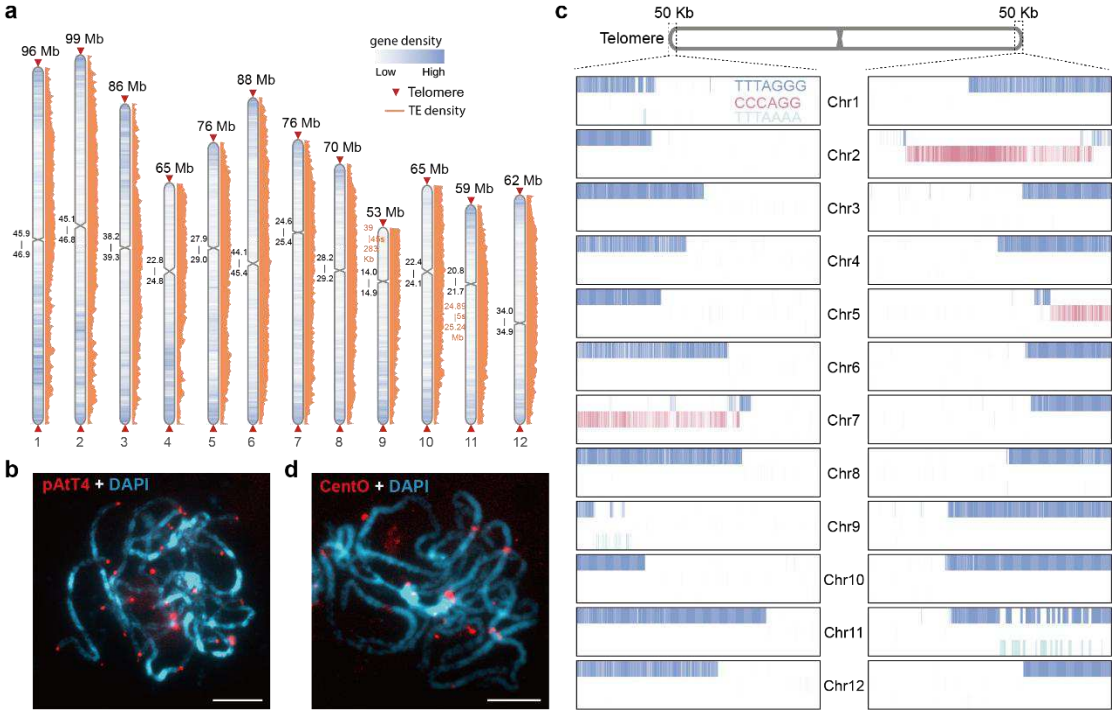


**Fig. 1 Genomic characteristics of *O. australiensis.*** **a,** Chromosome plots of the *O. australiensis* genome. The black numerical intervals represent the centromere regions, measured in megabases (Mb). The pink numerical intervals represent the 45S and 5S rDNA regions. **b,d,** FISH with the telomeric probe pAtT4 (TTTAGGG) (**b**) and CentO probe (**d**) was performed on chromosomes of *O. australiensis*. Chromosomes were stained with DAPI (blue). Scale bar: 5 μm. **c,** Typical 7-nucleotide (TTTAGGG) telomere repeats and nucleotide variants (CCCAGG; TTTAAAA) are distributed in the 50 Kb region at the end of chromosomes.

**Table 1 Statistics of Oryza australiensis genome assembly and annotation**

| Genome feature | Statistic |
|---|---|
| **Genome size (Mb)** | 901 |
| **Number of contigs** | 12 |
| **Number of telomeres** | 24 |
| **Masked repeat sequence length (Mb)** | 611.6 |
| **BUSCO genes (%) (genome)** | 99.2 |
| **LTR assembly index (LAI)** | 22.53 |
| **Consensus quality value (QV)** | 35.65 |
| **G + C (%)** | 45.02 |
| **Number of genes** | 36753 |

## Annotation of the *O australiensis* genome assembly

In *O. australiensis* genome, a total of 36,753 protein-coding genes has been annotated, using *ab initio*, protein homology, and RNA-seq assembly prediction methods (Supplementary Table 2). These genes are predominantly located in distal

regions of the chromosome arms, with a total length of 110.4 Mb, occupying 12.2% of the genome (Fig. 1a; Supplementary Figure 3; Supplementary Table 2). The average lengths of these protein coding genes and corresponding proteins are 3004 bp and 362 aa, respectively.

The *O. australiensis* genome is predominantly composed of repetitive sequences. A total of 611.6 Mb of the sequence has been annotated as repetitive, which represents 67.8% of the *O. australiensis* genome. The long terminal repeat retrotransposon (LTR-RTs) including *Copia*, *Gypsy*, and unclassified LTR are the dominant components of repeats, making up 91.5% of all repeats and 62.1% of the assembled genome; non-LTR-RTs account for less than 0.25% of the genome (Supplementary Table 3). The length of DNA transposons in *O. australiensis* genome is only 55.7 Mbp, occupying 6.1% of the genome (Supplementary Table 3). Among DNA transposons, *En-Spm* emerged as the predominant type, representing 5.73% of all repetitive elements and 4.5% of the assembled genome.

Eukaryotic telomeres represent a specialized class of microsatellite sequences characterized by 4-10 bp core motifs. The TTTAGGG telomeric repeat predominates in flowering plants[21]. *O. sativa* telomeres comprise 5.1-10.8 kb repetitive regions containing canonical arrays of TTAGGG monomers [22,23]. The telomere length of *O. australiensis* was ranging from 9-42 kb (Supplementary Table 4). Unlike *O. sativa*, the telomeric region of the *O. australiensis* genome exhibits both a canonical array composed of TTTAGGG monomers (Fig. 1b) and a non-canonical array composed of derivative (CCCAGG and TTTAAAA) monomers (Fig. 1c). Their non-canonical array is present in multiple chromosome-ends, including Chr2L (long arm of chromosome 2), Chr5L, Chr7S (short arm of chromosome 7), and Chr1S, Ch9S, Chr11L (Fig. 1c). These telomere derivative sequences also occur in *Chenopodium*[24]. These results indicate the diversity of telomere sequences in the genus of *Oryza* and the universality of the non-canonical arrays of telomere sequences.

Satellite repeats, also known as highly repetitive tandem repeats, are the primary components of heterochromatin in most higher eukaryotes, and they are often organized as long arrays in the pericentromeric, centromeric, and subtelomeric regions of

chromosomes[25,26]. The rice genome has been identified to contain numerous satellites repeat sequences, including 45S rDNA, 5S rDNA, and Os48 (a subtelomeric repeat)[27,28]. Similar to japonica rice, 45S rDNA and 5S rDNA of the *O. australiensis* genome are located at the end of Chr9S and the precentromeric region of Chr11 (Fig. 1a), respectively[29]. However, unlike *O. sativa*, the *O. australiensis* genome does not contain any Os48 repeats.

**Centromere Mapping Reveals Repositioning and Transposon-Driven Neocentromere Formation in *O. australiensis***

The centromere is a constricted region on a chromosome that plays a critical role in properly segregating chromosomes during cell division. In plants, functional centromeres are epigenetically defined by the localization of the histone H3 variant CENH3[30]. To investigate the centromere landscape within the *O. australiensis* genome, we used the rice CENH3 antibody for chromatin immunoprecipitation and sequencing (ChIP-seq) of the captured DNA fragments. The CENH3 binding sites precisely demarcated functional centromere boundaries on each chromosome within the EE genome. High-quality centromere assembly was supported by a 98.96% alignment rate of CENH3 ChIP-seq reads, absence of Oxford Nanopore (ONT) sequence breakpoints in centromeric regions, and consistent assembly results across multiple software tools (Supplementary Figure 4-5; Supplementary Table 1). Centromeres in the EE genome ranged from 0.8 to 2 Mb in size, comparable to those in *O. sativa* (Fig. 2a; Supplementary Table 5). Tandem arrays of highly repetitive satellite sequences make up most of the plant and animal centromeres[28]. In *O. sativa*, the tandem repeats satellite sequence of 155 bp (CentO) is found in each centromere region [28,31]. In EE genome, we identified and experimentally validated via Fluorescence in situ hybridization (FISH) a 155 bp satellite sequence within centromeric regions, which exhibits a high sequence identity of 87.09% to CentO (Fig. 1d; Fig. 2b). To explore CentO sequence divergence across *Oryza* species, we identified centromeric satellite repeats using RepeatExplorer[32]. By RepeatExplorer software, we obtained the centromeric satellite sequences of 155 bp of BB genome, 366 bp and 126 bp of CC genome, 155 bp and 125 bp of CCDD genome, and 154 bp of FF genome, respectively (Supplementary Table 6). The 155 bp and 125

bp satellite repeats of the CCDD genome were most similar to 155 bp satellite repeats of the EE genome (similarity: 94.2%) and 126 bp satellite repeats of the CC genome (similarity: 92.9%), respectively (Fig. 2b-2c). The high similarity of satellite sequences suggests that the DD subgenome in CCDD may have directly inherited the centromeric characteristics of the EE-like ancestor, supporting the common ancestral hypothesis.

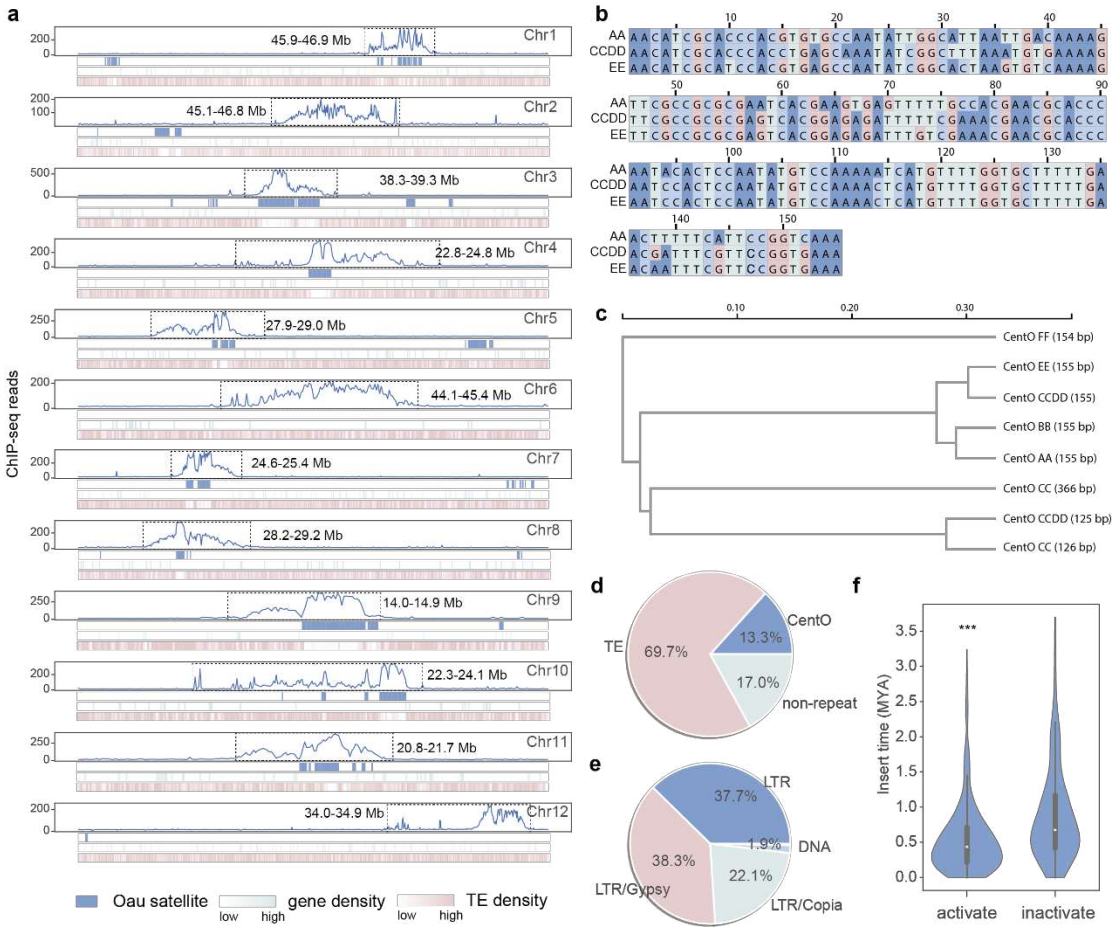


**Fig. 2 Location and evolution analyses of *O. australiensis* centromeres. a,** The delimiting of *O. australiensis* centromeres. The layers of each chromosome graph indicate (1) the density of read mapping from CENH3 ChIP-seq (2) the CentO satellite distribution; (3) gene distribution; and (4) TE distribution, respectively. The number interval represents the defined centromere region. **b,** Multiple alignment of 155 bp centromere satellite sequences of EE, CCDD and AA. **c,** Phylogenetic tree of centromere satellite repeats of *Oryza* genomes. **d-e,** The proportion of repeat sequences in functional centromere region (**d**) and the classification of transposon in functional centromere region (**e**). **f,** Preference of CENH3 to bind younger centromeric repeats. *** represents a statistically significant difference.

Unlike *O. sativa*, the EE genome exhibits centromere repositioning and neocentromere formation on multiple chromosomes. The neocentromere is present in Chr2, Chr6, and Chr12, where Chr2 has a large number of satellite repeats at a distance

of 1.5 Mb from the functional centromere, Chr12 contains fewer satellite repeats, while Chr6 contains almost no satellite repeats (Fig. 2a; Supplementary Table 7). These patterns may represent sequential stages of neocentromere evolution. Additionally, functional neocentromere formation also occurs on chromosomes 1, 3, 5, 7, 8, and 10, where CENH3-binding assays revealed that large structural domains containing two distinct satellite repeat clusters serve as novel functional centromeric regions. Recent studies have indicated that transposon insertions may initiate centromere repositioning events, leading to the formation of neocentromere[33]. In the EE genome, approximately 69.7% of centromeric sequences are composed of TEs (Fig. 2d). Among the TEs found in all centromeres, 98% are long terminal repeat (LTR) elements, with 38.3% belonging to the *Gypsy* family and 22.1% to the *Copia* family (Fig. 2e). We further compared the insertion times of LTRs in active and inactive centromeres, and the results showed that active centromeres contain younger LTRs (Fig. 2f), suggesting that TE insertions may contribute to the formation of neocentromere.

Studies have shown that plant centromere regions contain transcriptional genes[30,34,35]. There were 81 genes in the centromere region of the EE genome, 34 of which had transcriptional activity (Supplementary Table 8). Further studies showed that many of these expressed genes were conserved, and all expressed genes were located at non-CENH3 binding sites.

**Centromere Repositioning Reshapes Chromosomal Arm Morphology in *Oryza* Species**

Chromosomes are divided into short arms and long arms by centromeres, and their structural features provide critical insights for gene mapping, evolutionary studies, and functional genomics[36-38]. To compare the chromosomal arm distribution patterns between *O. australiensis* and Nipponbare, we oriented the chromosomes of *O. australiensis* based on the syntenic alignment of homologous genes with Nipponbare, ensuring consistent chromosomal orientation prior to analyzing the relative lengths of short-arms and long-arms. Comparative analysis revealed conserved long-short arm ratios (>1 for most chromosomes) but species-specific divergences (Fig. 3a). For example, Chr4 in Nipponbare showed a long-short arm ratio of 2.73, compared to 1.73

in *O. australiensis*; Chr7 in Nipponbare had a long-short arm ratio of 1.44 versus 2.07 in *O. australiensi* (Fig. 3a). Strikingly, Chr3 and Chr12 exhibited reversed long-short arm trends between species (Fig. 3a). In Nipponbare, Chr3 had a long-short arm ratio <1 (short-arm longer than long-arm), whereas *O. australiensis*, displayed long-short arm ratio = 1.22 (long-short arm ratio >1); Chr12 showed a long-short arm ratio of 1.3 in Nipponbare but 0.7 in *O. australiensis*. These reversed patterns indicate that *Oryza* chromosomes may have experienced unique structural rearrangements or centromere positional shifts during speciation.

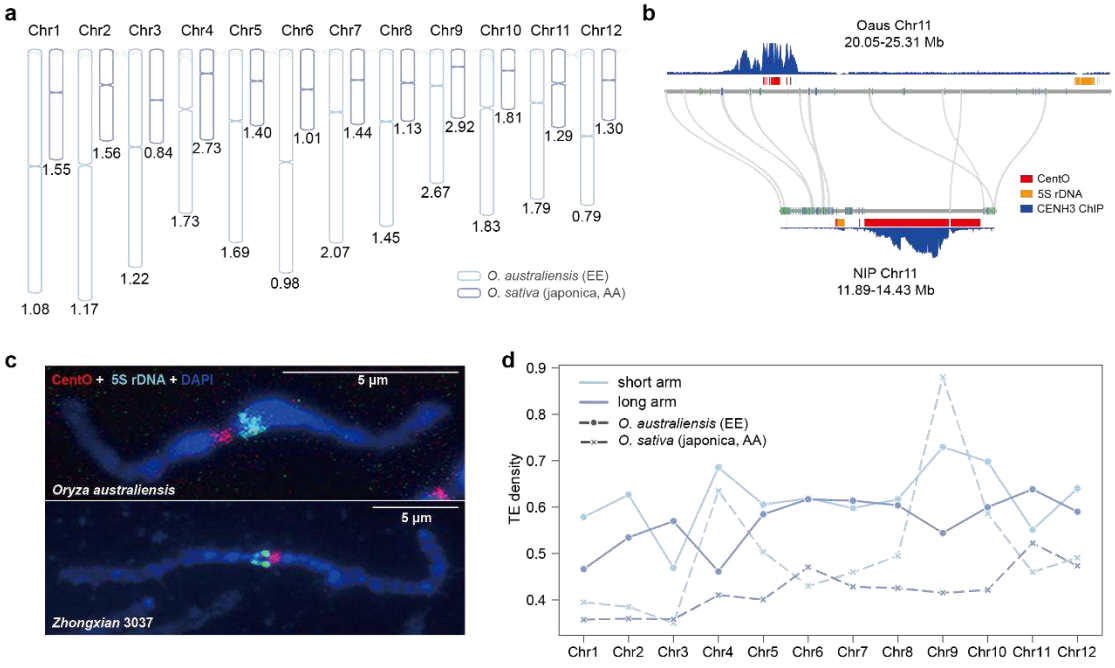


**Fig. 3 Dynamic centromere positional shifts drive structural remodeling of chromosomal arm in *Oryza* species. a,** Chromosome long and short arm distribution of *O. australiensis* and *O. sativa* (NIP). The numbers represent long-short arm ratio. **b,** The pairwise synteny visualization of centromere regions between *O. australiensis* and *O. sativa* (NIP). **c,** FISH with the CentO probe and 5S rDNA was performed on Chr11 of *O. australiensis* and Zhongxian 3037. Chromosomes were stained with DAPI (blue). Scale bar: 5 μm. **d,** Statistics on the TE density of q and p arms of chromosomes of *O. australiensis* and *O. sativa* (NIP).

To elucidate the mechanisms underlying these arm ratio differences, we performed whole-genome synteny analysis. Results demonstrated that the centromere of Chr3 in *O. australiensis* was shifted ~4 Mb toward the p-arm compared to Nipponbare, resulting in short-arm shortening and long-arm elongation (Supplementary Figure 6). Similarly, the centromere of Chr12 in *O. australiensis* shifted ~6 Mb toward the long-arm, leading to short-arm elongation and long-arm shortening (Supplementary Figure

7). Furthermore, comparative syntenic analysis revealed that the centromere on chromosome 11 of *O. australiensis* underwent a short-armward shift relative to Nipponbare, resulting in increased arm ratios in the *O. australiensis*. Strikingly, the 5S rDNA locus—positioned immediately adjacent to the centromere within the short-arm of Nipponbare—was shifted to the long-arm region in *O. australiensis* (Fig. 3b). FISH with CentO and 5S rDNA probes confirmed these structural rearrangements, directly demonstrating both the centromere shift and the 5S rDNA relocated on chromosome 11 of *O. australiensis* (Fig. 3c). This concerted repositioning of both centromere and repetitive elements strongly suggests their co-regulation by a pericentric inversion event. In general, the directional consistency between centromere shifts and long-short arm ratio alterations across chromosome confirms that centromere repositioning is the primary driver of arm ratio divergence.

In humans, chromosomal short arms typically harbor higher proportions of repetitive sequences[39,40]. We further investigated the relationship between chromosomal arm lengths and transposon distribution in *Oryza*. The transposon density was significantly higher($P$<0.032) in short-arms than long-arms across most chromosomes (Fig. 3d). Remarkably, despite the long-short arm ratio reversal in Chr3 and 12 between species, the transposon density disparities persisted. For instance, the p-arm of Chr12 in *O. australiensis*, elongated to 34.5 Mb due to centromere shifts, still exhibited higher transposon density than its long-arm (TE-length/arm-length: short-arm: 0.64 vs. long-arm: 0.58). This phenomenon suggests that inherent chromosomal arm properties—such as chromatin 3D architecture or epigenetic modifications—predominantly govern transposon distribution preferences, while arm length changes exert limited direct influence on transposon density.

**Evolutionary Trajectory of Centromeres Uncovers Three Distinct Repositioning Mechanisms in *Oryza***

Research on centromere evolution in *Oryza* species provides crucial insights into the fundamental mechanisms and driving forces behind chromosomal structural remodeling. To reconstruct the dynamic evolutionary pathways of ancestral centromeres, we took *Leersia perrieri* as an outgroup and constructed a multi-genome

synteny map based on the centromeric core sequences and epigenetic characteristics of the cultivated species (Nipponbare and MH63) and wild species (*O. australiensis* and *O. brachyantha*). Ancestral centromeres were identified by syntenic conservation in at least two independent evolutionary lineages, with cultivated rice genotypes (Nipponbare, MH63) considered collectively as the AA genome lineage. Comparative genomic analysis revealed positional synteny of centromeric location across chromosomes 1, 2, 4, 5, 6, 10, and 11 in both AA and FF genomes, indicating retention of ancestral centromere positions. In contrast, the EE genome exhibited pronounced dynamic remodeling of centromeric positions across these chromosomes (Fig. 4a-4b; Supplementary Figure 6-7). Specifically, a pericentric inversion of Chr11 resulted in centromere repositioning, while Chr1, 2, 4, 5, 6 and 10 form neocentromeres at locations far from the ancestral position (2.3-5 Mb). It is worth noting that Chr1, 2, 4, 5, and 10 of the EE genomes still retain satellite repeats in the ancestral centromere position, and the ancestral centromere archaeological site of Chr1, 2, and 5 have undergone small-scale inversion. For Chr9, syntenic conservation of functional centromeres between EE and AA genomes identifies this locus as ancestral, whereas two pericentric inversions in FF repositioned the centromere toward the q-arm, leaving archaeological traces of residual satellite repeats at the ancestral position (Supplementary Figure 10). In Chr3 and 8, the ancestral centromere position can be determined by the synteny between the inactive centromere loci of the EE genome and the functional centromere loci of the AA genome. This suggests that the AA genome retains the ancestral centromere loci, whereas the EE genome has formed neocentromeres ~3.4 Mb and ~2.3 Mb away from the ancestral loci, respectively (Supplementary Figure 6-7). In contrast, the centromere position of FF genome in Chr3 and Chr8 has shifted due to pericentric inversions. In Chr12, the functional centromere locus of the AA genome and the centromeric satellite repeats locus of *Leersia perrieri* exhibit synteny, supporting this region as the ancestral centromere position in *Oryza* (Fig. 4c). This conclusion is further strengthened by the presence of residual satellite repeats at the ancestral centromeric position in both EE and FF genomes. The neocentromere position on Chr12 in the EE and FF genomes is located at ~6 Mb and

~400 Kb away from their ancestral positions, respectively.

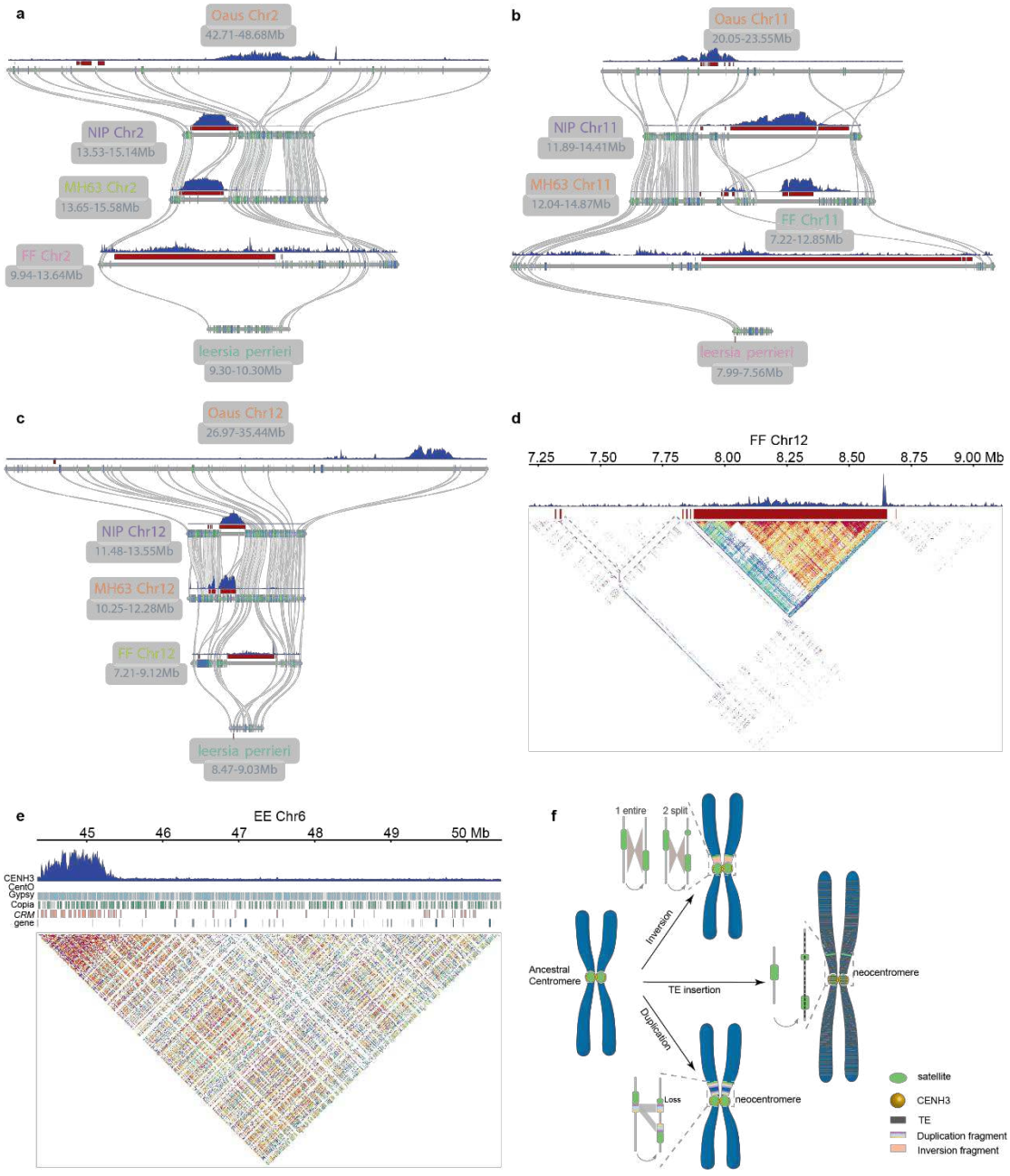


**Fig. 4 Evolution of Centromeres in *Oryza.* a-c**, The pairwise synteny visualization of centromere regions among *O. australiensis*, *O. sativa* (NIP and MH63), *O. brachyantha* and *Leersia perrieri*. **d**, Characteristics of centromeres across *O. brachyantha* Chr12:7.25-9.25 Mb, including the information on CENH3 and satellites. StainedGlass heat maps show sequence identity between 2-kb bins within peri/centromeric regions. The dotted lines represent the sites where large fragment duplication are located. **e**, Characteristics of centromeres across *O. australiensis* Chr6:42-54 Mb, including the information on CENH3, satellites, *Gypsy*, *Copia*, *CRM*, and gene. **f**, Three patterns of centromere repositioning during *Oryza* evolution were discovered.

To investigate the factors driving neocentromere formation, we performed a detailed sequence analysis of the ancestral centromere position and neocentromere position in the EE and FF genomes. In the FF genome, a segment duplication exists

near the ancestral and new centromeric positions of Chr12 (Fig. 4d), suggesting that this segment duplication may drive the formation of the neocentromere on Chr12, consistent with previous reports[41]. In contrast, segment duplication driving neocentromere formation was not observed in the EE genome. Neocentromeres in the EE genome exhibit a high density of TEs, with *CRM* LTR specifically enriched in centromeric regions (Fig. 4e; Supplementary Figure 9-10), suggesting that *CRM* insertion may facilitate CENH3 binding and drive neocentromere formation, similar observations have been reported in *Capsicum* [42].

Here, three centromere repositioning model during *Oryza* evolution were proposed (Fig. 4f): 1. Pericentric inversion-driven local relocation: Inversion events alter the genomic structure of the original centromeric region, leading to small-scale shifts in functional centromere positions. 2. Duplication-mediated neocentromerization: Duplication of large genomic segments near centromeres induces the emergence of new centromere loci. 3. Transposon bursts-drive neocentromere formation: Large-scale transposon amplification induces rapid genome expansion while disrupting epigenetic stability at ancestral centromeres, ultimately driving neocentromere formation at adjacent loci through epigenetic reprogramming.

**Phylogenomic Reconstruction Resolves *Oryza* Divergence Times and EE/DD Clade Affiliation**

To investigate the evolutionary history of the *Oryza* genus, we generated a phylogenetic tree using OrthoFinder software (Emms and Kelly 2019) based on 1,460 single-copy genes shared among 18 chromosome-level *Oryza* genomes, with *Brachypodium distachyon* as the outgroup. The tree resolved six highly supported monophyletic clades (AA, BB, CC, DD/EE, FF, and GG; Fig. 5a), consistent with previous studies [6,8,13,43]. Notably, the diploid CC genome (*O. rhizomatis*, *O. eichingeri*, *O. officinalis*) clustered with the CC subgenome of *O. alta* (CCDD), while the DD subgenome of *O. alta* clustered with the EE genome. Next, the divergence time of *Oryza* species was estimated by calculating the coding sequence (CDS) of 1,460 single-copy genes using MCMCtree software[44]. The estimated divergence time between the CC subgenome and *O. rhizomatis* genome around 3.9 Mya (95% highest posterior density (HPD): 1.57–

6.09 Mya) (Fig. 5a), which is longer than the 1.7 Mya estimated by Ammiraju JSS et al. [6] and the 2.71 Mya estimated by Long et al [13]. The divergence time between the DD subgenome of *O. alta* and EE genome occurred at 5.13 Mya (95% HPD: 1.71–8.89 Mya) (Fig. 5a) placing it intermediate between the 7.08 Mya [6] and 4.44 Mya [13] estimates.

**Shared Transposon Amplification History Implicates EE as Direct Ancestor of the DD Subgenome**

Transposons play a crucial role in shaping eukaryotic genomes, particularly in plants, contributing to genome size expansion and evolution through amplification and insertion into different genomic locations[45-47]. To analyze the evolutionary histories of EE genome and DD subgenomes based on transposon characteristics, the transposon sequence of the *Oryza* genomes was annotated using mipsREdat_9.3p and Viridiplantae_version_3.0 database. LTR-RTs is the main composition of *Oryza* transposon sequences, with a proportion of 91.36% found in the transposon sequences of the EE genome, and proportions of 87.03% and 81.92% found in the DD subgenome and CC subgenomes, respectively (Supplementary Table 9). DNA transposons constitute a low percentage of transposon sequences in the *Oryza* genus, with proportions of 8.3%, 12.5%, and 17.5% found in the EE genome, DD subgenome, and CC subgenomes, respectively. Intriguingly, although the DD and CC subgenomes of the allotetraploid CCDD are similar in size, the DD subgenome harbored larger LTR size and proportions (Fig. 5b; Supplementary Table 9). Given that subgenomic transposons are activated to balance genome size and evolutionary rate in polyploid plants[48,49], these inter-subgenomic differences likely predate CCDD formation.

To reconstruct the transposon-driven evolutionary history of EE and DD, we calculated LTRs insertion time across EE and CCDD genomes. The EE genome displays a high proportion of young *Copia* and *Gypsy* LTR insertions, with amplification peaks occurring at 0.32 Mya and 0.43 Mya, respectively (Fig. 5c). Notably, LTR, particularly *Gypsy* LTR, exhibited an additional amplification peak at 1.7 Mya. These results indicate that the EE genome may have experienced two transposon amplification events at approximately 0.4 Mya and 1.7 Mya, consistent with

previous reports of two transposon amplifications in the EE genome[17,50]. In the CC and DD subgenomes, *Copia* and *Gypsy* have an amplification peak at 0.25 and 0.14 Mya (Fig. 5c), respectively. The DD subgenome, unlike CC subgenome, showed an additional *Copia* and *Gypsy* amplification peak at 1.7 Mya (Fig. 5c), mirroring the EE genome pattern. These results suggest convergent transposon amplification in the DD and EE genomes prior to ~1.7 Mya.

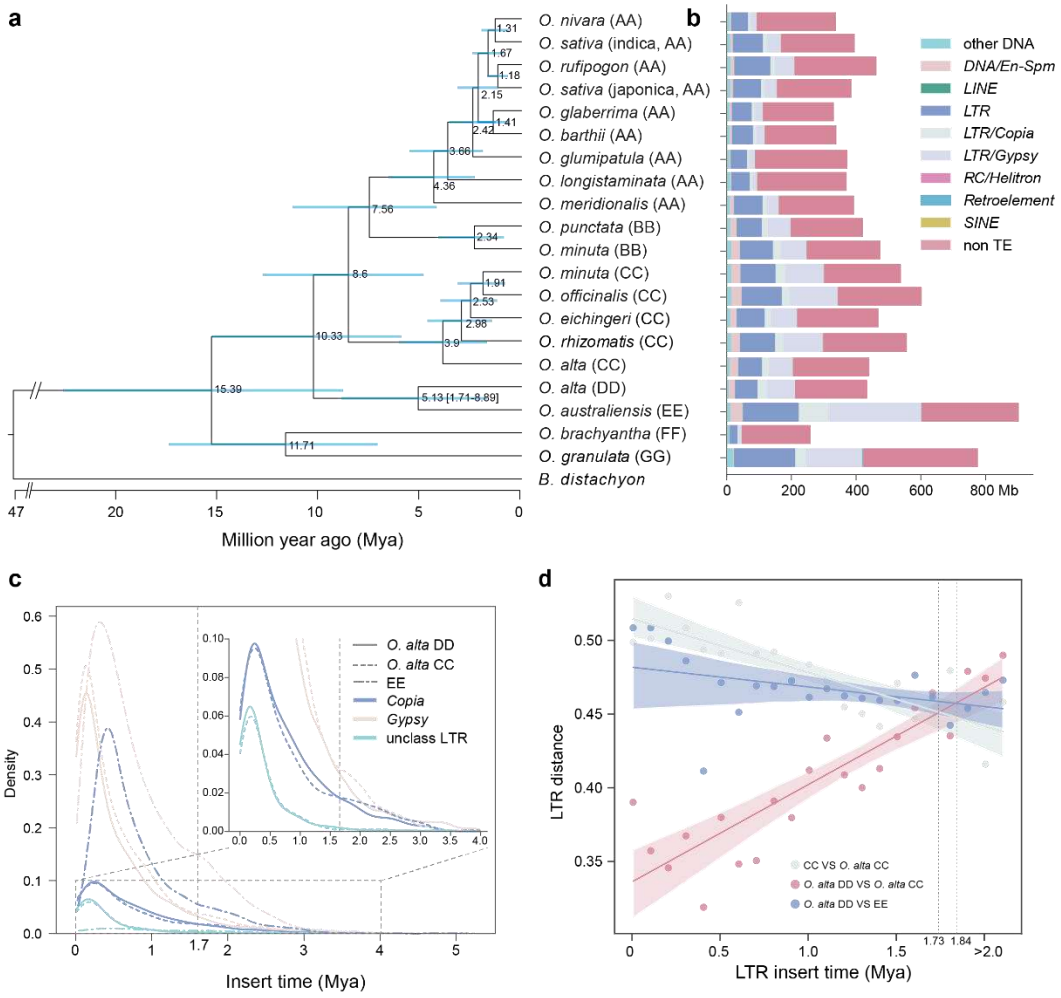


**Fig. 5 EE and DD genomes undergo a similar transposon amplification history. a**, Phylogenetic tree of 17 species based on orthologs of single-copy genes. Blue boxes reflect the 95% highest posterior density (HPD) interval for the age of the respective nodes. The number represents divergent time (Mya) from the common ancestor. **b**, Abundance of the main classes of TEs (Mb) in *Oryza* genome. **c**, Temporal patterns of LTR-RT insertion time in EE genome and tetraploid CCDD genome. **d**, The dynamic changes in the similarity of repetitive sequences between the polyploid genome and its progenitor genome over evolutionary time.

The similarity of repetitive sequences between species has been demonstrated to be effective in studying the evolutionary history of plant and animal groups[51,52]. Here, we specifically focus on the dynamic changes in the similarity of repetitive sequences

between the polyploid genome and its donor genome over evolutionary time. To investigate this, we calculated the sequence similarity of LTRs between CC genome and CC subgenome, CC subgenome and DD subgenome, and EE genome and DD subgenome within different time intervals. As expected, LTR sequence divergence between the CC subgenome and DD subgenomes increased with time, whereas divergence between the CC genome and CC subgenome and between EE genome and DD subgenome decreased (Fig. 5d). Within the time interval of approximately 1.52 to 2.1 Mya, the fitted curves of CC genome and CC subgenome, as well as EE genome and DD subgenome, intersected with the fitted curve of CC subgenome and DD subgenome at 1.73 MYA and 1.84 MYA, respectively. These intersection points indicate the approximate formation period of CCDD at around 1.7 Mya. Although the formation time of the CCDD predicted by transposon sequence features is significantly earlier than the divergence time of CC genome and CC subgenome, as well as EE genome and DD subgenome predicted by single-copy genes, it still falls within the 95% HPD predicted by single-copy gene prediction. What's more, the nearly-parallel trends in the fitted curves of CC genome and CC subgenome, as well as EE genome and DD subgenome, suggest that EE genome may be the direct donor of the DD subgenome.

**Angela LTR Retrotransposons and a Conserved 275-bp Repeat Validate EE as the Donor of DD**

To identify evolutionarily informative LTR sequences linking EE genome and DD subgenomes, we constructed a standardized dataset containing 100,0000 Illumina sequencing reads per genome (AA, BB, CC, EE, FF, GG, CCDD) and performed comparative analysis using RepeatExplorer to detect shared/lineage-specific LTR patterns [32]. Comparative analysis of repetitive sequences across rice species revealed three *Copia/Angela* family clusters enriched in EE and DD genomes but rare/absent in others (Fig. 6a). This lineage-restricted *Angela* LTR enrichment implicates their role in the distinct evolutionary trajectories of the EE genome and DD genomes.

Complete LTR retrotransposons were annotated across 16 *Oryza* genomes using TEsorter [53]. Seven LTR subfamilies from *Copia* and six LTR subfamilies from *Gypsy* exist in the *Oryza* genomes (Supplementary Table 10). Of them, *Ale*, *Ikeros* and *Ivana*

are lineages enriched in the number of intact LTR-RTs of the *Copia* superfamily among AA, BB, CC, CCDD, EE, and GG genotypes. *Retand* is the lineage with the most abundant intact LTR-RTs of the *Gypsy* superfamily among most diploid *Oryza* genotypes, while *Orge* lineage is the most abundant intact LTR-RTs in CC subgenome and DD subgenome (Supplementary Table 10). Among all LTR, the Angela LTR number varies greatly among the *Oryza* genomes, with rare or absent in AA, CC, and FF genomes, and the most abundant in EE genome, followed by DD subgenome (Fig. 6b). This suggests that *Angela* LTR was gradually lost during the evolution of *Oryza*, and similar transposon amplification events before 1.7 Mya of EE genome and DD subgenomes led to the rapid expansion of *Angela* LTR in the genome. As expected, the insertion events of *Angela* LTRs occurred between 1.7-3 Mya in the EE genome and DD subgenomes, which were not observed in the CC subgenome (Fig. 6c).

To elucidate the evolutionary relationship between the EE genome and DD subgenomes at the transposon dynamics level, we conducted a phylogenetic analysis of the integrase (INT) domain sequences of *Angela* LTR retrotransposons inserted prior to 1.7 Mya. The results revealed that *Angela* LTR sequences from EE genome and DD subgenome formed multiple highly supported clades on the phylogenetic tree (Fig. 6d), with a significantly smaller genetic distance (average p-distance = 0.06) compared to their divergence from the GG genome (average p-distance = 0.12), indicating high homology between EE genome and DD subgenome. Further whole-genome synteny analysis demonstrated conserved chromosomal positions and flanking sequences of *Angela* LTRs inserted >1.7 Mya in both EE genome and DD subgenome. For instance, two LTRs on chromosome 4 of the DD subgenome and their counterparts on chromosome 4 of the EE genome were located within syntenic regions (Fig. 6e), suggesting that these LTR insertion events occurred prior to the divergence of the EE genome and DD subgenome and remained unaffected by subsequent chromosomal rearrangements. Collectively, the sequence conservation, synteny, and synchronized amplification history of *Angela* LTRs prior to 1.7 Mya strongly support the hypothesis that the pre-1.7 Mya EE genome served as the direct donor of the DD subgenome in the CCDD polyploid species.

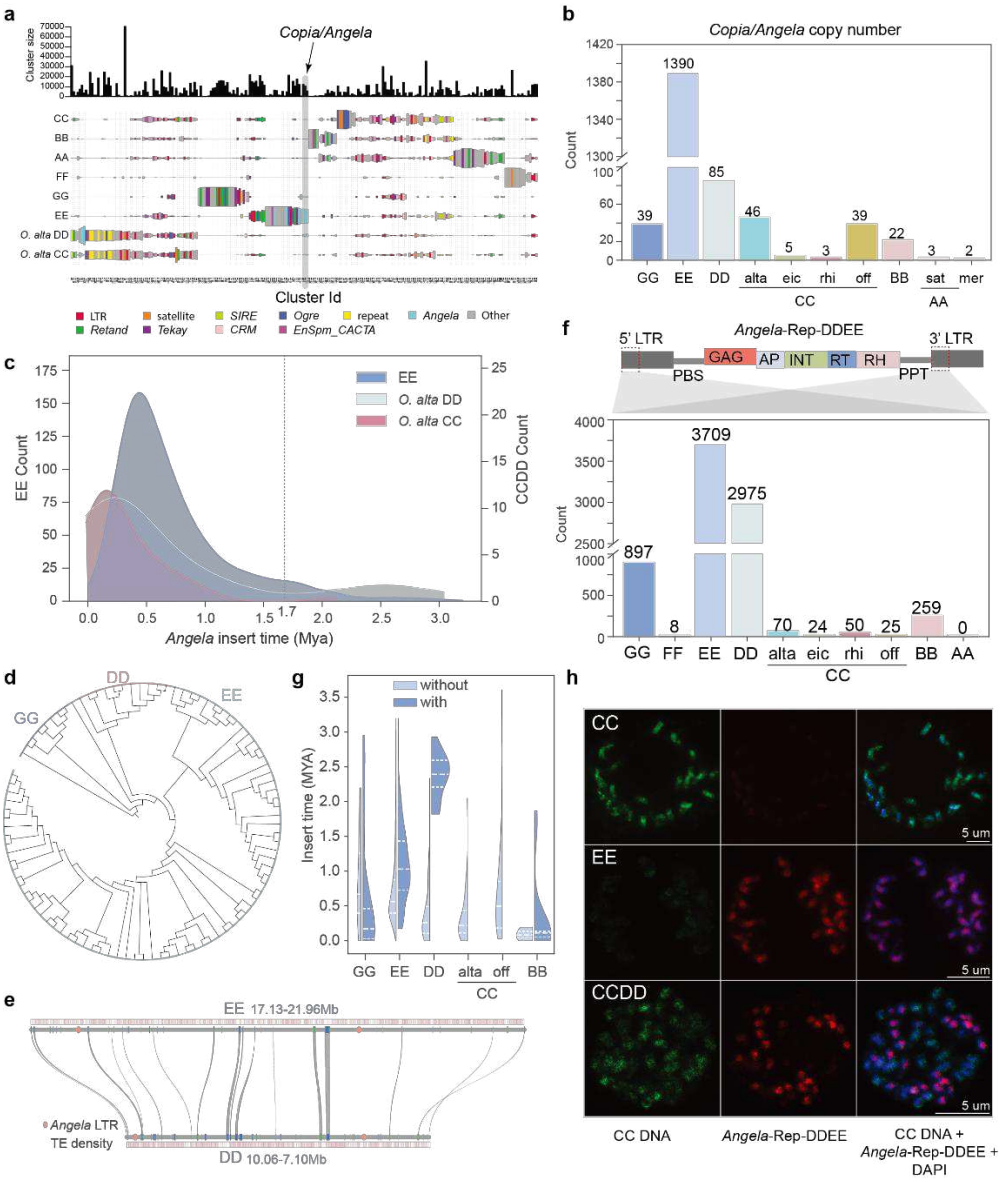


**Fig. 6 *Angela* LTR corrects the divergence time of EE and DD. a**, Comparative repeat analysis of the *Oryza*. The top 228 clusters with a proportion greater than 0.1% of the analyzed reads are shown. Bar plot shows the sizes (numbers of reads) of individual top clusters. Rectangle size is proportional to the number of reads in a cluster for each species. The three clusters in the shaded area are repeats of the *Angela* family that are specifically enriched in the EE and DD genomes. **b**, Copy number of *Angela* LTR in *Oryza* genomes. **c**, *Angela* LTR insertion time for EE genome, CC and DD subgenomes. **d**, Construction of evolutionary trees of *Angela* LTR sequences with insertion times greater than 1.7 Mya in GG (n=7), EE (n=77) and DD (n=11) genomes. **e**, The pairwise synteny visualization of *Angela* LTR between *O. australiensis* and *O. sativa* (NIP). **f**, The diagram at the top shows the structure of *Angela* LTR; The dashed boxes represent repeats that are significantly enriched in the EE and DD genomes, known as *Angra*-Rep-DDEE; The graph at the bottom shows the number of copies of *Angela*-Rep-DDEE in the *Oryza* genomes. **g**, The insertion time of the complete LTR associated with the *Angela*-Rep-DDEE sequence. **h**, FISH of mitotic metaphase chromosomes in *O. australiensis*, CC and *O. alta* using repeat sequence maker (red) and CC DNA (green) as probe with chromosomes counterstained with DAPI (blue).

Transposon characterization was employed to further test the pre-1.7 Mya EE genome as the direct donor of the polyploid DD subgenome. Comparative analysis identified 275-bp repetitive sequences uniquely enriched in the EE genome and DD subgenome when contrasted with the CC subgenome of *O. alta* (Fig. 6f; Supplementary Figure 11; Supplementary Table 11). FISH results confirmed the specific enrichment of this repetitive sequence only in the EE genome and DD subgenome (Fig. 6h). As expected, this specific sequence is a repeat sequence of the LTR terminal in the *Angela* LTR structure and was named the *Angela*-specific repeat sequence in DD and EE (*Angela*-Rep-DDEE) (Fig. 6f). Genome-wide screening revealed maximal *Angela*-Rep-DDEE abundance in the EE genome (3709 loci) and DD subgenome (2974 loci), with secondary enrichment in GG (897) and BB (259) genomes (Fig. 6f). The number of *Angela*-Rep-DDEE loci in the rice genome is significantly higher than the number of *Angela* LTRs because many *Angela* LTRs are structurally incomplete due to transposition. The insertion time of intact *Angela* LTRs that match *Angela*-Rep-DDEE in the DD subgenome ranges from 1.7 to 3 Mya, while the insertion time of intact *Angela* LTRs that do not match *Angela*- Rep-DDEE is between 0 and 1.5 Mya (Fig. 6g). Similar to the DD subgenome, the insertion time of intact *Angela* LTRs that match *Angela*-Rep-DDEE in the EE genome is significantly older than that of intact *Angela* LTRs that do not match *Angela*-Rep-DDEE. These results suggest that the Angela LTRs inserted in the DD subgenome before 1.7 Mya share similarities structural with those in the EE genome. However, with the formation of CCDD, the evolutionary rate of the DD subgenome significantly accelerated, leading to substantial changes in the *Angela* LTR sequences. In the EE genome, continuous amplification events of *Angela* LTRs occurred between 0 and 3 Mya, and as the evolutionary time progressed, significant differences emerged between the younger *Angela* LTR sequences and the older ones. In general, *Angela* LTR in EE and DD genomes had similar amplification events and structural characteristics before 1.7 Mya, and the sequence of *Angela* LTR in EE and DD genomes changed significantly with transposon burst or polyploid formation. This suggests the EE genome served as the direct donor of the DD subgenome following divergence from their common ancestor before 1.7 Mya.

## Discussion

The telomere-to-telomere (T2T) assembly of the *O. australiensis* genome (901 Mb, 12 chromosomes) represents a critical resource for elucidating genetic diversity, functional gene annotation, and evolutionary dynamics in wild rice. This assembly achieved high continuity (BUSCO completeness: 99.2%; LAI score: 20.67) and accuracy (QV score: 35.2), enabling the comprehensive characterization of centromeres and telomeres in the EE genome lineage. This advance address long-standing gaps in utilizing wild rice genetic resources for crop improvement.

The extinction of diploid DD genome species and restriction of DD genomes to three CCDD polyploids have prompted multiple hypotheses regarding DD genome origins and relationships to other *Oryza* genomes[1,3,10,12,54-56]. *O. australiensis* is the sole representative of the EE genome clade in *Oryza* genus. Several studies have suggested that the EE genome is a potential donor to the DD genome[10,12,56]. Ammiraju et al. reported an EE-DD divergence of 7.08 Mya, refuting the donor hypothesis and supporting classification of the *O.* alta DD genome as an independent lineage[6]. The T2T genome of *O. australiensis* allows us to understand the evolutionary relationship between DD and EE more comprehensively. The 1,460 rice single-copy genes' phylogenetic tree clearly shows rice's evolutionary relationship. The EE and DD subgenomes are located in the same branch and the divergence time is 5.13 Mya (95% HPD: 1.72–8.89 Mya), which is shorter than the 7.08 Mya predicted by Ammiraju JSS et al. and longer than 4.4 Mya predicted by Long et al. [6,13]. In addition, the divergence time of CC subgenome and diploid CC genomes obtained in this study was 3.9 Mya (95% HPD: 1.29–5.47 Mya), which is longer than the 1.7 Mya predicted by Ammiraju et al. and 2.71 Mya predicted by Long et al. The significant differences in divergence times of the EE genome and DD subgenome, diploid CC genome and CC subgenome observed in various studies are ambiguous. In recent study, plant polyploidy has been shown to increase meiotic recombination frequency, facilitate speciation and adaptation, and lead to increased genetic variation[14]. In paniceae tribe, the genome evolution rate of allotetraploid *Panicum miliaceum* was significantly faster than that of diploid *Panicum hallii*[15]. These factors combined suggest that plant polyploidy can indeed lead

to an increase in evolution rates. To study the evolutionary rates of tetraploid subgenomes and diploid genomes, we calculated the Tajima's relative rate score of homologous genes between the *B. distachyon* and CCDD subgenomes, as well as between the *B. distachyon* and diploid genomes (CC and EE genomes). The results show that the formation of the tetraploid CCDD significantly accelerated the evolutionary rate of the CC and DD subgenomes (Supplementary Table 12), which may lead to an overestimation of divergence times between DD subgenome and EE, CC subgenome and diploid CC based on single-copy gene sequences. It is noteworthy that Ammiraju JSS et al. did not find significant differences in evolutionary rates between the DD subgenome and the EE genome in the *Adh1* region, which may be attributed to the relatively small number of genes in the *Adh1* region[6]. In general, the increased evolutionary rate of the CCDD subgenome introduces bias in divergence time estimation based on single-copy genes. Divergence times between EE and DD need further correction using other features, such as transposons.

Being a transposon of neutral evolution provides independent evidence for reconciling these conflicts[57]. The evolutionary relationship between EE and DD genomes was analyzed through transposon evolutionary history. *Angela* LTR retrotransposons were identified as marker repeats in both genomes, indicating that the pre-1.7 Mya EE genome served as the direct donor of the CCDD polyploid's DD subgenome. Three evidences support this conclusion: 1, The DD subgenome underwent similar transposon amplification events during 1.7-3 Mya as the EE genome; 2, Similar transposon amplification events lead to significant enrichment of *Angela* LTR in EE and DD; 3, The structure of the tetraploid DD genome *Angela* LTR can be distinguished according to the formation time of CCDD. The *Angela* LTRs inserted between 1.7 and 3 Mya exhibit sequence conservation, collinearity, and synchronous amplification histories with the EE genome. These observations strongly support that the genomic divergence between EE and DD occurred prior to 1.7 Mya, with both lineages sharing a common ancestral genome before this critical timepoint. In summary, this analysis delineates a specific EE-DD evolutionary relationship, resolving a key gap in rice DD genome history.

585 Centromere repositioning and neocentromere formation are increasingly
586 documented across plant species[33,58,59]. In wheat, the formation of polyploids results in
587 the repositioning of centromere regions of multiple chromosomes[59]. The change of
588 centromere position on Chr6 and 7 of cucumber (*Cucumis sativus* L.) and melon
589 (*Cucumis melo* L.) was revealed by comparing FISH localization[58]. 67 high-quality
590 functional centromeric maps of the *Oryza* AA group revealed reposition and
591 neocentromere formation in multiple genomes[33]. In cotton, an inversion near the D08
592 centromere of Zhongmian 113 led to the emergence of a neocentromere[60].
593 Centromere evolutionary dynamics were investigated using L. perrieri as an
594 outgroup, with syntenic relationships compared across AA, EE, and FF genomes.
595 Ancestral centromeres were defined by syntenic conservation across more than two
596 evolutionary lineages, enabling systematic identification across the *Oryza* genus.
597 Minimal centromere shifts were observed in the AA genome, whereas five FF
598 chromosomes exhibited repositioning or neocentromere formation. Further studies
599 revealed that pericentric inversions drove most FF genome centromere shifts, consistent
600 with mechanisms in humans and einkorn wheat[61,62]. Additionally, the formation of a
601 neocentromere on chromosome 12 of the FF genome was attributed to the segmental
602 duplication near the centromere, consistent with the findings of Liao et al[41]. In the EE
603 genome, centromere shifts on chromosome Chr11 were caused by pericentric
604 inversions. For other chromosomes, except for chromosome 9, neocentromere loci
605 emerged on all remaining chromosomes. Unlike the FF genome, the formation of
606 neocentromeres in the EE genome is not caused by inversions or segmental duplications,
607 but rather demonstrates a rare transposon burst-driven mechanism.

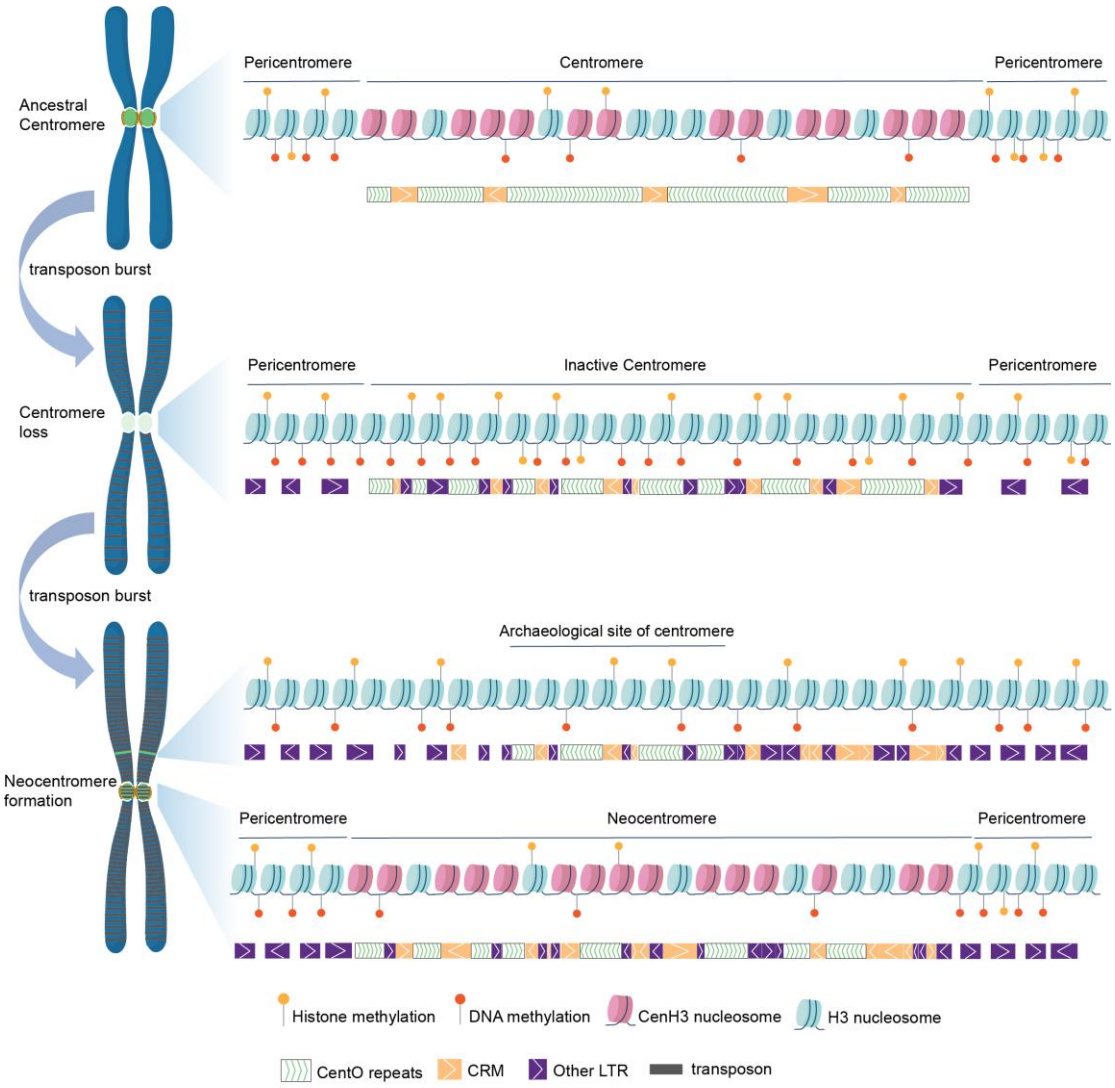


**Fig. 7 The model of neocentromere formation in the *O. australiensis*.** Drawings are not drawn precise to the scale. Functional centromeres exhibit an alternating distribution pattern of CENH3 and H3 nucleosomes. During EE genome evolution, a massive transposon burst disrupted CENH3 nucleosome assembly and compromised the epigenetic homeostasis of centromeric regions. The functional collapse of the original centromere triggered the establishment of a neocentromere within 6 Mb of the ancestral locus. This neocentromere positioning might be driven by high-density insertion of *CRM* LTR. Concurrently, satellite repeats in the original centromere were progressively eliminated, ultimately degenerating into an archaeological site of centromere.

The evolutionary trajectory of EE neocentromeres (Fig. 7) involves: massive *CRM* LTR retrotransposition over the last 3 Mya, causing 2-fold genome expansion through preferential integration into ancestral centromeres. While this targeting mechanism closely resembles the invasion of centromeric satellite arrays by ATHILA transposons in *Arabidopsis*[63], the exceptionally intense transposon bursts in the EE genome directly disrupted CENH3 nucleosome assembly. The subsequent interference with ancestral centromere organization drove the establishment of new centromeric sites within 6 Mb

of the original loci in the EE genome. These neocentromeres likely formed through *CRM* LTR insertions. Following neocentromere establishment, epigenetic silencing mechanisms suppressed the original centromeres, gradually eliminating satellite sequences. The three centromere repositioning mechanisms we delineated inversion-driven shifts, duplication-mediated neocentromerization, and transposon burst-induced neocentromere formation-represent fundamental evolutionary strategies for chromosomal restructuring across eukaryotes. While inversion-driven shifts and duplication-mediated neocentromerization align with established models in taxa like wheat and cotton, transposon burst-induced neocentromere formation unveils transposons as potent architects of centromere plasticity through epigenetic disruption. We propose that non-B DNA structures from LTR termini may induce DNA breaks, with break-induced replication (BIR) as a potential repair contributor to neocentromere establishment—though BIR remains unvalidated in plant centromere repositioning[64]. This tripartite framework resolves long-standing questions about karyotype diversity in polyploid plants and provides a predictive model for centromere evolution. Future studies should explore: (i) how epigenetic reprogramming interfaces with transposon dynamics to stabilize neocentromeres, (ii) whether these mechanisms drive speciation in other taxa with high TE activity, and (iii) their implications for centromere engineering in synthetic genomes for crop improvement.

In summary, the telomere-to-telomere assembly of *O. australiensis* was completed, identifying the pre-1.7 Mya EE genome as the direct donor of the DD subgenome in CCDD tetraploids based on transposon characteristics. This work resolves a key gap in *Oryza* genus evolutionary history concerning DD genome origins and yields novel insights into tetraploid genome evolution.

## Method

### Plant materials and sequencing

Fresh young seed (14 days) tissue was collected from *O. australiensis* plants. For ONT and Illumina sequencing, genome DNA was extracted using the CTAB method, and libraries were prepared according to the manufacturer's instructions. The genome was

sequenced on the ultra-long Oxford Nanopore technology, Nanopore PromethION, and Illumina HiSeq X Ten platforms according to standard ONT and Illumina (Illumina, San Diego, CA, USA) protocols. We generate ~20 Gb ultra-long ONT (~25 coverage), ~100 Gb ONT reads (~100 coverage) and ~84 Gb (~94 coverage) 150-bp paired-end reads.

**mRNA-seq library preparation**

For RNA-seq, 0.5 g sample of young leaves, roots, flag leaf, and panicles were collected from *O. australiensis* plants. The total RNA for each sample was extracted using TRIzol reagent and sequenced on an Illumina HiSeq sequencing system.

**ChIP-seq library preparation**

Young leaves were collected from *O. australiensis* plants. Chromatin immunoprecipitation (ChIP) was performed using published protocols with minor modifications[65]. As above, 2% PVP40 was added to the nuclei isolation buffer and nuclei washing buffer. The commercial antibodies directed against CenH3 were used to obtain immunoprecipitated DNA. ChIP-seq libraries were constructed as previously described and sequenced using the paired-end 150 base pairs (PE150) strategy[65].

**Genome de novo assembly and assessment**

We remove the organelle-related reads and use Nextdenovo v2.3.1 (https://github.com/Nextomics/NextDenovo) with parameters "task=all, read_cut-off=1k, genome_size=900m, rerun=3, input_type=raw, pa_correction=3, min-imap2_options_cns = -x ava-ont -t 8 -k17 -w17" to assemble Ultra-long reads and PromethION reads. At this point, we obtained 13 contig sequences larger than 1 Mb. We extracted telomere-related reads from Nanopore long reads and performed telomere polish on contig. Both ends of 11 contigs were enriched with telomere repeat sequences, indicating that these contigs were complete chromosomes. We connect contig ends that lack telomere repeat sequences and use LR_Gapcloser[66] to fill gaps. Finally, the assembled genome size was ~901 Mb and consisted of 12 telomere-to-telomere chromosomes. We used Arang Rhie's method to polish the assembled genome[20]. We used automated polishing.sh (https://github.com/arangrhie/T2TPolish/tree/master/automated_polishing) to perform 3 rounds of polishing using ONT reads and WGS short reads. Then

NextPolish[67] software was used for polishing the assembled genome three times. The quality of the assembled genome was assessed using BUSCO[68] with parameters "-l viridiplantae_odb10 -m geno". LTR assemble index (LAI) was assessed using LTR_retriever[69] with default parameters. Consensus quality value (QV) was assessed using Merqury (k-mers from WGS reads)[70]. In addition, we mapped the ONT reads with Minimap2 v.2.24[71], Illumina reads with BWA v.0.7.17[72], and RNA-seq reads with HISAT2 v.2.2.1[73] to the assembled genome.

**Gene and repeat annotations**

We used the Genome Annotation Pipeline (https://github.com/meiyang12/Genome-annotation-pipeline) to annotate the genome. First, we downloaded the protein sequences of Nipponbare (http://rice.uga.edu/) and R498 (https://www.mbkbase.org/R498/), with the command "Oryza_sativa_protein.fa". BRAKER2 (v.2.1.5) was used to generate novel gene models with parameters "--prot_seq= Oryza_sativa_protein.fa –etpmode –gff3 --softmasking". Second, we used HISAT2[73] and StringTie[74] for transcripts assembling. Third, homology-based evidence was generated by GenomeThreader[75]. Fourth, we integrated three types of evidence by EVidenceModeler[76] to obtain the official gene sets (OGS). Finally, we used IGV-GSAman (https://gitee.com/CJchen/IGV-sRNA) to individually validate the annotated genes. The quality of the annotated genes was assessed using BUSCO with parameters "-l viridiplantae_odb10 -m tran". Repeats in the genome were annotated using RepeatMasker (http://repeatmasker.org/) with mipsREdat_9.3p and Viridiplantae _version_3.0. Full-length LTR were identified using LTR_retriever[69] and further classified by TEsorter[53]. The insertion time of the full-length LTR was calculated using LTR_retriever. We predicted domains of genes using InterProScan[77] with parameters "-appl TIGRFAM,SMART,CDD,SUPERFAMILY, Pfam" and eggnog-mapper[78] with parameters " --evalue 0.001 --score 60 --pident 40 --query_cover 20 --subject_cover 20 --itype proteins --tax_scope 33090 --target_orthologs all --go_evidence non-electronic --pfam_realign none ". The transcription factors and transcription regulators were annotated using iTAK[79] with default parameters. We use RIdeogram[80] to visualize chromosomes and genes.

**Centromere analysis**

For identification of the centromere region, we used BWA[72] to align the CENH3 ChIP-seq and Input-seq reads to *O. australiensis* genome and used Samtools[72] to obtain reads with map quality greater than 30. We used MACS2[81] to call the peaks of CENH3 with parameters "-f BAM -g 9e8 --broad --broad-cutoff 0.1". We picked peaks with ChIP signals greater than 50 and used Bedtools[82] to merge peaks within 150 kb. Finally, the start and end points of each chromosomal centromeric region based on the CENH3 ChP-seq signal were obtained.

**Identification of centromeric satellite repeat sequences**

To identify the centromere satellite sequence, we used a previously described method[83]. First, 1,000,000 reads were randomly selected from the input data of GG (SRR6175405), FF (DRR053294), EE (CRR1936088), CCDD (CRR123261), CC (DRR014203), BB (DRR056215) and AA (SRR25203306). The selected reads are uploaded to RepeatExplorer and the putative satellites' repeat sequence is calculated using Tandem Repeat Analyzer (TAREAN)[32]. No satellite repeats were detected in GG genome. In order to compare the differences in repetitive sequence composition within the *Oryza* genus, we randomly extracted 1,000,000 reads of AA, BB, CC, tetraploid CC (reads uniquely mapped to the CC subgenome), tetraploid DD (reads uniquely mapped to the DD subgenome), EE, FF, and GG, and wrote them into a file with species labels attached to the reads name. Next, we used RepeatExplorer2[51] to analyze the merged file and obtain the results on differences in repetitive sequence composition within the *Oryza* genus.

**Phylogenetic analysis**

Protein sequences of *B. distachyon* (GCA_000005505.4), *O. granulate* (GWHAAEL00000000), *O. brachyantha* (GCA_000231095.2), *O. alta* (GWHAZTO 00000000), *O. eichingeri* (https://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Shenton_OofficinalisAnnotation_2019), *O. rhizomatis* (https://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Shenton_OofficinalisAnnotation_2019), *O. officinalis* (https://figshare.com/s/32d 0ac68cee7f647e1e3), *O. minuta* (https://figshare.com/s/32d0ac68cee7f647e1e3), *O.*

*punctata* (GCA_000573905.1), *O. meridionalis* (GCA_000338895.2), *O. longistaminata* (GCA_000789195.1), *O. glumipatula* (GCA_000576495.1), *O. glaberrima* (GCA_000147395.1), *O. barthii* (GCA_000182155.2), *O. nivara* (GCA_000576065.1), *O. rufipogon* (GCA_000817225.1), *O. sativa* (NIP: GWHDRIN00000000; MH63: CP054676–CP054688) and *O. australiensis* were calculated by OrthoFinder[84]. To construct phylogenetic relationships, 1460 single-copy orthologues were extracted from all 19 species, and multiple alignment analysis was performed with MUSCLE v.5.1[85]. All alignments were combined into one supergene, and a phylogenetic tree was analyzed with IQ-TREE 2[86]. The divergence time was estimated using MCMCTree from the PAML package v.4.10.7[87], and the following constraints were used for time calibrations: 15-24 Mya for the origin and divergence of *Oryza*[88] and 6-17 Mya for GG and FF genome split[88].

**Differential Analysis of LTR Sequences**

In order to study the evolutionary relationship between tetraploid and diploid genomes, we calculated the sequence differences of LTRs with different insertion times. Firstly, we obtained the fLTR sequences of LTRs at different insertion time intervals in both tetraploid subgenomes and diploid genomes. The fLTR annotation results were derived from LTR_retriever [69]. Next, we randomly paired fLTR sequences from different species at the same insertion time interval and selecting 1000 pairs randomly. We calculated the sequence differences between pairs using Biopython's pairwise2[89]. To mitigate the impact of significant sequence differences caused by different sequence types, we performed a binary classification on the calculation results using gaussian_kde, selecting the average score of the class with low sequence differences as the final evaluation result.

**Evolutionary rate test**

To compare the relative evolutionary rates of EE genome with CC subgenome and DD subgenomes, 1460 orthologue proteins were aligned and concatenated into a super alignment. *B. distachyon* as an out-group, and Tajima's relative rate test analysis was conducted using MEGAX[90].

**Identification of DD and EE genomic repeat markers**

The following steps were used to identify whether conserved repeat sequence markers existed between DD and EE genomes. We split the *O. alta* CCDD genome into CC subgenome and DD subgenome. KMC[91] software was used to generate 21 bp short sequences of the CC subgenome, DD subgenome, and EE genome. Next, we screened out short sequences enriched between EE and CC subgenomes and short sequences enriched between EE and DD subgenomes, respectively. The WGS short reads related to the short sequence are extracted, and the short reads are assembled by RepeatExplorer[32] software. The EE genome and CC subgenome enriched short reads did not obtain assembly results, while the EE genome and DD subgenome enriched short reads assembled consistent repeats. The assembled repeats were mapped to CC subgenome, DD subgenome, and EE genome, respectively. It was found that a 275 bp repeat sequence was enriched only in the DD subgenome and EE genome but was lacking in the CC subgenome. Primers were designed according to the matching sequence of *Angela*-Rep-DDEE of the EE genome (Table S13), and amplified in the EE genome. The biotin labeling was added to the amplified sequence, and FISH experiments were performed.

**Chromosome preparation and fluorescence in situ hybridization (FISH)**

Chromosome preparations were performed according to established methods[92]. For mitotic chromosome analysis, fresh root tips (1-2 cm) were collected from either field-grown plants or germinated seeds. The roots underwent pretreatment with 0.002 M 8-hydroxyquinoline at 20°C for 2 hours to arrest cells at prometaphase, followed by fixation in Carnoy's solution (3:1 ethanol:acetic acid) and storage at -20°C. After thorough washing, samples were enzymatically digested in 2.5% cellulase at 37°C for 1.5 hours, then fixed overnight before being squashed onto pre-chilled slides.

For meiotic chromosome analysis, young panicles at various developmental stages were fixed in Carnoy's solution. Anthers containing microsporocytes at pachytene stage were carefully dissected from spikelets, treated with acetocarmine stain, and gently squashed under coverslips. Slides were briefly heated over a flame, then destained with 45% acetic acid. After freezing in liquid nitrogen for coverslip removal, samples were

dehydrated through an ethanol series.

For FISH analysis, chromosome slides were denatured at 80°C for 2 minutes in 70% formamide/2×SSC buffer, followed by dehydration in an ethanol series. The hybridization mixture (20 μL per slide), containing 50-100 ng of labeled probe in 2×SSC buffer with 50% formamide and 10% dextran sulfate, was applied to denatured chromosomes. Hybridization proceeded overnight at 37°C in a humidified chamber. Post-hybridization washes consisted of 2×SSC at 42°C for 10 minutes followed by two 5-minute room temperature washes.

Three DNA clones were employed as probes: (1) pAtT4 containing telomeric repeats (TTTAGGG), (2) pTa794 harboring 5S rRNA genes, and (3) pRCS2 carrying centromeric CentO repeats. Whole-genome DNA probes from *O. australiensis* and *O. eichingeri* were additionally used. Biotin-labeled probes were detected with Alexa Fluor 488-conjugated streptavidin, while digoxigenin-labeled probes were visualized using rhodamine-conjugated anti-digoxigenin antibodies. Chromosomal DNA was counterstained with 4',6-diamidino-2-phenylindole (DAPI) in antifade mounting medium. Fluorescence imaging was performed using a Zeiss A2 fluorescence microscope equipped with a high-resolution CCD camera.

**Data availability**

The raw sequencing data (ONT, Illumina paired-end and RNA-seq) and genome assembly generated in this study have been deposited in the National Genomics Data Center under BioProject PRJCA041880.

**Acknowledgements**

Not applicable.

**Authors contributions**

T.Z., and ZK.C. provided overall direction for the project. Y.B., S.L., GQ.L, YC.W and Q.Y. performed analysis and assembled the genome, HL.Y., WL.Z., QQ.Y., P.L. performed the experiment, CD. Y provide the seeds, Y.B., T.Z and ZK.C wrote and

revised the manuscript. All authors read and approved the final manuscript.


## Funding

This work was financially supported by the Key R&D Program of Jiangsu Province (Modern Agriculture) (BE2022335), the Project of Zhongshan Biological Breeding Laboratory (BM2022008-02), the National Natural Science Foundation of China (grant no. 32270585), the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) to T.Z, and the grant from the National Key R&D Program of China (2023YFD1200800) to ZK.C. and CD.Y.


## Reference


1 Nayar, N. M. Origin and Cytogenetics of Rice. *Advances in Genetics* **17**, 153-292, doi:10.1016/S0065-2660(08)60172-8 (1973).

2 Aggarwal, R. K., Brar, D. S. & Khush, G. S. Two new genomes in the Oryza complex identified on the basis of molecular divergence analysis using total genomic DNA hybridization. *Mol Gen Genet* **254**, 1-12, doi:10.1007/s004380050384 (1997).

3 Ge, S., Sang, T., Lu, B. R. & Hong, D. Y. Phylogeny of rice genomes with emphasis on origins of allotetraploid species. *Proc Natl Acad Sci U S A* **96**, 14400-14405, doi:10.1073/pnas.96.25.14400 (1999).

4 Ammiraju, J. S. S. *et al.* Dynamic evolution of Oryza genomes is revealed by comparative genomic analysis of a genus-wide vertical data set. *Plant Cell* **20**, 3191-3209, doi:10.1105/tpc.108.063727 (2008).

5 Lu, F. *et al.* Comparative sequence analysis of MONOCULM1-orthologous regions in 14 Oryza genomes. *Proc Natl Acad Sci U S A* **106**, 2071-2076, doi:10.1073/pnas.0812798106 (2009).

6 Ammiraju, J. S. S. *et al.* Spatio-temporal patterns of genome evolution in allotetraploid species of the genus. *Plant J* **63**, 430-442, doi:10.1111/j.1365-313X.2010.04251.x (2010).

7 Wing, R. A., Purugganan, M. D. & Zhang, Q. The rice genome revolution: from an ancient grain to Green Super Rice. *Nat Rev Genet* **19**, 505-517, doi:10.1038/s41576-018-0024-z (2018).

8 Fornasiero, A. *et al.* Oryza genome evolution through a tetraploid lens. *Nat Genet* **57**, 1287–1297, doi:10.1101/2024.05.29.596369 (2025).

9 Chen, E. W., Huang, X. H., Tian, Z. X., Wing, R. A. & Han, B. The Genomics of Oryza Species Provides Insights into Rice Domestication and Heterosis. *Annual Review of Plant Biology, Vol 70* **70**, 639-665, doi:10.1146/annurev-arplant-050718-100320 (2019).

10 Bao, Y. & Ge, S. Origin and phylogeny of Oryza species with the CD genome based on multiple-gene sequence data. *Plant Systematics and Evolution* **249**, 55-66, doi:10.1007/s00606-004-0173-8 (2004).

11 Ammiraju, J. S. S. *et al.* The Oryza BAC resource: a genus-wide and genome scale tool for

exploring rice genome evolution and leveraging useful genetic diversity from wild relatives. *Breeding Sci* **60**, 536-543, doi:10.1270/jsbbs.60.536 (2010).

12 Gong, P. C. & Bao, Y. Nucleotide diversity of the homoeologous adh1 loci in the American allotetraploid Oryza species. *Plant Systematics and Evolution* **276**, 243-253, doi:10.1007/s00606-008-0093-0 (2008).

13 Long, W. *et al.* Genome evolution and diversity of wild and cultivated rice species. *Nat Commun* **15**, 9994, doi:10.1038/s41467-024-54427-3 (2024).

14 Pecinka, A., Fang, W., Rehmsmeier, M., Levy, A. A. & Mittelsten Scheid, O. Polyploidization increases meiotic recombination frequency in Arabidopsis. *BMC Biol* **9**, 24, doi:10.1186/1741-7007-9-24 (2011).

15 Sun, Y. *et al.* Biased mutations and gene losses underlying diploidization of the tetraploid broomcorn millet genome. *Plant J* **113**, 787-801, doi:10.1111/tpj.16085 (2023).

16 Yu, H. *et al.* A route to de novo domestication of wild allotetraploid rice. *Cell* **184**, 1156-1170 e1114, doi:10.1016/j.cell.2021.01.013 (2021).

17 Henry, R. J. in *The Wild Oryza Genomes* (eds Tapan K. Mondal & Robert J. Henry) Ch. Chapter 5, 61-66 (Springer International Publishing, 2018).

18 Piegu, B. *et al.* Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. *Genome Res* **16**, 1262-1269, doi:10.1101/gr.5290206 (2006).

19 Phillips, A. L. *et al.* The first long-read nuclear genome assembly of Oryza australiensis, a wild rice from northern Australia. *Sci Rep* **12**, 10823, doi:10.1038/s41598-022-14893-5 (2022).

20 Mc Cartney, A. M. *et al.* Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nat Methods* **19**, 687-695, doi:10.1038/s41592-022-01440-3 (2022).

21 Richards, E. J. & Ausubel, F. M. Isolation of a higher eukaryotic telomere from Arabidopsis thaliana. *Cell* **53**, 127-136, doi:10.1016/0092-8674(88)90494-1 (1988).

22 Mizuno, H. *et al.* Sequencing and characterization of telomere and subtelomere regions on rice chromosomes 1S, 2S, 2L, 6L, 7S, 7L and 8S. *Plant J* **46**, 206-217, doi:10.1111/j.1365-313X.2006.02684.x (2006).

23 Shang, L. *et al.* A complete assembly of the rice Nipponbare reference genome. *Mol Plant* **16**, 1232-1236, doi:10.1016/j.molp.2023.08.003 (2023).

24 Belyayev, A. *et al.* Telomere sequence variability in genotypes from natural plant populations: unusual block-organized double-monomer terminal telomeric arrays. *BMC Genomics* **24**, 572, doi:10.1186/s12864-023-09657-y (2023).

25 Heslop-Harrison, J. S. & Schwarzacher, T. Organisation of the plant genome in chromosomes. *Plant J* **66**, 18-33, doi:10.1111/j.1365-313X.2011.04544.x (2011).

26 Koo, D. H., Zhao, H. & Jiang, J. Chromatin-associated transcripts of tandemly repetitive DNA sequences revealed by RNA-FISH. *Chromosome Res* **24**, 467-480, doi:10.1007/s10577-016-9537-5 (2016).

27 Cheng, Z., Stupar, R. M., Gu, M. & Jiang, J. A tandemly repeated DNA sequence is associated with both knob-like heterochromatin and a highly decondensed structure in the meiotic pachytene chromosomes of rice. *Chromosoma* **110**, 24-31, doi:10.1007/s004120000126 (2001).

28 Cheng, Z. *et al.* Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell* **14**, 1691-1704, doi:10.1105/tpc.003079 (2002).
29 Ohmido, N., Kijima, K., Akiyama, Y., de Jong, J. H. & Fukui, K. Quantification of total genomic DNA and selected repetitive sequences reveals concurrent changes in different DNA families in indica and japonica rice. *Mol Gen Genet* **263**, 388-394, doi:10.1007/s004380051182 (2000).
30 Nagaki, K. *et al.* Sequencing of a rice centromere uncovers active genes. *Nat Genet* **36**, 138-145, doi:10.1038/ng1289 (2004).
31 Zhang, T. *et al.* The CentO satellite confers translational and rotational phasing on cenH3 nucleosomes in rice centromeres. *Proc Natl Acad Sci U S A* **110**, E4875-4883, doi:10.1073/pnas.1319548110 (2013).
32 Novak, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**, 792-793, doi:10.1093/bioinformatics/btt054 (2013).
33 Xie, L. *et al.* Genetic diversity and evolution of rice centromeres. *bioRxiv*, doi:10.1101/2024.07.28.605524 (2024).
34 Zhao, H. *et al.* Gene Expression and Chromatin Modifications Associated with Maize Centromeres. *G3 (Bethesda)* **6**, 183-192, doi:10.1534/g3.115.022764 (2015).
35 Xue, C. *et al.* De novo centromere formation in pericentromeric region of rice chromosome 8. *Plant J* **111**, 859-871, doi:10.1111/tpj.15862 (2022).
36 Sharakhova, M. V., Xia, A., Leman, S. C. & Sharakhov, I. V. Arm-specific dynamics of chromosome evolution in malaria mosquitoes. *BMC Evol Biol* **11**, 91, doi:10.1186/1471-2148-11-91 (2011).
37 Kakui, Y. *et al.* Chromosome arm length, and a species-specific determinant, define chromosome arm width. *Cell Reports* **41**, doi:10.1016/j.celrep.2022.111753 (2022).
38 Therizols, P., Duong, T., Dujon, B., Zimmer, C. & Fabre, E. Chromosome arm length and nuclear constraints determine the dynamic relationship of yeast subtelomeres. *Proc Natl Acad Sci U S A* **107**, 2025-2030, doi:10.1073/pnas.0914187107 (2010).
39 McStay, B. The p-Arms of Human Acrocentric Chromosomes Play by a Different Set of Rules. *Annu Rev Genomics Hum Genet* **24**, 63-83, doi:10.1146/annurev-genom-101122-081642 (2023).
40 Antonarakis, S. E. Short arms of human acrocentric chromosomes and the completion of the human genome sequence. *Genome Res* **32**, 599-607, doi:10.1101/gr.275350.121 (2022).
41 Liao, Y. *et al.* Comparison of Oryza sativa and Oryza brachyantha Genomes Reveals Selection-Driven Gene Escape from the Centromeric Regions. *Plant Cell* **30**, 1729-1744, doi:10.1105/tpc.18.00163 (2018).
42 Chen, W. *et al.* Two telomere-to-telomere gapless genomes reveal insights into Capsicum evolution and capsaicinoid biosynthesis. *Nat Commun* **15**, 4295, doi:10.1038/s41467-024-48643-0 (2024).
43 Stein, J. C. *et al.* Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. *Nat Genet* **50**, 285-296,

doi:10.1038/s41588-018-0040-0 (2018).
44 Rannala, B. & Yang, Z. Inferring speciation times under an episodic molecular clock. *Syst Biol* **56**, 453-466, doi:10.1080/10635150701420643 (2007).
45 Lisch, D. How important are transposons for plant evolution? *Nature Reviews Genetics* **14**, 49-61, doi:10.1038/nrg3374 (2013).
46 Akakpo, R., Carpentier, M. C., Ie Hsing, Y. & Panaud, O. The impact of transposable elements on the structure, evolution and function of the rice genome. *New Phytol* **226**, 44-49, doi:10.1111/nph.16356 (2020).
47 Bennetzen, J. L. & Wang, H. The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. *Annual Review of Plant Biology* **65**, 505-530, doi:10.1146/annurev-arplant-050213-035811 (2014).
48 Chen, Z. J. *et al.* Genomic diversifications of five Gossypium allopolyploid species and their impact on cotton improvement. *Nat Genet* **52**, 525-533, doi:10.1038/s41588-020-0614-5 (2020).
49 Meca, E., Diez, C. M. & Gaut, B. S. Modeling transposable elements dynamics during polyploidization in plants. *J Theor Biol* **579**, 111701, doi:10.1016/j.jtbi.2023.111701 (2024).
50 Tikapunya, T., Fox, G., Furtado, A. & Henry, R. Grain physical characteristic of the Australian wild rices. *Plant Genet Resour-C* **15**, 409-420, doi:10.1017/S1479262116000083 (2017).
51 Vitales, D., Garcia, S. & Dodsworth, S. Reconstructing phylogenetic relationships based on repeat sequence similarities. *Mol Phylogenet Evol* **147**, 106766, doi:10.1016/j.ympev.2020.106766 (2020).
52 Johnson, W. E. & Coffin, J. M. Constructing primate phylogenies from ancient retrovirus sequences. *Proceedings of the National Academy of Sciences* **96**, 10254-10260, doi:doi:10.1073/pnas.96.18.10254 (1999).
53 Zhang, R. G. *et al.* TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic Res* **9**, doi:10.1093/hr/uhac017 (2022).
54 Fukui, K., Shishido, R. & Kinoshita, T. Identification of the rice D-genome chromosomes by genomic in situ hybridisation. *Theoretical and Applied Genetics* **95**, 1239-1245, doi:10.1007/s001220050687 (1997).
55 Li, C. B., Zhang, D. M., Ge, S., Lu, B. R. & Hong, D. Y. Differentiation and inter-genomic relationships among C, E and D genomes in the Oryza officinalis complex (Poaceae) as revealed by multicolor genomic in situ hybridization. *Theoretical and Applied Genetics* **103**, 197-203, doi:DOI 10.1007/s001220100562 (2001).
56 Guo, Y. L. & Ge, S. Molecular phylogeny of Oryzeae (Poaceae) based on DNA sequences from chloroplast, mitochondrial, and nuclear genomes. *Am J Bot* **92**, 1548-1558, doi:10.3732/ajb.92.9.1548 (2005).
57 Lawlor, M. A. & Ellison, C. E. Evolutionary dynamics between transposable elements and their host genomes: mechanisms of suppression and escape. *Curr Opin Genet Dev* **82**, doi:10.1016/j.gde.2023.102092 (2023).
58 Han, Y. *et al.* Centromere repositioning in cucurbit species: implication of the genomic impact from centromere activation and inactivation. *Proc Natl Acad Sci U S A* **106**, 14937-14941, doi:10.1073/pnas.0904833106 (2009).
59 Zhao, J. *et al.* Centromere repositioning and shifts in wheat evolution. *Plant Commun* **4**, 100556, doi:10.1016/j.xplc.2023.100556 (2023).

60 Hu, G. *et al.* A telomere-to-telomere genome assembly of cotton provides insights into centromere evolution and short-season adaptation. *Nat Genet* **57**, 1031-1043, doi:10.1038/s41588-025-02130-4 (2025).
61 Ahmed, H. I. *et al.* Einkorn genomics sheds light on history of the oldest domesticated wheat. *Nature* **620**, 830-838, doi:10.1038/s41586-023-06389-7 (2023).
62 Logsdon, G. A. *et al.* The variation and evolution of complete human centromeres. *Nature* **629**, 136-145, doi:10.1038/s41586-024-07278-3 (2024).
63 Wlodzimierz, P. *et al.* Cycles of satellite and transposon evolution in Arabidopsis centromeres. *Nature* **618**, 557-565, doi:10.1038/s41586-023-06062-z (2023).
64 Talbert, P. B. & Henikoff, S. Centromeres drive and take a break. *Chromosome Research* **33**, doi:10.1007/s10577-025-09777-z (2025).
65 Nagaki, K. *et al.* Chromatin Immunoprecipitation Reveals That the 180-bp Satellite Repeat Is the Key Functional DNA Element of Arabidopsis thaliana Centromeres. *Genetics* **163**, 1221-1225, doi:10.1093/genetics/163.3.1221 (2003).
66 Xu, G. C. *et al.* LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *Gigascience* **8**, doi:10.1093/gigascience/giy157 (2019).
67 Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253-2255, doi:10.1093/bioinformatics/btz891 (2020).
68 Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212, doi:10.1093/bioinformatics/btv351 (2015).
69 Ou, S. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol* **176**, 1410-1422, doi:10.1104/pp.17.01310 (2018).
70 Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245, doi:10.1186/s13059-020-02134-9 (2020).
71 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).
72 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
73 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907-915, doi:10.1038/s41587-019-0201-4 (2019).
74 Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290-295, doi:10.1038/nbt.3122 (2015).
75 Gremme, G., Brendel, V., Sparks, M. E. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Inform Software Tech* **47**, 965-978, doi:10.1016/j.infsof.2005.09.005 (2005).
76 Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7, doi:10.1186/gb-2008-9-1-r7 (2008).
77 Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics*

**30**, 1236-1240, doi:10.1093/bioinformatics/btu031 (2014).

78 Cantalapiedra, C. P., Hernandez-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol* **38**, 5825-5829, doi:10.1093/molbev/msab293 (2021).

79 Zheng, Y. *et al.* iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases. *Mol Plant* **9**, 1667-1670, doi:10.1016/j.molp.2016.09.014 (2016).

80 Hao, Z. *et al.* RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Comput Sci* **6**, e251, doi:10.7717/peerj-cs.251 (2020).

81 Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, doi:10.1186/gb-2008-9-9-r137 (2008).

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

83 Gong, Z. *et al.* Repeatless and repeat-based centromeres in potato: implications for centromere evolution. *Plant Cell* **24**, 3559-3574, doi:10.1105/tpc.112.100511 (2012).

84 Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238, doi:10.1186/s13059-019-1832-y (2019).

85 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797, doi:10.1093/nar/gkh340 (2004).

86 Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534, doi:10.1093/molbev/msaa015 (2020).

87 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591, doi:10.1093/molbev/msm088 (2007).

88 Tang, L. *et al.* Phylogeny and biogeography of the rice tribe (Oryzeae): evidence from combined analysis of 20 chloroplast fragments. *Mol Phylogenet Evol* **54**, 266-277, doi:10.1016/j.ympev.2009.08.007 (2010).

89 Cock, P. J. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423, doi:10.1093/bioinformatics/btp163 (2009).

90 Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution* **35**, 1547-1549, doi:10.1093/molbev/msy096 (2018).

91 Kokot, M., Dlugosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759-2761, doi:10.1093/bioinformatics/btx304 (2017).

92 You, H. *et al.* Chromosome ends initiate homologous chromosome pairing during rice meiosis. *Plant Physiology* **195**, 2617-2634, doi:10.1093/plphys/kiae152 (2024).

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx