

**Telomere-to-telomere genome assembly of *Oryza australiensis* reveals
transposon-driven centromere repositioning and shared EE–DD**

ancestry

Bao *et al.*

Supplementary Method 1. TE density calculation

To calculate TE density for each chromosome arm and the whole chromosome, we first determined the centromere position based on CENH3 ChIP-seq data. The midpoint of the centromeric region was used as the boundary to divide each chromosome into short arm (from chromosome start to the midpoint) and long arm (from the midpoint to chromosome end). The lengths of the short and long arms were calculated accordingly. For each chromosome, transposable element (TE) coverage was calculated separately for the short arm and long arm by summing the lengths of all TE intervals (TE end position minus start position) located within each arm, based on their midpoint positions. TE density for each arm was then computed as the total TE coverage length divided by the total length of that arm. Whole-chromosome TE density was calculated as the sum of TE coverage lengths across both arms divided by the total chromosome length.

Supplementary Method 2. Comparative synteny analysis of centromere evolution

The evolutionary analysis of centromeres was performed by projecting centromeric features onto a whole-genome synteny map. A multi-genome synteny map was inferred and constructed using the JCVI software packages. This was achieved by first identifying putative orthologous gene pairs (anchors) among the protein-coding genes of the included genomes, and then defining systemic blocks based on the conserved order and density of these anchors. This process involved several sequential steps implemented through specific JCVI modules. First, the genomic coordinates of protein-coding genes were extracted from the genome annotation files (GFF format) using the `jcvi.formats.gff bed` command to generate BED files for each genome. Orthologous gene pairs between any two genomes were then identified by comparing their coding sequences (CDS) using BLASTN with default parameters (`jcvi.compara.catalog ortholog`). High-confidence anchors were defined as reciprocal best hits. Subsequently, the `jcvi.compara.synteny mcsan` command was employed to delineate systemic blocks

based on the conservation of order and density among these anchor gene pairs. For multi-genome comparison, the systemic blocks from pairwise analyses were merged using `jcvl.formats.base.join`. Finally, the resulting synteny maps were visualized using the `jcvl.graphics.synteny` module.

The genomic coordinates of the core centromeric satellite repeats (CentO) and the CENH3-binding domains (from ChIP-seq peaks) for each species were then integrated into this syntenic map. The CENH3 ChIP-seq and CentO tracks were plotted using `pyGenomeTracks`. This integrated map allowed for the direct visual comparison and identification of conserved, repositioned, and nascent centromeres across the *Oryza* genus.

Supplementary Method 3. Inference of ancestral centromere positions

The inference of ancestral centromere positions was based on the conservation of functional centromeres across multiple divergent lineages. Syntenic blocks among *O. sativa* (AA), *O. australiensis* (EE), and *O. brachyantha* (FF) were identified using MCScan, with orthology defined by reciprocal best-hit BLASTP analysis (E-value < 1e-10). A genomic locus was designated as an ancestral centromere if the functional CENH3-binding domain was located within a syntenic block shared by at least two of the three lineages (AA, EE, FF). This approach relies on the conservation of the flanking genic regions to anchor the centromeric position phylogenetically.

Supplementary Method 4. Differential analysis of LTR sequences

To investigate the evolutionary relationships between tetraploid and diploid genomes, we compared the sequence divergence of Long Terminal Repeat retrotransposons (LTR-RTs) across different insertion time windows. The analysis focused on the 5' Long Terminal Repeat (5' LTR) sequences, as they provide a homologous and representative marker for each intact LTR-RT element. First, full-length LTR (fLTR) sequences and their estimated insertion times were obtained using `LTR_retriever`¹. These 5' LTRs were

then binned into consecutive 100-thousand-year (kyr) intervals based on their insertion age. For each time interval and each pair of genomes compared, we randomly sampled LTR sequences for pairwise alignment: 1,000 pairs were randomly drawn from all possible combinations, and this sampling process was independently repeated three times to enhance robustness. Sequence alignment was performed using the Needleman-Wunsch global alignment algorithm (implemented via `Bio.pairwise2.align.globalxx`)², and sequence divergence was calculated as below:

$$\text{Sequence divergence} = 1 - (\text{alignment score} / \text{length of the longer sequence}) \quad (1)$$

Because random alignments mix signals from homologous LTR pairs (low divergence) and non-homologous LTR pairs (high divergence), we applied K-means clustering (with the number of clusters set to 2) to partition all computed divergence values into two classes. The class with the smaller mean was selected as representing the homologous comparison signal, and the mean of this class was taken as the final quantitative measure of LTR sequence divergence between the two genomes in that time interval. Finally, divergence values for all genome pairs across time intervals were visualized as a scatter plot (Fig. 5d).

Supplementary Method 5. Classification of intact *Angela* LTR based on *Angela*-Rep-DDEE similarity

To investigate the distinct evolutionary dynamics of *Angela* LTR retrotransposons, all intact elements identified across the *Oryza* genomes (BB, CC, DD, EE, GG) were categorized based on their sequence similarity to the previously defined *Angela*-Rep-DDEE domain. A BLASTN search (E-value $\leq 1 \times 10^{-4}$) was performed using the *Angela*-Rep-DDEE sequence as a query against each target genome. Only hits with >80% sequence identity and alignment length >100 bp were considered significant. An intact *Angela* LTR was classified as "similar to *Angela*-Rep-DDEE" if its annotated genomic coordinates contained at least one significant BLASTN hit (i.e., the hit fell entirely

within the LTR region); intact LTRs without any such hit within their boundaries were classified as "dissimilar to *Angela*-Rep-DDEE". This binary classification enabled a comparative analysis of the insertion time distributions between the two groups to assess potential correlations with specific evolutionary periods.

Supplementary Method 6. 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_cutoff=1k, genome_size=900m, rerun=3, input_type=raw, pa_correction=3, minimap2_options_cns = -x ava-ont -t 8 -k17 -w17" to assemble Ultra-long reads and PromethION reads. This initial assembly generated 14 contigs, of which 13 were longer 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_Gapcloser³ to fill gaps. After gap filling, we performed a final round of quality assessment. One contig (148 kb) was found to be fully contained within a larger assembled region based on whole-genome alignment and was therefore removed as redundancy. 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 genome⁴. 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 NextPolish⁵ software was used for polishing the assembled genome three times. The quality of the assembled genome was assessed using BUSCO⁶ with parameters "-l viridiplantae_odb10 -m geno". LTR assemble index (LAI) was assessed using LTR_retriever¹ with default parameters. Consensus quality value (QV) was assessed using Merqury (k-mers=17 from WGS reads)⁷. In addition, we mapped the ONT reads with Minimap2 v.2.24⁸, Illumina reads with BWA v.0.7.17⁹, and RNA-seq reads with

HISAT2 v.2.2.1¹⁰ to the assembled genome.

Supplementary Method 7. Identification and annotation of repetitive sequences

Repetitive sequences were identified and annotated with RepeatMasker (<http://repeatmasker.org/>) using the mipsREdat_9.3p and Viridiplantae_version_3.0 libraries. Full-length LTR were identified using LTR_retriever¹ and further classified by TESorter¹¹; their insertion times were also estimated using LTR_retriever. Specifically, the *O. australiensis* genome assembly was indexed using suffixerator (GenomeTools) with parameters: -tis -suf -lcp -des -ssp -sds -dna. Two structure-based prediction tools were then employed in parallel: LTRharvest was run with -similar 90 -vic 10 -seed 20 -minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1, and LTR_FINDER was executed with -D 20000 -d 500 -L 7000 -l 100 -p 20 -C -M 0.9. The raw predictions from both tools were integrated, filtered, and deduplicated using LTR_retriever with the options -inharvest LTR.harvest.scn -infinder LTR_finder.scn -threads 150 to generate the final high-confidence set of full-length LTR retrotransposons.

Supplementary Method 8. Gene prediction and functional annotation

We performed gene annotation following Genome Annotation Pipeline (<https://github.com/meiyang12/Genome-annotation-pipeline>). Briefly, protein sequences of Nipponbare (<http://rice.uga.edu/>) and R498 (<https://www.mbkbase.org/R498/>) were downloaded and combined as "Oryza_sativa_protein.fa". BRAKER2 (v.2.1.5) was then run with the parameters "--prot_seq=Oryza_sativa_protein.fa --etpmode --gff3 --softmasking" to generate *ab initio* gene models. Transcript assembly was conducted using HISAT2¹⁰ and StringTie¹². Homology-based gene prediction was used in GenomeThreader¹³ using protein evidence. The three sets of evidence—*ab initio* predictions, transcript assemblies, and

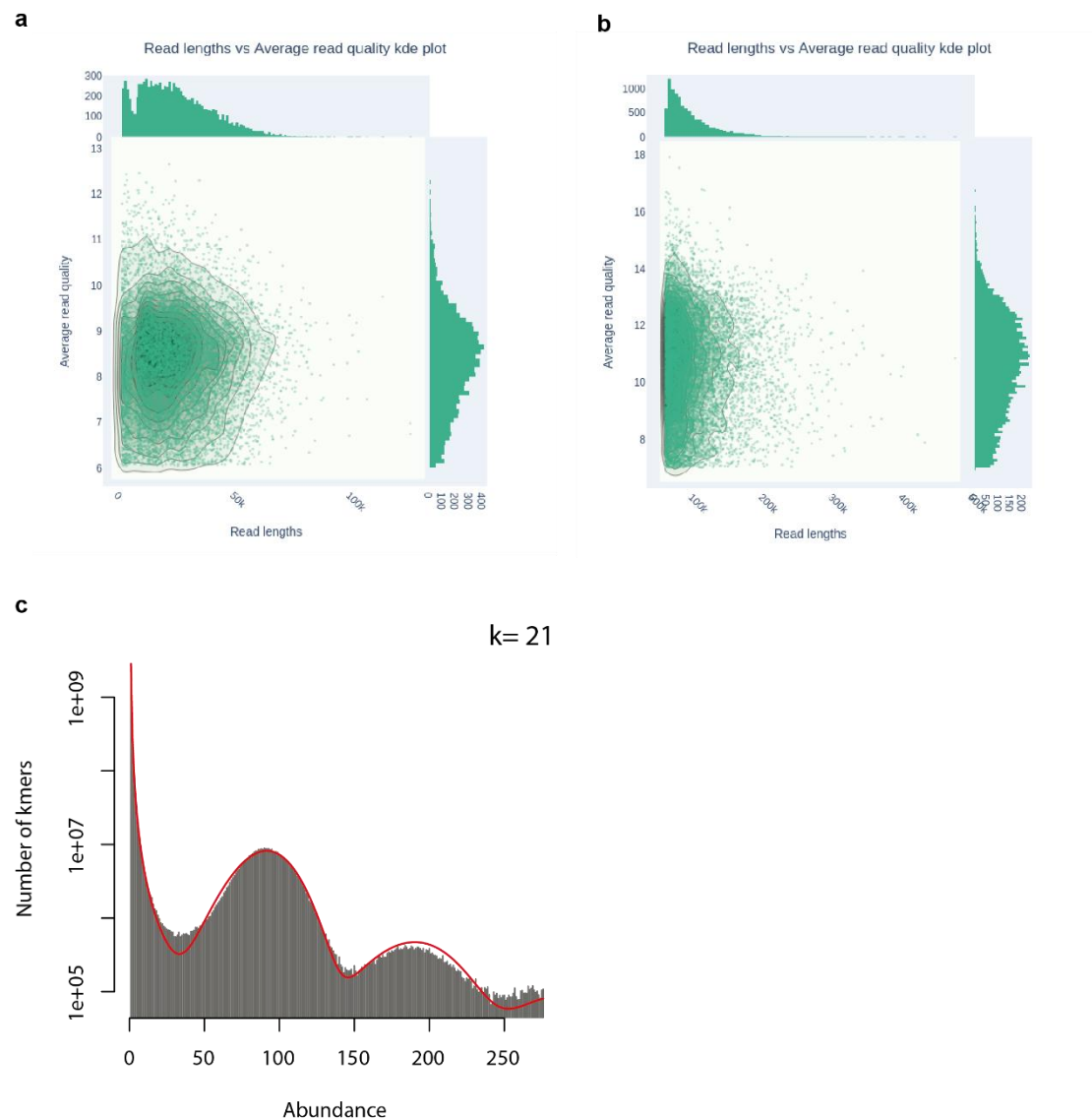
homology-based predictions—were integrated using EVIDENCEModeler¹⁴ to produce the official gene set (OGS). Each gene in the OGS was manually validated using IGV-GSaman (<https://gitee.com/CJchen/IGV-sRNA>). Gene annotation quality was assessed with BUSCO under the lineage dataset "viridiplantae_odb10" in -m tran mode.

Functional annotation of protein-coding genes included domain prediction with InterProScan¹⁵ (-appl TIGRFAM,SMART,CDD,SUPERFAMILY,Pfam) and orthology assignment with eggNOG-mapper¹⁶ (--evaluate 0.001 --score 60 --pidient 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 iTAK¹⁷ with default parameters. We use RIdiogram¹⁸ to visualize chromosomes and genes.

Supplementary Method 9. Assessment of centromeric assembly accuracy by comparative dot plot analysis

To evaluate the structural concordance of the centromere regions in our final assembly, we performed a comparative dot plot analysis against an independent assembly. First, the genome was assembled independently using two strategies: 1) a de novo assembly was generated from the ultra-long and PromethION reads using NECAT (v0.0.1) with the following parameters: MIN_READ_LENGTH=3000, PREP_OUTPUT_COVERAGE=40, OVLP_FAST_OPTIONS=-n 500 -z 20 -b 2000 -e 0.5 -j 0 -u 1 -a 1000, OVLP_SENSITIVE_OPTIONS=-n 500 -z 10 -e 0.5 -j 0 -u 1 -a 1000, CNS_FAST_OPTIONS=-a 2000 -x 4 -y 12 -l 1000 -e 0.5 -p 0.8 -u 0, CNS_SENSITIVE_OPTIONS=-a 2000 -x 4 -y 12 -l 1000 -e 0.5 -p 0.8 -u 0, TRIM_OVLP_OPTIONS=-n 100 -z 10 -b 2000 -e 0.5 -j 1 -u 1 -a 400, ASM_OVLP_OPTIONS=-n 100 -z 10 -b 2000 -e 0.5 -j 1 -u 0 -a 400, CNS_OUTPUT_COVERAGE=30; and 2) the final telomere-to-telomere (T2T) assembly used throughout this study, which was initially constructed using NextDenovo (v2.3.1). The centromeric regions were extracted from both the NECAT

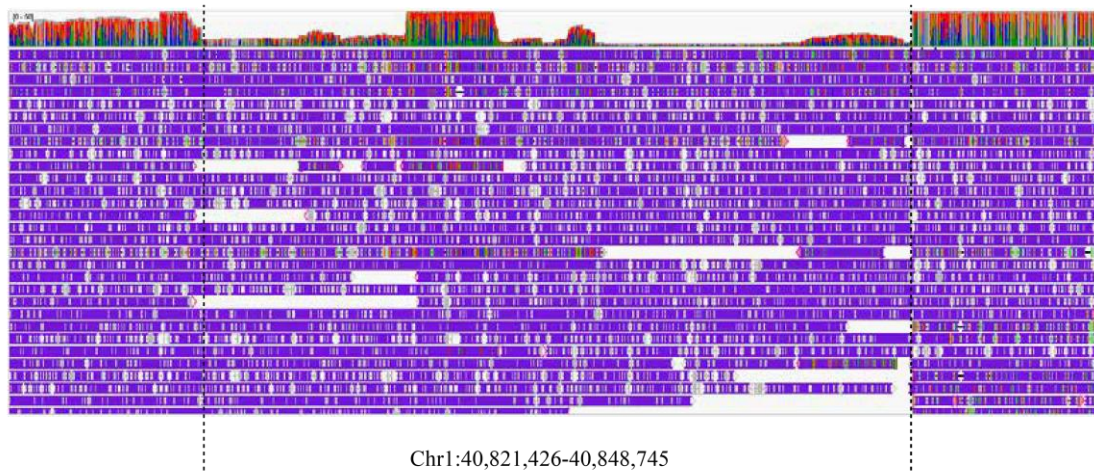
assembly and the final T2T assembly based on their annotated coordinates. To assess large-scale structural agreement, the centromeric sequence from the NECAT assembly (query) was aligned to the corresponding region from the final T2T assembly (target) using minimap2 (v2.24) with its default parameters. The resulting alignment was visualized as a dot plot using dotPlotly. The primary criterion for assessing assembly accuracy and continuity in these repetitive regions was the presence of continuous, unbroken diagonal lines across the entire plotted centromeric span, which indicates conserved collinearity and the absence of major structural gaps between the two independent assemblies.



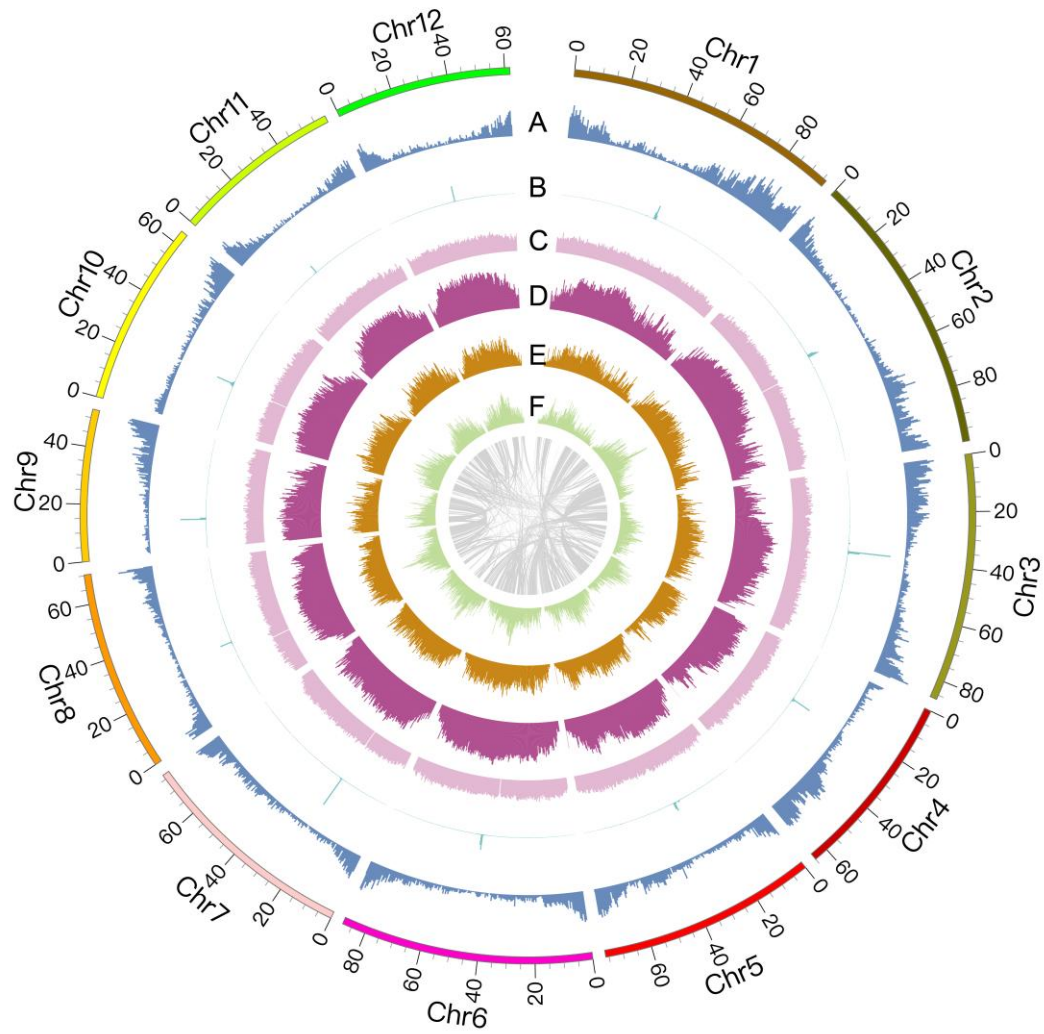
Supplementary Figure 1. ONT sequence length information and genome size and heterozygosity estimation. **a-b**, Bivariate plot of base call quality against percent identity with a kernel density estimate and marginal density plots. **a**: PromethION reads; **b**: Ultra-long reads. **c**, K-mer ($k = 21$) distribution of *O. australiensis*. The solid line and shadows represent the ratio of k-mer species with difference depths. Source data are provided as a Source Data file.



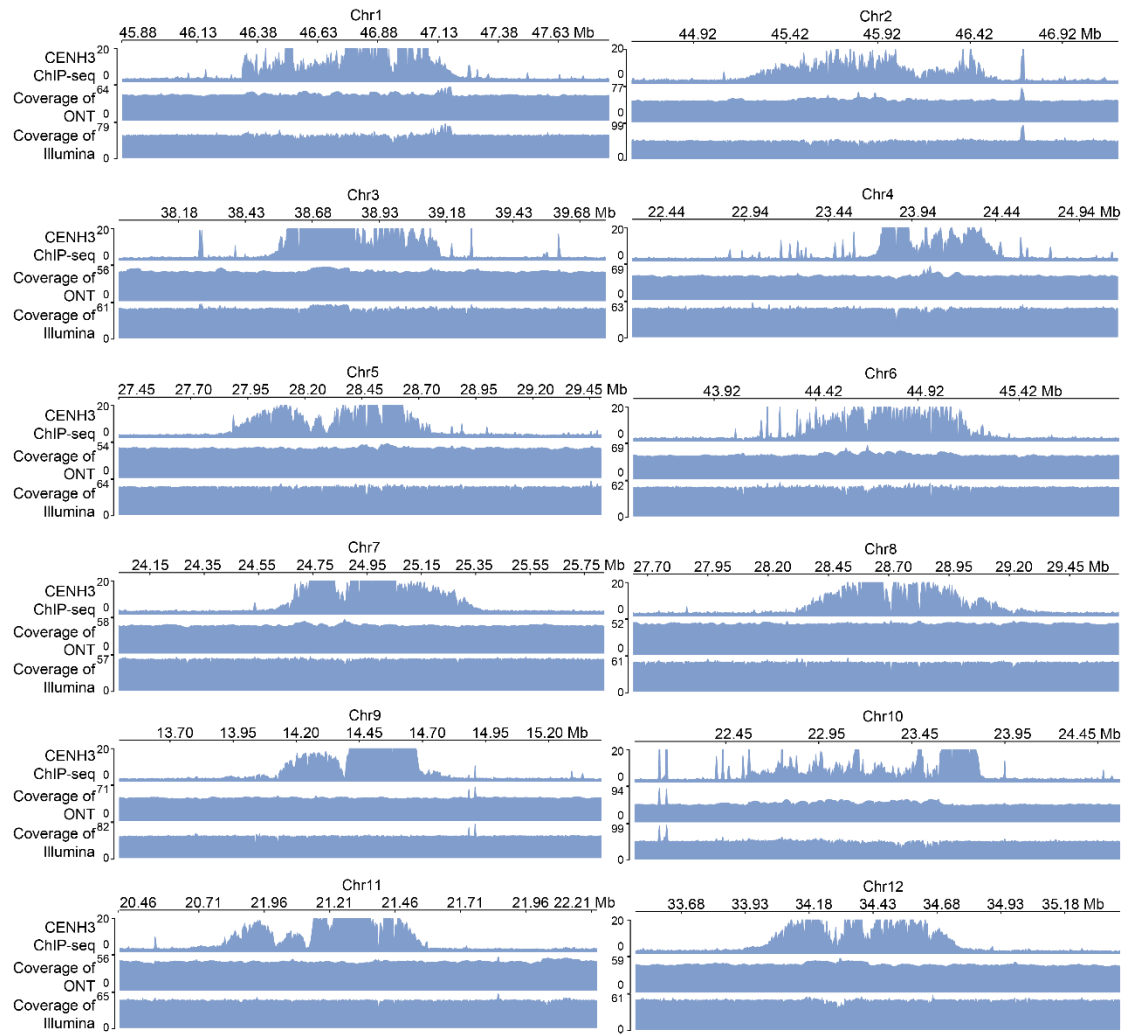
Supplementary Figure 2. The pipeline of genome assembly.



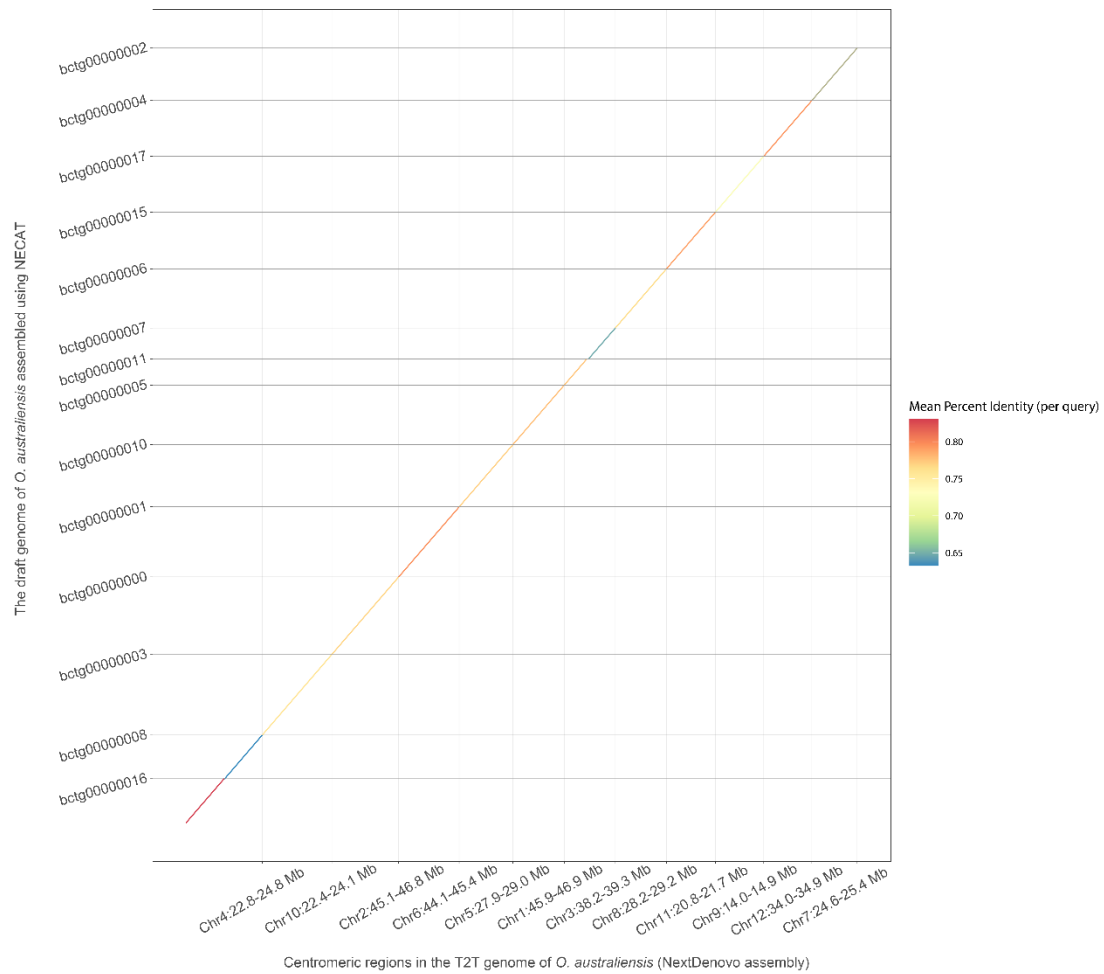
Supplemental Figure 3. Nanopore long reads coverage of closed gap regions in the *O. australiensis* genome.



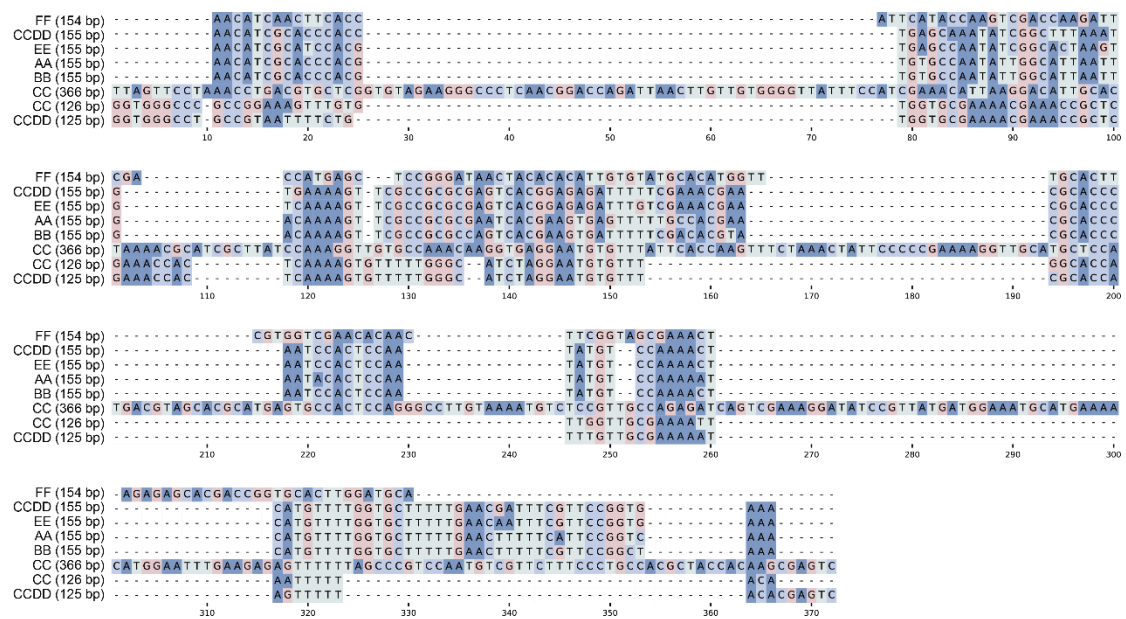
Supplementary Figure 4. Characteristics of the *O. australiensis* genome. The circle from outside to inside represents chromosome (A) gene density, (B) CENH3 ChIP-seq reads density, (C) GC content, (D) TE density, (E) *Gypsy* density, (F) *Copia* density, and gene collinearity connected by curved lines.



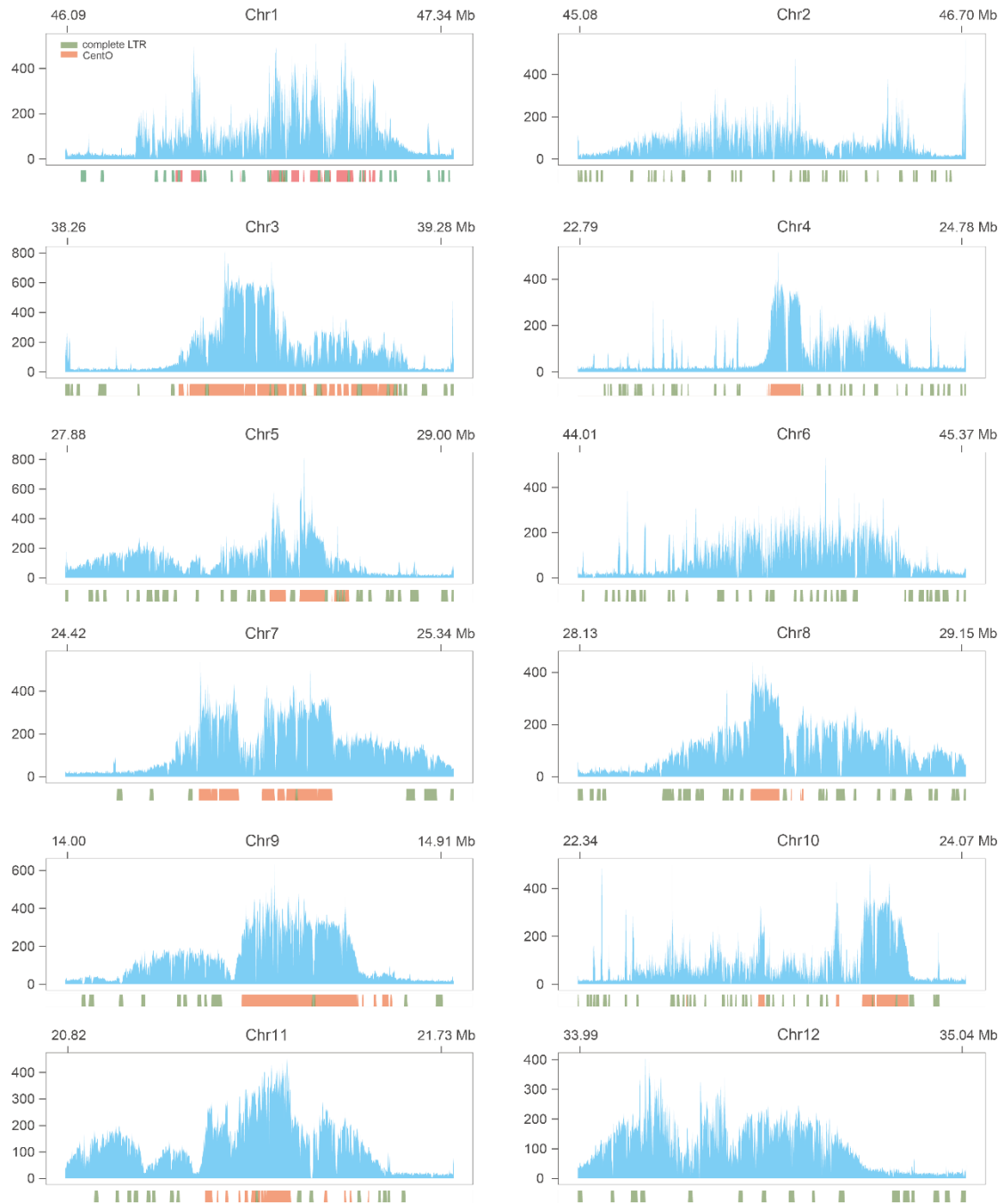
Supplementary Figure 5. Coverage of CENH3 ChIP-seq, ONT, and Illumina reads in the centromere on chromosome of *O. australiensis*.



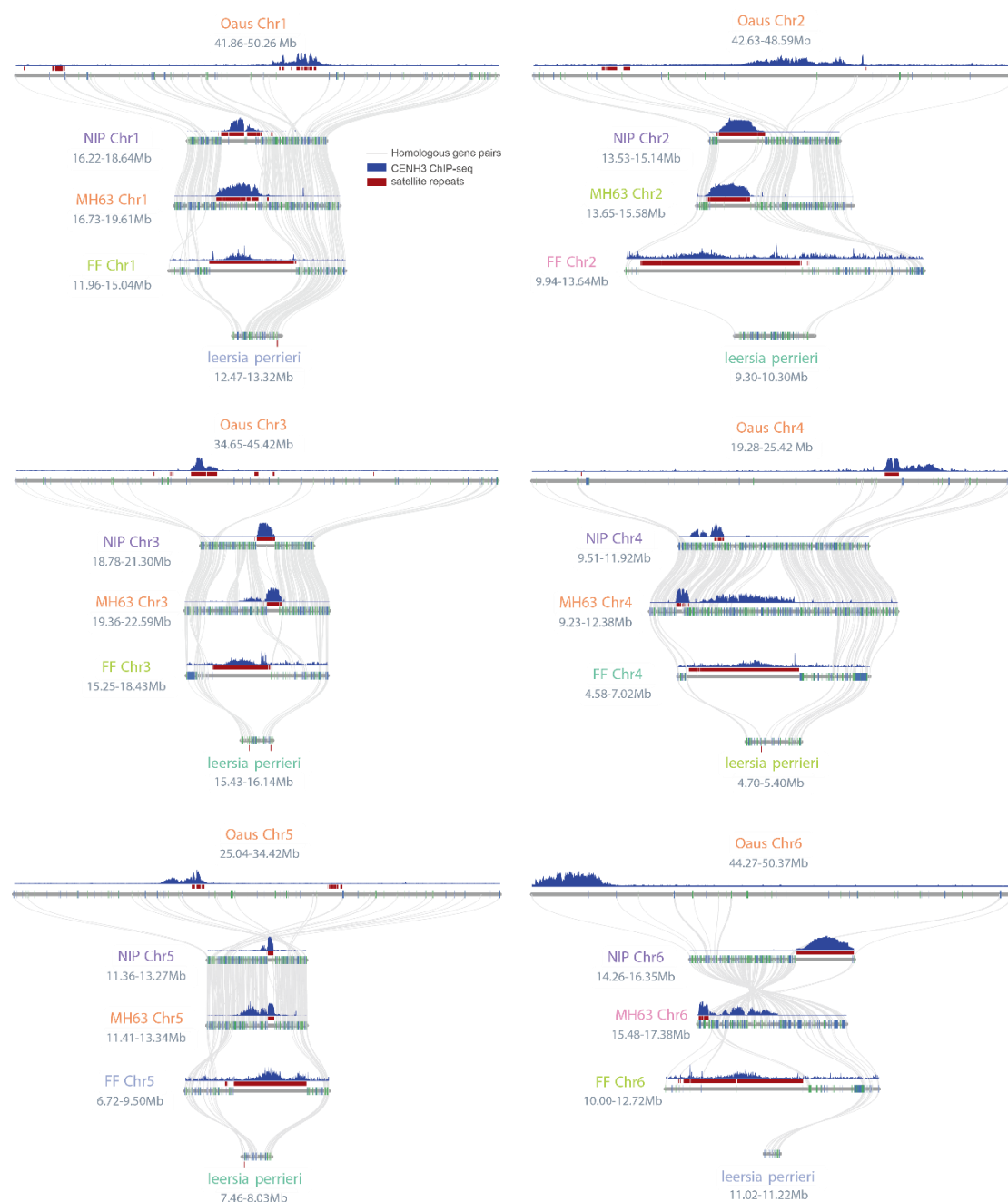
Supplementary Figure 6. Sequence alignment of centromere regions assembled by Necat and Nextdenovo software.



Supplementary Figure 7. Multiple alignment centromere satellite sequences of FF, EE, CCDD, CC, BB and AA.



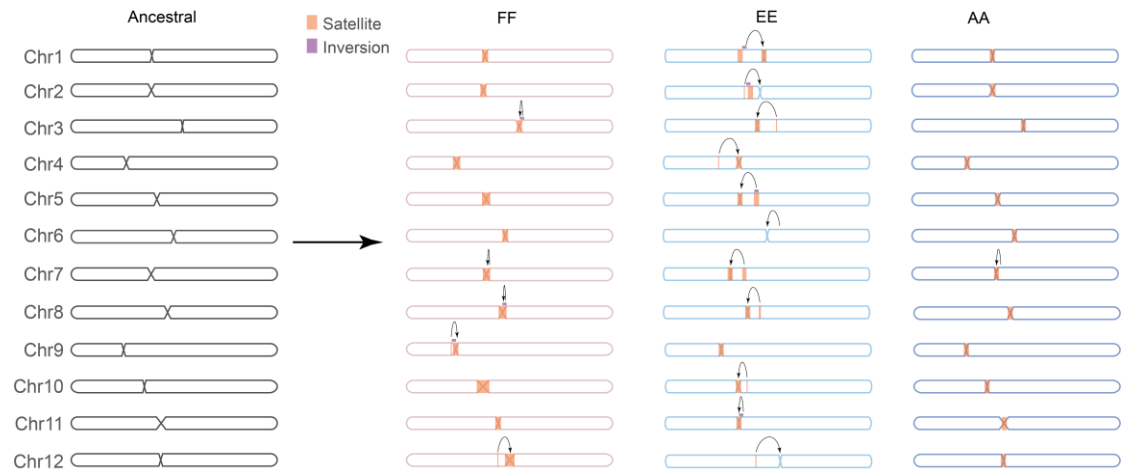
Supplementary Figure 8. Distribution of complete LTRs and satellite repeats in centromeric regions.



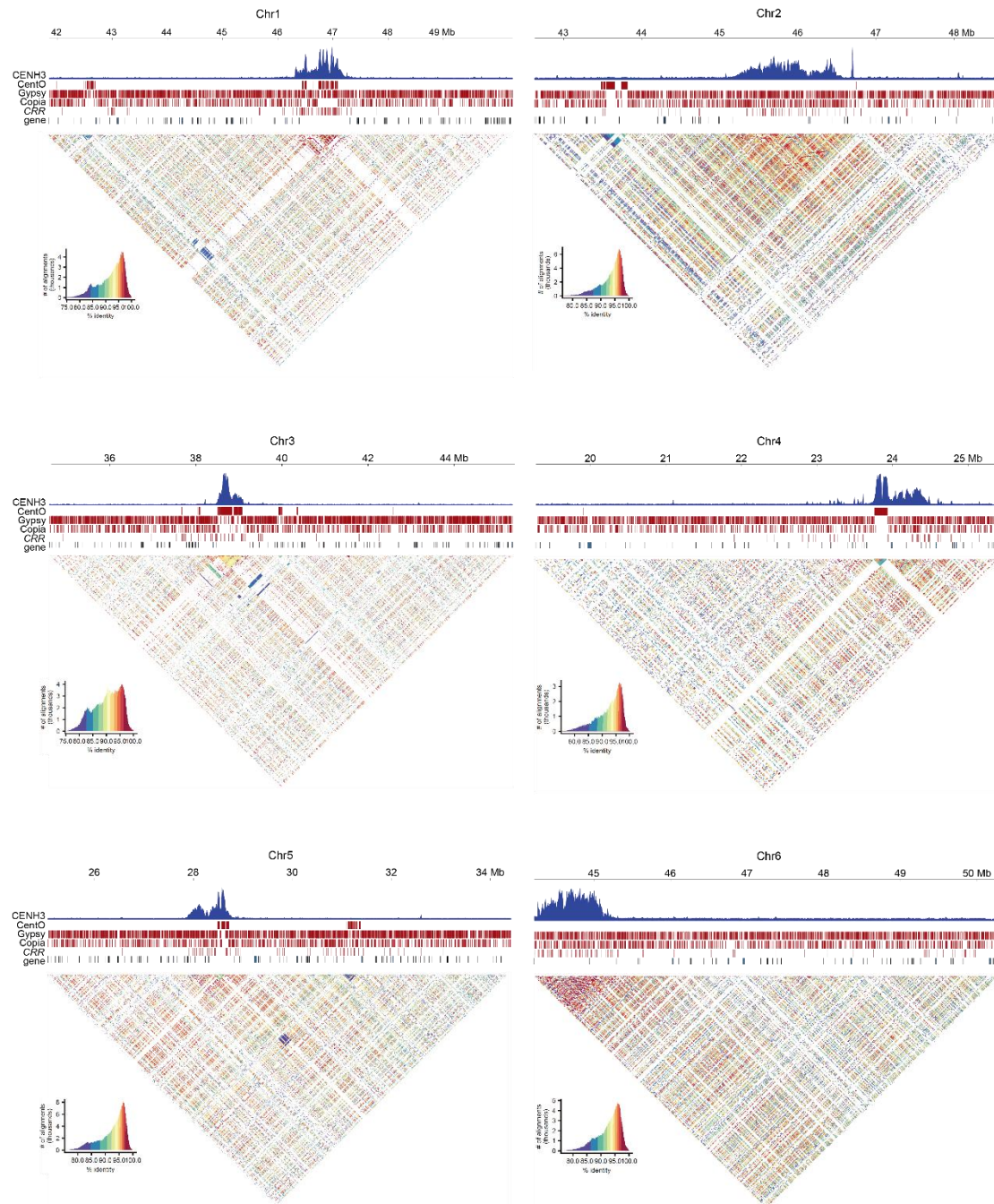
Supplementary Figure 9. The pairwise synteny visualization of centromere regions (Chr1-Chr6) among *O. australiensis*, *O. sativa* (NIP and MH63), *O. brachyantha* and *Leersia perrieri*. Source data are provided as a Source Data file.



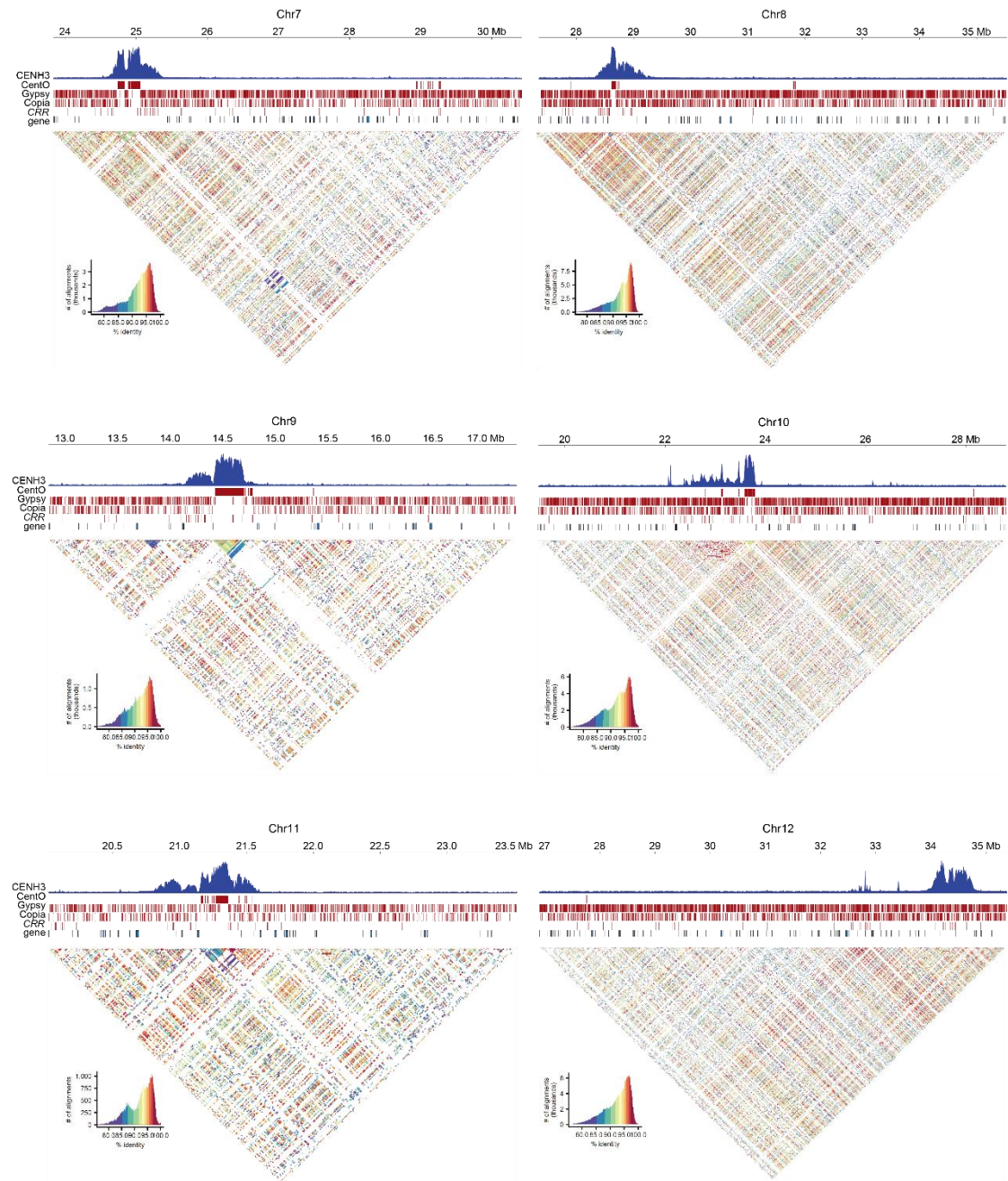
Supplementary Figure 10. The pairwise synteny visualization of centromere regions (Chr7-Chr12) among *O. australiensis*, *O. sativa* (NIP and MH63), *O. brachyantha* and *Leersia perrieri*. Source data are provided as a Source Data file.



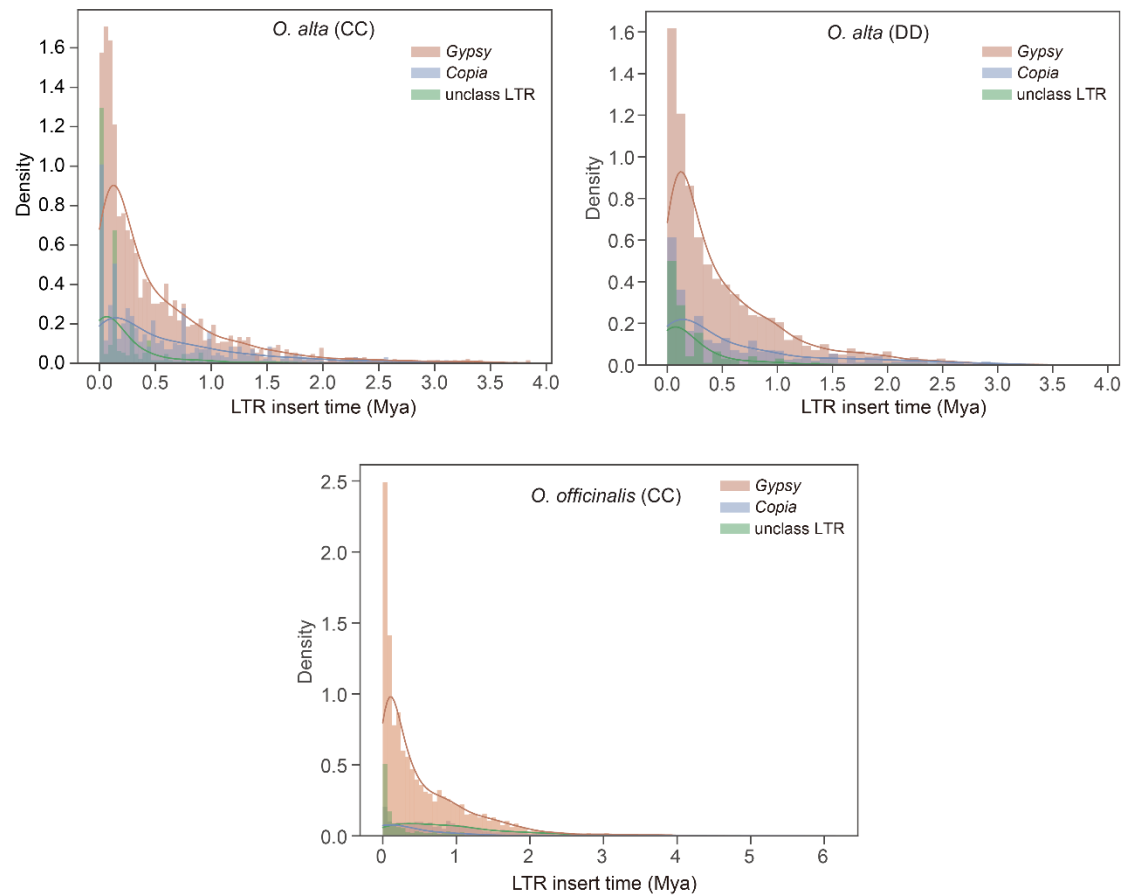
Supplementary Figure 11. Schematic diagram of the dynamic changes of centromere sites in *O. australiensis*, *O. sativa*, and *O. brachyantha*.



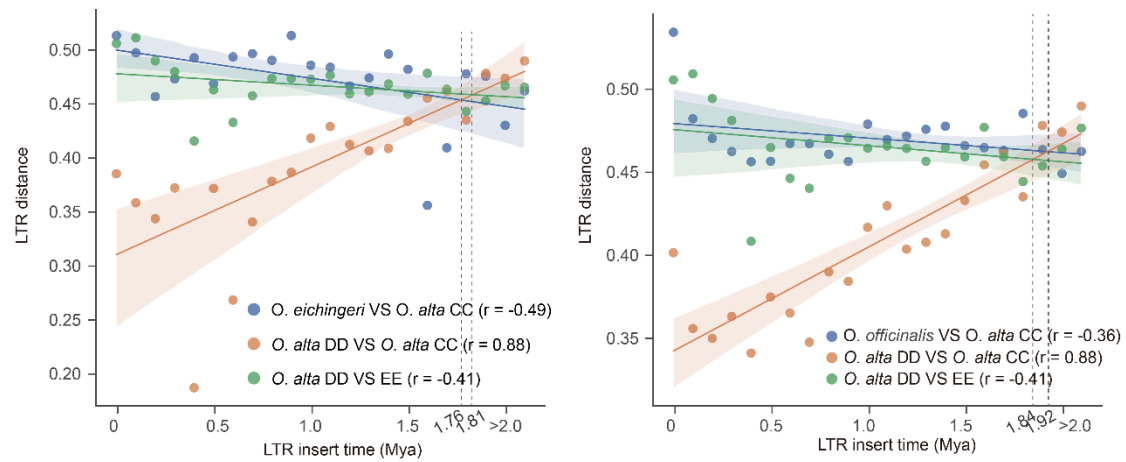
Supplementary Figure 12. Characteristics of centromeres across *O. australiensis* Chr1-Chr6, including the information on CENH3, satellites, *Gypsy*, *Copia*, *CRR* and gene.



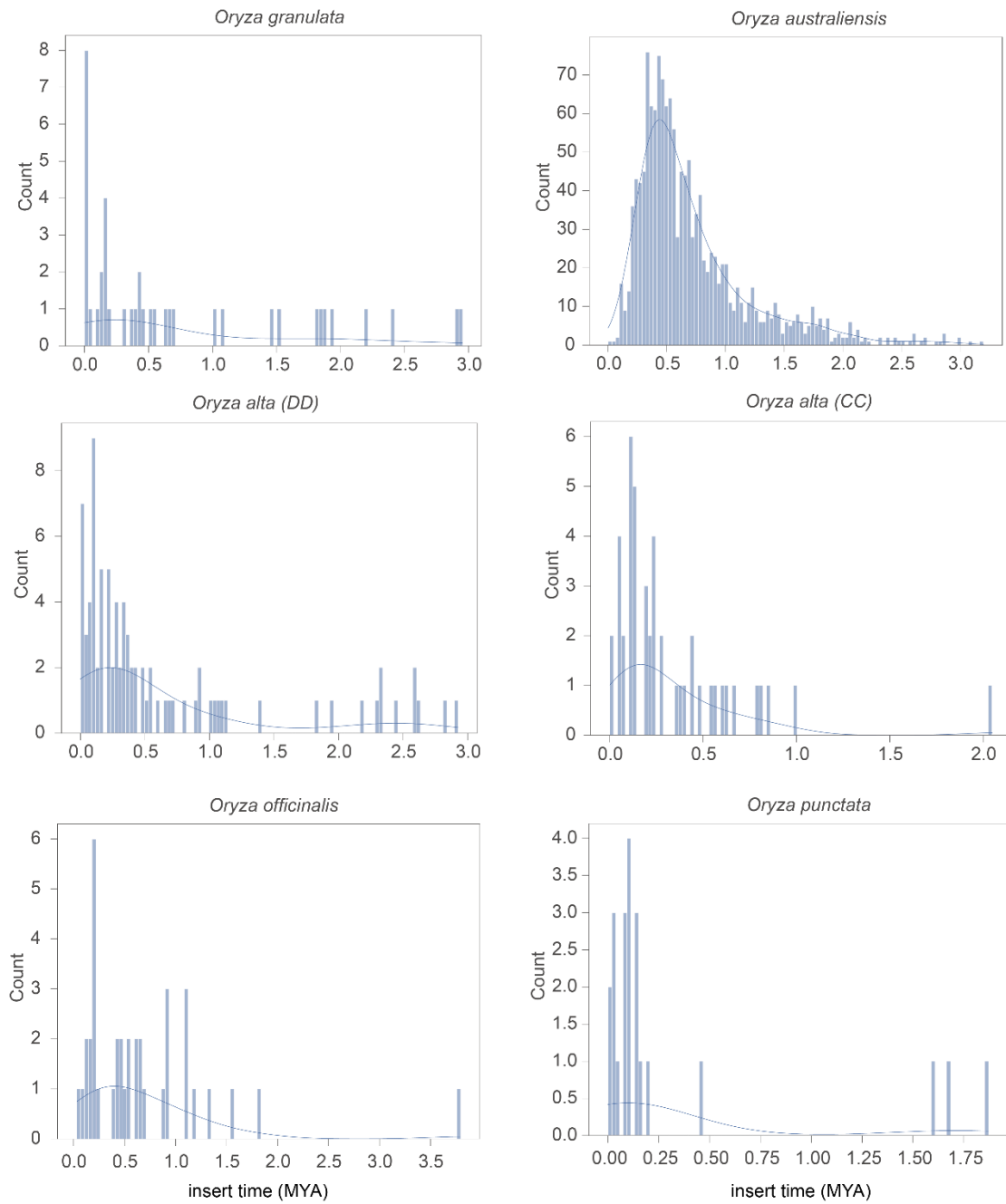
Supplementary Figure 13. Characteristics of centromeres across *O. australiensis* Chr7-Chr12, including the information on CENH3, satellites, *Gypsy*, *Copia*, *CRR*, and gene.



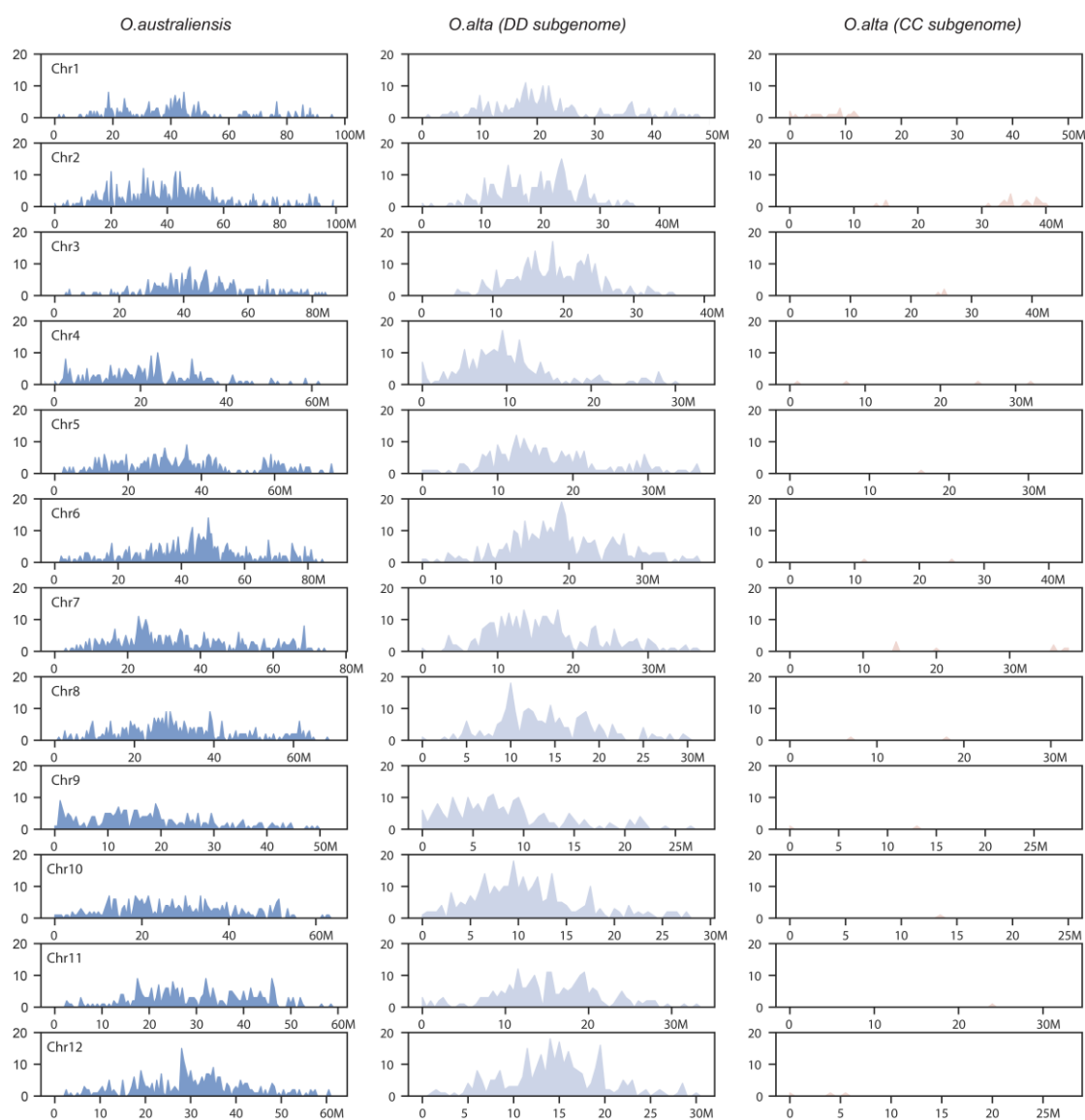
Supplementary Figure 14. LTR insertion time for CC and DD subgenomes, CC genome. Source data are provided as a Source Data file.



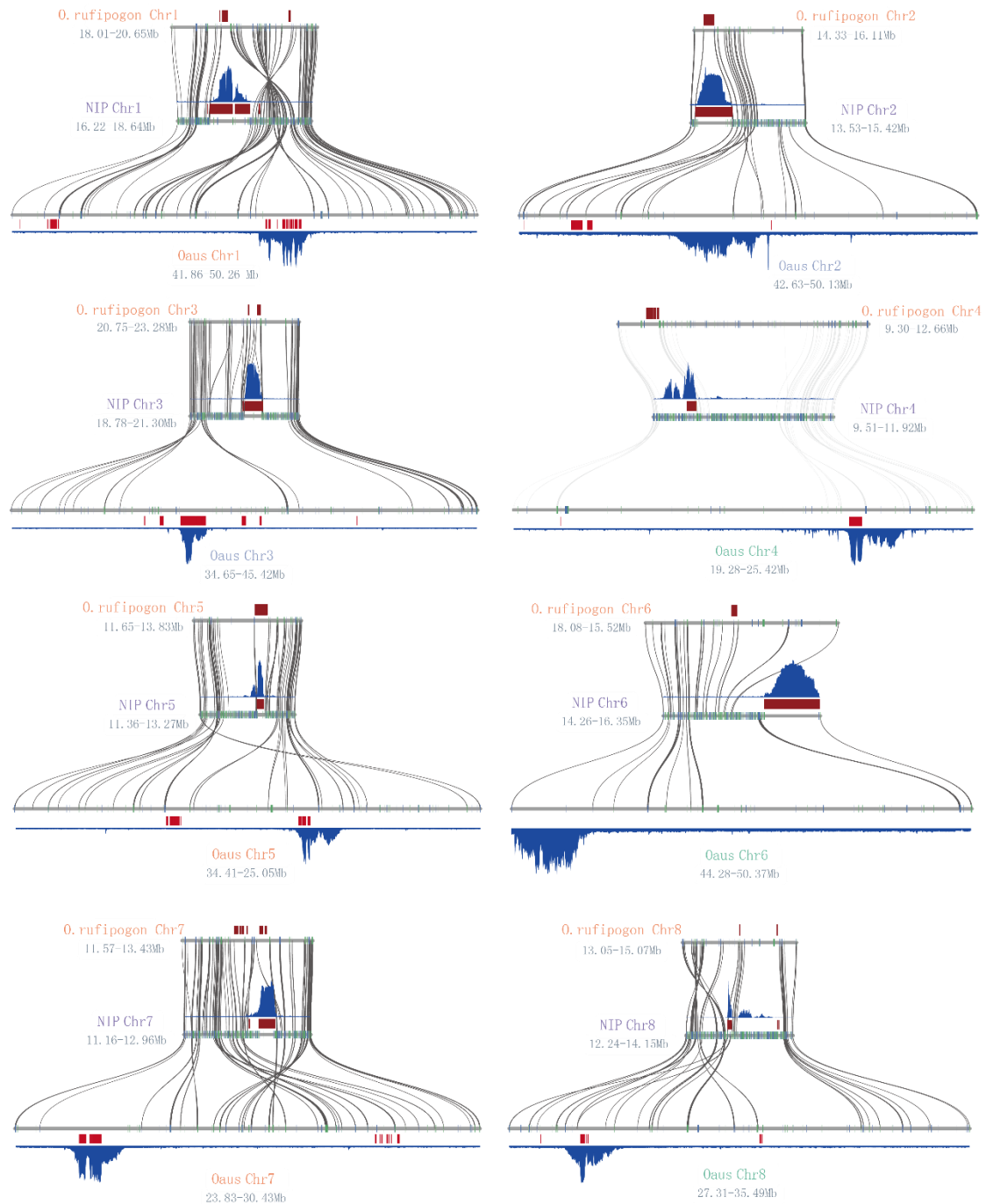
Supplementary Figure 15. The dynamic changes in the similarity of repetitive sequences between the polyploid genome and its progenitor genome over evolutionary time. “r” represents the Pearson correlation between LTR distance and LTR insertion time ($O. alta$ DD VS $O. alta$ CC: $P < 1 \times 10^{-12}$; $O. alta$ DD VS EE: $P < 0.02$; $O. eichingeri$ VS $O. alta$ CC: $P < 0.02$; $O. officinalis$ VS $O. alta$ CC: $P < 0.05$). Shaded bands indicate the 95% confidence interval of the linear regression. $n = 22$ LTR pairs per group. Source data are provided as a Source Data file.



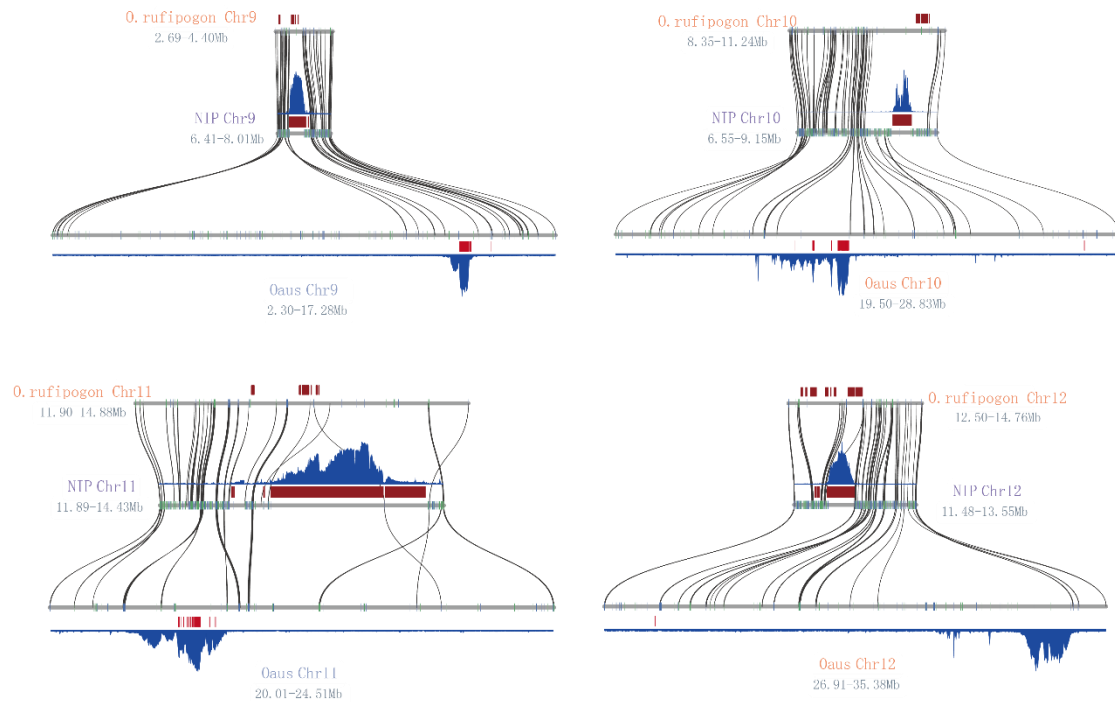
Supplementary Figure 16. *Angela* LTR insertion time for GG genome, EE genome, CC and DD subgenomes, CC genome, BB genome. Source data are provided as a Source Data file.



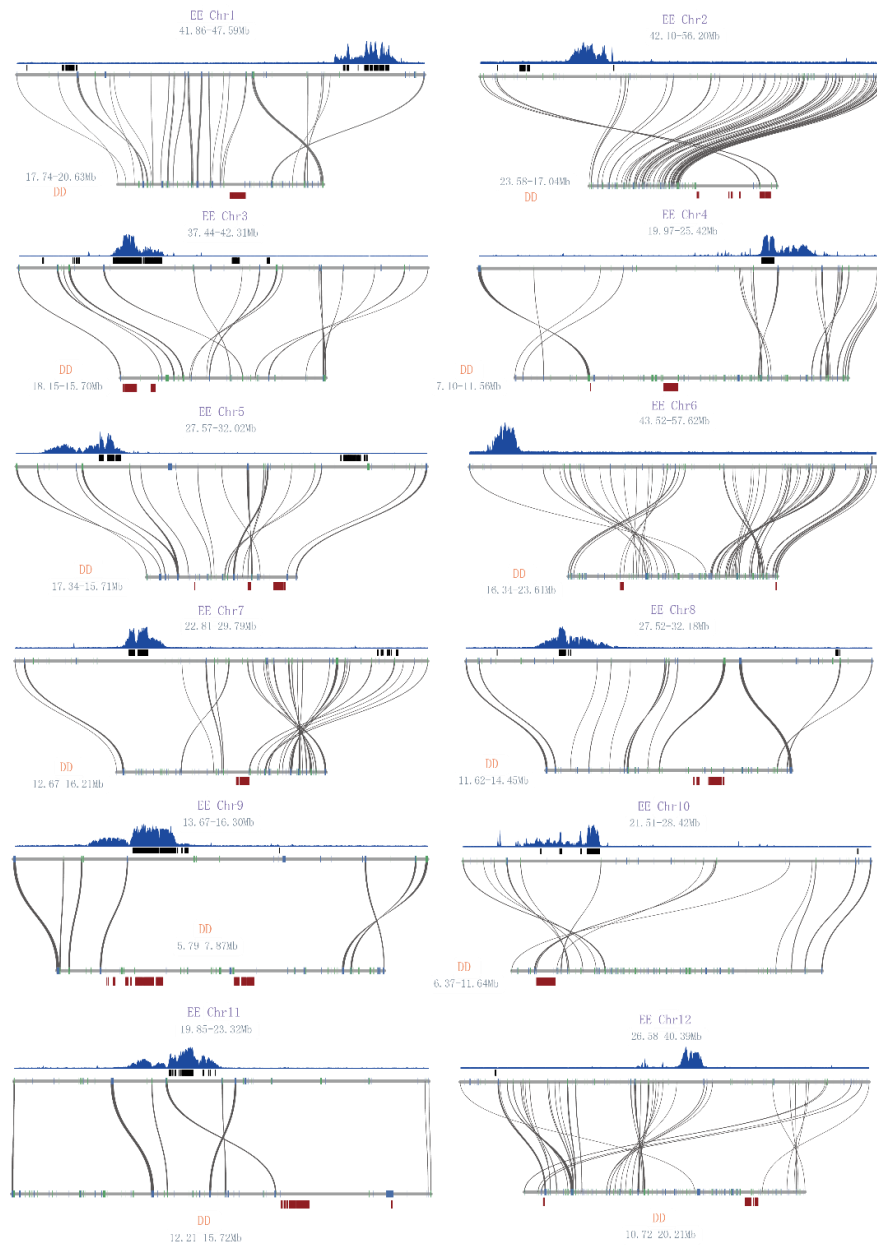
Supplementary Figure 17. Distribution of repeat sequence markers on the genomes of *O. australiensis* and *O. alta*. Source data are provided as a Source Data file.



Supplementary Figure 18. The pairwise synteny visualization of centromere regions (Chr1-Chr8) among *O. australiensis*, *O. sativa* (NIP) and *O. rufipogon*. Red lines represent the distribution of centromeric satellite repeats, while gray lines connect syntenic gene pairs to indicate homologous regions. Histograms (blue peaks) show the distribution of CENH3 ChIP-seq signals, defining functional centromeres in *O. sativa* and *O. australiensis*. Due to the absence of CENH3 data, centromere loci in *O. rufipogon* were inferred based on CentO enrichment. The analysis reveals conserved centromere positioning between *O. rufipogon* and *O. sativa* across Chromosomes 1–8, with notable positional shifts on Chr6. Source data are provided as a Source Data file.



Supplementary Figure 19. The pairwise synteny visualization of centromere regions (Chr9-Chr12) among *O. australiensis*, *O. sativa* (NIP) and *O. rufipogon*. Red lines represent the distribution of centromeric satellite repeats, while gray lines connect syntenic gene pairs to indicate homologous regions. Histograms (blue peaks) show the distribution of CENH3 ChIP-seq signals, defining functional centromeres in *O. sativa* and *O. australiensis*. Due to the absence of CENH3 data, centromere loci in *O. rufipogon* were inferred based on CentO enrichment. The analysis reveals conserved centromere positioning between *O. rufipogon* and *O. sativa* across Chromosomes 9 to 12, with notable positional shifts on Chr11. Source data are provided as a Source Data file.



Supplementary Figure 20. Comparative analysis of centromere positions between the diploid EE genome and the DD subgenome of *O. alta* (CCDD). For each of the 12 chromosomes, the centromeric regions are depicted. Black and red lines represent the distribution of the centromeric satellite repeat in the EE and DD subgenome, respectively. Grey lines connect syntenic gene pairs, indicating collinear regions. The comparative analysis reveals four distinct patterns: (1) positionally conserved centromeres (Chr 2, 9, 10); (2) positionally divergent centromeres (Chr 1, 4, 7, 11); (3) EE-specific neocentromere formation with a conserved ancestral CentO locus aligned to the DD centromere (Chr 3, 5, 8); and (4) independent neocentromere formation in both lineages with a shared ancestral CentO locus (Chr 6, 12). The prevalence of shared ancestral loci strongly suggests that the EE and DD genomes possessed congruent centromere architecture prior to their divergence, with observed differences likely driven by post-divergence, lineage-specific evolution. Source data are provided as a Source Data file.

Supplementary Table 1. The mapping rate of sequencing reads to assembled genome.

Platform	Total reads	Map reads	Map rate	Covered genome	Covered depth
Ultra-long ONT	368,744	368,740	100%	99.98%	22.65
PromethION	4,241,979	4,233,154	99.79%	99.97%	88.71
WGS	690,826,394	681,262,720	98.62%	99.12%	113.15
CENH3 ChIP-seq	68,427,736	67,719,140	98.96%	/	/
Input	56,262,947	56,075,605	99.67%	/	/

Supplementary Table 2. Comparison of genome assembly metrics between this study and previously published.

	This study	Long <i>et al.</i>¹⁹	Fornasiero <i>et al.</i>²⁰	Phillips <i>et al.</i>²¹	Morrison <i>et al.</i>²²
Gaps	0	7	136	838	4
LAI	23.79	22.9	23.77	18.35	21.55
QV	41.76	57.62	36.99	25.18	-
BUSCO	99.8	99.8	99.5	99.3	99.3
N50	76.6 Mb	78.1 Mb	73.8 Mb	2.6 Mb	76.6 Mb
N90	59.5 Mb	56.3 Mb	52.5 Mb	437 Kb	61.6 Mb
L50	6	6	6	83	6
L90	11	11	12	401	11
Number of contigs	12	16	60	850	13

Note: In the genome assembly project of *O. australiensis* completed by Morrison *et al.*²², the QV is not provided due to the lack of Illumina whole-genome sequencing data.

Supplementary Table 3. Statistics of *O. australiensis* genome annotation.

Genome feature	Statistic
Number of genes	36742
Annotation genes	33426
Number of non-TE gene loci	33565
Gene length (Mb)	110.3
Average lengths of genes	2999 bp
Average lengths of proteins	362 aa
BUSCO genes (%) (transcript)	97.2%

Supplementary Table 4. Repetitive elements in *O. australiensis* genome assembly.

Class	Superfamily	Abbr.	Number	Total length (Mb)	Repeat percentage (%)
Class I	Gypsy	RLG	396,423	338.0	47.13
	Copia	RLC	114,241	100.4	13.99
	Unclassified LTR	RLX	311,487	217.7	30.35
	LINE	RIX	1,984	1.2	0.16
	SINE	RSX	1,847	0.26	0.036
	Retroelement	RXX	2,129	0.8	0.11
Class II	En-Spm	DTC	71,971	41.1	5.73
	Harbinger	DTH	1,960	0.57	0.079
	hAT	DTA	9,298	3.7	0.51
	Mite	DXX	15,069	2.2	0.30
	MuDR	DTM	3,734	3.6	0.50
	Stowaway	DXX	338	0.03	0.0041
	TcMar	DTT	281	0.05	0.0069
	Tourist	DXX	5,380	0.8	0.11
	Helitron	DHH	296	0.05	0.0069
	Unclassified	DTX	16,217	3.6	0.50
	microsatellite	XXX	156,698	7.7	1.07
	satellite	XXX	14,349	2.1	0.30
Others		XXX	10,119	3.1	0.43

Supplementary Table 5. Identified regions of telomeres in *O. australiensis*.

Chr	Short arm start (bp)	Short arm end (bp)	Length (bp)	Long arm start (bp)	Long arm end (bp)	Length (bp)	TTTAGGG length (short arm, long arm)	CCCAGG length (short arm, long arm)	TTTAAAA length (short arm, long arm)
Chr1	1	24500	24500	97036611	97063211	26600	(9002, 17577)	(18, 30)	(497, 70)
Chr2	1	15500	15500	99592022	99621822	29800	(14441, 1953)	(42, 10626)	(140, 63)
Chr3	1	25900	25900	86430823	86447723	16900	(23135, 14994)	(72, 18)	(84, 210)
Chr4	1	21800	21800	65027376	65050376	23000	(14637, 19621)	(72, 30)	(49, 182)
Chr5	1	17300	17300	76622403	76633803	11400	(14896, 2142)	(132, 3204)	(91, 70)
Chr6	1	30800	30800	88672772	88690072	17300	(18809, 15631)	(72, 18)	(77, 182)
Chr7	1	34400	34400	76776646	76793046	16400	(2471, 14686)	(9204, 48)	(105, 140)
Chr8	1	33100	33100	70179292	70200392	21100	(26579, 18270)	(102, 84)	(49, 112)
Chr9	1	6500	6500	53109967	53143067	33100	(2086, 28210)	(138, 24)	(1330, 119)
Chr10	1	14100	14100	65752900	65786800	33900	(13713, 31073)	(48, 6)	(147, 84)
Chr11	1	38700	38700	59422878	59455478	32600	(34748, 22862)	(42, 24)	(0, 6748)
Chr12	1	27700	27700	61902462	61920362	17900	(18781, 17675)	(96, 42)	(56, 217)

Supplementary Table 6. Identified regions of active and inactive centromeres in *O. australiensis*.

Chromosome	Start (active)	End (active)	Length (active)	Start (inactive)	End (inactive)	Length (inactive)
Chr1	46,097,419	47,340,910	1,243,491	42,505,679	42,711,648	205,969
Chr2	45,082,502	46,702,041	1,619,539	43,480,322	43,832,213	351,891
Chr3	38,266,308	39,284,242	1,017,934	39,976,941	40,064,465	87,524
Chr4	22,792,056	24,785,959	1,993,903			
Chr5	27,889,253	29,002,464	1,113,211	31,076,137	31,367,011	290,874
Chr6	44,015,182	45,371,557	1,356,375			
Chr7	24,425,980	25,348,116	922,136	28,934,089	29,285,826	351,737
Chr8	28,136,910	29,150,676	1,013,766	31,764,899	31,808,779	43,880
Chr9	14,008,332	14,919,937	911,605			
Chr10	22,343,345	24,076,763	1,733,418			
Chr11	20,822,886	21,737,080	914,194			
Chr12	33,997,897	35,044,144	1,046,247	27,773,331	27,780,000	6,669

Supplementary Table 7. Satellite repetitive sequences in the centromeric region of the *Oryza* genus.

Genus	Satellite repetitive sequences
EE	AACATCGCATCCACGTGAGCCAATATCGGCACTAAGTGTCAAAA GTTCGCCGCGCGAGTCACGGAGAGATTTGTCGAAACGAACGCA CCCAATCCACTCCAATATGTCCAAAACATCATGTTTTGGTGCTTTT TGAACAATTTTCGTTCCGGTGAAA
FF	GACTATGCACACCGAGATAACTACACATATTGTGCGTCCACTTGG TTTACACTTCGTGGTTCGAACACAACCTTCGGTAGCAAAACTCCCC AGCATGAAGAGTACACTTGGATTCCACCATCAACTTCACCTTTC ATGCAATGGCAACAAAGATTC
CCDD	AACATCGCACCCACCTGAGCAAATATCGGCTTTAAATGTGAAAA GTTCGCCGCGCGAGTCACGGAGAGATTTTTTCGAAACGAACGCA CCCAATCCACTCCAATATGTCCAAAACATCATGTTTTGGTGCTTTT TGAACGATTTTCGTTCCGGTGAAA
CCDD	GGTGGGCCTGCCGTAATTTTCTGTGGTGCGAAAACGAAACCGCT CGAAACCACTCAAAAGTGTTTTTGGGCATCTAGGAATGTGTTTC GCACCATTGTTGCGAAAAATAGTTTTTACACGAGTC
CC	GGTGGGCCCCGCCGAAAGTTTGTGTGGTGCGAAAACGAAACCG CTCGAAACCACTCAAAAGTGTTTTTGGGCATCTAGGAATGTGTT TGGCACCATTTGGTTGCGAAAATTAATTTTTACACGAGTC
CC	AGTCGAAACGACATCCGTTATGATGGAAATGCATGACAACATGG AATTTGAAGAGAGTTTTTTAGCCCGTCCAATGTCGTTCTTTCCCT GCCACGCTACCACAAATTAGTTCCTAAACCTGACGTGCTCGGTG TAGAAGGGCCCTCAACGGACCAGATTAACTTGTTGTGGGGTTAT TTCCATCGAAACTTTAAGGACATTGCCCTAAAACGCATCGCTTAT CCAAAGGTTGTGCCAAACAAGGTGAGGAATGTGTTTATTCACCA AGTTTCTAAACTTTTCCCCTGAAAAGGTTGCATGCTCCATGACGT AGCACGCATGAGTGCCACTCCAGGGCCTTGTAGAATGTCTCCGT TGCCAGATATC
BB	AACATCGCACCCACGTGTGCCAATATTGGCATTAAATTGACAAAA GTTCGCCGCGCCAGTCACGAAGTGATTTTTCGACACGTACGCAC CCAATCCACTCCAATATGTCCAAAACATCATGTTTTGGTGCTTTTT GAACTTTTTTCGTTCCGGCTAAA

Supplementary Table 8. The total length of satellite repetitive sequences in *O. australiensis* and *O. sativa*.

Chromosome	EE	AA (NIP)	AA (MH63)
Chr1	241,639	429,115	475,620
Chr2	156,120	405,318	457,714
Chr3	437,876	215,446	146,711
Chr4	139,630	42,969	55,356
Chr5	167,034	85,443	87,529
Chr6	133	678,325	74,754
Chr7	243,250	150,254	458,364
Chr8	86,778	48,273	49,913
Chr9	261,722	374,089	84,963
Chr10	225,206	213,438	45,768
Chr11	78,884	1,208,069	285,500
Chr12	4,545	400,245	151,980
Total	2,038,272	4,250,984	2,374,172

Supplementary Table 9. Proportions of TEs in the genomes of *Oryza*.

Species	other DNA	DNA/En-Spm	LINE	LTR	LTR/Copia	LTR/Gypsy	RC/Helitron	Retroelement	SINE
<i>O. rufipogon</i>	6.21482738	5.4313746	0.26818129	52.7658028	6.71735982	27.9949236	0.1755121	0.15279434	0.27922412
<i>O. nivara</i>	11.0065866	4.13071028	0.49802682	55.9960438	8.39993185	19.0990915	0.1115764	0.22465209	0.53338079
<i>O. meridionalis</i>	6.67273127	7.81020095	0.32686286	54.178829	8.2728488	22.2024213	0.06603258	0.15509258	0.31498069
<i>O. punctata</i>	5.70681207	10.5145728	0.47593851	38.412829	9.02884551	35.3869825	0.0588278	0.19563633	0.21955555
<i>O. brachyantha</i>	18.748716	3.01634821	0.08063069	51.7891097	10.7768174	14.6676298	0.05920277	0.15779948	0.70374591
<i>O. eichingeri</i>	5.19618756	9.43216025	0.10884865	39.3203454	9.30223484	36.0670003	0.08758111	0.18451346	0.30112837
<i>O. granulata</i>	4.76337678	0.8891655	0.20821331	44.5909101	7.67969986	41.0887907	0.05755343	0.68183177	0.04045853
<i>O. sativa</i> (NIP)	7.38867763	5.79873609	0.34334465	54.6950082	8.00685717	23.0629743	0.11987221	0.25274192	0.33178784
<i>O. rhizomatis</i>	4.55400025	9.21331469	0.0762061	36.4619389	8.76280287	40.4923483	0.04753283	0.16811221	0.22374391
<i>O. sativa</i> (MH63)	6.76041125	4.85792346	0.30490106	54.8360297	6.94680864	25.6784833	0.11118285	0.20079313	0.30346653
<i>O. glumipatula</i>	11.0474746	3.71014613	0.55141266	56.1219751	8.20051781	19.4860741	0.09769709	0.24402149	0.54068097
<i>O. barthii</i>	8.76726192	5.37521363	0.43210572	54.3882984	9.34976545	20.9672136	0.09426545	0.21145896	0.41441684
<i>O. glaberrima</i>	9.1016339	5.48818016	0.43690876	54.3425109	9.11320786	20.7841943	0.09684859	0.20742571	0.42908981
<i>O. longistaminata</i>	12.610028	3.78062359	0.59453128	58.1572625	8.11461902	15.854799	0.10224474	0.12945452	0.65643733
<i>O. alta</i> (CC)	4.80812507	12.6436245	0.10046657	35.9016779	10.3472037	35.681671	0.03491565	0.23568539	0.24663024
<i>O. alta</i> (DD)	3.57169812	8.95471028	0.12271083	32.3535158	13.9274974	40.7488348	0.01640029	0.1632557	0.14137671
<i>O. australiensis</i>	1.98557464	6.33486937	0.18503048	28.6087934	15.2131523	47.5466409	0.00741482	0.08574884	0.03277532
<i>O. officinalis</i>	4.41786263	9.15462197	0.06724743	36.0831698	6.85890185	43.0611076	0.03779713	0.11175103	0.20754053
<i>O. minuta</i> (BB)	5.97878168	10.8785192	0.38851426	40.6440174	9.71523861	32.032292	0.0410981	0.15256502	0.16897374
<i>O. minuta</i> (CC)	4.86761965	9.58759795	0.07881134	35.7184548	10.3040614	39.0718549	0.04274257	0.11660438	0.21225295

Supplementary Table 10. Abundance of *Copia* and *Gypsy* LTR subfamilies in the *Oryza* genus.

	Copia							Gypsy					
	Bianca	Angela	Tork	Ale	Ikeros	SIRE	Ivana	Athila	Reina	Ogre	Retand	Tekay	CRM
<i>O. australiensis</i>	27	1390	13	65	92	100	50	3	59	468	1478	989	125
<i>O. alta</i> DD	21	85	47	51	66	4	45	0	31	657	22	109	34
<i>O. alta</i> CC	27	46	55	56	64	17	54	0	20	585	47	87	21
<i>O. granulate</i>	14	39	0	95	68	42	85	0	101	305	743	135	130
<i>O. officinalis</i>	35	39	114	84	139	194	211	3	29	2090	1037	744	300
<i>O. punctata</i>	15	22	17	36	32	40	20	0	14	197	240	93	47
<i>O. rhizomatis</i>	6	5	22	18	32	39	43	0	3	73	117	10	17
<i>O. sativa</i>	14	3	17	118	52	53	82	3	111	52	574	354	37
<i>O. eichingeri</i>	4	3	12	31	10	13	24	2	1	5	149	6	16
<i>O. meridionalis</i>	8	2	15	27	17	18	34	5	24	38	86	298	14
<i>O. glaberrima</i>	4	0	13	29	6	25	26	1	27	38	65	35	7
<i>O. longistaminata</i>	0	0	2	7	4	1	2	0	14	2	5	1	0
<i>O. rufipogon</i>	8	0	13	45	37	21	34	1	64	30	612	500	24
<i>O. nivara</i>	1	0	2	19	0	0	2	0	21	1	2	0	0
<i>O. barthii</i>	0	0	0	25	0	2	6	0	28	4	11	0	0
<i>O. brachyantha</i>	0	0	2	7	0	1	2	0	7	0	1	0	0
<i>O. glumipatula</i>	0	0	0	7	1	0	1	1	11	0	15	0	0

Supplementary Table 11. *Angela* LTRs inserted prior to 1.7 Mya in the DD subgenome and their syntenic counterparts in the EE genome.

DD genomic location	EE genomic location	LTR integrity status
ChrD3:23350000-23357133	Chr3:51407729-51408018	Non-intact
ChrD4:7719158-7727244	Chr4:20268737-20277171	Intact
ChrD4:9966132-9975565	Chr4:17586300-17595273	Intact
ChrD5:14028699-14035959		
ChrD5:16178273-16186361	Chr5:29958153-29958441	Non-intact
ChrD6:12717174-12724985		
ChrD8:9808543-9824011	Chr8:23573962-23574148	Non-intact
ChrD10:7584126-7591865	Chr10:21855593-21855882	Non-intact
ChrD12:7645460-7653337	Chr12:15475537-15483961	Intact
ChrD12:18123460-18125326		
ChrD12:19947852-19955789	Chr12:37493762-37494050	Non-intact
ChrD10:14352701-14359150		

Supplementary Table 12. The sequence of *Angela*-Rep-DDEE.

	Sequence
<i>Angela</i> -Rep-DDEE	ATATGTGTCCATATTATCGATGTGCTATCTCCATACCATGTTCTAATG AGACATGATCATCAATCAATACATGTGCTGATCGAGTGGTCATATT GTTTCATATATGATATATGATCAGGGATCATTTAGACATAATAGCATA ACATAGATAGTCTCATGAATGAGTCACGTACTCATTGAACAATAA AGTTACTTATATCAAGGAACACTTTAAAATATATAAACATAGTTTG TGATACAACCTCTCCATACTGTCTCTAGGGCATATCACTTACA

Supplementary Table 13. Statistics of evolutionary rate in CC, EE and CCDD genomes.

Ingroup1	Ingroup2	Outgroup	Genes	Identical	Divergent	Ingroup1 specific	Ingroup2 specific	Outgroup specific	Fast-evolving	CHI ² statistics	test	P- value
<i>O. alta</i> CC	<i>O. alta</i> DD	<i>B. distachyon</i>	1,460	572,828	10,714	13,205	12,994	114,973	equality	1.7		0.19
<i>O. alta</i> CC	EE	<i>B. distachyon</i>	1,460	596,932	10,377	13,845	10,547	119,060	<i>O. alta</i> CC	445.92		0
<i>O. alta</i> DD	EE	<i>B. distachyon</i>	1,460	600,837	6,050	9,691	6,555	128,950	<i>O. alta</i> DD	605.35		0
<i>O. alta</i> CC	<i>O. eichingeri</i>	<i>B. distachyon</i>	1,460	586,027	4,305	7,789	5,473	129,088	<i>O. alta</i> CC	404.45		0
<i>O. alta</i> CC	<i>O. rhizomatis</i>	<i>B. distachyon</i>	1,460	583,766	4,985	7,742	7,434	128,538	<i>O. alta</i> CC	6.25		0.012

Note: Statistical significance of rate differences was assessed using Tajima's relative rate test with a two-sided χ^2 test. *P* values were determined with 1,000 bootstrap replicates. No adjustments were made for multiple comparisons.

Supplementary Table 14. *Angela*-Rep-DDEE sequence primers.

Primer	Sequence (5'-3')
Primer1	TCATGCACATCCCACTCCATGT
Primer2	ACCCACCACGAGACACTACA

Supplementary references

1. Ou S, Jiang N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol* **176**, 1410-1422 (2018).
2. Cock PJ, *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423 (2009).
3. Xu GC, *et al.* LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *Gigascience* **8**, giy157 (2019).
4. Mc Cartney AM, *et al.* Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nat Methods* **19**, 687-695 (2022).
5. Hu J, Fan J, Sun Z, Liu S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253-2255 (2020).
6. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212 (2015).
7. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245 (2020).
8. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
9. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
10. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907-915 (2019).
11. Zhang RG, *et al.* TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic Res* **9**, uhac017 (2022).
12. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290-295 (2015).
13. Gremme G, Brendel V, Sparks ME, Kurtz S. Engineering a software tool for gene structure prediction in higher organisms. *Inform Software Tech* **47**, 965-978 (2005).
14. Haas BJ, *et al.* Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7 (2008).
15. Jones P, *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014).
16. Cantalapiedra CP, 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 (2021).

17. 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 (2016).
18. Hao Z, *et al.* RIdiogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Comput Sci* **6**, e251 (2020).
19. Long W, *et al.* Genome evolution and diversity of wild and cultivated rice species. *Nat Commun* **15**, 9994 (2024).
20. Fornasiero A, *et al.* *Oryza* genome evolution through a tetraploid lens. *Nat Genet* **57**, 1287–1297 (2025).
21. Phillips AL, *et al.* The first long-read nuclear genome assembly of *Oryza australiensis*, a wild rice from northern Australia. *Sci Rep* **12**, 10823 (2022).
22. Morrison S, Masouleh AK, Constantin L, Panaud O, Wing RA, Henry RJ. The *Oryza australiensis* genome reveals potential as a source of genes for rice improvement. *Scientific Reports* **15**, 27826 (2025).