scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A complete telomere-to-telomere assembly of *Oryza sativa* L. subsp. indica Kato T197 genome

Tianyuan Zhang[1,2,5], Yi Zhang[1,3,5], Xiao Tang[1,3], Wenyan Peng[1,3], Hongjun Xie[1,3], Yangqin Xie[4], Yuchen Zhao[4], Lulu Yang[4], Yinghong Yu[1,3] ✉ & Mingdong Zhu[1,3] ✉

Here, we report the complete Telomere-to-Telomere (T2T) genome of *Oryza sativa* L. subsp. indica Kato T197 from China. The 12 chromosomes were assembled employing a length of 395 Mbp, and a G + C content of 43.62% with 0 gap. We identify the 56,908 protein-coding genes, of which 56,016 (98.42%) were functionally annotated. Repetitive elements accounted for 218.8 Mb (55.38%) of the genome. Compared to the gap-free genome of ZS97, 878,662 genomic variants were identified. Our findings provide a high-quality genomic resource for functional genomics, evolutionary studies, and molecular breeding of indica rice.

Background & Summary

Early-maturing, high-yielding double-cropped rice (*Oryza sativa* L.) with diverse panicle types is well suited for cropping in the Yangtze River region[1]. Several high-quality *indica* rice genome assemblies, such as SZ97 and MH63[2], have been released in recent years, providing valuable resources for genomic and breeding studies. *Oryza sativa* L. subsp. indica Kato T197 is an ultra-early maturing rice germplasm developed by the Hunan Academy of Agricultural Sciences. This cultivar exhibits a moderately compact plant architecture, with relatively slender stems and strong tillering capacity. The total growth duration is 91 days, with an average plant height of 65 cm (Fig. 1A). The average number of grains per panicle is 132.6, with a seed-setting rate of 83.1% and a 1000-grain weight of 25.2 g (Fig. 1B).

The Oxford Nanopore Technologies (ONT) have been around since 2014 and can produce reads that are more than 2 Mb long, which shows promise in genomics research. ONT sequencing can be used to assembly the genomes of highly heterozygous species[3,4]. By employing ultra-long sequencing of ONT, the researchers were able to obtain the T2T or gapfree-level genome, such as rice[2], watermelon[5], and Neosalanx taihuensis[6]. The R10.4.1 flowcell, in combination with Q20 + chemistry and SQK-LSK114 kit, can produce the sequencing data with a model read accuracy of about 99%[7]. In additional, the PacBio high-fidelity (HiFi) sequencing generates long reads with higher accuracy (>99.9%)[8].

In this study, we first reported a complete level genome of *O. Sativa* L. T197 using both the Oxford nanopore R10.4.1 long-read and PacBio HIFI sequencing. Additionally, we validated the reliability of chromosomes using the High-throughput chromosome conformation capture (Hi-C) approach. The complete genome sequence of the cultivar T197 offers a valuable genomic resource for downstream functional genomics research and facilitates the identification of key genes and molecular markers, thereby accelerating genetic improvement and breeding efforts in rice.

Methods

Sample preparation and genomic DNA sequencing. Tissue samples from the T197 cultivar were collected after two months of growth to facilitate genome extraction. High-quality genomic DNA was isolated from the fresh leaves using the CTAB method and the DNA quality and concentration were tested by 0.75% agarose gel electrophoresis, NanoDrop One spectrophotometer (Thermo Fisher Scientific) and Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). For Oxford Nanopore Ultra-long R10.4.1 flowcell sequencing, the

[1]State Key Laboratory of Hybrid Rice, Hunan Hybrid Rice Research Center, Changsha, 410125, China. [2]Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, 518120, China. [3]Yuelushan Laboratory, Changsha, 410128, China. [4]Wuhan Benagen Technology Co., Ltd, Wuhan, 430000, China. [5]These authors contributed equally: Tianyuan Zhang, Yi Zhang. ✉e-mail: yyh30678@163.com; uhz_uhz@hotmail.com

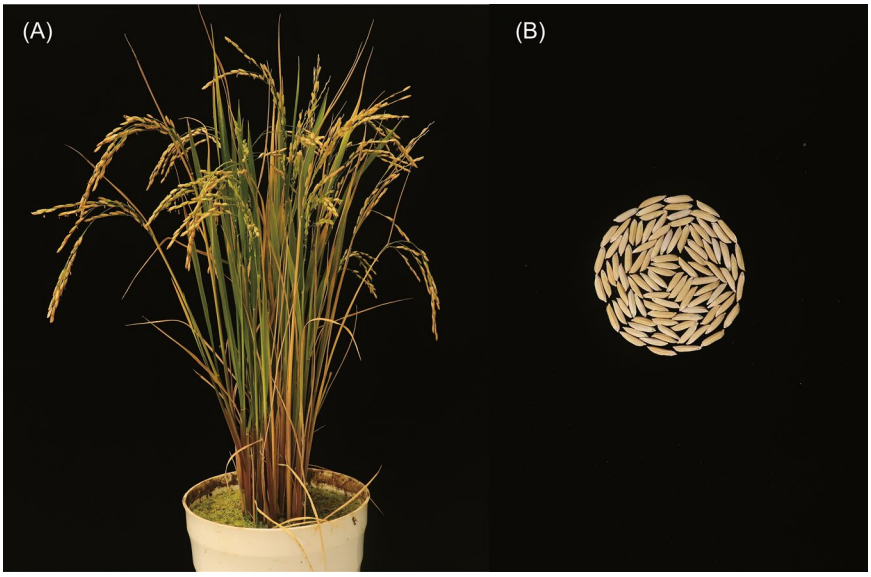

Fig. 1 Morphological characteristics of Oryza sativa L. subsp. indica Kato T197. (**A**) Plant architecture of Kato T197; (**B**) Grain morphology of Kato T197.

Type	Platform	Usage	Clean data (Gb)	Depth
short-reads	DNBSEQ T7	Genome survey and evaluation	47.40	119.1×
Long-reads	Nanopore R10.4.1	Assembly	23.21	58.8×
Long-reads	PacBio Revio	Assembly	24.89	62.4×
Hi-C	DNBSEQ T7	Hi-C verification	44.03	111.4×
RNA-seq(stem)	DNBSEQ T7	Annotation	11.93	
RNA-seq(root)	DNBSEQ T7	Annotation	11.24	
RNA-seq(leaf)	DNBSEQ T7	Annotation	10.21	
RNA-seq(bud)	DNBSEQ T7	Annotation	11.04	—
RNA-seq	Nanopore R10.4.1	Annotation	12.59	—

Table 1. Statistics of sequencing data.

libraries were prepared using the SQK-LSK114 ligation kit and using the standard protocol. The purified library was loaded onto primed R10.4.1. Spot-On Flow Cells and sequenced using a PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK) with 72-h runs at Wuhan Benagen Technology Co., Ltd. (Wuhan, China). A total of 23.21 Gb (58.8×) of ultra-long-reads (N50 > 100 kb) data were generated, with max length of 803,837 bp, and mean quality score(Q) of 23.27 (Table 1 and Fig. 2). For PacBio HiFi sequencing, the libraries were prepared using Pacific Biosciences SMRT bell express template prep kit 2.0 (Pacific Biosciences, USA) and using the standard protocol. Then, a SMRT cell sequencing library containing about 15 kb cut fragment was constructed and sequenced using PacBio sequel II sequencing platform with 30-h runs at Wuhan Benagen Technology Co., Ltd. (Wuhan, China). A total of 24.89 Gb (62.4×) of HIFI sequencing data were generated, with length N50 of 17,029 bp. After the DNA quality and integrity were tested, it was randomly sheared by covaris ultrasonic disruptor. DNBSEQ T7 sequencing pair-end libraries with an insert size of 300 bp were prepared using VAHTS® Universal Plus DNA Library Prep Kit for MGI V2 (Vazyme, Nanjing, China). A total of 47.40 Gb (119.1×) short-reads data were generated (Table 1).

Hi-C library preparation and sequencing. T197 were used to create the Hi-C library to capture genome-wide chromatin interactions. Chromatin digestion was carried out using the restriction enzyme DpnII. The Hi-C samples were then extracted through biotin labelling, fat-end ligation, and DNA purification. The Hi-C library was sequenced using the DNBSEQ T7 platform with paired-end 150-bp reads. A total of 44.03 Gb (111.4×) of clean data were generated (Table 1).

Transcriptome sequencing. For transcriptome sequencing, RNA purity was checked using the kaiaoK5500® Spectrophotometer (Kaiao, Beijing, China) from root, stem, leaf, and bud. RNA integrity and concentration was assessed using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system. A total amount of 2 μg RNA per sample was used as input material for the RNA sample preparations. Sequencing libraries were generated using NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (#E7530L, NEB, USA) following the manufacturer's recommendations. The libraries were then sequenced on the DNBSEQ T7 platform with PE reads

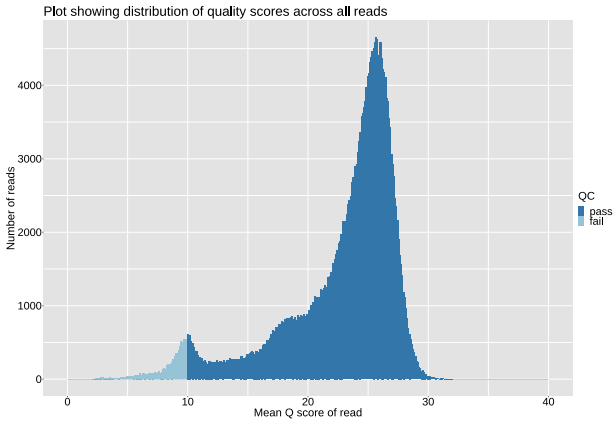

Fig. 2 Q-value distributions of ONT R10.4.1 reads.

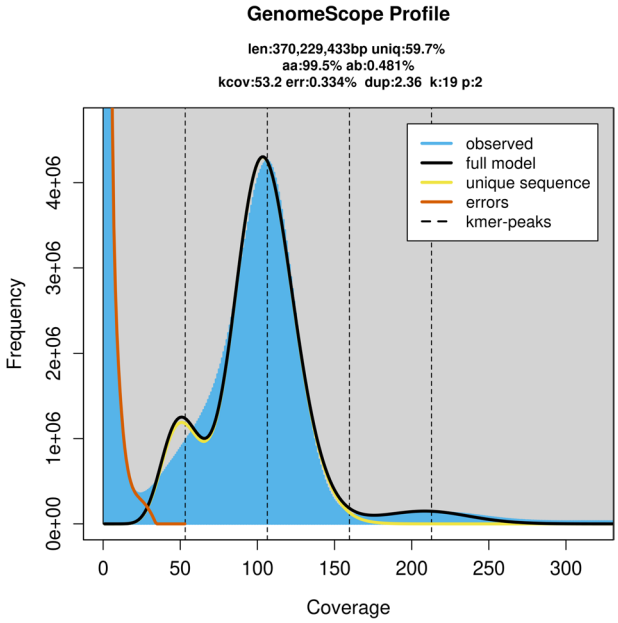

Fig. 3 The estimated characteristics of T197 genome based on short-read data using 19-mers count histogram. Genome size was estimated to be 370.22 Mb, with a duplication rate of 2.36% and heterozygosity rate of 0.48%.

of 150 bp. A total of 44.43 Gb of clean data were generated from 4 tissues (Table 1). For Oxford Nanopore cDNA sequencing, mixed RNA from for tissues were pooled and used a strand-switching method and the cDNA-PCR Sequencing Kit (SQK-PCS109) to carry out sequencing of cDNA by PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK). A total of 12.59 Gb of pass data were generated (Table 1).

Estimation of genomic characteristics. The k-mer method was utilized to survey the genome features of T197 with the short reads. The k-mer count histogram was calculated from Illumina short reads using Jellyfish[9] version 2.2.10 with the parameters: '-m 19'. Genome size, heterozygosity, and duplication rate were estimated using GenomeScope[10] version 2.0 with the parameters: '-n 2'. The genome size is estimated to be approximately 370.22 Mb, with a low percentage of duplication (2.36%) and heterozygosity (0.48%) (Fig. 3). This is consistent with the genome size of rice reported in previous studies[11].

Genome assembly and Telomere patching. The hybrid assembly of the ONT ultra-long data, PacBio HIFI data was performed using hifiasm[12] (version:0.24.0-r702, parameters: -l 2) software. Preliminary assembly results showed that there was only 1 gap in one contig (Chr11). Then, align the ONT ultra and PacBio HIFI reads with the genome after gap filling using Winnowmap2[13] (v2.03,parameters: k = 15, greater-than distinct = 0.9998 b, –MD), other software was used to filter comparison information, and racon was used to correct errors (v1.6.0, -L -u, https://github.com/isovic/racon/tree/liftover), in order to complete the gap-free assembly of the genome. Finally, we analyzed the terminal 20 kb of each chromosome to detect telomeric areas, as telomeres are generally situated there. The standard 7-base telomeric repeat motif (5': CCCTAAA; 3' TTTAGGG) was examined, and

Chr	Gap	Upstream	DownStream	UpStream_TelStart-TelEnd	DownStream_TelStart-TelEnd
chr1	0	2728	2784	1–19995	44717001–44736994
chr2	0	1956	1703	1–14388	37621690–37637425
chr3	0	1093	2799	3–19372	39711287–39731284
chr4	0	2389	2738	1–17626	36497219–36517212
chr5	0	2340	2823	1–17273	30887621–30907615
chr6	0	2442	2512	1–18240	32321657–32340560
chr7	0	2148	2832	1–16115	30939212–30959206
chr8	0	2790	2844	1–19995	30997137–31017136
chr9	0	2270	1223	1–18498	24644644–24653870
chr10	0	2418	2658	1–17581	26536907–26556905
chr11	0	1807	2692	1–15479	31510042–31530035
chr12	0	1801	2849	1–14262	28476037–28496034

Table 2. Statistical table of genome telomere.

Item	T197	SZ97R3
Total_length(bp)	395,084,276	391,561,630
total_length_withoutN(bp)	395,084,276	391,561,630
Total_number	12	12
GC_content	43.63%	43.62%
N50(bp)	31,530,035	32,119,910
N90(bp)	26,556,905	25,797,731
Average(bp)	32,923,689.67	32,630,135.83
Median(bp)	31,273,585.50	31,480,230.00
Min(bp)	24,653,870	22,989,350
Max(bp)	44,736,994	44,754,788
Gaps	0	0
Telomere	24	19

Table 3. Features of T197 genome and SZ97 genome.

	TE proteins		De novo + repbase		Combined TEs	
Type	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome
DNA	5,145,073	1.30	88,438,896	22.38	88,468,785	22.39
LINE	1,699,412	0.43	10,596,041	2.68	10,645,013	2.69
SINE	0	0.00	549,628	0.14	549,628	0.14
LTR	26,380,235	6.68	115,677,356	29.28	116,834,852	29.57
LTR-Gypsy	21,064,376	5.33	78,068,732	19.76	79,607,814	20.15
LTR-Copia	5,148,882	1.30	10,779,138	2.73	11,384,260	2.88
Other	0	0.00	631	0.00	631	0.00
Unknown	0	0.00	5,346,883	1.35	5,346,883	1.35
Total	33,223,888	8.41	217,595,939	55.08	218,801,348	55.38

Table 4. Repeats elements statistics in genomes of T197. Note: SINEs, short interspersed nuclear elements; LINEs, long interspersed nuclear elements; LTR, long terminal repeat.

areas containing a minimum of 100 repeats were designated as telomeric regions. Only sequences with a mapped length of ≥50 bp and a mapping identity of ≥80% were deemed suitable for high-quality identification. Regions that satisfy all these criteria were classified as full telomeres.

Finally, 12 chromosome sequences were obtained by Hifiasm assembler, among which 12 chromosomes contained complete telomeres (Table 2). The T197 genome has a G + C content of approximately 43.63%. we assembled the genome into 395 Mb, the contig N50 is 31.53 Mb. The length of each chromosome is range of 24,653,870 to 44,736,994 bp (Table 3). Noticeable, the difference between the estimated genome size (370.22 Mb) and the final assembled genome size can be attributed to the presence of repetitive elements within the genome, such as transposable elements (TEs), tandem repeats, rRNA, and other repetitive sequences. For published SZ97RS3 genome[2], genome size is 391 Mb with G + C content of approximately 43.62%. Compared to the gap-free SZ97 genome, the telomeres on chr5, chr8, chr9, chr10, and chr12 have all been completed.

Type	Number	Length (bp)	Rate (%)
Microsatellite (1–9 bp)	158,235	3,333,990	0.844
Minisatellite (10–99 bp)	112,675	8,083,768	2.046
Satellite (≥100 bp)	12,505	11,459,615	2.901
Total	283,415	21,355,661	5.405

Table 5. Statistics of tandem repeat in T197 genome.

Type	Subtype	Copy	Average length (bp)	Total length (bp)	% of genome
miRNA	miRNA	8,587	198	1,703,502	0.4312
tRNA	RNA	627	75	47,201	0.0119
rRNA	rRNA	737	243	179,251	0.0454
	18S	19	1735	32,963	0.0083
	28S	15	4197	62,953	0.0159
	5.8S	13	153	1,988	0.0005
	5S	690	118	81,347	0.0206
snRNA	snRNA	701	116	81,295	0.0206
	CD-box	551	109	60,154	0.0152
	HACA-box	64	130	8,289	0.0021
	splicing	86	149	12,852	0.0033

Table 6. Statistics of non-coding RNAs in T197 genome.

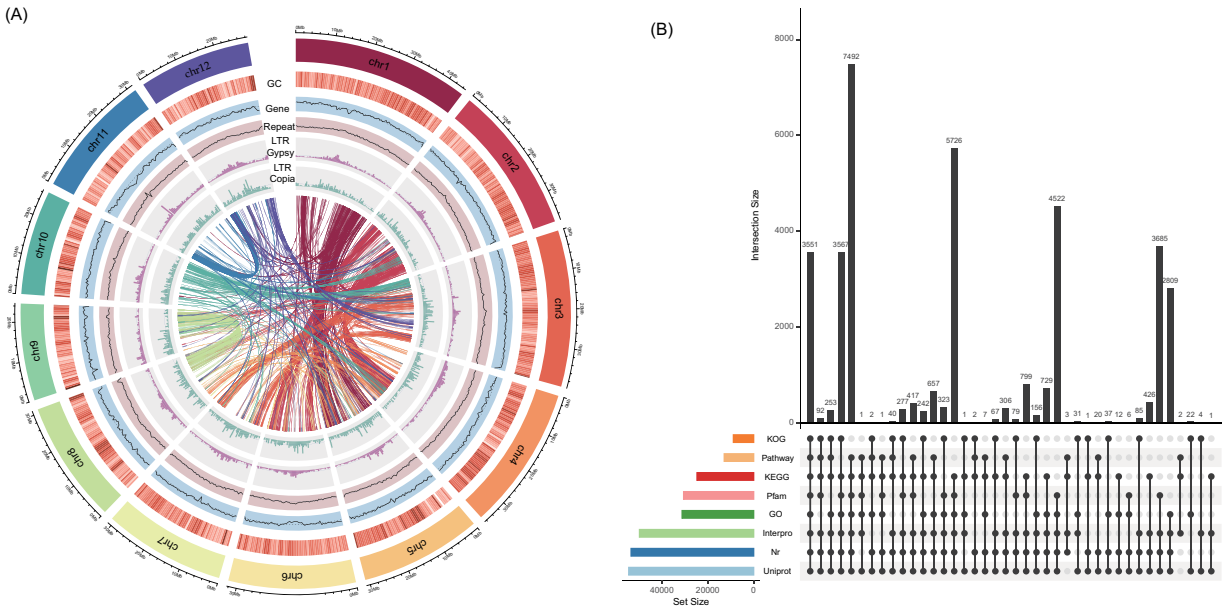

Fig. 4 Statistics of genomic features and functional annotation in T197 genome. (**A**) The Circos plot depicts the physical locations and features of chromosomes, including GC content, gene density, repeat regions, LTR Copia, and LTR Gypsy, as well as syntenic connections between different chromosomes. (**B**) The UpSet plot summarizes functional annotation of genes across different databases (e.g., KOG, Pathway, KEGG, Pfam, GO, InterPro, NR, and UniProt).

Repeat element and non-coding RNA annotation. The software RepeatModeler (version:2.0.6, parameters: deafult) is employed to predict the repeat sequences, relying on its sequence characteristics. The software RepeatMasker[14] (version: open-4.0.9, parameters: deafult) was utilized to make predictions for both the denovo library and the RepBase library. RepeatModeler software performs Repeat annotation using the RepBase Proteins library. The final compilation of repeat sequences is obtained by merging and removing redundancy. In the T197 genome, a total of 218 Mb sequences (55.38%) were identified as repetitive elements (Table 4). A total of 283,415 tandem repeats were identified in the genome, including 158,235 microsatellites (1–9 bp), 112,675 minisatellites (10–99 bp), and 12,505 satellites (≥100 bp), collectively accounting for 5.41% of the genome (Table 5).

Item	Count	Percentage
All	56,908	100.00%
Annotation	56,016	98.43%
KEGG	24,718	43.44%
Pathway	12,846	22.57%
Nr	53,461	93.94%
UniProt	54,480	95.73%
GO	31,203	54.83%
KOG	8,852	15.55%
Pfam	30,550	53.68%
InterPro	49,800	87.51%

Table 7. Statistics of functional annotation in T197 genome.

Number	SNP	InDel
Total	714,626	156,777
Gene	331,272	80,360
Intergenic	383,354	76,417
Structure		
UTR3	12,028	3,741
UTR5	7,520	3,345
Downstream	65,850	17,614
Upstream	80,078	20,686
Upstream/downstream	11,412	3,507
Intronic	95,392	27,250
Exonic	58,772	4,146
Intergenic	383,354	76,417
Splicing	170	59
Function		
Synonymous SNV	24,832	0
Nonsynonymous_SNV	33,307	0
Stoploss	110	9
Stopgain	523	59
Nonframeshift_insertion	0	976
Nonframeshift_deletion	0	971
Frameshift_insertion	0	1,064
Frameshift_deletion	0	1,068

Table 8. Analysis of SNP and InDel mutations.

We predicted 737 rRNAs,627 tRNAs, 701 small nuclear RNAs, and in the T197 genome based on Rfam database (Table 6) (Fig. 4A).

Gene and functional predictions. The gene structure prediction performed by integrating of homologous prediction, de novo prediction, and transcript prediction. Initially, miniprot (version: 0.13-r248, parameters: default) was employed using a protein sequence set from similar species. Homology prediction(*Oryza rufipogon*, *Oryza sativa* Indica Group, *Oryza sativa*, *Sorghum bicolor*, *Triticum aestivum*, Zea mays, and *Oryza sativa* indica ZS97) was performed using the software Augustus (version: 35.0, parameters:–uniqueGeneId = true–noInFrameStop = true–gff3 = on–strand = both)[15], based on the training set received from Benchmarking Universal Single-Copy Orthologs (BUSCO)[16]. Next, the RNA-seq data was processed and converted into transcription data using the software Stringtie2 (version: 2.2.3, parameters: -l ONT -R -L)[17] and Scallop2 (version:1.1.2, parameters:–min_mapping_quality 10)[18]. In addition, employ the TransDecoder software (version: v5.7.0, parameters: default) to predict the coding sequences. Ultimately, through the process of integrating and eliminating duplication. Furthermore, the proteins were assessed against UniProt[19] Nr[20], GO[21], KOG[22], Pfam[23], InterPro[24], and KEGG[25] databases to functional information based on similarities in sequence and motif. The tRNA sequences were searched using tRNAscan-SE (version: 2.0.12, parameters: -Iut 50)[26], while rRNA prediction was performed using the rnammer (version: 1.2, parameters: -S euk)[27]. A total of 56,908 protein-coding genes were annotated in the T2T-level assembly, in which 56,016 genes (98.43%) were functionally annotated (Table 7). Among them, 3,551 genes were annotated to KOG, KEGG, GO, PFAM, InterPro, UniProt, and Nr databases. UniProt database has the highest proportion (95.73%) of annotations (Fig. 4B).

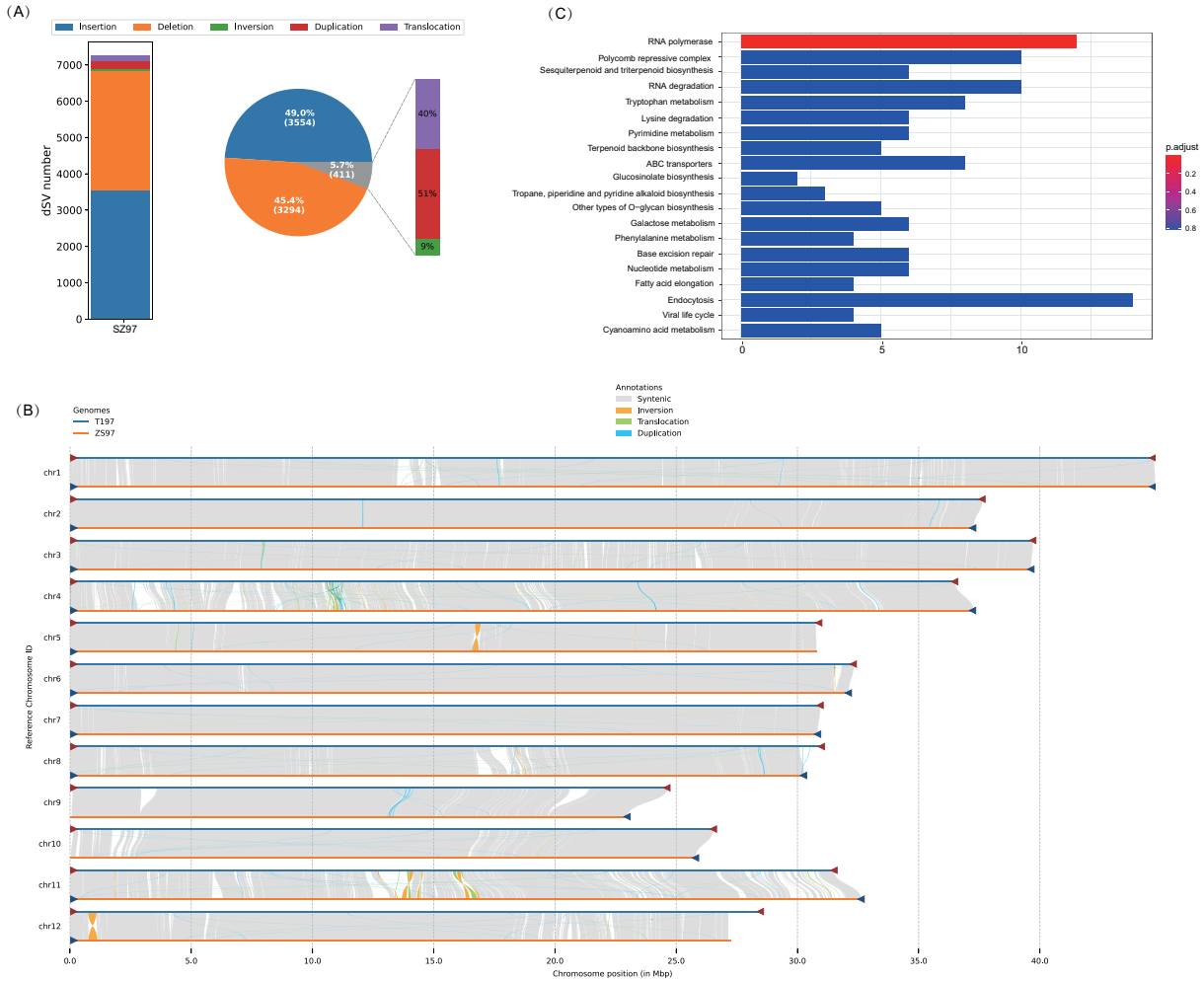

Fig. 5 Whole-genome level structural variation (SV) display. (**A**) Statistics of SV between T197 and SZ97 genomes; (**B**) Collinearity between T197 and ZS97 genomes. blue chromosome: T197; orange chromosome: ZS97; (**C**) KEGG pathway enrichment analysis of SVs affects genes.

Genomic structural variants analysis. The genome was aligned using MUMmer[28] (version 4.0.0rc, parameters:–mum–mincluster 500). Subsequently, SyRI[29] (version 1.6.3, parameters:–nc 5–invgaplen 20000) was used to identify collinear regions and structural variations in the genome. Between T197 and ZS97 genome, a total of 878,662 variants were identified, including 714,626 single nucleotide polymorphisms, indels, and SVs (Table 8). Among 7,259 SVs, including 3,249 deletions (44.7%), 3,554 insertions (48.9%), 211 duplications (2.9%), 213 inversion (2.9%) and 163 translocations (2.2%) (Fig. 5A). Large SVs, such as duplications and inversions, were observed in the ZS97 genome (Fig. 5B). Overlapping of SVs and gene annotation revealed that 2,361 SVs were located within 2 kb upstream of genes (32.5%), 495 SVs were located within the coding region of genes (6.8%), 1,182 SVs were located within introns (16.2%), 2207 SVs were located within 2 kb downstream of genes (30.4%), and 2,346 SVs were located within intergenic regions (32.3%). These results suggest that SVs located in upstream regions (32.5%) may have the greatest impact on gene expression by altering promoter or enhancer activity, while SVs in coding regions (6.8%) could directly disrupt protein function. Further functional validation is required to confirm the roles of these SVs in regulating key traits. KEGG enrichment of SV-affected genes, were performed to link functions and metabolic pathways with these genes using the R package ClusterProfiler[30] (version 3.14.3, parameters: default). Based on the KEGG enrichment analyses, RNA polymerase pathway was identified among the top enriched groups (Fig. 5C). These results implied that SV plays important roles in the multiple biological pathways. Impotrance, this pathway including RNA polymerase subunit gene family, including Rpb1 (OsatChr10G489330.1, OsatChr8G384180.1), Rpb2 (OsatChr5G256160.1), Rpb8 family protein (OsatChr3G187410.1), RNA polymerase III RPC4 (OsatChr4G199390.1, OsatChr4G199550.1, OsatChr4G199570.1, OsatChr4G199640.1, OsatChr4G199700.1, OsatChr4G199780.1, OsatChr4G199810.1, OsatChr4G209390.1, OsatChr4G199550.1). Previous study reporting that RNA-directed DNA methylation at Miniature Inverted-repeat Transposable Elements regulates rice tillering[31]. In addition, cytoplasmic male sterility (CMS) and its restoration are essential mechanisms in plant hybrid breeding, facilitating the effective generation of hybrid seeds. The interaction between nuclear genes, including those that code for RNA polymerases, and mitochondrial genes is fundamental to the processes of male sterility and fertility restoration in various

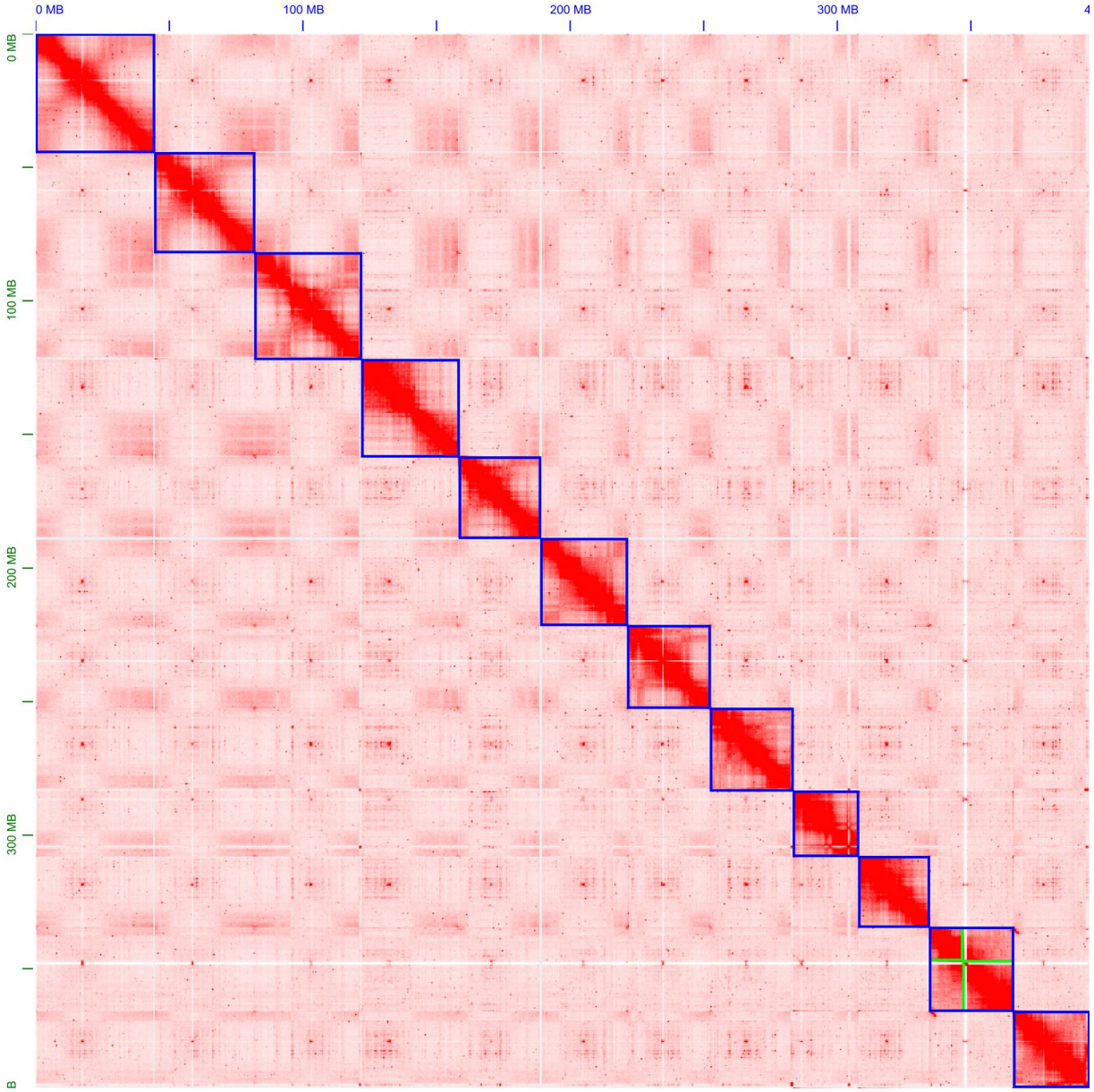

Fig. 6 Hi-C-maps of T197 genome.

crops[32]. RNA editing of in rice and sorghum is linked to the restoration of fertility[33,34]. Restorer genes (Rf) have been found in wheat, rice, maize *et al.*[35–38]. These genes frequently encode proteins including pectinesterases, pentatricopeptide repeat (PPR) proteins, or enzymes associated with DNA repair or ubiquitin-like processes. Moreover, FEM3 encodes the predominant subunit of DNA-dependent RNA polymerase V (Pol V). Mutations in Pol V result in male and female gametophyte infertility in rice, demonstrating that the RNA-directed DNA methylation (RdDM) molecular pathway is essential for reproductive development in rice[39]. These findings suggest that structural variants (SVs) affecting RNA polymerase-related genes could play a pivotal role in regulating gene expression and impacting phenotypic traits, particularly in the context of plant reproductive development. Future investigations focusing on the functional validation of these SVs will further elucidate their role in shaping phenotypic traits and enhancing crop breeding strategies.

Data Records

Data Records T197 genome project was deposited at China National Center for Bioinformation (CNCB) under BioProject No. PRJCA039529[40]. Short-read Genomic data, Nanopore ultra-long DNA data, genomic HIFI data Hi-C data, and RNA-seq were deposited in the Genome Sequence Archive (GSA) at CNCB under accession number CRA025323[41]. The final complete chromosome assembly was deposited in GenBank under the accession JBRALX010000000[42] and Genome Warehouse (GWH) at CNCB under accession number

chr	QV	E-value
chr1	49.7153	1.07E-05
chr2	50.1776	9.60E-06
chr3	50.0406	9.91E-06
chr4	50.0426	9.90E-06
chr5	50.0813	9.81E-06
chr6	49.587	1.10E-05
chr7	49.824	1.04E-05
chr8	50.0301	9.93E-06
chr9	48.8934	1.29E-05
chr10	48.5033	1.41E-05
chr11	47.7503	1.68E-05
chr12	48.8969	1.29E-05

Table 9. Statistics of functional annotation in T197 genome.

GWHGDIJ00000000.1[43]. The genome assembly and genome annotation files are available in figshare under a DOI number of https://doi.org/10.6084/m9.figshare.28912358[44].

The dataset package includes the following files:

OsT197.T2T_chr.fasta (380.55 Mb): This file provide assembled genome sequence of the *Oryza sativa* L. T197 genome.

Oryza_sativaL.gene.gff (32.59 Mb): This file provide gene structure annotation for the T197 genome, including gene loci.

Oryza_sativaL.rRNA.gff (90.53 kb): Includes annotations of rRNA (ribosomal RNA) loci in the T197 genome.

Oryza_sativaL.snRNA.gff (104.81 kb): Provides annotations of snRNA (small nuclear RNA) in the T197 genome.

Oryza_sativaL.Tandem_Repeats.gff (36.79 kb): Describes tandem repeat loci within the T197 genome, suitable for studying repetitive sequence structures.

Oryza_sativaL.tRNA.gff (107.03 kb): Contains annotations of tRNA (transfer RNA) loci in the T197 genome.

Technical Validation

The Hi-C interaction is employed to assess the degree of connections between distinct contigs in reliable data and group them into clusters. The contigs were grouped, arranged, and aligned to chromosomes using the ALLHiC method[45]. Subsequently, the contigs HIC map was visualized using the Juicebox (version: 2.15) software[46]. The chromosome level map includes 12 groups, consisting of 12 contigs. This result turns that the contig we assembled is chromosome-level (Fig. 6).

To assess the accuracy of the final genome assembly, we aligned Illumina short reads to the T197 genome using BWA-MEM version 0.7.17-r1188 (https://github.com/lh3/bwa). The analysis indicated that 99.46% of the short reads were mapped to the genome with 99.82% of coverage and 119.10 of average depth.

Base level accuracy (QV) and k-mer completeness assessments were conducted using the pre-existing meryl database and merqury[47]. The results of the k-mer statistical analysis showed that the QV value of the genome was 49.46, and the QV value of each chromosome was between 47.75 and 50.17, indicating high accuracy of the genome assembly (Table 9).

For the genome assembly and gene annotation, we assessed the quality using BUSCO (v5.7.0, embryophyta_odb10). The BUSCO analysis showed that for the chromosome level assemblies, 98.8% of the evaluated Complete BUSCOs were identified as complete (single-copied gene: 96.6%, duplicated gene: 2.2%). For gene level, 99.1% of evaluated Complete BUSCOs (v5.7.0, embryophyta_odb10).

Data availability

The raw sequencing data this study have been deposited in the China National Center for Bioinformation (CNCB) under BioProject No. PRJCA03952939 with GSA number CRA02532340. The final complete chromosome assembly is available in the GenBank under the accession JBRALX010000000 and GWH under accession number GWHGDIJ00000000.1. The genome assembly and annotation files are shared on Figshare under the https://doi.org/10.6084/m9.figshare.28912358.

Code availability

No custom scripts or code were used in this study.

Received: 21 May 2025; Accepted: 30 September 2025;
Published online: 17 November 2025

References

1. Zhu, M. *et al*. Single-cell transcriptome sequencing reveals the mechanism regulating rice plumule development. *The Crop Journal* **12**, 688–697, https://doi.org/10.1016/j.cj.2024.04.009 (2024).
2. Song, J. M. *et al*. Two gap-free reference genomes and a global view of the centromere architecture in rice. *Mol Plant* **14**, 1757–1767, https://doi.org/10.1016/j.molp.2021.06.018 (2021).

3. Zhang, T. *et al*. Nanopore sequencing: flourishing in its teenage years. *J Genet Genomics* **51**, 1361–1374, https://doi.org/10.1016/j.jgg.2024.09.007 (2024).
4. Zhang, T. *et al*. Computational Tools and Resources for Long-read Metagenomic Sequencing Using Nanopore and PacBio. Genomics, Proteomics & Bioinformatics, qzaf075, https://doi.org/10.1016/10.1093/gpbjnl/qzaf075 (2025).
5. Deng, Y. *et al*. A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding. *Mol Plant* **15**, 1268–1284, https://doi.org/10.1016/j.molp.2022.06.010 (2022).
6. Zhou, Y. *et al*. Gap-free genome assembly of Salangid icefish Neosalanx taihuensis. *Scientific data* **10**, 768, https://doi.org/10.1038/s41597-023-02677-z (2023).
7. Zhang, T. *et al*. The newest Oxford Nanopore R10.4.1 full-length 16S rRNA sequencing enables the accurate resolution of species-level microbial community profiling. *Applied and environmental microbiology* **89**, e0060523, https://doi.org/10.1128/aem.00605-23 (2023).
8. Duan, H. *et al*. Physical separation of haplotypes in dikaryons allows benchmarking of phasing accuracy in Nanopore and HiFi assemblies with Hi-C data. *Genome Biol.* **23**, 84, https://doi.org/10.1186/s13059-022-02658-2 (2022).
9. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
10. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun* **11**, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
11. Qin, P. *et al*. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* **184**, 3542–3558. e3516, https://doi.org/10.1016/j.cell.2021.04.046 (2021).
12. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
13. Jain, C. *et al*. Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36**, i111–i118, https://doi.org/10.1093/bioinformatics/btaa435 (2020).
14. Tempel, S. Using and understanding RepeatMasker. *Methods Mol Biol* **859**, 29–51, https://doi.org/10.1007/978-1-61779-603-6_2 (2012).
15. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644, https://doi.org/10.1093/bioinformatics/btn013 (2008).
16. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: Assessing Genomic Data Quality and Beyond. *Curr Protoc* **1**, e323, https://doi.org/10.1002/cpz1.323 (2021).
17. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* **20**, 278, https://doi.org/10.1186/s13059-019-1910-1 (2019).
18. Zhang, Q., Shi, Q. & Shao, M. Accurate assembly of multi-end RNA-seq data with Scallop2. *Nature computational science* **2**, 148–152, https://doi.org/10.1038/s43588-022-00216-1 (2022).
19. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **46**, 2699, https://doi.org/10.1093/nar/gky092 (2018).
20. Deng, Y. *et al*. Integrated nr database in protein annotation system and its localization. *Comput Eng* **32**, 71–74, https://doi.org/10.1109/INFOCOM.2006.241 (2006).
21. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29, https://doi.org/10.1038/75556 (2000).
22. Koonin, E. V. *et al*. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* **5**, R7, https://doi.org/10.1186/gb-2004-5-2-r7 (2004).
23. Finn, R. D. *et al*. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* **44**, D279–D285, https://doi.org/10.1093/nar/gkv1344 (2016).
24. Mulder, N. J. *et al*. Florence Servant, Christian JA Sigrist, and InterPro Consortium. Interpro: an integrated documentation resource for protein families, domains and functional sites. *Brief Bioinform* **3**, 225–235, https://doi.org/10.1093/bib/3.3.225 (2002).
25. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30, https://doi.org/10.1093/nar/28.1.2 (2000).
26. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res* **49**, 9077–9096, https://doi.org/10.1093/nar/gkab688 (2021).
27. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935, https://doi.org/10.1093/bioinformatics/btt509 (2013).
28. Marçais, G. *et al*. MUMmer4: A fast and versatile genome alignment system. *PLoS Comp. Biol.* **14**, e1005944, https://doi.org/10.1371/journal.pcbi.1005944 (2018).
29. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 1–13, https://doi.org/10.1186/s13059-019-1911-0 (2019).
30. Wu, T. *et al*. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**, 100141, https://doi.org/10.1016/j.xinn.2021.100141 (2021).
31. Xu, L. *et al*. Regulation of rice tillering by RNA-directed DNA methylation at miniature inverted-repeat transposable elements. *Mol. Plant* **13**, 851–863, https://doi.org/10.1016/j.molp.2020.02.009 (2020).
32. Kuwabara, K. *et al*. Artificial mutations in the nuclear gene encoding mitochondrial RNA polymerase restore pollen fertility in cytoplasmic male sterile tomato. bioRxiv, 2025.2002. 2018.638787, https://www.x-mol.com/paperRedirect/1894107329233248256 (2025).
33. Ngangkham, U., Parida, S. K., Singh, A. K. & Mohapatra, T. Differential RNA Editing of Mitochondrial Genes in WA-Cytoplasmic Based Male Sterile Line Pusa 6A, and Its Maintainer and Restorer Lines. *Rice Science* **26**, 282–289, https://doi.org/10.1016/j.rsci.2019.08.002 (2019).
34. Pring, D., Tang, H., Howad, W. & Kempken, F. A unique two-gene gametophytic male sterility system in sorghum involving a possible role of RNA editing in fertility restoration. *J. Hered.* **90**, 386–393, https://doi.org/10.1093/jhered/90.3.386 (1999).
35. Niu, F. *et al*. Rfd1, a restorer to the Aegilops juvenalis cytoplasm, functions in fertility restoration of wheat cytoplasmic male sterility. *J. Exp. Bot.* **74**, 1432–1447 (2023).
36. Xu, Z. *et al*. Identification and fine mapping of a fertility restorer gene for wild abortive cytoplasmic male sterility in the elite indica rice non-restorer line 9311. *The Crop Journal* **11**, 887–894, https://doi.org/10.1016/j.cj.2022.11.009 (2023).
37. Liu, Y. *et al*. Mapping of Rf20 (t), a minor fertility restorer gene for rice wild abortive cytoplasmic male sterility in the maintainer line 'Zhenshan97B. *Theor. Appl. Genet.* **136**, 248, https://doi.org/10.1007/s00122-023-04486-9 (2023).
38. Liu, Y. *et al*. Identification of maize Rf4-restorer lines and development of a CAPS marker for Rf4. *Agronomy* **12**, 1506, https://doi.org/10.3390/agronomy12071506 (2022).
39. Zheng, K. *et al*. The effect of RNA polymerase V on 24-nt siRNA accumulation depends on DNA methylation contexts and histone modifications in rice. *Proceedings of the National Academy of Sciences* **118**, e2100709118, https://doi.org/10.1073/pnas.2100709118 (2021).
40. *NGDC BioProject* https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA039529 (2025).
41. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA025323 (2025).
42. Zhu, M. Oryza sativa Indica Group cultivar T197, whole genome shotgun sequencing project. *GenBank* https://identifiers.org/ncbi/insdc:JBRALX010000000 (2025).

43. *NGDC Genome Warehouse* https://ngdc.cncb.ac.cn/gwh/Assembly/95113/show (2025).
44. Zhu, M. Genome assembly and annotation of Oryza sativa L. subsp. indica Kato T197. *Figshare* https://doi.org/10.6084/m9.figshare.28912358 (2025).
45. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat Plants* **5**, 833–845, https://doi.org/10.1038/s41477-019-0487-8 (2019).
46. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**, 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
47. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 1–27, https://doi.org/10.1186/s13059-020-02134-9 (2020).

Acknowledgements

The work was supported by the Changsha Natural Science Foundation (kq2208128) and Biological Breeding-National Science and Technology Major Project (2023ZD04022).

Author contributions

M.Z., Y.H. and T.Z. designed the study and led the research. X.T., W.P. and H.X. contribute to the materials of this study. T.Z., Y.X., Y.Z. and L.Y. analyzed the data. Y.X. and Y.Z. contribute to the genome assembly and annotation. T.Z. and Y.Z. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.Y. or M.Z.

Reprints and permissions information is available at www.nature.com/reprints.

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