Scientific Data

https://doi.org/10.1038/s41597-026-07833-9

Article in Press

A telomere-to-telomere gap-free genome assembly of the indica rice (*Oryza sativa L.*) cytoplasmic male sterile line Funong A.

Received: 23 December 2025

Accepted: 1 July 2026

Published online: 08 July 2026

Cite this article as: Hong, Y., Zhou, P., Zhang, H. *et al.* A telomere-to-telomere gap-free genome assembly of the indica rice (*Oryza sativa L.*) cytoplasmic male sterile line Funong A.. *Sci Data* (2026). https://doi.org/10.1038/s41597-026-07833-9

Yonghe Hong, Peng Zhou, Hui Zhang, Feiyan Zheng, Ronghua Hu, Jiatuan Zheng & Shihang Tu

A telomere-to-telomere gap-free genome assembly of the indica rice (*Oryza sativa L.*) cytoplasmic male sterile line Funong A

Shihang Tu[1,#], Peng Zhou[1], Hui Zhang[2], Feiyan Zheng[1], Ronghua Hu[1], Jiatuan Zheng[1], Yonghe Hong[1,#,*]

[1] Rice Research Institute, Fujian Academy of Agricultural Sciences, Fuzhou 350018, China

[2] Crop Research Institute, Fujian Academy of Agricultural Sciences, Fuzhou 350013, China

[#] These authors contributed equally: Shihang Tu, Yonghe Hong.

[*] Corresponding authors: Yonghe Hong, edwardhyh@163.com.

Authors Email:

Shihang Tu, tushihang@faas.cn

Yonghe Hong, edwardhyh@163.com

Peng Zhou, zpl2009@yeah.net

Hui Zhang, zhanghui-zws@faas.cn

Feiyan Zheng, faaszfy@163.com

Ronghua Hu, huon2000@foxmail.com

Jiatuan Zheng, superrice63@163.com

Abstract

Rice (*Oryza sativa* L.) is a staple crop crucial for global food security, yet its productivity is limited by climate change, diseases, and genetic erosion. The indica cytoplasmic male sterile (CMS) line Funong A (FNA) is an elite breeding resource, exhibiting strong blast resistance, favorable floral traits, and enhanced stem strength through high silica and hemicellulose deposition. Its hybrids demonstrate >10% yield advantages with medium blast resistance, positioning FNA as a pivotal parent for subtropical hybrid breeding. breeding. Here, we present a fully gap-free, telomere-to-telomere (T2T) chromosome-scale genome assembly of FNA. By integrating PacBio HiFi, Oxford Nanopore long reads, and Hi-C sequencing, we produced a 394.7 Mb gap-free assembly with a contig N50 of 32 Mb, capturing all 24 telomeric regions. The assembly demonstrates exceptional accuracy and completeness, including fully resolved centromeres across all 12 chromosomes. We annotated 39,453 protein-coding genes and 1,376 lncRNAs, with the majority of genes functionally characterized. Repetitive elements comprise 51.1% of the genome, reflecting lineage-specific bursts of long terminal repeat retrotransposons. This high-quality T2T genome provides a valuable foundation for dissecting CMS mechanisms, centromere biology, and structural evolution, while offering critical resources for hybrid rice improvement.

Background & Summary

Rice (*Oryza sativa* L.) is a cornerstone of global food security, feeding over half the world's population[1-4]. Despite its agronomic importance, rice productivity faces escalating threats from climate change-induced heat stress, diseases like rice blast, and the erosion of genetic diversity in modern cultivars[5-10]. These challenges necessitate the development of resilient cultivars through genomic-driven breeding, particularly leveraging cytoplasmic male sterile (CMS) lines—the maternal foundation of three-line hybrid systems[11-14]. Among these, the indica CMS line 'Funong A' (FNA) has emerged as an ideal genetic resource[15,16]. Developed through pyramiding critical alleles from the blast-resistant donor 'Huahang Simiao' and the agronomically elite but disease-susceptible maintainer 'Fudao B', FNA exhibits exceptional blast resistance (integrating *Pi5*, *Pid2*,

and *Pid3* genes), optimal floral traits (85% morning flowering rate, 57.12% stigma exertion), and enhanced stem strength due to high silica and hemicellulose deposition (stem diameter: 4.8 mm, 18% thicker than average). Its hybrids demonstrate >10% yield advantages with medium blast resistance, positioning FNA as a pivotal parent for subtropical hybrid breeding[15,16].

Recent advances in telomere-to-telomere (T2T) assembly have revolutionized rice genomics by resolving previously inaccessible regions such as centromeres, telomeres, and ribosomal DNA arrays[17-19]. Study for the first gap-free genomes of two elite indica rice varieties (Zhenshan 97 (ZS97) and Minghui 63 (MH63))[20] revealed conserved centromere structures with satellite motifs, identifying over 395 actively transcribed non-TE genes in centromere regions, and discovering large structural variants on chromosome 11 linked to disease resistance. Analysis of gap-free genomes for four key hybrid rice parents (Xiangling 628S, Jing 4155S, Longke 638S, and Huazhan) demonstrated that structural variations (SVs), particularly presence/absence variations (PAVs) affecting 2,182 functional genes, play critical roles in heterosis and transcriptional advantages in hybrid breeding[21]. A major milestone was reached with the release of the NIP T2T rice genome (AGIS-1.0), representing the first complete T2T genome sequence in rice[22]. The AGIS-1.0 genome closed all 72 gaps in the IRGSP-1.0 reference and added 12.5 Mb of novel sequences, including functional centromeric repeats, and corrected 1,324 protein coding gene annotations. Recently, the gap-free genome of an early-season indica variety Jiafuzhan (JFZ) provided insights into thermotolerance and grain quality, highlighting the role of phased assemblies in breeding[23]. Collectively, these advances underscore T2T genomes as indispensable tools for dissecting genetic mechanisms underlying agronomic traits.

In this study, we reported a telomere-to-telomere, chromosome-level reference genome assembly of the indica CMS line FNA. Using PacBio HiFi, Oxford Nanopore long reads, and Hi-C data, we generated a 394.7 Mb gap-free assembly with a contig N50 of 32 Mb and complete representation of all 24 telomeric regions. The assembly demonstrates outstanding accuracy, with 99.99% of HiFi reads and 99.85% of ONT reads successfully remapped, and exhibits high completeness

(BUSCO >99%, LTR Assembly Index (LAI) of 21.99). A total of 39,453 high-confidence protein-coding genes were predicted, of which more than 81% were functionally annotated, together with 1,376 lncRNAs. Repetitive elements comprise 51.1% of the genome, with long terminal repeat retrotransposons (LTR-RTs) as the dominant class, reflecting multiple lineage-specific amplification bursts. Complete centromere assemblies revealed chromosome-specific divergence in repeat composition and organization, suggesting evolutionary specialization of centromeres in FNA. The near-complete homozygosity of FNA further reflects its breeding history through recurrent backcrossing. Collectively, this high-quality reference genome provides a valuable resource for elucidating CMS mechanisms, centromere biology, and evolutionary dynamics in rice.

Methods

Rice Sample Collection

The CMS line FNA, used as the experimental material, was developed by the Rice Research Institute of Fujian Academy of Agricultural Sciences. Plants were grown under controlled greenhouse conditions. Healthy, disease-free seedlings at the 5- to 6-leaf stage were carefully selected for sampling. Collected tissues were immediately snap-frozen in liquid nitrogen to preserve RNA integrity. To ensure experimental consistency, all samples were obtained from plants grown under comparable environmental conditions.

Genomic DNA Extraction, Library Construction and Sequencing

Genomic DNA was extracted from rice cultivar FNA using the CTAB DNA extraction method. Briefly, rice leaves were harvested and quickly ground into powder in liquid nitrogen. The ground samples were placed in an Eppendorf tube containing 1 ml of pre-warmed 1.5 × CTAB buffer (40 mM CTAB, 75 mM Tris–HCl pH 8.0, 15 mM EDTA, and 1M NaCl), vortexed thoroughly, and incubated at 65°C for 30 minutes. After cooling, an equal volume of chloroform/isoamyl alcohol (V/V = 24:1) was added, mixed thoroughly, and centrifuged at 10,391 × g for 15 minutes. The

supernatant was transferred to a new Eppendorf tube, and 1/10 volume of 10% CTAB (0.27 M CTAB, 0.7 M NaCl, pre-warmed at 65°C) was added. After mixing, an equal volume of chloroform/isoamyl alcohol (V/V = 24:1) was added, and the solution was centrifuged at 10,391 × g for 15 minutes. The supernatant was transferred to another tube, and 2.5 volumes of 1% CTAB (27 mM CTAB, 50 mM Tris–HCl pH 8.0, and 10 mM EDTA) was added. The mixture was centrifuged again at 10,391 × g for 15 minutes, and the supernatant was discarded. The precipitated DNA was dissolved in 0.5 ml TE buffer containing 2 mg RNase and incubated at 37°C for 1 hour. An equal volume of chloroform/isoamyl alcohol solution (V/V = 24:1) was added, mixed thoroughly, and centrifuged at 10,391 × g for 15 minutes. The supernatant was transferred to a new tube, and two volumes of pre-cooled ethanol were added. The mixture was incubated for 2 hours at -20°C, then centrifuged at 10,391 × g for 15 minutes, and the supernatant was discarded. The DNA pellets were washed with 70% ethanol and air-dried. Finally, the dried DNA pellets were resuspended in 100 μl TE buffer.

High-quality genomic DNA extracted from the FNA was used for sequencing on multiple platforms, including PacBio HiFi, Oxford Nanopore Technologies (ONT), Hi-C, and RNA-seq (**Table 1**). For HiFi sequencing, approximately 15–20 kb high-molecular-weight DNA fragments were size-selected and used to construct SMRTbell libraries with the SMRTbell Express Template Prep Kit 2.0 (PacBio, Menlo Park, USA). Libraries were sequenced on the PacBio Sequel II platform in circular consensus sequencing (CCS) mode, generating 20.97 Gb of data from 1,420,823 reads with an N50 read length of ~15 kb, corresponding to ~52.8× genome coverage. For ONT, high-molecular-weight DNA (~5–10 μg) was used to construct long-read libraries with the Ligation Sequencing Kit (SQK-LSK109, Oxford Nanopore Technologies, Oxford, UK). Libraries were loaded onto PromethION flow cells (R9.4.1) and sequenced until 16.25 Gb of data were obtained from 166,654 reads, with an N50 read length of 100 kb, achieving ~39.9× genome coverage. Hi-C libraries were prepared from young leaf tissues using the Arima-HiC kit (Arima Genomics, USA). Chromatin was crosslinked with formaldehyde, digested with a restriction enzyme cocktail, end-labeled with biotin, and ligated to generate proximity ligation products. DNA

was purified, sheared to ~350 bp fragments, and enriched for biotin-labeled fragments. Libraries were sequenced on the Illumina NovaSeq 6000 platform (Illumina, San Diego, USA) to generate 44.51 Gb of paired-end 150 bp reads, representing ~82.3× genome coverage.

Table 1. Summary of sequencing data for FNA rice

Data Type	Total Bases	Total Reads	Reads N50	Average Cov.
CCS (HiFi)	20.97 Gb	1,420,823	15 kb	52.8 X
ONT	16.25 Gb	166,654	100 kb	39.9 X
Hi-C	44.51 Gb	148,594,582	150 bp	82.3 X
RNA-seq	14.31 Gb	47,988,016	150 bp	102.7 X

RNA-seq Library Construction and Sequencing

Total RNA was extracted from a mixture of tissues, including root, stem, leaf, and flower, of the FNA, using the Eastep Total RNA Extraction Kit (Promega, Wisconsin, USA) according to the manufacturer's instructions. The RNA quality and quantity were assessed using a NanoDrop 2000 spectrophotometer, and integrity was verified by agarose gel electrophoresis. First-strand cDNA synthesis was performed using the PrimeScript RT Reagent Kit with gDNA Eraser (Takara, Japan) to remove residual genomic DNA. Strand-specific RNA-seq libraries were prepared and sequenced on the Illumina NovaSeq 6000 platform, generating 14.31 Gb of paired-end 150 bp reads from 47,988,016 reads, corresponding to ~102.7× transcriptome coverage (**Table 1**). The RNA-seq data were used to assist genome annotation, providing evidence for gene structure and transcript models across diverse tissues.

FNA T2T Genome Assembly and Completeness Evaluation

To construct a high-quality reference genome for FNA, we generated 20.97 Gb of PacBio HiFi reads (~52.8× coverage, with reads N50 of 15 kb), 16.25 Gb of Oxford Nanopore (ONT) long

reads (~40×, with reads N50 of 100 kb), and 44.51 Gb of Hi-C data (~82.3×). The HiFi reads, together with ONT long reads and Hi-C data, were assembled using Hifiasm software (version 0.19.5-r587) with default settings [24]. After gap-filling using TGS-GapCloser (v1.2.1) [25] with parameters of '-m 2000 -c 2 -l 100 -s minimap2 -p 80', we achieved a gap-free, chromosome-scale assembly comprising 12 chromosomes, with genome size of 394.7 Mb (**Fig. 1A, B**, **Table 2**). The genome assembly was visualized using Circos [26]. The assembly exhibits exceptional continuity with a contig N50 of 32 Mb. Hi-C sequencing data were processed using HiC-Pro (v3.1.0) [27], which performs quality control, read alignment, and generation of contact maps. For visualization, contact matrices generated by HiC-Pro were imported into HiCPlotter (v0.92)[28] to generate genome-wide Hi-C heatmaps and visualize chromosomal interactions. Chromatin interaction maps derived from Hi-C data confirmed precise chromosome assignment and strong organizational continuity (**Fig. 1B**).

The telomeric regions were predicted using Tandem Repeat Finder (v 4.09) [29] with parameters of '2 6 6 80 10 50 200', and the simple repeat with the repeat unit of either 'TAAACCC' or 'GGGTTTA' was considered as the telomeric repeat. All 24 telomeric regions at the both ends of 12 chromosomes were identified with telomeric repeat unit copy numbers ranging from 100 to 1080 (**Table 3**), indicative of telomere-to-telomere (T2T) level of FNA genome. The completeness of the genome was assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO, version 5.5.0) with fungi_odb10, ascomycota_odb10 databases [30]. Assessment with BUSCO using three lineage-specific datasets including Viridiplantae (n=425), Embryophyta (n=1,614), and Poales (n=3,386) revealed completeness rates of 99.5%, 98.9%, and 98.7%, respectively, indicating outstanding genome completeness (**Fig. 1C**). The HiFi reads were realigned to the assembly using Minimap2 and Samtools software [31,32]. Remarkably, 99.99% of HiFi reads and 99.85% of ONT reads could be remapped to the assembly, demonstrating exceptional assembly accuracy. Collectively, the FNA reference genome assembly demonstrates excellent quality in terms of accuracy, completeness, continuity and integrity.

Consistent with its breeding history as a CMS line developed through repeated backcrossing with its maintainer line 'Funong B', FNA exhibits near-complete genomic homozygosity. Comparative analysis of the two haplotype-resolved assemblies using Minimap2 and SyRI [31,33] software with default parameters revealed only 53 heterozygous SNVs/Indels (0.0134 heterozygous sites per Mb), further confirming its genetic uniformity. This exceptionally low heterozygosity underscores the efficacy of the backcrossing protocol in eliminating residual genetic variation.

Table 2. Genomic characteristics of FNA and other rice T2T genomes.

Genome Features	FNA	NIP	T197	ZH11
Genome size (bp)	394,697,788	385,710,679	395,084,276	379,330,463
Contig Number	92	-	-	-
Contig N50 (Mb)	32.10	31.11	31.53	31.07
Contig N90 (Mb)	25.62	27.47	26.56	23.89
Chr Number	12	12	12	12
Repeat Content	51.10%	51.10%	55.38%	50.66%
Gene Number	39,453	57,359	56908	57,178
Transcript Number	50,965	57,359	56908	57,178
GC Content	43.64%	43.71%	43.63%	43.52%
LTR Assembly Index (LAI)	21.99	21.68	20.78	21.61
BUSCO	99.00%	99.88%	98.80%	99.01%

Table 3. Telomeric regions in 12 chromosomes of FNA genome

Telomere Type	Chr	Start	End	Repeat Unit	Copy Num
5'Telomere	Chr1	1	3,866	CCCTAAA	558
3'Telomere	Chr1	45,324,362	45,329,017	TTTAGGG	669
5'Telomere	Chr2	1	2,184	CCTAAAC	317
3'Telomere	Chr2	37,460,464	37,462,194	TTAGGGT	250
5'Telomere	Chr3	282	3,664	CCTAAAC	486
3'Telomere	Chr3	40,069,290	40,071,035	AGGGTTT	252

5'Telomere	Chr4	5	5,744	ACCCTAA	832
3'Telomere	Chr4	35,539,245	35,543,193	TTTAGGG	568
5'Telomere	Chr5	1	5,124	CCCTAAA	737
3'Telomere	Chr5	31,379,812	31,383,567	TTAGGGT	536
5'Telomere	Chr6	1	2,316	CCCTAAA	331
3'Telomere	Chr6	32,415,086	32,418,284	AGGGTTT	462
5'Telomere	Chr7	4	1,514	ACCCTAA	216
3'Telomere	Chr7	31,136,607	31,138,558	AGGGTTT	281
5'Telomere	Chr8	1	2,843	TAAACCC	411
3'Telomere	Chr8	31,789,185	31,792,190	TAGGGTT	436
5'Telomere	Chr9	1	866	CCTAAAC	124
3'Telomere	Chr9	24,791,355	24,798,814	TTAGGGT	1080
5'Telomere	Chr10	5	704	TAAACCC	100
3'Telomere	Chr10	25,618,219	25,622,470	TTTAGGG	609
5'Telomere	Chr11	8	2,111	CCCTAAA	303
3'Telomere	Chr11	32,097,286	32,099,204	TAGGGTT	276
5'Telomere	Chr12	1	1,928	CCTAAAC	278
3'Telomere	Chr12	27,038,466	27,039,246	TAGGGTT	112

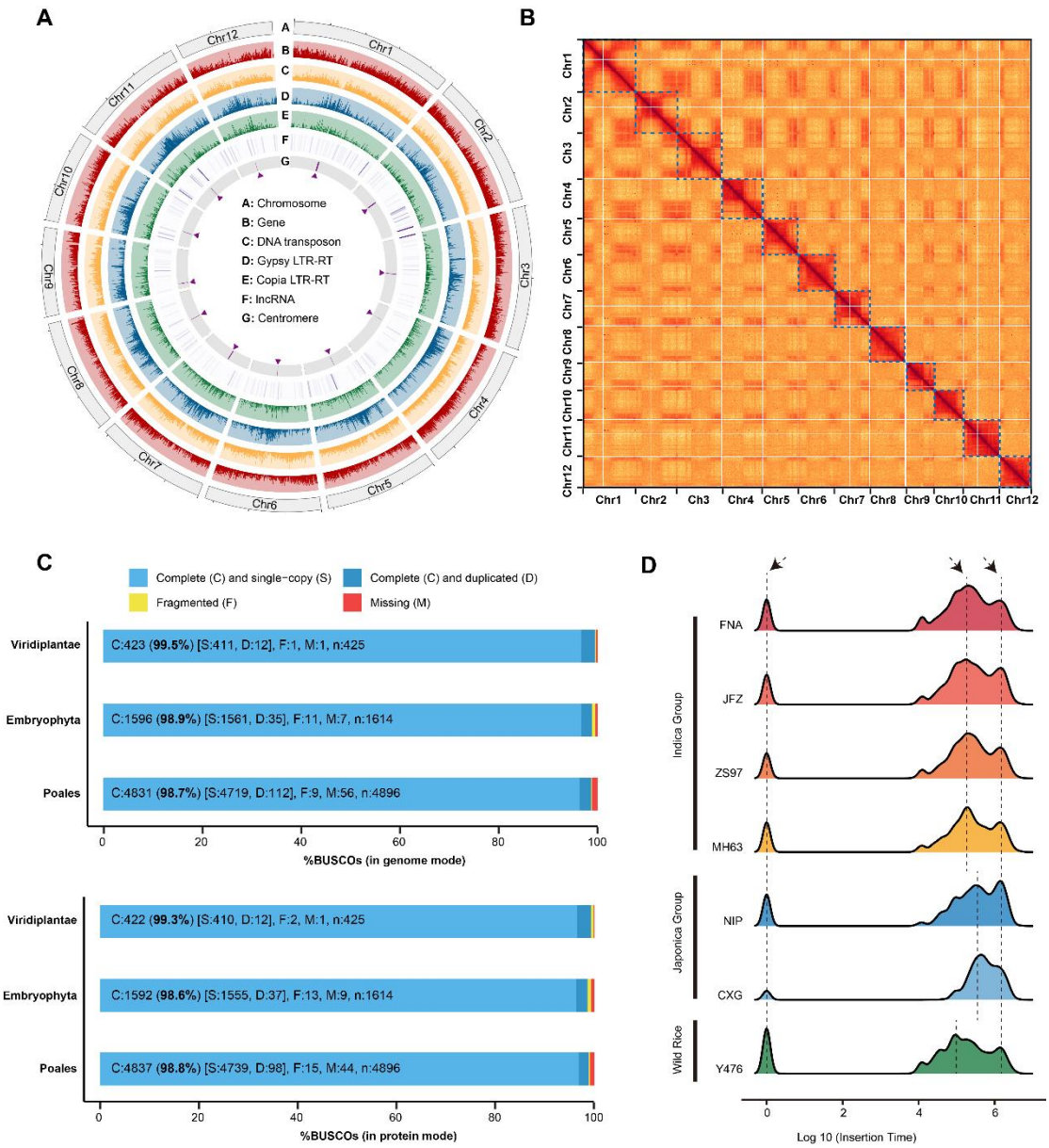

Figure 1. **A**) Circos plot displays the chromosomes (circle A), genomic density of gene (circle B), DNA transposon (circle C), Gypsy LTR retrotransposon (LTR-RT) (circle D), Copia LTR-RT (circle E), lncRNA (circle F) and centromere (circle G) using 100-kb windows. The positions marked with purple triangles indicate regions enriched for centromeric repeat sequences. **B**) Genome-wide Hi-C heatmap showing chromatin interactions across all 12 chromosomes of FNA. The intensity of the color represents the frequency of interactions, with higher intensity indicating stronger chromatin contacts. Chromosomes are ordered along both axes according to their genomic

positions. **C**) Completeness of the assembled genome and predicted proteins was evaluated with BUSCO (v5.0) in both genome and protein modes. Three lineage-specific datasets were used: Viridiplantae (n = 425), Embryophyta (n = 1,614), and Poales (n = 3,386). **D**) Ridge plot shows the distribution of LTR retrotransposon (LTR-RT) insertion times across FNA and six reference rice genomes. Three major bursts of LTR-RT activity are evident across all genomes. Notably, the second burst exhibits significant differences in peak insertion times among japonica, indica, and wild rice species, indicating lineage-specific retrotransposon activity

Gene Structure Prediction and Functional Annotation

Gene structure prediction was carried out using the GETA software version 2.5.7, which is accessible at its GitHub repository (https://github.com/chenlianfu/GETA). GETA integrates three different gene prediction approaches: the de novo method, homology-based prediction, and transcriptome-based prediction. Throughout this process, a suite of tools was applied, including PASA [34] , genewise [35], AUGUSTUS [36], Hisat2 [37], hmmsearch [38], BGM2AT, exonerate, and others. This comprehensive approach enabled the identification of 39,453 high-confidence protein-coding genes (**Table 1**). Further refinement yielded annotations for 50,965 mRNA transcripts, revealing that 6,129 genes (15.5% of the total) possess alternative transcript isoforms, indicative of complex regulatory potential. Additionally, 1,376 long non-coding RNA (lncRNA) loci were annotated using the GETA pipeline. The exceptional completeness and accuracy of this annotation were quantitatively validated through BUSCO assessment against different lineage-specific datasets, achieving a near-complete score of 99.3% for Viridiplantae_odb10 database (**Fig. 1C**).

To infer gene functions, interpro-scan and eggNOG-Mapper were used, and a range of gene function databases were consulted for detailed functional annotations. These databases included Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Pfam, Clusters of Orthologous Groups (COG), and the Swiss-Prot protein database [39-45]. This analysis successfully assigned putative functions to over 81% of the predicted protein-coding genes, providing robust orthogonal validation of both the structural annotation's reliability and the biological relevance of

the gene models. The exceptionally high functional annotation rate, coupled with the benchmark BUSCO completeness score, strongly supports the accuracy and comprehensiveness of the final genome annotation.

Repeat Landscape in FNA Genome

For repeat annotation in the two genomes, the de novo identification of repeat elements was conducted with RepeatModeler (version 2.0.3) [46]. The repeat consensus sequences identified were then used as seeds for RepeatMasker (version 4.1.0) [47] to locate and analyze all repeat regions within the genome. De novo repeat annotation revealed that repetitive elements constitute 202 Mb (51.1%) of the FNA genome assembly (**Table 4**). Long terminal repeat retrotransposons (LTR-RTs), predominantly of the Gypsy lineage, represent the major repeat class (26.13% of the genome). DNA transposons comprise the second-largest category (16.44%), dominated by MULE-MuDR (5.3%), CMC-EnSpm (3.42%), PIF-Harbinger (3.25%), and TcMar-Stowaway (2.22%). Chromosomal distribution analysis indicates association between DNA transposons and genic regions (**Fig. 1A**), suggesting their potential influence on gene structure and regulatory mechanisms. Conversely, LTR-RTs exhibit an inverse correlation with gene density and were significantly enriched in centromeric regions, implicating their involvement in centromere architecture and 3D genome organization.

Table 4. Repeat contents in FNA genome

Class	Number	Length	Percentage
Total	679903	201718471	51.10%
Non-interspersed Repeats	119248	11554042	2.92%
Satellite	2111	5565188	1.40%
Simple_repeat	104939	5116559	1.29%
Interspersed Repeats	560655	191286071	48.46%
Retroelements	137605	110296920	27.94%

LTR	108761	103153624	26.13%
Gypsy	81543	89321272	22.63%
Copia	18646	12724897	3.22%
LINE	10725	4562217	1.15%
SINE	12615	1078696	0.27%
DNA transposons	350478	64893978	16.44%
MULE-MuDR	87508	20958219	5.30%
CMC-EnSpm	32472	13515074	3.42%
PIF-Harbinger	91200	12857939	3.25%
TcMar-Stowaway	79705	8799131	2.22%
hAT-Ac	16081	3158623	0.80%
hAT-Tip100	9968	1991183	0.50%

To predict full-length LTR retrotransposons (LTR-RTs), a combination of tools was used, including LTR_finder (version 1.07) [48], LTRharvest (version 1.6.2) [49], and LTR_retriever (version 2.9.0) [50]. Analysis of intact LTR-RTs identified 2,738 full-length elements. Assessment of genome assembly quality using these elements yielded an LTR Assembly Index (LAI) value of 21.99, confirming high assembly continuity. Analysis of LTR-RT insertion time by estimating divergence between 5'/3' LTR pairs revealed 285 elements with identical LTRs, signifying recent insertion events. Temporal density analysis resolved three distinct amplification bursts of LTR-RTs, including a pronounced recent event (**Fig. 1D**). Comparative analysis of intact LTR-RT insertion times across six additional gap-free rice genomes – three *indica* (JFZ, ZS97, MH63), two *japonica* (NIP, CXG), and one wild rice (Y476) – identified three corresponding burst events in all seven genomes (**Fig. 1D**). The oldest burst (occurring millions of years ago) was highly conserved across all accessions, predating the divergence of these rice lineages. The most recent burst was also ubiquitous, potentially linked to domestication processes and environmental adaptation pressures. Interestingly, the timing of the second burst event varied among groups: it occurred most recently

in wild rice (~70-100 thousand years ago, kya), followed by *indica* (~150-200 kya), and was oldest in *japonica* (~250-300 kya). This differential timing likely reflects lineage-specific evolutionary pressures and demographic histories influencing retrotransposon activity after the initial divergence but before domestication.

Centromeric regions were completely assembled and annotated on all 12 FNA chromosomes. These regions were characterized by significant enrichment of 155/165 bp centromeric monomer repeats and LTR-RTs. Centromere lengths varied substantially, ranging from 17.9 kb (Chr4) to 950 kb (Chr1). Compositional heterogeneity was observed: chromosomes 2-8 were primarily enriched with the 155 bp monomer repeat, chromosomes 9 and 10 predominantly contained the 165 bp repeat unit, while chromosomes 1, 11, and 12 harbored a mixture of both 155 bp and 165 bp repeat units. This chromosome-specific centromeric repeat composition indicates evolutionary divergence in the organization and potential functional specialization of centromeres across the FNA genome.

Data Records

The raw sequencing reads generated for the FNA genome assembly have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject PRJNA1373546. Specifically, these include ONT ultra-long reads (SRR36326080), in situ Hi-C data (SRR36326079), PacBio HiFi reads (SRR36326081), and RNA-seq data from mixed tissues (SRR36326078) [51]. The assembled genome sequences have been submitted to NCBI GenBank under accession GCA_054286825.1 [52]. The genome annotation files, including the gene annotation (GFF3), cDNA sequences, and protein sequences, have been deposited in Figshare and are publicly available at (https://doi.org/10.6084/m9.figshare.32244801) [53].

Technical Validation

The quality of the Funong A (FNA) genome assembly was rigorously validated through multiple

complementary approaches. All 24 telomeric regions were successfully identified across the 12 chromosomes, with telomeric repeat copy numbers ranging from 100 to 1,080, and centromeric regions were fully resolved on every chromosome, confirming the telomere-to-telomere (T2T) and gap-free nature of the assembly. Read remapping demonstrated exceptional accuracy, with 99.99% of PacBio HiFi reads and 99.85% of Oxford Nanopore reads aligning successfully to the assembly. Completeness assessments using BUSCO (v5.5.0) against Viridiplantae, Embryophyta, and Poales datasets yielded scores of 99.5%, 98.9%, and 98.7%, respectively, while the LTR Assembly Index (LAI) of 21.99 further corroborated high continuity and integrity. Hi-C chromatin interaction maps exhibited strong intra-chromosomal signals, validating correct chromosome-scale scaffolding and organization. Genome annotation was supported by multi-tissue RNA-seq data, enabling the identification of 39,453 protein-coding genes and 1,376 lncRNAs, with over 81% of genes functionally annotated and BUSCO completeness exceeding 99% in protein mode. Comparative analyses with other high-quality rice genomes confirmed annotation accuracy and orthology assignments. Finally, only 53 heterozygous variants were detected across the entire genome, consistent with the near-complete homozygosity expected from the recurrent backcrossing history of this CMS line. Collectively, these metrics establish the FNA assembly as a high-quality, accurate, and complete reference genome suitable for downstream genomic and breeding studies.

Code availability

The bioinformatic tools, including version and parameters, that were used in this study have been described in the Method section. No customized script was used.

Acknowledgements, Author Contributions, Funding & Competing Interests

Acknowledgements

We thank all members of the Rice Research Institute at Fujian Academy of Agricultural Sciences for their assistance in plant cultivation, sample collection, and valuable discussions throughout this

project. We are particularly grateful to Zuyin Biology for their expert assistance in bioinformatics analysis.

Author Contributions

S.T. and Y.H. conceived the study. P.Z., F.Z., H.Z. and R.H. prepared the rice samples and performed experiments. S.T. and J.Z. conducted the bioinformatic analyses. S.T., Y.H., H.Z. and J.Z. wrote the manuscript. S.T. and Y.H. supervised the project. All authors reviewed and provided comments on the manuscript.

Funding

This work was supported by the Special Program for Outstanding Scientific and Technological Innovation Talent of Fujian Academy of Agricultural Sciences (YCZX202402), the Natural Science Foundation of Fujian Province (2024J01323), the Fujian Provincial Basic Research Fund for Nonprofit Research Institutes (2024R1022003 and 2025R1021002), and the Technology Innovation Project of Fujian Academy of Agricultural Sciences (ZYTS202402).

Competing interests

The authors declare no conflict of interest.

Reference

1 Bouman, B. *et al.* in *Water for Food Water for Life* 515-550 (Routledge, 2013).

2 Rezvi, H. U. A. *et al.* Rice and food security: Climate change implications and the future prospects for nutritional security. **12**, e430 (2023).

3 Mohidem, N. A., Hashim, N., Shamsudin, R. & Che Man, H. J. A. Rice for food security: Revisiting its production, diversity, rice milling process and nutrient content. **12**, 741 (2022).

4 Fukagawa, N. K., Ziska, L. H. J. J. o. n. s. & vitaminology. Rice: Importance for global nutrition. **65**, S2-S3 (2019).

5 Jagadish, S. V., Craufurd, P. Q. & Wheeler, T. R. High temperature stress and spikelet fertility in rice (Oryza sativa L.). *Journal of experimental botany* **58**, 1627-1635 (2007). https://doi.org/10.1093/jxb/erm003

6 Zeigler, R. S., Leong, S. A. & Teng, P. S. *Rice blast disease*. (Int. Rice Res. Inst., 1994).

7 Asibi, A. E., Chai, Q. & Coulter, J. A. Rice blast: A disease with implications for global food security. *Agronomy* **9**, 451 (2019).

8 Sun, T. *et al.* Current rice models underestimate yield losses from short-term heat stresses. **27**, 402-416 (2021).

9 Kumar, K. P., Pushpam, R., Manonmani, S., Raveendran, M. & Senthil, A. J. E. J. o. P. B. Deciphering gene action for traits contributing to submergence tolerance and yield in rice (Oryza sativa L.). **15**, 63-69 (2024).

10 Xu, C.-C. *et al.* Simulating climate change impacts and potential adaptations on rice yields in the Sichuan Basin, China. **22**, 565-594 (2017).

11 Bhattacharya, J., Nitnavare, R. B., Bhatnagar-Mathur, P. & Reddy, P. S. J. P. Cytoplasmic male sterility-based hybrids: mechanistic insights. **260**, 100 (2024).

12 Bohra, A., Jha, U. C., Adhimoolam, P., Bisht, D. & Singh, N. P. J. P. C. R. Cytoplasmic male sterility (CMS) in hybrid breeding in field crops. **35**, 967-993 (2016).

13 Chapagain, S. J. A. S. M. A Comprehensive Review On-Application And Future Prospect Of Cytoplasmic Male Sterility. **8**, 21-27 (2024).

14 Wu, Z. *et al.* Breeding and Application of Quality Indica CMS Line Taonong 1A in Rice. **18**, 760-762 (2017).

15 Shi-Hang, T. *et al.* Breeding of the Three-line Indica Cytoplasmic Male Sterile Line 'Funong A' and Its Comprehensive Agronomic Trait Analysis. **32**, 701-711 (2024).

16 Ling, L. *et al.* Disease and Insect Resistance Genes in Premium Sterile Line Funong A. **39**, 1007-1016 (2024).

17 Li, H. & Durbin, R. J. N. R. G. Genome assembly in the telomere-to-telomere era. **25**, 658-670 (2024).

18 Garg, V. *et al.* Unlocking plant genetics with telomere-to-telomere genome assemblies. **56**, 1788-1799 (2024).

19 Miga, K. H. J. E. c. r. Centromere studies in the era of 'telomere-to-telomere'genomics. **394**, 112127 (2020).

20 Song, J.-M. *et al.* Two gap-free reference genomes and a global view of the centromere architecture in rice. **14**, 1757-1767 (2021).

21 Zhang, Y. *et al.* The telomere-to-telomere gap-free genome of four rice parents reveals SV and PAV patterns in hybrid rice breeding. **20**, 1642-1644 (2022).

22 Shang, L. *et al.* A complete assembly of the rice Nipponbare reference genome. **16**, 1232-1236 (2023).

23 Tao, W. *et al.* A gap-free and haplotype-resolved genome of an early-season indica variety provides insights into rice thermotolerance and grain quality. S1673-8527 (1625) 00129-00128 (2025).

24 Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods* **18**, 170-175 (2021). https://doi.org/10.1038/s41592-020-01056-5

25 Xu, M. *et al.* TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. **9**, giaa094 (2020).

26 Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome research* **19**, 1639-1645 (2009). https://doi.org/10.1101/gr.092759.109

27 Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome biology* **16**, 259 (2015). https://doi.org/10.1186/s13059-015-0831-x

28 Akdemir, K. C. & Chin, L. HiCPlotter integrates genomic data with interaction matrices. *Genome biology*

16, 198 (2015). https://doi.org/10.1186/s13059-015-0767-1

29 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**, 573-580 (1999). https://doi.org/10.1093/nar/27.2.573

30 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 (2015). https://doi.org/10.1093/bioinformatics/btv351

31 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018). https://doi.org/10.1093/bioinformatics/bty191

32 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009). https://doi.org/10.1093/bioinformatics/btp352

33 Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome biology* **20**, 277 (2019). https://doi.org/10.1186/s13059-019-1911-0

34 Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**, R7 (2008). https://doi.org/10.1186/gb-2008-9-1-r7

35 Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Research* **14**, 988-995 (2004). https://doi.org/10.1101/gr.1865504

36 Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435-439 (2006). https://doi.org/10.1093/nar/gkl200

37 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* **37**, 907-915 (2019). https://doi.org/10.1038/s41587-019-0201-4

38 Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC bioinformatics* **11**, 431 (2010). https://doi.org/10.1186/1471-2105-11-431

39 Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution* **34**, 2115-2122 (2017). https://doi.org/10.1093/molbev/msx148

40 Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014). https://doi.org/10.1093/bioinformatics/btu031

41 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28**, 27-30 (2000). https://doi.org/10.1093/nar/28.1.27

42 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25-29 (2000). https://doi.org/10.1038/75556

43 Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* **31**, 365-370 (2003). https://doi.org/10.1093/nar/gkg095

44 Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Research* **40**, D290-301 (2012). https://doi.org/10.1093/nar/gkr1065

45 Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* **28**, 33-36 (2000). https://doi.org/10.1093/nar/28.1.33

46 Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families.

Proceedings of the National Academy of Sciences of the United States of America **117**, 9451-9457 (2020). https://doi.org/10.1073/pnas.1921046117

47 Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* **Chapter 4**, Unit 4 10 (2004). https://doi.org/10.1002/0471250953.bi0410s05

48 Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic acids research* **35**, W265-268 (2007). https://doi.org/10.1093/nar/gkm286

49 Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC bioinformatics* **9**, 18 (2008). https://doi.org/10.1186/1471-2105-9-18

50 Ou, S. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant physiology* **176**, 1410-1422 (2018). https://doi.org/10.1104/pp.17.01310

51 NCBI Sequence Read Archive: https://identifiers.org/ncbi/insdc.sra:SRR36326078, SRR36326079, SRR36326080, SRR36326081 (2026)

52 NCBI GenBank https://identifiers.org/ncbi/insdc.gca: GCA_054286825.1 (2026)

53 Gene annotation of FNA T2T genome. figshare. https://doi.org/10.6084/m9.figshare.32244801 (2026)