scientific data

OPEN

DATA DESCRIPTOR

Chromosome-scale and haplotype-resolved genome assembly of the autotetraploid *Misgurnus anguillicaudatus*

Bing Sun[1,3], Qingshan Li[1,3], Yihui Mei[1], Yunbang Zhang[1], Yuxuan Zheng[1], Yuwei Huang[1], Xinxin Xiao[1], Jianwei Zhang[2], Gao Jian[1] ✉ & Xiaojuan Cao[1] ✉

In nature, diploids and tetraploids are two common types of polyploid evolution. *Misgurnus anguillicaudatus* (mud loach) is a remarkable fish species that exhibits both diploid and tetraploid forms. However, reconstructing the four haplotypes of its autotetraploid genome remains unresolved. Here, we generated the first haplotype-resolved, chromosome-level genome of autotetraploid *M. anguillicaudatus* with a size of 4.76 Gb, contig N50 of 6.78 Mb, and scaffold N50 of 44.11 Mb. We identified approximately 2.9 Gb (61.03% of genome) of repetitive sequences and predicted 91,485 protein-coding genes. Moreover, allelic gene expression levels indicated the absence of significant dominant haplotypes within the autotetraploid loach genome. This genome will provide a valuable biological model for unraveling the mechanisms of polyploid formation and evolution, adaptation to environmental changes, and benefit for aquaculture applications and biodiversity conservation.

Background & Summary

Polyploidy is widely recognized as a pivotal mechanism that contributes significantly to genetic diversity, facilitates genomic restructuring and the evolution of novel traits in organisms[1,2]. In plants, comprehensive studies in polyploidy organisms have been investigated, including some model autotetraploid organisms, such as *Saccharum spontaneum* L[3]., *Solanum tuberosum*[4] and "Zhongmu No.1" alfalfa[5]. However, our understanding of polyploidy in vertebrates remains limited, as it is comparatively infrequent in this group of organisms[6]. *Misgurnus anguillicaudatus* is a small-size freshwater teleost, widely distributed in China, Japan, Korea and other Southeast Asian countries[7]. Cytogenetic investigations conducted on *M. anguillicaudatus* have revealed the presence of naturally occurring diploid, triploid, tetraploid, pentaploid, and hexaploid loach populations in China[8]. Among these polyploid loach, the natural tetraploid loach (4n = 100 chromosomes) appear with sympatric diploid loach (2n = 50 chromosomes) and they are regarded as a autotetraploid formed by chromosome doubling[9,10]. The coexistence of diploid and autotetraploid individuals within this species makes it an exceptional model organism for investigating the evolutionary history and trajectory of polyploidization in vertebrates. However, the genomic data available for the loach is currently insufficient.

Here, we have successfully constructed a high-quality, haplotype-resolved, chromosome-level reference genome for the autotetraploid *M. anguillicaudatus*, employing a combination of Illumina short-read sequencing, PacBio long-read sequencing (Circular Consensus Sequencing, CCS), and Hi-C technologies. The reference genome created for autotetraploid *M. anguillicaudatus* bears significant importance in the fields of evolutionary biology, genomic stability research, aquaculture applications, and biodiversity conservation.

[1]College of Fisheries, Engineering Research Center of Green development for Conventional Aquatic Biological Industry in the Yangtze River Economic Belt, Ministry of Education, Huazhong Agricultural University, Wuhan, 430070, China. [2]National Key Laboratory of Crop Genetic Improvement, Hubei Hongshan Laboratory, Huazhong Agricultural University, Wuhan, 430070, China. [3]These authors contributed equally: Bing Sun, Qingshan Li. ✉e-mail: gaojian@mail.hzau.edu.cn; caoxiaojuan@mail.hzau.edu.cn

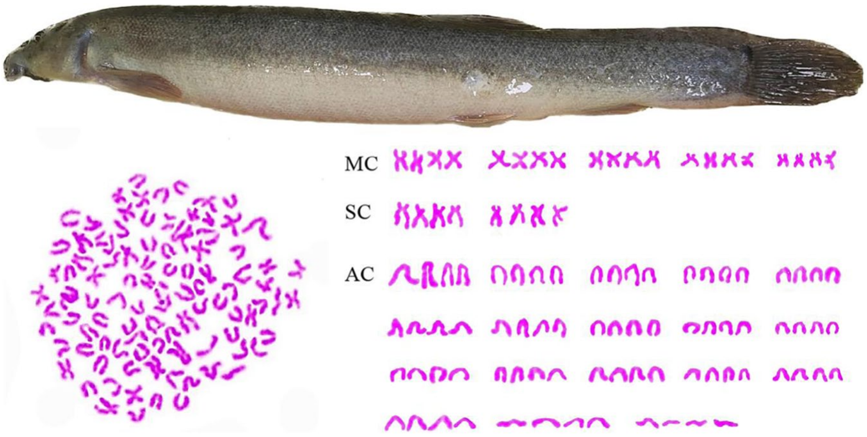

Fig. 1 Autotetraploid *Misgurnus anguillicaudatus* (Female) and karyotype analysis.

Library types	Insert size (bp)	Raw data (Gb)	Clean data (Gb)	Read length (bp)	Sequence coverage (X)
Illumina reads	350	253.85	242.72	150	105.5
PacBio reads (HiFi reads)	15000	92.9	/	14351.39	26.29
Hi-C reads	100–500	423.69	423.61	150	184.2
RNA reads	350	5.17	4.98	150	

Table 1. Sequencing data used for the genome autotetraploid *Misgurnus anguillicaudatus* assembly.

Methods

Experimental fish and sequencing.

We firstly conducted karyotype analysis of an autotetraploid *M. anguillicaudatus* (female) obtained from the Aquaculture Base of College of Fisheries, Huazhong Agricultural University in Wuhan City, Hubei Province, China, and 100 chromosomes were divided into three types (metacentric chromosome (MC), submetacentric chromosome (SC) and acrocentric chromosome (AC)) according to the position of centromere (Fig. 1). Genomic DNA was extracted from muscle tissues of the autotetraploid *M. anguillicaudatus* for sequencing. All experimental protocols in this study were approved by the Animal Experimental Ethical Inspection of Laboratory Animal Center, Huazhong Agricultural University, Wuhan, China (HZAUFI-2022-0025). All efforts were made to minimize the suffering of the fish.

For genome survey, we utilized the Illumina TruSeq Nano DNA Library Prep Kit to construct a paired-end (PE) library, which was subsequently sequenced on the Illumina HiSeq X Ten platform, generating 150-bp read length sequences[11]. fastp (v 0.23.4) was performed to eliminate adaptors and low-quality reads from the raw data, resulting in a total of 242.72 Gb clean data (105.5 × coverage). These clean reads were then utilized to estimate the genome size and heterozygosity (Table 1).

To perform *de novo* assembly of the autotetraploid *M. anguillicaudatus* genome, SMRT bell Template Prep Kit was employed to construct the library, which was subsequently sequenced on the PacBio Sequel II platform. After removing redundant reads, a total of 92.9 Gb HiFi reads (26.29 × coverage) with an average length of 14,351.39 bp were obtained (Table 1).

For chromosome-level assembly of the autotetraploid *M. anguillicaudatus* genome, a Hi-C library was generated using the DpnII restriction enzyme and sequenced on the MGISEQ-T7 platform, resulting in 423.61 Gb (184.2 × coverage) of clean data after[12] (Table 1).

For genome structure annotation, samples were taken from nine tissues of autotetraploid *M. anguillicaudatus*, including muscle, liver, fin, skin, eye, brain, gill, intestine, and ovary. These samples were mixed and subjected to RNA sequencing for genome annotation. RNA-seq libraries were constructed following the protocol and sequenced on the Illumina NovaSeq6000 platform, generating a total of 5.17 Gb raw data. (Table 1).

De novo assembly of autotetraploid *M. anguillicaudatus* genome.

K-mer method was used to survey the genomic features of the autotetraploid *M. anguillicaudatus*[13]. Based on the total number of 216,747,902,623 17-mers and a peak 17-mer depth of 93, estimated genome size and heterozygosity rate of the *M. anguillicaudatus* was 2,315 Mb and 1.38%, respectively. (Fig. 2 and Table s1). Subsequently, a total of 92.9 Gb HiFi and 423.61 Gb Hi-C reads were used for genome assembly using HiFiasm (v 0.16.1, default parameters)[14], and then polished by Pilon (v 1.23)[15] using the 242.72 Gb of Illumina HiSeq clean reads. Finally, we obtained a 4,761 Mb genome with a contig N50 size of 6.78 Mb, consisting of 215 contigs (Table 2). Subsequently, the completeness of the assembly was assessed using BWA software (v 0.7.17, default parameters)[16], which revealed that 99.85% of the Illumina short reads could be mapped and covered 99.83% of the assembled genome (Table s2). We also assessed the assembly using Pacbio long reads with the minimap2 software (v 2.23, default parameters). The analysis showed that all reads were successfully mapped, covering 99.98% of the assembled genome[17] (Table s2).

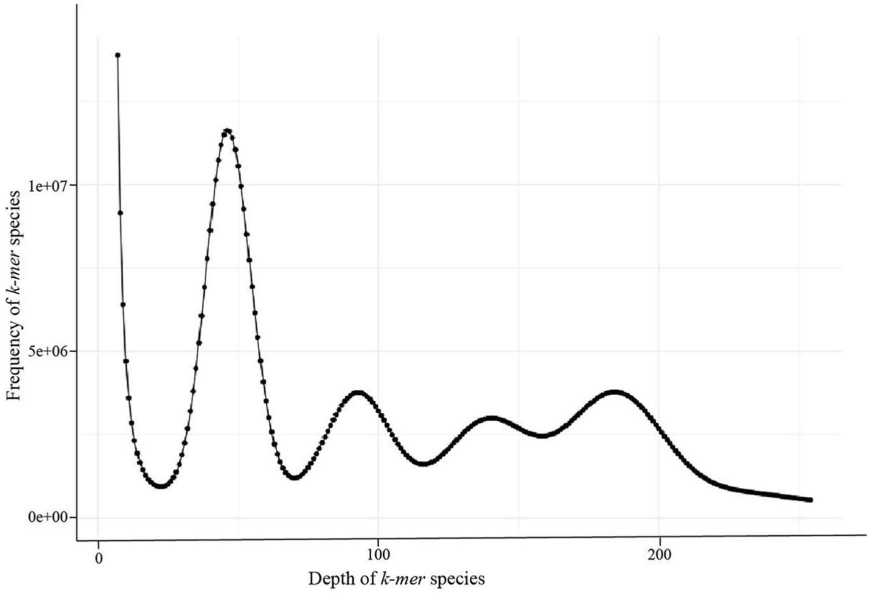

Fig. 2 17-kmer distribution in the autotetraploid *Misgurnus anguillicaudatus* genome.

	Contig Length (bp)	Contig Number	Scaffold Length (bp)	Scaffold Number
N90	13,60,784	772	3,61,33,691	93
N80	28,33,278	537	3,83,49,178	80
N70	39,67,880	395	4,09,09,013	68
N60	54,54,833	293	4,25,04,299	57
N50	67,82,996	215	4,41,14,842	46
Total length	4,76,07,08,595	—	4,76,08,41,395	—
Number(>=100 bp)	—	5,257	—	3,929
Number(>=2 kb)		5,257	—	3,929
Max length	3,42,59,278	—	7,32,11,041	—

Table 2. The statistics of length and number for the de novo assembled autotetraploid *Misgurnus anguillicaudatus* genome.

Finally, the completeness of the genome assembly was assessed using BUSCO (v 4.1.3) with the actinopterygii_odb10 database[18]. The analysis revealed that 96.81% of the genes were identified as complete, including 3.85% of single-copy genes and 92.97% of duplicated genes. Additionally, 0.36% of the genes were fragmented, and 2.83% were missing from the assembled genome (Table 3).

The Hi-C technique was employed to assemble the chromosome-level genome of autotetraploid *M. anguillicaudatus*[19]. Hi-C clean reads were mapped to the contig-level genome with an end-to-end algorithm implemented in Bowtie (v 2.4.1, default parameters)[20] following the Hi-C-Pro (v 2.11.1, default parameters)[21] strategy. To anchor these contigs into chromosomes, Juicer (v 1.5.6, default parameters)[22] and 3d-DNA (v 201008, default parameters)[23] were utilized. Subsequently, juicebox (v 1.11.08, default parameters) was utilized to fix error-joins and remove duplicated contigs. Finally, the assembled sequences were anchored and orientated onto 100 pseudo-chromosomes, covering 95.43% of the entire genome. The final chromosome-scale genome assembly of the autotetraploid *M. anguillicaudatus* was 4.76 Gb with a contig N50 of 6.78 Mb and scaffold N50 of 44.11 Mb (Figs. 3, 4; Table 3).

Repeat annotation. *De novo* and homology-based approaches were used to identify the repetitive elements of the autotetraploid *M. anguillicaudatus* genome. Tandem Repeats Finder (http://tandem.bu.edu/trf/trf.html)[24] was used to extract tandem elements. For homology prediction, RepeatMasker (v 4.1.0, default parameters)[25] and RepeatProteinMask (v 1.36, default parameters)[26] were employed, using based the RepBase database (http://www.girinst.org/repbase) as a reference. For *ab initio* prediction, RepeatModeler (v 2.0.1, default parameters)[27] (http://www.repeatmasker.org/RepeatModeler.html) and LTR-FINDER (v 1.07, default parameters) (http://tlife.fudan.edu.cn/ltr_finder/)[28] were utilized to detect repetitive elements based on *de novo* repetitive element databases. Finally, we identified about 2.9 Gb (61.03%) of repetitive sequences in autotetraploid *M. anguillicaudatus* genome. Among these elements, DNA sequences were found to be the most abundant, comprising 38.48% of the genome. In contrast, SINEs were the least prevalent, making up only 0.84% of the genome (Table s3).

Type	Proteins	Percentage (%)
Complete BUSCOs (C)	3,524	96.81
Complete and single-copy BUSCOs (S)	140	3.85
Complete and duplicated BUSCOs (D)	3,384	92.97
Fragmented BUSCOs (F)	13	0.36
Missing BUSCOs (M)	103	2.83
Total BUSCO groups searched	3,640	100
Hi-C analysis		
Total Sequence Length (bp)	4,76,08,41,395	
Contig N50 (bp)	67,82,996	
Scaffold N50 (bp)	4,41,14,842	
Chromosome anchoring rate (%)	95.43	

Table 3. Statistics of the assembled genome for the autotetraploid *Misgurnus anguillicaudatus*.

Gene prediction and annotation.

Gene predictions were based on *de novo* prediction, homology-based annotation and transcriptome-based annotation. For the *de novo* prediction, Genscan (v 2003, parameters: HumanIso.smat)[29] and Augustus (v 3.3.3, parameters:–species = zebrafish–uniqueGeneId = true–noInFrameStop = true–gff3 = on–strand = both)[30] were used to predict coding regions. For homology-based prediction, the protein sequences from *Sinocyclocheilus graham*, *Danio rerio*, *Carassius auratus* and *Cyprinus carpio* genome were download from the public NCBI database (release 75) for aligning to the autotetraploid *M. anguillicaudatus* genome by BLAST + (v 2.9.0, *e*-value ≤ 10^{-5})[31]. Subsequently, GeneWise (v 2.4.1, default parameters)[32] was used to identify gene structure of each protein region. For transcriptome-based annotation, the clean RNA-seq reads were from nine tissues mixed RNA samples (including muscle, liver, fin, skin, eye, brain, gill, intestine, and ovary) mapped onto the autotetraploid *M. anguillicaudatus* genome by using Tophat (v 2.1.1, default parameters)[33]. Cufflinks (v 2.2.1, default parameters) (http://cole-trapnell-lab.github.io/cufinks/)[34] was applied for genome-based transcript assembly. Finally, MAKER (v 3.01.03)[35] was used to integrate the results from the three methods (above-mentioned) and generate a final non-redundant gene set. In the autotetraploid *M. anguillicaudatus* genome, 92,500 protein-coding genes were predicted. The average gene length was 19,975 bp, while the average coding sequence (CDS) length was 1,668 bp. Each gene contains 9.44 exons with an average exon length of 220.16 bp and the introns length of 21,121 bp. It was worth mentioned that these gene model statistics, encompassing gene, CDS, intron, and exon lengths, were observed to be similar to those observed in closely related species (Table s4 and Fig. s1). For non-coding RNAs annotation, tRNAscan-SE (v 1.3.1)[36] was applied to search the tRNA. Based on the Rfam (v 14.0, parameters: cmscan–rfam–nohmmonly) database[37], BLASTN was used to detect the microRNA and rRNA. Finally, we totally identified 5,742 miRNAs, 31,484 rRNAs and 38,281 tRNAs (Table s5).

For gene functional annotations, InterProscan (v 5.55–88.0, parameters:–applications ProDom, PRINTS, Pfam, SMART, PANTHER, ProSiteProfiles–goterms–pathways)[38] was used to screen proteins against InterPro database. GO (v 20171220, parameters: blastp -*e*-value 10^{-5})[39], KEGG (v kobas-3.0, parameters: blastp -*e*-value 10^{-5})[40], NR (diamond:v 0.9.27, parameters: blastp -*e*-value 10^{-5})[41], SwissProt (diamond:v 0.9.27, parameters: blastp -*e*-value 10^{-5})[42] and TrEMBL databases (diamond:v 0.9.27, parameters: blastp -*e*-value 10^{-5})[43] were used for gene functional annotations using BLAST + (v 2.9.0, *e*-value ≤ 10^{-5}). A total of 91,485 genes (98.9%) were successfully annotated (Table s6). In addition, the completeness of the gene annotation of autotetraploid *M. anguillicaudatus* was further assessed by using BUSCO (actinopterygii_odb9 database) (v 4.1.3). Our analysis revealed that 98.4% of the complete genes and 0.3% of the fragmented genes out of the total 3,640 BUSCO genes were detected in the genome (Table s7).

Haplotype assembly.

Firstly, the diploid loach genome was used here[44,45]. By comparing the homologous gen sets from the autotetraploid loach and the diploid loach with jcvi (v 1.0.6, parameters: jcvi.compara.catalog ortholog-dbtype = prot-cscore 0.7; jcvi.compara.synteny screen-minspan = 30)[46], we identified the single-copy homologous genes from corresponding chromosomes in five genomes (four from autotetraploid (h1, h2, h3 and h4 for short), and one from diploid (Mis for short)). For example, if we obtained about 500 groups of homologous genes on chromosome 1 (chr1 for short), it meant that each group consisted of five homologous genes (from autotetraploid (chr1-h1, chr1-h2, chr1-h3, chr1-h4) and diploid (Mis)). Secondly, all single-copy homologous genes were used to construct gene trees by raxml (v 1.2.0, parameters: raxmlHPC PTHREADS-f a-N 100-m GTRGAMMA)[47]. Finally, according to the corresponding chromosome, the gene trees were integrated by Astral (v 5.6.3, default parameters), and the final consensus tree was used to distinguish haplotype chromosomes (Fig. s2)[48]. Moreover, we performed the collinearity analysis among diploid and four haplotype genomes using jcvi (v 1.0.6, parameters: jcvi.compara.catalog ortholog-dbtype = prot-cscore 0.99; jcvi.compara.synteny screen-minspan = 30). A total of 65,387 gene pairs from diploid and autotetraploid genomes were identified and showed a high collinearity (Fig. s3). In order to evaluate the assembly quality of four haplotypes, we analyzed the GC content distribution, short-read depth distribution, long-read depth distribution, homozygous snp density distribution, heterozygous snp density distribution, homozygous indel density distribution and heterozygous indel density distribution (Figs. s4–s8). Above all results showed that we obtained four high-quality chromosome-level haplotypes from autotetraploid loach genome. Moreover, the Ks values of syntenic gene pairs

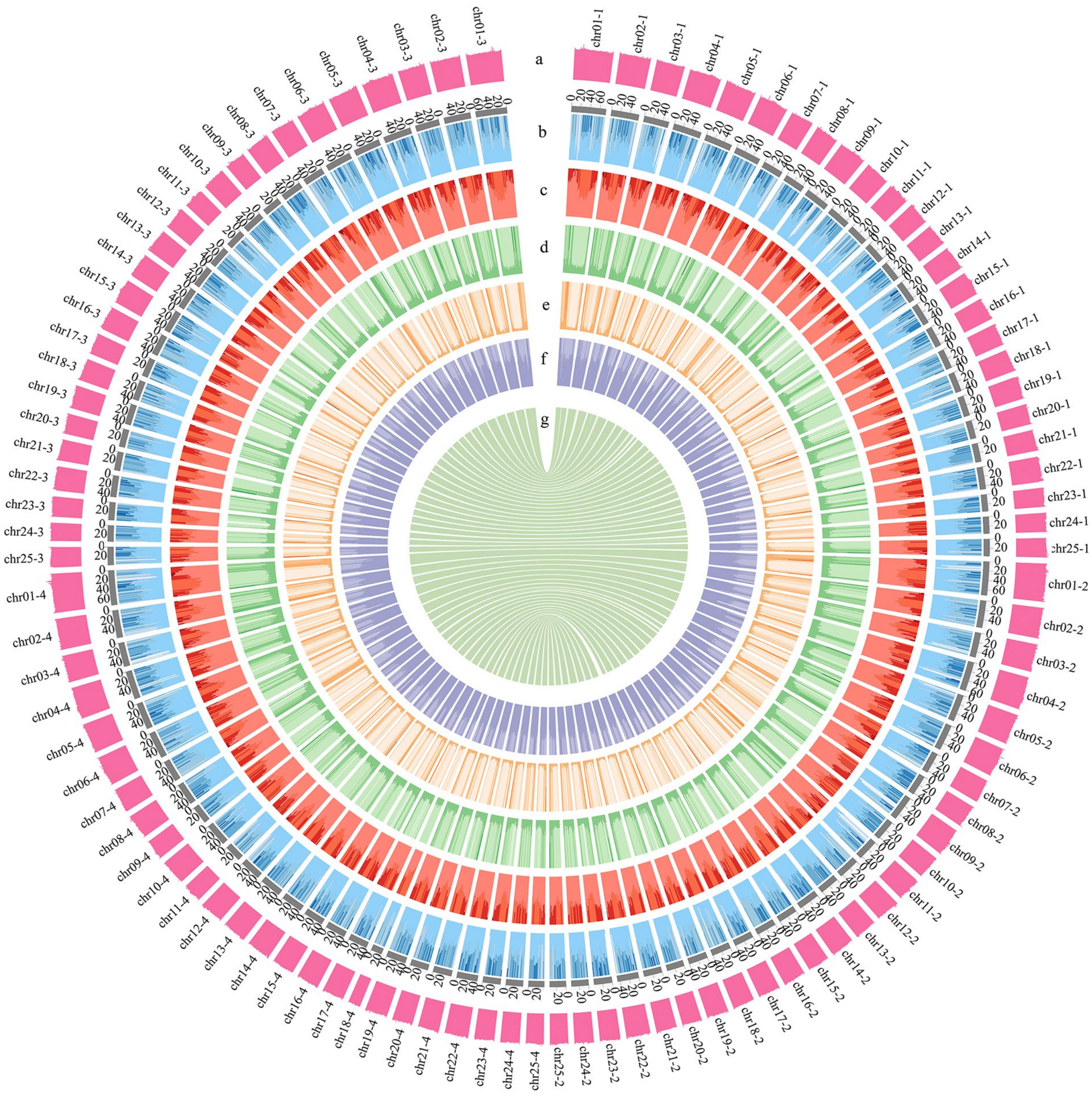

Fig. 3 Circos plot of the autotetraploid *Misgurnus anguillicaudatus* genome assembly. The circos plot showing the 100 chromosomes in autotetraploid *M. anguillicaudatus* genome, h1-4 means the four haplotype genomes. From the outer to the inner layers: GC content, gene density, repeat coverage, LTR (long terminal repeats), LINE (long interspersed nuclear elements), DNA-TE (DNA-transposable elements) and genome collinearity.

between four haplotypes of autotetraploid loach and the diploid loach were calculated using ParaAT2.0 and KaKs_calculator 3.0 (default parameters), revealed a peak value of 0.03 for all four haplotypes[49,50]. Consistency in haplotype divergence was observed (Fig. s9).

Haplotype collinearity analysis. Jcvi (v 1.0.6, parameters: jcvi.compara.catalog ortholog-dbtype = prot-cscore 0.99; jcvi.compara.synteny screen-minspan = 30) was used for collinearity analysis and visualization[46] (Fig. 5). All haplotype genomes showed high collinearity, which indicated our successful haplotype phasing and a high-quality chromosomal genome assembly. To detect the structural variations among the four haplotype genomes, whole-genome alignments was performed using MUMmer (v 4.0.0beta2)[51]. All alignment results between two allelic chromosome pairs were obtained using Nucmer (parameters: -c 500, -b 500, and -l 100) with the -maxmatch option[51]. Then, the alignment results were filtered using delta-filter (parameters: delta-filter -m -i 90 -l 100) and the results were converted into a tab-separated file using the show-coords (parameters: show-coords -THrd) subprogram. Finally, SyRI (v 1.6) was used to identify structural variations among the four haplotype genomes (Fig. s10)[52].

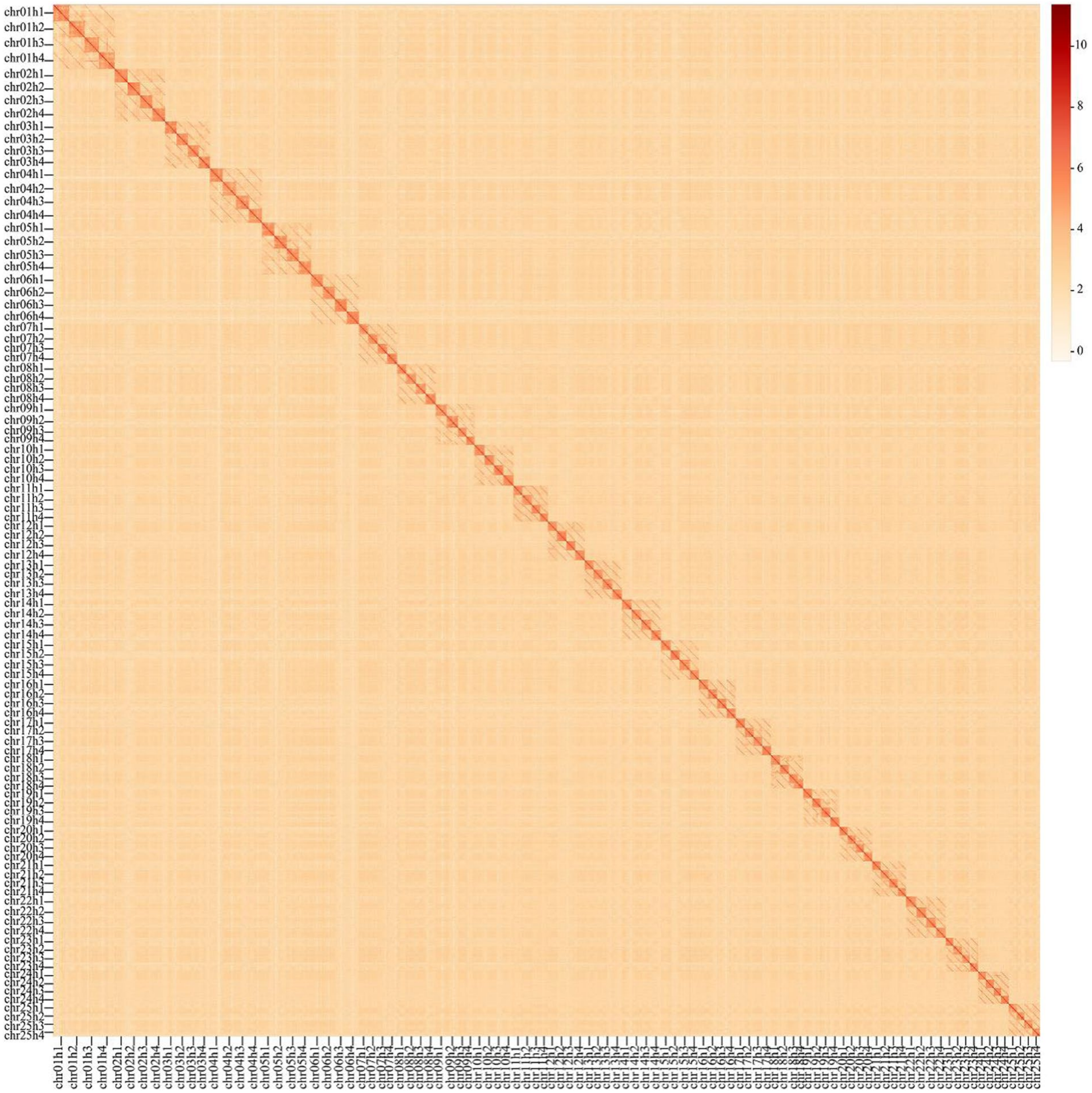

Fig. 4 Hi-C intra-chromosomal contact map of the autotetraploid *Misgurnus anguillicaudatus* genome assembly.

Haplotype expression analysis. To identify collinear block gene pairs between allelic chromosomes, Jcvi (v 1.0.6, parameters: jcvi.compara.catalog ortholog-dbtype = prot-cscore 0.99; jcvi.compara.synteny screen-minspan = 30) was employed[46]. Manual curation was performed to eliminate collinear blocks that likely originated from whole genome duplication (WGD). Consequently, a total of 4,371 alleles were identified across all four haplotype genomes. For allelic expression analysis, the raw data of the brain, muscle, liver, and ovary tissues from female autotetraploid loach were downloaded from NCBI database (SRP accession number SRP293717, Table s8)[53,54]. RNA-seq clean reads were aligned to the four haplotype genomes separately using STAR (v 2.7.8a, default parameters)[55]. The normalized TPM values of each sample were estimated with RSEM (v 1.3.3, default parameters)[56]. Notably, our analysis revealed that no dominant haplotype was found among the four haplotype genomes. (Fig. 6, Figs. s11–s14).

Data Records

All raw data of the whole genome and assembled genome reported in this paper have been deposited in the Genome Warehouse in National Genomics Data Center[57,58], Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformation, under BioProject accession number PRJCA024750[59]. The genomic PacBio sequencing data accession number is CRR1184674[60], the Hi-C sequencing data accession number is CRR1184675[60]. The assembled genome accession number are GWHERQI00000000 (haplotype 1)[61], GWHERQL00000000(haplotype 2)[62], GWHERQH00000000 (haplotype 3)[63], GWHERQG00000000

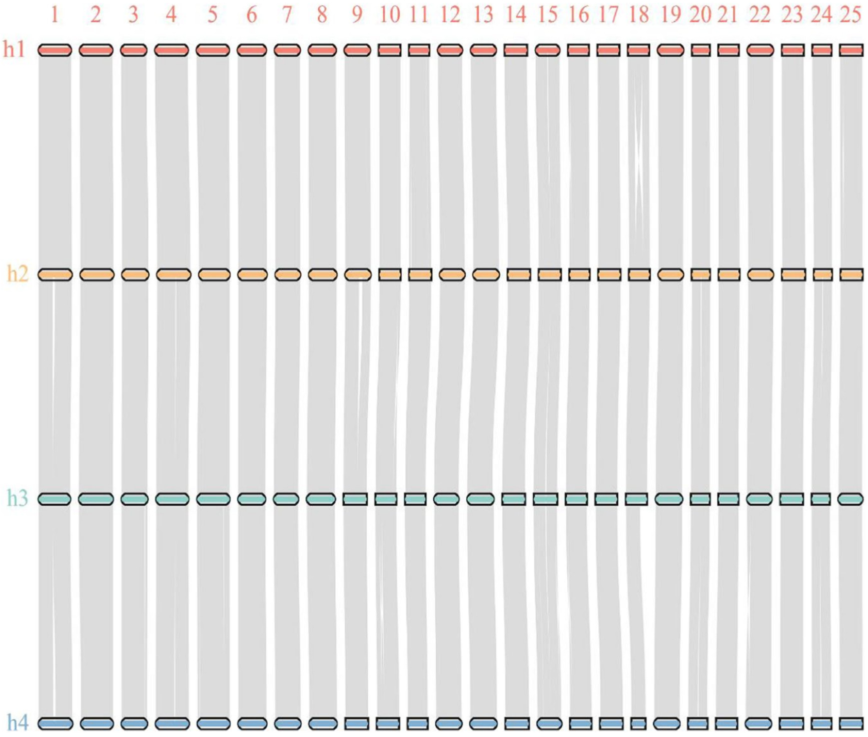

Fig. 5 Haplotype collinearity analysis. h1-h4 representing the four haplotype genomes.

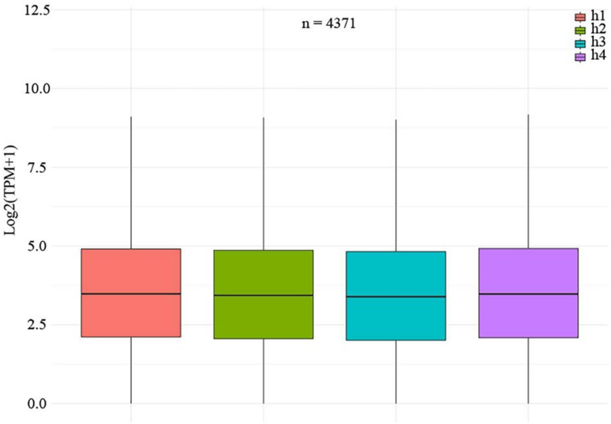

Fig. 6 Haplotype expression analysis. h1-h4 representing the four haplotype genomes. N = 4371 means all allelic genes.

(haplotype 4)[64], that are publicly accessible at Genome Warehouse. The assembled genome was also deposited in the NCBI Genome with the accession number JBGGOFO00000000 (haplotype 1)[65], JBGGOG000000000 (haplotype 2)[66], JBGGOHO00000000 (haplotype 3)[67], JBGGOI000000000 (haplotype 4)[68]. The assembled genome can also be obtained from the figshare dataset at https://doi.org/10.6084/m9.figshare.26340437.v1[69]. RNA sequencing data for transcriptome-based gene prediction were deposited in the SRA at NCBI SRR29303864[70].

Technical Validation

RNA integrity. In this study, nine tissues (including muscle, liver, fin, skin, eye, brain, gill, intestine, and ovary) were sampled from the autotetraploid *M. anguillicaudatus* and mixed for transcriptome sequencing. For constructing RNA-Seq libraries, the purity of RNA was analyzed by using NanoPhotometer Spectrophotometer (Implen, USA). Subsequently, the concentration and integrity of RNA were quantified by using Qubit 2.0 Fluorometer (Life Technologies, USA) and Agilent Bioanalyzer 2100 (Agilent Technologies, USA), respectively. In this study, the spectrophotometer ratios (260 nm/280 nm) of all DNA were over 1.8. The quality of all purified RNA was evaluated by absorbance (260 nm/280 nm) >1.7 and the RNA integrity number >8. These high-quality DNA was finally subjected to construct the sequencing library.

Code availability

All analyses were conducted on Linux systems. All softwares used in this work were in the public domain, with parameters being clearly described in Methods. If no detail parameters were mentioned for a software, default parameters were used as suggested by developer.

Received: 9 April 2024; Accepted: 17 September 2024;
Published online: 28 September 2024

References

1. Soltis, P. S., Liu, X., Marchant, D. B., Visger, C. J. & Soltis, D. E. Polyploidy and novelty: Gottlieb's legacy. *Philos Trans R Soc Lond B Biol Sci.* **369**, 20130351 (2014).
2. Otto, S. P. & Whitton, J. Polyploid incidence and evolution. *Annu Rev Genet.* **34**, 401–437 (2000).
3. Zhang, F., Qu, L., Gu, Y., Xu, Z. H. & Xue, H. W. Resequencing and genome-wide association studies of autotetraploid potato. *Mol Hortic.* **2**, 6 (2022).
4. Sun, H. *et al.* Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nat Genet.* **54**, 342–348 (2022).
5. Shen, C. *et al.* The Chromosome-Level Genome Sequence of the Autotetraploid Alfalfa and Resequencing of Core Germplasms Provide Genomic Resources for Alfalfa Research. *Mol Plant.* **13**, 1250–1261 (2020).
6. Mable, B. K. Why polyploidy is rarer in animals than in plants': myths and mechanisms. *Biol J Linn Soc.* **82**, 453–466 (2004).
7. Yu, Y. Y., Li, Y. H., Li, R. W., Wang, W. M. & Zhou, X. Y. Mitochondrial genome of the natural tetraploid loach *Misgurnus anguillicaudatus*. *Mitochondrial DNA.* **25**, 115–6 (2014).
8. Zhong, J., Yi, S., Ma, L. & Wang, W. Evolution and phylogeography analysis of diploid and polyploid *Misgurnus anguillicaudatus* populations across China. *Proc Biol Sci.* **286**, 20190076 (2019).
9. Arai, K. Genetics of the loach, *Misgurnus anguillicaudatus*: recent progress and perspective. *Folia Biol (Krakow).* **51**, 107–17 (2003).
10. Yin, J., Zhao, Z. S., Chen, X. Q., Li, Y. Q. & Zhu, L. Y. Karyotype comparison of diploid and tetraploid loach, *Misgurnus anguillicanudatus*. *Acta Hydrob Sin.* **29**, 469–72 (2005).
11. Peng, Y. *et al.* Chromosome-level genome assembly of the Arctic fox (*Vulpes lagopus*) using PacBio sequencing and Hi-C technology. *Mol Ecol Resour.* **21**, 2093–2108 (2021).
12. Belton, J. M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–76 (2012).
13. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–70 (2011).
14. 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 (2021).
15. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
16. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
17. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
18. Simão, 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).
19. Belton, J. M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
20. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* **9**, 357–359 (2012).
21. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
22. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98 (2016).
23. Dudchenko, O. *et al.* De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
24. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
25. Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics*. Chapter 4, Unit 4.10, (2004).
26. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res.* **110**, 462–467 (2005).
27. Abrusán, G., Grundmann, N., DeMester, L. & Makalowski, W. TEclass–a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329–1330 (2009).
28. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–268 (2007).
29. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol.* **268**, 78–94 (1997).
30. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–467 (2005).
31. Gertz, E. M., Yu, Y. K., Agarwala, R., Schäffer, A. A. & Altschul, S. F. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol.* **4**, 41 (2006).
32. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* **14**, 988–995 (2004).
33. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
34. Ghosh, S. & Chan, C. K. Analysis of RNA-Seq Data Using TopHat and Cufflinks. *Methods Mol Biol.* **1374**, 339–361 (2016).
35. Cantarel, B. L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
36. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (2019).
37. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A. & Bateman, A. J. N. A. R. Rfam: Annotating Non-Coding RNAs in Complete Genomes. **33**, D121–124 (2005).
38. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1640 (2014).
39. Dimmer, E. C. *et al.* The UniProt-GO Annotation database in 2011. *Nucleic Acids Res.* **40**, D565–570 (2012).
40. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
41. Marchler-Bauer, A. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **39**, D225–229 (2011).
42. Rolf, A. *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **32**, D115–119 (2004).
43. Boeckmann, B. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
44. Sun, B. *et al.* The chromosome-level genome and key genes associated with mud-dwelling behavior and adaptations of hypoxia and noxious environments in loach (*Misgurnus anguillicaudatus*). *BMC Biol.* **21**, 18 (2023).
45. Sun, B. *et al.* Loach (*Misgurnus anguillicaudatus*) genome data. *GenBank* https://ncbi.nlm.nih.gov/datasets/genome/GCF_027580225.1/ (2023).
46. Tang, H. *et al.* Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).

47. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
48. Rabiee, M., Sayyari, E. & Mirarab, S. Multi-allele species reconstruction using ASTRAL. *Mol Phylogenet Evol.* **130**, 286–296 (2019).
49. Zhang, Z. KaKs_Calculator 3.0: Calculating Selective Pressure on Coding and Non-coding Sequences. *Genomics Proteomics Bioinform.* **20**, 536–540 (2022).
50. Zhang, Z. *et al*. ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochem Biophys Res Commun.* **419**, 779–81 (2012).
51. Marçais, G. *et al*. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol.* **14**, e1005944 (2018).
52. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
53. Luo, L.F. *et al*. Tissues sequencing data of female autotetraploid loach. *GenBank* https://trace.ncbi.nlm.nih.gov/Traces/?view=study&acc=SRP293717.
54. Luo, L. F. *et al*. Comparative transcriptome analysis revealed genes involved in sexual and polyploid growth dimorphisms in loach (*Misgurnus anguillicaudatus*). *Biology (Basel)* **10**, 935 (2021).
55. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
56. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
57. Chen, M. *et al*. Genome Warehouse: A Public Repository Housing Genome-scale Data. *Genomics Proteomics Bioinformatics* **4**, 584–589 (2021).
58. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. *Nucleic Acids Res*, **52**, D18–D32 (2024).
59. Li, Q. *et al*. Chromosome-scale and haplotype-resolved genome assembly of the autotetraploid *Misgurnus anguillicaudatus*. *Genome Warehouse in National Genomics Data Center* https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA024750 (2024).
60. *The raw reads data for PacBio and Hi-C sequencing of autotetraploid Misgurnus anguillicaudatus archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA016966 (2024).
61. *Hap1 genome accession number* https://ngdc.cncb.ac.cn/gwh/Assembly/GWHERQI00000000 (2024).
62. *Hap2 genome accession number* https://ngdc.cncb.ac.cn/gwh/Assembly/GWHERQL00000000 (2024).
63. *Hap3 genome accession number* https://ngdc.cncb.ac.cn/gwh/Assembly/GWHERQH00000000 (2024).
64. *Hap4 genome accession number* https://ngdc.cncb.ac.cn/gwh/Assembly/GWHERQG00000000 (2024).
65. Sun, B. & Li, Q. *NCBI GenBank (Haplotype 1)* https://identifiers.org/ncbi/insdc:JBGGOF000000000 (2024).
66. Sun, B. & Li, Q. *NCBI GenBank (Haplotype 2)* https://identifiers.org/ncbi/insdc:JBGGOG000000000 (2024).
67. Sun, B. & Li, Q. *NCBI GenBank (Haplotype 3)* https://identifiers.org/ncbi/insdc:JBGGOH000000000 (2024).
68. Sun, B. & Li, Q. *NCBI GenBank (Haplotype 4)* https://identifiers.org/ncbi/insdc:JBGGOI000000000 (2024).
69. Li, Q. *et al*. Whole genome assembly of autotetraploid *Misgurnus anguillicaudatus*. *figshare. Dataset.* https://doi.org/10.6084/m9.figshare.26340437.v1 (2024).
70. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR29303864 (2024).

Acknowledgements

The authors would like to thank Wuhan Onemore-tech Co., Ltd. for their assistance with genome sequencing and analysis. The authors would like to thank the National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University for providing the computing platform. Most bioinformatic analyses and computations in this paper were run on the platform. This study was supported by National Natural Science Foundation of China (U23A20245 and 32172962), and Key R & D Program Project in Hubei Province of China (2023BBB054).

Author contributions

X.J.C. and J.G. designed the research; B.S., Q.S.L., Y.H.M. and Y.B.Z. performed the research; B.S., Q.S.L., Y.H.M., Y.B.Z., Y.X.Z., X.X.X. and J.W.Z. analyzed the data; B.S. and Y.W.H. wrote the paper; X.J.C. and J.G. obtained the research funding; X.J.C. and J.G. involved in the discussion; X.J.C. and J.G. revised the paper. All authors have reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-03891-z.

Correspondence and requests for materials should be addressed to G.J. or X.C.

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.