scientific data

OPEN

DATA DESCRIPTOR

Near telomere-to-telomere diploid genome assembly of *Acrossocheilus wenchowensis*

Lingzhan Xue[1,4], Mingkun Luo[2,4], Haoyu Wang[3,4], Wenbin Zhu[2], Duhuang Chen[1], Gaoxiong Zeng[1], Mengxiang Liao[1], Ji Zhao[1], Bin Wu[1] ✉, Luohao Xu[3] ✉ & Zaijie Dong[2] ✉

Acrossocheilus wenchowensis **is a lukewarm-water fish found in southern Chinese mountain streams, valued for both ornamental and edible purposes. We assembled a near telomere-to-telomere (T2T) genome using HiFi, ONT, Hi-C and Illumina data. The assembly is approximately 870.69 Mb with a contig N50 of about 21.28 Mb. Among these, 14 chromosomes in Hap1 and 15 chromosomes in Hap2 have reached T2T levels. A total of 24,909 protein-coding genes were predicted in Hap1 and 24,496 in Hap2, with BUSCO scores of 97.4% and 97.6%, respectively. A conserved centromeric satellite sequence (262 bp) derived from an LTR transposon was identified. Comparative genomics showed that *Acrossocheilus* and *Onychostoma* diverged approximately 13.7 million years ago (Mya), while *A. wenchowensis* diverged from *A. fasciatus* about 5.25 Mya. Resequencing of four geographic populations of *A. wenchowensis* revealed distinct genetic structure in the LY group compared to the other populations based on SNP and InDel analysis. This genome provides a framework for diploid T2T studies in fish and supports further functional genomics research.**

Background & Summary

Acrossocheilus wenchowensis, commonly known as the freshwater grouper, belongs to the order Cypriniformes, family Cyprinidae, genus Barbinae, is a lukewarm-water fish that lives in mountain streams and rivers south of the Yangtze River in China[1]. It is popular with many consumers because its meat is tender and tasty, and the muscle contains a variety of polyunsaturated fatty acids (PUFA) as well as many amino acids that are beneficial to humans[2]. Meantime, there was also an obvious sexual dimorphism in growth phenomenon between the two sexes, i.e. the growth rate of the females was significantly faster than that of the males[3]. This phenomenon has important economic implications for aquaculture, as rearing single-sex populations is an effective strategy for increasing yields, reducing farming costs, improving feed conversion rates, and achieving uniform size consistency[4]. High-quality whole-genome reference sequences, particularly those with T2T (telomere-to-telomere) gap-free coverage, are indispensable foundational tools for research into sex determination mechanisms[5]. There are about 26 species in the genus *Acrossocheilus*, mainly found in Laos, Vietnam and China, with different populations exhibiting diverse morphological characteristics and ecological habits, providing a valuable model for studying the origin and geographical distribution of freshwater fishes[6]. In addition, previous research has examined rearing environments such as temperature and heavy metals, embryonic development, population structure and the mitochondrial genome was carried out[7–9], and the genomes of *A. fasciatus* has been assembled[10,11]. However, no high-quality, chromosome-level T2T reference genome for *A. wenchowensis* has been reported.

With the development of sequencing technology, an increasing number of high-quality genomes have been reported, especially T2T genomes with even higher quality[12]. For example, the assembly of the human T2T genome fills gaps in previously unknown genes and duplicate variants, expanding the critical repository of human genomic sequences and revealing the impact of specific gene duplications on the genome[5]. Completion

[1]Freshwater Fisheries Research Institute of Fujian Province, Fuzhou, Fujian, China. [2]Key Laboratory of Freshwater Fisheries and Germplasm Resources Utilization, Freshwater Fisheries Research Center of Chinese Academy of Fishery Sciences,, Ministry of Agriculture and Rural Affairs, Wuxi, Jiangsu, China. [3]Integrative Science Center of Germplasm Creation in Western China (Chongqing) Science City, MOE Key Laboratory of Freshwater Fish Reproduction and Development, School of Life Sciences, Southwest University, Chongqing, 400715, China. [4]These authors contributed equally: Lingzhan Xue, Mingkun Luo, Haoyu Wang. ✉e-mail: wubinfire@126.com; luohaox@gmail.com; dongzj@ffrc.cn

of the T2T genome of sheep revealed that a 168 bp insertion in the 5'UTR of the *hoxb*13 gene is associated with the long-tail trait[13]. The publication of the chicken T2T genome fills the previous research gap regarding the dot chromosome and the epigenetic landscape of sex chromosomes[14]. However, current research on T2T genomes in animals has primarily focused on mammals and poultry, while studies involving fish remain significantly behind. The T2T Y chromosome of the zig-zag eel (*Mastacembelus armatus*) represents a significant advancement in the T2T genome of fish[15]. Nevertheless, further efforts are needed to improve the T2T genomes of fish species. This will provide a genetic foundation for the analysis of biological phenotypes.

In species classification and identification, researchers have traditionally relied on morphological characteristics such as body coloration, patterns, and countable or measurable traits for differentiation. However, the limitations of this approach have become increasingly apparent. Genetic exchange between species, convergent evolution, and gene mutations associated with specific traits often result in transitional or highly similar problematic phenotypes, leading to uncertainty in identification[16]. Increasing research also indicates that these variable morphological traits alone are insufficient to resolve complex issues, including the delineation of closely related species, clarification of phylogenetic relationships, and understanding of population evolutionary histories[17]. In this study, we present the first haplotype-resolved T2T genome assemblies of *A. wenchowensis*, with each haplotype anchored to 25 pseudo-chromosomes. Hap1 spans approximately 856.99 Mb with a contig N50 of 31.12 Mb, while Hap2 spans approximately 868.36 Mb with a contig N50 of 31.07 Mb. 14 chromosomes in Hap1 and 15 chromosomes in Hap2 achieved T2T resolution, with BUSCO completeness rates of 97.4% and 97.6%, respectively. We also conducted the first resequencing of *A. wenchowensis* across four distinct geographic populations, integrating whole-genome genetic variation and diversity information. Comparative genomic analysis subsequently revealed the origin history of *A. wenchowensis*. The constructed genome information, approaching T2T level, provides valuable foundational data for subsequent molecular studies of this species.

Methods

Sample collection, library construction and sequencing. The use and handling of *A. wenchowensis* complied with animal welfare laws, guidelines and policies approved by the Scientific Ethics Committee of the Freshwater Fisheries Research Institute of Fujian, Fuzhou, China. Healthy adult male *A. wenchowensis* was anesthetized with 15~20 mg/L MS-222 buffer at pH 7.0-7.5. Muscle, intestine, blood, gills, heart, brain, eye, skin, spleen, kidney, caudal fin, testis and liver were sampled, immediately frozen in liquid nitrogen, and stored at −80 °C.

High molecular weight genomic DNA (gDNA) was isolated from sufficient muscle tissue for Illumina, ONT ultra-long and HiFi sequencing using a standard sodium dodecyl sulfate extraction method[18], while blood was used for Hi-C sequencing using a Blood & Cell Culture DNA Kit (Qiagen 13323). DNA quality is measured using three methods: (1) NanoDrop 2000 spectrophotometry (Thermo Fisher Scientific, USA) for concentration and purity (OD 260/280 ratio); (2) pulsed-field electrophoresis to analyze the degree of DNA degradation and the presence of RNA contamination; and (3) Qubit fluorometer (Invitrogen, USA) for precise quantification of concentration. Additionally, we selected 12 tissue samples preserved in liquid nitrogen for RNA sequencing. Total RNA was extracted using TRIzol™ reagent (Thermo Fisher Scientific, USA). The extracted RNA was treated with DNase I (NEB, USA) to remove genomic DNA. RNA quality was assessed by 1.5% agarose gel electrophoresis and with NanoDrop 2000 spectrophotometer.

A paired-end genomic library with an insert size of 350 bp was generated using gDNA according to the manufacturer's instructions using the NEB Next® Ultra™ DNA Library Prep Kit (NEB, USA), and then sequenced on the Illumina NovaSeq6000 platform. For HiFi sequencing, a standard SMRTbell library was constructed using the SMRTbell Express Template Prep Kit 2.0 following the manufacturer's guidelines and sequenced on a PacBio Sequel II platform (Pacific Biosciences, CA, USA). For ONT sequencing, a library was prepared using the Oxford Nanopore SQK-ULK001 kit according to the instructions and sequenced on a PromethION flow cell. For Hi-C library construction, we followed the standard protocol using the Illumina HiSeq2500 platform. Additionally, the RNA-seq library was constructed using the Illumina standard protocol (San Diego, CA, United States) and sequenced on the Illumina HiSeq6000 platform, and analyzed by the Fastp v0.20 software[19].

Genome survey, gap-free genome assembly and quality assessment. The generated adapter sequences and low-quality reads were filtered using Fastp v0.20 software (Fig. s1)[20], and the genome size of *A. wenchowensis* was estimated using the *k*-mer method[21]. The *k*-mer distribution was estimated using jellyfish[22] with the parameters -m 17-C. The heterozygosity ratio was estimated using GenomeScope[23]. The preliminary genomic estimated a genome size of 863.32 Mb, with a heterozygosity of 0.55% and a repeat rate of 55.95% (Fig. s2A)[20].

The PacBio SMRT-analysis package (https://www.pacb.com) was used for quality control of the raw reads, and the circular consensus sequence (CCS) reads were generated by the SMRTLink v11.0 (https://www.pacb.com/support/software-downloads/) with the parameters–min-passes = 3--min-rq = 0.99. To filter Hi-C reads, index reads were removed using the following SOAPNUKE v2.0[24] parameters: (1) reads with 3nt unidentified nucleotides (N); (2) reads aligned at the adaptor; (3) reads with ≥ 20% bases with Phred quality ≤ 5. The Hifiasm software[25] was used to create the assembly design with default parameters. The draft genome was trimmed by purge_dups software[26] for purging haplotigs and overlaps in draft assembly based on read depth. Subsequently, the remaining contigs were aligned to the NT database by BLAST[27] to remove the contaminated sequences. The ONT reads were used for whole-genome assembly with NextDenovo[28], and the sequences of the original ONT assembly were polished four times with HiFi data using Racon[29]. Furthermore, the filtered Hi-C reads were aligned to the initial genome using HiC-Pro[30]. Valid paired-end reads were used to construct the pseudochromosome by 3D DNA[31]. Juicebox[32] was used to manually correct the pseudochromosome based on the Hi-C

Library type	Tissue	Platform	Data size (Gb)	Average Length (bp)	N50 Read Length (bp)	Sequencing Coverage (×)
Illumina	Muscle	Illumina NovaSeq6000	67.2	150	2 × 150	77.8
PacBio SMRT	Muscle	Pacbio Sequel II	42.18	16,318	16,276	47
ONT ultra-long	Muscle	PromethION	53.63	94,720	100,000	63
Hi-C	Blood	Illumina HiSeq2500	143.09	150	2 × 150	159

Table 1. Statistics of the sequencing data used for A. wenchowensis genome assembly.

contact map. Finally, the contact map was drawn using HiCExplorer[33]. Each have initial haplotype assembled sequences were anchored onto 25 pseudo-chromosomes after anchoring.

The contigs assembled by NextDenovo were aligned to the contigs from Hifiasm with the help of Winnowmap2[34]. The ONT contigs that were close to or spanned the gaps were selected to fill the gaps in the chromosome genome using MUMmer[35]. Appropriate telomere sequences (TTAGGGG) were selected according to species characteristics for telomere analysis on the genome. For chromosomes lacking telomeres or with excessively short telomere length, HiFi reads was used to expand them, and then telomere analysis was conducted on the genome after completion. The pyTanFinder[36] screens candidate centromere monomer based on following criteria: (1) the centromere monomer shows a unimodal distribution on each chromosome; (2) reduced gene abundance in the centromere region, with increased abundance of transportable elements and tandem repeats; (3) weaker Hi-C interaction signals in the candidate centromere regions. StainedGlass[37] is used to visualize tandem replication sequences in the genome.

Finally, the Illumina short reads were obtained for genome size estimation, HiFi reads for initial genome assembly, and ONT and Hi-C reads for near-complete T2T genome assembly (Table 1) and (Table s1)[20]. Notably, the initial HiFi assembly produced a genome draft of 895.74 Mb, with a GC content of ~37.57% (Table s2)[20]. We also assessed the assembly quality using data match rates, GC base depth and BUSCO completeness (Figs. s2A-2B and Table s3)[20].

We then performed haplotype assembly and found that Hap1 was ~867.57 Mb in size with a contig N50 of 24.42 Mb, while Hap2 was ~882.22 Mb with a contig N50 of 26.06 Mb (Table s4)[20]. Based on the collinear with Hap1, we renamed the composite Hap2 chromosome and obtained a detailed sequence distribution (Fig. 1A) and (Table s5)[20], and heatmaps of the interactions further confirm the accuracy of the assembly results (Figs. s2C-2D)[20]. We also assembled the ONT reads and found that the genome size was ~870.69 Mb and contig N50 was ~21.28 Mb (Table s6)[20]. ONT data further filled gaps in the haplotype genomes, with 14 chromosomes in Hap1 and 15 chromosomes in Hap2 reaching T2T-level assembly (Table 2). Of the two haplotype genomes, 44 chromosomes had at least one end assembled to telomeric structure (Fig. s3 and Table s7)[20]. Furthermore, we analyzed the synteny between the two T2T haplotype genomes and their respective synteny with two published genomes[38–40] of *A. fasciatus* (Fig. 1B) and (Fig. s4)[20].

Furthermore, we identified a 262 bp centromere monomer with a unimodal distribution across all chromosomes in *A. wenchowensis* (Cen262, GAAGCTCTAACATGACTTGCTCTGC AAAA-AAGCTCTAACACATGAAAATCTGTCTTCTGCATTGCAAATAACATTTCATTTCATGAG AAAGAAACATCTTTTATGTACTTGAGCTTCACTCTTAAAATTGCATTTAGCAACTTTCT CTCACTGAGAAAGAAGCTTTTCAGCATCCGTGTCCTTGCATTCTCTTCATTTGCTTAG TTGCACTAAACACAGCATTAGTGTTCTTAGAAGCTCTAAACTAGTGATTCTGTGCATT). In the centromere regions, we observed lower gene density, higher densities of transposable elements and tandem repeats, and reduced Hi-C interaction information. Notably, our findings revealed that among all transposable elements, the LTR/Gypsy element consistently co-localized with the centromere, while the LINE/Rex-babar element was predominantly enriched at the periphery of the centromere (Fig. 1C) and (Fig. s5)[20].

Genome annotation. A combination of *de novo* and homology-based methods was used for the prediction of repeat elements. For *de novo* annotation, MITE-Hunter[41] was used to identify mini-inverted repeat transposable elements (MITEs), which are widely distributed in the genome and belong to type II transposable elements. For homolog detection, RepeatMasker[42] was used to search the genome sequence for the sequence similar to the known repetitive sequence in the RepBase database (http://www.girinst.org/repbase). Gene structures were subsequently annotated using AUGUSTUS[43], SNAP[44], GlimmerHMM[45] and GeneMark-ET[46] software through *de novo* prediction, protein homology searches and RNA sequencing analysis. GeMoMa[47] was used for homology prediction, and exon and intron information were determined by comparing the transcript with the genome. For RNA-Seq, HISAT2[48] was used to align reads to the genome sequence, and Cufflinks[49] was used to assemble transcripts to obtain full-length transcript sequences. SMRTLink (https://www.pacb.com/support/software-downloads/) was used to generate high-quality full-length transcripts. Open reading frame (ORF) were predicted using PASA[50], and the results were integrated using EvidenceModler[51].

By aligning with three protein sequence databases: NR (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz), SwissProt (https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz) and eggNOG (http://eggnog5.embl.de/#/app/home), all protein-coding genes were identified. Protein domains were annotated using InterPro, and Gene Ontology terms for each gene were obtained from the corresponding eggnog-mapper annotation entry. Pathways in which the genes might be involved were assigned by blast against the KEGG databases (https://www.genome.jp/kegg/brite.html). tRNAscan-SE[52] was used to predict tRNA in the genome sequence. The Rfam (https://rfam.org/) was used to annotate other types of ncRNA using BLAST[27]. 47.02% and 47.56% of the repetitive sequences were detected in Hap1 and Hap2, respectively

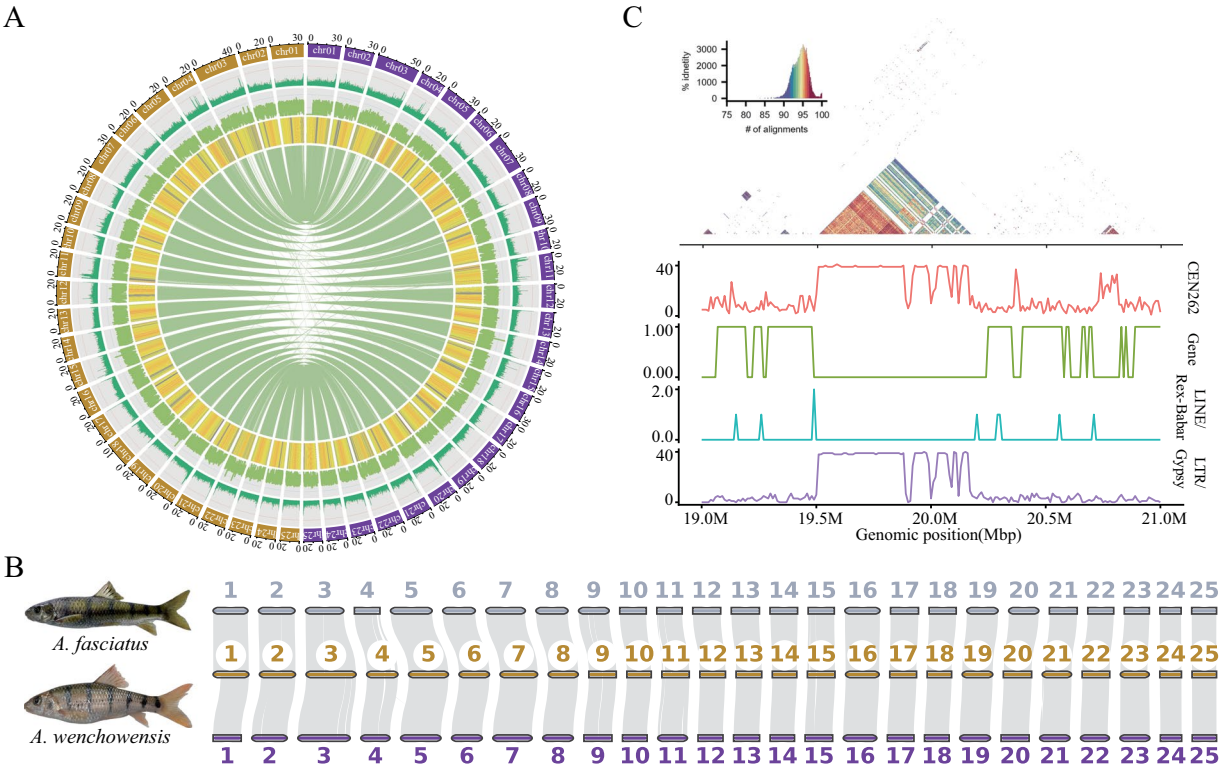

Fig. 1 Genomic characterization of *A. wenchowensis*. (**A**) Circos plot of two haplotype genomes of *A. wenchowensis*. Showed GC content, gene density, TE density, and collinearity of haplotype from outside to inside, respectively. (**B**) Colinear map of two haplotype genomes of *A. wenchowensis* and *A. fasciatus*. (**C**) The centromere characteristics on chr01. Showed TE heat map, CEN262 density; Gene density; LINE/Rex-Babar density; LTR/Gypsy density from top to bottom.

(Table s8)[20]. We identified extensive inversion regions on chromosomes 21 and 23, translocation regions on chromosome 11, and several duplication regions on chromosomes 2, 16 and 20 (Fig. s6)[20]. In Hap1 and Hap2, a total of 24,909 and 24,496 protein-coding genes were identified, respectively (Table 3). Using the NR, SwissProt, KEGG, GO and eggNOG databases, 98.39% (24,508 genes) of the Hap1 and 98.47% (24,120 genes) of the Hap2 genome were annotated (Fig. s7 and Table s9)[20]. We also counted the predicted non-coding RNAs and found 772 miRNAs in Hap1 and 615 miRNAs in Hap2 (Table s10)[20].

Structural variants identification between haploids. Whole-genome alignments of haplotype assemblies were performed using the nucmer module in MUMmer with optimized parameters (-maxmatch -c 100 -b 500 -l 50)[35]. The results were then filtered with delta-filter (-m -i 90 -l 100) to retain high-quality mappings. Further detection of structural variations such as single nucleotide polymorphisms (SNPs), deletions and insertions (INDELs), was performed using SyRI v1.7.0[53]. Based on fragment size, we define small-scale variations (<50 bp) as INDELs and large-scale variations (≥50 bp) as presence-absence variations (PAVs). All identified structural variants (SVs) are functionally annotated using SnpEff v5.1[54] and categorized into four impact levels:

1. High-impact variants: Cause critical disruptions, including premature stop codons, translation initiation failure, frameshift mutations and large-scale duplications.
2. Moderate-impact variants: Affect protein structure but do not result in complete loss of function.
3. Low-impact variants: Induce non-disruptive sequence alterations.
4. Modifier variants: Alterations in non-coding regions.

A total of 3,854,977 SNPs, 749,189 INDELs, 18,926 PAVs, 3,773 Copy Number Variations (CNVs), 49 Inversions (INVs), and 1,090 Translocations (TRANSs) were identified. Approximately 56% of SVs overlapped with transposable elements, with nearly all INVs and TRANSs coinciding with these elements (Fig. 2A). Among variants overlapping gene structures, the majority (82%) were located within the gene body, while a smaller proportion occurred in promoter regions (Fig. 2B). Compared to other variant types, INVs showed a higher proportion in promoter regions (Fig. 2C). Nearly all INDELs variants overlapping the gene body resulted in non-synonymous mutations (Fig. 2D).

Hap1			Hap2		
Chromosome	Gap Num	Chromosome Len (bp)	Chromosome	Gap Num	Chromosome Len (bp)
Chr01	0	42,706,579	Chr01	0	42,874,082
Chr02	0	35,650,767	Chr02	0	38,009,898
Chr03	1	55,174,287	Chr03	0	55,100,322
Chr04	1	36,429,557	Chr04	1	33,314,700
Chr05	0	44,688,630	Chr05	0	44,177,792
Chr06	0	33,801,162	Chr06	1	34,234,028
Chr07	0	47,278,654	Chr07	0	47,713,612
Chr08	0	30,787,172	Chr08	1	32,110,405
Chr09	1	35,494,943	Chr09	1	36,934,699
Chr10	2	26,646,443	Chr10	0	26,349,852
Chr11	1	32,559,851	Chr11	1	34,752,183
Chr12	0	31,208,481	Chr12	0	31,072,192
Chr13	0	31,484,802	Chr13	0	32,696,485
Chr14	0	31,121,667	Chr14	0	31,409,711
Chr15	1	30,708,846	Chr15	0	30,371,960
Chr16	0	34,938,419	Chr16	0	36,061,702
Chr17	0	32,228,176	Chr17	0	32,980,074
Chr18	0	33,620,301	Chr18	0	34,178,730
Chr19	1	32,222,154	Chr19	1	32,442,124
Chr20	1	33,489,200	Chr20	1	33,596,834
Chr21	1	32,259,612	Chr21	2	33,178,922
Chr22	2	31,782,840	Chr22	1	33,986,604
Chr23	1	29,092,317	Chr23	0	29,877,091
Chr24	0	26,554,631	Chr24	0	26,535,539
Chr25	0	25,061,622	Chr25	1	24,403,828

Table 2. Distribution statistics of each chromosome sequence after gap filling.

Statistics	Number & Length	
	Hap1	Hap2
Number of genes	24,909	24,496
Total genic length	474,756,628	469,144,218
Mean gene length	19,509	19,151
Number of transcripts	27,210	26,797
Total transcript length	79,503,190	79,382,667
Mean transcript length	2921	2962
Number of exons	297,960	295,680
Mean exon length	266	268
Number of coding exons	290,550	288,290
Number of introns	270,750	268,883
Mean intron length	1674	1667
Total CDS length	51,616,005	50,977,005
Mean CDS length	1896	1902

Table 3. Statistics of the results of genome structure prediction.

Data Records

All relevant DNA and RNA sequencing raw data have been submitted to the NCBI Sequence Read Archive (SRA) database, with reference information at https://identifiers.org/ncbi/insdc.sra:SRP510699[55], which is associated with the BioProject accession number PRJNA1117189. The genome assembly has been deposited at GenBank under the accession JBTMPJ000000000[56] and JBTMPK000000000[57]. The genome annotation data have been deposited in the *Figshare* repository (https://doi.org/10.6084/m9.figshare.25911631.v1)[58]. Table s1 to s14 and Figures s1 to s9 referenced in this manuscript have been uploaded to the *Figshare* database (https://doi.org/10.6084/m9.figshare.29400050.v3)[20]. Additionally, the raw SNP and INDEL data used for resequencing analysis are deposited in the Zenodo database (https://doi.org/10.5281/zenodo.17861395)[59].

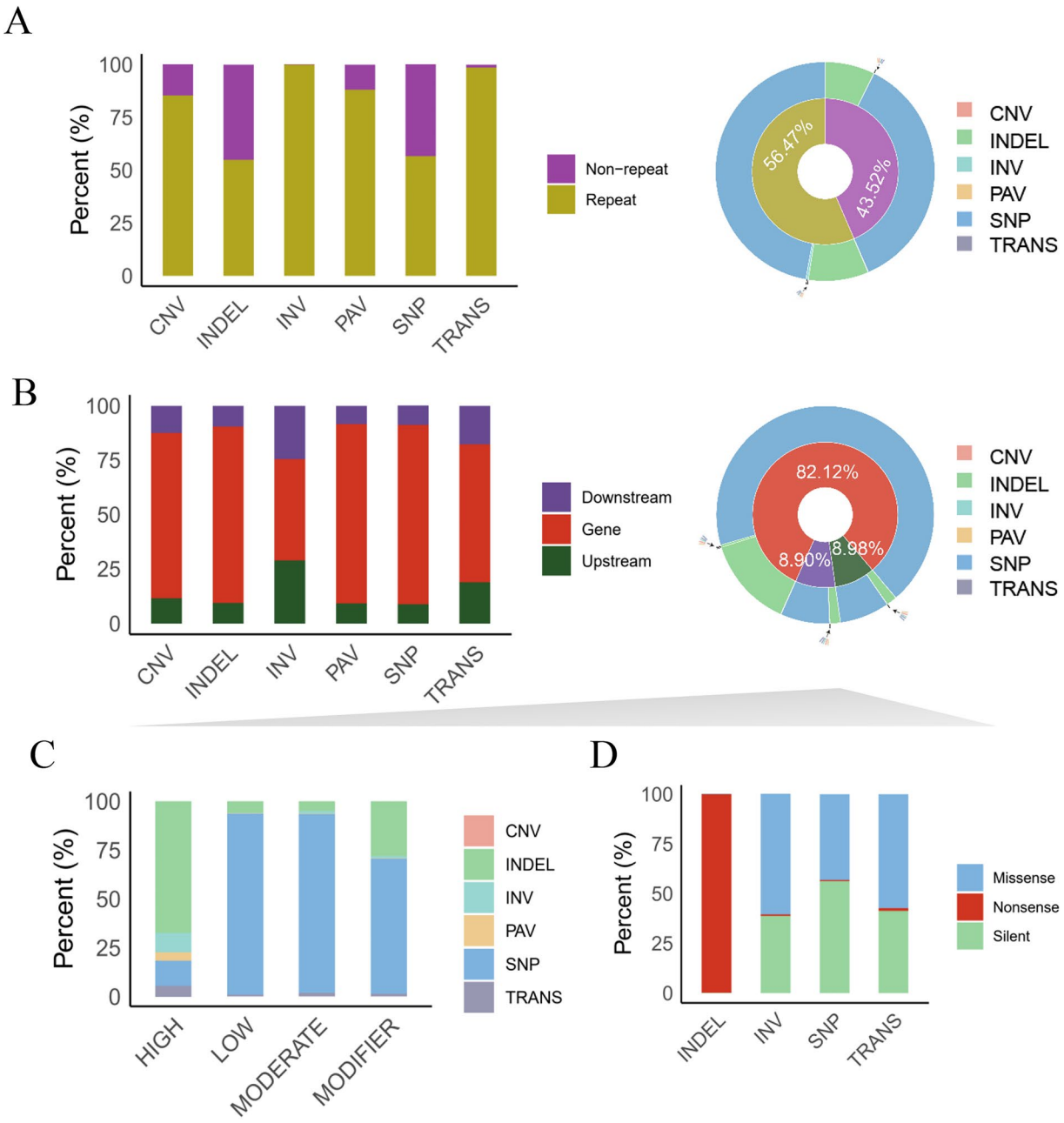

Fig. 2 Structural variation between two haplotype genomes of *A. wenchowensis*. (**A**) The intersection of various structural variation types and repeats. Stacked bar plots represent the proportion of repeat in each structural variation. The circle chart shows the proportion of repeat and various structural variations from the outside to the inside. (**B**) The intersection of various structural variation types and gene structure. Stacked bar plots represent the proportion of repeat in each structural variation. The circle chart shows the proportion of gene structure and various structural variations from the outside to the inside. (**C**) The proportion of various structural variations in four different degrees of gene function. (**D**) Four major structural variants affect gene function: the proportion of Missense, Nonsense, and Silent in INDEL, INV, TRANS, and SNP.

Technical Validation

The quality of the *A. wenchowensis* genome assembly was evaluated to check its continuity, completeness and accuracy. For continuity evaluation, the HiFi and ONT reads were aligned to the final genome assembly using the minimap2[60]. Based on the alignment results, SAMtools[61] and BEDtools[62] were used to calculate the depth and GC content in a 10Kb window. The mean depths were 47× and 63×, and the GC contents were 37.5% and 37.4%. Completeness was assessed using conserved genes from the BUSCO[63] database, and the complete BUSCOs were found to be 97.4% and 97.6%, including 93.9% single-copy genes for Hap1 and 96.2% single-copy genes for Hap2 (Table s14)[20]. Additionally, we compared the two newly assembled haplotype genomes with previously published genomes[38,39] and found a significant improvement in assembly quality including Contig number, Contig N90 and Scaffold N90 (Table 4). These results demonstrate that the *A. wenchowensis* genome has high continuity, completeness and accuracy.

Statistics	*A. fasciatus* (GCA_033556955.1)	*A. fasciatus* (GCA_036784905.1)	*A. wenchowensis* (hap1)	*A. wenchowensis* (hap2)
Total length (bp)	879,520,627	899,129,631	856,991,113	868,363,369
Contig number	134	120	38	36
Pseudo-chromosomes number	25	25	25	25
Contig N50	32,702,747	32,348,965	31,121,667	31,072,192
Contig N90	11,417,557	17,285,356	22,733,280	20,680,241
Scaffold N50	32,702,747	33,862,838	33,489,200	33,986,604
Scaffold N90	11,417,557	27,383,622	29,092,317	29,877,091
GC (%)	37.5	37.5	37.4	37.4

Table 4. Comparative statistics of genome assembly results.

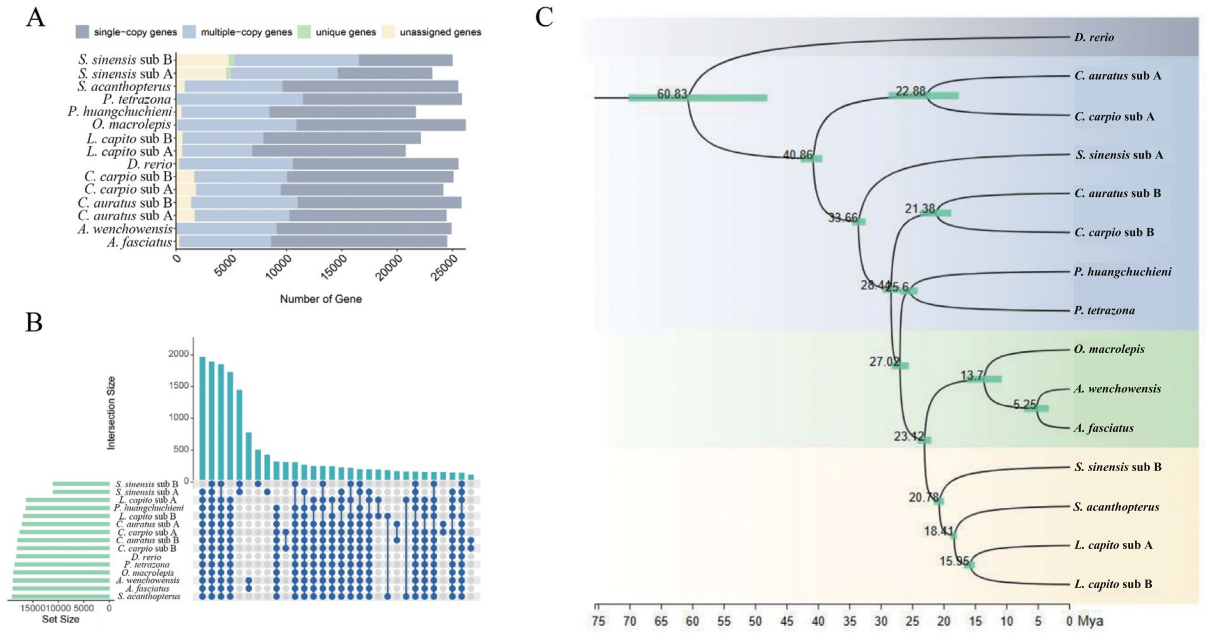

Fig. 3 Phylogenetic analysis. (**A**) Statistical information on gene families. Single: Gene of single-copy orthologs, Multiple: Gene of multiple-copy orthologs, Unassigned: Unassigned genes, Unique: Unique families, Other: Gene of other orthologs. (**B**) Upset relationship of orthologs genes of 6 species (*A. wenchowensis*, *C. auratus*, *C. carpio*, *C. idella*, *D. rerio*, *M. amblycephala*). (**C**) Divergence time and contraction, expansion of gene families for different species.

Gene sets of 10 fish species (*A. fasciatus*, *Danio rerio*, *Carassius auratus*, *Cyprinus carpio*, *Spinibarbus sinensis*, *Poropuntius huangchuchieni*, *Puntigrus tetrazona*, *Luciobarbus capito*, *Scaphiodonichthys acanthopterus*, *Onychostoma macrolepis*) were downloaded from public database (Table s11)[20]. OrthoFinder v2.3[64] was used to construct ortholog groups, including *A. wenchowensis*. Cluster analysis of gene families was performed using Diamond v2.0.6[65]. Single-copy orthologous genes shared among the 11 genomes were aligned using MAFFT v7.427[66], followed by construction of a maximum likelihood phylogenetic tree with RAxML v8.2.12[67] (parameters: bootstrap = 1000, model PROTGAMMAJTT). Species divergence times were estimated using MCMCTREE v4.9[68], with the Crown age of *Cyprinidae* (40.4~48.6 Mya), Stem age of *Barbinae* and *Cypininae* (27.82~33.9 Mya), and the split between Barbus and its sister group, *Luciobarbus* (18.0~19.0 Mya). A total of 27,137 orthologous genes were identified across the genomes of the *A. wenchowensis* and the *A. fasciatus*, and nine other fish species. Of these, 25,698 orthologs were shared by all species, including 695 single-copy orthologs, and 6,311 to 11,355 multi-copy orthologs (Fig. 3A). Additionally, we identified four unique orthologs in the *A. wenchowensis* and 1,893 shared orthologs (Fig. 3B). We selected 3,400 single-copy orthologs for phylogenetic tree construction, and the results indicate that the genera *Acrossocheilus* and *Onychostoma* diverged approximately 13.7 Mya, while *A. wenchowensis* differentiated from *A. fasciatus* around 5.25 Mya (Fig. 3C).

We further conducted resequencing analysis on 80 (20 from each locality) *A. wenchowensis* samples collected from Huangshan City, Anhui Province (AH, 29°79′N, 118°12′E), Ningde City, Fujian Province (ND, 27°19′N,119°84′E), Longyan City, Fujian Province (LY, 25°56′N, 116°29′E), and Shaoxing City, Zhejiang Province (Sx, 30°25′N, 120°69′E) (Fig. 4A). gDNA extraction and sequencing methods followed those described previously. BWA[69] was used for genome alignment (Hap 1 was selected), SAMtools[61] and Picard (http://broadinstitute.github.io/picard/) were used to

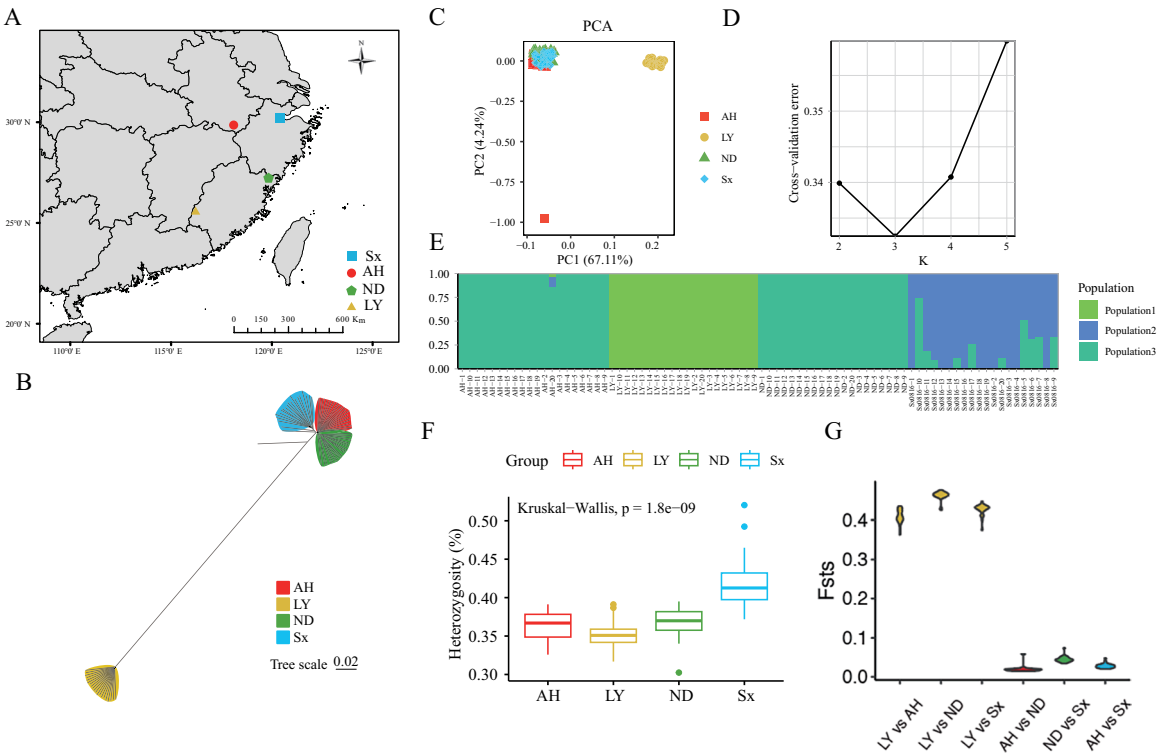

Fig. 4 Distribution map, population structure and selective sweep analysis. (**A**) Distribution map showing the locations of the four populations in China. (**B**) Phylogenetic trees of different populations, with samples from the same subpopulation shown in the same color, and evolutionary branching trends of samples with closer relatives clustered together. (**C**) PCA clustering diagram of four populations. (**D**) and (**E**) Analysis of the genetic structure of the population, where each histogram represents a sample and the colors of the columns and their proportions represent the different genetic backgrounds and their proportions in this sample. (**F**) Heterozygosity distribution in four populations. (**G**) *Fst* analysis for the four populations.

calculate whole-exome sequencing depth and chromosome coverage. GATK[70] was used to identify SNP and INDEL variations, which were annotated with SnpEff. Population genetic structure was analyzed using Admixture v1.3.0[71]. TreeBest v.1.9.2 (http://treesoft.sourceforge.net/treebest.shtml) was used to estimate individual-based adjacency trees based on *p*-distance, with 1000 bootstrap replications. *Fst* values calculated with VCFtools[72]. The results showed that the average sequencing size and depth for each sample were 12.85 Gb and 14.19×, respectively, with an average genome alignment rate of 98.46% and coverage of 95.21% (Table s12)[20]. A total of 52,309,888 SNPs were detected, with 1.79% being synonymous SNPs and 3.43% located in exons (Table s13)[20]. Meanwhile, 14,198,848 INDELs were identified, predominantly annotated in intron and intergenic regions (Fig. s8)[20]. Based on the detected SNPs, a phylogenetic tree (Fig. 4B) and principal components analysis (PCA) (Fig. 4C) were constructed. Population genetic structure analysis revealed three distinct subpopulations when $K = 3$. Some samples from the Sx population exhibited gene flow with the AH and ND (Fig. 4D-4E), and the Sx population showed significantly higher heterozygosity than the other populations (Fig. 4F). Pairwise *Fst* values among the four populations showed that the average *Fst* values between the AH, ND, and Sx populations were all below 0.1, while *Fst* values between the LY and the other three populations approached 0.5 (Fig. 4G). π values and XP-CLR analysis also indicated significant genetic distance between the LY and ND populations (Fig. s9)[20].

Data availability

The DNA and RNA sequencing raw data have been deposited in the NCBI Sequence Read Archive (SRA) database[55]. The genome assembly was submitted to GenBank with the accession number JBTMPJ000000000[56] and JBTMPK000000000[57]. The genome annotation was deposited in the *Figshare* database[58]. Supplementary tables and figures for this study are available in the *Figshare* repository[20]. The raw SNP and INDEL data for the four different groups are deposited in the Zenodo repository[59].

Code availability

All experimental procedures were carried out using commercially available standard, tools, so that no custom scripts had to be created. In addition, all bioinformatics tools and pipelines used are publicly available. Unless specific parameters are stated otherwise, the default settings recommended by the developers were used throughout.

Received: 5 September 2025; Accepted: 27 January 2026;
Published online: 14 February 2026

References

1. Han, J. *et al*. Identification and phylogenetic analysis of the mitogenome from Acrossocheilus wenchowensis. *Mitochondrial DNA B* **4**, 1277–1279 (2019).
2. Zhang, Y. & Jiang, J. A study on the artificial propagation of Acrossocheilus fasciatus. *Journal of Zhejiang Ocean University (Natural Science)* **29**, 211–214 (2010).
3. Wei, W. *et al*. Integrated mRNA and miRNA expression profile analysis of female and male gonads in Acrossocheilus fasciatus. *Biology* **11**, 1296 (2022).
4. Baroiller, J. F. *et al*. Environmental effects on fish sex determination and differentiation. *Sex Dev* **3**, 118–135 (2009).
5. Hoyt, S. J. *et al*. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022).
6. Zhao, D., Guo, Y. & Gao, Y. Natural selection drives the evolution of mitogenomes in Acrossocheilus. *PloS One* **17**, e0276056 (2022).
7. Wei, Z. *et al*. Transcriptional modulation reveals physiological responses to temperature adaptation in Acrossocheilus fasciatus. *Int J Mol Sci* **24**, 11622 (2023).
8. Ren, Y. *et al*. dmrt3, nom1, abce1, and pkmyt1 play key roles in gonadal sex determination in Acrossocheilus fasciatus. *Aquacult Int* **31**, 317–332 (2023).
9. Wang, C. *et al*. Protective effects of berberine in chronic copper-induced liver and gill injury in freshwater grouper (Acrossocheilus fasciatus). *Ecotoxicol Environ Saf* **267**, 115672 (2023).
10. Zheng, J. *et al*. Chromosome-level genome assembly of Acrossocheilus fasciatus using PacBio sequencing and Hi-C technology. *Sci Data* **11**, 166 (2024).
11. Yuan, Y. *et al*. Chromosome-scale genome assemblies of sexually dimorphic male and female Acrossocheilus fasciatus. *Sci Data* **11**, 653 (2024).
12. Wang, B. *et al*. Long and Accurate: How HiFi Sequencing is Transforming Genomics. *Genomics Proteomics Bioinformatics* **23**, qzaf003 (2025).
13. Li, R. *et al*. A sheep pangenome reveals the spectrum of structural variations and their effects on tail phenotypes. *Genome Res* **33**, 463–477 (2023).
14. Huang, Z. *et al*. Evolutionary analysis of a complete chicken genome. *Proc. Natl Acad. Sci. USA* **120**, e2216641120 (2023).
15. Xue, L. *et al*. Telomere-to-telomere assembly of a fish Y chromosome reveals the origin of a young sex chromosome pair. *Genome Biol* **22**, 203 (2021).
16. Schoon, I. Towards an integrative taxonomy of social-emotional competences. *Front Psychol* **12**, 515313 (2021).
17. Yang, L. *et al*. Phylogeny and polyploidy: resolving the classification of cyprinine fishes (Teleostei: Cypriniformes). *Mol Phylogenet Evol* **85**, 97–116 (2015).
18. Green, M. R. Sambrook, J. Isolation of High-Molecular-Weight DNA using organic solvents. *Cold Spring Harb Protoc* **4**, pdb. prot093450 (2017).
19. Chen, S. *et al*. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
20. Luo, M. Supplementary files for Acrossocheilus wenchowensis genome. *Figshare. Dataset*. https://doi.org/10.6084/m9.figshare.29400050.v3 (2025).
21. Liu, B. *et al*. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *Quantitative Biol* **35**, 62–67 (2013).
22. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
23. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
24. Chen, Y. *et al*. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience* **7**, 1–6 (2018).
25. Cheng, H. *et al*. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).
26. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
27. Ye, J. *et al*. BLAST: Improvements for better sequence analysis. *Nucleic Acids Res* **34**, w6–9 (2006).
28. Hu, J. *et al*. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol* **25**, 107 (2024).
29. Fang, L. & Wang, K. Polishing high-quality genome assemblies. *Nat Methods* **19**, 649–650 (2022).
30. Servant, N. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* **16**, 259 (2015).
31. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
32. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst* **3**, 99–101 (2016).
33. Wolff, J. *et al*. Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res* **48**, W177–W184 (2020).
34. Jain, C. *et al*. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat Methods* **19**, 705–710 (2022).
35. Marçais, G. *et al*. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol* **14**, e1005944 (2018).
36. Kirov, I. *et al*. Pilot satellitome analysis of the model plant, Physcomitrella patens, revealed a transcribed and high-copy IGS related tandem repeat. *Comp Cytogenet* **12**, 493–513 (2018).
37. Vollger, M. R. *et al*. StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinformatics* **38**, 2049–2051 (2022).
38. Zheng, J. *GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_033556955.1 (2023).
39. Yuan, Y., Ren, J. & Zhong, T. *GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_036784905.1 (2024).
40. Yuan, Y., Ren, J. & Zhong, T. *GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_039880705.1 (2024).
41. Grativol, C. *et al*. A miniature inverted-repeat transposable element, AddIn-MITE, located inside a WD40 gene is conserved in Andropogoneae grasses. *PeerJ* **7**, e6080 (2019).
42. Chen, N. Using Repeat Masker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* **5**, 4–10 (2004).
43. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435–9 (2006).
44. Korf, I. Gene finding in novel genomes. *Bioinformatics* **5**, 59 (2004).
45. Majoros, W. H. *et al*. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–9 (2004).
46. Lomsadze, A. *et al*. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res* **42**, e119 (2014).
47. Keilwagen, J. *et al*. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods Mol Biol* **1962**, 161–177 (2019).

48. Kim, D. *et al*. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915 (2019).
49. Trapnell, C. *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511–515 (2010).
50. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**, 5654–5666 (2003).
51. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7 (2008).
52. 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 (1997).
53. Goel, M. *et al*. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol* **20**, 277 (2019).
54. Cingolani, P. *et al*. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
55. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP510699 (2024).
56. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/nuccore/JBTMPJ000000000.
57. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/nuccore/JBTMPK000000000.
58. Luo, M. Telomere-to-Telomere and Haplotype-Phased Genome Assemblies of the *Acrossocheilus wenchowensis*. *Figshare. Dataset.* https://doi.org/10.6084/m9.figshare.25911631.v1 (2025).
59. Luo, M. Indel and Snp data for *Acrossocheilus wenchowensis* from four different populations. *Zenodo. Dataset.* https://doi.org/10.5281/zenodo.17861395 (2025).
60. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
61. Li, H. *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009).
62. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
63. Seppey, M. *et al*. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol Biol.* **1962**, 227–245 (2019).
64. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238 (2019).
65. Buchfink, B. *et al*. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60 (2015).
66. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772–80 (2013).
67. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–3 (2014).
68. Puttick, M. N. MCMCtreeR: functions to prepare MCMCtree analyses and visualize posterior ages on trees. *Bioinformatics* **35**, 5321–5322 (2019).
69. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).
70. Wang, K. *et al*. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
71. Alexander, D. H. *et al*. *Admixture 1.3 software manual*. (2015).
72. Danecek, P. *et al*. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–8 (2011).

Acknowledgements

This work was funded by the Fujian Provincial Bureau of Ocean and Fisheries (2021FJSCZY04).

Author contributions

Lingzhan Xue, Mingkun Luo, Haoyu Wang: Investigation, Writing - Original draft. Wenbin Zhu, Gaoxiong Zeng, Mengxiang Liao, Ji Zhao: Resources. Duhuang Chen: Data curation. Lingzhan Xue, Bin Wu: Funding acquisition. Luohao Xu: Conceptualization, Writing - Review & Editing. Zaijie Dong: Conceptualization, Writing - Review & Editing, Supervision.

Competing interests

The authors declare that they have no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.W., L.X. or Z.D.

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.