



OPEN

DATA DESCRIPTOR

Chromosome-level assembly and annotation of the yellow-shelled fish (*Barbodes Wynaadensis*)

Qi Shen^{1,4}, Junjie Wu^{2,4}, Chunxiang Yang², Qiang Sheng¹, Mei Xu³, Shaokui Yi¹✉ & Jianyu Song²✉

Barbodes wynaadensis, a unique cyprinid species native to Yunnan Province in China, stands out as an allotetraploid (AABB) fish with a complex evolutionary history. Leveraging a multi-platform sequencing strategy combining MGI short-read, PacBio long-read, and Hi-C scaffolding technologies, we assembled the first chromosome-level genome for *B. wynaadensis*. The final assembled genome spans 1.76 Gb in length with a contig N50 of 33.53 Mb, demonstrating high assembly continuity. Hi-C scaffolding enabled the reconstruction of 50 pseudochromosomes, representing 99.94% of the total genome assembly. Genome annotation identified 46,121 protein-coding genes, with a functional annotation rate of 99.76%. Repetitive elements constituted 48.26% of the genomic sequences, including lineage-specific expansions of DNA transposons (29.26%) and LTRs (6.36%). This high-quality assembly resolves challenges in polyploid genome reconstruction and provides a critical resource for investigating Cyprinidae evolution, particularly subgenome divergence and adaptation. The dataset also enables practical applications, such as molecular marker development for population monitoring, supporting conservation efforts for this threatened endemic species amid habitat degradation in the Nujiang River basin.

Background & Summary

The freshwater cyprinid species *Barbodes wynaadensis* belongs to the family Cyprinidae, subfamily Barbinae, and genus *Barbodes*. This species was first formally described by Wu *et al.*¹, with subsequent taxonomic revisions and a comprehensive synopsis of the genus *Barbodes*, including an identification key, provided by Chen *et al.*². *B. wynaadensis* is an endemic species exclusively distributed within the Nujiang River basin in Yunnan Province, China. However, with increasing human activities and environmental degradation, *B. wynaadensis* faces significant threats. In recent years, the species has been increasingly targeted for its perceived value, leading to over-exploitation and indiscriminate capture driven by economic interests. Coupled with the serious environmental problems, these factors have contributed to a sharp decline in the wild population of *B. wynaadensis*³. Due to its restricted geographic range and relatively small population size, it has been listed as a protected aquatic wildlife species in Yunnan Province⁴.

Morphologically, *B. wynaadensis* is characterized by a fusiform body shape, with adults typically reaching a substantial size. The species exhibits a predominant cyan-blue body coloration, with lateral line scales that are primarily yellow, although some individuals display black lateral line scales, a phenotypic variation potentially influenced by local water quality and environmental factors. In its natural habitat, the yellow lateral line scales form a distinct longitudinal band, earning it the common name “yellow-shelled fish.” Ecologically, *B. wynaadensis* is an omnivorous species. Reproductive studies indicate that males generally achieve gonadal maturity by two years of age, while females reach maturity by three years, suggesting a one-year difference in reproductive development between sexes⁵. The species is highly valued for its large body size, tender flesh, and sweet taste, which contribute to its economic importance. Additionally, its large scales, blue body coloration, and the striking yellow longitudinal band formed by the lateral line scales enhance its ornamental appeal. *B. wynaadensis* predominantly inhabits the upper and middle layers of the water column. Beyond its economic and ornamental value, the species is also recognized for its potential medicinal properties. Given its favorable aquaculture potential,

¹School of Life Sciences, Huzhou University, Huzhou, 313000, China. ²Yunnan Institute of Fishery Sciences Research, Kunming, 650111, China. ³Yunnan Agricultural Broadcast and Television School, Kunming, 650000, China. ⁴These authors contributed equally: Qi Shen, Junjie Wu. ✉e-mail: yishakui@foxmail.com; sjy020829@126.com



Fig. 1 Overview of the *B. wynaadensis* specimen used for sequencing in the study.

Sequencing strategy	Platform	Usage	Clean data (Gb)	Sequencing depth (×)
Short-reads	Illumina NovaSeq 6000	Genome survey	139.25	79
Hi-Fi	PacBio Sequel II	Assembly	70.9	40
Hi-C	Illumina HiSeq X Ten	Hi-C assembly	186.7	106
ISO-seq	Illumina HiSeq X Ten	Annotation	11.0	—

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

B. wynaadensis represents a species of significant academic interest and practical value, particularly in the fields of conservation biology, aquaculture development, and traditional medicine research.

In 2016, Zhou *et al.*⁶ conducted the first mitochondrial genome sequencing of *B. hexagonolepis* and reconstructed a phylogenetic tree, which confirmed its distinct phylogenetic position relative to other genera within the subfamily Barbinae. Despite this advancement, genomic research on the genus *Barbodes* remains scarce. In this study, we present the first high-quality genome assembly of *B. wynaadensis*, which serves as a critical genomic resource for elucidating phylogenetic relationships and adaptive evolutionary mechanisms within the genus. The findings provide crucial insights for conservation of the *Barbodes* species, particularly subgenome divergence and adaptation. This genomic resource not only advances our understanding of adaptive evolution in cyprinids but also provides scientific guidance for both wild population conservation and sustainable aquaculture breeding practices.

Methods

Sample collection and sequencing. In September 2023, a 2-year-old male *B. wynaadensis* (Fig. 1) was collected in Nanpeng River, Yunnan Province, China (23.68°N, 98.94°E). High-quality, high-molecular-weight genomic DNA was extracted from muscle tissue using the cetyltrimethylammonium bromide (CTAB) method⁷. DNA purity was assessed through 1% agarose gel electrophoresis and quantified using a NanoDrop™ One UV-Vis spectrophotometer (Thermo Fisher Scientific, USA). DNA concentration was further determined using a Qubit 4.0 fluorometer (Invitrogen, USA). Following quality assessment, paired-end sequencing was performed on the Illumina HiSeq X Ten platform, generating 144.34 Gb of raw reads. Quality control of the sequencing data was conducted using fastp v0.23.2⁸, resulting in 139.25 Gb of clean reads. For long-read sequencing, a SMRTbell library was constructed and sequenced on the PacBio Sequel II system (Pacific Biosciences, USA), yielding 70.9 Gb (~40 × sequencing depth) of high-quality Circular Consensus Sequencing (CCS) reads after preprocessing with the CCS program⁹. Additionally, DNA extracted from liver tissue was subjected to Hi-C library construction and sequencing using the HiSeq X Ten platform (Illumina, USA), producing 186.7 Gb of clean reads. To support transcriptomic analysis, high-quality RNA was isolated from multiple tissues, including beard, spleen, gill, liver, hypothalamus, pituitary, heart, kidney, eye, and intestine. RNA sequencing libraries were constructed and sequenced on the Illumina NovaSeq6000 platform, generating 11.0 Gb of clean reads for downstream analyses (Table 1).

Genome survey. To characterize the genetic composition of the sample, 10,000 paired-end reads were randomly subsampled from the filtered high-quality sequencing data. These reads were subjected to homology-based alignment against the NCBI nucleotide database using BLAST software, with a focus on identifying the top 80% most abundant species. The analysis demonstrated that 86.18% of the reads aligned to the genus *Cyprinus*, confirming the absence of significant exogenous contamination in the sequencing library. Subsequently, Smudgeplot v1.0.0¹⁰ was employed to perform *K*-mer analysis on the entire high-quality dataset. Genome characteristics, including genome size, heterozygosity rate, and repetitive sequence content, were inferred using a *K*-mer length of 17 nucleotides. The results estimated a genome size of approximately 1,665 Mb, a heterozygosity rate of 0.38%, and a repetitive sequence content of 65.30% (Fig. 2a). Additionally, Smudgeplot was utilized to extract and

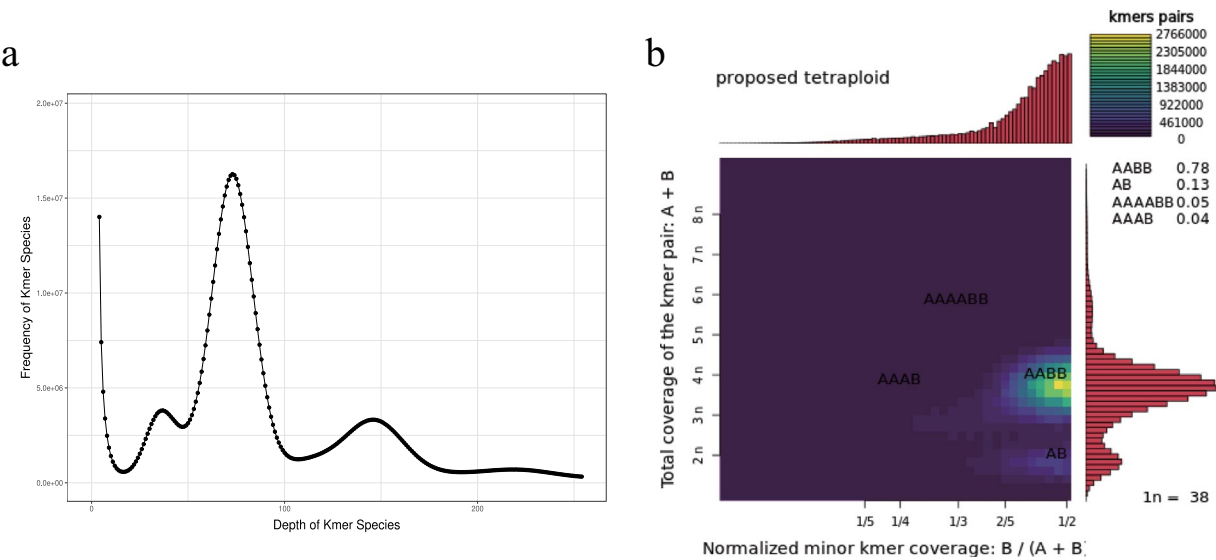


Fig. 2 Genome size estimation and ploidy determination. **(a)** Frequency distribution of *K*-mer depth and species-specific *K*-mer patterns; **(b)** Smudgeplot output showing tetraploid (AABB) configuration, where dot clusters represent distinct subgenome coverage ratios (CovB/[CovA + CovB]).

Assembly	<i>B. wynaadensis</i>
Estimated genome size by <i>k</i> -mer (Mb)	1665
Total scaffolded assembly size (Mb)	1756
Contigs N50 (Mb)	32.42
Completeness BUSCOs (%)	97.61
Chromosome number	50
Anchoring ratio (%)	99.94

Table 2. Genome assembly and assessment of *B. wynaadensis* genome.

analyze heterozygous *K*-mer pairs from the *K*-mer database generated from the sequencing data with the parameter of “-k21-m200-ci1-cs10000”. By evaluating the total coverage (CovA + CovB) and relative coverage (CovB / (CovA + CovB)) of *K*-mer pairs, the number of heterozygous *K*-mer pairs was quantified, and the genome structure was determined. The analysis revealed a tetraploid (AABB) genome configuration with a confidence score of 0.78 (Fig. 2b).

Genome assembly and Hi-C scaffolding. The initial genome assembly was performed using Hifiasm¹¹ due to its demonstrated efficacy in generating haplotype-phased assemblies from long-read sequencing data. The Hifiasm yielded a draft assembly spanning 1.77 Gb with an N50 of 32.49 Mb, comprising 169 contigs. To address redundancy in the preliminary assembly, purge_haplotigs v1.2.5¹² was utilized to identify and remove redundant heterozygous contigs based on read depth distribution and sequence similarity. The refined assembly consisted of 137 contigs, totaling 1.76 Gb (Table 2), which served as the basis for downstream analyses. Short-read and long-read data were aligned to the genome using BWA v0.7.12¹³ and minimap2¹⁴, respectively. Alignment statistics revealed a short-read mapping rate of 99.76%, an average sequencing depth of 81.11 ×, and 98.84% of sequences achieving a depth of 20 × or greater. For long-read data, the mapping rate was 99.99%, with an average depth of 39.65 × and 95.26% of sequences exceeding 20 × depth. Hi-C data were aligned to the genome using Juicer v1.6¹⁵ to facilitate chromosome-level scaffolding. Error correction was subsequently performed using 3D-DNA v180922¹⁶ and Juicebox v1.11.08¹⁷. During the scaffolding and error correction processes, the original 137 contigs were reorganized based on interaction maps. To further split the assembled chromosomes into subgenomes, we mapped the genome sequences against the reference genome of *Cyprinus carpio* (version: GCA_027406505.1, 2n = 100), which is a representative species of the Cyprinidae family. The final assembly comprised 50 chromosomes and 28 scaffolds, with a total length of 1.76 Gb and a N50 of 33.53 Mb (Fig. 3a, Table 3). The genome size is close to that of its congeneric species, *Sinocyclocheilus graham* (1.75 Gb)¹⁸, consistent with their phylogenetic proximity within Cyprinidae. The chromosome anchoring rate reached 99.94%. Assembly completeness was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO, v5.3.1¹⁹). BUSCO analysis showed that among 3640 single copy homologues in the actinopterygii_odb10 database, 97.61% were complete BUSCOs (24.40% single copy genes, 73.21% repeat genes), 0.25% were fragmented and 2.14% were missing. These results collectively confirm the successful generation of a high-quality, chromosome-level genome assembly (Fig. 3b).

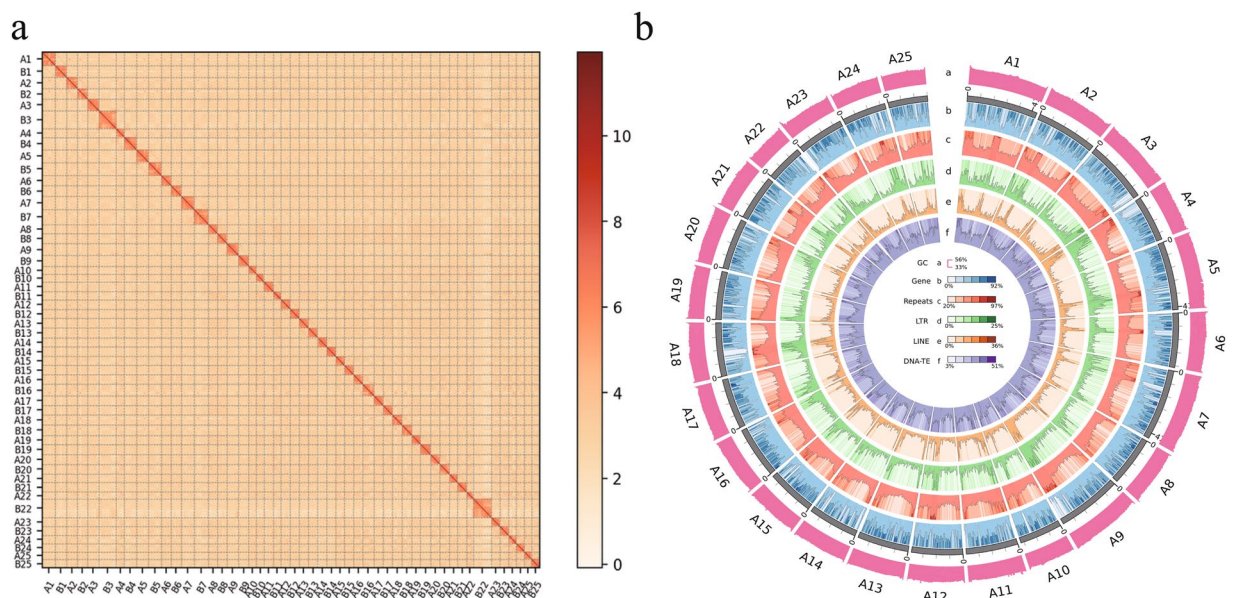


Fig. 3 Summary of the *B. wynaadensis* genome assembly. (a) Hi-C interaction heatmap illustrating chromosomal interactions, the color blocks represent the interaction strength from yellow (low) to red (high); (b) Chromosomal characteristics. The window size, calculated as the genome size divided by 3,000, corresponds to approximately 587 kb.

Repeat and protein-coding gene annotation. Repetitive sequences were annotated through an integrated approach combining three methodologies: *de novo* prediction using TRF v4.09²⁰; homology-based prediction employing RepeatMasker v4.09²¹ and RepeatProteinMask²² with the RepBase database; and a hybrid approach utilizing RepeatModeler v1.0.11²³, LTR_FINDER_parallel v1.0.7²⁴, and RepeatMasker. This comprehensive analysis identified 847.46 Mb of repetitive sequences, representing 48.26% of the assembled genome. The annotated repetitive elements included 29.26% DNA transposons, 4.99% long interspersed nuclear elements (LINEs), and 6.36% long terminal repeats (LTRs) (Table 4). DNA transposons, as the dominant repeat type, may have contributed to gene regulatory diversification through exaptation as cis-regulatory elements.

The structural annotation of protein-coding genes was performed using an integrative multi-method approach. Homology-based prediction was conducted using Exonerate v2.2.0²⁵ with parameter of “--model protein2genome--showtargetgff 1” and LiftOff v1.6.3²⁶, utilizing genomic data from closely related species (*Carassius auratus*, *C. gibelio*, *Ctenopharyngodon idella*, *Cyprinus carpio*, and *Danio rerio*). *De novo* prediction was implemented through Augustus v3.3.2²⁷ and Genscan. For transcript-based prediction, RNA-seq data were aligned using HISAT2 v2.1.0²⁸, followed by transcript assembly with StringTie v1.3.5²⁹. Full-length transcript isoforms were reconstructed using ISOseq3, and coding sequences were predicted with TransDecoder v5.0³⁰. Additionally, BUSCO5 was employed to optimize the Augustus self-training model and validate predictions against species-specific conserved gene sets. The gene sets generated from these diverse approaches were integrated into a non-redundant, comprehensive gene set using MAKER2 v2.31.10³¹. Final refinement and integration were performed using the in-house HiFAP software (Wuhan OneMore Tech Co., Ltd., <https://www.onemore-tech.com/>), resulting in a high-confidence gene set containing 46,121 protein-coding genes. Functional annotation of the predicted proteins was conducted by aligning against multiple databases, including the non-redundant protein database (NR), Kyoto Encyclopedia of Genes and Genomes (KEGG), Swiss-Prot, TrEMBL, eukaryotic Orthologous Groups (KOG), and Gene Ontology (GO), using Diamond v2.0.14 with the parameter of “--evaluate 1e-05”. KEGG pathway mapping was performed using KOBAS v3.0³² with the parameter of “-t blastout:tab -s ko”. Conserved domain analysis was carried out using InterProScan v5.61-93.0³³ (parameter: “--seqtype p--formats TSV--goterms--pathways -dp”) and HMMER3 v3.3.1³⁴, achieving functional annotation for 99.76% (46,076) of the predicted genes (Table 5). Non-coding RNA annotation included tRNA identification using tRNAscan-SE v1.3.1³⁵, rRNA alignment with BLASTN v2.11.0, and miRNA/snRNA annotation using Rfam v14.8³⁶. This analysis identified 1,923 miRNAs, 3,337 rRNAs, 7,912 tRNAs, and 2,066 snRNAs (Table 6). Gene prediction completeness was assessed using BUSCO, and the result showed that 99.0% were complete BUSCOs, 0.3% were fragmented and 0.7% were missing, which confirms the high accuracy and reliability of the annotation results.

Data Records

The DNA sequencing data from the PacBio HiFi library are available under the SRA accession number SRR33118795³⁷, the Hi-C library data under SRR33053330³⁸, the short-read genomic sequencing data under SRR33038255³⁹, and the RNA-seq data under SRR33038300⁴⁰. The assembled genome sequences have been deposited in the figshare⁴¹ and NCBI GenBank with the accession number GCA_050231085.1⁴². The genome annotation results have been deposited in the figshare database⁴³.

Pseudo-chromosomes	Length (bp)	Percentage (%)
ChrA1	44,942,959	2.56%
ChrB1	39,671,520	2.26%
ChrA2	38,658,994	2.20%
ChrB2	35,629,333	2.03%
ChrA3	39,669,728	2.26%
ChrB3	61,535,814	3.50%
ChrA4	29,278,426	1.67%
ChrB4	42,018,561	2.39%
ChrA5	43,472,899	2.48%
ChrB5	44,796,331	2.55%
ChrA6	34,541,378	1.97%
ChrB6	35,251,886	2.01%
ChrA7	45,220,695	2.57%
ChrB7	51,006,105	2.90%
ChrA8	31,515,007	1.79%
ChrB8	33,530,384	1.91%
ChrA9	40,737,619	2.32%
ChrB9	36,401,475	2.07%
ChrA10	26,948,553	1.53%
ChrB10	26,755,777	1.52%
ChrA11	32,755,848	1.87%
ChrB11	27,684,329	1.58%
ChrA12	32,347,550	1.84%
ChrB12	31,301,146	1.78%
ChrA13	32,511,389	1.85%
ChrB13	32,972,587	1.88%
ChrA14	31,577,642	1.80%
ChrB14	32,478,696	1.85%
ChrA15	30,197,892	1.72%
ChrB15	32,490,658	1.85%
ChrA16	32,639,146	1.86%
ChrB16	38,805,466	2.21%
ChrA17	31,951,385	1.82%
ChrB17	32,835,873	1.87%
ChrA18	34,174,382	1.95%
ChrB18	35,360,686	2.01%
ChrA19	30,477,323	1.74%
ChrB19	35,653,970	2.03%
ChrA20	32,409,301	1.85%
ChrB20	33,534,364	1.91%
ChrA21	28,943,742	1.65%
ChrB21	32,415,148	1.85%
ChrA22	23,518,999	1.34%
ChrB22	63,129,461	3.59%
ChrA23	29,979,761	1.71%
ChrB23	30,656,317	1.75%
ChrA24	27,897,507	1.59%
ChrB24	28,263,262	1.61%
ChrA25	25,001,006	1.42%
ChrB25	29,561,550	1.68%
Total	1,756,165,279	100.00%

Table 3. Summary of chromosome length of *B. wynaadensis* genome.

Technical Validation

The corrected high-quality short-read data were aligned to the assembled genome sequence using BWA, achieving a read mapping rate of 99.76%, with an average sequencing depth of $81.11 \times$. Notably, 98.84% of the sequences exhibited a depth of $20 \times$ or greater. For long-read data alignment, minimap2 was employed, resulting

	RepBase TEs		TE Proteins		De novo		Combined TEs	
	Length (bp)	%in Genome	Length (bp)	%in Genome	Length (bp)	%in Genome	Length (bp)	%in Genome
DNA	345,380,795	19.67	36,700,591	2.09	320,840,106	18.27	513,839,090	29.26
LINE	69,903,788	3.98	47,748,699	2.72	43,532,674	2.48	87,654,741	4.99
SINE	6,951,332	0.4	0	0	5,879,063	0.33	11,443,567	0.65
LTR	65,437,335	3.73	35,725,239	2.03	73,998,706	4.21	111,700,746	6.36
Satellite	42,061,381	2.4	0	0	7,789,470	0.44	48,345,108	2.75
Simple_repeat	0	0	0	0	271,965	0.02	271,965	0.02
Other	23,031	0	0	0	0	0	23,031	0
Unknown	7,963,636	0.45	4,785	0	93,017,950	5.3	98,590,755	5.61
Total	508,631,607	28.96	120,148,736	6.84	536,404,070	30.54	821,475,608	46.78

Table 4. Statistics of repetitive sequence classification results. Note: TEs, transposable elements; LINE, long interspersed nuclear elements; SINE, short interspersed nuclear elements; LTR, long terminal repeats.

Database	Number of annotated genes	Percentage (%)
NR	46,012	99.76
SwissProt	40,577	87.98
TrEMBL	45,991	99.72
KOG	37,569	81.46
TF	11,986	25.99
InterPro	44,444	96.36
GO	33,633	72.92
KEGG	45,928	99.58

Table 5. Number of genes annotated using different databases.

Type of ncRNA	Type	Number	Total length (bp)	Percentage of genome (%)
miRNA		1,923	169,063	0.009627
tRNA		7,912	593,099	0.033772
rRNA	rRNA	3,337	458,838	0.026127
	18S	43	72,232	0.004113
	28S	0	0	0
	5.8S	37	5,698	0.000324
	5S	3,257	380,908	0.02169
snRNA	snRNA	2,066	303,050	0.017256
	CD-box	311	40,118	0.002284
	HACA-box	150	22,712	0.001293
	splicing	1,584	235,559	0.013413
	scaRNA	21	4,661	0.000265

Table 6. Number of the annotated non-coding RNA.

in a mapping rate of 99.99%, an average sequencing depth of $39.65 \times$, and 95.26% of sequences achieving a depth of $20 \times$ or higher. BUSCO analysis demonstrated that 97.61% of the BUSCO genes were identified as complete, underscoring the high integrity of the genome assembly. To investigate the genomic synteny between *B. wynaadensis* and related cyprinid species (*Puntigrus tetrazona*, *Labo rohita*, and *C. auratus*), we conducted a comprehensive comparative genomic analysis. Initially, an all-versus-all protein sequence alignment was performed using BLASTP v2.11.0+⁴⁴ with a stringent e-value cutoff of $1e-5$. Subsequently, chromosomal location data and chromosome length information were extracted for syntenic block identification. These data were analyzed using JCVI v1.1.22⁴⁵, facilitating the interpretation of conserved genomic architecture among the studied species. Interspecies synteny analysis identified substantial gene pair conservation: *B. wynaadensis* shared 47,713 syntenic pairs with *P. tetrazona*, 47,194 with *L. rohita*, and notably higher conservation (83,381 pairs) with *C. auratus*, suggesting significant homology at both the sequence and structural levels (Fig. 4). Furthermore, the Hi-C interaction heatmap revealed significantly higher interaction intensities along the diagonal positions compared to non-diagonal regions, providing strong evidence for the high quality of the chromosome-level genome assembly.

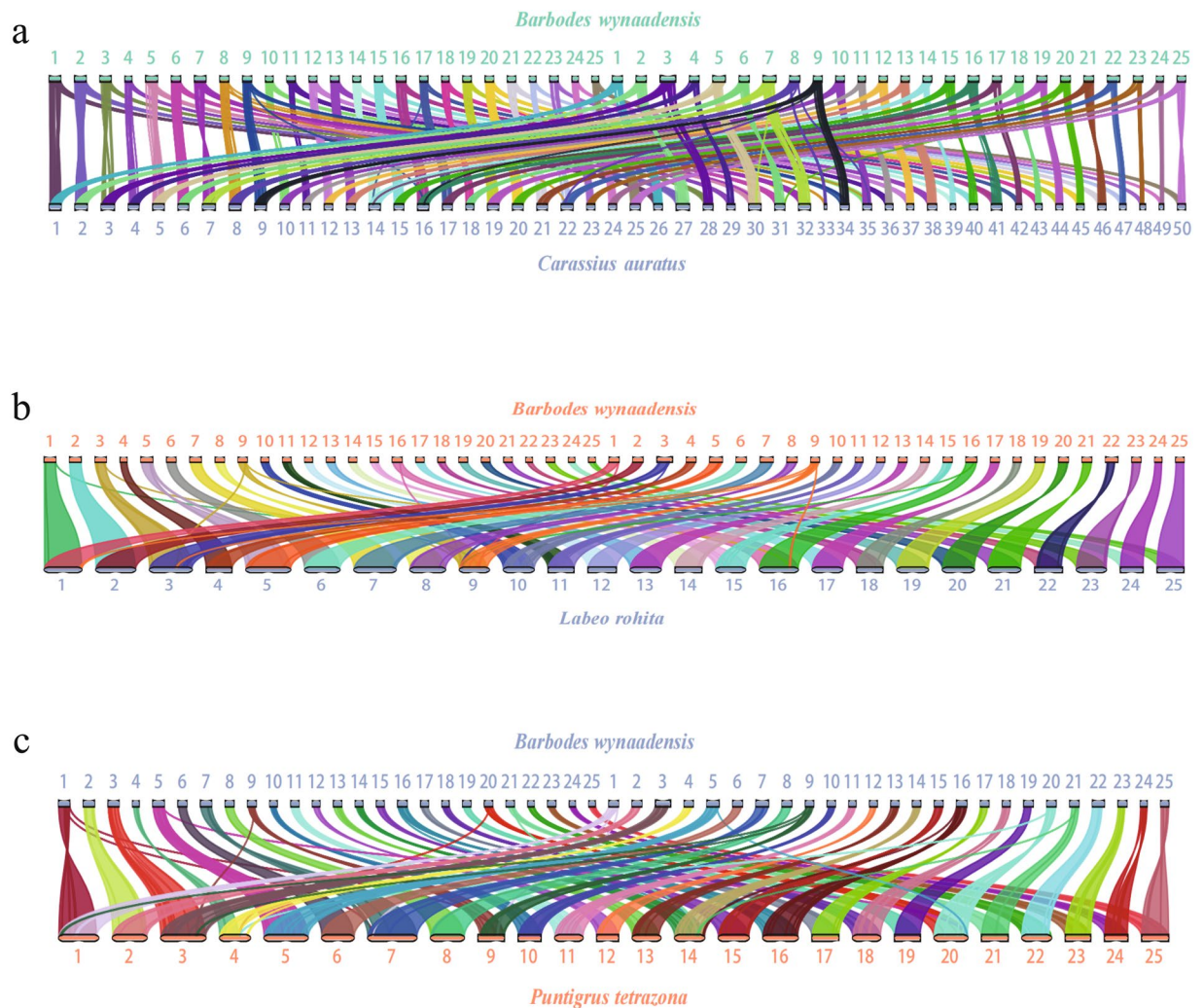


Fig. 4 Comparative synteny analysis between *B. wynaadensis* and related cyprinidae species. (a) *B. wynaadensis* vs *C. auratus*; (b) *B. wynaadensis* vs *L. rohita*; (c) *B. wynaadensis* vs *P. tetrazona*.

Code availability

The versions and parameters of bioinformatic tools used in this study have been described in the Method section. If no parameter is provided, the default is used. No custom code was used.

Received: 28 April 2025; Accepted: 3 July 2025;

Published online: 11 July 2025

References

1. Wu, X. W. The cyprinid fishes of China. in vol. 2 236–252 (People's Press, Shanghai, 1977).
2. Chen, X., Yang, J. & Chen, Y. A Review of the Cyprinoid Fish Genus *Barbodes* Bleeker, 1859, from Yunnan, China, with Descriptions of Two New Species. *Zool. Stud.* **38**, 82–88 (1999).
3. Yu, Z. Research on the Aquaculture Techniques of Yellow-shelled Fish (*Barbodes wynaadensis*). *World Trop. Agric. Inf.* **74**, 76 (2024).
4. Kong, L. *et al.* Artificial propagation technology of *Barbodes wynaadensis*. *J. Aquac.* **41**, 52–53+55 (2020).
5. Zhang, J. *et al.* Study on the biological characteristics of *Barbodes wynaadensis*. *Freshw. Fish.* 52–53 (2003).
6. Zhou, C. *et al.* The complete mitochondrion genome of the *Barbodes hexagonolepis* (Cypriniformes, cyprinidae). *Mitochondrial DNA Part B* **1**, 158–159 (2016).
7. Healey, A., Furtado, A., Cooper, T. & Henry, R. J. Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. *Plant Methods* **10**, 21 (2014).
8. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
9. Travers, K. J., Chin, C.-S., Rank, D. R., Eid, J. S. & Turner, S. W. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* **38**, e159 (2010).
10. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
11. 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–+ (2021).
12. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).
13. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
14. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

15. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98 (2016).
16. Zhang, Y., Xiong, Y. & Xiao, Y. 3dDNA: A Computational Method of Building DNA 3D Structures. *Molecules* **27**, 5936 (2022).
17. Robinson, J. T. *et al.* Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cell Syst.* **6**, 256–258.e1 (2018).
18. Yang, J. *et al.* The *Sinocyclocheilus* cavefish genome provides insights into cave adaptation. *BMC Biol.* **14**, 1 (2016).
19. 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).
20. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
21. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma.* **4**, 4.10.1–4.10.14 (2009).
22. Liu, Z. *et al.* Chromosome-level genome assembly of the deep-sea snail *Phymorhynchus buccinoides* provides insights into the adaptation to the cold seep habitat. *BMC Genomics* **24**, 679 (2023).
23. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA.* **117**, 9451–9457 (2020).
24. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–268 (2007).
25. Guy, S. & Ewan, B. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
26. Shumate, A. & Salzberg, S. L. LiftOff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).
27. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
28. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
29. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
30. Almeida, D. *et al.* Data Employed in the Construction of a Composite Protein Database for Proteogenomic Analyses of Cephalopods Salivary Apparatus. *Data* **5**, 110 (2020).
31. Carson, H. & Mark, Y. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
32. Bu, D. *et al.* KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res.* **49**, W317–W325 (2021).
33. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
34. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121–e121 (2013).
35. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096 (2021).
36. Sellés Vidal, L., Ayala, R., Stan, G.-B. & Ledesma-Amaro, R. rfaRm: An R client-side interface to facilitate the analysis of the Rfam database of RNA families. *PLOS ONE* **16**, e0245280 (2021).
37. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR33118795> (2025).
38. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR33053330> (2025).
39. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR33038255> (2025).
40. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR33038300> (2025).
41. The assembled genome sequences of *Barbodes Wynaadensis*. figshare <https://doi.org/10.6084/m9.figshare.29224193> (2025).
42. NCBI Assembly https://identifiers.org/ncbi/insdc.gca:GCA_050231085.1 (2025).
43. *Barbodes wynaadensis*. gene.gff. figshare <https://doi.org/10.6084/m9.figshare.28793867> (2025).
44. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
45. Tang, H. *et al.* Synteny and Collinearity in Plant Genomes. *Science* (2008).

Acknowledgements

This study was funded by the funding of “Selection of sex-specific molecular markers using chromosomal characteristics in the *Parachromis managuensis* and new variety breeding” (202401AT070091).

Author contributions

S. Yi and J. Song conceived the study. J. Wu and C. Yang collected samples. Bioinformatics analysis was performed by Q. Shen, Q. Sheng and M. Xu, Q. Shen and J. Wu wrote and revised the original manuscript. All authors have read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.Y. or J.S.

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.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025