scientific data

OPEN

DATA DESCRIPTOR

Chromosome-scale genome assembly and annotation of *Xenocypris argentea*

Yidi Wu[1], Hang Sha[1] & Hongwei Liang[1,2] ✉

Xenocypris argentea **is a small to medium-sized freshwater cyprinid fish. It distributes widely in the rivers and lakes of China, and is often used as a tool fish for water quality improvement and optimizing aquaculture structures. In recent years, natural populations of *X. argentea* have decreased rapidly due to human activities, yet little is known about the genetics and genomics of this fish. In the present work, we reported a chromosome-level reference genome of *X. argentea* based on PacBio HiFi, Hi-C and Illumina paired-end sequencing technologies. The assembled genome was 984.96 Mb in length, with a contig N50 of 36.02 Mb. Using Hi-C interaction information, 99.47% of the contigs were anchored onto 24 chromosomes, and 18 of the chromosomes were gap-free. Further analysis identified 560.27 Mb of repeat sequences and 28,533 protein-coding genes in the genome, of which, 95.62% (27,284) genes were functionally annotated. This high-quality genome offers an invaluable resource for population genetics and phylogeny, comparative genomics, adaptive evolution and functional exploration of *X. argentea*.**

Background & Summary

The *Xenocypris argentea* is a small to medium-sized freshwater fish which belongs to the family Cyprinidae, subfamily Xenocyprinae and genus *Xenocypris*. It distributes widely in the major river systems and affiliated lakes of China. This fish mainly inhabits the middle-to-lower layers of waters and feeds on decaying sediment, diatoms, and attached algae[1]. Therefore, it is often used as a tool fish for purifying water and optimizing aquaculture structures in reservoirs and ponds[2]. With the advantages of fast growth, low disease incidence, high reproductive capacity and strong adaptability, *X. argentea* can live in complex and diverse waters and often forms different geographical populations due to geographical isolation, resulting in abundant genetic resources[1,3]. However, the impact of human activities has led to a rapid decline in the natural population of *X. argentea* in recent years[2,4].

Currently, research on *X. argentea* is still very limited, mainly focusing on its ecology and biology[5,6], breeding techniques[7], genetic diversity[3,8,9] and molecular marker development[2,4,10]. However, data and studies about the genomics, adaptive evolution and genetic analysis of *X. argentea* are still scarce, which limits our understanding, protection and utilization of this species. In addition, the phylogenetic relationships of the subfamily Xenocyprinae are still unclear in molecular systematics research of the cyprinid fishes[11]. Most of the previous studies are analyzed based on a few mitochondrial or nuclear genes, and the resulted phylogenetic relationships of some species always varied with different molecular markers[11–13]. Nowadays, the development of sequencing technology has made it possible to conduct phylogenetic studies at whole-genome level. However, only the genomes of 2 fish species (*Pseudobrama simoni*[14] and *Plagiognathops microlepis*[15]) have been reported in the subfamily Xenocyprinae. Given these aspects, it is necessary to provide a high-quality genome of *X. argentea* to help in exploring the phylogenetic relationships of the subfamily Xenocyprinae, as well as to provide a basis for the genetic and evolutionary studies, germplasm resource protection, development and utilization of this fish.

In this study, we constructed a chromosome-level genome of *X. argentea* using 61.85 Gb of Illumina short-reads (66×), 38.37 Gb of PacBio HiFi long-reads (41×) and 181.19 Gb of Hi-C reads (192×). The final genome size was 984.96 Mb, and 979.72 Mb of which was anchored to 24 chromosomes. This genome contains 79 contigs, 18 gap-free chromosomes, 56.88% (560.27 Mb) of repeat elements and 28,533 protein-coding genes. The contig N50, scaffold N50, BUSCO completeness, short-reads mapping rate and quality value (QV) score

[1]Yangtze River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Wuhan, 430223, China. [2]Key Laboratory of Aquatic Genomics, Ministry of Agriculture and Rural Affairs, Yangtze River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Wuhan, 430223, China. ✉e-mail: lianghw@yfi.ac.cn

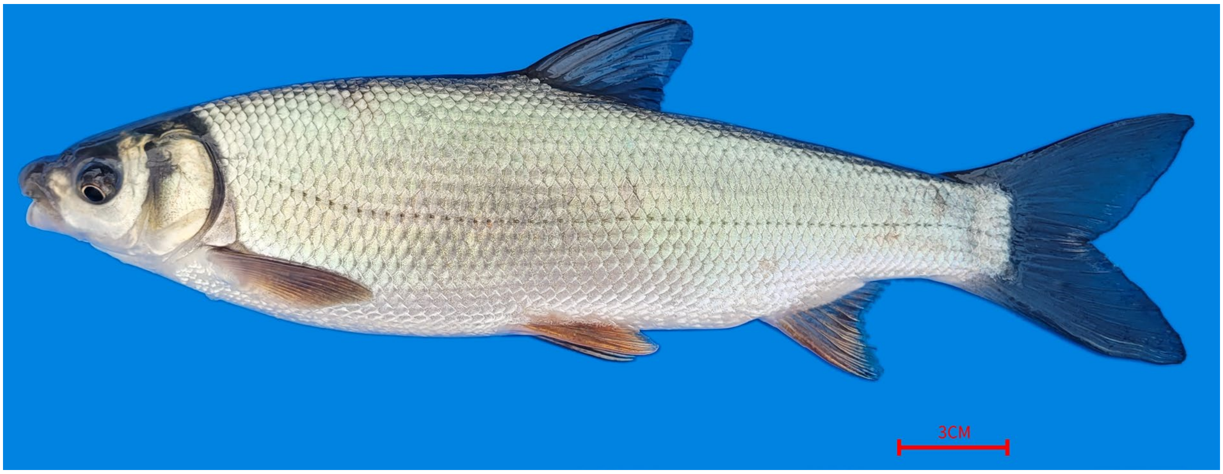

Fig. 1 The morphological characters of the *Xenocypris argentea* used for genome sequencing.

Library types	Sample	Platform	Bases (Gb)	Reads Count	Mean length (bp)	N50 (bp)
SMRT Bell	muscle	PacBio Sequel II (HiFi)	38.38	1,811,106	21,191	21,128
Hi-C	muscle	Illumina Novaseq 6000	181.19	1,207,999,640	150	150
Short-read	muscle	Illumina Novaseq 6000	61.85	412,396,578	150	150
RNA-seq	mix	Illumina Novaseq 6000	11.55	77031496	150	150

Table 1. Sequencing data and related information used in the assembly and annotation of *Xenocypris argentea* genome. Note: Sample used for RNA-seq are mix of the muscle, heart, liver, brain, gill and spleen tissues.

were estimated to be 36.02 Mb, 38.12 Mb, 97.2%, 99.89% and 52.22 respectively, indicating the high quality of our assembly. These data and results will not only contribute to the genetic basis exploration and genetic conservation of *X. argentea*, but also facilitate phylogenetic and evolutionary research within the subfamily Xenocyprinae.

Methods

Sampling and sequencing.

A healthy male *X. argentea* (body weight: 103.60 g, Fig. 1) collected from the original breeding farm in Liling, Zhuzhou City, Hunan Province, China, was used for genome sequencing. The fish was anesthetized with MS222, and the muscle, brain, heart, spleen, gill and liver tissues were sampled and immediately frozen in liquid nitrogen, followed by storage at −80 °C for further use. Genomic DNA was extracted from muscle using a modified cetyltrimethyl ammonium bromide (CTAB) method[16]. According to the standard protocol of PacBio, a PCR-free SMART-bell library was constructed after shearing the genomic DNA into 15–20 Kb fragments, followed by being sequenced on the PacBio Sequel II platform in CCS mode to produce HiFi long reads. A short-read library with 350 bp insert size was prepared following the manufacturer's instructions of Nextera DNA Flex Library Prep Kit (Illumina, San Diego, CA, USA). For Hi-C sequencing, the muscle tissue was first cross-linked and digested with *Dpn II* restriction enzyme. After biotinylating 5′-overhang and blunt-end ligation, the DNA was sheared into 300–700 bp fragments to obtain a Hi-C library[17]. The total RNAs of all sampled tissues were isolated separately with Omega Bio-tek's E.Z.N.A.®Total RNA Kit I (R6834, Omega, USA). For RNA-seq, the qualified RNA from each tissue (2 μg) was equally pooled and a library was constructed using NEBNext® Ultra™ RNA Library Prep Kit (#E7530L, NEB, USA). Finally, all the constructed short-read, Hi-C and RNA-seq libraries were sequenced with PE-150 paired-end strategy on an Illumina Novaseq 6000 platform of Wuhan Benagen Technology Co., Ltd (Wuhan, China). As a result, the Illumina-based sequencing obtained 61.85 Gb of raw short reads, 181.19 Gb of Hi-C reads and 11.55 Gb of RNA-seq data (Table 1).

Genome size estimation and assembly.

For genome size estimation, the raw short reads were first filtered with FastQC (v 0.20.1) to remove sequencing adaptors and low-quality reads. Based on the obtained 60.50 Gb clean data, the 19-mer frequency depth distribution was constructed with Jellyfish (v 2.2.10)[18], and the genome size was estimated using Jellyfish and Genomescope (v 2.0)[19]. As a result, the genome size of *X. argentea* was estimated to be 949.39 Mb, with a heterozygous ratio of 0.81% (Fig. 2a).

Before *de novo* assembly, the PacBio sequencing data was filtered to remove low-quality polymerase reads, and SMRTLink (v 8.0, parameter:–min-passes = 3–min-rq = 0.99) was used to process the remaining subreads, resulting in 38.38 Gb of HiFi reads with a mean read length of 21.19 Kb (Table 1). The generated HiFi long reads were subsequently subjected to HiFiasm (v 0.16.1)[20] with default parameters, and a contig-level genome with a total length of 984.96 Mb and contig N50 of 36.02 Mb was obtained, which contained 79 contigs (Table 2).

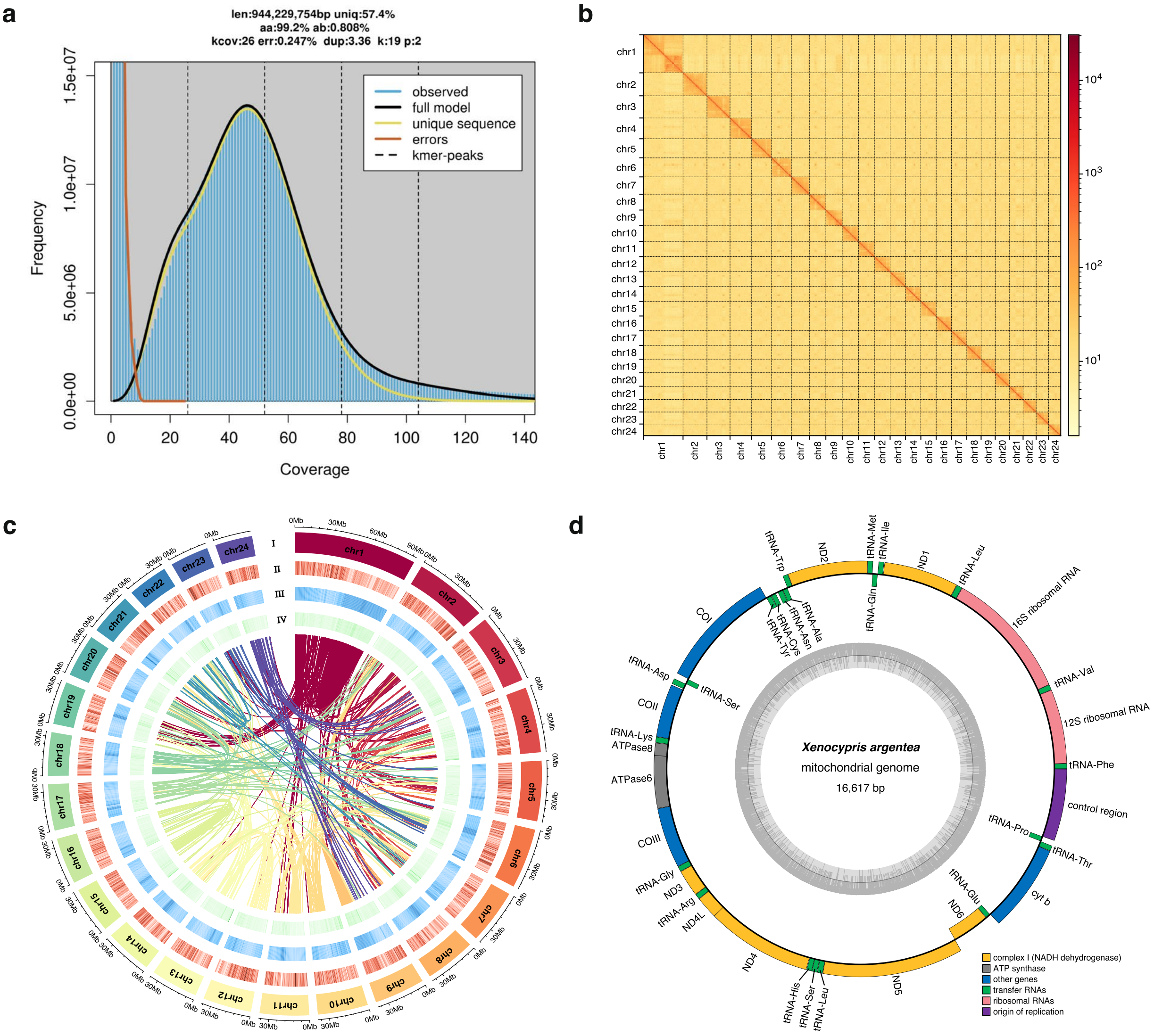

Fig. 2 The characteristics of *X. argentea* genome. (**a**) The K-mer (k = 19) distribution and genome size estimation of *X. argentea* genome. (**b**) The genome-wide chromosomal interactions heatmap based on Hi-C data. (**c**) Circos ideogram of *X. argentea* genome, tracks from outer to inner represent the 24 chromosomes assembly (I); gene frequency (II), repeat elements density (III), GC content (IV) and links of intragenomic syntenic blocks within 100Kbp sliding windows. (**d**) The distribution of annotated genes in mitochondrial genome, the inner ring shows the GC content, and the circle inside marks the 50% threshold.

	Contig	Scaffold
Total length (bp)	984,965,711	984,966,611
GC content (%)	37.55	37.55
Number	79	70
N50 (bp)	36,020,545	38,125,623
N90 (bp)	26,695,896	32,373,224
Average length (bp)	12,467,920	14,070,951.59
Maximum length (bp)	71,544,306	92,977,404

Table 2. Genome assembly statistics for *X. argentea*.

The size of this genome is slightly larger than our genome survey result, but falls within the range of previously reported genome sizes of 2 related species, *P. simoni* (940.9 Mb)[15] and *P. microlepis* (1004.34 Mb)[16].

To gain a chromosome-level genome, the raw Hi-C data was filtered using fastp (version 0.21.0), and the retained 178.40 Gb clean reads were aligned to the preliminary assembly with HICUP (v 0.8.0)[21] to obtain valid interaction pairs. Because the chromosome number of *X. argentea* was reported to be 2n = 48 in previous

Chr ID	Length (bp)	Contig number	Chr ID	Length (bp)	Contig number
Chr1	92,977,404	4	Chr13	36,094,866	3
Chr2	56,128,976	2	Chr14	36,041,041	1
Chr3	54,192,188	1	Chr15	36,020,545	1
Chr4	50,756,658	1	Chr16	35,984,277	1
Chr5	46,736,115	1	Chr17	35,871,634	1
Chr6	46,491,836	1	Chr18	33,617,906	1
Chr7	42,255,504	1	Chr19	33,164,918	1
Chr8	38,711,419	2	Chr20	33,023,601	1
Chr9	38,465,267	2	Chr21	32,373,224	1
Chr10	38,125,623	1	Chr22	30,648,589	1
Chr11	37,366,885	2	Chr23	29,363,522	1
Chr12	36,943,989	1	Chr24	28,371,514	1

Table 3. Statistics of the 24 anchored chromosomes in *X. argentea* genome.

	TE protiens		De novo + repbase		Combined TEs	
Type	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome
DNA	17,808,194	1.81	278,233,420	28.25	279,993,046	28.43
LINE	16,228,820	1.65	29,102,031	2.95	31,096,927	3.16
SINE	0	0	2,906,674	0.3	2,906,674	0.3
LTR	24,324,286	2.47	120,328,518	12.22	122,134,038	12.4
Satellite	0	0	7,085,265	0.72	7,085,265	0.72
Simple repeat	0	0	1,203,366	0.12	1,203,366	0.12
Other	0	0	5,135	0	5,135	0
Unknown	3,747	0	140,716,717	14.29	140,720,464	14.29
Total	58,346,591	5.92	541,808,753	55.01	560,270,208	56.88

Table 4. Summary of the repetitive elements in the *X. argentea* genome.

karyotype study[22], we used ALLHiC (v 0.9.8)[23], 3D-DNA (v 1.8.0419)[24] and Juicer (v 1.6)[25] to assemble the contigs onto 24 chromosomes according to the Hi-C chromatin interaction data. The anchored chromosomes were further reviewed manually using Juicebox (v 1.11.08)[26]. Ultimately, 99.47% of the contigs were anchored to 24 chromosomes, with a scaffold N50 of 38.12 Mb (Fig. 2b,c, Tables 2, 3). In the final genome assembly, the size of each chromosome ranged from 28.37 Mb to 92.97 Mb, and 18 chromosomes were gap-free.

In addition, we assembled and annotated the mitochondrial genome of *X. argentea* using MitoZ (v 3.6) and Getorganelle (v 1.7.1a), and a circular mitochondrial genome (16,617 bp in size) with 13 unique protein-coding genes, 22 tRNAs and 2 rRNAs was obtained (Fig. 2d).

Repeat sequences annotation. Repeat and LTR sequences in the genome of *X. argentea* were *de novo* predicted using RepeatModeler (v 1.0.11, parameters: BuildDatabase -name mydb ;RepeatModeler -database mydb -pa 10)[27] and LTR_finder (Official release of LTR_FINDER_parallel, parameters: -threads 16 -harvest_out -size 1000000 -time 300). After being deduplicated with LTR_retriever (v 2.9.0, parameter: -threads 16), the predicted LTR sequences were merged with detected repeat elements and RepBase library (v 20181026) to obtain the repeat library. RepeatMasker (v 4.0.9, parameters: -nolow -no_is -norna -parallel 2)[28] and RepeatProteinMask (v 4.0.9, parameters: -noLowSimple -pvalue 0.0001) were subsequently used to search repetitive elements against the repeat library and identify TE_protein class repeat sequences. After merging and deduplicating all the predicted results, we obtained a total of 560.27 Mb repeat sequences, which occupied 56.88% of our assembled genome (Table 4, Fig. 3a).

Structural and functional annotation of protein coding genes. For the structural annotation of protein-coding genes, a method integrating transcript-assisted prediction, homolog evidence-based prediction and *ab initio* prediction was applied. The RNA sequencing data was filtered with fastp (v 0.21.0, parameter: -j out.json -h out.html), aligned onto the genome with Hisat2 (v 2.1.0)[29] and assembled with Stringtie (v 2.1.4)[30]. The obtained transcripts were subsequently subjected to TransDecoder (v 5.1.0) to predict coding regions in *X. argentea* genome. In the homolog-based prediction, the protein sequences of *Danio rerio* (GCF_000002035.6), P. *microlepis* (GCA_040144785.1), *P. simoni*[14], *Megalobrama amblycephala* (GCF_018812025.1) and *Hypophthalmichthys molitrix* (GCA_041475455.1) were aligned to the assembled genome using tblastn (v 2.7.1, parameter: -t 16 -q 7), and Exonerate (v 2.4.0, parameter: -model protein2genome -showtargetgff 1)[31] was used to predict transcripts and coding regions. For the *ab initio* prediction, Augustus (v3.3.2, parameter: -uniqueGeneId = true -noInFrameStop = true -gff3 = on -strand = both)[32], Genscan (v1.0, parameter: HumanIso.smat)[33] and GlimmerHMM (v3.0.4, parameter: -f -g)[34] were used based on the repeat masked-genome. Ultimately, all the

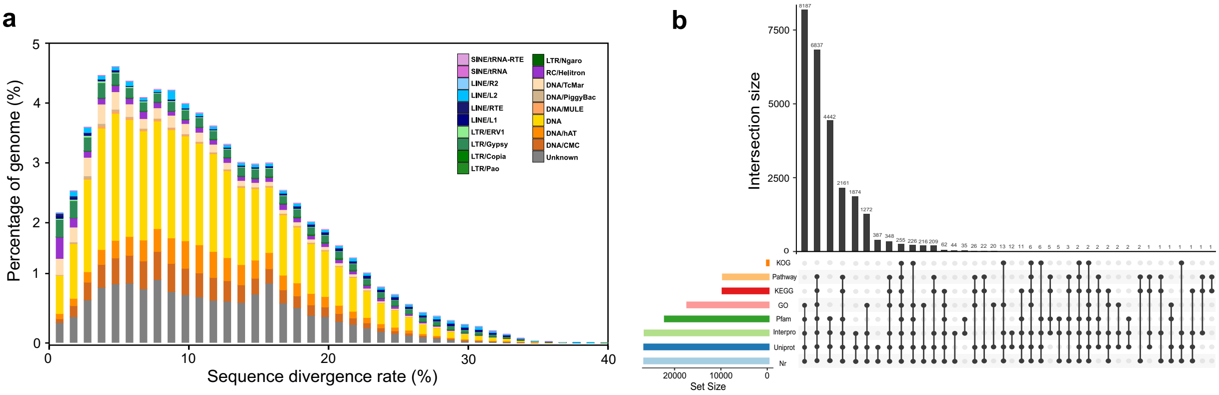

Fig. 3 Distribution of the predicted repeat elements and functionally annotated genes in *X. argentea* genome. (**a**) The transposon activity distribution in the genome. (**b**) The upset bar plot visualizing the functional annotation results based on different databases.

Method	Gene set	Gene number	Average gene length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length (bp)	Average intron length (bp)
Ab initio	*Genscan*	35,117	18,578.82	1,540.77	7.83	196.69	2,493.26
	AUGUSTUS	10,750	59,670.37	4,223.91	24.62	171.58	2,347.69
Homology-based	*P. simoni*	29,400	14,263.66	873.37	4.9	178.23	3,433.29
	H. molitrix	65,910	23,539.48	937.77	4.89	191.76	5,809.71
	P. microlepis	75,608	18,561.90	803.62	4.13	194.73	5,679.29
	D. rerio	54,685	27,465.48	1,200.81	6.13	195.97	5,122.25
	M. amblycephala	62,185	24,775.84	1,143.91	5.84	195.74	4,878.65
RNAseq		21,649	20,030.40	1,518.11	10.04	373.04	1,800.35
Integration		25,035	24,132.15	1,996.31	10.97	258.11	2,136.74
Final set		28,533	19,661.53	1,639.41	9.57	310.71	1,946.78

Table 5. Statistic of the predicted protein coding genes by different methods.

Item	Annotated number of putative genes	Percentage
KEGG	9,960	34.91%
Pathway	9,882	34.63%
Nr	26,648	93.39%
Uniprot	26,633	93.34%
GO	17,475	61.24%
KOG	511	1.79%
Pfam	22,244	77.96%
Interpro	26,560	93.09%
Annotated	27,284	95.62%
Predicted genes	28,533	100.00%

Table 6. Summary of the functionally annotated genes in the genome of *X. argentea*.

predicted results were combined and processed using MAKER (v 2.31.10, parameter: maker_exe.ctl maker_opts.ctl maker_bopts.ctl -ignore_nfs_tmp -fix_nucleotides)[35], and a high-confidence gene set including 28,533 protein-coding genes was generated. The average exon number, exon length and CDS length in each gene were 9.57, 310.71 bp and 1639.41 bp, respectively (Table 5).

The predicted protein models were functionally annotated by aligning them against the UniProt, NR and KEGG databases using Diamond (v 2.0.11.149, parameter: --evalue 1e-5)[36] and KOBAS (v 3.0)[37]. Motifs, domains and conservative sequences of proteins were also annotated using Hmmscan (v 3.3.2, parameter: -E 0.01) and InterProScan (v 5.52–86, parameters: -goterms -pa -dp -verbose -cpu 20)[38]. In summary, 95.62% (27,284) of the predicted genes were functionally annotated in public databases (Fig. 3b, Table 6).

In addition, we annotated the non-coding RNAs using RNAmmer (v 1.2, parameters: -S euk -m tsu,lsu,ssu)[39], tRNAscan-SE (v 2.0.12, parameters: -E -j tRNA.gff -o tRNA.result -f tRNA.struct -thread 16)[40] and INFERNAL

Type		Copy	Average length(bp)	Total length(bp)	% of genome
miRNA		1,683	72	120,594	0.0122
tRNA		14,772	76	1,116,021	0.1133
rRNA	rRNA	5,954	175	1,042,919	0.1059
	18S	52	1892	98,410	0.01
	28S	50	4977	248860	0.0253
	5.8S	512	120	61,299	0.0062
	5S	5,340	119	634,350	0.0644
snRNA	snRNA	1,805	149	268,316	0.0272
	CD-box	320	163	52,075	0.0053
	HACA-box	78	151	11,801	0.0012
	splicing	1,382	146	201,246	0.0204
	scaRNA	10	238	2,381	0.0002

Table 7. Statistic of the identified non-coding RNAs in the genome.

	Genome		Annotation	
Type	Number	Percentage (%)	Proteins	Percentage (%)
Complete BUSCOs (C)	3,537	97.2	3,399	93.4
Single-Copy BUSCOs (S)	3,479	95.6	3,327	91.4
Duplicated BUSCOs (D)	58	1.6	72	2
Fragmented BUSCOs (F)	13	0.4	79	2.2
Missing BUSCOs (M)	90	2.4	162	4.4
Total BUSCOs	3,640	100	3,640	100
Shore-reads mapping rate	—	99.89	—	—
QV	52.22	—	—	—

Table 8. The BUSCO integrity, consistency and consensus quality value (QV) evaluation of the genome and annotation.

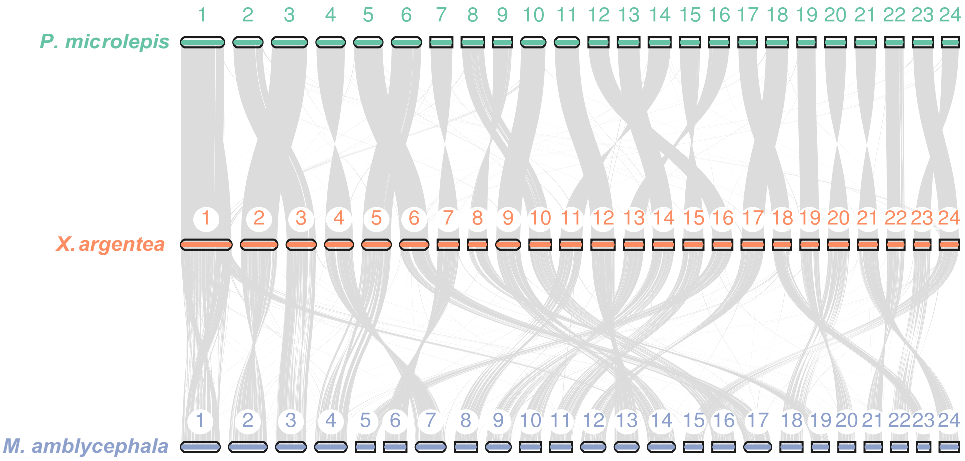

Fig. 4 Synteny analysis of the chromosomal genomes among *X. argentea* and two related species.

(v 1.1.4, parameter: -cut_ga -rfam -nohmmonly -fmt 2)[41]. As a result, 1,683 miRNAs, 14,772 tRNAs, 5,954 rRNAs and 1,805 snRNAs were identified in the genome (Table 7).

Data Records

All the raw sequencing data have been submitted to Sequence Read Archive (SRA) database with accession numbers of SRR28431798[42], SRR28431799[43], SRR28431800[44] and SRR28431801[45]. The chromosome-level genome and mitochondrial genome are available in Genbank database with the accession number GCA_046562865.1[46] and PQ824064.1[47]. And the genome annotation results have been deposited in *figshare* database[48].

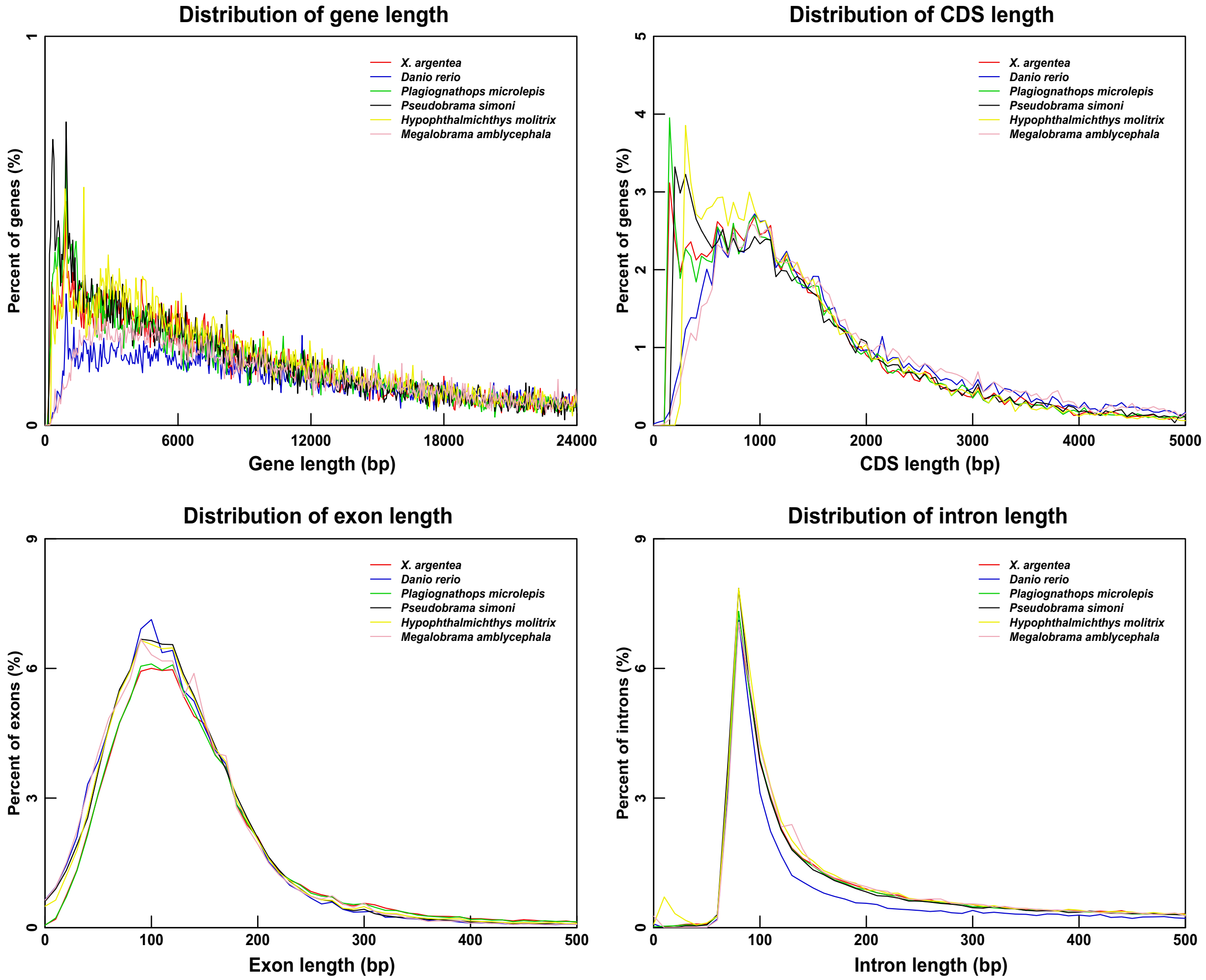

Fig. 5 Comparison of the distribution characteristics for the lengths of gene (**a**), CDS (**b**), exon (**c**) and intron (**d**) in the assembled *X. argentea* genome and 5 related species.

Technical validation

Evaluation of the genome assembly.

The integrity, consistency and consensus quality value (QV) of the assembled genome were evaluated by BUSCO (v 5.3.0, parameter: -m prot -c 40 -long -f), BWA (v 0.7.17)[49] and Merqury (v 1.3)[50], respectively. Results showed that 97.2% of complete BUSCOs were found in the Actinopterygii_odb10 database, and the mapping ratio of Illumina short-reads to the assembly and QV of the genome reached 99.89% and 52.22 respectively (Table 8). These indicators reflected high completeness and good quality of our assembly.

For the quality assessment of chromosome assembly, strong interactive signals were first found alongside the diagonals of Hi-C heatmap and no obvious noise in other areas (Fig. 2b). Besides, 18 chromosomes in the assembly contained only one contig for each chromosome. Moreover, chromosomal collinearity analysis with Last (v1.17.0)[51] and JCVI (v0.9.13)[52] revealed a strong consistency across the genome of *X. argentea*, *P. microlepis* and *M. amblycephala* (Fig. 4). All these results supported the precision of our chromosome assembly.

Quality assessment of genome annotation.

The quality of genome annotation was also assessed using BUSCO analysis, and 93.4% of complete BUSCOs and 2.2% fragmented BUSCOs were identified (Table 8). In addition, the length distributions of different gene elements in the *X. argentea* genome were compared with 5 related species (*D. rerio*, *P. microlepis*, *P. simoni*, *H. molitrix* and *M. amblycephala*), and the results showed high similarity, indicating the reliability of our genome annotation (Fig. 5).

Code availability

The bioinformatic analyses were conducted following the manual and protocols of the corresponding software. No specific codes were developed in this work. The versions and main parameters of the software have been described in the Method section. Default parameters were used for those without detailed information.

Received: 25 December 2024; Accepted: 26 March 2025;
Published online: 04 April 2025

References

1. He, G. *et al.* Present research situation of biology and genetic diversity of *Xenocypris argentea*. *Jiangxi Fishery Science and Technology* **3**, 43–44 (2013).
2. Zhao, L., Peng, X. & Guo, X. Screening of microsatellite markers in *Xenocypris argentea* using transcriptome sequencing. *Freshwater Fisheries* **48**, 23–29 (2018).
3. Liu, J., Zhao, L., Liu, Q. & Zhang, H. Genetic variation of *Xenocypris argentea* between different populations based on mitochondrial COI gene. *Freshwater Fisheries* **45**, 3–8 (2015).
4. Peng, X., Zhao, L., Liu, J. & Guo, X. Development of SNP markers for *Xenocypris argentea* based on transcriptomics. *Conservation Genetics Resources* **10**, 679–684 (2018).
5. Li, P., Liu, J., Lu, W., Sun, S. & Wang, J. Age, growth, reproduction and mortality of *Xenocypris argentea* (Günther, 1868) in the lower reaches of the Tangwang River, China. *PeerJ* **12**, e16673 (2024).
6. Xiang, C. *et al.* Biology and breeding techniques of *Xenocypris argentea*. *Inland Fisheries* **5**, 18–19 (2003).
7. Gu, J. & Zhang, W. Preliminary exploration of artificial breeding technology for *Xenocypris argentea* seedlings. *Aquaculture* **8**, 36–37 (2019).
8. Xiao, W. & Zhang, Y. Mitochondrial DNA diversity in populations of *Xenocypris argentea* as revealed by restriction analysis. *Acta Hydrobiologica Sinica* **24**, 1–10 (2000).
9. Hu, Y., Yang, S., Li, M., Cao, W. & Liu, H. Population differentiation of *Xenocypris argentea* in Poyang lake and adjacent drainages. *Sichuan Journal of Zoology* **31**, 696–703 (2012).
10. Peng, X., Liu, J., Zhao, L. & Guo, X. The transcriptome sequencing and functional analysis of liver tissue of *Xenocypris argentea*. *Genomics and Applied Biology* **39**, 1471–1477 (2020).
11. Li, L. *et al.* Molecular systematics of Xenocyprinae (Cypriniformes, Cyprinidae). *Acta Hydrobiologica Sinica* **47**, 628–636 (2023).
12. Xiao, W., Zhang, Y. & Liu, H. Molecular systematics of Xenocyprinae (Teleostei: Cyprinidae): taxonomy, biogeography, and coevolution of a special group restricted in East Asia. *Molecular Phylogenetics and Evolutionis* **18**, 163–173 (2001).
13. Chen P. Phylogenetic pattern and macroevolutionary characteristics of the East Asian endemic group of Cyprinids. *Wuhan: Institute of Hydrobiology, Chinese Academy of Sciences*, 46–49 (2015).
14. Fan, G. *et al.* Genomic data of *Pseudobrama simoni*. *GigaScience* https://doi.org/10.5524/102191 (2020).
15. Wu, Y. *et al.* Chromosome-level genome assembly of *Plagiognathops microlepis* based on PacBio HiFi and Hi-C sequencing. *Scientific Data* **11**, 802 (2024).
16. Shan, G., Jin, W., Lam, E. & Xing, X. Purification of total DNA extracted from activated sludge. *Journal of Environmental Sciences* **20**, 80–87 (2008).
17. Zeng, Q. *et al.* Chromosome-level haplotype-resolved genome assembly for *Takifugu ocellatus* using PacBio and Hi-C technologies. *Scientific Data* **10**, 22 (2023).
18. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
19. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* **11**, 1432 (2020).
20. Cheng, H. Y. *et al.* Haplotype-resolved assembly of diploid genomes without parental data. *Nature Biotechnology* **40**, 1332–1335 (2022).
21. Wingett, S. *et al.* HiCUP: Pipeline for mapping and processing Hi-C data. *F1000Research* **4**, 1310 (2015).
22. Li, K., Li, Y., Zhou, M. & Zhou, D. Studies on the karyotypes of Chinese Cyprinid fishes II. karyotypes of four species of Xenocyprininae. *ACTA Zoologica Sinica* **29**, 207–213 (1983).
23. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nature Plants* **5**, 833–845 (2019).
24. Dudchenko, O. *et al. De novo* assembly of the genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
25. Durand, N. C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* **3**, 95–98 (2016).
26. Durand, N. C. *et al.* Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems* **3**, 99–101 (2016).
27. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *PNAS* **117**, 9451–9457 (2020).
28. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocals in Bioinformatics* **25**, 4.10.11–14.10.14 (2009).
29. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**, 907–915 (2019).
30. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology* **20**, 1–13 (2019).
31. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 1–11 (2005).
32. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* **24**, 637–644 (2008).
33. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268**, 78–94 (1997).
34. Delcher, A. L., Bratke, K. A., Powers, E. C. & Salzberg, S. L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–679 (2007).
35. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 1–14 (2011).
36. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59–60 (2015).
37. Xie, C. *et al.* KOBAS 2.0: A web server for annotation and identification of enriched pathways and diseases. *Nucleic acids research* **39**, W316–W322 (2011).
38. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research* **49**, D344–D354 (2021).
39. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic acids research* **35**, 3100–3108 (2007).
40. Chan, P., Lin, B., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research* **49**, 9077–9096 (2021).
41. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
42. *NCBI Sequence Read Archive.* https://identifiers.org/ncbi/insdc.sra:SRR28431798 (2024).
43. *NCBI Sequence Read Archive.* https://identifiers.org/ncbi/insdc.sra:SRR28431799 (2024).
44. *NCBI Sequence Read Archive.* https://identifiers.org/ncbi/insdc.sra:SRR28431800 (2024).
45. *NCBI Sequence Read Archive.* https://identifiers.org/ncbi/insdc.sra:SRR28431801 (2024).
46. *NCBI Genbank.* http://identifiers.org/assembly:GCA_046562865.1 (2024).
47. *NCBI GenBank.* https://identifiers.org/ncbi/insdc:PQ824064.1 (2024).
48. Wu, Y. The genome annotation of *Xenocypris argentea*. *Figshare.* https://doi.org/10.6084/m9.figshare.25000685.v2 (2024).
49. Manni, M., Berkeley, M. R., Seppey, M., Simao, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution* **38**, 4647–4654 (2021).

50. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, 1–27 (2020).
51. Frith, M. C., Hamada, M. & Horton, P. Parameters for accurate genome alignment. *BMC Bioinformatics* **11**, 1–14 (2010).
52. Tang, H. B. *et al*. Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).

Acknowledgements

This work was financially supported by Central Public-interest Scientific Institution Basal Research Fund, CAFS (NO. 2023TD36), the earmarked fund for CARS (CARS-45), National Freshwater Aquatic Germplasm Resource Center (FGRC18537). We would also like to thank Dr. Zhixiong Zhou in Tianjin Agricultural University for his kindly helps in data analysis.

Author contributions

H.L. conceived and supervised this study. H.S. collected the samples and extracted the genomic DNA. Y.W. designed the experiment, performed data analysis and drafted the manuscript. H.L. revised this manuscript. All authors have read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.L.

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.