scientific data



OPEN

DATA DESCRIPTOR

# Chromosome-scale assembly of the *Xenocypris davidi* using PacBio HiFi reads and Hi-C technologies

Tiezhu Yang[1,2,5], Liangjie Zhao[1,2,5], Chaoqun Su[1,2], Xusheng Guo[1,2,3] ✉, Xinliang Peng[1,2] ✉, Shijie Yang[4] & Gaoyou Yao[1,2]

***Xenocypris davidi*** **is a benthic fish species widely distributed in the water systems south of the Yellow River in China, playing a significant role in aquatic ecosystems. Despite its ecological and economic importance, genomic resources for *X. davidi* are limited, hindering a comprehensive understanding of its evolutionary adaptations and genetic improvements. This study presents the first chromosome-level genome assembly of *X. davidi*, utilizing PacBio long-reads, Illumina short reads, and Hi-C sequencing data. The genome assembly spans 1.05 Gb with a scaffold N50 length of 33.99 Mb, and 95.12% of the genome sequence was successfully anchored onto 24 pseudochromosomes. We identified 27,360 protein-coding genes, of which 26,672 were functionally annotated. This genome sequence provides a valuable resource for exploring the molecular basis of agronomic traits in *X. davidi* and will facilitate its genetic enhancement.**

## Background & Summary

*Xenocypris davidi* Bleeker, 1871, a species of fish extensively distributed throughout the water systems south of the Yellow River in China, is commonly found in rivers, lakes, and reservoirs, and is classified as a benthic species. It primarily feeds on bottom-dwelling algae and detritus. These fish typically inhabit the mid-to-upper reaches of rivers and migrate upstream to shallow rapids for spawning during the breeding season[1]. Owing to its identity as a benthic scraper, it actively contributes to the enhancement of aquatic ecological environments and the amelioration of water quality, rendering it an ideal candidate for artificial propagation and release, as well as for pond polyculture practices. Research on the dietary habits of *X. davidi* in a certain reservoir has revealed that adult fish primarily consume a substantial amount of detritus, supplemented by a small number of sessile diatoms and oscillatorian cyanobacteria. This diet does not compete with that of other economically important fish species such as silver carp and bighead carp in the reservoir, indicating that *X. davidi* is a preferred species for integrated aquaculture[2]. Studies by Tang *et al.* have demonstrated that *X. davidi* has significant potential in managing the overgrowth of filamentous algae[3]. Appreciated for its delicate texture and savory taste, *X. davidi* has become one of the specialty fish species favored by consumers. Recent research has also explored the dietary nutritional requirements of *X. davidi* broodstock and the impact of adding bee pollen and glutamine to their feed on the growth of this species[4–6]. In the field of toxicology, it has been discovered that *X. davidi* exhibits a low tolerance to heavy metal copper ions. A hepatic transcriptome database for *X. davidi* has been constructed to analyze gene expression profiles, providing molecular insights into the species' response to environmental pollutants[7].

*X. davidi*, classified within the order Cypriniformes, family Cyprinidae, subfamily Xenocyprininae, and genus Xenocypris, has been the subject of evolutionary studies primarily based on the analysis of the complete mitochondrial genome[8,9]. Assessments in the field of population genetics, utilizing microsatellite markers, have evaluated the impact of stocking activities on wild populations of *X. davidi* in the Qiantang River, with results indicating that these activities have not posed genetic risks to the wild populations[10]. Additionally, research has identified genetic diversity variations among different aquaculture populations of *X. davidi*[11]. When analyzing

[1]College of Fisheries, Xinyang Agriculture and Forestry University, Xinyang, 464000, China. [2]Fishery Biological Engineering Technology Research Center of Henan Province, Xinyang, 464000, China. [3]Key Laboratory of Fishery Ecosystem Regulation, Xinyang, 464000, China. [4]Xinyang Nanwan Reservoir Affairs Center, Xinyang, 464000, China. [5]These authors contributed equally: Tiezhu Yang, Liangjie Zhao. ✉e-mail: gxs1968@xyafu.edu.cn; 2008210024@xyafu.edu.cn

| Kmer | Depth | N Kmer | Genome size (M) | Revised genome size (M) | Heterozygous rate (%) | Repeat_rate (%) |
|---|---|---|---|---|---|---|
| 17 | 37 | 37,775,268,357 | 1,020.95 | 1,004.32 | 0.52 | 53.91 |

**Table 1.** Kmer = 17 Analysis of genomic characteristic statistics.

| Title | Total_length | Total_number | Max_length | N50_length | N90_length |
|---|---|---|---|---|---|
| Contig | 876,892,788 | 1,361,752 | 79,659 | 2,166 | 183 |
| Scaffold | 899,312,128 | 1,111,026 | 104,922 | 3,222 | 258 |

**Table 2.** Statistical summary of the genome survey assembly results for *X. davidi*.

wild populations from various water systems, significant genetic divergence was observed between populations in Qiandao Lake and the Yangtze River[12]. Nonetheless, the scant genomic data resources limit the understanding of *X. davidi*'s evolutionary adaptation and molecular mechanisms of its traits, restricting our full appreciation and exploitation.

In this research, we have achieved the construction of a chromosome-level genome assembly for *X. davidi*, representing the initial case of jointly employing PacBio long-reads, Illumina short reads, and Hi-C sequencing data. The obtained genome assembly has a total length of 1.05 Gb. The scaffold N50 length reaches 33.99 Mb, and a remarkable 95.12% of the genome sequence is effectively anchored to 24 pseudochromosomes. By using a combined method including de novo gene predictions, RNA-seq data, and homologous protein evidence, a sum of 27,360 protein-coding genes have been identified, among which 26,672 have been functionally annotated. The genome sequence is a valuable asset for understanding *X. davidi*'s agronomic trait molecular basis and facilitating its genetic improvement.

## Methods

**Ethics statement.** The Experimental Animal Care and Ethics Committee of Xinyang Agriculture and Forestry University approved all fish sampling experiment procedures.

**Samling and genome survey.** A healthy female specimen of *X. davidi*, weighing 659.36 grams, was collected from the Pohe Reservoir in Guangshan County, Xinyang City, Henan Province, China. Following euthanasia with eugenol at an anesthetic concentration of 16 mg L^-3[13], the fish was rapidly rinsed three times with sterile physiological saline and dabbed dry with sterile cotton. Tissues including those from the muscle, heart, spleen, liver, and kidney were promptly dissected and promptly immersed in liquid nitrogen for conservation, and then stored at -80 °C until the DNA extraction process. High-quality genomic DNA from the muscle was extracted with the PureLink™ Genomic DNA Mini Kit (K182001, Thermo Fisher Scientific, USA). Meanwhile, RNA from diverse tissues was isolated by TRIzol™ Reagent (15596026CN, Thermo Fisher Scientific, USA). The quality and concentration of DNA were evaluated via 1% agarose gel electrophoresis and a Qubit 2.0 Fluorometer (Invitrogen, Thermo Fisher Scientific, USA). In contrast, for RNA, its purity and integrity were further appraised using a NanoDropTM One spectrophotometer (Thermo Fisher Scientific, USA) and Agilent 2100 Bioanalyzer (Agilent Technologies, Inc., USA).

In the context of the genomic survey, a 10 μg DNA sample was utilized to construct a 350 bp library, employing a paired-end 150 bp (PE150) sequencing strategy on the Illumina Novaseq 6000 platform. A total of 50.51 Gb of raw data and 168,353,928 read pairs were generated. The Jellyfish[14] (version 2.2.7) was used to construct the 17-mer frequency depth distribution and estimate the genome size. Genome assembly was performed using SOAPdenovo2[15]. Ultimately, survey analysis conducted with a Kmer size of 17 estimated the genome size to be 1,020.95 Mbp, which was refined to 1,004.32 Mbp. The heterozygosity rate was determined to be 0.52%, with the proportion of repetitive sequences constituting 53.91% (Table 1). Assembly with a Kmer size of 41 yielded a contig N50 of 2,166 bp, with a total length of 876,892,788 bp, and a scaffold N50 of 3,222 bp, with a total length of 899,312,128 bp. (Table 2).

**PacBio and Hi-C based whole-genome sequencing.** Regarding PacBio sequencing, high-quality DNA samples (with the main band being greater than 30 kb) were fragmented into segments ranging from 15 to 18 kb by utilizing a Covaris ultrasonic disruptor. Afterwards, the large DNA fragments were enriched and purified through magnetic beads. Subsequently, the fragmented DNA underwent damage repair as well as end repair. Hairpin sequencing adapters were then ligated to both ends of the DNA fragments, and exonucleases were used to eliminate any fragments that had unsuccessful ligation. The properly prepared library was then sequenced on the PacBio Sequel IIe platform in the CCS mode. After filtering polymerase reads, raw subreads were acquired, and these were further processed using SMARTLink (version 11.0, parameters: filter_min_qv = 20) to generate HiFi reads. Ultimately, a total of 27.86 Gb HiFi reads were obtained, along with a read number of 1,908,298 and a mean read length of 14,601 bp (as shown in Table 3 and Table 4).

In the case of Hi-C sequencing, muscle tissue was processed with paraformaldehyde to stabilize the intracellular DNA conformation. After cell lysis, the crosslinked DNA was digested by the restriction enzyme MboI to create sticky ends. Subsequently, the DNA termini were biotinylated, and DNA ligase was used to connect the DNA fragments. Thereafter, proteases were applied to reverse the crosslinking of DNA. The purified DNA was then randomly fragmented into pieces ranging from 300 to 500 bp and utilized to construct a Hi-C library[16,17].

| Library types | Sample | Platform | Bases (Gb) | Reads Count | Mean Length (bp) | N50 (bp) |
|---|---|---|---|---|---|---|
| SMRT Bell | muscle | PacBio Sequel IIe (HiFi) | 27.86 | 1,908,298 | 14,601 | 15,148 |
| Hi-C | muscle | Illumina Novaseq 6000 | 49.22 | 344,604,444 | 150 | 150 |
| Short-read | muscle | Illumina Novaseq 6000 | 50.51 | 168,353,928 | 150 | 150 |
| RNA-seq | muscle | Illumina Novaseq 6000 | 7.01 | 23,352,108 | 150 | 150 |
| RNA-seq | heart | Illumina Novaseq 6000 | 6.33 | 21,090,726 | 150 | 150 |
| RNA-seq | spleen | Illumina Novaseq 6000 | 6.48 | 21,598,661 | 150 | 150 |
| RNA-seq | liver | Illumina Novaseq 6000 | 6.99 | 23,304,023 | 150 | 150 |
| RNA-seq | kidney | Illumina Novaseq 6000 | 7.22 | 24,060,364 | 150 | 150 |

**Table 3.** Statics of different types of sequencing reads.

| Type | HiFi reads |
|---|---|
| Read_base (bp) | 27,863,295,623 |
| Read_Number | 1,908,298 |
| Read_length(max) (bp) | 56,349 |
| Read_length(mean) (bp) | 14,601 |
| Read_length(N50) (bp) | 15,148 |

**Table 4.** A statistical analysis of the sequencing data obtained from Hi-Fi.

| Type | Counts |
|---|---|
| Total Reads Pairs | 5,564,221 |
| Total Paired (mapped) | 2,754,693 |
| Total Paired Ratio (%) | 49.51 |
| Valid Pairs | 2,334,774 |
| Unique di-Tags | 2,237,596 |
| Effect Rate (%) | 40.21 |

**Table 5.** Statistical analysis of sequencing data from Hi-C.

| Type | Contig |
|---|---|
| Total length (bp) | 1,048,230,699 |
| Total number | 134 |
| Average length (bp) | 7,822,617 |
| Max length (bp) | 57,517,301 |
| Min length (bp) | 17,878 |
| N50 length (bp) | 33,985,674 |
| N50 number | 12 |
| N90 length (bp) | 12,284,600 |
| N90 number | 32 |

**Table 6.** Statistics for assembly at the contig level.

Hi-C sequencing was carried out on the Illumina Novaseq 6000 platform following a paired-end 150 bp (PE-150) sequencing strategy. The raw sequence data obtained were processed through the HiCUP[18] (v 0.8.3), which includes hicup_truncater for identifying and mapping chimeric sequences, hicup_filter for filtering mapped reads, and hicup_deduplicator for removing duplicate contacts. After conducting the aforementioned quality assessments on the Hi-C data, a total of 5,564,221 Total Reads Pairs were obtained, with 2,754,693 Total Paired (mapped) reads (accounting for 49.51%), 2,334,774 Valid Pairs, and 2,237,596 Unique di-Tags (Table 5).

**Genome assembly.** The genomic assembly of *X. davidi* was facilitated by employing the default settings within the Hifiasm software[19] (v 0.16.1-r375). This assembly approach commences from the uncollapsed genomic data, thereby maximizing the retention of haplotype information. Hifiasm was provisioned with HiFi long reads to produce a monoploid assembly and a pair of contig graphs that resolve haplotypes. Consequently, the assembly process resulted in the construction of 134 contigs with a combined length of 1.05 Gb. This genomic assembly size is marginally larger than what was initially anticipated based on the survey results. The average length, maximum contig size, and N50 were 7.8, 57.52, and 33.99 Mb (Table 6), respectively.

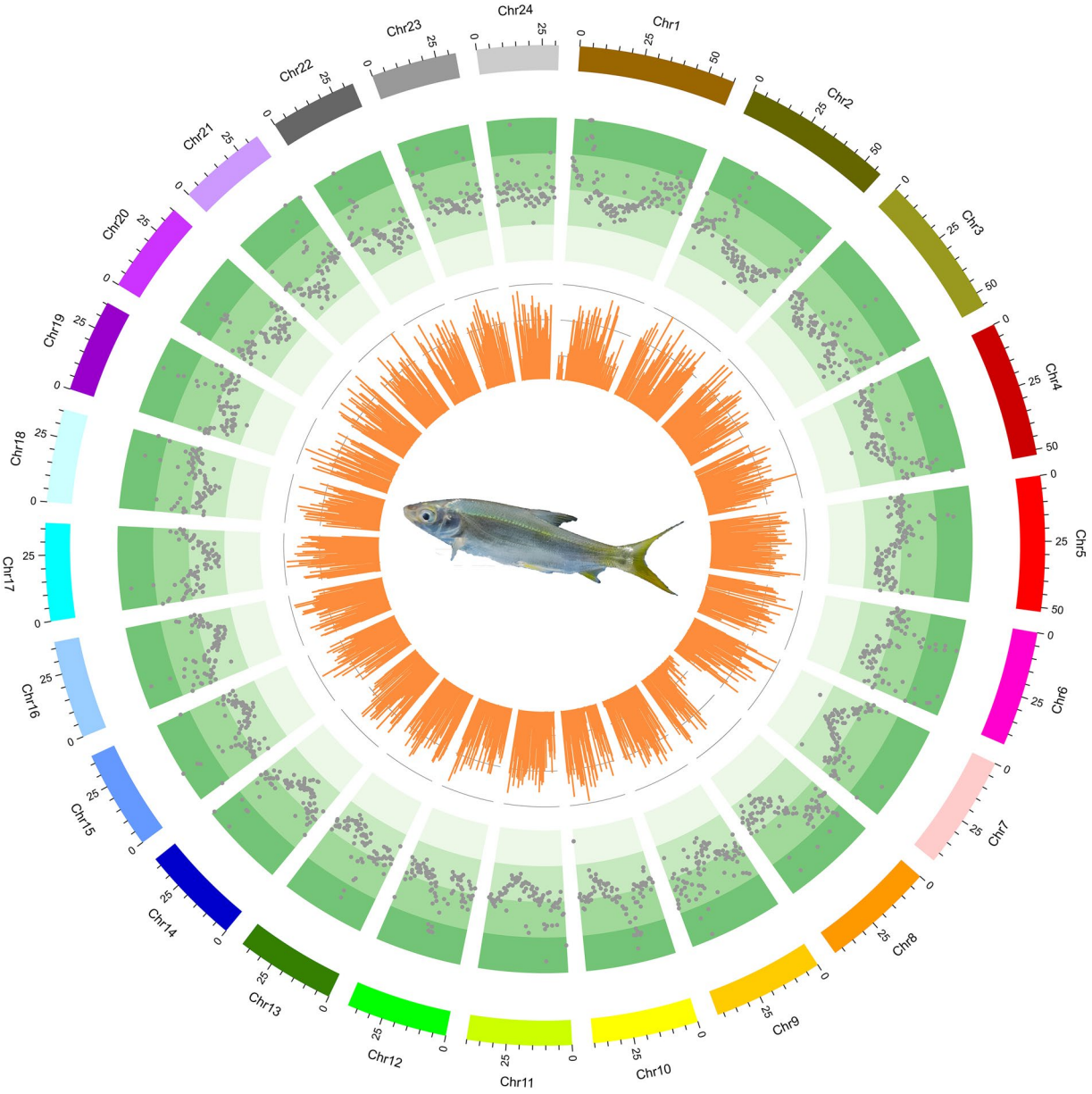


**Fig. 1** A circus plot of 24 chromosomes of *X. davidi*. The tracks from outside to inside are: 24 chromosomes, the distributions of transposable element and bar plot for gene density profile.

**Hi-C assembly and Chromosome anchoring.** The Hi-C technology was employed to differentiate contigs or scaffolds into distinct chromosomes based on the higher probability of intra-chromosomal interactions compared to inter-chromosomal interactions. Additionally, it facilitated the ordering and orientation of contigs or scaffolds on the same chromosome, as the interaction probability decreases with increasing interaction distance along the chromosome. After the Hi-C corrected contigs were integrated into the ALLHiC pipeline[20] for pruning, partition, rescue, optimization, and building, a substantial 95.12% of the assembled sequences were anchored to 24 pseudochromosomes[21] (Fig. 1), with chromosome lengths varying from 31.29 Mb to 60.28 Mb (Table 7). This outcome aligns with the karyotype results derived from cytological observations[22], which are consistent with the chromosome numbers of 2n = 48 observed in several Xenocyprididae fish species, such as *Plagiognathops microlepis*[23], *Chanodichthys erythropterus*[24], *M. amblycephala*[25], and *Ctenopharyngodon idella*[26]. Moreover, we manually refined the Hi-C scaffolding based on the chromatin contact matrix within Juicebox[27] (Fig. 2). The 24 pseudochromosomes can be clearly identified on the heatmap, and there is a strong signal intensity around the diagonal, which suggests the high quality of the genome assembly. After Hi-C correction, the final assembled genome had a total span of 1.05 Gb, along with a scaffold N50 of 40.13 Mb (Table 8).

**Genome annotation.** In the annotation of repetitive sequences within the genome of *X. davidi*, this study employed two complementary approaches: homology-based alignment[28] and *de novo* prediction[29]. The

| Chromosome | Cluster Number | Sequeues Length (bp) |
|---|---|---|
| Chr1 | 2 | 60,275,181 |
| Chr2 | 2 | 57,625,182 |
| Chr3 | 1 | 55,586,653 |
| Chr4 | 1 | 52,022,816 |
| Chr5 | 6 | 51,677,919 |
| Chr6 | 4 | 44,187,633 |
| Chr7 | 4 | 43,435,225 |
| Chr8 | 2 | 43,302,729 |
| Chr9 | 2 | 42,370,478 |
| Chr10 | 3 | 40,516,972 |
| Chr11 | 2 | 40,131,624 |
| Chr12 | 4 | 38,782,997 |
| Chr13 | 2 | 38,256,806 |
| Chr14 | 2 | 38,096,289 |
| Chr15 | 2 | 38,035,212 |
| Chr16 | 3 | 37,432,084 |
| Chr17 | 2 | 37,282,341 |
| Chr18 | 3 | 35,654,014 |
| Chr19 | 2 | 35,628,516 |
| Chr20 | 3 | 35,330,893 |
| Chr21 | 3 | 34,045,325 |
| Chr22 | 1 | 33,397,843 |
| Chr23 | 3 | 32,692,433 |
| Chr24 | 2 | 31,287,097 |

**Table 7.** Statistics of the 24 anchored chromosomes of *X. davidi* genome.

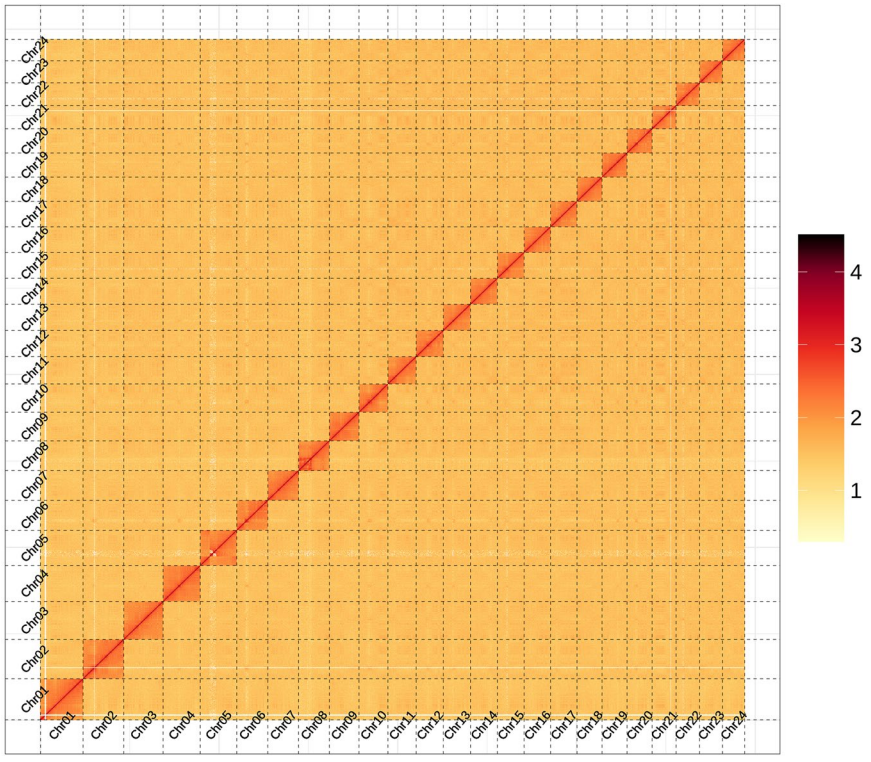


**Fig. 2** Hi-C chromatin interaction heatmap of the *X. davidi* assembly.

homology-based alignment method was predicated on the RepBase database[30] and leveraged two software tools, RepeatMasker[31] (version 4.1.0, parameters: -a -nolow -no_is -norna -parallel 4) and RepeatProteinMask (version 4.1.0, parameters: -noLowSimple -pvalue 0.0001 -engine ncbi), to identify sequences with similarity to known

| Sample ID | Contig length | Scaffold length | Contig number | Scaffold number |
|---|---|---|---|---|
| Total | 1,048,230,699 | 1,048,234,399 | 142 | 105 |
| Max | 57,517,301 | 60,275,181 | — | — |
| Number > = 2000 | — | — | 142 | 105 |
| N50 | 31,018,029 | 40,131,624 | 13 | 11 |
| N60 | 27,366,945 | 38,096,289 | 17 | 14 |
| N70 | 24,045,337 | 37,282,341 | 21 | 17 |
| N80 | 13,754,553 | 35,330,893 | 26 | 20 |
| N90 | 10,369,524 | 32,692,433 | 34 | 23 |

**Table 8.** Assembly statistics for Hi-C.

| Type | Denovo + Repbase | | TE Proteins | | Combined TEs | |
|---|---|---|---|---|---|---|
| | Length(bp) | % in Genome | Length(bp) | % in Genome | Length(bp) | % in Genome |
| DNA | 39,993,470 | 3.82 | 15,440,641 | 1.47 | 47,841,916 | 4.56 |
| LINE | 7,002,432 | 0.67 | 14,991,786 | 1.43 | 19,012,975 | 1.81 |
| SINE | 172,417 | 0.02 | 0 | 0 | 172,417 | 0.02 |
| LTR | 475,525,141 | 45.36 | 23,810,950 | 2.27 | 477,886,955 | 45.59 |
| Unknown | 13,215,147 | 1.26 | 0 | 0 | 13,215,147 | 1.26 |
| Total | 527,762,304 | 50.35 | 54,235,314 | 5.17 | 533,336,825 | 50.88 |

**Table 9.** Summary of the transposable elements in *X. davidi* genome.

repetitive elements. Conversely, the de novo prediction approach commenced with the establishment of a *de novo* repetitive sequence library utilizing LTR_FINDER[32] (version 1.06, parameters: -C -w 2), RepeatScout (version 1.0.5), and RepeatModeler[33] (version 2.0.1, parameters: -engine ncbi -pa 15), followed by the application of RepeatMasker for prediction. Additionally, within the de novo prediction methodology, the software TRF[34] (version 4.09, parameters: 2 7 7 80 10 50 2000 -d -h -ngs) was employed to detect tandem repeats within the *X. davidi* genome. Ultimately, all predicted results were consolidated and duplicates were eliminated, yielding the identification of 533.34 Mb of repetitive sequences, constituting 50.88% of the assembled genome. The predominant element among these repetitive sequences was long terminal repeats (LTR), which accounted for 45.59% (477.87 Mb) of the assembled genome (Table 9), a significant departure from the genome of the fine-scaled gudgeon, where DNA transposons were the most abundant, comprising 31.55%[23]. Long interspersed nuclear elements (LINE) constituted 1.81% of the genome, short interspersed nuclear elements (SINE) constituted 0.02% of the genome, and DNA elements constituted 4.56% of the genome, respectively (Table 9).

For the prediction of gene structures within the genome of *X. davidi*, this manuscript employed a triad of methodologies: de novo prediction, homology-based prediction, and annotation using transcriptome data[35]. The *de novo* prediction outcomes were derived from the utilization of Augustus[36] (version 3.2.3, parameters: --species = pasa1--uniqueGeneId = TRUE --noInFrameStop = TRUE --GFF3 = on --genemodel = complete --strand = both), GlimmerHMM[37] (version 3.0.4, parameters: -d pasa1 -f -g), SNAP (version 2013.11.29, parameters: -gff pasa1.hmm), Geneid (version 1.4, parameters: -P homo_sapiens.param -v -G -p geneid), and Genscan (version 1.0, parameters: HumanIso.smat) software. In the homology-based prediction approach, protein sequences from *Carassius auratus*[38] *(GenBank: GCA_003368295.1)*, *Cyprinus carpio*[39] *(GenBank: GCA_000951615.2)*, *Ctenopharyngodon idellus*[26] *(GenBank: GCA_019924925.1)*, *Danio rerio*[40] *(GenBank: GCA_000002035.4)*, *Onychostoma macrolepis*[41] *(GenBank: GCA_012432095.1)*, *M. amblycephala*[25] *(GenBank: GCA_018812025.1)*, and *Opsariichthys bidens*[42] *(GenBank: GCA_037194245.1)* were downloaded from the NCBI database and used to predict gene structures within the genome through alignment software such as Blastall[43] (version 2.2.26, parameters: -e 1e-05 -F T -m 8), Solar (version 0.9.6, parameters: -a prot2genome2 -z -f m8), and Genewise[44] (version 2.4.1, parameters: -tfor -genesf -gff -sum) (Fig. 3). For the transcriptome data annotation method, high-quality RNA from muscle, heart, spleen, liver, and kidney were used to construct RNAseq libraries. Subsequently, these libraries were sequenced on the Illumina Novaseq 6000 platform, and 150 bp paired-end reads were obtained as a result. Post-sequencing, 37.73 Gb of raw data was generated, which was filtered to yield 34.03 Gb of clean data (Table 3). Subsequent de novo assembly was performed using Trinity (version 2.1.1, parameters:–normalize_reads–full_cleanup–min_glue 2–min_kmer_cov 2–KMER_SIZE 25), alignment analysis with Hisat2 (version 2.0.4), and assembly annotation with Stringtie (version 1.3.3). The gene sets predicted by the above-mentioned three methods were combined into a non-redundant gene set with the help of EvidenceModeler (EVW)[45] (version 1.1.1, parameters:–segmentSize 200000 –overlapSize 20000–min_intron_length 20). Finally, PASA (http://pasa.sourceforge.net/) was utilized, in conjunction with transcriptome assembly results, to refine the EVW annotation, incorporating UTR and alternative splicing information, to arrive at the final gene set. The genomic prediction for *X. davidi* resulted in 27,360 genes, with an average transcript length of 10,053.32 bp, an average coding region length of 1,121.54 bp, an average of 6.28 introns per gene, an average intron length of 178.57 bp, and an average exon length of 1,691.46 bp (Table 10, Fig. 4A).

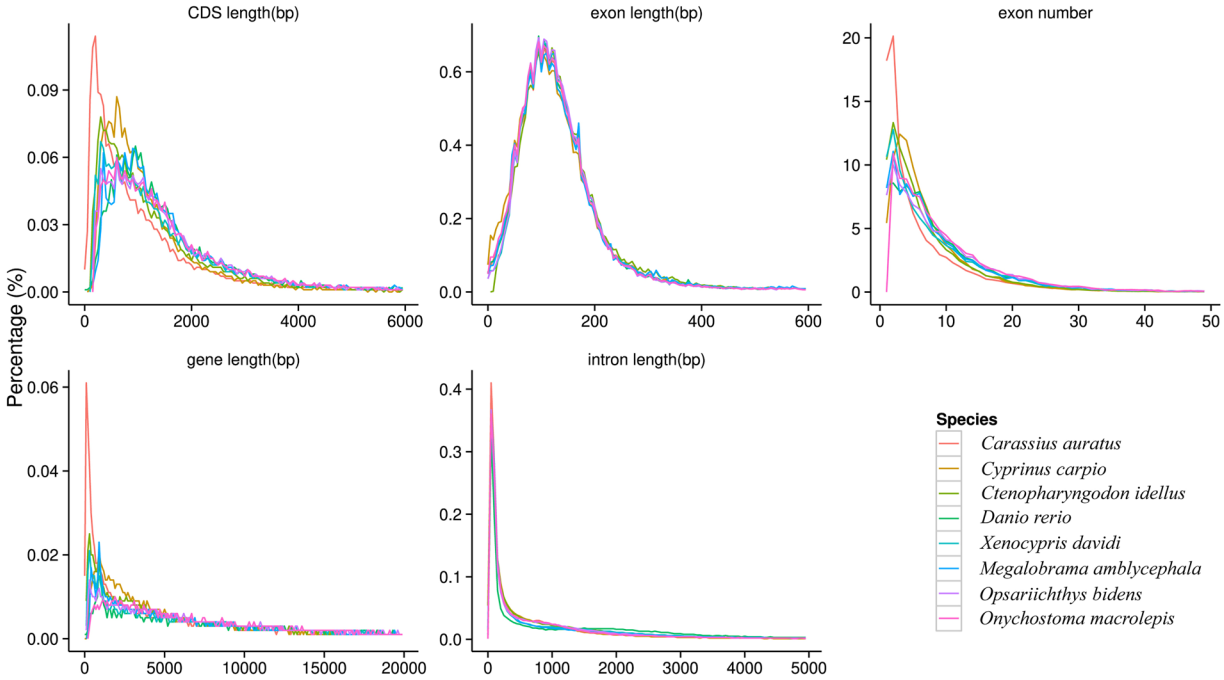


**Fig. 3** Comparisons of the genomic elements of closely related species.

| | Gene set | Number | Average transcript length(bp) | Average CDS length(bp) | Average exons per gene | Average exon length(bp) | Average intron length(bp) |
|---|---|---|---|---|---|---|---|
| *De novo* | Augustus | 41,765 | 10,053.32 | 1,121.54 | 6.28 | 178.57 | 1,691.46 |
| | GlimmerHMM | 106,335 | 8,754.46 | 573.66 | 3.89 | 147.46 | 2,830.35 |
| | SNAP | 59,634 | 12,134.09 | 673.63 | 4.87 | 138.4 | 2,963.41 |
| | Geneid | 32,058 | 19,612.36 | 1,344.29 | 6.3 | 213.33 | 3,445.94 |
| | Genscan | 31,763 | 22,650.39 | 1,559.14 | 8.21 | 189.9 | 2,925.14 |
| Homolog | *Ctenopharyngodon idellus* | 29,826 | 9,959.04 | 1,251.03 | 6.87 | 182.16 | 1,484.06 |
| | *Opsariichthys bidens* | 26,232 | 13,465.66 | 1,449.43 | 7.96 | 182.16 | 1,727.23 |
| | *Onychostoma macrolepis* | 24,228 | 14,824.04 | 1,520.60 | 8.58 | 177.18 | 1,754.54 |
| | *Megalobrama amblycephala* | 24,907 | 14,497.76 | 1,576.73 | 8.68 | 181.66 | 1,682.52 |
| | *Carassius auratus* | 29,056 | 10,456.97 | 1,301.02 | 7.02 | 185.33 | 1,520.97 |
| | *Cyprinus carpio* | 26,439 | 10,919.42 | 1,317.09 | 7.12 | 184.91 | 1,568.31 |
| | *Danio rerio* | 23,752 | 14,024.94 | 1,547.84 | 8.39 | 184.57 | 1,689.21 |
| RNAseq | PASA | 26,100 | 15,024.86 | 1,453.96 | 8.96 | 162.26 | 1,704.73 |
| | Transcripts | 33,355 | 26,488.96 | 2,770.83 | 9.87 | 280.77 | 2,674.37 |
| EVM | | 38,936 | 12,542.04 | 1,232.01 | 6.93 | 177.83 | 1,907.85 |
| Pasa-update* | | 38,720 | 12,882.32 | 1,250.40 | 7.03 | 177.78 | 1,927.93 |
| Final set* | | 27,360 | 16,345.93 | 1,566.90 | 8.87 | 176.69 | 1,878.35 |

**Table 10.** Statistical analyses of the gene structure annotation of the *X. davidi*.

To conduct the functional annotation of protein-coding genes, this study employed a dual strategy utilizing Blastp[46] (version 2.2.26, parameters: -max_target_seqs. 1 -evalue 1e-4) and Diamond[47] (version 0.8.22, parameters:–more-sensitive -k 10 -e 1e-5 -f 6 qseqid qlen qstart qend sseqid slen sstart send pident ppos qcovhsp bitscore evalue–salltitles–threads 10) for aligning the protein-coding genes against the SwissProt[48], NCBI Non-redundant protein (NR) (https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/), KEGG[49], InterPro[50], Gene Ontology (GO)[51], and Pfam[52] protein databases. The identification of protein domains and motifs was facilitated through the application of InterProScan[53] (version 5.35–74.0, parameters: -cpu 20 -format tsv -appl ProDom, SMART, ProSiteProfiles, PRINTS, Pfam, Panther -iprlookup -dp -goterms). Ultimately, 26,672 (97.50%) of the 27,360 predicted genes received annotations from at least one of the databases (Table 11). Among the functionally annotated proteins, 20,599 (75.29%) were corroborated by annotations from all four databases (Fig. 4B).

The annotation of non-coding RNAs encompasses tRNA, rRNA, miRNA, and snRNA. To identify tRNA sequences within the genome, the structural characteristics of tRNA were leveraged using the tRNAscan-SE

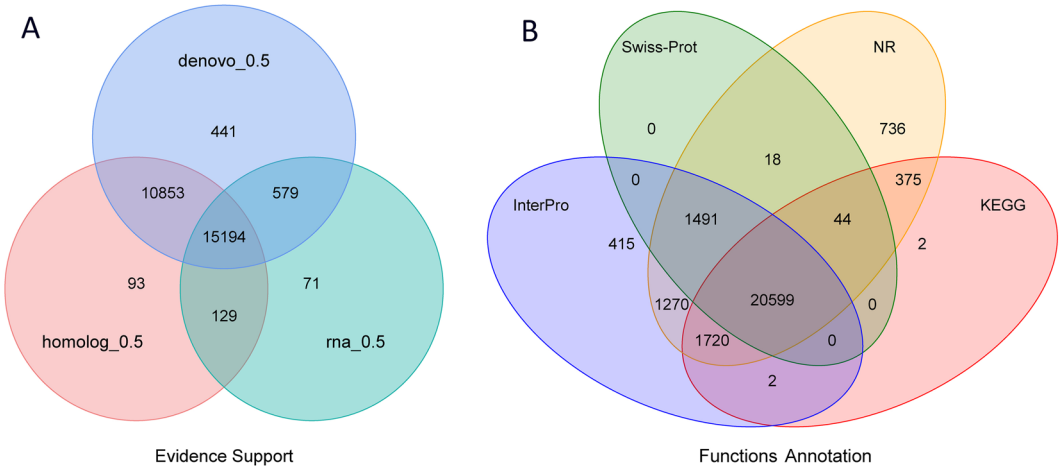


**Fig. 4** Gene prediction and functional annotation of the *X. davidi* genome. (**A**) Venn diagram of the gene set prediction. (**B**) Venn diagram of functional annotation based on different databases.

| | Type | Copy number | Average length(bp) | Total length(bp) | % of genome |
|---|---|---|---|---|---|
| miRNA | | 2,009 | 105.11 | 211,175 | 0.020146 |
| tRNA | | 11,325 | 76.03 | 861,093 | 0.082147 |
| rRNA | rRNA | 13,213 | 147.93 | 1,954,630 | 0.19 |
| | 18S | 322 | 732.25 | 235,783 | 0.022493 |
| | 28S | 790 | 387.85 | 306,403 | 0.02923 |
| | 5.8S | 116 | 156.56 | 18,161 | 0.001733 |
| | 5S | 11,985 | 116.34 | 1,394,283 | 0.13 |
| snRNA | snRNA | 2,412 | 160.5 | 387,133 | 0.036932 |
| | CD-box | 249 | 152.23 | 37,906 | 0.003616 |
| | HACA-box | 92 | 148.27 | 13,641 | 0.001301 |
| | splicing | 2,031 | 162.25 | 329,527 | 0.031436 |
| | scaRNA | 26 | 194.35 | 5,053 | 0.000482 |
| | Unknown | 14 | 71.86 | 1,006 | 0.000096 |

**Table 11.** Non-coding RNA statistical results of *X. davidi*.

| | Number | Percent (%) |
|---|---|---|
| Total | 27,360 | — |
| Swissprot | 22,152 | 81 |
| Nr | 26,253 | 96 |
| KEGG | 22,742 | 83.1 |
| InterPro | 25,497 | 93.2 |
| GO | 16,878 | 61.7 |
| Pfam | 20,899 | 76.4 |
| Annotated | 26,672 | 97.5 |
| Unannotated | 688 | 2.5 |

**Table 12.** Statistical analysis of the functional gene annotations of the *X. davidi* genome.

software[54] (version 1.4). Given the high conservation of rRNA, the rRNA sequences of closely related species were chosen as reference sequences. Moreover, the identification of rRNA within the genome was made easier through BLAST alignment (v 2.2.26, parameters: -e 1e-10 -v 10000 -b 10000). Furthermore, employing the covariance models from the Rfam family, the INFERNAL software, which is integrated within the Rfam[55] (v 14.1) suite, can be utilized to predict microRNA (miRNA) and small nuclear RNA (snRNA) sequence information on the *X. davidi* genome. Ultimately, four types of non-coding RNAs were identified from the X. davidi genome, including 2,009 miRNAs, 11,325 tRNAs, 13,213 rRNAs, and 2,412 snRNAs (Table 12).

| Type | Number | Percentage (%) |
|---|---|---|
| Complete BUSCOs (C) | 3,578 | 98.3 |
| Single-copy BUSCOs (S) | 3,502 | 96.2 |
| Duplicated BUSCOs (D) | 76 | 2.1 |
| Fragmented BUSCOs (F) | 25 | 0.7 |
| Missing BUSCOs (M) | 36 | 1.0 |
| Total BUSCOs | 3,640 | — |
| Short-reads mapping rate | — | 99.36 |
| Genome covered by reads | — | 99.96 |
| Quality value (QV) | 50.458 | — |

**Table 13.** Completeness and accuracy evaluation of the genome.

## Data Records

The raw sequence data of RNAseq data, HiC data, PacBio data and Illumina short reads data reported in this paper have been deposited in the Genome Sequence Archive in National Genomics Data Center, China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences (GSA: CRA020814[56], CRA020817[57], CRA020818[58], CRA020819[59], CRA020820[60], CRA020821[61], CRA020822[62], CRA020823[63]). The whole genome sequence data reported in this paper have been deposited in the GenBank (JBLRZY000000000)[64] and figshare database[65]. The genome annotation files will already be uploaded and shared publicly in the figshare database[66].

## Technical Validation

**Evaluation of the genome assembly and annotation.** The quality assessment of the genome assembly and annotation was conducted with meticulous attention to detail. The completeness of the assembled genome was assessed by utilizing BUSCO[67] (version 5.4.3) with the actinopterygii_odb10 database, yielding a 98.3% complete BUSCO score within the assembled genomes (Table 11), which is a testament to the high degree of completeness of our genomic assemblies. The genomic consistency was further appraised by aligning Illumina short-reads to the assembled genomes with BWA[68] (version 0.7.17), resulting in exceptionally high mapping rates (99.36%) and coverage (99.96%) against the assembled genomes (Table 13). Employing Merqury[69] (version 1.4.1), the consensus quality value (QV), indicative of per-base consensus accuracy, was calculated to be 50.458 for the assembled genomes. Additionally, a comparative analysis of the length distributions of genes, coding sequences (CDSs), introns, and exons across the genomes of *C. idellus*, *O. bidens*, *O. macrolepis*, *M. amblycephala*, *C. auratus*, *C. carpio*, and *D. rerio* was performed, revealing similarities (Fig. 3), which substantiates the reliability of our genome annotation. Collectively, the outcomes from these four methodologies demonstrate the high accuracy and completeness of the final genome assembly.

## Code availability

In this study, no custom-written codes were used. All data processing operations were conducted in accordance with the manuals and protocols of the relevant software. The specific parameters for different software and tools were described in the Methods section. For cases where detailed parameters were not specified, default parameters were adopted.

Received: 18 December 2024; Accepted: 11 March 2025;
Published online: 18 March 2025

## References

1. Ding, D. Introduction to Four Species of Culter Fish. *Hunan Agricuture* **2**, 30+24 (2013).
2. Xu, D. A Preliminary Analysis on the Food Habie of Xenocypris davidi Bleeker in Reservoir Guangting. *Acta Hydrobiologica Sinica* **01**, 43–53 (1988).
3. Tang, Y. *et al.* Grazing Effects of Xenocypris davidi Bleeker (Cyprinidae, Cypriniformes) on Filamentous Algae and the Consequent Effects on Intestinal Microbiota. *Aquaculture Research* **2023**, 1–14 (2023).
4. Li, C. *et al.* Research on the Dietary Protein and Fat Requirements of Parental Xenocypris davidi. *Jiangsu Agricultural Sciences* **47**, 220–223 (2019).
5. Li, C. *et al.* Effects of Different Levels of Bee Pollen in Feed on Reproductive Performance of Xenocypris davidi Bleeker. *Agricultural Science and Technology* **20**, 48–52 (2019).
6. Wang, Y. *et al.* Effects of dietary glutamine supplementation on growth performance, intestinal digestive ability, antioxidant status and hepatic lipid accumulation in Xenocypris davidi (Bleeker,1871). *Aquacult Int* **32**, 725–743 (2024).
7. Peng, X., Zhao, L., Liu, J., Guo, X. & Ding, Y. Comparative transcriptome analyses of the liver between Xenocypris microlepis and Xenocypris davidi under low copper exposure. *Aquatic Toxicology* **236**, 105850 (2021).
8. Xu, H., Zhu, Y., Zheng, D. & Yang, S. Molecular identification and phylogenetic analysis of mitogenome of the Xenocypris davidi from Caoè. *Mitochondrial DNA Part B Resources* **4**, 3998–3999 (2019).
9. Liu, Y. The complete mitochondrial genome sequence of *Xenocypris davidi* (Bleeker). *Mitochondrial DNA* **25**, 374–376 (2014).
10. Guo, A. *et al.* Stock enhancement effect and potential genetic risks of Xenocypris davidi by molecular markers in the upper reaches of Qiantang River, China. *Journal of Fisheries of China* **46**, 2349–2356 (2022).
11. Liu, S. *et al.* Genetic Diversity Analysis of Four Cultured Xenocypris davidi Populations Based on Mitochondrial D-loop Sequences. *Guangdong Agricultural Sciences* **50**, 139–145 (2023).
12. Zhang, H., Zhao, L., Hu, Z. & Liu, Q. Genetic variation analysis of Xenocypris davidi populations from Qiandao Lake and Yangtze River. *Journal of Shanghai Ocean University* **24**, 12–19.

13. Wang, W. Study on the mechanism and protection of anaesthesia injury in Lateolabrax maculatus. (Shanghai Ocean University, Shanghai, 2020).
14. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
15. Luo, R. *et al*. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 2047-217X-1–18 (2012).
16. van Berkum, N. L. *et al*. Hi-C: A Method to Study the Three-dimensional Architecture of Genomes. *JoVE (Journal of Visualized Experiments)* e1869 (2010).
17. Rao, S. S. P. *et al*. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).
18. Wingett, S. W. *et al*. HiCUP: pipeline for mapping and processing Hi-C data. *F1000research* **4**, 1310 (2015).
19. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).
20. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
21. Hu, J. *et al*. Characteristics of diploid and triploid hybrids derived from female Megalobrama amblycephala Yih × male Xenocypris davidi Bleeker. *Aquaculture* **364–365**, 157–164 (2012).
22. Zhang, H., Xu, X., Zhang, Y. & Wang, S. Chromosomal Karyotype Analysis of Xenocypris davidi. *Jiangxi Fishery Science and Technology* **20**, 22 (2018).
23. Wu, Y., Sha, H., Luo, X., Zou, G. & Liang, H. Chromosome-level genome assembly of Plagiognathops microlepis based on PacBio HiFi and Hi-C sequencing. *Sci Data* **11**, 802 (2024).
24. Zhao, S. *et al*. A chromosome-level genome assembly of the redfin culter (Chanodichthys erythropterus). *Sci Data* **9**, 535 (2022).
25. Liu, H. *et al*. A Chromosome-Level Assembly of Blunt Snout Bream (Megalobrama amblycephala) Genome Reveals an Expansion of Olfactory Receptor Genes in Freshwater Fish. *Mol Biol Evol* **38**, 4238–4251 (2021).
26. Wu, C.-S. *et al*. Chromosome-level genome assembly of grass carp (Ctenopharyngodon idella) provides insights into its genome evolution. *BMC Genomics* **23**, 271 (2022).
27. Robinson, J. T. *et al*. Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *cels* **6**, 256–258.e1 (2018).
28. Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21**, i152–i158 (2005).
29. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
30. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* **110**, 462–467 (2005).
31. Nishimura, D. RepeatMasker. *Biotech Software & Internet Report* **1**, 36–39 (2000).
32. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**, W265–W268 (2007).
33. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457 (2020).
34. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580 (1999).
35. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654–5666 (2003).
36. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435–W439 (2006).
37. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
38. Chen, Z. *et al*. De novo assembly of the goldfish (Carassius auratus) genome and the evolution of genes after whole-genome duplication. *Science Advances* **5**, eaav0547 (2019).
39. Xu, P. *et al*. Genome sequence and genetic diversity of the common carp, Cyprinus carpio. *Nat Genet* **46**, 1212–1219 (2014).
40. Howe, K. *et al*. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498–503 (2013).
41. Sun, L. *et al*. Chromosome-level genome assembly of a cyprinid fish Onychostoma macrolepis by integration of nanopore sequencing, Bionano and Hi-C technology. *Molecular Ecology Resources* **20**, 1361–1371 (2020).
42. Xu, X. *et al*. Chromosome-Level Assembly of the Chinese Hooksnout Carp (Opsariichthys bidens) Genome Using PacBio Sequencing and Hi-C Technology. *Front. Genet*. **12** (2022).
43. Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656–664 (2002).
44. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
45. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7 (2008).
46. Altschul, S. F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402 (1997).
47. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**, 366–368 (2021).
48. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research* **28**, 45–48 (2000).
49. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27–30 (2000).
50. Hunter, S. *et al*. InterPro: the integrative protein signature database. *Nucleic Acids Research* **37**, D211–D215 (2009).
51. Ashburner, M. *et al*. Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29 (2000).
52. Mistry, J. *et al*. Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**, D412–D419 (2021).
53. Zdobnov, E. M. & Apweiler, R. InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).
54. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Research* **25**, 955–964 (1997).
55. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research* **33**, D121–D124 (2005).
56. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA020814 (2024).
57. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA020817 (2024).
58. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA020818 (2024).
59. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA020819 (2024).
60. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA020820 (2024).
61. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA020821 (2024).
62. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA020822 (2024).
63. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA020823 (2024).
64. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBLRZY000000000 (2025).
65. Yang, T. Genome annotation files of Xenocypris davidi. *figshare* https://doi.org/10.6084/m9.figshare.28287308.v1 (2025).
66. Yang, T. Genome annotation files of Xenocypris davidi. *figshare* https://doi.org/10.6084/m9.figshare.27932985.v1 (2024).

67. Manni, M., Berkeley, M. R., Seppey, M., Simão, 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).
68. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **1303**, 3997 (2013).
69. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, 245 (2020).

## Acknowledgements

The study was supported by the Youth Fund Project of Xinyang Agriculture and Forestry University (No. QN2021020), the Natural Science Foundation of Henan (No. 232300421273, No. 242300420175), the Key Scientific Research Projects of Colleges and Universities in Henan Province (No. 23B240003, No. 24B240001), and the Henan Province Science and Technology Research Project (No. 252102110075, No. 202102110263, No. 162102110053).

## Author contributions

X. G. and X. P. conceived the research project. C. S. and S. Y. collected the samples. L. Z. designed the experiment. T. Y., G. Y. and L. Z. performed data analysis. T. Y and C. S. drafted the manuscript. L. Z. and G. Y. 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 X.G. or X.P.