scientific **data**

OPEN

DATA DESCRIPTOR

A high-quality chromosome-level genome assembly of the Chinese medaka *Oryzias sinensis*

Zhongdian Dong[1,2,6], Jiangman Wang[3,6], Guozhu Chen[4], Yusong Guo[1], Na Zhao[1,5], Zhongduo Wang[1,2 ✉] & Bo Zhang [1,5 ✉]

Oryzias sinensis, also known as Chinese medaka or Chinese ricefish, is a commonly used animal model for aquatic environmental assessment in the wild as well as gene function validation or toxicology research in the lab. Here, a high-quality chromosome-level genome assembly of *O. sinensis* was generated using single-tube long fragment read (stLFR) reads, Nanopore long-reads, and Hi-C sequencing data. The genome is 796.58 Mb, and a total of 712.17 Mb of the assembled sequences were anchored to 23 pseudo-chromosomes. A final set of 22,461 genes were annotated, with 98.67% being functionally annotated. The Benchmarking Universal Single-Copy Orthologs (BUSCO) benchmark of genome assembly and gene annotation reached 95.1% (93.3% single-copy) and 94.6% (91.7% single-copy), respectively. Furthermore, we also use ATAC-seq to uncover chromosome transposase-accessibility as well as related genome area function enrichment for *Oryzias sinensis*. This study offers a new improved foundation for future genomics research in Chinese medaka.

Background & Summary

Chinese medaka, or *Oryzias sinensis*, is a teleost fish closely related to Japanese medaka (*Oryzias latipes*), which has been used as a model organism in many genomic studies as well as in aquatic toxicology research[1]. Both fish belong to the family Adrianichthyidae, commonly referred to as the Medaka family. Similar to its relative, Chinese medaka is also attracting the attention of scientists due to its small size and short generation interval[2]. There are several differences between these two fishes, including their vertebrae and pectoral fin strip; most importantly, Chinese medaka is mainly distributed in freshwater while Japanese medaka can adapt to a certain level of salinity. Because of this, Chinese medaka could be more suitable for freshwater quality evaluation.

Although the Chinese medaka genome has been released[3], the completeness and genome annotations still need to be further improved. The reported genome was only released at the scaffold level with alignments to the chromosomes of Japanese medaka. Several phylogenetic analyses and taxonomic revisions for medaka have already found that the Chinese medaka is different from its Japanese relative in chromosome constitution; the diploid chromosome number is 46 in *O. sinensis*, and 48 in *O. latipes*[4,5]. Therefore, a high-quality reference genome for *O. sinensis* is increasingly important to support future research.

In the present study, an improved high-quality chromosome-level genome assembly of *O. sinensis* was generated using single-tube long fragment read (stLFR) reads, Nanopore long-reads, and the Hi-C sequencing data. The genome size was 796.58 Mb with a scaffold N50 length of 30.38 Mb (Table 1). A total of 712.17 Mb (89.34%) of assembled sequences were anchored to 23 pseudo-chromosomes, with lengths ranging from 58.48 Mb to 18.87 Mb (Table 2). Based on this improved genome assembly, repeat elements and gene structure annotations were conducted combining *de novo* prediction, homolog-based alignment, and transcriptome-assisted methods.

[1]Key Laboratory of Aquaculture in the South China Sea for Aquatic Economic Animals of Guangdong Higher Education Institutes, College of Fishery, Guangdong Ocean University, Zhanjiang, 524088, China. [2]Guangdong Provincial Key Laboratory of Aquatic Animal Disease Control and Healthy Culture, College of Fishery, Guangdong Ocean University, Zhanjiang, 524088, China. [3]Qingdao Marine Management Support Center, Qingdao, Shandong, China. [4]National Plateau Wetland Research Center, College of Wetlands, Southwest Forestry University, Kunming, 650224, China. [5]Southern Marine Science and Engineering Guangdong Laboratory-Zhanjiang, Zhanjiang, 524000, China. [6]These authors contributed equally: Zhongdian Dong, Jiangman Wang. ✉e-mail: aduofa@hotmail.com; zb611273@163.com

Statistical level:	New chromosome-level genome			Published genome (GCA_008586565.1)	
	chromosome	scaffold	contig	scaffold	contig
Total number (>):	23	17,780	24,324	68,189	98,653
Total length of (bp):	712,173,063	797,156,861	765,499,752	813,986,518	734,065,487
N50 Length (bp):	31,744,625	30,386,322	373,585	991,358	28,180
N90 Length (bp):	24,137,528	36,705	13,816	2,913	2,405
Maximum length (bp):	58,480,650	58,480,650	2,853,130	7,585,817	287,113
Minimum length (bp):	18,868,712	714	48	747	1
GC content is (%):	40.69			40.45	

Table 1. Genome assembly statistics of Chinese medaka *Oryzias sinensis*.

Chromosome ID	Length (bp)	Percentage (%)
1	58,480,650	7.34
2	39,505,166	4.96
3	39,054,318	4.9
4	33,217,277	4.17
5	32,964,224	4.14
6	32,856,377	4.12
7	32,552,571	4.08
8	32,162,499	4.03
9	31,899,507	4
10	31,744,625	3.98
11	31,538,233	3.96
12	30,386,322	3.81
13	29,554,195	3.71
14	29,246,493	3.67
15	28,794,247	3.61
16	28,705,415	3.6
17	27,734,809	3.48
18	27,467,262	3.45
19	24,846,853	3.12
20	24,137,528	3.03
21	23,984,531	3.01
22	22,471,249	2.82
23	18,868,712	2.37
Unplaced*	84,983,798	10.66

Table 2. Chromosome length distribution of assembled genome. *Unplaced referring the sequences which did not mount to known chromosome.

	Genome		Annotation	
	Gene numbers	Percentage	Gene numbers	Percentage
Complete BUSCOs	4,356	95.1	4,335	94.6
Complete Single-Copy BUSCOs	4,275	93.3	4,202	91.7
Complete Duplicated BUSCOs	81	1.8	133	2.9
Fragmented BUSCOs	123	2.7	153	3.3
Missing BUSCOs	105	2.2	96	2.1
Total BUSCO groups searched	4,584	100	4,584	100

Table 3. Genome and Annotation BUSCO evaluation result.

Benchmarking Universal Single-Copy Orthologs (BUSCO) evaluation result showed that the final assembly was benchmarked at 95.1% and the annotation reached 94.6% (Table 3).

In addition, we conducted ATAC-seq try to find out the chromosome accessibility of *O. sinensis*. After we got the peak results we also performed function enrichment of related genome areas and obtain some clue of transcriptional activity. Taken together, this study could provide a new reliable foundation for research on the Chinese medaka, as well as for its use as a model organism.
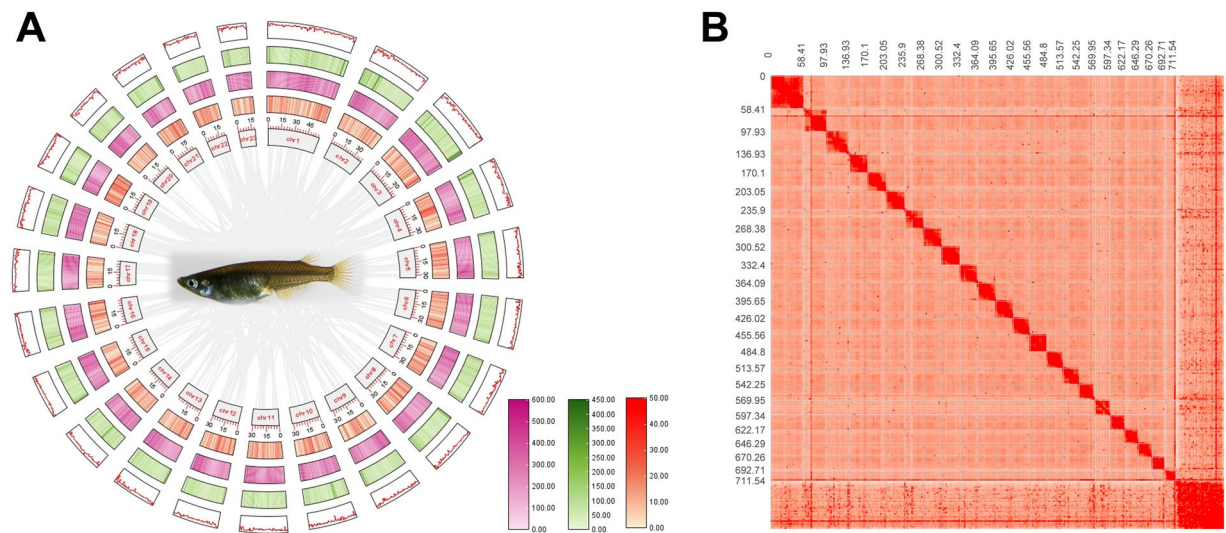
Fig. 1 (**A**) Global genome landscape of Chinese medaka. From inner to outer circles: Density of genes with 500 kbp windows, ranging from 0 to 50; Density of TE with 500 kbp windows, ranging from 0 to 600; Density of TRF with 500 kbp windows, ranging from 0 to 450; Distributions of GC content with 500 kbp windows. (**B**) Hi-C interaction heat map and scatter plot of genome assembly GC content and sequencing depth.

Methods

***De novo* genome assembly.** The samples of *O. sinensis* were obtained from the National Plateau Wetland Research Center, Southwest Forestry University; muscle tissue was used for nucleic acid extraction. High-quality purified RNA was used to construct a transcript sequencing library and DNA was used to construct stLFR, Nanopore long-reads, Hi-C sequencing, and ATAC-seq libraries. stLFR technology added the same barcode sequence to subfragments of the original long DNA molecule, allowing these co-barcoded sub-fragments to be subsequently sequenced using a second-generation platform[6]. This method generates long reads with high accuracy, enabling high-quality assembly and subsequent analysis. Most importantly, stLFR was cost-effective and has been widely used in aquatic genome projects[7,8].

The genome size and heterozygosity of *O. sinensis* were estimated using Jellyfish v2.2.6[9] and GenomeScope v1.0.0[10] by k-mer analysis with clean stLFR data. The clean data was then used in *de novo* genome assembly with the stlfr2supernova pipeline (https://github.com/BGI-Qingdao/stlfr2supernova_pipeline). This pipeline conducts *de novo* assembly using stLFR data with Supernova Assembler, which refers to the de novo software from 10X Genomics[11]. Nanopore long-reads were employed to carry out further scaffolding, gap-closing, and polishing at the same time using TGS-GapCloser[12]. The size of the assembly after these steps was bigger than the k-mer result (approximately 906 Mb) so purge_haplotigs[13] was applied to remove redundancy. Finally, a draft assembly that covered approximately 796 Mb of the genome with a contig N50 length of 373.71 Kb was obtained.

Hi-C analysis and chromosome assembly. The Hi-C reads were aligned to the draft assembly using HiCPro[14] to find valid read pairs. Juicer[15] and 3D-DNA[16] were then used to finish the construction of chromosomes, with manual correction of misjoins, wrong order, and opposite orientation using Juicebox[15]. The Hi-C scaffolding resulted in 23 pseudo-chromosomes with a total length of 712.17 Mb. We offer a chromosome circus map using TBtools[17] and heatmap using Juicebox[18] (Fig. 1A,B). Genome assembly statistics are shown in Table 1 and Table 2. BUSCO (Benchmarking Universal Single-Copy Orthologs)[19] evaluation of the assembly reached 95.1% using actinopterygii_odb9 database, indicating a well-covered fish genome assembly (Table 3).

Repeat annotation. Tandem repeats and interspersed repeats identification was essential before protein-coding gene and function annotations. RepeatMasker[20] and RepeatProteinMask were applied to find repeat elements as homology predictions based on RepBase[21]. For the *de-novo* method, RepeatModeler[22] was used to predict repeat elements; LTR_FINDER[23], and TRF tool[24] were also used to predict repeat elements based on sequences features. Overall, 311.32 Mb of the *O. sinensis* genome assembly were identified as repetitive elements, accounting for 39.05% of the whole genome (Table 4).

Gene prediction and annotation. Protein coding genes were predicted with multiple sources of evidence including homology-based alignments, *de novo* prediction, and transcriptome-assisted methods. In homology alignments, the protein sequences of *Danio rerio* (GRCz11), *Ictalurus punctatus* (IpCoco_1.2), *Oryzias javanicus* (OJAV_1.1), *Oryzias latipes* (ASM223467v1), *Oryzias melastigma* (Om_v0.7), *Poecilia formosa* (PoeFor_5.1.2), and *Takifugu rubripes* (FUGU5) were mapped to a repeat soft masked draft genome using blat[25], and Genewise was applied to define gene models[26]. In *de-novo* prediction, Augustus[27] was used to predict the coding regions of genes. In transcriptome-assisted methods, two different methods were used to obtain predicted gene sets. First, RNA reads were mapped to the genome assembly using Hi-SAT2[28]; the result was used to build a transcript gene model using Stringtie[29] and TransDecoder. Second, a transcriptome was first assembled with RNA reads using

Type	Repeat Size(bp)	% of genome
TRF	12,455,455	1.56
RepeatMasker	104,455,645	13.10
RepeatProteinMask	31,952,649	4.01
De novo	293,230,747	36.78
Total	311,329,140	39.05

Table 4. Prediction of repeat elements.

	Number	Percentage (%)
Total	22,461	100
Swissprot-Annotated	20,906	93.08
KEGG-Annotated	19,352	86.16
TrEMBL-Annotated	22,135	98.55
Interpro-Annotated	20,291	90.34
Overall	22,162	98.67

Table 5. Genome function annotation result.

Trinity, before PASA was used to get a gene model. Finally, all evidence was merged to form a consensus gene structure annotation result using GLEAN[30]. A total of 22,461 protein-coding genes were successfully predicted. BUSCO[19] assessments reached 94.6% using the actinopterygii_odb9 database, indicating relatively complete gene annotation coverage (Table 3).

All predicted genes were compared with public biological function databases including KEGG[31], Swissprot[32], and TrEMBL[33] (https://www.uniprot.org/statistics/TrEMBL) using BLASTp[34] with an E-value cutoff of 1e-5 for functional annotation. Interpro was applied to provide functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites, followed by Gene Ontology annotation[35–37]. Overall, 22,162 protein-coding genes (98.67%) were successfully functionally annotated (Table 5).

Non-coding RNAs (ncRNAs) are an important part of genome annotation as they can be active in transcriptional and translational regulation of gene expression as well as in the modulation of protein function[38]. Since ribosomal RNA (rRNA) was highly conservative, vertebrate rRNA data was used as a reference to map to the draft genome using BLASTn with an E-value of 1e-5. For the discovery of transfer RNA (tRNA), tRNAscan-SE v1.3.1 was applied with eukaryotic parameters according to the characteristics of tRNA[39]. Rfam database was applied to find microRNAs (miRNA) and small nuclear RNA (snRNA)[40,41]. In total, 676 rRNA, 740 tRNA, 231 miRNA, and 1,168 snRNA genes were identified from the *O. sinensis* genome.

Chromosome accessibility analysis using ATAC-seq. ATAC-seq is a widely used method used to identify regions of open chromatin in genomes. SOAPnuke (v1.5.2)[42] was used to filter out reads with low quality original data or high content of unknown bases. Bowtie2 (v2.2.5)[43] was used to align reads to the *O. sinensis* genome assembly and peak calling was performed with MACS2 software (v2.1.0)[44]. Peaks related to genes were used to carry out GO and KEGG enrichment analysis (Fig. 2).

Data Records

The Nanopore long reads, stLFR genomic sequencing data, Hi-C data as well as the assembled genome have been deposited at China National GeneBank DataBase (CNGBdb) under the accession CNP0003475[45]. The whole sequencing dataset of *O. sinensis* was deposited in the NCBI Sequence Read Archive database (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA895195) under project identification number PRJNA895195. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JAUDJI000000000. The version described in this paper is version JAUDJI010000000[46]. Sequence Read Archive (SRA) project number was SRP410304[47]. DNA sequencing data from the WGS library were deposited in the SRA at SRR22435960[48]. DNA sequencing data from the Hi-C library were deposited in the SRA at SRR22435961[49]. The ATAC-sequencing data were also deposited at Figshare[50].

Technical Validation

To evaluate the quality of the genome assembly, stLFR reads were mapped to the final reference genome assembly using BWA (v0.7.12). A total of 98.27% of reads were mapped, covering 98.29% of the genome sequence. Genome average sequencing depth reached 172x, and 97.34% of the genome had a sequencing depth of over 20x. The scatter plot between assembly sequencing depth and GC content found no abnormal GC content, recognized as no exogenous pollution. The completeness of the genome assembly and annotation was assessed using BUSCO (v3.0) with the actinopterygii_odb9 database. The BUSCO benchmark of genome assembly and gene annotation reached 95.1% and 94.6%, respectively.
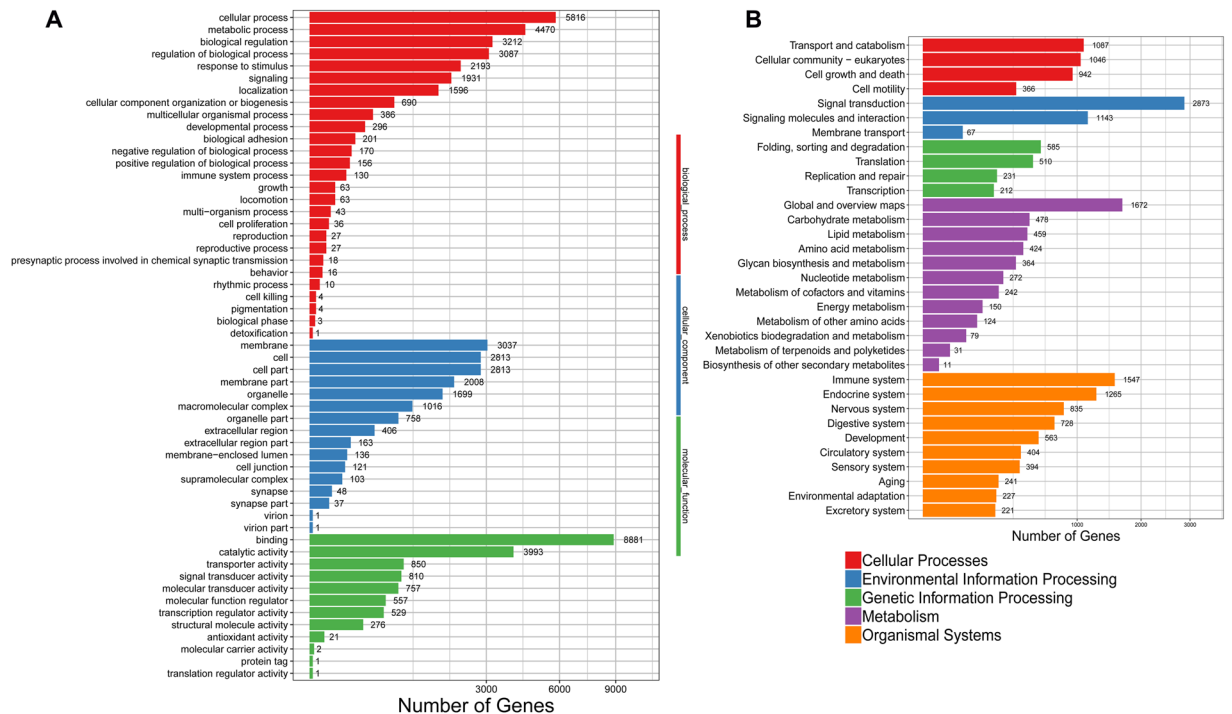
Fig. 2 ATAC-seq peaks result from related genes' functional enrichment. (**A**) GO enrichment; (**B**) KEGG enrichment.

Code availability

All software used in this study are in the public domain, with parameters described in Methods and this section. If no detailed parameters were mentioned for the software, default parameters were used according to the software introduction.

References

1. Wittbrodt, J., Shima, A. & Schartl, M. Medaka–a model organism from the far East. *Nat Rev Genet.* **3**, 53–64 (2002).
2. Cui, L. *et al*. *Oryzias sinensis*, a new model organism in the application of eco-toxicity and water quality criteria (WQC). *Chemosphere.* **261**, 127813 (2020).
3. Wang, Y. *et al*. Genome and transcriptome of Chinese medaka (*Oryzias sinensis*) and its uses as a model fish for evaluating estrogenicity of surface water. *Environ Pollut.* **317**, 120724 (2023).
4. Kasahara, M. *et al*. The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**, 714–719 (2007).
5. Parenti, L. R. A phylogenetic analysis and taxonomic revision of ricefishes, *Oryzias* and relatives (Beloniformes, Adrianichthyidae). *Zool J Linn.* **154**, 494–610 (2008).
6. Wang, O. *et al*. Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res.* **29**, 798–808 (2019).
7. Zhao, N. *et al*. Genome assembly and annotation at the chromosomal level of first Pleuronectidae: Verasper variegatus provides a basis for phylogenetic study of Pleuronectiformes. *Genomics.* **113**, 717–726 (2021).
8. Zhao, N. *et al*. High-quality chromosome-level genome assembly of redlip mullet (*Planiliza haematocheila*). *Zool Res.* **42**, 796–799 (2021).
9. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.* **27**, 764–770 (2011).
10. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics.* **33**, 2202–2204 (2017).
11. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res.* **27**, 757–767 (2017).
12. Xu, M. Y. *et al*. TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *Gigascience.* **9**, giaa094 (2020).
13. 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).
14. Servant, N. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
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. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science.* **356**, 92–95 (2017).
17. Chen, C. C. *et al*. TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol Plant.* **13**, 1194–1202 (2020).
18. Robinson, J. T. *et al*. Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cell Syst.* **6**, 256–258 (2018).
19. Waterhouse, R. M. *et al*. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol Biol Evol.* **35**, 543–548 (2018).

20. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics.* Chapter 4, 4.10.1–4.10.14 (2009).
21. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA.* **6**, 11 (2015).
22. Tempel, S. Using and understanding RepeatMasker. *Methods Mol Biol.* **859**, 29–51 (2012).
23. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–8 (2007).
24. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573–580 (1999).
25. Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
26. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome Res.* **14**, 988–995 (2004).
27. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–7 (2005).
28. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods.* **12**, 357–60 (2015).
29. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol.* **33**, 290–295 (2015).
30. Elsik, C. G. *et al.* Creating a honey bee consensus gene set. *Genome Biol.* **8**, R13 (2007).
31. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
32. UniProt Consortium. T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46**, 2699 (2018).
33. Consortium, U. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res* **51**, D523–D531 (2023).
34. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol.* **215**, 403–410 (1990).
35. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).
36. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
37. Mitchell, A. L. *et al.* InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360 (2019).
38. Szymanski, M., Erdmann, V. A. & Barciszewski, J. Noncoding RNAs database (ncRNAdb). *Nucleic Acids Res.* **35**, D162–D1644 (2007).
39. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
40. Kalvari, I. *et al.* Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* **46**, D335–D342 (2018).
41. Kalvari, I. *et al.* Non-Coding RNA Analysis Using the Rfam Database. Curr Protoc. *Bioinformatics.* **62**, e51 (2018).
42. Chen, Y. *et al.* SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience.* **7**, 1–6 (2018).
43. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* **9**, 357–359 (2012).
44. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome.* **9**, R137 (2008).
45. *China national gene bank* https://db.cngb.org/search/project/CNP0003475/
46. Dong, Z. D. *et al.* Genbank https://identifiers.org/ncbi/insdc.gca:GCA_037389245.1 (2024).
47. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP410304 (2023).
48. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR22435960 (2023).
49. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR22435961 (2023).
50. Dong, Z. D. *et al.* A high-quality chromosome-level genome assembly of the Chinese rice fish *Oryzias sinensis*. *figshare* https://doi.org/10.6084/m9.figshare.24171774.v3 (2023).

Acknowledgements

Author contributions

B.Z. and Z.W. designed the study. J.W., G.C., Y.G. and N.Z. performed the experiments and analyzed the data. Z.D. and B.Z. wrote the paper. Z.W. and N.Z. revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.W. or B.Z.

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.