scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Achieving Chromosome-Level Genome Assembly of *Onychostoma macrolepis* with PacBio Sequencing and Hi-C Technologies

Lu Zhang[1], Ying Zhu[1], Lanzhu Zou[2], Min Cao[1] & Chao Li[1] ✉

Onychostoma macrolepis **is an emerging commercial fish with high edible value. Due to overfishing and environmental pollution, the wild population of** ***Onychostoma macrolepis*** **is declining sharply. In this study, we successfully assembled a chromosome-level genome for** ***Onychostoma macrolepis*** **using a comprehensive approach that integrated Illumina sequencing, PacBio sequencing, and Hi-C technology. The genome spans is 944.02 MB, the contig N50 is 32.36 Mb, 99.15% of the assembled sequences are anchored on 25 chromosomes. A total of 26,294 protein-coding genes were predicted, among which 96.30% were functionally annotated. The high-quality genome data of** ***O. macrolepis*** **obtained in this study provide reliable resources for further research and restoration of wild germplasm resources.**

Background & Summary

Largescale shoveljaw fish (*Onychostoma macrolepis*, Bleeker, 1871, NCBI Taxonomy ID: 369639), Cypriniformes, Cyprinidae, Barbinae, *Onychostoma*[1], is an emerging commercial fish, distributed mainly in the river system of the Yangtze River, the Huaihe River, the Weihe River and the Haihe River in China. *Onychostoma macrolepis* is popular among children, the elderly and especially pregnant women because of its tender meat, delicious taste and medicinal value. It is twenty times more expensive than common carp (*Cyprinus carpio*). In recent decades, because of overfishing and environmental pollution, wild populations of *O. macrolepis* have declined sharply, and the wild population of *O. macrolepis* is listed as a second-class protected animal in the List of Key Protected Wild Animals of China[2]. Therefore, to meet the market demand, *O. macrolepis* farming has received more and more attention.

In order to protect the germplasm resources of this fish and develop its industrialization, there are many fundamental research areas including immunity[3,4], metabolism[5], feed nutrition[6,7], reproduction mechanism[8,9] and so on. In addition, we found that the fish grew slowly, with the body length of adult fish being only about 20 cm, which showed sexual dimorphism. Adult female *O. macrolepis* grow twice as large as their male siblings[10]. For the further development of relevant studies, especially in molecular breeding fields like all-female breeding and the establishment of fast-growing strains, possessing high-quality assembled genomes of *O. macrolepis* serves as a crucial and indispensable foundation.

In our current study, a comprehensive strategy was adopted, which combined PacBio long-read sequencing, chromosome conformation capture (Hi-C), and Illumina paired-end sequencing for initial assembly and polishing, with Oxford Nanopore Technologies (ONT) ultra-long reads used for gap filling. As a result, we successfully assembled a high-quality reference genome of *O. macrolepis* at the chromosomal level. The assembly spans 944.02 Mb with contig N50 of 32.36 Mb and scaffold N50 of 36.04 Mb, organized into 118 contigs and 79 scaffolds, of which 936.02 Mb (99.15%) were anchored onto 25 chromosomes in size from 26.85 to 61.86 Mb. According to the BUSCO assessment (97.9%) and CEGMA assessment (95.56%), along with the QV value of 48.86 and the re-mapping coverage of 99.59% for Illumina paired reads, it was shown that the assembly of *O. macrolepis* exhibited good completeness. In total, 26,294 protein-coding genes were predicted in the genome, and ~96.30% of the genes were functionally annotated. In addition, we identified 444.62 Mb (47.07% of the genome) of repetitive sequence, and the non-coding RNA prediction enabled the identification of 25,255 tRNAs,

[1]School of Marine Science and Engineering, Qingdao Agricultural University, Qingdao, 266109, People's Republic of China. [2]Institute of Fisheries, Taian Academy of Agricultural Sciences, Shandong, China. ✉e-mail: chaoli@qau.edu.cn

Sample	Raw Reads	Clean Reads	Raw Base (G)	Clean Base (G)	Q20 (%)	Q30 (%)	GC Content (%)
SM1-sp-A	25,370,225	23,729,684	7.61	7.12	97.22	92.38	44.12
SM1-G-A	39,240,210	35,746,754	11.77	10.72	97.16	92.31	45.48
SM1-br-A	51,598,038	47,140,625	15.48	14.14	96.91	91.73	45.16
SM1-Sk-A	48,783,650	43,486,340	14.64	13.05	96.83	91.5	46.33
SM1-Gi-A	21,513,134	21,018,129	6.45	6.31	97.16	92.13	46.27
SM1-Sw-A	21,539,751	21,049,229	6.46	6.31	97.48	92.85	48.34
SM1-In-A	37,950,784	35,102,149	11.39	10.53	97.09	92.14	47.14
SM1-K-A	24,020,715	22,059,731	7.21	6.62	97.41	92.78	45.27
SM1-bl-A	24,832,153	21,761,636	7.45	6.53	97.78	93.61	47.49

Table 1. The output quality of *O. macrolepis* transcriptome data.

platform	PacBio-HiFi	ONT ultra-long	HiC	Illumina
Total read bases (Gb)	46.04	42.67	112.29	55.10
Read number	2,794,360	1,207,111		—
Read N50 (bp)	16,905	52,217		—
Mean read length (bp)	17,051	35,349		—
Coverage depth (×)	48.64	45.08	118.63	58.27

Table 2. Statistics of genome sequence of *O. macrolepis*.

19,839 rRNAs, 2,926 miRNAs, and 2,695 snRNAs, in the genome. The quality of our *O. macrolepis* genome shows notable enhancement compared with the genome of the previous study, and more protein-coding genes and non-coding RNAs have been annotated[11]. The genomic data in this study provide a reliable resource for the further study of fish biology and genetics, restoration of wild germplasm resources and other aspects.

Methods

Sample collection and construction of sequencing libraries. The fish was dissected promptly following anesthesia with MS-222. Muscle sample was collected from a healthy *O. macrolepis*. Genomic DNA was extracted using Qiagen DNeasy Kit (Qiagen, Shanghai, China). In line with the protocols corresponding to different sequencing strategies, one Illumina paired-end sequencing library with an insert size of 350 bp, one SMRT bell sequencing library with a 15 kb insert size, and one Hi-C library were constructed. Subsequently, the Illumina paired-end sequencing library and the Hi-C library were sequenced on the Illumina HiSeq platform, while the SMRT bell sequencing library was sequenced on the PacBio Sequel IIe platform. The raw sequence reads generated by the Illumina platform, which had adapters or contained N bases exceeding 10% or had low-quality bases (≤5) accounting for more than 20%, were filtered out[12,13]. Finally, a total of 55.10 Gb Illumina paired-reads and 112.29 Gb Hi-C data were obtained. The subreads of PacBio data were filtered using default parameters. Subsequently, the ccs software (https://github.com/PacificBiosciences/ccs) was employed to generate high-precision HiFi reads. In total, 2,794,360 reads with an average read length of 17,051 bp and N50 of 16,905 bp were obtained and used for further assembly of *O. macrolepis* genome (Table 1). Nine tissues including blood, brain, kidney, gonad, skin, gill, swim bladder, intestine and spleen were sampled for the same fish individual. Subsequently, RNA was extracted from all nine samples with the RNAiso Pure RNA Isolation Kit (TaKaRa, Japan), after which DNaseI treatment was carried out. The concentration and quality of the obtained RNAs were evaluated using Nanodrop (Thermo Fisher Scientific, USA). Afterwards, all the RNA samples were sequenced on the Illumina NovaSeq. 6000 to generate 150 bp paired-end reads (Table 1).

For Oxford Nanopore Technologies (ONT) ultra-long reads (which are based on the Single molecule nanopore DNA sequencing method developed by ONT), DNA required for Nanopore sequencing was isolated from muscle samples of a healthy *O. macrolepis*. The Qiagen DNeasy kit was used for extraction, following the manufacturer's provided instructions. The extracted genomic DNA underwent quality assessment for purity, concentration, and integrity using Nanodrop, Qubit, and 1% agarose gel electrophoresis. Subsequently, the DNA was fragmented with the Diagenode Megaruptor. Fragment selection was performed using BluePippin. Finally, the constructed library was sequenced on the PromethION-specific Flow Cell (R9.4). Fastq files were generated from raw Nanopore data using Guppy and quality control was subsequently performed on the data using Nanoplot (v1.38.1). A total of 1,207,111 reads (42.67 Gb) with an average read length of 35,349 bp and N50 of 52,217 bp were passed for further assembly (Table 2).

Genome size estimation and assembly. According to the k-mer analysis (k-mer = 17) conducted using the software Jellyfish 2.2.7[14], the genome size of *O. macrolepis* was estimated based on the paired-end Illumina sequencing data. The results indicated that the genome size of *O. macrolepis* was 945.67 Mb, the heterozygosity rate of the genome was 0.28%, and the proportion of repetitive sequences was 51.55%. Utilizing the long HiFi reads produced by the PacBio Sequel lle platform, the contig assembly of *O. macrolepis* genome was carried out with Hifiasm 0.16.1[15]. Subsequently, the primary contigs underwent polishing with PacBio reads via Arrow Smartlink 8.0[16]. Next, the Purge Haplotigs software[17] was employed to eliminate redundant contigs from the

Parameter	*O. macrolepis* (this study)	*O. macrolepis* (previous study)
Assembly strategy	PacBio-HiFi + ultra-ONT + Hi-C	PacBio + Hi-C
Total Size (bp)	944,024,602	894,945,186
Anchored onto chromosomes (bp)	936,022,107	889,754,504
Sequences assigned to chromosemes (%)	99.15	99.42
Gap Number	4	—
No. of contig	118	689
Contig N50 length (bp) /rank	32,362,784/13	20,336,701/16
Contig N90 length (bp) /rank	11,335,856/30	4,120,503/56
No. of scaffold	79	241
Scaffold N50 length (bp) /rank	36,044,644/11	35,346,331/11
Scaffold N90 length (bp) /rank	32,328,402/22	31,713,302 /21
BUSCO (%)	98.20	97.20
CEGMA (%)	95.56	94.35
Quality Value (QV)	48.86	—
Mapping rate of Illumina data (%)	99.59	98.49
Mapping covarge of Illumina data (%)	99.95	99.23
Protein coding genes	26294	25779
Percentage of functionally annotated genes	96.3	94.70
Repeat bases (Mb)	444.62	424.37
Percentage of repeats (%)	47.07	47.42
tRNAs	25255	3098
rRNAs	19839	410
miRNAs	2926	2654
snRNAs	2695	708

Table 3. Summary of our *O. macrolepis* genome and published *O. macrolepis* genome assembly.

initial assembly, yielding a non-redundant contig assembly. In the end, the short reads from Illumina were utilized to rectify any remaining errors in the contig assembly through Pilon 1.22 software[18]. These processes would yield a final draft genome of *O. macrolepis* assembly with a total length of 944.02 Mb and contig N50 of 32.36 Mb (Table 3). To generate a chromosome-level genome assembly, Hi-C technology was employed to anchor the contigs onto chromosomes. By means of the hierarchical clustering algorithm, the contigs were clustered, ordered and oriented. Subsequently, they were anchored into 25 pseudo-chromosome linkage groups using ALLHiC 0.9.8, and a total of 936.02 Mb assembled genome sequences were anchored on 25 chromosomes, with an average chromosome length of 37.44 Mb. Finally, gap filling was carried out using the ultra-long read length of Oxford Nanopore Technologies (ONT) through the TGS-GapCloser[19], and as a result, the number of gaps was reduced from 39 to 4. (Table 3, Fig. 1).

The quality of the assembled genome was evaluated through four methods. Specifically, the Benchmarking Universal Single-Copy Orthologs analysis (BUSCO, v5.2.2)[20] and the Core Eukaryotic Genes Mapping Approach analysis (CEGMA, V2.5)[21] were utilized to assess the integrity of the genome. For BUSCO, this was accomplished by searching against 3640 BUSCOs of the actinopterygii_odb10, while for CEGMA, it was performed by selecting the conserved genes (248 genes) existing in six eukaryotic model organisms. The Illumina paired reads were realigned to the assembled genome using the BWA 0.7.8[22] software. The realignment ratio and coverage depth were then computed to evaluate the integrity of the assembly and the uniformity of the sequencing; and the package Merqury[23] was employed for assessing the accuracy of genome assembly using a reference-free and k-mer-based approach.

Genome annotation of *O. macrolepis*. In our repeat annotation pipeline, a combined strategy relying on homology alignment and de novo search was applied to identify the whole genome repeats. Tandem Repeat was extracted using TRF v4.09[24] by ab initio prediction. The homolog prediction utilized Repbase database[25] employing RepeatMasker v4.1.0[26] software and its in-house scripts (RepeatProteinMask v4.1.0) with default parameters to extracted repeat regions. For ab initio prediction, a de novo database of repetitive elements was built using LTR_FINDER v1.06[27], RepeatScout v1.0.5[28], and RepeatModeler v2.0.1[29] with default parameters. Subsequently, all repeat sequences with lengths > 100 bp and gap 'N' less than 5% constituted the raw transposable element (TE) library. A custom library, which combined Repbase with our de novo TE library (processed by uclust to make it non-redundant), was provided to RepeatMasker v4.1.0 for identifying DNA-level repeats. Based on these analyses, approximately 444.62 Mb of repeat sequences were ultimately uncovered, making up 47.07% of the entire *O. macrolepis* genome (Table 4).

Three methods (ab initio prediction, homology-based prediction and RNA-Seq assisted prediction) were employed to predict the protein-coding genes in the *O. macrolepis* genome. For homology-based prediction, the

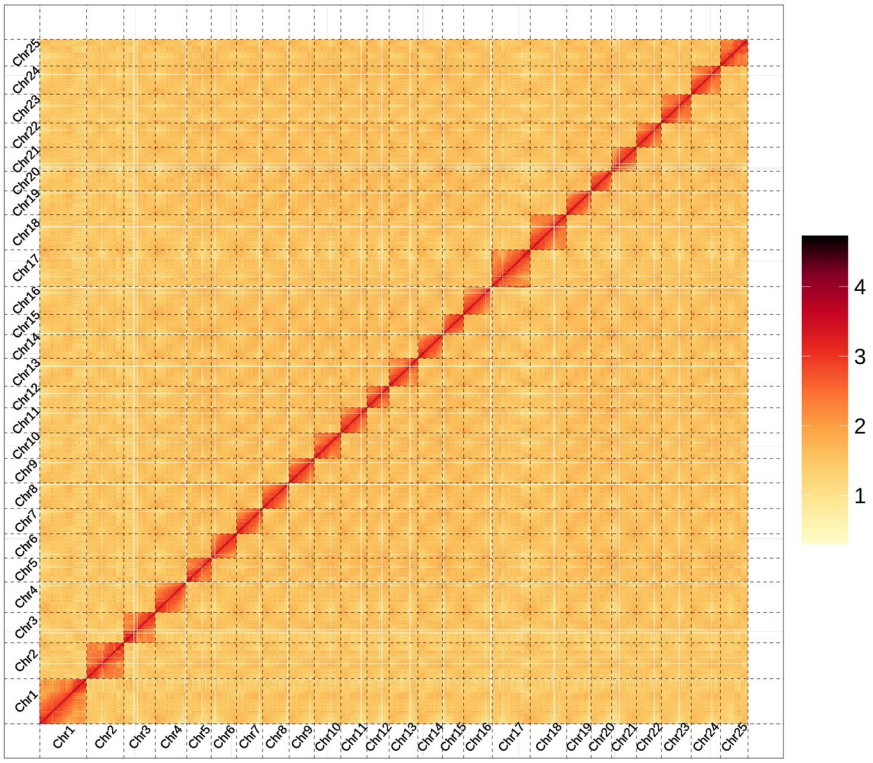

Fig. 1 Hi-C assembly of chromosome interactive heatmap of *O. macrolepis* genomes. Chromosome interactive heatmap: Chr1 to Chr25 are the abbreviations for the 25 Chromosomes. The abscissa and ordinate represent the order of each bin on the corresponding chromosome group; The colour block illuminates the intensity of interaction from yellow (low) to red (high). Window size = 500 kb.

	Repbase + Denovo		TE Protiens		Combined TEs(all without trf)	
Type	**Length (Bp)**	**% in genome**	**Length (Bp)**	**% in genome**	**Length (Bp)**	**% in genome**
DNA	70,667,895	7.48193	6,797,579	0.71969	73,713,681	7.804401
LINE	20,795,259	2.201688	26,746,203	2.831742	38,099,063	4.03372
SINE	858,828	0.090928	0	0	858,828	0.090928
LTR	331,980,268	35.148255	15,783,509	1.671072	333,349,095	35.293179
Satellite	2,349,268	0.248728	0	0	2,349,268	0.248728
Simple_repeat	3,552	0.000376	0	0	3,552	0.000376
Unknown	18,131,052	1.919617	0	0	18,131,052	1.919617
Total	430,957,298	45.627402	49,317,394	5.221456	434,587,509	46.011749

Table 4. Summary of the TEs in *O. macrolepis* genome.

reference protein sequences of four species (Table 5) were downloaded from the NCBI database. Subsequently protein sequences were aligned to the *O. macrolepis* genome using TblastN(v2.2.26; E-value $\leq 1*e^{-5}$ [30]), the gene statistics, including gene length, coding sequence (CDS), intron length, and exon length, were comparable between these species and the *O. macrolepis* genomes (Fig. 2).

The matching proteins were aligned to the homologous genome sequences via GeneWise (v2.4.1)[31] software for accurate spliced alignments and to predict the gene structure within each protein region. Regarding ab initio gene prediction, five tools, namely Augustus (v3.2.3)[32], Geneid (v1.4)[33], Genescan (v1.0)[34], GlimmerHMM (v3.04)[35] and SNAP (V2013.11.29)[36] were utilized in the gene prediction pipeline. For RNA-Seq assisted prediction, transcriptome reads assemblies were generated with Trinity (v2.1.1) for the genome annotation. To optimize the genome annotation, the RNA-Seq reads from nine different tissues were aligned to the genome fasta using TopHat (v2.0.11) with default parameters[37] to identify exons region and splice positions. Subsequently, Cufflinks (v2.2.1) was used for reconstruction and assembly of transcripts[38]. The non-redundant reference gene set was generated by merging genes predicted by three methods with EvidenceModeler (EVM,v1.1.1) using PASA(Program to Assemble Spliced Alignment)[39] terminal exon support and including masked transposable elements as input into gene prediction. Overall, 26,294 protein-coding genes were successfully identified in the *O. macrolepis* genome, respectively (Table 3).

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	38567	9562.43	1149.98	6.53	176.23	1522.45
	GlimmerHMM	66212	9490.28	620.71	4.22	147.23	2757.95
	SNAP	63070	13538.17	712.09	5.32	133.92	2970.85
	Geneid	30989	18493.87	1361.98	6.42	212.12	3160.48
	Genscan	31303	20538.98	1539.5	8.23	187.09	2628.42
Homolog	Drer	22811	13389.39	1592.77	8.73	182.35	1525.17
	Ccar	23679	13539.57	1579.98	8.74	180.74	1544.83
	Anig	28372	11209.33	1310.14	7.48	175.21	1528.28
	Amex	22060	13271.45	1573.65	8.55	183.97	1548.62
RNAseq	PASA	74320	10969.31	1263.94	7.75	163.13	1438.24
	Transcripts	51883	27841.98	4318.61	12.79	337.74	1995.75
EVM		35920	12348.78	1269.24	7.37	172.3	1740.28
Pasa-update*		35627	12688.85	1291.61	7.43	173.75	1771.53
Final set*		26294	15557.19	1575.97	9.13	172.7	1720.71

Table 5. Statistics of gene structure in *O. macrolepis* genomes.

Annotation Database	Number of Annotated Genes	Percentage (%)
Swissprot	21,539	81.9
Nr	24,617	93.6
KEGG	21,609	82.2
InterPro	24,533	93.3
GO	16,620	63.2
Pfam	20,509	78
Annotated	25,327	96.3
Unannotated	967	3.7
Total	26,294	—

Table 6. Summary of functional annotations for predicted genes of *O. macrolepis* genome.

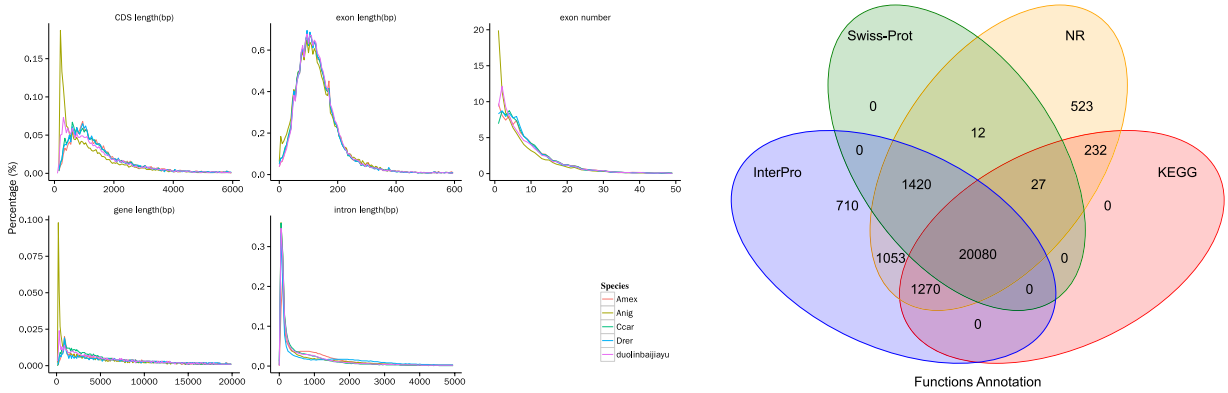

Fig. 2 Comparisons of the lengths of genes, CDSs, exons, and introns between *O. macrolepis* and closely related species, along with a Venn diagram showing the number of genes annotated by different gene databases in the *O. macrolepis* genome. Amex represents *A. mexicanus*, Anig represents *Anchorythroculter nigrocauda*, Ccar represents *C. carpio*, Drer represents *D. rerio*, and duolinbaijiayu represents the studied species *O. macrolepis*.

Gene functions were assigned according to the best match by aligning the protein sequences to the Swiss-Prot using Blastp (with a threshold of E-value $\leq 1e^{-5}$). The motifs and domains were annotated using InterProScan70 (v5.31)[40] by searching against publicly available databases, including ProDom, PRINTS, Pfam, SMRT, PANTHER, and PROSITE. The Gene Ontology (GO) IDs[41] for each gene were assigned according to the corresponding InterPro entry. The protein function was predicted by transferring annotations from the closest BLAST hit (E-value $< 10^{-5}$) in the Swissprot database[42] and BLAST hit (E-value $< 10^{-5}$) in the NR database. Additionally, the gene set was mapped to a KEGG pathway and identified the best match for each gene[43]. In summary,

a total of 25,327 genes (96.30%) were successfully annotated by function implications or the conserved functional motifs in *O. macrolepis* genomes (Table 6, Fig. 2).

For non-coding RNA annotation, the tRNAs were predicted using the program tRNAscan-SE v1.4[44]. Due to rRNAs are highly conserved, the relative species' rRNA sequences were chosen as references to predict rRNA sequences by using Blast with an E-value of $1*E^{-10}$[45]. Other ncRNAs, including miRNAs and snRNAs, were identified by searching against the Rfam database with default parameters using the infernal software[46].

Data Records

The genomic data of *O. macrolepis* have been submitted to NCBI with the respective accession numbers, PRJNA1212228[47] (BioProject) and SAMN46295721[48] (BioSample). Furthermore, various data from Illumina sequencing, Hi-C sequencing, PacBio sequencing, RNA-seq data and ONT reads have been submitted to NCBI SRA with the accession numbers SRR32040837[49], SRR32040935[50], SRR32355807[51], SRR32040922[52]-SRR32040930[53], SRR33977354[54] respectively. Additionally, the final chromosome-level genome assembly of *O. macrolepis* has been uploaded to GenBank in NCBI, bearing the accession number JBMGJU000000000[55].

Technical Validation

The quality of the assembled genome was evaluated through four methods. Specifically, the Benchmarking Universal Single-Copy Orthologs analysis (BUSCO, v5.2.2) and the Core Eukaryotic Genes Mapping Approach analysis (CEGMA, V2.5) were utilized to assess the integrity of the genome. For BUSCO, this was accomplished by searching against 3640 BUSCOs of the actinopterygii_odb10, while for CEGMA, it was performed by selecting the conserved genes (248 genes) existing in six eukaryotic model organisms, the results showed that 98.20% and 95.56% of BUSCOs and CEGMAs recovered in genome assemblies of *O. macrolepis*, respectively (Table 2). The Illumina paired reads were realigned to the assembled genome using the BWA 0.7.8 software. 99.59% Illumina short-reads were mapped the genome assembly of *O. macrolepis* and the coverage reached to 99.95%. The package Merqury was employed for assessing the accuracy of genome assemblies using a reference-free and k-mer-based approach, which resulted in a high assembly quality value (QV) of 48.86 for *O. macrolepis* genome and exceed the Vertebrate Genome Project (VGP) standard of QV40[56]. Overall, the genome assemblies of *O. macrolepis* in the present study showed excellent completeness and superior accuracy.

Code availability

All data processing commands and pipelines were carried out in strict accordance with the manuals and protocols of the relevant bioinformatics software. No custom script or code was used in the study.

Received: 27 March 2025; Accepted: 10 July 2025;
Published online: 13 August 2025

References

1. Chen, S. W. Study on the biological characteristics of wild *onychostoma macrolepis* in renhe, ziyang, Shaanxi. *Shaanxi J Agricultural Sci* **65**, 63–66 (2019).
2. Qu, G. S. *et al*. Investigation and protection of germplasm resources of *Varicorhinus macrolepis* in Qinba Mountain area. *Hubei Agricultural Sciences* **18**, 93–97 (2019).
3. Xu, H. Z. *et al*. Molecular characterization and antibacterial ability of galectin-3 and galectin-9 in *Onychostoma macrolepis*. *Dev Comp Immunol* **128**, 104333 (2022).
4. Shang-Guan, X. Y. *et al*. A C-type lectin (OmCTL) in *Onychostoma macrolepis*: Binding ability to LPS, PGN and agglutinating activity against bacteria. *Mol Immunol* **132**, 21–29 (2021).
5. Deng, W. *et al*. Energy response and fatty acid metabolism in *Onychostoma macrolepis* exposed to low-temperature stress. *J Therm Biol* **94**, 102725 (2020).
6. Gou, N., Wang, K. F., Jin, T. Z. & Yang, B. First Insights on the Administration of Insect Oil (Black Soldier Fly Larvae) in the Diet of Juvenile *Onychostoma macrolepis*. *Animals (Basel)* **13**, 518 (2023).
7. Zhong, M. Z. *et al*. Effects of dietary dextrin level on growth performance, feed utilization, biochemical composition, digestive enzyme activity, antioxidant capacity and tissue structure of juvenile *Onychostoma macrolepis*. *Animal Feed Science and Technology* **312**, 115967 (2024).
8. Peng, G. F. *et al*. Testicular miRNAs and tsRNAs provide insight into gene regulation during overwintering and reproduction of *Onychostoma macrolepis*. *Fish Physiology and Biochemistry* **48**, 481–499 (2022).
9. Ding, Y. B. *et al*. Analysis of morphology, histology characteristics, and circadian clock gene expression of *Onychostoma macrolepis* at the overwintering period and the breeding period. *Fish Physiology and Biochemistry* **50**, 1265–1279 (2024).
10. Song, H., Wang, S. Y., Peng, K. & Yin, X. H. Histological studies on gonadiai differentiation and development of *Varicorhinus macrolepis*. *Journal of Fishery Sciences of China* **13**, 723–731 (2006).
11. Sun, L. *et al*. Chromosome-level genome assembly of a cyprinid fish *Onychostoma macrolepis* by integration of nanopore sequencing, Bionano and Hi-C technology. *Mol Ecol Resour* **20**(5), 1361–1371 (2020).
12. Zhang, B. D., Li, Y. L., Xue, D. X. & Liu, J. X. Population genomic evidence for high genetic connectivity among populations of small yellow croaker (*Larimichthys polyactis*) in inshore waters of China. *Fisheries Research* **225**, 105505 (2020).
13. Zhang, B. D., Li, Y. L., Xue, D. X. & Liu, J. X. Population Genomics Reveals Shallow Genetic Structure in a Connected and Ecologically Important Fish From the Northwestern Pacific Ocean. *Frontiers in Marine Science* **7**, 374 (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. Cheng, H. Y., 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).
16. Chin, C. S. *et al*. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**, 563–569 (2013).
17. 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).
18. Walker, B. J. *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).

19. 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).
20. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
21. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
22. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
23. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245 (2020).
24. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573–580 (1999).
25. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11 (2015).
26. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **4**, 4.10.1–4.10.14 (2009).
27. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* **35**, W265–268 (2007).
28. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**(Suppl 1), 351–358 (2005).
29. Bergman, C. M. & Quesneville, H. Discovering and detecting transposable elements in genome sequences. *Brief Bioinform* **8**, 382–392 (2007).
30. Gertz, E. M., Yu, Y. K., Agarwala, R., Schäffer, A. A. & Altschul, S. F. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol* **4**, 41 (2006).
31. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* **14**, 988–3995 (2004).
32. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* **33**, W465–467 (2005).
33. Guigó, R., Knudsen, S., Drake, N. & Smith, T. Prediction of gene structure. *J Mol Biol* **226**, 141–157 (1992).
34. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 78–94 (1997).
35. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
36. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
37. Kim, D. *et al*. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).
38. Trapnell, C. *et al*. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562–578 (2012).
39. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7 (2008).
40. Mulder, N. & Apweiler, R. InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol Biol* **396**, 59–70 (2007).
41. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29 (2000).
42. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28**, 45–48 (2000).
43. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30 (2000).
44. 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).
45. 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).
46. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33**, D121–124 (2005).
47. *NCBI BioProject* https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1212228 (2025).
48. *NCBI BioSample* https://www.ncbi.nlm.nih.gov/biosample/SAMN46295721 (2025).
49. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR32040837 (2025).
50. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR32040935 (2025).
51. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR32355807 (2025).
52. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR32040922 (2025).
53. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR32040930 (2025).
54. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33977354 (2025).
55. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBMGJU000000000 (2025).
56. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357–360 (2015).

Acknowledgements

This study was supported by the Shandong Technical System of Fish Industry (SDAIT-12-02), and the Shandong Natural Science Foundation (ZR2024QC027).

Author contributions

Chao Li conceived the research. Lanzhu Zou was involved in sample collection. Lu Zhang and Ying Zhu contributed to genome assembly and annotation, data analysis and wrote the manuscript. Min Cao and Chao Li revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

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