scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly and annotation of Chinese herring (*Ilisha elongata*)

Bingjian Liu[1,2,5], Xinyi Niu[1,2,5], Chi Zhang[3,5], Sixu Zheng[1,2], Luxiu Gao[1,2], Mingzhe Han[1], Taobo Feng[1], Jinghua Wu[1], Chaoxuan Jiang[1], Shuaishuo Kang[1], DongDong Xu[4] ✉ & Yifan Liu[1,2] ✉

The Chinese herring (*Ilisha elongata*) is a commercially and scientifically significant fish species. In this study, we conducted high-precision whole-genome sequencing using two high-throughput platforms: second-generation MGI and third-generation PacBio. Hi-C technology assisted in assembling the contig sequences onto 24 chromosomes, resulting in a high-quality chromosome-level genome map with excellent continuity and coverage. The completed genome size was approximately 815 Mb, with a contig N50 of 4.82 Mb, scaffold N50 of 32.61 Mb, and a chromosome mounting rate of 95.32%. SNP and InDel purity rates were 0.003% and 0.012%, respectively, and the genome assembly completeness was 96.68%, assessed by BUSCO. Repetitive sequences were annotated via ab initio and homology predictions, identifying 295.7 Mb of repetitive sequences, constituting 35.08% of the genome. A total of 26,381 protein-coding genes were predicted, with 24,596 functionally annotated.

Background & Summary

I. elongata is an important species in clupeoid fisheries across Asian countries[1], significantly contributing to local economies. Ecologically, it occupies a crucial position in the marine food chain, linking primary producers to larger predators and contributing to ecosystem stability. During the day, it is mainly active in the middle and lower layers of the water, but in the evening, it is more likely to live in the middle and upper layers of the water[2]. Spawning occurs near river mouths in summer, with peak activity observed in June, as indicated by ichthyoplankton data collected in March, May, July, and November[3].

Environmental pressures, including climate change, habitat destruction, and overfishing, have adversely impacted *I. elongata* populations, leading to observable trends like miniaturization and early sexual maturity. These changes threaten the sustainability of the species and underscore the urgent need for conservation measures. In-depth exploration in this field not only provides scientific support for the sustainable development of fisheries, but also plays a key role in protecting wild fish stocks and maintaining the ecological balance of waters[4]. The study of fish genomes, including that of *I. elongata*, helps solve biological questions related to reproduction, adaptation, and population structure.

Despite significant progress, there remains a gap in understanding the specific genetic mechanisms behind the environmental adaptability and reproductive biology of *I. elongata*. These studies not only solve the biological problems of fish themselves, but also provide new perspectives and important clues to reveal the entire genetic evolution of vertebrates[5]. Advances in high-throughput sequencing and assembly technologies have enabled genome studies to expand beyond evolution into areas such as gene function, ecology, and environmental adaptation[6].

In this context, genome sequencing of *I. elongata* is timely and essential. It will provide insights to support sustainable fisheries management, preserve genetic resources, and ensure the long-term viability of this valuable species.

[1]College of Marine Science and Technology, Zhejiang Ocean University, 316022, Zhoushan, China. [2]National engineering research center for facilitated marine aquaculture, Zhejiang Ocean University, 316022, Zhoushan, China. [3]Institute of Fisheries Science, Tibet Academy of Agricultural and Animal Husbandry Sciences, Lhasa, China. [4]Key Laboratory of Mariculture and Enhancement, Zhejiang Marine Fisheries Research Institute, 316021, Zhoushan, China. [5]These authors contributed equally: Bingjian Liu, Xinyi Niu, Chi Zhang. ✉e-mail: xudong0580@163.com; et999927@163.com

Methods

Sample collected for genome assembly. In this study, the *I. elongata* samples used for genome assembly were collected from the offshore waters of Zhoushan, Zhejiang Province, China. This fish was a one-year-old female. Various tissue samples of *I. elongata*, including muscles, livers, gills, intestines and hearts, were collected and placed in separate cryopreservation tubes, which were subsequently stored in liquid nitrogen to facilitate subsequent DNA/RNA extraction.

DNA and RNA extraction. Classical phenol/chloroform extraction method was used to extract DNA from liver tissue of *I. elongata* in this study. After extraction, the integrity of the DNA sample was assessed by 1% agarose gel electrophoresis, while the concentration was determined using Nanodrop[7] and quantified with Qubit2.0 (Invitrogen)[8] to ensure that a high quality DNA sample was obtained. Other tissue samples of *I. elongata*, including back muscles, livers, gills, intestines and hearts, were extracted using a TRIzol kit (Invitrogen, USA). The NanoDropND-1000 spectrophotometer (LabTech) and 2100 Bioanalyzer (Agilent Technologies) were then used for quality assessment. The qualified RNA samples were individually sequenced, with 2 μg extracted for transcriptome library preparation and sequencing.

Library construction, high throughput sequencing and raw data quality control. Following the standardized operation procedure of MGI sequencing, a double-ended sequencing library with an insert fragment length of 350 bp was first constructed using qualified DNA samples. After the quality of the library was qualified, DNBSEQ-T7 second-generation sequencing technology was used for sequencing. Fastp (v0.12.4)[9] and Trimmomatic (v0.39)[10] tools were used to perform quality control on the original sequencing data generated by the MGI platform, removing the splicing sequence and screening out low-quality read segments. Then, according to the official specification of SMRTbell Template Prep Kit, a SMRTbell long-read sequence library of about 20 kb fragments was constructe. After quality verification, the library was sequenced on the PacBio Sequel IIe platform. The SequelQC[11] tool was used to remove joints and low quality areas. The construction of the Hi-C library was made by *MboI* enzyme digestion and formaldehyde cross-linking treatment of liver cells of *I. elongata*. The high quality Hi-C library was constructed and verified. The 150 bp double-ended read segment was sequenced by DNBSEQ-T7 platform, and the quality control and filtering of the original data were carried out. Ensure that pairs of reads are available for comparison analysis. For each RNA sample, a clean data set was generated through a series of steps. First, a cDNA library was constructed for each sample individually, followed by 150 bp double-end sequencing using the DNBSEQ-T7 platform. During the later stages of genome sequencing analysis, data containing splices, poly-N sequences, or low-quality reads were removed to ensure data integrity. The overall process included several critical steps, such as DNA and RNA library preparation, sequencing, and Hi-C library construction and sequencing. Each stage adhered to strict quality control protocols to guarantee high-quality data, providing a robust foundation for subsequent bioinformatics analysis.

Genome survey analysis. GCE (v1.0.2)[12] software was used to pre-assess the genome size of *I. elongata* based on sequencing data from the MGI platform. GCE software estimated the genome characteristics by simulating the Poisson distribution law of K-mer in the genome. In this study, the size of K-mer selected was 17, and the total number and coverage of K-mer were calculated. The process of genome survey analysis was mainly divided into two stages: First, the kmerfreq program was used to calculate the frequency distribution of K-mer in the second-generation sequencing data; Secondly, according to the distribution of K-mer, the gce program was used to estimate the genome size, heterozygosity, repeat sequence proportion and other key indicators. These two stages of analysis provided insight into the characteristics of the *I. elongata* genome. These preliminary assessments provided important reference data for subsequent genome splicing, helping to select the most appropriate assembly strategies and computational tools to ensure more accurate genome sequences.

Genome assembly process. The NextDenovo (v2.5.2)[13] software was selected for genome assembly of long-read sequence data generated by PacBio. After assembly, in order to further improve the accuracy of the splicing sequence, racon (v1.5.0)[14] software was used to modify the contigs obtained by splicing. Subsequently, in order to further refine these sequences, meticulous error correction was performed on contigs using Pilon (v1.23)[15] software based on second-generation sequencing data. Comparing the initial assembly results with the estimated genome size from the previous survey analysis, the preliminary assembly of the *I. elongata* genome showed low heterozygosity and proportion of repetitive sequences, demonstrating the high quality of genome assembly.

A series of specialized tools, including Juicer (v1.5)[16], 3D-DNA[17], and Juicebox (v1.9.8)[18], were used in the later stages of genome assembly to attach the initially assembled sequence to the chromosome level. The key operation steps were as follows: (1) Hi-C data comparison: Firstly, Juicer software was used to efficiently compare Hi-C sequencing data to the assembled genome sketch, which provides basic data for the subsequent horizontal assembly of chromosomes; (2) 3D-DNA analysis and correction: Subsequently, 3D-DNA software was used to analyze the Hi-C comparison results, identify and correct the possible wrong connections in contigs, and generate optimized and corrected genome files through Polish, Split, Seal and Merge. (3) JuiceBox correction and visualization: According to the genome assembly manual provided by Aiden Laboratory, Juicebox software was used for graphical manual correction, which not only further improves the accuracy of the genome, but also ensures a high level of reference genomes assembled at the chromosome level. After performing the above steps, the sequence was successfully promoted from contigs to chromosome level, forming a high-quality reference genome.

After the *de novo* assembly of the *I. elongata* genome, in order to ensure the high quality of genomic data, a comprehensive quality evaluation of the assembly results must be carried out, mainly in the following three core aspects: (1) evaluation of sequence continuity: The continuity of genome assembly was assessed using QUAST (v5.0.2, http://cab.cc.spbu.ru/quast/)[19], and the Matplotlib library was used for graphical plotting, and the assembled genome sequence was used as an input file to run the software to contig N50, scaffold N50 and other metrics were calculated and visualized; (2) Sequence consistency evaluation: Using the BWA (v0.7.17)[20] and Minimap2[21] tools, quality controlled MGI and PacBio sequencing data were compared to the assembled *I. elongata* reference genome. Qualimap (v2.2.2)[22] comparison rate (mapping rate) and genome coverage were evaluated. (3) Genome integrity assessment: The genome integrity was analyzed using BUSCO software[23] (https://gitlab.com/ezlab/busco). As *I. elongata* belongs to the radial fins subclass, the actinopterygii_odb10 database was selected as the reference gene set to evaluate the completeness of the genome assembly.

Genome annotation. In this study, two methods, *de novo* prediction and homologous prediction, were used to annotate the repeated sequence of the genome of *I. elongata*. Together, they provide complementary approaches for genome annotation. The specific steps are as follows: (1) A repetitive sequence library was established using RepeatModeler (v2.0.1)[24] and LTR-FINDER (v1.07)[25], and the database was searched by RepeatMasker (v4.1.0)[26] to predict the repetitive sequences in the *I. elongata* genome. Tandem Repeat Finder (v4.09)[27] (https://tandem.bu.edu/trf/trf.whatnew.html) was used to identify tandem repeats in the genome. (2) The identification of repeat sequences in the *I. elongata* genome, similar to those in the ReBase repetitive sequence database (http://www.girinst.org/repbase), is performed using RepeatMasker (v4.1.0)[26] and RepeatProteinMask (v4.1.0)[28] (http://www.repeatmasker.org) software; Finally, the results based on homologous prediction and *de novo* prediction were integrated to obtain a comprehensive annotation of the genomic repeat sequence of *I. elongata*. This strategy effectively captures repetitive elements in the genome, offering essential insights for subsequent functional annotation and evolutionary studies.

Based on the covariance model of Rfam family, miRNA, snRNA and rRNA sequences were predicted using INFERNAL (v1.1.3)[29] software that comes with Rfam. The specific process was as follows: (1) download and verify the location of the Rfam database file; (2) based on the Rfam database file (Rfam.cm), apply the cmpress script to build the library; (3) apply the cmscan script to search the I. elongata genome against the Rfam database, and ultimately obtain the sequence information from the comparison. BLASTN[30] software was employed to align sequences with the known rRNA sequences of related species, leveraging the high conservation of rRNA, to identify rRNA sequences within the genome. Furthermore, tRNAscan-SE (v2.0.9)[31] software was utilized to detect tRNA sequences in the genome of *I. elongata*. The tRNAscan-SE tool integrated multiple analysis programs to analyze the tRNA secondary structure, conserved sequence pattern of promoter elements and transcriptional control elements.

Gene structure prediction of the *I. elongata* genome was performed using three different methods: *de novo* prediction, homology prediction and transcriptome-based prediction. *De novo* prediction uses algorithms to identify genes directly from genomic sequences without external references. Meanwhile, homology prediction detects genes by comparing sequences to known reference genomes, leveraging similarities. In addition, transcriptome-based prediction incorporates RNA data to confirm gene expression and refine gene annotations, making these methods complementary for accurate genome annotation. The following steps are detailed: (1) gene structure was predicted *de novo* using GENSCAN[32] and Augustus (v3.3.3)[33] software; (2) the *I. elongata* genome was compared to the genomes of related species, including *Clupea harengus, Denticeps clupeoides, Alosa alosa, Oncorhynchus keta, and Danio rerio*, using the TBLASTN[34] program (E-value $\leq$ 1e-05). The steps involved are as described: (1) gene structure was predicted; (2) the *I. elongata* genome was compared to the genomes of related species. The results of the comparison were analyzed using GeneWise (v2.4.1)[35] software to predict the protein structures of the coding genes; (3) the transcriptome data was aligned to the reference genome using HISAT2 (v2.2.1)[36] to generate SAM files. Transcriptome assembly was performed using Trinity (v2.1.1)[37], and transcriptome-based annotation is conducted using the PASApipeline (v2.4.1)[38] via the Launch_PASA_pipeline.pl script; finally, the results obtained from the three prediction methods were integrated using the EvidenceModeler (v1.1.1)[39] software, and then the results were filtered using the GFF3Clear script under the GETA (v2.5.6) (https://github.com/chenlianfu/geta) software to Filtering was performed to obtain a comprehensive gene set. Next, the InterPro[40], GO[41], NR[42], SwissProt[43], TrEMBL[44], KEGG[45] and KOG[46] databases were downloaded. The blastp subcommand of Diamond (v2.0.2)[47] software was then used to compare and annotate the predicted gene set with each database (the comparison parameter was set to E-value $\leq$ 1e-09). InterProScan[40] was used to compare the gene set of the *I. elongata* genome with the InterPro protein database. In addition, Blast2GO (v5.2.5)[48] software was used to perform GO annotation on genes of *I. elongata* based on GO database.

The raw sequencing data from the DNBSEQ-T7 platform were subjected to quality control and filtering, resulting in 78.92 Gb of MGI second-generation data, which were used for genome survey analysis and assembly error correction. Additionally, 69.52 Gb of Hi-C data were obtained to assist in the assembly of the genome at the chromosome level. Furthermore, 28.15 Gb of transcriptome data were collected for gene annotation of the genome. The HiFi data from the PacBio Sequel IIe platform were also subjected to quality control and filtering, ultimately yielding 35.37 Gb of long-read data, which were used for genome assembly (Table 1). These long reads comprised 1,995,442 reads with an average length of 17,692 bp and an N50 of 16,795 bp. Of these, 434,775 reads exceeded 20 kb in length (Fig. 1).

Prior to genome assembly, the characteristics of the genome can be estimated from the quality-controlled MGI second-generation data. K-mer analysis is employed to predict the genome size, heterozygosity, and repetitive sequence information. The kmerfreq program was used for statistics, and the K-mer analysis showed a total K-mer count of 61,709,089,203, with the estimated genome size of the *I. elongata* being approximately 793 Mb, corrected to 785 Mb. The genome heterozygosity was 1.2%, and the proportion of repetitive sequences

Data type	Sequencing technology	Sequencing platform	Library size (bp)	Clean data (Gb)
Genome	MGI	DNBSEQ-T7	150	78.92
Genome	PacBio	PacBio Sequel IIe	20,000	35.37
Genome	Hi-C	DNBSEQ-T7	150	69.52
RNA	MGI	DNBSEQ-T7	150	28.15

Table 1. The statistical information of the sequenced genome data of *I. elongata*.

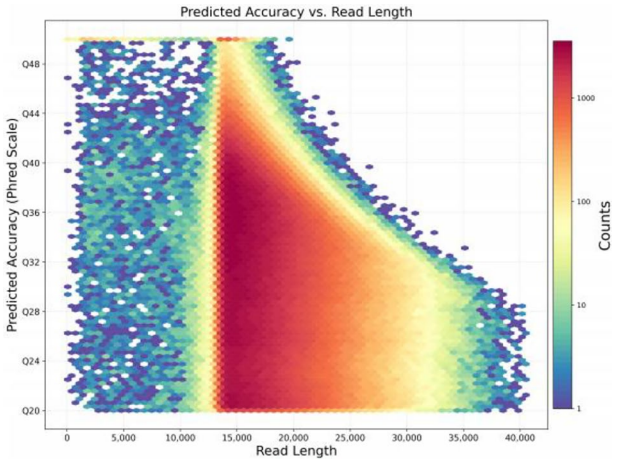

Fig. 1 2D statistical chart of the quality and length of the third-generation PacBio sequencing results of *I. elongata* genome.

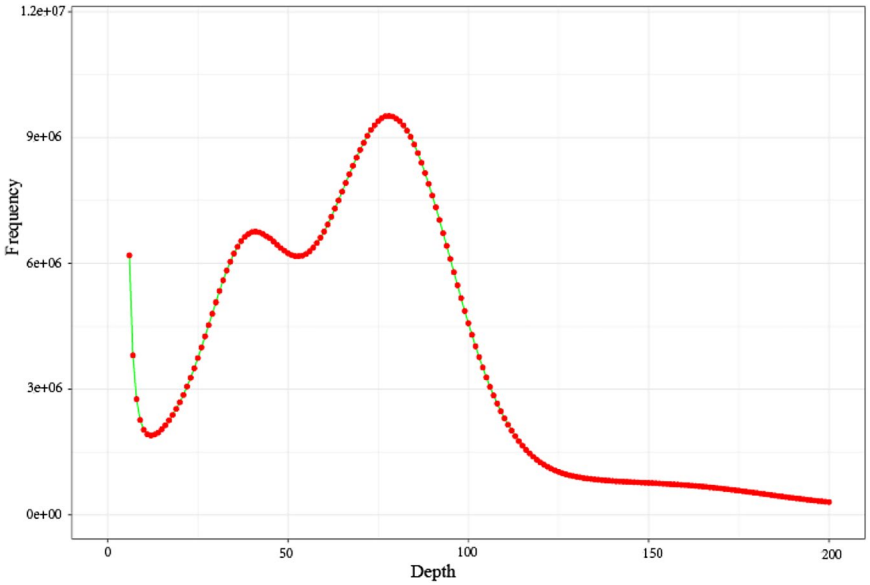

Fig. 2 Frequency and depth distribution plot of the *I. elongata* genome surveyed through genome analysis.

was 39.6%. The K-mer frequency distribution results were visualized (Fig. 2), revealing two distinct peaks in the K-mer distribution curve, which indicates a high level of heterozygosity in the *I. elongata* genome.

In this study, the PacBio sequencing data were assembled using the nextDenovo software. The initial genome assembly of the *I. elongata* was obtained, with a size of 831.89 Mb and containing 980 contigs, with a contig N50 of 4.73 Mb (Table 2). Subsequently, the genome was corrected and refined using both third-generation and second-generation data, resulting in the final corrected genome, which had a size of 839.57 Mb. In the corrected genome, the contig N50 reached 4.82 Mb, and the longest sequence was 16,686,348 bp (Table 2).

In this study, Hi-C sequencing technology was employed to assist in the chromosomal-level assembly of the *I. elongata* genome. The quality-controlled Hi-C paired-end data were aligned with the contig-level assembly, and a chromosome interaction map was constructed for visualization. The results showed that 769.94 Mb of

Assembly method	Genome length(bp)	Sequence number	Max length(bp)	N50(bp)	N90(bp)	GC content (%)
nextDenovo	831,898,742	980	16,476,865	4,739,510	357,820	45.56
nextDenovo + racon	842,533,391	975	16,740,204	4,824,529	362,298	45.6
nextDenovo + racon + pilon	839,572,073	975	16,686,348	4,816,715	360,601	45.61
nextDenovo + racon + pilon + Hi-C	815,087,891	914	45,402,468	32,609,743	27,815,500	44.59

Table 2. Assembly results of *I. elongata* genome.

Chromosome name	Sequence length (bp)	Contig number
Chr1	11,959,159	25
Chr2	41,992,466	36
Chr3	27,815,499	42
Chr4	33,666,499	32
Chr5	32,235,275	43
Chr6	33,220,880	45
Chr7	32,872,489	26
Chr8	45,402,467	29
Chr9	29,294,746	50
Chr10	32,246,499	41
Chr11	30,230,675	39
Chr12	34,854,648	43
Chr13	28,158,525	35
Chr14	33,954,452	53
Chr15	33,671,872	44
Chr16	34,975,499	38
Chr17	28,646,126	59
Chr18	35,268,841	62
Chr19	34,716,029	22
Chr20	32,609,742	46
Chr21	31,107,250	53
Chr22	28,707,775	68
Chr23	32,105,143	37
Chr24	29,381,265	47

Table 3. Statistics of the chromosome-level assembly results of *I. elongata* genome.

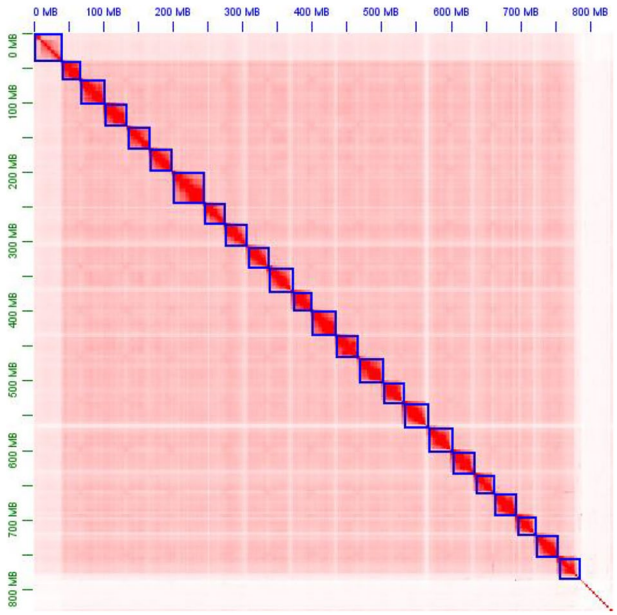

Fig. 3 Hi-C interaction heatmap of *I. elongata* chromosome genome.

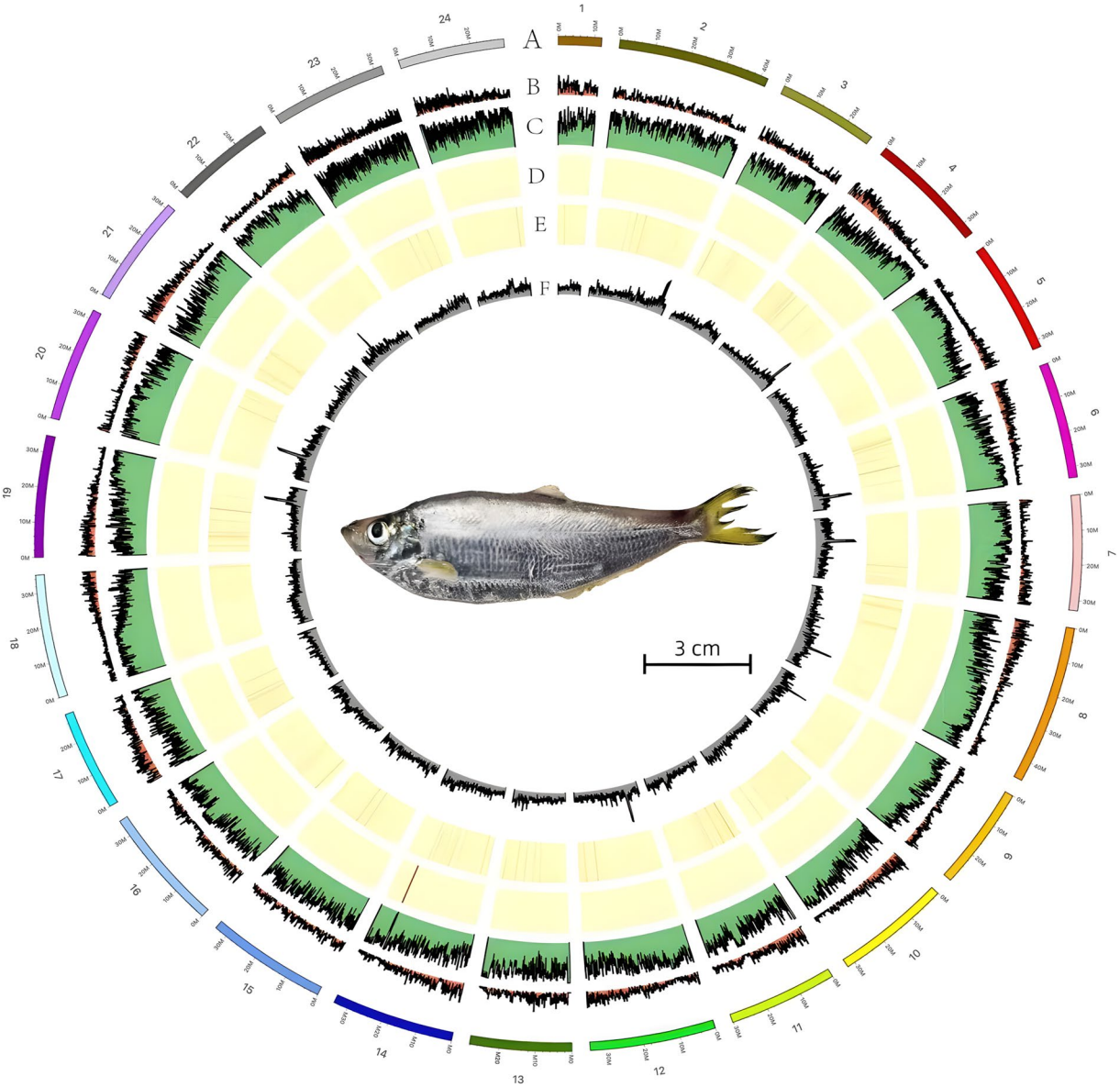

Fig. 4 Circle map of *I. elongata* genome at the chromosome level. From outside to inside: (**A**) Chromosome length, (**B**) GC content distribution, (**C**) Second-generation sequencing depth distribution, (**D**) Third-generation sequencing depth distribution, (**E**) Gene density distribution, (**F**) Distribution of long tandem repeat sequences.

Data	Mapping rate (%)	Average sequencing depth (×)	Coverage at least 20X (%)
MGI	96.54	99.87	99.65
PacBio	97.56	44.53	99.86

Table 4. Alignment statistics of second-generation and third-generation sequencing data.

contig sequences were anchored to 24 chromosomes (Fig. 3), with chromosome lengths ranging from 11.95 Mb to 45.42 Mb (Table 3; Fig. 4). Further statistical analysis of the assembly, conducted with Python scripts, revealed that the *I. elongata* genome size is 815 Mb, the scaffold N50 is 32.61 Mb, and the chromosome anchoring rate is 96.54%.

Assembly quality evaluation. The genome of *I. elongata* was assembled at the chromosome level with a contigN50 and scaffoldN50 of 4.82 Mb and 32.61 Mb respectively, showing good sequence continuity. At the same time, the comparison rate of the second- and third-generation data, after quality assessment, to the assembled genome was 96.54%, and the comparison rate of the third-generation data was 97.56%, with coverage of 99.65%

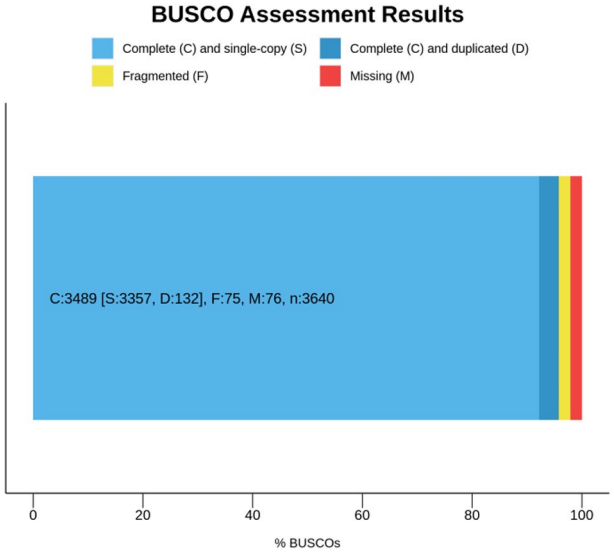

Fig. 5 BUSCO evaluation results of *I. elongata* genome at the chromosome level.

Type	Repbase TEs		TE protiens		*De novo*		Combined TEs	
	Length (bp)	Percentage in genome (%)	Length(bp)	Percentage in genome (%)	Length (bp)	Percentage in genome (%)	Length (bp)	Percentage in genome (%)
DNA	164,064,903	19.46	30,669,674	3.63	104,296,070	12.37	201,582,341	23.91
LINE	2,861,067	0.33	6,943,639	0.82	20,491,761	2.43	6,931,114	0.82
SINE	1,238,190	0.14	3,849,582	0.45	3,854,493	0.46	3,762,411	0.44
LTR	1,902,236	0.22	5,537,862	0.65	18,559,790	2.2	5,531,364	0.65
Satellite	212,188	0.02	647,787	0.07	1,038,808	0.12	646,759	0.07
Other	157,728	0.01	471,952	0.05	255,573	0.03	913,536	0.08
Unknow	122,583	0.01	414,838	0.04	637,718	0.08	411,325	0.04
Total	244,069,805	28.94	90,743,711	10.76	168,418,176	19.98	295,757,806	35.08

Table 5. The statistical analysis of repetitive sequences in the genome of *I. elongata*.

and 99.86% respectively (Table 4). In addition, BUSCO software was used to compare the genome of *I. elongata* with the phylogenetic database, and 3,640 conserved genes of Actinopterygii were selected for comparison. The results showed that 3,489 genes (95.8%) could be completely matched to the genome of *I. elongata*, of which 92.2% were single-copy genes and 3.6% were multi-copy genes. Besides, 75 genes (2.1%) were only partially matched and 76 genes (2.1%) were not matched (Fig. 5).

Repeat sequences and non-coding RNA. The repeat sequence obtained by homologous prediction strategy annotation was 90.74 Mb (10.76%), and the repeat sequence obtained by *de novo* prediction strategy annotation was 244.06 Mb (28.94%). Combining the results of the two strategies annotation, a total of 295.7 Mb was obtained, accounting for 35.08% of the genome sequence of *I. elongata* (Table 5). Common repeats fall into two main categories: Interspersed repeats and Tandem Repeats (TRs). In the genome of *I. elongata*, tandem repeats accounted for 7.41%, Simple repeat 7.12% and Satellite 0.07%; In scattered repeats, 23.91% of DNA transposons (TEs), 0.82% of long scattered transposons (lines), 0.44% of short scattered transposons (SINE) and 0.65% of long terminal repeats (LTR) were found (Table 5). In addition, some repeat sequences have not been successfully compared with existing databases and are classified as Unknown repeat types. The length of such sequences is 0.41 Mb, accounting for 0.04% in the whole genome. With the help of tRNAscan-SE, 705 tRNAs were identified through structural prediction. Based on the Rfam database, 794 miRNAs, 105 rRNAs and 564 snRNAs were annotated; 2,826 lncRNAs were annotated using GETA (Table 6).

Gene structure and function annotation. Through annotating 46,804 genes by using AUGUATUS software based on the *de novo* prediction method, the AUGUSTUS model was trained with an accuracy of 66.3%, and at the nucleotide level, its sensitivity reached 88.2% and specificity reached 79.5%. A total of 33,951 genes were predicted through homology, including 10,747 complete genes, 7,800 5′ partial genes, 3,434 3′ partial genes, and 11,970 internal genes. The alignment rate of the transcriptome data reached 99.05%, and 25,923 genes were predicted based on the transcriptome, including 14,022 complete genes, 5,437 5′ partial genes, 3,157 3′ partial genes, and 3,307 internal genes. By combining the three gene prediction methods and filtering the gene models, a total of 26,381 structural genes were identified, of which 5,777 had variable splicing, 12,363 were from AUGUSTUS

Type		Copy	Average length (bp)	Total length (bp)	Percentage in genome (%)
miRNA		794	89	115,152	0.01354
tRNA		705	76	306,757	0.03635
rRNA	rRNA	105	125	289,145	0.03422
	18S	5	1376	33,156	0.00394
	5.8S	3	162	11,757	0.00135
	5S	364	117	723,064	0.08516
snRNA	snRNA	564	138	88,521	0.00105
	CD-box	114	143	212,188	0.02517
	HACA-box	105	136	157,728	0.01873
	splicing	291	145	800,034	0.0949
	scaRNA	8	248	87,630	0.00103
lncRNA		2,826	252	3,762,411	0.44631

Table 6. Annotation of non-coding RNA genes in *I. elongata* genome.

Gene set		Number	Average gene length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length (bp)	Average intron length (bp)
De novo	Genscan	27,542	19,862	1,628	7.63	207.72	2,768
	AUGUSTUS	46,804	18,867	1,573	7.12	168.23	2,546
Homolog	*Clupea harengus*	48,653	19,256	1,093	6.35	175.42	3,232
	Denticeps clupeoides	45,276	20,358	1,386	6.1	178.31	3,165
	Alosa alosa	43,265	19,382	1,036	6.32	182.54	4,625
	Oncorhynchus keta	44,895	18,644	1,156	5.69	172.49	3,896
	Danio rerio	47,618	20,753	1,148	5.14	181.32	3,278
RNAseq		25,923	18,952	1,489	11.36	215.63	2,654
BUSCO		4536	17,254	1,536	13.58	156.14	1,962
EVM		25,693	17,356	1,263	8.62	235.38	2,143
GETA		26,381	17,243	1,365	9.53	242.42	1,854

Table 7. Predicted protein-coding genes in *I. elongata* genome.

Category	Number	Percent (%)
InterPro	23,352	88.52
GO	17,999	68.23
KEGG	22,791	86.39
Swissprot	17,206	65.22
TrEMBL	24,099	91.35
NR	21,057	79.82
KOG	21,801	82.64
Annotated	24,596	93.23
Unannotated	1,786	6.77
Total	26,381	100

Table 8. Table of functional annotations for protein-coding genes.

predictions, 10,790 were from transcripts, and 3,228 were from homolog. By calculation, the average length of CDS was 182.6 bp, the average length of genes is 13,451.3 bp, each gene contains an average of 8.65 introns and 9.64 exons, the average length of introns was 1,290.81 bp, and the average length of exons was 235.86 bp (Table 7). Functional annotation was carried out based on the predicted results of genomic structure of *I. elongata*, and the predicted protein sequences were compared with InterPro, GO, KEGG, Swissprot, TrEMBL, NR and KOG databases. The results showed that the functions of 24,596 genes were successfully annotated, accounting for 93.23% of the total number of genes (Table 8).

Data Records

The genomic Illumina, PacBio, Hi-C, and transcriptomic sequencing data were deposited in the Sequence Read Archive (SRA) at NCBI under the accession numbers SRP530534[49].

The final chromosome assembly was deposited in the GenBank under the accession JBLRXZ000000000[50].

The annotation files have been uploaded to the figshare database, with the DOI number: https://doi.org/10.6084/m9.figshare.28151450.v1[51].

Technical Validation

High-throughput sequencing was used to obtain the raw genome data, which were then subjected to quality control and filtering to generate clean data. Prior to assembly, a survey analysis was conducted on the second-generation sequencing data of the *I. elongata* genome. The results of the survey analysis indicated that the *I. elongata* genome size is 785 Mb, with a heterozygosity rate of 1.2% and a repetitive sequence proportion of 39.6%.

In this study, third-generation PacBio sequencing data were used for genome assembly, followed by sequence correction using both second- and third-generation data to generate an initial genome draft, with a contig N50 size of 4.82 Mb. Subsequently, Hi-C data were employed for assisted assembly, enabling the *I. elongata* genome to be assembled to the chromosomal level. The final genome size was 815 Mb, and the construction of the chromosome interaction map clearly revealed that contig sequences were anchored to 24 chromosomes (Fig. 3), with an anchoring rate of 96.54%. The scaffold N50 was 32.61 Mb (Table 2). Genome evaluation using BUSCO indicated a gene completeness rate of 95.8% (Fig. 5).

Based on the *I. elongata* genome assembly, further annotation analysis was conducted, which identified 295.7 Mb of repetitive sequences, accounting for 35.08% of the *I. elongata* genome. This includes 7.41% of tandem repeats, 3.34% of interspersed repeats, and 0.04% of repetitive sequences that could not be annotated to any homologous databases (Table 5). Structural annotation of the *I. elongata* genome revealed a total of 26,381 protein-coding genes (Table 7), of which 93.23% were successfully annotated with functional information (Table 8).

Code availability

No special codes or scripts were used in this work, and data processing was done based on the protocols and manuals of the corresponding bioinformatics software.

Received: 26 December 2024; Accepted: 7 March 2025;
Published online: 21 April 2025

References

1. Zhang, J., Takita, T. & Zhang, C. Reproductive biology of Ilisha elongata (Teleostei: Pristigasteridae) in Ariake Sound, Japan: Implications for estuarine fish conservation in Asia. *Estuarine, Coastal and Shelf Science* **81**, 105–113, https://doi.org/10.1016/j.ecss.2008.10.013 (2009).
2. Whitehead, P. J. Clupeoid fishes of the World (Suborder Clupeoidei): An annotated and illustrated catalogue of the herrings, sardines, pilchards, sprats, shads, anchovies and wolf herrings. Part 1. Chirocentridae, Clupeidae and Pristigasteridae. *FAO Fisheries Synopsis* **125**, 1, https://doi.org/10.1093/ices/fct001 (1985).
3. Wang, X. *et al*. Early life history of Ilisha elongata (Pristigasteridae, Clupeiformes, Pisces) in Ariake Sound, Shimabara Bay, Japan. *Plankton and Benthos Research* **16**, 210–220, https://doi.org/10.3800/pbr.16.210 (2021).
4. Cossins, A. R. & Crawford, D. L. Fish as models for environmental genomics. *Nature Reviews Genetics* **6**, 324–333 (2005).
5. Kasahara, M. *et al*. The medaka draft genome and insights into vertebrate genome evolution. *Nature* 447, 714-719 (2007).
6. Star, B. *et al*. The genome sequence of Atlantic cod reveals a unique immune system. *Nature* **477**, 207–210, https://doi.org/10.1038/nature10342 (2011).
7. Desjardins, P. & Conklin, D. NanoDrop microvolume quantitation of nucleic acids. *Journal of visualized experiments: JoVE*, e2565, https://doi.org/10.3791/2565 (2010).
8. Ponti, G. *et al*. The value of fluorimetry (Qubit) and spectrophotometry (NanoDrop) in the quantification of cell-free DNA (cfDNA) in malignant melanoma and prostate cancer patients. *Clinica Chimica Acta* **479**, 14–19, https://doi.org/10.1016/j.cca.2018.01.007 (2018).
9. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890, https://doi.org/10.1093/bioinformatics/bty560 (2018).
10. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120, https://doi.org/10.1093/bioinformatics/btu170 (2014).
11. Hufnagel, D. E., Hufford, M. B. & Seetharam, A. S. SequelQC: Analyzing PacBio Sequel Raw Sequence Quality. *bioRxiv* (2019).
12. Liu, B. *et al*. Estimation of genomic characteristics by analyzing k-mer frequency in *de novo* genome projects. *arXiv: Genomics*, https://doi.org/10.1016/j.gpb.2013.05.002 (2013).
13. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255, https://doi.org/10.1093/bioinformatics/btz891 (2020).
14. Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome research* **27**, 737–746, https://doi.org/10.1101/gr.214270.116 (2017).
15. Walker, B. J. *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963, https://doi.org/10.1371/journal.pone.0112963 (2014).
16. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell systems* **3**, 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
17. Dudchenko, O. *et al*. *De novo* assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95, https://doi.org/10.1126/science.aal3327 (2017).
18. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell systems* **3**, 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
19. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075, https://doi.org/10.1093/bioinformatics/btt086 (2013).
20. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arxiv preprint arxiv:1303.3997* (2013).
21. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
22. García-Alcalde, F. *et al*. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* **28**, 2678–2679 (2012).
23. 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).

24. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
25. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic acids research* **35**, W265–268, https://doi.org/10.1093/nar/gkm286 (2007).
26. Chen, N. Using Repeat Masker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* **5**, 4.10. 11–14.10. 14 (2004).
27. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**, 573–580, https://doi.org/10.1093/nar/27.2.573 (1999).
28. Tempel, S. Using and understanding RepeatMasker. *Mobile genetic elements* **859**, 29–51, https://doi.org/10.1007/978-1-61779-603-6_2 (2012).
29. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335–1337, https://doi.org/10.1093/bioinformatics/btp157 (2009).
30. Chen, Y., Ye, W., Zhang, Y. & Xu, Y. High speed BLASTN: an accelerated MegaBLAST search tool. *Nucleic acids research* **43**, 7762–7768, https://doi.org/10.1093/nar/gkv784 (2015).
31. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic acids research* **49**, 9077–9096, https://doi.org/10.1093/nar/gkab688 (2021).
32. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59, https://doi.org/10.1186/1471-2105-5-59 (2004).
33. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research* **34**, W435–439, https://doi.org/10.1093/nar/gkl200 (2006).
34. Altschul, S. F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389–3402, https://doi.org/10.1093/nar/25.17.3389 (1997).
35. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome research* **14**, 988–995, https://doi.org/10.1101/gr.1865504 (2004).
36. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nature methods* **12**, 357–360, https://doi.org/10.1038/nmeth.3317 (2015).
37. Haas, B. J. *et al*. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* **8**, 1494–1512, https://doi.org/10.1038/nprot.2013.084 (2013).
38. Zhu, W. & Buell, C. R. Improvement of whole-genome annotation of cereals through comparative analyses. *Genome research* **17**, 299–310 (2007).
39. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
40. Zdobnov, E. M. & Apweiler, R. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).
41. Consortium, G. O. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research* **32**, D258–D261 (2004).
42. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* **35**, D61–65, https://doi.org/10.1093/nar/gkl842 (2007).
43. Gasteiger, E., Jung, E. & Bairoch, A. SWISS-PROT: connecting biomolecular knowledge via a protein database. *Current issues in molecular biology* **3**, 47–55 (2001).
44. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research* **31**, 365–370, https://doi.org/10.1093/nar/gkg095 (2003).
45. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27–30, https://doi.org/10.1093/nar/28.1.27 (2000).
46. Tatusov, R. L. *et al*. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41, https://doi.org/10.1186/1471-2105-4-41 (2003).
47. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature methods* **18**, 366–368 (2021).
48. Conesa, A. *et al*. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676, https://doi.org/10.1093/bioinformatics/bti610 (2005).
49. Gao, L. *et al*. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRP530534 (2024).
50. Niu, X. Chromosome assembly of Ilisha elongata. *GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_048126385.1 (2025).
51. Niu, X. Annotation results(Ilisha elongata). *Figshare* https://doi.org/10.6084/m9.figshare.28151450.v1 (2025).

Acknowledgements

The study was supported by the Zhejiang Provincial Natural Science Foundation of China (LY22D060001&LY20C190008); Science and Technology Programme of Zhejiang Institute of Marine and Fisheries Research (No. HYS-CZ-202409); National Natural Science Foundation of China (NSFC) (NO.41806156); Key research and development projects in Xizang (XZ202301ZY0012N); Science and Technology Project of Zhoushan (2020C21016).

Author contributions

B.J.L., S.T., J.S.L., Y.F.L. D.D.X. conceived and designed the research. B.J.L., X.Y.N., C.Z., S.X.Z., L.X.G., M.Z.H., T.B.F., J.H.W., C.X.J., S.S.K., Y.F.L., D.D.X. conducted experiments, analyzed data, and wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.X. or Y.L.