scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly and annotation of pawak croaker (*Pennahia pawak*)

Lihua Jiang[1,3], Peng Zheng[2,3], Jialang Zheng[1], Yifan Liu[1], Weihua Song[1], Shun Chen[1], Wangyang Jin[1] & Xiaojun Yan[1,2] ✉

The diversity of fish coloration is a fascinating scientific question. The pawak croaker (*Pennahia pawak*) is distinguished by its silver-white colour, setting it apart from other Sciaenidae species. However, the genomic information of this genus remains largely unexplored. This study aims to advance our understanding by assembling and annotating the genome of *P. pawak* at the chromosome level using multi-platform sequencing data. The assembled genome size of *P. pawak* is 613.02 Mb, closely matching the estimated size of 570.32 Mb from 21-mer analysis. The assembly features a scaffold N50 of 27.09 Mb, with 24 chromosomes successfully constructed using Hi-C technology, achieving a mounting rate of 99.31%. Genome annotation revealed that 26.20% of the genome consists of repetitive sequences and identified 26,361 protein-coding genes, of which 25,885 have functional annotations. This chromosome-level genome assembly of *P. pawak* provides valuable resources for comparative genomic studies within the Sciaenidae family and offers foundational data for researchers to understand its unique silver-white pigmentation.

Background & Summary

The Sciaenidae family, within the order Perciformes, comprises approximately 270 species across 70 genera distributed globally[1] Fig. 1. These species are of significant commercial importance, primarily serving as food fish. A unique characteristic of sciaenids is their ability to produce croaking sounds[1], which play various ecological roles. While Sciaenidae species share similar morphological traits, their body color is highly variable, including black, yellow, multi-colored, and color-changing patterns[2]. In contrast to other Sciaenidae, species in the genus *Pennahia* generally exhibit a monotone and stable body color, primarily presenting a silver-white hue[3]. This distinct coloration makes them ideal candidates for studying fish body colour variation, particularly in the context of countershading camouflage and light response adaptations. To better understand these traits, we sequenced and investigated a high-quality genome of this species.

We integrated multiple sequencing technologies, including Illumina sequencing, PacBio circular consensus sequencing (CCS), and Hi-C techniques. The final assembly size of the *P. pawak* genome was 613.02 Mb, with 99.31% of the contigs anchored to 24 chromosomes. The contig N50 length was 26.06 Mb, and the scaffold N50 length was 27.09 Mb. The genome contained 26.20% repetitive sequences. We predicted a total of 25,885 protein-coding genes, with 98.19% functionally annotated, reflecting the high quality of the assembled genome. These comprehensive genomic resources will facilitate future research into the molecular mechanisms of fish body color polymorphism and the adaptive mechanisms of *P. pawak*.

Methods

Sampling. The *P. pawak* sample was collected from the intertidal zone of Zhoushan, Zhejiang Province, China (30.65°N, 122.78°E). The dissection of the sample was conducted in a sterilized environment, where organs including muscle, liver, and brain were sampled and snap-frozen in liquid nitrogen for nucleic acid extraction. All anatomical procedures complied with ethical regulations set by the Institutional Animal Care and Use Committee of Zhejiang Ocean University, Zhejiang, China (Protocol Number: 2023082). The size and integrity of

[1]National Engineering Research Center for Marine Facilities Aquaculture, Zhejiang Ocean University, Zhoushan, 316022, China. [2]Marine Science and Technology College, Zhejiang Ocean University, Zhoushan, 316022, China. [3]These authors contributed equally: Lihua Jiang, Peng Zheng. ✉e-mail: yanxj@zjou.edu.cn

Fig. 1 Pawak croaker, *P. pawak* was collected from the intertidal zone of Zhoushan, Zhejiang Province, China.

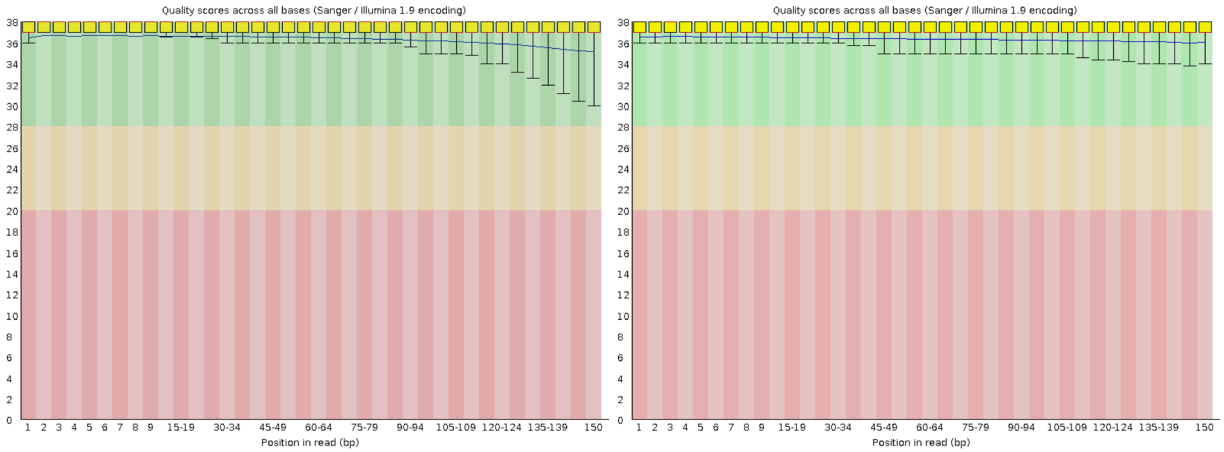

Fig. 2 Base quality examined for Illumina PE150 reads using FastQC software. Note: X-axis represents the position in reads, and the Y-axis represents the quality scores across the bases. The colored area illuminates the low (pink), middle (yellow) and high (green) quality scores of the bases.

Library name	Illumina reads number	Illumina reads base (bp)	Sequencing strategy
LN312137207	560,228,144	84,034,221,600	PE150

Table 1. Statistics of the genome sequencing data generated from Illumina NovaSeq 6000 platform.

HiFi reads number	HiFi reads base (bp)	Max HiFi reads length (bp)	Average HiFi reads length (bp)	HiFi reads N50 (bp)
2,830,726	50746554202	60,280	17,927	18,399

Table 2. Statistics of the sequencing reads generated from PacBio Sequel II platform.

Library name	Raw reads number	Raw Bases (bp)	Sequencing strategy
LN312137209	492,989,010	73,948,351,500	PE150

Table 3. Statistics of the Hi-C sequencing data generated from Illumina NovaSeq 6000 platform.

the extracted DNA and RNA were evaluated using 1% agarose gel electrophoresis. Additionally, the concentration and purity of DNA and RNA were analyzed using a Nanodrop 2000c ultraviolet spectrophotometer.

Library construction & sequencing. Genomic DNA was extracted from muscle samples using the E.Z.N.A.® Tissue DNA Kit (Omega Bio-tek, USA). For HiFi library preparation and sequencing, the SMRTbell prep kit 3.0 was used, and sequencing was performed in CCS mode on the PacBio Sequel II system (Pacific Biosciences, USA), following the manufacturer's protocols. For Illumina sequencing, a short-fragment library with an insert size of 300–500 bp was prepared using the TruSeq™ Nano DNA Sample Prep Kit (Illumina,

Read_number	Len_all(bp)	Source
216,665,402	32,499,810,300	Brain
217,310,464	32,596,569,600	Liver
244,561,412	36,684,211,800	Muscle
678,537,278	101,780,591,700	Total

Table 4. Statistics of RNA-seq data generated from Illumina NovaSeq 6000 platform.

Term	HiFi contigs		Hi-C scaffolds	
	Size (bp)	Number	Size (bp)	Number
N90	18,082,980	21	20,986,166	21
N80	23,115,463	18	23,332,860	18
N70	24,980,551	15	23,889,661	16
N60	25,136,919	13	25,856,863	13
N50	26,061,895	11	27,090,935	11
Max length (bp)	50,796,522		32,647,052	
Total length (bp)	617,293,045		617,295,437	
Total number	53		35	
Number ≥10 Kbp	53		35	

Table 5. Statistics of the assembled genome based on the HiFi and Hi-C data. Note: HiFi contigs represent the contigs assembled using HiFi data, and Hi-C scaffolds represent the scaffolds after the chromosome assembly.

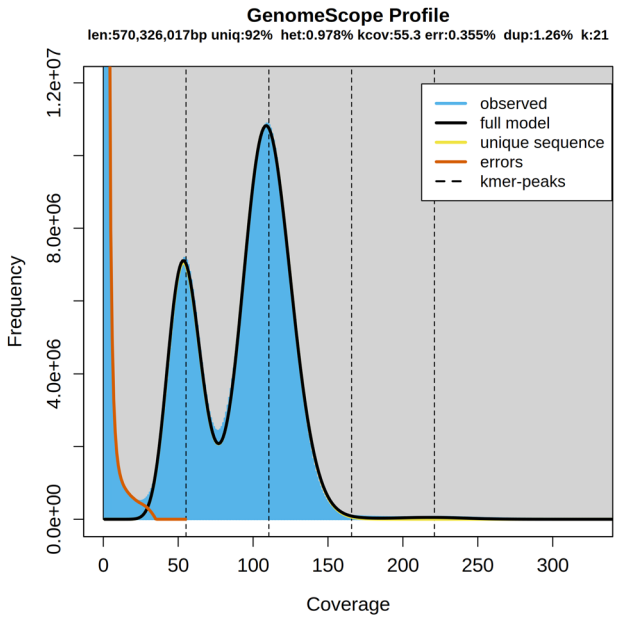

Fig. 3 K-mer analysis for the genome size evaluation in *P. pawak*. Note: The distribution of 21-mer frequency in *P. pawak* genome was shown. The X-axis represents the k-mer depth, and the Y-axis represents the frequency of the k-mer for a given depth. The first, second and third peaks in the figure corresponded to the heterozygous homozygous, and repeated Kmers, respectively.

USA). The library was purified with AMPure XP Beads (Beckman Coulter, USA) and sequenced on an Illumina NovaSeq 6000 platform (Illumina, USA) to generate 150-bp paired-end (PE150) reads. Hi-C library preparation and sequencing were performed with muscle tissue using a Dovetail Hi-C Core Kit (Dovetail Genomics, USA) following protocol instructions. The library was assessed with an Agilent 2100 Bioanalyzer (Agilent, USA) to ensure sufficient concentration and an insert size of 300–500 bp. Sequencing was done on an Illumina NovaSeq. 6000 platform (Illumina, USA) to generate PE150 reads. For transcriptome analysis, muscle, brain, and liver tissues of *P. pawak* were processed individually to generate separate transcriptomic data for each tissue. Total RNA was extracted using an E.Z.N.A.® HP Total RNA Kit (Omega Bio-tek, USA) following the manufacturer's instructions. The quality and concentration of the extracted RNA were assessed using a NanoDrop® Series (Thermo Scientific, USA) and an Agilent 2100 Bioanalyzer. Three cDNA libraries(muscle, liver, brain; Table 4)

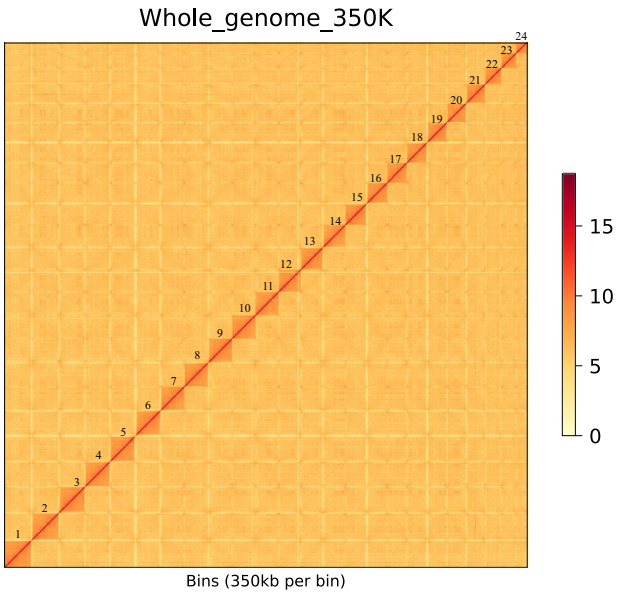

Fig. 4 A heatmap of chromosome interaction in *P. pawak*, the blocks represent the 24 pseudochromosomes. Note: The color bar illuminates the contact density from white (low) to red (high).

were prepared from total RNA using the NEBNext® Ultra™ RNA Library Prep Kit (New England Biolabs, USA) and sequenced on an Illumina NovaSeq. 6000 platform (Illumina, USA).

Quality control of raw sequencing data. All raw sequencing data were processed to remove adaptors, low-quality bases, and duplicate reads using various strategies based on the platform. For Illumina PE150 reads, Trimmomatic v0.39[4] was used with parameters set as "ILLUMINACLIP:adapters.fa:2:30:10 SLIDINGWINDOW:4:15 MINLEN:75" to remove adaptor sequences, low-quality reads, and short sequences. The quality assessment of the cleaned data was conducted using FastQC software v0.11.3[5], which showed very high base scores, indicating high quality of the sequencing data(Fig. 2). The same filtering methods and parameter settings were applied to the Hi-C and RNA-seq data. The sequencing data output included 84.03 Gb of Illumina PE150 reads (Table 1), 50.75 Gb of HiFi long reads with an N50 length of 18.40 Kb (Table 2), 73.95 Gb of Hi-C sequencing data (Table 3), and transcriptome data consisting of 32.60 Gb from the liver, 36.68 Gb from muscle, and 32.50 Gb from the brain (Table 4).

Genome size estimation. Illumina PE150 reads were used for genome size estimation based on k-mer analyses. Filtered high-quality Illumina PE150 reads were analyzed using jellyfish software v2.2.6[6] with parameters set as "-C -m 21".The genome size was calculated using the formula: genome size = TKN21-mer/PKFD21-mer, where TKN21-mer represents the total number of k-mers and PKFD21-mer represents the peak frequency depth of the 21-mer. The analysis revealed an estimated genome size of approximately 570.33 Mb for *P. pawak*. We used GenomeScope software v2.0.1[7] with parameters set as "-k 21 -p2" to generate the k-mer frequency distribution plot. The k-mer distribution of the genome exhibited three peaks (Fig. 3), which likely correspond to heterozygous, homozygous, and repeated k-mers, as commonly observed in many other teleost fishes[8,9].

Genome assembly. Illumina PE150 reads, which have a relatively higher error rate at the single-base level compared to HiFi long reads, underwent error correction using BLESS software v0.6.3[10] with parameters set as "-kmerlength 21" to produce clean reads.These corrected reads were then assembled into the genome using Hifiasm software v0.19.8-r603[11] with parameters set to "-t 50 -l 0". Redundant heterozygous contigs were identified and removed using Purge_haplotigs software v1.1.2[12]. The preliminary assembly resulted in a genome size of 617.29 Mb, with 53 contigs and a contig N50 of 26.06 Mb (Table 5). Hi-C sequencing data were employed for chromosome assembly using ALLHIC software v0.9.8[13], using parameters set to "-e GATC -k 8 -m 50" Interaction maps were generated and error-corrected using Juicer software[14] and JuiceBox software v1.11.08[15] (Fig. 4). This process produced 24 chromosomes with a scaffold N50 of 27.09 Mb (Table 5; Figs. 4, 5)) and an assembly rate of contigs into chromosomes reaching 99.31% (Table 6). The chromosome number aligns with closely related species such as *Collichthys lucidus*[8], *Miichthys miiuy*[16], and *Larimichthys crocea*[17]. Heatmap analysis indicated that all 24 pseudochromosomes were distinguishable (Fig. 4), with strong interaction signals around the diagonal, showcasing the high quality of the genome assembly.

Genome evaluation. The completeness and accuracy of the *P. pawak* genome assembly were thoroughly assessed. The Contig N50 was 26.06 Mb, and the Scaffold N50 was 27.09 Mb. BUSCO analysis, conducted using BUSCO software v5.4.7[18] against the Actinopterygii single-copy orthologous gene library, identified 3,640 core genes, including 3,547 complete genes (97.45%), 3,515 single-copy genes (96.57%), 32 multi-copy genes (0.88%), 23 fragmented genes (0.63%), and 70 missing genes (1.92%)(Table 7). Short reads mapping ratio analysis involved

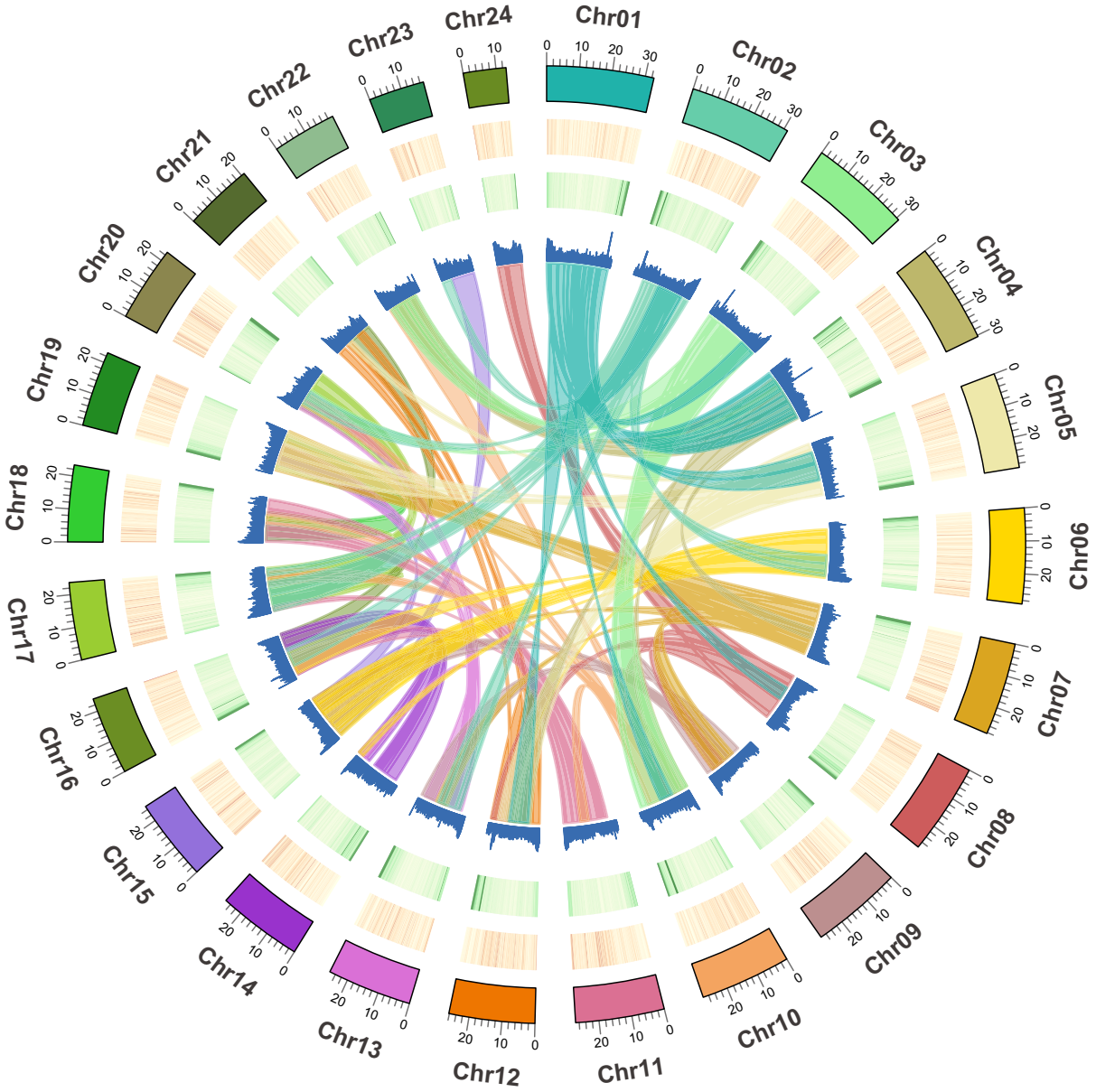

Fig. 5 Circos plot indicating gene density, repetitive sequences, GC content, and colinear relationship among chromosomes of the *P. pawak* genome assembly.

building a genome index using BWA-MEM v0.7.17-r1188[19] with the parameters set to "-a bwtsw" resulting in a total mapping rate of 97.82%, a paired mapping rate of 96.76%, and a properly paired mapping rate of 83.98% for Illumina PE150 reads(Table 8). Additionally, transcript mapping ratio analysis, performed by building a genome index with HISAT2 v2.1.0[20] and mapping transcripts to the genome, showed that 624,504,850 reads were mapped with a mapping rate of 97.13% (Table 9). These analyses indicated a high-quality chromosome-level assembly of the *P. pawak* genome.

Annotation of repetitive sequences. Repetitive elements in the *P. pawak* genome were annotated using a combination of homology prediction with the Repbase library[21] and *de novo* prediction based on self-sequence alignment and repetitive sequence features. Tandem repeats were annotated using Tandem Repeat Finder v4.09[22] with parameters set as "Match = 2, Mismatch = 7, Delta = 7, PM = 80, PI = 10, Minscore = 50, MaxPeriod = 2000 -d -h". Transposable elements (TEs) were identified through *de novo* prediction at both DNA and protein levels. For the DNA level, RepeatModeler v1.0.11[23] (-database redb -pa 10) was employed to build a *de novo* repeat library, and RepeatMasker v4.0.7[24] (-nolow -no_is -norna -pa 2) was used to identify homologous repeats against the *de novo* library and Repbase. At the protein level, RepeatProteinMask v4.0.7 was utilized to search for TEs in its protein database. The annotation results of all repetitive sequences were merged into a final dataset. This comprehensive annotation revealed that 160.64 Mb of sequences, accounting for 26.20% of the *P. pawak* genome, were

Sequence ID	Sequence Length (bp)	Sequence ID	Sequence Length (bp)
Chr1	32,647,052	Chr13	25,856,863
Chr2	31,171,955	Chr14	25,850,940
Chr3	30,838,672	Chr15	25,109,490
Chr4	30,427,248	Chr16	23,889,661
Chr5	29,282,019	Chr17	23,702,058
Chr6	29,089,116	Chr18	23,332,860
Chr7	28,304,032	Chr19	23,115,463
Chr8	27,890,477	Chr20	23,050,519
Chr9	27,803,222	Chr21	20,986,166
Chr10	27,746,100	Chr22	19,555,964
Chr11	27,090,935	Chr23	17,054,624
Chr12	26,174,222	Chr24	13,054,061
Total contig length at chromosomes (bp)		613,023,719	
Total contig length (bp)		617,295,437	
Chromosome/total		99.31%	

Table 6. Summary of the chromosome assemblies for *P.Pawak* based on Hi-C data.

Type	Proteins	Percentage (%)
Complete BUSCOs (C)	3,547	97.45
Complete and single-copy BUSCOs (S)	3,515	96.57
Complete and duplicated BUSCOs (D)	32	0.88
Fragmented BUSCOs (F)	23	0.63
Missing BUSCOs (M)	70	1.92
Total BUSCO groups searched	3,640	100.00

Table 7. Results of the BUSCO assessment for genome assembly in *P.Pawak*.

Mapping rate (%)	Paired mapping rate (%)	Properly paired rate (%)
97.82	96.76	83.98

Table 8. The mapping ratio of the short reads to the assembled genome of *P.Pawak*.

Mapping rate (%)	Total reads number	Number of alignments
97.13	624,504,850	606,563,445

Table 9. The mapping ratio of transcript to the assembled genome of *P.Pawak*.

repetitive (Table 10). Of these, 82.26 Mb were transposable elements, making up 13.42% of the genome (Table 10). The identified types of TEs included DNA elements, which accounted for 6.81% (41.75 Mb) of the genome, long interspersed nuclear elements (LINEs) at 3.03% (18.56 Mb), short interspersed nuclear elements (SINEs) at 0.35% (2.17 Mb), and long terminal repeats (LTRs) at 3.23% (19.79 Mb).

Prediction of protein-coding genes. A high-confidence gene set for the *P. pawak* genome was annotated using a combination of three strategies: *de novo* prediction, homology-based prediction, and transcript-based prediction. *De novo* prediction was performed using Augustus software v3.3.3[25] was used with parameters set to "-species = pasa.gb.train -noInFrameStop = true -gf3 = on -strand = both" to ensure accuracy. The gene model file (*pasa.gb.train*) was generated using transcriptome data to provide a species-specific model for more precise gene predictions. Homology-based prediction involved aligning protein-coding gene sequences from closely related species such as *L. crocea* (GCF_000972845.2)[26], *Nibea albiflora* (GCA_014281875.1)[27], *C. lucidus* (GCA_004119915.2)[28], *Scatophagus argus* (GCF_020382885.2)[29], *Chelmon rostratus* (GCF_017976325.1)[30], *Cheilinus undulatus* (GCF_018320785.1)[31], *Cyprinus carpio* (GCF_018340385.1)[32], and *Danio rerio* (GCF_000002035.6)[33] to the *P. pawak* genome using BLAST software v2.14.0[34] with a stringent e-value of 1e-9. The BLAST results were then processed using Wise software v2.4.1[35] with parameters set to "-init local -alg 623 -gap 12 -ext 2 -pretty -genes -quiet" to identify homologous genes. For transcript-based prediction, transcripts were assembled using StringTie software v2.1.3[36] with parameters set to "-f 0.1 -m 200 -a 10 -c 2.5 -g 50 -M 1.0" and RNA-seq data were mapped using HISAT2 software v2.1.0[20] with parameters set to "-summary-file -S -x -1 -2". Coding regions were predicted using TransDecoder software v5.5.0 with parameters set to "-retain_long_orfs_length 150 -T 500". The results from all three strategies were integrated using EVidenceModeler software

Methods	RepeatMasker	RepeatProteinMask	TRF	Combined
(I) Transposon elements	81.93 (13.36%)	17.32 (2.83%)	—	82.26 (13.42%)
DNA	41.59 (6.78%)	3.25 (0.53%)	—	41.75 (6.81%)
LTR	19.73 (3.21%)	6.59 (1.07%)	—	19.79 (3.23%)
LINE	18.45 (3.01%)	7.49 (1.22%)	—	18.56 (3.03%)
SINE	2.16 (0.35%)	—	—	2.16 (0.35%)
(II) Tandem repeats	—	—	36.80 (6.00%)	36.80 (6.00%)
Minisatellite	—	—	14.63 (2.39%)	14.63 (2.39%)
Microsatellite	—	—	7.67 (1.25%)	7.67 (1.25%)
(III)Other	5.09 (0.83%)	0.59 (0.10%)	14.50 (2.37%)	19.60 (3.20%)
(IV)Unknow	46.60 (7.60%)	0.05 (0.01%)	—	47.37 (7.37%)
Total	123.50 (20.15%)	17.91 (2.90%)	36.80 (6.00%)	160.64 (26.20%)

Table 10. Repeat sequences of the *P. pawak* genome annotated using different methods. Note: Total length (Mb) and percentage (within bracket) of the *P. pawak* genome for each type of repeat sequence are shown.

Term	Gene number	Percentage (%)
Nr	25,792	97.84%
Swiss-Prot	22,000	83.45%
KOG	13,140	49.84%
eggNOG	24,559	93.16%
InterPro	24,021	91.12%
Pfam	22,711	86.15%
GO	21,074	79.94%
KEGG	15,938	60.46%
Annotated	25,885	98.19%
Unannotated	476	1.81%
Total	26,361	100.00%

Table 11. Functional annotation of the predicted protein-coding genes in *P.Pawak* genome.

v1.1.1[37] with parameters set to "-segmentSize 500000 -overlapSize 10000", resulting in the prediction of 26,361 protein-coding genes in the *P. pawak* genome(Table 11). Quality validation showed similarities in the distributions of mRNA length, CDS length, exon length, and intron length between the *P. pawak* genome and those of closely related species (Fig. 6), indicating similar patterns of gene structure distribution.

Functional annotation of protein-coding genes. The quality of the gene annotation and functional information of the *P. pawak* genome was evaluated by comparing the predicted protein sequences with existing public protein databases. The databases used for comparison included InterPro[38], Gene Ontology (GO)[39], Kyoto Encyclopedia of Genes and Genomes (KEGG)[40], SwissProt[41], Pfam[42], the NCBI Non-Redundant Protein Sequence Database (NR) (https://www.ncbi.nlm.nih.gov/refseq/about/non-redundantproteins/), eggNOG[43], and Eukaryotic Orthologous Groups of Proteins (KOG)[44]. Functional information analysis was conducted using BLAST software v2.14.0[34]. The annotation results showed that a total of 25,885 genes were annotated, accounting for 98.19% of the protein-coding genes. Only 476 genes, representing 1.81% of the protein-coding genes, could not be annotated (Table 11). These findings suggest a reliable assembly and annotation of the *P. pawak* genome.

Data Records

The Illumina (SRR30242865), PacBio HiFi (SRR30242864), Hi-C (SRR30242861), and RNA-seq (SRR30242854-SRR30242860, SRR30242862- SRR30242863) data used for the genome assembly of *P. pawak* were deposited in the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI)[45]. Chromosome-level genome assembly of *P. pawak* was deposited in the NCBI genome database under accession number GCA_042310395.1[46]. The genomic annotation results of *P. pawak* genome assembly can be found in the Figshare database under DOI code: https://doi.org/10.6084/m9.figshare.26793379[47].

Technical Validation

Genome evaluation. Multiple evaluation metrics were utilized to assess the quality and robustness of the *P. pawak* genome assembly. The assembly was meticulously evaluated using N50, BUSCO, the short reads mapping ratio, and the transcripts mapping ratio. The results confirmed the assembly's strong contiguity, with a high percentage of complete and single-copy genes detected. Additionally, the high mapping rates of both short reads and transcripts underscored the accuracy and reliability of the assembly. These results collectively indicate that the *P. pawak* genome assembly is of high quality, making it a reliable resource for future genomic studies and bioinformatics applications.

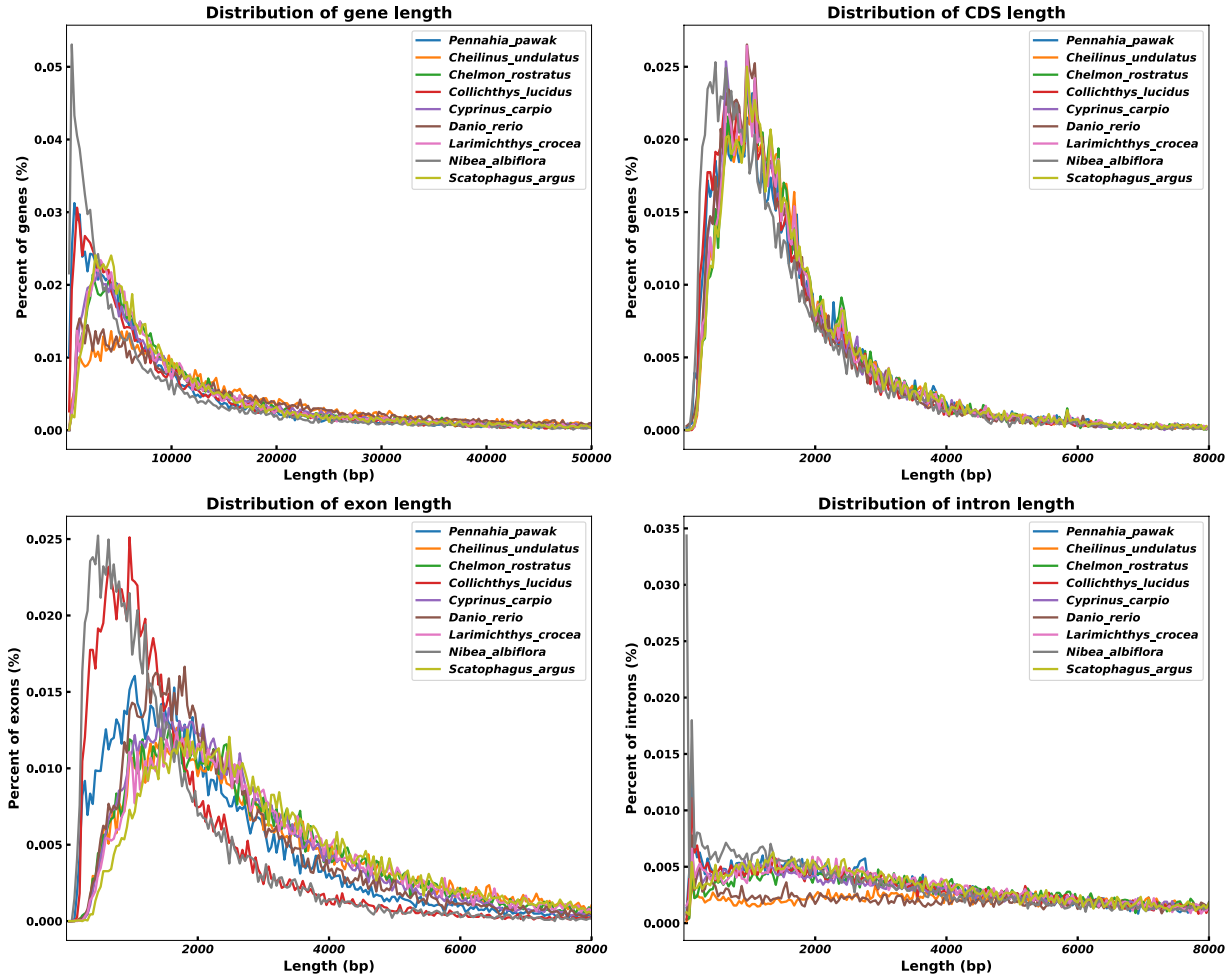

Fig. 6 The length of protein-coding genes in *P. pawak* and their closely related species. Note: The gene length, CDS length, exon length, and intron length were compared among *P. pawak* and the other eight teleost species to verify the quality of gene annotation.

Code availability

The software utilized in this study is publicly available, and the parameters have been clearly outlined in the Methods section. Where specific parameters are not mentioned, the default settings recommended by the software developers were applied. No custom scripts or code were employed in this analysis.

Received: 3 October 2024; Accepted: 3 March 2025;
Published online: 10 March 2025

References

1. Ramcharitar, J., Higgs, D. M. & Popper, A. N. Sciaenid inner ears: a study in diversity. *Brain Behav Evol* **58**, 152–162, https://doi.org/10.1159/000047269 (2001).
2. Lal Mohan, R. Scianidae. *FAO Species Identification Sheets* **51**, 1–9 (1983).
3. Nakabō, T. *Fishes of Japan: with pictorial keys to the species*. **Vol. 2** (Tokai University Press, 2002).
4. 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).
5. Andrews, S. FastQC A Quality Control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/ citeulike-article-id:11583827 (2010).
6. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
7. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204, https://doi.org/10.1093/bioinformatics/btx153 (2017).
8. Cai, M. *et al*. Chromosome assembly of Collichthys lucidus, a fish of Sciaenidae with a multiple sex chromosome system. *Sci Data* **6**, 132, https://doi.org/10.1038/s41597-019-0139-x (2019).
9. Zhang, K. *et al*. A chromosome-level reference genome assembly of the Reeve's moray eel (Gymnothorax reevesii). *Sci Data* **10**, 501, https://doi.org/10.1038/s41597-023-02394-7 (2023).
10. Heo, Y., Wu, X. L., Chen, D., Ma, J. & Hwu, W. M. BLESS: bloom filter-based error correction solution for high-throughput sequencing reads. *Bioinformatics* **30**, 1354–1362, https://doi.org/10.1093/bioinformatics/btu030 (2014).
11. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).

12. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460, https://doi.org/10.1186/s12859-018-2485-7 (2018).
13. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat Plants* **5**, 833–845, https://doi.org/10.1038/s41477-019-0487-8 (2019).
14. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
15. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**, 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
16. Xu, T., Li, Y., Zheng, W. & Sun, Y. A chromosome-level genome assembly of the miiuy croaker (Miichthys miiuy) using nanopore sequencing and Hi-C. *Aquaculture and Fisheries* **9**, 218–225, https://doi.org/10.1016/j.aaf.2021.06.001 (2024).
17. Chen, B. *et al*. The sequencing and de novo assembly of the Larimichthys crocea genome using PacBio and Hi-C technologies. *Sci Data* **6**, 188, https://doi.org/10.1038/s41597-019-0194-3 (2019).
18. Simao, 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, https://doi.org/10.1093/bioinformatics/btv351 (2015).
19. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).
20. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915, https://doi.org/10.1038/s41587-019-0201-4 (2019).
21. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462–467, https://doi.org/10.1159/000084979 (2005).
22. 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).
23. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
24. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* Chapter 4, 4 10 11–14 10 14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
25. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435–439, https://doi.org/10.1093/nar/gkl200 (2006).
26. *NCBI GenBank assembly* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000972845.2/ (2018).
27. *NCBI GenBank assembly* https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_014281875.1/ (2020).
28. *NCBI GenBank assembly* https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_004119915.2/ (2019).
29. *NCBI GenBank assembly* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_020382885.2/ (2021).
30. *NCBI GenBank assembly* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_017976325.1/ (2021).
31. *NCBI GenBank assembly* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_018320785.1/ (2021).
32. *NCBI GenBank assembly* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_018340385.1/ (2021).
33. *NCBI GenBank assembly* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000002035.6/ (2017).
34. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403–410, https://doi.org/10.1016/s0022-2836(05)80360-2 (1990).
35. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* **14**, 988–995, https://doi.org/10.1101/gr.1865504 (2004).
36. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295, https://doi.org/10.1038/nbt.3122 (2015).
37. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
38. Finn, R. D. *et al*. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res* **45**, D190–D199, https://doi.org/10.1093/nar/gkw1107 (2017).
39. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. *Nature genetics* **25**, 25–29, https://doi.org/10.1038/75556 (2000).
40. 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).
41. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic acids research* **28**, 45–48, https://doi.org/10.1093/nar/28.1.45 (2000).
42. Mistry, J. *et al*. Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**, D412–D419, https://doi.org/10.1093/nar/gkaa913 (2021).
43. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**, D309–D314, https://doi.org/10.1093/nar/gky1085 (2019).
44. Tatusov, R. L. *et al*. The COG database: an updated version includes eukaryotes. *BMC bioinformatics* **4**, 1–14, https://doi.org/10.1186/1471-2105-4-41 (2003).
45. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP526444 (2024).
46. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBGOQM000000000 (2024).
47. Jiang, L. H. & Zheng, P. Chromosome-level genome assembly and annotation of pawak croaker (Pennahia pawak). *figshare. Dataset.* https://doi.org/10.6084/m9.figshare.26793379.v1 (2024).

Acknowledgements

This work was supported by the Aquaculture Breeding and Seedling Technology Innovation Center Project of Zhoushan (2024Y001-1) [Yan] and Nanji National Marine Natural Reserves Administration (NJKJ2023002) [Jiang].

Author contributions

Lihua Jiang and Peng Zheng co-led the research project, designing experiments and collecting samples. Lihua Jiang, Jialang Zheng, Wangyang Jin and Peng Zheng performed the analyses. Xiaojun Yan supervised the study and provided guidance. Jialang Zheng, Yifan Liu, Weihua Song, Shun Chen and Peng Zheng collaborated on writing and revising the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

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