scientific data

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of Jaguar guapote (*Parachromis manguensis*) by massive parallel sequencing

Jianmeng Cao[1,4 ✉], Yannan Tong[2,4], Zhigang Xiao[3], Huizi Chen[1] & Zhigang Liu[1]

Parachromis managuensis is a native cichlid fish from Central America and is the most commonly traded species within its genus. This study presents the first chromosome-scale genome assembly of _P. managuensis_ using PacBio HiFi long reads and Hi-C sequencing data. The size of the _P. managuensis_ genome is approximately 896.66 Mb, with a scaffold N50 of 38.19 Mb. The assembled genome demonstrates high quality in terms of completeness and accuracy, with a BUSCO score of 98.85% and a quality value (QV) of 50.95. A total of 888.60 Mb (99.10%) sequences were anchored to 24 pseudochromosomes. Additionally, 21,145 protein-coding genes and 325.58 Mb (~36.31%) repetitive sequences were identified. This chromosome-level genome assembly provides a crucial reference for studying the evolution and ecological adaptability of _P. managuensis_.

Background & Summary

Parachromis managuensis (NCBI Taxonomy ID: 172535), commonly known as the jaguar guapote, belongs to the order Perciformes and the family Cichlidae within the genus *Parachromis*. This species is native to rivers and lakes in Nicaragua, Honduras, Costa Rica, and other Central American countries[1]. *P. managuensis* is a predominantly carnivorous, bottom-dwelling fish that helps regulate the overpopulation of smaller fish in freshwater ecosystems. It is recognized for its rapid growth, tolerance to low oxygen levels, high yield, strong disease resistance, and excellent meat quality[2]. These attributes have contributed to its increasing prominence as an economic fish species, leading to a significant expansion in its aquaculture. It is therefore regarded as an important tropical fish.

Research on *P. managuensis* at the molecular level is limited. In 2015, the mitochondrial genome of *P. managuensis* was sequenced[3]. Subsequently, in 2018, investigations into the adaptive evolution of *P. managuensis* at the transcriptomic level were conducted[4]. However, studies of its genome have not yet been reported, which significantly hinders the advancement of breeding programs and the development of desirable aquacultural traits. Recent advancements in sequencing technologies have markedly improved genomic research. Notably, Pacific BioSciences (PacBio)'s Circular Consensus Sequencing (CCS) mode, available on the Sequel II platform, offers long read lengths (10–20 kb) and high accuracy (over 99%), thereby greatly benefiting *de novo* assembly studies of both plant and animal genomes[5,6]. High-quality reference genome sequences serve as the foundational prerequisite for advancing both fundamental and applied research in plants and animals. For example, the recent cat genome study, AnAms1.0, revealed over 1,600 novel protein-coding genes when compared to the previously utilized felCat9 assembly[7]. Furthermore, AnAms1.0 exhibits a higher count of structural variants and intact olfactory receptor genes compared to felCat9. These advancements play a crucial role in the discovery of novel genetic traits and in fostering advancements in veterinary medicine.

In this study, we report a chromosome-level genome assembly of *P. managuensis* by combining short sequencing reads, PacBio high-fidelity (HiFi) long reads, and chromosomal conformational capture (Hi-C)

[1]Key Laboratory of Tropical and Subtropical Fishery Resources Application and Cultivation, Ministry of Agriculture and Rural Affairs, Pearl River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Guangzhou, 510380, China. [2]Hainan Academy of Ocean and Fisheries Science, Haikou, 570206, China. [3]Menghai County Fisheries Technology Extension Station of Xishuangbanna Dai Autonomous Prefecture, Menghai, 666200, China. [4]These authors contributed equally: Jianmeng Cao, Yannan Tong. ✉e-mail: caojianmeng@aliyun.com

Types	Reads Number	Biosample accession	SRA accession	Total length (Gb)	Genome depth*	N50 length of reads (bp)
Short sequencing raw reads	422,982,962	SAMN43277827	SRR30324166	63.45	70.76	150
Hi-C raw data	1,008,510,144		SRR30324165	151.28	168.72	150
PacBio HiFi reads	1,255,776		SRR30324164	21.7	24.2	17,831

Table 1. Summary of DNA sequencing data of *P. managuensis* genome. *Estimated based on the assembly size of 896.66 Mb.

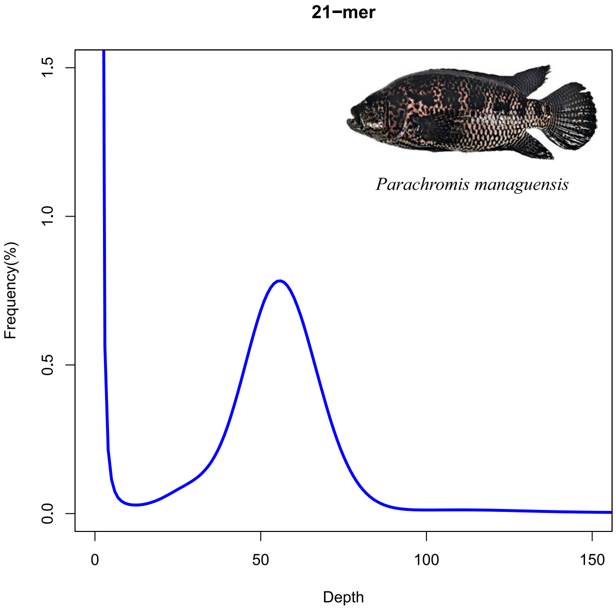

Fig. 1 *K*-mer distribution of *P. managuensis* genome sequencing reads. The X-axis is the depth of *K*-mer derived from the sequenced reads, and the Y-axis is the frequency of the *K*-mer depth.

sequencing data. This high-quality assembled genome not only facilitates population genetic research and evolutionary analysis of *P. managuensis* but also provides important resources for optimizing genetic breeding.

Methods

Sampling, DNA and RNA extraction.

Samples of *P. managuensis* were collected from Dagun Village, Lancang Lahu Autonomous County, Pu'er City, Yunnan Province (coordinates: E 100°27′95.11″, N 22°93′99.27″). Tissue samples were promptly collected, snap-frozen in liquid nitrogen, and then stored at −80 °C. DNA and RNA extraction, library construction, and sequencing in this study were performed using standard experimental and analytical protocols provided by NextOmics Biosciences (Wuhan, China).

Long read DNA preparation and sequencing.

A total of 8 μg of high-quality genomic DNA was extracted from muscle tissue using a Qiagen DNeasy Blood and Tissue Kit (Qiagen, USA) according to the manufacturer's instructions. The quality and concentration of the extracted DNA were assessed using a NanoDrop One spectrophotometer (Thermo Scientific, USA) and 1% agarose gel electrophoresis. PacBio long insert libraries were prepared using the SMRTbell Express Template Prep Kit 2.0 according to manufacturers' instructions, with an insert size of approximately 20 kb. The libraries were sequenced on the PacBio Sequel II system in CCS mode. Subreads were processed with SMRTLink (v11.1.0)[8] using the parameters "–minPasses 3 –minPredictedAccuracy 0.99 –minLength 500", producing approximately 21.70 Gb HiFi reads with an N50 size of 17,831 (Table 1). The parameter "minPredictedAccuracy" set to 0.99 in the context of PacBio SMRTLink software means that, during the data processing of sequencing reads, only those reads that have a predicted accuracy of 99% or higher will be retained for further analysis. This filtering step helps ensure that only high-quality reads, with a high likelihood of correct base calls, are included in the downstream processes, thereby improving the overall accuracy of the genome assembly.

Short read DNA preparation and sequencing.

The extracted DNA (~5 μg) was randomly sheared into approximately 350 bp fragments, and a short fragment library was constructed using the MGIEasy Universal DNA Library Prep Set (MGI, China). Sequencing was conducted on the MGISEQ T7 platform (MGI, China), resulting in a total of 63.45 Gb of short sequencing reads, each 150 bp in length (Table 1).

Item	Parachromis managuensis
Size of assembly (Mb)	896.66
Contig N50 (Mb)	5.14
Scaffold N50 (Mb)	38.19
Chromosome number	24
Hi-C anchored ratio	99.10%
GC content	40.90%
Genome complete BUSCOs	98.85%
Completenes	98.28%
Quality value	50.95
Repetitive sequences	36.31%
Number of protein-coding genes	24,145

Table 2. Summary statistics of *P. managuensis* assembly. Note: The lineage dataset used in BUSCO is actinopterygii_odb10.

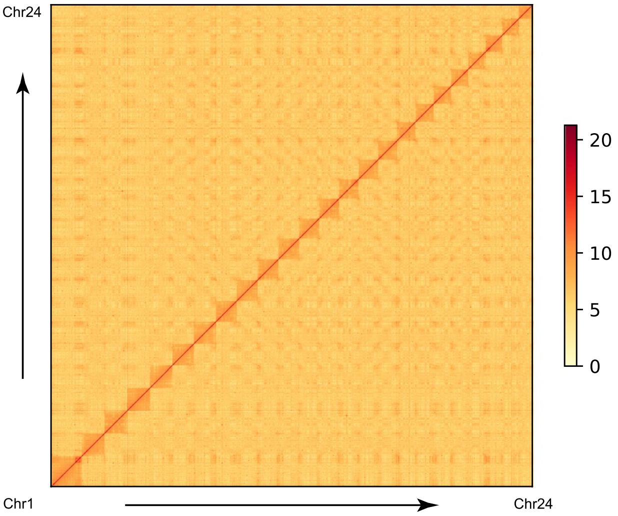

Fig. 2 Hi-C assembly of chromosome interactive heat map. The abscissa and ordinate represent the order of each bin on the corresponding chromosome group. The colour block illuminates the intensity of interaction from yellow (low) to red (high).

Hi-C DNA library preparation and sequencing. A Hi-C library was generated using the DpnII restriction enzyme (GrandOmics, China). Muscle tissue samples were treated with 1% formaldehyde at room temperature for 10–30 minutes to crosslink chromatin-interacting proteins. Subsequently, the DNA was digested with the restriction enzyme, and the 5′ overhangs were repaired with a biotinylated residue. A paired-end library with insert sizes of approximately 300 bp was prepared and then sequenced on the Illumina NovaSeq platform (Illumina, USA). A total of 151.09 Gb of clean data was obtained from 151.28 Gb of sequencing data using the software fastp (v0.19.5)[9] with parameters "-w 20 --length_required 100" (Table 1). The "-w" parameter specifies the number of threads, while "--length_required 100" ensures that reads shorter than 100 bp are discarded.

RNA library preparation and sequencing. For the purpose of RNA sequencing, we extracted total RNA from gill and ovary tissues using the TRIzol reagent (Invitrogen, USA) following the manufacturer's protocol. RNA purity was assessed with a NanoPhotometer spectrophotometer (IMPLEN, CA, USA), while RNA concentration was quantified using the Qubit RNA Assay Kit with a Qubit 2.0 Fluorometer (Life Technologies, CA, USA). RNA-seq libraries were prepared using the TruSeq Stranded mRNA Library Prep Kit (Illumina, USA) according to the manufacturer's instructions. Sequencing was performed on a HiSeq X-Ten instrument, generating 150 bp paired-end reads.

Genome size estimation. The genome size of *P. managuensis* was estimated through *K*-mer profiling. First, raw short sequencing reads underwent quality control using fastp (v0.19.5)[9]. Subsequently, a *K*-mer depth distribution curve was generated with Jellyfish (v1.1)[10] using clean short sequencing reads. Base on the method described in the orange-spotted grouper study[11], the genome size, G, can be calculated from the formula G = K_num/K_depth, where the K_num is the total number of *K*-mer, and K_depth denotes the frequency occurring more frequently than the other frequencies. In this study, K was 21, K_num was 52,215,706,815 and K_depth was 56. Therefore, the genome size of *P. managuensis* was estimated to be 932,423,335 bp (Fig. 1). The estimated

Pseudomolecule	Length (bp)	GC content	Gap number
Chr01	56,130,088	41.31%	34
Chr02	42,772,040	40.78%	13
Chr03	42,111,548	40.53%	9
Chr04	41,935,750	40.95%	15
Chr05	40,403,500	40.71%	11
Chr06	40,170,400	40.79%	10
Chr07	39,892,501	40.97%	11
Chr08	38,657,400	40.70%	8
Chr09	38,511,500	40.64%	11
Chr10	38,423,550	40.94%	11
Chr11	38,189,600	40.76%	12
Chr12	37,101,300	40.86%	7
Chr13	36,889,607	40.76%	9
Chr14	36,246,500	41.07%	10
Chr15	35,470,451	40.67%	11
Chr16	34,719,576	41.01%	11
Chr17	34,254,647	40.67%	11
Chr18	33,212,549	41.03%	9
Chr19	32,206,250	41.18%	5
Chr20	31,928,291	40.65%	14
Chr21	31,667,733	40.98%	13
Chr22	30,514,871	41.58%	10
Chr23	30,337,444	40.98%	9
Chr24	26,858,500	40.97%	10

Table 3. Pseudo-chromosome length statistics after Hi-C assisted assembly.

Type		Length (bp)	% in genome
Transposable elements	LINE	88,419,952	9.86
	LTR	30,121,487	3.36
	SINE	15,397,702	1.72
	DNA	140,258,935	15.64
	MITE	20,429,993	2.28
	RC	3,137,261	0.35
	Total	297,765,330	33.21
Tandem Repeats		6,221,436	0.69
SSR		1,413,745	0.16
Other		3,235,558	0.36
Unknown		14,675,851	1.64
Total		325,584,473	36.31

Table 4. Statistics of repetitive sequences in *P. managuensis* assembly. Note: LINE, long interspersed nuclear elements; SINE, short interspersed nuclear elements; LTR, long terminal repeat; MITE, miniature inverted repeat transposable element.

heterozygosity rate was about 0.15% and the repeat content was approximately 29.31%, suggesting that it is not a complex genome.

De novo assembly and Hi-C assembly.

The primary contigs were assembled from HiFi reads using Hifiasm (v0.16.1)[12] with default parameters. Three rounds of error correction were performed with NextPolish (v1.4.1)[13] with both HiFi reads and short reads. This process resulted in 346 contigs, a total size of 896.65 Mb, and an N50 length of 5.14 Mb. To anchor contigs onto chromosomes, BWA (v0.7.12)[14] was used to align the Hi-C clean data to the assembled contigs. Low-quality reads were filtered using the HiC-Pro pipeline[15] with default parameters. The remaining valid reads were employed to anchor chromosomes using Juicer[16] and the 3d-dna pipeline[17], followed by manual correction with Juicebox (v2.13.07)[18]. Ultimately, approximately 99.10% of the contig sequences were anchored to 24 pseudochromosomes, with only 55 contigs remaining unanchored to chromosomes (Table 2 and Fig. 2). The longest and shortest pseudochromosomes were 56.13 Mb and 26.86 Mb in length, respectively (Table 3). The final genome assembly harbored a total size of 896.66 Mb with a scaffold N50 of 38.19 Mb (Table 2).

Sample	Tissue	Biosample accession	SRA accession	Raw data	Clean data	Clean Q20	GC rate	Total MappingRatio	Uniquely MappingRatio
ds-G-1	Gill	SAMN43277828	SRR30324163	19,875,577,800	19,684,053,144	97.64%	48.82%	95.54%	91.95%
ds-Lc-1	Ovary	SAMN43277829	SRR30324162	36,704,199,900	36,167,468,862	97.81%	49.78%	93.86%	90.12%

Table 5. Summary of RNAseq sequencing data of *P. managuensis* genome.

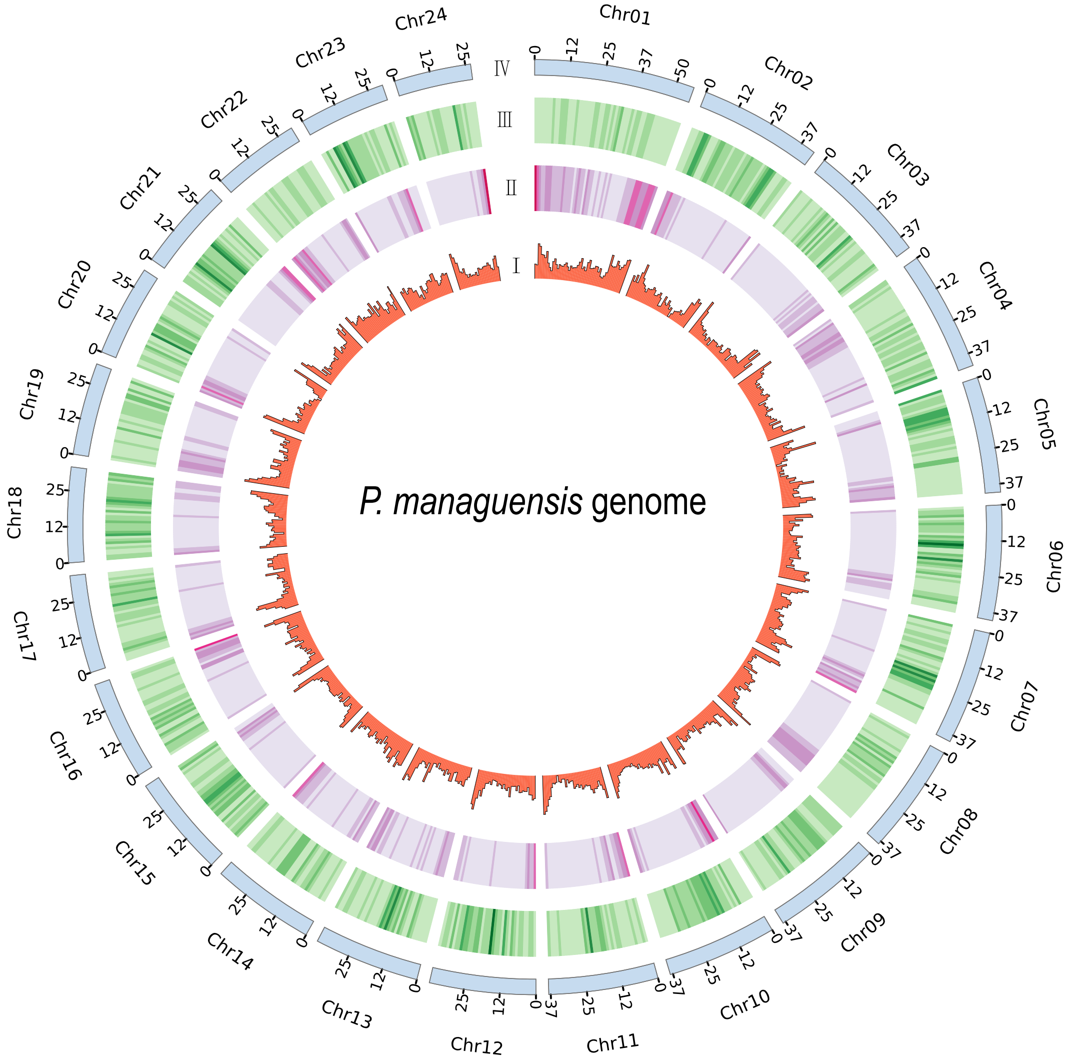

Fig. 3 Circos plot showing the features of the assembled *P. managuensis* genome. From the inside to the outside, GC content, repetitive sequences density, and gene density, respectively.

Annotation of repetitive sequences. Tandem repeats and interspersed repeats were identified in this study. Tandem Repeat Finder (v4.10.0)[19] and GMATA (v2.2.1)[20] were used to discern tandem repeats with default parameters. The analysis revealed that simple sequence repeats (SSRs) account for 0.16% of the total genome length, while tandem repeat sequences comprised 0.69% of the genome length (Table 4). Subsequently, RepeatModeler (v1.0.4)[21] and MITE-hunter[22] were employed to construct a *de novo* repeat sequence library, which was then used to detect interspersed repeats and low-complexity sequences with RepeatMasker (v4.0.7)[23]. DNA transposable elements (TEs) were identified through RepeatMasker (v4.0.7). In detail, a total of 325.58 Mb (~36.31%) of repetitive sequences were obtained (Table 4). Among the interspersed repeats, DNA elements were the most prevalent type, accounting for 15.64% of the genome (Table 4).

Protein-coding genes prediction and function assignment. Gene prediction was performed using a multifaceted approach incorporating transcriptome-based, homology-based, and *ab initio* methods. For the transcriptome-based prediction, a total of 55.85 Gb of RNA-seq clean reads were aligned to the *P. managuensis* assembly using STAR (v2.7.3a)[24] (Table 5). Stringtie (v1.2.2)[25] was then utilized to assemble transcripts based

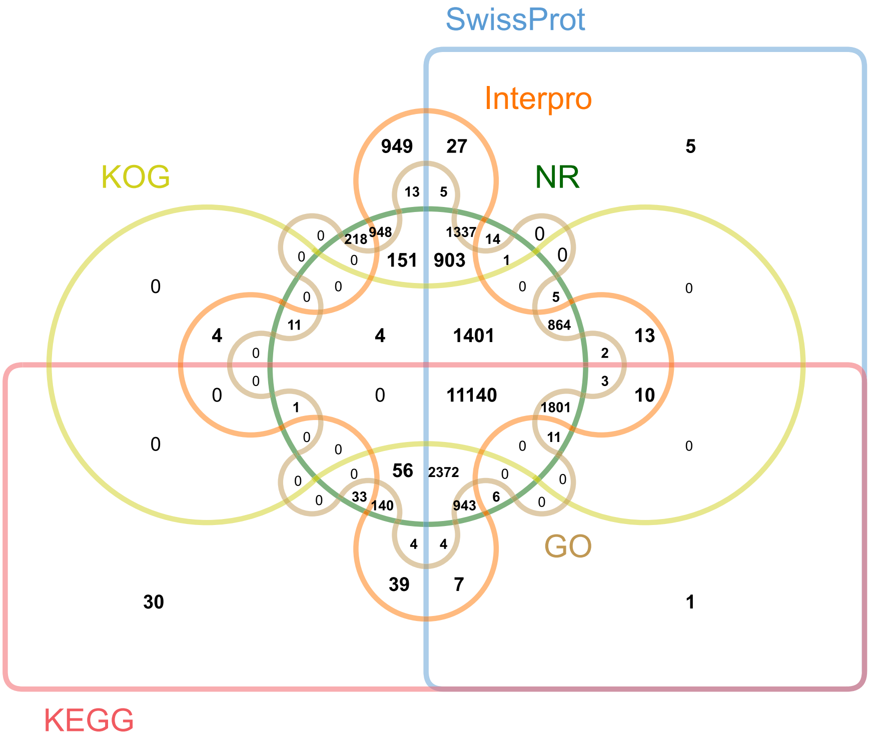

Fig. 4 Venn diagram showing the number of genes with homology or functional classification by each method.

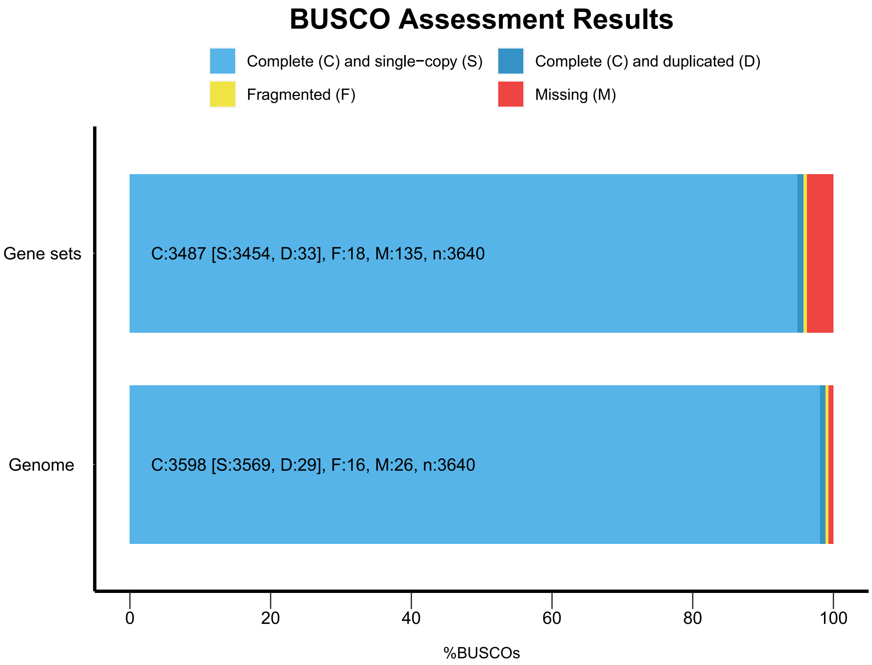

Fig. 5 BUSCO assessments of *P. managuensis* genome and gene sets.

on the alignment results. Afterwards, these assembled transcripts were aligned against the *P. managuensis* assembly using Program to Assemble Spliced Alignment (PASA) (v2.4.1)[26]. This process led to the identification of 22,712 genes, which were designated as the "RNAseq-set". Genomes from five fish species were retrieved from the National Center for Biotechnology Information (NCBI) database for homology-based gene prediction. The species included *Oreochromis niloticus* (GCF_001858045.2), *Maylandia zebra* (GCF_000238955.4), *Neolamprologus brichardi* (GCF_000239395.1), *Astatotilapia calliptera* (GCF_900246225.1), and *Archocentrus* *centrarchus* (GCF_007364275.1). *O. niloticus* is a well-studied fish with abundant genomic information and holds an important position in the field of fish biology and genomics research. It can provide a broad foundation of reference sequences and functional information for gene annotation of *P. managuensis*. *M. zebra*, *N. brichardi*, *A. calliptera*, and *A. centrarchus* all belong to the Cichlidae family and have a relatively close phylogenetic relationship with *P. managuensis*. These genomes were used as queries to search against the *P. managuensis* assembly using GeMoMa (v1.9)[27], yielding 30,878 genes. Homology predictions were denoted as "Homology-set". For *ab initio* prediction, 3,000 high-quality genes from PASA were randomly selected as the training set for model training with AUGUSTUS (v3.2.3)[28]. AUGUSTUS (v3.2.3)[28] was then employed to predict coding regions in the repeat-masked genome, identifying 26,490 genes. The gene models generated by AUGUSTUS were labeled as the AUGUSTUS-set. Finally, all gene models were integrated using EvidenceModeler (v2.1.0)[29]. The weight of each evidence was set as follows: RNAseq-set > Homology-set > AUGUSTUS-set. The final comprehensive gene set comprised 24,145 genes, with an average of 10.53 exons per gene, an exon length of 167.11 bp, and a coding sequence (CDS) length of 1,760.01 bp. Circos (v0.69)[30] was used to visualize the 24 pseudochromosomes, GC content, repetitive sequence density, and gene density (Fig. 3). The values for GC content were determined using SeqKit2 (v2.9.0)[31]. For the calculation of repetitive sequence density and gene density, bedtools (v 2.29.2)[32] was employed and a window size of 1 Mb was set.

The integrated gene set was translated into amino-acid sequences and annotated using six public databases. Briefly, amino-acid sequences were aligned to EuKaryotic Orthologous Groups (KOG)[33], SwissProt[34], Kyoto Encyclopedia of Genes and Genomes (KEGG)[35], and the NCBI nonredundant database (NR) using the BLSAT program (v2.7.1+)[36] with an E-value cutoff of 1e-05. Protein domains were identified using the InterProScan (v5.30)[37] program, and Gene Ontology (GO) terms for each gene were also extracted through InterProScan. Overall, 23,476 (97.23%) of the predicted protein-coding genes were functionally annotated (Fig. 4).

Data Records

All the raw sequencing data utilized in this study were submitted to the NCBI Sequence Read Archive database with accession numbers SRP527692[38] under BioProject number PRJNA1150327. The genome assembly has been deposited GenBank under the WGS accession GCA_040437545.1[39]. Additionally, the genome assembly and annotation were deposited at Figshare[40].

Technical Validation

Genome assembly and gene prediction quality assessment.

The precision and integrity of the *P. managuensis* assembly were evaluated using several methods. We utilized Merqury (v1.3)[41] with PacBio HiFi long reads, employing a *K*-mer value of 17-bp to infer the consensus quality value (QV). Merqury assesses genome completeness by first defining reliable *K*-mers based on *K*-mer count histograms to set a threshold. Then, it calculates the fraction of these reliable *K*-mers in the read set that are also in the assembly. The results revealed a QV of 50.95 and a genome completeness of 98.28% (Table 2). Next, minimap2 (v2.24-r1122)[42] and BWA (v0.7.12)[14] were employed for aligning the HiFi reads and short sequencing clean reads to the *P. managuensis* assembly, respectively. Notably, 99.99% of HiFi reads and 99.51% of short sequencing reads were aligned to the genome. Referring to the genome study of *Epinephelus coioides*[11] and *Neosalanx taihuensis*[43], the completeness evaluation of the *P. managuensis* assembly was assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO) (v5.5.0)[44] with the actinopterygii_odb10 database. Results showed that 98.85% of the expected actinopterygii genes (including 98.05% single and 0.80% duplicated ones) had complete gene coverage, while 0.44% were identified as fragmented (Fig. 5). Furthermore, BUSCO analysis was also conducted to validate the genome annotation quality, which indicated that 95.80% of recognized BUSCOs, comprising 94.89% single-copy and 0.91% duplicated genes, were complete (Fig. 5). In conclusion, the *P. managuensis* genome assembly has achieved high quality.

Code availability

No specific code or script was used in this work. Commands used for data processing were all executed according to the manuals and protocols of the corresponding software.

Received: 11 September 2024; Accepted: 4 March 2025;
Published online: 10 March 2025

References

1. Barros, L. C., Santos, U., Zanuncio, J. C. & Dergam, J. A. Plagioscion squamosissimus (Sciaenidae) and *Parachromis managuensis* (Cichlidae): a threat to native fishes of the Doce River in Minas Gerais, Brazil. *PloS one* **7**, e39138 (2012).
2. Jian, W., Haitao, S., Kai, M. & Chuang, L. C. An invasive fish *Parachromis managuensis* found in Wanquan River, Hainan, China. *Chinese Journal of Zoology* **47**, 124–126 (2012).
3. Liu, L. *et al*. Complete mitochondrial genome of *Parachromis managuensis* (Perciformes: Cichlidae). *Mitochondrial DNA. Part A, DNA mapping, sequencing, and analysis* **27**, 2533–2534 (2016).
4. Zhong, H. *et al*. Evidence for natural selection of immune genes from *Parachromis managuensis* by transcriptome sequencing. *Biotechnology & Biotechnological Equipment* **32**, 1431–1439 (2018).
5. Wenger, A. M. *et al*. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology* **37**, 1155–1162 (2019).
6. Lovell, J. T. *et al*. Four chromosome scale genomes and a pan-genome annotation to accelerate pecan tree breeding. *Nature Communications* **12**, 4125 (2021).
7. Matsumoto, Y. *et al*. Chromosome-scale assembly with improved annotation provides insights into breed-wide genomic structure and diversity in domestic cats. *Journal of Advanced Research* (2024).
8. Chin, C. *et al*. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* **10**, 563–569 (2013).

9. Chen, S., Zhou, Y., Chen, Y. & Jia, G. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
10. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
11. Li, S. *et al*. Mechanisms of sex differentiation and sex reversal in hermaphrodite fish as revealed by the Epinephelus coioides genome. *Molecular Ecology Resources*. **23**, 920–932 (2023).
12. 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 (2021).
13. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics (Oxford, England)* **36**, 2253–2255 (2020).
14. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
15. Servant, N. *et al*. HiC-Pro: An optimized and flexible pipeline for Hi-C data processing. *Genome Biology* **16** (2015).
16. Durand, N. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**, 95–98 (2016).
17. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, eaal3327 (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 (2016).
19. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580 (1999).
20. Wang, X. & Wang, L. GMATA: An Integrated Software Package for Genome-Scale SSR Mining, Marker Development and Viewing. *in Frontiers in plant science* **7**, 1350 (2016).
21. Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Current Protocols in Bioinformatics* **5** (2004).
22. Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic acids research* **38**, e199 (2010).
23. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics (Oxford, England)* **21**(Suppl 1), i351–8 (2005).
24. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15–21 (2013).
25. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology* **20**, 278 (2019).
26. Haas, B. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654–5666 (2003).
27. Jens *et al*. GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. *Methods in Molecular Biology* (2019).
28. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic acids research* **33**, W465–7 (2005).
29. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**, R7 (2008).
30. Krzywinski, M. *et al*. Circos: an information aesthetic for comparative genomics. *Genome research* **19**, 1639–1645 (2009).
31. Shen, W., Sipos, B. & Zhao, L. SeqKit2: A Swiss army knife for sequence and alignment processing. *in iMeta* **3**, e191 (2024).
32. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**, 841–842 (2010).
33. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic acids research* **28**(1), 33–6 (2000).
34. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Research* **27**, 49–54 (1999).
35. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27–30 (2000).
36. 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 (1990).
37. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
38. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP527692 (2024).
39. Cao, J. *et al*. *Parachromis managuensis* isolate HL-2024b, whole genome shotgun sequencing project. *GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_040437545.1 (2024).
40. Cao, J. Chromosome-level genome assembly of *Parachromis managuensis*. *figshare. Dataset.* https://doi.org/10.6084/m9.figshare.26793697.v1 (2024).
41. Rhie, A., Walenz, B.P., Koren, S. & Phillippy, A.M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biology **21** (2020).
42. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2017).
43. Zhou, Y. *et al*. Gap-free genome assembly of Salangid icefish Neosalanx taihuensis. *in Scientific data* **10**, 768 (2023).
44. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods in molecular biology (Clifton, N.J.)* **1962**, 227–245 (2019).

Acknowledgements

We thank Mr. Yunqing Wang and Mr. Minqi Wang in Xishuangbanna Yunbo Aquaculture Development Co., Ltd for their kindly helps in sample collection. This work is supported by Key R&D project in Hainan Province (ZDYF2023XDNY063); Guangzhou Science and Technology Plan Project (202201010049); Central Public-interest Scientific Institution Basal Research Fund, CAFS (2023SJHX3).

Author contributions

J.C. conceived this study, designed the experiment and provided the funding. Y.T. and Z.X. collected samples and performed data analysis. H.C. and Z.L. drafted and revised the manuscript. All authors have 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 J.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025