scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A complete telomere-to-telomere chromosome-level genome assembly of X-ray tetra (*Pristella maxillaris*)

Chao Bian[1,3], Changxing Hu[1,2,3], Zhe He[1], Zigang Li[2] ✉ & Qiong Shi[1] ✉

X-ray tetra (*Pristella maxillaris*) originates from the lower Amazon basin in South America. It is renowned for its strikingly transparent body, which has drawn significant interests in biomedical research and the world ornamental fish industry. Nevertheless, genomic resources for this interesting fish species remains scarce, hindering exploration of the molecular basis behind its unique transparency. To address this gap, we constructed the first complete telomere-to-telomere (T2T) chromosome-scale genome assembly of the X-ray tetra by integration of PacBio HiFi, ONT ultra-long, and Hi-C sequencing technologies. This haplotypic assembly spans approximately 1.1 Gb, with a contig N50 of 42.8 Mb. It is anchored onto 25 chromosomes, highlighting a complete set of 50 telomeres and 25 centromeres. We predicted 514.3 Mb of repetitive sequences and annotated 28,456 protein-coding genes in the assembled genome. Subsequent BUSCO analysis discovered high genome completeness (98.0%). This high-quality T2T genome assembly provides a valuable genetic resource for investigating the molecular mechanisms underlying transparency, and supporting in-depth studies on functional genomics, genetic diversity, and selective breeding for this economically important species.

Background & Summary

X-ray tetra (*Pristella maxillaris*, Ulrey 1894) has attracted considerable attention for its nearly transparent body since its initial description in 1894. Native to the lower Amazon River basin in Brazil, Venezuela, and Guyana, it is widely distributed across the coastal river systems of northern South America, east of the Andes[1]. It is classified under the phylum Chordata, class Actinopterygii, order Characiformes, family Characidae, and genus Pristella. This specie spotlights an albino variant with full-body transparency at the adult stage, permitting skeletal and internal structures to be visible, particularly in abdominal regions[2]. This distinct transparency has made it highly popular in the world ornamental fish market, especially after its introduction to China in the 1990s[3].

Adapted to both freshwater and brackish environments, the X-ray tetra can tolerate water temperatures from 22 and 30 °C and survives in cold conditions up to 12 °C[3]. It presents a silvery-white body with yellow tinged, black-spotted dorsal and anal fins bordered by white, and a deeply forked caudal fin with red markings. Its adipose fin lacks bony support, positioned between the dorsal and caudal fins[1]. Due to its remarkable transparency, the X-ray tetra is widely used as a model organism for extensive studies on chromatophore development, transparency mechanism, and molecular basis of adaptive coloration.

Despite its popularity for research and trade, international investigations on the X-ray tetra have focused primarily on its ecological adaptability[4], morphological features[5,6], and behavioral traits[7,8]. The transparent phenotype of X-ray tetra is attributed to the absence or reduction of melanophores, iridophores, and xanthophores, coupled with a thin and translucent dermal layer. Previous transcriptomic studies showed that some genes associated with melanin synthesis (i.e., *tyr* and *tyrp1*) are significantly down-regulated, and the expression of purine metabolism pathway genes associated with iridescent cells (such as *gart* and *hprt*) is reduced[3]. However, the molecular basis of its unique transparency remains largely unexplored, with availability of limited genomic

[1]Laboratory of Aquatic Genomics, College of Life Sciences and Oceanography, Shenzhen University, Shenzhen, 518057, China. [2]State Key Laboratory of Chemical Oncogenomics, School of Chemical Biology and Biotechnology, Peking University Shenzhen Graduate School, Shenzhen, 518055, China. [3]These authors contributed equally: Chao Bian, Changxing Hu. ✉e-mail: lizg@pkusz.edu.cn; shiqiong@genomics.cn; shiqiong@szu.edu.cn

Parameter	Value
HiFi reads (Gb)	70.2
Hi-C reads (Gb)	117.3
ONT reads (Gb)	30.0
RNA-seq data (Gb)	11.7
Genome size (Gb)	1.1
Contig N50 (Mb)	42.8
BUSCO	98.0%
Protein BUSCO	96.5%
Repeat ratio	46.4%
Gene number	28,456
Average gene length (bp)	21311.9
Functional gene number	27,991

Table 1. Statistics of the T2T genome assembly and annotation.

resources. To date, only mitochondrial genome[9] and transcriptome data[3] have been reported, while no complete genome assembly exists. This lack of genomics data limits our understanding of its genetic regulatory mechanisms, adaptive evolution, and genetic diversity, hereby restricting its potential for conservative and aquaculture practices.

Recent advancements in high-throughput sequencing technologies have permitted molecular research in evolutionary analysis, functional gene discovery, and genomics-based breeding[10]. In this study, we constructed a solid T2T chromosome-level genome assembly at the first time for X-ray tetra by integrating PacBio, ONT, and high-throughput chromatin conformation capture (Hi-C) technologies[11,12]. This haplotypic assembly spans approximately 1.1 Gb and is anchored to 25 chromosomes. These genomics data lay the foundation for in-depth investigations on genomic architecture, functional genes, and genetic traits of this economically important fish.

Methods

Sample collection and DNA extraction. A single X-ray tetra individual was collected from a local company in Pudong District, Shanghai, China (121°46′28″E, 31°11′7″N). Muscle tissue was immediately flash-frozen in liquid nitrogen and stored at −80 °C for further use. Genomic DNA (gDNA) was extracted using a TIANamp Genomic DNA Extraction Kit (Tiangen, Beijing, China), following the manufacturer's protocol to obtain high-molecular-weight DNA for subsequent whole genome sequencing[13].

DNA concentration and purity were measured on a Qubit 3.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) and a Nanodrop spectrophotometer (Thermo Fisher Scientific)[14] respectively, and integrity was verified on a 1.0% agarose gel. Animal experiments complied with the Ministry of Science and Technology of China's humane treatment guidelines and were approved by Shenzhen University's Animal Care and Use Committee.

Library preparation for genome and transcriptome sequencing. *PacBio HiFi long-read sequencing.* DNA libraries were prepared using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, CA, USA) following the PacBio's standard protocol, and then sequenced on a PacBio Sequel II platform[15]. High-accuracy consensus reads were generated using the Circular Consensus Sequencing (CCS, SMRT Link v11.0) software[16], yielding ~70.2 Gb of HiFi data with an N50 length of 21.3 kb (Table 1).

ONT ultra-long sequencing. A ultra-long library was prepared using an SQK-ULK001 kit according to the guidance from the manufacturer (Oxford Nanopore Technologies, UK). Sequencing on a PromethION flow cell (Oxford Nanopore Technologies) yielded 30.0 Gb of ultra-long reads after being processed with NECAT v200221 with default parameters[17].

Hi-C Sequencing. Hi-C libraries were prepared from the fresh muscle sample using a GrandOmics Hi-C kit and DpnII restriction enzyme according to the manufacturer's protocol (GrandOmics, Wuhan, China). In brief, gDNA was cross-linked, digested, biotin-labeled, ligated, and fragmented to 500 bp, followed by purification with streptavidin magnetic beads. Library quality and insert size were assessed using a Qubit 3.0 Fluorometer and an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA), respectively. Libraries were sequenced on an Illumina NovaSeq. 6000 platform (Illumina, San Diego, CA, USA), generating ~117.3 Gb of high-quality 150 bp paired-end reads after being processed with Fastp v0.20.0 (https://github.com/OpenGene/fastp)[18].

RNA sequencing. Total RNA was extracted from muscle tissue by using TRIzol reagent (Invitrogen, Carlsbad, CA, USA) in accordance with the manufacturer's instructions. RNA integrity was verified using an Agilent 2100 Bioanalyzer, and only those samples with an RNA Integrity Number (RIN) above 7.0 were selected for subsequent library preparation. Libraries were constructed using the DNA nanoball (DNB) technology on a DNBSEQ T7 platform (MGI, BGI Shenzhen, China), and sequencing was performed on a MGISEQ-2000 platform (MGI) to generate 11.7 Gb of 150-bp paired-end reads.

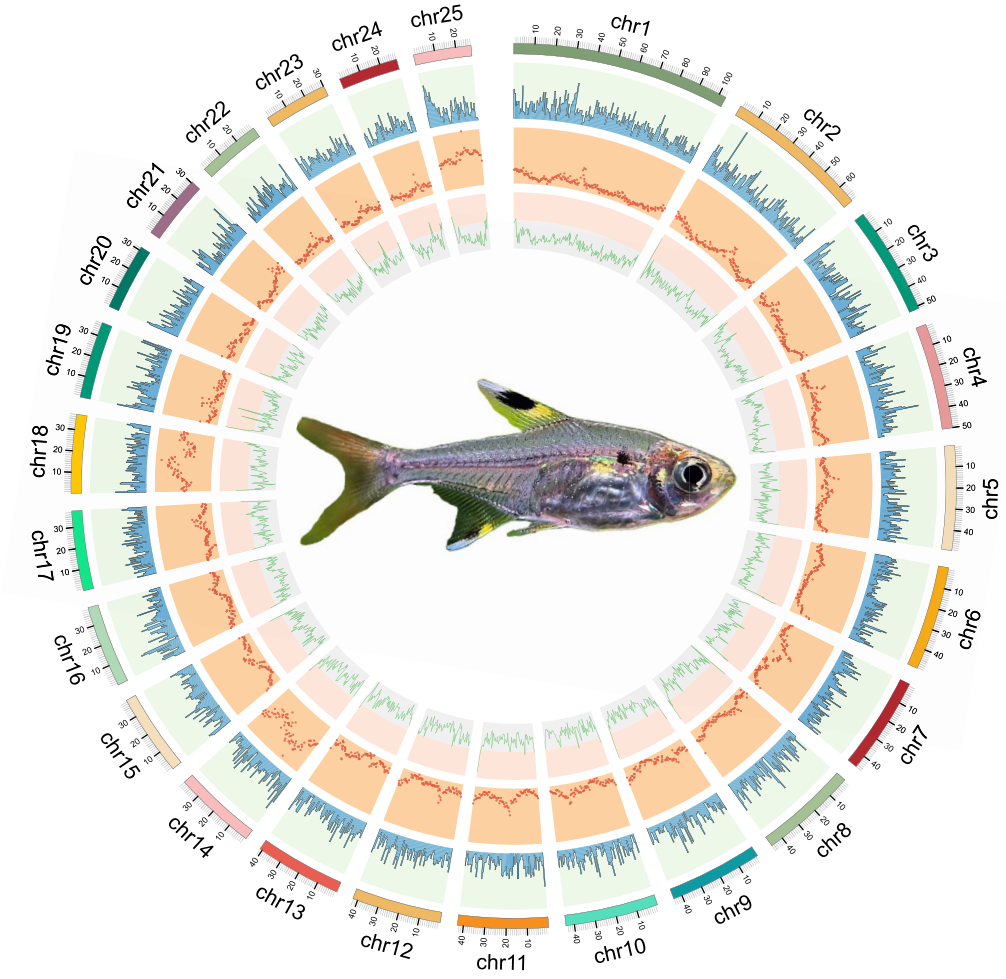

Fig. 1 A Circos image of the 25 anchored chromosomes within the haplotypic genome. Data from the outside to the inside include the length and number of each chromosome, gene density distribution (bin = 1 Mb), repetitive sequence density (bin = 1 Mb), GC content (bin = 1 Mb), and a fish photo (in the center).

Genome assembly and quality evaluation. A genome assembly of X-ray tetra was initially completed using NextDenovo[19] (v2.2-beta.0, https://github.com/Nextomics/NextDenovo) with the following parameters: -rerun 3 -read_cutoff 1k -pa_correction 2, resulting in primary contigs with a total size of ~1.1 Gb and a contig N50 of 14.9 Mb. Subsequently, Hi-C sequencing data were mapped to these assembled contigs using Bowtie2 v2.2.5 (parameters: --very-sensitive -L 30 --score-min L,-0.6,-0.2 --end-to-end –reorder)[20], and then effective linkage products were detected using HiC-Pro v2.8.1[21] under default settings, retaining only valid contact pairs to support the anchoring of contigs to chromosomes. To orient, order, and cluster contigs into chromosome-level scaffolds, we applied Juicer v1.5 (parameter: chr num 25)[22] and 3D-DNA v170123 (parameters: -m haploid -r 2)[23]. Visualization and manual corrections were performed with Juicebox v1.11.08[24] to adjust mis-assemblies and eliminate redundant contigs.

However, the primary chromosome-level genome assembly contained 188 gaps. To achieve a gap-free and T2T level assembly, LR_GapCloser v1.0 (parameters: -t 35 -m 1000000 -v 10000)[25] and TGS-GapCloser v1.0.1 (parameter: -min_match 2000)[26] were sequentially employed to fill these gaps. Centromere and telomere sequences were identified by using the QuarTeT software[27]. The final haplotypic genome assembly anchored 25 chromosomes (Figs. 1 and 2) with a total length of 1.1 Gb (Fig. 1, Table 2), which represents ~97.8% of the assembled contigs. The N50 value of these anchored chromosomes was increased to 43.9 Mb.

To evaluate genome completeness and accuracy, Benchmarking Universal Single-Copy Orthologs v5.2.2 (BUSCO)[28] was utilized, employing the Actinopterygii odb10 dataset as the reference. Moreover, genome quality was assessed using the K-mer-based tool Merqury v1.3 (https://github.com/marbl/merqury)[29] with a K-mer size of 19. This analysis compared K-mer frequency distributions between the genome assembly and whole-genome HIFI sequencing data, producing good metrics for genome completeness and quality values (QV). Our results demonstrated that 98.0% of complete BUSCO genes were discovered (Table 1), with 97.1% as single-copy genes. The assembly's QV score was 41.8, indicating that the assembled X-ray tetra genome exhibits high continuity, reliability, and overall high quality.

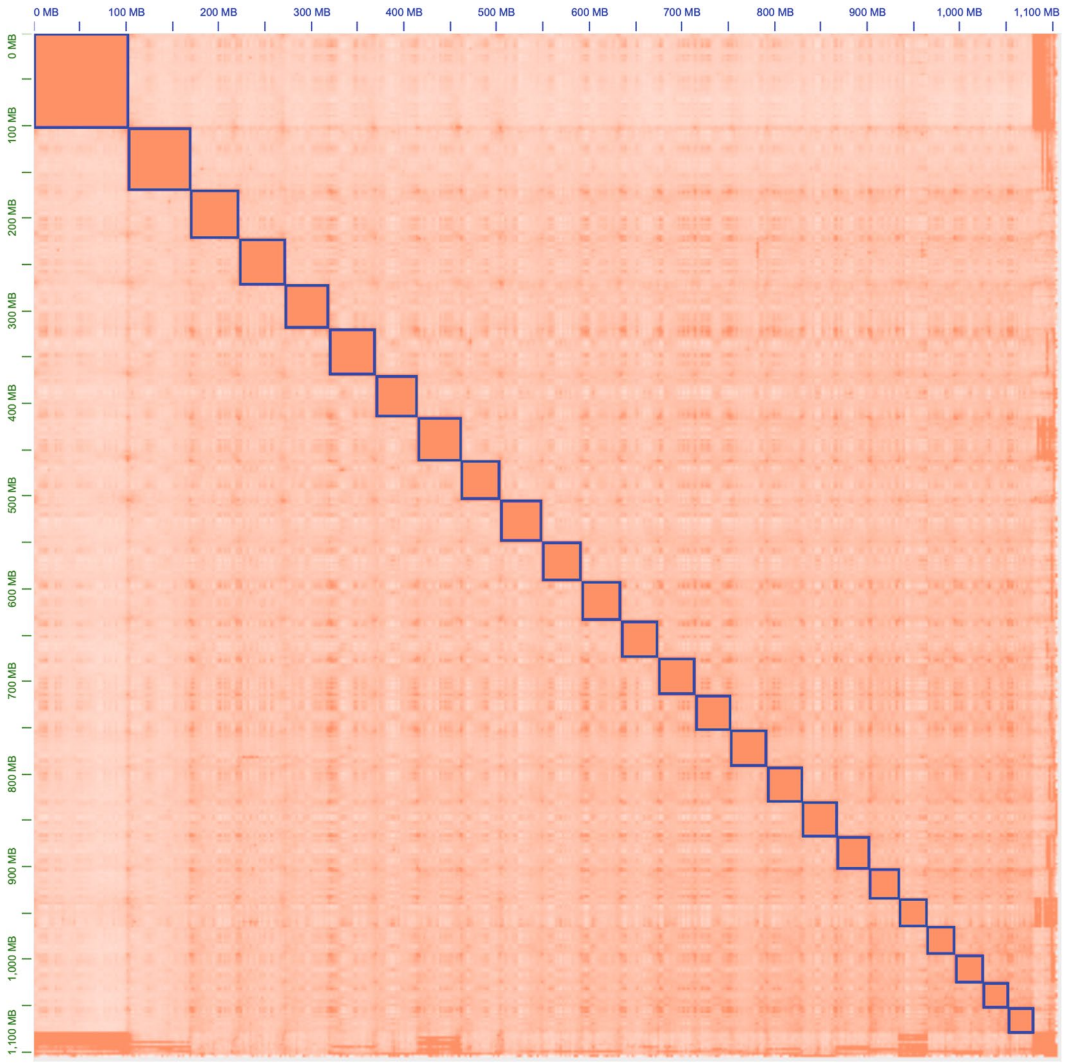

Fig. 2 Chromosome heatmap of the Hi-C data. A total of 25 chromosomes were anchored for the hyplotypic genome.

Repeat element annotation. In the chromosome-scale genome assembly of X-ray tetra, repeat elements (REs) were characterized by using a combination of *de novo* and homology-based approaches. For the *de novo* prediction, RepeatModeler v2.0.1[30] and LTR-FINDER v1.0.6[31] were applied to develop a custom repeat library. Subsequently, RepeatMasker v4.1.0[32] was employed with default parameters to annotate REs in the genome based on the Repbase TE library v21.01[33]. For the homology-based approach, Tandem Repeats Finder v4.07 with optimized parameters (2 7 7 80 10 50 2000 -d -h)[34] was utilized to identify tandem repeat sequences, while RepeatMasker v4.1.0[32] and RepeatProteinMask v4.1.0[32] were applied to annotate transposable elements (TEs) in the assembled genome with default settings. Centromere and telomere sequences were identified by using the quarTeT[27] software.

The final results discovered that ~514.3 Mb of the X-ray tetra genome comprised repetitive sequences, accounting for 46.4% of the entire genome (Table 1). A complete set of 50 telomeres (at both ends of each chromosome with telomeric sequences) were identified, and centromeres were present on the all 25 chromosomes (see more details in Fig. 3). This comprehensive repeat annotation provides a solid foundation for subsequent gene prediction and functional analysis.

Gene prediction and functional annotation. To construct a comprehensive gene set for X-ray tetra, we utilized a combination of *ab initio* prediction, homology-based annotation, and transcriptome-based annotation to predict protein-coding genes[14]. For the homology-based annotation, protein sequences from four representative teleost species, including *Danio rerio* (zebrafish; GCF_000002035.6), *Oryzias latipes* (medaka; GCF_002234675.1), *Takifugu rubripes* (Japanese puffer; GCF_901000725.2), and *Tetraodon nigroviridis* (green spotted puffer; GCA_000180735.1) were downloaded from the NCBI database. These sequences were aligned to the assembled X-ray tetra genome using tBLASTn (with an e-value of 10^{-5})[35], followed by further gene structure refinement with GeneWise v2.2.0[36] (parameters: –blast_eval 1e-5 –align_rate

Chr ID	Chr Length (bp)	Gap number	Telomere number	GC (%)	Centromere start	Centromere end	Left number of telomere motifs	Right number of telomere motifs
Chr1	103,378,760	0	2	38.32	75,632,438	75,762,961	68	35
Chr2	68,543,615	0	2	38.65	218,964	586,362	159	34
Chr3	51,255,812	0	2	38.68	48,013,427	48,171,079	313	36
Chr4	50,169,974	0	2	38.39	46,583,736	46,777,535	191	100
Chr5	49,437,457	0	2	38.65	13,894,232	14,284,261	146	65
Chr6	49,378,772	0	2	38.59	12,605,947	13,780,125	285	69
Chr7	46,097,283	0	2	38.82	20,127,014	20,261,987	144	319
Chr8	46,052,650	0	2	38.66	19,846,217	20,120,613	213	49
Chr9	44,084,817	0	2	38.72	9,379,270	10,726,203	80	101
Chr10	43,935,422	0	2	38.61	31,390,091	31,734,292	270	117
Chr11	42,828,344	0	2	39.02	21,401,988	21,656,115	136	307
Chr12	42,365,502	0	2	38.74	28,490,231	28,685,355	43	32
Chr13	41,184,826	0	2	38.61	31,295,681	31,493,960	48	43
Chr14	39,148,021	0	2	38.91	13,501,201	13,612,152	68	29
Chr15	38,797,547	0	2	38.77	34,806,423	34,917,167	31	111
Chr16	38,869,987	0	2	38.71	17,895,531	18,044,413	89	90
Chr17	37,822,013	0	2	38.99	5,199,257	5,351,441	60	166
Chr18	37,229,363	0	2	38.90	18,676,837	18,814,355	496	102
Chr19	36,416,246	0	2	39.12	15,114,627	15,252,090	433	39
Chr20	31,634,620	0	2	38.87	7,930,687	8,214,206	27	195
Chr21	30,622,564	0	2	38.96	26,468,809	26,761,178	48	56
Chr22	29,995,524	0	2	38.99	13,270,294	13,394,048	64	96
Chr23	30,150,453	0	2	39.20	17,301,026	17,510,512	25	251
Chr24	28,239,184	0	2	39.19	18,636,580	19,320,411	103	37
Chr25	27,287,464	0	2	38.67	16,643,226	17,036,456	98	46

Table 2. Summary of the complete T2T chromosomes (Chr) in the assembled *P. maxillaris* genome.

0.5–extend_len 500). The transcriptome-based annotation utilized ~13.91 Gb of RNA-seq data to map the assembled genome with HISAT[37]. These RNA-seq alignments were then analyzed using Cufflinks v2.2.1 (http://cole-trapnell-lab.github.io/cufflinks/) to identify gene structures. These results from the three methods were then integrated by using MAKER (parameters: max_dna_len = 300000, min_contig = 500, pred_flank = 500, AED_threshold = 1, split_hit = 30000, single_exon = 1, single_length = 250, tries = 2)[38] to create a non-redundant gene set.

For subsequent functional annotation, predicted protein sequences were compared against six public databases, including NCBI NR[39], Swiss-Prot[40], GO[41], TrEMBL[39], KOG[42], and KEGG[43], using BLASTP (e-value < $1e^{-5}$). Gene Ontology (GO) terms were also assigned using InterProScan[44]. Finally, a total of 28,456 protein-coding genes were annotated, with an average mRNA length of 21,311.9 bp. Each gene had an average of 9.7 exons, and the average exon length was 191.4 bp. Approximately 98.4% (27,991 genes) of the predicted genes were annotated with at least one database, highlighting the high completeness and reliability of the X-ray tetra gene annotation (Table 1, Fig. 3).

Data Records

The final genome assembly, gene set, and raw sequencing reads for *P. maxillaris* are publicly accessible on the NCBI and GenBank under accession numbers PRJNA1190323 and GCA_045781885.1[45]. Annotated coding sequences and protein sequences have been archived on Figshare (https://doi.org/10.6084/m9.figshare.27901167)[46]. Additionally, raw reads obtained by PacBio, ONT and Illumina sequencing are available on the NCBI under accession number SRP552708[47].

Technical Validation

We constructed the 25 chromosomes for *P. maxillaris* with a high-quality Hi-C assembly. PacBio HiFi reads aligned to the genome assembly using Minimap2 achieved a mapping rate of 100.0%. Completeness was further evaluated with BUSCO v.5.2.2, using the Actinopterygii database (3,640 single-copy orthologs; OrthoDB v.10) as the reference. In short, our results indicated that 98.0% (3,567) of BUSCO genes were complete, with 97.1% (3,536) identified as single-copy, 0.9% (31) as duplicated, and 0.6% (22) as fragmented. A protein-level BUSCO assessment using the Actinopterygii dataset revealed 96.5% (3,512) completeness, with 95.2% (3,466) single-copy, 1.3% (46) duplicated, and 1.3% (48) fragmented orthologs, confirming high-quality gene annotation. The Merqury QV score of 41.8 further validated the high quality and reliability of our genome assembly and annotation.

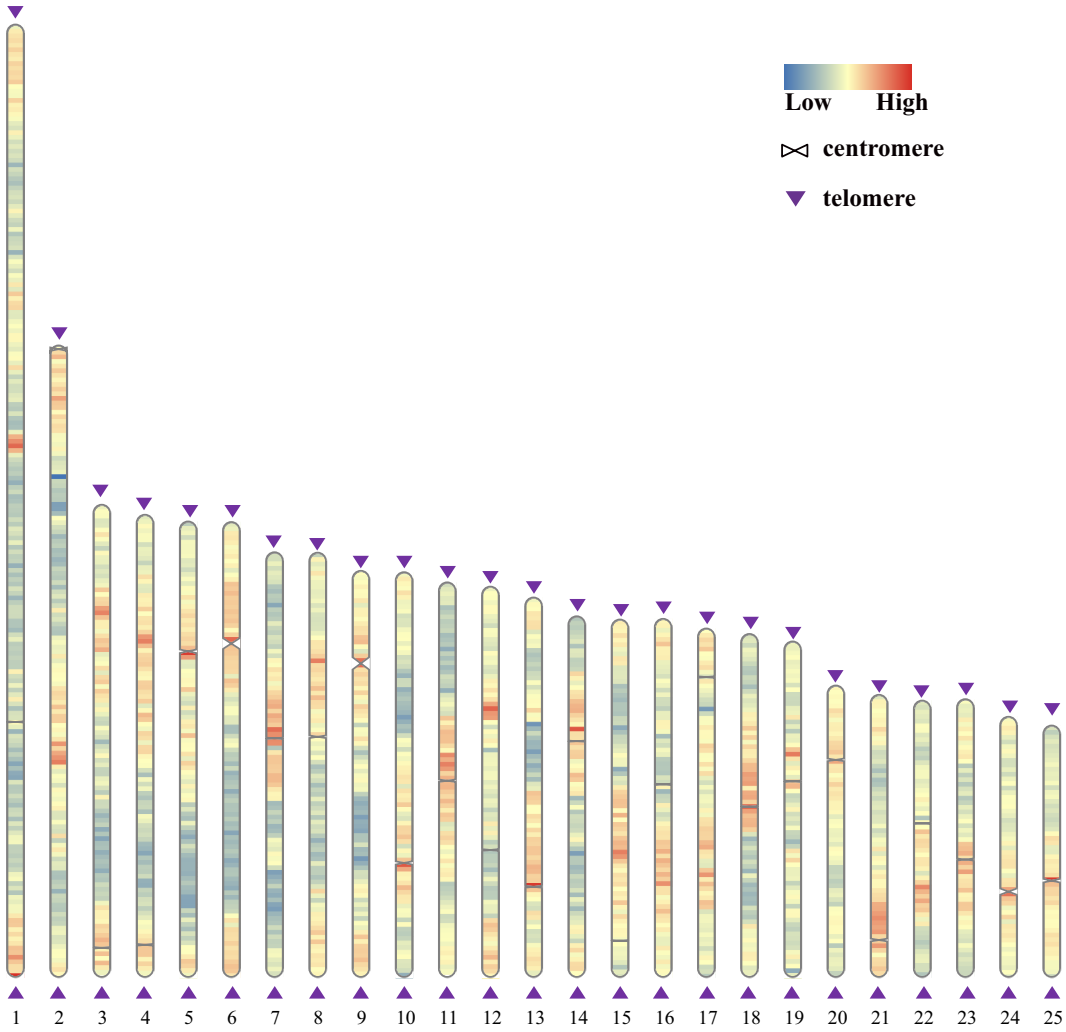

Fig. 3 Repeat distribution, telomere and centromere locations in the *P. maxillaris* genome. Arrows mark the complete set of 50 telomeres.

Code availability

All scripts and pipelines used for the genome assembly and gene annotation followed the standard manuals and protocols of the applied bioinformatics software. No specific code was developed for this study.

Received: 30 December 2024; Accepted: 13 March 2025;
Published online: 24 March 2025

References

1. Lima, F. C. T. *et al.* A new miniature *Pristella* (Actinopterygii: Characiformes: Characidae) with reversed sexual dimorphism from the rio Tocantins and rio Sao Francisco basins, Brazil. *Can J Zool* **99**, 339–348 (2021).
2. Ma, K. *et al.* Cloning and characterization of nicotinic acetylcholine receptor γ-like gene in adult transparent *Pristella maxillaris*. *Gene* **769**, 145193 (2021).
3. Bian, F. F. *et al.* Morphological characteristics and comparative transcriptome analysis of three different phenotypes of *Pristella maxillaris*. *Front Genet* **10**, 698 (2019).
4. Ponpornpisit, A. *et al.* Experimental infections of a ciliate Tetrahymena pyriformis on ornamental fishes. *Fisheries Sci* **66**, 1026–1031 (2000).
5. Conde-Saldana, C. C. *et al.* A New *Pristella* (Characiformes: Characidae) from the Rio Orinoco Basin, Colombia, with a Redefinition of the Genus. *Copeia* **107**, 439–446 (2019).
6. Ward, A. J. W. *et al.* The physiology of leadership in fish shoals: leaders have lower maximal metabolic rates and lower aerobic scope. *J Zool* **305**, 73–81 (2018).
7. Schaerf, T. M. *et al.* The effects of external cues on individual and collective behavior of shoaling fish. *Sci Adv* **3**, e1603201 (2017).
8. Wilson, A. D. M. *et al.* Conformity in the collective: differences in hunger affect individual and group behavior in a shoaling fish. *Behav Ecol* **30**, 968–974 (2019).
9. Tang, C. B. *et al.* First determination and analysis of the complete mitochondrial genome of X-ray tetra *Pristella maxillaris* (Ulrey, 1894) (Actinopteri, Characidae). *Mitochondrial DNA B* **7**, 253–254 (2022).
10. Huang, Y. *et al.* Fish genomics and its application in disease-resistance breeding. *Rev Aquacult* **17**, e12973 (2025).

11. Bian, C. *et al*. A chromosome-level genome assembly for the astaxanthin-producing microalga *Haematococcus pluvialis*. *Sci Data* **10**, 511 (2023).
12. Zhang, K. *et al*. A chromosome-level reference genome assembly of the Reeve's moray eel (*Gymnothorax reevesii*). *Sci Data* **10**, 501 (2023).
13. Yang, Y. X. *et al*. Gap-free chromosome-level genomes of male and female spotted longbarbel catfsh *Hemibagrus guttatus*. *Sci Data* **11**, 572 (2024).
14. Wen, Z. Y. *et al*. Chromosome-level genome assemblies of vulnerable male and female elongate loach (*Leptobotia elongata*). *Sci Data* **11**, 924 (2024).
15. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genom Proteom Bioinf* **13**, 278–289 (2015).
16. Chin, C. S. *et al*. Nonhybrid, fnished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**, 563–569 (2013).
17. Chen, Y. *et al*. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nat Commun* **12**, 60 (2021).
18. Chen, S. F. *et al*. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, 884–890 (2018).
19. Hu, J. *et al*. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol* **25**, 107 (2024).
20. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–U354 (2012).
21. Servant, N. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* **16**, 259 (2015).
22. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98 (2016).
23. Dudchenko, O. *et al*. De novo assembly of the genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
24. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**, 99–101 (2016).
25. Xu, G. C. *et al*. LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *Gigascience* **8**, giy157 (2019).
26. Xu, M. *et al*. TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *Gigascience* **9**, giaa094 (2020).
27. Lin, Y. *et al*. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic Res* **10**, uhad127 (2023).
28. Simao, F. A. *et al*. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
29. Rhie, A. *et al*. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21** (2020).
30. Abrusán, G. *et al*. TEclass-a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329–1330 (2009).
31. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* **35**, W265–W268 (2007).
32. Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **Chapter 4**, Unit 4 10 (2004).
33. Jurka, J. *et al*. Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462–467 (2005).
34. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573–580 (1999).
35. Mount, D. W. Using the Basic Local Alignment Search Tool (BLAST). *CSH Protoc* **2007**, pdb top17 (2007).
36. Birney, E. *et al*. GeneWise and Genomewise. *Genome Res* **14**, 988–995 (2004).
37. Kim, D. *et al*. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915 (2019).
38. Cantarel, B. L. *et al*. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* **18**, 188–196 (2008).
39. Kulikova, T. *et al*. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res* **32**, D27–30 (2004).
40. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365–370 (2003).
41. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29 (2000).
42. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
43. Ogata, H. *et al*. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **27**, 29–34 (1999).
44. Hunter, S. *et al*. InterPro: the integrative protein signature database. *Nucleic Acids Res* **37**, D211–215 (2009).
45. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_045781885.1 (2025).
46. Bian, C. *et al*. The genome and annotation of *Pristella maxillaris*. *Figshare* https://doi.org/10.6084/m9.figshare.27901167 (2024).
47. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP552708 (2025).

Acknowledgements

This study was supported by Shenzhen Special Fund for Sustainable Development (no. KCXFZ20211020164013021), Guangxi Major Program for Science and Technology (no. GuikeAA24263042), and the Engineering Laboratory Support Program from Development and Reform Commission of Shenzhen Municipality (no. XMHT20220104019).

Author contributions

Q.S. and C.B. conceived this project; C.H and C.B. collected samples; C.B., C.H, Z.H., and Q.S assembled the genome and performed annotation; C.B., C.H and Z.L analyzed the data. C.B., C.H, Z.H. and Q.S wrote the manuscript; Q.S. and Z.L revised the manuscript. All authors have read and approved the final manuscript for publication.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.L. or Q.S.