Scientific Data

https://doi.org/10.1038/s41597-026-06849-5

Article in Press

Near telomere-to-telomere genome assembly of the stone loach (*Traccatichthys pulcher*)

Received: 26 August 2025

Accepted: 6 February 2026

Cite this article as: Du, L.-N., Wang, Z.-C., Chen, Z.-N. *et al.* Near telomere-to-telomere genome assembly of the stone loach (*Traccatichthys pulcher*). *Sci Data* (2026). https://doi.org/10.1038/s41597-026-06849-5

Li-Na Du, Zhuo-Cong Wang, Zhuo-Ni Chen, Zhi-Xian Qin & Chen-Hong Li

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Near telomere-to-telomere genome assembly of the stone loach (*Traccatichthys pulcher*)

Li-Na Du[1,2*], Zhuo-Cong Wang[3], Zhuo-Ni Chen[1,2], Zhi-Xian Qin[1,2], Chen-Hong Li[4]

1, Key laboratory of Ecology of Rare and Endangered Species and Enviornmental Protection, Guangxi Normal University, Ministry of Education, Guilin, Guangxi 541004, China

2, Guangxi Key Laboratory of Rare and Endangered Animals Ecology, College of Life Sciences, Guangxi Normal University, Guilin, Guangxi 541004, China

3, Changbaishan Academy of Science, Erdaobaihe, Jilin 133613, China

4, College of Fisheries and Life Science, Shanghai Ocean University, Shanghai 201306, China

*, Corresponding authors: e-mail: Li-Na Du (dulina@mailbox.gxnu.edu.cn)

Abstract. *Traccatichthys pulcher* is an ornamental loach species recognized for its vibrant body coloration, characteristic black dorsal fin margin, and iridescent green lateral stripes. To advance genomic research on this species, a high-quality, near telomere-to-telomere (T2T) genome assembly was generated using PacBio HiFi, ONT ultra-long, and Hi-C sequencing technologies. The resulting haplotype-resolved assembly spanned approximately 623.68 Mb, with a contig N50 of 22.9 Mb, and was anchored onto 24 chromosomes. Telomeric sequences were detected at both ends of eight chromosomes and at one end of 13 chromosomes. Twenty-three chromosomes were entirely gapless, while a single gap was identified in the remaining chromosome. The assembly contained 119.1 Mb of repetitive elements, and 23 967 protein-coding genes were annotated. BUSCO analysis indicated high completeness, with 98.6% of conserved genes recovered. This high-quality, near T2T genome assembly offers a valuable and robust genetic resource for investigating molecular mechanisms, evolutionary processes, conservation biology, and selective breeding of *T. pulcher*.

Background & Summary

The loach genus *Traccatichthys* (family Nemacheilidae), established by Freyhof and Serov in 2001, comprises six recognized species distributed across southern China—Guangdong, Guizhou, Fujian, Hainan Island, and Guangxi Zhuang Autonomous Region—as well as parts of Vietnam and Laos[1-4]. Among them, *Traccatichthys pulcher* occupies a geographically fragmented range encompassing Hainan Island, the Pearl River basin in Guangxi, Guangdong, and Guizhou, and coastal watersheds in Guangxi and Guangdong[5-6]. Although the species was originally described from a single specimen collected in Nada (=Nodoa), Hainan Island[7], subsequent field investigations have questioned its native status in that locality. Du et al.[6] posited that the early records from Nada likely refelect anthropogenic translocation from mainland. More recent surveys by Qin et al.[4] identified individuals with diagnostic *T. pulcher* morphology in Dongfang City and Baisha Li Autonomous County, indicating that native populations may occur in these regions of Hainan Island.

Traccatichthy pulcher exhibits striking phenotypic traits, including vivid lateral striping and a sharply defined black margin on the dorsal fin, which contribute to its high demand in ornamental markets and its growing relevance in evolutionary and developmental research[8] (Fig. 1A). However, the absence of a high-quality reference genome has impeded progress in population-level and phylogenomic analyses, and its establishment is essential for advancing taxonomic resolution within the genus *Traccatichthys.*

In this study, a high-quality, near telomere-to-telomere (T2T) genome assembly of *T. pulcher*, constructed using 50× PacBio HiFi data and 100× Hi-C data. The final assembly spans 623.68 Mb and comprised 24 chromosome-level scaffolds, 23 of which were completely gap-free. Telomeric repeats were resolved at both ends of eight chromosomes and at one end of 13 chromosomes, indicating high assembly completeness. A total of 23 967 protein-coding genes were annotated, with 98.6% assigned functional annotations. The near T2T genome generated in this study represents the first genome assembly for any species within the genus *Traccatichthys*. It fills the current gap in high-quality genomic resources for this genus, provides critical references for molecular and evolutionary research, and supports conservation and breeding programs targeted at restoring natural populations and advancing sustainable aquaculture. Additionally, it directly lays a foundation for identification and selective breeding of ornamental traits (e.g., genes associated with pigment synthesis).

Methods

Ethics statement. All procedures complied with the Implementation Rules of the Fisheries Law of the People's Republic of China and the Laboratory Animal Guidelines for the Ethical Review of Animal Welfare (GB/T 35892–2018).

Sample collection and sequencing. A specimen of *T. pulcher* was obtained from the Xia Village, Gancheng Town, Dongfang City, Hainan Island, China. Genomic DNA was extracted using a Blood & Tissue DNA Kit (Qiagen 19086), and its quality and concentration were assessed using a NanoDrop One spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA).

For PacBio HiFi sequencing, a SMRTbell library was prepared using the SMRTbell Express Template Prep Kit 3.0 (Pacific Biosciences, CA, USA, 102-182-700). Sequencing was performed on the PacBio REVIO paltform, generating 51.57 Gb of HiFi data (~50×) (Table 1).

For Oxford Nanopore sequencing, genomic libraries were constructed using a SQK-LSK110 Ligation Kit (Oxford Nanopore Technologies, Oxford, UK), following the manufacturer's standard protoclos. The purified library was loaded onto primed R9.4 Spot-On Flow Cells and sequenced using a PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK) with 48-hours runs conducted at Wuhan Benagen Technology Co. Ltd. (Wuhan, China). Base-calling was performed using GUPPY (v0.3.0), yielding 38.65 Gb ONT ultra-long reads.

Hi-C sequencing was conducted using high-quality DNA extracted from muscle tissue. Chromatin was crosslinked with formaldehyde and processed using an *in situ* Hi-C protocol based on Dnase digestion, as established in previous research[9]. The resulting libraries were sequenced on the Illumina NovaSeq (Illumina, San Diego, CA, USA) with 150 bp paired-end reads. Raw data quality was assessed using Juicer (v1.6)[10] with default parameters, producing 100.69 Gb of raw data (~100× coverage).

For transcriptome sequencing, total RNA was extracted from three tissues, including brain, skeletal tissue, and muscle, and pooled for library preparation. A strand-switching protocol was employed using a cDNA-PCR Sequencing Kit (SQK-PCS109), followed by adapter ligation. The libraries were sequenced on the Illumina NovaSeq 6000 platform, yielding 14.33 Gb of transcriptomic data.

Genome survey and assembly. To assess genome characteristics of *T. pulcher*, a K-mer frequency distribution was generated using the kmer_freq module in GenomeScope (v2.0). The estimated genome size was 534.7 Mb, with heterozygosity inferred from K-mer complexity patterns (Fig. 1B).

De novo assembly was performed using a combination of PacBio Hifi reads, ultra-long ONT reads, and Hi-C data. Two haplotype-resolved contig-level assemblies were generated with Hifiasm (v0.19.9-r616) under default parameters (Fig. 1C). Hi-C data were processed through the 3D-DNA pipeline[11] for scaffold clustering, sorting, and orientation. Manual curation and error correction were conducted using JuiceBox (v2.13.07)[10], with chromatin interaction heatmaps guiding the correction of misassemblies and structural inconsistencies in contig positioning (Fig. 1D). Gap filling was carried out using TGS-GapCloser (v1.2.1), which integrated ONT ultra-long reads and Hifiasm-assembled contigs to span unresolved regions. Gaps were closed by selecting the longest and most consistent aligned sequences across reads, replacing scaffold gaps with high-confidence sequences. The final assembly yielded a high-quality reference genome with total length of 623.68 Mb, a scaffold N50 of 22.9 Mb, and a GC content of 37.05%. Approximately 97.02% of the assembled sequences were anchored to 24 pseudochromosomes, including of 23 chromosomes that were entirely gap-free and one chromosome (chromosome 24) containing a single 500 bp gap (Table 2). The gap is at positions 11 465–11 964 on chromosome 24. This region corresponds to a telomeric repeat region (spanning positions 12–15 942) and containing approximately 15 kb of highly repetitive non-coding DNA, with "CTAACC" as the repeating unit. A single sequencing read cannot fully span this repetitive region, nor can it distinguish the specific positions of these fragments within the region. Consequently, the non-repetitive sequences flanking the repetitive region cannot be accurately connected, ultimately leading to an uncloseable gap.

→ Figure 1 goes here.

Genome annotation. Repetitive element annotation was conducted using HiTE[12], which systematically identified and masked transposable elements (TEs) across the assembled genome. This analysis revealed that

19.1% of the genome was annotated as repetitive sequences, a proportion consistent with estimates from genome survey data (Table 3). The distribution of genes and repeats on chromosomes followed the typical pattern observed in vertebrate genomes, with higher gene densities in GC-rich region and lower gene densities in repeat-rich distal and pericentromeric regions (Fig. 1C).

Protein-coding gene annotation was achieved through an integrative framework that combined *de novo* prediction, homology-based alignment, and transcriptome-based-guided reconstruction. *De novo* gene models were predicted using Augustus (v3.5.0)[13] and GALBA (v1.0.11)[14]. For homology-based inference, Miniport (v0.13)[15] was used to align the *T. pulcher* genome against protein sequences from closely related species, including *Dano rerio*, *Misgurnus anguillicaudatus*, *Triplophysa lixianensis*, *Triplophysa rosa*, and *Triplophysa tibetana*, generating a comprehensive set of orthologous gene models. Transcriptomic data were integrated using BRAKER3 (v3.0.3), which harmonized RNA-seq evidence with homology-derived models for enhanced structural accuracy. The final gene models were constructed by merging evidence from all three annotation approaches using EvidenceModeler (v2.10)[16], followed by structural refinement with PASA (v2.5.3)[17]. This process resulted in the annotation of 23 967 protein-coding genes.

Data Records

The raw sequence data of *T. pulcher* in this paper have been deposited in the Genome Sequence Archive[18] in National Genomic Data Center[19], China National Center for Bioinfomation/Beijing Institute of Genomics, Chinese Academy of Sciences (GSA: CRA029066)[20]. Associated resources—including genome assemblies, genome annotation, coding sequences (CDS), and protein datasets—have been submitted to Figshare[21] and the European Nucleotide Archive (ENA) at EMBL-EBI under accession number GCA_977009865[22]

Technical Validation

A multi-tiered validation framework was implemented to assess the accuracy and integrity of the *T. pulcher* genome. Hi-C interaction heatmaps revealed well-defined chromatin contact patterns across all chromosomes, supporting accurate scaffold ordering and orientation. Comparative collinearity analysis using LAST (v1.17.0)[23] and JCVI (v0.9.13)[24] demonstrated strong syntenic conservation with the genome *of Oreonectes platycephalus* (Fig. 2), reinforcing the structural reliability of the assembly. Assembly completeness was evaluated using BUSCO against Actinopterygii_odb10 in the ortholog database, yielding 97% completeness, indicative of a nearly complete gene set. Base-level accuracy was quantified with Merqury (v1.3), which estimated a consensus quality value (QV) of 37.8. Structurally, 23 chromosomes were completely gap-free, and telomeric sequences were detected on 21 chromosomes—present at both termini of eight chromosomes and at one terminus of 13 chromosomes. The absence of high-density terminal telomeric signals at both ends of few chromosomes is not necessarily due to the poor assembly of these region. Telomeres could also be lost or gradually shortened on these chromosomes. In conclusion, this highly accurate near T2T genome offers critical insights into the genetic architecture of a stream-adapted freshwater loach and establishes a platform for dissecting the evolutionary, ecological, and physiological responses of riverine fish to dynamic montane environments.

→ Figure 2 goes here.

Data Avaliability

All data generated or analyzed during this study are openly available. Raw sequencing data—including PacBio Hifi reads, ultra-long ONT reads, Hi-C data, and three transcriptome sequence—have been deposited in Genome Sequence Archive[18] at National Genomic Data Center[19], China National Center for Bioinfomation/Beijing Institute of Genomics, Chinese Academy of Sciences at accession number GSA: CRA029066 that are publicly accessible at https://ngdc.cncb.ac.cn/gsa/s/5ea4tmv2. Associated resources, including genome assemblies, genome annotation, coding sequences (CDS), and protein datasets, have been submitted to Figshare[21] (https://doi.org/10.6084/m9.figshare.29607215.v2), and the European Nucleotide Archive (ENA) at EMBL-EBI under accession number GCA_977009865[22] (https://www.ebi.ac.uk/ena/browser/view/GCA_977009865).

Code availability

All computational tools used in this study are open-source and publicly available. Software versions and parameter settings are described in the Methods section. When unspecified, programs were executed using default settings according to the developers' documentation.

Funding information

National Natural Science Foundation of China (NSFC32560121). Investigation of Fishery Resources and Habitat in the Pearl River Basin (ZJZX-10). Investigation and Monitoring Project of Biodiversity Backgroud Resources in Guilin Nature Reserve (GLZC2024-C3-990742-GXXL), Comprehensive Scientific Investigation Project of Guangxi Daguishan Crocodile Lizard National Nature Reserve, and Jilin Provincial Key R&D Project (20210202082NC).

Acknowledgement

We would like to thank Yuan-Sheng Huang from Wuhan Benagen Technology Co., Ltd for his assistance in project management. We are grateful to Yu-Fan Liang from Sanya Boao Biotechnology Co., Ltd for his help in genome annotation. We greatly thank Jia-Jun Zhou from Zhejiang Forest Resource Monitoring Center for collecting and taking living photo of *T. pulcher*. We thank C. Watts for English corrections and suggestions.

Author Contributions & Competing Interests

Li-Na Du and Zhuo-Cong Wang analyzed the data, prepared the manuscript, and providing funding acquisition. Zhuo-Ni Chen and Zhi-Xian Qin assembly and annotation of genome. Chen-Hong Li conceived and designed the study. All authors read and approved the final version of the manuscript. The authors declare that they have no competing interests.

Reference

1. Jin, X. B. *The Fishes of Fujian Province*. Fujian Science and Technology Press, 385–386 (1984). [In Chinese]
2. Kottelat, M. Fishes of Laos. Wildlife Heritage Trust, Colombo, 198 pp (2001).
3. Kottelat, M. The fishes of the inland waters of southeast Asia: a catalogue and core bibliography of the fishes known to occur in freshwaters, mangroves and estuaries. *Raffles Bulletin of Zoology*, Supplement 27: 1–663

(2013).

4. Qin, Z. X., Zhou, J. J., Du, L. N. & Lin F. *Traccatichthys punctulatus* sp. nov., a new species of stone loach (Pisces, Nemacheilidae) from Guangxi, southern China. Zoosyst. Evol., 101(3): 1013 – 1021 https://doi.org/10.3897/zse.101.146077 (2025).
5. Qiu, C. F., Lin, Y. G., Qin, N., Zhao, J. & Chen, X. L. Genetic variation and phylogeography of *Micronoemacheilus pulcher* populations among basin systems between western South China and Hainan Island. *Acta Zoologica Sinica*, 54: 805–813 (2008).
6. Du, C. X., Zhang, E. & Chan, B. P. L. *Traccatichthys tuberculum*, a new species of Nemacheiline loach from Guangdong Province, South China (Pisces, Balitoridae). *Zootaxa*, 3586: 304 – 312 https://doi.org/10.11646/zootaxa.3586.1.28 (2012).
7. Nichols, J. T. & Pope, C. H. The fishes of Hainan. *Bulletin of the American Museum of Natural History*, 54: 321–394 (1927).
8. Zheng, S. M., Wu, Q., He, L. J. & Zhang, Z. G. *Native ornamental fish of China illustrated book*. Beijing: Science Press; 219 pp (2016). [in Chinese]
9. Ramani, V., Deng, X. X., Qiu, R. L., Lee, C. L., Disteche, C. M., Noble, W. S., Shendure, J. & Duan, Z. J. Sci-Hi-C: a single-cell Hi-C method for mapping 3D genome organization in large number of single cells. *Methods*, 170: 61–68 https://doi.org/10.1016/j.ymeth.2019.09.012 (2020).
10. Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S. & Aiden, E. L. 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).
11. Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., Aiden, A. P. & Aiden, E. L. 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).
12. Hu, K. et al. HiTE: a fast and accurate dynamic boundary adjustment approach for full-length transposable element detection and annotation. Nature Communications 15, https://doi.org/10.1038/s41467-024-49912-8 (2024).
13. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic acids research 34, W435–W439, https://doi.org/10.1093/nar/gkl200 (2006).
14. Bruna, T. et al. Galba: genome annotation with miniprot and AUGUSTUS. BMC bioinformatics 24, https://doi.org/10.1186/s12859023-05449-z (2023).
15. Li, H. Protein-to-genome alignment with miniprot. Bioinformatics (Oxford, England) 39, https://doi.org/10.1093/bioinformatics/btad014 (2023).
16. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome biology 9, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
17. Do, V. H. et al. Pasa: leveraging population pangenome graph to scaffold prokaryote genome assemblies. Nucleic acids research 52, https://doi.org/10.1093/nar/gkad1170 (2024).
18. The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data Types. Genomics, Proteomics & Bioinformatics 2021, 19(4):578-583. https://doi.org/10.1016/j.gpb.2021.08.001 [PMID=34400360]
19. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2025. Nucleic Acids Res 2025, 53(D1):D30-D44. https://doi.org/10.1093/nar/gkae978 [PMID=39530327]

20. Genome Sequence Archive (GSA). Du, L. N. *Traccatichthys pulcher*, whole genome sequencing project. National Genomic Data Center https://ngdc.cncb.ac.cn/gsa/s/5ea4tmv2: CRA029066 (2025).
21. Du, L. N. *Traccatichthys pulcher* (Nichols and Pope, 1921). Figshare. Dataset. https://doi.org/10.6084/m9.figshare.29607215.v2 (2025).
22. European Nucleotide Archive (ENA). https://www.ebi.ac.uk/ena/browser/view/GCA_977009865 (2025).
23. Frith, M. C., Hamada, M. & Horton, P. Parameters for accurate genome alignment. BMC Bioinformatics 11, 1–14 http://doi.org/10.1186/1471-2105-11-80 (2010).
24. Tang, H. B. et al. Synteny and collinearity in plant genomes. Science 320, 486–488 http://doi.org/10.1126/science.1153917 (2008).

Figure and Table captions

Fig. 1. The characteristics of *Traccatichthy pulcher* morphology and genome. A. Morphological characters of *Traccatichthy pulcher* used for genome sequencing. B. GenomeScope *k*-mer analysis (*k* = 19) of whole-genome sequencing reads. C. The circos plot of genomic features: arranged from outside to inside, (I) 24 chromosomes assembly; (II) GC density; (III) TE density; (IV) gene density, and links of intragenomic syntenic blocks within 100Kbp sliding windows. D. The Hi-C heatmap of chromosome interaction, Chr1-Chr24 is an abbreviation for 24 chromosomes. The abscissa and ordinate represent the order of each bin on the corresponding chromosome. The color from light to dark indicates the strength of the interaction from low to high.

Fig. 2. Synteny analysis of the chromosomes between *Traccatichthy pulcher* and *Oreonectes platycephalus*.

Table 1. Statistics of the sequencing data

Table 2. Assembly statistics of chromosomes.

Unplaced: sequences that could not be anchored to any known chromosome.

Table 3. Statistic results of different types of annotated repeat content.

Table 1. Statistics of the sequencing data

Reads Type	Platform	Tissue	Total Reads	Data size (Gb)	Average depth (×)	Average length (bp)	Max length (bp)	N50 length of reads (bp)
ONT ultra-long data	PROMETHion sequencer	muscle	378,241	33.18	50	102,182.79	645,345	100,100
PacBio CCS reads	PacBio REVIO	muscle	2,578,309	51.57	50	20,000	59,988	20,176
Hi-C data	Illumina Novaseq 6000	muscle	335,664,963	100.7	100	150	150	150
RNA-Seq data	Illumina Novaseq 6000	brain	30,758,222	4.61	10	150	150	150
RNA-Seq data	Illumina Novaseq 6000	bone	28,776,449	4.32	10	150	150	150
RNA-Seq data	Illumina Novaseq 6000	muscle	35,985,584	5.4	10	150	150	150

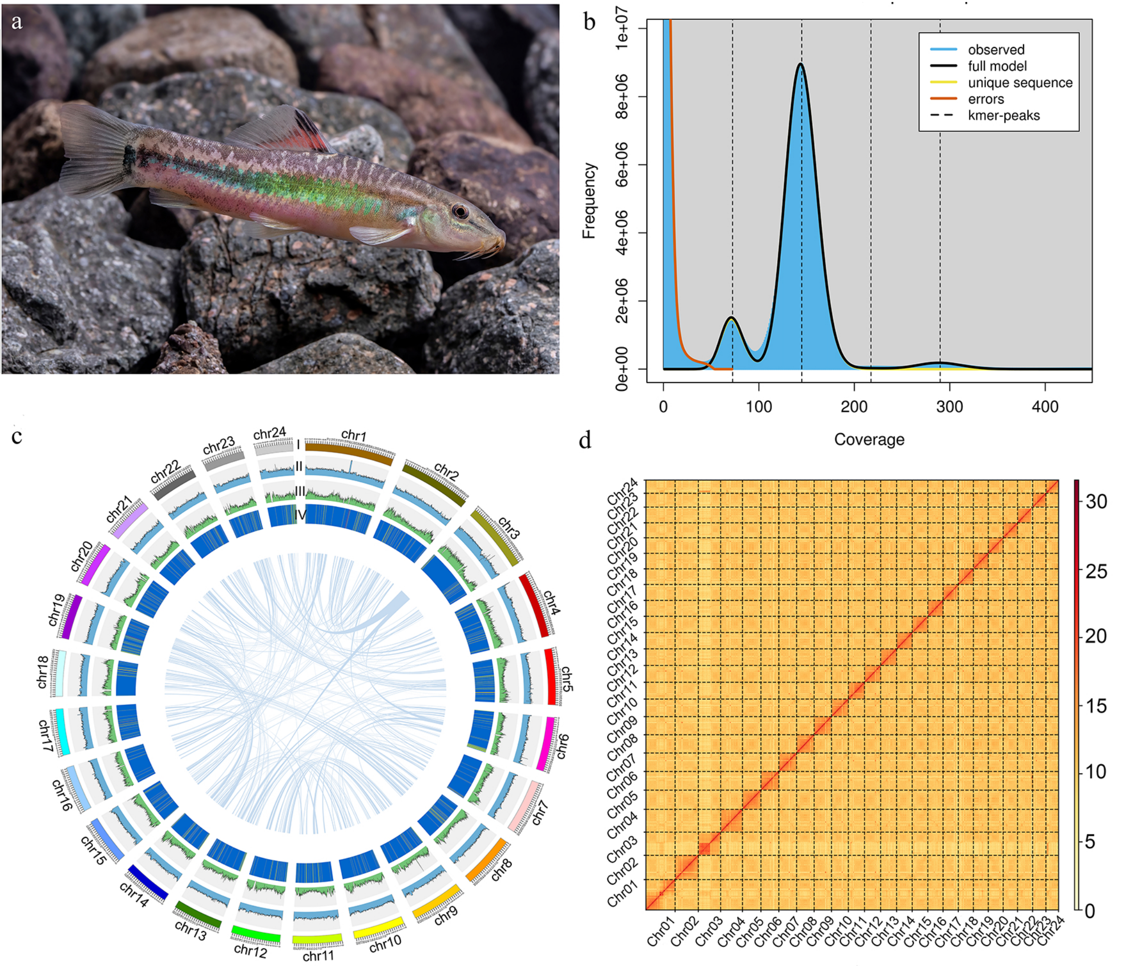
Table 2. Assembly statistics of chromosomes.

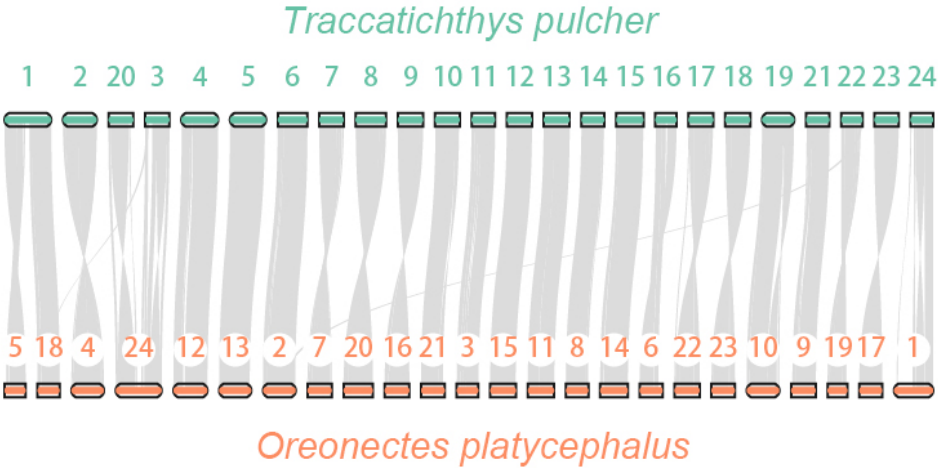
Chromosome number	Length (Mb)	Number of gaps	Number of telomeres
Chr1	41.16	0	0
Chr2	32.87	0	Right
Chr3	31.50	0	Left
Chr4	30.64	0	Rigth
Chr5	26.11	0	0
Chr6	25.18	0	0
Chr7	25.14	0	Both
Chr8	24.70	0	Both
Chr9	24.40	0	Left
Chr10	23.72	0	Right
Chr11	22.90	0	Right
Chr12	22.77	0	Right
Chr13	22.52	0	Right
Chr14	21.93	0	Both
Chr15	21.91	0	Left
Chr16	21.52	0	Right
Chr17	21.58	0	Both
Chr18	21.32	0	Right
Chr19	20.96	0	Both
Chr20	20.81	0	Left
Chr21	20.32	0	Both
Chr22	21.35	0	Left
Chr23	18.78	0	Both
Chr24	17.37	1	Both
Unplaced	4.22		

Unplaced: sequences that could not be anchored to any known chromosome.

Table 3. Statistic results of different types of annotated repeat content.

Type	Length (bp)	% of genome
DNA	101,086,059	16.21
LINE	332,027	0.05
SINE	570,104	0.09
LTR	4,276,705	0.69
Unknown	12,835,421	2.06
Total	119,100,316	19.1


a
b
Frequency
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
0
100
200
300
400
Coverage
observed
full model
unique sequence
errors
kmer-peaks
c
I
II
III
IV
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chr23
chr24
d
Chr01
Chr02
Chr03
Chr04
Chr05
Chr06
Chr07
Chr08
Chr09
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22
Chr23
Chr24
0
5
10
15
20
25
30


Traccatichthys pulcher
1 2 20 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 21 22 23 24
5 18 4 24 12 13 2 7 20 16 21 3 15 11 8 14 6 22 23 10 9 19 17 1
Oreonectes platycephalus