scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly and annotation of largemouth bronze gudgeon (*Coreius guichenoti*)

Xuemei Li, Xingbing Wu, Yongjiu Zhu, Xiaoli Li, Zihao Meng, Nian Wei, Miao Xiang, Deguo Yang & Tingbing Zhu ✉

***Coreius guichenoti*, mainly distributed in upstream regions of the Yangtze River China, is currently on the brink of extinction and listed as national secondary protected animal. In this study, we aimed to obtain the chromosome-level genome of *C. guichenoti* using PacBio and Hi-C techniques. According to the PacBio sequencing, *C. guichenoti* genome was successfully assembled to 1100.1 Mb size, with a Contig N50 size of 25.0 Mb, and containing 731.0 Mb of repeats. Hi-C sequencing data was utilized for chromosome assembly and 25 chromosome sequences were ultimately yielded, with a total length of 1076.8 Mb. Moreover, a total of 22,506 protein-coding genes were predicted with average intron length of 2293 bp. Evolutionary analysis and divergence time prediction revealed that *C. guichenoti* was closely related to *C. heterodon* and they phylogenetically diverged from common ancestor ~20.7 million years ago (Mya), following the separation of Cyprinidae at 28.3 Mya. In the future, the utilization of comparative genomics research is important in elucidating the molecular mechanisms of Ichthyophthirius disease and ensuring the conservation of biological resources.**

Background & Summary

The largemouth bronze gudgeon (*Coreius guichenoti*) distributed in upstream regions of the Yangtze River, was ever an important economic fish in southwest China and the maximum weight it is able to reach was 4.0 kg[1,2]. However, the population of this species has significantly decreased in recent decades and was close to extinction as a result of overfishing, construction of hydropower stations, habitat destruction and so on[3,4]. In 2021, *C. guichenoti* (wild population only) was listed as a secondary protected animal according to the Catalog of Wildlife under Key State Protection of China. At present, the introduction of juvenile *C. guichenoti* into natural habitat is a crucial method for protecting this species and a successful program for artificial breeding has been implemented for this species with the scientists' efforts for more than ten years[5–7]. In that case, the conservation of their germplasm resources and preventing any potential decline was obviously important.

As a protected species, *C. guichenoti* has been widely studied on reproduction and development, dietary habits, physiological and biochemical aspects, artificial breeding and stress relief[4,8–11]. However, research on its whole genome is limited. The whole genome information of an organism serves as the foundation for investigating the species germplasm resources and assessing population genetic structure and diversity and evolutionary history, which plays a crucial role in the management of fishery resources and in the advancement of their sustainable utilization, especially in the field of aquaculture for new variety breeding[12–14]. As the cost of next-generation technology decreases, an increasing number of organism genomes have been decoded recently, such as pelagic fish (*Decapterus maruadsi*)[12], Schizothoracine fish (*Gymnocypris eckloni*)[13], freshwater mussel (*Sinosolenaia oleivora*)[15], dark sleeper (*Odontobutis potamophila*)[16] and big-head schizothorcin (*Aspiorhynchus laticeps*)[17].

In the present study, we conducted a novel assembly of the genome of *C. guichenoti* by employing a variety of sequencing techniques, such as PacBio long-reads and high-throughput chromosome conformation capture (Hi-C) technology. Subsequently, we carried out genome annotation, performed evolutionary analysis, and characterized genome features based on the assembled genome sequences. Our results provide a foundation for

Key Laboratory of Freshwater Biodiversity Conservation, Ministry of Agriculture and Rural Affairs, Yangtze River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Wuhan, 430223, China. ✉e-mail: zhutb@yfi.ac.cn

Library	Sequencing platform	Clean data (Gb)	Read N50 (bp)	Sequence coverage (×)	Contig N50 (Mb)	GC content (%)	Q20 (%)	Q30 (%)
Short reads	BGI T7	114.7	150	104.3	—	38.8	98.7	95.5
Long reads	PacBio Sequel II	37.4	13,504	34.0	25.0	39.1	—	—
Hi-C	BGI T7	94.2	150	85.6	—	41.5	98.3	94.3

Table 1. Sequencing data used for the genome assembly of *C. guichenoti*.

further genome-wide studies on *C. guichenoti* such as development of new breeding varieties and will be useful for studying the evolution of *C. guichenoti*.

Methods

Ethics statement. The fish management and sampling were conducted in accordance with animal care protocols (No. 2022YFI-ZTB-02) approved by the Committee on Animal Ethics of Yangtze River Fisheries Research Institute, Chinese Academy of Fishery Sciences.

Sequencing libraries. A four-year-old adult female *C. Guichenoti* (body weight 180.6 g, total length 25.7 cm) was collected from the domestication and breeding base of endemic fishes of the Yangtze River, Yangtze River Fisheries Research Institute, Chinese Academy of Fishery Sciences (Jingzhou, Hubei, China). Genomic DNA sequencing libraries was constructed using muscle tissue, while transcriptome sequencing libraries were created using muscle, heart, kidney, gill, liver, and spleen tissues. All the tissues were preserved in liquid nitrogen until they were utilized. High-quality genomic DNA were extracted from the muscle samples by using the cetyltrimethylammonium bromide (CTAB) method. The quality and quantity of the extracted DNA were examined using electrophoresis on a 1% agarose gel, and the results were analyzed with a spectrophotometer, respectively.

Library preparation is performed by using MGIEasy Universal DNA Library Prep Set (Frasergen-Wuhan, China). A certain amount of genomic DNA is taken and subjected to fragmentation processing. Fragmented samples are size selected through magnetic beads. Then the size selected DNA fragments are converted to blunt-end DNA with end repair reaction. A single adenosine is added on the 3′ end of DNA through the A tailing reaction. Later the library adaptors are connected to the two ends of DNA by adaptor ligation. Finally, the library products are amplified through PCR reaction and subjected to quality control process. Next, the final double strand library products are denatured to generate the single stranded library product. Then, the circularization reaction is set up to get single stranded circularized DNA products. Any single strand linear DNA will be digested to remove. The final single strand circularized library is amplified with phi29 and rolling circle amplification (RCA) to generate the DNA nano ball (DNB) which carries about 300 copies of the initial single stranded library molecule. The DNBs are loaded into the patterned nanoarray and sequencing reads of PE150 bases length are generated with DNBSEQ-T7 platform (Frasergen-Wuhan, China).

For SMRT sequencing, a standard HiFi library was prepared according to the SMRTbell Express Template Prep Kit 2.0 manual (Pacific Biosciences, CA, USA). A total amount of 10 μg DNA per muscle sample was used for the DNA library preparations. Sequencing was performed on a PacBio Sequel II platform.

Hi-C libraries were constructed according to previous studies[18]. Briefly, samples were cross-linked with 1% formaldehyde for 10 min at room temperature and quenched with 0.125 M final concentration glycine for 5 min. The cross-linked cells were subsequently lysed. Endogenous nuclease was inactivated with 0.3% SDS, then chromatin DNA was digested by 100U MboI (NEB, MASS, USA), and marked with biotin-14-dCTP (Invitrogen) and then ligated by 50U T4 DNA ligase (NEB, MASS, USA). After reversing cross-links, the ligated DNA was extracted through QIAamp DNA Mini Kit (Qiagen, Düsseldorf, Germany) according to manufacturers' instructions. Purified DNA was sheared to 300- to 500-bp fragments and were further blunt-end repaired, A-tailed and adaptor-added, followed by purification through biotin-streptavidin-mediated pull-down and PCR amplification. Finally, the Hi-C libraries were quantified and sequenced on the MGIseq platform (BGI, China).

Genome survey and assembly. The short-reads from DNBSEQ-T7 platform were quality filtered by SOAPnuke[19] using the following method. Firstly, the adaptors were removed from the sequencing reads. Second, read pairs were excluded if any one end has an average quality lower than 20. Third, ends of reads were trimmed if the average quality lower than 20 in the sliding window size of 5 bp. Finally, read pairs with any end was shorter than 75 bp were removed. A total of 114.7 Gb of high-quality filtered data were produced, providing 104.3 times coverage of the genome, with a Q20 of 98.7% and a Q30 of 95.5%, and GC content of 38.8% (Table 1). The quality filtered reads were used for genome size estimation. We generated the 17-mer occurrence distribution of short reads using the k-mer method GCE (v1.0.2)[20]. Then, we estimated the genome size to be about 1028.9 Mb, and the proportion of repeat sequences and heterozygosity rate of the genome were determined to be approximately 58.4% and 0.47%, respectively.

The HiFi reads were generated using the ccs software (https://github.com/PacificBiosciences/ccs) with the parameter '-minPasses 3'. A total of 37.4 Gb of PacBio HiFi reads were generated to enhance the quality and validate the assemblies. These long (reads N50 = 13.5 kb) and highly accurate (>99%) HiFi reads were assembled using hifiasm (v.0.16)[21] with default parameters, and the gfatools (https://github.com/lh3/gfatools) was used to convert sequence graphs in the GFA to FASTA format. Therefore, *C. guichenoti* genome was successfully assembled to a total size of around 1100.1 Mb, comprising 391 contigs, with a Contig N50 size of 25.0 Mb (Table 2).

Name	Length (bp)		Number	
	Scaffold	Contig	Scaffold	Contig
Max	57,690,069	41,139,736	—	—
Total	1,100,050,391	1,099,921,391	133	391
N10	54,694,860	37,098,204	2	3
N20	47,937,386	35,979,706	5	6
N30	44,861,498	33,378,796	7	10
N40	43,679,610	29,861,093	9	13
N50	41,997,361	25,015,297	12	17
N60	41,331,550	21,063,355	15	21
N70	40,757,113	13,067,915	17	28
N80	38,760,418	7,132,181	20	40
N90	35,515,902	1,873,774	23	72

Table 2. The statistics of length and number for the *de novo* assembled *C. guichenoti* genome.

Hi-C sequencing data were utilized for chromosome assembly of *C. guichenoti*. A total of 94.2 Gb of clean reads was generated from the Hi-C library (Table 1). Hi-C clean reads were mapped to the reference genome by Juicer (https://github.com/aidenlab/juicer, v3)[22] which allowed for the identification of the corresponding chromosome contigs and the determination of their arrangement order based on Hi-C interaction signals. The ab initio 3D-DNA pipeline[23] was applied to scaffold the *C. guichenoti* genome, and Juicebox[22] was used to adjust the scaffolds manually. We have successfully completed the genome assembly, which consists of 25 chromosomes, with a total length of 1076.8 Mb. The Hi-C anchoring rate refers to the proportion of the genome that has been anchored after Hi-C mapping compared to the pre-mapping genome. The Hi-C anchoring rate in this study is 97.9%. The sizes of chromosome ranged from 32.7 to 57.7 Mb, with an average chromosome length of 43.1 Mb (Fig. 1).

To assess the quality of the assembled chromosomes, BGI short-reads were mapped to the assembled genome using BWA (v0.7.8)[24] to assess coverage rate (97.4%) and mapping rate (97.6%) and HiFi reads were mapped to the assembled genome using minimap2 (v2.21)[25] with coverage of 98.4% and mapping rate of 100.0%. At the whole-genome level, Benchmarking Universal Single-Copy Orthologues (BUSCO) evaluation revealed that 96.2% of the 3640 gene sets were complete BUSCO genes (Table 3), implying that our assembly was complete.

Repeats prediction. In this study, the repetitive sequences, including tandem repeats and TEs, were searched. First, we used Tandem Repeats Finder (TRF, v4.09.1)[26] to annotate the tandem repeats using the following parameters: 2 7 7 80 10 50 2000. Then TEs were identified at both the DNA and protein levels using a combination of *de novo* and homology-based approaches. At the DNA level, LTR_FINDER (v1.0.7)[27] was first used to identify LTR-RTs and RepeatModeler (v2.0.1)[28] was utilised to construct a *de novo* repeat library, which comprised a repeat consensus database with classification information. We employed RepeatMasker (v4.1.2)[29] to search for similar TEs in the known Repbase TE library[30,31] and *de novo* repeat library. At the protein level, RepeatProteinMask within the RepeatMasker package was used to search against the TE protein database using a WU-BLASTX engine.

A total of 731.0 Mb of repeats in the *C. guichenoti* genome were identified, accounting for 66.46% of the genome, according to our analyses. Of these, transposable elements totaled 699.6 Mb, representing 63.40% of the genome (Table 4).

Annotation of non-coding RNA genes. We used tRNAscan-SE (v2.0.9) algorithms[32] with default parameters to identify the genes associated with tRNA, which is an adaptor molecule composed of RNA used in biology to bridge the three-letter genetic code in messenger RNA (mRNA) with the twenty-letter code of amino acids in proteins. RNAmmer (v1.2)[33] was used to predict rRNA sequences. snoRNAs are a class of small RNA molecules that guide chemical modifications of other RNAs, mainly ribosomal RNAs, transfer RNAs and small nuclear RNAs. MiRNAs and snRNAs were identified by Infernal (v1.1.2)[34] software against the Rfam (v14.6) database[35] with default parameters (Table 5).

Genes prediction and annotation. Protein-coding genes were identified based on two different strategies: ab initio gene prediction and homology-based gene prediction. Prior to denovo gene prediction, the assembled *C. guichenoti* genome was hard masked using RepeatMasker. For ab initio gene prediction, Augustus (v. 3.3.3)[36–38] and Genscan(v1.0)[39]. Models used for each gene predictor were trained from a set of high-quality proteins generated from the Exonerate dataset below. We used Exonerate (v2.2.0)[40] to conduct homology-based gene prediction. Finally, Maker (v3.00)[41] was used to integrate the prediction results of the three methods to predict genes models. The output included a set of consistent and non-overlapping sequence assemblies, which were used to describe the gene structures. In summary, a total of 22,506 protein-coding genes were predicted with average gene length of 20,588 bp, coding sequence of 1,659 bp, exon length of 180 bp and intron length of 2,293 bp (Table 6).

Gene functions were inferred according to the best match of the alignments to the National Center for Biotechnology Information (NCBI) Non-Redundant (NR), Kyoto Encyclopedia of Genes and Genomes (KEGG) database[42], Gene Ontology (GO)[43], TrEMBL[44] and Swiss-Prot[44] protein databases using Diamond BLASTP

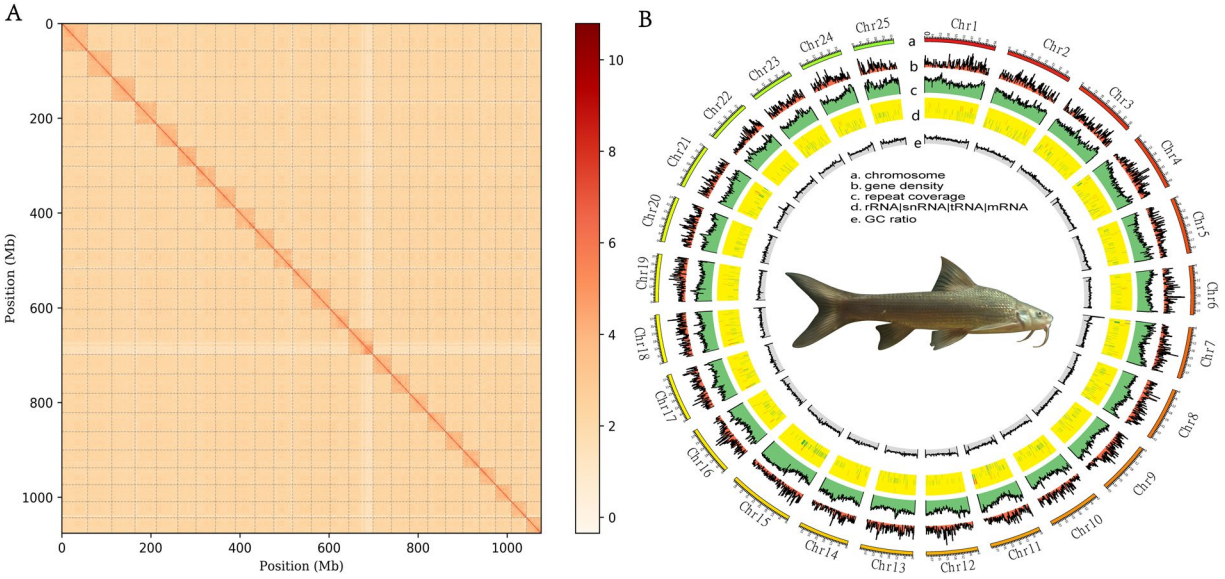

Fig. 1 Characteristics of *C. guichenoti* genome. (**A**) Hi-C intra-chromosomal contact map of C. guichenoti genome assembly. (**B**) Circos plot of the *C. guichenoti* genome assembly. a. chromosome; b. gene density; c. repeat coverage; d. rRNA/SnRNA /tRNA/mRNA; e. GC ratio.

	Assembly No.	Assembly ratio (%)	Annotation No.	Annotation ratio (%)
Complete BUSCOs	3500	96.2	3208	88.2
Complete and single-copy BUSCOs	3447	94.7	3115	85.6
Complete and duplicated BUSCOs	53	1.5	93	2.6
Fragmented BUSCOs	14	0.40	126	3.5
Missing BUSCOs	126	3.5	306	8.4
Total BUSCO groups searched	3640	100	3640	100

Table 3. Assessment of the gene coverage rate in *C. guichenoti* genome using BUSCO.

Type		DNA	LINE	SINE	LTR	Other	Unknown	Total TE
Length (bp)	RepeatMasker TEs	280,528,430	37,928,001	4,336,031	60,195,745	10,903	4,703,982	37,291,4845
	RepeatProteinMask TEs	73,537,466	24,048,789	—	32,770,033	—	—	130,315,393
	De novo	414,574,286	66,984,150	1,908,285	94,015,792	—	28,373,680	567,559,130
	Combined TEs	529,126,692	85,508,852	5,693,806	123,673,751	10,903	32,810,035	699,611,493
% in genome	RepeatMasker TEs	25.5	3.45	0.39	5.47	0	0.43	33.9
	RepeatProteinMask TEs	7	2	—	3	—	—	12
	De novo	38	6	0	9	—	3	52
	Combined TEs	48	8	1	11	—	3	64

Table 4. The proportion of TE classifications in the *C. guichenoti* genome.

(v2.0.7)[45] with an E-value threshold of 1E-5. The protein domains were annotated using InterProScan (v5.50–84.0)[46] based on InterPro[47] protein databases. In total, about 99.6% (22,405) predicted protein-coding genes in *C. guichenoti* were successfully annotated (Table 7).

Evolutionary analysis and divergence time prediction. The protein data of some representative fish species, including *Oryzias latipes*[48], *Misgurnus anguillicaudatus*[49], *Triplophysa tibetana*[50], *Triplophysa dalaica*[51], *Myxocyprinus asiaticus*[52], *Danio rerio*[53], *Ctenopharyngodon idella*[54], *Megalobrama amblycephala*[55], *Rhinogobio ventralis*[56], *Coreius heterodon* (unpublished data from the authors), *Paracanthobrama guichenoti*[57], *Labeo rohita*[58], *Puntigrus tetrazona*[59], *Onychostoma macrolepis*[60], *Cyprinus carpio*[61], *Carassius auratus*[62], and *Carassius gibelio*[63] were retrieved from the NCBI database and used for gene family clustering. By software OrthoFinder (v2.5.4)[64], all protein sequences were pooled and clustered into different kinds of homologs.

Type		Copy	Average length (bp)	Total length (bp)	% of genome
miRNA		820	87.12073	71439	0.006494
tRNA		9185	76.44007	702102	0.063825
rRNA	rRNA	18837	149.1009	2808613	0.255317
	18S	100	1877.51	187751	0.017067
	28S	104	4598.817	478277	0.043478
	5S	18633	114.9887	2142585	0.194772
snRNA	snRNA	1283	156.0935	200268	0.018205
	CD-box	160	176.65	28264	0.002569
	HACA-box	55	158.5818	8722	0.000793
	splicing	1051	152.9553	160756	0.014614

Table 5. Number of the annotated non-coding RNA in the *C. guichenoti*.

Method	Gene set	Number	Average length (bp)				Average exon per gene
			gene	CDS	exon	intron	
De novo	Augustus	27280	18755.48	1413.97	180.03	2530.14	7.85
	Genscan	24218	31198.51	1631.22	186.68	3821.13	8.74
Homolog	*Danio rerio*	49658	15494.77	1177.86	211.52	3133.82	5.57
	Labeo rohita	50050	13770.15	1090.13	198.53	2823.39	5.49
	Paracanthobrama guichenoti	67827	8704.73	745.68	179.79	2528.59	4.15
	Puntius tetrazona	48688	10745.2	900.83	187.25	2583.24	4.81
Maker		22506	20588.33	1658.55	179.55	2292.64	9.26

Table 6. Gene annotation of *C. guichenoti* genome via three methods. Note: data sources for other species see Supplementary Table S1.

Database	Number	Percent (%)
InterPro	19262	85.6
GO	17976	79.9
KEGG_ALL	22213	98.7
KEGG_KO	15491	68.8
Swissprot	20403	90.7
TrEMBL	22045	97.9
NR	22378	99.4
Annotated	22405	99.6

Table 7. Protein-coding gene prediction for *C. guichenoti* genome.

According to the OrthoFinder gene family clustering analysis, a total of 18,249 gene families consisting of 22,506 genes were revealed in the *C. guichenoti* genome (Fig. 2A). Among the 18 analyzed species, 308 strictly single-copy ortholog gene sets were identified by OrthoFinder clustering. Analysis of Orthologs showed that *R. ventralis*, *C. heterodon*, *P. guichenoti* and *C. guichenoti* had 14,440 gene families in common and 162 gene families specific to *C. guichenoti* (Fig. 2B). Single-copy ortholog gene families of each species were performed multiple sequence alignment using Muscle[65], while RaxML[66] was used to construct a phylogenetic tree by the Maximum Likelihood method (Fig. 2C).

The expansion and contraction of gene family were further analyzed with CAFE. The results showed that 329 gene families were expanded while 1272 gene families were contracted in *C. guichenoti* (Fig. 2D). In comparison, the close relative *C. heterodon* revealed 252 expanded gene families and 692 losing gene families, *C. guichenoti* obtained more gene families. These expanded genes exhibit a variety of functions, such as binding, catalytic activity and molecular transducer activity. Analysis of evolutionary relationships revealed that *C. guichenoti* was closely related to *C. heterodon* and they phylogenetically diverged from common ancestor ~20.7 million years ago (Mya), following the separation of Cyprinidae at 28.3 Mya (Fig. 2E).

Data Records

Sequencing reads for *C. guichenoti* are available on the NCBI Sequence Read Archive (SRA) https://identifiers.org/ncbi/insdc.sra: SRR28210515 for Hi-C data; SRR28210513 for T7 data; SRR28210516 for PacBio data. Genome assembly for *C. guichenoti* on *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_039654105.1[67].

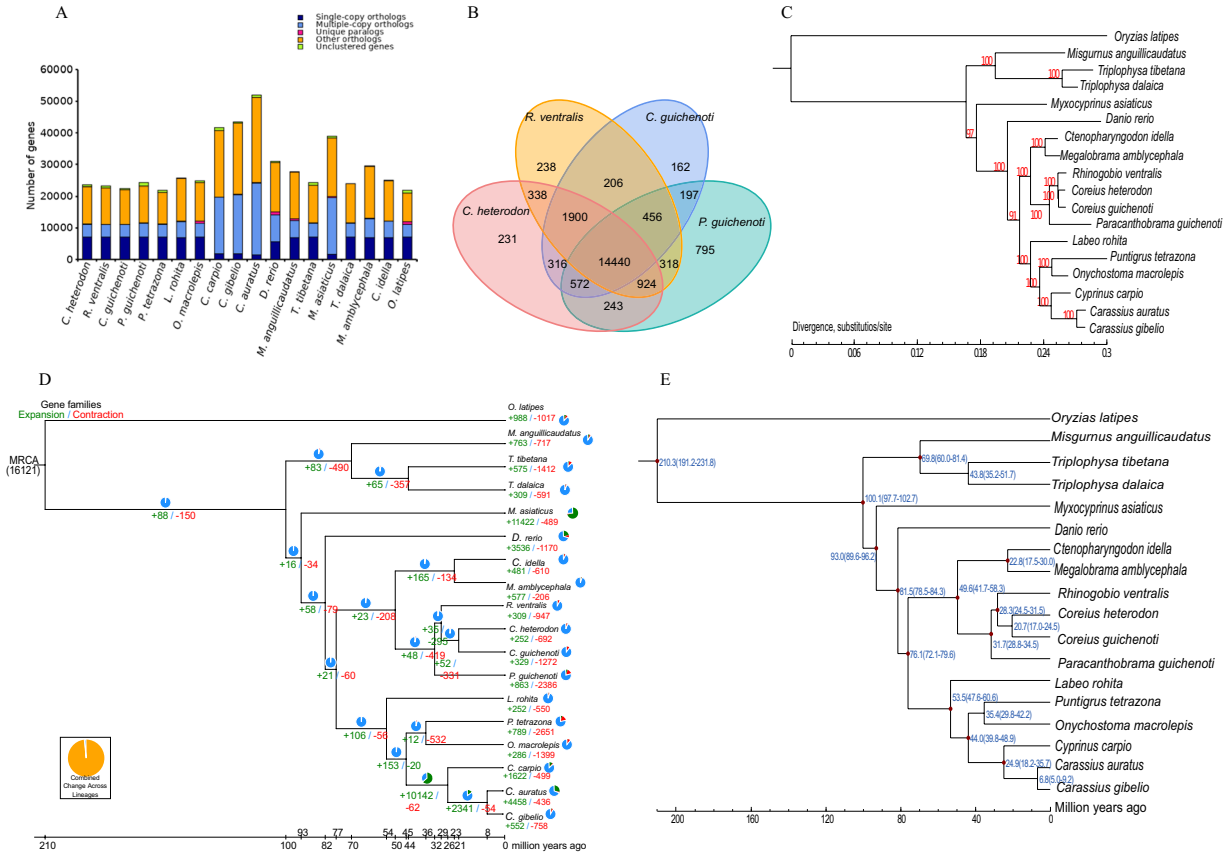

Fig. 2 Evolutionary analysis and divergence time prediction of *C. guichenoti*. (**A**) Clusters of orthologous and paralogous gene families in *C. guichenoti* and 17 other sequenced fish genomes. (**B**) Venn diagram representing the distribution of shared gene families among *C. guichenoti* and three other fishes (*R. ventralis*, *C. heterodon* and *P. guichenoti*). (**C**) Phylogenetic tree of *C. guichenoti* and 17 other fish species. (**D**) Estimates of gene family expansions and contractions. The green and red numbers indicate expanded and contracted gene families, respectively. Conserved gene families are indicated in blue in the pi charts. MRCA represents the most recent common ancestor. (**E**) Estimates of divergence time among *C. guichenoti* and 17 other fish species.

Technical Validation

To precisely measure the quality and quantity of genomic DNA, NanoDrop 2000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA), and Qubit dsDNA HS Assay Kit on a Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) were used and the DNA integrity was assessed through electrophoresis on a 0.8% agarose gel. BUSCO (v5.2.2) was used to assess the completeness of assembled genome, and 95.7% completeness of the BUSCOs (Actinopterygii_odb10) was identified.

Code availability

No custom code was used during this study for the curation and/or validation of the dataset. All instructions and sequences of actions carried out during data processing were performed in accordance with the guidelines and procedures outlined in the manual and protocols of the relevant bioinformatics software.

Received: 16 May 2024; Accepted: 3 January 2025;
Published online: 15 January 2025

References

1. Liu, L. H., Wu, G. X. & Wang, Z. L. Reproduction ecology of *Coreius eterodon* and *Coreius guichenoti* in the mainstream of the Changjiang River after the constructions of Gezhouba Dam. *Acta Hydrobiol. Sin.* **14**, 205–215 (1990). (In Chinese)
2. Liao, X. *et al.* Polymorphic microsatellites in largemouth bronze gudgeon (*Coreius guichenoti*) developed from repeat-enriched libraries and cross-species amplifications. *Mol. Ecol. Notes.* **7**, 1104–1107 (2007).
3. Zhao, J. H. *et al.* Effects of temperature reduction and MS-222 on water quality and blood biochemistry in simulated transport experiment of largemouth bronze gudgeon, *Coreius guichenoti*. *J. World Aquacult. Soc.* **45**, 493–507 (2014).
4. Li, X. M. *et al.* The influence of weight and gender on intestinal bacterial community of wild largemouth bronze gudgeon (*Coreius guichenoti*, 1874). *BMC Microbiol.* **16**, 191 (2016).
5. Liu, G. X. *et al.* Cryopreservation and its effects on spermatozoa quality of *Coreius guichenoti*. *J. Fish. Sci. China* **27**, 44–52 (2020). (In Chinese)
6. Li, X. L. *et al.* Effects of water temperature on growth performance, digestive enzymes activities, and serum indices of juvenile *Coreius guichenoti*. *J. Therm. Biol.* **115**, 103595 (2023).

7. Sun, Z. Y. *et al*. Progress and prospect on the artificial domestication and reproduction of *Coreius guichenoti*. *Freshw. Fish.* **50**, 107–112 (2020). (In Chinese).
8. Zhang, X. F. *et al*. Preliminary studies on ovary development and oogenesis in *Coreius guichenoti*. *Journal of Southwest Agriculture University (Natural Science)* **27**, 892–897 (2005). (In Chinese)
9. Chen, G. L., Zhang, X. Biological characteristics and disease control techniques of *Coreius guichenoti*. *Jilin Agriculture*, (**4**), 289 (2011). (In Chinese)
10. Zhu, T. B. *et al*. Domestication comparison of *Coreius guichenoti* in the closed recirculating aquaculture system and the boat -net trunk. *Freshwater Fisheries* **45**, 97–101(2015). (In Chinese)
11. Sun, B. Z. *et al*. Studies on the oxygen consumption rate and asphyxiant point of *Megalobrama pellegrini* and *Coreius guichenoti*. *Acta Hydrobiol. Sin*. **34**, 88–93 (2010). (In Chinese)
12. Chen, L. *et al*. Chromosome-level assembly and gene annotation of *Decapterus maruadsi* genome using Nanopore and Hi-C technologies. *Sci Data* **11**, 69 (2024).
13. Wang, F. *et al*. Chromosome-level assembly of *Gymnocypris eckloni* genome. *Sci. Data* **9**, 464 (2022).
14. Qu, M. *et al*. Seadragon genome analysis provides insights into its phenotype and sex determination locus. *Sci. Adv.* **7**, eabg5196 (2021).
15. Ma, X. *et al*. Chromosome-level genome assembly of the freshwater mussel *Sinosolenaia oleivora* (Heude, 1877). *Sci. Data* **11**, 606 (2024).
16. Jia, Y. *et al*. A Chromosome-level genome assembly of the dark sleeper *Odontobutis potamophila*. *Genome Biol Evol.* **13**, evaa271 (2021).
17. Niu, J. *et al*. Chromosomal-scale genome assembly of the near-extinction big-head schizothorcin (*Aspiorhynchus laticeps*). *Sci. Data* **9**, 556 (2022).
18. Rao, S. S. *et al*. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
19. Chen, Y. *et al*. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput quencing data. *Gigascience* **7**, 1–6 (2018).
20. Liu, B. *et al*. Estimation of genomic characteristics by analyzing k-mer frequency in *de novo* genome project. arXiv.org arXiv: 1308.2012. (2013)
21. Driguez, P. *et al*. LeafGo: Leaf to Genome, a quick workflow to produce high-quality *de novo* plant genomes using long-read sequencing technology. *Genome Biol.* **22**, 1–18 (2021).
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 *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
24. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv e-prints* http://arxiv.org/abs/1303.3997v2 (2013).
25. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
26. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
27. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, 265–268 (2007).
28. Flynn, J. M. *et al*. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2007).
29. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* Chapter 4, Unit 4.10 (2009).
30. Jurka, J. Repbase Update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000).
31. Jurka, J. *et al*. Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
32. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
33. Lagesen, K. *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
34. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335–1337 (2009).
35. Kalvari, I. *et al*. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* **49**, 192–200 (2021).
36. Stanke, M. *et al*. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, 309–312 (2004).
37. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, 465–467 (2005).
38. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, 435–439 (2006).
39. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
40. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinf.* **6**, 31 (2005).
41. Cantarel, B. L. *et al*. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
42. Kanehisa, M. *et al*. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, 109–114 (2012).
43. Ashburner, M. *et al*. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
44. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
45. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
46. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
47. Mitchell, A. *et al*. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* **43**, 213–221 (2015).
48. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_002234675.1 (2017).
49. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_027580225.1 (2023).
50. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_008369825.1 (2019).
51. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_015846415.1 (2020).
52. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_019703515.2 (2022).
53. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000002035.6 (2017).
54. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_019924925.1 (2021).
55. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_018812025.1 (2021).
56. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_039654625.1 (2024).
57. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_018749465.1 (2021).
58. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_022985175.1 (2022).
59. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_018831695.1 (2021).
60. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_012432095.1 (2020).
61. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_018340385.1 (2021).
62. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_003368295.1 (2018).

63. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_023724105.1 (2022).
64. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
65. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
66. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
67. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_039654105.1 (2023).

Acknowledgements

This work was funded by China Agriculture Research System of MOF and MARA (No. CARS-46), the Central Public-Interest Scientific Institution Basal Research Fund, CAFS (No. 2023TD61), and the Key Research and Development Program of Jiangxi Province (No. 20223BBF61010). We are grateful to Wuhan Frasergen Bioinformatics Co., Ltd for assisting in sequencing and bioinformatics analysis.

Author contributions

Xuemei Li, X.W., D.Y. and T.Z. conceived the study. Y.Z. and Xiaoli Li collected the samples. Z.M. and M.X. extracted the genomic DNA and conducted sequencing. Xuemei Li, T.Z. and N.W., performed bioinformatics analysis. Xuemei Li and T.Z. wrote the manuscript, all authors read and approved the final manuscript, T.Z. is the lead contact for this paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04416-y.

Correspondence and requests for materials should be addressed to T.Z.

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.