Scientific Data

https://doi.org/10.1038/s41597-026-07960-3

Article in Press

A near telomere to telomere genome assembly of red crucian carp (*Carassius auratus* red var.)

Received: 5 January 2026

Accepted: 21 July 2026

Published online: 28 July 2026

Cite this article as: Liu, Z., Wang, C., Zhang, K. *et al.* A near telomere to telomere genome assembly of red crucian carp (*Carassius auratus* red var.). *Sci Data* (2026). https://doi.org/10.1038/s41597-026-07960-3

Zhengkun Liu, Chongqing Wang, Kun Zhang, Yuhan Yang, Zeyang Li, Qian Xiao, Xu Huang, Xiaowei Xu, Xidan Xu, Yue Zhou, Jinhai Bai, Yuchen Jiang, Wanjing Peng, Ling Xiong, Enkui Hu, Siyang Li & Qinbo Qin

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.

ARTICLE IN PRESS

© The Author(s) 2026. **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/.

ARTICLE IN PRESS

A near telomere to telomere genome assembly of red crucian carp (*Carassius auratus* red var.)

Zhengkun Liu[1], Chongqing Wang[1], Kun Zhang[1], Yuhan Yang[1], Zeyang Li[1], Qian Xiao[1]
Xu Huang[1], Xiaowei Xu[1], Xidan Xu[1], Yue Zhou[1], Jinhai Bai[1], Yuchen Jiang[1], Wanjing Peng[1], Ling Xiong[1,]
Enkui Hu[1], Siyang Li[1], Qinbo Qin[1,2]*

[1] Engineering Research Center of Polyploid Fish Reproduction and Breeding of the State Education, Ministry, College of Life Sciences, Hunan Normal University, Changsha, 410081, Hunan, People's Republic of China.

[2] Yuelushan Laboratory, Changsha 410128, China.

*Correspondence:

Qinbo, Qin, Engineering Research Center of Polyploid Fish Reproduction and Breeding of the State Education, Ministry, College of Life Sciences, Hunan Normal University, Changsha, 410081, Hunan, People's Republic of China. Email: qqb@hunnu.edu.cn. Telephone number, 186-7074-7358.

Authors E-mail:

Zhengkun Liu, Email: 1530697267@qq.com.

Chongqing Wang, Email: 1436236021@qq.com.

Kun Zhang, Email: 1975322578@qq.com.

Yuhan Yang, Email: 2879909647@qq.com.

Zeyang Li, Email: 1095921866@qq.com.

Qian Xiao, Email: 3092823443@qq.com.

Xu Huang, Email: 2696177754@qq.com.

Xidan Xu, Email: 1912533911@qq.com.

Xiaowei Xu, Email: 707792491@qq.com.

ARTICLE IN PRESS

ARTICLE IN PRESS

Yue Zhou, Email: 809745303@qq.com.

Jinhai Bai, Email: b13807253562@outlook.com.

Yuchen Jiang, Email: yc990213@163.com.

Wanjing Peng, Email: 1411772587@qq.com.

Ling Xiong, Email: xiongling@hunnu.edu.cn.

Enkui Hu, Email: 1650403773@qq.com.

Siyang Li, Email: 2439728223@qq.com.

Qinbo Qin, Email: qqb@hunnu.edu.cn.

ARTICLE IN PRESS

ARTICLE IN PRESS

Abstract

Red crucian carp, a variant of *Carassius auratus*, has a long history of cultivation as both a food and an ornamental fish. Here, we integrated MGI short-reads, PacBio HiFi long-reads, ONT ultra-long reads, and Hi-C sequencing data to assemble a nearly telomere-to-telomere (T2T) genome for this species, containing only four gaps. This assembly, which consists of a primary haplotype containing 50 chromosomes, covers 99.88% of the sequence, with a total length of 1.55 Gb, and reveals 93 telomeres and 50 centromeres. Key quality metrics for the assembly include a contig N50 of 29.92 Mb (contiguity), a BUSCO score of 99.40% (completeness), and a consensus quality value of 64.44 (correctness), collectively reflecting a high-quality genome. Annotation identified 46,061 protein-coding genes and 679.10 Mb (43.72% of the assembly) of repetitive sequences. Furthermore, 22,454 tRNAs, 42,628 rRNAs, 3,309 miRNAs, and 3,380 snRNAs were annotated. This represents the first near-T2T genome assembly for any species within the genus *Carassius*, offering a valuable genomic resource for investigating the biological traits of the red crucian carp and advancing research in evolutionary genomics.

ARTICLE IN PRESS

ARTICLE IN PRESS

Background & Summary

Crucian carp (*Carassius auratus*), widely distributed across Eurasia and also known as goldfish in its ornamental form[1], is an omnivorous freshwater species of significant value for both aesthetic and economic purposes[2, 3]. It is widely accepted that both crucian carp and their close relative, the common carp, underwent the same allotetraploidy event (involving hybridisation between the two species followed by a whole-genome duplication)[4-6]. This process gave rise to two subgenomes and substantially increased genomic complexity[7]. In contrast to the common carp, certain crucian carp populations have undergone additional whole-genome duplication events[8, 9]. Notably, the Dongting Lake system is home to crucian carp complexes that coexist and exhibit different ploidy levels[10]. This provides an ideal model for studying the key evolutionary mechanisms involved in the genome duplication process of polyploid vertebrates, including gene variation[11, 12], subgenomic evolution[7], expression differentiation[13, 14], and chromosomal rearrangements[15]. These characteristics also reveal the genomic tractability of crucian carp, offering potential for engineering polyploid genomes with desirable traits[16].

Red crucian carp (*Carassius auratus* red var.), one of the oldest varieties of goldfish, has been domesticated in China for a period exceeding one millennium[17, 18]. A key significance of this species is its high phenotypic variability, derived from extensive artificial selection[2, 19]. Furthermore, it serves as a significant model organism for lower vertebrates in research contexts pertaining to physiology[20], neurobiology[21], disease response[22], and environmental adaptation[23]. Despite its significance in both basic and applied research, and the publication of a draft scaffold-level genome assembly in 2020, the current assembly contains an excessive number of gaps (contig N50 = 1.11 Mb) and remains limited in continuity, completeness, and chromosomal mounting accuracy[7].

The complexity of the red crucian carp genome, such as its dual subgenomes, poses considerable challenges to achieving high-quality assembly. Despite the recent advancements in sequencing and assembly algorithms, which have facilitated the construction of telomere-to-telomere (T2T) genomes, as evidenced by the multiple published T2T koi carp genomes[4, 24], such comprehensive genomic resources remain scarce for red crucian carp and the broader genus *Carassius*. In this study, we performed de novo sequencing and assembly of the red crucian carp using PacBio HiFi long reads, Oxford Nanopore ultra-long reads, and Hi-C sequencing data, ultimately yielding a high-quality near-T2T reference genome. This assembly has been demonstrated to significantly enhance the completeness and utility of red crucian carp genetic data. This will facilitate the elucidation of the genetic underpinnings that govern the morphological and color diversity observed among its varieties. Moreover, it will furnish a pivotal resource for the study of the rediploidization process in allopolyploid vertebrates.

Methods

Ethics statement All animal experiments were conducted with the approval of the Animal Care Committee of Hunan Normal University (Approval No. 380; 2025) and in accordance with ARRIVE guidelines (https://arriveguidelines.org).

ARTICLE IN PRESS

Sample collection and genome sequencing. A one-year-old red crucian carp was obtained for sampling from the Fisheries Research Institute of Hunan Province, Changsha, China. Prior to dissection, red crucian carp were euthanized with 2-phenoxyethanol (Sigma) at the Polyploid Fish Protection Station of Hunan Normal University. Muscle and liver tissue were collected for the purpose of whole-genome DNA extraction. The former included MGI short-read, PacBio HiFi long-read, and ONT (Oxford Nanopore Technologies) ultra-long-read sequencing, while the latter employed Hi-C sequencing. Concurrently, a total of twelve tissue samples were used for transcriptome sequencing, including fin rays, gills, ovaries, liver, spleen, head kidney, trunk kidney, heart, skin, brain, pituitary gland, and eyes. Tissue samples were briefly rinsed with DEPC-treated water (Sangon Biotech, Shanghai, China) to remove residual blood and adipose tissue, then flash-frozen in liquid nitrogen and stored at -80°C.

The construction of DNA libraries was undertaken from genomic DNA using the MGIEasy Universal DNA Library Prep Kit, and the subsequent sequencing was conducted on the DNBSEQ-T7RS platform. This process yielded 152.56 Gb of raw data. After the filtration and deduplication of the data using Fastp (v0.23.41, parameter: -l 50)[25], 151.73 Gb of clean data were obtained for sequence assembly (Table 1).

For the PacBio HiFi sequencing, genomic DNA was sheared and used to construct SMRTbell libraries (Pacific Biosciences SMRTbell® prep kit 3.0). Libraries were size-selected for ≥20 kb fragments, prepared with the Revio™ polymerase kit, and sequenced on a PacBio Revio platform. This process yielded 151.51 Gb of HiFi reads with an N50 of 20,469 bp (Table 1).

The ultra-long read library was constructed using the Oxford Nanopore SQK-LSK110 Ligation Kit. Following purification, the library was sequenced on a PromethION instrument (Oxford Nanopore Technologies, Oxford, UK), yielding 44.87 Gb of high-quality data. The resulting reads had a mean length of 73,004 bp and an N50 length of 100,000 bp (Table 1).

Hi-C libraries were prepared using the DpnII restriction enzyme. In brief, Hi-C libraries were prepared from cross-linked cells/tissues through a series of steps, including nuclear extraction, MboI chromatin digestion, adjacent ligation, and biotin-labeled fragment enrichment. The resulting library was sequenced on the DNBSEQ-T7RS platform, generating paired-end 150 bp reads. As previously stated, the raw data for Hi-C was filtered using the Fastp (v0.23.2, parameter: -l 50)[25]. A total of 170.12 Gb of clean data was obtained (Table 1).

Total RNA extraction was performed using TRIzol (TIANGEN). Subsequent to the pooling of equal amounts of total RNA extracted from different tissues, the pooled RNA is utilised for the preparation of iso-seq transcriptomic libraries. Briefly, Poly(A)+ RNA was enriched, fragmented, and used to construct sequencing libraries via the following process: cDNA synthesis, end repair, adapter ligation, PCR amplification, and circularization. The final Iso-Seq cDNA libraries were subjected to sequencing on the DNBSEQ-T7RS platform. A total of 6.14 Gb of cleaned transcriptome reads were consequently generated for the purpose of protein-coding gene annotation (Table 1).

ARTICLE IN PRESS

Table 1. Summary of sequencing data generated for the red crucian carp genome assembly

Library type	Raw Data (Gb)	Clean Data (Gb)	Mean length/Read Length (bp)	N50/ Read Length(bp)	GC content (%)
MGI	152.56	151.72	150	150	41.59
PacBio HiFi	/	151.51	20,502	20,469	37.44
ONT ultra-long	/	44.87	73,004	100,000	37.60
Hi-C	171.35	170.12	150	150	39.16
RNA-seq	6.20	6.14	150	150	49.58

Polyploidy evaluation and genome assembly

Estimation of haplotype-ploidy Based on MGI data, the haplotype structures were extracted from heterozygous k-mer pairs using smudgeplots (v0.2.3dev, parameter: -k21 -m100 -ci1 -cs10000)[26] to estimate the ploidy levels of the red crucian carp. The results indicate that 75% of heterozygous k-mer pairs support the AB haplotype, while 23% and 2% of heterozygous k-mer pairs support the AAABB and AAAAB haplotypes, respectively (Fig. 1). This pattern of expression aligns with the known haplotype characteristics of allotetraploid[6, 27].

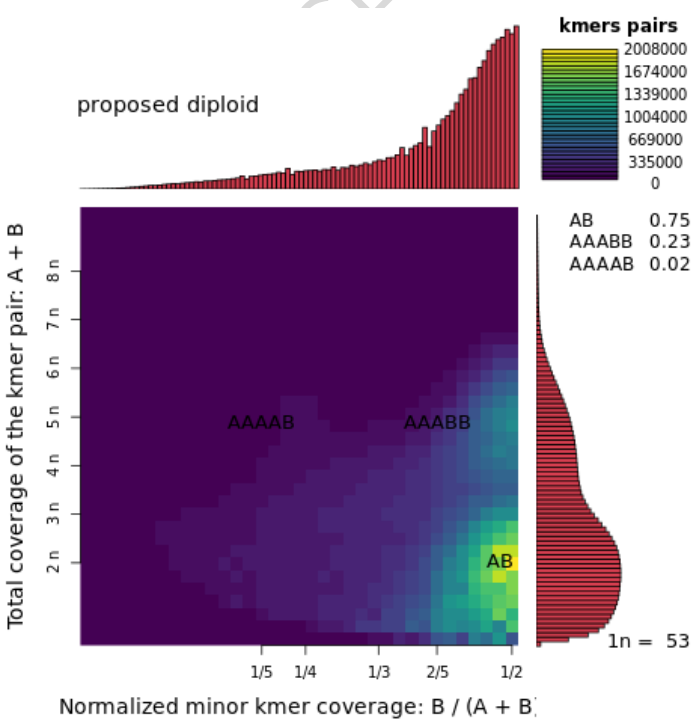


Fig. 1 Smudgeplot analysis of K-mer distribution for ploidy assessment in red crucian carp.

Chromosome-level genome assembly Initially, Hifiiasm (v0.25.0-r726, default parameters)[28] and purge_haplotigs (v1.1.3, default parameters)[29] were utilised to assemble HiFi reads, ONT reads, and Hi-

ARTICLE IN PRESS

C data into contigs and remove redundancy. The assembled contig-level genome size was 1.56 Gb, comprising 119 contigs with an N50 size of 29.93 Mb. Clean Hi-C reads were aligned to contigs using HiCUP (v0.7.2, parameter: --NM 3)[30]. The resulting valid pairs were processed with Juicer (v1.6, default parameters)[31] and the 3D-DNA pipeline (v180992, parameter: -r 0)[32] to map contigs into chromosomes. Visual error correction was performed in JuiceBox (v1.11.08, default parameters)[33], guided by the expectation that cis-interactions (within the same chromosome) occur more frequently and with greater intensity than trans-interactions (between chromosomes), and that interaction strength decays with increasing genomic distance (Fig. 2a). The assembled contigs were ultimately anchored to 50 chromosomes, achieving an anchoring rate of 99.88%.

T2T genome assembly ONT ultra-long reads (>100 kb) were aligned to the chromosomes using minimap2 (v2-2.24, parameter: ont:-ax map-ont ccs:-ax map-hifi)[34]. Reads that aligned exactly once within the 100 bp terminal regions of the chromosomes were collected. These reads were subsequently reassembled into consensus sequences using medaka consensus (v1.7.2, parameter: -m r941_prom_fast_g507, https://github.com/nanoporetech/medaka). Finally, the resulting consensus sequences were aligned back to both ends of each chromosome via BLASTN (v2.11.0+, parameter: -outfmt 7)[35] to extend the telomeric regions. We employed TGS-GapCloser (v1.2.0, parameter: --min_nread 10)[36] to perform gap filling and contig extension, utilizing the coverage relationships between the ONT ultra-long reads and the pre-assembled contigs. Subsequently, the resulting genome assembly was polished using Pilon v1.23[37] with short-read sequencing data to correct potential errors. The final assembly of the red crucian carp genome was 1.55 Gb, spanning 50 chromosomes with only four remaining gaps and achieving a contig N50 of 29.92 Mb (Fig. 2a-b). Compared to the previously published scaffold-level assembly of the red crucian carp (contig N50 = 1.11 Mb)[7], the new assembly exhibits significant improvements in both completeness and contiguity (Table 2).

Table 2. Comparison of assembly statistics for the red crucian carp genome.

Statistics	This study	GWHAAIA00000000[7]
Total length (Gb)	1.55	1.64
Number of chromosomes	50	/
Number of scaffolds	63	5,474
N50 length (scaffold, bp)	29,937,171	2,933,731

ARTICLE IN PRESS

N90 length (scaffold, bp)	25,796,168	96,045
Number of contig	70	6,141
N50 length (contig, bp)	29,923,908	1,109,331
N90 length (contig, bp)	25,796,168	88,816
GC Content (%)	37.60	35.40
Maximum scaffold length (bp)	53,167,710	11,063,794
Mean scaffold length (bp)	24,656,872	299,489

Telomere and centromere identification The genome-wide identification of telomeric repeat motifs (CCCTAA/TTAGGG) was performed, and occurrences of four or more consecutive repeats were quantified within the terminal 100-kb regions at both ends of each chromosome. Candidate centromeric regions were identified by detecting tandem repeats and satellite DNA using Tandem Repeats Finder (v4.09, parameter: 2 7 7 80 10 50 2000 -d -h)[38] and RepeatMasker (v4.0.9, parameter: -nolow -no_is -norna, http://www.repeatmasker.org), followed by validation with the quarTeT (v1.1.36; parameter: '-m 1')[39]. In conclusion, a total of 93 telomeres and 50 centromeres were screened on the extremities of 50 chromosomes (Fig. 3, Table 3).

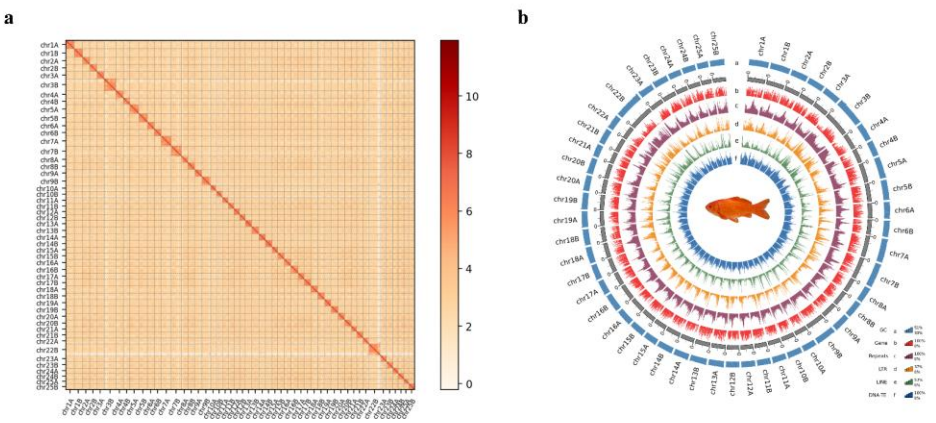


Fig. 2 The near T2T genome assembly of red crucian carp. (**a**) Whole-genome Hi-C interaction map, where darker colours indicate stronger interactions (Resolution 500 kb). (**b**) Genomic Feature Map of the Red Crucian Carp T2T genome (Window Size 500 kb).

Table 3. Summary of chromosomal assembly features, including length, contig and gap counts, telomere repeat units, and centromere positions.

Chromosome	Length (bp)	Number of contigs	Number of gaps	Telomere		Centromere
				Number of start	Number of end	

ARTICLE IN PRESS

				repeat unit	repeat unit	Start positions	End positions
chr1A	37,024,658	1	0	345	5	19,500,001	20,200,000
chr1B	37,461,668	1	0	324	527	20,250,001	20,800,000
chr2A	31,640,104	1	0	903	351	22,300,001	22,750,000
chr2B	31,429,280	1	0	242	0	21,550,001	21,850,000
chr3A	32,363,325	1	0	0	244	2,950,001	3,350,000
chr3B	53,167,710	1	0	286	518	18,554,031	19,055,418
chr4A	32,738,954	1	0	55	254	16,900,001	17,300,000
chr4B	29,050,366	1	0	456	610	11,650,001	12,500,000
chr5A	40,121,352	1	0	33	137	30,300,001	31,200,000
chr5B	37,653,559	1	0	115	283	25,700,001	26,250,000
chr6A	31,268,934	1	0	97	563	11,950,001	12,400,000
chr6B	30,433,736	1	0	123	343	11,350,001	11,900,000
chr7A	44,516,968	1	0	66	270	31,450,001	31,700,000
chr7B	45,289,524	1	0	438	450	31,850,001	32,550,000
chr8A	29,076,607	1	0	342	354	3,600,001	4,150,000
chr8B	30,724,334	2	1	127	174	3,300,001	3,900,000
chr9A	30,993,022	1	0	173	187	21,050,001	21,550,000
chr9B	38,127,005	1	0	487	301	33,550,001	34,100,000
chr10A	25,860,356	1	0	119	387	22,250,001	22,900,000
chr10B	27,252,	1	0	326	4	23,700,	24,300,

ARTICLE IN PRESS

ARTICLE IN PRESS

	266					001	000
chr11A	26,266,400	1	0	64	401	20,500,001	20,900,000
chr11B	25,327,274	1	0	17	191	3,900,001	4,350,000
chr12A	25,859,859	1	0	561	337	16,450,001	16,950,000
chr12B	25,325,959	1	0	252	233	15,550,001	16,150,000
chr13A	28,855,095	1	0	332	68	4,950,001	5,450,000
chr13B	31,399,862	1	0	222	310	4,650,001	5,100,000
chr14A	28,317,754	1	0	29	45	5,250,001	5,850,000
chr14B	29,897,911	2	1	0	78	5,500,001	6,150,000
chr15A	25,796,168	1	0	238	214	6,200,001	6,550,000
chr15B	28,068,999	1	0	371	157	7,000,001	7,600,000
chr16A	29,923,281	1	0	118	61	8,800,001	9,350,000
chr16B	32,577,400	1	0	324	208	9,007,969	9,824,885
chr17A	27,100,380	1	0	0	889	5,400,001	5,850,000
chr17B	28,797,187	1	0	0	1085	5,500,001	6,000,000
chr18A	30,635,318	1	0	222	13	17,750,001	18,400,000
chr18B	29,923,908	1	0	323	263	18,600,001	19,300,000
chr19A	29,714,164	2	1	265	0	3,400,001	3,800,000
chr19B	31,179,160	1	0	247	367	3,150,001	3,750,000
chr20A	28,939,850	1	0	228	234	4,950,001	5,350,000
chr20B	29,937,171	1	0	95	236	5,600,001	6,050,000
chr21A	23,401,111	1	0	72	124	16,400,001	16,900,000

ARTICLE IN PRESS

chr21B	28,448,803	1	0	354	326	19,250,001	19,600,000
chr22A	26,264,173	1	0	46	180	20,750,001	21,150,000
chr22B	51,124,777	1	0	313	4	38,900,001	39,300,000
chr23A	27,667,331	1	0	105	551	4,450,001	4,750,000
chr23B	28,313,256	2	1	398	253	3,000,001	3,750,000
chr24A	23,755,085	1	0	130	277	8,550,001	9,000,000
chr24B	25,022,399	1	0	46	347	8,450,001	8,950,000
chr25A	20,902,603	1	0	0	353	17,950,001	18,500,000
chr25B	26,894,553	1	0	195	877	23,050,001	23,650,000

ARTICLE IN PRESS

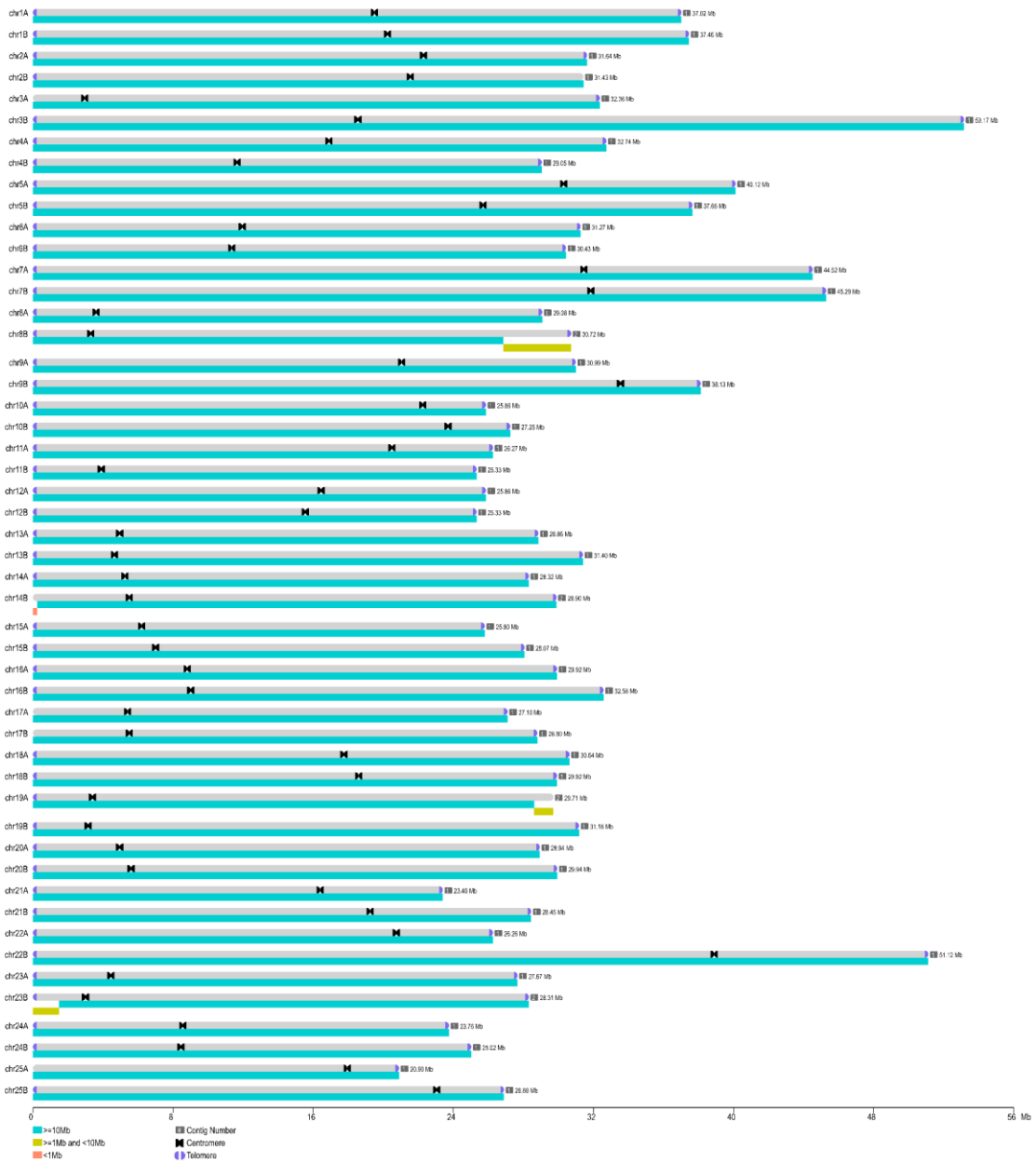


Fig. 3 Chromosome, telomere, and centromere visualization of the red crucian carp near T2T genome

Genome annotation

Repeat elements annotation We performed a comprehensive annotation of repetitive sequences in the red crucian carp genome using complementary approaches. Initially, MISA (v2.1, default parameters)[40] was employed to forecast SSR (Fig. 4). De novo prediction with Tandem Repeats Finder (v4.09, parameter: 2 7 7 80 10 50 2000 -d -h)[38] identified 124.33 Mb of tandem repeats, accounting for 8.00% of the genome. Homology-based searches against the RepBase library (http://www.girinst.org/repbase) using RepeatMasker (v4.0.9, parameter: -nolow -no_is -norna, http://www.repeatmasker.org) and RepeatProteinMask identified 412.12 Mb and 110.51 Mb of repeats at the DNA and protein levels, representing 26.60% and 7.11% of the genome, respectively. Furthermore, a de novo repeat library was constructed using RepeatModeler (v1.0.11, default parameters)[41] and LTR_FINDER _parallel (v1.0.7, default parameters)[42], and subsequent masking with RepeatMasker

ARTICLE IN PRESS

revealed 456.49 Mb of repeats, constituting 29.39% of the assembly. Following the integration and removal of redundancies from these results, a final set of 679.10 Mb of repetitive sequences was obtained, representing 43.72% of the genome. Among these, DNA transposons were the most abundant category, comprising 24.02% of the genomic sequence (Table 4).

Table 4. Statistics on classification results of repetitive sequences in red crucian carp genome.

Type	RepBase TEs		TE Proteins		De novo		Combined TEs	
	Length (bp)	%in Genome	Length (bp)	%in Genome	Length (bp)	%in Genome	Length (bp)	%in Genome
DNA	237,606,338	15.3	19,511,281	1.26	236,674,837	15.24	373,189,367	24.02
LINE	72,256,285	4.65	49,241,697	3.17	61,935,525	3.99	95,750,006	6.16
SINE	8,594,395	0.55	0	0	6,525,852	0.42	14,053,682	0.90
LTR	68,615,546	4.42	41,799,995	2.69	85,035,259	5.47	118,032,692	7.60
Satellite	43,515,696	2.80	0	0	3,707,029	0.24	45,969,422	2.96
Simple_repeat	0	0	0	0	200,270	0.01	200,270	0.01
Other	31,621	0	0	0	0	0	31,621	0
Unknown	3,719,431	0.24	8,319	0	68,781,501	4.43	70,779,506	4.56
Total	413,132,503	26.6	110,506,559	7.11	456,488,195	29.39	679,103,224	43.72

Protein-coding genes structural and functional annotation The structural annotation of protein-coding genes was performed by integrating multiple approaches. (1) The de novo predictions were conducted using Augustus (v3.3.2, default parameters)[43] and Genscan (default parameters)[44]. (2) For homology-based prediction, genomic data from four closely related fish species were downloaded (*Carassius auratus*[45]; *Carassius carassius*[46]; *Danio rerio*[47]; *Cyprinus carpio*[48]), and gene structures were projected using miniprot (v0.11-r234, parameter: -gff-only -O 11 -E 1 -F 23 -C 1 -B 5 -G 200000 -j 1)[49] and Liftoff (v1.6.3, parameter: -a 0.5 -s 0.5)[50]. (3) Predictions based on Iso-seq transcriptomic data were obtained through a pipeline involving HISAT2 (v2.1.0, parameter: --dta)[51] for read alignment, StringTie (v1.3.5, default parameters)[52] for transcript assembly, and TransDecoder (v5.5.0, default parameters, https://github.com/TransDecoder/TransDecoder) for coding region prediction. (4) Prediction was also performed using BUSCO (v5.7.1, parameter: -m protein -f –offline)[53] with a conserved species gene set. (5) The gene sets generated by the above methods were integrated into a non-redundant and more complete set using MAKER2 (v2.31.10, parameter: max_dna_len=3000000 min_contig=10000 pred_flank=500 min_protein=30)[54]. (6) Finally, HiFAP (Wuhan OneMore, https://www.onemore-tech.com) was used to consolidate all results, yielding a final high-quality set of 46,424 protein-coding genes (Table 5). Similar distribution patterns in the length distribution characteristics of gene set elements were revealed by comparisons between red crucian carp and the other four closely related species (Fig. 5).

The functional annotation of the gene set was achieved through the implementation of two approaches. The alignment of protein sequences was conducted against public databases, encompassing Non-Redundant Protein Sequence (20240207)[55], SwissProt (20250423)[56], TrEMBL (20250423)[56], EuKaryotic Orthologous Groups (KOG, 20030301, https://ftp.ncbi.nih.gov/pub/COG/KOG/), AnimalTF (v4.0)[57], InterPro (v5.74-105.0)[58], Gene Ontology (GO, 20250605)[59], Kyoto Encyclopedia of Genes and

ARTICLE IN PRESS

Genomes (KEGG, v105.0, 20230101)[60], and Pfam[61], utilising the DIAMOND (v2.1.11.165, parameter:--evalue 1e-05 --max-hsps 1)[62]. This approach was based on the sequence similarity. Based on domain or motif similarity, conserved sequences, motifs, and domains were identified using InterProScan (v5.74-105.0, parameter: --seqtype p --formats TSV --goterms --pathways -dp)[63]. In total, 46,061 genes (99.22%) were functionally annotated, each supported by at least one database (Table 6).

Non-coding RNA annotation According to the characteristic features of different RNA types, we predicted non-coding RNAs in the red crucian carp genome using tRNAscan-SE (v1.3.1, parameter: -q)[64] for tRNAs, BLASTN (v2.11.0+, parameter: -evalue 0.01) for rRNAs, and Rfam (v14.8, parameter: cmscan --rfam --nohmmonly)[65] for miRNAs and snRNAs. The final results are summarized in the Table 7.

ARTICLE IN PRESS

Table 5. Basic statistical results of gene prediction in red crucian carp genome.

Type	Gene set	Protein coding gene number	Average gene length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length (bp)	Average intron length (bp)
De novo	Genscan	50,059	19,442.51	1,563.57	8.37	186.87	2,426.82
	AUGUSTUS	41,764	12,731.50	1,472.85	8.48	173.74	1,505.73
Homolog	Carassius auratus	116,083	18,362.10	2,040.62	11.11	183.61	1,613.77
	Carassius carassius	91,300	20,478.31	1,917.32	11	174.25	1,855.51
	Danio rerio	66,061	15,672.37	1,552.13	8.44	183.99	1,898.89
	Cyprinus carpio	190,674	21,185.01	1,998.76	11.8	169.32	1,775.78
Liftoff	Carassius carassius	41,041	15,367.26	1,721.06	9.87	174.43	1,539.04
	Carassius auratus	44,220	14,098.36	1,698.84	9.55	177.8	1,449.40
Trans ORF	RNAseq	18,085	14,931.13	1,627.57	10.81	250	1,247.05
BUSCO		5,614	11,770.72	1,630.06	10.09	161.56	1,115.61
MAKER		44,781	14,363.17	1,550.66	9.32	231.51	1,467.04

ARTICLE IN PRESS

HiFAP	46,424	14,592.74	1,691.16	9.39	229.18	1,482.88

Table 6. Statistical results of functional annotation for protein-coding genes in red crucian carp genome.

	Gene Number	Percent (%)	mRNA Number	Percent (%)
Total	46,424	100.00	54,784	100.00
Annotated	46,061	99.22	54,293	99.10
NR	46,026	99.14	54,236	99.00
SwissProt	39,087	84.20	46,552	84.97
TrEMBL	45,918	98.91	54,121	98.79
KOG	36,153	77.88	43,038	78.56
TF	11,850	25.53	13,741	25.08
InterPro	43,638	94.00	51,369	93.77
GO	38,538	83.01	45,564	83.17
KEGG_ALL	45,633	98.30	53,800	98.20
KEGG_KO	30,229	65.12	36,150	65.99
Pfam	41,425	89.23	48,631	88.77
Unannotated	363	0.78	491	0.90

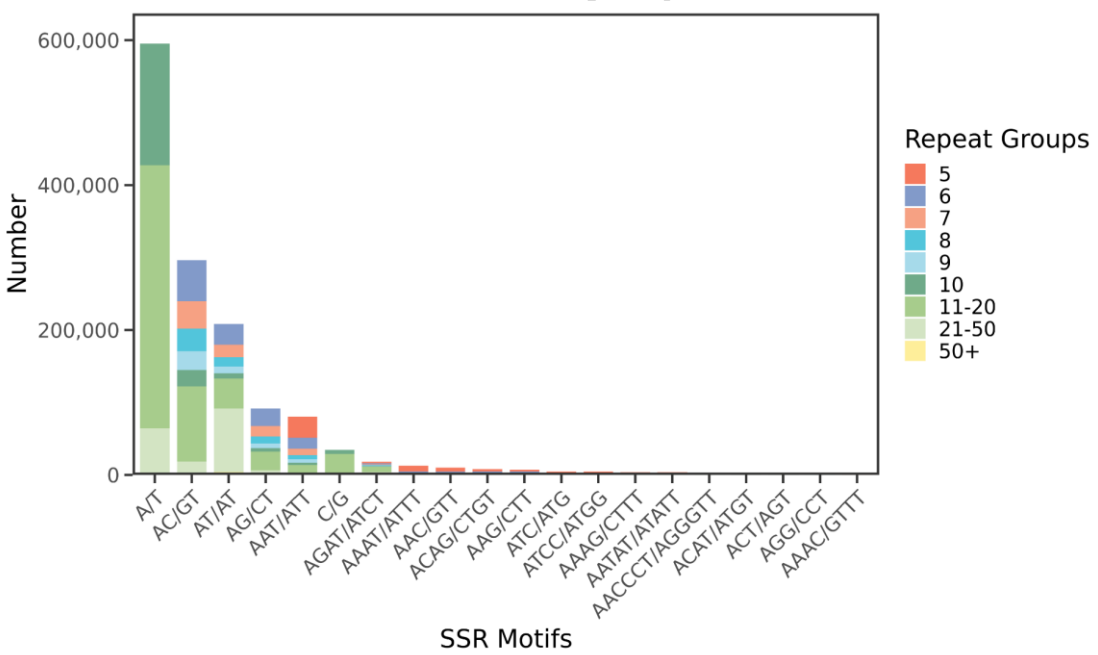


Fig. 4 Stacked bar plot of SSR motif repeat counts in the red crucian carp near T2T genome. (including complementary sequences)

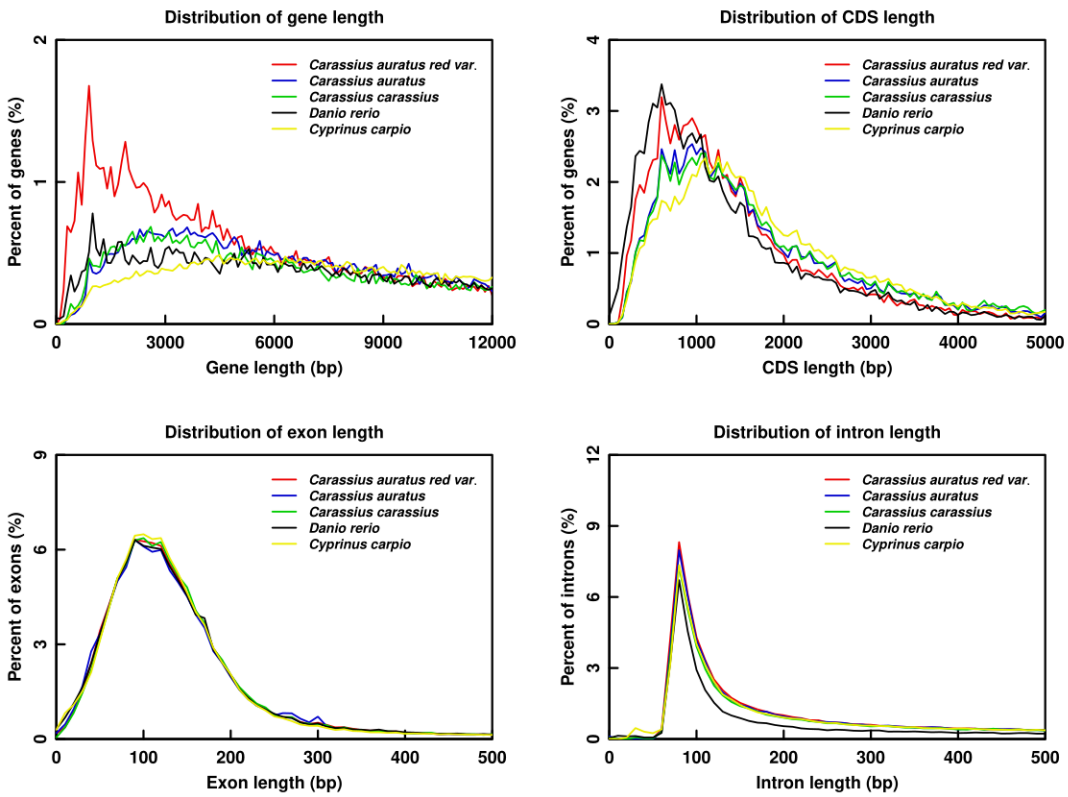


Fig. 5 Distribution of gene set element lengths in five closely related species, including gene length, CDS length, exon length, and intron length among *Carassius auratus* red var., *Carassius auratus*, *Carassius carassius*, *Danio rerio*, and *Cyprinus carpio*.

Table 7. Statistical results of non-coding RNA annotation in red crucian carp genome.

Type		Copy	Average length (bp)	Total length (bp)	% in genome
miRNA		3,309	89	293,417	0.018889
tRNA		22,454	76	1,706,907	0.109883
rRNA		42,628	115	4,913,931	0.316337
snRNA	Total	3,380	148	500,627	0.032228
	CD-box	387	145	55,933	0.003601
	HACA-box	136	152	20,683	0.001331
	splicing	2,842	148	420,279	0.027056
	scaRNA	15	249	3,732	0.000240

Data Records

The raw data (including MGI, PacBio Hi-Fi, Hi-C, ONT ultra-long, and RNA sequencing data) has been submitted to the NCBI Sequence Read Archive under BioProject number PRJNA1468686, with accession numbers SRR38809693[66], SRR38809692[67], SRR38809691[68], SRR38809690[69], and SRR38809689[70].

ARTICLE IN PRESS

Chromosome-level genome assembly for the red crucian carp has been submitted to NCBI GenBank, accession number: GCA_058768465.1[71]. Annotation files and FASTA-formatted genomic sequence data are available via the Figshare platform: https://doi.org/10.6084/m9.figshare.30959696[72].

In the meantime, the red crucian carp genome assembly is also stored in the Genome Warehouse (GWH) of the National Genome Data Center (NGDC) in China, with the accession number: GWHHJOS00000000.1[73]. Raw sequencing data (including MGI, HiFi, Hi-C, ONT, and RNA sequencing data) have been submitted and deposited in the Genome Sequence Archive (GSA) of the China National Center for Biological Information (CNCB, CRA036282)[74].

Technical Validation

Assessment of genome assembly To verify the species identity of the assembly, the genome was divided into non-overlapping 10-kb windows. These sequences were then aligned against the NCBI nucleotide database (NT) using BLAST. The highest percentage of significant matches was to the genus *Carassius*, confirming the correctness of the assembly (Table 8).

The assessment of genome completeness was initially conducted utilising the BUSCO (v5.7.1, parameter: --miniprot -m genome -f –offline)[53] and Compleasm (v0.2.6, parameter: compleasm run -m busco)[75], which were employed against the highly conserved orthoDB database (https://www.orthodb.org/). The utilisation of these tools respectively resulted in the identification of 99.40% and 99.86% complete BUSCOs. Subsequently, consensus quality values (QV) were evaluated with Merqury (v1.3, default parameters)[76]. K-mers derived from high-accuracy unassembled reads—including short-read and PacBio long-read data—were compared to the assembled genome, yielding QVs of 36.48 and 64.44, respectively. Additionally, sequencing data were aligned to the genome with the following tools: BWA (v0.7.12-r1039, parameter: -M)[77] for DNA short reads; minimap2 (v2-2.24, parameter: -ax map-ont / map-hifi / map-pb)[34] for PacBio long reads and ONT ultra-long reads; and HISAT2 (v2.1.0, parameter: --dta --new-summary)[51] for transcriptomic data. The resulting mapping rates were 99.79%, 99.99%, 100.00%, and 97.20%, respectively. Collectively, these analyses confirm the high quality of the red crucian carp genome assembly.

Forty-one chromosomes were assembled to telomere-to-telomere (T2T) completeness, while the remaining chromosomes contained telomeres detected only at one end or retained unresolved gaps. The failure to achieve full T2T status for these chromosomes is likely attributable to complex, highly repetitive genomic regions that remain challenging to resolve even with ONT ultra-long sequencing reads. Ongoing progress in both sequencing technology and T2T assembly algorithms is poised to address these gaps, paving the way for a fully assembled red crucian carp genome.

Table 8. NT database alignment summary (top five genera by alignment count for 10-kb genomic segments).

Genus	BLAST number	Genome split number	Hit number	Percentage of genome (%)	Percentage of hits (%)	Median identity (%)

ARTICLE IN PRESS

Carassius	109,987	155,368	155,302	70.79	70.82	97.3
Zebrafish	15,523	155,368	155,302	9.99	10	84.83
Cyprinus	10,768	155,368	155,302	6.93	6.93	89.17
Carp	3,579	155,368	155,302	2.3	2.3	91.06
Onychostoma	3,120	155,368	155,302	2.01	2.01	86.7

Table 9. BUSCO completeness and accuracy evaluation of the red crucian carp carp genome annotations.

Type	BUSCO		Compleasm	
	Proteins	Percentage (%)	Proteins	Percentage (%)
Complete BUSCOs	3,616	99.34	3,616	99.34
Complete Single-Copy BUSCOs	1,553	42.66	1,346	36.98
Complete Duplicated BUSCOs	2,063	56.68	2,270	62.36
Fragmented BUSCOs	6	0.16	6	0.16
Missing BUSCOs	18	0.49	18	0.49
Total BUSCO groups searched	3,640	100	3,640	100

Genome annotation evaluation Genome annotation completeness was validated using BUSCO (v5.7.1, parameters: --miniprot -m genome -f --offline)[53] and Compleasm (v0.2.6, parameter: compleasm run -m busco)[75]. Both tools identified 3,616 complete BUSCOs (99.34%, Table 9). A Circos plot (Circos v0.69-6, default parameters[78]) was generated to provide an overall visualisation of the genomic distributions, such as GC content, gene density, and repetitive sequence density (Fig. 2b).

Code availability

This study did not involve the development of any custom code. All bioinformatics tool versions and parameters are described in the Methods section.

References

1. Ren, L. et al. Genomes reveal pervasive distant hybridization in nature among cyprinid fishes. *GigaScience* **14**, http://10.1093/gigascience/giae117 (2025).
2. Yu, P. et al. Causal gene identification and desirable trait recreation in goldfish. *Sci China Life Sci* **65**, 2341-2353, http://10.1007/s11427-022-2194-7 (2022).
3. Liu, Q. et al. The chimeric genes in the hybrid lineage of *Carassius auratus* cuvieri (♀)×*Carassius auratus* red var. (♂). *Science China Life Sciences* **61**, 1079-1089, http://10.1007/s11427-017-9306-7 (2018).
4. Bian, C. et al. Telomere-to-telomere chromosome-scale genome assemblies of black and golden koi carp variants support construction of an ancient karyotype of Cypriniformes. *GigaScience* **14**, http://10.1093/gigascience/giaf073 (2025).
5. Xu, P. et al. The allotetraploid origin and asymmetrical genome evolution of the common carp Cyprinus carpio. *Nature Communications* **10**, 4625, http://10.1038/s41467-019-12644-1 (2019).
6. Wang, Y. et al. Comparative genome anatomy reveals evolutionary insights into a unique amphitriploid fish. *Nature Ecology & Evolution* **6**, 1354-1366, http://10.1038/s41559-022-01813-z (2022).

ARTICLE IN PRESS

7. Luo, J. et al. From asymmetrical to balanced genomic diversification during rediploidization: Subgenomic evolution in allotetraploid fish. **6**, eaaz7677, http://doi:10.1126/sciadv.aaz7677 (2020).
8. Qin, Q. et al. The Autotetraploid Fish Derived from Hybridization of Carassius auratus red var. (Female) × Megalobrama amblycephala (Male)1. *Biology of Reproduction* **91**, http://10.1095/biolreprod.114.122283 (2014).
9. Lian, K. et al. Chromosome-level genome assembly of Qihe gibel carp. *Scientific data* **12**, 1290, http://10.1038/s41597-025-05636-y (2025).
10. Qin, Q. et al. Autotriploid origin of Carassius auratus as revealed by chromosomal locus analysis. *Sci China Life Sci* **59**, 622-626, http://10.1007/s11427-016-5040-7 (2016).
11. Šimková, A. et al. MHC genes and parasitism in Carassius gibelio, a diploid-triploid fish species with dual reproduction strategies. *BMC evolutionary biology* **13**, 122, http://10.1186/1471-2148-13-122 (2013).
12. Liu, Z. et al. Gynogenetic offspring from distant hybridization of red crucian carp × blunt snout bream: Insights into barbel origin. *Aquaculture* **609**, http://10.1016/j.aquaculture.2025.742791 (2025).
13. Chen, Z. et al. De novo assembly of the goldfish (Carassius auratus) genome and the evolution of genes after whole-genome duplication. *Science advances* **5**, eaav0547, http://10.1126/sciadv.aav0547 (2019).
14. Qin, Q.B. et al. Differential expression of HPG-axis genes in autotetraploids derived from red crucian carp Carassius auratus red var., ♀ × blunt snout bream Megalobrama amblycephala, ♂. *J. Fish Biol.* **93**, 1082-1089, http://10.1111/jfb.13818 (2018).
15. Zhang, K. et al. Unveiling potential sex-determining genes and sex-specific markers in autotetraploid Carassius auratus. *Sci China Life Sci* **67**, 2444-2458, http://10.1007/s11427-023-2694-5 (2024).
16. Lu, M. et al. Manipulating Unisexual-Sexual Reproduction Transition to Engineer Genome-Reconstructed Polyploids. *Advanced science (Weinheim, Baden-Wurttemberg, Germany)* **12**, e06024, http://10.1002/advs.202506024 (2025).
17. Zoological, C.J.A.Z.S. HISTORY OF DOMESTICATION AND FACTORS OF VARIETAL FORMATION OF THE COMMON GOLDFISH, CARASSIUS AURATUS. (1954).
18. Ota, K.G. & Abe, G. Goldfish morphology as a model for evolutionary developmental biology. *Wiley interdisciplinary reviews. Developmental biology* **5**, 272-295, http://10.1002/wdev.224 (2016).
19. Omori, Y. & Kon, T. Goldfish: an old and new model system to study vertebrate development, evolution and human disease. *Journal of biochemistry* **165**, 209-218, http://10.1093/jb/mvy076 (2019).
20. Sánchez-Bretaño, A. et al. The liver of goldfish as a component of the circadian system: Integrating a network of signals. *General and comparative endocrinology* **221**, 213-216, http://10.1016/j.ygcen.2015.05.001 (2015).
21. Blanco, A.M. et al. Why goldfish? Merits and challenges in employing goldfish as a model organism in comparative endocrinology research. *General and comparative endocrinology* **257**, 13-28, http://10.1016/j.ygcen.2017.02.001 (2018).
22. Filice, M. et al. The goldfish Carassius auratus: an emerging animal model for comparative cardiac research. *Journal of comparative physiology. B, Biochemical, systemic, and environmental physiology* **192**, 27-48, http://10.1007/s00360-021-01402-9 (2022).
23. Shi, W. et al. Microplastics impair olfactory-mediated behaviors of goldfish Carassius auratus. *Journal of hazardous materials* **409**, 125016, http://10.1016/j.jhazmat.2020.125016 (2021).
24. Yuan, J. et al. A telomere-to-telomere genome assembly of koi carp (Cyprinus carpio) using long reads and Hi-C technology. *GigaScience* **14**, http://10.1093/gigascience/giaf087 (2025).
25. Chen, S. et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics (Oxford, England)* **34**, i884-i890, http://10.1093/bioinformatics/bty560 (2018).
26. Ranallo-Benavidez, T.R. et al. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* **11**, 1432, http://10.1038/s41467-020-14998-3 (2020).

ARTICLE IN PRESS

27. Shen, F. et al. The allotetraploid horseradish genome provides insights into subgenome diversification and formation of critical traits. *Nature Communications* **14**, 4102, http://10.1038/s41467-023-39800-y (2023).
28. Cheng, H. et al. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170-175, http://10.1038/s41592-020-01056-5 (2021).
29. Roach, M.J. et al. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC bioinformatics* **19**, 460, http://10.1186/s12859-018-2485-7 (2018).
30. Wingett, S. et al. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* **4**, 1310, http://10.12688/f1000research.7334.1 (2015).
31. Durand, N.C. et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell systems* **3**, 95-98, http://10.1016/j.cels.2016.07.002 (2016).
32. Dudchenko, O. et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science (New York, N.Y.)* **356**, 92-95, http://10.1126/science.aal3327 (2017).
33. Durand, N.C. et al. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell systems* **3**, 99-101, http://10.1016/j.cels.2015.07.012 (2016).
34. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)* **34**, 3094-3100, http://10.1093/bioinformatics/bty191 (2018).
35. Mount, D.W. Using the Basic Local Alignment Search Tool (BLAST). *CSH protocols* **2007**, pdb.top17, http://10.1101/pdb.top17 (2007).
36. 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**, http://10.1093/gigascience/giaa094 (2020).
37. Walker, B.J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963, http://10.1371/journal.pone.0112963 (2014).
38. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**, 573-580, http://10.1093/nar/27.2.573 (1999).
39. Lin, Y. et al. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Horticulture research* **10**, uhad127, http://10.1093/hr/uhad127 (2023).
40. Beier, S. et al. MISA-web: a web server for microsatellite prediction. *Bioinformatics (Oxford, England)* **33**, 2583-2585, http://10.1093/bioinformatics/btx198 (2017).
41. Price, A.L. et al. De novo identification of repeat families in large genomes. *Bioinformatics (Oxford, England)* **21 Suppl 1**, i351-358, http://10.1093/bioinformatics/bti1018 (2005).
42. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mobile DNA* **10**, 48, http://10.1186/s13100-019-0193-0 (2019).
43. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research* **34**, W435-439, http://10.1093/nar/gkl200 (2006).
44. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA11Edited by F. E. Cohen. *Journal of Molecular Biology* **268**, 78-94, http://https://doi.org/10.1006/jmbi.1997.0951 (1997).
45. NCBI GenBank https://identifiers.org/ncbi/insdc.gca:GCA_003368295.1 (2018).
46. NCBI GenBank https://identifiers.org/ncbi/insdc.gca:GCA_963082965.1 (2023).
47. NCBI GenBank https://identifiers.org/ncbi/insdc.gca:GCA_000002035.4 (2017).
48. NCBI GenBank https://identifiers.org/ncbi/insdc.gca:GCA_905221575.1 (2021).
49. Slater, G.S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics* **6**, 31, http://10.1186/1471-2105-6-31 (2005).
50. Shumate, A. & Salzberg, S.L. Liftoff: accurate mapping of gene annotations. *Bioinformatics (Oxford, England)* **37**,

ARTICLE IN PRESS

1639-1643, http://10.1093/bioinformatics/btaa1016 (2021).

51. Kim, D. et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907-915, http://10.1038/s41587-019-0201-4 (2019).
52. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290-295, http://10.1038/nbt.3122 (2015).
53. Manni, M. et al. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular biology and evolution* **38**, 4647-4654, http://10.1093/molbev/msab199 (2021).
54. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC bioinformatics* **12**, 491, http://10.1186/1471-2105-12-491 (2011).
55. Sayers, E.W. et al. Database resources of the national center for biotechnology information. *Nucleic acids research* **50**, D20-d26, http://10.1093/nar/gkab1112 (2022).
56. Coudert, E. et al. Annotation of biologically relevant ligands in UniProtKB using ChEBI. *Bioinformatics (Oxford, England)* **39**, http://10.1093/bioinformatics/btac793 (2023).
57. Shen, W.K. et al. AnimalTFDB 4.0: a comprehensive animal transcription factor database updated with variation and expression annotations. *Nucleic acids research* **51**, D39-d45, http://10.1093/nar/gkac907 (2023).
58. Paysan-Lafosse, T. et al. InterPro in 2022. *Nucleic acids research* **51**, D418-d427, http://10.1093/nar/gkac993 (2023).
59. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25-29, http://10.1038/75556 (2000).
60. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27-30, http://10.1093/nar/28.1.27 (2000).
61. Mistry, J. et al. Pfam: The protein families database in 2021. *Nucleic acids research* **49**, D412-d419, http://10.1093/nar/gkaa913 (2021).
62. Buchfink, B. et al. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* **18**, 366-368, http://10.1038/s41592-021-01101-x (2021).
63. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)* **30**, 1236-1240, http://10.1093/bioinformatics/btu031 (2014).
64. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research* **25**, 955-964, http://10.1093/nar/25.5.955 (1997).
65. Griffiths-Jones, S. et al. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic acids research* **33**, D121-124, http://10.1093/nar/gki081 (2005).
66. NCBI Sequence Read Archive https://identifiers.org/ncbi/insdc.sra:SRR38809693 (2026).
67. NCBI Sequence Read Archive https://identifiers.org/ncbi/insdc.sra:SRR38809692 (2026).
68. NCBI Sequence Read Archive https://identifiers.org/ncbi/insdc.sra:SRR38809691 (2026).
69. NCBI Sequence Read Archive https://identifiers.org/ncbi/insdc.sra:SRR38809690 (2026).
70. NCBI Sequence Read Archive https://identifiers.org/ncbi/insdc.sra:SRR38809689 (2026).
71. NCBI GenBank https://identifiers.org/ncbi/insdc.gca:GCA_058768465.1 (2026).
72. Liu, Z. Annotation file for the *Carassius auratus* red var. t2t genome. *figshare* http://10.6084/m9.figshare.30959696.v4 (2025).
73. Ma, Y. et al. The Updated Genome Warehouse: Enhancing Data Value, Security, and Usability to Address Data Expansion. *Genomics Proteomics Bioinformatics* **23**, http://10.1093/gpbjnl/qzaf010 (2025).
74. Zhang, S. et al. The GSA Family in 2025: A Broadened Sharing Platform for Multi-omics and Multimodal Data. *Genomics, Proteomics & Bioinformatics* **23**, http://10.1093/gpbjnl/qzaf072 (2025).

ARTICLE IN PRESS

ARTICLE IN PRESS

75. Huang, N. & Li, H. compleasm: a faster and more accurate reimplementation of BUSCO. *Bioinformatics (Oxford, England)* **39**, http://10.1093/bioinformatics/btad595 (2023).
76. Rhie, A. et al. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245, http://10.1186/s13059-020-02134-9 (2020).
77. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754-1760, http://10.1093/bioinformatics/btp324 (2009).
78. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-1645, http://10.1101/gr.092759.109 (2009).

Author Contributions

Qinbo Qin conceived and designed the project. Zhengkun Liu conducted the writing, data analysis and sample collection. Yuhan Yang, Zeyang Li, Qian Xiao, Wanjing Peng, Ling Xiong, Enkui Hu, and Siyang Li participated in data analysis and visualization. Chongqing Wang, Kun Zhang, Xu Huang, Xiaowei Xu, Xidan Xu, Yue Zhou, Jinhai Bai, Yuchen Jiang, and Qinbo Qin provided guidance on experimental design and manuscript revision. Qinbo Qin also provided project support.

Competing interests

The authors declare no competing interests.

Acknowledgements

This work was financially supported by Yuelushan Laboratory Breeding Program (Grant No. YLS-2025-ZY02043), the earmarked fund for HARS (Grant No. HARS-07) and the Aid Program for Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province.

ARTICLE IN PRESS