scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly and annotation of the *Triplophysa pappenheimi*

Ying Li[1,2,4], Dan Liu[1,4], Zhanwen Yao[1,2], Qiang Gao[1], Cunfang Zhang[1], Wei Wang[1], Junmei Jia[1], Sijia Liu[3], Fei Tian[3] & Delin Qi[1] ✉

***Triplophysa pappenheimi*, an ecologically important species within the plateau freshwater ecosystem, predominantly inhabits the rapids of the upper Yellow River, its tributaries, and associated lakes, playing a crucial role in maintaining local aquatic biodiversity and serving as a model for plateau fish adaptation studies. In this study, we utilized a combination of Illumina short-read sequencing, PacBio HiFi long-read sequencing, and Hi-C technology to generate the first high-quality, chromosome-level genome assembly of *T. pappenheimi*. The assembled genome spans 626.33 Mb, organized into 25 pseudochromosomes, with a scaffold N50 of 24.41 Mb. We annotated 24,952 genes, representing 98.82% of the predicted protein-coding genes. In addition, we used GetOrganelle v1.7.7.1 to isolate and assemble the complete mitochondrial genome from the Illumina sequencing data, with a total length of 16,574 bp, and successfully achieved circularization. This high-quality genome assembly provides a valuable genomic resource for understanding the evolutionary history, adaptive mechanisms, and biological characteristics of this species.**

Background & Summary

The Qinghai-Tibet Plateau (QTP), recognized as one of the world's highest and most expansive plateaus on Earth, covers an impressive area of 2.5 million square kilometers, with an average elevation ranging between 3,000 and 5,000 meters above sea level. The plateau's dramatic uplift has profoundly influenced both regional and global climate systems, as well as broader environmental dynamics[1,2]. This unique geographical setting, combined with its harsh natural conditions, makes the QTP an exceptional natural laboratory for studying adaptive evolution in organisms. Species inhabiting the plateau are exposed to extreme environmental stressors, including hypoxia, subzero temperatures, and intense ultraviolet radiation. These formidable conditions have driven the evolution of distinctive physiological and genetic adaptations among the plateau's diverse biota[3,4].

The genus *Triplophysa*, commonly known as Tibetan loaches, comprises small, bottom-dwelling, omnivorous fishes of the family Nemacheilidae within the order Cypriniformes[5]. *Triplophysa* is among the most diverse genera of loaches, with nearly 160 species described to date (http://researcharchive.calacademy.org/research/ichthyology/catalog/fishcatmain.asp), over 126 of which are distributed across China. Speciation within this genus shows a remarkable correlation with climatic upheavals associated with the geodynamic evolution of the QTP including paleo-drainage reorganizations, extreme hypothermia, and chronic hypoxia[6]. Widely distributed across river-lake systems of the plateau and its surrounding regions, *Triplophysa* has become a key model for understanding the origins of plateau biodiversity and the mechanisms of environmental adaptation[7]. *T. pappenheimi*, one of the larger species within the genus—reaching up to 28 cm in body length—is endemic to the Yellow River, primarily inhabiting its upper reaches and tributary rapids. As a Yellow River endemic species, *T. pappenheimi* exhibits remarkable adaptations to the plateau's seasonal extremes, including chronic hypoxia and low temperatures. However, it is currently listed as Endangered (EN) on China's Red List of Vertebrates[6,7]. Genomic studies of *T. pappenheimi* may reveal the molecular mechanisms underlying the ecological adaptation of plateau-endemic fishes to high-altitude stressors, offering crucial insights for conserving evolutionary resilience in extreme environments.

[1]State Key Laboratory of Plateau Ecology and Agriculture, Qinghai University, Xining, 810016, China. [2]College of Eco-Environmental Engineering, Qinghai University, Xining, 810016, China. [3]Key Laboratory of Adaptation and Evolution of Plateau Biota, Northwest Institute of Plateau Biology, Chinese Academy of Sciences, Xining, 810001, China. [4]These authors contributed equally: Ying Li, Dan Liu. ✉e-mail: delinqi@126.com

Library types	Insert size (bp)	Clean data (Gb)	Read length (bp)	Sequence coverage (X)
Illumina reads	300	63.70	150	107.1
PacBio reads	20000	83.95	20126	141.0
Hi-C reads	—	81.50	150	275.8
RNA reads	300	88.53	150	—
Total	—	317.68	—	—

Table 1. Sequencing data used for the genome *T. pappenheimi* assembly.

In this study, we present the first high-quality, chromosome-scale genome assembly of *T. pappenheimi*, generated using an integrative multi-platform sequencing approach that combines Illumina short reads, PacBio HiFi long reads, and Hi-C chromatin interaction data. This comprehensive genomic resource provides unprecedented resolution of the genome organization of *T. pappenheimi*, enabling detailed characterization of its genetic architecture and functional elements. The assembly not only facilitates the investigation of fundamental biological questions related to genome structure and regulation in this species, but also offers novel insights into its evolutionary history and ecological specialization. As a foundational genomic framework, this resource will greatly advance future research on the organism's biology and its ecological adaptations.

Methods

Sample collection and DNA extraction. Specimens of *T. pappenheimi* (NCBI Taxonomy ID: 1792974) were collected from the Zequ River, a tributary in the upper reaches of the Yellow River, located in Zeku County, Huangnan Tibetan Autonomous Prefecture, Qinghai Province, China. Species identification was established based on pelvic fin and gonad morphology, with reference to the original description by Fang (1935) and subsequent taxonomic revisions by Kottelat[8].

Genomic DNA was extracted from dorsal muscle tissue of a healthy adult female individual and used for whole-genome sequencing. The sequenced individual was preliminarily identified as female based on the morphology of the pelvic fins and gonads at the time of sample collection. However, its sex has not been further confirmed through genome-based analysis using sex chromosome-linked markers (e.g., W/Z or X/Y systems). Due to limited research resources and the primary objective of this study being the construction of a high-quality reference genome, sequencing was conducted on a single individual. The specimen is currently stored at the State Key Laboratory of Plateau Ecology and Agriculture, Qinghai University, Xining, China, and is identified as specimen Ge2023.

Genome sequencing. For genome assembly, two distinct libraries with insert sizes of 300 bp and 20 kb were constructed using the Illumina TruSeq Nano DNA Library Prep Kit and the SMRTbell Template Prep Kit (Pacific Biosciences, USA), respectively. The 300 bp library was sequenced on the Illumina NovaSeq 6000 platform, generating raw reads that were subjected to rigorous quality filtering. The Illumina short reads were quality-filtered using fastp[9] with the following parameters: -q 10 -u 50 -y -g -Y 10 -e 20 -l 100 -b 150 -B 150. This process yielded 63.70 Gb (107.1 × coverage) of high-quality Illumina reads, which were subsequently used for genome size estimation, heterozygosity analysis, and repeat content characterization. Concurrently, the 20 kb library was sequenced on the PacBio Sequel II platform, producing 83.95 Gb (141.0 × coverage) of high-quality subreads with an average length of 20,126 bp after adapter removal and quality control (Table 1).

To achieve chromosome-level genome assembly, Hi-C libraries were prepared using the *MboI* restriction enzyme according to established protocols[10]. The libraries were sequenced on an Illumina HiSeq X Ten platform with 150 bp paired-end reads. After removing adapter sequences and low-quality reads through stringent filtering, a total of 81.50 Gb of high-quality Hi-C data were obtained for downstream scaffolding and genome assembly (Table 1).

RNA extraction and transcriptome sequencing. To support genomic structural annotation with transcriptomic evidence, RNA sequencing was performed on multiple tissues, including dorsal muscle, skin, gills, liver, intestine, spleen, kidney, heart, brain, eye, and blood, collected from adult specimens. Total RNA was extracted from each tissue samples using TRIzol reagent (Kangwei Century, China), followed by DNase I treatment to eliminate genomic DNA contamination. RNA quality was assessed by measuring concentration with a Nanodrop 2000 spectrophotometer (Thermo Fisher Scientific, USA) and evaluating integrity using an Agilent 2100 Bioanalyzer system equipped with LabChip GX (PerkinElmer, USA). For library construction, cDNA libraries were prepared using the Hieff NGS Ultima Dual-mode mRNA Library Prep Kit for Illumina (Yeasen Biotechnology, China). Library quality was evaluated through several metrics: concentration was measured with a Qubit 2.0 Fluorometer (Thermo Fisher Scientific, USA), insert size distribution was analyzed using the Agilent 2100 Bioanalyzer, and library quantification was performed by quantitative PCR (qPCR). High-quality libraries were sequenced on the Illumina NovaSeq 6000 platform (Illumina, USA) using 150 bp paired-end reads. A total of 88.53 Gb of clean data was generated. The clean data for each sample reached 10.19 Gb, and all samples exhibited a Q30 quality score of ≥95.49% (Table 1).

Contamination assessment. To assess potential contamination in the extracted samples, we randomly selected 10,000 single-end reads (150 bp) from a 350 bp library and aligned them against the NT database using BLAST[11] (ncbi-blast + v2.2.29; parameters: -num_descriptions 100, -num_alignments 100, -evalue 1e-5). According to standard evaluation criteria, significant alignment of reads to evolutionarily distant species (e.g., plants

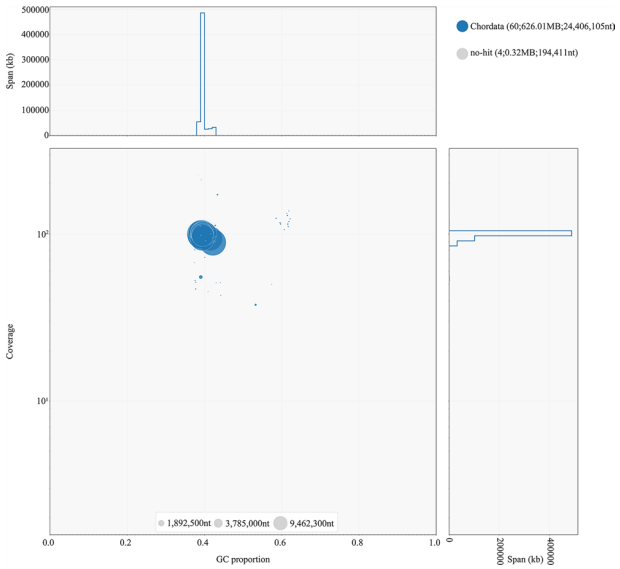

Fig. 1 Contamination assessment of genome assembly using BlobTools. The figure displays the distribution characteristics of assembled sequences in three dimensions: GC content (X-axis), sequencing coverage (Y-axis), and sequence length (bubble size). Each dot represents a scaffold, with size proportional to scaffold length and color indicating taxonomic annotation (blue: Chordata; gray: no-hit). The top and right panels show GC content distribution histograms and coverage distribution histograms of scaffold lengths, respectively.

or microorganisms) would indicate potential contamination requiring further investigation. The results showed that 14.31% of the sequencing reads aligned to non-target species, specifically: *Cyprinus carpio* (7.09%), *Danio rerio* (1.32%), *Danio kyathit* (0.75%), and other species (5.15%). Comprehensive NT database comparison confirmed no substantial contamination in the sample. These non-target alignments were likely due to spurious alignment in conserved genomic regions.

In addition, we employed BlobTools v1.1.1 to assess contamination in the genome assembly. This tool integrates GC content, sequencing depth, and taxonomic annotation based on BLASTN alignment against the NT database (E-value $\leq$ 1e−25). The analysis revealed that 99.95% (626.01 Mb) of the assembled sequences were confidently annotated to Chordata, with an average length of 24.41 Mb, GC content of approximately 42%, and uniform sequencing depth. Only 0.32 Mb (0.05%) of sequences remained unannotated (no-hit), with an average length of 194.41 kb, indicating minimal contamination and suitability for subsequent genomic analyses (Fig. 1).

To assess extranuclear DNA contamination levels, we aligned Illumina sequencing data against the plastid genome using SOAP v2.21[12] with parameters -m 260 -x 440. The plastid genome data referred to was sourced from the NCBI nt database (2021 release), covering all complete chloroplast and mitochondrial genomes of green plants (Viridiplantae). The alignment results showed that, out of 426,306,324 total sequencing reads, only 1,380 paired-end reads and 3,087 single-end reads were successfully aligned. The total proportion of extranuclear reads is <0.01%, with mitochondrial-derived reads accounting for <0.01%, far below the empirical threshold of 5%, confirming that extranuclear DNA contamination is extremely low and has no significant impact on subsequent nuclear genome assembly.

Genome survey. The Illumina short reads were quality-filtered using fastp[9]. The processed reads were subsequently used for genome size estimation. A k-mer (k = 21) frequency distribution was generated from a 300 bp insert-size library. K-mer counting was performed using Jellyfish v2.1.4[13], and genome characteristics were estimated using Genomescope v2.0[14] under a diploid model. The primary k-mer peak corresponded to a depth of 43, yielding an estimated haploid genome size of ~594.75 Mb. Additionally, the k-mer distribution indicated a repeat content of 24.58% and a heterozygosity rate of 0.46%, suggesting a relatively simple genome structure. This low complexity facilitates high-resolution genome assembly in downstream analyses.

De novo assembly of *T. pappenheimi* genome. High-fidelity (HiFi) reads were assembled using Hifiasm v0.19.11[15], producing a draft genome comprising 133 contigs with a total length of 651.63 Mb. The assembly demonstrated high continuity, with a contig N50 of 21.49 Mb and a maximum contig length of 29.81 Mb. Following decontamination of plastid-derived sequences using BLAST and removal of redundant sequences based on Hi-C data, the final set of contigs totaled 626.33 Mb in length, with the contig N50 remaining at 21.49 Mb.

To anchor the contigs to chromosomes, a total of 272,408,169 paired-end reads were generated from the Hi-C library and aligned to the polished genome using BWA v0.7.17[16] with default parameters. Paired reads with mate mapped to a different contig were retained for Hi-C-based scaffolding. Invalid reads—such as self-ligated fragments, non-ligated fragments, Start NearRsite, PCR duplicates, random breaks, large/small fragments, and extreme fragments—were removed. Contig clustering was performed using an agglomerative hierarchical clustering algorithm implemented in Lachesis[17], with the following parameters: CLUSTER_MIN_RE_SITES = 321,

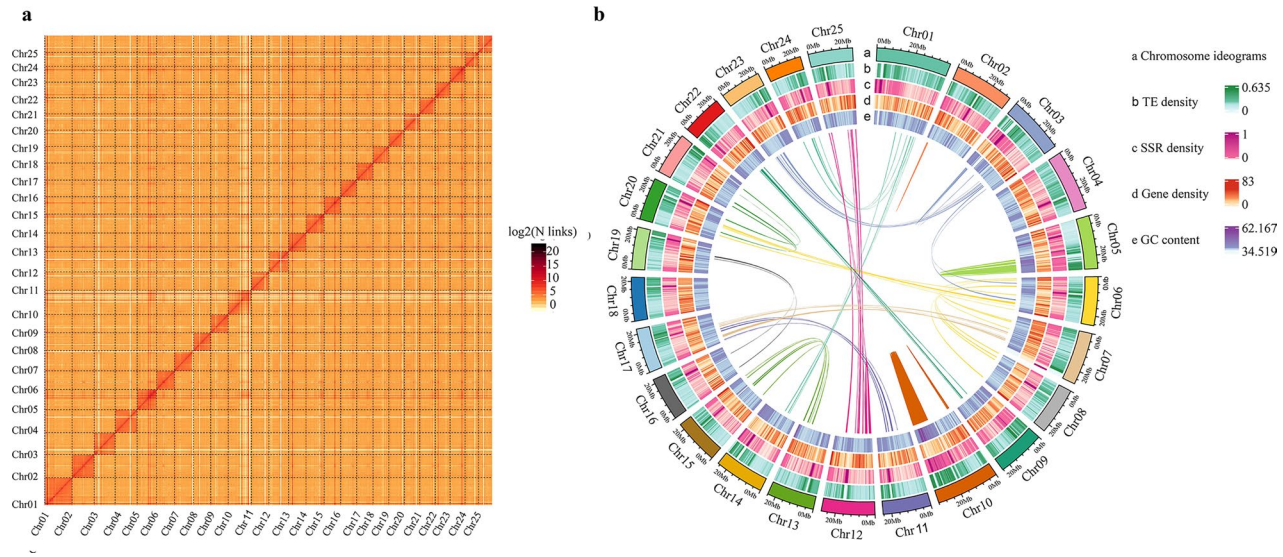

Fig. 2 Characteristics of the *T. pappenheimi*. (**a**) Hi-C intra-chromosomal contact map of the *T. pappenheimi* genome assembly. (**b**) Circos plot of the *T. pappenheimi* genome assembly. (**a**) Chromosome ideograms; (**b**) TE density; (**c**) SSR density; (**d**) Gene density; (**e**) GC content. The inner connecting lines represent syntenic relationships between conserved homologous gene pairs across chromosomes, revealing intra-genomic structural conservation or rearrangements.

Pseudo-Chromosomes	Cluster Num	Cluster Len	Order Num	Order Len
Chr01	11	38200413	6	37848595
Chr02	4	30349898	4	30349898
Chr03	6	28907091	5	28874142
Chr04	5	30630270	4	30591952
Chr05	5	26982053	4	26750733
Chr06	3	24712553	3	24712553
Chr07	7	27267533	5	26560836
Chr08	1	23935560	1	23935560
Chr09	3	24282718	3	24282718
Chr10	27	32950994	14	32092528
Chr11	3	24405905	3	24405905
Chr12	2	27076240	2	27076240
Chr13	2	24000994	2	24000994
Chr14	6	25396434	2	25144009
Chr15	1	22986202	1	22986202
Chr16	4	22238980	2	22131019
Chr17	4	22584990	4	22584990
Chr18	3	22034587	3	22034587
Chr19	2	21331643	2	21331643
Chr20	1	21485114	1	21485114
Chr21	1	21367219	1	21367219
Chr22	1	20442792	1	20442792
Chr23	1	20846711	1	20846711
Chr24	4	18761051	3	18704724
Chr25	4	22515489	3	22465925
Total (Ratio%)	111 (93.28)	625693434 (99.9)	80 (72.07)	623007589 (99.57)

Table 2. Statistics of the chromosome assemblies using Hi-C data in *T. pappenheimi*.

CLUSTER_MAX_LINK_DENSITY = 2, ORDER_MIN_N_RES_IN_TRUNK = 631, and ORDER_MIN_N_RES_IN_SHREDS = 647, which was subsequently employed to order and orient the clustered contigs. After error correction, chromosome anchoring, and final redundancy filtering, we generated a chromosome-level genome assembly. The final assembly consisted of 64 scaffolds and 119 contigs, with a contig N50 of 21.49 Mb, a

Type	Number	Percentage
Complete BUSCOs (C)	3,480	95.60%
Complete and single-copy BUSCOs (S)	3,421	93.98%
Complete and duplicated BUSCOs (D)	59	1.62%
Fragmented BUSCOs (F)	14	0.38%
Missing BUSCOs (M)	146	4.01%
Total BUSCO groups searched	3,640	—

Table 3. BUSCO analysis of the genome assembly and genes.

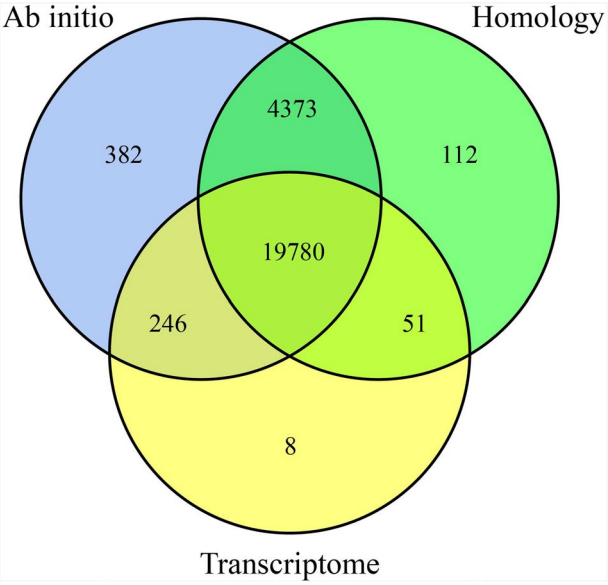

Fig. 3 Distribution chart of genes integrated from three prediction methods.

scaffold N50 of 24.41 Mb, the longest scaffold spanning 37.85 Mb, and the longest contig reaching 29.81 Mb. The genome's GC content was 39.66%. In total, 625.69 Mb of sequence (99.9%) was anchored to 25 chromosomes, among which 623.01 Mb (99.57%) could be confidently ordered and oriented. In addition, we used the R package Circlize to generate a genome-wide Circos plot to visualize the genomic features of *T. pappenheimi*, including gene density, transposable element (TE) density, SSR density, and GC content. The colored lines within the Circos plot represent syntenic relationships between homologous gene pairs located on different chromosomes, highlighting potential structural rearrangements or conserved collinear blocks (Fig. 2b). This resulted in the first high-quality chromosome-level assembly of *T. pappenheimi*, with individual chromosome lengths ranging from 18.70 Mb to 37.85 Mb, comprising 99.57% of the total genome sequence (Table 2 and Fig. 2a,b).

Mitochondrial genome assembly. The mitochondrial genome was assembled using both second-generation and third-generation sequencing data. First, Illumina second-generation sequencing data was used to isolate and assemble the mitochondrial genome using the GetOrganelle v1.7.7.1 tool with parameters set to -F animal_mt -R 42. A complete circular mitochondrial genome sequence was obtained, with a total length of 16,574 bp. Additionally, PacBio Revio third-generation sequencing data was employed to improve the accuracy and completeness of the assembly. By combining both second- and third-generation sequencing data, we successfully assembled a high-quality mitochondrial genome sequence[18].

Assessment of the genome assembly. To evaluate the completeness and accuracy of the *T. pappenheimi* genome assembly, we we conducted a series of comprehensive quality assessments. First, the assembly was analyzed using BUSCO v5.2.218 with the OrthoDB database to assess gene content completeness. Out of 3,640 conserved single-copy orthologous genes, 95.60% were identified as complete, with 93.98% being single-copy and 1.62% duplicated. Only 0.38% of the genes were fragmented, and 4.01% were missing (Table 3). Second, genome integrity was evaluated using CEGMA v2.5[19], which revealed that 98.91% (453 out of 458) of highly conserved eukaryotic genes were present in the assembly. Furthermore, 230 of the 248 ultra-conserved core genes (92.74%) were successfully detected, reinforcing the conclusion of high assembly completeness. To assess assembly continuity and accuracy, Illumina short reads were mapped back to the assembled genome using BWA[20]. The mapping results showed a high alignment rate, with 98.78% of reads successfully aligned to the genome and 99.99% of the genome covered, indicating a high degree of agreement between the assembled genome and the raw sequencing data (Table 3).

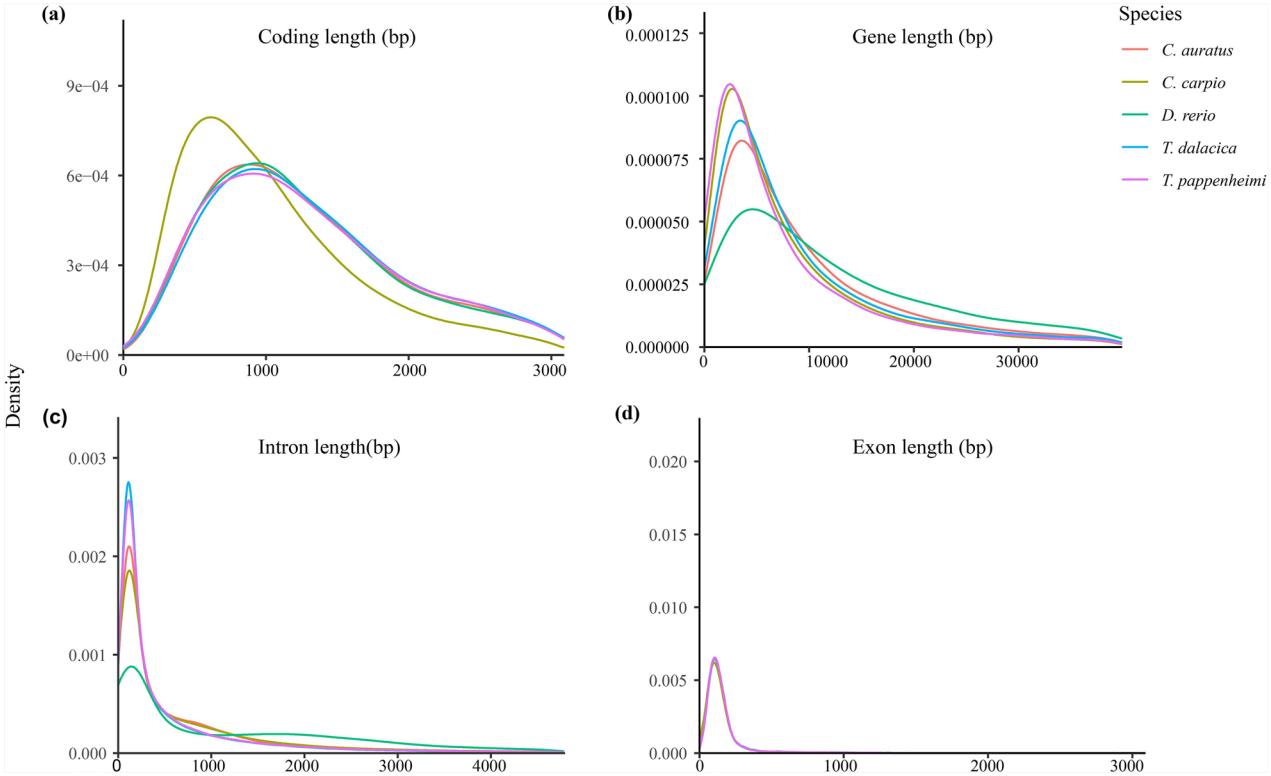

Fig. 4 Comparison of the distribution characteristics for the lengths of CDS (**a**), gene (**b**), intron (**c**) and exon (**d**) in the assembled *T. pappenheimi* genome and four related species.

Repetitive element and non-coding gene annotation in the *T. pappenheimi* genome. TE and tandem repeats in the *T. pappenheimi* genome were annotated using a combination of homology-based and de novo approaches. A custom de novo repeat library was first generated using RepeatModeler2 (http://www.repeatmasker.org/RepeatModeler/) v2.0.1[21], which integrates the outputs of RECON v1.0.8[22] and RepeatScout v1.0.6[23]. To further identify full-length long terminal repeat retrotransposons (fl-LTR-RTs), we applied LTRharvest v1.5.10[24] and LTR_finder v1.07[25], and the results were integrated and refined using LTR_retriever[26] to produce a high-quality, non-redundant LTR library. A comprehensive, non-redundant species-specific TE library was then constructed by merging the de novo repeat library with entries from the Dfam v3.5 database. Prior to gene prediction and annotation, the genome was subjected to soft-masking using RepeatMasker v4.1.2[27] with the curated repeat library, ensuring that repeat sequences—particularly transposable elements—were identified and masked to prevent false positives during gene structure prediction. In total, approximately 150.02 Mb of TE sequences were identified, comprising 23.95% of the genome. Retrotransposons were the predominant class, accounting for 65.91 Mb (or 10.52% of the genome). Among these, long terminal repeats (LTRs) were the most prevalent, totaling 34.95 Mb (5.59%), while short interspersed nuclear elements (SINEs) were the least represented, with 1.75 Mb (0.28%) (Table s1). Tandem repeats were annotated using Tandem Repeats Finder (TRF 4.09)[28] and the MIcroSAtellite identification tool MISA v2.1[29]. In the end, approximately 49.97 Mb of tandem repeat sequences were obtained, accounting for 7.98% (Table s2).

To identify various types of non-coding RNAs (ncRNAs) based on their structural characteristics, multiple specialized prediction strategies were employed. tRNAs were identified using tRNAscan-SE version 1.3.1[30], rRNAs were primarily predicted with barrna version 0.9, and miRNAs, snoRNAs, and snRNAs were predicted based on the Rfam version 14.5[31] database utilizing Infernal version 1.1[32]. In total, 24,480 tRNAs, 14,390 rRNAs, and 180 miRNAs were predicted in the *T. pappenheimi* genome.

Annotation of protein-coding genes. Following repeat masking, protein-coding gene annotation was performed using three primary approaches: homology-based prediction, de novo prediction, and transcriptome-based prediction. For de novo prediction, Augustus version 3.1.0[33] and SNAP 2006-07-28[34] were employed. Homology-based prediction was carried out using GeMoMa version 1.7[35], leveraging protein sequences from homologous species. Transcriptome-based prediction using second-generation sequencing involved two strategies. The first employed Hisat version 2.1.0[36] and StringTie version 2.1.4[37] to assemble transcripts, followed by gene prediction using GeneMarkS-T version 5.1[38]. The second strategy used Trinity version 2.11[39] for transcript assembly, and PASA version 2.4.1[40] for gene structure prediction. For third-generation transcriptome data, transcript alignment was performed using GMAP 2020-06-30, followed by splicing site refinement and gene prediction with PASA. To integrate predictions from all three approaches, EVM version 1.1.1[41] was used, followed by refinement with PASA. For homology-based evidence, protein sequences from four related fish species—*D. rerio*,

Carassius auratus, *C. carpio*, and *Triplophysa dalacica*—were included. As a result, a total of 24,952 protein-coding genes were predicted in the *T. pappenheimi* genome. Among them, most were supported by at least two of the three prediction approaches, demonstrating the high quality of the gene annotation (Fig. 3,Table s3). The total length of the predicted gene set was 300,834,996 bp, with an average gene length of 12,056.55 bp and an average CDS length of 1,754.75 bp. On average, each gene contained 9.93 exons, with an average exon length of 1,755.08 bp and an average intron length of 10,301.47 bp. The statistical data of the gene models for *T. pappenheimi*—covering gene, CDS, exon, and intron lengths—were comparable to those observed in closely related species. (Fig. 4, Table s4).

Gene functions were inferred based on the best alignment matches against multiple public protein databases, including the National Center for Biotechnology Information (NCBI), Non-Redundant (NR), EggNOG[42], KOG, TrEMBL[43], InterPro[44] and Swiss-Prot[43] protein databases using diamond blastp and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database with an E-value threshold of 1E-5. Protein domains were annotated using InterProScan v5.34–73.0[45], referencing the InterPro protein databases. Motifs and domains within gene models were identified using the PFAM databases[46]. Gene Ontology (GO) IDs for each gene were assigned based on annotations from TrEMBL, InterPro and EggNOG. As a result, 24,657 protein-coding genes (98.82%) were successfully annotated by at least one of the aforementioned databases.

Pseudogene prediction. Pseudogenes usually have similar sequences to functional genes, but may have lost their biological function because of some genetic mutations, such as insertion and deletion. The GenBlastA v1.0.4[47] program was used to scan the whole genomes after masking predicted functional genes. Putative candidates were then analyzed by searching for premature stop codons and frame-shift mutations using GeneWise v2.4.1[48]. A total of 73 pseudogenes were ultimately predicted. The total length of the pseudogenes was 233,390 bp and the average length of the pseudogenes was 3,197.12 bp.

Regulatory motif annotation. A motif, also known as a "module" is a locally conserved region within a sequence or a small sequence pattern shared by a set of sequences. It generally refers to the basic structure that constitutes any characteristic sequence, but in most cases, it refers to any sequence pattern that may have molecular function, structural properties, or is related to family members. The motif annotation was performed using InterProScan v5.34–73.0[45]. Genomic analysis revealed the identification of 2,046 putative regulatory short sequence motifs, along with the annotation of 61,781 functional domains within the genome.

Comparative genomics and phylogenetic resources. Protein-coding gene sequences of *T. pappenheimi* and eight other species (*C. carpio*, *D. rerio*, *Misgurnus anguillicaudatus*, *Oryzias latipes*, *Siniperca grahami*, *T. dalacica*, *Triplophysa rosa*, and *T. tibetana*) were collected from the NCBI Genome Database for orthology and gene family clustering analysis. For each gene, only the longest isoform was retained. Orthogroups were identified using OrthoFinder v2.4.0[49]. A total of 29,463 gene families were obtained, including 1,021 conserved single-copy orthologous groups shared across all nine species. These orthologous genes were used for phylogenetic reconstruction and gene family evolution analysis.

Data Records

All raw sequencing data generated from the whole-genome project have been deposited in the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) under the accession number SRP578206. The associated BioProject is registered under the accession PRJNA1248722[50]. The final chromosome assembly has been deposited at DDBJ/ENA/GenBank under the accession JBPJAZ000000000[51].

The corresponding BioSample is registered under the accession SAMN47866966. The dataset includes multiple sequencing libraries, with the following SRA accession numbers corresponding to each library: Whole-genome short-read sequencing (Illumina): SRR33108354 – Ge2023_L1_Illumina (paired-end); Long-read sequencing (PacBio Sequel II): SRR33108355 – Ge2023_pacbio (single-end); Hi-C sequencing (Illumina NovaSeq 6000): SRR33108352 – Ge2023_L2_Hic (paired-end); RNA-seq libraries (Illumina HiSeq 4000) were constructed from eight tissues, with corresponding SRA accession numbers: SRR33108345–SRR33108351 (eye, brain, heart, kidney, liver, skin, muscle) and SRR33108353 (intestine).

The assembled genome, data of the genome annotations, predicted coding sequences, and protein sequences have been deposited at Figshare[52]. The dataset comprises several key files that provide comprehensive information about the genome. Specifically, the final chromosome-level genome assembly sequence is available in FASTA format under the file name "Lachesis_assembly_changed.fa". Additionally, the structural annotation of genes is detailed in the GFF3 format file "Chr_genome_final_gene.gff3". The predicted coding sequences (CDS) are provided in FASTA format within the file "Chr_genome_final_gene.gff3.cds", while the predicted protein sequences are included in the FASTA format file "Chr_genome_final_gene.pep". Furthermore, the annotation of repetitive elements is documented in the GFF3 format file "Chr_genome_repeat.gff3".

Technical Validation

RNA integrity. RNA sequencing samples were collected from multiple tissues (dorsal muscle, skin, gills, liver, intestine, spleen, kidneys, heart, brain, eyes, and blood) of an adult female fish. Prior to library construction, RNA quality was assessed using standard methods to ensure the suitability of the material for sequencing.

Gene annotation validation. To validate the gene predictions, we assessed their completeness using BUSCO with the vertebrata database, which comprises 3,354 conserved core genes. The analysis revealed that 3,269 (97.47%) of these BUSCO genes were successfully identified in our predicted gene set. Among these, 3,202 (95.47%) were complete single-copy genes, while 67 (2.00%) were complete duplicates. In addition, the length

distributions of different gene elements in the *T. pappenheimi* genome were compared with 4 related species (*D. rerio*, *C. auratus*, *C. carpio*, and *T. dalacica*), and the results showed high similarity, indicating the reliability of our genome annotation.

Data availability

The raw sequencing data generated in this study have been deposited in the in the NCBI Sequence Read Archive (SRA) under the BioProject accession number PRJNA1248722[50]. The chromosome-level genome assembly has been deposited at DDBJ/ENA/GenBank under the accession JBPJAZ000000000[51]. The assembled genome, annotation files, coding sequences (CDS), and protein sequences are available on Figshare under https://doi.org/10.6084/m9.figshare.28783331.v1[52].

Code availability

No custom code or in-house scripts were used in this study. All data processing and analyses were conducted using publicly available bioinformatics software. The software versions and parameters are detailed in the Methods section. For tools where specific parameters were not explicitly stated, default settings recommended by the developers were applied.

Received: 22 April 2025; Accepted: 6 October 2025;
Published online: 21 November 2025

References

1. Li, J. & Fang, X. Uplift of the Tibetan Plateau and environmental changes. *Chinese Science Bulletin* **44**, 2117–2124, https://doi.org/10.1007/BF03182692 (1999).
2. Favre, A. *et al*. The role of the uplift of the Qinghai-Tibetan Plateau for the evolution of Tibetan biotas. *Biological Reviews* **90**, 236–253, https://doi.org/10.1111/brv.12107 (2015).
3. Qiu, Q. *et al*. The yak genome and adaptation to life at high altitude. *Nature Genetics* **44**, 946-+, https://doi.org/10.1038/ng.2343 (2012).
4. Scheinfeldt, L. B. & Tishkoff, S. A. Living the high life: high-altitude adaptation. *Genome Biology* **11**, 133–133 (2010).
5. Liu, S. Q. *et al*. Phylogenetic relationships of the Cobitoidea (Teleostei: Cypriniformes) inferred from mitochondrial and nuclear genes with analyses of gene evolution. *Gene* **508**, 60–72, https://doi.org/10.1016/j.gene.2012.07.040 (2012).
6. Jiang, Z., Jiang, J., Wang, Y., Zang, E. & Zhang, Y. Red List of China's Vertebrates. *Biodiversity Science* **24**, 501–551+615 (2016).
7. He, D., Chen, Y. & Chen, Y. Molecular Phylogeny and Biogeography of the Genus *Triplophysa* (Cypriniformes: Nemacheilidae) in the Qinghai-Xizang Plateau Region. *Advances in Natural Sciences*, 1395–1404 (2006).
8. Kottelat, M. Conspectus Cobitidum: An Inventory of The Loaches of The World (Teleostei: Cypriniformes: Cobitoidei). *Raffles Bulletin of Zoology*, 1–175 (2012).
9. Chen, S. F., Zhou, Y. Q., Chen, Y. R. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, 884–890, https://doi.org/10.1093/bioinformatics/bty560 (2018).
10. Belton, J. M. *et al*. Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276, https://doi.org/10.1016/j.ymeth.2012.05.001 (2012).
11. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic Local Alignment Search Tool. *Journal of Molecular Biology* **215**, 403–410, https://doi.org/10.1006/jmbi.1990.9999 (1990).
12. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714, https://doi.org/10.1093/bioinformatics/btn025 (2008).
13. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
14. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* **11**, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
15. Cheng, H. Y., Concepcion, G. T., Feng, X. W., Zhang, H. W. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170-+, https://doi.org/10.1038/s41592-020-01056-5 (2021).
16. Wang, F. Y. *et al*. Chromosome-level assembly of *Gymnocypris eckloni* genome. *Scientific Data* **9**, https://doi.org/10.1038/s41597-022-01595-w (2022).
17. Burton, J. N. *et al*. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nature Biotechnology* **31**, 1119-+, https://doi.org/10.1038/nbt.2727 (2013).
18. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212, https://doi.org/10.1093/bioinformatics/btv351 (2015).
19. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genornes. *Bioinformatics* **23**, 1061–1067, https://doi.org/10.1093/bioinformatics/btm071 (2007).
20. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).
21. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
22. Bao, Z. R. & Eddy, S. R. Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Research* **12**, 1269–1276, https://doi.org/10.1101/gr.88502 (2002).
23. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**, I351–I358, https://doi.org/10.1093/bioinformatics/bti1018 (2005).
24. Ellinghaus, D., Kurtz, S. & Willhoeft, U. *LTRharvest*, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9**, https://doi.org/10.1186/1471-2105-9-18 (2008).
25. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**, W265–W268, https://doi.org/10.1093/nar/gkm286 (2007).
26. Ou, S. J. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiology* **176**, 1410–1422, https://doi.org/10.1104/pp.17.01310 (2018).
27. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Current Protocols in Bioinformatics* **25**, 4.10.11–14.10.14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
28. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580, https://doi.org/10.1093/nar/27.2.573 (1999).
29. Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**, 2583–2585, https://doi.org/10.1093/bioinformatics/btx198 (2017).
30. 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, https://doi.org/10.1093/nar/25.5.955 (1997).

31. Griffiths-Jones, S. Annotating Non-Coding RNAs with Rfam. *Current Protocols in Bioinformatics* **9**, 12.15.11–12.15.12, https://doi.org/10.1002/0471250953.bi1205s9 (2005).
32. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935, https://doi.org/10.1093/bioinformatics/btt509 (2013).
33. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* **24**, 637–644, https://doi.org/10.1093/bioinformatics/btn013 (2008).
34. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59, https://doi.org/10.1186/1471-2105-5-59 (2004).
35. Keilwagen, J. *et al*. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res* **44**, e89, https://doi.org/10.1093/nar/gkw092 (2016).
36. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357–360, https://doi.org/10.1038/nmeth.3317 (2015).
37. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295, https://doi.org/10.1038/nbt.3122 (2015).
38. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res* **43**, e78, https://doi.org/10.1093/nar/gkv227 (2015).
39. Grabherr, M. G. *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644–652, https://doi.org/10.1038/nbt.1883 (2011).
40. Haas, B. J. *et al*. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654–5666, https://doi.org/10.1093/nar/gkg770 (2003).
41. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
42. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**, D309–d314, https://doi.org/10.1093/nar/gky1085 (2019).
43. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365–370, https://doi.org/10.1093/nar/gkg095 (2003).
44. Mitchell, A. *et al*. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research* **43**, D213–D221, https://doi.org/10.1093/nar/gku1243 (2015).
45. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240, https://doi.org/10.1093/bioinformatics/btu031 (2014).
46. Finn, R. D. *et al*. The Pfam protein families database. *Nucleic Acids Research* **36**, D281–D288, https://doi.org/10.1093/nar/gkm960 (2008).
47. She, R., Chu, J. S., Wang, K., Pei, J. & Chen, N. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res* **19**, 143–149, https://doi.org/10.1101/gr.082081.108 (2009).
48. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* **14**, 988–995, https://doi.org/10.1101/gr.1865504 (2004).
49. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**, https://doi.org/10.1186/s13059-019-1832-y (2019).
50. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP578206 (2025).
51. Qi, D. L. *GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_051294615.1 (2025).
52. Qi, D. Chromosome-level genome assembly and annotation of the *Triplophysa pappenheimi*. *figshare. Dataset*. https://doi.org/10.6084/m9.figshare.28783331.v1 (2025).

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (No. 32371578) and Chief Scientist Program of Qinghai Province (2024-SF-102).

Author contributions

D.L.Q. and F.T. planned the project. Y.L., D.L., and Z.W.Y. performed the experiments. D.L.Q., Q.G., C.F.Z. and S.J.L. performed the data analyses. D.L., J.M.J. and S.J.L. assisted with sampling and experimentation. Y.L., D.L. and D.L.Q. wrote and revised the manuscript. Also, all authors read, edited and approved the final manuscript.

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/s41597-025-06111-4.

Correspondence and requests for materials should be addressed to D.Q.

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.