scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-scale assembly of *Artemia tibetiana* genome, first aquatic invertebrate genome from Tibet Plateau

Panpan Chen[1,2,3], Zhaoyan Zhong[2,3], Wei-Xin Jin[1,2], Jin Sun[2] ✉ & Shi-Chun Sun[1,2] ✉

Genomic-level studies on the adaptive evolution of animals in the Qinghai-Tibet Plateau have been rapidly increasing. However, most studies are concentrated on vertebrates, and there are few reports on invertebrates. Here, we report the chromosome-level genome assembly for the brine shrimp *Artemia tibetiana* from Kyêbxang Co, a high-altitude (4620 m above sea level) salt lake on the plateau, based on the combination of Illumina, Nanopore long-reads and Hi-C sequencing data. The assembled genome is 1.69 Gb, and 94.83% of the assembled sequences are anchored to 21 pseudo-chromosomes. Approximately 75% of the genome was identified as repetitive sequences, which is higher than most crustaceans documented so far. A total of 17,988 protein-coding genes were identified, among them 14,388 were functionally annotated. This genomic resource provides the foundation for whole-genome level investigation on the genetic adaptation of *Artemia* to the harsh conditions in the Qinghai-Tibet Plateau.

Background & Summary

The Qinghai-Tibet Plateau, which is known as the "Roof of the World" and the "Third Pole of the World"[1], is the highest and largest plateau on Earth, with an average altitude over 4,000 m above sea level[2]. It is characterized by cold, low oxygen, and strong ultraviolet radiation, which has driven striking phenotypic and genetic adaptations in local species[3–5]. Rapid advancements of genomic technologies have greatly promoted the acquisition of genome data of Qinghai-Tibet Plateau organisms and researches on adaptive evolution at genome level. However, most of the existing studies have been concentrated on vertebrates, such as mammals[6–12], birds[13,14], reptiles[15,16], amphibians[17] and some fishes unique to the plateau[18–21], genome level adaptation study about invertebrates was only documented for Tibetan migratory locusts[22].

The brine shrimp *Artemia* (Branchiopoda: Anostraca: Artemiidae), which is a small halophilic planktonic crustacean[23], is distributed in hypersaline habitats all over the world except Antarctica, and the altitude range can be from −154 m to more than 5,000 m[24,25]. At least 352 salt lakes are distributed on the Qinghai-Tibet Plateau[26], and more than 70 of these salt lakes are known to be inhabited by *Artemia*[25]. Kyêbxang Co, located in Shuanghu County, Tibet, with an elevation of 4620 m above sea level and an area of 188 km^2 (Fig. 1A), is the highest salt lake currently with commercial harvesting of *Artemia* resting eggs in the world, and harbors the largest *Artemia* population on the plateau[27]. The *Artemia* from Kyêbxang Co and the surrounding salt lakes (e.g., Yangnapen Co, located about 10 km southwest to Kyêbxang Co) are bisexual. Previous studies identified these populations as *Artemia tibetiana* Abatzopoulos, Zhang & Sorgeloos, 1998[24,28] or *Artemia urmiana*[29,30]. Recently, Asem *et al.*[31] identified the *Artemia* from Qixiang Lake (= Kyêbxang Co) as *Artemia sorgeloosi* Asem, Eimanifar, Hontoria, Rogers & Gajardo, 2023. Each of these three species represents a distinct clade in trees of mitochondrial DNA marker[27,31]. However, they could not be separated from each other in trees of the nuclear *ITS1*[27,30], and populations of "*A. sorgeloosi*" and "*A. tibetiana*" linages were clustered together in analyses of 13 microsatellite loci[27]. Therefore, comprehensive investigations remain imperative to clarify the taxonomic assignment of the Kyêbxang Co *Artemia* population, as well as many other populations from the Qinghai-Tibet

[1]Fisheries College, Ocean University of China, Qingdao, 266000, China. [2]MOE Key Laboratory of Evolution & Marine Biodiversity and Institute of Evolution and Marine Biodiversity, Ocean University of China, Qingdao, 266000, China. [3]These authors contributed equally: Panpan Chen, Zhaoyan Zhong. ✉e-mail: jin_sun@ouc.edu.cn; sunsc@ouc.edu.cn

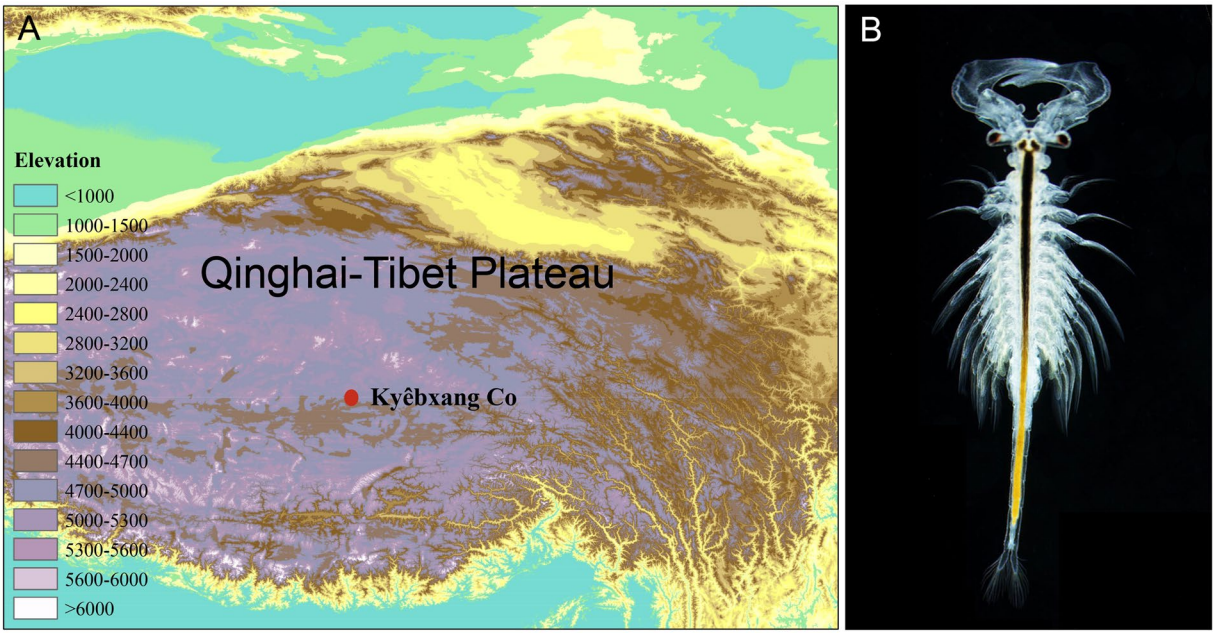

Fig. 1 (**A**) Location of Kyêbxang Co on the Qinghai-Tibet Plateau. (**B**) A photo of *Artemia tibetiana* (male) from Kyêbxang Co.

Plateau; as a provisional solution, the name *Artemia tibetiana* is used for the *Artemia* from Kyêbxang Co in this study. *Artemia* in the plateau area has formed its unique reproduction and survival methods, which are obviously a result of adaptation to the plateau environment[32]. To date, the molecular mechanism for the adaptation of *Artemia* in the plateau has only been explicated at mitochondrial genome level[24,33], and no whole genome are documented. To provide referential resource for future adaptive evolution studies on high-altitude *Artemia*, as well as other aquatic invertebrates, the chromosome-level genome of *A. tibetiana* from Kyêbxang Co was generated with a combination of Illumina, Nanopore long-reads and Hi-C sequencing in this study.

Methods

Sample collection. *Artemia tibetiana* was initially sampled in the form of resting eggs from Kyêbxang Co (32°27′N, 89°57′E), Tibet, China in 2019. In laboratory, resting eggs were hatched at temperature of 20 °C, salinity of 30 (natural seawater). Nauplii were cultured at 20 °C, salinity of 40 (about the salinity level of lake water), and continuous light (1500 lx). In order to reduce the heterozygosity of the genome and considering that its sex chromosomes are ZW type, we chose male adults (Fig. 1B) from the inbreeding F3 generation for subsequent analyses. To reduce bacterial DNA contamination, *Artemia* were treated with 25 mg/L chloramphenicol (in culture media) for 48 hours, during which no food was supplied. After washing with sterile double-distilled water, samples were frozen in liquid nitrogen and stored at −80 °C until DNA extraction.

Hi-C analysis was based on 40 F1 males from the same parents. They were starved for 24 hours before collection, washed three times with sterile double distilled water, and stored in liquid nitrogen until DNA extraction.

To assist in gene annotation, a total of 11 *Artemia* samples were collected for RNA sequencing. Among them 10 samples were individuals at five developmental stages (nauplius, post-metanauplius, post-larva, adult and ovigerous female) reared under 40 and 100 salinity conditions, respectively, and the other was decapsulated resting eggs incubated for 12 hours at the aforementioned hatching condition. All samples were frozen in liquid nitrogen and stored at −80 °C until RNA extraction.

Genomic DNA and RNA sequencing. For the Illumina sequencing, genomic DNA was extracted individually by using classical SDS extraction method. The gDNA quality was evaluated through 1% agarose gel electrophoresis, Nanodrop microvolume spectrophotometer (Thermo Fisher Scientific, MA, USA), and Qubit® DNA Assay Kit (Invitrogen, USA) in the Qubit® 3.0 fluorometer. The paired-end libraries with an insert size of 300 bp were constructed using the NEB Next® Ultra™ DNA Library Prep Kit (NEB, USA) following manufacturer's instructions and sequenced on the Illumina Hiseq 2500 platform (PE 150 bp) at Novogene Bioinformatics Technology (Beijing, China). We obtained 96 Gb and 380 Gb Illumina clean reads for genome survey and assembly, respectively (Table 1).

For the ONT sequencing, high molecular weight DNA was extracted by using classical SDS extraction method. Quality evaluation was performed through 1% agarose gel electrophoresis, Nanodrop microvolume spectrophotometer (Thermo Fisher Scientific, MA, USA), and Qubit® DNA Assay Kit (Invitrogen, USA) in the Qubit® 3.0 fluorometer. A total of 6 μg DNA from an inbreeding F3 individual was used as input material for the ONT library preparation. We referred to the company's standard protocol to construct the ONT library and used PromethION Flow Cell (R9.4, ONT) by using the PromethION sequencer (ONT) from Novogene Bioinformatics Technology (Beijing, China) to obtain raw sequence reads. The raw Nanopore sequencing signal

Sequencing strategy	Platform	Usage	Insert size	Clean data (Gb)	Depth (×)
Short-read	Illumina	Genome survey	300 bp	96.0	56
Short-read	Illumina	Genome assembly	300 bp	380.0	224
Long-read	Nanopore	Genome assembly	10–20 kb	97.3	57
Hi-C	Illumina	Hi-C assembly	300 bp	268.7	159
RNA-seq	Illumina	Annotation-evidence	300 bp	64.9	38

Table 1. Sequencing data of *Artemia tibetiana* and their usage in this study.

data (fast5) were then base-called with Guppy v5 (Oxford Nanopore Technologies Ltd.) in high accuracy mode, and 97.3 Gb long reads with average quality of 21.37 ± 3.72 were generated (Table 1).

For the Hi-C analysis, the gDNA was extracted from the admixture of 40 F1 male adults (total weight 255 mg) crushed in liquid nitrogen using the traditional SDS extraction procedure. After the sample passed the quality inspection, the libraries were constructed according to the standard protocol described previously[34] at Novogene Bioinformatics Technology (Beijing, China). After adding A-tails to the fragment ends and following ligation by the Illumina paired-end (PE) sequencing adapters, Hi-C sequencing libraries were amplified by PCR (12–14 cycles) and sequenced on Illumina HiSeq 2500 platform (PE 150 bp). Finally, a total of 268.7 Gb (159×) of clean data was generated (Table 1).

For transcriptome sequencing, the total RNA from each sample was extracted using TRIzol (Thermo Fisher Scientific, USA). A total of 11 RNA libraries were constructed with Illumina NEBNext Ultra RNA Library Prep Kit (New England Biolabs, MA, USA) according to the standard protocols, and sequenced on the NovaSeq 6000 platform (PE 150 bp) (Novogene Bioinformatics Technology, Beijing, China). A total of 64.9 Gb transcriptome clean data was obtained (Table 1).

Genome survey. The Illumina raw reads were checked and filtered by Trimmomatic v0.39[35] to remove reads with adapter sequences, unknown nucleotides (Ns), and >20% low quality bases. K-mer analysis[36] based on Jellyfish v2.1.4[37] and GenomeScope[38] software combined with manual correction was used to estimate the genome size, heterozygosity, and repetitive sequences. According to the formula Genome Size = K-mer number/peak depth, the genome size of *A. tibetiana* was estimated to be about 2.17 Gb, with a heterozygosity rate of 0.84% and a repetitive sequence ratio of 79.70%, thus suggesting a high complex genome (Supplementary Table 1).

***De novo* genome assembly and polishing.** The whole genome assembly protocol follows our former genome assembly strategy[39]. The Illumina raw reads were filtered with Trimmomatic v0.39[35] and Kraken[40] before assembly. Two strategies were used to assemble the genome, the assembly using the ONT data alone and the hybrid assembly of ONT and Illumina data. In the assembly of high quality ONT data, multiple assemblers were used, including Flye v2.9[41], Shasta v0.8.0[42], wtdbg2 v2.4[43], and Raven v1.0[44]. Their assembly quality was evaluated with Quast v5.0.2[45] and BUSCO v5.2.2[46], and the best assembly result was chosen for subsequent analysis. The hybrid assembly of ONT and Illumina was conducted using MaSuRCA v4.0.5[47].

The assembled genomic sequences were further polished by Flye v2.9[41] following the "polish-target" parameter with two iterations based on all the error-corrected ONT data, followed by two iterations of Purge_dups pipeline v1.2.5[48] to remove the heterozygous contigs. The resultant genome was polished twice using the POLCA pipeline from MaSuRCA v4.0.5[47] with Illumina short-reads. Finally, BlobTools v1.1.1[49] and MEGAN v6.21.7[50] were combined to remove bacterial contamination (Fig. 2). After comparison and evaluation (Supplementary Table 2), we merged the results from Flye and MaSuRCA using Quickmerge[51], and redid the above polishing analysis, obtaining the final contig level draft genome with a size of 1684.45 Mb, a contig N50 of 363,426 bp (Table 2) and a GC content of 35.29%. The completeness of genome assembly was assessed by the single-copy ortholog set (BUSCO v.5.2.2[46]) against arthropoda_odb10 with 1013 conserved genes. The result indicated that 82.8% ([S:77.5%, D:5.3%]) of genes were completed, 7.1% of genes were fragmented, and 10.1% of genes were missing (Table 3).

Genome scaffolding by Hi-C sequencing. Low-quality and adapter sequences through Hi-C sequencing were removed using Trimmomatic v0.39[35]. We used HiC-Pro pipeline v3.1.0[52] to select valid-ligation reads with the restriction enzyme of MboI. Juicer v1.6[53] and 3D-DNA (3D-de novo assembly)[54] were then used to scaffold contigs onto chromosomes, with the misjoins, wrong order, and opposite orientation manually corrected using Juicebox v1.11.08[55]. The contiguity of the genome was assessed by Quast v5.0.2[45], and the completeness of the genome was evaluated by BUSCO v5.2.2[46] against the Arthropoda conserved orthologous gene set (arthropoda_odb10). As a result, 94.83% of the contigs were anchored to 21 pseudo-chromosomes, which were presented in the heatmap of the chromatin contact matrix (Fig. 3A and Table 4). The final chromosome-scale genome has a size of 1689.12 Mb and a GC content of 35.29%, with the N50 scaffold of 76.94 Mb and the maximum 106.53 Mb in length (Table 4). It is similar to the genome size (1.7 Gb) and GC content (35.5%) of *Artemia sinica*[56]. BUSCO (Benchmarking Universal Single-Copy Orthologs) evaluation of the assembly reached 85.0% (Table 3), which ranks third in the *Artemia* genomes documented so far[56,57].

Repeat sequence and non-coding RNA annotation. We annotated repetitive sequences by both the *de novo* and homologue methods. Firstly, RepeatModeler v2.0.2a[58] was used to generate a *de novo* species-specific repeat library, and RepeatMasker v4.1.4[59] software was run to map with the repeat library. Secondly, RepeatMasker v4.1.4[59] combined with the Repbase v20181026[60] database was run for homology prediction. Thirdly, the two

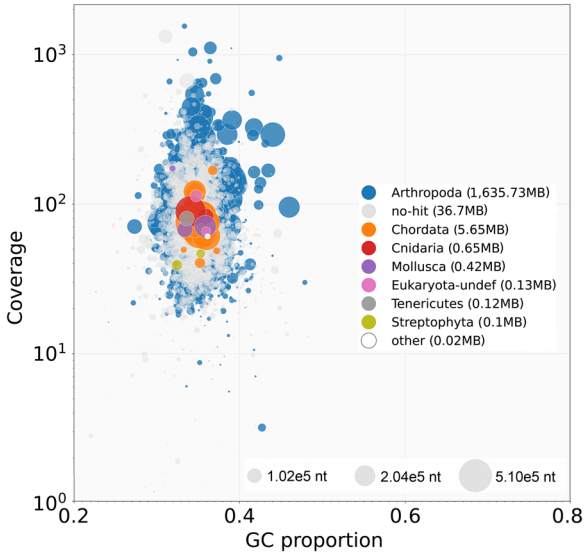

Fig. 2 BlobTools view on the mean GC content, sequencing coverage and NCBI species hit of the genomic assembly of *Artemia tibetiana*. In this result, one contig corresponding to Tenericutes and two contigs corresponding to Streptophyta were further confirmed to belong to Metazoa through BLAST against NCBI nr database.

Labels	Statistics
Draft genome assembly	
Number of contigs	10,645
Contig N50 (bp)	363,426
Contig N90 (bp)	85,207
Maximum contig size (bp)	2,038,326
Number ≥ 5 kb	9,496
Total size (bp)	1,684,446,245
Chromosome-level genome assembly	
Number of scaffolds	3,420
Scaffold N50 (bp)	76,939,700
Scaffold N90 (bp)	55,754,734
Maximum scaffold size (bp)	106,533,452
Number ≥ 5 kb	1,726
Total size (bp)	1,689,109,743

Table 2. Genome assembly statistics of *Artemia tibetiana*.

BUSCO parameters	Contig	Hi-C	Annotation
Complete BUSCOs (C)	82.8	85.0	77.1
Complete and single-copy BUSCOs (S)	77.5	79.7	72.8
Complete and duplicated BUSCOs (D)	5.3	5.3	4.3
Fragmented BUSCOs (F)	7.1	5.6	9.1
Missing BUSCOs (M)	10.1	9.4	13.8

Table 3. BUSCO evaluation (%) of genome assemblies and annotation.

results were integrated to complete the annotation of the final repeats, and the repeated sequences were masked as N bases to reduce the calculation amount of subsequent analyses. About 1262.65 Mb of repeat sequences were finally annotated, accounting for 74.96% of the whole genome, which is higher than that documented for most crustaceans. The main type of repeat sequences are interspersed repeats, including long interspersed repeat sequences, DNA transposons, and long terminal repeat retrotransposons, which account for 10.55%, 10.07%, and 1.82% of the genome, respectively (Table 5). However, still 47.03% of the repeated sequences were not classified

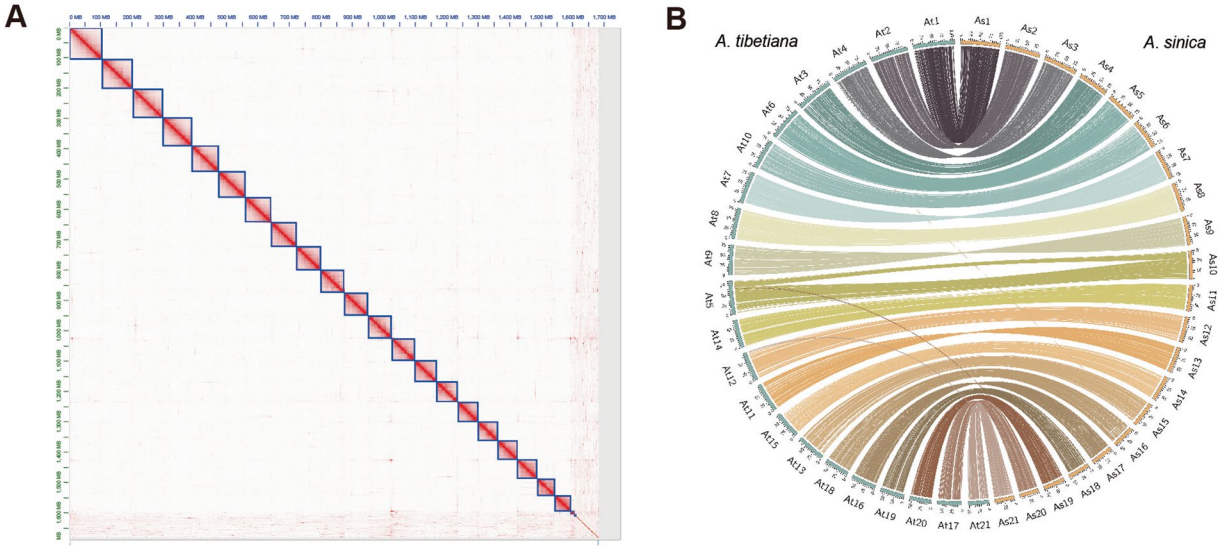

Fig. 3 The assembly of *Artemia tibetiana* genome and syntenic analysis between genomes of two *Artemia* species. (**A**) The Hi-C contact matrix of *A. tibetiana* genome. Blue squares represent pseudo-chromosomes. Color bars illuminate the Hi-C contact density in the plot. (**B**) Chromosome-level synteny analysis between *A. tibetiana* and *Artemia sinica*. Lines in different colors depict interchromosomal synteny.

Chromosomes ID	Length (bp)	% of genome
HiC_chr_1	106,533,452	6.31
HiC_chr_2	97,240,591	5.76
HiC_chr_3	95,445,780	5.65
HiC_chr_4	92,305,618	5.46
HiC_chr_5	85,875,895	5.08
HiC_chr_6	84,981,423	5.03
HiC_chr_7	82,585,736	4.89
HiC_chr_8	79,452,315	4.70
HiC_chr_9	77,704,023	4.60
HiC_chr_10	76,939,700	4.56
HiC_chr_11	74,862,970	4.43
HiC_chr_12	74,578,640	4.42
HiC_chr_13	72,633,413	4.30
HiC_chr_14	71,322,918	4.22
HiC_chr_15	68,059,180	4.03
HiC_chr_16	63,420,823	3.75
HiC_chr_17	63,155,552	3.74
HiC_chr_18	62,415,307	3.70
HiC_chr_19	62,295,147	3.69
HiC_chr_20	55,754,734	3.30
HiC_chr_21	54,144,441	3.21
Total	1,601,707,658	94.83

Table 4. Statistics on the length of the 21 pseudo-chromosomes in *Artemia tibetiana* genome.

(Table 5), which might be due to the specificity of the species' own repeated sequences or the incompleteness of the genome assembly. Additionally, tRNAscan-SE v1.4[61] with default parameters was used to predict tRNA genes, and Barrnap (barrnap 0.9: rapid ribosomal RNA prediction; https://github.com/tseemann/barrnap) was used for rRNA prediction. In total, 13,732 tRNAs and 188 rRNAs were identified, accounting for 0.06% and 0.01% of the genome, respectively (Table 6).

Gene prediction and functional annotation. Three evidences, including transcripts-based, ab initio and homology-based, were employed to predict the protein-coding genes, and MAKER v3.01.04[62] pipeline was used for final annotation. For transcripts-based evidence, Illumina clean data filtered by Trimmomatic v0.39[35] was used to perform transcriptome assembly. *De novo* and genome-guided transcriptome assemblies of the RNA-seq

Type	Length (bp)	% of genome
DNA	169,684,943	10.07
DNA/hobo-Activator	36,904,082	2.19
DNA/Tc1-IS630-Pogo	6,682,343	0.40
DNA/MULE-MuDR	22,346,651	1.33
DNA/PiggyBac	2,732,021	0.16
DNA/Tourist/Harbinger	829,935	0.05
DNA/Other (Mirage, P-element, Transib)	225,521	0.01
LINE	177,767,371	10.55
LINE/CRE/SLACS	1,555	0.00
LINE/L2/CR1/Rex	53,477,239	3.17
LINE/R1/LOA/Jockey	19,341,844	1.15
LINE/R2/R4/NeSL	229,136	0.01
LINE/RTE/Bov-B	64,311,591	3.82
LINE/L1/CIN4	998,582	0.06
SINE	1,847,455	0.11
SINE/Penelope	19,588,373	1.16
LTR	30,688,493	1.82
LTR/BEL/Pao	91,777	0.01
LTR/Ty1/Copia	223,984	0.01
LTR/Gypsy/DIRS1	29,758,349	1.77
LTR/Retroviral	518,891	0.03
Rolling-circles	22,254,474	1.32
Unknown	792,142,669	47.03
Total interspersed repeats	1,172,130,931	69.59
Small RNA	14,592,581	0.87
Satellites	4,583,938	0.27
Simple repeats	48,848,127	2.90
Low complexity	1,948,703	0.12
Total	1,262,654,504	74.96

Table 5. Statistics of *de novo* annotated repeat sequences in *Artemia tibetiana* genome.

Type	Number	Average length (bp)	Total length (bp)	% of genome
rRNA	188	1,338.3	251,603	0.01
tRNA	13,732	78.4	1,076,656	0.06

Table 6. Statistics of non-coding RNAs in *Artemia tibetiana* genome.

data from samples of different development stages were performed using Trinity v2.15.1[63] with and without the '--jaccard_clip' parameter, respectively. BUSCO v5.2.2[46] was used to evaluate the assembly results, and the best results of *de novo* and genome-guided assemblies were merged. The cd-hit-est program in the CD-HIT v4.8.1[64] was then used to remove redundancy and obtain the final transcripts. For ab initio-based prediction, RNA-seq reads with quality above Q30 were aligned to the modified genome of *A. tibetiana* (repeat regions were masked and sequences with a length less than 100 kb were removed) using Hisat2 v0.6.1[65], and Braker2 v2.1.6[66] was then used to train Augustus v3.4.0[67]. For homology-based prediction, protein-coding genes (except mitochondrial genes) of crustaceans were downloaded from the NCBI and UniProt databases. Their redundant sequences were filtered using CD-HIT v4.8.1[64], and the integrated results were used as the input file for subsequent annotation analyses. A final annotation was obtained by writing the above three annotations into the configuration file of MAKER v3.01.04[62], according to the weight of 10, 7 and 3 for transcriptome prediction, homology prediction and ab initio prediction, respectively. Furthermore, the Liftoff[68] was used to lift the genome to chromosome-level annotation. Using BUSCO v5.2.2[46] with the parameter "-m protein" and the Arthropoda conserved gene set "arthropoda_odb10", the completeness of gene annotation was evaluated finally. In total, 17,988 protein-coding genes were predicted in the *A. tibetiana* genome with the BUSCO completeness score of 77.1% (Table 3), which may be due to the high proportion of repetitive sequences affecting the completeness of genome assembly[69,70].

For functional annotation, we have tried several methods. Firstly, we searched the NR database with BLASTP v2.14.0[71], and the result file was imported into the functional analysis module of OmicsBox v3.0.30 (https://www.biobam.com/omicsbox/#functional) to perform the Gene Ontology (GO) annotation. Secondly, Pannzer2 online website (http://ekhidna2.biocenter.helsinki.fi/sanspanz/) was used to perform the GO annotation based on the Uniprot database. Thirdly, eggNOG-mapper[72] online website (http://eggnog-mapper.embl.de/) was used to predict orthologous genes for protein-coding genes. Finally, KEGG functional annotation of protein-coding

Type	Number
NR	14,355
BlastKOALA	5,157
eggNOG-mapper	7,744
Pannzer2_uniprot	7,625
Pannzer2_GO	7,199
Annotated	14,388
Unannotated	3,600

Table 7. Functional annotation statistics of *Artemia tibetiana* genome.

genes was completed at the BlastKOALA online website (https://www.kegg.jp/blastkoala/). Overall, 14,388 protein-coding genes were functionally annotated (Table 7).

Synteny analysis. Protein coding sequences of *A. tibetiana* and *A. sinica* genomes plus the gene coordinate file (*.gff* or.*gtf* format) were used for the inter-species synteny analysis. Syntenic genomic blocks between two genomes were identified using MCScanX[73]. The Circos plot was generated using the Circos software[74]. The genomes of *A. tibetiana* and *A. sinica* exhibited complete one-to-one chromosomal synteny with no detectable large-scale rearrangements (fissions, fusions, or deletions) (Fig. 3B), demonstrating both the high-quality assembly of the *A. tibetiana* genome and the remarkable conservation of genomic architecture between these two species.

Data Records

The genome project was deposited at NCBI under BioProject number PRJNA1136257. The raw Illumina sequencing data for genome survey were deposited in the Sequence Read Archive at NCBI under accession number SRR30106271[75]. The raw Illumina sequencing data for genome assembly were deposited in the Sequence Read Archive at NCBI under accession number SRR30106270[76]. Hi-C sequencing data were deposited in the Sequence Read Archive at NCBI under accession number SRR30106263[77]. Nanopore sequencing raw data were deposited in the Sequence Read Archive at NCBI under accession number SRR30106264[78]. RNA-seq data were deposited in the Sequence Read Archive at NCBI under accession numbers SRR30106257–62[79–84], SRR30106265–69[85–89]. The final chromosome assembly was deposited in GenBank under accession number JBGFSX000000000[90]. Furthermore, for broader accessibility, we have deposited the assembled genome, gene annotation and amino acid sequences to Figshare repository under the DOI of https://doi.org/10.6084/m9.figshare.26503894.v1[91].

Technical Validation

Quality assessment of the genomic materials. After the genomic material extraction, we performed a thorough quality assessment. The quality and quantity of gDNA was checked using 1% agarose gel electrophoresis, Nanodrop microvolume spectrophotometer (Thermo Fisher Scientific, MA, USA), and Qubit® DNA Assay Kit (Invitrogen, USA) in the Qubit® 3.0 fluorometer. RNA integrity was assessed using the Fragment Analyzer 5400 (Agilent Technologies, CA, USA), and the sample used in our study had an RNA integrity number (RIN) larger than 9.

Sufficient sequencing data for genome assembly. For Illumina sequencing, about 115 Gb (67×), and 408 Gb (240×) raw data were obtained for genome survey and assembly, respectively. For Hi-C sequencing, 926 Gb of compressed fast5 data and 130 Gb fastq data (76×, average quality is 9.6) were obtained from Novogene Bioinformatics Technology, Beijing, China. In order to obtain higher quality ONT reads, raw ONT reads in the. *fast5* format were re-base-called using Guppy5 (Oxford Nanopore Technologies Ltd.) with the high-accuracy mode, generating a total of 97.3 Gb (57×) ONT data and improving the quality to 21.37 ± 3.72. For Hi-C sequencing, 334.14 Gb raw data (196×) was obtained. For gene annotation, a total of 70.12 Gb raw RNA-seq data from *Artemia* at different development stages and culture conditions was obtained.

Draft genome assembly and assessment. Multiple assemblers were used to achieve the best draft genome, with Quast v5.0.2 and BUSCO v5.2.2 being used for evaluating the contiguity and completeness, respectively (Supplementary Tables 2, 3). The assembly result from Flye v2.9 was the best one with a total length of 2.16 Gb, N50 of 187 Kb and a completed BUSCO score of 85.0% ([S:77.3%, D:7.7%]). The result from MaSuRCA v4.0.5 was the second best overall with a total length of 1.62 Gb, N50 of 160 Kb and a completed BUSCO score of 81.2% ([S:76.3%, D:4.9%]). In order to improve the assembly quality, we integrated the results of Flye and MaSuRCA, generating a completed BUSCO score of 82.8% ([S:77.5%, D:5.3%]) (Table 3), which showed an improvement in comparison with the results of the above assemblers.

Chromosomal clustering assessment. Hi-C sequencing was used for obtaining chromosome-level assembly, and the heatmap exhibits the accuracy of genome assembly, with relatively independent Hi-C signals observed among the 21 pseudo-chromosomes (Fig. 3A), which is consistent with the chromosome number of *A. tibetiana* determined by cytological preparation[92], as well as that assembled from *A. sinica*[56] and *Artemia franciscana*[57]. This clear pattern affirms the successful clustering of chromosomal-length scaffolds. The final chromosome-level

assembly was evaluated using the conserved Arthropoda gene set "arthropoda_odb10" from BUSCO v5.2.2. The genome assembly was found to have a relatively high level of completeness (85.0%) among the crustaceans with high repetitive sequences published so far. The results of gene annotation indicated a completed BUSCO score of 72.8% complete and single-copy, 4.3% complete and duplicated (Table 3), which was similar to the results of *A. franciscana* (76.9% complete and single-copy, 13.3% complete and duplicated)[57]. Synteny analysis between *A. tibetiana and A. sinica* also indicated high quality of *A. tibetiana* genome (Fig. 3B).

Code availability

All data processing commands and pipelines are executed according to instructions and guidelines provided by relevant bioinformatics software. There were no custom scripts or code utilized in this study. The codes used along with versions and parameters of the primary software/tools are available in the supplementary information.

Received: 7 January 2025; Accepted: 1 May 2025;
Published online: 12 May 2025

References

1. Qin, Y. M. *et al*. Testate amoebae (Protozoa) in lakes of the Qinghai-Tibet Plateau: biodiversity, community structures, and protozoic biosilicification in relation to environmental properties and climate warming. *Sci. Total Environ.* **913**, 169661 (2024).
2. Wu, T. Y. The Qinghai-Tibetan plateau: how high do Tibetans live? *High Alt. Med. Biol.* **2**, 489–499 (2001).
3. Storz, J. F. High-altitude adaptation: mechanistic insights from integrated genomics and physiology. *Mol. Biol. Evol.* **38**, 2677–2691 (2021).
4. Storz, J. F. & Cheviron, Z. A. Physiological genomics of adaptation to high-altitude hypoxia. *Annu. Rev. Anim. Biosci.* **9**, 149–171 (2021).
5. Hao, Y. *et al*. Divergent contributions of coding and noncoding sequences to initial high-altitude adaptation in passerine birds endemic to the Qinghai-Tibet Plateau. *Mol. Ecol.* **32**, 3524–3540 (2023).
6. Qiu, Q. *et al*. The yak genome and adaptation to life at high altitude. *Nat. Genet.* **44**, 946–949 (2012).
7. Li, M. Z. *et al*. Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nat. Genet.* **45**, 1431–1438 (2013).
8. Ge, R. L. *et al*. Draft genome sequence of the Tibetan antelope. *Nat. Commun.* **4**, 1858 (2013).
9. Yu, L. *et al*. Genomic analysis of snub-nosed monkeys (*Rhinopithecus*) identifies genes and processes related to high-altitude adaptation. *Nat. Genet.* **48**, 947–952 (2016).
10. Zeng, L. *et al*. Genomes reveal selective sweeps in kiang and donkey for high-altitude adaptation. *Zool. Res.* **42**, 450 (2021).
11. Cai, C. M. *et al*. Comparative genomic analysis of high-altitude adaptation for Mongolia Mastiff, Tibetan Mastiff, and *Canis Lupus*. *Genomics* **114**, 110359 (2022).
12. Zhou, C. *et al*. Chromosome-level genome assembly of the kiang (*Equus kiang*) illuminates genomic basis for its high-altitude adaptation. *Integr. Zool.* **0**, 1–12 (2023).
13. Wang, M. S. *et al*. Genomic analyses reveal potential independent adaptation to high altitude in Tibetan chickens. *Mol. Biol. Evol.* **32**, 1880–1889 (2015).
14. Li, X. J. *et al*. A de novo assembled genome of the Tibetan Partridge (*Perdix hodgsoniae*) and its high-altitude adaptation. *Integr. Zool.* **18**, 225–236 (2023).
15. Li, J. T. *et al*. Comparative genomic investigation of high-elevation adaptation in ectothermic snakes. *Proc. Natl. Acad. Sci.* **115**, 8406–8411 (2018).
16. Yan, C. C. *et al*. Genome of *Laudakia sacra* provides new insights into high-altitude adaptation of ectotherms. *Int. J. Mol. Sci.* **23**, 10081 (2022).
17. Sun, Y. B. *et al*. Whole-genome sequence of the Tibetan frog *Nanorana parkeri* and the comparative evolution of tetrapod genomes. *Proc. Natl. Acad. Sci.* **112**, E125–7E1262 (2015).
18. Liu, H. P. *et al*. Draft genome of *Glyptosternon maculatum*, an endemic fish from Tibet Plateau. *GigaScience* **7**, giy104 (2018).
19. Liu, H. P. *et al*. The sequence and de novo assembly of *Oxygymnocypris stewartii* genome. *Sci Data* **6**, 1–11 (2019).
20. Yuan, D. Y. *et al*. Chromosomal genome of *Triplophysa bleekeri* provides insights into its evolution and environmental adaptation. *GigaScience* **9**, giaa132 (2020).
21. Xiao, S. J. *et al*. Genome of tetraploid fish *Schizothorax o'connori* provides insights into early re-diploidization and high-altitude adaptation. *Iscience* **23**, 1–12 (2020).
22. Ding, D. *et al*. Genetic variation in PTPN1 contributes to metabolic adaptation to high-altitude hypoxia in Tibetan migratory locusts. *Nat. Commun.* **9**, 4991 (2018).
23. Asem, A. *et al*. An overview on the nomenclatural and phylogenetic problems of native Asian brine shrimps of the genus *Artemia* Leach, 1819 (Crustacea, Anostraca). *ZooKeys* **902**, 1–15 (2020).
24. Zhang, H. X. *et al*. Mitochondrial genome sequences of *Artemia tibetiana* and *Artemia urmiana*: assessing molecular changes for high plateau adaptation. *Sci. China Life Sci.* **56**, 440–452 (2013).
25. Zheng, B. & Sun, S. C. Review of the biogeography of *Artemia* Leach, 1819 (Crustacea: Anostraca) in China. *Int. J. Artemia Biology* **3**, 20–50 (2013).
26. Zheng, M. P., Xiang, J., Wei, X. J. & Zheng, Y. *Saline Lakes On The Qinghai-Xizang (Tibet) Plateau* (Beijing Scientific and Technical Publishing House, 1989).
27. Li, W. J., Chen, P. P., Sui, L. Y. & Sun, S. C. Temporal genetic variation mediated by climate change-induced salinity decline, a study on *Artemia* (Crustacea: Anostraca) from Kyêbxang Co, a high altitude salt lake on the Qinghai-Tibet Plateau. *Gene* **902**, 148160 (2024).
28. Han, X., Xu, R., Zheng, Y., Gao, M. & Sui, L. Development of EST-SSR markers and genetic diversity analysis among three *Artemia* species from different geographic populations. *Crustaceana* **92**, 841–851 (2019).
29. Hou, L. *et al*. Molecular systematics of bisexual *Artemia* populations. *Aquac. Res.* **37**, 671–680 (2006).
30. Sainz-Escudero, L., López-Estrada, E. K., Rodríguez-Flores, P. C. & García-París, M. Settling taxonomic and nomenclatural problems in brine shrimps, *Artemia* (Crustacea: Branchiopoda: Anostraca), by integrating mitogenomics, marker discordances and nomenclature rules. *PeerJ* **9**, e10865 (2021).
31. Asem, A. *et al*. Tibetan *Artemia* (Crustacea: Anostraca) mitogenomic biodiversity and population demographics. *Zool. J. Linn. Soc.* **201**, 32–56 (2024).
32. Van Stappen, G., Sui, L. Y., Xin, N. H. & Sorgeloos, P. Characterisation of high-altitude *Artemia* populations from the Qinghai-Tibet Plateau, PR China. *Aquatic Biodiversity: A Celebratory Volume in Honour of Henri J. Dumont*, 179–192 (2003).
33. Wang, W. W. *et al*. Phylogenetic analysis of brine shrimp (*Artemia*) in China using DNA barcoding. *Genomics, Proteomics and Bioinformatics* **6**, 155–162 (2008).
34. Belton, J. M. *et al*. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).

35. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
36. Liu, B. H. *et al*. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv* (2013).
37. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
38. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
39. Sun, J., Li, R., Chen, C., Sigwart, J. D. & Kocot, K. M. Benchmarking Oxford Nanopore read assemblers for high-quality molluscan genomes. *Philos. T. R. Soc. B* **376**, 20200160 (2021).
40. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, 1–12 (2014).
41. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
42. Shafin, K. *et al*. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.* **38**, 1044–1053 (2020).
43. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **17**, 155–158 (2020).
44. Vaser, R. & Šikić, M. Raven: a *de novo* genome assembler for long reads. *BioRxiv* (2020).
45. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150 (2018).
46. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
47. Zimin, A. V. *et al*. The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).
48. Guan, D. F. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
49. Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies. *F1000Research* **6**, 1287 (2017).
50. Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res.* **17**, 377–386 (2007).
51. Chakraborty, M., Baldwin-Brown, J. G., Long, A. D. & Emerson, J. J. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res.* **44**, e147 (2016).
52. Servant, N. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 1–11 (2015).
53. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
54. Dudchenko, O. *et al*. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
55. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
56. Elkrewi, M. *et al*. ZW sex-chromosome evolution and contagious parthenogenesis in *Artemia* brine shrimp. *Genetics* **222**, iyac123 (2022).
57. Bett, V. K., Macon, A., Vicoso, B. & Elkrewi, M. Chromosome-level assembly of *Artemia franciscana* sheds light on sex chromosome differentiation. *Genome Biol. Evol.* **16**, evae006 (2024).
58. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA.* **117**, 9451–9457 (2020).
59. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* **4**, 4–10 (2009).
60. Bao, W., Kojima, K. K. & Kohany, O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
61. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096 (2021).
62. Cantarel, B. L. *et al*. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
63. Grabherr, M. G. *et al*. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* **29**, 644–652 (2011).
64. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **8**, 3150–3152 (2012).
65. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
66. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom. Bioinfrom.* **3**, lqaa108 (2021).
67. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
68. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).
69. Zhang, X. J. *et al*. Penaeid shrimp genome provides insights into benthic adaptation and frequent molting. *Nat. Commun.* **10**, 356–370 (2019).
70. Polinski, J. M. *et al*. The American lobster genome reveals insights on longevity, neural, and immune adaptations. *Sci. Adv.* **7**, eabe8290 (2021).
71. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
72. Huerta-Cepas, J. *et al*. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
73. Wang, Y. *et al*. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49–e49 (2012).
74. Krzywinski, M. *et al*. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
75. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30106271 (2025).
76. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30106270 (2025).
77. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30106263 (2025).
78. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30106264 (2025).
79. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30106257 (2025).
80. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30106258 (2025).
81. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30106259 (2025).
82. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30106260 (2025).
83. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30106261 (2025).
84. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30106262 (2025).
85. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30106265 (2025).
86. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30106266 (2025).
87. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30106267 (2025).
88. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30106268 (2025).

89. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30106269 (2025).
90. Chen, P. P., Zhong, Z. Y., Jin, W. X. & Sun, S. C. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_049724085.1 (2025).
91. Chen, P. P. Chromosomal-scale genome sequencing and assembly of *Artemia tibetiana* from Kyêbxang Co. *figshare* https://doi.org/10.6084/m9.figshare.26503894.v1 (2025).
92. Abatzopoulos, T. J., Zhang, B. & Sorgeloos, P. *Artemia tibetiana*: preliminary characterization of a new *Artemia* species found in Tibet (People's Republic of China). International Study on *Artemia*. LIX. *Int. J. Salt Lake Res.* **7**, 41–44 (1998).

Acknowledgements

This research was supported by the Science and Technology Plan Project of Nagqu City (NQKJ-2024-10), and Science and Technology Plan Project of Tibet Autonomous Region (No. XZ202102YD0022C). We appreciate the computing resources provided on IEMB-1, a high-performance computing cluster operated by the Institute of Evolution and Marine Biodiversity. We acknowledge the High-Performance Biological Supercomputing Center at the Ocean University of China for invaluable support in this research. We thank Mengdi Chu for assistance in preparing Fig. 1.

Author contributions

P.C. and S.C.S. designed the study. P.C. performed the experiments and wrote the draft. W.X.J., P.C. and Z.Z. analysed the data. All authors revised and approved the 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/s41597-025-05136-z.

Correspondence and requests for materials should be addressed to J.S. or S.-C.S.