The chromosome-level genome of Triplophysa dalaica (Cypriniformes: Cobitidae) provides insights into its survival in extremely alkaline environment

Chuanjiang Zhou[1], Bo Hu[1], YongTao Tang[1], Changxing Yang[1], Wenwen Ma[1], Xi Wang[1], Ruyao Liu[1], Xuemeng Yan[1], Jing Dong[1], Xianfeng Wang[1], and Guoxing Nie[1]

[1]Henan Normal University College of Fisheries

July 16, 2020

Abstract

Lake Dali Nur, located in Inner Mongolia, North China, is alkaline, with Triplophysa dalaica one of the three fish species that not only survive, but thrive, in the lake. To investigate the presence of molecular mutations potentially responsible for this adaptation, the whole genome sequence of the species endemic to the lake was sequenced. A total of 126.5 Gb and 106 Gb data, covering nearly 200X of the estimated genome, were generated using long-read sequencing and Hi-C technology, respectively. De novo assembly generated a genome totalled 607.91 Mb, with a contig N50 of 9.27 Mb. Nearly all whole genome sequences were anchored and oriented onto 25 chromosomes, with telomeres for most chromosomes also being recovered. Repeats comprised approximately 35.01% of the whole genome. A total of 23,925 protein-coding genes were predicted, within which, 98.62% could be functionally annotated. Through comparisons of T. dalaica, T. tibetana, and T. siluroides gene models, a total of 898 genes were identified as likely being subjected to positive selection, with several of them potentially associated with alkaline adaptation, such as sodium bicarbonate cotransporter, SLC4A4. Demographic analyses suggested that the Dali population might have diverged from endemic freshwater Hai River populations, approximately 1 million years ago. The high-quality T. dalaica genome, sequenced in this study, not only aids in the analyses of alkaline adaptation, but may also assist in revealing the mysteries of the highly divergent genus Triplophysa.

The chromosome-level genome of *Triplophysa dalaica* (Cypriniformes: Cobitidae) provides insights into its survival in extremely alkaline environment

Chuanjiang Zhou* , Bo Hu, Yongtao Tang, Changxing Yang, Wenwen Ma, Xi Wang, Ruyao Liu, Xuemeng Yan, Jing Dong, Xianfeng Wang, Guoxing Nie*

College of Fisheries, Engineering Technology Research Center of Henan Province for Aquatic Animal Cultivation, Henan Normal University, Xinxiang 453007, Henan, China

* Corresponding authors:

College of Fisheries, Henan Normal University, 46 Jian She Dong Road, Xinxiang, Henan, China

E-mail: niegx@htu.cn or chuanjiang88@163.com **Abstract**

Lake Dali Nur, located in Inner Mongolia, North China, is alkaline, with *Triplophysa dalaica* one of the three fish species that not only survive, but thrive, in the lake. To investigate the presence of molecular mutations potentially responsible for this adaptation, the whole genome sequence of the species endemic to the lake was sequenced.

A total of 126.5 Gb and 106 Gb data, covering nearly 200X of the estimated genome, were generated using long-read sequencing and Hi-C technology, respectively. De novo assembly generated a genome totalled 607.91 Mb, with a contig N50 of 9.27 Mb. Nearly all whole genome sequences were anchored and oriented onto 25 chromosomes, with telomeres for most chromosomes also being recovered. Repeats comprised approximately 35.01% of the whole genome. A total of 23,925 protein-coding genes were predicted, within which, 98.62% could be functionally annotated. Through comparisons of *T. dalaica* ,*T. tibetana* , and *T. siluroides* gene models, a total of 898 genes were identified as likely being subjected to positive selection, with several of them potentially associated with alkaline adaptation, such as sodium bicarbonate cotransporter, *SLC4A4* . Demographic analyses suggested that the Dali population might have diverged from endemic freshwater Hai River populations, approximately 1 million years ago.

The high-quality *T. dalaica* genome, sequenced in this study, not only aids in the analyses of alkaline adaptation, but may also assist in revealing the mysteries of the highly divergent genus *Triplophysa* .

Introduction

Triplophysa dalaica is a member of the family Cobitidae: Cypriniformes. It is widely distributed in northern China, predominately in the trunk and tributaries of the Yellow River, and Inner Mongolia, in artesian waters, such as Lake Dali Nur. This lake is located in an endorheic basin on the eastern Inner Mongolian Plateau, North China, where its alkalinity can rise to 53.57 mmol/L (pH 9.6). It is believed to have expanded to 1,600 km^2 during the early Holocene (11,500–7,600 calibrated years BP), due to a mass influx of glacial meltwater from the Greater Khingan Range. Its water level fluctuated dramatically during the middle Holocene and constantly shrank during the late Holocene (3,450 calibrated years BP to present) (Xiao, Si, Zhai, Itoh, & Lomtatidze, 2008). Due to both the perennial sandstorms and dry weather of Inner Mongolia, water of the Dali Lake has been gradually concentrated and salinized. This has resulted in the long-term directional selection of fish in this habitat, with only a small number of special species, such as *Carassius auratus* Linnaeus,*Leuciscus waleckii* , and *T. dalaica* , able to adapt to the highly salinized environment (J. Xu et al., 2013). As *T. dalaica*can thrive in the highly alkaline Lake Dali Nur, thus demonstrating their ability to adapt to high basicity tolerance and stress resistance, they are optimal candidates for investigating alkaline resistance mechanisms in fish.

In recent years, salt-alkali water development has intensified significantly in China. Due to their economic value, *C. auratus*and *L. waleckii* have always attracted greater attention than*T. dalaica* . In *L. waleckii* , a set of genomic regions, under selective sweeps harboring genes involved in ion homoeostasis, acid-base regulation, unfolded protein response, reactive oxygen species elimination, and urea excretion were detected (J. Xu et al., 2017).

Compared with other saline-tolerant or high-altitude acclimatized fish, *T. dalaica* reach sexual maturity at an early age, generally at two years of age. Therefore, for alkali-resistant gene inheritance and screening, the research cycle can be significantly shortened. Thirteen candidate genes, including *ηιφ-1aB* and *ηιφ-2aA,* have been reported as likely involved in the hypoxia response of *T. dalaica* , suggesting that genetic mechanisms of adaptation to high-altitude hypoxia could be resolved using RNA sequencing technology (Wang, Yang, Wu, Song, & He, 2015). *Triplophysa* fish are a strongly diverged group of fish encompassing 137 valid species, as recorded in FishBase (Froese & Pauly, 2014). These fish not only inhabit freshwater environments, but can also survive in saline and alkaline lakes, or at high altitudes, which is ideal for investigating local adaptation and ecological speciation. In recent years, the evolution of high-throughput sequencing techniques has provided new opportunities for elucidating the genetic basis of adaptation and speciation in *Triplophysa* fish. For example, the whole genome sequence of *T. tibetana* laid a solid foundation for further investigation into the environmental adaptation mechanisms of endemic fishes in the Tibetan Plateau (X. Yang et al., 2019). The genome assembly of *T. siluroides* provided genomic resources to better understand Tibetan loach biology, and set a stage for comparative analysis of classification, diversification, and adaptation of fishes in Cobitoidea (L. Yang et al., 2019). However, the molecular mechanisms underlying adaptation to high salt environments in populations of*T. dalaica* remain relatively poorly understood.

In this study, we present the chromosome-level genome of *T. dalaica,* which inhabits the extremely alkaline waters of Lake Dali Nur, using both PacBio long-read sequencing and Hi-C technology. The size of the resultant genome was approximately 607.91 Mb, with a contig N50 size of 9.27 Mb, and a final gene set consisting of 23,925 genes. To detect possible genes related to alkaline adaptation, the final gene set was compared with publicly available gene sets of *T. siluroides* and *T. tibetana* . Effective population size dynamics between the alkaline Lake Dali Nur and the freshwater region were also discussed.

MATERIALS AND METHODS

Sample collection and sequencing

A female *T. dalaica* (Figure 1A) was collected from Lake Dali Nur, in Inner Mongolia (43°22'43"N, 116°39'24"E, sampling site 1 of Figure 1B), and subjected to DNA sequencing. The fish was immediately dissected following treatment with the anesthetic, tricaine MS-222. Total genomic DNA was extracted from muscle tissue using the standard phenol/chloroform extraction method. A paired-end library, with an insert size of 400 bp, was constructed according to Illumina standard procedures (Illumina, San Diego, CA, USA). The library was sequenced on a HiSeq 2500 system, using the 150 bp PE mode. Extracted DNA was also used to construct two 20 kb libraries, according to PacBio manufacturing protocols (Pacific Biosciences, CA, USA). Libraries were then sequenced using one cell of the PacBio Sequel II sequencing platform. For comparison, another freshwater *T. dalaica,* from the Hai River in Henan province (35°54'28.0"N, 113°51'26.0"E, sampling site 2 of Figure 1B), was also subject to both Illumina library construction and sequencing using the HiSeq 2500 platform; resulting in the generation of millions of 150 bp paired-end reads.

To obtain as many expression evidence as possible for gene prediction, total RNA from eight tissues, including intestine, liver, brain, heart, muscle, gill, spleen, and ovary, of the aforementioned individual used for de novo sequencing, were extracted using a total RNA purification kit (Takara Bio). For each tissue, one RNA-sequencing library, with an insert size of 350 bp, was constructed and sequenced on the Illumina HiSeq 2500 platform, using the 150 bp PE mode.

K-mer analysis and evaluation of genome size

All paired-end reads generated in this study were cleaned using fastp (Chen, Zhou, Chen, & Gu, 2018), under default settings. About 50X of the estimated genome, totaling 37.5 Gb clean data were randomly selected from the whole genome sequencing data to estimate *T. dalaica*genome size, using k-mer analysis. The depth distribution of effective 17-mers was estimated using Jellyfish version 2.2.10 (Marcais & Kingsford, 2011), with depth of the main peak selected as Kdepth, and genome size estimated using the following formula$Genome_Size = Knum/Kdept$. Genome heterozygosity was estimated using GenomeScope 2.0 (Vurture et al., 2017).

Assembly and correction of the Genome

De novo assembly was performed on the basis of PacBio long reads using FALCON v0.3.0 (https://github.com/PacificBiosciences/falcon). Error correction and pre-assembly were initially performed following the FALCON pipeline, with a length_cutoff of 28 kb and a length_cutoff_pr of 27 kb chosen for the assembly step. The draft assembly was corrected using Arrow, based on the mapping results of the PacBio reads to the assembled genome when following the BLASR pipeline (Chaisson & Tesler, 2012). Next, all filtered Illumina reads were mapped to the corrected contigs using BWA-mem (H. Li & Durbin, 2009), to polish the genome for 3-rounds of Pilon iteratively (Walker et al., 2014).

Scaffolding of the genome using Hi-C technology

The Hi-C technique has been widely applied to construct chromosome-level genome assemblies. By capturing all DNA interaction patterns in chromatin, high-throughput sequencing technology, combined with bioinformatic analyses, allows whole chromosome DNA relationships, in spatial position within the entire genome, to be studied (Lieberman-Aiden et al., 2009). Muscle tissue, from the aforementioned *T. dalaica,* was used for Hi-C library construction, with the library sequenced on the Illumina HiSeq X Ten platform, using the

150 bp PE mode. Raw reads were cleaned using fastp (Chen et al., 2018), and then mapped to the polished genome using Bowtie 2 (version 2.2.5) (Langmead & Salzberg, 2012). After removing low-quality alignments and duplicate reads, an inter/intra chromosomal contact map was constructed and used to anchor and orientate contigs to the chromosomes, using an agglomerative hierarchical clustering method implemented in Lachesis (Burton et al., 2013). Gaps between adjacent contigs were filled using 100 Ns.

Annotations of the genome

Repeat sequences in the genome comprised of simple sequence repeats (SSRs), moderately repetitive sequences, and highly repetitive sequences. The MISA tool (Thiel, Michalek, Varshney, & Graner, 2003) was used to search for SSRs in the *T. dalaica* genome, with default parameters applied. Tandem repeats were identified using Tandem Repeats Finder 4.07b (Benson, 1999). RepeatMasker (Tarailo-Graovac & Chen, 2009) was used to identify known transposable elements (TEs) present in the *T. dalaica* genome, with Repbase (v. 22.11) as the query (Bao, Kojima, & Kohany, 2015). RepeatModeler v.1.0.11 and LTR_finder were also used to identify possible transposable elements, de novo, with default settings applied (Tarailo-Graovac & Chen, 2009; Z. Xu & Wang, 2007). Next, all the TEs were identified using RepeatMasker (Tarailo-Graovac & Chen, 2009).

Homology-based ncRNA annotation was performed by scanning the covariance models of rRNA, miRNA, and snRNA genes deposited in the Rfam database (release 13.0) (Kalvari et al., 2018), with candidate regions residing in the *T. dalaica* genome, preliminary detected using BLASTN (E-value [?] 1e-5) (Camacho et al., 2009). The tRNAscan-SE (v1.3.1) search server (Lowe & Eddy, 1997) and the RNAmmer v1.2 server (Lagesen et al., 2007) were also used to predict tRNAs and rRNAs, respectively, with default settings applied.

De novo, together with homology- and transcriptome-based strategies were used to predict possible protein coding genes. For de novo prediction, a variety of software, including Augustus v3.3 (Stanke, Steinkamp, Waack, & Morgenstern, 2004), GeneID v1.4.4 (Blanco, Parra, & Guigo, 2007), GlimmerHMM (Majoros, Pertea, & Salzberg, 2004), and SNAP (Korf, 2004), was used. For homology-based prediction, the proteome of each of the six species, including *Astyanax mexicanus* , *Danio rerio* ,*Ictalurus punctatus* , *Takifugu rubripes* , *Triplophysa siluroides* , and *Xiphophorus maculatus,* was mapped onto the*T. dalaica* genome, and each of the protein sequences compared against the best aligned regions; with possible coding regions predicted using GeneWise v2.2.0 (Birney & Durbin, 2000). For transcriptome-based prediction, RNA-seq reads were mapped to the genome using TopHat (Trapnell et al., 2012), and subsequently assembled into gene models (Cufflinks-set) using Cufflinks (Roberts, Trapnell, Donaghey, Rinn, & Pachter, 2011). The Cufflinks-set was then fed into PASA (Haas et al., 2003) to identify the donor and acceptor sites of possible exon regions; with resultant coding regions predicted using TransDecoder (https://github.com/TransDecoder/TransDecoder/wiki). Finally, to generate a consensus gene set, EVidenceModeler (EVM) v1.1.1 (Haas et al., 2008) was used to integrate all predicted gene models using de novo, homology-, and transcriptome-based strategies. Low-quality genes, fewer than 50 encoded amino acids and/or harboring premature termination or frameshifts, were also removed from the gene set.

Functional annotations of *T. dalaica* predicted genes were performed by searching the nr, KOG, Uniprot (release 2018_10), and KEGG (release 84.0) databases, using Blast with an E-value of 1e-5 (Camacho et al., 2009). Descriptions and KEGG pathways were then extracted from the best hit sequence. Next, InterProScan (Quevillon et al., 2005) was used to annotate predicted genes based on the InterPro database (5.21-60.0), with GO terms assigned according to the best hits.

Phylogenetic analysis and divergence time estimation

To ascertain the evolutionary position of *T. dalaica* , gene models of its genome were compared with that of six other relative fish in Cypriniformes. These included two in Cobitidae (*T. siluroides* and*T. tibetana*), and four in Cyprinidae (*Anabarilius grahami* ,*D. rerio* , *Danionella translucida* , and *Megalobrama amblycephala*). Gene models of these species were downloaded from the NCBI Database and then were used to identify potential orthologous gene families following the OrthoMCL v2.0.9 pipeline (L. Li, Stoeckert, & Roos, 2003) under default settings.

Single copy orthologous genes were used to perform both molecular phylogenetic analysis and subsequent divergence time estimations. Briefly, deduced protein sequences were aligned using MUSCLE (Edgar, 2004), where they were transformed back into nucleotide sequences; with highly variable regions filtered using Gblocks v0.91b (Talavera & Castresana, 2007). Alignments were then concatenated and fed into RAxML v8.2.11 (Stamatakis, 2014), to perform phylogenetic analysis using the GTRGAMMA model. To assess topological robustness, 100 bootstrap replicates were performed. Divergence time estimations, for *T. dalaica* and relative fishes, were implemented using MCMCTree, included in PAML package v4.9e (Yang, 2007) with known divergence times downloaded from Timetree (http://www.timetree.org/) set as calibration points. The MCMCTree parameters were set as follows: clock = 2, RootAge [?] 1.73, model = 7, BDparas = 110, kappa_gamma = 62, alpha_gamma = 11, rgene_gamma = 23.18, and sigma2_gamma = 14.5.

Gene family dynamics and positively selected T. dalaica candidate genes

Based on divergence times estimated for *T. dalaica* and relative species, as well as gene families identified using OrthoMCL, the possible expansion and contraction of gene families residing in these genomes were detected using Computational Analysis of gene Family Evolution, v4.0.1 (CAFE) (De Bie, Cristianini, Demuth, & Hahn, 2006).

To precisely detect positively selected genes, specifically those likely to be related to alkaline acclimation, a new set of orthologous gene groups were identified. A reciprocal best hits strategy, based on gene model comparison between three public *Triplophysa* species generated using BLAST v2.2.30, with an E-value cutoff [?] 1e-05, was used. Multiple sequence alignments were performed using GUIDANCE2 (Sela, Ashkenazy, Katoh, & Pupko, 2015) for all ortholog groups, with parameters set as follows: seqType = codon, seqCutoff = 0.3, and msaProgram = muscle. Next, codeml, included in PAML v4.9e (Yang, 2007), was used to estimate dN/dS ratios (ω), to deduce selection pressures leading to the current evolution of *T. dalaica* . To achieve this, branch-site models (model = 2 and NSsite = 2) were used. For the null hypothesis, parameters 'fix_omega' and 'omega' were each set to 1, while for the alternative hypothesis, 'fix_omega' and 'omega' were set to 0 and 1.5, respectively. To check convergence, analyses were performed twice for each ortholog group, with final p -values obtained through comparison of both the chi2 distribution and twice the LRT values, between the two models.

Demographic history

Based on short reads generated for whole genome sequencing, using next-generation technology, demographic histories for two *T. dalaica* were deduced following the PSMC pipeline (H. Li & Durbin, 2011), and subsequently compared. Briefly, after cleaning and filtering using fastp (Chen et al., 2018), all short reads were mapped to the de novo assembly using BWA-mem, with default settings applied (H. Li & Durbin, 2009). Based on both mapping results and the assembly, SAMtools (H. Li et al., 2009), followed by BCFtools (Danecek & McCarthy, 2017), were used to generate a consensus sequence of the heterozygous genome. This was then used to deduce the demographic history using PSMC, with parameters set as follows: -N25 -t15 -r5 -p "4+25*2+4+6". A previously reported cyprinid mutation rate of 4.13e-9 (Fu, Chen, Zou, Long, & He, 2010), and a 2-year generation time, were used to infer the effective population size (EPS) of *T. dalaica* .

RESULTS AND DISCUSSION

De novo assembly of the *T. dalaica* genome

All next-generation data generated in this report are summarized in Table S1. To estimate sample genome size and heterozygosity, a total of 37.5 Gb short reads, covering approximately 50-fold of the estimated genome, were selected from the generated NGS genome data. These reads were then subject to 17-mer analysis. As the main peak was located at a depth of 53 (Table S2), the genome size was estimated to be 631 Mb, which was similar to the previously reported genome sizes of other *Triplophysa* species, such as *T. siluroides* and *T. tibetana,* which were estimated to be 638.07 Mb (L. Yang et al., 2019) and 652 Mb (X. Yang et al., 2019), respectively. Moreover, the estimated species heterozygosity rate was approximately 0.375%.

Data used for the de novo assembly of the *T. dalaica* genome were generated using PacBio Sequel II.

Following adaptor and low-quality read removal, nearly 200X, totaling 126.5 Gb, subreads remained, with an average read length and N50 length of 19.7 kb and 32.4 kb, respectively. Preliminary genome assembly was performed in FALCON v0.3, with ultimate genome assembly, approximately 607.91 Mb with a contig N50 size of 9.27 Mb, obtained after several polishing rounds using Arrow and Pilon. Contig N50 lengths of the formerly reported *T. siluroides* and *T. tibetana* were 2.87 Mb and 3.1 Mb, respectively (L. Yang et al., 2019; X. Yang et al., 2019). Thus, the contig N50 length of *T. dalaica* was much longer than that formerly reported. This may be because we used the newest sequel II technology available; thus, throughput and subread N50 length were significantly improved from previous platforms.

The quality of the assembled genome was thoroughly scrutinized. GC analysis was conducted to assess potential contamination before sequencing. As a result, a unimodal distribution of GC content was detected, with an average GC content of 38.77% for the assembly; suggesting no bacterial contamination. To evaluate coverage of the assembly, all RNA-seq reads were mapped to the *T. dalaica* genome using HISAT2 (Kim, Langmead, & Salzberg, 2015), with default parameters applied. The percentage of aligned reads ranged from 84.78% to 91.08% (Table S3). Moreover, Benchmarking Universal Single-Copy Orthologs (BUSCO) (Simao, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) were used to estimate the coverage of the 4,584 single-copy genes conserved among all Actinopterygii, with approximately 93.7% of the complete BUSCOs found in the assembly (Table S4). Taken together, these results suggest that the genome assembly was robust and nearly complete.

Scaffolding of the *T. dalaica* genome

It is thought that interactions occur more frequently between closer locations on the chromosome, than farther. Based on next-generation sequencing technology, a total of 106 Gb of data, covering nearly 200X of the estimated genome size, were generated for the Hi-C library. Based on interaction relationships detected using Hi-C technology, 141 contigs, covering approximately 96.3% of the whole genome length, were anchored and orientated onto 25 chromosomes, resulting in a scaffold N50 length of 23.6 Mb. The interaction relationships along each chromosome are shown in Figure 1. Notably, each chromosome had less than 10 introduced gaps, except chromosome 5. Moreover, our assembly captured long stretches of telomeric sequence (TAACCC/TTAGGG)n at all ends of 14 chromosomes, and at one single end for eight chromosomes (Table 1). To assess the accuracy of the scaffolding results, based on Hi-C technology, gene collinearity between *T. dalaica* and its two relative species, *D. rerio* and *T. tibetana* , were compared, with results indicating nearly perfect collinearity (Figure 2). Thus taken together, the high anchoring rate, the minimal gap numbers, the presence of telomeric sequences for most chromosomes, and the nearly perfect collinearity, all indicate the high level of quality of the present *T. dalaica* genome assembly.

Genome annotations

Generally, annotations of a newly assembled genome include repeat, gene model, and gene function annotations. For repeat annotations, a total of 197,396 SSRs were obtained using MISA. Combined homology and de novo based results showed that repeat sequences accounted for 35.01% of the *T. dalaica* genome assembly (Table 2), of which, DNA transposons made up the greatest proportion (16.01%), followed by LRT (8.9%) and LINEs (4.24%).

The final set of protein-coding genes was obtained by integrating the results of ab initio, homologue based, RNA-seq based predictions. This set consisted of 23,925 genes, with average gene length, average CDS length, and number of exons, per gene, 12,128.58 bp, 1,715.77 bp and 9.9, respectively. Distribution of these parameters was similar among *T. dalaica* and the species used for annotation (Figure S1), suggesting both gene conservation and annotation robustness. On the contrary, the homology-based ncRNA annotation showed a total of 1,664 miRNAs, 11,504 tRNAs, 684 rRNAs, and 1,207snRNAs residing in the genome.

Functional annotations, including function descriptions, KEGG pathways, and GO term assignments, as well as database summaries, are shown in Table S5. In total, 23,594, 98.62% of the total 23,925 genes, could be annotated as having potential functions.

Phylogenetic analysis and divergence time estimation

The phylogenetic position of *T. dalaica,* within the order Cypriniformes, and its estimated time of divergence were deduced based on single copy ortholog genes from related species. Gene family clustering showed that a total of 23,925 genes in *T. dalaica*could be divided into 19,271 gene families, of which 116 were unique to*T. dalaica* and 4,126 were single copy orthologs across all related species. Based on both the supermatrix constructed using these single copy genes and the maximum likelihood method, a phylogenetic tree of these species was reconstructed with high confidence (Figure 3). It showed that *T. dalaica* is closer to *T. tibetana,* with*T. siluroides* identified as the basal species within the*Triplophysa* species. Divergence time estimation showed that*T. dalaica* and *T. tibetana* diverged approximately 8.31 million years ago (Mya), with the history of *Triplophysa* fish likely to be at least 14.6 Mya (Figure 4).

Gene family dynamics in the genome

Significant expansion and contraction of gene families may denote environmental adaptation of the species. Based on both divergence time estimation and OrthoMCL gene family clustering, CAFE was used to detect dynamics of gene families residing in the genome. The analysis detected 431 expanded gene families, as well as 491 contracted gene families, in the *T. dalaica* genome. The top three enriched pathways for expanded family genes were "Gap junction", "Relaxin signaling pathway", and "Bile secretion", while for contracted family genes, the top three were "Estrogen signaling pathway", "Dopaminergic synapse", and "Parathyroid hormone synthesis, secretion, and action" (Table S6).

Positively selected genes in the genome

In addition to significant gene family dynamics, genes subject to positive selection also denote environmental adaptation. When comparing with the other two *Triplophysa* species, genes with evidences of positive selection were identified. As a result, a set of 898 candidate genes, likely subject to positive selection, were identified (Table S7).

Salinity change is a key driving force for the adaptation of the fishes endemic to Lake Dali Nur, with ion channels and aquaporins thought to be pivotal players in salinity stress conditions. Two piezo-type mechanosensitive genes (FAM38: Tdalaica.Chr01.107; Tdalaica.Chr08.21) associated with ion channels, and the enriched calcium signaling pathway (KEGG: map04020, Table S8), associated with positively selected genes, may partially explain the adaptation of *T. dalaica* to extreme alkaline environments. Paracellular transportation is another mechanism key to the maintenance of both ion homeostasis and acid-base balance. KEGG analyses suggest that positively selected genes participate in a number of ways to paracellular permeability pathways, including ECM-receptor (KEGG: map04512), cell adhesion molecules (CAMs) (KEGG: map04514), vascular endothelial growth factor (VEGF) (KEGG: map04370), and focal adhesion (KEGG: map04510), each of which was significantly enriched (Table S8).

Previous studies have reported that both carbonic anhydrase (CA) and sodium bicarbonate cotransport carriers (Na^+/ HCO_3^- cotransporter, Solute Carrier4A4, *SLC4A4*) can mediate HCO_3^- transport; an important mechanism through which fish tolerate high alkali environments and export HCO_3^- from the body (Purkerson & Schwartz, 2007; Romero, Fulton, & Boron, 2004). CA is a zinc-containing metal enzyme predominantly involved in various biological osmotic processes, including permeability regulation, ion regulation, and acid-base regulation, such as CO_2rehydration: $CO_2+H_2O = H_2CO_3 = H^+ + HCO_3^-$. A previous study conducted a similar investigation in *Oncorhynchus mykiss,* finding that expression levels of the CA gene were up-regulated under saline-alkali stress; indicating an important relationship between the CA gene and ion regulation (Goss, Wood, Laurent, & Perry, 1994). We found that two CA genes (Tdalaica.Chr17.484: *CAV* , Tdalaica.Chr24.506: *CAXVI*), and one *SLC4A4* gene (Tdalaica.Chr04.806:*SLC4A4*) were in a positive selection cluster (Table S7). The natriuretic peptide (NP) system is a key endocrine system involved in osmoregulation and ion homoeostasis in vertebrates, while atrial natriuretic peptide (ANP) has been confirmed as the primary sodium regulating hormone in eels (Cao et al., 2009). Two ANP genes (Tdalaica.Chr18.955, Tdalaica.Chr02.422) are in a positive selection cluster (Table S7), which may play a pivotal role in the response to regulate sodium in *T. dalaica.*

Hypoxia inducible factor 1 alpha (HIF-1a) acts as a key transcriptional activation factor in the hypoxic response regulation mechanism, which enhances the body's tolerance to low oxygen by controlling the expression of genes related to physiological processes, such as glucose transport and sugar leaven (Haase, 2013). HIF-1 signaling pathway (KEGG: map04066) was significantly enriched in positively selected genes, suggesting it may play a pivotal role in the response to high altitude hypoxia (Table S8).

In a previous study, growth hormone (GH) and insulin-like growth factor I (IGFI) were involved in the regulation of ion homeostasis and salinity acclimation (Cao et al., 2009), with fibroblast growth factors (FGFS) behaving as important modulators of paracellular permeability regulation. We also found one IGFI gene, and three FGFS genes, subject to positive selection, providing further evidence of the involvement of these hormones in the regulation of ion homeostasis and salinity acclimation (Table S7).

Demographic history

Demographic history deduced from the genomic data is shown in Figure 5. It shows that the two EPSs are similar by more than 1 Mya. The EPS of the Hai River population, endemic to freshwater, is almost constant until 20,000 years ago. Since then, the EPS has expanded in a short time. This corresponds to the onset of Northern Hemisphere deglaciation, approximately 19,000-20,000 years ago (Clark et al., 2009). However, the EPS of the population has been slowly expanding since approximately 1 Mya, reaching nearly 400,000 at approximately 0.4 Mya. Since then, the EPS has gradually declined to an extremely low level. According to the geological research of Gao (1988), Lake Dali Nur may have begun developing approximately 1 Mya, with the whole area providing plenty of shallow water for the *T. dalaica* ; thus, its EPS gradually expanded. However, along with lake enlargement, *T. dalaica*survival areas were gradually drowned by deep waters, possibly resulting in decreased EPS. This process of lake expansion may have concluded approximately 10,000 years ago, at which point the Lake begun to shrink, and EPS increased proportionally.

Conclusion

Through application of both long-read sequencing and Hi-C technology, we obtained a whole genome sequence assembly of an individual *T. dalaica* endemic to the alkaline Lake Dali Nur, located in Inner Mongolia, northeast China. The contig N50 length of the assembly exceeded 9 Mb, with nearly all contigs anchored to chromosomes; thus, the assembly was at the chromosome level. Importantly, telomeres were obtained for majority of the chromosomes. A number of assessments, including mapping of the RNA-Seq short reads, BUSCO appraisal, and collinearity with *D. rerio* and *T. tibetana,* all suggest high assembly accuracy. Through gene model comparisons of *T. dalaica* with other fish belonging to Cypriniformes, we found that *T. dalaica* was closer to *T. tibetana* than *T. siluroides* , with divergence between *T. dalaica* and *T. tibetana* occurring more than 8 Mya. Positive selection analyses identified a number of pivotal genes likely related to alkaline adaptation of the species. Demographic analyses suggested that the population of *T. dalaica* , endemic to Lake Dali Nur, might have diverged approximately 1 Mya from the Hai River population, with different EPS fluctuations likely resulting from different environmental factors.

AUTHOR CONTRIBUTIONS

Z.C.J. and N.G.X. conceived and designed the research. L.R.Y., Y.X.M., D.J., H.B., W.X. and M.W.W. performed the experiments. Z.C.J., T.Y.T., Y.C.X. and W.X.F analyzed and interpreted the assembly and annotations. Z.C.J., H.B. and N.G.X. wrote the manuscript with input from all authors.

ORCID Chuanjiang Zhou https://orcid.org/0000-0002-6433-737X

DATA AVAILABILITY STATEMENT

The DNA sequencing data have been deposited into the NCBI Sequence Read Archive under BioProject: PRJNA624716.

FUNDING INFORMATION

This work was financially supported by the following funding: the National Natural Science Foundation

of China (31401964, 31872199). The Science and Technology Innovation Team supported the project in Henan province, China (CXTD2016043). Projects from Henan Province, Department of Science and Technology (182102110046, 182102110237, 182102110007), and Training program for young excellent teachers in Universities of Henan province in 2019, China (2019GGJS063), also supported this study.

ACKNOWLEDGEMENTS

We appreciate the help from Xupan and Dr. Zouming, Diggers Biotechnology Co, Ltd, Wuhan, who kindly provided help with drawing Figures. Study support was also provided by The High Performance Computing Center of Henan Normal University.

TABLE LEGENDS

Table 1 Summary of the *T. dalaica* genome assembly.

Table 2 Summary of repeat annotations based on different strategies.

Table S1 Summary of clean reads generated for the whole genome and transcriptome, based on Illumina Hiseq technology.

Table S2 Estimation of genome size based on 17-mer statistics.

Table S3 Summary of the transcriptome mapping rate to the de novo genome assembly.

Table S4 BUSCO assessments of the de novo genome assembly.

Table S5 Summary of functional annotations for all gene models predicted for the genome.

Table S6 KEGG pathway enrichment analyses of genes families subject to expansion and contraction in the genome.

Table S7 Summary of genes subject to positive selection residing in the genome.

Table S8 KEGG pathway enrichment analyses of genes subject to positive selection residing in the genome.

FIGURE LEGENDS

Figure 1 Sample information of *T. dalaica* individuals used in this study.

A Image of the *T. dalaica* specimen

B Sampling sites of the two individuals

Figure 2 Interaction relationships between chromosome regions across the genome. The color bar illuminates contact density from red (high) to white (low).

Figure 3 Gene model collinearity between *T. dalaica* and the other two relative fishes, *D. rerio* and *T. tibetana* .

Figure 4 Phylogenetic relations and divergence times among *T. dalaica* and relative species. The phylogeny was recovered using the maximum likelihood method, and all nodes were fully supported using 100 bootstrap replicates.

Figure 5 PSMC plot depicting the demographic history of the *T. dalaica* individuals collected from Lake Dali Nur (Dali, alkaline water) and the Hai River (Hai, freshwater), using genomic data.

Figure S1 Distribution of gene length, CDS, and the numbers of exons and introns between *T. dalaica* and relative species.

Reference

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA, 6* , 11. doi:10.1186/s13100-015-0041-9

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res, 27* (2), 573-580. doi:10.1093/nar/27.2.573

Birney, E., & Durbin, R. (2000). Using GeneWise in the Drosophila annotation experiment. *Genome Res, 10* (4), 547-548. doi:10.1101/gr.10.4.547

Blanco, E., Parra, G., & Guigo, R. (2007). Using geneid to identify genes. *Curr Protoc Bioinformatics, Chapter 4* , Unit 4 3. doi:10.1002/0471250953.bi0403s18

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., & Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol, 31* (12), 1119-1125. doi:10.1038/nbt.2727

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics, 10* , 421. doi:10.1186/1471-2105-10-421

Cao, Y. B., Chen, X. Q., Wang, S., Chen, X. C., Wang, Y. X., Chang, J. P., & Du, J. Z. (2009). Growth hormone and insulin-like growth factor of naked carp (Gymnocypris przewalskii) in Lake Qinghai: expression in different water environments. *Gen Comp Endocrinol, 161* (3), 400-406. doi:10.1016/j.ygcen.2009.02.005

Chaisson, M. J., & Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics, 13* , 238. doi:10.1186/1471-2105-13-238

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics, 34* (17), i884-i890. doi:10.1093/bioinformatics/bty560

Clark, P. U., Dyke, A. S., Shakun, J. D., Carlson, A. E., Clark, J., Wohlfarth, B., . . . McCabe, A. M. (2009). The Last Glacial Maximum.*Science, 325* (5941), 710-714. doi:10.1126/science.1172873

Danecek, P., & McCarthy, S. A. (2017). BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics, 33* (13), 2037-2039. doi:10.1093/bioinformatics/btx100

De Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution.*Bioinformatics, 22* (10), 1269-1271. doi:10.1093/bioinformatics/btl097

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res, 32* (5), 1792-1797. doi:10.1093/nar/gkh340

Fu, B., Chen, M., Zou, M., Long, M., & He, S. (2010). The rapid generation of chimerical genes expanding protein diversity in zebrafish.*BMC Genomics, 11* , 657. doi:10.1186/1471-2164-11-657

Gao, Z. (1988). Changes of the Dali Nur Lake. *Geographical Research, 7* (4).

Goss, G. G., Wood, C. M., Laurent, P., & Perry, S. F. (1994). Morphological responses of the rainbow trout (Oncorhynchus mykiss) gill to hyperoxia, base (NaHCO3) and acid (HCl) infusions. *Fish Physiol Biochem, 12* (6), 465-477. doi:10.1007/BF00004449

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Jr., Hannick, L. I., . . . White, O. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies.*Nucleic Acids Res, 31* (19), 5654-5666. doi:10.1093/nar/gkg770

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., . . . Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol, 9* (1), R7. doi:10.1186/gb-2008-9-1-r7

Haase, V. H. (2013). Regulation of erythropoiesis by hypoxia-inducible factors. *Blood Rev, 27* (1), 41-53. doi:10.1016/j.blre.2012.12.003

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., . . . Petrov, A. I. (2018). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res, 46* (D1), D335-D342. doi:10.1093/nar/gkx1038

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat Methods, 12* (4), 357-360. doi:10.1038/nmeth.3317

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics, 5* , 59. doi:10.1186/1471-2105-5-59

Lagesen, K., Hallin, P., Rodland, E. A., Staerfeldt, H. H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res, 35* (9), 3100-3108. doi:10.1093/nar/gkm160

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods, 9* (4), 357-359. doi:10.1038/nmeth.1923

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25* (14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature, 475* (7357), 493-496. doi:10.1038/nature10231

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25* (16), 2078-2079. doi:10.1093/bioinformatics/btp352

Li, L., Stoeckert, C. J., Jr., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res, 13* (9), 2178-2189. doi:10.1101/gr.1224503

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., . . . Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science, 326* (5950), 289-293. doi:10.1126/science.1181369

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res, 25* (5), 955-964. doi:10.1093/nar/25.5.955

Majoros, W. H., Pertea, M., & Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders.*Bioinformatics, 20* (16), 2878-2879. doi:10.1093/bioinformatics/bth315

Marcais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.*Bioinformatics, 27* (6), 764-770. doi:10.1093/bioinformatics/btr011

Purkerson, J. M., & Schwartz, G. J. (2007). The role of carbonic anhydrases in renal physiology. *Kidney Int, 71* (2), 103-115. doi:10.1038/sj.ki.5002020

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., & Lopez, R. (2005). InterProScan: protein domains identifier. *Nucleic Acids Res, 33* (Web Server issue), W116-120. doi:10.1093/nar/gki442

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., & Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol, 12* (3), R22. doi:10.1186/gb-2011-12-3-r22

Romero, M. F., Fulton, C. M., & Boron, W. F. (2004). The SLC4 family of HCO 3 - transporters. *Pflugers Arch, 447* (5), 495-509. doi:10.1007/s00424-003-1180-2

Sela, I., Ashkenazy, H., Katoh, K., & Pupko, T. (2015). GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res, 43* (W1), W7-14. doi:10.1093/nar/gkv318

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics, 31* (19), 3210-3212. doi:10.1093/bioinformatics/btv351

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30* (9), 1312-1313. doi:10.1093/bioinformatics/btu033

Stanke, M., Steinkamp, R., Waack, S., & Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res, 32* (Web Server issue), W309-312. doi:10.1093/nar/gkh379

Talavera, G., & Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol, 56* (4), 564-577. doi:10.1080/10635150701472164

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics, Chapter 4* , Unit 4 10. doi:10.1002/0471250953.bi0410s25

Thiel, T., Michalek, W., Varshney, R. K., & Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theor Appl Genet, 106* (3), 411-422. doi:10.1007/s00122-002-1031-0

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., . . . Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc, 7* (3), 562-578. doi:10.1038/nprot.2012.016

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics, 33* (14), 2202-2204. doi:10.1093/bioinformatics/btx153

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., . . . Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One, 9* (11), e112963. doi:10.1371/journal.pone.0112963

Wang, Y., Yang, L., Wu, B., Song, Z., & He, S. (2015). Transcriptome analysis of the plateau fish (Triplophysa dalaica): Implications for adaptation to hypoxia in fishes. *Gene, 565* (2), 211-220. doi:10.1016/j.gene.2015.04.023

Xiao, J., Si, B., Zhai, D., Itoh, S., & Lomtatidze, Z. (2008). Hydrology of Dali Lake in central-eastern Inner Mongolia and Holocene East Asian monsoon variability. *Journal of Paleolimnology, 40* (1), 519-528. doi:10.1007/s10933-007-9179-x

Xu, J., Ji, P., Wang, B., Zhao, L., Wang, J., Zhao, Z., . . . Sun, X. (2013). Transcriptome sequencing and analysis of wild Amur Ide (Leuciscus waleckii) inhabiting an extreme alkaline-saline lake reveals insights into stress adaptation. *PLoS One, 8* (4), e59703. doi:10.1371/journal.pone.0059703

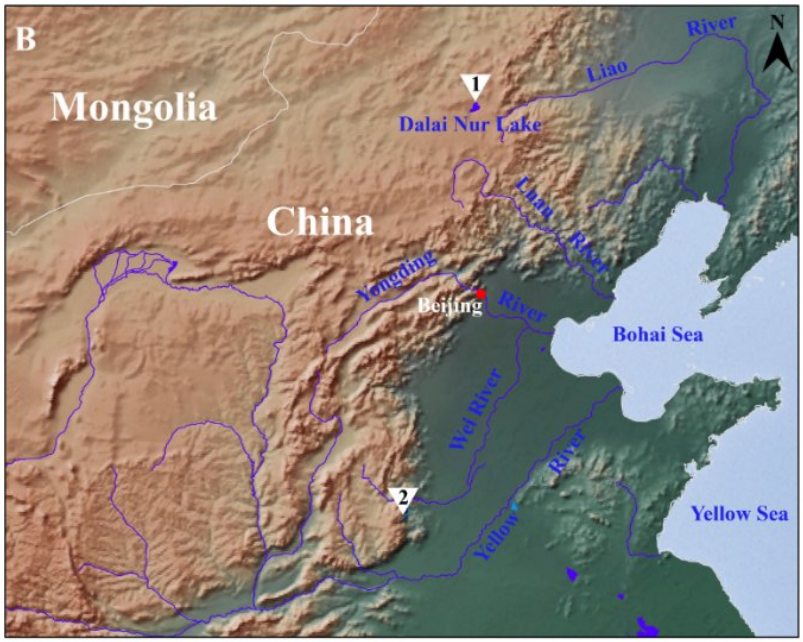
Xu, J., Li, J. T., Jiang, Y., Peng, W., Yao, Z., Chen, B., . . . Xu, P. (2017). Genomic Basis of Adaptive Evolution: The Survival of Amur Ide (Leuciscus waleckii) in an Extremely Alkaline Environment. *Mol Biol Evol, 34* (1), 145-159. doi:10.1093/molbev/msw230

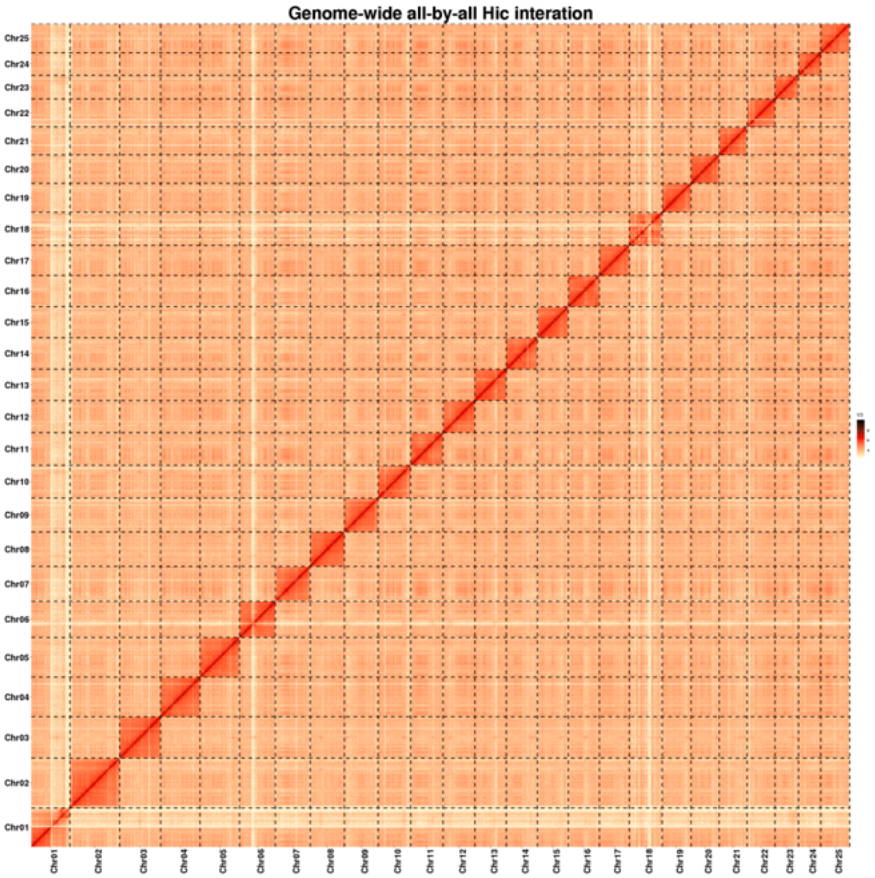
Xu, Z., & Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res, 35* (Web Server issue), W265-268. doi:10.1093/nar/gkm286

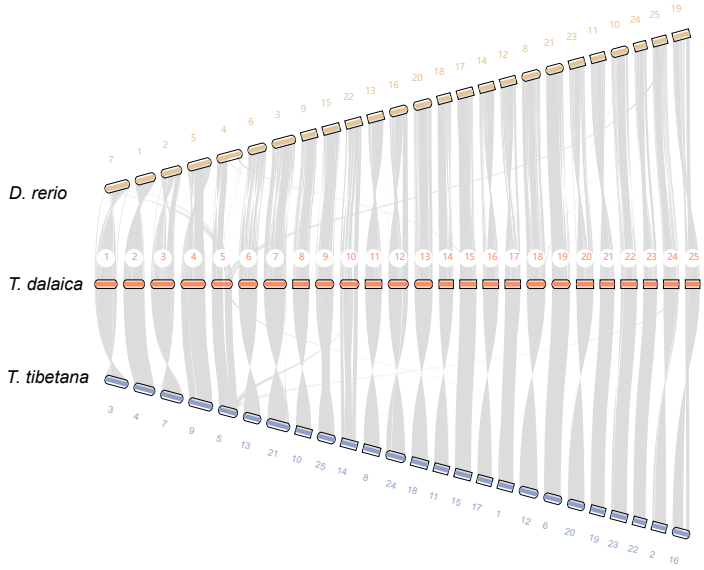
Yang, L., Wang, Y., Wang, T., Duan, S., Dong, Y., Zhang, Y., & He, S. (2019). A Chromosome-Scale Reference Assembly of a Tibetan Loach, Triplophysa siluroides. *Front Genet, 10* , 991. doi:10.3389/fgene.2019.00991

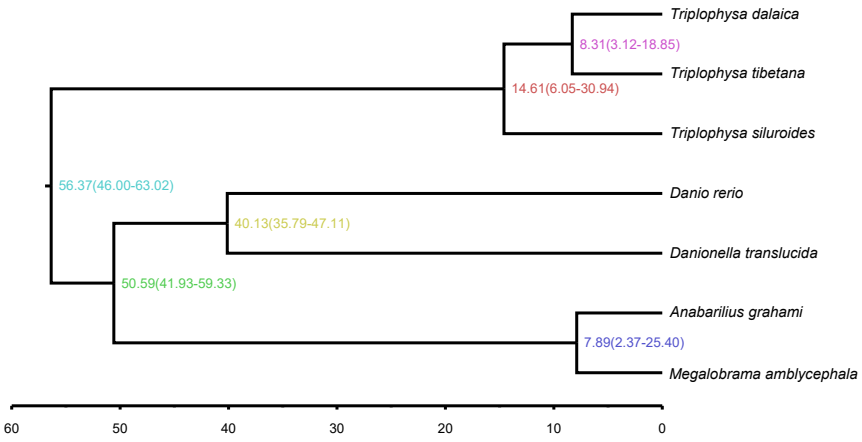
Yang, X., Liu, H., Ma, Z., Zou, Y., Zou, M., Mao, Y., . . . Yang, R. (2019). Chromosome-level genome assembly of Triplophysa tibetana, a fish adapted to the harsh high-altitude environment of the Tibetan Plateau.*Mol Ecol Resour, 19* (4), 1027-1036. doi:10.1111/1755-0998.13021

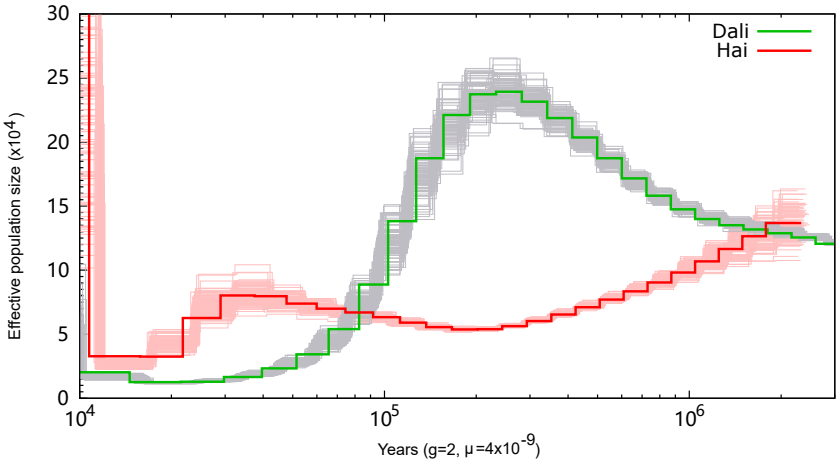
Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol, 24* (8), 1586-1591. doi:10.1093/molbev/msm088











Hosted file

Table 1.docx available at https://authorea.com/users/343092/articles/469773-the-chromosome-level-genome-of-triplophysa-dalaica-cypriniformes-cobitidae-provides-insights-into-its-survival-in-extremely-alkaline-environment

Hosted file

Table 2.docx available at https://authorea.com/users/343092/articles/469773-the-chromosome-level-genome-of-triplophysa-dalaica-cypriniformes-cobitidae-provides-insights-into-its-survival-in-extremely-alkaline-environment