De novo assembly of a chromosome-level genome of naked carp (Gymnocypris przewalskii) reveals geographic isolation of Schizothoracine fishes in Qinghai-tibet plateau lift

Liu Yimeng[1], Zongli Yao[1], zongyi sun[2], Hongfang Qi[3], Kai Zhou[1], Jianquan Shi[4], Pengcheng Gao[1], Zhen Sun[5], and Qifang Lai[1]

[1]East China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences
[2]Nextomics Biosciences Institute
[3]The Rescue and Rehabilitation Center of Naked Carps of Qinghai Lake
[4]The Rescue and Rehabilitation Center of Naked Carps of Qinghai Lake, Qinghai Key Laboratory of Qinghai-Lake Naked Carps Breeding and Conservation
[5] East China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences

August 28, 2020

Abstract

In this study, we present a de novo genome assembly of Gymnocypris przewalski using long-read PacBio sequencing technology with Genome-wide high throughput chromosome conformation capture correction. The assembly resulted in a high sequence contiguity and accuracy with 23 chromosomes and a genome size of 945 Mb. This work is the first chromosome-level genome study of Schizothoracine fishes from the Qinghai-tibet plateau (QTP). Phylogenetic analysis showed that the species differentiation events between G. przewalskii and Cyprinus carpio occurred at 24.06 Mya ago, which reveals geographic isolation of Schizothoracine fishes in QTP lift; the unique gene analysis found that most of these genes enriched in membrane transport and immune defensive pathways such as ABC transporters, bile secretion and cell adhesion molecules. Moreover, this genome resource is a key to understanding evolutionary adaptation to high altitudes, disentangling complex evolutionary histories, and will be useful for research on species diversity and resource conservation of Schizothoracine fishes on the QTP.

Introduction

Qinghai-Tibet Plateau (QTP), formed by the collision of the Indian Plate with the Eurasian Plate beginning in Eocene ˜50 million years ago (Mya), is often deemed as the Roof of the world or the Third Pole due to its average altitude of over 4, 000 meters and massive glaciers (Harrison, et al. 1992; Molnar, et al. 1993; Yin and Harrison 2000; Tapponnier, et al. 2001; Molnar, et al. 2010). High absolute elevation, high radiation, severe cold, and hypoxia are the main characteristics of extreme environmental conditions on QTP, which greatly threaten the survival of the local creatures, especially the wild plateau fish (Li 1991). Originated from a common ancestor Barbinae, Schizothoracine fishes (Teleostei: Cyprinidae), which distributed extensively in the high-altitude streams, rivers and lakes scattered throughout QTP (Zan, et al. 1985; Yue 2000; Li, et al. 2013), evolve gradually and become highly adapted to life in high-altitude environmental stress, are spectacularly diverse in ploidy, physiological plasticity, and morphological innovations, such as numbers of scales, pharyngeal teeth and barbels (Zhou, et al. 2020). Polyploidy can also be commonly observed in Schizothoracine fishes and is regarded as a successful evolutionary transition of organisms to these extreme environments (Zan, et al. 1985; Leggatt and Iwama 2003; Wang, et al. 2016).

Gymnocypris przewalskii (Kessler 1876, NCBI: txid 369649, Fishbase ID: 55381, Figure 1), otherwise known as the Przewalski's naked carp or scale-less carp is an endemic tetraploid Schizothoracine fish to Qinghai Lake (3196 m above sea level, a.s.l.) in the northeast margin of QTP (Wood, et al. 2007; Xiong, et al. 2010; Tian, et al. 2019). Relative to organisms at higher altitudes, *G. przewalskii* is an important link in the evolution of biological plateau adaptation. Qinghai Lake is the largest closed-basin lake in China with no surface water outflows, owning to blocking the channel from Qinghai Lake to Yellow River around the end of the Middle Pleistocene (An, et al. 2006). The lake not only has high-altitude environmental characteristics, such as a low concentration of dissolved oxygen (4.16-6.08 mg/L) and unyielding cold (average annual air temperature ˜ -0.1), but also has its own unique characteristics of strong alkalinity (carbonate alkalinity approximately 29 mM, pH 9.1-9.5) and a high salinity level (approximately 13 ppt) (Wang, et al. 2003; Xu, et al. 2010; Zhang, et al. 2010; Fu, et al. 2013; Cui, et al. 2016). Due to high evaporative water loss, decreasing water supplies, and extensive water diversion for agricultural use, the salinity and the alkalinity levels are increasing by 7% and 0.5% per year, respectively (Yao, et al. 2016) . Unlike most stenohaline cyprinids, *G. przewalskii* undergoes an annual spawning migration between the saline-alkaline Qinghai Lake and the freshwater tributary rivers. From March to July, fish migrate to freshwater rivers to spawn and, upon completion of reproduction, return to the lake for the rest of the year (Walker, et al. 1995). This transition from river water to lake water causes profound changes for *G. przewalskii* , which include an acid-base disturbance, a rapid rise in osmolality, and a rise in the concentrations of all measured plasma ions (Wood, et al. 2007; Yao, et al. 2016). Although historically abundant, the naked carp is facing collapse due to overfishing and destruction of spawning habitat through dam building for irrigation, and has been listed as a national class II endangered species and assessed as being vulnerable (VU) in the new Red List of China's Vertebrates (Wang and Xie 2009; Jiang, et al. 2016).

G. przewalskii is an important and unique animal model for studying aquatic biological development, genetics, evolution and physiology in highland aquatic areas. This unique environmental adaptability is very important for the study of low-oxygen, high saline-alkaline and low-temperature tolerance of high-altitude fish and the long-term evolution mechanism of aquatic organism with tetraploid genome structure in this extreme environment. We therefore chose to sequence, assemble and annotate the genome and transcriptome of G. przewalskii as an example to reveal geographic isolation of Schizothoracine fishes in Qinghai-tibet plateau lift and understand evolutionary adaptation of *G. przewalskii* to high altitudes, stressful saline-alkaline water environment.

Materials and Sequencing

The experimental animals in this project were collected from the spawning field of *G. przewalskii* (Figure 1C) in Quanji river, a tributary of Qinghai Lake, with an altitude of 3,194 meters (Figure 1A and B). Muscle was used for DNA extraction and then high quality DNA was constructed libraries and sequenced on Pacbio Sequel and Illumnia X-ten Platform for genome *de novo* assembly. DNA from blood was extracted and used for constructing Hi-C libraries. For genome annotation, we sequenced RNA by Illumnia X-ten and Pacbio sequel for RNA-seq and Iso-seq data. All the RNA was extracted from brain and gill.

Genome size estimation

Genome size of *G. przewalski* was estimated by using the K-mer method (Liu, et al. 2013) based on the Illumina X-ten next-generation sequencing data. As for the method, 17-mer was selected to do this analysis and the information of total k-mer and average k-mer depth was obtained from the quality-filtered reads. Finally, the genome size was obtained by calculation based on the formula: genome size = total k-mer number / average k-mer depth.

Genome assembly and assessment

As for the genome assembly, all of the subreads was corrected by Falcon v1.8.7, (https://github.com/falconry/falcon/releases) with specific parameters (length_cutoff = 18,000; length_cutoff_pr = 19,000) to generate the preads. And the initial genome was assembled with smartdenovo (wtpre -J 3000, wtzmo -k 21 -z 10 -Z 19 -U -1 -m 0.1 -A 1000, https://github.com/ruanjue/smartdenovo) by using the corrected preads. In order to produce more precise genome sequence, initial genome was polished by

Arrow with all of subreads based on default parameters. All high-quality NGS data was used to polish the Arrow-correct genome by nextpolish with specific parameters (task=12121212) to obtain the polished genome (Walker, et al. 2014; Hu, et al. 2020). Finally, to acquire non-redundant haploid genome, some short and redundant sequences were removed from the polished genome by using redundans (Pryszcz and Gabaldon 2016) with some parameters (identity=0.824; coverage=0.8).

To assess the precise and non-redundant of genome, we carried out four methods as follows: (1) RNA-seq data were mapped to *G. przewalski*genome by using hisat2 (Pertea, et al. 2016) with default parameters for the accuracy of gene regions (2) the genome of subreads data and NGS data were mapped to genome with minimap2 (-x pb) and bwa based on default parameters, respectively for the accuracy of assembly sequences (Li and Durbin 2009; Li 2018) . (3) the NGS mapping file was utilized to analysis the genome single-base accuracy by calling SNPs and Indels. (4) BUSCO database (https://busco.ezlab.org/) was employed to assess the completeness of genome with default parameters.

Chromosome assembly

Adapter sequences of Hi-C raw reads were trimmed, and low-quality paired-end reads were removed for clean data by using fastp v0.12.6 with default parameters (Chen, et al. 2018). Then clean reads were aligned to contig sequences using Bowtie2 (2.3.2, -end-to-end –very-sensitive -L 30) (Langmead and Salzberg 2012). Valid interaction paired reads were identified and retained by HiC-Pro v2.8.1 from unique mapped paired-end reads for further analysis (Burton, et al. 2013). Invalid read pairs, including dangling-end, self-cycle, re-ligation, and dumped products were filtered by HiC-Pro v2.8.1 (https://github.com/nservant/HiC-Pro). And then, the congtis were clustered, ordered, and oriented onto chromosomes by LACHESIS (https://github.com/shendurelab/LACHESIS), with parameters CLUSTER_MIN_RE_SITES=100, CLUSTER_MAX_LINK_DENSITY=2.5, CLUSTER NONINFORMATIVE RATIO = 1.4, ORDER MIN N RES IN TRUNK=60, ORDER MIN N RES IN SHREDS=60. Finally, placement and orientation errors exhibiting obvious discrete chromatin interaction patterns were manually adjusted.

Repeat analysis

There are different repeats types in genome sequences. So repeat sequences analysis was performed with different methods to find different repeat types. Firstly, simple sequence repeats (SSRs) were identified using the MIcroSAtellite Identification Tool (MISA) (Beier, et al. 2017). MISA can distinguish and locate both simple and compound microsatellites. Next, a combination of *de novo* -based and homology-based strategy was utilized to search other repeat sequences. RepeatModeler (v1.0.8) was applied in detecting repeat sequences as the *de novo* -based method and then, repeat sequences, which were found by RepeatModeler, were classified by TEclass (Abrusan, et al. 2009). These classified sequences were merged with Repbase sequences to construct a custom TE library (Jurka, et al. 2005). Finally, *G. przewalskii* genome took advantage of the custom TE library to annotate repeat sequences with RepeatMasker (http://repeatmasker.org).

Gene prediction and functional annotation

As for gene prediction, three methods were used in prediction, which are ab initio prediction, homolog protein mapping, transcripts annotation, respectively. About ab initio prediction, AUGUSTUS was selected to predict *G. przewalskii'* s gene and the training set was produced by transcriptome (Stanke, et al. 2008; Hoff and Stanke 2019). As for homolog protein mapping, six relative species (*Carassius auratus* ,*Cyprinus carpio* , *Oryzias latipes* , *Takifugu rubripes* , *Gasterosteus aculeatus* , *Danio rerio*) data were download from National Center for Biotechnology Information (NCBI) and Ensembl to construct homolog-protein database (Table S1), which was utilized by GeMoMa to annotate gene (Keilwagen, et al. 2016). Next, Pacbio sequel transcript data was corrected to produce transcriptome by IsoSeq2 (https://github.com/PacificBiosciences/IsoSeq) and illumina X-ten RNA-seq data were mapped to *G. przewalskii'* s genome to assembly transcriptome by hisat2 and stringtie (Pertea, et al. 2016). And then RNA-seq and Iso-seq data was used to predict gene by PASA (−ALIGNERS gmap -f) (Haas, et al. 2003). Then, EVidenceModeler (Haas, et al. 2008) was employed in integrating above-mentioned prediction gene to obtain a raw gene set. Finally, we used the PSI database to search annotated gene and removed the hit gene from raw gene set to obtain the precise final gene set

(Altschul, et al. 1997).

In order to know gene function, all of predicted gene was searched against five databases, which are KEGG, KOG, NR, Swissprot, GO (Ashburner, et al. 2000; Kanehisa and Goto 2000). About the first four databases, gene sequence, which would translate to protein sequence, was mapped to different database by using BASTP (e-value 0.00001) (Altschul 1990). As for GO database, InterProScan was applied for annotation (Hunter, et al. 2009).

Gene family and phylogenetic analysis

To identify gene families in the *G. przewalskii* genome, we selected genomes of 14 other fish data and these data was downloaded from the open-source database (Table S2). We performed the OrthoMCL (v2.0.9) pipeline to identify gene families between genomes of these species (Li, et al. 2003). All-to-all BASTP with an E-value threshold of 1e-5 was applied to determine the similarities between protein sequences of the longest transcript of each gene for these species, and genes were classified into orthologues, paralogues and single copy orthologues (only one gene in each species), respectively.

Molecular phylogenetic analysis was performed using single copy orthologous genes, and each gene family for multiple sequence alignment used Mafft and curated the alignments with Gblocks v0.91b (Castresana 2000; Katoh and Standley 2013). We constructed the phylogenetic tree based on the GTRGAMMA model and a bootstrap of 100 by RAxML (v 8.2.11) (Stamatakis 2006). MCMCTREE in PAML v4.9e was used to estimate the divergence times (Yang 1997). Three fossil calibration times were obtained from the TimeTree database (http://www.timet ree.org/).

Unique Gene Analysis

The protein sequences of each gene between the 15 species were used to identify orthologous genes by using all-to-all BASTP with an E-value threshold of 1e-5. All of genes were classified to five groups, which are single, multiple, other, unique and uncluster. As for *G. przewalskii,* the gene of unique and uncluster group is peculiar, which was carried out to do enrichment analysis about KEGG by clusterProfiler (v3.10.0) with p-value < 0.05.

Results

Genome and Transcriptome sequencing

We adopted different sequence platforms to get different omic's data. As for the genome data, pacbio sequel data, Illumina X-ten data, Hi-C data were sequenced 148,752,476,139 bp, 90,421,902,000 bp and 221,620,204,800 bp, respectively. As for the transcriptome data, the brain and gill of RNAseq was obtained 13,482,126,600 bp and 10,463,033,100 bp, respectively. The Isoseq sequenc data was sequenced 22,037,012,283 bp. All of those omic's data was employed at this paper (Table S3).

Genome Size Estimation

87,165,732,691 bp Illumina clean reads were used to estimate*G. przewalskii* genome size. The K-mer analysis result showed that three peaks arised in the depth distribution curve of 17-mers are 20, 40, 80, respectively. The different depth peak is representative of different genome size. Because *G. przewalskii* is autotetraploid, the peak is representative of tetraploid, diploid, haploid genome size, respectively. So 949,878,192 bp is haploid genome size of *G. przewalskii* genome size (Figure S1 and Table S4).

Genome Assembly and Evaluation

The 148.75 Gb subreads data were obtained from Pacbio sequel sequencing technology to apply in genome assembly. The initial genome is 1,217,613,942 bp and then polished genome is 1,223,597,352 bp, which were larger than the 17-mer estimating size (949,878,192 bp). As for the autotetraploid, there were some redundant sequences product during assembly. So 278,014,088 bp redundant sequences were removed from the polished genome. Finally, ultimate genome is the non-redundant haploid genome and the size is 945,583,264 bp and N50 is 1,645,408 bp, including 967 contigs (Table 1).

For the quality of ultimate genome, some information can indicate that the genome is very precise. As for the precision of genome,the coverage of TGS and NGS is nearly 100% and the accuracy of genome is more than 99.99%. As for the completeness of genome, BUSCO database searched the genome and 96.06% complete genes were detected in the genome (Table 1).

Chromosome assembly by Hi-C Data

The 215,637,539,004 bp clean data yielded from Hi-C library were used to anchore ultimate genome sequences (945,583,264 bp, including 967 contigs) to 23 chromosomes (Figure 2 and Figure S2). Finally, *G. przewalskii* genome had been successfully reconstructed the first chromosomal-level assembly. The chromosomal-level genome size is 945,068,080 bp (Table S5).

Genome Repeat Analysis and Gene Prediction

Overall, repetitive sequences accounted for 66.32% of *G. przewalskii* genome. LTR elements consisted 53.22% of the *G. przewalskii* genome as the most abundant repeat class (Table S6). Used the gene prediction method and obtaining gene structure (Table 2). Finally, 27,224 protein-coding genes were identified in *G. przewalskii* genome, with an average of 19,299.98 bp in length, 10.13 exons per gene (Table S7). The completeness of the annotation was evaluated by BUSCO (v3.0.1)(Simao, et al. 2015). The result of BUSCO analysis proved that our annotation covered 88% complete BUSCOs. The distribution of genes and repeats was show in the circos (Figure 2).

Gene Family and Phylogenetic Analysis

To investigate the extent of genetic conservation among teleost fishes, fifteen vertebrate genomes was compared in this study, including*G. przewalskii, Carassius auratus, Cyprinus carpio, Ctenopharyngodon idella, Danio rerio, Ictalurus punctatus, Gadus morhua, Boleophthalmus pectinirostris, Oryzias latipes, Larimichthys crocea, Tetraodon nigroviridis, Takifugu rubripes, Paralichthys olivaceus, Oncorhynchus mykiss and Salmo salar.* These 15 species data were preformed to indentify the gene family and the single copy gene number is 241, the gene number of multiple groups is 9,832, the gene number of other groups is 16,273, the gene number of unique groups is 593 and the unclustered gene is 1,400 (Table S8 and Figure 3). The phylogenetic results showed that the 15 fish clustered into three groups, marine, freshwater and migratory fish and revealed that *G. przewalskii*diverged 24 Mya from its common ancestor with *Carassius auratus*and *Cyprinus carpio* , closely relating to Cyprinidae family (Figure 3).

Unique Gene Analysis

The unique genes of *G. przewalskii* is 1,993 (unique groups gene: 593 and unclustered gene 1,400) and the enrichment of KEGG indicates that the significant enrichment pathways are ATP-binding cassette (ABC) transporters (map02010), Cell adhesion molecules (map04514) and Bile secretion (map04976) (Table S9 and Figure S3).

Discussion

The species differentiation events between *G. przewalskii* and*C. carpio* occurred at 24.06 Mya ago based on the phylogenetic analysis. This period coincides with the second tectonic uplift of the Qinghai-tibet plateau (25˜17 Mya) affected by Himalayan orogeny, resulting in an average elevation of up to 2000m (Shi, et al. 1999; Wang, et al. 2014). In addition, the complex geography, such as the appearance of sedimentary basins (Qaidam basins, 27Mya; Yumen basins, 25Mya; Linxia basins, 29Mya) surrounding the Qinghai lake, had intensified the barriers to communication among species before the Early Yellow River formation (from late Middle Pleistocene to Holocene) (Fang, et al. 2003; Han, et al. 2013; Ding, et al. 2014; Wang, et al. 2014). We advanced a hypothesis that some primordial ancestor of Cyprinidae fish which were isolated from the population and remained on the plateau side had evolved to Schizothoracine fish.

The unique gene analysis found that most of these genes enriched in pathways participating in membrane transport and immune defense, such as bile secretion, ABC transporters, cell adhesion molecules. The migration between freshwater river and saline-alkaline lake requires *G. przewalskii* deal with profound physiology changes in acid–base balance and osmoregulation. For example, the bile secretion pathway plays an important role in bicarbonate secretion and proton uptake in the bile and liver, which may aid naked carp in compensating for a respiratory alkalosis during saline-alkaine water adaptation (Yao, et al. 2016). With respect to the immune response, the innate immune system provides a fundamental barrier to preventing pathogen entry into the fish body (Whyte 2007). As one of the physical components of the innate immune system, fish scales provide a protective barrier against viruses and bacteria in a variety of environments, effectively preventing infection and fighting off disease (Ellis 2001; Magnadottir 2006, 2010). However, scaleless fishes have lost this protective barrier, so they might have to enhance other biological functional pathways such as cell adhesion molecules to replenish the protective deficiency associated with the loss of scales.

Author contributions

Yimeng Liu : Funding Acquisition, Project Administration, Validation, Visualization, Writing – Original Draft Preparation.**Zongli Yao** : Writing – Review & Editing, Investigation.**Zongyi Sun** : Methodology, Software, Visualization.**Hongfang Qi** and **Jianquan Shi** : Resources. **Kai Zhou** , **Pengcheng Gao** and **Zhen Sun** : Writing – Review & Editing. **Qifang Lai** : Funding Acquisition, Supervision.

Availability of supporting data

The project has been deposited in the US National Center for Biotechnology Information (NCBI) under accession number PRJNA644469.

Declaration of competing interest

The authors declare that they have no competing financial interests.

Acknowledgment

This work was supported by National key R&D program of China (NO. 2018YFD0900603) and special research fund for the National non-profit institutes of China (No.2018Z02-01).

References

Abrusan G, Grundmann N, DeMester L, Makalowski W. 2009. TEclass–a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics 25:1329-1330.

Altschul S. 1990. Basic Local Alignment Search Tool. Journal of Molecular Biology 215:403-410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

An Z, Wang P, Shen J, Zhang Y, Zhang P, Wang S, Li X, Sun Q, Song Y, Al L, et al. 2006. Geophysical survey on the tectonic and sediment distribution of Qinghai Lake basin. Science in China Series D: Earth Sciences 49:851-861.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25-29.

Beier S, Thiel T, Munch T, Scholz U, Mascher M. 2017. MISA-web: a web server for microsatellite prediction. Bioinformatics 33:2583-2585.

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol 31:1119-1125.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540-552.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884-i890.

Cui B-L, Li X-Y, Wei X-H. 2016. Isotope and hydrochemistry reveal evolutionary processes of lake water in Qinghai Lake. Journal of Great Lakes Research 42:580-587.

Ding L, Xu Q, Yue Y, Wang H, Cai F, Li S. 2014. The Andean-type Gangdese Mountains: Paleoelevation record from the Paleocene–Eocene Linzhou Basin. Earth planet. sc. lett. 392:250-264.

Ellis AE. 2001. Innate host defense mechanisms of fish against viruses and bacteria. Developmental & Comparative Immunology 25:827-839.

Fang X, Garzione C, Van der Voo R, Li J, Fan M. 2003. Flexural subsidence by 29 Ma on the NE edge of Tibet from the magnetostratigraphy of Linxia Basin, China. Earth and Planetary Science Letters 210:545-560.

Fu C, An Z, Qiang X, Bloemendal J, Song Y, Chang H. 2013. Magnetostratigraphic determination of the age of ancient Lake Qinghai, and record of the East Asian monsoon since 4.63 Ma. Geology 41:875-878.

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr., Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, et al. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31:5654-5666.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol 9:R7.

Han J, Shao Z, Zhu D, Meng X, Yu J, Wang J, Lv R, Qian C, He C. 2013. Characteristics of river terraces and formation of the Yellow River in the source region of Yellow River. Geology in China 40:1531-1541.

Harrison TM, Copeland P, Kidd WS, Yin A. 1992. Raising tibet. Science 255:1663-1670.

Hoff KJ, Stanke M. 2019. Predicting Genes in Single Genomes with AUGUSTUS. Curr Protoc Bioinformatics 65:e57.

Hu J, Fan J, Sun Z, Liu S. 2020. NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics 36:2253-2255.

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, et al. 2009. InterPro: the integrative protein signature database. Nucleic Acids Res 37:D211-215.

Jiang Z, Jiang J, Wang Y, Zhang E, Zhang Y, Li L, Xie F, Cai B, Cao L, Zhang G. 2016. Red List of China's Vertebrates. Biodiversity Science 24:500-551.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462-467.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27-30.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-780.

Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, Hartung F. 2016. Using intron position conservation for homology-based gene prediction. Nucleic Acids Res 44:e89.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357-359.

Leggatt RA, Iwama GK. 2003. Occurrence of polyploidy in the fishes. Reviews in Fish Biology and Fisheries 13:237-246.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094-3100.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760.

Li J. 1991. The environmental effects of the uplift of the Qinghai-Xizang Plateau. Quaternary Science Reviews 10:479-483.

Li L, Stoeckert CJ, Jr., Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178-2189.

Li Y, Ren Z, Shedlock AM, Wu J, Sang L, Tersing T, Hasegawa M, Yonezawa T, Zhong Y. 2013. High altitude adaptation of the schizothoracine fishes (Cyprinidae) revealed by the mitochondrial genome analyses. Gene 517:169-178.

Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, Li Z, Chen Y, Mu D, Fan W. 2013. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. Quantitative Biology.

Magnadottir B. 2010. Immunological control of fish diseases. Mar Biotechnol (NY) 12:361-379.

Magnadottir B. 2006. Innate immunity of fish (overview). Fish Shellfish Immunology 20:137-151.

Molnar P, Boos WR, Battisti DS. 2010. Orographic controls on climate and paleoclimate of Asia: Thermal and mechanical roles for the Tibetan Plateau. Annual Review of Earth and Planetary Sciences 38:77-102.

Molnar P, England P, Martinod J. 1993. Mantle dynamics, uplift of the Tibetan Plateau, and the Indian Monsoon. Reviews of Geophysics 31.

Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc 11:1650-1667.

Pryszcz LP, Gabaldon T. 2016. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res 44:e113.

Shi Y, Li J, Li B, Yao T, Wang SL, Shijie, Cui Z, Wang F, Pan B, Fang X, Zhang Q. 1999. Uplift of the Qinghai-Xizang (Tibetan) plateau and East Asia environmental change during late cenozoic. Journal of Geographical Sciences 54:12-22.

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210-3212.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688-2690.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24:637-644.

Tapponnier P, Zhiqin X, Roger F, Meyer B, Arnaud N, Wittlinger G, Jingsui Y. 2001. Oblique stepwise rise and growth of the Tibet plateau. Science 294:1671-1677.

Tian F, Liu S, Shi J, Qi H, Zhao K, Xie B. 2019. Transcriptomic profiling reveals molecular regulation of seasonal reproduction in Tibetan highland fish, Gymnocypris przewalskii. BMC Genomics 20:2.

Tong C, Li M. 2020. Genomic signature of accelerated evolution in a saline-alkaline lake-dwelling Schizothoracine fish. International Journal of Biological Macromolecules 149:341-347.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9:e112963.

Walker KF, Dunn IG, Edwards D, Petr T, Yang HZ. 1995. A fishery in a changing lake environment: The naked carp *Gymnocypris przewalskii* (Kessler) (Cyprinidae: schizothoracinae) in Qinghai Hu, China. International Journal of Salt Lake Research 4:169-222.

Wang C, Dai J, Zhao X, Li Y, Graham SA, He D, Ran B, Meng J. 2014. Outward-growth of the Tibetan Plateau during the Cenozoic: A review. Tectonophysics 621:1-43.

Wang S, Xie Y. 2009. China species red list: Higher education press, Beijing.

Wang X, Gan X, Li J, Chen Y, He S. 2016. Cyprininae phylogeny revealed independent origins of the Tibetan Plateau endemic polyploid cyprinids and their diversifications related to the Neogene uplift of the plateau. Sci China Life Sci 59:1149-1165.

Wang YS, Gonzalez RJ, Patrick ML, Grosell M, Zhang C, Feng Q, Du J, Walsh PJ, Wood CM. 2003. Unusual physiology of scale-less carp, Gymnocypris przewalskii, in Lake Qinghai: a high altitude alkaline saline lake. Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology 134:409-421.

Whyte SK. 2007. The innate immune response of finfish–a review of current knowledge. Fish Shellfish Immunol 23:1127-1151.

Wood CM, Du J, Rogers J, Brauner CJ, Richards JG, Semple JW, Murray BW, Chen XQ, Wang Y. 2007. Przewalski's naked carp (*Gymnocypris przewalskii*): an endangered species taking a metabolic holiday in Lake Qinghai, China. Physiological and biochemical zoology 80:59-77.

Xiong F, Chen D, Duan X. 2010. Threatened fishes of the world: *Gymnocypris przewalskii* (Kessler, 1876) (Cyprinidae: Schizothoracinae). Environmental Biology of Fishes 87:351-352.

Xu H, Hou Z, An Z, Liu X, Dong J. 2010. Major ion chemistry of waters in Lake Qinghai catchments, NE Qinghai-Tibet plateau, China. Quaternary International 212:35-43.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555-556.

Yao Z, Guo W, Lai Q, Shi J, Zhou K, Qi H, Lin T, Li Z, Wang H. 2016. *Gymnocypris przewalskii*decreases cytosolic carbonic anhydrase expression to compensate for respiratory alkalosis and osmoregulation in the saline-alkaline lake Qinghai. Journal of Comparative Physiology B 186:83-95.

Yin A, Harrison TM. 2000. Geologic Evolution of the Himalayan-Tibetan Orogen. Annual Review of Earth and Planetary Sciences 28:211-280.

Yue P. 2000. Fauna Sinica Osteichthyes: Cypriniformes(III): Beijing: Science Press.

Zan R, Liu W, Song Z. 1985. Tetraploid-hexaploid relationship in Schizothoracinae. Acta. Genet. Sin. 12:137-142 (in Chinese with English abstract).

Zhang K, Sun Y, Ju S, Ma S, Yu J. 2010. The neotectonic process causing the conversion of the Qinghai Lake from an outflow lake into an interior lake. Remote Sensing for Land Resources 11:77-81.

Zhou C, Xiao S, Liu Y, Mou Z, Zhou J, Pan Y, Zhang C, Wang J, Deng X, Zou M, et al. 2020. Comprehensive transcriptome data for endemic Schizothoracinae fish in the Tibetan Plateau. Sci Data 7:28.

Hosted file

`FIGURE.docx` available at https://authorea.com/users/353675/articles/477463-de-novo-assembly-of-a-chromosome-level-genome-of-naked-carp-gymnocypris-przewalskii-reveals-geographic-isolation-of-schizothoracine-fishes-in-qinghai-tibet-plateau-lift

Hosted file

`TABLE.docx` available at https://authorea.com/users/353675/articles/477463-de-novo-assembly-of-a-chromosome-level-genome-of-naked-carp-gymnocypris-przewalskii-reveals-geographic-isolation-of-schizothoracine-fishes-in-qinghai-tibet-plateau-lift