


Resource Article: Genomes Explored

# Chromosome-level genome of Tibetan naked carp (*Gymnocypris przewalskii*) provides insights into Tibetan highland adaptation

Fei Tian<sup>1,2,†</sup>, Sijia Liu<sup>1,†</sup>, Bingzheng Zhou<sup>1,2</sup>, Yongtao Tang <sup>1,3</sup>, Yu Zhang<sup>1,2</sup>, Cunfang Zhang<sup>1,4</sup>, and Kai Zhao<sup>1,\*</sup>

<sup>1</sup>Qinghai Provincial Key Laboratory of Animal Ecological Genomics, Key Laboratory of Adaptation and Evolution of Plateau Biota, Northwest Institute of Plateau Biology, Chinese Academy of Sciences, Xining, Qinghai, China, <sup>2</sup>University of Chinese Academy of Sciences, Beijing, China, <sup>3</sup>Henan Normal University, Xinxiang, China, and <sup>4</sup>State Key Laboratory of Plateau Ecology and Agriculture, Qinghai University, Xining, Qinghai, China

\*To whom correspondence should be addressed. Tel: 0086-0971-6103697; Email: zhaokai@nwipb.cas.cn

†These authors contributed equally to this work.

Received 2 May 2022; Editorial decision 6 July 2022

## Abstract

*Gymnocypris przewalskii*, a cyprinid fish endemic to the Qinghai-Tibetan Plateau, has evolved unique morphological, physiological and genetic characteristics to adapt to the highland environment. Herein, we assembled a high-quality *G. przewalskii* tetraploid genome with a size of 2.03 Gb and scaffold N50 of 44.93 Mb, which was anchored onto 46 chromosomes. The comparative analysis suggested that gene families related to highland adaptation were significantly expanded in *G. przewalskii*. According to the *G. przewalskii* genome, we evaluated the phylogenetic relationship of 13 schizothoracine fishes, and inferred that the demographic history of *G. przewalskii* was strongly associated with geographic and eco-environmental alterations. We noticed that *G. przewalskii* experienced whole-genome duplication, and genes preserved post duplication were functionally associated with adaptation to high salinity and alkalinity. In conclusion, a chromosome-scale *G. przewalskii* genome provides an important genomic resource for teleost fish, and will particularly promote our understanding of the molecular evolution and speciation of fish in the highland environment.

**Key words:** *Gymnocypris przewalskii*, genome, whole-genome duplication, gene family expansion, adaptation

## 1. Introduction

The Qinghai-Tibetan Plateau (QTP) is characterized by an extreme environment with low temperature, hypoxia, strong UV radiation and limited food resources. Native species of the QTP have undergone significant genetic, physiological and morphological changes to adapt to this harsh environment.<sup>1–4</sup> Schizothoracinae is the exclusive Cyprinidae endemic to the QTP, including 12 genera and over 100 species distributed in the QTP and surrounding area.<sup>5</sup> It has been

reported that schizothoracine fish evolved in response to geographic and eco-environmental transformations in the uplift to the QTP.<sup>6,7</sup> Extant species of Schizothoracinae and their altitude distribution reflect the phased uplift of the QTP, thus providing an ideal case for studying the adaptation and diversification of fish under palaeoenvironmental changes. Additionally, several studies revealed the association between the evolutionary history of Schizothoracinae and geomorphological changes in the QTP, which provided biotic

evidence for the reconstruction of the paleoelevation of the QTP.<sup>8–11</sup> Therefore, deciphering the genome of schizothoracine fish will advance our knowledge on the adaptive mechanisms of teleosts to the highland environment, and facilitate the understanding of evolutionary processes of the major drainages in the QTP.

Whole-genome duplication (WGD) is considered one of the driving forces for evolution, and provides genetic materials for evolutionary novelty, species diversification and environmental adaptation.<sup>12–14</sup> Most cyprinid fish experienced an additional round of WGD after the teleost-specific WGD,<sup>15</sup> and varied in their ploidy level, ranging from diploids ( $2n = 50$ ) to high polyploids ( $2n = 470$ ).<sup>16–18</sup> As a sub-family of Cyprinidae, polyploidization was observed in all examined schizothoracine species based on karyotype examinations.<sup>18–20</sup> The phylogenetic analyses showed that a single lineage of the nuclear marker in Schizothoracinae compared with two paralogs in allotetraploid Cyprinini, which indicated possible autotetraploidization in schizothoracine fish.<sup>19,21</sup> *Gymnocypris przewalskii* is a representative species in Schizothoracinae, which resides in Lake Qinghai, the largest inland lake with high salinity and alkalinity (pH up to 9.2, ~14‰ salinity and altitude of ~3,200 m) in China.<sup>22</sup> As a primary food source for migratory birds from Siberia, *G. przewalskii* plays a crucial role in maintaining the stability of ecological processes and biodiversity of the QTP.<sup>23,24</sup>

The polyploid genomes of Schizothoracinae complicate the sequencing and assembly of a high-quality genome. Although the genome of *G. przewalskii* has been reported (preprint), the assembly information has not been released and the adaptive mechanism has not been comprehensively discussed.<sup>25</sup> By combining PacBio long-read sequencing, Illumina short-read sequencing and Hi-C technology, we present a chromosome-scale *G. przewalskii* genome, which can serve as the reference genome for schizothoracine fish. The comparative genomics analyses were conducted between *G. przewalskii* and related cyprinid fish to characterize the gene family evolution and WGD in *G. przewalskii*. The high-quality genome of *G. przewalskii* provided new insights into the evolution, speciation and diversification of teleosts, which would advance our understanding on the adaptive mechanisms at the genomic level.

## 2. Materials and methods

### 2.1. Field investigation and ethics statement

A female *G. przewalskii* was net-captured in Lake Qinghai. The fieldwork was authorized and supervised by the Qinghai Provincial Bureau of Fishery. All animal experiments were conducted following the procedures described in the ‘Guidelines for animal care and use’ manual approved by the Animal Care and Use Committee, Northwest Institute of Plateau Biology, Chinese Academy of Sciences.

### 2.2. Genome survey

Two methods were applied to predict the genome size of *G. przewalskii*, including flow cytometry and the *k*-mer method. For flow cytometry, blood samples (0.2–0.5 ml) were centrifuged at <500 (rpm) to obtain red blood cells. Red blood cells were fixed in 70% ethanol at 4°C, and stained with CyStain UV Precise P (Partec, Germany) for 30 min at 4°C. The cell number was counted and reached over  $1 \times 10^4$ . Samples and controls were run in CyFlow Space (Partec). A chicken blood sample ( $2C = 2.30$  pg/N) was used as a control.<sup>26</sup> Second, the blood sample was collected from a female *G. przewalskii*, which was used for the *k*-mer method and subsequent *de novo*

sequencing. For the *k*-mer method, genomic DNA was purified from blood samples using a TIANamp Blood DNA Kit (TIANGEN, China) according to the manufacturer’s description. The quantity and quality of DNA was assessed by 1% agarose gel electrophoresis, Qubit<sup>®</sup> 2.0 Fluorometer (Life Technologies, USA) and Agilent Bioanalyzer 2100 system (Agilent Technologies, USA). Genome survey sequencing was performed to estimate genome size and determine the sequencing depth. Blood DNA was randomly sheared for library construction and sequencing using the Illumina HiSeq X ten platform. The map of *k*-mer distribution with  $k = 17$  (Supplementary Fig. S1) was constructed using Jellyfish (v2.2.6).<sup>27</sup> The genome size and the heterozygosity were estimated using GenomeScope (v1.0.0).<sup>28</sup>

### 2.3. Library construction and whole-genome sequencing

(Color online) Genomic DNA was sheared to a size range of 15–40 kb, enzymatically repaired and converted into SMRTbell template libraries as recommended by Pacific Biosciences. The resulting SMRTbell templates were sequenced on a PacBio Sequel instrument. A total of 40 SMRT cells were sequenced, producing 225.68 Gb SMRT raw data. Genomic DNA was used to construct paired-end libraries with a 350 bp insert size using a Paired-End DNA Sample Prep kit (Illumina, San Diego, CA). These libraries were sequenced using the Illumina HiSeq X ten platform, producing 192.53 G raw data. Genome sequencing was conducted by Genedenovo Inc. (Guangzhou, China).

### 2.4. *De novo* assembly of the genome using PacBio and Illumina data

Primary contigs were assembled from PacBio long reads by MECAT (v1.0).<sup>29</sup> Overlaps of long reads were found using the command `mecat2pw` (parameters: `-k 4 -a 2000`) and were corrected using the command `mecat2cns` (parameters: `-r 0.9 -c 6 -l 5000`). The resulting contigs were polished using more than 100× coverage of Illumina short reads by two rounds of Pilon (version 1.22) with default parameters.<sup>30</sup>

### 2.5. Hi-C assembly

To improve the assembly, we performed an Hi-C experiment. Muscle tissue from female *G. przewalskii* was cross-linked and digested for *de novo* sequencing. DNA fragments with 300–700 bp insert sizes were used for Hi-C library construction<sup>31</sup> and were sequenced with an Illumina HiSeq 4000 platform. The clean Hi-C reads were first truncated at the putative Hi-C junctions, and the resulting trimmed reads were realigned to the assembly results with BWA-MEM (v0.7.16a-r1181).<sup>32</sup> Only uniquely aligned paired reads whose mapping quality was more than 20 were used for further analysis. Invalid read pairs, including dangling-end, self-cycle, relegation and dumped products, were filtered by HiC-Pro (v2.8.1).<sup>33</sup> The uniquely mapped read pairs were used for scaffolds clustered, ordered and oriented onto chromosomes by ALLHiC (v0.9.8).<sup>34</sup> The genome-wide Hi-C interaction matrix was generated using ALLHiC\_plot program and visualized as heatmap using R. Merqury (v 1.3)<sup>35,36</sup> was performed to evaluate the completeness to the genome assembly with parameters of tolerable collision rate of 0.001 and a specific Kmer of 21.

## 2.6. Genome annotation

We applied RNAmmer<sup>37</sup> for rRNA prediction, and tRNAscan-SE (v2.0)<sup>38</sup> for the identification of tRNAs, and miRNAs and siRNAs were identified through the Rfam database.<sup>39</sup> Repeat sequences were annotated using both homology-based and *de novo* approaches. LTR\_FINDER (v1.07)<sup>40</sup> and MGEscan (v1.1)<sup>41,42</sup> were used to identify transposable elements. PILER (v1.0),<sup>43</sup> RepeatScout (v1.05)<sup>44</sup> and RepeatModeler (v2.01)<sup>45</sup> were applied for the *de novo* prediction of repetitive elements, which were merged with the REPBASE database<sup>46</sup> to construct a repetitive sequence database. The final repetitive sequences of the *G. przewalskii* genome were obtained by prediction using RepeatMasker (v4.0.5)<sup>47</sup> using the constructed database.

Three strategies were adopted to annotate genes in the *G. przewalskii* genome. First, AUGUSTUS (v3.3.3),<sup>48</sup> SNAP,<sup>49</sup> GlimmerHMM (v3.0.4)<sup>50</sup> and GeneMark (v4.35)<sup>51</sup> were applied for *ab initio* gene prediction. Second, proteins from *Danio rerio*, *Ctenopharyngodon idellus*, *Cyprinus carpio*, *Simocyclocheilus anshuiensis*, *Simocyclocheilus grahami*, and *Simocyclocheilus rhinoceros* were aligned to *G. przewalskii* using BLASTX with an e-value of  $1e-6$ , and then gene models were defined using Exonerate (v2.2).<sup>52</sup> Third, we produced 11 RNA-seq libraries from the gill, kidney, heart and intestine of four *G. przewalskii*. RNA was purified using MiniBEST Universal RNA extraction kit (TaKaRa, China). RNA quality was evaluated by 1% gel electrophoresis, Qubit<sup>®</sup> 4.0 Fluorometer (Life Technologies), and Agilent Bioanalyzer 2100 system (Agilent Technologies, USA). A total of 1.5  $\mu$ g RNA was used for library construction based on the manufacturer's description. Briefly, polyT oligo-attached magnetic beads were used to capture mRNA from total RNA, and then fragmented. The synthesis of first-strand cDNA was carried out using random hexamer primer and MMuLV Reverse Transcriptase (RNase H<sup>-</sup>), and followed by the synthesis of second-strand cDNA. After end repair, sequence adaptors were added. The purified cDNA was amplified using PCR with universal PCR primers and index primers. The library was quantified using Qubit<sup>®</sup> 4.0 Fluorometer (Life Technologies), which were sequenced on the Illumina HiSeq 4000 platform (Novogene, China). Additionally, 19 RNA-seq datasets of *G. przewalskii* were downloaded from the NCBI Sequence Read Archive (SRA) database (Supplementary Table S16). In total, 30 RNA-seq data were filtered by fastp<sup>53</sup> to obtain clean data, which were mapped to the *G. przewalskii* genome using HISAT (v2.1.0)<sup>54</sup> to identify splicing junctions. StringTie (v1.3.4)<sup>55</sup> was used to obtain transcriptomes using aligned reads. Finally, the outputs from *ab initio* prediction, homology search and RNA-seq were combined as inputs for the MAKER pipeline (v2.3.1)<sup>56</sup> for protein-coding gene annotation, which were then functionally annotated by searching in the SwissProt, UniProt, NCBI non-redundant protein databases, Gene Ontology (GO) and KEGG databases using BLASTP (E value  $< 1 \times 10^{-5}$ ) and Blast2GO (v2.3.5).<sup>57</sup>

## 2.7. Identification of orthologous genes and phylogenetic analysis

Protein sequences of nine cyprinid fish, including *D. rerio*, *C. idellus*, *Poropuntius huangchuchieni*, *Oxygymocypris stewartii*, *Onychostoma macrolepis*, *C. carpio*, *S. anshuiensis*, *S. grahami* and *S. rhinoceros* were downloaded (Supplementary Table S17). First, all-versus-all BLASTP (e-value  $< 1 \times 10^{-7}$ ) comparison of all protein sequences was performed between nine cyprinid fish and *G. przewalskii*, and orthologous genes were clustered by OrthoMCL (v1.4).<sup>58</sup> CAFÉ (v3.1) was applied to identify orthologous genes that had

undergone expansion and/or contraction.<sup>59</sup> GO enrichment analysis was conducted in the expanded gene families in *G. przewalskii* using the clusterProfiler package in R.<sup>60</sup>

Single-copy gene families were identified in 10 cyprinid fish, which were used to construct a phylogenetic tree. MUSCLE<sup>61</sup> was used to generate a multiple sequence alignment of protein sequences for each single-copy family with default parameters. The alignments of each family were concatenated to a superalignment matrix that was used for phylogenetic tree reconstruction using MrBayes (v3.2).<sup>62</sup> The divergence time between species was estimated using MCMCTree in PAML (v4.9)<sup>63</sup> with the options 'correlated rates clock' and 'HKY85' model. A Markov chain Monte Carlo analysis was run for 1,000,000 generations using a burn-in of 100,000 iterations. The convergence was evaluated using Tracer (v1.7.1) with ESS  $\geq 200$ . The divergence time for *D. rerio* and *C. idella* [51.9–81.8 million years ago (Mya)] as well as *D. rerio* and *C. carpio* (59.3–122.1 Mya) were obtained from the TimeTree database (<http://www.timetree.org/>) and were used as the calibration points.

## 2.8. Positively selected genes analysis

To find genes that potentially experienced positive selection, the improved branch-site model (model = 2, Nsites = 2) in the codeml program in the PAML package (v4.9) was used to detect signatures of positive selection on individual codons in a specific branch.<sup>64</sup> For the detection of positively selected genes (PSGs) in *G. przewalskii*, the branch of *G. przewalskii* was used as the foreground branch, and all other branches in the phylogenetic tree were used as background branches.

## 2.9. Whole-genome duplication

The synonymous substitution rate ( $K_s$ ) density plot was generated to determine the timing of WGD in *G. przewalskii*. BLASTP was used to obtain the best-match gene pair between *G. przewalskii* and *D. rerio*, *P. huangchuchieni*, *O. macrolepis* and *C. carpio* based on the protein sequences. The protein alignments of gene pairs were then converted into codon comparisons using ParaAT (v2.0),<sup>65</sup> which were used for  $K_s$  calculation using KaKs Calculator (v2.0).<sup>66</sup> MCScanX analysis<sup>67</sup> was carried out to search genome-wide duplications between *G. przewalskii* itself and with *D. rerio* and *C. carpio*. Based on the  $K_s$  rate ( $r$ ) of  $3.51 \times 10^{-9}$  per year per nucleotide,<sup>21,68,69</sup> the timing of WGD was estimated according to the formula  $T = K_s/2r$ .<sup>68</sup>

## 2.10. Phylogenetic analysis of Schizothoracinae

The reduced representative sequencing data of 13 Schizothoracinae species were downloaded from NCBI SRA (Supplementary Table S16). The raw reads were processed using fastp<sup>53</sup> to obtain clean reads according to the description by Tang *et al.*<sup>70</sup> Clean reads were mapped to the *G. przewalskii* genome using BWA with the MEM algorithm with parameters of  $t$  and  $k$  equivalent to 4 and 32.<sup>32</sup> Variants were identified using GATK (4.0).<sup>71</sup> SNPs with a missing rate  $< 0.20$  and a minimum allelic frequency  $> 0.05$  were considered for phylogenetic tree construction. SNPs were aligned across all the samples, which were applied to ML analysis with *O. stewartii* as the root. An ML tree was constructed using RAxML-HPC2 with the GTRGAMMA model and 1,000 bootstrap replicates.<sup>72</sup>

### 2.11. Inference of effective population size history

We used the pairwise sequentially Markovian coalescent (PSMC) to infer historical changes in effective population size based on sequence data of *G. przewalskii*,<sup>73</sup> with parameter settings of  $-p\ 4 + 25 * 2 + 4 + 6 -r\ 4 -t\ 15 -N\ 30$ . For plot settings, we used a generation time of 3 (*G. przewalskii* reaches its sexual maturation at 3) and the mutation rate of  $3.51 \times 10^{-9}$  per year per nucleotide.

### 2.12. RNA-seq analysis

RNA-seq data for genome annotation (see Section 2.5) were used to measure the transcription for duplicate genes. To clearly distinguish the expression levels of duplicate copies, clean reads that were uniquely mapped to a single gene without a mismatch were counted to calculate FPKM (Supplementary Table S16). Differentially expressed genes (DEGs) were defined as having an FDR <0.01 and fold change >2.

## 3. Results

### 3.1. Genome assembly and annotation

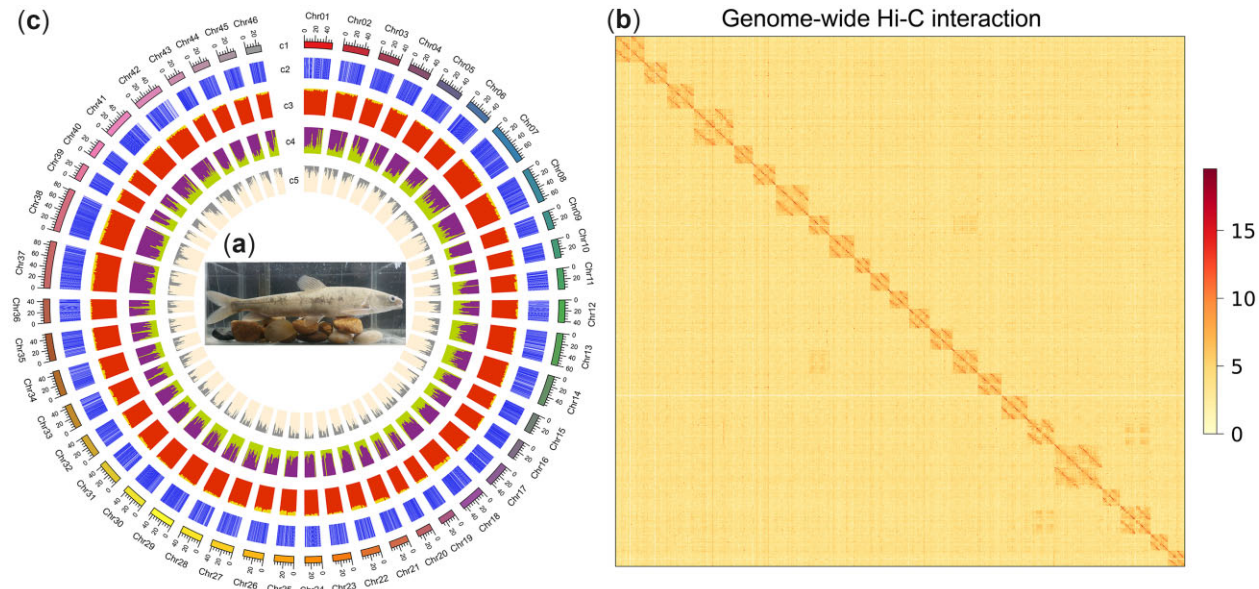
The *G. przewalskii* (Fig. 1a) genome size was estimated to be 1.90 Gb by flow cytometry (Supplementary Fig. S1a and b) and 2.19 Gb with heterozygosity of 0.96% by the *k*-mer method ( $k=17$ ) (Supplementary Fig. S1c and Table S1). We thus generated sequence data consisting of 230 Gb of PacBio Sequel long reads (~100x), 110 Gb of Illumina paired-end reads (~50x) and 350 Gb of Hi-C reads (~150x). The initial PacBio-based assembly produced 22,550 contigs with an N50 of 136.1 kb (Supplementary Table S2). According to a reported karyotype,<sup>74,75</sup> primary contigs were scaffolded into 46 chromosomes, which anchored 96.98% sequences. We developed the final assembly of 2.03 Gb with a scaffold N50 reaching 44.98 Mb (Table 1 and Fig. 1b). This assembly was similar to the *Schizothorax o'connori*

genome of 2.07 Gb and larger than the *O. stewartia* genome of 1.85 Gb.<sup>69,76</sup> The *G. przewalskii* genome had a GC content of 38.34%. Gaps comprised 0.05% of the genome. The evaluation of the assembly by BUSCO (v3.0.1, vertebrata\_odb9) resulted in the identification of 89.4% complete (duplicated: 43.5%; single-copy: 45.9%) and 2.6% fragmented genes from 2,586 vertebrate core single-copy orthologs (Supplementary Table S3). The reference free Merqury evaluation showed a QV (consensus quality) of 31.4 and a completeness of 80.36%, indicating the relatively complete and accurate assembly of the *G. przewalskii* genome.

Combining prediction, homology searching and RNA-seq data, we obtained 56,397 genes in the *G. przewalskii* genome, among which 50,660 genes (89.83%) were annotated in public databases (Supplementary Table S4). Additionally, non-coding RNAs, including 24,705 tRNAs, 1,653 microRNAs, 2,649 small nuclear RNAs and 2,068 rRNAs, accounted for 0.19% of the genome (Supplementary Table S5). We also identified that repetitive elements occupied 42.89% of the *G. przewalskii* genome by *de novo* searching and searching in repeat databases. DNA transposons (20.49%), long-interspersed nuclear elements (14.05%) and long-terminal repeats (21.31%) were the most abundant repetitive sequence types (Supplementary Table S6). In summary, we generated a chromosome scale of the *G. przewalskii* genome with relatively high completeness and accuracy.

**Table 1.** The statistics of *G. przewalskii* genome assembly

Statistics	Contig	Chromosome	Unanchored
Total number	22,550	46	1,022
Total length (bp)	2,085,351,307	2,029,228,326	63,250,882
Gap (bp)	—	1,084,502	—
Average length (bp)	185,814.94	44,073,952.24	61,889.32
N50 length (bp)	136,081	44,980,694	66,094



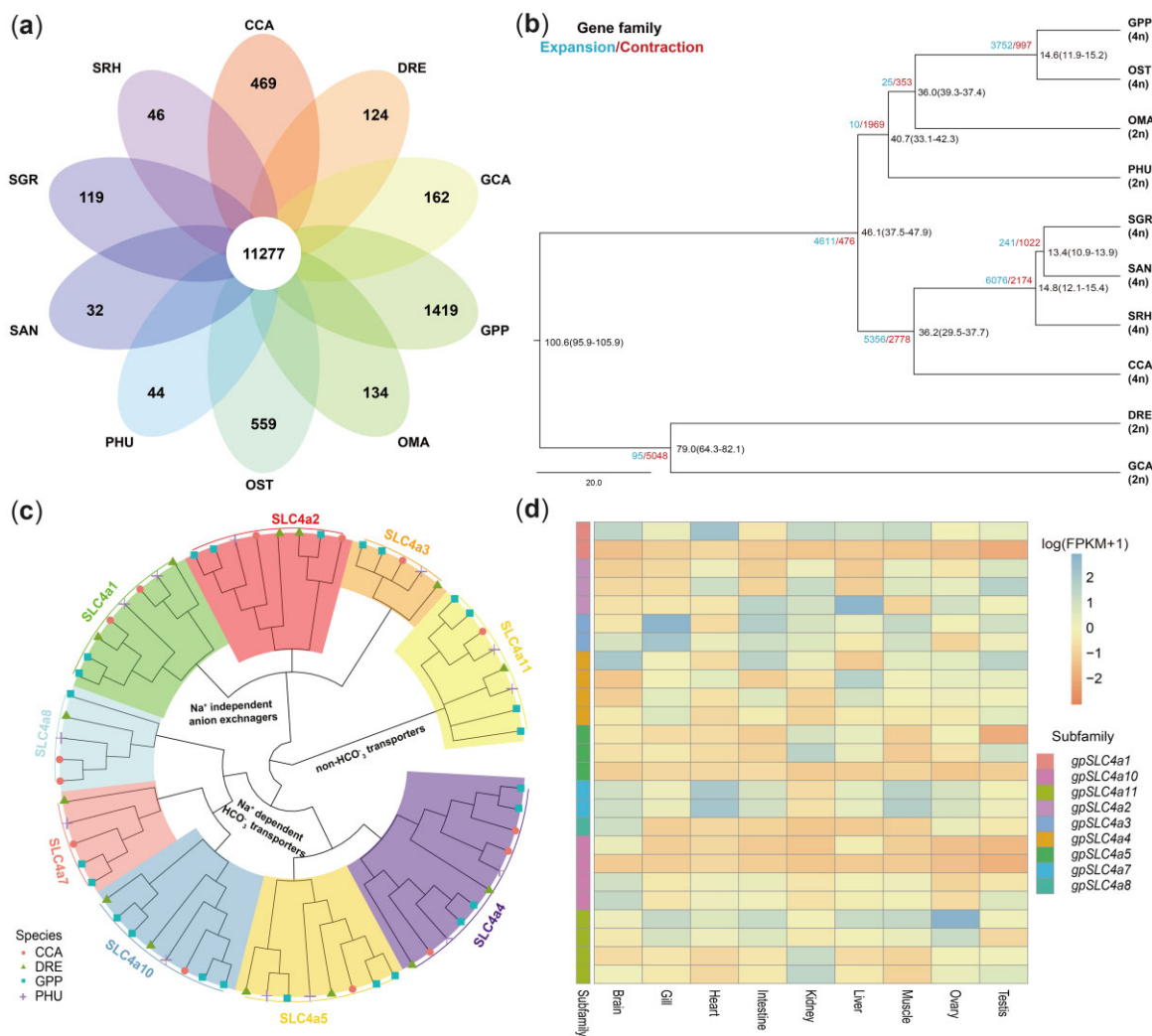
**Figure 1.** Genome of *G. przewalskii*. (a) Photo of *G. przewalskii* that was net-captured in Lake Qinghai and sequenced in the current study. (b) Chromosomal contact map of *G. przewalskii* using Hi-C data. (c) The *G. przewalskii* genome landscape created using Circos. From outer to inner circles: c1, *G. przewalskii* chromosomes at Mb scale; c2, coding sequences; c3, GC content; c4, repetitive elements drawn in a 100 kb sliding window with a 100 kb step; b5, SNP density in a 100 kb sliding window with a 50 kb step.

### 3.2. Comparative analysis between *G. przewalskii* and Cyprinidae

Comparative genomic analysis detected that 11,277 orthologous genes with 204,732 genes were shared by *G. przewalskii* and 9 cyprinid fish (*D. rerio*, *O. stewartii*, *P. huangchuchieni*, *O. macrolepis*, *C. carpio*, *S. grahami*, *S. rhinoceros*, *S. anshuiensis* and *C. idellus*), among which all fish (Fig. 2a; Supplementary Tables S7 and S17). There were 1,419 *G. przewalskii*-specific clusters with 3,468 genes. *Gymnocypris przewalskii* harboured 8,918 expanded and 4,508 contracted gene families, and expanded genes were significantly enriched in GO functions of ion and cation transport and homeostasis (Supplementary Table S8). The phylogenetic tree was constructed using 991 single-copy orthologs identified among *G. przewalskii* and nine cyprinid fish (*D. rerio*, *C. carpio*, *S. grahami*, *S. rhinoceros*, *S. anshuiensis* and *C. idellus*). The phylogenetic tree indicated that the schizothoracine fish, *G. przewalskii* and *O. stewartii* ( $4n = 92$ ), had the closest genetic relationship with *O. macrolepis* ( $2n = 50$ ),<sup>77</sup> which

is currently distributed in the middle reaches of the Yangtze River and the Yellow River in the adjacent areas of the QTP. The split between two schizothoracine fish, *G. przewalskii* in the northeast QTP and *O. stewartii* in the Yarlung Zangbo River of the southern QTP, occurred 14.6 Mya [95% highest posterior density (HPD): 11.9–15.2 Mya] (Fig. 2b), which was related to the early uplift of the QTP.<sup>78,79</sup>

Interestingly, we noticed that the *SLC4* family was extensively expanded to 25 members in *G. przewalskii*, compared with 14 members in zebrafish, 14 members in common carp and 13 members in *P. huangchuchieni*. The phylogenetic analysis classified *gpSLC4A* members into nine sub-families, all of which were grouped with their counterparts in zebrafish, common carp and *P. huangchuchieni* (Fig. 2c). This classification was consistent with *SLC4* transporter types,<sup>80</sup> including  $\text{Na}^+$ -independent anion exchangers (*SLC4A1-3*),  $\text{Na}^+$ -dependent  $\text{HCO}_3^-$  transporters (*SLC4A4*, *SLC4A5*, *SLC4A7*, *SLC4A8* and *SLC4A10*) and the non- $\text{HCO}_3^-$  transporter (*SLC4A11*). The expanded members, including *gpSLC4A2-5*, *gpSLC4A7*,



**Figure 2.** Comparative genomics between *G. przewalskii* and cyprinid fish. (a) Venn diagram of shared and unique gene families among 10 cyprinid fish species. The number represents the number of gene families. GPP, *G. przewalskii*; DRE, *D. rerio*; CCA, *C. carpio*; OST, *O. stewartii*; PHU, *P. huangchuchieni*; OMA, *O. macrolepis*; GCA, *C. idellus*; SGR, *S. grahami*; SAN, *S. anshuiensis*; SRH, *S. rhinoceros*. (b) Phylogenetic tree of 10 cyprinid fish based on single-copy gene families. The species divergence times were estimated and labelled in the branch with 95% HPD. The number of gene family expansions and contractions are labelled in blue and red, respectively. (c) Phylogenetic analysis of the *SLC4a* gene family in *G. przewalskii* (blue square), *P. huangchuchieni* (purple plus sign), *C. carpio* (red circle) and *D. rerio* (green triangle) based on the protein sequences. (d) Transcription patterns of *gpSLC4a* family members in the brain, gill, heart, intestine, kidney, liver, muscle, ovary and testis (A color version of this figure appears in the online version of this article).

*gpSLC4A10* and *gpSLC4A11*, covered each transporter type and displayed more variable expression patterns among tissues (Fig. 2d).

### 3.3. Positively selected genes

Based on single-copy orthologs, we identified 141 PSGs in *G. przewalskii* by comparison with nine cyprinid fish (Supplementary Table S9). Although these PSGs were not significantly enriched in any particular GO terms, we found that the hypoxia-related genes *EPAS1* and *HIF1a* were also detected as PSGs in *G. przewalskii* (Supplementary Table S9), which were shown to contain beneficial mutations and contribute to hypoxia adaptation in many Tibetan species.<sup>1,3,4,81</sup> Since *G. przewalskii* adapts to the high saline and alkaline environment (12–14‰ salinity) of Lake Qinghai, we paid particular attention to genes for iono- and osmoregulation. Evidence of positive selection was identified in *SLC22a23*, a member of the solute carrier superfamily involved in organic cation transport.<sup>82</sup> Two positively selected sites were identified in *AQP3*, a gene mainly expressed in the gill and kidney as a water channel function<sup>83</sup> (Supplementary Fig. S2). Genes in cell junctions were upregulated in *G. przewalskii* under high salinity.<sup>84</sup> We found a calcium-dependent cell–cell adhesion glycoprotein, *CDH4*, harbours three positively selected sites (Supplementary Fig. S3), which might play an important role in preventing the diffusion of water and/or salts in high salinity. The positively selected sites in genes related to iono- and osmoregulation might lead to structural and functional changes, which would contribute to the adaptation of *G. przewalskii* to high salinity in Lake Qinghai.

### 3.4. Population demography

Population history reconstructed using the PSMC model suggested a strong correlation between the demography of *G. przewalskii* and the upheaval of the QTP (Fig. 3a). First, ancient *G. przewalskii* experienced population expansion (~13–8 Mya) and reached a maximum effective population size from the middle to late Miocene (~27–8 Mya). This period was coincident with the early lifting of the QTP (~25–17 Mya) and the tectonic process occurring 13–8 Mya that steadily lifted the QTP to 1–2 km.<sup>85–87</sup> Then, the population size appeared to dramatically decline during the Qingzang movement (3.6–1.7 Mya)<sup>78</sup> and Kunhuang movement (1.2–0.6 Mya),<sup>88</sup> when the QTP was intensively elevated from ~2 km to its present altitude.<sup>79,89,90</sup>

### 3.5. Phylogenomic reconstruction of Schizothoracinae

The SLAF-seq data of 13 species in Schizothoracinae were obtained from NCBI (Supplementary Table S10), including the *gymnocypris* genus, *schizopygopsis* genus and *oxygymnocypris* genus,<sup>70</sup> which were aligned to the *G. przewalskii* genome. The mapping ratios ranged from 80.85% in *Schizopygopsis thermalis* (distributed in the southwestern QTP) to 91.57% in *Gymnocypris eckloni* (Supplementary Table S10), which produced 287,222 informative SNPs. The reconstructed phylogenetic tree classified 13 schizothoracine fish into three major clades (Fig. 3b), consistent with their geographic distributions. Species from the genera *Gymnocypris* and *Schizopygopsis* formed polyphyletic relationships, indicating the incongruence between taxonomic classification and molecular phylogeny in Schizothoracinae. This result indicated that the *G. przewalskii* genome could serve as a reference for schizothoracine fish, which could promote our understanding of the taxonomic classification and nomenclature of cyprinid fish.

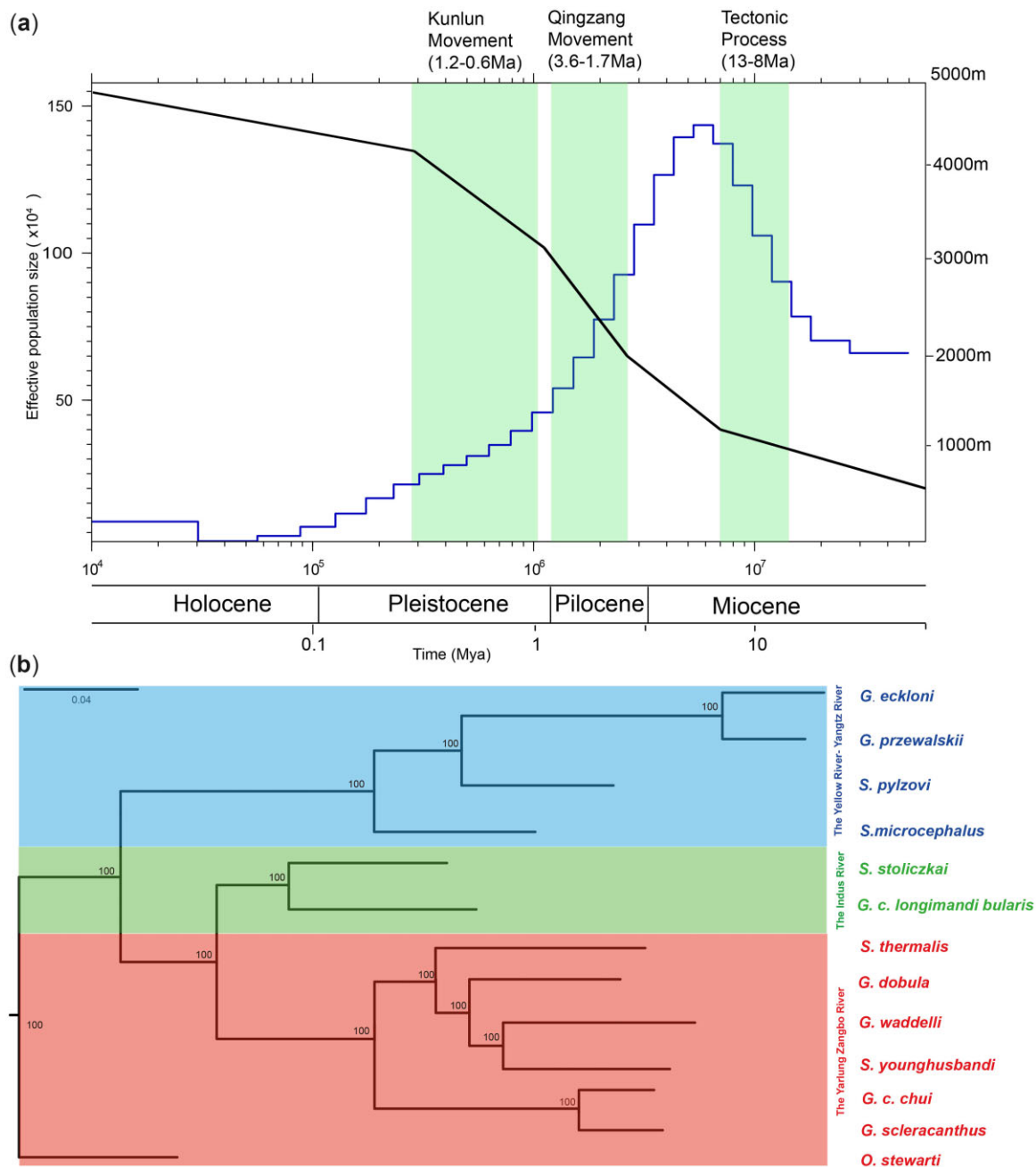
### 3.6. Genome duplication in *G. przewalskii*

WGD is considered one of the driving forces of evolution. Cyprinidae experienced an additional round of genome doubling after the teleost-specific genomic duplication event.<sup>14,74,75</sup> A small *Ks* peak of 0.018 was observed in *G. przewalskii*, suggesting that it experienced a recent WGD and is a young tetraployploid fish. Based on the *Ks* distribution, we determined that WGD in *G. przewalskii* occurred at ~2.55 Mya based on the *Ks* rate of  $3.51 \times 10^{-9}$  per year per nucleotide (Fig. 4a), which was close to the WGD of *S. o'connori* at 1.23 Mya.<sup>69</sup>

To identify genes produced by WGD in *G. przewalskii*, we performed intra- and inter-syteny analyses. Homology was clearly observed in *G. przewalskii* chromosome pairs, which revealed 754 collinear blocks covering 55.24% of coding genes (31,156/56,397) (Fig. 4b; Supplementary Table S11). A total of 8.64% of genes (4,870) were classified as singleton genes, which were considered the genes that lost one copy (Supplementary Table S11). Notably, ~64% of *G. przewalskii* specific genes (3,380/5,280) originated from WGD, suggesting that WGD might also have contributed to *G. przewalskii* speciation. GO enrichment suggested that duplicate genes were overrepresented in pathways related to highland adaptation, such as voltage-gated sodium channel activities (GO:0005248), regulation of respiratory gaseous exchange (GO: 0043576) and cellular response to UV (GO:0034644) (Supplementary Table S12). The comparisons of homologous gene pairs between *G. przewalskii* and zebrafish showed clearly collinear patterns, in which two *G. przewalskii* chromosomes roughly corresponded to one zebrafish chromosome (Fig. 4c; Supplementary Fig. S4a). Since the arm in zebrafish chromosome 4 contained potential regions specific to zebrafish,<sup>91</sup> we did not observe collinearity with any regions in *G. przewalskii* chromosomes. In total, 11,942 *G. przewalskii* genes had a two-to-one relationship in *D. rerio* (Supplementary Table S13), accounting for 54% of the *G. przewalskii* genome (Supplementary Fig. S4b). These triplet gene pairs were enriched in cold acclimation (GO:0009631), regulation of respiratory gaseous exchange (GO: 0043576) and response to UV (GO:0071493) (Supplementary Table S14), suggesting that WGD in *G. przewalskii* might contribute to its adaptation to the QTP highland. We also identified 4,835 and 4,183 genes in *G. przewalskii* with one-to-one and two-to-one copies in *C. carpio*, respectively (Fig. 4c; Supplementary Fig. S5 and Table S13). We found that 2,944 *G. przewalskii* genes had two-to-one relationship with both *D. rerio* and *C. carpio*, which might be due to allotetraploid in *C. carpio* and autotetraploid in *G. przewalskii*. Genes with two paralogs in *G. przewalskii* but one in *C. carpio* were enriched in GO terms, such as response to caloric restriction (GO:0061771), sequestering of metal ion (GO:0051238) and adaptive thermogenesis (GO:1990845) (Supplementary Table S15). These results indicated that duplicate genes in *G. przewalskii* contributed to highland adaptation.

### 3.7. The evolution and transcription of duplicate gene pairs

The evolutionary outcomes of gene duplication generally fall into three categories: non-functionalization of one copy by degenerative mutations, neofunctionalization of one copy and sub-functionalization of both copies.<sup>92–94</sup> We analysed the transcription of duplicate pairs in the brain, gill, heart, intestine, kidney, liver, muscle, ovary and testis of *G. przewalskii* (Supplementary Fig. S6 and Table S16), and the DEGs were identified between tissue comparisons. Among 31,156 duplicate genes, 26,036 were transcribed in nine organs. These expressed gene pairs were classified into two categories, either showing equal

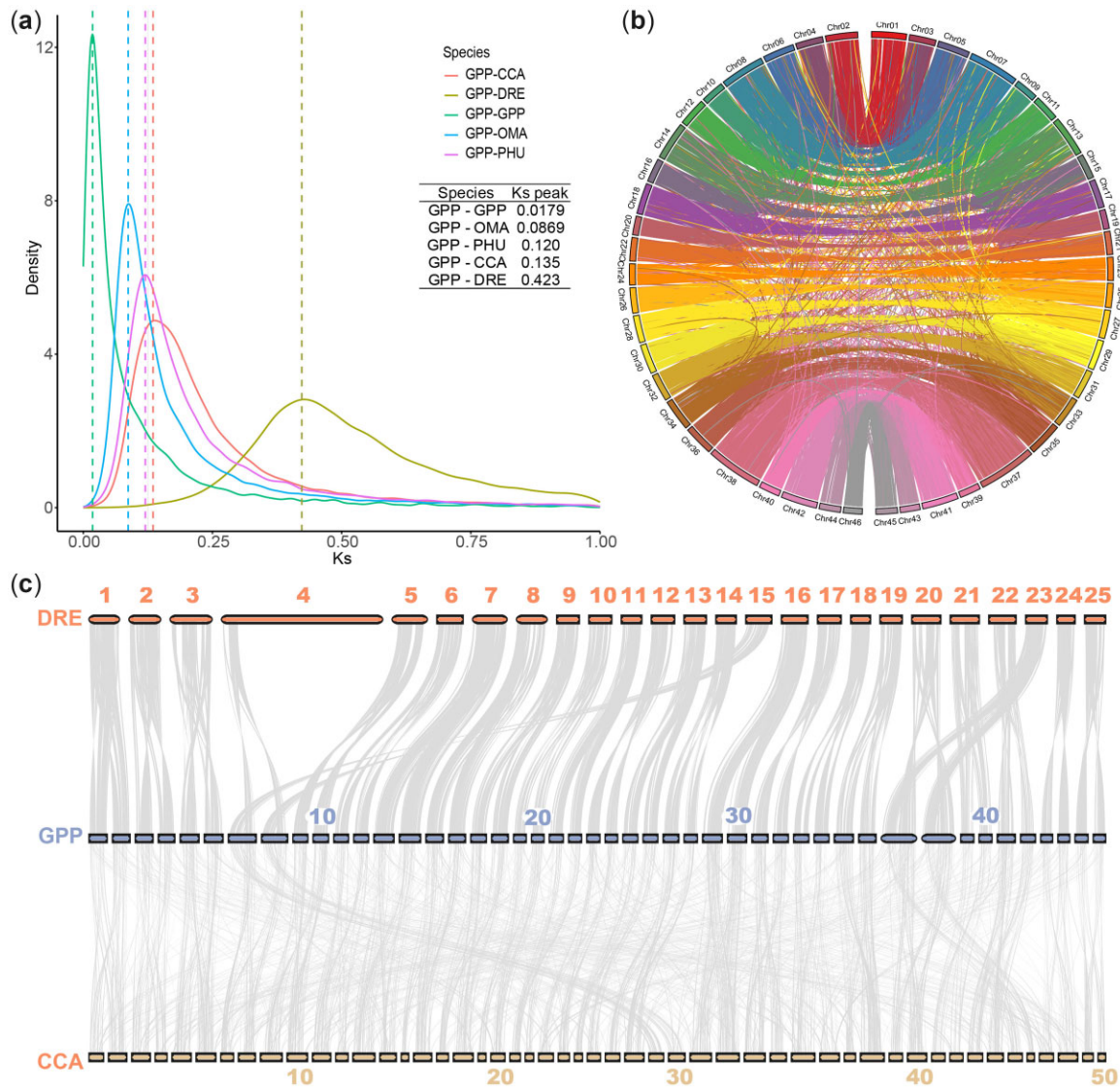


**Figure 3.** Phylogenetic analysis and dimorphic history. (a) Phylogenetic tree of schizothoracine fish. The phylogenetic tree of 13 schizothoracine fish was inferred by the maximum likelihood (ML) method based on SNPs. Bootstrap support values are labelled in the branch. Fishes from drainages of the Yellow River-Yangtze River, Indus River and the Yarlung Zangbo River are shown in blue, green and red, respectively. (b) Inference of *G. p. przewalskii* population history. The left y axis denotes the effective population size of *G. p. przewalskii* in a time-series manner (blue line). The right y-axis represents the approximate altitude in the northeastern QTP according to the estimation from Li *et al.*, Wu *et al.* and Deng *et al.* (black line). The periods of important geographic movements are shaded in green (A color version of this figure appears in the online version of this article).

expression among all tissues (1,424 pairs) or displaying different transcription between at least two tissues (10,322 pairs). In the latter category, 2,515 pairs exhibited tissue-dependent expression patterns of one copy and equal transcription of the other copy. Among the remaining gene pairs (7,807), 7,150 gene pairs showed that two copies had partitioned transcription patterns between at least two tissues (Supplementary Fig. S7). Diverse transcriptional profiles of duplicate pairs highlighted possible neo- or sub-functionalization of duplicate genes in *G. przewalskii*.

#### 4. Discussion

As an endemic fish of the QTP, *G. przewalskii* exhibits great adaptation to severe natural conditions, such as high salinity and alkalinity, chronic cold and hypoxia, which represents a remarkable model to study the evolutionary process in the extreme environment. In the present study, we assembled the chromosome-scale genome of *G. przewalskii* (2.03 Gb) with 56,397 protein-coding genes, which were comparable to the tetraploid cyprinid fish, including *C. carpio*



**Figure 4.** WGD of *G. przewalskii*. (a) Density distribution of Ks (synonymous substitution rate) of homolog gene pairs of GPP–GPP, GPP–PHU, GPP–OMA, GPP–CCA and GPP–DRE. Dashed lines indicate the peak values in each species pair. GPP, *G. przewalskii*; DRE, *D. rerio*; CCA, *C. carpio*; PHU, *P. huangchuchieni*; OMA, *O. macrolepis*. (b) Chord diagram displaying the synteny of 46 *G. przewalskii* chromosomes on the basis of 754 syntenic blocks, as indicated by lines. The chromosomes are ordered by size, and homologous chromosomes are labelled with the same colour. (c) Collinearity among 46 *G. przewalskii* chromosomes (middle), 50 *C. carpio* chromosomes (bottom) and 25 *D. rerio* chromosomes (top).

(1.83 Gb with 52,610 genes) and schizothoracine fish of *S. o'connori* (2.07 Gb with 43,731 genes) and *O. stewartii* (1.85 with 46,400 genes).<sup>69,76,95</sup> The genome sizes and gene numbers of these tetraploid fish were doubled compared with potential diploid ancestors *P. huangchuchieni* (1.02 Gb with 24,099 genes) and *O. macrolepis* (928 Mb with 24,770 genes), presumably due to WGD (Supplementary Table S18).<sup>77,96</sup> The chromosome numbers in schizothoracine fish ranged from 92 to 96, which might result from the fusion of some chromosomes in the potential autotetraploidization.<sup>69</sup>

The origination of Schizothoracinae is still under debate, and their diploid ancestors have not yet been identified.<sup>18</sup> Based on the genomic data, the phylogenetic analysis suggests that it was evolved from diploid *O. macrolepis* (sub-family Barbinae) that were currently distributed in the middle reaches of the Yangtze River in sub-tropical areas. This result was consistent with fish fossils in the QTP, which

were morphologically similar to Cyprinidae in warm regions. This finding was not only significant for the phylogeny and zoogeography of fish, but also had implications for understanding the paleoelevations of the QTP.<sup>9</sup> The demographic history of ancient *G. przewalskii* confirmed that the evolution of schizothoracine fish was influenced by environmental changes induced by the uplift of the QTP.<sup>6</sup> The period of population expansion (~13 to 8 Mya) was coincident with a pulse of rapid uplift in the northeastern QTP,<sup>97,98</sup> which also overlapped with adaptive radiation of the Cyprininae fishes ~23 to 16 Mya.<sup>18</sup> A growing body of evidence suggests that QTP uplift leads to extensive eco-environmental alterations, such as the onset of the Asian monsoon and the formation of a temperate continental climate.<sup>9,79,89,99,100</sup> The transition to temperate and arid weather might favour the expansion of ancestral *G. przewalskii*, allowing them to replace fishes living in warm and humid areas.<sup>7</sup>



The continuous population declines (from ~6 to 0.1 Mya) are considered consequences of the comprehensive geological, drainage, climatic and ecosystem changes during the major phase of QTP uplift, which restricted the survival and colonization of *G. przewalskii*.

The impact of WGD on evolution was extensive, providing genomic materials for the origin of evolutionary novelty and facilitating diversification.<sup>13</sup> The impact of WGD on evolution was extensive, providing genomic materials for the origin of evolutionary novelty and facilitating diversification.<sup>13</sup> Our analysis suggested that the WGD of *G. przewalskii* occurred at 2.25 Mya, which might confer their adaptive advantages to the environmental changes in the extensive uplift in the Qingzang movement (3.6–1.7 Mya).<sup>89,101,102</sup> Genes preserved post-WGD were not random, and were functionally associated with adaptation to extreme highland environments. Additionally, expression variations within gene pairs may have resulted in enhanced vigour and rapid adaptation to the novel conditions,<sup>103</sup> allowing ancient *G. przewalskii* to cope better than its diploid relatives with the highland environment. Therefore, the evolutionary mechanisms may have considerably shaped the tetraploid genome of *G. przewalskii* by preserved duplicate genes related to environmental adaptation from being lost. Therefore, deciphering the *G. przewalskii* genome will improve our knowledge on the genome evolution of polyploid fish species during eco-environmental changes.

## 5. Conclusions

In the current study, we generated a high-quality chromosome-level assembly of the *G. przewalskii* genome, providing a valuable genomic resource for comparative analysis across teleost fish. Our analysis underscored the importance of WGD, which could grant *G. przewalskii* adaptive advantages to the environmental changes during the QTP uplift. The extensive genomic changes, including positive selection and gene family expansion, facilitated the evolution of *G. przewalskii* to the extreme aquatic environment in the QTP. Additionally, we demonstrated that our assembly could serve as a reference for schizothoracine fishes, which will decipher their origination and evolution. Conclusively, the availability of the genomic resources of *G. przewalskii* will benefit future evolutionary and speciation studies on cyprinids, particularly regarding molecular adaptation, genomic conservation and phylogeny.

## Acknowledgements

We appreciate the editor and anonymous reviewers for their constructive suggestions and comments. We thank Dr. Delin Qi for his suggestions in revising the manuscript. We thank Ms Hongfang Qi for her help in sample collection.

## Supplementary data

Supplementary data are available at DNARES online.

## Ethics approval and consent to participate

The field investigation was authorized and supervised by the Qinghai Provincial Bureau of Fishery [The Authorized Capturing Permission of Aquatic Wildlife in the People's Republic of China (Qing) Wildlife Capturing (2018) No. 2]. All animal experiments were conducted following the procedures described in the 'Guidelines for animal care and use' manual approved by the Animal Care and Use Committee, Northwest Institute of Plateau Biology, Chinese Academy of Sciences.

## Authors' contributions

K.Z. conceived, designed and supervised the study. F.T. and S.L. conducted phylogenomic analysis, WGD and RNA-seq analysis. Z.B.Z. and Y.Z. performed gene analysis. Y.T. carried out DNA and RNA purification and flow cytometry. F.T. wrote the manuscript. K.Z. and S.L. revised manuscript. All authors commented on the manuscript and were involved in the interpretation of the preliminary data.

## Funding

This work was supported by National Science Foundation of China (31700325 and 31870365), The Genebank of Qinghai-Tibet Plateau Biota (2021-SF-SW1), Qinghai Science and Technology Major program 'Sanjiangyuan National Park Genome Project' and joint foundation from Chinese Academy of Sciences – People's Government of Qinghai Province on Sanjiangyuan National Park (LHZX-2020-01). Drs Fei Tian and Sijia Liu were supported by CAS 'Light of West China' Program. The field investigation was supported by Sino BON-Inland Water Fish Diversity Observation Network.

## Conflict of interest

None declared.

## Data availability

Sequencing raw data of PacBio sequence, Illumina sequence and Hi-C, and whole genome assembly of *G. przewalskii* were deposited in CNGB database under project number CNP0002001. RNA-seq data produced by the present study were deposited under accession numbers SRP284182 and SRP284021. Published *G. przewalskii* RNA-seq data used in this manuscript were deposited in NCBI SRA under accession Nos SRP136464, SRP045343, SRP045396, SRP045343, SRP045396 and SRP254169.

## References

1. Qiu, Q., Zhang, G., Ma, T., et al. 2012, The yak genome and adaptation to life at high altitude, *Nat. Genet.*, **44**, 946–9.
2. Cai, Q., Qian, X., Lang, Y., et al. 2013, Genome sequence of ground tit *Pseudopodoces humilis* and its adaptation to high altitude, *Genome Biol.*, **14**, R29.
3. Ge, R.-L., Cai, Q., Shen, Y.-Y., et al. 2013, Draft genome sequence of the Tibetan antelope, *Nat. Commun.*, **4**, 1858.
4. Gou, X., Wang, Z., Li, N., et al. 2014, Whole genome sequencing of six dog breeds from continuous altitudes reveals adaption to high-altitude hypoxia, *Genome Res.*, **24**, 1308–15.
5. Wu, Y. and Tan, Q. 1991, Characteristics of the fish-fauna of the characteristics of Qinghai-Xizang Plateau and its geological distribution and formation, *Curr. Zool.*, **37**, 135–52.
6. Cao Wenxuan, C.Y. and Wu Yunfei, Z.S. 1981, *Origin and Evolution of Schizothoracine Fishes in Relation to the Upheaval of the Qinghai-Tibetan Plateau*. Beijing, China: Science Press.
7. Mee-mann Chang, D.M. 2010, Ning Wang. In: Gu, H.Y., Long, M.Y. and Zhou, Z.H., eds. *DARWIN's Heritage Today*, p.60–75. Beijing, China: Higher Education Press.
8. He, D.K. and Chen, Y.F. 2007, Molecular phylogeny and biogeography of the highly specialized grade schizothoracine fishes (Teleostei: Cyprinidae) inferred from cytochrome b sequences, *Chin. Sci. Bull.*, **52**, 777–88.

9. Deng, T., Wang, X., Wu, F., et al. 2019, Review: implications of vertebrate fossils for paleo-elevations of the Tibetan Plateau, *Glob. Planet. Change*, **174**, 58–69.
10. Wanghe, K., Tang, Y., Tian, F., et al. 2017, Phylogeography of *Schizopygopsis stoliczkae* (Cyprinidae) in Northwest Tibetan Plateau area, *Ecol. Evol.*, **7**, 9602–12.
11. Zhao, K., Duan, Z., Peng, Z., et al. 2011, Phylogeography of the endemic *Gymnocypris chilianensis* (Cyprinidae): sequential westward colonization followed by allopatric evolution in response to cyclical *Pleistocene glaciations* on the Tibetan Plateau, *Mol. Phylogenet. Evol.*, **59**, 303–10.
12. Soltis, P.S. and Soltis, D.E. 2012, *Polyploidy and Genome Evolution*. Berlin: Springer Verlag.
13. Van de Peer, Y., Maere, S. and Meyer, A. 2009, The evolutionary significance of ancient genome duplications, *Nat. Rev. Genet.*, **10**, 725–32.
14. Glasauer, S.M.K. and Neuhauss, S.C.F. 2014, Whole-genome duplication in teleost fishes and its evolutionary consequences, *Mol. Genet. Genomics*, **289**, 1045–60.
15. Jaillon, O., Aury, J.-M., Brunet, F., et al. 2004, Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype, *Nature*, **431**, 946–57.
16. Arai, R. 2011, *Fish Karyotypes: A Check List*, Copeia.
17. Leggatt, R.A. and Iwama, G.K. 2003, Occurrence of polyploidy in the fishes, *Rev. Fish Biol. Fisher.*, **13**, 237–46.
18. Wang, X.Z., Gan, X.N., Li, J.B., Chen, Y.Y. and He, S.P. 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**, 1149–65.
19. Yang, L., Sado, T., Vincent Hirt, M., et al. 2015, Phylogeny and polyploidy: resolving the classification of cyprinine fishes (Teleostei: Cypriniformes), *Mol. Phylogenet. Evol.*, **85**, 97–116.
20. Wu, Y.F., Kang, B., Men, Q. and Wu, C.Z. 1999, Chromosome diversity of Tibetan fishes, *Zool. Res.*, **20**, 7.
21. Xu, P., Xu, J., Liu, G., et al. 2019, The allotetraploid origin and asymmetrical genome evolution of the common carp *Cyprinus carpio*, *Nat. Commun.*, **10**, 4625.
22. Lanzhou Branch of Chinese Academy of Sciences. 2008, *Evolution and Prediction of the Modern Environmental of Lake Qinghai*, p.10–19. Beijing, China: Science Press.
23. Tian, F., Tong, C., Feng, C.G., Wanghe, K.Y. and Zhao, K. 2017, Transcriptomic profiling of Tibetan highland fish (*Gymnocypris przewalskii*) in response to the infection of parasite ciliate *Ichthyophthirius multifiliis*, *Fish Shellfish Immun.*, **70**, 524–35.
24. Jiang, Y., Jiang, M., Sun, Y., et al. 2020, Comparisons of fecundity and reproductive hormone levels in three populations of *Gymnocypris przewalskii*, *Aquaculture*, **514**, 734449.
25. Liu, Y., Yao, Z., Sun, Z., et al. 2020, De novo assembly of a chromosome-level genome of naked carp (*Gymnocypris przewalskii*) reveals geographic isolation of Schizothoracine fishes in Qinghai, *Tibet Plateau Lift.*, doi:10.22541/au.159863199.96614555.
26. Dolezel, J., Bartos, J., Voglmayr, H. and Greilhuber, J. 2003, Nuclear DNA content and genome size of trout and human, *Cytometry A.*, **51**, 127–8.
27. Marcais, G. and Kingsford, C. 2011, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, *Bioinformatics*, **27**, 764–70.
28. Vurture, G.W., Sedlazeck, F.J., Nattestad, M., et al. 2017, GenomeScope: fast reference-free genome profiling from short reads, *Bioinformatics*, **33**, 2202–4.
29. Xiao, C.-L., Chen, Y., Xie, S.-Q., et al. 2017, MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads, *Nat. Methods*, **14**, 1072–4.
30. Walker, B.J., Abeel, T., Shea, T., et al. 2014, Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement, *PLoS One*, **9**, e112963.
31. Rao, S.S.P., Huntley, M.H., Durand, N.C., et al. 2014, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping, *Cell*, **159**, 1665–80.
32. Li, H. and Durbin, R. 2009, Fast and accurate short read alignment with Burrows–Wheeler transform, *Bioinformatics*, **25**, 1754–60.
33. Servant, N., Varoquaux, N., Lajoie, B.R., et al. 2015, HiC-Pro: an optimized and flexible pipeline for Hi-C data processing, *Genome Biol.*, **16**, 259.
34. Burton, J.N., Adey, A., Patwardhan, R.P., et al. 2013, Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions, *Nat. Biotechnol.*, **31**, 1119–25.
35. Rhie, A., Walenz, B.P., Koren, S. and Phillippy, A.M. 2020, Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies, *Genome Biol.*, **21**, 245.
36. Belser, C., Baurens, F.-C., Noel, B., et al. 2021, Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing, *Commun. Biol.*, **4**, 1047.
37. Lagesen, K., Hallin, P., Rødland, E.A., et al. 2007, RNAmmer: consistent and rapid annotation of ribosomal RNA genes, *Nucleic Acids Res.*, **35**, 3100–8.
38. Chan, P.P. and Lowe, T.M. 2019, tRNAscan-SE: searching for tRNA genes in genomic sequences, *Methods Mol. Biol.*, **1962**, 1–14.
39. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. and Eddy, S.R. 2003, Rfam: an RNA family database, *Nucleic Acids Res.*, **31**, 439–41.
40. Xu, Z. and Wang, H. 2007, LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons, *Nucleic Acids Res.*, **35**, W265–8.
41. Rho, M.N. and Tang, H.X. 2009, MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes, *Nucleic Acids Res.*, **37**, e143.
42. Lee, H., Lee, M., Mohammed Ismail, W., et al. 2016, MGEScan: a Galaxy-based system for identifying retrotransposons in genomes, *Bioinformatics*, **32**, 2502–4.
43. Edgar, R.C. and Myers, E.W. 2005, PILER: identification and classification of genomic repeats, *Bioinformatics*, **21**, 1152–8.
44. Price, A.L., Jones, N.C. and Pevzner, P.A. 2005, De novo identification of repeat families in large genomes, *Bioinformatics*, **21**(Suppl 1), i351–8.
45. Chen, N. 2004, Using RepeatMasker to identify repetitive elements in genomic sequences, *Curr. Protoc. Bioinformatics*, **Chapter 4**, Unit 4.10.
46. Bao, W.D., Kojima, K.K. and Kohany, O. 2015, Repbase update, a database of repetitive elements in eukaryotic genomes, *Mobile DNA-UK.*, **6**, 11.
47. Jurka, J., Kapitonov, V.V., Pavlicek, A., et al. 2005, Repbase update, a database of eukaryotic repetitive elements, *Cytogenet. Genome Res.*, **110**, 462–7.
48. Stanke, M., Keller, O., Gunduz, I., et al. 2006, AUGUSTUS: ab initio prediction of alternative transcripts, *Nucleic Acids Res.*, **34**, W435–9.
49. Korf, I. 2004, Gene finding in novel genomes, *BMC Bioinformatics*, **5**, 59.
50. Majoros, W.H., Pertea, M. and Salzberg, S.L. 2004, TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders, *Bioinformatics*, **20**, 2878–9.
51. Borodovsky, M. and Lomsadze, A. 2011, Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES, *Curr Protoc Bioinformatics*, **Chapter 4**, Unit 4.6.1.
52. Altschul, S.F., Madden, T.L., Schäffer, A.A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–402.
53. Chen, S.F., Zhou, Y.Q., Chen, Y.R. and Gu, J. 2018, fastp: an ultra-fast all-in-one FASTQ preprocessor, *Bioinformatics*, **34**, i884–90.
54. Kim, D., Langmead, B. and Salzberg, S.L. 2015, HISAT: a fast spliced aligner with low memory requirements, *Nat. Methods*, **12**, 357–60.
55. Pertea, M., Kim, D., Pertea, G.M., Leek, J.T. and Salzberg, S.L. 2016, Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown, *Nat. Protoc.*, **11**, 1650–67.

56. Campbell, M.S., Holt, C., Moore, B. and Yandell, M. 2014, Genome annotation and curation using MAKER and MAKER-P, *Curr. Protoc. Bioinformatics*, **48**, 4.11.11–39.
57. Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. and Robles, M. 2005, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, *Bioinformatics*, **21**, 3674–6.
58. Li, L., Stoeckert, C.J. and Roos, D.S. 2003, OrthoMCL: identification of ortholog groups for eukaryotic genomes, *Genome Res.*, **13**, 2178–89.
59. De Bie, T., Cristianini, N., Demuth, J.P. and Hahn, M.W. 2006, CAFE: a computational tool for the study of gene family evolution, *Bioinformatics*, **22**, 1269–71.
60. Yu, G.C., Wang, L.G., Han, Y.Y. and He, Q.Y. 2012, clusterProfiler: an R package for comparing biological themes among gene clusters, *Omic*, **16**, 284–7.
61. Edgar, R.C. 2004, MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.*, **32**, 1792–7.
62. Huelsenbeck, J.P. and Ronquist, F. 2001, MRBAYES: Bayesian inference of phylogenetic trees, *Bioinformatics*, **17**, 754–5.
63. Yang, Z. 1997, PAML: a program package for phylogenetic analysis by maximum likelihood, *Comput. Appl. Biosci.*, **13**, 555–6.
64. Zhang, J.Z., Nielsen, R. and Yang, Z.H. 2005, Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level, *Mol. Biol. Evol.*, **22**, 2472–9.
65. Zhang, Z., Xiao, J., Wu, J., et al. 2012, ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments, *Biochem. Biophys. Res. Commun.*, **419**, 779–81.
66. Wang, D., Zhang, Y., Zhang, Z., Zhu, J. and Yu, J. 2010, KaKs\_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies, *Genomics Proteomics Bioinformatics*, **8**, 77–80.
67. Tang, H., Bowers, J.E., Wang, X., et al. 2008, Synteny and collinearity in plant genomes, *Science*, **320**, 486–8.
68. Yang, J., Chen, X., Bai, J., et al. 2016, The Sinocyclocheilus cavefish genome provides insights into cave adaptation, *BMC Biol.*, **14**, 1.
69. Xiao, S., Mou, Z., Fan, D., et al. 2020, Genome of Tetraploid Fish *Schizothorax o'connori* provides insights into early re-diploidization and high-altitude adaptation, *Iscience*, **23**, 101497.
70. Tang, Y., Li, C., Wanghe, K., et al. 2019, Convergent evolution misled taxonomy in schizothoracine fishes (Cypriniformes: Cyprinidae), *Mol. Phylogenet. Evol.*, **134**, 323–37.
71. McKenna, A., Hanna, M., Banks, E., et al. 2010, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, *Genome Res.*, **20**, 1297–303.
72. Stamatakis, A. 2006, RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models, *Bioinformatics*, **22**, 2688–90.
73. Li, H. and Durbin, R. 2011, Inference of human population history from individual whole-genome sequences, *Nature*, **475**, 493–6.
74. De-Lin, Q.I. 2004, Preliminary studies on chromosome karyotype and polyploidy of Qinghai-lake Naked Carp, *J. Qinghai Univ.*, **22**, 44–8.
75. Yan, X., Shi, J., Sun, X. and Liang, L. 2007, Study on the karyotype of *Gymnocypris przewalskii*, *J. Northeast Agric. Univ.*, **38**, 645–8.
76. Xu, P., Zhang, X., Wang, X., et al. 2014, Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*, *Nat. Genet.*, **46**, 1212–9.
77. Sun, L., Gao, T., Wang, F., et al. 2020, Chromosome-level genome assembly of a cyprinid fish *Onychostoma macrolepis* by integration of nanopore sequencing, Bionano and Hi-C technology, *Mol. Ecol. Resour.*, **20**, 1361–71.
78. Li, J.J. and Fang, X.M. 1999, Uplift of the Tibetan Plateau and environmental changes, *Chin. Sci. Bull.*, **44**, 2117–24.
79. Deng, T. and Ding, L. 2015, Paleolimetry reconstructions of the Tibetan Plateau: progress and contradictions, *Natl. Sci. Rev.*, **2**, 417–37.
80. Romero, M.F., Fulton, C.M. and Boron, W.F. 2004, The SLC4 family of HCO3- transporters, *Pflügers Archiv-Eur. J. Physiol.*, **447**, 495–509.
81. Bai, L., Liu, B., Ji, C., et al. 2019, Hypoxic and cold adaptation insights from the Himalayan Marmot Genome (vol 11, pg 519, 2019), *Iscience*, **11**, 505–7.
82. Verri, T., Terova, G., Romano, A., Barca, A. and Saroglia, M. 2012, *The SoLute Carrier (SLC) Family Series in Teleost Fish*. Oxford: Functional Genomics in Aquaculture.
83. Cutler, C.P. 2017, *Water Balance and Aquaporin*. Elsevier: Reference Module in Life Sciences.
84. Zhou, B., Qi, D., Liu, S., et al. 2022, Physiological, morphological and transcriptomic responses of Tibetan naked carps (*Gymnocypris przewalskii*) to salinity variations, *Comp. Biochem. Physiol. Part D Genomics Proteomics*, **42**, 100982.
85. Harrison, T.M., Copeland, P., Kidd, W.S.F. and Yin, A. 1992, Raising Tibet, *Science*, **255**, 1663–70.
86. Mulch, A. and Chamberlain, C.P. 2006, Earth science - the rise and growth of Tibet, *Nature*, **439**, 670–1.
87. Erchie, W. 2013, Evolution of the Tibetan Plateau: as constrained by major tectonic-thermo events and a discussion on their origin, *Chin. J. Geol.*, **48**, 334–53.
88. Wu, Y., Cui, Z., Liu, G., et al. 2001, Quaternary geomorphological evolution of the Kunlun Pass area and uplift of the Qinghai-Xizang (Tibet) Plateau, *Geomorphology*, **36**, 203–16.
89. Zhisheng, A., Kutzbach, J.E., Prell, W.L. and Porter, S.C. 2001, Evolution of Asian monsoons and phased uplift of the Himalayan Tibetan plateau since Late Miocene times, *Nature*, **411**, 62–6.
90. Zhou, Z., Deng, T., Wu, F., Su, T. and Wang, S. 2019, Significant shift in the terrestrial ecosystem at the Paleogene/Neogene boundary in the Tibetan Plateau, *Chin. Sci. Bull.*, **64**, 2894–906.
91. Howe, K., Clark, M.D., Torroja, C.F., et al. 2013, The zebrafish reference genome sequence and its relationship to the human genome, *Nature*, **496**, 498–503.
92. Lynch, M. and Conery, J.S. 2000, The evolutionary fate and consequences of duplicate genes, *Science*, **290**, 1151–5.
93. Force, A., Lynch, M., Pickett, F.B., et al. 1999, Preservation of duplicate genes by complementary, degenerative mutations, *Genetics*, **151**, 1531–45.
94. Conant, G.C. and Wolfe, K.H. 2008, Turning a hobby into a job: how duplicated genes find new functions, *Nat. Rev. Genet.*, **9**, 938–50.
95. Liu, H.-P., Xiao, S.-J., Wu, N., et al. 2019, The sequence and de novo assembly of *Oxygymnocypris stewartii* genome, *Sci. Data*, **6**, 190009.
96. Chen, L., Li, B., Chen, B., et al. 2021, Chromosome-level genome of *Poropuntius huangchuchieni* provides a diploid progenitor-like reference genome for the allotetraploid *Cyprinus carpio*, *Mol. Ecol. Resour.*, **21**, 1658–69.
97. Ritts, B.D., Yue, Y., Graham, S.A., et al. 2008, From sea level to high elevation in 15 million years: uplift history of the northern Tibetan Plateau margin in the Altun Shan, *Am. J. Sci.*, **308**, 657–78.
98. Yuan, W.M., Dong, J.Q., Wang, S.C. and Carter, A. 2006, Apatite fission track evidence for Neogene uplift in the eastern Kunlun Mountains, northern Qinghai-Tibet Plateau, China, *J. Asian Earth Sci.*, **27**, 847–56.
99. Li, Q., Xie, G., Takeuchi, G.T., et al. 2014, Vertebrate fossils on the roof of the world: biostratigraphy and geochronology of high-elevation Kunlun Pass Basin, northern Tibetan Plateau, and basin history as related to the Kunlun strike-slip fault, *Palaeogeogr. Palaeoclimatol. Palaeoecol.*, **411**, 46–55.
100. Fang, X., Wang, J., Zhang, W., et al. 2016, Tectonosedimentary evolution model of an intracontinental flexural (foreland) basin for paleoclimatic research, *Glob. Planet. Change*, **145**, 78–97.
101. Li, J.J., Fang, X.M. and Ma, H.Z. 2001, Late Cenozoic intensive uplift of Qinghai-Xizang Plateau and its impacts on environments in surrounding area, *Quat. Ences.*, **21**, 381–391.
102. Zhijiu, C., Yun, H., Quanzhou, G. and Huailu, C. 1996, The process and environment of Palaeokarst in the Northeast area of Qinghai-Xizang Plateau, *Acta Geograph. Sin.*, **28**, 53–59.
103. Rieseberg, L.H., Raymond, O., Rosenthal, D.M., et al. 2003, Major ecological transitions in wild sunflowers facilitated by hybridization, *Science*, **301**, 1211–6.