Chromosome-level Genome Assembly Provides Insight Into the Evolution of Chromosomes and the Globin Gene Superfamily in *Gymnocypris Eckloni*

Fayan Wang
Qinghai University
Lihan Wang
Qinghai University
Dan Liu
Qinghai University
Qiang Gao
Qinghai University
Miaomiao Nie
Qinghai University
Shihai Zhu
Qinghai University
Yan Chao
Qinghai University
Xi Yang
Qinghai University
Chaojie Yang
Qinghai University
Cunfang Zhang
Qinghai University
Rigui Yi
Qinghai University
Weilin Ni
Qinghai University
Guangxin Li
Qinghai University
Qichang Chen
Qinghai University
Fei Tian
Northwest Institute of Plateau Biology, Chinese Academy of Sciences
Kai Zhao

Northwest Institute of Plateau Biology, Chinese Academy of Sciences
Delin Qi (delinqi@126.com)
Qinghai University

Research Article

Keywords: Gymnocypris eckloni, chromosome-level genome assembly, whole-genome duplication, chromosome evolution, globin gene superfamily

Posted Date: January 19th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1227329/v1

Abstract

Background: *Gymnocypris eckloni* is widely distributed in isolated lakes and the upper reaches of the Yellow River and is well adapted to the Qinghai-Tibetan plateau (QTP) aqueous environment.

Results: Here, we report the chromosome-level genome of *G. eckloni*, which is the first genome reported for this highly specialized schizothoracine fish. The final genome was assembled into 918.68 Mb sequences using PacBio long-read and Illumina short-read sequences, including 711 scaffolds with a contig N50 length of 4.19 Mb and a scaffold N50 length of 43.54 Mb. The assembled sequences were further anchored and orientated onto 23 pseudo-chromosomes. In total, 23,157 genes were annotated, representing 94.80% of the total predicted protein-coding genes. We conducted comparative genomic analyses and found that species-specific and rapidly expanding gene families in *G. eckloni* are mainly involved in Dorso-ventral axis formation, ABC transporters, Peroxisome and Herpes simplex virus 1 infection. The phylogenetic analysis showed that *G. eckloni* was most closely related to *C. carpio* with an estimated divergence time of ~34.8 million years ago (Ma), which coincides with the first and second tectonic uplifts of the QTP. We also found that *G. eckloni* underwent an additional whole-genome duplication (WGD) event (4R) at ~14.2 Ma, which had a significant effect on the evolution of chromosomes and the globin gene superfamily in *G. eckloni*, thereby facilitating the adaptation of *G. eckloni* to the plateau water environment.

Conclusions: The high-quality genome assembly of *G. eckloni* will serve as a valuable genomic resource for future research on the evolution and ecology of the schizothoracine fish.

Background

The Qinghai-Tibetan plateau (QTP) is characterized by high altitude, low oxygen partial pressure (hypoxia), low temperatures, dramatic temperature fluctuations, and high UV radiation. It is the highest and one of the youngest plateaus on Earth. Schizothoracine fish (Teleostei: Cyprinidae) are the largest and most diverse taxon within the QTP ichthyofauna [1, 2]. These fish have been accepted as ideal models for studying the molecular mechanisms underlying the adaptation to harsh environments [3–5]. Previous studies have shown that the karyotypes of Schizothoracinae range from 90 to ± 446 and that almost all species were polyploid [6–9]. A recent genomic study confirmed that *Schizothorax o'connori* of Schizothoracinae was a young tetraploid that underwent a fourth whole-genome duplication (4R WGD) after the teleost-specific third WGD (3R WGD) [10]. Other studies indicated that the globin gene superfamily, toll-like receptor family, and interferon regulatory factors in a representative species from this subfamily underwent adaptive evolution in response to the plateau environment, specifically gene loss, and gain events as a result of genome and/or gene duplications [5, 11–13]. *Gymnocypris eckloni* is a representative species of the highly specialized schizothoracine fish that is widely distributed in isolated lakes and the upper reaches of the Yellow River, and is very well adapted to the plateau's aqueous environment [1, 2]. Investigating the genomic evolution of *G. eckloni* may shed light on the underlying molecular mechanisms involved in high-altitude adaptations in schizothoracine fish of the QTP.

In the present study, we integrated PacBio long-read sequencing, Illumina short-read sequencing, and high-throughput chromosome conformation capture (Hi-C) technology to generate a high-quality chromosome-level reference genome for *G. eckloni*. Comparative analyses with 13 other vertebrate genomes were conducted, focusing particularly on phylogenetic relationships, divergence time, and the expansion and contraction of gene families. Then, the *G. eckloni* WGD event (4R) was determined, and the evolution of chromosomes and the globin gene superfamily in *G. eckloni* were analyzed. The reference genome obtained in this study will provide a foundation for future investigations on the evolution and adaptation of schizothoracine fish.

Results And Discussion

Genome sequencing and assembly of G. eckloni

A total of 312.2 Gb PacBio reads, 215.7 Gb Illumina short reads, and 253.7 Gb Hi-C reads were generated in a female *G. eckloni* (Supplementary Fig. 1, Additional file 1; Supplementary Table 1, Additional file 2). After filtering, 215.2 Gb (231.2× coverage) of clean Illumina data were retained to perform a genome survey (Supplementary Table 1, Additional file 2). Based on the k-mer (k = 17) depth frequency distribution analysis, the genome size of *G. eckloni* was estimated to be 927.13 Mb (Supplementary Fig. 2, Additional file 1; Supplementary Table 2, Additional file 2). For genome assembly, a total of 239 Gb clean PacBio long reads (334.6× coverage) with an average length of 23,706 bp were generated (Supplementary Table 1, Additional file 2). The data were assembled using wtdbg2 followed by Quiver and Pilon polishing using the 215.2 Gb (231.2× coverage) of Illumina HiSeq clean reads, which produced a 918.45 Mb genome assembly with a contig N50 size of 4.19 Mb (Supplementary Table 3, Additional file 2). The assembled genome size was slightly larger than most other teleosts (~700 Mb) [14, 15], but obviously smaller than the genomes of *Cyprinus carpio* (1.83 Gb) and *Carassius auratus* (1.85 Gb) [16, 17]. The assembled sequences were further anchored and orientated onto 23 pseudo-chromosomes using Hi-C data. The 23 pseudo-chromosomes ranged in size from 15.91 to 89.39 Mb (Fig. 1a, Table 1), covering ~98.52% of the whole genome. The heatmap of the pseudo-chromosome crosstalk confirmed the high quality of our genome assembly (Fig. 1a, b). Finally, the *G. eckloni* genome was obtained with 711 scaffolds and a total length of 918,681,488 bp, a contig N50 of 4.19 Mb, and scaffold N50 of 43.54 Mb (Supplementary Table 3, Additional file 2).

Table 1
Statistics of the chromosome assemblies using Hi-C data.

Pseudo-Chromosomes	Sequence IDs	Cluster Numbers	Sequences Lengths
Chr01	Hic_asm_0	170	44,161,858
Chr02	Hic_asm_1	20	25,669,045
Chr03	Hic_asm_2	102	34,715,927
Chr04	Hic_asm_3	125	61,913,669
Chr05	Hic_asm_4	6	28,665,261
Chr06	Hic_asm_5	66	43,543,958
Chr07	Hic_asm_6	154	61,756,641
Chr08	Hic_asm_7	118	47,622,867
Chr09	Hic_asm_8	62	27,783,661
Chr10	Hic_asm_9	69	32,896,108
Chr11	Hic_asm_10	82	21,739,851
Chr12	Hic_asm_11	29	32,200,481
Chr13	Hic_asm_12	42	29,129,546
Chr14	Hic_asm_13	56	28,882,651
Chr15	Hic_asm_14	232	70,525,535
Chr16	Hic_asm_15	195	89,391,071
Chr17	Hic_asm_16	193	16,898,605
Chr18	Hic_asm_17	54	15,913,822
Chr19	Hic_asm_18	68	33,789,651
Chr20	Hic_asm_19	190	46,494,833
Chr21	Hic_asm_20	100	32,874,627
Chr22	Hic_asm_21	96	39,527,105
Chr23	Hic_asm_22	101	39,007,358

We conducted a sequence consistency assessment using BWA software. The results showed that 93.40% of the reads could be mapped, covering 96.34% of the assembled genome when the Illumina short reads were mapped to the assembled genome (Supplementary Table 4, Additional file 2). The completeness of

the final assembled genome was assessed using CEGMA and BUSCO analyses. The CEGMA analysis revealed that 221 conserved genes (89.11% of the core eukaryotic genes) supported the completeness of the assembled genome (Supplementary Table 5, Additional file 2). The BUSCO analysis showed that 88.4% (single-copy genes: 82.3%, duplicated genes: 6.1%) of the 2,586 single-copy genes were identified as complete, 2.4% were fragmented, and 9.2% were missing from the assembled genome (Supplementary Table 6, Additional file 2). Overall, these analyses suggested that we assembled a high quality and chromosome-level genome of *G. eckloni*.

Genome annotation

The repeat sequences prediction analysis revealed that 47.63% of the *G. eckloni* genome was annotated as repetitive elements (Supplementary Table 7, Additional file 2), of which LTRs were the most abundant with a total length of 356.79 Mb, accounting for 38.84% of the whole genome. SINEs were the rarest with a total length of 2.37 Mb and represented 0.26% of the whole genome (Supplementary Table 8, Additional file 2).

Using a comprehensive strategy based on homologous sequence searches, *ab initio* gene predictions, and RNA-seq-derived evidence from nine tissue and blood samples, a total of 24,430 protein-coding genes were predicted in the genome of *G. eckloni*. The average transcript length was 1,530.87 bp with an average coding sequence (CDS) length of 1,536.71 bp. The average exon number per gene was 8.88 with an average exon length of 173.00 bp and average intron length of 1,862.69 bp (Supplementary Fig. 3, Additional file 1; Supplementary Table 9, Additional file 2). The comparison of gene numbers, average coding sequence lengths, and average gene lengths with eight other fish species indicated that our annotations were comprehensive (Supplementary Fig. 3, Additional file 1; Supplementary Table 10, Additional file 2).

A total of 23,157 genes were annotated using at least one public database and represented 94.80% of the total predicted protein-coding genes (Table 2). The functional annotations found that 20,432 (83.60%), 23,110 (94.60%), and 21,539 (88.20%) genes had significant hits with proteins catalogued in Swissprot, NR, and InterPro, respectively. A total of 15,281 (62.60%), 20,593 (84.30%), and 19,290 (79.00%) genes were annotated to GO, KEGG, and Pfam, respectively (Supplementary Fig. 4, Additional file 1; Table 2). Additionally, 1,717 miRNAs, 12,157 tRNAs, 1,780 rRNAs, and 1,152 snRNAs were identified, which had average lengths of 116.58, 75.00, 178.65, and 133.32 bp, respectively (Supplementary Table 11, Additional file 2). Taken together, these analyses suggested a satisfying level of completeness and accuracy of genome annotation.

Table 2
The number of genes with functional classifications in *G. eckloni*.

Database	Number	Percent (%)
SwissProt	20,432	83.60
NR	23,110	94.60
KEGG	20,593	84.30
InterPro	21,539	88.20
GO	15,281	62.60
Pfam	19,290	79.00
Annotated	23,157	94.80
Unannotated	1,273	5.20
Total	24,430	94.80

Gene family analysis and phylogenetics of G. eckloni

To examine *G. eckloni* evolution, we conducted a comparative genomics analysis using the *G. eckloni* genome obtained from this study and 13 other vertebrate genomes downloaded from the Ensembl database (*Astyanax mexicanus, Ictalurus punctatus, Danio rerio, C. carpio, Ctenopharyngodon idella, Oreochromis niloticus, Oryzias latipes, Takifugu rubripes, Gallus gallus, Homo sapiens, Mus musculus, Xenopus tropicalis, and Petromyzon marinus*). A total of 24,619 gene families were identified among the 14 species (Supplementary Fig. 5, Additional file 1; Supplementary Table 12, Additional file 2), of which 2,739 core gene families were shared by all 14 species (Supplementary Fig. 6, Additional file 1). Compared to other species, we found 1,488 *G. eckloni*-specific genes, classified into 856 gene families. Further analysis revealed that 210 genes of the species-specific genes were significantly enriched five KEGG pathways, including Dorso-ventral axis formation, Peroxisome, ABC transporters, Endocytosis and Herpes simplex virus 1 infection.

A phylogenetic tree was constructed using 597 single-copy orthologs with one-to-one correspondence in the different genomes, indicating that *G. eckloni* clustered together with *C. carpio* (Supplementary Fig. 7, Additional file 1). According to the time-calibrated phylogeny, the age of the most recent common ancestor (MRCA) of the teleost fish was estimated to be 211.8–254.1 Ma. The *G. eckloni* with the closest relationship to *C. carpio* shared an MRCA at 34.8 Ma (26.0–41.4, Fig. 2a). Fluctuations in the ecogeographical environment and major hydrographic formation occurred during uplifts of the QTP and are generally accepted to have significantly affected speciation events of schizothoracine fish [2, 18]. Our estimated divergence time for *G. eckloni* and *C. carpio* is similar to those obtained for other speciation

events of congeners in family Cyprininae [19, 20], which is consistent with the first and second tectonic uplifts of the QTP [21].

Analysis of the expansion and contraction of the gene families revealed that there were 464 (1650 genes) expanded and 743 (192 genes) contracted gene families in *G. eckloni* when compared to its MRCA (Fig. 2a). Further functional enrichment analysis of the expanded gene families highlighted 62 significantly enriched GO terms ($p < 0.05$) and 25 KEGG pathways ($p < 0.05$) (Supplementary Table 13, Additional file 3), while the contracted gene families highlighted 43 significant GO terms ($p < 0.05$) and 23 KEGG pathways ($p < 0.05$) (Supplementary Table 13, Additional file 3). The 25 KEGG pathways, including 625 genes from these expanded gene families, revealed that they were mainly classified as ABC transporters, Peroxisome, Herpes simplex virus 1 infection, Staphylococcus aureus infection, Axon guidance, Dorso-ventral axis formation, Pertussis, Legionellosis, Rap1 signaling pathway and so on (Supplementary Table 13, Additional file 3). Intriguingly, some of these pathways overlapped with pathways that species-specific *G. eckloni* genes were involved in, such as the Dorso-ventral axis formation, ABC transporters, Peroxisome and Herpes simplex virus 1 infection. The 23 KEGG pathways, including 189 genes from these contracted gene families, revealed that they were mainly classified as Tight junction, Systemic lupus erythematosus, Pathogenic Escherichia coli infection, Gap junction, Alcoholism, Pertussis, Ascorbate and aldarate metabolism, NOD-like receptor signaling pathway and so on (Supplementary Table 14, Additional file 3). Thus, it is likely that the expanded and contracted gene families played important roles in the adaptation of *G. eckloni* to the plateau water environment.

WGD events and chromosome evolution in G. eckloni

To determine the date of the *G. eckloni* WGD event (4R), we used inter- and intra-genomic colinear genes and calculated their synonymous substitution rates (*K*s values) (Fig. 2b). The *K*s distribution of homologous gene pairs in syntenic blocks indicated that the peak *K*s for the *G. eckloni* genome was 0.10. Based on a *K*s rate of 3.51×10^{-9} substitutions per synonymous site per year [22], we estimated that the latest WGD (4R) happened at ~14.2 Ma, which was consistent with the most recent genome duplication time of goldfish and later than the divergence time between *G. eckloni* and *C. carpio*. Similarly, an analysis of the fourfold synonymous third codon transversion (4dTv) provided additional evidence for an extra WGD event (Supplementary Fig. 8, Additional file 1).

Previous studies showed that there are several typical karyotypes in schizothoracine fish, including 2n = 90, 92, 94, 98, 148, and ± 446 [7]. Almost all species from Schizothoracinae that have been karyologically investigated are polyploid, including tetraploid and hexaploid [6, 7, 9]. The karyotype of *G. eckloni* was reported as 2n = 94 [7]. Along with the chromosome number, which is roughly twice as large as diploid cyprinid, including *D. rerio* (2n = 50), *C. idellus* (2n = 48), *Hypophthalmichthys molitrix* (2n = 48), and *Hypophthalmichthys nobilis* (2n = 48) [23, 24], *G. eckloni* is likely a tetraploid species. The *D. rerio* genome allowed us to identify potential fusion and fission events that shaped the *G. eckloni* genome since the ancestral karyotype. The chromosome synteny comparison of *G. eckloni* with *D. rerio* as a reference confirmed that four chromosome fusion events occurred in *G. eckloni* (Fig. 1c). Notably, chromosome 22

in *D. rerio* involved both fusion and fission events. The synteny analysis revealed a high level of collinearity between the chromosome-level genomes of *G. eckloni* and *D. rerio*, and although some were small, inter-chromosomal translocations indicated that the overall gene order in the *G. eckloni* genome remained very stable after its divergence from zebrafish (Fig. 1c). Previous studies corroborated that the *C. carpio* and *C. auratus* genome have tetraploidized due to an additional round of genome duplication (4R) and, thus, have 50 pairs of chromosomes (2n = 4x = 100) [16, 17, 22, 25], which is twice as many chromosomes as the ancestral cyprinid, *D. rerio* (2n = 50). All things considered, we propose that the latest WGD (4R), chromosome fusions, fissions, and deletions were a result of the formation of karyotypes with 94 chromosomes in *G. eckloni*.

Evolution of the globin gene superfamily in G. eckloni

The globin repertoire of extant vertebrates is the product of successive genome and gene duplication events followed by differential gene retention among lineages [26–28]. Fish endemic to the QTP is comparatively well adapted to aquatic environments with low oxygen partial pressure (hypoxia), in which the globin superfamily has played an important role [5, 29, 30]. Analyses of the *G. eckloni* genome revealed 25 globin genes distributed on five chromosomes (Supplementary Table 12, Additional file 2). An NJ phylogenetic tree was constructed based on the alignment of all the globins from *G. eckloni* and *D. rerio*, which showed well-supported monophyletic clades of *Ngb*, *Cygb*, *Mb*, *XGb*, and *α-* and *β-Hb* (Fig. 3a). Sequence comparisons and phylogenetic analyses identified one *Ngb*, one *Mb*, one *XGb*, and two paralogous *Cygb* genes (*Cygb*1 and *Cygb*2) in *G. eckloni*, which was consistent with a previous study on *D. rerio* [26] (Supplementary Fig. 9, Additional file 1). Moreover, the *Hb* gene repertoires of *G. eckloni* differed from *D. rerio*, based on the phylogenetic relationships and previous studies [31–33]. *Hb* genes expressed at the embryonic/larval and adult stages were denoted by “e” and “a,” respectively. In the *G. eckloni* genome, we identified 11 *a-Hb* genes, including seven intact embryonic *a-Hb* (*Hbae*1.1, *Hbae*1.2, *Hbae*1.4, *Hbae*1.5, *Hbae*1.6, *Hbae*4, and *Hbae*5), two intact adult *a-Hb* (*Hbaa*1 and *Hbaa*2), six intact embryonic *β-Hb* (*Hbbe*1.1, *Hbbe*1.2, *Hbbe*1.4, *Hbbe*1.5, *Hbbe*1.6, and *Hbbe*3), two intact adult *β-Hb* (*Hbba*1and *Hbba*2), and three *a-* and *β-Hb* incomplete or pseudogenes (*Hbae*1.3, *Hbbe*1.3, and *Hbaa*-like). Like zebrafish, all *Hb* genes of *G. eckloni* were distributed on two distinct chromosomes (Fig. 3b). Chromosome 20 contained the major globin locus with 17 *α-* and *β-Hb* genes and chromosome 05 housed the minor globin locus with three *α-* and *β-Hb* genes. There were six duplicate copies in the *Hbae*1 and *Hbbe*1 genes, which was three more genes than in the zebrafish genome. Compared to the zebrafish genome, *Hbae*3, *Hbbe*3, and one copy of *Hbba*1-*Hbbe*1 gene pairs were lost in the *G. eckloni* genome. Phylogenetic analyses revealed that all the *α-* and *β-Hb* genes, except the *Hbaa*-like gene, in the *G. eckloni* genome appeared to have 1:1 orthologs in zebrafish (*Hbae*1, *Hbae*4, *Hbae*5, *Hbaa*1, *Hbaa*2, *Hbbe*1, *Hbbe*3, *Hbba*1, and *Hbba*2). These findings suggested that the latest WGD (4R) and a recent small-scale gene duplication event after WGD contributed to gene duplication and deletion in the *α-* and *β-Hb* gene families in *G. eckloni*, which facilitated the adaptation of *G. eckloni* to the plateau water environment.

Conclusions

Using PacBio long-read sequencing data, Illumina short-read sequences, RNA-seq data and high-throughput chromosome conformation capture technology we have generated a high-quality draft genome assembly and annotation of the *G. eckloni* genome. Comparative analyses focused particularly on phylogenetic relationships, divergence time, and the expansion and contraction of gene families, together with the synteny analysis could deepen our understanding of the origin and speciation, as well as species polyploidization, and the evolution of chromosomes and the globin gene superfamily of the Schizothoracine fish endemic to the QTP. In conclusion, these results lay the foundation for future research on the evolution and ecology of the schizothoracine fish.

Materials And Methods

Experimental fish and sequencing

G. eckloni genomic DNA were extracted from the muscle samples of healthy female individuals obtained from the Native Fish Artificial Proliferation and Release Station, Xunhua, Qinghai Province, China (Supplementary Fig. 1, Additional file 1). For genome assembly, two libraries with insert sizes of 300 bp and 20 kb were separately constructed using an Illumina TruSeq Nano DNA Library Prep Kit and SMRT bell Template Prep Kit. The two libraries were subsequently sequenced using an Illumina HiSeq X Ten instrument and a PacBio Sequel platform [34]. To conduct chromosome-level assembly of the *G. eckloni* genome, a Hi-C library was generated using the *Mbo* I restriction enzyme following previously described standard protocol with minor modifications [35].

De novo assembly and evaluation

We used the k-mer method to survey the genomic features of the *G. eckloni*. The k-mer count histogram was obtained from Illumina paired-end sequencing data using Jellyfish v2.99 [36]. Wtdbg2 v2.5 was used to obtain the best assembly with the highest continuity and accuracy of the genome [37]. Quiver [38] was used to polish the genome using long sequencing data followed by polishing with short reads using Pilon [39].

Hi-C technology was applied to conduct the chromosome-level genome assembly of *G. eckloni*. Clean reads sequenced from the Hi-C library were aligned to the contig-level genome with an end-to-end algorithm implemented in Bowtie v2.3.5 according to the Hi-C-Pro strategy [40, 41]. Juicer v1.6.2 and 3D *de novo* assembly (3D-DNA) pipelines were used to assemble the contigs into the chromosome-level genome [42, 43]. The completeness of the genome was evaluated by analyzing single-copy orthologues, which was implemented by benchmarking universal single-copy orthologues using Benchmarking Universal Single-Copy Orthologs (BUSCO) v3.0.2 [44]. Synteny between the *G. eckloni* and *D. rerio* genomes was analyzed using TBtools v1.09854 [45].

Assessment of genome assembly

The assembled genome was evaluated by mapping reads from paired-end (PE) libraries with short insert-sizes to the scaffolds using BWA-MEM v0.7.12 [46]. BUSCO v4.0.5 was used to evaluate the integrity of the *G. eckloni* genome [44]. The completeness of conserved genes in the *G. eckloni* genome was assessed by Core Eukaryotic Genes Mapping Approach (CEGMA) [47].

Repeat annotation

A combined strategy using homology alignments and *de novo* searches to identify whole-genome repeats was applied in our repeat annotation pipeline. Tandem repeats were extracted using TRF(http://tandem.bu.edu/trf/trf.html)by *ab initio* prediction. For homolog prediction, Repbase (http://www.girinst.org /repbase) employing RepeatMasker (http://www.repeatmasker.org/) software and its in-house scripts (RepeatProteinMask) with default parameters was used to extract repeat regions. Additionally, *ab initio* prediction based on the *de novo* repetitive elements database was conducted by LTR_FINDER (http://tlife.fudan.edu.cn/ltr_finder/), RepeatScout (http://www.repeatmasker.org/), and RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html) with default parameters. Then, all repeat sequences with lengths > 100 bp and gap 'N' < 5% were used to construct the raw transposable element (TE) library. A custom library (a combination of Repbase and our *de novo* TE library, which was processed by uclust to yield a non-redundant library) was supplied to RepeatMasker for DNA-level repeat identification.

Annotation of protein-coding genes

Gene predictions were conducted through a combination of homology, *de novo*, and transcriptome-based prediction methods. For homology-based predictions, the protein sequences of seven fish species, including *Oryzias latipes*, *Ctenopharyngodon idellus*, *Ictalurus punctatus*, *Cyprinus carpio*, *Takifugu rubripes*, *Danio rerio*, and *Astyanax mexicanus*, were downloaded from Ensembl. Protein sequences were aligned to the genome using TblastN v2.2.26 with an e-value of $1e^{-5}$ [48]. Then, matching proteins were aligned to homologous genome sequences for accurate spliced alignments using GeneWise v2.4.1 [49], which was subsequently used to predict gene structure of each protein region. RNA-sequencing data derived from nine tissues and blood samples were assembled using Trinity v2.1.1 [50]. To optimize genome annotation, RNA-seq reads from different tissues were aligned to genome fasta using TopHat package v2.0.11 with default parameters to identify exons region and splice positions [51]. The alignment results were then used as inputs for Cufflinks package v2.2.1 with default parameters for genome-based transcript assembly [52]. Finally, EvidenceModeler v1.1.1 was used to combine the gene models into weighted consensus gene structures with masked repetitive elements [53]. Additionally, PASA was used to update the final gene models, thereby adding information of alternatively spliced sites and untranslated regions (UTR).

The NR, SwissProt [54], InterPro [55], and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases were used for the functional annotation of protein-coding genes using BLASTX and BLASTN utilities with an e-value threshold of $1e^{-5}$ [56].

Non-coding RNA annotations

The tRNAs were predicted using tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscan-SE/). Because rRNAs are highly conserved, we chose relative species' rRNA sequences as references, and the rRNA sequences were predicted using BLAST. Other ncRNAs, including miRNAs and snRNAs, were identified by searching against the Rfam database with default parameters using infernal software (http://infernal.janelia.org/).

Comparative genomic analyses

The protein sequences of 13 vertebrates, including *A. mexicanus, I. punctatus, D. rerio, C. carpio, C. idella, O. niloticus, O. latipes, T. rubripes, G. gallus, H. sapiens, M. musculus, X. tropicalis, and P. marinus*, were downloaded from the Ensembl database (Release 98). Orthologous relationships between the genes from *G. eckloni* and the 13 other vertebrates were inferred through all-against-all protein sequence similarity searches using OthoMCL (http://orthomcl.org/orthomcl/). Only the longest predicted transcript per locus was retained. In the all-against-all BLASTP comparisons, a cutoff e-value of $1e^{-5}$ was used. The MCL inflation index was set to 1.5. A Venn diagram showing the shared orthologous groups (duplicated genes were counted as one) was based on the presence of a representative gene in at least one of the grouped species using a bioinformatics webtool (http://bioinformatics.psb.ugent.be/webtools/Venn/).

For each gene family, an alignment was produced using Muscle (http://www.drive5.com/muscle/), and ambiguously aligned positions were trimmed using Gblocks (http://molevol.cmima.csic.es/castresana/Gblocks.html). The tree was inferred using RAxML v7.2.9 (http://sco.h-its.org/exelixis/software.html). The best-scoring ML tree was inferred by a rapid bootstrap algorithm and ML searches after performing 1000 rapid bootstrap replications. Divergence times between species were calculated using the MCMC tree(http://abacus.gene.ucl.ac.uk/software/paml.html)program implemented by PAML v4.6 [57]. The divergence times for *D. rerio* vs *C. idella* (48–75 Ma), *A. mexicanus* vs *C. carpio* (137–174 Ma), *C. carpio* vs *T. rubripes* (206–252 Ma), *G. gallus* vs *X. tropicalis* (347.6–358.3 Ma), *T. rubripes* vs *G. gallus* (413–443 Ma), and *G. gallus* vs *P. marinus* (515–646 Ma) were obtained from the TimeTree database then used to calibrate divergence dates of other nodes on the phylogenetic tree [58].

According to the divergence times and phylogenetic relationships, CAFÉ was used to analyze the expansion and constriction of gene families in the *G. eckloni* genome based on the gene families identified by OrthoMCL [59]. The phylogenetic tree topology and branch lengths were taken into account when inferring the significance of change in the gene family size of each branch. Enrichment analyses based on the Gene Ontology (GO) and KEGG annotations were performed to identify the functional implications of expanded and contracted genes (Fisher's exact test, adjusted p-value < 0.05).

Abbreviations

QTP: Qinghai-Tibetan plateau; Ma: Million years ago; WGD: whole-genome duplication; 4R WGD: fourth whole-genome duplication; Hi-C: high-throughput chromosome conformation capture; CEGMA: Core Eukaryotic Genes Mapping Approach; BUSCO: Benchmarking Universal Single-Copy Orthologs; CDS: coding sequence; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; MRCA: most

recent common ancestor; Hb: hemoglobin; Ngb: neuroglobin; Cygb: cytoglobin; Mb: myoglobin; XGb: globin X.

Declarations

Acknowledgments

We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

Data Accessibility

All raw data have been deposited into the CNGB Sequence Archive (CNSA) of China National GeneBank DataBase (CNGBdb) with a Project accession CNP0002087 and Sample accession CNS0401290. The PacBio long-read sequencing data, Illumina short-read sequencing data, Hi-C sequencing data and transcriptome data are also available under Accession no. CNX0335337- CNX0335353. The genome assembly is available under the Accession no. CNA0035945. The annotation files can be downloaded from CNSA of CNGBdb (https://ftp.cngb.org/pub/CNSA/data3/CNP0002087/CNS0401290/CNA0035945/).

Ethics approval and consent to participate

This study conformed to the Guidelines for the Care and Use of Experimental Animals established by the Ministry of Science and Technology of the People's Republic of China (Approved number: 2006-398). The research protocol was reviewed and approved by the Ethical Committee of Qinghai University. The study and all methods were carried out in accordance with relevant guidelines and regulations and in compliance with the ARRIVE guidelines (https://arriveguidelines.org/).

References

1. Chen YF, Cao WY: **Schizothoracinae**. In: *Fauna Sinica, Osteichthyes, Cypriniformes III.* Edited by Yue PQ. Beijing: Science Press; 2000: 273-390.
2. Wu YF, Wu CZ: **The fishes of the Qinghai – Xizang plateau**. Chengdu: Science and Technology Press; 1991.
3. Qi D, Chao Y, Guo S, Zhao L, Li T, Wei F, Zhao X: **Convergent, parallel and correlated evolution of trophic morphologies in the subfamily schizothoracinae from the Qinghai-Tibetan plateau**. *PLoS One* 2012, **7**(3):e34070.
4. Qi D, Chao Y, Wu R, Xia M, Chen Q, Zheng Z: **Transcriptome Analysis Provides Insights Into the Adaptive Responses to Hypoxia of a Schizothoracine Fish (Gymnocypris eckloni)**. *Front Physiol* 2018, **9**:1326.

5. Xia M, Chao Y, Jia J, Li C, Kong Q, Zhao Y, Guo S, Qi D: **Changes of hemoglobin expression in response to hypoxia in a Tibetan schizothoracine fish, Schizopygopsis pylzovi**. *J Comp Physiol B* 2016, **186**(8):1033-1043.
6. Zan RG, Liu WG, Song Z: **Tetraploid-hexaploid relationship in Schizothoracinae**. *Acta Genet Sin* 1985, **12**(2):137-142.
7. Yu XY, Li YC, Zhou T: **Karyotype studies of cyprinid fishes in China -Comparative study of the karyotypes of 8 species of schizothoracine fishes**. *Journal of Wuhan University* 1990, **2**:97-104.
8. Yang S, Yan T, Zhao LL, Fu HM, Xiao Q, Pu S, Du ZJ, He Z, Yan TM: **Morphogenesis of blood cell lineages in Ya-fish (Schizothorax prenanti)**. *Chinese Journal of Zoology* 2015, **50**(2):231-242.
9. Dai Y, Han H: **Karyological analysis of two species in the subfamily schizothoracinae (Cypriniformes: Cyprinidae) from China, with notes on karyotype evolution in schizothoracinae**. *Turkish Journal of Fisheries and Aquatic Sciences* 2018, **18**(1):175-186.
10. Xiao S, Mou Z, Fan D, Zhou H, Zou M, Zou Y, Zhou C, Yang R, Liu J, Zhu S *et al*: **Genome of Tetraploid Fish Schizothorax o'connori Provides Insights into Early Re-diploidization and High-Altitude Adaptation**. *iScience* 2020, **23**(9):101497.
11. Qi D, Chao Y, Liang J, Gao Q, Wu R, Mather I, Zhao Y, Chen Q: **Adaptive evolution of interferon regulatory factors is not correlated with body scale reduction or loss in schizothoracine fish**. *Fish Shellfish Immunol* 2018, **73**:145-151.
12. Qi D, Chao Y, Zhang C, Wang Z, Wang W, Chen Q, Zheng Z, Zhang Z: **Duplication of toll-like receptor 22 in teleost fishes**. *Fish Shellfish Immunol* 2019, **94**:752-760.
13. Chen QC, Zheng ZQ, Liu D, Wang FY, Chao Y, Zhang Z, QI DL: **A new pattern of hemoglobin switching in teleost fish-study of the embryonic hemoglobin in the Schizopygopsis pylzovi**. *Acta Hydrobiologica Sinica* 2020, **44**(6):1199-1207.
14. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y *et al*: **The medaka draft genome and insights into vertebrate genome evolution**. *Nature* 2007, **447**(7145):714-719.
15. Liu Z, Liu S, Yao J, Bao L, Zhang J, Li Y, Jiang C, Sun L, Wang R, Zhang Y *et al*: **The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts**. *Nat Commun* 2016, **7**:11757.
16. Xu P, Zhang X, Wang X, Li J, Liu G, Kuang Y, Xu J, Zheng X, Ren L, Wang G *et al*: **Genome sequence and genetic diversity of the common carp, Cyprinus carpio**. *Nat Genet* 2014, **46**(11):1212-1219.
17. Chen Z, Omori Y, Koren S, Shirokiya T, Kuroda T, Miyamoto A, Wada H, Fujiyama A, Toyoda A, Zhang S *et al*: **De novo assembly of the goldfish (Carassius auratus) genome and the evolution of genes after whole-genome duplication**. *Sci Adv* 2019, **5**(6):eaav0547.
18. Qi DL, Guo SC, Chao Y, Kong QH, Li CZ, Xia MZ, Xie BS, Zhao K: **The biogeography and phylogeny of schizothoracine fishes (Schizopygopsis) in the Qinghai-Tibetan Plateau**. *Zoologica Scripta* 2015, **44**(5):523-533.

19. Pasco-Viel E, Yang L, Veran M, Balter V, Mayden RL, Laudet V, Viriot L: **Stability versus diversity of the dentition during evolutionary radiation in cyprinine fish**. *Proc Biol Sci* 2014, **281**(1780):20132688.
20. Deng Y, Meng M, Fang J, Jiang H, Sun N, Lv W, Lei Y, Wang C, Bo J, Liu C *et al*: **Genome of the butterfly hillstream loach provides insights into adaptations to torrential mountain stream life**. *Mol Ecol Resour* 2021.
21. Shi Y, Li J, Li B, Yao TD, Wang SM, Li SJ, Cui ZJ, Wang FB, Pan B, Fang X *et al*: **Uplift of the Qinghai-Xizang (Tibetan) Plateau and east asia environmental change during Late Cenozoic**. *Acta Geographica Sinica* 1999, **54**:10-21.
22. David L, Blum S, Feldman MW, Lavi U, Hillel J: **Recent duplication of the common carp (Cyprinus carpio L.) genome as revealed by analyses of microsatellite loci**. *Mol Biol Evol* 2003, **20**(9):1425-1434.
23. Wang Y, Lu Y, Zhang Y, Ning Z, Li Y, Zhao Q, Lu H, Huang R, Xia X, Feng Q *et al*: **The draft genome of the grass carp (Ctenopharyngodon idellus) provides insights into its evolution and vegetarian adaptation**. *Nat Genet* 2015, **47**(6):625-631.
24. Jian J, Yang L, Gan X, Wu B, Gao L, Zeng H, Wang X, Liang Z, Wang Y, Fang L *et al*: **Whole genome sequencing of silver carp (Hypophthalmichthys molitrix) and bighead carp (Hypophthalmichthys nobilis) provide novel insights into their evolution and speciation**. *Mol Ecol Resour* 2021, **21**(3):912-923.
25. Yang L, Wang Y, Zhang Z, He S: **Comprehensive transcriptome analysis reveals accelerated genic evolution in a Tibet fish, Gymnodiptychus pachycheilus**. *Genome Biol Evol* 2014, **7**(1):251-261.
26. Hoffmann FG, Opazo JC, Storz JF: **Whole-genome duplications spurred the functional diversification of the globin gene superfamily in vertebrates**. *Mol Biol Evol* 2012, **29**(1):303-312.
27. Opazo JC, Butts GT, Nery MF, Storz JF, Hoffmann FG: **Whole-genome duplication and the functional diversification of teleost fish hemoglobins**. *Mol Biol Evol* 2013, **30**(1):140-153.
28. Schwarze K, Campbell KL, Hankeln T, Storz JF, Hoffmann FG, Burmester T: **The globin gene repertoire of lampreys: convergent evolution of hemoglobin and myoglobin in jawed and jawless vertebrates**. *Mol Biol Evol* 2014, **31**(10):2708-2721.
29. Chao Y, Xia M, Wu R, Chen Q, Zheng Z, Qi D: **Molecular characterization and expression changes of cytoglobin genes in response to hypoxia in a Tibetan schizothoracine fish, Schizopygopsis pylzovi**. *Fish Physiol Biochem* 2018.
30. Qi D, Chao Y, Zhao Y, Xia M, Wu R: **Molecular evolution of myoglobin in the Tibetan Plateau endemic schizothoracine fish (Cyprinidae, Teleostei) and tissue-specific expression changes under hypoxia**. *Fish Physiol Biochem* 2018, **44**(2):557-571.
31. Chan FY, Robinson J, Brownlie A, Shivdasani RA, Donovan A, Brugnara C, Kim J, Lau BC, Witkowska HE, Zon LI: **Characterization of adult alpha- and beta-globin genes in the zebrafish**. *Blood* 1997, **89**(2):688-700.
32. Brownlie A, Hersey C, Oates AC, Paw BH, Falick AM, Witkowska HE, Flint J, Higgs D, Jessen J, Bahary N *et al*: **Characterization of embryonic globin genes of the zebrafish**. *Dev Biol* 2003, **255**(1):48-61.

33. Ganis JJ, Hsia N, Trompouki E, de Jong JL, DiBiase A, Lambert JS, Jia Z, Sabo PJ, Weaver M, Sandstrom R *et al*: **Zebrafish globin switching occurs in two developmental stages and is controlled by the LCR**. *Dev Biol* 2012, **366**(2):185-194.
34. Peng Y, Li H, Liu Z, Zhang C, Li K, Gong Y, Geng L, Su J, Guan X, Liu L *et al*: **Chromosome-level genome assembly of the Arctic fox (Vulpes lagopus) using PacBio sequencing and Hi-C technology**. *Mol Ecol Resour* 2021.
35. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J: **Hi-C: a comprehensive technique to capture the conformation of genomes**. *Methods* 2012, **58**(3):268-276.
36. Marcais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers**. *Bioinformatics* 2011, **27**(6):764-770.
37. Ruan J, Li H: **Fast and accurate long-read assembly with wtdbg2**. *Nat Methods* 2020, **17**(2):155-158.
38. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE *et al*: **Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data**. *Nat Methods* 2013, **10**(6):563-569.
39. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK *et al*: **Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement**. *PLoS One* 2014, **9**(11):e112963.
40. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**(4):357-359.
41. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, Barillot E: **HiC-Pro: an optimized and flexible pipeline for Hi-C data processing**. *Genome Biol* 2015, **16**:259.
42. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL: **Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments**. *Cell Syst* 2016, **3**(1):95-98.
43. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP *et al*: **De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds**. *Science* 2017, **356**(6333):92-95.
44. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs**. *Bioinformatics* 2015, **31**(19):3210-3212.
45. Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, Xia R: **TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data**. *Mol Plant* 2020, **13**(8):1194-1202.
46. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.
47. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes**. *Bioinformatics* 2007, **23**(9):1061-1067.
48. Gertz EM, Yu YK, Agarwala R, Schaffer AA, Altschul SF: **Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST**. *BMC Biol* 2006, **4**:41.

49. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise**. *Genome Res* 2004, **14**(5):988-995.
50. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome**. *Nat Biotechnol* 2011, **29**(7):644-652.
51. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**. *Genome Biol* 2013, **14**(4):R36.
52. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B: **AUGUSTUS: ab initio prediction of alternative transcripts**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W435-439.
53. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR: **Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments**. *Genome Biol* 2008, **9**(1):R7.
54. UniProt Consortium T: **UniProt: the universal protein knowledgebase**. *Nucleic Acids Res* 2018, **46**(5):2699.
55. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztanyi Z, El-Gebali S, Fraser M *et al*: **InterPro in 2017-beyond protein family and domain annotations**. *Nucleic Acids Res* 2017, **45**(D1):D190-D199.
56. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M: **Data, information, knowledge and principle: back to metabolism in KEGG**. *Nucleic Acids Res* 2014, **42**(Database issue):D199-205.
57. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood**. *Mol Biol Evol* 2007, **24**(8):1586-1591.
58. Kumar S, Stecher G, Suleski M, Hedges SB: **TimeTree: A Resource for Timelines, Timetrees, and Divergence Times**. *Mol Biol Evol* 2017, **34**(7):1812-1819.
59. De Bie T, Cristianini N, Demuth JP, Hahn MW: **CAFE: a computational tool for the study of gene family evolution**. *Bioinformatics* 2006, **22**(10):1269-1271.

Figures

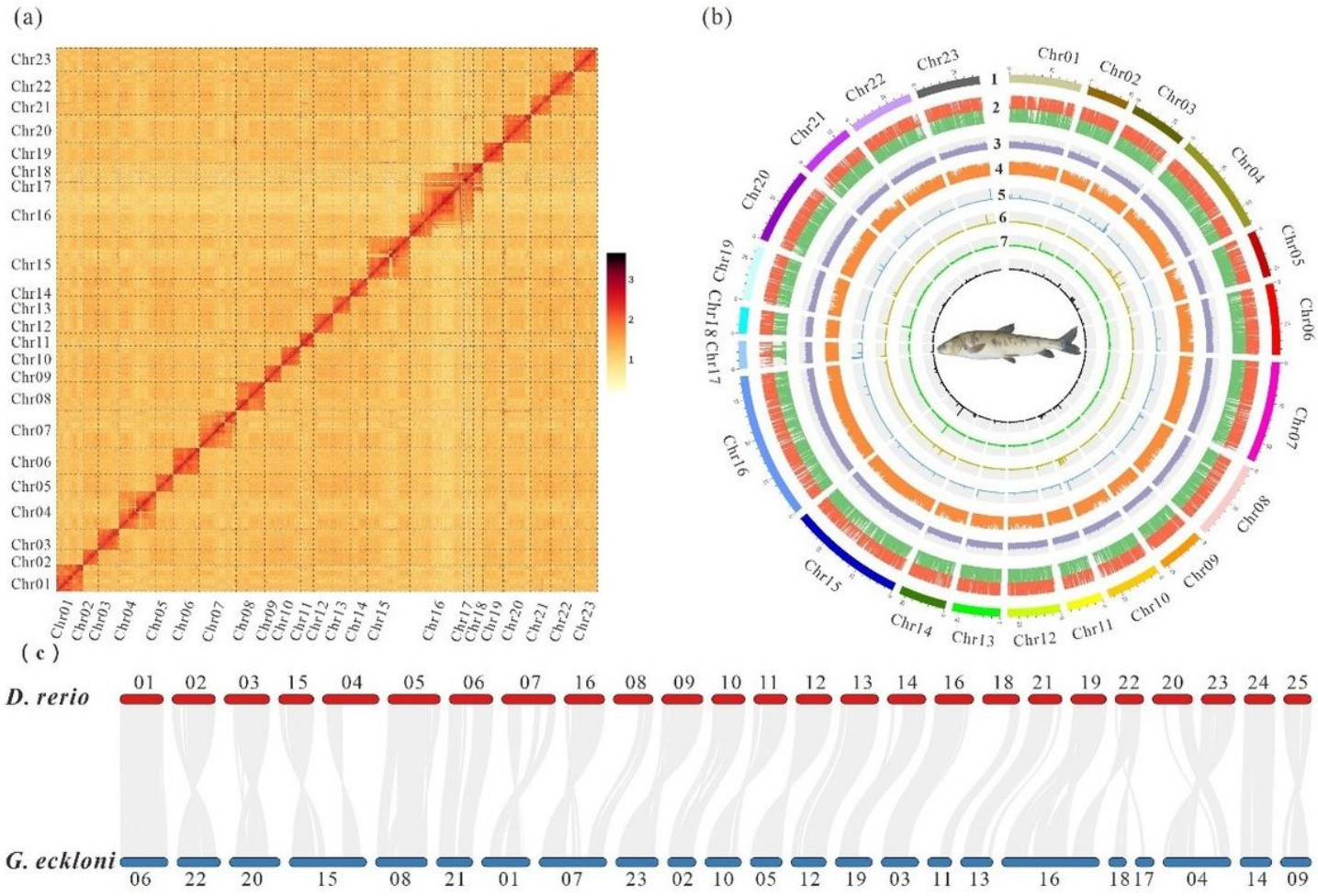

Figure 1

Characteristics of the *G. eckloni* genome. (a) Hi-C intra-chromosomal contact map of the *G. eckloni* genome assembly. (b) Circos plot of the *G. eckloni* genome assembly. 1) Pseudo-chromosomes; 2) gene distribution; 3) GC content; 4) repeat distribution; 5) rRNA distribution; 6) tRNA distribution; 7) miRNA distribution; 8) snRNA distribution. All data were obtained using a sliding window of 10 Kb. (c) Synteny between *G. eckloni* and *D. rerio* genomes.

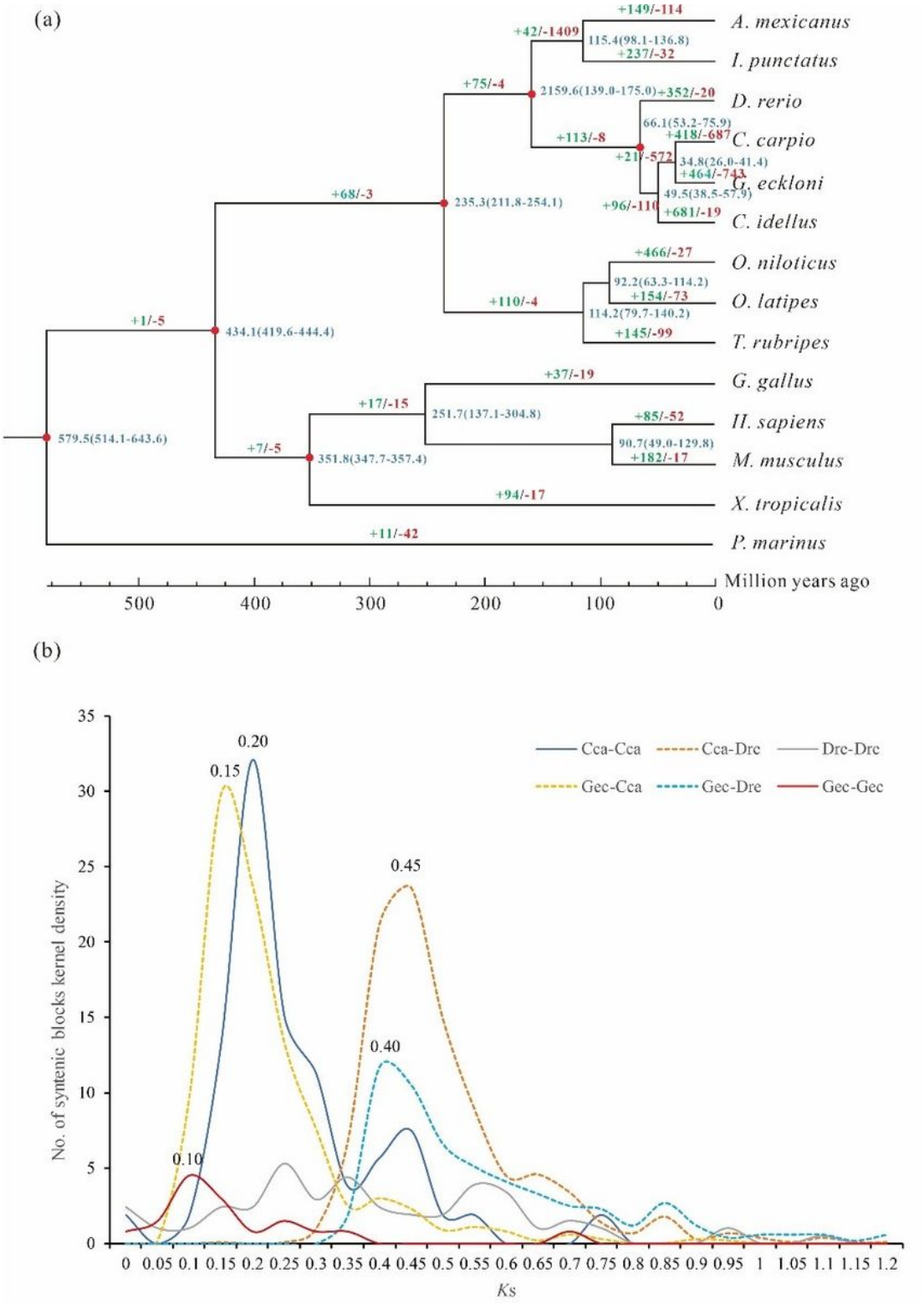

Figure 2

Evolutionary analysis of *G. eckloni*. (a) Phylogenetic tree based on single-copy genes from 14 species shows the estimated divergence time (blue numbers), topology and expansion (green numbers), and contraction (red numbers) of gene families. (b) Kernel density distribution of the number of synonymous substitutions per synonymous site (*K*s) among colinear genes within (continuous lines) and between genomes (dashed lines).

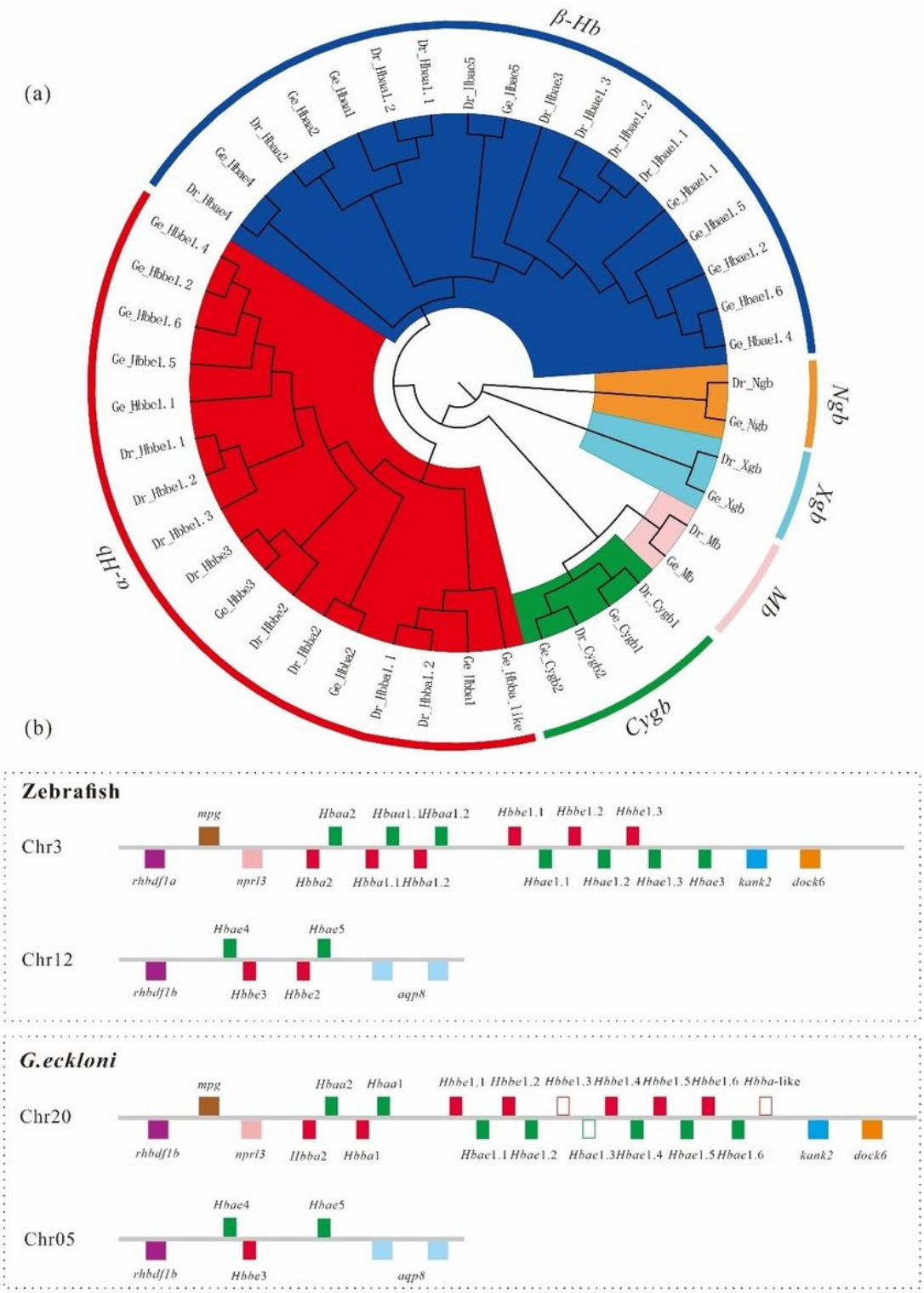

Figure 3

Evolution of the globin gene superfamily in *G. eckloni*. (a) Circular phylogenetic tree of the globin gene superfamily based on coding sequences. (b) Unscaled depiction of the genomic structures of the *α-* and *β-Hb* gene families in zebrafish and *G. eckloni*. To facilitate comparisons, all clusters are presented in the same orientation as the zebrafish. Genes in the forward orientation are shown on top of the chromosome, while genes in the reverse orientation are shown below.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx
- Additionalfile2.docx
- Additionalfile3.xlsx