

Polyploidization, hybridization, and maternal and paternal lineages in Cyprinids (Teleostei: Cypriniformes)

Yan ling Wen

Beijing Institute of Genomics Chinese Academy of Sciences <https://orcid.org/0000-0001-8488-1666>

Jing Chai

Yunnan University

Wei Ma

Yunnan University

Robert W. Murphy

Kunming Institute of Zoology Chinese Academy of Sciences

Shunping He

Institute of Hydrobiology Chinese Academy of Sciences

Ziming Chen

Yunnan University

Ya-ping Zhang

Kunming Institute of Zoology Chinese Academy of Sciences

Xuemei Lu (✉ xuemeilu@mail.kiz.ac.cn)

Kunming Institute of Zoology Chinese Academy of Sciences

Research article

Keywords: Allopolyploidization, transcriptomes, Cyprinidae, matrilineal relatives

Posted Date: December 4th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-119099/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background

The reconstruction of phylogenetic relationships for allopolyploids using genomic data is challenging because hybridization and polyploidy can blur history. In cyprinids, one allopolyploidization event involves goldfish and common carp, yet little is known about the origins of other cyprinid polyploid lineages, including their maternal and paternal ancestral lineages.

Results

Herein we employ 10 *HOX* genes, genomic and transcriptomic data of representative species from seven subfamilies in Cyprinidae to investigate the origins of polyploid lineages. Analyses of the nuclear and mitochondria genomes reveal that the Schizothoracinae and Cyprininae share the same maternal common ancestor, and identify the candidate genes for identifying maternal gene-copies from Cyprininae. Gene-trees show a close relationship between *Sinocyclocheilus grahami* and Cyprininae, and indicate that they share the same whole genome duplication (WGD) event about 12.10 ~ 14.03 Ma. Another allopolyploidization event involves Torinae; one duplication clustering with Schizothoracinae and Schizopygopsinae identifies them as paternal siblings. Further, Labeoninae has a history of recurrent hybridization, which is supported by both gene-trees from genomes and *HOX* genes.

Conclusions

Collectively, the diverse WGD history makes Cyprinidae a candidate system for investigating the origin of hybridization and polyploidization in vertebrates. Further investigations should enlarge the representatives, combine morphological traits, and reconstruct gene-trees based on whole-genome markers.

Background

The Cyprinidae has about 400 closely related polyploid species whose evolution involves auto- or allopolyploidization [1]. Many species have ploidies of 4n, 6n, 8n, and one species even has over 400 chromosomes. Within Cyprinidae, great contradiction exists between traditional and molecular classifications. Traditionally, cyprinids were divided into 10 subfamilies (www.fishbase.org), but 12 subfamilies have been recognized in China [2]. Past molecular phylogenies scatter species in one subfamily across several subfamilies. The topologies did not resolve monophyletic relationships for ten traditional subfamilies whose species have hybrid origins, and, thus, reject the traditional taxonomy [3–9]. Yang *et al.* [9] examined the phylogenetic relationships of Cyprinidae by using mitogenome sequences and cloned nuclear *RAG1*. Consequently, they reassigned some genera to subfamily Cyprininae, and subdivided it into 11 tribes [9]. Subsequently, Tan and Armbruster [10] upgraded these tribes to

subfamilies. However, phylogenetic hypotheses revealed in several studies suggest recurrent, blurred origins via hybridization and polyploidization within Cyprinidae, which may be responsible for the controversial classification among subfamilies. The history of polyploidization in the Cyprininae remains uncertain and this poses significant challenges for phylogenetic systematics.

The ploidy of many cyprinids is clear; yet the history of polyploidization remains poorly understood. Uncertainty involves the timing of whole genome duplications, origins of polyploidization and so on. Ma *et al.* [11] resolved the phylogenetic positions of the paternal and maternal ancestors of Cyprininae and estimated the dates of whole genome duplication to be around 10.71–12.42 Ma based on mitochondrial and nuclear DNA datasets. However, the evolution of polyploids in the other subfamilies await clarification. Further, few studies use large-scale molecular datasets to explore the origin of polyploidization of Cyprinidae.

Analyses of mitochondrial and nuclear DNA marks can yield insights on the history of polyploidization. On one hand, nuclear DNA sequences generally evolve slower than the mitochondrial DNA because of difference in effective population sizes [12–14]. On the other hand, nuclear DNA of polyploids can yield signals of biparental contributions, while mitochondrial DNA only provide that of matrilineal genealogy [15, 16]. An understanding of complicated polyploidization requires discerning signals from both mitochondrial and nuclear DNAs.

Whole genome markers are necessary to assess the recurrent, complex origins of taxa via hybridization and polyploidization. Polyploid cyprinids have very large genomes and numbers of chromosomes, and the technical challenges and costs preclude whole-genome sequencing, yet it is possible to assess variation by using transcriptomic data, which are affordable. Herein, we generate transcriptomes from liver from one representative species of Poropuntinae and Schizopygopsinae, and four species of Schizothoracinae (C-values: 2.19–3.24 pg). Analyses use downloaded mitochondrial genomes, genomic and transcriptomic data from NCBI and Ensembl for representative species in Cyprininae, Labeoninae, Schizopygopsinae and outgroup. Analyses construct gene- and gene-copy trees, which track the evolution history of cyprinids. Because transcriptome assemblies have potential mistakes, we also clone and sequence 10 *HOX* genes and gene-copies from 14 representative species and one species complex, which covers seven subfamilies and one genus of Cyprinidae, and one outgroup to help clarify the large-scale history of polyploidization.

Results

(a) Transcriptomes

Combined cDNA sequences were obtained from six species and two were downloaded from NCBI (Supplementary Table S2). A total of 51.73 Gb raw paired-end reads were produced (Supplementary Table S7). After removing adaptors and low-quality reads, 49.24 Gb of cleaned data were assembled into contigs for *P. huangchuchieni*, *L. rohita*, *Sc. kozlovi*, *Sc. waltoni*, *Sc. macropogon*, *Sc. oconnori*, *O. stewartii*

and *G. pachycheilus* (Supplementary Table S8 and 9). The number of unigenes ranged from 38,785 to 66,439 (Supplementary Table S10) and clustered unigenes with length from 200 to 400 bp had length-distributions ranging from 30.08% to 45.69% (Supplementary Table S11, Fig. S2).

All assembled protein-coding genes of seven species obtained in the clustered unigenes were used to evaluate the assembly quality by comparing the assembled mitochondrial protein-coding sequence with downloaded sequences. The proportions of mismatch nucleotides relative to mitochondrial protein-coding genes for the unigenes of seven species ranged from 0.10% to 1.60%. Total coverage was at least 98.83%. More than 87.76% of paired reads mapped back to the unigenes after alignment using Bowtie2 [17]. The low ratio of mismatch nucleotides, high coverage and reads utilization confirmed assemblies were reliable.

(b) Matrilineal Genealogy

The matrilineal genealogy of 32 species (Supplementary Table S4) and assembled sequences (Supplementary Table S12) were consistent with previous study [9] and included Acrossocheilinae, Barbinae, Cyprininae, Labeoninae, Poropuntiinae, Schizopygopsinae, Schizothoracinae, Smiliogastrinae, and Torinae (Fig. S3). The MCMCtree analysis dated the origin of Cyprininae to about 11.37 Ma (Fig. 1, node I), and the origin of *Schizothorax* to about 3.10 Ma (Fig. 1, node II).

(c) Identification of ORFs and Homologous Genes Clusters

The number of ORFs ranged from 34,202 to 56,298 for *G. pachycheilus*, *L. rohita*, *O. stewartia*, *P. huangchuchieni*, *Sc. kozlovi*, *Sc. macropogon*, *Sc. oconnori* and *Sc. waltoni* (Supplementary Table S13). After filtering homologous gene clusters, 3,671 eligible homologous gene clusters were used to construct nuclear gene-trees to infer the phylogenetic relationships using multiple gene-copies.

(d) Gene-trees Resulting from Multi-omics

The matrilineal genealogy and nuclear genes-trees revealed the maternal and paternal lines, respectively. These trees clearly identified potential diploid progenitors, or at least narrowed parental origins. After filtering out the gene-trees having bootstrap values of less than 50 for key nodes (Fig. S1), 136 trees were retained. Further reduction was based on conflict with gene classification criteria.

(i) Phylogeny of Genera in Schizothoracinae and Schizopygopsinae

The gene-trees mainly depicted classes “Schizo_A” and “Schizo_B” (Fig. 2a, b). The first, supported by 77.94% (106/136) of trees for Schizo_A resolved a monophyletic Schizothoracinae and a monophyletic Schizopygopsinae (Fig. 2a). At least two gene-copies within Schizothoracinae *indicated* a recent

intergeneric WGD events. In Schizo_B, supported by 22.06% (30/136) of gene-trees (Table 1), the topology of gene copies from Schizothoracinae and Schizopygopsinae scattered with each other, and thus indicated introgression between these subfamilies. We used the 106 gene-trees supporting monophyly of both subfamilies to identify the evolutionary relationships of the lineages of Cyprininae, Poropuntiinae and Labeoninae (Fig. 2b).

(ii) Phylogenetic relationship between Sinocyclocheilus and Cyprininae

For *S. grahami*, the 106 gene-trees resolved three topologies. Among the gene-trees, 23.58% (25/106) resolved single-copy genes for *S. grahami* (Fig. 2c, d), 73.59% (78/106) resolved two copies of genes (Fig. 2e, f) and 2.83% (3/106) had no genes for *S. grahami*. Given the different placements of *S. grahami*, trees with single-copy genes were divided into two classes: either “Sinocyclocheilus_single_A” (64.00%, 16/25, Fig. 2c) and “Sinocyclocheilus_single_B” (36.00%, 9/25, Fig. 2d) united single-copy genes of *S. grahami* and one clade of Cyprininae. Two trees based on two-copy genes from *S. grahami* (Fig. 2e, f) differed as follows: Sinocyclocheilus_double_A (62.82%, 49/78) resolved two clades of *S. grahami* that clustered with the two divergent clades of Cyprininae (Fig. 2e); and Sinocyclocheilus_double_B (37.18%, 29/78) had one clade of *S. grahami* clustering with one clade of Cyprininae with high confidence (Fig. 2f). Altogether, 61.32% (65/106) of gene-trees strongly clustered a gene-copy (or copies) of *S. grahami* with either/both divergent gene copies of Cyprininae as sister-groups, and 27.36% (29/106) clustered at least one copy of *S. grahami* with Cyprininae as a sister-group. The matrilineal genealogy and nuclear gene phylogenies for Cyprininae and *S. grahami* resolved a close relationship between them. Thus, at least 65 gene-trees strongly indicated that *S. grahami* might have shared the allopolyploidization origin with *Ca. auratus* and *Cy. carpio*.

(iii) Placement of *L. rohita* and the maternal progenitor of Cyprininae

Four species of *Schizothorax* formed a clade labeled as “Schizothoracinae”, *O. stewartia* and *G. pachycheilus* were labeled as “Schizopygopsinae”, *S. grahami*, *Ca. auratus* and *Cy. carpio* were labeled as “Cyprininae”. Regarding *L. rohita*, 68 gene-trees formed three classes (Fig. 2g, h, i, Supplementary Fig. S4–6). Labeo_A (23.53%, 16/68, Fig. 2g) clustered *L. rohita* with one clade of Cyprininae, Labeo_B (32.35%, 22/68, Fig. 2h) indicated the placement of *L. rohita* was between two clades of Cyprininae, and Labeo_C (44.12%, 30/68, Fig. 2i) rooted *L. rohita* outside of both clades of Cyprininae. Although the placement of *L. rohita* was uncertain, both Labeo_B and Labeo_C indicated that Schizothoracinae and Schizopygopsinae were matrilineal siblings of Cyprininae after comparing placements in the matrilineal genealogy and nuclear gene-tree.

(e) Gene-trees of 10 *HOX* genes

Considering the potential mistakes caused by transcriptomic assemblies, we also tested our hypothesis by analyses of 10 *HOX* genes after PCR amplification and cloning, including *hoxa3a*, *hoxa4a*, *hoxa9a*, *hoxa13a*, *hoxb1b*, *hoxb5b*, *hoxb6b*, *hoxc4a*, *hoxc6b*, *hoxd10a* from 14 species covering seven subfamilies, and outgroup *Opsariichthys bidens* (Supplementary Table S3). One to three copy-numbers of representative species were identified after the alignment, haplotype classification and phylogenetic reconstruction (Supplementary Table S6). A total of six cladograms were simplified by the topologies of 10 *HOX* genes (Fig. 3). According to the simplified topologies, the results indicated that: (1) two diverged gene-copies exist in both Torinae and Cyprininae (Fig. 3a, b); (2) Acrossocheilinae, Poropuntiinae, Schizothoracinae and Schizopygopsinae clustered with one copy of Cyprininae, as maternal siblings, while the other copy clustered with Torinae as paternal siblings, although positions among four subfamilies and one copy of Cyprininae varied in different gene-trees (Fig. 3a–c, g; Supplementary Fig. S7); (3) placement of *L. rohita* varied among gene-trees (Fig. 3, Fig. S7), mainly in three positions including outside allopolyploid subfamilies (Fig. 3a, c and g), in either Cyprininae or Torinae (Fig. 3b, e), and with matrilineal/patrilineal siblings of Cyprininae/Torinae (Fig. 3d). The results of *HOX* gene-trees were consistent with the RNA-seq analyses.

(f) Time-Estimation of WGD in Cyprininae

The distributions of *Ks* values for homologous gene clusters from *Sc. kozlovi*, *Sc. waltoni*, *Sc. oconnori* and *Sc. macropogon* did not present a clear secondary peak. Thus, WGDs did not appear to accumulate (Fig. 4). For *S. grahami*, the distributions presented a clear peak (Fig. 4) with a *Ks* value of about 0.155 to 0.160. Assuming the synonymous substitution mutation rate was about $5.7\text{--}6.4 \times 10^{-9}$ substitutions/synonymous site/year [18], the duplication episode was estimated to have occurred 12.10~14.03 Ma.

Discussion

To guarantee candidate genes meet the criteria for reconstructing their relationships given polyploidization, we employ several strict filtering steps, which result in a dramatic reduction in gene markers. The reduction owes to defective transcriptomes, which might owe to (1) cDNA being transcribed from partial gene sequences, or two cDNA sequences might be transcribed from non-overlapped regions of a same gene, and/or (2) multiple cDNAs are from different spliceosomes of the same gene. These phenomena make it difficult to assemble full-length genes from transcriptomic data [19, 20]. Further, the species of *Schizothorax* may be autopolyploids with multiple copies of genes [21, 22], and accordingly high homology of gene assemblies may create sequences not in the genome. Considering the potential mistakes and bias caused by *de novo* assembly of transcriptomic data, we use cloning and Sanger sequencing of 10 *HOX* genes to reconstruct the gene-trees.

Several representative species within Cyprinidae reveal in detail the hybridization resulting in WGDs in the Cyprininae, Labeoninae, Poropuntiinae, Torinae, Schizopygopsinae and Schizothoracinae. By integrating

mitogenomes, nuclear genomes and transcriptomes, tree topologies indicate that species of Schizothoracinae and Schizopygopsinae are the maternal siblings of Cyprininae and paternal siblings of Torinae. This advances knowledge of the origins of polyploidy within Cyprinidae, especially when compared to previous studies [9, 11]. Analyses of the RNA-seq data cluster gene-copies together in 77.94% of nuclear gene-trees and 76.47% of the nuclear gene-trees for matrilineal copies place Schizothoracinae in the Cyprininae. The phylogenies depict species of these two subfamilies as having polyploid origins and their matrilineal progenitor came from Cyprininae. The 10 *HOX* gene-trees are concordant with the conclusion indicated by RNA-seq data. This narrows the origins of the polyploids and shows that genes from species in these two subfamilies provide reliable makers for identifying maternally or paternally inherited copies of duplicated genes in Cyprinidae and Torinae. Analyses identify complicated gene fates and unstable topologies among seven subfamilies and one genus of Cyprinidae, which is inconsistent with previous findings[23, 24]. Further, although the chromosome numbers in *Schizothorax* suggest that they might have experienced a recent WGD event, no statistical test can technically estimate the significance of a secondary peak (Fig. 4a-d). Nevertheless, the initial peak of recent duplications may hide secondary peaks. This may be a limitation of the unigene method, which requires a secondary peak to dissociate sufficiently from another corresponding large-scale duplication, and that the two peaks have obvious distributions. Our four species of *Schizothorax* may be autopolyploids, because the divergence of duplicate gene-pairs is too slight to separate them from the first peak (Fig. 4a-d). In contrast, a peak emerges above the background level for *S. graham* (Fig. 4e). The second peak reflects a period of increased accumulation of duplicate genes. Thus, analysis of complete genome sequences can detect large-scale duplications that are older or encompass fewer genes than events detected by the unigene method [5, 25–28].

Although *Sinocyclocheilus* assigns to the Cyprininae, its relationship with other subfamilies requires further exploration. Our results suggest that *S. grahami* is an allotetraploid, and shares the same allopolyploid events with goldfish and common carp, same to our result published previously [29]. Lastly, the estimated time of allopolyploidization is about 12.10 ~ 14.03 Ma, which is consistent with a previous estimation in Cyprininae based on mtDNA and nDNA analyses [11]. Thus, the shared allopolyploidization events unite the species by assigning copies to different subfamilies on the same gene-tree; this rejects their taxonomic hypothesis of monophyly.

The phylogenetic positions of polyploid subfamilies of Cyprinidae are complex, especially for Labeoninae. Two phenomena might lead to this. On one hand, species in the subfamily might have experienced hybridization. These species readily crossbreed. Our differing gene-tree distributions agree with the “genomic historical mosaicism” model of Folk *et al.* [30]; their origin began with hybridization and ended with fixation of parental histories, i.e. coalescence of all alleles after the hybridization. On the other hand, species in the subfamily experienced genome downsizing (rediploidization) after its WGD [31]. This process removed redundant DNA (repetitive sequences) and reduced chromosome numbers via complex chromosomal rearrangements [32]. These complex processes likely result in its variable phylogenetic positions in gene-trees. However, this possibility awaits further testing from phylogenomic analyses.

Conclusions

In summary, our work identifies the maternal and paternal lineage of allopolyploids as being in Cyprininae and Torinae, respectively. This suggests that species of *Sinocyclocheilus* are also allopolyploids belonging to Cyprininae, and indicates that Labeoninae fits the model of “genomic historical mosaicism”. However, the limited genomic/transcriptomic data for few representative species of cyprinids fish results in speculation on the history of hybridization and polyploidization in eight cyprinid subfamilies. Further strategies might help clarify the history of hybridization and WGDs by enlarging the number of representative species, combining phenomic and molecular data, and constructing gene-trees based on whole-genome markers.

Methods

(a) Samples and downloaded data

One species each of Acrossocheilinae, Poropuntiinae and Torinae, four species of Cyprininae and Schizothoracinae, three species of Labeoninae, and two species of Schizopygopsinae were sampled. The detailed information on ploidy/genome size and sampling location were listed in Supplementary Table S1–3. Voucher specimens were deposited in the State Key Laboratory for Conservation and Utilization of Bio-resources, Yunnan University.

Mitochondria genomes from 32 species were downloaded from NCBI, including for one species of Acrossocheilinae, Smiliogastrinae and Torinae, two species of Barbinae, eight species of Cyprininae, five species of Labeoninae and Schizothoracinae, four species of Poropuntiinae and Schizopygopsinae, and outgroups including *Danio rerio* and *Ctenopharyngodon idella* (Supplementary Table S4). Genome sequences from five species of *Danio rerio*, *Carassius auratus*, *Ctenopharyngodon idella*, *Cyprinus carpio*, and *Sinocyclocheilus graham*, and RNA-seq data from *Gymnoditychus pachycheilus* and *Labeo rohita* were also downloaded from NCBI (Supplementary Table S2).

(b) Ploidy Confirmation, mRNA Isolation from Liver and Sequencing

Flow cytometry was used to measure the DNA content of blood cells for all four individuals of Schizothoracinae. Ploidy information for *Poropuntius huangchuchieni* was taken from Zheng *et al.* [33] (Supplementary Table S2).

Livers were excised carefully from one individual of *P. huangchuchieni*, four species of *Schizothorax* and one *Oxygymnocypris stewartii*. Samples were stored in RNALater (Ambion, Austin, TX, USA) at –80°C. RNA was extracted from all six samples following instructions for RNeasy Mini Kit (Qiagen, Chatsworth, CA, USA). After isolating mRNA, preparation of the transcriptome library was started with 10µg of total RNA from each sample using an mRNA-Seq Sample Prep Kit (Illumina Inc., San Diego, CA, USA). After

purifying and fragmenting the mRNA, first strand cDNA was used to synthesize the second strand. Following end repair and adding “A” bases, the adaptors were ligated to the ends of the cDNA fragments. After fragment size-selection on gels, suitable cDNA fragments were enriched by PCR. A MinElute PCR Purification Kit (Qiagen, Chatsworth, CA, USA) was used to purify the products at all steps. Finally, quality control and quantification of the library was assessed using Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, Calif., USA) and Qubit (v1.0, Invitrogen, Carlsbad, CA, USA) qPCR. Clusters were generated in a cluster station system and sequenced using the HiSeq 2000. The library was denatured and hybridized to a flowcell. Captured DNA was used as a template for second strand synthesis and amplified into a clonal cluster. Subsequently, clusters were linearized, blocked and hybridized with a sequencing primer, which provided a site for sequencing by synthesis. Finally, the flowcell was transferred to and sequenced by the HiSeq 2000 platform.

(c) RNA-seq Data Quality Control and Transcriptome *de novo* Assembly

Raw reads of downloaded and our sequencing were filtered by excluding the adaptor sequence, the reads with more than 10% unknown nucleotides and low quality reads (more than 50% bases with quality value less than 10), by using in-house scripts.

After filtering, Trinity (v2.3.4) [34] was used for the *de novo* assembly of transcriptomes with parameters set to “–min_kmer_cov 2 –min_glue 3”. The results were clustered into unigenes using TGICL (<https://sourceforge.net/projects/tgicl/>) [35] with default parameters, and unigenes were compared to 13 mitochondrial protein-coding genes of seven species downloaded from NCBI for evaluating the quality of the assemblies. The candidate coding regions of unigenes were identified by TransDecoder (v3.0.0) (<https://transdecoder.github.io/>) [34] with parameters set to “–retain_long_orfs 150 –single_best_orf”. To further maximize sensitivity for capturing open reading frames (ORFs) that may have functional significance, all unigenes were scanned for all possible ORFs to identify homologies to known proteins in the UniProt database (<http://www.uniprot.org/>) and common protein domains in the Pfam database (<http://pfam.xfam.org/>).

(d) Definition of Homologous Gene Clusters and Construction of Gene-trees

First, all–against–all protein sequence alignment were performed with BLASTP (v2.5.0) [36] using the protein sequences annotated by available genomes and ORFs annotated by TransDecoder (v3.0.0). Only the proteins with alignment length over 60 bp and identity more than 50% were retained. Next, homologous genes were clustered using OrthoMCL (v2.0.9) [37]. Most important, homologous gene-clusters were selected as follows: 1) having at least one marker of outgroup (*D. rerio* or *C. idella*); 2) at least one species of *Schizothorax* with two paralogs; and 3) either *Cy. carpio* or *Ca. auratus* with two

paralogs. Subsequently, gene-trees based on homologous gene clusters were constructed by PhyML (v3.1) [38, 39], while employing both *D. rerio* and *C. idella* as outgroup taxa. Lastly, homologous gene-trees were filtered as follow: 1) having at least one marker of outgroup (*D. rerio* or *C. idella*); 2) at least one species of *Schizothorax* with two paralogs; 3) either *Cy. carpio* or *Ca. auratus* with two paralogs from different origins; and 4) bootstrap value of key node was more than 50 (Fig. S1). Setting the conditions for filtering the homologous gene clusters and the homologous gene trees involved the following: 1) either *D. rerio* or *C. idella* should have one marker; 2) to identify the maternal and parental origins of Cyprininae, *Cy. carpio* or *Ca. auratus*, or *Cy. carpio* or *Ca. auratus* should have two homologous gene occurring on different branches in the gene-tree; and 3) all species of *Schizothorax* are polyploids, but their origins are unknown. We select at least one species of *Schizothorax* with two paralogs to explore the origin of *Schizothorax*.

Eventually, in total, 136 gene-trees were reserved. These included the outgroup and at least one species of *Schizothorax* with two copies, clear divergence of maternal and paternal copies in Cyprininae, and markers of *L. rohita* and *P. huangchuchieni*, which served to distinguish maternal and paternal copies.

(e) Construction of Matrilineal Genealogy and Divergence Time Estimation

The mitochondria DNA of *P. huangchuchieni* were locally assembled using the transcriptomic data with Trinity (v2.3.4) [40]. Next, 13 contigs were obtained and connected to a pseudo-mitochondria genome using the mitochondrial DNA of *Barbonymus gonionotus* as reference. Published mitogenome sequences for 32 species and our data were aligned using MUSCLE (v3.8.31)[41]. Then, unreliably aligned regions were removed by using Gblocks (v0.91b) [42] with parameters set to “-b1=4 -b2=5 -b5=a -t=d”.

Subsequently, the retained data were used to construct phylogenetic trees. The maximum likelihood (ML) tree was constructed using PHyML (v3.1) [38, 39] with the GTRGAMMAIG model and 1,000 random sequence addition replicates. Bayesian trees were constructed by using MrBayes (v3.2.2) [43] incorporating optimized models of sequence evolution selected by jModelTest2 [44, 45]. MCMC analyses were run for 5×10^7 generations, with sampling every 1,000th generation and discarding the first 1.25×10^7 generations (12,500 samples) as burn-in. The standard deviation of split frequencies was 0.007550.

The Bayesian MCMC divergence time was computed by the mcmctree program in PAML v4.8 [46] using the mitochondrial DNA sequences and the substitution model GTR. Analyses used the relaxed clock with an uncorrelated log-normal distribution. The Birth-Death model served as the tree prior. The first 50,000 iterations were discarded as burn-in and every 2,000th iteration was sampled until obtaining 50,000 samples. Effective sample sizes estimated with Tracer (v1.6) (<http://tree.bio.ed.ac.uk/software/tracer/>) exceeded 200.

Three calibration points were used. (i) For the age of root of Xenocyprididae, Danionidae and Cyprinidae, the oldest reliable fossil record of 55.8 Ma was used as the maximum bound [47]. (ii) The earliest fossil

specimens of Labeoninae were from the Early Miocene, 11.63 Ma, and this was used as the minimum bound for its node and 16 Ma was used as maximum bound [48, 49]. (iii) The Cyprinidae-Xenocypridae split had a minimum bound of 18.86 Ma and maximum of 22.14 Ma [50].

(f) Estimation of Ks Value and Date of Whole Genome Duplication

For the four species of *Schizothorax* and *S. grahami*, each pair of paralogs in the homologous gene clusters were aligned with MUSCLE v3.8.31 [41]. Then, the levels of synonymous substitutions (Ks) between the pairs of paralogs were estimated by using PAML. Median Ks values were used for clusters with more than two homologous genes [51]. The distribution of Ks values of each species was obtained by combining the Ks values of all homologous gene-clusters. The time of WGD was estimated using the formula: $T = Ks/2r$, where r represented the substitution rate and given values ranging from 5.7 to 6.4×10^{-9} substitutions per site per year among fishes [18].

(g) DNA Extraction, PCR Amplification, Cloning, Sequencing, and Analysis for *HOX* Genes

Total genomic DNA was isolated from frozen or ethanol-fixed tissue samples using the standard phenol-chloroform procedure. The concentration and quality of DNA were assessed by agarose gel electrophoresis and then the samples were stored at 4°C. Ten *HOX* genes were amplified from total DNA extracts via polymerase chain reaction (PCR) using published and/or optimized primers (Supplementary Table S5). Reaction mixtures contained approximately 25–50 ng of DNA template, 1x PCR buffer solution, 0.15 mM MgCl₂ (TaKaRa), 0.25 mM dNTPs (TaKaRa) and 1U Taq DNA polymerase, in a final volume 50 µL. PCR included an initial denaturation step at 95°C for 5 min, followed 33 cycles consisting of 94°C for 30 sec, annealing for 30 sec at from 65°C to 58°C, extension for 60 sec at 72°C, and a final elongation step at 72°C for 7 min. PCR always included negative controls. Several independent PCR products were combined before purification to assure presence of DNA amplified for each allele and each copy in the duplicated genome. All DNA was fractionated by agarose gel extraction kit (Huashun, Shanghai, China) and then cloned into a pMD19-T vector (TaKaRa) following each manufacturer's protocols. After each vector was transformed, culturing was performed in LB solid medium. To increase the probability of detecting duplicated paralogs, 16–36 clones of each species were sequenced (Supplementary Table S6). Plasmid DNAs of 10 *HOX* genes were directly extracted and sequenced on a 3730 DNA Analyzer (Applied Biosystems, Foster City, CA).

All sequences were edited and initially aligned with DNASTAR Lasergene (v7.1.0) [52]. The sequences were checked manually adjusted when necessary. MEGA(v5.0) [53–55] was used to identify unique haplotypes in each species. A representative sequence of each haplotype was selected for further analysis. Phylogenetic reconstructions were performed by maximum likelihood (ML) using PHyML (v3.1)

[38, 39], and Bayesian inference (BI) employing MrBayes (v3.2.2) [43]. In the BI analysis, four independent Markov Chain Monte Carlo (MCMC) chains were run simultaneously for 2,000,000 generations while sampling one tree per 100 replicates and the best-fitting model was found using jModelTest [56]. Two independent runs were conducted. The average standard deviation of split frequencies was required less than 0.01 and the posterior probabilities diagnostic for convergence of branch lengths (potential scale reduction factor) approached 1 after 50,000,000 generations. In the ML analysis, the best-fitting model of nucleotide substitution from jModelTest and 1,000 random sequence addition replicates were used to construct tree.

Abbreviations

WGD

Whole genome duplication

Ma

Million years ago

mtDNA

Mitochondrial DNA

nDNA

Nuclear DNA

ML

Maximum likelihood

Declarations

Availability of data and materials

Transcriptomic data in this work have been submitted to GenBank and accession numbers these data were shown in the Supplementary Table S2. DNA sequences of *Ca. auratus* in this work have been submitted to GenBank under accession numbers of PRJNA289059. The accession numbers of downloaded genomic and transcriptomic data were listed in Supplementary Table S2. The accession number of PCR amplified *HOX* gene sequence were listed in the Supplementary Table S14.

Authors' Contributions

Y.P.Z., X.M.L. and Z.M.C. conceived and designed the experiments. Z.M.C. sampled and identified all the species. J.C. performed the experiments. Y.L.W. analyzed the data. Y.L.W. and J.C. contributed reagents/materials/analysis tools. Y.L.W., J.C., R.W.M., S.P.H., Z.M.C., X.M.L and Y.P.Z. wrote and revised the paper. All authors reviewed this manuscript.

Funding

The sample collection and sequencing were supported by the National Natural Science Foundation of China (91631305, 91331105, 31760306 and U1902201). Experiments and reagent supply were funded by National Postdoctoral Program for Innovative Talents (BX201600130) and China Postdoctoral Science Foundation (2017M613015). The cost of data analysis was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13000000), the National Natural Science Foundation of China (91631305, 91331105, 31760306 and U1902201).

Acknowledgements

We sincerely thank Jing Luo (State Key Laboratory for Conservation and Utilization of Bio-Resources in Yunnan, Yunnan University), Zuyu Yang (Beijing Institute of Genomics, Chinese Academy of Sciences) and Xue Lu (State Key Laboratory for Conservation and Utilization of Bio-Resources in Yunnan, Yunnan University) for suggestions on early drafts of this manuscript, Shanjin Wu and Feng Huang provided assistance in sample collection. Technical support from colleagues in the lab of YPZ is gratefully acknowledged.

Ethics declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. **FishBase**. World Wide Web electronic publication. www.fishbase.org, (02/2019).
2. Chen YY: **Fauna sinica, osteichthyes. Cypriniformes II**, vol. Cypriniformes II. Beijing: Science Press; 1998.
3. Zardoya R, Doadrio I: **Molecular Evidence on the Evolutionary and Biogeographical Patterns of European Cyprinids**. *J Mol Evol* 1999, **49**(2):227-237.
4. Gilles A, Lecointre G, Miquelis A, Loerstcher M, Chappaz R, Brun G: **Partial combination applied to phylogeny of European cyprinids using the mitochondrial control region**. *Mol Phylogenet Evol* 2001, **19**(1):22-33.

5. Cunha C, Mesquita N, Dowling TE, Gilles A, Coelho MM: **Phylogenetic relationships of Eurasian and American cyprinids using cytochrome b sequences.** *Journal of Fish Biology* 2002, **61**(4):929-944.
6. Liu HZ, Chen YY: **Phylogeny of the East Asian cyprinids inferred from sequences of the mitochondrial DNA control region.** *Can J Zool* 2003, **81**(12):1938-1946.
7. Wang X, Li J, He S: **Molecular evidence for the monophyly of East Asian groups of Cyprinidae (Teleostei : Cypriniformes) derived from the nuclear recombination activating gene 2 sequences.** *Mol Phylogenet Evol* 2007, **42**(1):157-170.
8. Wang X, Gan X, Li J, Mayden RL, He S: **Cyprinid phylogeny based on Bayesian and maximum likelihood analyses of partitioned data: implications for Cyprinidae systematics.** *Sci China-Life Sci* 2012, **55**(9):761-773.
9. Yang L, Sado T, Hirt MV, Pascoviel E, Arunachalam M, Li J, Wang X, Freyhof J, Saitoh K, Simons AM: **Phylogeny and polyploidy: Resolving the classification of cyprinine fishes (Teleostei: Cypriniformes).** *Mol Phylogenet Evol* 2015, **85**:97-116.
10. Tan M, Armbruster JW: **Phylogenetic classification of extant genera of fishes of the order Cypriniformes (Teleostei: Ostariophysi).** *Zootaxa* 2018, **4476**(1):6-39.
11. Ma W, Zhu ZH, Bi XY, Murphy RW, Wang S, Gao Y, Xiao H, Zhang Y, Luo J: **Allopolyploidization is Not So Simple: Evidence from the Origin of the Tribe Cyprinini (Teleostei: Cypriniformes).** *Curr Mol Med* 2014, **14**(10):1331-1338.
12. Kocher TD, Wilson AC: **Sequence Evolution of Mitochondrial DNA in Humans and Chimpanzees: Control Region and a Protein-Coding Region.** In: *Evolution of Life: Fossils, Molecules, and Culture.* Edited by Osawa S, Honjo T. Tokyo: Springer Japan; 1991: 391-413.
13. Moritz C, Dowling TE, Brown WM: **EVOLUTION OF ANIMAL MITOCHONDRIAL DNA: RELEVANCE FOR POPULATION BIOLOGY AND SYSTEMATICS.** *Annu Rev Ecol Syst* 1987, **18**(1):269-292.
14. Brown WM, George MR, Wilson AC: **Rapid evolution of animal mitochondrial DNA.** *Proc Natl Acad Sci USA* 1979, **76**(4):1967-1971.
15. Evans BJ, Kelley DB, Melnick DJ, Cannatella DC: **Evolution of RAG-1 in Polyploid Clawed Frogs.** *Mol Biol Evol* 2005, **22**(5):1193-1207.
16. Evans BJ, Kelley DB, Tinsley RC, Melnick DJ, Cannatella DC: **A mitochondrial DNA phylogeny of African clawed frogs: phylogeography and implications for polyploid evolution.** *Mol Phylogenet Evol* 2004, **33**(1):197-213.
17. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9**(4):357-359.
18. 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**:625-631.
19. Chopra R, Burow G, Farmer AD, Mudge J, Simpson CE, Burow MD: **Comparisons of *De Novo* Transcriptome Assemblers in Diploid and Polyploid Species Using Peanut (Arachis spp.) RNA-Seq Data.** *Plos One* 2014, **9**(12):e115055.

20. Gutierrez-Gonzalez JJ, Garvin DF: **De Novo Transcriptome Assembly in Polyploid Species**. In: *Oat: Methods and Protocols*. Edited by Gasparis S. New York, NY: Springer New York; 2017: 209-221.
21. Chi W, Ma X, Niu J, Zou M: **Genome-wide identification of genes probably relevant to the adaptation of schizothoracins (Teleostei: Cypriniformes) to the uplift of the Qinghai-Tibet Plateau**. *Bmc Genomics* 2017, **18**(1):310.
22. Yu M, Zhang D, Hu P, Peng S, Li W, He S, Zhai W, Xu Q, Chen L: **Divergent adaptation to Qinghai-Tibetan Plateau implicated from transcriptome study of *Gymnocypris dobula* and *Schizothorax nukiangensis***. *Biochem Syst Ecol* 2017, **71**:97-105.
23. Cao WX, Chen YY, Wu YF, Zhu SQ: **Relationship between the uplift of Qinghai-Tibet Plateau and the origin and evolution of Schizothoracinae fishes**. In: *The problem of the era , amplitude and form about uplift of the Qinghai-Tibet Plateau*. Beijing: Science Press; 1981: 118-130.
24. Hora SL: **Comparison of the Fish-Faunas of the Northern and the Southern Faces of the Great Himalayan Range**. *SIL Proceedings, 1922-2010* 1938, **8**(3):95-107.
25. Vandepoele K, Simillion C, De Peer YV: **Evidence That Rice and Other Cereals Are Ancient Aneuploids**. *Plant Cell* 2003, **15**(9):2192-2202.
26. Initiative TAG: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana***. *Nature* 2000, **408**(6814):796-815.
27. Wolfe KH, Shields DC: **Molecular evidence for an ancient duplication of the entire yeast genome**. *Nature* 1997, **387**(6634):708-713.
28. Wendel JF, Cronn R: **Polyploidy and the Evolutionary History of Cotton**. *Adv Agron* 2003, **78**:139-186.
29. Luo J, Chai J, Wen Y, Tao M, Lin G, Liu X, Ren L, Chen Z, Wu S, Li S *et al*: **From asymmetrical to balanced genomic diversification during rediploidization: Subgenomic evolution in allotetraploid fish**. *Science Advances* 2020, **6**(22):eaaz7677.
30. Folk RA, Soltis PS, Soltis DE, Guralnick RP: **New prospects in the detection and comparative analysis of hybridization in the tree of life**. *Am J Bot* 2018, **105**(3):364-375.
31. Leitch IJ, Bennett MD: **Genome downsizing in polyploid plants**. *Biol J Linn Soc* 2004, **82**(4):651-663.
32. Dodsworth S, Chase MW, Leitch AR: **Is post-polyploidization diploidization the key to the evolutionary success of angiosperms?** *Bot J Linnean Soc* 2016, **180**(1):1-5.
33. Zheng L, Yang J, Chen X: **Molecular phylogeny and systematics of the Barbinae (Teleostei: Cyprinidae) in China inferred from mitochondrial DNA sequences**. *Biochem Syst Ecol* 2016, **68**:250-259.
34. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M *et al*: **De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis**. *Nat Protoc* 2013, **8**(8):1494-1512.
35. Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B *et al*: **TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large est datasets**. *Bioinformatics* 2003, **19**(5):651-652.

36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic Local Alignment Search Tool.** *J Mol Biol* 1990, **215**(3):403-410.
37. Li L, Stoeckert CJ, Roos DS: **OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes.** *Genome Res* 2003, **13**(9):2178-2189.
38. Guindon S, Lethiec F, Duroux P, Gascuel O: **PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference.** *Nucleic Acids Res* 2005, **33**:557-559.
39. Guindon S, Dufayard J, Lefort V, Anisimova M, Hordijk W, Gascuel O: **New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0.** *Syst Biol* 2010, **59**(3):307-321.
40. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng QD *et al*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nat Biotechnol* 2011, **29**(7):644-U130.
41. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**(5):1792-1797.
42. Castresana J: **Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis.** *Mol Biol Evol* 2000, **17**(4):540-552.
43. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17**(8):754-755.
44. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52**(5):696-704.
45. Darriba D, Taboada GL, Doallo R, Posada D: **jModelTest 2: more models, new heuristics and parallel computing.** *Nat Methods* 2012, **9**(8):772-772.
46. Yang ZH: **PAML 4: Phylogenetic Analysis by Maximum Likelihood.** *Mol Biol Evol* 2007, **24**(8):1586-1591.
47. Cavender TM: **The fossil record of the Cyprinidae.** In: *Cyprinid Fishes: Systematics, biology and exploitation.* Edited by Winfield IJ, Nelson JS. Dordrecht: Springer Netherlands; 1991: 34-54.
48. Van Couvering JAH: **Early Records of Freshwater Fishes in Africa.** *Copeia* 1977, **1977**(1):163.
49. Stewart KM: **Fossil Fish from the Nile River and Its Southern Basins.** In: *The Nile: Monographiae Biologica.* Edited by Dumont HJ. Dordrecht: Springer Netherlands; 2009: 677-704.
50. 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.
51. Blanc G, Wolfe KH: **Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes.** *Plant Cell* 2004, **16**(7):1667-1678.
52. Burland TG: **DNASTAR's Lasergene sequence analysis software.** In: *Bioinformatics Methods and Protocols Methods in Molecular Biology.* Edited by Misener. S, Krawetz SA, vol. 132. Totowa, NJ.: Humana Press; 2000: 71-91.

53. Kumar S, Nei M, Dudley JT, Tamura K: **MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences.** *Briefings Bioinf* 2008, **9**(4):299-306.
54. Tamura K, Peterson DS, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods.** *Mol Biol Evol* 2011, **28**(10):2731-2739.
55. Xia X, Xie Z: **DAMBE: Software Package for Data Analysis in Molecular Biology and Evolution.** *J Hered* 2001, **92**(4):371-373.
56. Posada D: **jModelTest: Phylogenetic Model Averaging.** *Mol Biol Evol* 2008, **25**(7):1253-1256.

Tables

Table 1 The classification of copy clades from all the gene trees				
	Type	Copy Clade	Number	Percent (%)
<i>Schizothorax</i>	A	"Schizo_A"	106	77.94
	B	"Schizo_B"	30	22.06
	Total	-	136	100.00
<i>Sinocyclocheilus</i>	Single	"Sinocyclocheilus_single_A"	16	15.09
	Single	"Sinocyclocheilus_single_B"	9	8.49
	Double	"Sinocyclocheilus_double_A"	49	46.23
	Double	"Sinocyclocheilus_double_B"	29	27.36
	None	-	3	2.83
	Total	-	106	100.00
<i>Labeo</i>	A	"Labeo_A"	16	23.53
	B	"Labeo_B"	22	32.35
	C	"Labeo_C"	30	44.12
	Total	-	68	100.00

Figures

Fig1

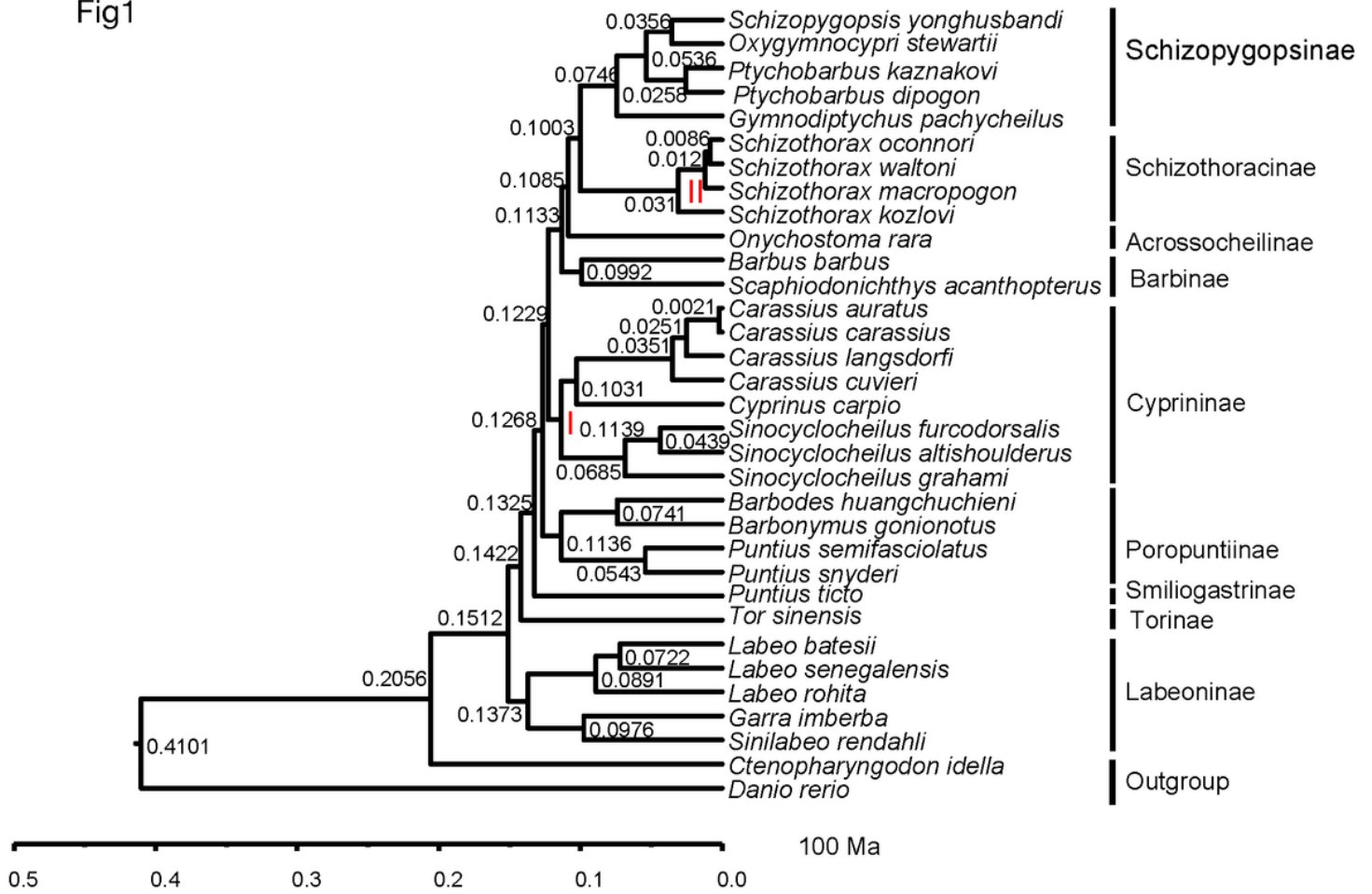


Figure 1

Estimated divergence time among the subfamilies Acrossocheilinae, Barbinae, Cyprininae, Labeoninae, Poropuntiinae, Schizopygopsinae, Schizothoracinae, Smiliogastrinae, and Torinae in Cyprinidae, calculated by concatenated mitochondrial DNA sequences of 33 representative species. Numbers above branches represent divergence times in 100 millions of years.

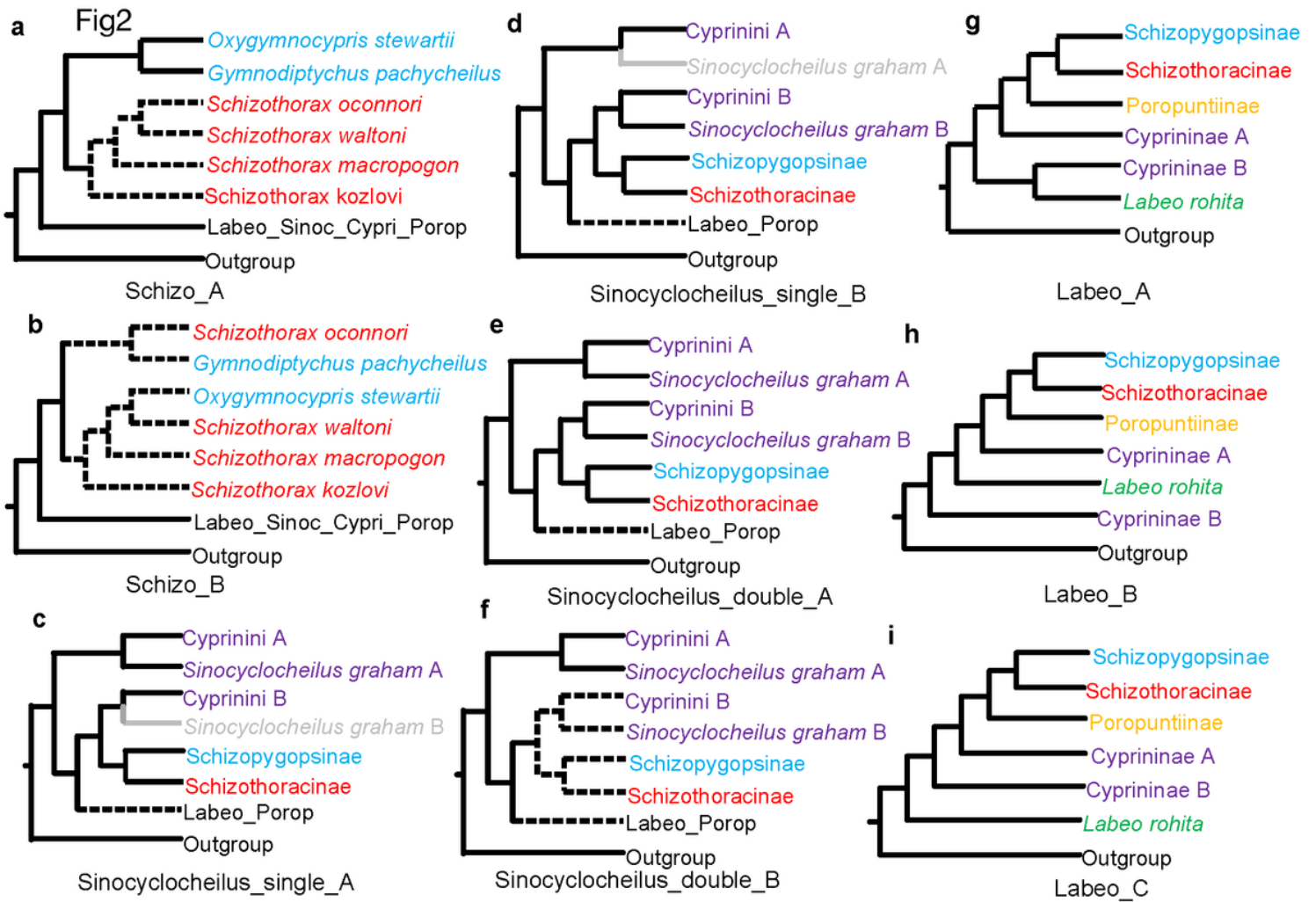


Figure 2

Classification of two gene-copies in polyploid fishes. (a) Schizo_A; (b) Schizo_B; (c, d) tree diagrams of a single clade of *Sinocyclocheilus grahami* involving either A or B; (e, f) tree diagrams of two clades of *S. grahami*; (g) *Labeo rohita* and Cyprininae B are clustering as sister-groups; (h) *L. rohita* locates between Cyprininae A and B; (i) *L. rohita* clusters out of Cyprininae A and B. Dashed line indicates an unstable branch; gray line shows the corresponding clade does not exist; “*Labeo_Sinoc_Cypri_Poropuntius*” denotes *L. rohita*, *S. grahami*, *Cyprinus carpio*, *Carassius auratus* and *Poropuntius huangchuchieni*; “Outgroup” includes *Danio rerio* and *Ctenopharyngodon idella*; “*Labeo_Porop*” denotes *L. rohita* + *P. huangchuchieni*; “*Schizopygopsinae*” indicates *Oxygymnocypris stewartii* + *Gymnodiptychus pachycheilus*; “*Cyprinini A*” indicates one copy of *Cy. carpio* and *Ca. auratus*; “*Cyprinini B*” indicates the other copy of *Cy. carpio* and *Ca. auratus*; “*Sinocyclocheilus grahami A*” denotes one copy of *S. grahami*, and “*Sinocyclocheilus grahami B*” shows the other copy; “*Schizothoracinae*” is the clade including *Schizothorax kozlovi*, *Sc. waltoni*, *Sc. macropogon* and *Sc. oconnori*; “*Poropuntiinae*” indicates *P. huangchuchieni*; “*Cyprininae*” is *S. grahami* + *Cy. carpio* + *Ca. auratus*.

Fig3

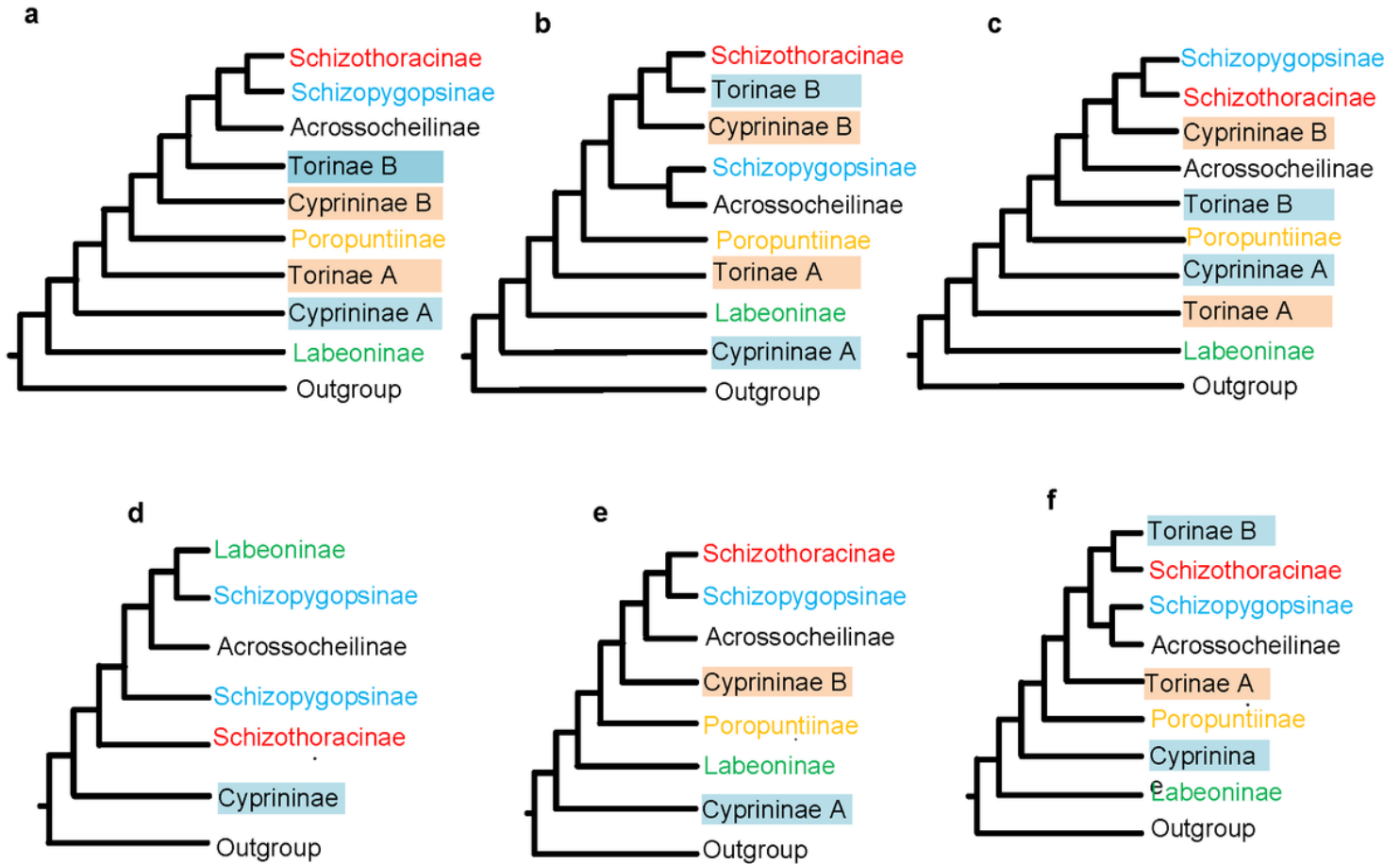


Figure 3

Seven cladograms according to the topologies of gene-trees reconstructed by 10 Hox genes including *hoxa3a*, *hoxa4a*, *hoxa9a*, *hoxa13a*, *hoxb1b*, *hoxb5b*, *hoxb6b*, *hoxc4a*, *hoxc6b*, *hoxd10a* (detailed topologies of gene-trees shown in Fig.7). (a) Simplified topology from *hoxa4a*, *hoxb5b*, *hoxb6b* and *hoxc6b*; (b) simplified topology from *hoxa3a* and *hoxa13a*; (c) simplified topology from *hoxc4a*; (d) simplified topology from *hoxb1b*; (e) simplified topology from *hoxa9a*; (f) simplified topology from *hoxd10a*. “Outgroup” denotes *Opsariichthys bidens gunthe*, *Danio rerio* and *Megalobrama amblycephala*; “Acrossocheilinae” indicates *Onychostoma gerlachi*; “Cyprininae” denotes *Carassius auratus* + *Cyprinus carpio* + *Sinocyclocheilus grahami* + *Procypris merus* + *Procypris merus* + *Cyprinus carpio*; “Labeoninae” includes *Sinilabeo laticeps* Wu et Lin + *Pseudogyrincheilus procheilus* + *Garra pingi*; “Poropuntiinae” denotes *Poropuntius huangchuchieni*; “Schizopygopsinae” indicates *Oxygymnocypris stewartia* + *Ptychobarbus dipogon*; “Schizothoracinae” includes *Schizothorax*; “Torinae” includes *Tor sinensis*. Different putative alleles of allopolyploid subfamilies are denoted by A and B following the subfamily names. The orange boxes marks the maternal copies, and the blue ones marks the paternal copies of allopolyploid subfamilies.

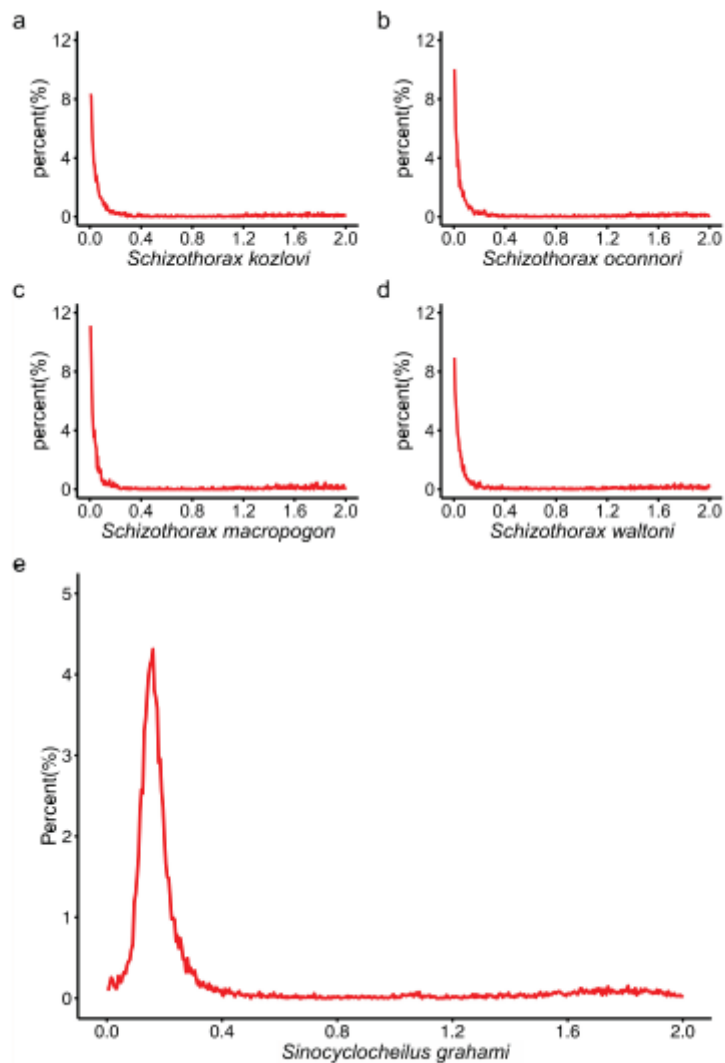


Figure 4

Frequency distributions of Ks values estimated for duplicated genes in homologous clusters for *Schizothorax kozlovi* and *Sc. waltoni*, *Sc. oconnori*, *Sc. macropogon* and *Sinocyclocheilus grahami*. X-axis is the Ks value and y-axis indicates the percent. (a) *Schizothorax kozlovi*; (b) *Sc. waltoni*; (c) *Sc. oconnori*; (d) *Sc. macropogon*; (e) *Sinocyclocheilus grahami*.

Supplementary Files

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

- [SuppFigurecaptions201027.docx](#)
- [SuppTables201027.docx](#)
- [SupplementaryTableS14Hoxgenelocation201027xlsx.xlsx](#)
- [figspp201027.pdf](#)