# Genomes reveal pervasive distant hybridization in nature among cyprinid fishes

Li Ren [1,†], Xiaolong Tu [2,4,†], Mengxue Luo [1,†], Qizhi Liu [1,†], Jialin Cui [1], Xin Gao [1], Hong Zhang [1], Yakui Tai [1], Yiyan Zeng [1], Mengdan Li [1], Chang Wu [1], Wuhui Li [1], Jing Wang [1], Dongdong Wu [2,3,*], and Shaojun Liu [1,*]

[1]State Key Laboratory of Developmental Biology of Freshwater Fish, Engineering Research Center of Polyploid Fish Reproduction and Breeding of the State Education Ministry, College of Life Sciences, Hunan Normal University, Changsha 410081, China
[2]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650201, China
[3]Kunming Natural History Museum of Zoology, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China
[4]Kunming College of Life Science, University of the Chinese Academy of Sciences, Kunming 650204, China
*Correspondence address. Shaojun Liu, College of Life Sciences, 36 Lushan Rd., Yuelu District, Changsha 410081, Hunan, China. E-mail: lsj@hunnu.edu.cn;
Dongdong Wu, Kunming Institute of Zoology, the Chinese Academy of Sciences, 17 Longxin Rd., Kunming 650201, Yunnan, China. E-mail: wudongdong@mail.kiz.ac.cn
[†]These authors contributed equally to this work.

## Abstract

**Background:** Genomic data have unveiled a fascinating aspect of the evolutionary past, showing that the mingling of different species through hybridization has left its mark on the histories of numerous life forms. However, the relationship between hybridization events and the origins of cyprinid fishes remains unclear.

**Results:** In this study, we generated *de novo* assembled genomes of 8 cyprinid fishes and conducted phylogenetic analyses on 24 species. Widespread allele sharing across species boundaries was observed within 7 subfamilies of cyprinid fishes. Based on a systematic analysis of multiple tissues, we found that the testis exhibited a conserved pattern of divergence between the herbivorous *Megalobrama amblycephala* and the carnivorous *Culter alburnus*, suggesting a potential link to incomplete reproductive isolation. Significant differences in the expression of 4 genes (*dpp2*, *ctrl*, *psb7*, and *ppce*) in the liver and intestine, accompanied by variations in enzyme activities, indicated swift divergence in digestive enzyme secretion. Moreover, we identified introgressed genes linked to organ development in sympatric fishes with analogous feeding habits within the *Cultrinae* and *Leuciscinae* subfamilies.

**Conclusions:** Our findings highlight the significant role played by incomplete reproductive isolation and frequent gene flow events, particularly those associated with the development of digestive organs, in driving speciation among cyprinid fishes in diverse freshwater ecosystems.

**Keywords:** diet divergence, genetic introgression, phylogenomics, incomplete reproductive isolation

## Introduction

Cyprinidae (order Cypriniformes) is the largest and most diverse family of ray-finned fish, comprising about 13 subfamilies and 370 genera [1, 2]. This family includes popular aquarium fishes like goldfish and koi, as well as the valuable vertebrate model organism, the zebrafish [2]. As a family of freshwater fish, the origin time of Cyprinidae was estimated at 154 million years ago (MYA) [3], and they are now widely distributed in almost all types of water around the world [2]. Their great diversities in feeding and reproductive behaviors, as well as morphology, including body length (ranging from about 8 mm for *Paedocypris progenetica* to approximately 3 m for *Catlocarpio siamensis*) [4, 5] and digestive organs [6, 7], are intriguing to evolutionary biologists due to their phylogenetic relationships and adaptive radiation evolution. However, the narrow distribution ranges or small population sizes of Cyprinidae fish now face threats from human activity, such as overfishing, damming of upland rivers, pollution, habitat destruction, and novel viral infections [8].

Introgression plays a significant and frequent role in adaptive evolution [9]. At least 10% of animal species are involved in hybridization, although most hybrid individuals have low viability or are sterile [10]. While reproductive isolation (RI) is frequently observed in intergeneric hybridization among birds and mammals, it is less common in cyprinid fish species, as evidenced by the documented instances of disrupted isolation [11–13]. Natural selection, including variable water environments and deleterious homozygosity for small population sizes, brings great pressures to the survival of freshwater fish [14, 15]. Hybridization has the potential to generate genetic diversity and create opportunities for novel adaptive radiations, although it has been considered a breakdown of isolating mechanisms [16]. The low viability or sterility of hybrids could reinforce RI through selection for assortative mating and result in adaptive introgression [10]. The rate of introgression depends on the pressures of the freshwater environment and affects fish biodiversity [17]. Natural hybridization involving intergeneric hybridization was always observed among cyprinid fishes [18–20], while bisexual fertile progenies were detected in the laboratory experiments of various hybrid groups [13]. This evidence suggests that prezygotic isolation evolves more rapidly than postzygotic isolation in cyprinid fishes, challenging the assumption of the criticality of variation in dietary niche breadth for speciation. Now, the relationship between biodiversity and introgres-

sive hybridization in cyprinid fishes is still obscure. Phylogenomics from whole genome sequences could provide us with more detailed evidence of it than the fragmented detective technologies, including rDNA [20], mitochondrial DNA, and microsatellites [21].

Diet plays a crucial role in the biodiversity and habitat distribution of fishes and is influenced by differences in foraging behavior and digestive organ morphology [22]. The selection of diet is particularly important for sympatric species [23]. In the case of cyprinid fishes, 4 main categories of diet have been identified. These include herbivorous fishes (e.g., *Megalobrama amblycephala* and *Ctenopharyngodon idella*), carnivorous fishes (e.g., *Culter alburnus* and *Elopichthys bambusa*), filter-feeding fishes (e.g., *Hypophthalmichthys nobilis*), and omnivorous fishes (e.g., *Cyprinus carpio* and *Carassius auratus*) [24]. These fish species are distributed across different water layers to acquire various types of food resources. Cyprinid fishes have evolved unique adaptations for food digestion, such as advanced protrusible pharyngeal teeth, despite the absence of jaw teeth and stomachs [25]. The number and shape of teeth in cyprinid fishes exhibit significant variation and are used as phenotypic features for species classification [2]. In East African cichlids, the number of tooth rows on both jaws has been associated with specific feeding ecologies [26]. However, it remains unclear whether variations in the width of the dietary niche are critical for the biodiversity of cyprinid fishes.

In this study, we obtained the *de novo* assembled genome sequences of 8 cyprinid fish species and conducted comparative genome analyses using a set of 24 high-quality assembled genome sequences. Through gene flow analyses, we investigated hybridization events and their contributions to the speciation of cyprinid fishes. Furthermore, we conducted evolutionary constraints analyses on various tissues and organs and investigated their divergence between *M. amblycephala* and *C. alburnus*. Our findings emphasize the significance of gene flow events in the origin of cyprinid fishes and the functional divergence that drives speciation.

## Methods

### Sample collection

Three sexually mature male *C. alburnus* (NCBI:txid194366) and *M. amblycephala* (NCBI:txid75352), raised in identical controlled conditions for 24 months posthatching, were bred at the Engineering Center of Polyploid Fish Breeding, National Education Ministry, Changsha, Hunan, China. The broodstock were sourced from the Yangtze River (30°25′56″ N, 114°50′32″ E). The parents of *Gobiocypris rarus* (NCBI:txid143606) were obtained from the Liu Sha River, Sichuan Province (coordinates 29°19′31″ N, 102°40′38″ E). We also collected 1 individual of each fish species (*Cirrhinus molitorella* [NCBI:txid172907], *Pseudorasbora parva* [NCBI:txid51549], *Xenocypris davidi* [NCBI:txid291826], *Elopichthys bambusa* [NCBI:txid238031], and *Ctenopharyngodon idella* [NCBI:txid7959]) from Dongting Lake, Hunan, China (coordinates 29°15′8″ N, 112°50′24″ E). These individuals were deeply anesthetized with 300 mg/L tricaine methanesulfonate (Sigma-Aldrich) for 10 minutes (20°C) in a separation tank. After confirming their deaths, the muscle, brain, liver, intestine, kidney, and testis of all samples were collected after dissection.

### DNA isolation and whole genome sequencing

High-quality and high-molecular-weight genomic DNA was isolated from muscle based on the DNA extraction methods. The purification was performed using the QIAGEN Genomic Kit based on the standard operating procedure. The degradation and contamination of the extracted DNA were detected using 1% agarose gels. Then, DNA purity was determined using the NanoDrop One UV-Vis spectrophotometer (Thermo Fisher Scientific) with 260/280 and 260/230 ratios. DNA concentration was measured by the Qubit 4.0 Fluorometer (Invitrogen).

After quality checking, the genomic DNA of *C. alburnus* and *M. amblycephala* was randomly sheared using Megaruptor (Diagenode). DNA was size-selected using an SPRI bead protocol. The purity of the extracted DNA was determined using a NanoDrop spectrophotometer (Thermo Fisher Scientific). All procedures were carried out at room temperature. Large DNA fragments were separated using BluePippin DNA Size Selection System. DNA damage repair and end repair were performed. Barcoded overhang hairpin adapters were ligated to the fragment ends. The connection reaction was performed using the Ligation Sequencing Kit (Oxford Nanopore, SQK-LSK108). A constructed DNA library was quantified using Qubit. Lastly, sequencing was performed using Nanopore sequencing.

The genomic DNA of *C. alburnus* and *M. amblycephala* was utilized for whole genome resequencing. First, high-quality DNA samples were used to prepare single-stranded circular libraries. Subsequently, the circular libraries were transformed into DNA nanoballs (DNBs), which are spherical structures containing millions of copies of the circular DNA templates. Once the DNBs were formed, they were loaded onto patterned nanoarrays. Following the loading of DNBs onto the nanoarrays, combinatorial probe anchor synthesis sequencing was conducted. Finally, DNBSEQ-T7 sequencing was performed using a paired-end approach (150 bp × 2) in accordance with the standard protocol [27].

In total, 15 $\mu$g DNA for the 6 fishes (*C. idella*, *C. molitorella*, *P. parva*, *X. davidi*, *G. rarus*, and *E. bambusa*) was used for the preparation of SMRTbell target-size libraries, which were constructed using PacBio's standard protocol (Pacific Biosciences) with 15-kb preparation solutions. The main steps for library preparation are listed as follows: (i) The genomic DNA was sheared using g-TUBEs (Covaris), (ii) an A-tailing reaction was used to form an overhang, (iii) the fragments were ligated with the hairpin adapter using the SMRTbell Express Template Prep Kit 2.1 (Pacific Biosciences), (iv) the library was treated with nuclease and purified using AMPure PB Beads, and (v) the SMRTbell library was purified using PB beads. The high-quality library was checked for fragment size using the Agilent 2100 Bioanalyzer (Agilent Technologies). Sequencing Primer V2 and Sequel II Binding Kit 2.1 were used for PacBio Sequel II sequencing.

### Genome assembly and chromosomal organization

The adapter and low-quality bases of the 2 species were filtered before assembly using Fastp (RRID:SCR_016962) (v. 0.21.0) [28]. All clean reads of *C. alburnus* and *M. amblycephala* were used for genome assembly using Nextdenovo (RRID:SCR_025033) (v. 2.3.0) [29]. The parameters "random_round = 20, minimap2_options_cns = -x ava-ont -t 40 -k17 -w17, and nextgraph_options = -a 0" were used in genome assembly. The base errors in the genome generated were fixed using Nextpolish (RRID: SCR_025232) (v. 1.3.0) [30]. For the 6 fishes (*C. idella*, *C. molitorella*, *P. parva*, *X. davidi*, *G. rarus*, and *E. bambusa*), the HIFI data was used for genome assembly using hifiasm (0.15.4-r347) software (RRID:SCR_021069) [31].

Hi-C libraries of *C. alburnus* and *M. amblycephala* were created from muscle cells. Briefly, cells were fixed with formaldehyde and

lysed, and the cross-linked DNA was digested with MobI. Sticky ends were biotinylated and proximity ligated to form chimeric junctions that were enriched for and then physically sheared to a size of 300–700 bp, as illustrated in Rao et al. [32]. Chimeric fragments representing the original cross-linked long-distance physical interactions were then processed into paired-end sequencing libraries. The clean reads of Hi-C were obtained from trimming adapter sequences and low-quality pair-end reads, which were truncated at the putative Hi-C junctions, and then the resulting trimmed reads were aligned to the assembly results with BWA (RRID:SCR_010910) (v. 0.7.17) [33]. Invalid read pairs, including Dangling-End and Self-cycle, Re-ligation, and Dumped products, were filtered by HiC-Pro (RRID:SCR_017643) (v. 2.8.1) [34]. They were used for the correction of scaffolds and the clustering, ordering, and orientation of scaffolds onto chromosomes by LACHESIS (release: 2017-12-21) [35]. After this step, placement and orientation errors exhibiting obvious discrete chromatin interaction patterns were manually adjusted.

## Gene prediction and annotation

For protein-coding gene prediction in the genomes of *C. alburnus* and *M. amblycephala*, we employed 3 integrated methods: *de novo* prediction, homology search, and cDNA-based prediction (muscle, brain, liver, intestine, kidney, and testis). *De novo* gene models were predicted using Augustus (RRID:SCR_008417) (v. 3.4.0) [36] with default parameters. In the homolog-based analysis, protein genes from 5 species (*C. carpio*: GCF_018340385.1, *C. auratus*: GCF_003368295.1, *O. macrolepis*: GCA_012432095.1, *P. tetrazona*: GCF_018831695.1, and *D. rerio*: GCF_000002035.6) obtained from NCBI were used to predict gene regions using GeneWise (v. 2.4.1) with default parameters. The cDNA-based approaches involved using Hisat2 (RRID:SCR_015530) (v. 2.1.0) [37] and TransDecoder (RRID:SCR_017647) (v. 5.5.0) software to predict open reading frames (ORFs). Subsequently, we integrated the results of genome annotation using GETA (v. 2.5.7). Gene functional predictions were assigned using Blast-2.11.0+ against public databases, including Swiss-Prot, the Non-Redundant Protein Sequence Database (NR), and the EuKaryotic Orthologous Groups database; KEGG Orthology annotations were conducted with Kofamscan software; and motifs and domains were predicted using Hmmer [38] software against the PFAM database.

To identify repetitive sequences, we utilized both *de novo*–based and homology-based methods. First, LTR_FINDER_parallel (RRID:SCR_018969) (v. 1.1) [39], LTRharvest (GenomeTools, v. 1.6.1) [40], LTR_retriever (RRID:SCR_017623) (v. 2.9.0) [41], and RepeatModeler (RRID:SCR_015027) (v. 2.0.1) software were employed to build a *de novo* repeat library, which was then merged with the Repbase database. RepeatMasker (RRID:SCR_012954) was subsequently used to predict repeat sequences using the new repeat library database. Tandem repeats were detected using Tandem Repeats Finder (TRF). For transfer RNA (tRNA) identification, we used tRNAscan-SE (RRID:SCR_008637) (v 2.0.7) [42], while ribosomal RBA (rRNA) was annotated using Blastn (RRID:SCR_001598) (BLAST v. 2.2.26, e-value: 1e$^{-5}$) against the human rRNA sequence from the Rfam database. The small nuclear RNA (snRNA) and microRNA (miRNA) were searched using the Rfam database and the Infernal (RRID:SCR_011809) (v. 1.0.2) software.

## Comparative phylogenomics

For the phylogeny analyses, we performed multiple whole genome alignments (WGAs) for 17 (nonpolyploid species) and 24 (including 7 polyploid species) species using cactus (v. 2.1.1), respectively [43]. The WGAs were utilized to construct a phylogenetic tree with *Beaufortia kweichowensis* as the root. To facilitate the analysis, syntenic blocks were concatenated into 10-kb windows. Subsequently, a file containing 6 Mb (17 species) and 4.75 Mb (24 species) sequences for each genome was generated, respectively. To build the maximum likelihood tree, we employed RAxML (RRID:SCR_006086) (v. 8.2.12) [44] with the following parameters: -p 12,345 -# 100 -m GTRGAMMA -s all.phylip -o B.kweichowensis -f a -x 12,345 -k -n tree -T 10. The coalescent species tree estimations were performed using Astral (RRID:SCR_001886) (v. 5.15.5) [45]. For estimating divergence times, we used the MCMCTree (RRID:SCR_025348) in PAML (4.9j) [46] with 4 fossil calibration time points. The conserved scores of the 17 nonpolyploid species were estimated using the phastCons tool from the phast packages [47]. VCF files for each species were generated by aligning whole genomes to the zebrafish genome using Chen's methods [48]. To investigate gene flow, we exclusively analyzed the 17 nonpolyploid genomes from non-inbred populations using the ABBA-BABA test implemented in the Dsuite (0.4 r38) software [49] with the D-statistic method. The results were visualized using the Fbranch and dtools.py programs in Dsuite. To ensure a sufficient number of informative sites for analysis within each examined window, we employed a Python script named "ABBABABAwindows.py" to detect window D values. We used a window size of 20 kb with a step size of 10 kb, implemented through the script's parameters "-w 20,000 -m 100 -s 10,000."

## RNA isolation and messenger RNA sequencing

Total RNA from the brain, liver, intestine, muscle, kidney, and testis organs of 3 individuals (*C. alburnus* and *M. amblycephala*) was isolated and purified according to the TRIzol extraction method, respectively [50]. The RNA concentration was measured using NanoDrop technology. Total RNA samples were treated with DNase I (Invitrogen) to remove any contaminating genomic DNA. The purified RNA was quantified using a 2100 Bioanalyzer system (Agilent). The isolated messenger RNA (mRNA) was fragmented with a fragmentation buffer. The resulting short fragments were reverse transcribed and amplified to produce cDNA. The transcriptome data of 36 samples (3 biological replicates) were obtained using DNA nanoball DNBSEQ-T7 (RRID:SCR_017981) technology according to the standard method [51]. The main steps are listed as follows: (i) single-stranded circular libraries were prepared using MGI Library Prep Kits; (ii) after the hybridization of a DNA anchor, a fluorescent probe is attached to the DNA nanoball using combinatorial probe anchor sequencing chemistry; (iii) the high-resolution imaging system captures the fluorescent signal; and (iv) after digital processing of the optical signal, the sequencer generates high-quality and accurate sequencing information. Low-quality bases and adapters were trimmed out using SOAPnuke with the thresholds "-n 0.01 -l 20 -q 0.4 -A 0.25 –cutAdaptor -Q 2 -G –polyX 50 –minLen 150" [52]. The high-quality reads were used in the next analyses.

## Gene expression profiling

All clean reads of *M. amblycephala* and *C. alburnus* were mapped to their corresponding reference genomes using HISAT2 (v. 2.1.0) [37] with default parameters. Then, the mapped files were handled with SAMtools/BCFtools (RRID:SCR_005227) (v. 1.10) [53], while the unique mapped reads were obtained using htseq-count (RRID:SCR_011867) (v. 0.12.4) [54]. The gene expression value of mRNA sequencing (mRNA-seq) was normalized and calculated based on the transcripts per million (TPM) values. Genes with mapped reads

<5 in each sample were not used in our next analyses. Differential expression (DE) analysis was performed using Deseq2 (RRID: SCR_015687) [55] of the R package with the following thresholds: $P < 0.001$ and $Padj < 0.001$. Organ-specific genes (OSGs) were identified using the following criteria: gene expression in the target tissue or organ differs significantly from that in the other 5 tissues and organs. Orphan genes (OGs) were detected based on the thresholds of BLASTx with an e-value of $1e^{-5}$ and tBLASTx with an e-value of $1e^{-5}$. The sequences with no BLAST result in the public database were considered potential OGs. Then, the expressed OGs (TPM >10) were considered OGs in the corresponding tissue or organ. GO analysis was performed with a significance threshold (false discovery rate of the Benjamini–Hochberg method <0.05).

## Diversifying selection analysis

In total, 17,337 orthologous gene pairs (OGPs) between *M. amblycephala* and *C. alburnus* were obtained using the all-against-all reciprocal BLASTP (v. 2.8.1) with an e-value of $1e^{-6}$ based on protein sequences (sequence alignment >70%). Then, transcripts that were shorter than 300 bp were discarded from OGPs. OGPs in the comparison of *M. amblycephala* and zebrafish and the comparison of *M. amblycephala* and zebrafish were obtained based on the above thresholds. We performed DE analyses on the OGPs between *M. amblycephala* and *C. alburnus* in the 6 tissues and organs. DE analysis was performed using Deseq2 [55] with the following thresholds: $P < 0.001$ and $Padj < 0.001$. The Ks and Ka/Ks values were calculated based on the following analysis process: (i) ParaAT2.0 and muscle software were used in sequence alignment of OGPs with the default parameters, and (ii) kaks_calculator3.0 program was used to calculate Ks and Ka/Ks values using the maximum likelihood method [56]. The threshold of $P < 0.05$ was used in our analyses.

## Measurement of enzymatic content

Equal amounts of liver (0.1 g) from *M. amblycephala* and *C. alburnus* (10 individuals in each species) were collected from the Engineering Center of Polyploid Fish Breeding of National Education Ministry in Hunan, China. Then, homogenates were used to determine the activity of trypsin and lipase. A trypsin assay kit (A080-2-2) and a lipase assay kit (A054-1-1) were purchased from Nanjing Jiancheng Bioengineering Institute (Jiangsu, China), and the experimental protocols followed the manufacturer's instructions. The significant difference was performed using Student's *t*-test.

## Hematoxylin and eosin staining

A 10-mm-thick section of skeletal muscle from 4 fish species, *M. amblycephala*, *C. alburnus*, *C. idella*, and *E. bambusa*, was dissected from the dorsum region. A 10-mm-thick section of intestine was dissected from the abdominal cavity of each of the 4 fish species after removing food residues. The tissue sections were then fixed in Bouin's solution for 24 hours. After fixation, the tissues were washed with distilled water for 4 hours at room temperature. The fixed tissues were then dehydrated using a series of alcohol concentrations (e.g., 70%, 80%, 90%, and 100%) and embedded in paraffin blocks. The paraffin-embedded tissue blocks were sectioned into 10-$\mu$m-thick slices using a microtome. The tissue sections were processed for hematoxylin and eosin (HE) staining according to the manufacturer's instructions using an HE staining kit. Digital images of the stained sections were captured using a microscope (DX8; Olympus). The samples obtained from 3 individuals were performed for each hybrid variety, and quantitative data on HE staining were collected from them.

# Results

## Genome assembly

A total of 8 species of cyprinid fishes from East Asia were sequenced using PacBio HiFi or Oxford Nanopore technology, resulting in over 602.21 Gb of raw data (Supplementary Table S1). *De novo* assembled genomes were obtained with contig N50 ranging from 7.57 to 38.12 Mb. Chromosome-scale genomes were assembled for blunt snout bream (*Megalobrama amblycephala*, BSB) and topmouth culter (*Culter alburnus*, TC) using 204.8 Gb Hi-C data. The resulting assemblies exhibited scaffold N50 values of 42.91 Mb and 39.60 Mb, respectively (Table 1 and Supplementary Tables S1–S2). Assembly quality was assessed using BUSCO, scoring between 94.5% and 98.7% (Supplementary Table S3). We significantly improved the genome assemblies for both *M. amblycephala* (contig N50 increased from 2.4 to 15.42 Mb [57]) and *C. alburnus* (contig N50 increased from 17.8 to 18.55 Mb [58]). High-quality genome data for *Cirrhinus molitorella*, *Pseudorasbora parva*, and *Xenocypris davidi* were presented for the first time. Through a combination of *de novo*, protein homology, and cDNA-based prediction, we annotated 26,550 and 27,303 protein-coding genes for *M. amblycephala* and *C. alburnus*, respectively (Supplementary Tables S4–S5). Repetitive elements comprised 51.81% (568.18 Mb) and 50.49% (544.26 Mb) of the assemblies for *M. amblycephala* and *C. alburnus*, respectively (Supplementary Table S6). Noncoding RNA was predicted in 8.76% and 7.21% of the genome assemblies for *M. amblycephala* and *C. alburnus*, respectively (Supplementary Table S7). Furthermore, we obtained high-quality assembled genomes for 15 cyprinid fishes (with an average scaffold N50 of 34.39 Mb) and *Beaufortia kweichowensis* from public databases (Supplementary Table S8). These 23 cyprinid fishes represent 7 nonpolyploid subfamilies (*Danioninae*, *Xenocyprinae*, *Gobioninae*, *Leuciscinae*, *Cultrinae*, *Labeoninae*, and *Hypophthalmichthyinae*) and 3 polyploid subfamilies (*Schizothoracinae*, *Barbinae*, and *Cyprininae*), with genome sizes ranging from 0.86 to 1.90 Gb and chromosome numbers varying widely from 48 to 150 (Table 1, Supplementary Tables S1–S2, and S8).

## Phylogenomic analyses and introgression

To investigate the evolutionary relationships among extant cyprinids, we analyzed 24 genomes with butterfly hillstream loach (*B. kweichowensis*) as an outgroup (Fig. 1A and Supplementary Fig. S1). Our results showed that *Gobiocypris rarus* belongs to the subfamily *Gobioninae*, even though from a morphological perspective, it appears similar to zebrafish (which belongs to the subfamily *Danioninae* of Cyprinidae). Molecular clock analysis with fossil calibration indicated their divergence time ranging from 41.5–61.3 MYA (Fig. 1A and Supplementary Table S9). Phylogenetic trees reconstructed the evolutionary history of subfamily *Leuciscinae*, showing that 1 group (including *Leuciscus idus*, *Abramis brama*, and *Rutilus rutilus*) diverged from another group (*Ctenopharyngodon idella* and *Elopichthys bambusa*) ranging from 23.8 to 35.1 MYA. This divergence occurred earlier than the divergence times observed among other subfamilies (*Cultrinae*, *Gobioninae*, *Xenocyprinae*, and *Hypophthalmichthyinae*) (Fig. 1A and Supplementary Table S9). Furthermore, phylogenomics analysis provided evidence regarding the divergence of common ancestors of extant cyprinids, which ranged from 81.9 to 100.0 MYA (Fig. 1A and Supplementary Table S9). The ancestor of *Danio rerio* (subfamily: *Danioninae*) diverged early in the evolution of extant cyprinids, while the ancestor of *Cirrhinus molitorella* and *Labeo rohita* (subfamily: *Labeoninae*) diverged between 36.5 and 53.6 MYA (Fig. 1A and Supplementary Table S9). These findings will assist us in understanding the evolutionary process of fish and constructing more reasonable clas-

**Table 1:** Assembly statistics of 8 species of cyprinid fishes

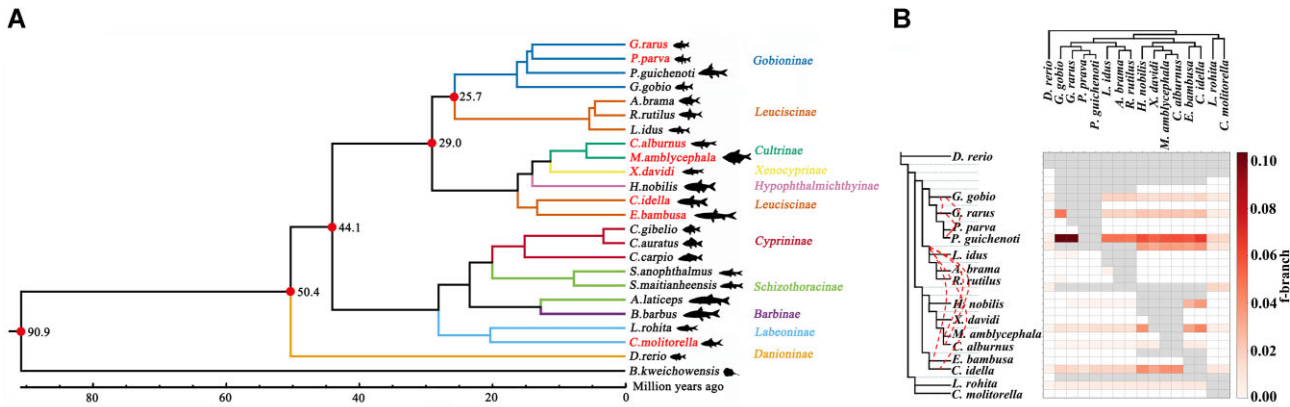| Species | Common name | Subfamily | Scaffold N50 (Mb) | Contig N50 (Mb) | Contig length (Mb) | BUSCO completeness (%) |
|---|---|---|---|---|---|---|
| *Megalobrama amblycephala* | Blunt snout bream | *Cultrinae* | 42.91 | 15.42 | 1,096.68 | 94.5% |
| *Culter alburnus* | Topmouth culter | *Cultrinae* | 39.60 | 18.55 | 1,077.98 | 98.3% |
| *Ctenopharyngodon idella* | Grass carp | *Leuciscinae* | / | 35.62 | 901.49 | 98.5% |
| *Cirrhinus molitorella* | Mud carp | *Labeoninae* | / | 38.12 | 1,066.79 | 98.6% |
| *Pseudorasbora parva* | Stone moroko | *Gobioninae* | / | 7.57 | 1,292.12 | 98.5% |
| *Xenocypris davidi* | Bleeker's yellow tail | *Xenocyprinae* | / | 38.11 | 1,044.52 | 98.7% |
| *Gobiocypris rarus* | Raregudgeon | *Danioninae* | / | 13.24 | 1,108.08 | 98.3% |
| *Elopichthys bambusa* | Yellow cheek carp | *Leuciscinae* | / | 30.38 | 863.54 | 98.4% |



**Figure 1:** Phylogenomic analyses of cyprinid fish. (A) Time-calibrated phylogenetic tree. Red pot represents fossil calibration time points, which were obtained from "Timetree of Life." Black number in the branch represents the median of the range of divergence times. The name marked in red represents the species sequenced in this study. (B) Gene flow determined using the f-branch method. Dashed red line represents gene flow (f4-ratio > 0.05) between 2 species.

sification relationships within the cyprinids, with the support of data from fields such as fossils, monsoons, and geography [59, 60].

Previous studies have reported phylogenetic discordance across genome regions in nonpolyploid cyprinid fishes from the East Asian region. This discordance has been attributed to incomplete lineage sorting, introgression, and the fish's demographic history [61]. To investigate this, we conducted gene flow analysis and observed pervasive introgression among the 17 nonpolyploid cyprinid fishes (f4-ratio > 0.0006, z-score > 3, and $P < 0.05$) (Fig. 1B). The ABBA-BABA tests [49, 62] revealed strong gene flow events among subfamily *Gobioninae*, including *Paracanthobrama guichenoti* and *Gobiocypris rarus* (f4-ratio = 0.1, z-score = 129, and $P < 0.001$), *G. rarus* and *Gobio gobio* (f4-ratio = 0.1, z-score = 120, and $P < 0.001$), and *P. guichenoti* and *G. gobio* (f4-ratio = 0.1, z-score = 96, and $P < 0.001$) (Fig. 1B and Supplementary Table S10). Notably, distinct gene flow signals (f4-ratio > 0.05) between 2 species were predominantly detected in 22 groups involving 12 species and 7 subfamilies (*Cultrinae*, *Danioninae*, *Gobioninae*, *Hypophthalmichthyinae*, *Leuciscinae_1*, *Leuciscinae_2*, and *Xenocyprinae*) (red dotted line in Fig. 1B and Supplementary Table S10). Recent studies utilizing mitochondrial genomes and *de novo* nuclear genomes have also highlighted frequent gene flow events during the radiation of cyprinid fishes [60]. Our findings, based on high-quality genome assemblies, support the hypothesis that gene flow is the primary driver of the observed phylogenetic incongruence among nonpolyploid cyprinid fishes.

## Conservation of the reproductive system in speciation

The frequent occurrence of gene flow events between cyprinid fishes, including *M. amblycephala* and *C. alburnus*, suggests incomplete reproductive isolation as a potential contributing factor to their speciation in the East Asian region. To investigate the underlying genetic mechanisms, we focused on comparing the genomes of these 2 species, which belong to different genera within the *Cultrinae* subfamily but share overlapping habitats in the middle and lower reaches of the Yangtze River Basin (Fig. 2A). While laboratory experiments have indicated some degree of postzygotic isolation, no natural hybrid populations have been identified in the wild [63]. A conserved synteny analysis revealed a high degree of gene conservation between *M. amblycephala* and *C. alburnus* (Supplementary Figs. S2–S3), making them a suitable model for studying the genetic basis of their speciation.

To investigate the genetic differences between *M. amblycephala* and *C. alburnus*, we conducted comparative gene expression analyses in 6 tissues: brain, liver, intestine, muscle, kidney, and testis (Supplementary Tables S11–S12). Our findings revealed that the testis exhibited a higher proportion of OSG, accounting for 6.67% in *C. alburnus* and 7.95% in *M. amblycephala*, compared to the intestine and kidney (Fig. 2B). We focused on OGs [64] and identified a greater number of OGs in the testis of both species, suggesting rapid divergence in this organ (Fig. 2C and Supplementary Table S13). However, the number of differentially expressed genes (DEGs) between the 2 species was lower in the testis than in
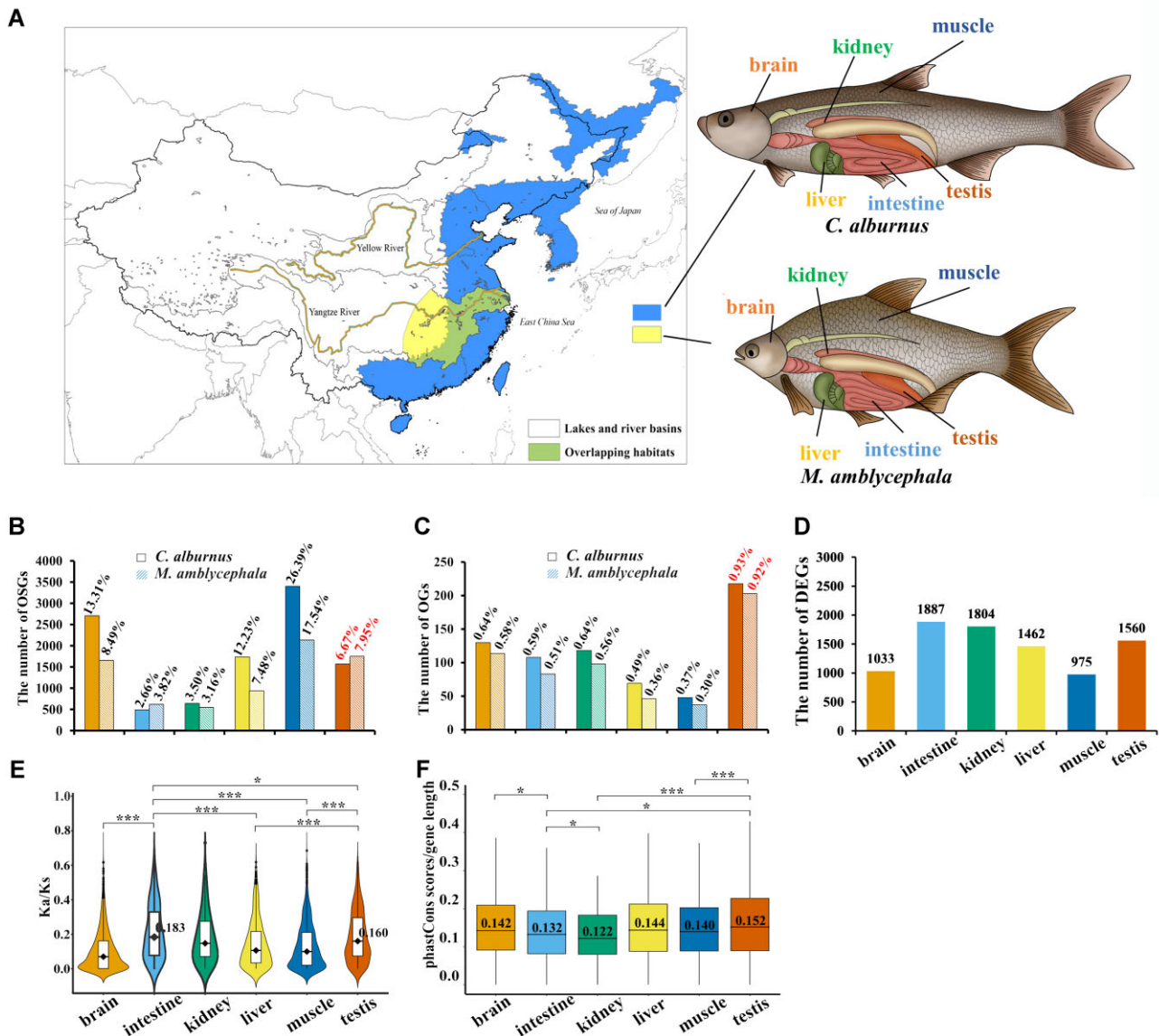
**Figure 2:** Divergent evolution of *M. amblycephala* and *C. alburnus* in East Asia. (A) Habitat distribution of extant *M. amblycephala* [65] and *C. alburnus*. Partial overlap in the habitats of the 2 species. (B) The number and percentage of OSGs in *M. amblycephala* and *C. alburnus*. (C) The number and percentage of OGs in *M. amblycephala* and *C. alburnus*. (D) Differential expression between *M. amblycephala* and *C. alburnus* in 6 tissues and organs. (E) The distribution of Ka/Ks values relating to OSGs in the 6 tissues and organs. The median value (black dot) was signed in the figure. (F) Conserved scores of OSGs in the 6 tissues and organs. The median value (black line and number) was signed in the figure. The comparisons involve the intestine and testis. "*" represents "0.01 < P ≤ 0.05," "**" represents "0.001 < P ≤ 0.01," and "***" represents "P ≤ 0.001."

the 3 tissues (Fig. 2D, Supplementary Fig. S4, and Supplementary Tables S14–S16). To explore functional divergence, we calculated Ka/Ks values for OSGs and found that the testis exhibited lower Ka/Ks values compared to the intestine (t-test: $P = 0.04$) but higher values compared to the 3 tissues (liver, brain, and muscle, t-test: $P < 0.001$) (Fig. 2E). Similar patterns were observed for DEGs and the shared genes (OSGs and DEGs) (Supplementary Fig. S5). Finally, to assess the degree of sequence conservation [47], we calculated phastCons scores of OSGs and observed the highest median value in the testis, which was higher than in the intestine, kidney, and muscle (t-test: $P < 0.05$) (Fig. 2F). A similar phenomenon was noted when analyzing DEGs using phastCons scores (Supplementary Fig. S6). *M. amblycephala* and *C. alburnus* exhibited lower genetic variation in their testes compared to their intestines.

## Rapid evolution of digestive system in speciation

The diverse digestive systems of cyprinid fishes allow them to adapt to various food sources, including plankton, aquatic plants, and benthic organisms, contributing to their ecological niche differentiation [66]. Differential expression analyses revealed that the intestine of *C. alburnus* had the lowest number of OSGs (752, 4.15%), while *M. amblycephala*'s intestine had the second lowest (942, 5.95%) (Fig. 2B). The number of OGs in the intestine (0.59% in *C. alburnus* and 0.51% in *M. amblycephala*) was lower compared to the testis, brain, and kidney (Fig. 2C). The study demonstrates that the genetic makeup of the intestine exhibits a higher degree of conservation compared to the other tissues and organs. However, rapid genetic divergence between the herbivorous *M. amblycephala* and the carnivorous *C. alburnus* was observed in their intestines. For instance, the highest number of DEGs be-
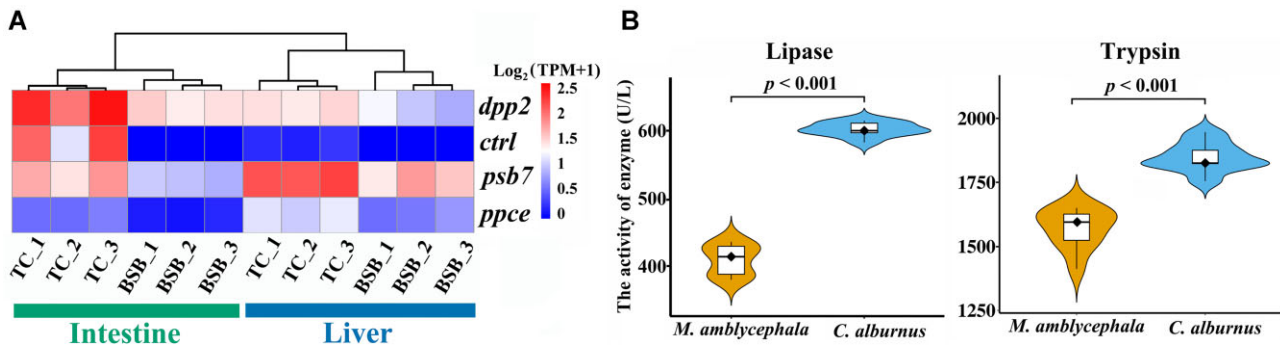
**Figure 3:** Diet divergences between *M. amblycephala* and *C. alburnus*. (A) The 4 genes relating to differential expression between *M. amblycephala* (BSB) and *C. alburnus* (TC) in both the intestine and liver (3 biological replicates showed "_1," "_2," and "_3"). (B) Significant differences in the enzyme activity of lipase and trypsin for the comparison between *M. amblycephala* and *C. alburnus*.

tween the 2 species was detected in their intestines (Fig. 2D and Supplementary Tables S14–S15). Moreover, the Ka/Ks values of OSGs were higher in the intestine compared to the brain, kidney, muscle, and testis (Fig. 2E). Lastly, the phastCons scores of OSGs in the intestine were lower than those in the brain, liver, muscle, and testis, although they were higher than in the kidney (Fig. 2F). Similar trends were observed in the phastCons scores of DEGs (Supplementary Fig. S6). Our findings provide preliminary evidence suggesting a potential rapid divergence in genetic diversity within the digestive organs of *M. amblycephala* and *C. alburnus*.

To investigate the genetic basis of diet divergence between herbivorous *M. amblycephala* and carnivorous *C. alburnus*, we conducted a functional analysis of their DEGs in the intestine. These genes associated with digestive enzymes were enriched for hydrolyzing O-glycosyl compounds (GO: 0004553) and peptidase activity (GO: 0008233) in terms of Molecular Function annotation, while carbohydrate metabolic process (GO: 0005975) and lipid catabolic process (GO: 0016042) were enriched for Biological Process annotation (Supplementary Fig. S7). Among these genes, *dpp2*, *ctrl*, *psb7*, and *ppce* were identified as potential genes involved in peptidase activity, exhibiting higher expression in the digestive organs (liver and intestine) of *C. alburnus* compared to *M. amblycephala* (Fig. 3A). After detecting the enzyme activities of trypsin and lipase in digestive organs, we found that the enzyme activities were higher in the carnivorous *C. alburnus* compared to the herbivorous *M. amblycephala* (Fig. 3B). We conducted analyses on positively selected genes (PSGs) (Ka/Ks > 1) between the 2 species and identified 30 of them that belong to OSGs in the 6 tissues and organs (Supplementary Table S17). Among the share genes of PSGs and OSGs in the intestine, *caspbl* and *vsig* were found to be associated with peptidase activity (GO: 0008233), apoptosis, and immune responses, which are closely related to the types of digested food (Supplementary Fig. S8) [67, 68]. These findings suggest that the observed genetic diversity may be related to adaptations in digestive enzyme secretion, reflecting potential dietary adjustments.

## Gene flow and its potential impact on feeding habits

Frequent gene flow events were observed among cyprinid fishes, including *M. amblycephala*, *C. alburnus*, *C. idella*, and *E. bambusa*. Among these, significant gene flow was detected between carnivorous *C. alburnus* (subfamily *Cultrinae*) and *E. bambusa* (subfamily *Leuciscinae*) (z-score > 45.8, f4-ratio = 0.039, and *P* < 0.001). Additionally, gene flow event was identified between the herbivorous *M. amblycephala* (subfamily Cultrinae) and *C. idella* (subfamily *Leu-*

*ciscinae*) (z-score > 38.3, f4-ratio = 0.044, and *P* < 0.001) (Fig. 1B and Supplementary Table S10). The overlapping habitats of these 4 species were primarily distributed in the eastern region of China (Fig. 4A).

To investigate the effects of gene flow events on diet diversity, we analyzed the 117 introgressed genomic regions (window size: 20 kb) between the 2 carnivorous fish species, which were associated with 69 genes (Supplementary Table S18). Similarly, the 102 introgressed regions (window size: 20 kb) between the 2 herbivorous fish species were associated with 68 genes (Supplementary Table S19). Among these genes, the top 3 molecular function annotations in both carnivores and herbivores were related to transcription regulator activity (GO: 0140110), DNA-binding transcription factor activity (GO: 0003700), and RNA polymerase II–specific activity (GO: 0000981) (Supplementary Tables S20–S21). Among these genes, 10 introgressed genes were shared between carnivores and herbivores (Fig. 4B). Moreover, 84 categories (38.36%) for biological processes and 16 categories (66.67%) for molecular functions were shared between carnivores and herbivores false discovery rate (FDR < 0.05, Fig. 4C and Supplementary Figs. S9–S10). These shared introgressed genes were associated with animal organ development, including skeletal muscle organ development (GO: 0060538; *sox6* and *ttn.2*) and head development (GO: 0060322; *zfhx3*, *tcf7l2*, and *meis1b*) (Supplementary Fig. S11). *Zbtb16a* (linked to osteogenic differentiation [69]) exhibited the broadest expression pattern among all introgressed genes, being detected in 5 different tissues and organs (Fig. 4D). This suggests that *zbtb16a* may be a hotspot for introgression events between *Cultrinae* and *Leuciscinae*.

There were potential relationships between diet habits and organ development, including mouth and pharyngeal tooth morphologies, intestinal morphology, and skeletal muscle structure (Fig. 4E). Comparative analyses revealed that carnivorous fishes exhibited larger and superior mouths, longer and sharper teeth, shorter intestines, thinner intestine linings, and smaller cross-sectional areas in skeletal muscle fibers compared to herbivorous fishes (Fig. 4). It is noteworthy that some of the introgressed genes have been experimentally validated in zebrafish to have functional associations with dietary traits. For instance, *tp53* and *tle3a* are implicated in intestinal morphology [70, 71], while *grin2bb* and *grin1a* are linked to food intake behavior [72, 73]. Additionally, *znf536*, *zfhx3*, *elavl4*, *hoxc11a*, *pik3r3b*, and *irf2bpl* have been identified as potential regulators of swimming behavior. These findings suggest that these introgressed genes may relate to organ development and feeding behavior may contribute to diet divergence for these fishes.
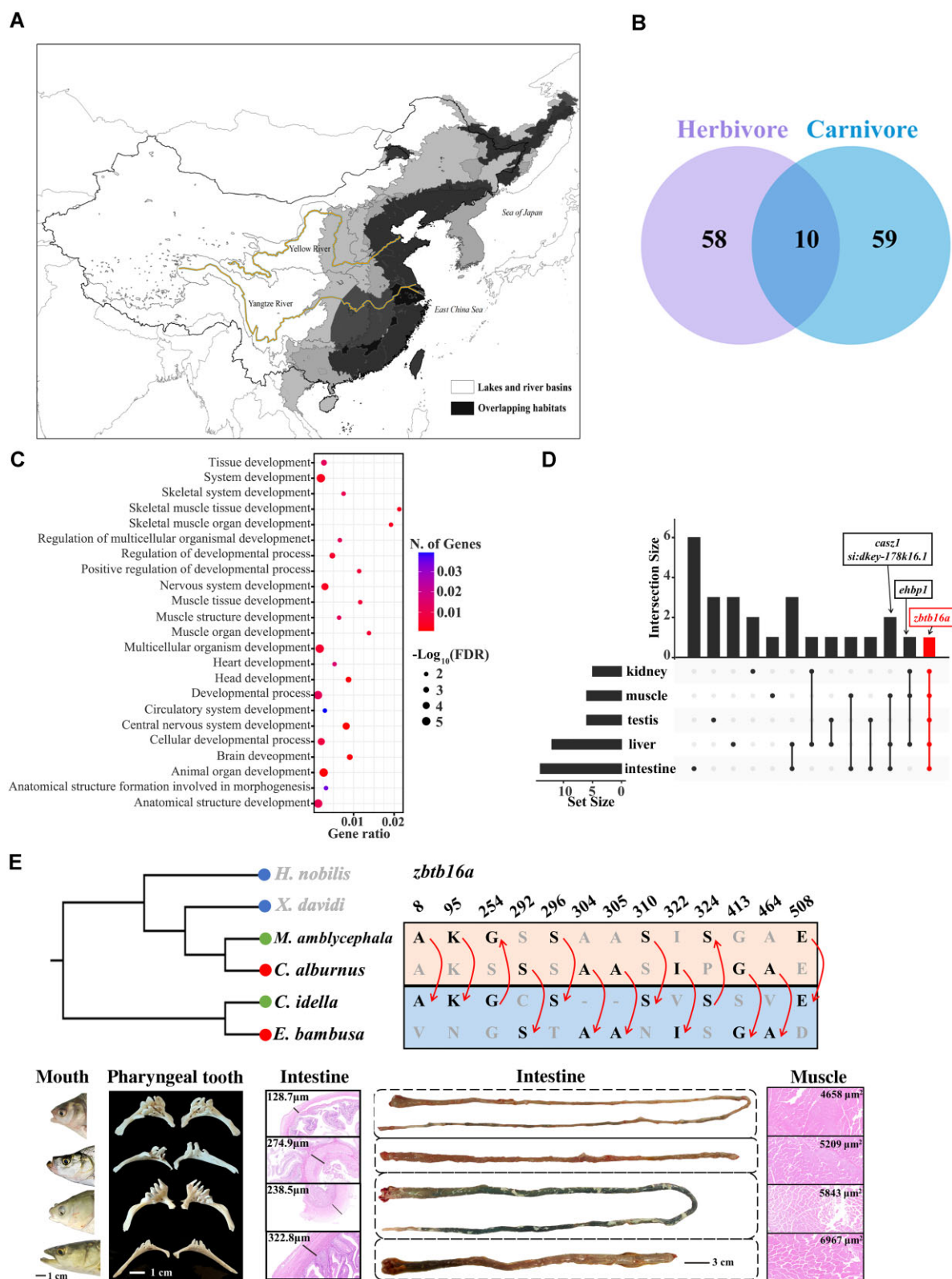
**Figure 4:** Gene flow in herbivorous and carnivorous fishes. (A) Habitat distribution of *M. amblycephala* [65], *C. alburnus*, *C. idella*, and *E. bambusa* in East Asia. Their overlap and unique habitats reflect their speciation of endemic East Asian cyprinid fishes in the river-lake ecosystems of East Asia. (B) Venn diagram showing introgressed genes between carnivores and herbivores. (C) The introgressed genes were expressed in different tissues and organs, while there was no introgressed gene expressed in the brain. (D) GO analyses of the shared 10 introgressed genes in carnivores and herbivores. (E) Phylogenomic analyses of the 6 endemic cyprinid species in China; introgressions visualizing at SNPs of *zbtb16a* (potential introgressed SNPs marking red arrow); the morphologies of mouth, pharyngeal teeth, and intestine; microstructure of intestine (thickness of intestinal wall marking in figure); and muscle (average area marking in figure) in the 4 endemic cyprinid species. Blue dot represents filter-feeding fish, green dot represents herbivorous fish, and red dot represents carnivorous fish.

## Discussion

The East Asian region, characterized by its unique topography, including the uplift of the Qinghai–Tibet Plateau, abundant rivers, and diverse climatic environments, is home to a multitude of freshwater fish species [60, 74, 75]. Cyprinids represent the largest and most diverse vertebrate group, with over 654 species, including 440 endemics in China [76]. Investigating the genetic mechanisms underlying their rapid speciation is vital for understanding evolutionary radiation in East Asian cyprinids. Our findings suggest that frequent gene flow events among cyprinid fishes have contributed to their rapid adaptive radiation, as evidenced by our analyses of 7 nonpolypoid subfamilies. This prompts the question: How does introgressive hybridization in East Asian cyprinids relate to rapid speciation?

Rapid evolutionary changes in the mammalian testis manifest at the molecular level, contributing to reproductive isolation. Comparative gene expression studies across various mammalian organs reveal that the testis exhibits the highest rates of evolutionary expression change [77, 78]. Therefore, comparative genomic analysis of *M. amblycephala* and *C. alburnus* revealed a lower degree of genetic divergence in the testis compared to the intestine. Additionally, laboratory experiments have demonstrated the production of fertile hybrids between various cyprinid species [13, 63], suggesting the possibility of incomplete RI and the prevalence of introgressive hybridization in East Asian cyprinids. However, further research is needed to elucidate the specific factors hindering and driving rapid speciation in these cyprinid fishes.

The complex and variable inland water ecosystem plays a crucial role in the adaptive evolution of fish [79, 80]. Among these factors, the diversity of food sources gradually influences the feeding habits of different populations, resulting in the adaptive evolution of their digestive and locomotion systems [81]. Our findings suggest that speciation in *M. amblycephala* and *C. alburnus* may have been driven by diet-dependent adaptations. Cyprinid fishes display significant diversity in behavior, habitat, geography, and morphology, including variations in feeding and digestive organs [60, 76]. Robust pharyngeal teeth and toothless jaws enable them to consume a wide range of foods [82, 83]. Our results reveal that rapid genetic divergence between *M. amblycephala* and *C. alburnus* occurs in the intestine. The variations in digestive enzyme secretion and digestive organs reflect their distinct feeding preferences and the effectiveness with which they metabolize different food types [84]. Considering the absence of postzygotic isolation and the overlapping habitat between *M. amblycephala* and *C. alburnus* [63, 85], our results suggest that ecological differentiation driven by dietary differences may be an important factor leading to the rapid formation of these 2 species.

When postzygotic reproductive isolation is no longer a significant barrier to gene flow among East Asian cyprinids, natural selection, including monsoon activities [60], and the uplift of the Qinghai–Tibet Plateau [75], can attenuate prezygotic isolation, providing opportunities for gene flow, thus promoting speciation in cyprinid fishes [86]. To adapt to diverse food supplies in different aquatic environments, feeding habits have diverged in the subfamilies *Cultrinae* (herbivorous *M. amblycephala* and carnivorous *C. alburnus*) and *Leuciscinae* (herbivorous *C. idella* and carnivorous *E. bambusa*). Does gene flow facilitate the divergence of diets for adaptive evolution? Our results demonstrate the introgression of genes associated with skeletal muscle and head development between fishes with the same diet. These changes play crucial roles in feeding and digestive efficiency. Fishes with the same diet in different subfamilies exhibit similar phenotypes involving the mouth, teeth, intestine, and muscle. These results suggest that coevolving interactions of diet habits occur in their speciation through introgressive hybridization. However, further evidence is needed to establish a definitive association between dietary convergent evolution and gene flow.

## Additional Files

**Supplementary Table S1.** Summary of whole genome sequencing in 8 species.

**Supplementary Table S2.** Genome assembly of 8 species.

**Supplementary Table S3.** Completeness of the 8 assembled genomes.

**Supplementary Table S4.** Statistics of gene prediction.

**Supplementary Table S5.** Gene function annotation of *M. amblycephala* and *C. alburnus*.

**Supplementary Table S6.** Summary of repeat contents.

**Supplementary Table S7.** The summary of predicted noncoding RNA in *M. amblycephala* and *C. alburnus*.

**Supplementary Table S8.** Information of downloaded genomes.

**Supplementary Table S9.** Range of divergence time in Fig. 1.

**Supplementary Table S10.** D statistic on the species tree based on genome-wide single-nucleotide polymorphisms (SNPs). The outgroup was fixed as *Beaufortia kweichowensis*.

**Supplementary Table S11.** Summary of transcriptome sequencing data.

**Supplementary Table S12.** Summary of transcriptome mapping data.

**Supplementary Table S13.** List of orphan genes (OGs) in *M. amblycephala* and *C. alburnus*.

**Supplementary Table S14.** Summary of differentially expressed genes (DEGs) between *M. amblycephala* and *C. alburnus* in 6 organs.

**Supplementary Table S15.** The gene number of orthologous gene pairs and differential expressed genes (DEGs) in 6 organs.

**Supplementary Table S16.** The summary of organ-specific genes (OSGs), positive selective genes (PSGs), and differential expressed genes (DEGs) in 6 organs.

**Supplementary Table S17.** Summary of expressed positive selective genes (PSGs) between *C. alburnus* (TC) and *M. amblycephala* (BSB) in the 6 organs.

**Supplementary Table S18.** Summary of gene flow between carnivorous *C. alburnus* and *E. bambusa*.

**Supplementary Table S19.** Summary of gene flow between herbivorous *M. amblycephala* and *C. idella*.

**Supplementary Table S20.** GO enrichment of introgressed genes between carnivorous *C. alburnus* and *E. bambusa*.

**Supplementary Table S21.** GO enrichment of introgressed genes between herbivorous *M. amblycephala* and *C. idella*.

**Supplementary Fig. S1.** Phylogenetic trees constructed using multiple whole-genome alignments of 17 (no polyploid species) and 24 (including 7 polyploid species) species with *Beaufortia kweichowensis* as the root, respectively. (A) Concatenation-based method for estimating a phylogenetic tree of 17 species with 10-kb length windows. (B) Coalescent method for estimating a phylogenetic tree of 17 species with 10-kb length windows. (C) Concatenation-based method for estimating a phylogenetic tree of 24 species with 10-kb length windows. (D) Coalescent method for estimating a phylogenetic tree of 24 species with 10-kb length windows. (E) Species tree with estimated divergence time.

**Supplementary Fig. S2.** The Hi-C interaction heatmap of 24 linkage groups in the genomes of *M. amblycephala* and *C. alburnus*.

**Supplementary Fig. S3.** The collinearity analysis between *M. amblycephala* and *C. alburnus*. Twenty-four pairs of homologous

chromosomes were determined based on 17,337 orthologous gene pairs.

**Supplementary Fig. S4.** The differential expression between *M. amblycephala* and *C. alburnus* in 6 organs. Upregulated genes in *M. amblycephala* are marked in blue, while the upregulated genes in *C. alburnus* are marked in red.

**Supplementary Fig. S5.** The distribution of Ka/Ks values in the 6 organs. (A) The Ka/Ks values of DEGs. (B) The Ka/Ks values of shared genes between DEGs and OSGs. The median value is indicated by a black dot, and the gene number is provided below each organ name. In the *t*-test, "∗" represents $0.01 < P \leq 0.05$, "∗∗" represents $0.001 < P \leq 0.01$, and "∗∗∗" represents $P \leq 0.001$.

**Supplementary Fig. S6.** Conserved scores of DEGs (*M. amblycephala* vs. *C. alburnus*) in the 6 organs. "∗" represents $0.01 < P \leq 0.05$, "∗∗" represents $0.001 < P \leq 0.01$, and "∗∗∗" represents $P \leq 0.001$.

**Supplementary Fig. S7.** The DEGs (*M. amblycephala* vs. *C. alburnus*) associated with diet habit. (A) The heatmap of the DEGs in the intestine. The hydrolyzing O-glycosyl compounds and peptidase activity in Molecular Function, as well as carbohydrate metabolic process and lipid catabolic process in Biological Process. (B) The gene distribution of DEGs in the intestine.

**Supplementary Fig. S8.** The alignment of 2 positively selected genes (PSGs) in the intestine.

**Supplementary Fig. S9.** The GO terms of introgressed genes in the carnivorous (*C. alburnus* and *E. bambusa*) and herbivorous (*M. amblycephala* and *C. idella*) fishes.

**Supplementary Fig. S10.** The distribution of enriched functional categories (FDR < 0.05) in Biological Process and Molecular Function for the introgressed genes.

**Supplementary Fig. S11.** Heatmap exhibiting the expression of introgressed genes in the carnivorous (*C. alburnus* and *E. bambusa*) and herbivorous (*M. amblycephala* and *C. idella*) fishes.

## Abbreviations

BSB: blunt snout bream; DE: differential expression; DEGs: differentially expressed genes; MYA: million years ago; OGPs: orthologous gene pairs; OGs: orphan genes; OSG: organ-specific gene; PSGs: positively selected genes; RI: reproductive isolation; TC: topmouth culter; TPM: transcripts per million; TRF: tandem repeats finder.

## Animal Ethics Declarations

All procedures performed on animals were approved by the academic committee at Hunan Normal University, Hunan, China (approval number: 2020C034).

## Acknowledgments

We thank Min Xie at Hunan Fisheries Science Institute for their invaluable assistance in collecting the fish samples.

## Author Contributions

Li Ren (Conceptualization [equal], Funding acquisition [equal], Formal analysis [equal], Writing—original draft [equal], Writing—original draft [equal], Writing—review & editing [equal]), Xiaolong Tu (Formal analysis [equal], Writing—original draft [equal]), Mengxue Luo (Formal analysis [equal]), Qinzhi Liu (Writing—review & editing [equal]), Jialin Cui (Resources [equal]), Xin Gao (Resources [equal]), Hong Zhang (Resources [equal]), Yakui Tai (Resources [equal]), Yiyan Zeng (Resources [equal]), Mengdan Li (Resources [equal]), Chang Wu (Resources [equal]), Wuhui Li (Resources [equal]), Jing Wang (Resources [equal]), Dongdong Wu (Conceptualization [equal], Writing—review & editing [equal]), and Shaojun Liu (Conceptualization [equal], Funding acquisition [equal], Writing—review & editing [equal]).

## Funding

## Data Availability

Genomic sequencing data obtained from PacBio HiFi, Oxford Nanopore, and DNBSEQ-T7 technologies, as well as Hi-C data, have been submitted to the National Center for Biotechnology Information (NCBI) (accession numbers: SRR26190421–SRR26190427, SRR26139312–SRR26139313, SRR26139214–SRR26139215, and SRR26319599–SRR26319600). The assembled genome and annotation files of 8 cyprinid fishes have been deposited on Figshare [87] and the National Genomics Data Center (NGDC) (accession numbers: GWHDOEU00000000, GWHDOEX00000000, GWHDOEV00000000, GWHDOEW00000000, GWHDOEB00000000, GWHDOEC00000000, GWHDOES00000000, and GWHDOET00000000). The raw reads of the mRNA-seq data have been submitted to NGDC (accession number: sub-CRA017373) and NCBI (accession numbers: SRR26087118–SRR26087153). All additional supporting data are available in the *GigaScience* repository, GigaDB [88–96].

## Competing Interests

The authors have declared that no competing interests exist.

## References

1. Froese R, Pauly D. FishBase. 2014. https://fishbase.mnhn.fr/summary/FamilySummary.php?ID=122.

2. Nelson J, Grande T, Wilson M. Fishes of the World. 5th ed. Hoboken, New Jersey, US: John Wiley & Sons; 2016.

3. Yang L, Sado T, Vincent Hirt M, et al. Phylogeny and polyploidy: resolving the classification of cyprinine fishes (Teleostei: cypriniformes). Mol Phylogenet Evol 2015;85:97–116. https://doi.org/10.1016/j.ympev.2015.01.014.

4. He FZ, Zarfl C, Bremerich V, et al. The global decline of freshwater megafauna. Global Change Biol 2019;25(11):3883–92. https://doi.org/10.1111/gcb.14753.

5. Jacquemin SJ, Pyron M. A century of morphological variation in Cyprinidae fishes. BMC Ecol 2016;16(1):48. https://doi.org/10.1186/s12898-016-0104-x.

6. German DP, Nagle BC, Villeda JM, et al. Evolution of herbivory in a carnivorous clade of minnows (teleostei: cyprinidae): effects on gut size and digestive physiology. Physiol Biochem Zool 2010;83(1):1–18. https://doi.org/10.1086/648510.

7. Yue PQ, Shan XH, Lin RD. Fauna sinica, osteichthyes, cypriniformes III. Beijing: Science Press; 2000.

8. Haenen O, Way K, Gorgoglione B, et al. Novel viral infections threatening cyprinid fish. Bull Eur Assoc Fish Pathol 2016;36(1):11–23.

9. Brauer CJ, Sandoval-Castillo J, Gates K, et al. Natural hybridization reduces vulnerability to climate change. Nat Clim Chang 2023;13:282–89. https://doi.org/10.1038/s41558-022-01585-1.

10. Mallet J. Hybridization as an invasion of the genome. Trends Ecol Evol 2005;20(5):229–37. https://doi.org/10.1016/j.tree.2005.02.010.

11. Schumer M, Powell DL, Delclós PJ, et al. Assortative mating and persistent reproductive isolation in hybrids. Proc Natl Acad Sci USA 2017;114(41):10936–41. https://doi.org/10.1073/pnas.1711238114.

12. Birkhead TR, Brillard JP. Reproductive isolation in birds: postcopulatory prezygotic barriers. Trends Ecol Evol 2007;22(5):266–72. https://doi.org/10.1016/j.tree.2007.02.004.

13. Wang S, Tang CC, Tao M, et al. Establishment and application of distant hybridization technology in fish. Sci China Life Sci 2019;62(1):22–45. https://doi.org/10.1007/s11427-018-9408-x.

14. Su GH, Logez M, Xu J, et al. Human impacts on global freshwater fish biodiversity. Science 2021;371(6531):835–38. https://doi.org/10.1126/science.abd3369.

15. Dias MS, Oberdorff T, Hugueny B, et al. Global imprint of historical connectivity on freshwater fish biodiversity. Ecol Lett 2014;17(9):1130–40. https://doi.org/10.1111/ele.12319.

16. Seehausen O. Hybridization and adaptive radiation. Trends Ecol Evol 2004;19(4):198–207. https://doi.org/10.1016/j.tree.2004.01.003.

17. Geiger MF, Herder F, Monaghan MT, et al. Spatial heterogeneity in the Mediterranean biodiversity hotspot affects barcoding accuracy of its freshwater fishes. Mol Ecol Resour 2014;14(6):1210–21. https://doi.org/10.1111/1755-0998.12257.

18. Costedoat C, Pech N, Salducci M-D, et al. Evolution of mosaic hybrid zone between invasive and endemic species of Cyprinidae through space and time. Biol J Linn Soc 2005;85(2):135–55. https://doi.org/10.1111/j.1095-8312.2005.00478.x.

19. Broughton RE, Vedala KC, Crowl TM, et al. Current and historical hybridization with differential introgression among three species of cyprinid fishes (genus Cyprinella). Genetica 2011;139:699–707. https://doi.org/10.1007/s10709-011-9578-9.

20. Pereira CSA, Aboim MA, Ráb P, et al. Introgressive hybridization as a promoter of genome reshuffling in natural homoploid fish hybrids (Cyprinidae, Leuciscinae). Heredity 2014;112(3):343–50. https://doi.org/10.1038/hdy.2013.110.

21. Aboim M, Mavárez J, Bernatchez L, et al. Introgressive hybridization between two Iberian endemic cyprinid fish: a comparison between two independent hybrid zones. J Evol Biol 2010;23(4):817–28. https://doi.org/10.1111/j.1420-9101.2010.01953.x.

22. Rønnestad I, Yufera M, Ueberschär B, et al. Feeding behaviour and digestive physiology in larval fish: current knowledge, and gaps and bottlenecks in research. Rev Aquacult 2013;5:S59–S98. https://doi.org/10.1111/raq.12010.

23. Kuang ZR, Li F, Duan QJ, et al. Host diet shapes functionally differentiated gut microbiomes in sympatric speciation of blind mole rats in Upper Galilee. Front Microbiol 2022;13:1062763. https://doi.org/10.3389/fmicb.2022.1062763.

24. Chen HY, Li CQ, Liu T, et al. A metagenomic study of intestinal microbial diversity in relation to feeding habits of surface and cave-dwelling sinocyclocheilus species. Microb Ecol 2020;79(2):299–311. https://doi.org/10.1007/s00248-019-01409-4.

25. Sibbing F. Food capture and oral processing. Cyprinid fishes: systematics, biology and exploitation. 1991;3:377–412. https://doi.org/10.1007/978-94-011-3092-9_13.

26. Hulsey CD, Machado-Schiaffino G, Keicher L, et al. The integrated genomic architecture and evolution of dental divergence in East African cichlid fishes (Haplochromis chilotes x H. nyererei). G3 (Bethesda) 2017;7(9):3195–202. https://doi.org/10.1534/g3.117.300083.

27. Jeon SA, Park JL, Park S-J, et al. Comparison between MGI and Illumina sequencing platforms for whole genome sequencing. Genes Genomics 2021;43:713–24. https://doi.org/10.1007/s13258-021-01096-x.

28. Chen SF, Zhou YQ, Chen YR, et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018;34(17):i884–90. https://doi.org/10.1093/bioinformatics/bty560.

29. Hu J, Wang Z, Sun Z, et al. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. Genome Biol 2024;25(1):107. https://doi.org/10.1186/s13059-024-03252-4.

30. Hu J, Fan JP, Sun ZY, et al. NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics 2020;36(7):2253–55. https://doi.org/10.1093/bioinformatics/btz891.

31. Cheng HY, Jarvis ED, Fedrigo O, et al. Haplotype-resolved assembly of diploid genomes without parental data. Nat Biotechnol 2022;40(9):1332–35. https://doi.org/10.1038/s41587-022-01261-x.

32. Rao SSP, Huntley MH, Durand NC, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 2014;159(7):1665–80. https://doi.org/10.1016/j.cell.2014.11.021.

33. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;26(5):589–95. https://doi.org/10.1093/bioinformatics/btp698.

34. Servant N, Varoquaux N, Lajoie BR, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol 2015;16:259. https://doi.org/10.1186/s13059-015-0831-x.

35. Burton JN, Adey A, Patwardhan RP, et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol 2013;31(12):1119–25. https://doi.org/10.1038/nbt.2727.

36. Stanke M, Diekhans M, Baertsch R, et al. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 2008;24(5):637–44. https://doi.org/10.1093/bioinformatics/btn013.

37. Kim D, Paggi JM, Park C, et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 2019;37(8):907–15. https://doi.org/10.1038/s41587-019-0201-4.

38. Eddy SR. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comput Biol 2008;4(5):e1000069. https://doi.org/10.1371/journal.pcbi.1000069.

39. Ou SJ, Jiang N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. Mobile DNA 2019;10:48. https://doi.org/10.1186/s13100-019-0193-0.

40. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinf 2008;9:18. https://doi.org/10.1186/1471-2105-9-18.

41. Ou SJ, Jiang N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol 2018;176(2):1410–22. https://doi.org/10.1104/pp.17.01310.

42. Chan PP, Lin BY, Mak AJ, et al. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res 2021;49(16):9077–96. https://doi.org/10.1093/nar/gkab688.

43. Armstrong J, Hickey G, Diekhans M. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature 2020;587(7833):246–51. https://doi.org/10.1038/s41586-020-2871-y.

44. Höhler D, Pfeiffer W, Ioannidis V, et al. RAxML Grove: an empirical phylogenetic tree database. Bioinformatics 2022;38(6):1741–42. https://doi.org/10.1093/bioinformatics/btab863.

45. Zhang C, Scornavacca C, Molloy EK, et al. ASTRAL-Pro: quartet-based species-tree inference despite paralogy. Mol Biol Evol 2020;37(11):3292–307. https://doi.org/10.1093/molbev/msaa139.

46. Yang ZH. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 2007;24(8):1586–91. https://doi.org/10.1093/molbev/msm088.

47. Cooper GM, Stone EA, Asimenos G, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 2005;15(7):901–13. https://doi.org/10.1101/gr.3577405.

48. Chen L, Qiu Q. Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. Science 2019;364(6446): eaav6202. https://doi.org/10.1126/science.aav6202.

49. Malinsky M, Matschiner M, Svardal H. Dsuite—fast D-statistics and related admixture evidence from VCF files. Mol Ecol Resour 2021;21(2):584–95. https://doi.org/10.1111/1755-0998.13265.

50. Rio DC, Ares M, Hannon GJ, et al. Purification of RNA using TRIzol (TRI reagent). Cold Spring Harb Protoc 2010;2010(6):pdb.prot5439. https://doi.org/10.1101/pdb.prot5439.

51. Patterson J, Carpenter EJ, Zhu Z, et al. Impact of sequencing depth and technology on de novo RNA-seq assembly. BMC Genomics 2019;20(1):604. https://doi.org/10.1186/s12864-019-5965-x.

52. Chen YX, Chen YS, Shi CM, et al. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. Gigascience 2018;7(1):1–6. https://doi.org/10.1093/gigascience/gix120.

53. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. Gigascience 2021;10(2):giab008. https://doi.org/10.1093/gigascience/giab008.

54. Srinivasan KA, Virdee SK, McArthur AG. Strandedness during cDNA synthesis, the stranded parameter in htseq-count and analysis of RNA-seq data. Brief Funct Genomics 2020;19(5–6):339–42. https://doi.org/10.1093/bfgp/elaa010.

55. Varet H, Brillet-Gueguen L, Coppee JY, et al. SARTools: a DESeq2- and EdgeR-based R pipeline for comprehensive differential analysis of RNA-seq data. PLoS One 2016;11(6):e0157022. https://doi.org/10.1371/journal.pone.0157022.

56. Zhang Z. KaKs_Calculator 3.0: calculating selective pressure on coding and non-coding sequences. Genomics Proteomics Bioinformatics 2022;20:536–40.https://doi.org/10.1016/j.gpb.2021.12.002.

57. Liu H, Chen C, Lv M, et al. A chromosome-level assembly of blunt snout bream (Megalobrama amblycephala) genome reveals an expansion of olfactory receptor genes in freshwater fish. Mol Biol Evol 2021;38(10):4238–51. https://doi.org/10.1093/molbev/msab152.

58. Jiang H, Qian Y, Zhang Z, et al. Chromosome-level genome assembly and whole-genome resequencing of topmouth culter (Culter alburnus) provide insights into the intraspecific varia-

tion of its semi-buoyant and adhesive eggs. Mol Ecol Resour 2023;23(8):1841–52. https://doi.org/10.1111/1755-0998.13845.

59. Chen F, Xue G, Wang YK, et al. Evolution of the Yangtze River and its biodiversity. Innovation 2023;4(3):100417. https://doi.org/10.1016/j.xinn.2023.100417.

60. Feng CG, Wang K, Xu WJ, et al. Monsoon boosted radiation of the endemic East Asian carps. Sci China Life Sci 2023;66(3):563–78. https://doi.org/10.1007/s11427-022-2141-1.

61. Feng SH, Bai M, Rivas-Gonzalez I, et al. Incomplete lineage sorting and phenotypic evolution in marsupials. Cell 2022;185(10):1646–60. https://doi.org/10.1016/j.cell.2022.03.034.

62. Martin SH, Davey JW, Jiggins CD. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Mol Biol Evol 2015;32(1):244–57. https://doi.org/10.1093/molbev/msu269.

63. Ren L, Li WH, Qin QB, et al. The subgenomes show asymmetric expression of alleles in hybrid lineages of Megalobrama amblycephala x Culter alburnus. Genome Res 2019;29(11):1805–15. https://doi.org/10.1101/gr.249805.119.

64. Tautz D, Domazet-Lošo T, The evolutionary origin of orphan genes. Nat Rev Genet 2011;12(10):692–702. https://doi.org/10.1038/nrg3053.

65. Chen J, Liu H, Gooneratne R, et al. Population genomics of megalobr ama pr ovides insights into e volutionary history and dietary adaptation. Biology 2022;11(2):186. https:// doi.org/ 10.3390/ biology11020186.

66. Hayden B, Palomares MLD, Smith BE, et al. Biological and environmental drivers of trophic ecology in marine fishes—a global perspective. Sci Rep 2019;9(1):11415. https://doi.org/10.1038/s41598-019-47618-2.

67. Zhou X, Khan S, Huang DB, et al. V-set and immunoglobulin domain containing (VSIG) proteins as emerging immune checkpoint targets for cancer immunotherapy. Front Immunol 2022;13:938470. https://doi.org/10.3389/fimmu.2022.938470.

68. Miao EA, Rajan JV, Aderem A. Caspase-1-induced pyroptotic cell death. Immunol Rev 2011;243(1):206–14. https://doi.org/10.1111/j.1600-065X.2011.01044.x.

69. Felthaus O, Gosau M, Morsczeck C. ZBTB16 induces osteogenic differentiation marker genes in dental follicle cells independent from RUNX2. J Periodontol 2014;85(5):e144–51. https://doi.org/10.1902/jop.2013.130445.

70. Sribudiani Y, Chauhan RK, Alves MM, et al. Identification of variants in RET and IHH pathway members in a large family with history of Hirschsprung disease. Gastroenterology 2018;155(1):118–29. https://doi.org/10.1053/j.gastro.2018.03.034.

71. Rai K, Sarkar S, Broadbent TJ, et al. DNA demethylase activity maintains intestinal cells in an undifferentiated state following loss of APC. Cell 2010;142(6):930–42. https://doi.org/10.1016/j.cell.2010.08.030.

72. Zoodsma JD, Keegan EJ, Moody GR, et al. Disruption of grin2b, an ASD-associated gene, produces social deficits in zebrafish. Mol Autism 2022;13(1):38. https://doi.org/10.1186/s13229-022-00516-3.

73. Zoodsma JD, Chan K, Bhandiwad AA, et al. A model to study NMDA receptors in early nervous system development. J Neurosci 2020;40(18):3631–45. https://doi.org/10.1523/jneurosci.3025-19.2020.

74. Li DY, Jiang XD, Gong W, et al. Tectonic uplift along the northeastern margin of the Qinghai–Tibetan Plateau: constraints from the lithofacies sequence and deposition rate of the Qaidam Basin. Tectonophysics 2022;827:229279. https://doi.org/10.1016/j.tecto.2022.229279.

75. Hren MT, Sheldon ND, Grimes ST, et al. Terrestrial cooling in Northern Europe during the eocene-oligocene transition. Proc

Natl Acad Sci USA 2013;110(19):7562–67. https://doi.org/10.1073/pnas.1210930110.

76. Xing YC, Zhang CG, Fan E, et al. Freshwater fishes of China: species richness, endemism, threatened species and conservation. Diversity Distributions 2016;22:358–70. https://doi.org/10.1111/ddi.12399.

77. Murat F, Mbengue N, Winge SB, et al. The molecular evolution of spermatogenesis across mammals. Nature 2023;613(7943):308–16. https://doi.org/10.1038/s41586-022-05547-7.

78. Brawand D, Soumillon M, Necsulea A, et al. The evolution of gene expression levels in mammalian organs. Nature 2011;478(7369):343–48. https://doi.org/10.1038/nature10532.

79. Carruthers M, Edgley DE, Saxon AD, et al. Ecological speciation promoted by divergent regulation of functional genes within African cichlid fishes. Mol Biol Evol 2022;39(11):msac251. https://doi.org/10.1093/molbev/msac251.

80. Olave M, Nater A, Kautt AF, et al. Early stages of sympatric homoploid hybrid speciation in crater lake cichlid fishes. Nat Commun 2022;13(1):5893. https://doi.org/10.1038/s41467-022-33319-4.

81. He S, Li L, Lv LY, et al. Mandarin fish (Sinipercidae) genomes provide insights into innate predatory feeding. Commun Biol 2020;3(1):361. https://doi.org/10.1038/s42003-020-1094-y.

82. Jawad LA, Agha GF, Abdullah SMA, et al. Morphology and morphometry of pharyngeal bone and teeth in cyprinid species from the Kurdistan Region. Anat Rec 2022;305(11):3356–66. https://doi.org/10.1002/ar.24906.

83. Gu QH, Yuan H, Zhong H, et al. Spatiotemporal characteristics of the pharyngeal teeth in interspecific distant hybrids of cyprinid fish: phylogeny and expression of the initiation marker genes. Front Genet 2022;13:983444. https://doi.org/10.3389/fgene.2022.983444.

84. Hartenstein V, Martinez P. Structure, development and evolution of the digestive system. Cell Tissue Res 2019;377(3):289–92. https://doi.org/10.1007/s00441-019-03102-x.

85. Xiao J, Kang XW, Xie LH, et al. The fertility of the hybrid lineage derived from female Megalobrama amblycephala x male Culter alburnus. Anim Reprod Sci 2014;151(1–2):61–70. https://doi.org/10.1016/j.anireprosci.2014.09.012.

86. Owens GL, Samuk K. Adaptive introgression during environmental change can weaken reproductive isolation. Nat Clim Chang 2020;10(1):58–62. https://doi.org/10.1038/s41558-019-0628-0.

87. Ren L. The assembled genomes and annotation files of eight cyprinid fishes. Figshare. 2023. https://doi.org/10.6084/m9.figshare.24125487.v1

88. Ren L, Tu X, Luo M, et al. Supporting data for "Genomes Reveal Pervasive Distant Hybridization in Nature among Cyprinid Fishes." GigaScience Database. 2024. https://doi.org/10.5524/102603.

89. Ren L, Tu X, Luo M, et al. Genome data of the cyprinid fish, Culter alburans (Topmouth Culter). GigaScience Database. 2024. https://doi.org/10.5524/102610.

90. Ren L, Tu X, Luo M, et al. Genome data of the cyprinid fish, Megalobrama amblycephala (blunt snout bream). GigaScience Database. 2024. https://doi.org/10.5524/102611.

91. Ren L, Tu X, Luo M, et al. Genome data of the cyprinid fish, Gobiocypris rarus (rare gudgeon). GigaScience Database. 2024. https://doi.org/10.5524/102612

92. Ren L, Tu X, Luo M, et al. Genome data of the cyprinid fish, Cirrhinus molitorella (mud carp). GigaScience Database. 2024. https://doi.org/10.5524/102613.

93. Ren L, Tu X, Luo M, et al. Genome data of the cyprinid fish, Pseudorasbora parva (stone moroko). GigaScience Database. 2024. https://doi.org/10.5524/102614.

94. Ren L, Tu X, Luo M, et al. Genome data of the cyprinid fish, xenocypris davidi (Bleekers yellow tail). GigaScience Database. 2024. https://doi.org/10.5524/102615.

95. Ren L, Tu X, Luo M, et al. Genome data of the cyprinid fish, Elopichthys bambusa (yellow cheek carp). GigaScience Database. 2024. https://doi.org/10.5524/102616.

96. Ren L, Tu X, Luo M, et al. Genome data of the cyprinid fish, Ctenopharyngodon idella (grass carp). GigaScience Database. 2024. https://doi.org/10.5524/102617.