Nucleic Acids Research, 2026, **54**, D1108–D1118
https://doi.org/10.1093/nar/gkaf1026
Advance access publication date: 14 October 2025
Database issue

iFish: a comprehensive multi-omics database for fish genetics, transcriptional regulation, and epigenetic research

Ruoyan Liu[1,†], Qin Tang[2,†], Peng Sun[2], Weiwei Jin[1], Yanbo Yang[1], Xiaohong Wu[1], Wen Cao[1], Xiaohui Niu[1], Jing Gong[1,3,*], Ze-Xia Gao[1,2,*]

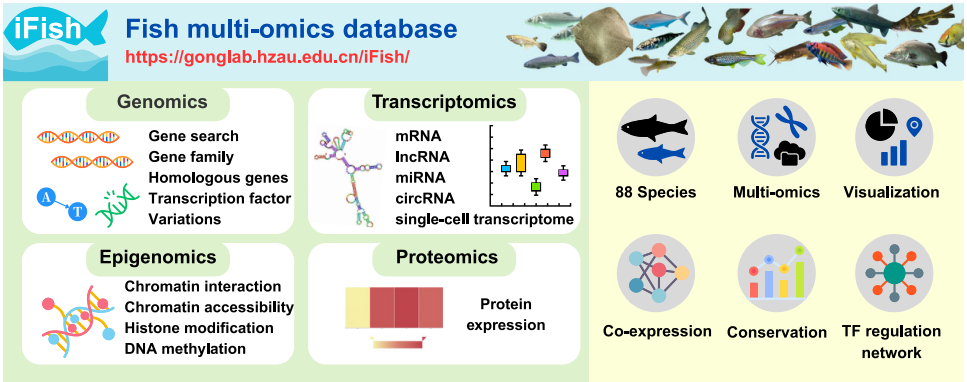
[1]Hubei Hongshan Laboratory, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China
[2]College of Fisheries, Huazhong Agricultural University, Wuhan 430070, China
[3]College of Biomedicine and Health, Huazhong Agricultural University, Wuhan 430070, China

[*]To whom correspondence should be addressed. Email: gong.jing@mail.hzau.edu.cn
Correspondence may also be addressed to Ze-Xia Gao. Email: gaozx@mail.hzau.edu.cn
[†]The first two authors should be regarded as Joint First Authors.

Abstract

The explosive expansion of fish multi-omics data is reshaping basic and applied research in fisheries and aquaculture. Here, we introduce iFish, a rigorously curated and comprehensive resource that integrates genome, transcriptome, epigenome, and proteome datasets from 88 fish species. From 884 whole-genome sequencing datasets, we identified 137.04 million single-nucleotide polymorphisms and 45.52 million insertions/deletions. Leveraging 9797 bulk RNA-seq, 293 single-cell RNA-seq, and 840 microRNA (miRNA)-seq datasets, we quantified 1.87 million messenger RNAs, 287 873 long noncoding RNAs, 197 554 circular RNAs, and 6068 mature miRNAs across different tissues of multiple species. By integrating 1563 epigenetics sequencing datasets, we provided genome-wide maps of histone modifications, chromatin accessibility, DNA methylation, and three-dimensional chromatin interactions. Furthermore, 50 proteomics projects were included for protein expression profiling. In addition, to facilitate gene analyses, we also curated 2.7 million gene annotation entries, identified homologous genes, annotated 192 107 transcription factors (TFs), constructed TF regulatory networks and gene co-expression networks, and performed cross-species conservation analyses of noncoding RNAs. Finally, we established the iFish database (https://gonglab.hzau.edu.cn/iFish/), a user-friendly platform that supports interactive browsing, visualization, and download. With comprehensive data and useful tools, iFish will advance genetic, transcriptional-regulatory, and epigenetic investigations, and accelerate genetic improvement and practical applications in aquaculture.

Graphical abstract



Introduction

Fish represent the most speciose group of vertebrates and play pivotal roles in global nutrition, economic development, ecological balance, and biodiversity conservation [1]. Driven by the increasing demand for high-quality protein sources, the global fisheries and aquaculture industries are experiencing rapid expansion. According to the State of World Fisheries and Aquaculture (FAO, 2024) [2], the global fish trade exceeds $160 billion annually, and the total production soared to 223.2 million tons in 2022, an increase of 33.7% compared with that in 2010, and a 4.4% rise compared with that in 2020. With the global population increasing, the demand for

Received: August 4, 2025. **Revised:** September 3, 2025. **Accepted:** September 19, 2025

protein from aquaculture will continue to show an upward trend. In addition to nutrition, fish contribute significantly to aquatic ecosystems through nutrient cycling, water purification, and population control of other species [3]. Moreover, fish biodiversity is essential for maintaining ecosystem balance and stability. Declines in fish diversity often signal pollution, habitat degradation, overfishing, and climate change [4].

To satisfy the growing demand for high-protein foods while also sustaining fish stocks and ecosystem health, contemporary aquaculture research is focusing on several key areas. These include, but are not limited to: optimizing feed efficiency and fish nutrition, improving genetic breeding and reproductive performance, regulating growth and development, unravelling the mechanisms behind disease resistance and stress tolerance, understanding sex determination and the genetic basis of sexual dimorphism, and addressing fundamental questions regarding fish origins, evolution, taxonomy, and environmental adaptation [5–10]. Tackling these challenges requires foundational insights into genomic architecture, gene regulatory networks, and epigenetic mechanisms, all of which can be advanced through the use of integrative multi-omics data. For example, Hi-C combined with RNA-seq has been widely applied in the high-quality genome assembly and sex chromosome evolution of fish, including *Siluriformes* species [10]. Huang *et al*. performed whole-genome sequencing (WGS) of yellow catfish (*Pelteobagrus fulvidraco*), conducted genome-wide association study (GWAS) combined with growth traits, identified single nucleotide polymorphism (SNP) markers related to body length and weight, and verified gene expression through RNA-seq, revealing the genetic basis of growth traits and laying a foundation for molecular marker-assisted breeding [5]. Xiong *et al*. conducted a combined analysis of whole-genome bisulfite sequencing (WGBS) and RNA-seq on the X and Y sperm of yellow catfish and found that the DNA methylation level of Y sperm was lower than that of X sperm, with a greater proportion of upregulated genes in Y sperm [11]. Through liver RNA-seq and Assay for Transposase-Accessible Chromatin with high throughput sequencing (ATAC-seq) analyses of channel catfish (*Ictalurus punctatus*) infected with *Aeromonas hydrophila*, researchers reported that cortisol can regulate the immune response at the early stage [12].

Currently, several fish-related databases have been developed, providing abundant resources and specialized information for fish research. For example, FishSED is a spatial expression database dedicated to zebrafish (*Danio rerio*), which has curated 56 spatial expression profile datasets from 10 research projects [13]. FishGET curates 1362 RNA-seq datasets of eight fish species [14]. FishCODE is an integrated database that incorporates genomic, transcriptomic, and methylome data for 35 fish species [15]. FishSCT compiles 129 single-cell transcriptome datasets from 44 projects across nine species [16]. FishmiRNA specializes in microRNA (miRNA) annotation and expression profiles of ray-finned fishes [17]. FishTEDB catalogues transposable elements (TEs) detected in the complete genomes of 63 fish species, facilitating studies on TE evolution and function [18]. HRGFish contains 818 hypoxia-responsive gene sequences representing 35 gene types from 38 fish species and provides tools for sequence analysis and primer design [19]. However, these databases are not sufficiently comprehensive in terms of fish species and multi-omics integration. The discovery of genetic variation, the comprehensive identification and quantification of noncoding RNAs (ncRNAs), and the systematic investigation of epigenomic datasets remain insufficient for most species. Moreover, other useful information, such as homologous genes, transcription factor (TF) annotation, TF binding genes, and ncRNA co-expressed genes, is largely absent from existing fish databases.

To address these limitations, we developed the iFish database (https://gonglab.hzau.edu.cn/iFish/), currently the most comprehensive multi-omics database for fish. It integrates data from diverse sources across genomics, transcriptomics, epigenetics, and proteomics for 88 fish species, encompassing both model species (e.g. zebrafish) and economically significant species such as the Atlantic salmon (*Salmo salar*), Atlantic cod (*Gadus morhua*), rainbow trout (*Oncorhynchus mykiss*), gilthead seabream (*Sparus aurata*), grass carp (*Ctenopharyngodon idella*), largemouth bass (*Micropterus salmoides*), and yellow catfish. Through unified processing and analysis of multi-omics data, iFish supports queries of diverse datasets, including genetic variations, expression profiles, and epigenetic profiles, and offers a range of visualization tools. By providing extensive data and tools, iFish aims to serve as a comprehensive and user-friendly platform for researchers to explore and analyze fish multi-omics data, thereby facilitating the discovery of new insights into fish biology.

Materials and methods

Genomic data collection and processing

To encompass a wide range of fish species, we comprehensively retrieved genome assembly datasets from the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/) [20] and Ensembl (https://www.ensembl.org/) [21] and obtained genomic data of 76 chromosome-level assemblies and 12 high-quality scaffold-level assemblies across 88 fish species (Fig. 1). The software and analytical parameters used in this study are listed in Supplementary Table S1.

Next, we performed standardized genomic functional annotation for each species. Homologous genes across fish species were identified using a sequence similarity-based approach [22]. Specifically, zebrafish protein sequences were used as queries in BLASTP [23] searches against the gene sequences of other fish species, with strict thresholds set as *E*-value $<1e-5$ and sequence identity $>50\%$ [22]. Reciprocal BLASTP searches were performed to confirm homology, retaining only the best-aligned sequence for each gene. Additionally, Superfamily is a structural classification database that groups evolutionary related proteins into super families based on shared ancestry and conserved functional domains [24]. We annotated protein-coding genes from other fish species to super families on the basis of sequence similarity, using zebrafish super families as references.

TFs across fish species were predicted via a comparative approach that uses zebrafish as the reference. We first constructed the hidden Markov model (HMM) profiles of DNA-binding domains (DBDs) from known zebrafish TF families using HMMER [25, 26]. These HMM profiles were subsequently used to scan the genomic sequences of other fish species via hmmsearch, with *E*-value thresholds optimized for each TF family based on zebrafish reference annotations. Reciprocal BLASTP searches against zebrafish TF sequences (*E*-value $<1e-5$; identity $>50\%$) were performed to validate

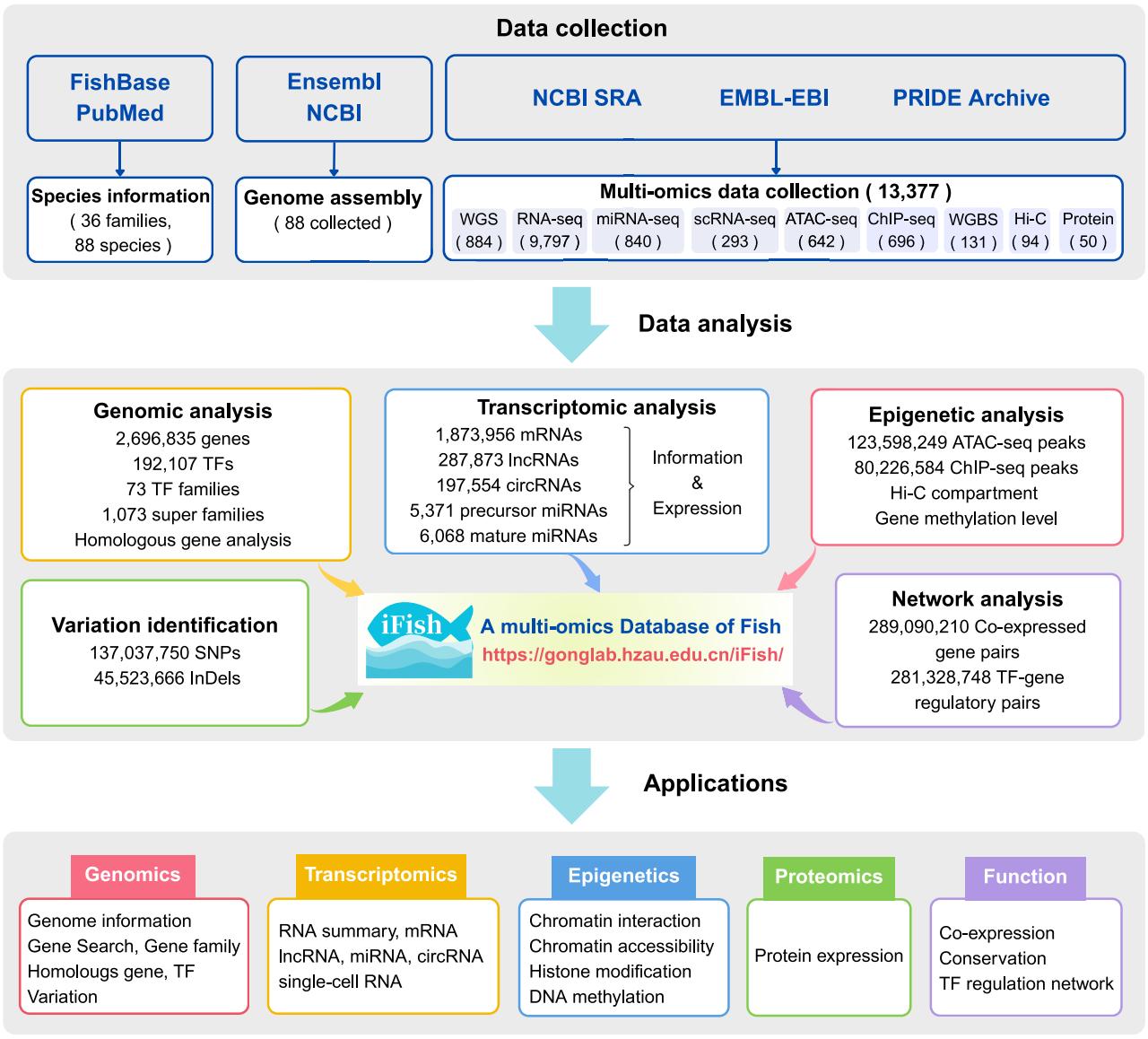


Figure 1. Overview of the construction of the iFish database.

putative TFs, retaining only high-confidence matches. Manual curation was further applied to remove non-TF proteins with homologous DBDs but unrelated functions, ensuring the reliability of the predicted TF repertoires across fish genomes.

To identify genetic variations, we further collected 884 WGS datasets of 23 species from NCBI SRA database. High-quality SNPs and insertions/deletions (InDels) were identified via a standardized pipeline for Illumina paired-end data. Raw FASTQ reads were quality-assessed with FastQC and then trimmed using fastp [27] to remove adapters and low-quality bases. Cleaned reads were aligned to the reference genome with BWA–MEM [28]. Sequence Alignment/Map (SAM) files were converted to sorted Binary Alignment/Map (BAM) files via SAMtools [29], with polymerase chain reaction (PCR) duplicates removed using the Genome Analysis Toolkit (GATK) MarkDuplicates [30]. Variant calling followed GATK best practices: HaplotypeCaller generated per-sample genomic Variant Call Format (gVCF) files, which were combined via CombineGVCFs and jointly genotyped with GenotypeGVCFs [31]. As an optional adaptive step, base quality score recalibration was performed against known variants to improve accuracy. Raw variants were filtered with GATK VariantFiltration and VCFtools. All retained variants were functionally annotated using SnpEff with a custom reference database, categorizing impacts on coding sequences [32].

Transcriptomic data collection and processing

We gathered 9797 bulk RNA-seq datasets of 32 species, applied to the identification and quantification of messenger RNA (mRNA), long noncoding RNA (lncRNA), and circular RNA (circRNA), as well as co-expression network construction. A total of 840 miRNA-seq datasets of 14 species were also gathered for profiling of precursor and mature miRNA. Additionally, we collected 293 single-cell RNA sequencing (scRNA-seq) datasets to analyze cell-type-specific

Table 1. Results of identified RNAs for 11 species with both miRNA-seq and RNA-seq data

Species	Genome version	No. of RNA-seq samples	No. of miRNA-seq samples	No. of tissues	No. of mRNAs	No. of lncRNAs	No. of circRNAs	No. of pre-miRNAs	No. of mature miRNAs
Danio rerio	GRCz11	4298	72	31	50 808	24 207	68 194	381	450
Gasterosteus aculeatus	BROADS1	305	64	17	24 438	10 604	8250	638	762
Haplochromis burtoni	AstBur1.0	24	6	1	35 585	1201	149	225	229
Lepisosteus oculatus	LepOcu1	11	16	15	21 254	6550	468	231	254
Maylandia zebra	M_zebra_UMD2a	24	6	1	31 870	1103	106	264	274
Nothobranchius furzeri	Nfu_20140520	253	193	9	48 737	4879	24 979	419	503
Oncorhynchus mykiss	USDA_OmykA_1.1	109	5	12	118 951	8368	3737	443	447
Oryzias melastigma	Om_v0.7.RACA	17	12	6	36 313	5854	635	306	329
Pundamilia nyererei	PunNye1.0	12	6	4	32 126	2925	140	234	240
Salmo salar	Ssal_v3.1	2698	221	21	139 693	51 738	23 893	1133	1452
Tachysurus fulvidraco	HZAU_PFXX_2.0	48	3	4	66 639	7083	1397	307	316

gene expression profiles. All these sequencing datasets were sourced from NCBI SRA database.

Following the same quality control pipeline used for the previously described FASTQ files, clean RNA-seq data were aligned to the reference genome using HISAT2 [33]. Transcripts were identified with StringTie [34] based on reference annotation files, followed by transcript assembly using TACO [35]. The assembled transcripts were compared with known annotations, and gffcompare was used to evaluate and compare annotation files, generating transcript assembly metrics to ensure data accuracy and completeness [34].

The identification of lncRNAs involved multiple filtering steps. First, transcripts with classcode "u" were retained, followed by selection of those containing >1 exon and exceeding 200 bp in length. Coding potential was predicted using CNCI [36] and CPC2 [37], retaining only transcripts that were jointly annotated as "noncoding." These sequences were translated into protein sequences and aligned against the Pfam database via pfam_scan (*E*-value $\leq 1e-5$) to exclude those with known protein domains [38]. Low-expression transcripts were finally filtered out.

CircRNAs were identified by find_circ [39]. Unmapped reads were filtered, and anchor sequences were extracted and re-aligned with Bowtie2 [40]. CircRNAs were detected via find_circ.py, and results from individual samples were merged. Merged sequences underwent strict filtering to retain those labeled "CIRCULAR," non-mitochondrial, with ≥2 unique junction reads, clear breakpoints, and those shorter than 100 kb [39]. Low-expression transcripts were further excluded.

The identification of miRNAs from miRNA-seq data was performed with miRDeep2 [41]. Cleaned data were aligned using mapper.pl (adapter sequence "TGGAATTC" and a minimum length threshold of 18 bp). Known and novel miRNAs were identified via miRDeep2.pl, with zebrafish miRNA sequences from miRBase used as references [42]. Low-expression transcripts were ultimately removed.

The expression levels of lncRNAs and mRNAs were calculated using featureCounts [43], whereas those of circRNAs were quantified using find_circ.py [39]. For miRNAs, expression levels were determined using quantizer.pl [41]. Transcripts per million values for all RNAs were computed to facilitate cross-sample comparisons. A summary of RNA-related results for the 11 species with both RNA-seq and miRNA-seq data is provided in Table 1.

For scRNA-seq data, quantification analysis was performed using CellRanger [44]. After quality control, cells were clustered and gene-level expression was calculated using SCANPY [45].

Epigenomic data collection and processing

We incorporated 696 ChIP-seq, 642 ATAC-seq, 131 WGBS, and 94 Hi-C datasets to profile histone modification peaks, open chromatin regions, DNA methylation levels, and chromatin compartment types and values, respectively.

For both ChIP-seq and ATAC-seq data, adapters and low-quality reads were removed using the same trimming pipeline applied to RNA-seq data. Clean reads were aligned to the genome assemblies using BWA with default parameters [28]. PCR duplicates were removed using Sambamba [46], followed by sorting and conversion to BAM file format using SAMtools [29]. Open chromatin regions and protein-bound regions were identified using the callpeak module of MACS2 based on aligned reads distribution and statistical models [47].

For WGBS data, clean reads were aligned to the reference genome using Bismark with duplicate removal [48]. The information of methylation sites was extracted from the alignment results, and the methylation levels of each gene were calculated as the number of methylated reads divided by the total reads at CpG sites within the gene.

For Hi-C sequencing data, clean reads were aligned to the reference genome using Bowtie2 [40], after which valid Hi-C pairs were filtered using HiC-Pro to construct the genome-wide interaction matrix [49]. The raw matrix was then normalized via an iterative correction method to eliminate technical biases [50]. A/B compartments were called using HOMER by classifying genomic bins based on the sign of the first eigenvector [51].

Proteomics data collection and processing

In proteomics study, 50 projects were retrieved from the PRIDE database [52]. Peptide-level signals in acquired datasets had been corrected using the label-free quantification (LFQ) normalization workflow, generating LFQ intensity values at the protein level. These processes enabled the compilation of intensity matrices at both the gene and protein levels.

Functional analysis

For co-expression analysis, the gene expression matrix was normalized using quantile normalization across samples. Genes with low expression levels were subsequently filtered

out to retain only those with robust expression signals, and pairwise Pearson correlation coefficients were computed between remaining genes. Results with Pearson correlation coefficients $\geq$0.8 and false-discovery rate (FDR) <0.05 were considered significant.

Conservation of RNAs was evaluated by sequence similarity using BLASTN. For each RNA in the focal species, sequence similarity to all RNAs in other species was calculated, and hits with sequence similarity identity $\geq$0.5 and *E*-value $\leq$1e − 5 were considered significant.

To construct TF regulatory networks, motifs within the 2 kb region upstream of the transcription start site of each gene were identified using FIMO [53]. These motifs were subsequently mapped to their corresponding TFs via the JASPAR database [54]. Promoter-proximal motifs were linked to their target genes using gene annotation files, thereby establishing TF–gene regulatory relationships.

Database implementation

All processed data in the iFish were stored in MongoDB (version 3.6.8). The iFish website was built with Flask (version 1.1.1) framework and AngularJS (version 1.6.1) and hosted on the Apache 2 webserver (version 2.4.18). iFish is freely available and does not require registration or login for access. Users can access it via https://gonglab.hzau.edu.cn/iFish/.

Database content and usage

Construction and overview of iFish

To provide comprehensive multi-omics resources for fish research, iFish was constructed by integrating multi-omics data from 88 fish species, encompassing 13 377 high-throughput sequencing samples with a total data volume exceeding 40 TB. iFish comprises nine core modules: species, genomics, transcriptomics, epigenetics, proteomics, function, JBrowse, download, and document. Genomics module includes 88 genome assembly datasets, 884 WGS datasets, 137 037 750 SNPs, 45 523 666 InDels, and annotations for protein-coding genes (including TFs and homologous genes). Transcriptomics module covers 9797 RNA-seq and 840 miRNA-seq libraries, with identified and quantified mRNAs, lncRNAs, circRNAs, and miRNAs. Single-cell transcriptomics module features 293 scRNA-seq datasets for cell-specific expression analysis. Epigenetics module incorporates ChIP-seq, ATAC-seq, WGBS, and Hi-C datasets for methylation, chromatin accessibility, and related analyses. Proteomics module includes 50 projects with protein expression profiles. iFish also offers functional tools for network analysis, cross-species RNA conservation analysis, and interactive visualization via JBrowse, supporting comparative and functional genomics studies in fish.

Species module

The species module compiles basic information of 88 fish species from FishBase (https://www.fishbase.se/) [55], including categorization, habitat, and biology (Fig. 2A). In the "Omics Data Portal" option bar, users can select functional modules to jump to the corresponding pages.

Genomics module

The genomics module is organized into six sections: genome information, gene search, gene family, homologous genes, TF, and genome variations (Fig. 2B). The genome information page provides data sources and genome versions for 88 fish genomes included in the database. In the gene search section, users can select a species and query by gene ID or symbol. The query results provide comprehensive annotations, including gene location information, visual displays of genes in JBrowse, gene expression in different tissues, and visualizations of TF–gene networks. The gene family module presents statistical information on 1073 super families across 88 fish species and supports retrieval of gene lists for any chosen super family within each species. Homologous gene section enables rapid identification of homologous genes using gene IDs. In the TF module, we have curated known TFs from 27 fish species and identified novel TFs for 61 fish species, with annotations to their corresponding TF families. Users can select specific TF families to access TF information for particular species.

To investigate genetic variations in fish species, we selected 884 high-quality WGS datasets for variation calling, resulting in a total of 137 037 750 SNPs and 45 523 666 InDels identified across 23 species. Users can input a gene name to retrieve all variations within the gene. Clicking the "details" button in the annotation column enables users to view detailed annotation information for each variation. Additionally, newly identified SNPs have been standardized with unique identifiers, enabling direct searches via SNP IDs.

Transcriptomics module

The transcriptomics module integrates 9797 RNA-seq datasets from 32 species, 840 miRNA-seq datasets from 14 species, and 293 scRNA-seq datasets from 6 species. In total, 1 873 956 mRNAs, 287 873 lncRNAs, 197 554 circRNAs, 5371 precursor miRNAs, and 6068 mature miRNAs were identified and quantified. This module is organized into six sections: RNA summary, mRNA, lncRNA, circRNA, miRNA, and scRNA. The RNA summary section presents bar charts illustrating statistics for each RNA type across species, as well as tables containing information such as reference genome version, sample size, and tissue count. Each of the mRNA, lncRNA, circRNA, and miRNA modules provides both information and expression sections for RNAs (Fig. 2C). The information section offers the specific location, sequence, and neighboring genes of the RNA. The expression section offers the expression levels of specific genes across different tissues, with a boxplot intuitively showing the differential expression between tissues. In the scRNA module, users can select species and biological projects. The results show a cell cluster map and the expression of the gene across different samples and cell types (Fig. 2D).

Epigenomics module

The epigenomics module contains four subpages: chromatin accessibility, histone modification, chromatin interactions, and DNA methylation. Based on 131 WGBS datasets, gene methylation levels were obtained and integrated into the DNA methylation subpage (Fig. 2E). Based on 642 ATAC-seq and 696 ChIP-seq datasets, 123 598 249 ATAC-seq peaks and 80 226 584 ChIP-seq peaks were obtained, respectively, and integrated into the chromatin accessibility and histone modification subpages (Fig. 2F). Based on 94 Hi-C datasets, chromatin compartments were obtained and integrated into the chromatin interaction module. These integrated datasets provide comprehensive epigenetic information, supporting the

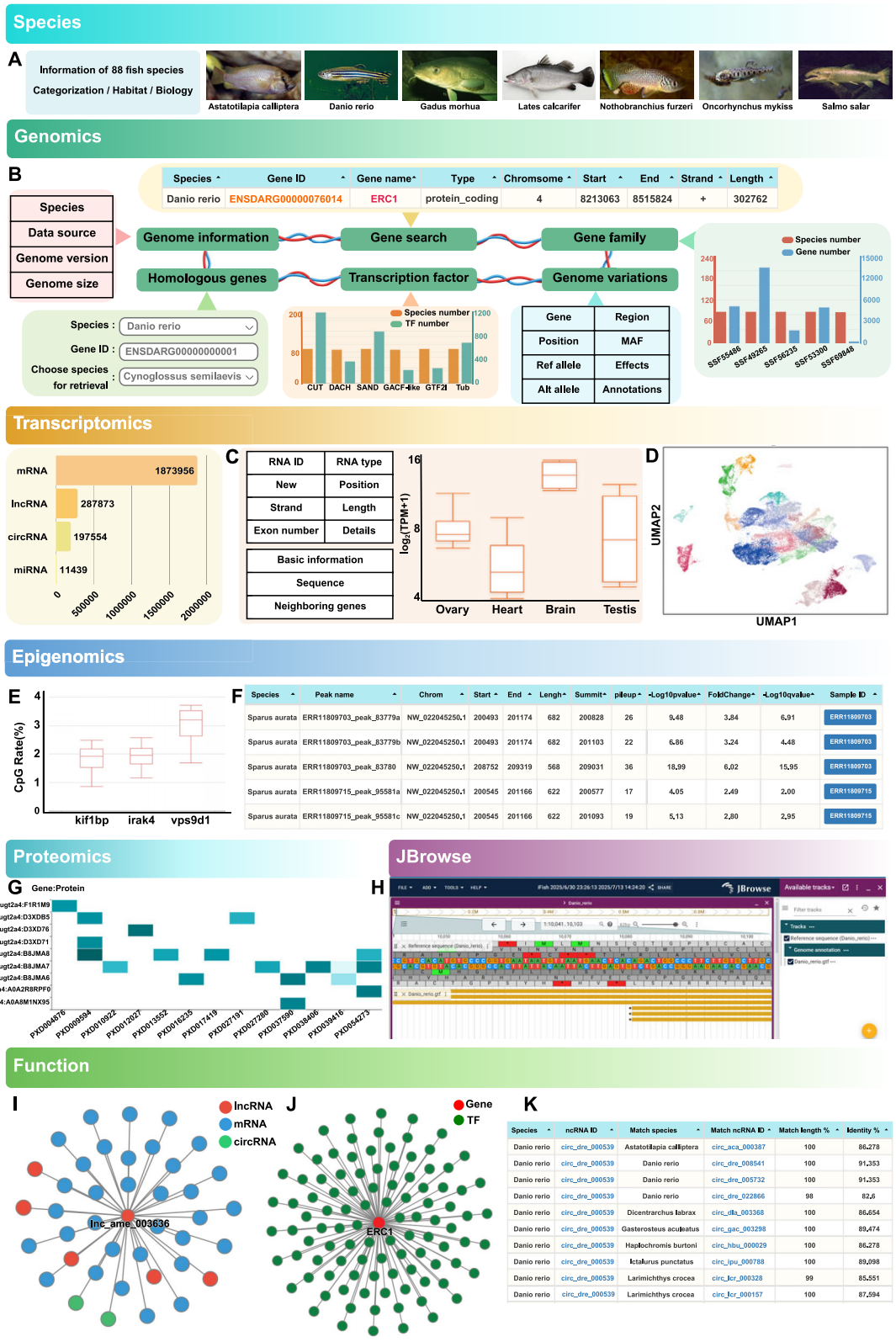


Figure 2. Key features and data in iFish. The iFish database encompasses seven key modules: species, genomics, transcriptomics, epigenomics, proteomics, function, and JBrowse. (**A**) Information displayed in the species module. (**B**) The genomics module provides six sections: genome information, gene search, gene family, homologous genes, TF, and genome variations. (**C**) The "mRNA," "lncRNA," "miRNA," and "circRNA" modules provide information on the RNAs of interest and display gene expression levels through data sheets and box plots. (**D**) Cell clustering information in "single-cell RNA" module. (**E**) On the "DNA methylation" page, the CpG rate of genes is displayed via box plots. (**F**) The "Chromatin accessibility" page presents information on ATAC peaks. (**G**) In proteomics module, the expression levels of different proteins of genes are shown through heatmaps. (**H**) JBrowse module is used for visualizing genes and regions of interest. (**I**) Gene co-expression network diagram. The central circle represents the queried gene, which is connected to related genes. (**J**) TF–gene regulation network. The central circle is the target gene, and the connected ones are the associated TFs. (**K**) The conservation module displays cross-species RNA sequence similarity.

exploration of regulatory elements and their mechanisms in fish genomes.

Proteomics module

In the proteomics module, users can input a gene symbol or protein ID to query protein expression across different projects, with a heatmap provided to facilitate a more intuitive understanding of expression differences (Fig. 2G).

Function module

The function module includes co-expression, conservation, and TF regulation network sections. The co-expression section is constructed using 289 090 210 co-expression pairs from 19 species, allowing users to input a gene to generate a gene co-expression network (Fig. 2I). The TF regulation network covers 46 species and 281 328 748 TF–gene regulatory pairs. This network enables users to explore potential transcriptional regulatory relationships between TFs and target genes (Fig. 2J). The conservation section contains BLAST results for different types of RNAs. Users can browse RNAs with sequence similarity across multiple species, with the percentage of matched length and identity percentage provided in the results (Fig. 2K).

JBrowse and download

The JBrowse tool integrates 88 fish genomes and gene annotations, enabling users to intuitively explore genomic sequences and gene information within specific regions (Fig. 2H). To enhance users' ability to conduct custom analysis and facilitate data mining, all queried tables in iFish and all images generated online can be downloaded. Additionally, in the download module, we provide downloadable gene expression files for each species.

Case study illustrating the utility of iFish

To illustrate the capabilities of iFish, we present a case by exploring sexual dimorphism and sex-determining genes in zebrafish. Sexual dimorphism of growth is widespread among fish and the cultivation of all-female or all-male populations targeting valuable economic traits holds significant importance for aquaculture. With the extensive application of functional genomic and epigenetic technologies, sex-determining genes in farmed fish have been successively identified, such as *gsdf*, *dmy*, *sox3*, *amhr2*, and *dmrt1* [56]. Among these genes, *dmrt1* (*dsx* and *mab-3* related transcription factor 1) has emerged as a significant player, which is widely recognized as a conserved and critical gene involved in sex determination and sexual differentiation across diverse fish species [56–58]. It acts as a key regulator of male sex development by promoting testicular differentiation, maintaining male germ cell identity, and suppressing female-specific pathways. For instance, in zebrafish and medaka (*Oryzias latipes*), *dmrt1* is essential for male gonad formation [57, 59]. Using multi-omics modules, we are able to explore *dmrt1* in iFish. First, users can enter "dmrt1" into the search box on the gene search page. After the search button is clicked, the gene information is displayed (Fig. 3A). Clicking the orange-highlighted gene ID will automatically navigate to the multi-modal information page of this gene, which integrates modules including basic gene information, genome JBrowse (Fig. 3B), gene expression (Fig. 3C), and TF–gene network (Fig. 3D). As expected, we observe high expression of *dmrt1* in testicular tissue from our database, consistent with previous reports of its testis-specific expression in fish (Fig. 3C). Furthermore, users can also query genes co-expressed with *dmrt1* in testis tissue of zebrafish on the co-expression page (Fig. 3E).

Discussion and future perspectives

With the advancement of sequencing technologies and extensive efforts among scientists, an increasing volume of fish omics data has been released, providing useful resources for fish research. However, multi-omics datasets of many fish species remain fragmented across various projects or publications and have not been effectively integrated, hampering functional and applied research in fish. To make full use of multi-omics data, we developed the iFish database, which integrates data from diverse sources across genomics, transcriptomics, epigenetics, and proteomics for 88 fish species spanning both model species and economically significant species. The sample number of iFish exceeds 10 000 and the processed data volume exceeds 40 TB. iFish not only encompasses information on genomic landscape, gene expression profiles, epigenetic modifications, and protein profiles, but also includes the identification of variation, TFs, and ncRNAs. It represents the most comprehensive platform that systematically integrates, analyses, and stores multi-omics data for fish to date, and will serve as an important resource for genetic, transcriptional regulation and epigenetic studies.

A distinguishing strength of iFish is its rich repository of ncRNA information. In recent years, ncRNAs have become a research hotspot due to their important roles in a wide range of biological processes. In fish, an increasing number of ncRNAs have been identified with pivotal roles in development and other functions. For example, an lncRNA named AANCR promotes antiviral responses by inhibiting miR-210 [60], and miR-722 regulates C5aR1 to modulate immune responses in half-smooth tongue sole (*Cynoglossus semilaevis*) to prevent bacterial diseases [61]. Given the important roles of ncRNAs, several ncRNA-related databases, such as Lnc2Cancer [62] and NONCODE [63], have been developed for humans and model animals. However, the expression profiles, regulatory networks, and potential functions of ncRNAs in fish remain poorly characterized, and a comprehensive platform for fish ncRNAs is still lacking. In iFish, we systematically identified and characterized the expression landscape of various ncRNAs (lncRNAs, circRNAs, pre-miRNAs, and mature miRNAs) in fish. Additionally, we examined co-expressed genes of ncRNAs and investigated the conservation of ncRNAs across multiple species. Our database provides strong support for fish ncRNA research.

Beyond ncRNAs, iFish also exhibits several advantages over existing data resources [13–19]. First, it encompasses a large number of fish species, providing comprehensive annotations spanning species metadata, gene families, homologous genes, and TFs within a single platform. Second, it provides extensive omics-derived results and downstream analysis data. At the genomic level, it includes genetic variants along with detailed annotations such as genomic coordinates, minor allele frequencies, and functional consequences, which will accelerate both genetic and evolutionary investigations. At the transcriptomics level, it curates not only bulk RNA-seq but also scRNA-seq and miRNA-seq datasets, coupled with systematic identification and quantification of lncRNAs and

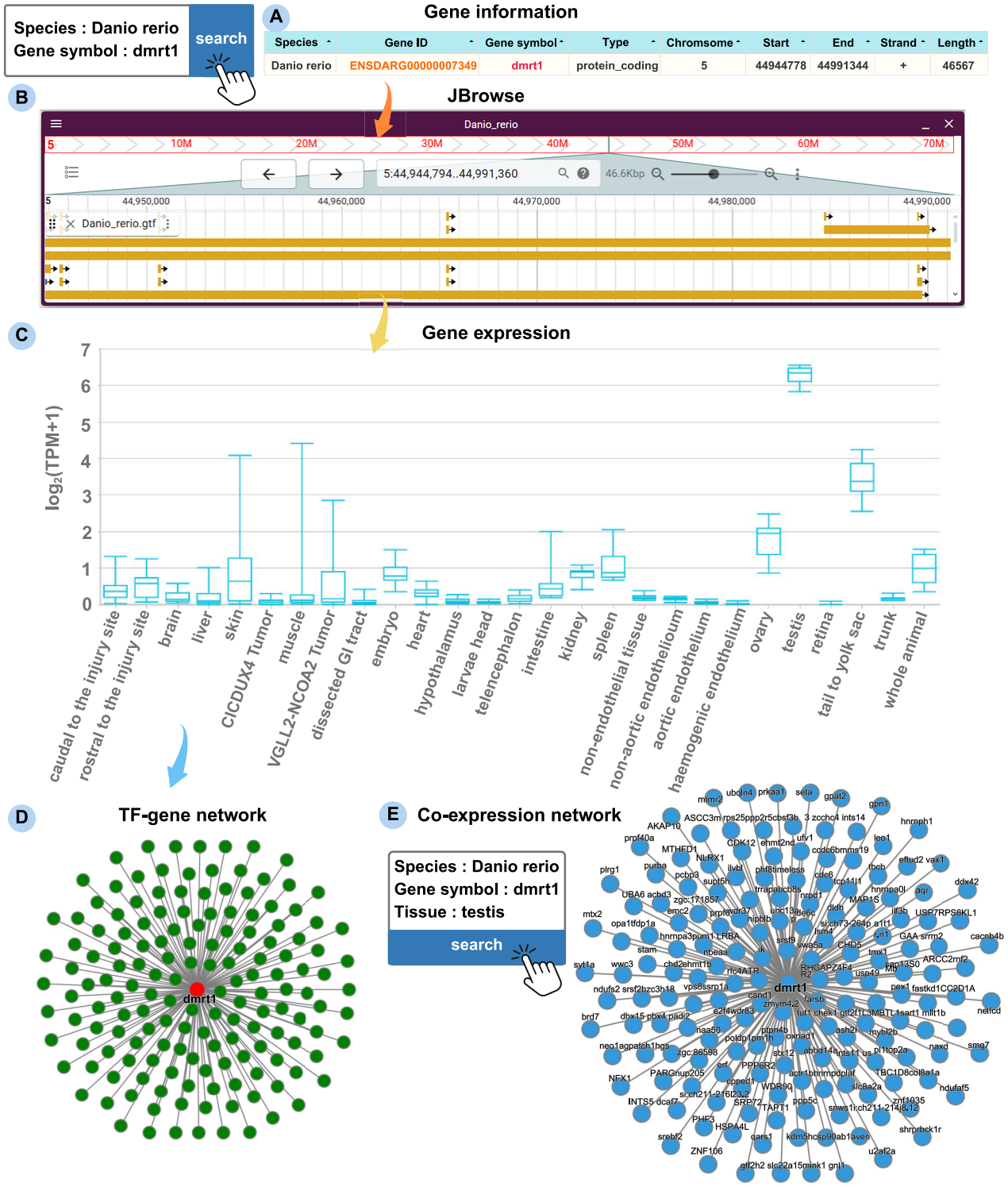


Figure 3. Case study illustrating the utility of iFish. (**A**) Gene information for *dmrt1* in *D. rerio* retrieved via the "Gene Search" module. (**B**) Genome Browse interface of *dmrt1*. (**C**) Interactive expression profile of *dmrt1* across different tissues. (**D**) Interface of the TF–*dmrt1* regulatory network. (**E**) Gene co-expression network of *dmrt1* in testis tissue of *D. rerio*.

circRNAs. At the epigenomic level, it encompasses WGBS DNA-methylation profiles, ChIP-seq histone modifications and TF binding, ATAC-seq chromatin accessibility maps, and Hi-C chromatin interaction matrices. Third, the platform offers multiple functionalities, enabling users to readily obtain TF regulatory networks and gene co-expression networks, as well as the cross-species conservation of ncRNAs.

Although the current iFish database has already integrated extensive multi-omics data, we recognize several avenues for continuous improvement. In the future, with the ongoing development of sequencing technologies, more omics data are expected to be generated, and we will continue updating our database to include more species and omics types. In particular, leveraging the development of third-generation long-read and Hi-C sequencing technologies, we will incorporate emerging telomere-to-telomere assemblies and high-quality pan genomes for relevant species to generate a pan-genome inversion index alongside refined and comprehensive annotations for genes and variants. At the current version, samples with paired omics data remain scarce, which limits the integration analyses across multiple omics types. For instance, expression quantitative trait locus analysis, which integrates genetic variants and gene expression profiles, is a powerful method for deciphering the function of genetic variation [64]. Therefore, we will prioritize the curation of jointly profiled genomes, transcriptomes, and epigenomes to enable powerful downstream applications. Additionally, rapid developments in artificial intelligence have resulted in unprecedented power for biological functional interpretation [65]. We will integrate AI-based multi-omics integration workflows and deploy large language model-assisted query engines to make it easier for users to obtain more knowledge.

In summary, we will continue to maintain and update the data within iFish, optimize its web interface, and promote the application of multi-omics data to advance fish research. The iFish database will undergo regular updates as new data becomes available. For long-term maintenance, our bioinformatics team will perform regular data backups, monitor the database's operational status in real time to prevent data loss caused by hardware failures or cyber threats, and update systems and software to optimize overall performance. We also welcome community members to submit data by contacting our team. All submissions will undergo rigorous quality assessment and standardization reviews prior to integration into iFish, and contributors will be appropriately acknowledged.

Acknowledgements

Author contributions: Ruoyan Liu (Data curation [lead], Formal analysis [lead], Investigation [equal], Methodology [equal], Writing—original draft [equal], Writing—review & editing [equal]), Qin Tang (Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Writing—original draft [equal], Writing—review & editing [equal]), Peng Sun (Data curation [equal], Formal analysis [equal], Methodology [equal], Visualization [equal]), Weiwei Jin (Investigation [equal], Methodology [equal]), Yanbo Yang (Methodology [equal], Visualization [equal]), Xiaohong Wu (Investigation [equal], Methodology [equal]), Wen Cao (Visualization [equal]), Xiaohui Niu (Investigation [equal]), Jing Gong (Conceptualization [equal], Funding acquisition [equal], Project administration [equal], Supervision [lead], Writing—review & editing [equal]), and Ze-Xia Gao (Conceptualization [equal], Funding acquisition [equal], Project administration [equal], Supervision [equal], Writing—review & editing [equal]).

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

Fundamental Research Funds for National Natural Science Foundation of China (32425054), the Fundamental Research Funds for the Central Universities (2662024XXPY002 and 2662025SCPY004), the Huazhong Agricultural University Scientific & Technological Self-innovation Foundation (11041810351), the earmarked fund for Agriculture Research System of China (CARS-45-01), and Hubei Province Significant Scientific Research (2023BBA001). Funding to pay the Open Access publication charges for this article was provided by National Natural Science Foundation of China (32425054).

Data availability

iFish is freely available to the public without registration or login requirements (https://gonglab.hzau.edu.cn/iFish/).

References

1. Martinez AS, Willoughby JR, Christie MR. Genetic diversity in fishes is influenced by habitat type and life-history variation. *Ecol Evol* 2018;**8**:12022–31. https://doi.org/10.1002/ece3.4661
2. FAO. *The State of World Fisheries and Aquaculture 2024—Blue Transformation in Action*. Rome: FAO, 2024.
3. Lisbeth L. Fish biodiversity and the health of aquatic ecosystems. *Poult Fish Wildl Sci* 2023;**11**:234.
4. Heilpern SA, DeFries R, Fiorella K *et al.* Declining diversity of wild-caught species puts dietary nutrient supplies at risk. *Sci Adv* 2021;**7**:eabf9967. https://doi.org/10.1126/sciadv.abf9967
5. Huang P, Guo W, Wang Y *et al.* Genome-wide association study reveals the genetic basis of growth trait in yellow catfish with sexual size dimorphism. *Genomics* 2022;**114**:110380. https://doi.org/10.1016/j.ygeno.2022.110380
6. Kause A, Nousiainen A, Koskinen H. Improvement in feed efficiency and reduction in nutrient loading from rainbow trout farms: the role of selective breeding. *J Anim Sci* 2022;**100**:skac214. https://doi.org/10.1093/jas/skac214
7. Liu Z, Zhou T, Gao D. Genetic and epigenetic regulation of growth, reproduction, disease resistance and stress responses in aquaculture. *Front Genet* 2022;**13**:994471. https://doi.org/10.3389/fgene.2022.994471
8. Torsabo D, Ishak SD, Noordin NM *et al.* Enhancing reproductive performance of freshwater finfish species through dietary lipids. *Aquacult Nutr* 2022;**2022**:7138012. https://doi.org/10.1155/2022/7138012
9. Figueroa RT, Goodvin D, Kolmann MA *et al.* Exceptional fossil preservation and evolution of the ray-finned fish brain. *Nature* 2023;**614**:486–91. https://doi.org/10.1038/s41586-022-05666-1
10. Gong G, Xiong Y, Xiao S *et al.* Origin and chromatin remodeling of young X/Y sex chromosomes in catfish with sexual plasticity.

Natl Sci Rev 2023;**10**:nwac239. https://doi.org/10.1093/nsr/nwac239
11. Xiong Y, Wang DY, Guo W *et al.* Sexually dimorphic gene expression in X and Y sperms instructs sexual dimorphism of embryonic genome activation in yellow catfish (*Pelteobagrus fulvidraco*). *Biology* 2022;**11**:1818.
12. Jiang H, Sun M, Zhao Y *et al.* The early function of cortisol in liver during *Aeromonas hydrophila* infection: dynamics of the transcriptome and accessible chromatin landscapes. *Front Immunol* 2022;**13**:989075. https://doi.org/10.3389/fimmu.2022.989075
13. Guo C, Ye W, Cao D *et al.* Unraveling the stereoscopic gene transcriptional landscape of zebrafish using FishSED, a fish spatial expression database with multispecies scalability. *Sci China Life Sci* 2024;**67**:843–6. https://doi.org/10.1007/s11427-023-2418-2
14. Guo C, Duan Y, Ye W *et al.* FishGET: A fish gene expression and transcriptome database with improved accuracy and visualization. *iScience* 2023;**26**:106539. https://doi.org/10.1016/j.isci.2023.106539
15. Li H, Zhang W, Ren K *et al.* FishCODE: a web-based information platform for comprehensive omics data exploration in fish research. *J Genet Genom* 2024;**51**:1525–8. https://doi.org/10.1016/j.jgg.2024.09.019
16. Guo C, Ye W, Shi M *et al.* FishSCT: a zebrafish-centric database for exploration and visualization of fish single-cell transcriptome. *Sci China Life Sci* 2023;**66**:2185–8. https://doi.org/10.1007/s11427-022-2293-4
17. Desvignes T, Bardou P, Montfort J *et al.* FishmiRNA: an evolutionarily supported microRNA annotation and expression database for ray-finned fishes. *Mol Biol Evol* 2022;**39**:msac004. https://doi.org/10.1093/molbev/msac004
18. Shao F, Zeng M, Xu X *et al.* FishTEDB 2.0: an update fish transposable element (TE) database with new functions to facilitate TE research. *Database (Oxford)* 2024;**2024**:baae044. https://doi.org/10.1093/database/baae044
19. Rashid I, Nagpure NS, Srivastava P *et al.* HRGFish: a database of hypoxia responsive genes in fishes. *Sci Rep* 2017;**7**:42346. https://doi.org/10.1038/srep42346
20. Sayers EW, Beck J, Bolton EE *et al.* Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Res* 2025;**53**:D20–9. https://doi.org/10.1093/nar/gkae979
21. Dyer SC, Austine-Orimoloye O, Azov AG *et al.* Ensembl 2025. *Nucleic Acids Res* 2025;**53**:D948–57. https://doi.org/10.1093/nar/gkae1071
22. Ren L, Gao X, Cui J *et al.* Symmetric subgenomes and balanced homoeolog expression stabilize the establishment of allopolyploidy in cyprinid fish. *BMC Biol* 2022;**20**:200. https://doi.org/10.1186/s12915-022-01401-4
23. Camacho C, Coulouris G, Avagyan V *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**:421. https://doi.org/10.1186/1471-2105-10-421
24. Pandurangan AP, Stahlhacke J, Oates ME *et al.* The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver. *Nucleic Acids Res* 2019;**47**:D490–4. https://doi.org/10.1093/nar/gky1130
25. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 2011;**39**:W29–37. https://doi.org/10.1093/nar/gkr367
26. Mistry J, Chuguransky S, Williams L *et al.* Pfam: the protein families database in 2021. *Nucleic Acids Res* 2021;**49**:D412–9. https://doi.org/10.1093/nar/gkaa913
27. Chen S, Zhou Y, Chen Y *et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;**34**:i884–90. https://doi.org/10.1093/bioinformatics/bty560
28. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA–MEM. arXiv, https://arxiv.org/abs/1303.3997, 26 May 2013, preprint: not peer reviewed.
29. Li H, Handsaker B, Wysoker A *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9. https://doi.org/10.1093/bioinformatics/btp352
30. McKenna A, Hanna M, Banks E *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**:1297–303. https://doi.org/10.1101/gr.107524.110
31. DePristo MA, Banks E, Poplin R *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;**43**:491–8. https://doi.org/10.1038/ng.806
32. Cingolani P, Platts A, Wang LL *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: sNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* 2012;**6**:80–92. https://doi.org/10.4161/fly.19695
33. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2015;**12**:357–60. https://doi.org/10.1038/nmeth.3317
34. Pertea M, Pertea GM, Antonescu CM *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 2015;**33**:290–5. https://doi.org/10.1038/nbt.3122
35. Niknafs YS, Pandian B, Iyer HK *et al.* TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nat Methods* 2017;**14**:68–70. https://doi.org/10.1038/nmeth.4078
36. Sun L, Luo H, Bu D *et al.* Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res* 2013;**41**:e166. https://doi.org/10.1093/nar/gkt646
37. Kang YJ, Yang DC, Kong L *et al.* CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res* 2017;**45**:W12–6. https://doi.org/10.1093/nar/gkx428
38. El-Gebali S, Mistry J, Bateman A *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res* 2019;**47**:D427–32. https://doi.org/10.1093/nar/gky995
39. Memczak S, Jens M, Elefsinioti A *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 2013;**495**:333–8. https://doi.org/10.1038/nature11928
40. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**:357–9. https://doi.org/10.1038/nmeth.1923
41. Friedländer MR, Mackowiak SD, Li N *et al.* miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* 2012;**40**:37–52. https://doi.org/10.1093/nar/gkr688
42. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res* 2019;**47**:D155–62. https://doi.org/10.1093/nar/gky1141
43. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**:923–30. https://doi.org/10.1093/bioinformatics/btt656
44. Zheng GXY, Terry JM, Belgrader P *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 2017;**8**:14049. https://doi.org/10.1038/ncomms14049
45. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 2018;**19**:15. https://doi.org/10.1186/s13059-017-1382-0
46. Tarasov A, Vilella AJ, Cuppen E *et al.* Sambamba: fast processing of NGS alignment formats. *Bioinformatics* 2015;**31**:2032–4. https://doi.org/10.1093/bioinformatics/btv098
47. Zhang Y, Liu T, Meyer CA *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 2008;**9**:R137. https://doi.org/10.1186/gb-2008-9-9-r137
48. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 2011;**27**:1571–2. https://doi.org/10.1093/bioinformatics/btr167
49. 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

50. Jin F, Li Y, Dixon JR *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* 2013;**503**:290–4. https://doi.org/10.1038/nature12644
51. Heinz S, Benner C, Spann N *et al.* Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol Cell* 2010;**38**:576–89. https://doi.org/10.1016/j.molcel.2010.05.004
52. Perez-Riverol Y, Bai J, Bandla C *et al.* The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res* 2022;**50**:D543–52. https://doi.org/10.1093/nar/gkab1038
53. Bailey TL, Johnson J, Grant CE *et al.* The MEME Suite. *Nucleic Acids Res* 2015;**43**:W39–49. https://doi.org/10.1093/nar/gkv416
54. Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2022;**50**:D165–73. https://doi.org/10.1093/nar/gkab1113
55. Boettiger C, Lang DT, Wainwright PC. rfishbase: exploring, manipulating and visualizing FishBase data from R. *J Fish Biol* 2012;**81**:2030–9. https://doi.org/10.1111/j.1095-8649.2012.03464.x
56. Mei J, Gui JF. Genetic basis and biotechnological manipulation of sexual dimorphism and sex determination in fish. *Sci China Life Sci* 2015;**58**:124–36. https://doi.org/10.1007/s11427-014-4797-9
57. Webster KA, Schach U, Ordaz A *et al.* Dmrt1 is necessary for male sexual development in zebrafish. *Dev Biol* 2017;**422**:33–46. https://doi.org/10.1016/j.ydbio.2016.12.008
58. Mustapha UF, Jiang DN, Liang ZH *et al.* Male-specific *Dmrt1* is a candidate sex determination gene in spotted scat (*Scatophagus argus*). *Aquaculture* 2018;**495**:351–8. https://doi.org/10.1016/j.aquaculture.2018.06.009
59. Chen L, Huang Y, Pan QH *et al.* The Chinese medaka (*Oryzias sinensis*) *dmrt1* gene converts females to males in medaka (*Oryzias latipes*). *Biochim Biophys Gen Sub* 2024;**1868**:130664. https://doi.org/10.1016/j.bbagen.2024.130664
60. Chu Q, Xu T, Zheng W *et al.* Long noncoding RNA AANCR modulates innate antiviral responses by blocking miR-210-dependent MITA downregulation in teleost fish. *Sci China Life Sci* 2021;**64**:1131–48. https://doi.org/10.1007/s11427-020-1789-5
61. Liu H, Tan S, Chen Y *et al.* Regulatory mechanism of miR-722 on C5aR1 and its functions against bacterial inflammation in half-smooth tongue sole (*Cynoglossus semilaevis*). *Int J Biol Macromol* 2023;**252**:126445. https://doi.org/10.1016/j.ijbiomac.2023.126445
62. Gao Y, Shang S, Guo S *et al.* Lnc2Cancer 3.0: an updated resource for experimentally supported lncRNA/circRNA cancer associations and web tools based on RNA-seq and scRNA-seq data. *Nucleic Acids Res* 2021;**49**:D1251–8. https://doi.org/10.1093/nar/gkaa1006
63. Zhao Y, Li H, Fang S *et al.* NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res* 2016;**44**:D203–8. https://doi.org/10.1093/nar/gkv1252
64. Gong J, Mei S, Liu C *et al.* PancanQTL: systematic identification of *cis*-eQTLs and *trans*-eQTLs in 33 cancer types. *Nucleic Acids Res* 2018;**46**:D971–6. https://doi.org/10.1093/nar/gkx861
65. Luo M, Yang W, Bai L *et al.* Artificial intelligence for life sciences: a comprehensive guide and future trends. *Innov Life* 2024;**2**:100105. https://doi.org/10.59717/j.xinn-life.2024.100105

Received: August 4, 2025. **Revised:** September 3, 2025. **Accepted:** September 19, 2025