A chromosomal-level genome assembly and the diet habit-specific amino acid mutation identification of the Cyprinidae fish Ancherythroculter nigrocauda

Yanhong Sun  
Fisheries Research Institute, Wuhan Academy of Agricultural Sciences

Guiying Wang  
Fisheries Research Institute, Wuhan Academy of Agricultural Sciences

Jianfang Gui  
State Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences

Jian Chen  
Fisheries Research Institute, Wuhan Academy of Agricultural Sciences

Pei Li  
Fisheries Research Institute, Wuhan Academy of Agricultural Sciences

Dongmei Zhu  
College of Fishery, Huazhong Agricultural University

Yingwu Liu  
Fisheries Research Institute, Wuhan Academy of Agricultural Sciences

Zongqun Zhang  
Wuhan Xianfeng Aquaculture Technology Co. Ltd

Wei Li  
Wuhan Xianfeng Aquaculture Technology Co. Ltd

Qing Li  
Fisheries Research Institute, Wuhan Academy of Agricultural Sciences

Research Article

Keywords: Ancherythroculter nigrocauda, Cyprinidae, genome assembly, genetic breeding, amino acid mutation

DOI: https://doi.org/10.21203/rs.3.rs-116303/v1
Abstract

Background

*Ancherythroculter nigrocauda* is an endemic Cyprinidae fish in China, it has many desirable traits for genetic breeding, including strong disease resistance, unusual stress tolerance and high efficiency in nutrition update, which have made it an emerging commercial aquaculture fish. With the publication of its close-related species’ genome sequence, we can study the diet-specific genomic mutations within Cyprinidae.

Results

Here we report whole genome assembly of a female *A. nigrocauda* individual constructed using the single molecule DNA sequencing platform PacBio Sequel. With the help of Hi-C anchoring, we successfully placed contigs to chromosome level (2n = 48), yielding a genome size of 1054.05 Mb with contig N50 of 3.40 Mb and scaffold N50 of 42.68 Mb. This genome assembly, which has reached a high base-level accuracy of 99.999%, harboring 33,606 annotated protein-coding genes. We also found 582 genes hold diet-specific amino acid mutation between herbivorous and carnivorous fishes and 26 of them showed significant different expression patterns in liver tissue of these two types of fishes.

Conclusions

The availability of the chromosome-level genome assembly of *A. nigrocauda* provides valuable resources for future in-depth comparative genomics studies and applications including genetic breeding. The diet-specific amino acid mutation can be used in breeding of new strains of carnivorous fishes which feed on herbivorous fodder.

1. Introduction

*Ancherythroculter nigrocauda* (NCBI Taxonomy ID: 263419) is a Cyprinidae fish endemic to the upper Yangtze River in China [1]. This species frequently inhabits in main channels and tributaries of the upper Yangtze River in Sichuan province and Chongqing city in southwest China. In recent decades, the natural resource of this species has suffered severely reduction both in population size and in genetic diversity due to habitat disruption caused by pollution from agriculture, chemical industry, overfishing, and damming [2]. As a result, there is a loss of genetic diversity of it [3]. Nevertheless, *A. nigrocauda* is a popular delicacy, and it can be easily kept alive and fresh because of its strong tolerance to hypoxia and stressful conditions, which make it a promising aquaculture fish in China. It has also been used as new stock in genetic breeding that has bred two new aquaculture varieties, such as the hybrid “Culter Xianfeng NO. 1” and “Bofang Xianfeng NO. 2”. The hybrid “Culter Xianfeng NO. 1” is the F1 offspring of *Culter alburnus* ♂ x *A. nigrocauda* ♀ and “Bofang Xianfeng NO. 2” is the selected variety from *Megalobrama*
amblycephala x A. nigrocauda through fourth-generation hybrid and selection. The excellent traits of A. nigrocauda are presented in both two new varieties.

Despite its high economic value and popularity as a resource in aquaculture breeding, genetic and genomic resources for this fish are still limited. Only a few of genomic nucleotide sequences [4, 5] and single nucleotide polymorphism (SNP) markers are publicly available [6]. Such a lack of resources seriously limited the progress of its genetic breeding and application. And the lack of a reference genome also caused some problems in studying gene expression using RNA-Seq approaches [7].

With the development of sequencing technology, the genome of more than 17 fishes in Cyprinidae have been deposited in NCBI database. However, this number is still faraway from enough compared with the 3,000 species in Cyprinidae. Nowadays, single-molecule real-time sequencing has been used more and more frequently in fish whole genome sequencing [8, 9]. Super long reads from it provided a much better assembly when compared with short reads from Illumina sequencing [10]. Besides, high-throughput chromatin conformation capture (Hi-C) also greatly foster the assembly quality of whole genome by translating genome-wide 3D proximity maps to contigs' clustering, ordering and pseudomolecule construction [8, 11]. Therefore, combination of long-reads sequencing and Hi-C has been used in many genome sequencing projects and proved its power in high-quality genome assembly [8, 12].

Here, we presented a highly contiguous reference genome for A. nigrocauda that used a series of technologies including the single-molecule sequencing platform PacBio Sequel and Hi-C analysis-based chromosome construction. The reference genome assembly of A. nigrocauda will be valuable for both comparative genomics research and genetic breeding research. With the help of this high-quality genome, we also inspected the amino acid mutation associated with diet for A. nigrocauda and its closely related species.

2. Method And Materials

2.1. Karyotype analyses

The karyotype analysis was conducted using the method as described in a previous study [13]. Phytohemagglutinin (PHA) was first injected near the base of the pectoral fin, followed by injecting colchicine in the kidney. Kidney tissues were taken and placed in a beaker, cut into small pieces, followed by mixing with saline. After settling for 15 min, cells in the supernatant were collected via centrifuge. These cells were resuspended using 0.075 M KCl solution for 35 min and then collected via centrifuge. The cells were fixed using Carnoy's Fluid and treated with Giemsa stain before observing using a microscope.

2.2. Samples and Illumina sequencing

As higher heterozygosity always causes difficulties in genome assembly, we want to choose one sample with the lowest heterozygosity in genome sequencing. Therefore, three female A. nigrocauda individuals
were chosen as candidates for genome sequencing because female samples always have lower heterozygosity than male ones for XY/XX sex determination system in many Cyprinidae fishes.

To obtain enough high-quality genomic DNA molecules, the adult female fish individual was dissected and fresh muscle tissues were used for DNA extraction using the traditional phenol/chloroform extraction method [14] and the extracted DNA was quality checked using agarose gel electrophoresis. A single band corresponding to high molecular weight was observed, indicating high integrity of DNA molecules for library construction for the Illumina HiSeq X Ten (Illumina Inc., San Diego, CA, USA) and the PacBio Sequel (Pacific Biosciences of California, Menlo Park, CA, USA) sequencing platforms.

A library with the insert size of 350 bp was constructed for Illumina HiSeq X Ten sequencing platform according to the manufacturer's protocol. To get high-quality Illumina reads for following analysis, ambiguous bases and low-quality reads were first trimmed and filtered using the HTQC package [15]. First, reads with adaptors or generated via duplication were discarded. Reads whose size of Ns or low-quality (≤ 5) bases reached 10% of the total read length were discarded. Second, reads were removed from further analysis if the average quality were smaller than 20 or the read length was shorter than 75 bp. Third, the mate reads were also removed if the corresponding reads had been removed. The processed reads were used for genome assessment analysis. With the K-mer analysis approach GCE software [16], we calculated the number of each 17-mer from the sequencing data to estimate the genome size and heterozygosity. The sample with the lowest heterozygosity was chosen to further PACBIO sequencing.

2.3. PacBio long reads sequencing and genome assembly

Eight libraries with 20Kb insert size were constructed for PacBio Sequel sequencing. PacBio data was assembled using the FALCON v2.0 assembler, which implements a hierarchical assembly approach [17]. The initial step is to error-correct long reads by aligning all reads to a subset of the longest reads. For the full raw data set, only reads longer than 5 kb (length_cut_off = 5 kb) were corrected for generating error-corrected reads. To identify overlaps between raw sequences, we used "p-read" (pre-assembled reads from error corrected reads) longer than 5 kb (length_cut_off_pr = 5 kb) for String Graph assembly. Then FALCON-unzip [17] operates from a completed FALCON job directory, it uses the phasing information of heterozygous positions within individual reads to group the reads into different phasing blocks and haplotypes within each block. In this study, we used primary contigs for further analysis.

2.4 In situ Hi-C library construction and chromosome assembly using Hi-C data

Liver sample taken from the same female A. nigrocauda individual with lowest heterozygosity was used for library construction for Hi-C analysis as described previously [18, 19]. The library was sequenced with 150 bp paired-end mode on the Illumina HiSeq X Ten platform (San Diego, CA, USA). The reads were mapped to assembled genome with Bowtie [20] with two ends of paired reads being mapped to the genome separately. To increase the interactive Hi-C reads ratio, an iterative mapping strategy was
performed as done previously [18, 19]. Only read pairs with both ends uniquely mapped were used for further analysis. From the alignment results of read pairs, self-ligation, non-ligation and other sorts of invalid reads, including Start Near Rsite, PCR amplification, random break, Large Small Fragments and Extreme Fragments, were filtered out by Hi-C lib as described previously [18]. Through tracking restriction sites, contact counts among contigs were calculated and normalized. By clustering the contigs using contig contact frequency matrix, we also corrected some minor errors in the FALCON-unzip assembly. Finally, uniquely mapped read pairs were used for clustering, ordering and orientating contigs to construct chromosomes using LACHESIS version c23474f [21].

2.5. PACBIO Iso-Seq Sequencing of mRNAs

We analyzed the full-length transcripts of *A. nigrocauda* using the Iso-Seq protocol [22]. Total RNA was extracted using the TRiZol reagent (Thermo Fisher Scientific, Waltham, MA, Cat. 15596018) from combined tissues including heart, liver, brain, spleen, kidney, gonads, gill, intestine, eyes, muscle and blood from the same individual used for PACBIO sequencing. Two SMRT cell libraries were constructed with size selection of 1–3 Kb and >3 Kb using the Blue Pippin (Sage Science, MA, USA) and sequenced with PacBio Sequel platform.

2.6 Annotation of repeats and protein-coding genes

First, Tandem Repeats Finder v4.09 (TRF) [23] was used with parameters as ‘Match = 2, Mismatching penalty = 7, Delta = 7, PM = 80, PI = 10, Minscore = 50, MaxPeriod = 2,000’ to annotate the tandem repeats in *A. nigrocauda*. Then we identified transposable elements (TEs) using a combination of de novo and homology-based approaches, at both the DNA and protein levels. At the DNA level, LTR_FINDER v1.0.5 [24] was first used to identify LTR and RepeatModeler version open-1.0.11 (parameters set as ‘–engine_db wublast’) [25, 26] was utilized to construct a de novo repeat library, which comprised a repeat consensus database with classification information. Then, RepeatMasker [25, 26] was used to search for TEs using Repbase [27] TE library and de novo repeat library. At the protein level, RepeatProteinMask (the parameter set as ‘-noLowSimple -pvalue 0.0001 -engine wublast’ ) within the RepeatMasker (the parameter set as ‘-nolow -no_is -norna -parallel 2 -e wublast–pvalue 0.0001’ ) package was used to search against the TE protein database.

Identification of protein-coding genes involved homolog-based prediction, de novo predictions, and the use of Iso-Seq data as follows.

(1) Homolog-based prediction. Homologous proteins of the closely related species including *Ctenopharyngodon idellus, Cyprinus carpio, Danio rerio, Gasterosteus aculeatus, Takifugu rubripes*, were aligned to the *A. nigrocauda* using tblastn [28] with an E-value cutoff of 1e-5. The aligned sequences, and their corresponding query proteins were then filtered and passed to Exonerate v2.2.0 [29] with parameter ‘–model protein2genome –showtargetgff 1’ to search for accurate spliced alignments.

(2) De novo prediction was performed on the transposons-masked genome. Augustus [30] using model training based on coding sequences from *Ctenopharyngodon idellus, Cyprinus carpio, Danio rerio,*
*Gasterosteus aculeatus* and *Takifugu rubripes* were used to predict coding genes.

(3) Iso-Seq assist prediction. Full-length transcripts obtained using Iso-Seq were mapped to the genome using Gmap (parameter `-n 5 --min-intronlength = 9 --max-intronlength-middle = 200000 -t 15 -A -f 2`) [31], and then the TransDecoder program [32] was used to predict the open reading frames (ORFs) in transcripts to define putative coding sequences (CDSs).

Finally, gene models predicted from all above methods were combined by MAKER [33].

To annotate functions of protein-coding genes in the *A. nigrocauda* genome, we searched all predicted gene sequences in NCBI non-redundant protein (NR), InterPro (InterProScan, RRID:SCR_005829) [34], GO (Gene Ontology), KEGG (RRID:SCR_012773) [35], Swissprot [36], TrEMBL databases [36] by tblastn [28] and BLASTX [37]. A threshold of e-value of $1 \times 10^{-5}$ was used for all BLAST applications.

### 2.7 Gene clustering by OrthoMCL

The sequences of protein-coding genes from *A. nigrocauda* and the closely related fishes, including *Oryzias latipes, Ictalurus punctatus, Hippocampus comes, Dicentrarchus labrax, Ctenopharyngodon idellus, Cyprinus carpio, Danio rerio, Takifugu rubripes, Tetraodon nigroviridis, Paralichthys olivaceus, Cynoglossus semilaevis, Lepisosteus oculatus, M. amblycephala* and *C. alburnus* were used for a gene family clustering analysis. For simplicity, only proteins encoded by transcripts with the longest alternative splicing sites were used in this analysis. First, we generated pairwise protein sequence alignments for all proteins by running blastp [37] with an e-value threshold of $1 \times 10^{-5}$. Second, OrthoMCL [38] was used to cluster similar proteins by setting the main inflation value at 1.5 and using the default settings for other parameters.

### 2.8 Phylogenetic analysis

Using single copy orthologs, we probed the phylogenetic relationships between *A. nigrocauda* and other fishes with sequenced genomes. To this end, protein sequences of single-copy genes were aligned using MUSCLE [39]. Guided by the protein multi-sequence alignment, the alignment of the coding DNA sequences (CDS) for those genes were generated and concatenated for following analyses. The phylogenetic relationships were constructed using PhyML [40] using the concatenated nucleotide alignment with the JTT + G + F model.

We further estimated divergent times for all pair using the phylogenetic tree using r8s [41], which were used as input, together with molecular clock data from the divergence time from the TimeTree database [42], to estimate species divergence time for all pairs of species in the phylogenetic tree using MCMCtree program (from PAML) [43].

### 2.9 Expansion and contraction of gene families

Based on the identified gene families and the constructed phylogenetic tree with predicted divergence time of the 14 species, we used CAFÉ [44] to analyze gene family expansion and contraction. In CAFÉ,
random birth and death model is proposed to study gene gain or loss in gene families across a specified phylogenetic tree. Then, conditional p-value was calculated for each gene family, and family with conditional p-value less than 0.05 was considered to have an accelerated rate for gene gain or loss. These expansion and contraction gene families in *A. nigrocauda* (p-value ≤ 0.05) were mapped to KEGG pathways for functional enrichment analysis, which was conducted using the enrichment methods. This method implemented hypergeometric test algorithms and the Q-value (FDR, False Discovery Rate) was calculated to adjust the p-value using R package (https://github.com/StoreyLab/qvalue).

### 2.10 Identification of diet-specific amino acid mutation and their expression pattern in different diet fishes.

To get *M. amblycephala* (blunt snout bream, BSB) and *C. idellus* (grass carp, GC) specific amino acid (AA) mutation, we first blasted proteins from BSB, GC, *C. albumus* (topmouth culter, TC) and *A. nigrocauda* to zebrafish proteome and get the corresponding relationship among them [45–47]. Then proteins mapped to the same zebrafish protein were considered as orthologs and aligned together with PRANK v.170427 (http://wasabiapp.org/software/prank/). Then in-house Python script was used to find AA presented in BSB and GC but absent in TC and *A. nigrocauda*.

To find the expression pattern of genes identified previously, we used liver transcriptome data from SRR8225322 (BSB), SRR8225320 (TC), SRR8225318 (BSB & TC hybrid F1) and SRR8225330 (BSB & TC hybrid F1) [46]. Then reads from SRR8225322 were mapped to BSB genome, SRR8225320 were mapped to TC genome with Hisat2 v2.1.0 [48]. Reads from SRR8225318 and SRR8225330 were mapped to both BSB and TC genome and get the expressed value for each gene with HTSEQ v 0.12.4 [49]. Only genes with average FPKM larger than 0.2 were left for following analysis. Then statistics analysis was carried out in R v3.6.1 (https://www.R-project.org/).

### 3. Results And Discussion

#### 3.1 Sample preparation

As *A. nigrocauda* had not been backcrossed to reduce heterozygosity, we chose three individuals to do genome survey with aim of finding one with the lowest heterozygosity, which will greatly benefit the assembly of genome. We sequenced 50.59, 56.64, and 47.06 Gb of Illumina reads for these three samples and found their heterozygosity levels varied remarkably (0.54%, 1.04% and 0.42%). However, they don’t have much difference on genome size or repeat contents (Table 1). Based on these results, the sample with the lowest heterozygosity (0.42%) was chosen for the following PACBIO genome sequencing (Fig. 1a.). Its k-mer distribution is shown in Supplementary Figure S1.

#### 3.2 PACBIO sequencing, genome assembly, and quality evaluation
Genomic DNA extracted from muscle tissues was used in library construction of PACBIO sequencing. In total, 101.50 Gb of long reads were sequenced on PACBIO Sequel platform (Table 2, Supplementary Figure S2). With the help of these long reads, we assembled the *A. nigrocauda* genome into 538 contigs with 1,053.82 Mb sequences, which is almost identical to that in k-mer analysis (Table 2. Supplementary Table S1). The contig N50 of this assembly is 3.40 Mb and the longest contig lasts for 14.79 Mb (Table 3). To correct the sequencing error in assembled genome, filtered Illumina reads were mapped to it with Bowtie2 and three rounds of consensus correction with Pilon were performed. The GC content of assembled genome is 37.6% and the average mapping depth is about 120X for the whole genome (Table 3, Supplementary Figure S3).

As no karyotype information about *A. nigrocauda* has been reported, we carried out Giemsa stain of dividing kidney cells under microscope and found each kidney cell has 24 pairs of chromosomes (Fig. 1b). To anchor assembled contigs to the 24 chromosomes, 102.34 Gb of Hi-C reads were generated. With the help from LACHESIS software, we successfully organized assembled contigs into 24 groups, with a combined size of 1,054.05 Mb (scaffold N50 was 42.68 Mb) (Fig. 1c). The 24 chromosomes ranged in size from 32.61 Mb to 66.38 Mb and they covered about 99.52% of the whole genome (Supplementary Table S2).

We evaluated the completeness and quality of the newly assembled *A. nigrocauda* genome with Illumina high-quality reads. We were able to align 98.21% of the NGS reads to the genome assembly, with insertion length distribution of read pairs peaking at around 300 bp, which was consistent with library construction designed for the Illumina sequencing. And we found 99.97% of assembled genome was covered by high-quality reads, among which 99.69% was covered by over 20X of reads, suggesting high quality level of the assembled genome (Supplementary Table S3). Above results indicate that our assembly is of high quality and continuity.

### 3.3 Genome annotation

To identify repeat elements in assembled genome, combined de novo prediction and homolog search against RepBase database was used in this study. In total, we identified 51.86% of the *A. nigrocauda* assembly as repetitive sequences, among which DNA transposons represented the most abundant fraction of transposable elements (35.02% of the whole genome), and LTR retrotransposons comprised 11.20% of the whole genome (Supplementary Table S4, Supplementary Table S5). Besides, we found the diversity rate for DNA transposon and LTR found in de novo prediction is much lower than that in RepBase (Supplementary Figure S4).

To identify protein-coding genes in our assembly, we combined results from *de novo* and homolog searching from 5 close-related fish genomes. Totally, we annotated 33,606 protein-coding genes in the *A. nigrocauda* genome assembly (Table 4). The distributions of gene density, repeat density, non-coding genes and GC density across the 24 chromosomes of *A. nigrocauda* genome were further illustrated in Fig. 2. We also found that the regions with low gene density typically had high repeat content, while the regions with high repeat content usually had high GC content. The gene number, gene length, CDS length,
exon length and intron length distribution were all comparable with the related fishes (Supplementary Figure S5).

We evaluated the genome annotation of the assembled A. nigrocauda genome using BUSCO v.3.0.2 [50] with actinopterygii_odb9 data set to assess the completeness of the genome. The A. nigrocauda genome assembly contained 94.50% of the BUSCO genes, indicating a high level of completeness of the genome assembly and high quality of annotation (Supplementary Table S6). Besides, most of the protein-coding genes (29,674, 88.30%) are homologous to genes with known functions (Supplementary Table S7). We also identified 6,712 non-coding RNA (ncRNA) including 975 microRNAs (miRNAs), 5,017 transfer RNAs (tRNAs), 356 ribosomal RNAs (rRNAs) and 364 small nuclear RNAs (Supplementary Table S8).

3.4 Evolutionary analysis of A. nigrocauda with other fish species

A phylogenetic tree was reconstructed with 945 single-copy orthologous genes identified among A. nigrocauda and 14 other vertebrate species (Fig. 3). The result showed that A. nigrocauda was most closely related to C. alburnus, and the two species diverged from their common ancestor around 7.3–9.7 million years ago (MYA). And M. amblycephala diverged from their three Cultrinae species common ancestors around 11.6–14.2 MYA (Fig. 3).

According to divergence times and phylogenetic relationships, 1,385 gene families out of 28,833 gene families were significantly expanded in the A. nigrocauda (P < 0.05, Supplementary Figure S6). Those expanded gene families included 148 significantly enriched (q-value < 0.05) KEGG pathways (Supplementary Table S9). In organismal systems, the digestive system is one of the top pathways in expanded gene families, and many KEGG pathways, including protein digestion and absorption (Fig. 4), which may contribute to the highly efficient protein conversion ability of A. nigrocauda (Lu et al., 2019; Tang et al., 2008). Meanwhile, we found that transferrin (TF) gene is one of the expanded genes in hypoxia induced factor-1 (HIF-1) signaling pathway (Supplementary Figure S7). TF gene participated in HIF-1 signaling pathway, which was the target gene in the downstream and played an important role in the hypoxia regulation (Rolfs et al., 1997).

Besides, 105 genes were subjected to positive selection and mainly classified into 62 KEGG pathways (Supplementary Table S10). Among these pathways, many genes related to protein digestion and absorption, amino acid metabolism and other metabolism were subjected to positive selection, which may contribute to the highly efficient protein conversion ability of A. nigrocauda.

3.5 Comparison between different diet habit within close-related species

A. nigrocauda and C. alburnus (TC) are both belong to Cyprinidae and are carnivorous fishes, while their close-related species in Cyprinidae, M. amblycephala (BSB) and C. idellus (GC) are pure herbivorous.
Therefore, we want to know whether there is some diet-specific amino acid (AA) mutation between herbivorous and carnivorous fishes. More than 2 diet-specific AA mutations were found in 582 genes out of 15,748 in total. GO enrichment analysis showed that 10 out of top 20 pathways were related to steroid biosynthesis and lipid transport (Supplementary Table S11). This pattern further verified how herbivorous fishes absorb nutrients from plants in GC [47].

Besides, inter-species hybridization between BSB and TC produces herbivorous offspring [46]. We used transcriptome data from the liver tissue of the parents and offspring to find diet-specific expression patterns for these 582 genes. After removing genes with low expression level, 297 genes were left for diet-associated analysis. Finally, we found 26 genes were significantly associated with diet (p < 0.05) (Supplementary Table S12). Among these genes, we also found some were associated with apolipoprotein's metabolism, such as ApoBb. These genes might be key functional elements for adapting to different diet in different fishes.

4. Conclusion

In this study, we constructed a high-quality chromosome-level assembly of the fish *A. nigrocauda* using combined methods including PacBio DNA sequencing, Illumina HiSeq X Ten DNA sequencing and Hi-C interaction analysis. Using the long reads from PacBio Sequel platform and short reads from the Illumina HiSeq X Ten platform, we successfully constructed contig assembly for *A. nigrocauda*. Leveraging contact information among contigs from Hi-C technology, we further improved the assembly to the chromosome-level quality. We annotated 33,606 protein-coding genes in the *A. nigrocauda* genome, 29,674 of which were functionally annotated, suggesting high conservation of genes in this economically important fish species. With 945 single-copy orthologs from *A. nigrocauda* and other related fish species, we constructed the phylogenetic relationship of these fishes, and found that *A. nigrocauda* might have diverged from its common ancestor of *C. alburnus* around 7.3–9.7 MYA.

Given the increasing interests in fish genome evolution and the economic importance of *A. nigrocauda* for aquaculture, our genomic data provide valuable genetic resources for functional genomics investigations for the research community.

Declarations

Ethics approval and consent to participate

All the fish used in this study were owned by authors of this study and reared in Fisheries Research Institute, Wuhan Academy of Agricultural Sciences, Wuhan, China (30°20′N, 114°14′E). All experimental steps in this study were approved by the Committee for Animal Experiments of the Wuhan Academy of Agricultural Sciences and were carried out in accordance with the Laboratory Animal Management Principles of China.

Consent for publication
All the authors have read the manuscript and consent for its publication.

**Availability of data and materials**

The raw data from our genome project was deposited in the NCBI Sequence database with Bioproject IDs PRJNA533477. The Illumina, PacBio and Hi-C sequencing data are available from NCBI via the accession number of SRR8929633, SRR8929629 and SRR8929634, respectively. The PacBio transcriptome sequencing data were deposited to NCBI via the accession number of SRR8929630 and SRR8929631. The final chromosome assembly was submitted to NCBI Assembly with the accession number of VTFQ00000000 in NCBI.

**Competing interests**

The authors declare no conflict of interest. The sponsors had no role in the design, execution, interpretation, or writing of the study.

**Funding**

This work was supported by the Technical innovation Project of Hubei Province (Grant No. 2018ABA105), the Huanghe Yingcai Project of Wuhan (Grant No. [2014]8 Wuhan Rencai).

**Authors contribution**

Q. Li conceived the project. Y. H. Sun and G. Y. Wang wrote this manuscript. Y. H. Sun and J. F. Gui revised the manuscript. J. Chen, P. Li and D. M. Zhu performed the genome assembly and data analysis. Y. W. Liu, Z. Q. Zhang and W. Li collected the samples and extracted the DNA and RNA. All authors have read, edited and approved the final version of the manuscript.

**Acknowledgements**

We thank Frasergen Bioinformatics for providing technique supports for this work.

**References**

1. Liu C, Gao X, Wang H, Liu H, Cao W, Danley PD: *Reproductive characteristics of Ancherythroculter nigrocauda, an endemic fish in the upper Yangtze River, China*. *Fisheries Science* 2013, 79(5):799–806.
2. Liu JK, Cao W: *Fish resources of the Yangtze River basin and the tactics for their conservation (in Chinese)*. *Resour Environ Yangtze Valley* 1992, 1:7.
3. Zhai D, Zhang Z, Zhang F, Liu H, Cao W, Gao X: *Genetic diversity and population structure of a cyprinid fish (Ancherythroculter nigrocauda) in a highly fragmented river*. *Journal of Applied Ichthyology* 2019.
4. Sun Y, Wang G, Zhu D, Chen J, Li P, Li Q: Development of polymorphic microsatellite loci isolated from the Ancherythoculter nigrocauda. 2014, 6(4):919–923.

5. Sun Y, Li Q, Wang G, Zhu D, Chen J, Li P: Development of transcript-associated microsatellite markers in Ancherythoculter nigrocauda and cross-amplification in Culter alburnus. Genet Mol Res 2015, 14(4):5.

6. Sun Y, Li Q, Wei H, Wang G, Chen J, Li P: Single nucleotide polymorphism identification in growth-related genes from the transcriptome of the fish Ancherythoculter nigrocauda. Conservation Genetics Resources 2017.

7. Lu X, Wen H, Li Q, Wang G, Li P, Chen J, Sun Y, Yang C, Wu F: Comparative analysis of growth performance and liver transcriptome response of juvenile Ancherythoculter nigrocauda fed diets with different protein levels. Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics 2019.

8. Sun LN, Gao T, Wang FL, Qin ZL, Yan LX, Tao WJ, Li MH, Jin CB, Ma L, Kocher TD et al: Chromosome-level genome assembly of a cyprinid fish Onychostoma macrolepisby integration of nanopore sequencing, Bionano and Hi-C technology. Molecular ecology resources 2020.

9. Suda A, Nishiki I, Iwasaki Y, Matsuura A, Akita T, Suzuki N, Fujiwara A: Improvement of the Pacific bluefin tuna (Thunnus orientalis) reference genome and development of male-specific DNA markers. Scientific reports 2019, 9(1):14450.

10. Jayakumar V, Ishii H, Seki M, Kumita W, Inoue T, Hase S, Sato K, Okano H, Sasaki E, Sakakibara Y: An improved de novo genome assembly of the common marmoset genome yields improved contiguity and increased mapping rates of sequence data. BMC genomics 2020, 21(Suppl 3):243.

11. Burton JN, Adey A, Patwardhan RP, Qiu RL, Kitzman JO, Shendure J: Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nature Biotechnology 2013, 31(12):1119–.

12. Jaworski CC, Allan CW, Matzkin LM: Chromosome-level hybrid de novo genome assemblies as an attainable option for non-model insects. Molecular ecology resources 2020.

13. Foresti F, Oliveira C, Almeida-Toledo LFD: A method for chromosome preparations from large fish specimens using in vitro short-term treatment with colchicine. Experientia 1993, 49(9):810–813.

14. Sambrook J, Fritsch EF, Maniatis T: Molecular Cloning: A Laboratory Manual. Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 1989.

15. Neff KL, Argue DP, Ma AC, Lee HB, Clark KJ, Ekker SC: Mojo Hand, a TALEN design tool for genome editing applications. BMC bioinformatics 2013, 14:1.

16. Binghang Liu YS, Jianying Yuan, Xuesong Hu, Hao Zhang, Nan Li, Zhenyu Li, Yanxiang Chen, Desheng Mu, Wei Fan: Estimation of genomic characteristics by analyzing k-mer frequency in denovo genome projects. Quantitative Biology 2013.

17. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O’Malley R, Figueroa-Balderas R, Morales-Cruz A et al: Phased diploid genome assembly with single-molecule real-time sequencing. Nature methods 2016, 13(12):1050–1054.
18. Gong G, Dan C, Xiao S, Guo W, Huang P, Xiong Y, Wu J, He Y, Zhang J, Li X et al: Chromosomal-level assembly of yellow catfish genome using third-generation DNA sequencing and Hi-C analysis. Gigascience 2018, 7(11).

19. Xu S, Xiao S, Zhu S, Zeng X, Luo J, Liu J, Gao T, Chen N: A draft genome assembly of the Chinese sillago (Sillago sinica), the first reference genome for Sillaginidae fishes. Gigascience 2018, 7(9).

20. Langmead B, Trapnell C, Pop M, Salzberg SL: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009, 10(3):R25.

21. Near TJ, Dornburg A, Eytan RI, Keck BP, Smith WL, Kuhn KL, Moore JA, Price SA, Burbrik FT, Friedman M: Phylogeny and tempo of diversification in the superradiation of spiny-rayed fishes. Proceedings of the National Academy of Sciences of the United States of America 2013, 110(31):12738.

22. Rhoads A, Au KF: PacBio Sequencing and Its Applications. Genomics, Proteomics & Bioinformatics 2015, 13(5):278–289.

23. Benson G: Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research 1999, 27(2):573–580.

24. Xu Z, Wang H: LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic acids research 2007, 35(Web Server issue):W265-268.

25. Chen N: Using RepeatMasker to identify repetitive elements in genomic sequences. In: Current protocols in bioinformatics. vol. 4; 2004: 14.

26. Tarailo-Graovac M, Chen N: Using RepeatMasker to identify repetitive elements in genomic sequences. Current protocols in bioinformatics 2009, Chap. 4:Unit 4 10.

27. Bao W, Kojima KK, Kohany O: Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 2015, 6:11.

28. Gertz EM, Yu YK, Agarwala R, Schaffer AA, Altschul SF: Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. BMC Biol 2006, 4:41.

29. Slater GS, Birney E: Automated generation of heuristics for biological sequence comparison. BMC bioinformatics 2005, 6:31.

30. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B: AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic acids research 2006, 34(Web Server issue):W435-439.

31. Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ: GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. Methods Mol Biol 2016, 1418:283–334.

32. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M et al: De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols 2013, 8(8):1494–1512.

33. Campbell MS, Holt C, Moore B, Yandell M: Genome Annotation and Curation Using MAKER and MAKER-P. Current protocols in bioinformatics 2014, 48:4 11 11–39.
34. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R: **InterProScan**: protein domains identifier. *Nucleic acids research* 2005, **33**(Web Server issue):W116-120.

35. Kanehisa M, Goto S: **KEGG**: kyoto encyclopedia of genes and genomes. *Nucleic acids research* 2000, **28**(1):27–30.

36. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O’Donovan C, Phan I et al: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research* 2003, **31**(1):365–370.

37. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications.** *BMC bioinformatics* 2009, **10**:421.

38. Li L, Stoeckert CJ, Jr., Roos DS: **OrthoMCL**: identification of ortholog groups for eukaryotic genomes. *Genome research* 2003, **13**(9):2178–2189.

39. Edgar RC: **MUSCLE**: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 2004, **32**(5):1792–1797.

40. Guindon S, Gascuel O: A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 2003, **52**(5):696–704.

41. Sanderson MJ: **r8s**: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 2003, **19**(2):301–302.

42. Kumar S, Stecher G, Suleski M, Hedges SB: **TimeTree**: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* 2017, **34**(7):1812–1819.

43. Yang Z: **PAML**: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* 1997, **13**(5):2.

44. De Bie T, Cristianini N, Demuth JP, Hahn MW: **CAFE**: a computational tool for the study of gene family evolution. *Bioinformatics (Oxford, England)* 2006, **22**(10):1269–1271.

45. Howe K, Clark MD, Torroja CF, Torrance J, Berthelon C, Hufford M, Collins JE, Humphray S, McLaren K, Matthews L et al: The zebrafish reference genome sequence and its relationship to the human genome. *Nature* 2013.

46. Ren L, Li WH, Qin QB, Dai H, Han FM, Xiao J, Gao X, Cui JL, Wu C, Yan XJ et al: The subgenomes show asymmetric expression of alleles in hybrid lineages of Megalobrama amblycephala x Culter alburnus. *Genome research* 2019, **29**(11):1805–1815.

47. Wang Y, Lu Y, Zhang Y, Ning Z, Li Y, Zhao Q, Lu H, Huang R, Xia X, Feng Q et al: The draft genome of the grass carp (Ctenopharyngodon idellus) provides insights into its evolution and vegetarian adaptation. *Nat Genet* 2015, **47**(6):625–631.

48. Kim D, Langmead B, Salzberg SL: **HISAT**: a fast spliced aligner with low memory requirements. *Nature methods* 2015, **12**(4):357–360.

49. Anders S, Pyl PT, Huber W: **HTSeq**: a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)* 2015, **31**(2):166–169.
50. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics* 2015, 31(19):3210–3212.

**Tables**

**TABLE 1. Sequencing and genome survey for three *A. nigrocauda* samples**

| ID    | Total Amount (Gb) | GC Content (%) | Revised Genome Size (Mb) | Heterozygous Ratio (%) | Repeat (%) |
|-------|-------------------|----------------|--------------------------|------------------------|------------|
| An-X4 | 50.598243         | 38.1           | 1072                     | 0.54                   | 57.61      |
| An-X5 | 56.6435877        | 38.1           | 1091                     | 1.04                   | 56.33      |
| An-X6 | 47.0688381        | 38.2           | 1054                     | 0.42                   | 58.17      |

**TABLE 2. Sequencing data generated for *A. nigrocauda* genome assembly and annotation**

| Library type | Platform       | Library size (bp) | Data size (Gb) | Application                              |
|--------------|----------------|-------------------|----------------|------------------------------------------|
| Short reads  | HiSeq X Ten    | 350               | 59.03          | Genome survey and genomic base correction|
| Long reads   | PacBio Sequel  | 20,000            | 101.50         | Genome assembly                          |
| Hi-C         | HiSeq X Ten    | 300-500           | 102.34         | Chromosome construction                   |
| Iso-Seq      | PacBio Sequel  | 1-3000            | 16.05          | Annotation                               |
|              |                | >3000             | 13.08          |                                          |

**TABLE 3. Statistics for genome assembly of *A. nigrocauda***
|                                | PacBio                      | Pacbio + Hi-C                |
|--------------------------------|-----------------------------|------------------------------|
| Total assembly size of contigs (bp) | 1,053,816,138              |                              |
| Number of contigs               | 538                         |                              |
| N50 contig length (bp)          | 3,397,711                   |                              |
| N90 contig length (bp)          | 1,214,437                   |                              |
| N50 contig count                | 86                          |                              |
| N90 contig count                | 287                         |                              |
| Longest contig (bp)             | 14,790,580                  |                              |
| Percentage of anchoring (%)     |                             | 99.54%                       |
| Total assembly size of scaffolds (bp) | 1,054,045,138              |                              |
| Number of scaffolds             | 80                          |                              |
| N50 scaffolds length (bp)       | 42,678,128                  |                              |
| N90 scaffolds length (bp)       | 36,063,290                  |                              |
| N50 scaffolds count             | 11                          |                              |
| N90 scaffolds count             | 21                          |                              |
| GC content (%)                  | 37.6                        |                              |
| Percentage of repeat sequences (%) | 51.86                      |                              |
| Number of genes (%)             | 33606                       |                              |

**TABLE 4. Statistics for genome annotation of *A. nigrocauda***
| Gene set  | Number | Average gene length(bp) | Average CDS length(bp) | Average exon num per gene | Average exon length(bp) | Average intron length(bp) |
|-----------|--------|--------------------------|------------------------|---------------------------|-------------------------|---------------------------|
| De novo   |        |                          |                        |                           |                         |                           |
| AUGUSTUS  | 35,028 | 14,923.75                | 1,339.69               | 7.53                      | 177.93                  | 2,080.43                  |
| GeneID    | 26,847 | 24,059.41                | 1,417.27               | 6.57                      | 215.66                  | 4,063.82                  |
| genscan   | 31,744 | 23,305.06                | 1,543.20               | 8.13                      | 189.73                  | 3,050.63                  |
| GlimmHMM  | 77,677 | 9,659.35                 | 651.19                 | 4.38                      | 148.55                  | 2,662.26                  |
| Homolog   |        |                          |                        |                           |                         |                           |
| C.idellus | 52,513 | 9,060.60                 | 917.60                 | 4.61                      | 199.02                  | 2,255.30                  |
| C.carpio  | 52,435 | 9,996.37                 | 953.38                 | 4.75                      | 200.63                  | 2,410.17                  |
| D.rerio   | 52,513 | 14,868.90                | 1,181.92               | 5.41                      | 218.48                  | 3,103.76                  |
| G.aculeatus | 40,760 | 11,456.61                | 943.22                 | 5.19                      | 181.70                  | 2,508.48                  |
| T.rubripes | 37,036 | 12,418.14                | 984.86                 | 5.56                      | 177.12                  | 2,507.06                  |
| ISOseq    | 808    | 16,979.80                | 1,482.66               | 10.04                     | 298.37                  | 1,546.50                  |
| MAKER     | 25,781 | 18,447.64                | 1,683.63               | 9.15                      | 184.22                  | 2,057.83                  |
| Final Set | 33,606 | 15,253.83                | 1,395.92               | 7.74                      | 185.79                  | 2,050.64                  |

Figures
Figure 1

Chromosomal construction and characterization (a) Image of A. nigrocauda; (b) Chromosomal characterization through karyotype analysis of A. nigrocauda; (c) Chromosomal construction using DNA interaction heatmap generated in Hi-C analysis (resolution: 500 Kb).
Figure 2

Distribution of protein-coding genes, repeats, non-coding genes, and GC ratio in the A. nigrocuada genome.
Figure 3

Phylogenetic relationship between A. nigrocauda and other fish species. The estimated species divergence time (million years ago) and the 95% confidential intervals were labeled at each branch site. The divergence used for time recalibration was illuminated as red dots in the tree.
Figure 4

Functional analysis of KEGG enrichment of A. nigrocauda expanded gene families

Supplementary Files

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

- SupplementaryFile1208.doc
- SupplementaryTabledataset.xlsx