ADVANCES IN GENOME ASSEMBLY FOR FISHERIES AND AQUACULTURE

EDITED BY: Liang Guo, Roger Huerlimann and Ka Yan Ma

PUBLISHED IN: Frontiers in Genetics and Frontiers in Veterinary Science

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88974-817-4
DOI 10.3389/978-2-88974-817-4

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

ADVANCES IN GENOME ASSEMBLY FOR FISHERIES AND AQUACULTURE

Topic Editors:
Liang Guo, Chinese Academy of Fishery Sciences (CAFS), China
Roger Huerlimann, Okinawa Institute of Science and Technology Graduate University, Japan
Ka Yan Ma, Sun Yat-sen University, China

Citation: Guo, L., Huerlimann, R., Ma, K. Y., eds. (2022). Advances in Genome Assembly for Fisheries and Aquaculture. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88974-817-4

Table of Contents

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 May 2021
doi: 10.3389/fgene.2021.658619

Genome Sequencing and Assembly Strategies and a Comparative Analysis of the Genomic Characteristics in Penaeid Shrimp Species

Jianbo Yuan[1,2,3], *Xiaojun Zhang*[1,2,3], *Fuhua Li*[1,2,3] *and Jianhai Xiang*[1,2,3]*

[1] CAS Key Laboratory of Experimental Marine Biology, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, China, [2] Laboratory for Marine Biology and Biotechnology, Qingdao National Laboratory for Marine Science and Technology, Qingdao, China, [3] Center for Ocean Mega-Science, Chinese Academy of Sciences, Qingdao, China

OPEN ACCESS

Edited by:
Liang Guo,
South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences (CAFS), China

Reviewed by:
Tanaporn Uengwetwanit,
National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand
Xinhai Ye,
Zhejiang University, China

***Correspondence:**
Jianhai Xiang
jhxiang@qdio.ac.cn

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: 26 January 2021
Accepted: 17 March 2021
Published: 03 May 2021

Citation:
Yuan J, Zhang X, Li F and Xiang J (2021) Genome Sequencing and Assembly Strategies and a Comparative Analysis of the Genomic Characteristics in Penaeid Shrimp Species. Front. Genet. 12:658619. doi: 10.3389/fgene.2021.658619

Penaeid shrimp (family Penaeidae) represents one of the most economically and ecologically important groups of crustaceans. However, their genome sequencing and assembly have encountered extreme difficulties during the last 20 years. In this study, based on our previous genomic data, we investigated the genomic characteristics of four penaeid shrimp species and identified potential factors that result in their poor genome assembly, including heterozygosity, polyploidization, and repeats. Genome sequencing and comparison of somatic cells (diploid) of the four shrimp species and a single sperm cell (haploid) of *Litopenaeus vannamei* identified a common bimodal distribution of K-mer depths, suggesting either high heterozygosity or abundant homo-duplicated sequences present in their genomes. However, penaeids have not undergone whole-genome duplication as indicated by a series of approaches. Besides, the remarkable expansion of simple sequence repeats was another outstanding character of penaeid genomes, which also made the genome assembly highly fragmented. Due to this situation, we tried to assemble the genome of penaeid shrimp using various genome sequencing and assembly strategies and compared the quality. Therefore, this study provides new insights about the genomic characteristics of penaeid shrimps while improving their genome assemblies.

Keywords: penaeid shrimp, genome, genome assembly, genomic characteristic, whole genome duplication

INTRODUCTION

Penaeid shrimp belongs to Penaeidae, a family of Decapoda (Crustacea), which comprise many aquatic animals with high ecological and economic values, such as the Pacific white shrimp *Litopenaeus vannamei*, Chinese shrimp *Fenneropenaeus chinensis*, giant tiger prawn *Penaeus monodon*, and kuruma prawn *Marsupenaeus japonicus* (Farfante and Kensley, 1997; Wilson et al., 2000; Koyama et al., 2010). These species are the subject of the most important group in fisheries and aquaculture and have therefore attracted considerable research attention (Dall et al., 1990). According to the statistics from the Food and Agriculture Organization of the United Nations

(FAO), shrimp and prawn (majorly penaeid shrimp) are the main groups of exported species that account for ∼16% of the total value of internationally traded fishery production in 2018, just less than that of salmon and trout (∼18%) (FAO, 2020). The production of farmed shrimp reached >6 million tonnes in 2018, valued at over US$38 billion. As the most important farmed crustacean species, *L. vannamei* alone contributed 53% of the total farmed crustacean production. Due to their high commercial values, genome-based selective breeding programs have been conducted to ensure sustainable and profitable production.

In addition to economical values, penaeid shrimp also exhibits some special biological features, including complex body plans, and novelties (Farfante and Kensley, 1997), high frequency of intermittent molting (about 50 molts during a lifetime) (Godin et al., 1996), and the fastest nerve signal conducting speed (∼200 ms^{-1}) in animals (Fan et al., 1961). However, the detailed mechanisms of these biological features are far from understood. Thus, numerous recent studies have tried to investigate these mechanisms through whole-genome sequencing (WGS) of penaeid shrimp (Yuan et al., 2018, 2021; Zhang et al., 2019; Uengwetwanit et al., 2021).

Due to their importance, decoding the genomes of these penaeid species has attracted global attention. As early as 1997, an international workshop on genome mapping of aquaculture animals was founded, aiming to construct complete genome maps of five economical important organisms, including penaeid shrimp, salmon, catfish, tilapia, and oyster (Alcivar-Warren et al., 1997). The genomes of the other four species have been published earlier before 2016 (Zhang et al., 2012; Berthelot et al., 2014; Lien et al., 2016; Liu et al., 2016). However, due to the high degree of genome complexity, penaeid shrimp has encountered extreme difficulties in genome sequencing and assembly, and the first penaeid shrimp genome was not completed until 2019 (Zhang et al., 2019). Nowadays, only three high-quality genomes of penaeid shrimp have been reported, namely, *L. vannamei*, *F. chinensis*, and *P. monodon* (Zhang et al., 2019; Uengwetwanit et al., 2021; Yuan et al., 2021). The draft assembly of *M. japonicus* is highly fragmented (Yuan et al., 2018). Even for the three high-quality genomes, the contig N50 lengths (<59 Kb) are significantly shorter than many newly published genomes (mostly > 1 Mb) (Shingate et al., 2020; Meyer et al., 2021) and genomes of many other crustaceans, e.g., *Eulimnadia texana* (18.07 Mb) (Baldwin-Brown et al., 2018) and *Portunus trituberculatus* (4.12 Mb) (Tang et al., 2020). Whereas the factors that cause poor assembly of the penaeid shrimp genome are still unclear, although their genomes are available.

In this study, we collected the genome sequencing data of four representative penaeid shrimp species, including *L. vannamei*, *F. chinensis*, *P. monodon*, and *M. japonicus*, and performed genome survey analyses to investigate their genomic characteristics. And then, we used various methods to conduct genome assembly using these sequencing data and tested how much data would be sufficient for the genome assembly. Based on this study, some clues may be provided for the future higher-quality genome assembly of penaeid shrimps.

MATERIALS AND METHODS

Genome Sequencing Data of Penaeid Shrimp

The Illumina paired-end sequencing data of four penaeid shrimp species (*L. vannamei*, *F. chinensis*, *P. monodon*, and *M. japonicus*) were collected from previous studies with the sequencing read length of 150 bp (PRJNA438564, PRJNA627295, PRJNA387410) (Yuan et al., 2018, 2021; Zhang et al., 2019). A total of 361.5 Gb data for *L. vannamei*, 160.9 Gb data for *F. chinensis*, 127.3 Gb data for *P. monodon*, and 127.5 Gb data for *M. japonicus* were collected. The raw sequencing data were trimmed to filter out low-quality data and adapter contaminants by using the NGS QC Toolkit with the parameters of "2 A -c 10" (Patel and Jain, 2012). The PacBio long-read sequencing data of *L. vannamei* and *F. chinensis* were collected from previous studies with the PacBio sequencing read N50 length of 11,205 and 9,813 bp, respectively (PRJNA438564 and PRJNA627295) (Zhang et al., 2019; Yuan et al., 2021). A total of 132.8 Gb PacBio data for *L. vannamei* and 160.3 Gb PacBio data for *F. chinensis* were collected. The final genome assembly sequences of *L. vannamei*, *F. chinensis*, and *P. monodon* were downloaded from the NCBI with the accession number of QCYY00000000, JABKCB000000000, and JABERT000000000, respectively (Zhang et al., 2019; Uengwetwanit et al., 2021; Yuan et al., 2021). These three genome assemblies were all assembled based on PacBio sequencing data. The contig N50 length are 57.65, 58.99, and 45.08 Kb and the scaffold N50 length are 31.300, 28.92, and 44.86 Mb for *L. vannamei*, *F. chinensis*, and *P. monodon*, respectively.

Genome Survey Analysis

In order to investigate genomic characteristics of penaeid shrimp, a K-mer (K represents the chosen length of substrings)-based genome survey was conducted to estimate the genome size and complexity. Based on the Illumina paired-end sequencing data, the K-mer frequency along the read was calculated (Li et al., 2010b). Jellyfish was used to calculate K-mer depth distribution (Marcais and Kingsford, 2011), which depends on the characteristic of the genome and follows a Poisson's distribution. Here, $K = 19$ was selected for the survey analysis.

An empirical formula, $G = N \times (L\text{-}K + 1)/(L \times M)$, was used to calculate the genome size (G), where N is the number of K-mers, L is the read length, K stands for the length of K-mer, and M stands for the observed peak of K-mer depth (Li et al., 2010a). The M values of the four shrimp species were calculated, namely, *L. vannamei*, $M = 37$; *F. chinensis*, $M = 66$; *P. monodon*, $M = 43$; and *M. japonicus*, $M = 47$. Besides, genome size can be determined using flow cytometry (approximately 1 pg = 978 Mb) (Dolezel et al., 2003). The genome size estimation results of penaeid shrimp species and other decapods were also downloaded from the Animal Genome Size Database[1]. The flow cytometry estimation of the four penaeid shrimp species were included, namely, *L. vannamei*, 2.50 pg; *F. chinensis*, 1.92 pg;

[1]www.genomesize.com

P. monodon, 2.53 pg; and *M. japonicus*, 2.83 pg. Combining the results above, the genome size of each penaeid shrimp species could be determined. Besides, the heterozygosity and repeat content of penaeid shrimp were estimated based on the K-mer depth distribution using GenomeScope 2.0[2].

Evaluation of Genome Duplication Events in *Litopenaeus vannamei*

To test whether penaeid shrimp has underwent whole-genome duplication, a series of analyses were carried out on the genome sequencing data of *L. vannamei*. Firstly, we sequenced the genome of a single sperm cell of *L. vannamei* and compared its K-mer depth distribution with WGS of somatic cells. Sperms were collected from spermatophore of a male *L. vannamei*. After continuous dilution, a single sperm cell was obtained by using a very thin straw under the microscope, and then the genomic DNA of the cell was subjected to PCR-based whole-genome amplification by the MALBAC® Single Cell WGA Kit (Yikon Genomics, Beijing, China). The amplified DNA fragments were directly used for sequencing on Illumina HiSeq2000 platform (Illumina, San Diego, CA, United States). A total of 1.60 Gb single sperm cell sequencing data were generated, and these data were deposited in NCBI SRA database with the accession number SRR13661692. Unlike somatic cells, single sperm cells are haploid and have low heterozygosity. Thus, the K-mer depth distribution of the single sperm cell sequencing will display some differences with that of WGS of somatic cells in the content of heterozygous K-mers. Jellyfish v2.2 was used to clarify all K-mers in single sperm cell genome sequencing, and the depth value of each K-mer was extracted from the K-mer depth distribution of WGS. The percentage of the K-mers in each depth was calculated to draw K-mer depth distribution plot of single sperm cell genome sequencing.

Next, according to previous studies (Berthelot et al., 2014; Xu et al., 2014), the plot of synonymous site divergence values (Ks) of paralogous genes was widely used to identify genome duplication events of *L. vannamei*. The homologous gene pairs were identified by using an all-to-all BLASTP comparison with *E*-value cutoff of 1E-07. The reciprocal best hit homologous gene pairs were selected to calculate Ks values using the CodeML program from the PAML package (Yang, 2007). The homologous pairs were aligned by MUSCLE (Edgar, 2004), and the well-aligned regions were extracted with Gblocks v0.91b (Talavera and Castresana, 2007).

In addition, the allele frequency distribution was calculated to identify genome duplication events (Pelin et al., 2015). All the Illumina sequencing reads were mapped to the *L. vannamei* genome using Burrows–Wheeler Aligner (BWA) (Li and Durbin, 2009), and all single-nucleotide polymorphisms (SNPs) were called by SAMTools-1.11 (Li et al., 2009). For each site of the SNP, the percentages of the four bases were calculated and sorted from most to least. Then, the allele frequency distribution was calculated based on these percentage values.

Repeat Annotation

The repeat annotation was performed on four genomes of penaeid shrimp species, *L. vannamei*, *F. chinensis*, *P. monodon*, and *M. japonicus*. Different from the other three species that assembled based on PacBio sequencing data, the *M. japonicus* genome was assembled based on Illumina sequencing data (Yuan et al., 2018). Both RepeatModeler v2.0[3] and RepeatMasker v4.1.0 were used for *de novo* identification of repeats. A local repeat database was constructed by RepeatModeler, and then, RepeatMasker was used to identify the transposable elements (TEs) by aligning the genome sequences against the local library and RepBase (RepBase21.04) with default parameters (Tarailo-Graovac and Chen, 2009). SciRoKo v3.4 was used to annotate simple sequence repeats (SSRs) in the three penaeid genomes (Kofler et al., 2007).

Genome Assembly and Comparison

Based on the Illumina sequencing data, the draft genomes of the four penaeid shrimp species (*L. vannamei*, *F. chinensis*, *P. monodon*, and *M. japonicus*) were assembled by SOAPdenovo2 with the *k* value set from 31 to 99 (Luo et al., 2012). Besides, the SOAPdenovo2 assembly was also performed on different amounts of sequencing data (genome coverage of 16 × to 135 ×) of *L. vannamei*.

Based on the PacBio sequencing data, various assembly approaches were used for the genome assembly of *L. vannamei* and *F. chinensis*, including FALCON v0.3.0 (Chin et al., 2016), HABOT2 (Zou et al., 2017), DBG2OLC (Ye et al., 2016), SMARTdenovo (Liu et al., 2020), and WTDBG2 (Ruan and Li, 2020). Due to the lack of raw PacBio sequencing data, genome assembly of *P. monodon* and *M. japonicus* have not been conducted using these assemblers. For FALCON assembly, the long sequencing subreads were firstly selected as the seed reads to be corrected by short subreads, and then, the error-corrected reads were assembled into contigs using FALCON with the parameters of "seed_coverage = 30, length_cutoff_pr = 1,000, length_cutoff = −1." For HABOT2 assembly, three main modules, namely, graph module, align module, and *Denovo* module, were used to get a hybrid assembly of the subreads with the parameters of "-k 17 -i 1 -m 3 -s 1." For DBG2OLC assembly, both long PacBio subreads and contigs obtained from a de Bruijn graph (DBG) assembly were used for genome assembly with the parameters of "k 17 MinOverlap 20 AdaptiveTh 0.01 Remove Chimera 1." The contigs are generated from SOAPdenovo assembly of Illumina sequencing data. For SMARTdenovo assembly, the raw PacBio sequencing subreads were directly used for the assembly follows the overlap-layout-consensus (OLC) paradigm with the parameters of "-c 1." For WTDBG2 assembly, the subreads were chopped into 1,024-bp segments, similar segments were merged into a vertex, and vertices were connected based on the segment adjacency on subreads. Since WTDBG2 had a better performance in penaeid shrimp genome assembly than the other four methods, it was used for the assembly of different amounts of PacBio sequencing data (genome coverage of 20× to 70×) of *L. vannamei*.

[2]http://qb.cshl.edu/genomescope/

[3]http://www.repeatmasker.org/RepeatModeler.html

Quality Assessment of Genome Assembly

To evaluate the quality of the genome assemblies of the penaeid shrimp species in this study and those published in previous studies (Zhang et al., 2019; Uengwetwanit et al., 2021; Yuan et al., 2021), several approaches were utilized to identify the completeness and accuracy of these assemblies. Firstly, Illumina sequencing reads were mapped back to the genome using Bowtie2 with the following parameters: –rdg 3,1 –rfg 3,1 –gbar 2, and the mapping rates were calculated (Langmead and Salzberg, 2012). Besides, according to previous study (Yuan et al., 2020), the unigenes that assembled from the transcriptome data were also mapped to the shrimp genomes using BLAT (Kent, 2002). The unigenes were downloaded from the shrimp gene database[4] with N50 lengths ranging from 1.40 to 2.34 Kb. In addition, BUSCO v4.0 tool suite was used to evaluate the quality of the genome assemblies by calculating the coverage of the eukaryotic single-copy core genes (BUSCOs, Eukaryota *odb9*) (Seppey et al., 2019).

Statistical Methods

The statistics for this study are conducted using Student's *t* test (between two groups) and one-way ANOVA (among three or more groups) using SPSS 22.0 software[5]. Significant differences are indicated when p value < 0.05.

RESULTS

Genome Survey of Penaeid Shrimp Species

In order to find the factors resulting in the poor assembly of the penaeid shrimp genomes, a comprehensive study of the general genomic characteristics, including genome size, heterozygosity, and repeat content, was conducted on these species. A K-mer-based genome survey was performed on the Illumina sequencing data of four representative penaeid shrimp species, *L. vannamei*, *F. chinensis*, *P. monodon*, and *M. japonicus*. Two peaks (Peak A and Peak B) were detected in the K-mer plot of all the four species (**Figure 1**), and the K-mer depth of Peak B was about twice of that for Peak A, e.g., K-mer depth of Peak A and Peak B were 37 and 74 in *L. vannamei*, respectively. Generally, according to previous studies (Zhang et al., 2012; Li et al., 2017; Shingate et al., 2020), Peak A represents the heterozygous single copy K-mers, while Peak B represents the homozygous single copy K-mers in the genome, which is also used for genome size estimation. However, the genome size estimated based on the K-mer depth of Peak B was half of that based on Peak A and also half of that estimated by flow cytometry methods (Alcivar-Warren et al., 1997; Zhang et al., 2019). Thus, it was confusing about which peak represents the homozygous single copy K-mers.

[4]http://www.genedatabase.cn/Decapoda.html

[5]https://www.ibm.com/analytics/spss-statistics-software

Genome Sizes of Penaeid Shrimp Species

According to the Animal Genome Size Database, the genome sizes of 145 decapods, covering 32 families, were recorded. Among them, the largest genome was 39.87 Gb (*Sclerocrangon ferox*) and the smallest genome was 1.04 Gb (*Carcinus maenas*). Genomes from Alpheidae (9.92 ± 4.99 Gb), Alvinocarididae (11.12 ± 2.04 Gb), Crangonidae (17.66 ± 12.73 Gb), and Palaemonidae (9.21 ± 4.98 Gb) have relatively larger sizes (**Figure 2A**). The genome sizes of Portunidae (1.86 ± 0.44 Gb, excluding *Necora puber*, as it has a singular genome size of 14.79 Gb), Penaeidae (2.51 ± 0.29 Gb), and Ocypodidae (2.45 ± 0.73 Gb) were smaller than those of other families.

The genome sizes of various penaeid shrimp species (Penaeidae) were around 2.5 Gb (**Figure 2B**). *Farfantepenaeus aztecus* has been identified to have the largest genome of 2.87 Gb, and the genome of *F. chinensis* was identified to be the smallest (1.87 Gb) among these penaeids. After combining the results of flow cytometry and K-mer analysis (Peak A), the genome sizes of the four penaeid shrimp species were estimated, namely, *L. vannamei*, 2.45 Gb; *F. chinensis*, 1.88 Gb; *P. monodon*, 2.66 Gb; and *M. japonicus*, 2.38 Gb.

Genome Duplication Evaluation

Generally, penaeid shrimp were considered diploid, which is supported by their karyotypes (Campos-Ramos, 1997; Mansouri et al., 2011). Here, we adopted a series of approaches to identify whether penaeid shrimp has undergone whole-genome duplication based on the genome data of *L. vannamei*.

Firstly, in order to identify which peak (Peak A or Peak B) represents homozygous K-mers, we sequenced the genome of a single sperm cell (haploid) of *L. vannamei* and compared the K-mer plot with that of WGS of somatic cells (diploid). Unexpectedly, the two peaks were also found in the K-mer plot of single cell sequencing, and Peak A highly fitted with that of WGS (**Figure 3A**). Besides, a lower trough was detected in front of Peak A, indicating the lower heterozygosity of the single sperm cell than somatic cells. Therefore, this result supported that Peak A represents the homozygous K-mers, and lots of genomic segments might be duplicated. As for the Peak B, the peak detected in the K-mer plot of single cell sequencing was higher than that of WGS, which may be due to the higher chance to be amplified and sequenced for duplicated sequences than homozygous sequences.

Ks analysis of the homologous gene pairs was also used to identify whole-genome duplication. Approximately 3,276 reciprocal best hit paralogous genes were identified in the *L. vannamei* genome, and the Ks values of these gene pairs were calculated. The spectrum of Ks showed L-shaped distribution with no obvious peak, indicating that the penaeid shrimp may not have undergone whole-genome duplication (**Figure 3B**).

Additionally, the allele frequency distribution of penaeid shrimp was calculated and compared with those of polyploid species (Pelin et al., 2015). After mapping the Illumina sequencing reads to the *L. vannamei* genome, a total of 4,185,110 SNPs were called. It was found that the allele frequency plot of

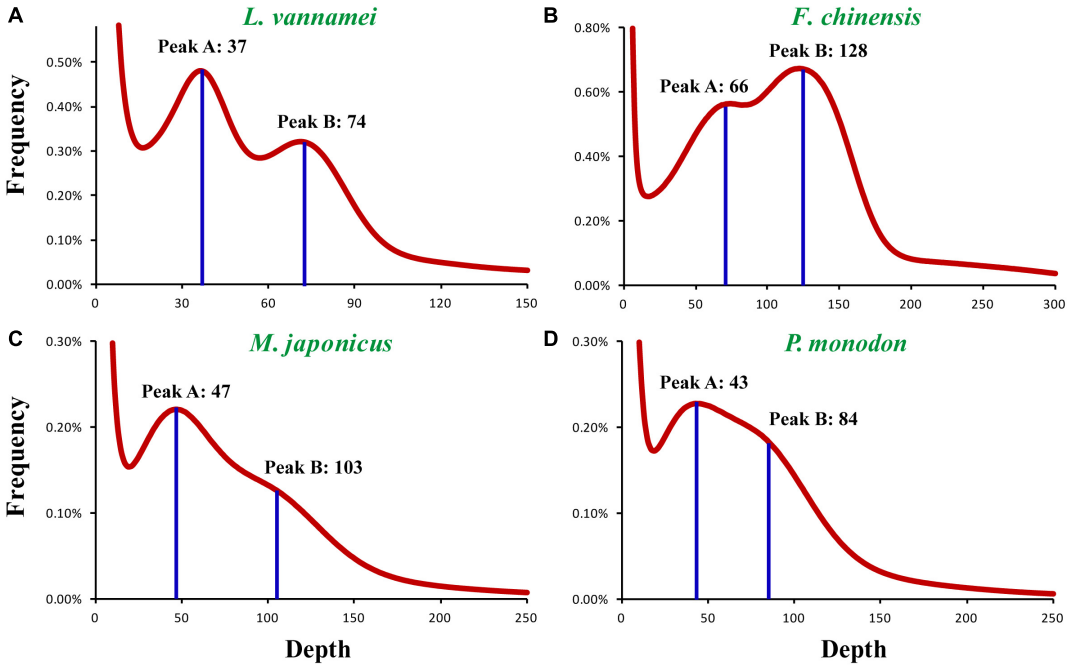

FIGURE 1 | K-mer distribution of four penaeid shrimp genomes. The K-mer distribution plots of the four shrimp species include **(A)** *L. vannamei*, **(B)** *F. chinensis*, **(C)** *M. japonicus*, and **(D)** *P. monodon*. *K* = 19 was selected for the K-mer frequency statistics.

all SNPs followed a unimodal distribution, with a peak at 0.5 (**Figure 3C**). The leftmost and rightmost truncated peaks may correspond to the variation between individuals in a population. According to previous study (Pelin et al., 2015), no peaks should be observed in haploids, unimodal distributions should be expected for diploids, and non-random trimodal distribution can be observed in polyploids. Unlike those of polyploidy genomes, peaks at 0.25 and 0.75 were not detected in the allele frequency plot of *L. vannamei*, but a unimodal distribution was found, suggesting that penaeid shrimp is diploid.

Overall, we speculated that penaeid shrimp was diploid without whole-genome duplication. Peak A in the K-mer plots represented the homozygous K-mers, thus, there might be a large amount of repetitive sequences in the penaeid shrimp genomes, and relatively low heterozygosity was expected. The heterozygosity of *L. vannamei*, *P. monodon*, and *M. japonicus* was estimated to be 0.26, 0.21, and 0.19% (model fit values ranged from 88.33 to 95.99%), respectively.

Repeats in the Penaeid Shrimp Genomes

The repeats were annotated in the genomes of four penaeid shrimp species that were published in previous studies, *L. vannamei*, *F. chinensis*, *P. monodon*, and *M. japonicus*. According to these studies (Yuan et al., 2018, 2021; Uengwetwanit et al., 2021), the first three genomes were assembled based on PacBio sequencing data, while the *M. japonicus* genome was assembled based on Illumina sequencing data, as no PacBio data were available. Repeats accounted for about 50% of the first three genomes, and the amount of TEs were varied among them that *L. vannamei* contained the least TEs (16.25%) and *P. monodon* contained the most TEs (22.01%) (**Table 1**). DNA transposons were highly expanded in the genomes of *L. vannamei* (9.33%) and *F. chinensis* (9.33%) compared to those in *P. monodon* (5.87%) and *M. japonicus* (5.66%) ($p < 0.05$). Whereas in the *P. monodon* genome, long interspersed nuclear elements (LINEs) were the most abundant TEs (9.26%) that was significantly higher than those in *L. vannamei* (2.82%), *F. chinensis* (3.27%), and *M. japonicus* (4.75%) ($p < 0.05$). Besides, short interspersed nuclear elements (SINEs) and long terminal repeats (LTRs) were also abundant in *P. monodon*.

Besides TEs, SSRs were also abundant in the first three shrimp genomes, *L. vannamei*, 23.93%; *F. chinensis*, 19.50%; and *P. monodon*, 15.01%), which have been identified as the most abundant among the species whose genomes are available (Zhang et al., 2019). As assembled based on Illumina sequencing data, the SSR content was possibly underestimated in the *M. japonicus* genome (9.79%). Similar results have been identified in the Illumina sequencing data assembly of *L. vannamei* (10.33%), *F. chinensis* (9.03%), and *P. monodon* (10.90%). When comparing with other decapods, the content of SSRs of penaeid shrimp were more abundant ($p < 0.05$), with significantly higher density (2,693–3,449 per Mb) and similar length distribution (56.54–72.21 bp in average) (**Figure 4A**). SSRs were densely distributed in the penaeid shrimp genomes, and thus, a large amount of compound SSRs, which are composed of different types of SSRs that linked head to tail, have been identified in these genomes. Among the total SSRs, approximately 60% of them were identified to form compound SSRs, which were significantly higher than those in many other crustaceans (<24%; **Supplementary Figure 1**). Besides, the lengths of compound

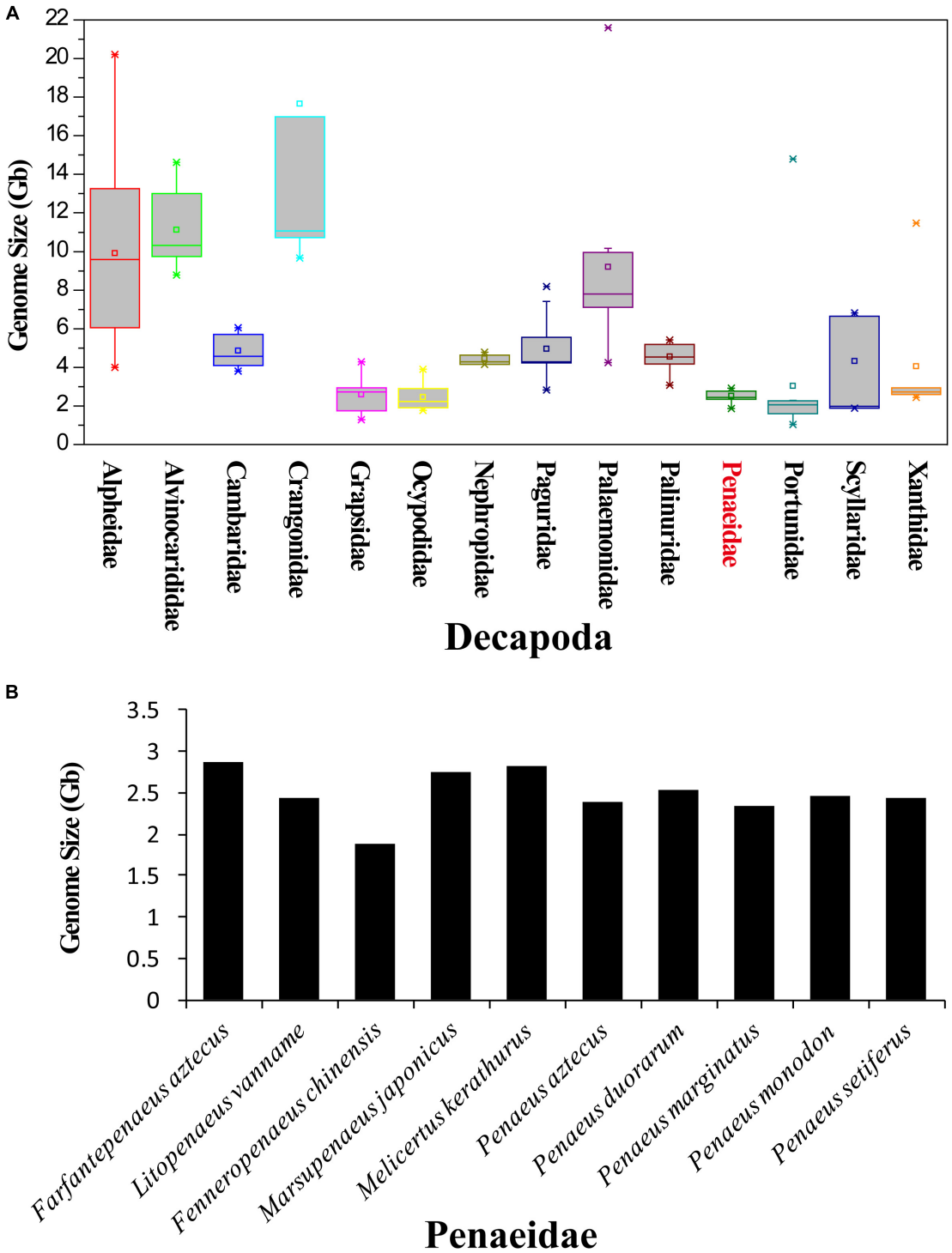

FIGURE 2 | Genome size of decapods and various penaeid shrimp species. **(A)** Genome sizes of various families of Decapoda. The information of genome sizes was obtained from the Animal Genome Size Database (www.genomesize.com). **(B)** Genome sizes of various penaeid shrimp species.

SSRs were significantly longer than those of single SSRs ($p < 0.05$) (**Figure 4B**).

Except for $(GC)_n$, dinucleotide SSRs [$(AT)_n$, $(AC)_n$, $(AG)_n$] were the most abundant SSRs in the penaeid shrimp genomes, which accounted for more than 73% of total SSRs (**Figure 4C**). The SSR compositions were quite similar among the three shrimp species, whereas some variations were also observed. *L. vannamei* had significantly higher amounts of $(AT)_n$ and $(AACCT)_n$ than

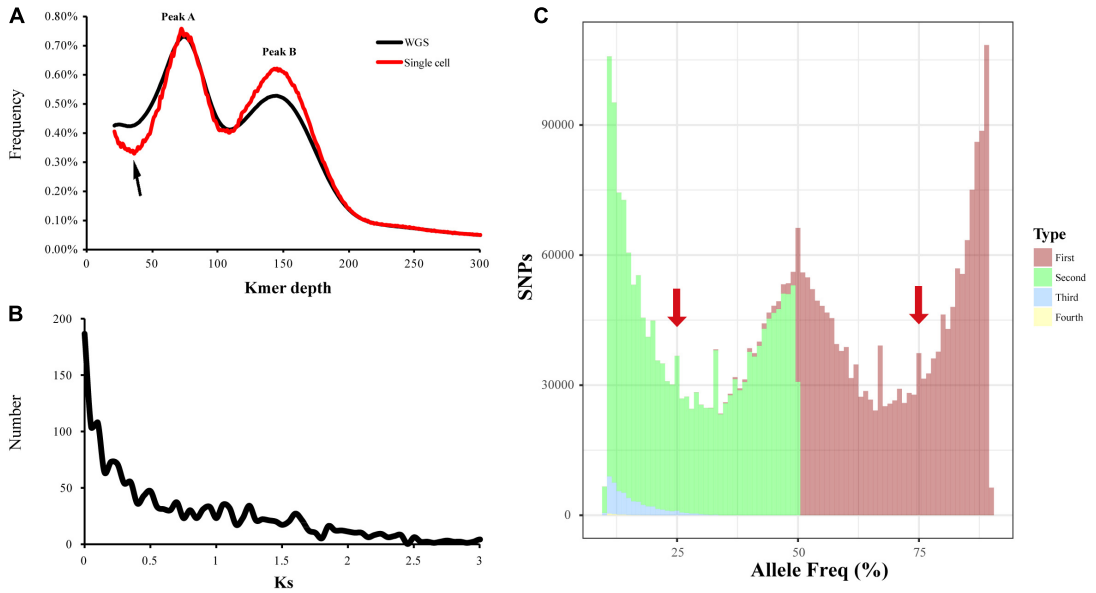

FIGURE 3 | Evaluation of genome duplication of *L. vannamei*. **(A)** K-mer distribution of the single sperm cell and whole-genome sequencing (WGS) data. **(B)** Ks frequency distribution for pairs of paralogous genes in the *Litopenaeus vannamei* genome. **(C)** Allele frequency spectra based on read counts of single-nucleotide polymorphisms (SNPs). For each SNP site, the appearance frequencies of the four bases were sorted from most (Type: First) to least (Type: Fourth). Leftmost and rightmost truncated peaks likely correspond to variations between individuals in a population. For polyploidy genome, the peaks at 25, 50, and 75% would be expected.

TABLE 1 | Summary of repetitive sequences in four penaeid shrimp genomes.

Repeats	*L. vannamei*	*F. chinensis*	*P. monodon*	*M. japonicus**
Genome length	1.66 Gb	1.57 Gb	2.39 Gb	1.79 Gb
Total repeats	49.39%	48.58%	42.83%	34.96%
DNA	9.33%	13.00%	5.87%	5.66%
LINE	2.82%	3.27%	9.26%	4.75%
SINE	0.06%	0.11%	1.30%	0.03%
LTR	0.62%	0.53%	1.42%	1.14%
Unknown	3.42%	3.52%	4.16%	7.19%
Satellite	0.10%	0.16%	0.00%	0.35%
Simple repeats	23.93%	19.50%	15.01%	9.79%
Low complexity	9.49%	8.49%	5.81%	6.28%

**The repeat annotation of Marsupenaeus japonicus was based on the assembly of Illumina sequencing data, and the annotation of the other three species were based on the genome assemblies published in previous studies (Zhang et al., 2019; Uengwetwanit et al., 2021; Yuan et al., 2021). LINE, long interspersed nuclear element; LTR, long terminal repeat; SINE, short interspersed nuclear element.*

those of the other two species, while *F. chinensis* and *P. monodon* had significantly higher amounts of $(AG)_n$, $(AAT)_n$, $(ATAC)_n$, and $(ACAG)_n$ than those of *L. vannamei*.

Genome Assembly of Penaeid Shrimp Species

Based on various sequencing data of penaeid shrimp, various genome assembly strategies have been carried out on these shrimp species. Firstly, based on the Illumina sequencing data, SOAPdenovo assembly was performed on the four penaeid shrimp species, *L. vannamei*, *F. chinensis*, *P. monodon*, and *M. japonicus*. However, these assemblies were rather poor in quality, similar to many previous studies (Yu et al., 2015; Yuan et al., 2018). The contig N50 lengths ranged from 301 bp (*P. monodon*) to 514 bp (*L. vannamei*) (**Supplementary Table 1**), which indicated that these assemblies were highly fragmented. Besides, after extending the contigs by filling gaps in scaffolds that assembled based on large insert sequencing libraries (insert size of 2, 5, and 10 Kb and read length of 100 bp, PRJNA438564), the contig N50 length could only reach 2.8 Kb in *L. vannamei* (**Table 2**). In addition, we performed genome assembly based on various amounts of sequencing data (16 ×–135 ×). When the sequencing depth reached 80 ×, the genome assembly size and N50 length tend to be stable (**Figure 5A**), and the assembly showed high completeness that covered more than 91% of the transcriptome unigenes. It indicated that the Illumina sequencing data are sufficient for assembly, while the poor assembly might be caused by the high complexity of the genome and/or the limitation of a short sequencing read length (150 bp).

As for the assembly of PacBio sequencing data, various amounts of the data (20 ×–70 ×) of *L. vannamei* were used to test the adequacy of the data first. When the sequencing data coverage reached 50 ×, the total length of the assembly tended to be stable, but when the sequencing coverage reached 70 ×, the N50 length became shorter (**Figure 5B**). However, the assembly of 70 × data showed higher completeness (94.45%) than that of 60 × (89.65%) and 50 × data (82.59%). Thus, 70 × PacBio sequencing data are sufficient for the genome assembly. As we only collected PacBio sequencing data of *L. vannamei* and *F. chinensis* in our previous studies (Zhang et al., 2019; Yuan et al., 2021), we only performed genome assembly of these two species herein. Based on the

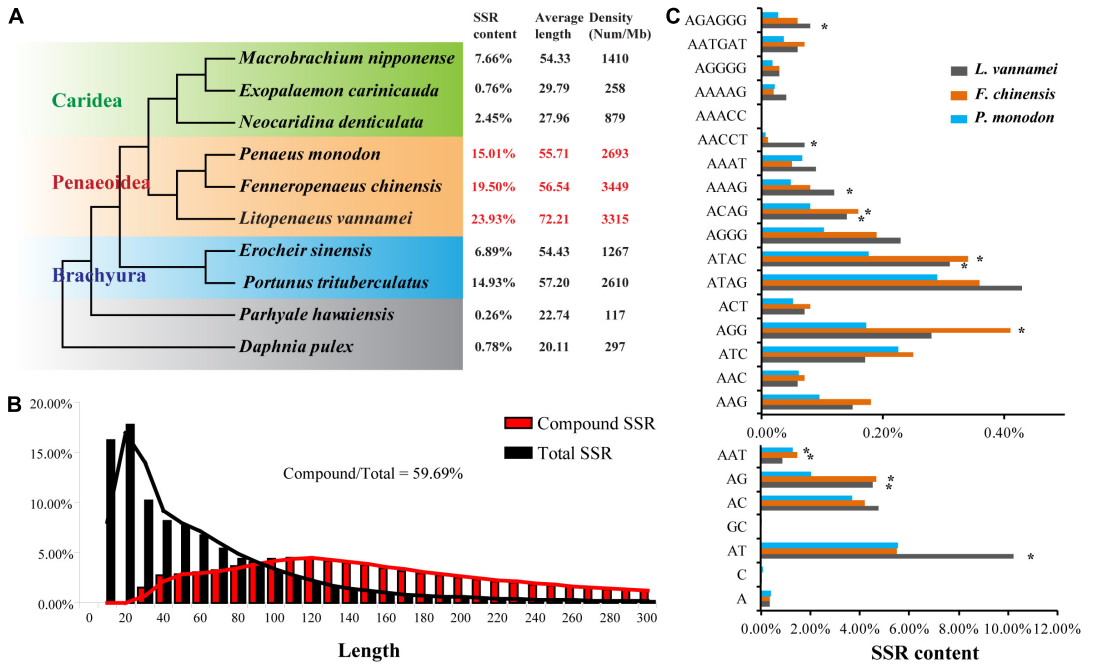

FIGURE 4 | Simple sequence repeats (SSRs) in penaeid shrimp genomes. **(A)** Content, average length, and density of SSRs in the genomes of various crustaceans. The phylogenetic tree was referred from a previous study (Yuan et al., 2018). The characteristics of SSRs in various crustaceans were referred from the study of Yuan et al. (2021). Since the SSR content may be underestimated in the *Marsupenaeus japonicus* genome, it was not shown and compared with the other three shrimp genomes. **(B)** The length distribution of compound SSRs and total SSR in the *Litopenaeus vannamei* genome. The compound SSR indicates numerous different types of SSRs connected head to tail. **(C)** Comparisons of the distributions of various types of SSRs in three penaeid shrimp genomes. A part of SSR types with relatively high content in the genome was shown in the plot.

TABLE 2 | Statistics of genome assembly of *Litopenaeus vannamei* using different methods.

	SOAPdenovo[§]	FALCON	HABOT2	DBG2OLC	SMARTdenovo	WTDBG2
Contig number	982,421	463,151	110,906	43,938	60,355	50,304
Total length (Gb)	1.35	1.59	1.69	1.30	1.78	1.62
Longest (Kb)	1,219	1,219	214	707	422	739
N50 (bp)	2,826	9,496	25,477	43,564	34,826	57,650
N90 (bp)	712	1,271	9,552	13,276	15,383	14,641
Unigene coverage	95.76%	89.56%	93.73%	83.16%	68.53%	94.45%
Unigene coverage (50%)*	85.50%	78.33%	84.85%	71.24%	50.12%	86.91%

[§] *SOAPdenovo genome assembly was conducted based on the Illumina sequencing data.* * *Unigene coverage (50%) indicates more than 50% of a unigene sequence covered by a single scaffold.*

total PacBio sequencing data, we assembled the *L. vannamei* and *F. chinensis* genomes using five methods, namely, FALCON, HABOT2, DBG2OLC, SMARTdenovo, and WTDBG2. Except for DBG2OLC, the total length of the assemblies by the other four methods was about 1.6 Gb in *L. vannamei* (**Table 2**). Different from the assembly based on Illumina sequencing data, the contigs assembled based on PacBio sequencing data showed significantly higher continuity ($p < 0.05$). In the two shrimp species, the contig N50 length was at least three times longer than the SOAPdenovo assembly (**Table 2**; **Supplementary Table 2**). The N50 length of the WTDBG2 assembly even reached 57,650 bp in *L. vannamei* and 58,996 bp in *F. chinensis*, which was more than 20 times longer than that in the SOAPdenovo assembly. Besides, WTDBG2 assembly not only has higher continuity than other methods but also has higher completeness (covering more than 94% of unigenes).

Although the assembly of PacBio sequencing data has higher continuity than that of SOAPdenovo assembly, it was still highly fragmented, as it was composed of more than 40,000 contigs, and the contig N50 lengths were significantly shorter than many recently published crustacean genomes, e.g., *Eulimnadia texana* (18.07 Mb) (Baldwin-Brown et al., 2018), *P. trituberculatus* (4.12 Mb) (Tang et al., 2020), and *Paralithodes platypus* (147.47 Kb) (Tang et al., 2021). Thus, we next investigated the factors that caused the high fragmentation of these assemblies. As for SOAPdenovo assembly, we mapped the contigs and Illumina sequencing reads on a complete bacterial artificial chromosome (BAC) (SHE003C23), which was previously sequenced by the

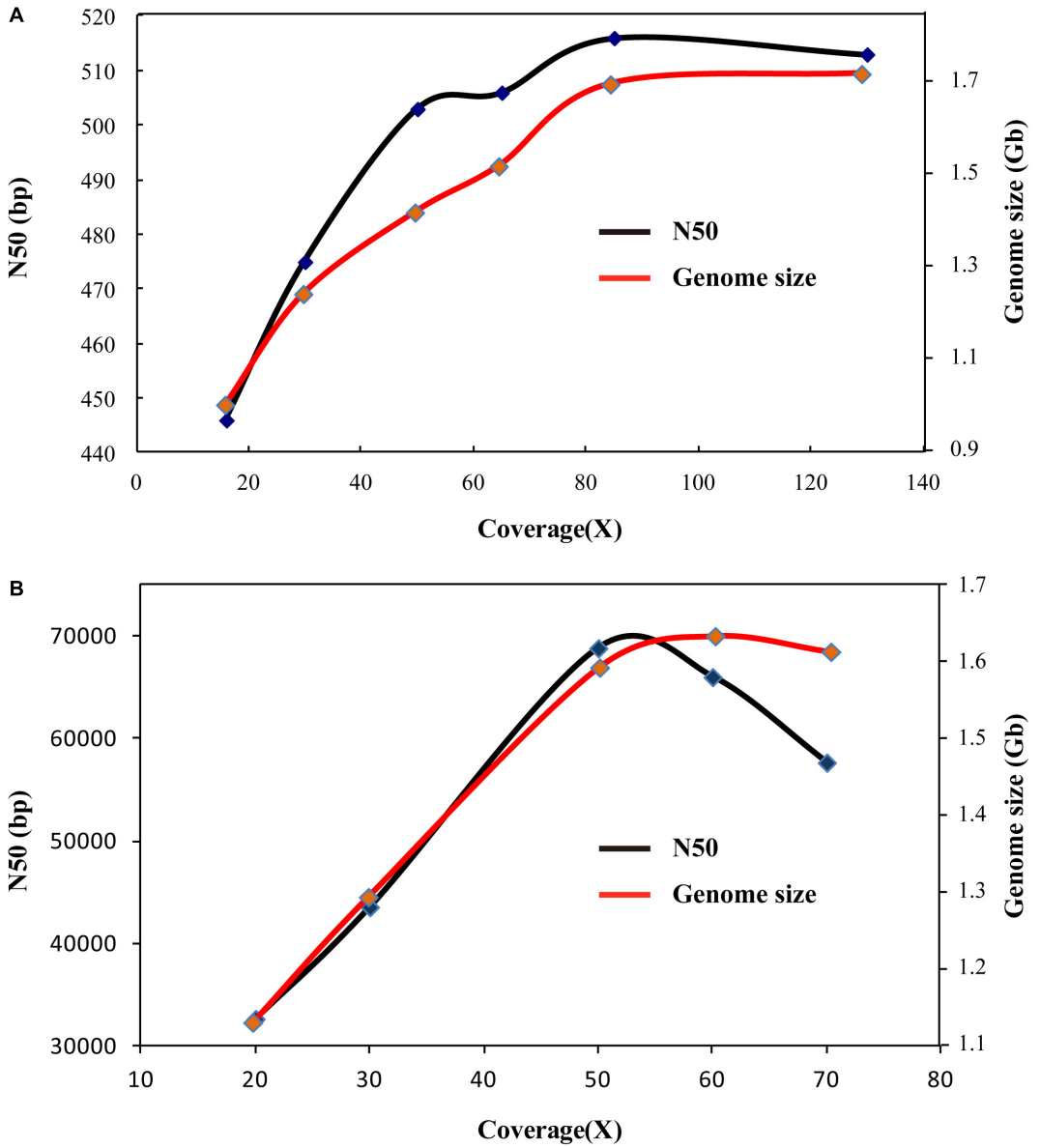

FIGURE 5 | Genome assembly using different coverages of sequencing data. **(A)** Genome assembly comparison using different coverages of Illumina sequencing data. **(B)** Genome assembly comparison using different coverages of PacBio sequencing data.

Sanger sequencing platform (Zhang et al., 2010, 2019). Low coverage of Illumina sequencing data was found in many regions, which was consistent with the lack of contigs in these regions (**Supplementary Figure 2**). When analyzing these low coverage regions, it was found that they were mainly composed of SSRs. Especially for the regions with extremely long single or compound SSRs, almost no sequencing reads were distributed in these regions. The read coverage of SSR regions (20.42 ×) was significantly lower than those of TE (190.32 ×) and other regions (202.98 ×) ($p < 0.05$) (**Supplementary Figure 3**).

As for the PacBio data assembly, we also mapped the assembly contigs to the sequenced BACs to find the factors that result in the assembly fragmentation. However, these BACs were aligned to the contigs in full length (Zhang et al., 2019), and thus the characteristics of the edges of these contigs could not be identified.

DISCUSSION

The study of the penaeid shrimp genome is attractive globally due to its high economic and biological values. Although several penaeid shrimp genomes have been published (Zhang et al., 2019; Uengwetwanit et al., 2021; Yuan et al., 2021), the factors that cause genome assembly difficulties and poor assembly quality are still ambiguous. In this study, two aspects have been identified

to be the potential causes for these problems. The first one was the high percentage of homo-duplicated repeats or high heterozygosity. Two peaks were identified in the K-mer depth distribution plots of all the four penaeid shrimp species, which were similar to those genomes with high heterozygosity, e.g., the Pacific oyster *Crassostrea gigas* (Zhang et al., 2012) and the Zhikong scallop *Chlamys farreri* (Li et al., 2017). Even for the genomes that underwent whole-genome duplication, e.g., the horseshoe crab *Limulus polyphemus* (Nossa et al., 2014) and the pineapple *Ananas comosus* (Ming et al., 2015), the former peak of the two peaks in the K-mer plots also represented heterozygous K-mers. If Peak A represents heterozygous K-mers, penaeid shrimp will have a high degree of heterozygosity that was estimated to be 2.43% in *L. vannamei*, 1.95% in *F. chinensis*, 4.95% in *P. monodon*, and 4.49% in *M. japonicus*. However, the results of genome size estimation and K-mer depth distribution of single sperm cell sequencing supported Peak A that represents homozygous K-mers, while Peak B represents homo-duplicated K-mers. Whereas no signature of whole-genome duplication has been identified in the penaeid shrimp genomes through Ks and allele frequency analyses. Furthermore, a single Hox gene cluster was identified in the penaeid shrimp genomes (Yuan et al., 2018; Zhang et al., 2019; Uengwetwanit et al., 2021), which also did not support the whole-genome duplication event. No matter what Peak A represents, the high heterozygosity and homo-duplication both will be responsible for a large number of polymorphic sites in genome sequencing, which will make the genome assembly very difficult.

The abundant SSRs in the penaeid shrimp genome appear to be the second aspect resulting in the poor assembly. In most sequenced species, SSRs only account for ∼1% of the genome (Oliveira et al., 2006; Zhang et al., 2019), whereas the penaeid shrimp genome is particularly notable for having the highest proportion of SSRs (>15%) among sequenced animal genomes up to now. Low coverage of Illumina sequencing data was detected around the SSR regions, which makes the SOAPdenovo assembly highly fragmented. Thus, no matter how much Illumina data are sequenced, the contig N50 lengths of these penaeid shrimp species were very short due to the assembly blocks at the edges of the SSR regions. Even though the third-generation sequencing could cover most of the SSR regions, the large number of SSRs also brings great difficulties to genome assembly. SSRs could be linked head to tail to form a compound SSR, which is much longer than a single SSR. The extremely long compound SSRs will also result in the blocks of the assembly based on PacBio sequencing data. Besides, the OLC paradigm and the DBGs are two major algorithms that are widely used in many genome assembly methods (Ruan and Li, 2020). Both algorithms need to perform sequence mapping and selecting the best hits for the assembly, whereas the simple composition of SSRs will make these processes more difficult. Therefore, even if the PacBio sequencing data are sufficient or excessive for genome assembly of the penaeid shrimp, the contig N50 length has not increased in expectation. Although it is still unclear what results in the fragmentation of the assemblies based on PacBio sequencing data, the sequences in the gaps between contigs will be more complex than we thought. And there may be many other potential factors affecting the genome assembly of penaeid shrimp, which need further investigation.

Before the development of PacBio sequencing technology, Illumina sequencing was widely used for most genome assemblies. As expected, the performance of the Illumina data assembly was worse than that of PacBio data in penaeid shrimp species. However, Illumina sequencing is still used for genome assembly in recent years (Li et al., 2017; Leclere et al., 2019) and also widely used for whole-genome resequencing nowadays. Since the development of the third-generation sequencing technology, many methods for genome assembly have been developed. Finding an effective method to assemble the target genome assembly is undoubtedly important. FALCON has been widely and firstly selected for the genome assembly of most species, whereas it seems unsuitable for the penaeid shrimp genome assembly because of its poor assembly results and extraordinarily long time for the error correction before the assembly. Similar assembly results were obtained through using HABOT2, DBG2OLC, SMARTdenovo, and WTDBG2, but the assembly quality of WTDBG2 was the highest. Thus, WTDBG2 was ultimately used for the genome assembly of the three penaeid shrimp species *L. vannamei*, *F. chinensis*, and *P. monodon* (Zhang et al., 2019; Uengwetwanit et al., 2021; Yuan et al., 2021). The final assembly of *L. vannamei* and *F. chinensis* was similar (Zhang et al., 2019; Yuan et al., 2021). The length of contig N50 was about 58 Kb, which was also similar to that of *P. monodon* (45 Kb) (Uengwetwanit et al., 2021). These three genomes showed high completeness that the coverages of unigenes, Illumina sequencing reads, and BUSCOs were all higher than 91% (**Supplementary Table 3**). Besides, in order to assemble genome into chromosomal level, Hi-C data were used for scaffolding the contigs of these shrimp species (2n = 88 chromosomes). Finally, these contigs were anchored onto 44 chromosomes, and their scaffold N50 lengths ranged from 30 to 45 Mb.

Besides the PacBio continuous long-read (CLR) sequencing, there are many other long-read sequencing technologies, such as Oxford Nanopore Technologies (ONT) (Feng et al., 2015). The ONT sequencing can generate long reads, with an average length of more than 40 Kb, which is 2–4 times longer than that of PacBio sequencing (Lang et al., 2020). The assembly based on longer sequencing reads will assemble longer contigs; thus, we have tried to conduct ONT sequencing on *L. vannamei*. However, due to the limitation of data generation and short sequencing reads, the ONT sequencing of penaeid shrimp failed. There are many other assembly methods that were not used herein, and they may also be suitable for genome assembly of penaeid shrimp species. For chromosomal-level assembly, besides Hi-C sequencing, Bionano genome mapping also supports individual chromosome physical mapping and assembly in complex genomes (Stankova et al., 2016). Further research on the strategies of genome sequencing and assembly will aid the construction of high-quality genomes of penaeid shrimp. Furthermore, with the development of new sequencing technologies and assembly methods, higher-quality genome assemblies of peaneid shrimp species can be obtained in the future. This study can provide some clues for the future genome assembly of penaeid shrimp species.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

JX and FL conceived and designed the study. JY conducted the genome assembly and bioinformatics analyses. XZ performed genome sequencing. JY wrote the manuscript. XZ revised the manuscript. All authors read and approved the final manuscript.

FUNDING

We acknowledge financial support from the National Natural Science Foundation of China (41876167 and 31830100), the National Key Research & Development Program of China (2018YFD0900103 and 2018YFD0900404), grants from Qingdao National Laboratory for Marine Science and Technology (MS2017NO04), and the China Agriculture Research system-48 (CARS-48).

ACKNOWLEDGMENTS

We acknowledge the support from High Performance Computing Center, Institute of Oceanology, CAS.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.658619/full#supplementary-material

Supplementary Figure 1 | Compound SSR length distribution in four crustaceans.

Supplementary Figure 2 | The coverage of Illumina sequencing reads and contigs in a complete BAC (SHE003C23).

Supplementary Figure 3 | The comparison of read coverages of various genomic regions.

Supplementary Table 1 | SOAPdenovo assembly of four penaeid shrimp species.

Supplementary Table 2 | Statistics of genome assembly of *F. chinensis* using different methods.

Supplementary Table 3 | Statistics of genome assembly of four penaeid shrimp species.

REFERENCES

Alcivar-Warren, A., Dunham, R., Gaffney, P., Kocher, T., and Thorgaard, G. (1997). First aquaculture species genome mapping workshop. *An. Genet.* 28, 451–452. doi: 10.1111/j.1365-2052.1997.00202.x

Baldwin-Brown, J. G., Weeks, S. C., and Long, A. D. (2018). A new standard for crustacean genomes: the highly contiguous, annotated genome assembly of the clam shrimp eulimnadia texana reveals HOX gene order and identifies the sex chromosome. *Genome Biol. Evolu.* 10, 143–156. doi: 10.1093/gbe/evx280

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noel, B., et al. (2014). The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat. Commun.* 5:3657.

Campos-Ramos, R. (1997). Chromosome studies on the marine shrimps *Penaeus vannamei* and P-californiensis (Decapoda). *J. Crustacean Biol.* 17, 666–673. doi: 10.2307/1549369

Chin, C. S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* 13, 1050–1054. doi: 10.1038/nmeth.4035

Dall, W., Hill, B., Rothlisberg, P., and Sharples, D. (1990). *The Biology of the Penaeidae*, Vol. 27. French: Bailliere, Tindall & Cox, 1–461.

Dolezel, J., Bartos, J., Voglmayr, H., and Greilhuber, J. (2003). Nuclear DNA content and genome size of trout and human. *Cytometry A* 51, 127–128; authorrely 129.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Fan, S., Hsu, K., Chen, F., and Ho, B. (1961). On the high conduction velocity of the giant nerve fiber of shrimp *Penaeus orientalis*. *Chin. Sci. Bull.* 12, 51–52.

FAO (2020). *Global Aquaculture Production 1950–2020*. Available online at: http://www.fao.org/fishery/statistics/global-aquaculture-production/query/en (accessed at October 20, 2020).

Farfante, I. P., and Kensley, B. (1997). *Penaeoid and Sergestoid Shrimps and Prawns of the World. Keys and Diagnoses for the Families and Genera*. Paris: Memories du Museum National D'Historie Naturelle.

Feng, Y., Zhang, Y., Ying, C., Wang, D., and Du, C. (2015). Nanopore-based fourth-generation DNA sequencing technology. *Genom. Prot. Bioinform.* 13, 4–16. doi: 10.1016/j.gpb.2015.01.009

Godin, D. M., Carr, W. H., Hagino, G., Segura, F., Sweeney, J. N., and Blankenship, L. (1996). Evaluation of a fluorescent elastomer internal tag in juvenile and adult shrimp *Penaeus vannamei*. *Aquaculture* 139, 243–248. doi: 10.1016/0044-8486(95)01174-9

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res.* 12, 656–664. doi: 10.1101/gr.229202.

Kofler, R., Schlotterer, C., and Lelley, T. (2007). SciRoKo: a new tool for whole genome microsatellite search and investigation. *Bioinformatics* 23, 1683–1685. doi: 10.1093/bioinformatics/btm157

Koyama, T., Asakawa, S., Katagiri, T., Shimizu, A., Fagutao, F. F., Mavichak, R., et al. (2010). Hyper-expansion of large DNA segments in the genome of kuruma shrimp, *Marsupenaeus japonicus*. *BMC Genomics* 11:141. doi: 10.1186/1471-2164-11-141

Lang, D., Zhang, S., Ren, P., Liang, F., Sun, Z., Meng, G., et al. (2020). Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific biosciences sequel II system and ultralong reads of Oxford Nanopore. *Gigascience* 9, 1–7.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Leclere, L., Horin, C., Chevalier, S., Lapebie, P., Dru, P., Peron, S., et al. (2019). The genome of the jellyfish clytia hemisphaerica and the evolution of the cnidarian life-cycle. *Nat. Ecol. Evol.* 3, 801–810. doi: 10.1038/s41559-019-0833-2

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., et al. (2010a). The sequence and de novo assembly of the giant panda genome. *Nature* 463, 311–317.

Li, R. Q., Fan, W., Tian, G., Zhu, H. M., He, L., Cai, J., et al. (2010b). The sequence and de novo assembly of the giant panda genome. *Nature* 463, 311–317.

Li, Y., Sun, X., Hu, X., Xun, X., Zhang, J., Guo, X., et al. (2017). Scallop genome reveals molecular adaptations to semi-sessile life and neurotoxins. *Nat. Commun.* 8:1721.

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., et al. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature* 533, 200–205.

Liu, H., Wu, S., Li, A., and Ruan, J. (2020). SMART denovo: a de novo assembler using long noisy reads. *[Preprints]* doi: 10.20944/preprints202009.200207.v202001

Liu, Z. J., Liu, S. K., Yao, J., Bao, L. S., Zhang, J. R., Li, Y., et al. (2016). The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. *Nat. Commun.* 7:11757.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAP denovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1:18.

Mansouri, S. M., Farahmand, H., and Khalilabadi, F. (2011). Chromosome studies on the marine shrimp *Penaeus* (fenneropenaeus) merguiensis from the Persian Gulf. *Iran. J. Fish. Sci.* 10, 734–741.

Marcais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Meyer, A., Schloissnig, S., Franchini, P., Du, K., Woltering, J. M., Irisarri, I., et al. (2021). Giant lungfish genome elucidates the conquest of land by vertebrates. *Nature* 590, 284–289. doi: 10.1038/s41586-021-03198-8

Ming, R., Vanburen, R., Wai, C. M., Tang, H., Schatz, M. C., Bowers, J. E., et al. (2015). The pineapple genome and the evolution of CAM photosynthesis. *Nat. Genet.* 47, 1435–1442.

Nossa, C. W., Havlak, P., Yue, J. X., Lv, J., Vincent, K. Y., Brockmann, H. J., et al. (2014). Joint assembly and genetic mapping of the Atlantic horseshoe crab genome reveals ancient whole genome duplication. *Gigascience* 3:9.

Oliveira, E. J., Padua, J. G., Zucchi, M. I., Vencovsky, R., and Vieira, M. L. C. (2006). Origin, evolution and genome distribution of microsatellites. *Genet. Mol. Biol.* 29, 294–307. doi: 10.1590/s1415-47572006000200018

Patel, R. K., and Jain, M. (2012). NGS QC toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* 7:e30619. doi: 10.1371/journal.pone.0030619

Pelin, A., Selman, M., Aris-Brosou, S., Farinelli, L., and Corradi, N. (2015). Genome analyses suggest the presence of polyploidy and recent human-driven expansions in eight global populations of the honeybee pathogen *Nosema ceranae*. *Environ. Microbiol.* 17, 4443–4458. doi: 10.1111/1462-2920.12883

Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg 2. *Nat. Methods* 17, 155–158. doi: 10.1038/s41592-019-0669-3

Seppey, M., Manni, M., and Zdobnov, E. M. (2019). BUSCO: assessing genome assembly and annotation completeness. *Methods Mol. Biol.* 1962, 227–245. doi: 10.1007/978-1-4939-9173-0_14

Shingate, P., Ravi, V., Prasad, A., Tay, B. H., Garg, K. M., Chattopadhyay, B., et al. (2020). Chromosome-level assembly of the horseshoe crab genome provides insights into its genome evolution. *Nat. Commun.* 11:2322.

Stankova, H., Hastie, A. R., Chan, S., Vrana, J., Tulpova, Z., Kubalakova, M., et al. (2016). Bio nano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. *Plant Biotechnol. J.* 14, 1523–1531. doi: 10.1111/pbi.12513

Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* 56, 564–577. doi: 10.1080/10635150701472164

Tang, B., Wang, Z., Liu, Q., Ren, Y., Guo, H., Qi, T., et al. (2021). Chromosome-level genome assembly of Paralithodes platypus provides insights into evolution and adaptation of king crabs. *Mol. Ecol. Res.* 21, 511–525. doi: 10.1111/1755-0998.13266

Tang, B., Zhang, D., Li, H., Jiang, S., Zhang, H., Xuan, F., et al. (2020). Chromosome-level genome assembly reveals the unique genome evolution of the swimming crab (*Portunus trituberculatus*). *Gigascience* 9:giz161.

Tarailo-Graovac, M., and Chen, N. (2009). Using repeatmasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* Chapter 4, Unit 4.10, 1–14.

Uengwetwanit, T., Pootakham, W., Nookaew, I., Sonthirod, C., Angthong, P., Sittikankaew, K., et al. (2021). A chromosome-level assembly of the black tiger shrimp (*Penaeus monodon*) genome facilitates the identification of growth-associated genes. *Mol. Ecol. Res.* doi: 10.1111/1755-0998.13357 [Epub ahead of print].

Wilson, K., Cahill, V., Ballment, E., and Benzie, J. (2000). The complete sequence of the mitochondrial genome of the crustacean *Penaeus monodon*: are malacostracan crustaceans more closely related to insects than to branchiopods? *Mol. Biol. Evol.* 17, 863–874. doi: 10.1093/oxfordjournals.molbev.a026366

Xu, P., Zhang, X., Wang, X., Li, J., Liu, G., Kuang, Y., et al. (2014). Genome sequence and genetic diversity of the common carp. Cyprinus carpio. *Nat. Genet.* 46, 1212–1219.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Ye, C., Hill, C. M., Wu, S., Ruan, J., and Ma, Z. S. (2016). DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci. Rep.* 6:31900.

Yu, Y., Zhang, X. J., Yuan, J. B., Li, F. H., Chen, X. H., Zhao, Y. Z., et al. (2015). Genome survey and high-density genetic map construction provide genomic and genetic resources for the pacific white shrimp *Litopenaeus vannamei*. *Sci. Rep.* 5:15612.

Yuan, J., Zhang, X., Gao, Y., Liu, C., Xiang, J., and Li, F. (2020). Adaptation and molecular evidence for convergence in decapod crustaceans from deep-sea hydrothermal vent environments. *Mol. Ecol.* 29, 3954–3969. doi: 10.1111/mec.15610

Yuan, J., Zhang, X., Liu, C., Yu, Y., Wei, J., Lia, F., et al. (2018). Genomic resources and comparative analyses of two economical penaeid shrimp species, *Marsupenaeus japonicus* and *Penaeus monodon*. *Mar. Genom.* 39, 22–25. doi: 10.1016/j.margen.2017.12.006

Yuan, J., Zhang, X., Wang, M., Sun, Y., Liu, C., Li, S., et al. (2021). Simple sequence repeats drive genome plasticity and promote adaptive evolution in penaeid shrimp. *Commun. Biol.* 4:186. doi: 10.1038/s42003-42021-01716-y

Zhang, G. F., Fang, X. D., Guo, X. M., Li, L., Luo, R. B., Xu, F., et al. (2012). The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* 490, 49–54.

Zhang, X., Zhang, Y., Scheuring, C., Zhang, H. B., Huan, P., Wang, B., et al. (2010). Construction and characterization of a bacterial artificial chromosome (BAC) library of Pacific white shrimp, *Litopenaeus vannamei*. *Mar. Biotechnol. (NY)* 12, 141–149. doi: 10.1007/s10126-009-9209-y

Zhang, X. J., Yuan, J. B., Sun, Y. M., Li, S. H., Gao, Y., Yu, Y., et al. (2019). Penaeid shrimp genome provides insights into benthic adaptation and frequent molting. *Nat. Commun.* 10:356.

Zou, C. S., Chen, A. J., Xiao, L. H., Muller, H. M., Ache, P., Haberer, G., et al. (2017). A high-quality genome assembly of quinoa provides insights into the molecular basis of salt bladder-based salinity tolerance and the exceptional nutritional value. *Cell Res.* 27, 1327–1340. doi: 10.1038/cr.2017.124

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 May 2021
doi: 10.3389/fgene.2021.671729

Integrated lncRNA and mRNA Transcriptome Analyses in the Ovary of *Cynoglossus semilaevis* Reveal Genes and Pathways Potentially Involved in Reproduction

Yani Dong[1,2], Likang Lyu[1], Daiqiang Zhang[2], Jing Li[2], Haishen Wen[1] and Bao Shi[2,3]**

[1] Key Laboratory of Mariculture, Ministry of Education, Fisheries College, Ocean University of China, Qingdao, China, [2] Key Laboratory of Sustainable Development of Marine Fisheries, Ministry of Agriculture and Rural Affairs, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao, China, [3] Laboratory for Marine Fisheries and Food Production Processes, Pilot National Laboratory for Marine Science and Technology, Qingdao, China

OPEN ACCESS

Edited by:
Roger Huerlimann, Okinawa Institute of Science and Technology Graduate University, Japan

Reviewed by:
Marcos Edgar Herkenhoff, São Paulo State University, Brazil
Yongshuang Xiao, Institute of Oceanology, Chinese Academy of Sciences (CAS), China

***Correspondence:**
Haishen Wen
wenhaishen@ouc.edu.cn
Bao Shi
shibao@ysfri.ac.cn

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *24 February 2021*
Accepted: *20 April 2021*
Published: *19 May 2021*

Citation:
Dong Y, Lyu L, Zhang D, Li J, Wen H and Shi B (2021) Integrated lncRNA and mRNA Transcriptome Analyses in the Ovary of Cynoglossus semilaevis Reveal Genes and Pathways Potentially Involved in Reproduction. Front. Genet. 12:671729. doi: 10.3389/fgene.2021.671729

Long non-coding RNAs (lncRNAs) have been reported to be involved in multiple biological processes. However, the roles of lncRNAs in the reproduction of half-smooth tongue sole (*Cynoglossus semilaevis*) are unclear, especially in the molecular regulatory mechanism driving ovarian development and ovulation. Thus, to explore the mRNA and lncRNA mechanisms regulating reproduction, we collected tongue sole ovaries in three stages for RNA sequencing. In stage IV vs. V, we identified 312 differentially expressed (DE) mRNAs and 58 DE lncRNAs. In stage V vs. VI, we identified 1,059 DE mRNAs and 187 DE lncRNAs. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses showed that DE mRNAs were enriched in ECM-receptor interaction, oocyte meiosis and steroid hormone biosynthesis pathways. Furthermore, we carried out gene set enrichment analysis (GSEA) to identify potential reproduction related-pathways additionally, such as fatty metabolism and retinol metabolism. Based on enrichment analysis, DE mRNAs with a potential role in reproduction were selected and classified into six categories, including signal transduction, cell growth and death, immune response, metabolism, transport and catabolism, and cell junction. The interactions of DE lncRNAs and mRNAs were predicted according to *antisense*, *cis*-, and *trans*-regulatory mechanisms. We constructed a competing endogenous RNA (ceRNA) network. Several lncRNAs were predicted to regulate genes related to reproduction including *cyp17a1*, *cyp19a1, mmp14, pgr*, and *hsd17b1.* The functional enrichment analysis of these target genes of lncRNAs revealed that they were involved in several signaling pathways, such as the TGF-beta, Wnt signaling, and MAPK signaling pathways and reproduction related-pathways such as the progesterone-mediated oocyte maturation, oocyte meiosis, and GnRH signaling pathway. RT-qPCR analysis showed that two lncRNAs (XR_522278.2 and XR_522171.2) were mainly expressed in the ovary. Dual-fluorescence *in situ* hybridization experiments showed that both XR_522278.2 and XR_522171.2 colocalized with their target genes *cyp17a1* and *cyp19a1*, respectively, in the follicular cell layer. The results further demonstrated

that lncRNAs might be involved in the biological processes by modulating gene expression. Taken together, this study provides lncRNA profiles in the ovary of tongue sole and further insight into the role of lncRNA involvement in regulating reproduction in tongue sole.

Keywords: *Cynoglossus semilaevis*, dual-fluorescence *in situ* hybridization, transcriptome, integrated analysis, reproduction, ovary, mRNA, lncRNA

INTRODUCTION

Transcriptome sequencing is an effective method for investigating the important functions of long non-coding RNAs (lncRNAs), a kind of non-coding RNAs without protein-coding potential with lengths longer than 200 nucleotides (Quinn and Chang, 2016). LncRNAs have been studied for their function in regulating the expression of genes through various transcriptional and posttranscriptional mechanisms (Dykes and Emanueli, 2017) such as epigenetic modification (Mercer and Mattick, 2013), chromatin remodeling (Tsai et al., 2010), repression or activation of translation (Carrieri et al., 2012) and miRNA sponging (Faghihi et al., 2010). With more sensitive sequencing technologies, a steadily increasing number of studies have shown that lncRNAs play critical roles in various biological processes including cell differentiation and development (Fatica and Bozzoni, 2014), immune responses (Chen et al., 2017), and diseases (Kopp, 2019). These studies concerning the identification and function of lncRNAs have mainly focused on humans and other mammals, while research in fish species, especially teleosts, which include the largest number of living species among all scientific classes of vertebrates, lags far behind that in mammals.

Reproduction, a limiting factor for aquaculture, is a complicated physiological process. In teleosts, the reproductive cycle is dominated by estradiol in oocyte growth and by progesterone in oocyte maturation prior to ovulation. Several enzymes, such as aromatase P450c19 (*cyp19a1*), cytochrome P450c17 (*cyp17a1*), and 17β-hydroxysteroid dehydrogenase type 1 (*hsd17b1*), are involved in the synthesis of key steroid hormones (Nagahama and Yamashita, 2008). P450c19 is the key enzyme that converts testosterone to estradiol. In this process, P450c17 is necessary of the steroid precursor shift that must occur prior to oocyte maturation (Nagahama and Yamashita, 2008). In addition, *hsd17b1* is an essential enzyme for converting estrone (E_1) to active, receptor-binding estradiol (E_2) in fish (Tokarz et al., 2015). Ovulation, a prerequisite for oocyte fertilization, is completed by follicular rupture, which requires extracellular matrix (ECM) dissolution (Ogiwara and Takahashi, 2019). The matrix metalloproteinase (MMP) system is required for the hydrolysis of ECM proteins present in the follicle layers of ovulating follicles. The nuclear progestin receptor (*pgr*), induced by luteinizing hormone, is an essential mediator of ovulation that complexed with maturation-inducing steroids in the nucleus to become an active transcription factor.

A number of studies have demonstrated that lncRNAs play a vital role in reproductive processes (Taylor et al., 2015), including sex hormone responses (Yang et al., 2013), oocyte meiosis (Yang et al., 2020), and ovulation (Lian et al., 2020). Emerging evidence from humans has identified the roles of lncRNAs in reproduction-related systems (Wang and Qin, 2019) and as key regulators of the normal development of granulosa cells (Tu et al., 2020). A novel lncRNA (lncRNA2193) that participates in oocyte meiosis and induces oocyte maturation is found in porcines (Yang et al., 2020). A previous study showed that oocyte-specific lncRNAs could control oocyte development and early embryogenesis in cattle (Wang et al., 2020). Additionally, several lncRNAs have been filtered and shown to regulate the synthesis of progesterone, oogenesis and oocyte maturation in goats (Liu Y. et al., 2018). The identification of lncRNAs involved in processes such as growth and development (Ali et al., 2018; Wu et al., 2020), stress responses (Dettleff et al., 2020; Quan et al., 2020), immune responses (Liu et al., 2019; Zheng et al., 2021), and sex differentiation (Cai et al., 2019; Feng et al., 2020) in fish species has begun to be reported in recent years. However, studies involving the systematic identification and characterization of lncRNAs or the roles of lncRNAs in tongue sole reproduction are lacking.

Tongue sole is marine flatfish that is an important aquaculture species in China with high economic value. Due to overfishing, its wild resources have decreased. A tongue sole aquaculture industry has developed in the last few years. However, there are some problems in culturing tongue sole. Its reproductive dysfunctions may be due to environmental and endogenous factors that restrict its reproduction (Shi et al., 2016; Song et al., 2020). In recent years, research on the reproduction of tongue sole has focused on the identification of candidate functional genes (Liu et al., 2016) and endocrine mechanisms. However, research on the lncRNA regulatory mechanism in the reproduction of tongue sole remains limited.

Thus, there is an urgent need to identify and characterize lncRNAs and investigate the regulatory mechanisms of oocyte growth, maturation and ovulation to improve the efficiency of reproduction in tongue sole. Therefore, in this study, three ovarian stages of tongue sole [stage IV: late vitellogenesis, stage V: maturation stage, and stage VI: after ovulation (Shi et al., 2016)] were chosen for sequencing to obtain their transcriptome profiles and investigate the potential involvement of lncRNAs in oocyte growth, maturation, and ovulation. This transcriptome sequencing analysis provides profiles that are useful for understanding the molecular regulatory mechanism of reproduction in tongue sole. The identification of the potential functions of lncRNAs provides a foundation for clearly understanding the factors that regulate oocyte growth, maturation and ovulation, which will allow the efficiency of tongue sole reproduction to be improved.

MATERIALS AND METHODS

Sample Preparation and Ethics Statement

We described the experimental fish and the applied feeding and sampling procedures previously (Shi et al., 2016). Briefly, we collected tongue sole ovary samples from three ovarian development stages in triplicate. The identification of ovarian development stages is detailed in our previous studies (Shi et al., 2016). In addition, we collected triplicate samples of 12 tissues (brain, pituitary, liver, heart, ovary, kidney, heart kidney, muscle, stomach, testis, spleen, and intestines) of mature tongue sole. The tissues were immediately frozen in liquid nitrogen and stored in a refrigerator at −80°C to prevent degradation until being used for RNA isolation.

The collection and handling of the animals used in this study were approved by the Animal Care and Use Committee at the Chinese Academy of Fishery Sciences, and all the experimental procedures were performed in accordance with the guidelines for the Care and Use of Laboratory Animals at the Chinese Academy of Fishery Sciences. No endangered or protected species were involved in this experiment.

RNA Isolation, Library Preparation, and RNA Sequencing

Total RNA was extracted using Total RNA Extraction Reagent (Vazyme Biotech, China) according to the manufacturer's recommendations. The concentration and purity of the RNA were evaluated with a biophotometer (OSTC, China). The quality was assessed via agarose gel electrophoresis. After total RNA was extracted, rRNAs were removed by Ribo-Zero^{TM} Magnetic Kit (Epicentre, Madison, WI, United States) to retain mRNAs and ncRNAs. Then the enriched mRNA and ncRNAs were fragmented into short fragments by using fragmentation buffer and reverse transcribed into complementary DNA (cDNA) with random primers from RNA Library Prep Kit (NEB, United States). After second-strand cDNA synthesis, the cDNA fragments were purified with a QiaQuick PCR extraction kit; they were then subjected to end repair, poly (A) addition, and Illumina sequencing adapter ligation. After the digestion of second-strand cDNA by uracil-N-glycosylase (UNG), the products were size selected by agarose gel electrophoresis, PCR amplified, and sequenced using an Illumina HiSeqTM 4000 system by Gene *Denovo* Biotechnology Co. (Guangzhou, China).

Sequence Data Processing and Analysis

To obtain high-quality clean reads, reads were further filtered by fastp (version 0.18.0) (Chen et al., 2018). Raw data containing adapters or low-quality (q-value $\leq$ 20) bases were removed from the subsequent analysis. Briefly, adapter sequences were removed according to the setting of retaining reads longer than 50 bp. Reads consisting of all "A" bases or containing more than 10% unknown nucleotides (N) were also removed. Reads were mapped to the rRNA database with Bowtie2 (2.2.8) (Langmead and Salzberg, 2012). After removing mapped rRNA reads, the remaining reads were aligned to the tongue sole reference genome by using TopHat2 (version 2.1.1) (Kim et al., 2013). Transcript reconstruction was carried out with Cufflinks (Trapnell et al., 2012), TopHat2 and Cuffmerge software. We used Cuffcompare to identify new transcripts with lengths greater than 200 bp, and more than 2 exons. We used the Coding-Non-Coding-Index (CNCI, version 2) (Sun et al., 2013) and Coding Potential Calculator (CPC[1]) (Kong et al., 2007) and the protein database SwissProt to predict the coding capacity of the new transcripts. The transcriptional intersections without coding potential and protein annotation information were defined as lncRNAs. According to their locations relative to protein coding genes, the lncRNAs were classified into five types: intergenic lncRNAs, bidirectional lncRNAs, intronic lncRNAs, antisense lncRNAs, and sense-overlapping lncRNAs.

Differentially Expressed Transcripts and Functional Enrichment Analysis

RSEM software (Li and Dewey, 2011) was used to quantify transcript abundances. The expression levels of all transcripts were normalized using the fragments per kilobase of transcript per million mapped reads (FPKM) method. The edgeR package[2] was used to identify significant differentially expressed (DE) mRNAs and lncRNAs among groups showing a false discovery rate (FDR) $<$ 0.05 and fold change (FC) $\geq$ 2. All DE genes were mapped to GO terms in the Gene Ontology database[3], gene numbers were calculated for every term, significantly enriched GO terms in DEGs comparing to the genome background were defined by hypergeometric test. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis identified significantly enriched metabolic pathways or signal transduction pathways in DE genes comparing with the whole genome background using a hypergeometric test. A p-value $<$ 0.05 was considered to indicate significant enrichment. Both GO and KEGG analyses were conducted by Gene *Denovo* Biotechnology Co. (Guangzhou, China).

Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) is a powerful and flexible analytical method that determines whether a previously defined set of genes, rather than a single gene, shows statistically significant, concordant differences between two biological states (Subramanian et al., 2005). Traditional enrichment analysis methods, such as GO and KEGG analyses, have focused on whether a group of DE genes is enriched for a pathway or ontology term by using overlap statistics such as the cumulative hypergeometric distribution. However, GSEA considers all genes, not only those that are significantly DE. In GSEA, significance is assessed by permuting the class labels, which preserves gene-gene correlations and thus provides a more accurate null model. The analysis was performed using GSEA software[4] and MSigDB to identify whether a set of genes associated with specific GO terms or pathways showed significant differences between the

[1]http://cpc.cbi.pku.edu.cn/
[2]http://www.r-project.org/
[3]http://www.geneontology.org/
[4]http://www.broadinstitute.org/gsea/

two groups. Briefly, we input a gene expression matrix and ranked genes via the signal-to-noise normalization method. Enrichment scores and *p*-values were calculated with default parameters. In this study, we used the following threshold criteria to obtain the significantly enriched pathways: significant *p*-value, $p < 0.05$; significant *q*-value, $q < 0.25$; and normalized enrichment score ($|\text{NES}|$) > 1.

Association Analysis of lncRNA-mRNA

The target genes of lncRNA transcripts were predicted through *antisense*, *cis*-, and *trans*-regulation analysis. Some *antisense* lncRNAs may regulate gene silencing, transcription and mRNA stability. To reveal the interaction between *antisense* lncRNAs and mRNAs, RNAplex software[5] (Tafer and Hofacker, 2008) was used to predict the complementary correlations of antisense lncRNAs and mRNAs. One of the functions of lncRNAs is the *cis*-regulation of their neighboring genes in the same allele. Upstream lncRNAs showing the intersection of promoters or other *cis*-elements may regulate gene expression at the transcriptional or posttranscriptional level. Downstream or 3′UTR lncRNAs may have other regulatory functions. Thus, lncRNAs that had been previously annotated as "unknown regions" were annotated again. LncRNAs located less than 100 kb upstream/downstream of a gene are likely *cis*-regulators. On the other hand, lncRNAs can have an effect on remote target genes. We analyzed the correlation of expression between lncRNAs and protein-coding genes to identify target genes of lncRNAs showing $p \geq 0.9$. Cytoscape (Shannon et al., 2003) software (v3.6.0) was used to analyze and visualize the interaction network relationships.

Construction of a Competing Endogenous RNA Network

The competing endogenous RNA (ceRNA) network was constructed based on ceRNA theory. The mRNA-miRNA and lncRNA-miRNA expression correlations were evaluated using the Spearman rank correlation coefficient (SCC). Pairs with an SCC < -0.7 were selected as negatively coexpressed lncRNA–miRNA pairs or mRNA-miRNA pairs. The expression correlation between lncRNAs and mRNAs was evaluated using the Pearson correlation coefficient (PCC). Pairs with a PCC > 0.9 were selected as coexpressed lncRNA–mRNA pairs, and both the mRNA and lncRNA in each pair were targeted and negatively coexpressed with a common miRNA. We used a hypergeometric cumulative distribution function test to determine whether the common miRNA sponges between the two genes were significant. As a result, only the gene pairs with a *p*-value of less than 0.05 were selected. Cytoscape software (v3.6.0) was used to analyze and visualize the interaction network relationships.

Validation by Real-Time Quantitative PCR

Eight transcripts, including four mRNAs and four lncRNAs, were selected to validate the RNA sequencing (RNA-Seq) results using real-time quantitative PCR (RT-qPCR). Gene-specific primers were designed with Primer 5 software (Premier Biosoft International). The β-*2-m* gene was used as the endogenous control for normalization, and all primers are shown in **Supplementary Table 1**. First-strand cDNA was synthesized with HiScript III RT SuperMix for qPCR (+gDNA wiper) (Vazyme Biotech, China) according to the manufacturer's instructions. The 20 μL RT-qPCR reaction mixture consisted of 2 μL of cDNA template, 0.4 μL of both primers, 10 μL of ChamQ SYBR Color qPCR Master Mix (2×), 0.4 μL of cDNA and 6.8 μL of RNase-free water. PCR amplification was performed as follows: incubation in a 96-well optical plate at 95°C for 30 s, followed by 40 cycles of 95°C for 10 s and 60°C for 30 s. Melting curves were also plotted (60–90°C) to ensure that a single PCR product was amplified for each pair of primers. All reactions were carried out in a StepOne Plus Real-Time PCR system (Applied Biosystems). The experiments were carried out in triplicate for each data point. The relative quantification of mRNA and lncRNA transcript expression was performed using the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001).

Tissue Distribution of Candidate lncRNAs by RT-qPCR

Real-time quantitative PCR was used to determine lncRNA expression levels in 12 tissues (brain, pituitary, ovary, liver, testis, kidney, head kidney, heart, spleen, muscle, stomach, and intestines). We chose two lncRNAs (XR_522278.2 and XR_522171.2) with a potential regulatory relationship with the *cyp17a1* and *cyp19a1* genes. Briefly, three parallel samples of different tissues from three adult tongue sole individuals were prepared for RNA isolation. The total RNA of the 12 tissues was reverse transcribed into cDNA. The methods of reverse transcription and RT-qPCR have been mentioned above. The lncRNA primers are listed in **Supplementary Table 1**.

Dual-Fluorescence *in situ* Hybridization of lncRNAs and mRNAs

To identify the regulation of target genes by lncRNAs, we performed a dual-fluorescence *in situ* hybridization (ISH) assay, which was modified with reference to previous studies (Qi et al., 2017). First, stage V ovary tissues of tongue sole were collected and fixed in 4% paraformaldehyde in phosphate-buffered saline (PBS) for 24 h and then preserved in paraffin. The sections for the ISH were 7 mm thick. In dual-fluorescence ISH, the probes for lncRNAs (XR_522278.2 and XR_522171.2) and mRNAs (*cyp17a1* and *cyp19a1*) were labeled with DIG and biotin (Roche Diagnostics, Mannheim, Germany), respectively. The sections were washed with PBS and blocked with blocking buffer (10% goat serum, Invitrogen, United States) after hybridization and posthybridization. Next, the sections were incubated with a peroxidase-conjugated anti-DIG secondary antibody (diluted 1:500 with blocking buffer, Roche Diagnostics, Mannheim, Germany) for 1 or 2 h. After two washes with PBS for 5 min each time, tyramide kits with Alexa Fluor 488 (Invitrogen, United States) were used for chromogenic reaction for 30 min. A fluorescence microscope (Olympus BX53F, Japan) was used to observe that the first

[5]http://www.tbi.univie.ac.at/RNA/RNAplex.1.html

round of staining was complete and that the samples were ready for the second round of fluorescein detection. The sections were incubated with 3% H_2O_2 for 20 min to remove HRP conjugated to the antibiotin, followed by two washes with PBS for 5 min each time. The sections were incubated with a peroxidase-conjugated streptavidin secondary antibody (Proteintech, United States) for 1 or 2 h. After washing similar to the previous washing step, tyramide kits with Alexa Fluor 594 (Invitrogen, United States) were used for the second chromogenic reaction for 30 min. Incubation was stopped by washing with PBS when the signal was detected. The sections were stained with DAPI (10 μg/mL, Solarbio, China) and then covered with coverslips with antifade mounting medium (Beyotime, China). Photos were obtained with a fluorescence microscope (Olympus, Japan).

RESULTS

Overview of RNA-Seq

To investigate the ovary transcriptome mechanism and identify the role of lncRNAs in the reproduction of tongue sole, nine cDNA libraries were constructed and sequenced, resulting in a total of 730.8 million (730,797,538) clean reads after primary filtering. After filtering out the low-quality reads, 720.5 billion (720,475,082, 98.59%) high-quality clean reads were further processed. On average, 99.84% of the high-quality reads were not mapped to the rRNA database. Among the unmapped reads, 78.35% were successfully mapped to the reference genome. Approximately 77.69% of the uniquely mapped reads were used for transcript construction (**Supplementary Table 2**).

According to the location of the assembled transcripts in the reference genome, the novel transcripts were filtered according to the following criteria: transcript length ≥ 200 bp and exon number ≥ 2. We considered the new transcripts without coding capacity and protein annotation information to be lncRNAs (**Figure 1A**). A total of 353,809 mRNAs (218,325 known and 135,484 novel) and 33,562 lncRNAs (10,201 known and 23,361 novel) were obtained. The detailed genomic information and functional annotations of the new transcripts are shown in **Supplementary Table 3**.

The novel lncRNA transcripts could be divided into five categories. We detected 1,941 intergenic lncRNAs, 355 bidirectional lncRNAs, 763 antisense lncRNAs, 384 sense-overlapping lncRNAs, and 455 other lncRNAs. There were no intronic lncRNAs (**Figure 1B**).

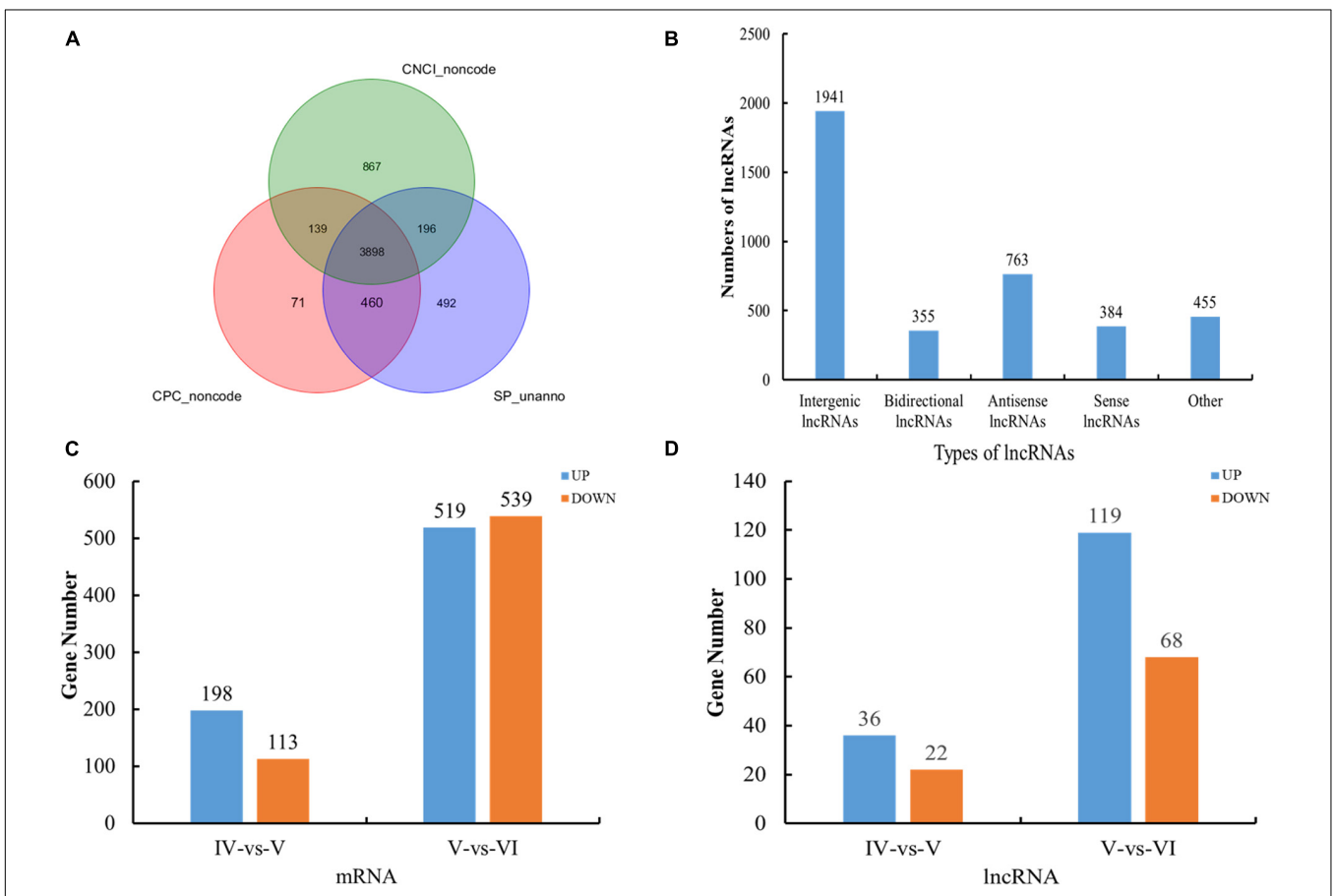

FIGURE 1 | Overview of RNA sequencing in the tongue sole ovary. **(A)** Venn diagrams of annotated results of CPC, CNCI, and SwissProt. **(B)** The categories of novel lncRNAs transcripts. **(C)** The histogram of differently expressed mRNA statistics. **(D)** The histogram of differently expressed lncRNA statistics.

Analysis of Differentially Expressed mRNAs and lncRNAs

A total of 311 and 1,058 mRNAs were DE in stage IV vs. V and V vs. VI, respectively, and 79 mRNAs were DE in both comparison groups (**Supplementary Table 4**). Among these DE mRNAs, 198 upregulated and 113 downregulated mRNAs were identified in stage IV vs. V, and 519 upregulated and 539 downregulated mRNAs were identified in stage IV vs. V (**Figure 1C**). A total of 58 and 187 lncRNAs were DE in stage IV vs. V and V vs. VI, respectively, and 13 lncRNAs were DE in both comparison groups (**Supplementary Table 4**). Among these DE lncRNAs, 36 upregulated and 22 downregulated lncRNAs were identified in stage IV vs. V, and 119 upregulated and 68 downregulated lncRNAs were identified in stage V vs. VI (**Figure 1D**). Circos software was used to visualize the DE mRNAs and lncRNAs. The DE mRNAs and lncRNAs in the two comparison groups (stage IV vs. V and stage V vs. VI) are shown in **Figures 2A,B**.

GO and KEGG Pathway Enrichment Analysis of DE mRNAs

We identified the biological functions of DE mRNAs through GO and KEGG pathway analysis. The GO terms with q-values ≤ 0.05 were considered significantly enriched (**Supplementary Table 5**). In stages IV vs. V, the identified molecular functions consisted mainly of binding functions, including metal ion binding, protein complex binding and purine nucleotide binding. Among biological processes, the significantly enriched GO terms included cell growth, collagen metabolic process, and DNA methylation (**Figure 3A**). In stage V vs. VI, the cellular component terms consisted of the cytoskeleton and microtubules. The most important molecular functions were related to steroid dehydrogenase activity, oxidoreductase activity, and protein hormone receptor activity. Among biological processes, the significantly enriched GO terms included steroid biosynthetic process, ovarian follicle development, reproductive structure development, and ovulation cycle process (**Figure 3B**).

In the KEGG pathway analysis, we used a q-value ≤ 0.05 as a threshold, and a total of 6 and 14 pathways were found to be significantly enriched in stage IV vs. V and V vs. VI, respectively (**Supplementary Table 5**). We selected the top 20 pathways in the KEGG enrichment analysis of stage IV vs. V and V vs. VI (**Figures 4A,B**). Among these pathways, the ECM-receptor interaction, PPAR signaling pathway, phagosome and focal adhesion pathways, which are associated with signaling molecules and interactions, the endocrine system, transport and catabolism and the cellular community, were significantly enriched in both comparison groups. Several genes were involved in the above pathways, such as thrombospondin-1 (*thbs1*), fibronectin-like (*fn1*), angiopoietin-related protein 4 (*angptl4*), cathepsin S (*ctss*), tenascin (*tnc*), and phosphoinositide 3-kinase regulatory subunit 5 (pik3r5) (**Supplementary Table 6**). In stages IV vs. V, oocyte meiosis was significantly enriched, which is related to oocyte development and maturation processes. Several genes, such as ribosomal protein S6 kinase alpha-1 (*rps6ka1*), serine/threonine-protein phosphatase 2 (*ppp2r5c*), F-box only protein 43 isoform X4 (*fbxo43*), and serine/threonine-protein kinase isoform X2 (*slk*), were involved in this pathway. In stage V vs. VI, several cellular process-related pathways were identified, including the regulation of the actin cytoskeleton, adherens junction and endocytosis pathways, which are associated with cell motility, proliferation, and catabolism. The steroid hormone

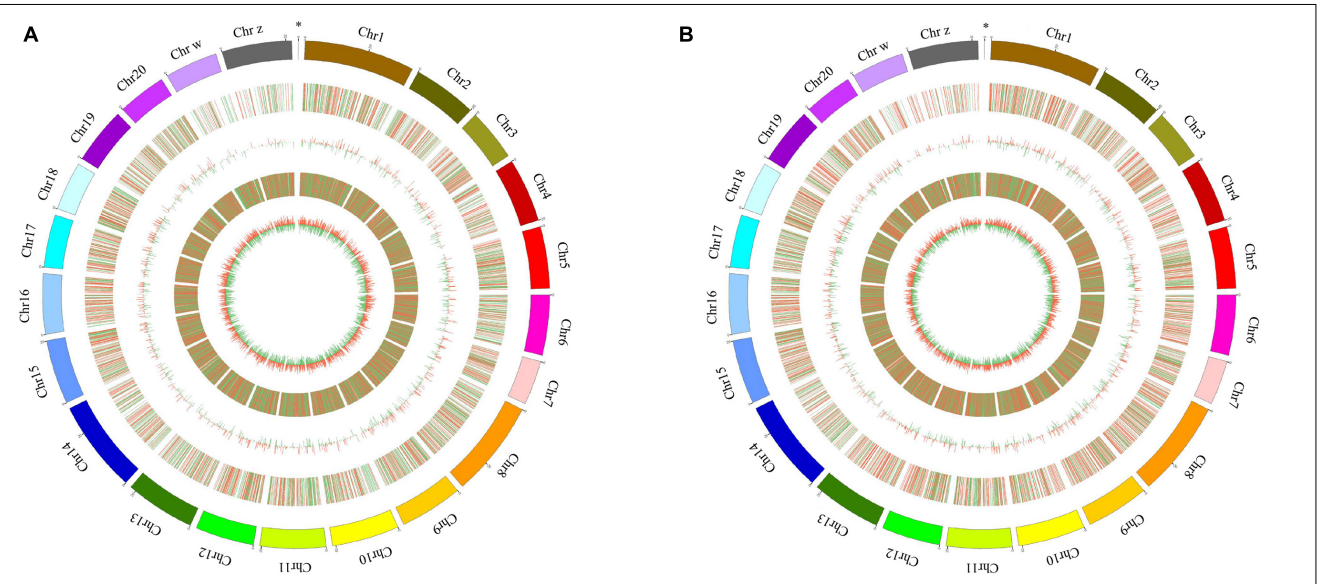

FIGURE 2 | Circos plot diagram of differential expressed (DE) lncRNA and mRNA in stage IV vs. V **(A)** and stage V vs. VI **(B)**. The outermost circle is the autosomal distribution of tongue sole; the second circle is the DE lncRNA on the chromosome, the red line represents up regulation, the green line represents down-regulation; the third circle is the histogram of DE lncRNAs at different positions. Red represents up regulation, green represents down regulation, and the higher column, indicates the more DE gene numbers. The fourth circle is the distribution of DE mRNAs on the chromosome, and the color distribution is the same as lncRNA; the innermost circle is the column with DE mRNAs at different positions, color distribution is the same as lncRNA. The "*" represents mitochondrion genome.

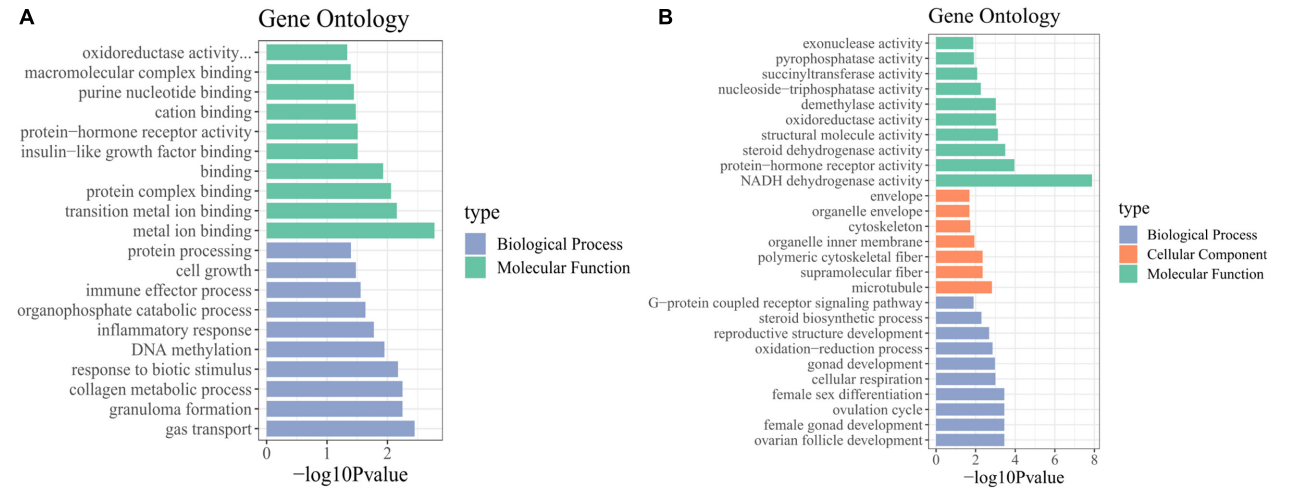

FIGURE 3 | GO enrichment analysis of DE mRNAs. GO terms enriched in **(A)** stage IV vs. V. and **(B)** stage V vs. VI.

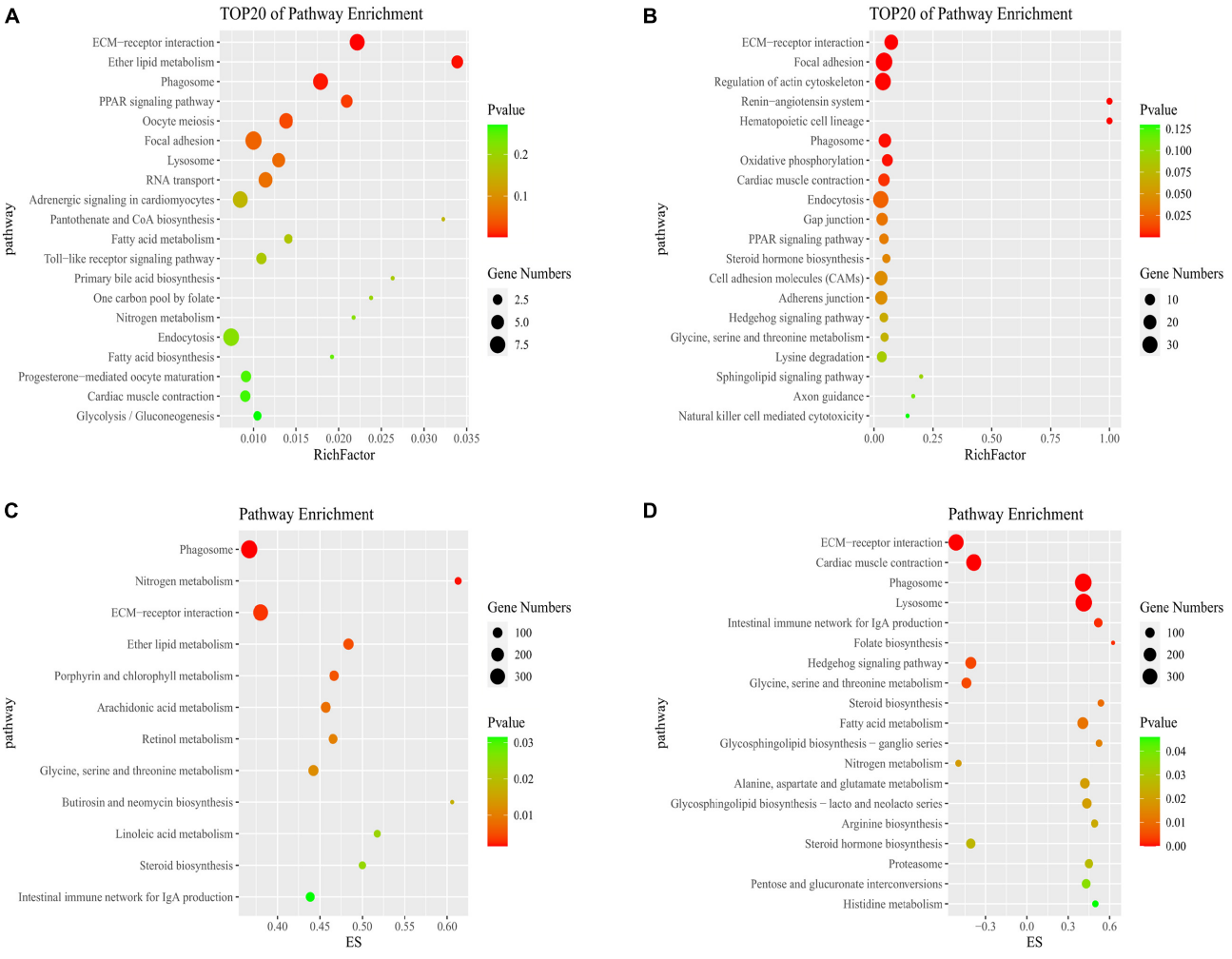

FIGURE 4 | Pathways enrichment analysis of DE mRNAs. KEGG pathway enriched in stage IV vs. V **(A)** and stage V vs. VI **(B)**. GSEA of KEGG pathway enriched in ovarian stage IV vs. V **(C)** and stage V vs. VI **(D)**.

biosynthesis pathway was also significantly enriched and included several genes, such as *cyp19a1, cyp17a1*, and *hsd17b1* (**Supplementary Table 6**).

According to the functional enrichment analysis and a literature search, a total of 100 candidate DE genes related to the reproductive cycle were filtered and divided into six categories by function, including signal transduction, metabolism, immune response, cell junction, transport and catabolism, and cell growth and death categories (**Supplementary Table 7**).

GSEA

To further study the biological functions of mRNAs and identify more undiscovered but very important pathways, we carried out GSEA. Generally, the gene sets were considered significant according to the following criteria: | NES| > 1, NOM p-value < 0.05, and FDR q-value < 0.25 (**Supplementary Table 8**). In stages IV vs. V, the identified cellular components were related to cell junctions. The molecular functions consisted of growth factor binding. In the biological process category, the regulation of cell development and differentiation was identified. In stage V vs. VI, the molecular functions consisted of oxidoreductase activity, and the biological processes consisted of muscle contraction and blood circulation (**Supplementary Table 8**). A total of 12 and 19 pathways were observed by GSEA to be significantly enriched in stage IV vs. V and V vs. VI, respectively (**Figures 4C,D**). Some pathways that we described above were also enriched according to GSEA, such as the ECM-receptor interaction and phagosome and steroid hormone biosynthesis pathways. In addition, steroid biosynthesis was significantly enriched in both stage IV vs. V and V vs. VI. The enrichment scores of the steroid biosynthesis pathway in the two comparison groups were driven by 16 and 13 genes, respectively. Some of these core genes were not significantly DE genes but contributed to significant pathway enrichment, such as 1,25-dihydroxyvitamin D (Mercer and Mattick, 2013) 24-hydroxylase (*cyp24a1*), sterol O-acyltransferase 1 (*soat1*), and delta (Yang et al., 2013)-sterol reductase (*tm7sf2*) (**Supplementary Table 8**). Furthermore, metabolic pathways such as glycine, serine and threonine metabolism, retinol metabolism and fatty acid metabolism were enriched according to GSEA.

Identification and Enrichment Analysis of Target Genes of lncRNAs

To predict the possible interaction of DE lncRNAs and mRNAs, we conducted *antisense, cis*-, and *trans*-regulatory relationship analyses. According to previous studies, lncRNAs can regulate gene expression through *antisense* and *cis*-regulatory mechanisms. These lncRNAs show relationships with genes according to their physical positions. In addition, lncRNAs can affect mRNAs that are not adjacent to the site of their transcription via *trans*-regulatory mechanisms. The Cytoscape platform was used to visualize these correlations (**Figure 5**).

Antisense Regulatory Relationship Analysis

A total of 1,767 pairs of *antisense* regulatory relationships were observed (**Supplementary Table 9**). Among these relationships, five lncRNAs exhibited *antisense* complementary correlations with four DE mRNAs, as shown in **Supplementary Table 7**. Among these genes, only one (*col4a1*) was an *antisense* target of a DE lncRNA (XR_001761307.1). Enrichment analysis indicated

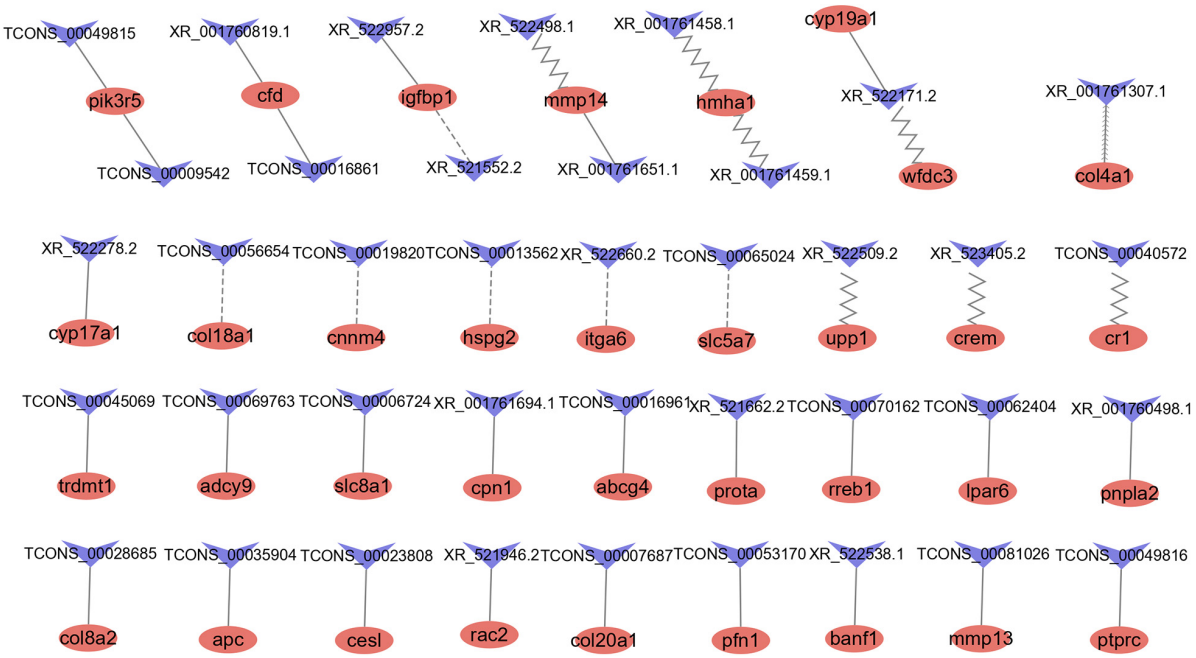

FIGURE 5 | LncRNAs and their potential *antisense, cis*-, and *trans*-target genes. Red ellipse represents mRNAs, blue "V" represents lncRNAs. The contiguous arrow, zigzag, dash, and solid line indicates the interaction between DE lncRNAs and its *antisense, cis*-, and *trans- positive* and *negative* target genes.

that *antisense* target genes were significantly enriched in 16 cellular component GO terms, 38 molecular function GO terms, 35 biological process GO terms, and 11 KEGG pathways (**Supplementary Table 10**).

Cis-Regulatory Relationship Analysis

A total of 6,019 pairs of *cis*-regulatory relationships were observed (**Supplementary Table 9**). Among these relationships, we identified 23 lncRNAs showing potential *cis*-target regulatory relationships with 14 DE mRNAs, as listed in **Supplementary Table 7**. Among these lncRNAs, only eight were DE. The *hmha1* gene was a potential *cis*-target gene of eight lncRNAs. Enrichment analysis indicated that *cis*-target genes were significantly enriched in 16 cellular component GO terms, 29 molecular function GO terms and 90 biological process GO terms and 15 KEGG pathways (**Supplementary Table 10**).

Trans Regulatory Relationship Analysis

We identified 5,359 coexpression pairs of lncRNAs and DE mRNAs (**Supplementary Table 9**). Of these, we identified 44 coexpression pairs of lncRNAs and mRNAs listed in **Supplementary Table 7**. Among these, there were 31 coexpression pairs of DE lncRNAs and DE mRNAs, 25 pairs exhibited a positive correlation, and six pairs exhibited a negative correlation. Enrichment analysis indicated that trans-target genes were significantly enriched for nine cellular component GO terms, 27 molecular function GO terms, and 93 biological process GO terms and 15 KEGG pathways (**Supplementary Table 10**).

The top 20 pathways identified in the KEGG enrichment analysis of the three types of regulatory relationships are shown in **Supplementary Figures 1A–C**. The pathways related to the cell cycle, signaling and reproduction included apoptosis, the TGF-beta pathway, the Wnt signaling pathway, the ErbB signaling pathway, the MAPK signaling pathway, the GnRH signaling pathway, and progesterone-mediated oocyte maturation. These results indicated that these lncRNAs may potentially play a role in reproduction.

mRNA-miRNA-lncRNA Interaction Analysis

We identified mRNAs and lncRNAs with an FC $\geq$ 2 and an FDR $<$ 0.05 in comparisons as significant DE transcripts and miRNAs with an FC $\geq$ 2 and p-value $<$ 0.05. Based on the DE mRNA, lncRNA and miRNA data, three software programs, MIREAP, miRanda, and TargetScan, were used to predict the targets of each miRNA. A ceRNA network was constructed after obtaining the mRNA-miRNA and lncRNA-miRNA pairs, in which the mRNA and lncRNA in each pair were targeted and negatively coexpressed with a common miRNA (**Supplementary Table 11**).

The interactions between lncRNAs and DE mRNAs predicted through ceRNA network analysis are listed in **Supplementary Table 1** (**Supplementary Figure 3**). The network consisted of 164 lncRNAs, 302 miRNAs, and 63 mRNAs. On the basis of combined ceRNA network and pathway enrichment analyses, five reproduction-related genes, *cyp17a1, cyp19a1, mmp14, pgr*, and *hsd17b1,* were selected to predict their interactions with lncRNAs. Thirty lncRNAs interacted with four mRNAs by competing for 35 miRNAs (**Figure 6**). Enrichment analysis was carried out to analyze the functions of the mRNAs involved in the ceRNA network. A p-value $<$ 0.05 was set as the threshold for significantly enriched GO terms and pathways. A total of 27 cellular component GO terms, 37 molecular function GO terms and 101 biological process GO terms were significantly enriched. In addition, a total of 27 pathways were significantly enriched, including reproduction-related pathways such as the GnRH signaling pathway and steroid hormone biosynthesis pathways (**Supplementary Figure 2** and **Supplementary Table 12**).

Validation by RT-qPCR

To validate the RNA-Seq results, eight transcripts, including four DE mRNAs and four DE lncRNAs, were selected for RT-qPCR. The results showed that the expression trends for all four DE mRNAs and four DE lncRNAs were consistent with the RNA-Seq results, with $R^2 = 0.9472$ and $R^2 = 0.8675$, respectively. Generally, the RNA-Seq results were mostly confirmed by qPCR analysis, implying the reliability and accuracy of the RNA-Seq analysis (**Supplementary Figure 4**).

Analysis of Expression Levels of Candidate lncRNAs in the Ovary

For the two lncRNAs (XR_522278.2 and XR_522171.2) showing a potential regulatory relationship with the *cyp17a1* and *cyp19a1* genes, respectively, we explored the expression distribution among 12 tissues of tongue sole by RT-qPCR. The lncRNAs were widely expressed in the 12 tissues but showed the highest expression in the ovary. The highest expression level of lncRNA XR_522278.2 was detected in the ovary (263.3-fold relative to control). The highest expression level of lncRNA XR_522171.2 was also detected in the ovary (646.3-fold relative to control) (**Figure 7**).

Colocalization of lncRNA and mRNA in Ovarian Tissue

Dual-fluorescence ISH of lncRNAs and mRNAs in stage V ovaries was performed to demonstrate the potential regulation of mRNAs by lncRNAs. As shown in **Figure 8**, positive XR_522278.2 and *cyp17a1* signals were present in the follicular cell layer, where *cyp17a1* plays a role in the shift in steroid precursor production. Similarly, positive XR_522171.2 and *cyp19a1* signals were present in the follicular cell layer, where cyp19a1 converts testosterone to E_2 (**Figure 8**).

DISCUSSION

Within the reproductive cycle, oocyte growth along with oocyte maturation and ovulation is a prerequisite for successful reproduction, which is important in the aquaculture industry (Das et al., 2017). To provide a comprehensive view of the changes in transcriptome levels during tongue sole ovary growth,

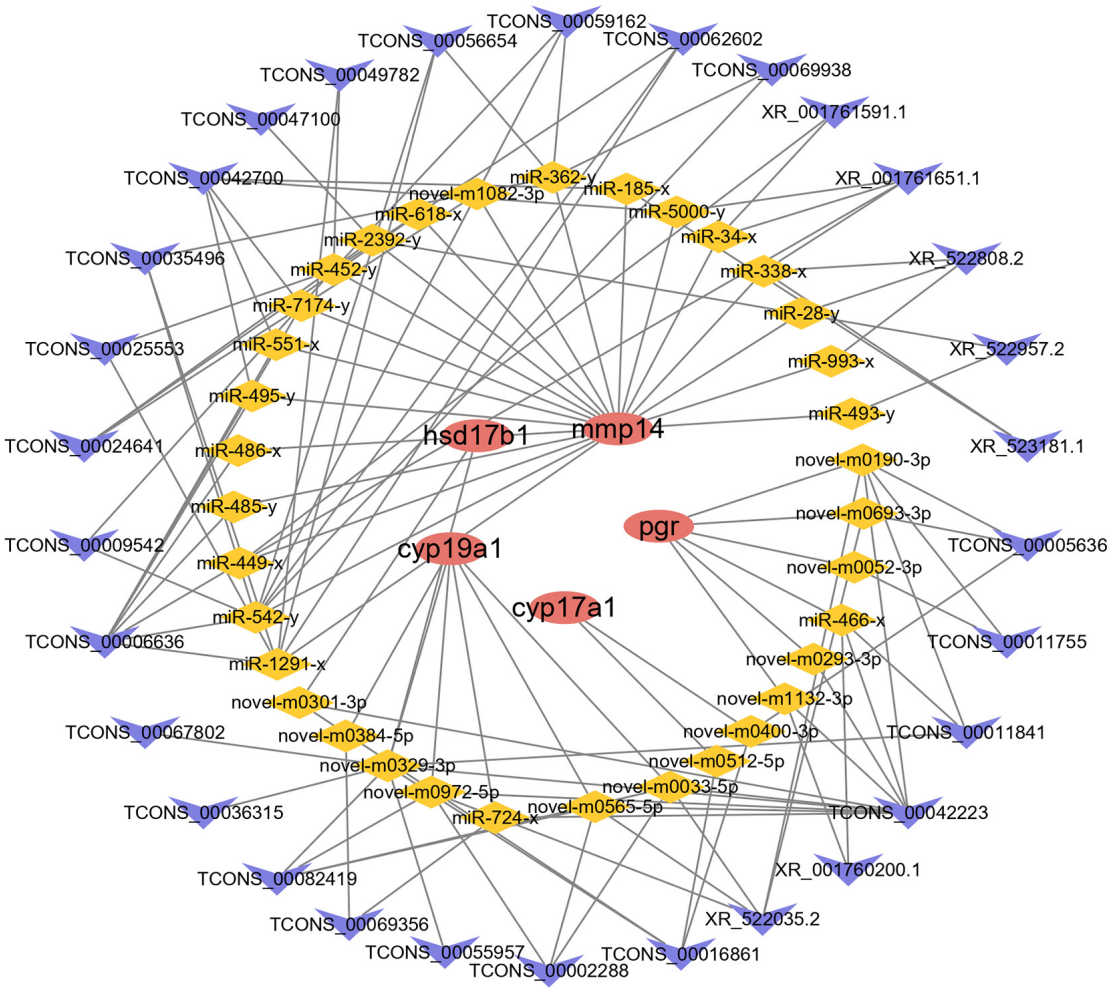

FIGURE 6 | The ceRNA network of *cyp17a1*, *cyp19a1*, *mmp14*, *pgr*, and *hsd17b1*. Red ellipse represents mRNAs, yellow diamond represents miRNAs, and blue "V" represents lncRNAs.

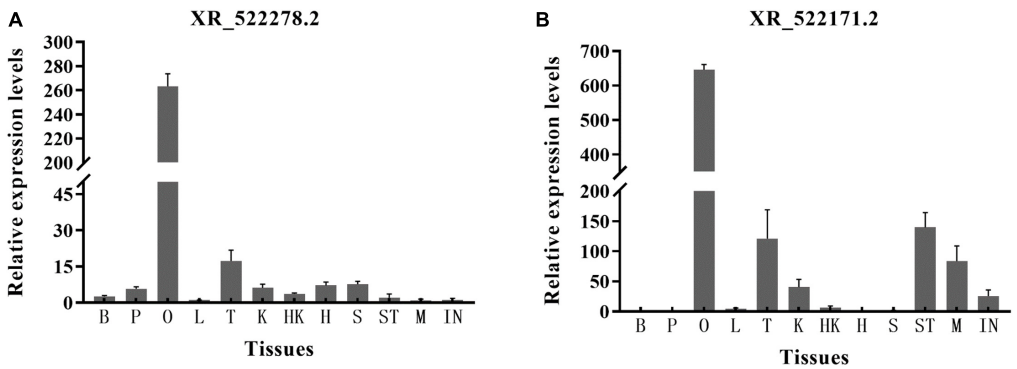

FIGURE 7 | The expression level of lncRNAs XR_522278.2 **(A)** and XR_522171.2 **(B)** in 12 tissues. B, brain; P, pituitary; O, ovary; L, liver; T, testis; K, kidney; HK, head kidney; H, heart; S, spleen; ST, stomach; M, muscle; IN, intestines.

maturation, and ovulation, RNA-Seq analysis was performed to investigate the regulatory mechanism involved and detect key gene functions and the corresponding regulatory lncRNAs. In this study, we identified 1,059 and 312 DE mRNAs in stage IV vs. V and V vs. VI, and total of 58 and 187 DE lncRNAs were observed in stage IV vs. V and V vs. VI, respectively.

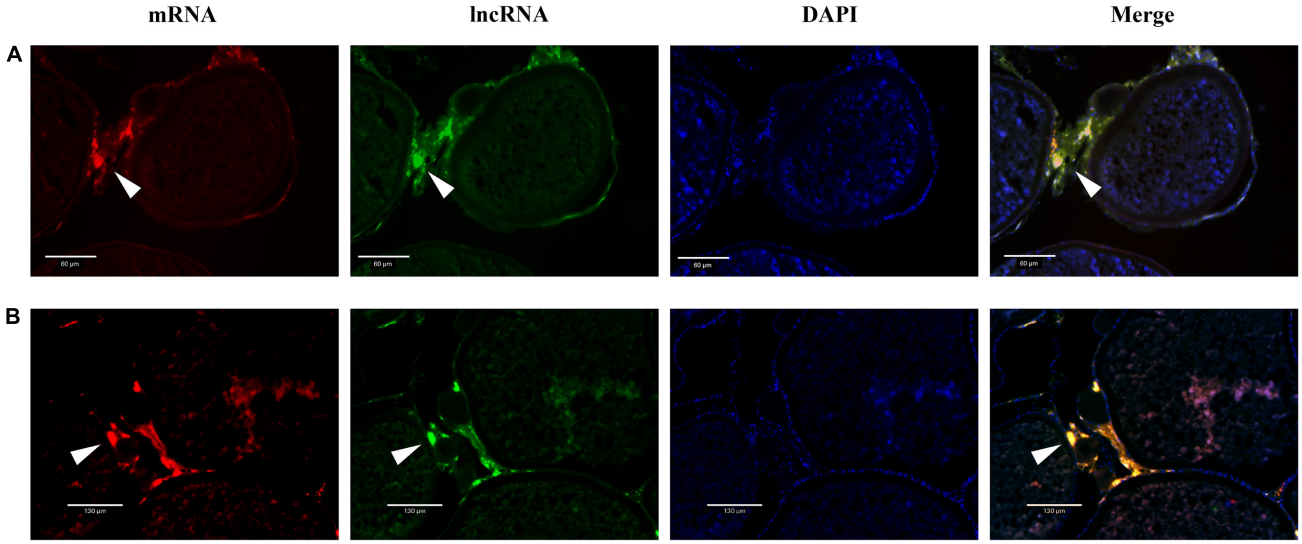

FIGURE 8 | The colocalization of lncRNA and mRNA. **(A)** The colocalization of cyp17a1 (red, stained with Alexa Fluor 594) and XR_522278.2 (green, stained with Alexa Fluor 488). **(B)** The colocalization of cyp19a1 (red, stained with Alexa Fluor 594) and XR_522171.2 (green, stained with Alexa Fluor 488).

It has been proven that steroid hormones are involved in the regulation of a variety of processes, such as embryonic development, sex differentiation, metabolism, immune responses, circadian rhythms, stress responses, and reproduction in vertebrates, including teleosts (Zhou et al., 2012). In the present study, we identified several DE genes that encode crucial enzymes in the biological synthesis of estradiol and progesterone, such as *cyp17a1*, *cyp19a1*, and *hsd17b1* (Tokarz et al., 2015). All of them were mainly expressed in stage V, when oocytes matured, and showed a significant decrease from stage V to VI. Moreover, the steroid hormone biosynthesis pathway was significantly enriched in stage V vs. VI, while GSEA showed that this pathway was downregulated in stage V vs. VI and that *cyp17a1*, *cyp19a1*, and *hsd17b1* were core genes in this pathway. These findings indicated that these genes and this pathway play key roles in stage V when oocytes mature.

Oocyte meiosis is an essential part of oocyte maturation in both mammals and teleosts (Nagahama and Yamashita, 2008; Zhao et al., 2020). This pathway, several genes involved, such as *fbxo43*, *slk*, and *rps6ka1*, were significantly enriched in stage IV vs. V. Early mitotic inhibitor 2 (EMI2), encoded by *fbxo43,* is a member of the F-box protein family that plays an important role in cytostatic factor arrest in *Xenopus* eggs (Tung et al., 2005; Vadhan et al., 2020). In the present study, *fbxo43* was found to be significantly DE in stage IV vs. V.

During the reproductive cycle, the ECM has been proven to participate in cell cycle processes during ovarian remodeling, ranging from follicular development and involution after ovulation to cell proliferation and differentiation (Curry and Osteen, 2003). The MMPs family is the major enzyme family responsible for ECM remodeling. Many studies in medaka have shown that MMPs participate in final ovulation, including oocyte maturation and follicle regression (Ogiwara et al., 2005; Takahashi et al., 2013). *Mmp14* is a membrane-type MMPs located at the cell surface (Pedersen et al., 2015). In this study, we found that *mmp14* was DE and exhibited the highest expression in stage VI in the ovary, after ovulation. Fibronectin and laminin are ECM proteins synthesized by follicular cells (Zhao and Luck, 1995; Yasuda et al., 2005). A study in *Pampus argenteus* identified laminin β2 as one of the main component of the basement membrane in ovarian follicles. Fibronectin is also detected in postovulatory follicular cells in *P. argenteus* (Thomé et al., 2010). In the present study, we found that *lamc3* exhibited the highest expression in stage IV but was significantly downregulated in stages V and VI. Fibronectin-like (*fn1*) increased from stage IV to stage V ovaries and reached a peak in stage VI in the ovary, when ovulation occurred. We also detected other cell adhesion-related genes, including *col4a1*, *ecm1*, *itga6*, *krt18*, and *tnc*, that were significantly DE in the ovary. Furthermore, ECM-receptor interaction and focal adhesion pathways were enriched in both stage IV vs. V and stage V vs. VI, and these pathways involved the greatest numbers of the above DE genes. These findings indicated that although endocrine hormones are necessary for the reproduction of tongue sole, the ECM is involved in all reproductive processes, especially ovulation, when follicular rupture causes the dissolution of the ECM of the follicular wall.

Some metabolic pathways were enriched in the two comparison groups according to GSEA, including retinol metabolism and fatty acid metabolism. Retinol plays an important role in mammalian placental and embryonic development (Marceau et al., 2007). Retinoic acid, a form of vitamin A, supports both male and female reproduction and embryonic development. In teleosts, retinotic acid is the key factor controlling the initiation of meiosis (Feng et al., 2015). Fatty acids may play important roles in the induction of oocyte maturation in marine teleosts, which requires n-3

long-chain PUFAs (LC-PUFAs) as essential fatty acids (Sorbera et al., 2001). In tongue sole, fatty acids have been proven to be required for gonadal steroidogenesis. An appropriate DHA:EPA ratio can induce high expression of the follicle-stimulating hormone receptor (FSHR), steroidogenic acute regulatory protein (STAR), 17α-hydroxylase (P450c17) and 3β-hydroxysteroid dehydrogenase (3β-HSD), which are involved in steroidogenesis (Xu et al., 2017). Thus, the retinol metabolism and fatty acid metabolism pathways may be involved in oocyte maturation, including steroidogenesis and oocyte meiosis, in tongue sole.

In recent years, lncRNAs have attracted increasing attention because of their involvement in modulating gene expression (Quinn and Chang, 2016). Research on the identification and characterization of lncRNAs involved in reproduction in fish, especially in tongue sole, is far behind that in mammals. Therefore, in the present study, we first obtained comprehensive expression profiles of lncRNA transcripts in three ovarian stages of tongue sole. A total of 33,562 lncRNAs (10,201 known and 23,361 novel) were obtained through high-throughput sequencing. These DE lncRNAs may have potential reproductive functions in tongue sole.

It is well established that lncRNAs can exert regulatory effects on mRNAs (Quinn and Chang, 2016; Tan-Wong et al., 2019; Elcheva and Spiegelman, 2020). The target genes of lncRNAs can be predicted on the basis of either their physical positions, in the case of *antisense* and *cis*-regulatory mechanisms, or their coexpression relationships, in the case of *trans*-regulation. In the present study, we showed that the *mmp14 gene* is a *cis*-target gene of two lncRNAs (XR_522498.1 and XR_522499.1) and that the *col4a1* and *crem genes* are the *cis*-target genes of lncRNAs XR_523405.2 and XR_001761307.1, respectively. We found that *antisense* and *cis*-target genes were enriched in the cell cycle, in processes including apoptosis and signaling pathways such as the TGF-beta signaling pathway and the Wnt signaling pathway. Several studies have shown that the TGF-beta signaling pathway and Wnt signaling pathway are involved in a broad spectrum of cellular functions, such as proliferation, apoptosis, differentiation, migration, and embryogenesis (Motomura et al., 2014; Yu et al., 2020). TGF-beta signals can lead to avian ovary granulosa cell proliferation mediated by signals transduced by the cytoplasmic signal transducer Smad (Schmierer et al., 2003). TGF-beta is involved in maintaining oocyte meiotic arrest regulated by FSH and LH in mice (Yang et al., 2019). Recent studies revealed that the inhibitory effect of FH535 (Wnt/β-catenin inhibitor) on the Wnt signaling pathway can promote the maturation of porcine oocytes and alter gene expression *in vitro* (Shi et al., 2018). In mice, cell cycle arrest is associated with the downregulation of Wnt signaling in granulosa cells (De Cian et al., 2020).

A total of 31 coexpressed pairs of DE lncRNAs and mRNAs were identified in this study, including 25 pairs with a positive correlation and six pairs with a negative correlation. Among these genes, *Igfbp1* exhibited a positive correlation with the lncRNA XR_522957.2. Previous studies have shown that different IGF-binding proteins exhibit marked expression in the preovulation period (Knight and Glister, 2006). The *cyp17a1 gene,* which encodes a key enzyme involved in steroidogenesis, exhibited a negative correlation with lncRNA (XR_522278.2). We also showed that *trans*-target genes were enriched in the MAPK signaling pathway, which plays a role in oocyte meiosis, meiotic resumption and maturation (Liang et al., 2007). Furthermore, lncRNAs have been reported to participate in the MAPK signaling pathway (Sulayman et al., 2019). Thus, the results indicated that these lncRNAs play a role in oocyte development and ovulation in tongue sole through *antisense, cis*-, and *trans*-regulatory mechanisms.

Through the analysis of mRNA, lncRNA and miRNA interactions, a ceRNA network was constructed in which lncRNAs competed for miRNA binding with mRNAs and thereby regulated the expression of mRNAs. In previous studies, ceRNA networks have been constructed to predict the mechanisms of immune infiltration in humans (Huang et al., 2019), muscle growth and development in Japanese flounder (Wu et al., 2020) and ovarian development in broody chickens (Liu L. et al., 2018). In the present study, we also constructed a ceRNA network to predict the biological functions of lncRNAs in reproduction. Interestingly, the genes closely associated with reproductive processes, including *mmp14*, *pgr, cyp17a1, cyp19a1*, and *hsd17b1,* were indicated to potentially be modulated by 30 lncRNAs via 35 miRNAs. Among these genes, *pgr* has been reported to play an important role in oocyte development and ovulation in mammalian reproduction (Akison and Robker, 2012). In addition, *pgr* knockout zebrafish exhibit normal oocyte growth and maturation but show defects in ovulation and fail to spawn, indicating the role of *pgr* in steroid-dependent genomic pathways leading to ovulation (Zhu et al., 2015). In the present study, *pgr* was found to be a DE gene that was enriched in pathways including oocyte meiosis and progesterone-mediated oocyte maturation. Twelve lncRNAs were shown to modulate *pgr* via 10 miRNAs. Moreover, we found that progesterone-mediated oocyte maturation, oocyte meiosis and the GnRH signaling pathway were enriched in DE mRNAs involved in the ceRNA network. This provides direct evidence that these lncRNAs, acting as regulators of gene expression, can modulate the expression of genes involved in ovarian growth, maturation, and ovulation in the process of reproduction.

Real-time quantitative PCR analysis revealed that the two lncRNAs (XR_522278.2 and XR_522171.2), potentially modulating *cyp17a1* and *cyp19a1* expression, respectively, were dominantly expressed in the ovary. As expected, both the XR_522278.2 and XR_522171.2 lncRNAs colocalized with their target genes *cyp17a1* and *cyp19a1*, respectively, in the follicular cell layer, which further demonstrated that the lncRNAs may play a role in the reproduction of tongue sole by regulating gene expression. Thus, lncRNAs may play roles in oocyte development, maturation and ovulation by regulating the expression of steroidogenic enzymes, transcription factors, and growth factors, which have been proven to be the key genes involved in reproduction in tongue sole.

CONCLUSION

This study first identified lncRNAs in the ovary and revealed the molecular mechanisms and pathways in the ovary involved in the reproduction of tongue sole. The enrichment analysis of DE mRNAs showed that pathways related to reproductive endocrine functions and the cell cycle drove oocyte growth, maturation, and ovulation. Several genes involved in these pathways, such as *cyp17a1*, *cyp19a1*, *mmp14*, *pgr*, and *hsd17b1*, were identified. We also identified lncRNAs that may exert regulatory effects on mRNAs related to the reproduction of tongue sole through *antisense*, *cis*- and *trans*-regulatory mechanisms. On the basis of the competitive relationships between mRNAs and lncRNAs, a ceRNA network was constructed to predict the potential mRNA targets of lncRNAs. The lncRNAs of critical genes were dominantly expressed in the ovary. Moreover, lncRNAs and their target genes were colocalized in the ovarian follicular cell layer. Taken together, the results of this study highlight the role of lncRNAs in regulating the expression of mRNAs that influence reproduction and provide deeper insight into the molecular mechanisms of reproduction in tongue sole. Moreover, further investigations involving functional analysis methods such as gene knockdown and overexpression are critical for elucidating the regulatory mechanism of lncRNAs in the fish ovary.

DATA AVAILABILITY STATEMENT

The sequencing data obtained from the RNA-seq were released to the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database under the accession numbers SRR14090692, SRR14090691, SRR14090690, SRR14090714, SRR14090713, SRR14090711, SRR14090700, SRR14090699 and SRR14090698.

ETHICS STATEMENT

The animal study was reviewed and approved by the Animal Care and Use Committee at the Chinese Academy of Fishery Sciences.

AUTHOR CONTRIBUTIONS

YD finished the experiment. YD and HW drafted and critically revised the manuscript. YD, LL, DZ, JL, and HW analyzed and interpreted the data. BS conceived the study and critically revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China (31772829), Central Public-Interest Scientific Institution Basal Research Fund, YSFRI, CAFS (20603022021004), National Key Research and Development Program (2019YFD0900503), Central Public-Interest Scientific Institution Basal Research, CAFS and Key Laboratory of Sustainable Development of Marine Fisheries, and Ministry of Agriculture and Rural Affairs, P. R. China (2019HY-XKQ01).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.671729/full#supplementary-material

REFERENCES

Akison, L. K., and Robker, R. L. (2012). The critical roles of progesterone receptor (PGR) in ovulation, oocyte developmental competence and oviductal transport in mammalian reproduction. *Reprod. Domest. Anim. Zuchthygiene* 47(Suppl. 4), 288–296. doi: 10.1111/j.1439-0531.2012.02088.x

Ali, A., Al-Tobasei, R., Kenney, B., Leeds, T. D., and Salem, M. (2018). Integrated analysis of lncRNA and mRNA expression in rainbow trout families showing variation in muscle growth and fillet quality traits. *Sci. Rep.* 8:12111. doi: 10.1038/s41598-018-30655-8

Cai, J., Li, L., Song, L., Xie, L., Luo, F., Sun, S., et al. (2019). Effects of long term antiprogestine mifepristone (RU486) exposure on sexually dimorphic lncRNA expression and gonadal masculinization in Nile tilapia (Oreochromis niloticus). *Aquat. Toxicol.* 215:105289. doi: 10.1016/j.aquatox.2019.105289

Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., et al. (2012). Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature* 491, 454–457. doi: 10.1038/nature11508

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/bioinformatics/bty560

Chen, Y. G., Satpathy, A. T., and Chang, H. Y. (2017). Gene regulation in the immune system by long noncoding RNAs. *Nat. Immunol.* 18, 962–972. doi: 10.1038/ni.3771

Curry, T. E. Jr., and Osteen, K. G. (2003). The matrix metalloproteinase system: changes, regulation, and impact throughout the ovarian and uterine reproductive cycle. *Endocrine Rev.* 24, 428–465. doi: 10.1210/er.2002-0005

Das, D., Khan, P. P., and Maitra, S. (2017). Endocrine and paracrine regulation of meiotic cell cycle progression in teleost oocytes: cAMP at the centre of complex intra-oocyte signalling events. *General Comparat. Endocrinol.* 241, 33–40. doi: 10.1016/j.ygcen.2016.01.005

De Cian, M. C., Gregoire, E. P., Le Rolle, M., Lachambre, S., Mondin, M., Bell, S., et al. (2020). R-spondin2 signaling is required for oocyte-driven intercellular communication and follicular growth. *Cell Death Different.* 27, 2856–2871. doi: 10.1038/s41418-020-0547-7

Dettleff, P., Hormazabal, E., Aedo, J., Fuentes, M., and Meneses, C. (2020). Identification and Evaluation of Long Noncoding RNAs in Response to Handling Stress in Red Cusk-Eel (Genypterus chilensis) via RNA-seq. *Mar. Biotechnol.* 22, 94–108. doi: 10.1007/s10126-019-09934-6

Dykes, I. M., and Emanueli, C. (2017). Transcriptional and Post-transcriptional Gene Regulation by Long Non-coding RNA. *Genomics Proteom. Bioinformat.* 15, 177–186. doi: 10.1016/j.gpb.2016.12.005

Elcheva, I. A., and Spiegelman, V. S. (2020). The Role of cis- and trans-Acting RNA Regulatory Elements in Leukemia. *Cancers* 12:3854. doi: 10.3390/cancers12123854

Faghihi, M. A., Zhang, M., Huang, J., Modarresi, F., Van der Brug, M. P., Nalls, M. A., et al. (2010). Evidence for natural antisense transcript-mediated inhibition of microRNA function. *Genome Biol.* 11:R56. doi: 10.1186/gb-2010-11-5-r56

Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.* 15, 7–21. doi: 10.1038/nrg3606

Feng, B., Li, S., Wang, Q., Tang, L., Huang, F., Zhang, Z., et al. (2020). lncRNA DMRT2-AS acts as a transcriptional regulator of dmrt2 involving in sex differentiation in the Chinese tongue sole (Cynoglossus semilaevis). *Comparat. Biochem. Physiol. Part B Biochem. Mol. Biol.* 253:110542. doi: 10.1016/j.cbpb.2020.110542

Feng, R., Fang, L., Cheng, Y., He, X., Jiang, W., Dong, R., et al. (2015). Retinoic acid homeostasis through aldh1a2 and cyp26a1 mediates meiotic entry in Nile tilapia (Oreochromis niloticus). *Sci. Rep.* 5:10131. doi: 10.1038/srep10131

Huang, R., Wu, J., Zheng, Z., Wang, G., Song, D., Yan, P., et al. (2019). The Construction and Analysis of ceRNA Network and Patterns of Immune Infiltration in Mesothelioma With Bone Metastasis. *Front. Bioengine. Biotechnol.* 7:257. doi: 10.3389/fbioe.2019.00257

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36. doi: 10.1186/gb-2013-14-4-r36

Knight, P. G., and Glister, C. (2006). TGF-beta superfamily members and ovarian follicle development. *Reproduction* 132, 191–206. doi: 10.1530/rep.1.01074

Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L., et al. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* 35, W345–W349. doi: 10.1093/nar/gkm391

Kopp, F. (2019). Molecular functions and biological roles of long non-coding RNAs in human physiology and disease. *J. Gene Med.* 21, e3104. doi: 10.1002/jgm.3104

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformat.* 12:323. doi: 10.1186/1471-2105-12-323

Lian, Z., Zou, X., Han, Y., Deng, M., Sun, B., Guo, Y., et al. (2020). Role of mRNAs and long non-coding RNAs in regulating the litter size trait in Chuanzhong black goats. *Reprod. Domest. Anim. Zuchthygiene* 55, 486–495. doi: 10.1111/rda.13642

Liang, C. G., Su, Y. Q., Fan, H. Y., Schatten, H., and Sun, Q. Y. (2007). Mechanisms regulating oocyte meiotic resumption: roles of mitogen-activated protein kinase. *Mol. Endocrinol.* 21, 2037–2055. doi: 10.1210/me.2006-0408

Liu, J., Zhang, W., Sun, Y., Wang, Z., Zhang, Q., and Wang, X. (2016). Molecular characterization and expression profiles of GATA6 in tongue sole (Cynoglossus semilaevis). *Comparat. Biochem. Physiol. Part B Biochem. Mol. Biol.* 198, 19–26. doi: 10.1016/j.cbpb.2016.03.006

Liu, L., Xiao, Q., Gilbert, E. R., Cui, Z., Zhao, X., Wang, Y., et al. (2018). Whole-transcriptome analysis of atrophic ovaries in broody chickens reveals regulatory pathways associated with proliferation and apoptosis. *Sci. Rep.* 8:7231. doi: 10.1038/s41598-018-25103-6

Liu, X., Li, W., Jiang, L., Lü, Z., Liu, M., Gong, L., et al. (2019). Immunity-associated long non-coding RNA and expression in response to bacterial infection in large yellow croaker (Larimichthys crocea). *Fish Shellf. Immunol.* 94, 634–642. doi: 10.1016/j.fsi.2019.09.015

Liu, Y., Qi, B., Xie, J., Wu, X., Ling, Y., Cao, X., et al. (2018). Filtered reproductive long non-coding RNAs by genome-wide analyses of goat ovary at different estrus periods. *BMC Genom.* 19:866. doi: 10.1186/s12864-018-5268-7

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

Marceau, G., Gallot, D., Lemery, D., and Sapin, V. (2007). Metabolism of retinol during mammalian placental and embryonic development. *Vitamins Hormon.* 75, 97–115. doi: 10.1016/s0083-6729(06)75004-x

Mercer, T. R., and Mattick, J. S. (2013). Structure and function of long noncoding RNAs in epigenetic regulation. *Nat. Struct. Mol. Biol.* 20, 300–307. doi: 10.1038/nsmb.2480

Motomura, E., Narita, T., Nasu, Y., Kato, H., Sedohara, A., Nishimatsu, S., et al. (2014). Cell-autonomous signal transduction in the Xenopus egg Wnt/β-catenin pathway. *Dev. Growth Different.* 56, 640–652. doi: 10.1111/dgd.12181

Nagahama, Y., and Yamashita, M. (2008). Regulation of oocyte maturation in fish. *Dev. Growth Differentiat.* 50(Suppl. 1), S195–S219. doi: 10.1111/j.1440-169X.2008.01019.x

Ogiwara, K., and Takahashi, T. (2019). Nuclear Progestin Receptor Phosphorylation by Cdk9 Is Required for the Expression of Mmp15, a Protease Indispensable for Ovulation in Medaka. *Cells* 8:8030215. doi: 10.3390/cells8030215

Ogiwara, K., Takano, N., Shinohara, M., Murakami, M., and Takahashi, T. (2005). Gelatinase A and membrane-type matrix metalloproteinases 1 and 2 are responsible for follicle rupture during ovulation in the medaka. *Proc. Natl. Acad. Sci. U S A.* 102, 8442–8447. doi: 10.1073/pnas.0502423102

Pedersen, M. E., Vuong, T. T., Rønning, S. B., and Kolset, S. O. (2015). Matrix metalloproteinases in fish biology and matrix turnover. *Matrix Biol.* 4, 86–93. doi: 10.1016/j.matbio.2015.01.009

Qi, X., Zhou, W., Wang, Q., Guo, L., Lu, D., and Lin, H. (2017). Gonadotropin-Inhibitory Hormone, the Piscine Ortholog of LPXRFa, Participates in 17β-Estradiol Feedback in Female Goldfish Reproduction. *Endocrinology* 158, 860–873. doi: 10.1210/en.2016-1550

Quan, J., Kang, Y., Luo, Z., Zhao, G., Ma, F., Li, L., et al. (2020). Identification and characterization of long noncoding RNAs provide insight into the regulation of gene expression in response to heat stress in rainbow trout (Oncorhynchus mykiss). *Comparat. Biochem. Physiol. Part D Genom. Proteom.* 36:100707. doi: 10.1016/j.cbd.2020.100707

Quinn, J. J., and Chang, H. Y. (2016). Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* 17, 47–62. doi: 10.1038/nrg.2015.10

Schmierer, B., Schuster, M. K., Shkumatava, A., and Kuchler, K. (2003). Activin a signaling induces Smad2, but not Smad3, requiring protein kinase a activity in granulosa cells from the avian ovary. *J. Biol. Chem.* 278, 21197–21203. doi: 10.1074/jbc.M212425200

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shi, B., Liu, X., Thomas, P., Pang, Y., Xu, Y., Li, X., et al. (2016). Identification and characterization of a progestin and adipoQ receptor (PAQR) structurally related to Paqr7 in the ovary of Cynoglossus semilaevis and its potential role in regulating oocyte maturation. *General Comparat. Endocrinol.* 237, 109–120. doi: 10.1016/j.ygcen.2016.08.008

Shi, M., Cheng, J., He, Y., Jiang, Z., Bodinga, B. M., Liu, B., et al. (2018). Effect of FH535 on in vitro maturation of porcine oocytes by inhibiting WNT signaling pathway. *Anim. Sci. J. Nihon Chikusan Gakkaiho* 89, 631–639. doi: 10.1111/asj.12982

Song, Y., Zheng, W., Zhang, M., Cheng, X., Cheng, J., Wang, W., et al. (2020). Out-of-season artificial reproduction techniques of cultured female tongue sole (Cynoglossus semilaevis): Broodstock management, administration methods of hormone therapy and artificial fertilization. *Aquaculture* 518:734866.

Sorbera, L. A., Asturiano, J. F., Carrillo, M., and Zanuy, S. (2001). Effects of polyunsaturated fatty acids and prostaglandins on oocyte maturation in a marine teleost, the European sea bass (Dicentrarchus labrax). *Biol. Reprod.* 64, 382–389. doi: 10.1095/biolreprod64.1.382

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U S A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Sulayman, A., Tian, K., Huang, X., Tian, Y., Xu, X., Fu, X., et al. (2019). Genome-wide identification and characterization of long non-coding RNAs expressed during sheep fetal and postnatal hair follicle development. *Sci. Rep.* 9:8501. doi: 10.1038/s41598-019-44600-w

Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., et al. (2013). Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* 41:e166. doi: 10.1093/nar/gkt646

Tafer, H., and Hofacker, I. L. (2008). RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics* 24, 2657–2663. doi: 10.1093/bioinformatics/btn193

Takahashi, T., Fujimori, C., Hagiwara, A., and Ogiwara, K. (2013). Recent advances in the understanding of teleost medaka ovulation: the roles of proteases and prostaglandins. *Zool. Sci.* 30, 239–247. doi: 10.2108/zsj.30.239

Tan-Wong, S. M., Dhir, S., and Proudfoot, N. J. (2019). R-Loops Promote Antisense Transcription across the Mammalian Genome. *Mol. Cell* 76, 600.e–616.e. doi: 10.1016/j.molcel.2019.10.002

Taylor, D. H., Chu, E. T., Spektor, R., and Soloway, P. D. (2015). Long non-coding RNA regulation of reproduction and development. *Mol. Reprod. Dev.* 82, 932–956. doi: 10.1002/mrd.22581

Thomé, R., dos Santos, H. B., Sato, Y., Rizzo, E., and Bazzoli, N. (2010). Distribution of laminin β2, collagen type IV, fibronectin and MMP-9 in ovaries of the teleost fish. *J. Mol. Histol.* 41, 215–224. doi: 10.1007/s10735-010-9281-7

Tokarz, J., Möller, G., Hrabě de Angelis, M., and Adamski, J. (2015). Steroids in teleost fishes: A functional point of view. *Steroids* 103, 123–144. doi: 10.1016/j.steroids.2015.06.011

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Tsai, M. C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J. K., Lan, F., et al. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 329, 689–693. doi: 10.1126/science.1192002

Tu, J., Chen, Y., Li, Z., Yang, H., Chen, H., and Yu, Z. (2020). Long non-coding RNAs in ovarian granulosa cells. *J. Ovarian Res.* 13:63. doi: 10.1186/s13048-020-00663-2

Tung, J. J., Hansen, D. V., Ban, K. H., Loktev, A. V., Summers, M. K., Adler, J. R. III, et al. (2005). A role for the anaphase-promoting complex inhibitor Emi2/XErp1, a homolog of early mitotic inhibitor 1, in cytostatic factor arrest of Xenopus eggs. *Proc. Natl. Acad. Sci. U S A.* 102, 4318–4323. doi: 10.1073/pnas.0501108102

Vadhan, A., Wang, Y. Y., and Yuan, S. F. (2020). EMI2 expression as a poor prognostic factor in patients with breast cancer. *Kaohsiung J. Med. Sci.* 36, 640–648. doi: 10.1002/kjm2.12208

Wang, J., Koganti, P. P., and Yao, J. (2020). Systematic identification of long intergenic non-coding RNAs expressed in bovine oocytes. *Reprod. Biol. Endocrinol. RB E* 18:13. doi: 10.1186/s12958-020-00573-4

Wang, X. Y., and Qin, Y. Y. (2019). Long non-coding RNAs in biology and female reproductive disorders. *Front. Biosci.* 24:750–764. doi: 10.2741/4748

Wu, S., Zhang, J., Liu, B., Huang, Y., Li, S., Wen, H., et al. (2020). Identification and Characterization of lncRNAs Related to the Muscle Growth and Development of Japanese Flounder (Paralichthys olivaceus). *Front. Genet.* 11:1034. doi: 10.3389/fgene.2020.01034

Xu, H., Cao, L., Wei, Y., Zhang, Y., and Liang, M. (2017). Effects of different dietary DHA:EPA ratios on gonadal steroidogenesis in the marine teleost, tongue sole (Cynoglossus semilaevis). *Br. J. Nutrit.* 118, 179–188. doi: 10.1017/s0007114517001891

Yang, C. X., Wang, P. C., Liu, S., Miao, J. K., Liu, X. M., and Miao, Y. L. (2020). Long noncoding RNA 2193 regulates meiosis through global epigenetic modification and cytoskeleton organization in pig oocytes. *J. Cell Physiol.* 235, 8304–8318. doi: 10.1002/jcp.29675

Yang, J., Zhang, Y., Xu, X., Li, J., Yuan, F., Bo, S., et al. (2019). Transforming growth factor-β is involved in maintaining oocyte meiotic arrest by promoting natriuretic peptide type C expression in mouse granulosa cells. *Cell Death Dis.* 10:558. doi: 10.1038/s41419-019-1797-5

Yang, L., Lin, C., Jin, C., Yang, J. C., Tanasa, B., Li, W., et al. (2013). lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs. *Nature* 500, 598–602. doi: 10.1038/nature12451

Yasuda, K., Hagiwara, E., Takeuchi, A., Mukai, C., Matsui, C., Sakai, A., et al. (2005). Changes in the distribution of tenascin and fibronectin in the mouse ovary during folliculogenesis, atresia, corpus luteum formation and luteolysis. *Zool. Sci.* 22, 237–245. doi: 10.2108/zsj.22.237

Yu, H., Wang, Y., Jin, C., Liu, Y., He, Y., and Zhang, Q. (2020). The functional differentiation of four smad4 paralogs in TGF-β signaling pathway of Japanese flounder (Paralichthys olivaceus). *Cell. Signal.* 71:109601. doi: 10.1016/j.cellsig.2020.109601

Zhao, H., Ge, J., Wei, J., Liu, J., Liu, C., Ma, C., et al. (2020). Effect of FSH on E(2)/GPR30-mediated mouse oocyte maturation in vitro. *Cell. Signal.* 66:109464. doi: 10.1016/j.cellsig.2019.109464

Zhao, Y., and Luck, M. R. (1995). Gene expression and protein distribution of collagen, fibronectin and laminin in bovine follicles and corpora lutea. *J. Reprod. Fertil.* 104, 115–123. doi: 10.1530/jrf.0.1040115

Zheng, W., Chu, Q., and Xu, T. (2021). Long noncoding RNA IRL regulates NF-κB-mediated immune responses through suppression of miR-27c-3p-dependent IRAK4 down-regulation in teleost fish. *J. Biol. Chem.* 2021:100304. doi: 10.1016/j.jbc.2021.100304

Zhou, X., Yi, Q., Zhong, Q., Li, C., Muhammad, S., Wang, X., et al. (2012). Molecular cloning, tissue distribution, and ontogeny of gonadotropin-releasing hormone III gene (GnRH-III) in half-smooth tongue sole (Cynoglossus semilaevis). *Comparat. Biochem. Physiol. Part B Biochem. Mol. Biol.* 163, 59–64. doi: 10.1016/j.cbpb.2012.04.010

Zhu, Y., Liu, D., Shaner, Z. C., Chen, S., Hong, W., and Stellwag, E. J. (2015). Nuclear progestin receptor (pgr) knockouts in zebrafish demonstrate role for pgr in ovulation but not in rapid non-genomic steroid mediated meiosis resumption. *Front. Endocrinol.* 6:37. doi: 10.3389/fendo.2015.00037

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 23 June 2021
doi: 10.3389/fgene.2021.671650

A Chromosome-Level Genome Assembly of the Mandarin Fish (*Siniperca chuatsi*)

Weidong Ding[1†], Xinhui Zhang[2†], Xiaomeng Zhao[2,3†], Wu Jing[1], Zheming Cao[1], Jia Li[2], Yu Huang[2,3], Xinxin You[2], Min Wang[4], Qiong Shi[2] and Xuwen Bing[1*]*

[1] Key Laboratory of Freshwater Fisheries and Germplasm Resources Utilization, Ministry of Agriculture, Freshwater Fisheries Research Center, Chinese Academy of Fishery Sciences, Wuxi, China, [2] Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, BGI Academy of Marine Sciences, BGI Marine, Shenzhen, China, [3] BGI Education Center, University of Chinese Academy of Sciences, Shenzhen, China, [4] BGI Zhenjiang Institute of Hydrobiology, Zhenjiang, China

The mandarin fish, *Siniperca chuatsi*, is an economically important perciform species with widespread aquaculture practices in China. Its special feeding habit, acceptance of only live prey fishes, contributes to its delicious meat. However, little is currently known about related genetic mechanisms. Here, we performed whole-genome sequencing and assembled a 758.78 Mb genome assembly of the mandarin fish, with the scaffold and contig N50 values reaching 2.64 Mb and 46.11 Kb, respectively. Approximately 92.8% of the scaffolds were ordered onto 24 chromosomes (Chrs) with the assistance of a previously established genetic linkage map. The chromosome-level genome contained 19,904 protein-coding genes, of which 19,059 (95.75%) genes were functionally annotated. The special feeding behavior of mandarin fish could be attributable to the interaction of a variety of sense organs (such as vision, smell, and endocrine organs). Through comparative genomics analysis, some interesting results were found. For example, olfactory receptor (OR) genes (especially the beta and delta types) underwent a significant expansion, and endocrinology/vision related *npy*, *spexin*, and *opsin* genes presented various functional mutations. These may contribute to the special feeding habit of the mandarin fish by strengthening the olfactory and visual systems. Meanwhile, previously identified sex-related genes and quantitative trait locis (QTLs) were localized on the Chr14 and Chr17, respectively. 155 toxin proteins were predicted from mandarin fish genome. In summary, the high-quality genome assembly of the mandarin fish provides novel insights into the feeding habit of live prey and offers a valuable genetic resource for the quality improvement of this freshwater fish.

Keywords: the mandarin fish (*Siniperca chuatsi*), whole-genome sequencing, feeding habit, molecular mechanism, chromosome-level genome assembly

OPEN ACCESS

Edited by:
Ka Yan Ma,
Sun Yat-sen University, China

Reviewed by:
Xinhai Ye,
Zhejiang University, China
Changwei Shao,
Chinese Academy of Fishery Sciences, China

****Correspondence:***
Xuwen Bing
bingxw@ffrc.cn
Xinxin You
youxinxin@genomics.cn

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *24 February 2021*
Accepted: *07 May 2021*
Published: *23 June 2021*

Citation:
Ding W, Zhang X, Zhao X, Jing W, Cao Z, Li J, Huang Y, You X, Wang M, Shi Q and Bing X (2021) A Chromosome-Level Genome Assembly of the Mandarin Fish (Siniperca chuatsi). Front. Genet. 12:671650. doi: 10.3389/fgene.2021.671650

INTRODUCTION

The mandarin fish, *Siniperca chuatsi*, belonging to the family Percichthyidae and order Perciformes, has a relatively high market value and widespread aquaculture throughout China (Liang and Cui, 1982; Liu et al., 1998). It has a special feeding habit, accepting only live prey fishes and refusing dead food items in the wild (Chiang, 1959; Liu et al., 1998). The feeding behaviors of

the mandarin fish require interactions of a variety of sense organs, such as eyes, mouth, lateral lines, and olfactory organs. Lateral-line may help alert the fish to vibrations that are made by nearby prey or approaching predators (Engelmann et al., 2000). Although the mandarin fish can feed properly on live prey fishes depending mainly on eyes and lateral-line, it can hunt prey fishes without these two organs (Liang et al., 1998). Researchers observed that the mandarin fish could recognize prey fishes using vision (Wu, 1988). A previous study (Liang et al., 1998) reported that the mandarin fish usually stayed more frequently near a perforated opaque cylinder containing live prey fishes rather than those without prey fishes, suggesting the importance of olfaction in searching for prey. However, this conclusion did not justify the food smells from other stimuli (such as hydromechanical stimulus).

Fish toxins have been poorly studied compared to venoms from other animals such as snakes, scorpions, spiders, and cone snails (Utkin, 2015). It is estimated that there are up to 2,900 venomous fishes (Xie et al., 2017) with venom systems convergently evolved 19 times (Harris and Jenner, 2019). Mandarin fish is one of those who can produce toxins in their hard spines to help them defense and prey, and cause pain and swelling at the site of the sting in human as well (Zhang F.-B. et al., 2019). However, apart from several antimicrobial peptides that can be regarded as toxins (Sun et al., 2007), there is no detailed report on venom genes and components of this fish yet.

In Mandarin fish species, females grow faster than males. Whether female mandarin fish have stronger predation ability is still unknown, so gender screening is of great significance to the cultivation of mandarin fish. So far, several gender-related molecular markers or functional genes have been screened, and even all-female mandarin fish have been bred (Guo et al., 2021; Liu et al., 2021). However, due to the lack of available genomic and transcriptome information, the mechanisms of sex differentiation remain poorly understood.

By far, genome data of the mandarin fish have been limited, which restricts genetic information for functional genomics studies. Therefore, in this study, we report a chromosome-level genome assembly of the mandarin fish using a combination of next-generation sequencing and previously reported genetic linkage map. The subsequent comparative genomic analysis provides novel insights into the feeding habit of live prey, toxin, and sex differentiation in the mandarin fish. This genome can not only serve as the genetic basis for in-depth investigations of fish evolution and biological functions but also offers a valuable genetic resource for quality improvement of this economically important fish.

MATERIALS AND METHODS

Sample Collection, Library Construction, and Sequencing

We collected muscle samples and extracted genomic DNA from a mandarin fish (**Figure 1**), which was obtained from Freshwater Fisheries Research Center of Chinese Academy of Fishery Sciences, Wuxi City, Jiangsu Province, China. The extracted DNA was used to construct seven libraries, including three short-insert (270, 500, and 800 bp) and four long-insert (2, 5, 10, and 20 kb) libraries. Subsequently, applying the routine whole-genome shotgun sequencing strategy, we sequenced these libraries on a Hiseq2500 platform (Illumina, San Diego, CA, United States). Those raw reads with adapters or low-quality sequences were filtered by a SOAPfilter (v2.2) (Luo et al., 2015).

All experiments were carried out following the guidelines of the Animal Ethics Committee and were approved by the Institutional Review Board on Bioethics and Biosafety of BGI, China (No. FT 18134).

Estimation of the Genome Size and Generation of a Genome Assembly

We performed a 17-mer distribution analysis to estimate the target genome size using the clean reads from the short-insert libraries (Liu et al., 2013). The calculation of genome size was based on the following formula: G = knum/kdepth. Here, knum is the sequenced k-mer number and kdepth is the k-mer sequencing depth. We set optimized parameters (pregraph-K 41 –d 1; contig –M 1; scaff –b 1.5) for the SOAPdenovo2 software (v2.04) to generate contigs and original scaffolds (Luo et al., 2012). Subsequently, we employed GapCloser (v1.12; with parameter settings of —t 8 —l 150) to fill the gaps of intra-scaffolds (Li et al., 2009) using the clean reads from short-insert libraries (270, 500, and 800 bp). Finally, we used BUSCO (Benchmarking Universal Single-Copy Orthologs; v1.22) to assess genome integrity (Simao et al., 2015).

Repeat Annotation, Gene Prediction, and Functional Annotation

Repetitive sequences including tandem repeats and transposable elements (TEs) were predicted from the assembled genome. We used Tandem Repeats Finder (v4.07) to search for tandem repeats (Benson, 1999). RepeatMasker (v4.0.6) and RepeatProteinMask (v4.0.6) were employed to detect known TEs based on the Repbase TE library (Tarailo-Graovac and Chen, 2009; Bao et al., 2015). Besides, we used RepeatModeler (v1.0.8) and LTR_FINDER (v1.0.6) with default parameters to generate the *de novo* repeat library (Xu and Wang, 2007; Abrusan et al., 2009), and RepeatMasker was applied to search the repeat regions against the built repeat library.

We utilized three different approaches to annotate structures of predicted genes in our assembled genome, including *de novo* prediction, homology-based prediction, and transcriptome-based prediction. For the *de novo* prediction, we employed AUGUSTUS (v3.2.1) and GENSCAN (v1.0) to identify protein-coding genes within the mandarin fish genome, using the repeat-masked genome as the template (Burge and Karlin, 1997; Stanke et al., 2006). For the homology-based prediction, we aligned the homologous proteins of five other fish species, including zebrafish (*Danio rerio*), three-spined stickleback (*Gasterosteus aculeatus*), Nile tilapia (*Oreochromis niloticus*), medaka (*Oryzias latipes*), and fugu (*Takifugu rubripes*) (downloaded from Ensembl 83 release), to the repeat-masked genome using tblastn (v2.2.26) with an E-value $\leq$ 1e-5 (Mount, 2007; Cunningham et al., 2015).

FIGURE 1 | The sampled mandarin fish (*Siniperca chuatsi*).

Subsequently, Solar (v0.9.6) and GeneWise (v2.4.1) (Birney and Durbin, 2000) were executed to define the potential gene structures for all alignments. The RNA-seq data from muscle tissues were aligned to the assembled genome also using tophat (v2.0.13) (Trapnell et al., 2009) and gene structures were predicted using cufflinks (v2.1.1) (Trapnell et al., 2012). Finally, we combined the above-mentioned three datasets to obtain a consistent and comprehensive gene set by GLEAN (v1.0) (Elsik et al., 2007).

These predicted coding proteins of the mandarin fish were aligned against public gene ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Swiss-Prot and TrEMBL databases for annotation of functions and pathways by using BLASTP (Kanehisa and Goto, 2000; Boeckmann et al., 2003; Harris et al., 2004). Subsequently, we applied InterProScan (v5.16−55.0) to identify functional motifs and domains through Pfam, PRINTS, ProDom, and SMART databases (Attwood, 2002; Letunic et al., 2004; Bru et al., 2005; Hunter et al., 2009; Finn et al., 2014).

Genome Evolution Analysis

Protein sequences of five ray-fin fishes, including spotted gar (*Lepisosteus osseus*), zebrafish, three-spined stickleback, medaka, and fugu were downloaded from the Ensembl (release-83) (Cunningham et al., 2015). Protein sequences of Asian arowana (*Scleropages formosus*, assembly fSclFor1.1) and giant-fin mudskipper (*Periophthalmus magnuspinnatus*, assembly fPerMag1.pri) were obtained from our recent works (You et al., 2014; Bian et al., 2016). Subsequently, OrthoMCL (v1.4) was executed to cluster the predicted mandarin fish genes into families (Li et al., 2003). We then identified and selected one-to-one orthologs from the above-mentioned eight teleost species, and finally used MUSCLE (v3.8.31) to perform multiple sequence alignment and PhyML (v3.0) to construct a phylogenetic tree (Edgar, 2004; Guindon et al., 2009, 2010).

Pseudo-Chromosome Construction

Single nucleotide polymorphisms (SNPs)-containing reads in the genetic linkage map of *S. chuatsi* (Sun et al., 2017) were mapped to our assembled mandarin fish genome, and the best hit reads were selected. Linkage groups (LGs) were assigned using the JoinMap4.1 software (Van Ooijen, 2006). Subsequently, a genetic linkage map of the mandarin fish was reconstructed, and SNPs in the genetic linkage map were used for assembling chromosomes. Based on genetic distances between these SNP markers, we determined the position and orientation of each scaffold and then anchored these scaffolds to construct pseudo-chromosomes.

To perform the genome synteny analysis, we downloaded genome sequences of European sea bass (*Dicentrarchus labrax*) from NCBI (Tine et al., 2014) as a reference. Genome-wide alignments were performed using lastz (Kurtz et al., 2004), and the best homology segments were selected using perl scripts. The final genomic synteny was visualized using the Circos software (Krzywinski et al., 2009).

Localization of Sex-Related Genes and Quantitative Trait Locis on Chromosomes

To identify candidate genes for underlying sex dimorphisms, we downloaded 81 putative sex-related genes from NCBI (Eshel et al., 2012, 2014; Zeng et al., 2016). The distribution of sex-related genes on chromosomes were determined by homologous sequence alignment.

Identification of *leptin*, Neuropeptide (*npy*), and *spexin* Genes From Teleost Fish Genomes

Protein sequences of three food intake genes, including leptin (NP_001122048.1), npy (AAI62071.1) and spexin (XP_010740053.1) protein sequences were downloaded from

NCBI. We performed tblastn (v2.26) (Mount, 2007) with default parameters searching against the mandarin fish, large yellow croaker (*Pseudosciaena crocea*; assembly L_crocea_2.0), grass carp (*Ctenopharyngodon idella*; assembly CI01), yellowtail (*Seriola dumerili*; assembly Sdu_1.0), kingfish (*Seriola lalandi*; assembly Sedor1), orange-spotted grouper (*Epinephelus coioides*), sea bass (*Lates calcarifer*; assembly ASM164080v1), and juvenile ovate pompano (*Trachinotus ovatus*; Zhang D.-C. et al., 2019) genome sequences. Subsequently, Genewise (v2.2.0) was applied to extract the best alignment results (Birney et al., 2004).

Identification of Olfactory Receptor and Taste Receptor Genes From Genome Sequences

We used zebrafish and pufferfish olfactory receptor (OR) protein sequences (**Supplementary Table 1**) as the queries to extract the OR genes in the Mandarin fish, zebrafish, fugu, stickleback, medaka, giant-fin mudskipper, Asian arowana and spotted gar (following the method mentioned in the previous section).

Protein sequences of five different types of taste receptor (TR) genes, including sour TR genes (*D. rerio*: ENSDARP00000119061), sweet TR genes (*D. rerio*: NP_001077325.1, NP_001034920.1, and *T. rubripes*: NP_001091094.1), umami TR genes (*D. rerio*: NP_001034614.2, NP_001034717.1 and *T. rubripes*: NP_001098687.1, NP_001072097.1), bitter TR genes (*D. rerio*: ENSDARG00000079880), salty TR genes (*Homo sapiens*: NP_001153048.1, NP_000327.2, and NP_001030.2), were downloaded from the NCBI or Ensembl database. We performed tblastn (Blast v2.26; Pevsner, 2005) with default parameters to search against these genome sequences. Subsequently, Genewise v2.2.0 (Birney et al., 2004) was employed to extract the best alignment results.

Proteins sequences were aligned by the MAFFT (v7.237) program (Katoh et al., 2002) with the einsi module. Phylogenetic trees were constructed using the PhyML (v3.0) program with bootstrap set to 1,000 (Guindon et al., 2010).

Identification of *opsin* Genes From Ray-Finned Fish Genomes

In this study, we chose eight teleost genomes to extract opsin protein sequences, including the mandarin fish, Asian arowana, mudskipper, spotted gar, medaka, stickleback, fugu, and zebrafish. Protein sequences of opsin genes (LWS-1: ENSDARP00000065940, LWS-2: ENSDARP00000149112, SWS-1: ENSDARP00000067159, SWS-2: ENSDARP00000144766, RH1: ENSDARP00000011562, RH2-1: ENSDARP00000001158, RH2-2: ENSDARP00000011837, RH2-3: ENSDARP00000001943, and RH2-4: ENSDARP00000000979) from zebrafish were downloaded from the Ensembl database as the queries. We performed tblastn (v2.2.28) (Mount, 2007) to align these sequences. Finally, Exonerate (v2.2.0) (Slater and Birney, 2005) was employed to predict the perfect alignment results. Multiple sequence alignment of these predicted *opsin* genes was performed with the Muscle module in MEGA (v 7.0) (Kumar et al., 2016). They were then translated into protein sequences for phylogenetic analyses. Phylogenetic trees were constructed using the PhyML (v3.0) program with bootstrap set to 1,000 (Guindon et al., 2010).

Venom Proteins Prediction

Protein sequences of animal venoms and toxins (**Supplementary Table 2**) were downloaded from UniProtKB/Swiss-Prot (UniProt Consortium, 2018) via the Animal Toxin Annotation Project (Jungo et al., 2012). These protein sequences were then filtered and only reviewed references (7,093 in total) were maintained as the trust-worthy input queries for searching. Firstly, we blasted the reviewed toxins against the coding sequences (CDS) predicted from the mandarin fish genome assembly using blastp (Camacho et al., 2009) with an e-value of 1e-10. Subsequently, the mapped sequences of mostly partial or fragmented genes with aligning ratios less than 75% were discarded, and the remaining 195 hits were further filtered manually according to the constrained lengths of the venom sequences within the same family, conserved patterns (e.g., disulfide bonds) and other post-translational modifications (PTMs).

RESULTS

Genome Sequencing and Assembly

Seven libraries including three short-insert (270, 500, and 800 bp) and four long-insert (2, 5, 10, and 20 kb) were constructed to generate a total of 327 Gb raw reads (**Supplementary Table 3**). Subsequently, these raw data were filtered, and 233 Gb clean data were obtained for subsequent genome assembly.

We calculated the genome size using the following formula: G = knum/kdepth. Here, the knum (i.e., k-mer number) was 43,888,350,480 and the kdepth (k-mer depth) was 59. Therefore, the estimated genome size of the mandarin fish is about 743.87 Mb (**Supplementary Table 4** and **Supplementary Figure 1**).

We generated contigs and original scaffolds by paired-end reads to assemble the mandarin fish genome. After filling the gaps of intra-scaffolds, we obtained a 758.78-Mb genome assembly for the mandarin fish, with contig and scaffold N50 values of 46.11 Kb and 2.64 Mb, respectively (**Table 1**).

Using BUSCO analysis to determine the completeness of our assembly, it is found that the assembly contained 86.1% complete, 3.0% duplicated, 9.3% fragmented, and 3.7% missed BUSCOs. Besides, 74.6% of the clean reads of RNAseq could be mapped to

TABLE 1 | Summary of the mandarin genome assembly and annotation.

Genome assembly	Contig N50 size (kb)	46.11
	Scaffold N50 size (Mb)	2.64
	Assembled genome size (Mb)	758.78
	Genome coverage (X)	430.83
	The longest scaffold (bp)	16,398,010
Genome characteristics	GC content	39.2%
	Gene number	19,904
	BUSCOs (complete in total)	86.1%

the genome assembly. These results suggested that our genome assembly was relatively complete.

Chromosome-Level Genome Assembly

Based on the previously reported genetic linkage map of the mandarin fish (Sun et al., 2017), we anchored a total of 518 scaffolds into 24 chromosomes (Chr). A total of 697.06 Mb was assembled, corresponding to 92.8% of the assembled genome and 18,752 genes (from a total of 19,904 genes). The largest chromosome was Chr10 with 37.87 Mb in length containing 56 scaffolds, and the smallest was Chr17 with 19.19 Mb containing 22 scaffolds. The average chromosome length was 29 Mb with 21 scaffolds (**Table 2** and **Figure 2A**).

There were 39,689 synteny blocks (>2 kb) between the assembled genomes of the mandarin fish and the reported European sea bass (Tine et al., 2014). We observed that almost all chromosomes showed the 1:1 synteny relationship, with an exception of Chr21 in the mandarin fish that aligned to two seabass chromosomes (**Figure 2B**). The results of collinearity between mandarin fish and European sea bass indicate that our chromosome assembly results are reliable.

Genome Annotation

Repeat sequences were identified based on homology search against the Repbase database and *de novo* prediction. We predicted that the mandarin fish genome contained 26.3% of repetitive elements. Compared with other perciforme fish, the mandarin fish was lower than red-spotted grouper (43.02%), giant grouper (45.1%), but much higher than large yellow croaker (18.1%) and golden pompano (20.25%) in the repeat sequence percentage. The most abundant TEs were long interspersed elements (13.96% of the genome), followed by DNA transposons (9.66%) and long terminal repeats (LTRs, 5.04%) (**Supplementary Table 5**).

Based on the genome with repeated elements masked, we integrated homology searching, *de novo,* and transcript methods to predict that the mandarin fish genome had 19,904 protein-coding genes (**Supplementary Table 6**), of which 19,059 (95.75%) genes were functionally annotated by at least one of the InterPro, GO, KEGG, Swiss-Prot, and TrEMBL protein databases (**Supplementary Table 7**). To estimate the completeness of our annotated genes, we determined that the annotated genes contained 88.9% complete, 3.2% duplicated, 6.5% fragmented, and 4.6% missed BUSCOs.

TABLE 2 | Summary of the assembled chromosomes of the mandarin fish.

Chr	Length (Mb)	Mapped scaffolds	Mapped genes
Chr1	30.26	14	930
Chr2	27.62	24	750
Chr3	26.07	42	505
Chr4	25.64	23	723
Chr5	22.01	39	480
Chr6	28.28	23	984
Chr7	29.34	35	906
Chr8	32.52	20	834
Chr9	30.54	15	710
Chr10	37.87	15	817
Chr11	30.42	22	843
Chr12	33.69	35	977
Chr13	30.45	16	922
Chr14	32.91	21	892
Chr15	36.19	17	1,030
Chr16	28.49	10	776
Chr17	19.19	22	405
Chr18	26.51	19	706
Chr19	29.45	21	892
Chr20	29.82	19	732
Chr21	29.59	16	960
Chr22	24.33	11	563
Chr23	25.99	23	554
Chr24	29.85	16	861
Total	697.06	518	18,752

Phylogenetic Analysis

To establish the phylogenetic position of the mandarin fish, we compared the genomes of the mandarin fish and seven other teleost fishes. We found that 16,922 orthologous gene families were shared among the eight teleost fishes, and identified 3,510 single-copy orthologs genes that were used to construct a phylogenetic tree (**Figure 3**). It appears that stickleback was most closely related to the mandarin fish. We selected the specific gene family in mandarin fish and they were functionally annotated by KEGG protein database (**Supplementary Table 8**). Finally, there were 74 specific families in mandarin fish, containing 194 genes. According to KEGG annotation, there were 45 genes associated to 205 pathways.

Localization of Sex-Related Genes and QTLs on Chromosomes

Of the 81 sex-related genes, 19 genes were located on Chr14 (13 clustered in **Figure 4A**). In a previous study (Sun et al., 2017), five QTLs for sex determination (SD) were detected on LG23 (Sun et al., 2017) and thereby localized on Chr17 of the mandarin fish genome (clustered between 0 and 4 Mb; **Figure 4B**). Both r2_42410 and r2_237649 were located within the receptor-type tyrosine-protein phosphatase-like N (*ptrpn*) and SH3 domain-containing YSC84-like protein 1 (*sh3yl1*), respectively. The other three QTLs were located in the intergenic regions (**Figure 4B**). Genotypes of all the male and female fishes on r1_33008 were homozygous and heterozygous respectively, which was reported previously (Sun et al., 2017). Subsequently, we validated the marker r1_33008 (Sun et al., 2017) in another group (**Figure 4B**) and found that there was no difference between male and female, which may be unique in the genetic linkage map population.

Analysis of Genes for Food Intake

In fishes, feeding behaviors are usually regulated by specific regions in the brain, the so-called feeding centers, which are under the influence of hormones produced by the brain and the periphery (Volkoff, 2016). The mandarin fish has a peculiar feeding habit of only accepting live prey

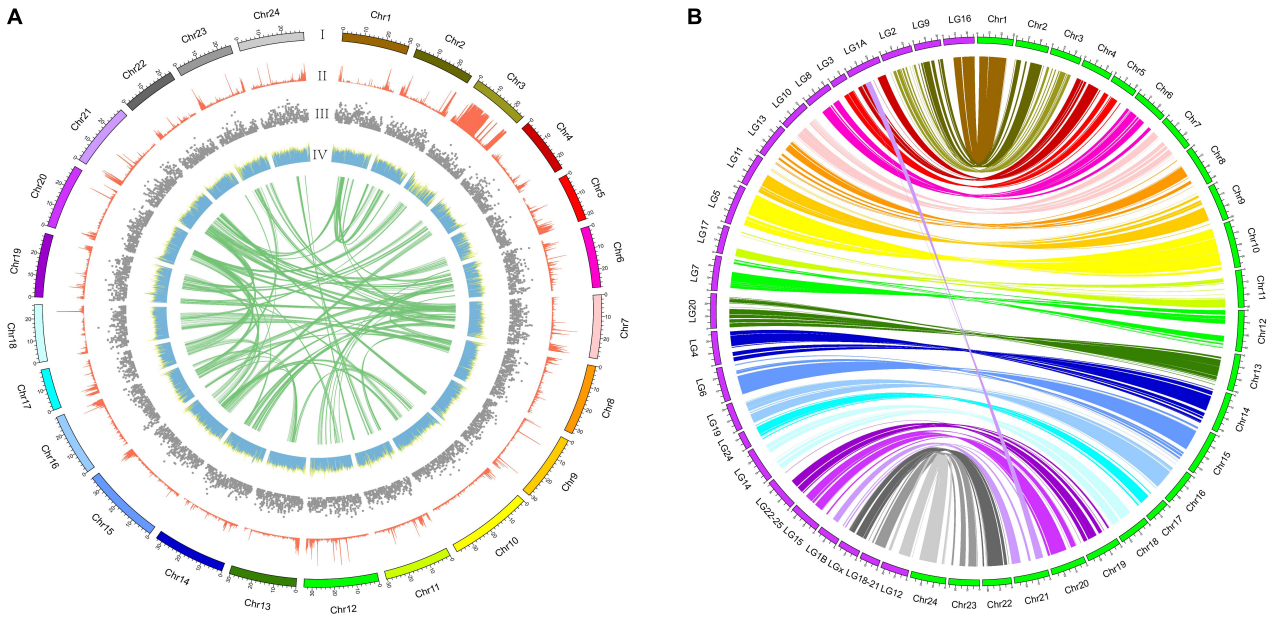

FIGURE 2 | Chromosome-level genome assembly of the mandarin fish. **(A)** A Circos atlas presents details of the pseudo-chromosome information from outside to inside: (I) the length of each pseudo-chromosome, (II) density of SNP distribution in each 100-kb genomic interval, (III) density of gene distribution in each 100-kb genomic interval, (IV) GC content of 100-kb genomic intervals, and (V) schematic presentation of major inter-chromosomal relationships in the mandarin fish genome. **(B)** A synteny comparison between mandarin fish and seabass genomes (chr: mandarin and LG: seabass) revealed high accuracy of our assembled genome for the mandarin fish.

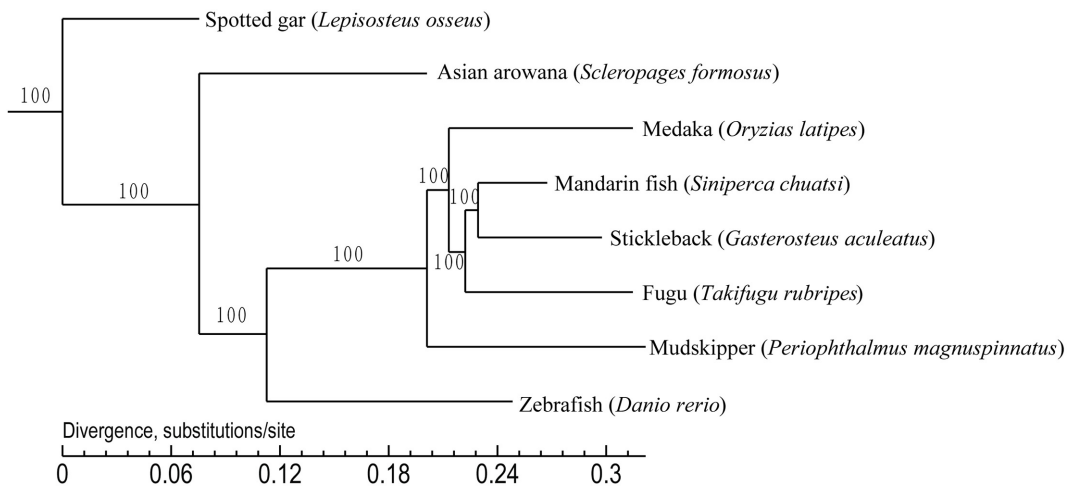

FIGURE 3 | Phylogenetic tree of eight teleost fishes. It was constructed using the shared 3,510 single copy orthologous genes. Spotted gar was used as the outgroup.

fishes and refusing artificial diets or dead prey fishes. It is almost unknown about any genes for regulation of this unique food preference (Liu et al., 1998; Li et al., 2013). In our present study, several candidate genes for food intake were analyzed.

leptin is an important hormone involved in the regulation of food intake and energy balance (Kurokawa et al., 2005). Our synteny analysis of four representative fishes (mandarin, large yellow croaker, grouper and grass carp) (**Figure 5**) indicated that the upstream and downstream of the *leptin* genes are proline-rich transmembrane protein 4 (*prrt4*), transmembrane protein 53 (*tmem53*), RNA-binding protein 28 (*rbm28*) and Leucine-rich repeat-containing protein 4 (*lrrc4*), hepatocyte growth factor (*hgf*), voltage-dependent calcium channel subunit alpha (*cacng-α*) genes respectively, which is consistent with a previous report (Kurokawa et al., 2005). Compared with the other three species, the upstream genes {[F-actin]-monooxygenase MICAL3 (*mical3*) and zinc finger BED domain-containing protein 1 (*zbed1*)} of

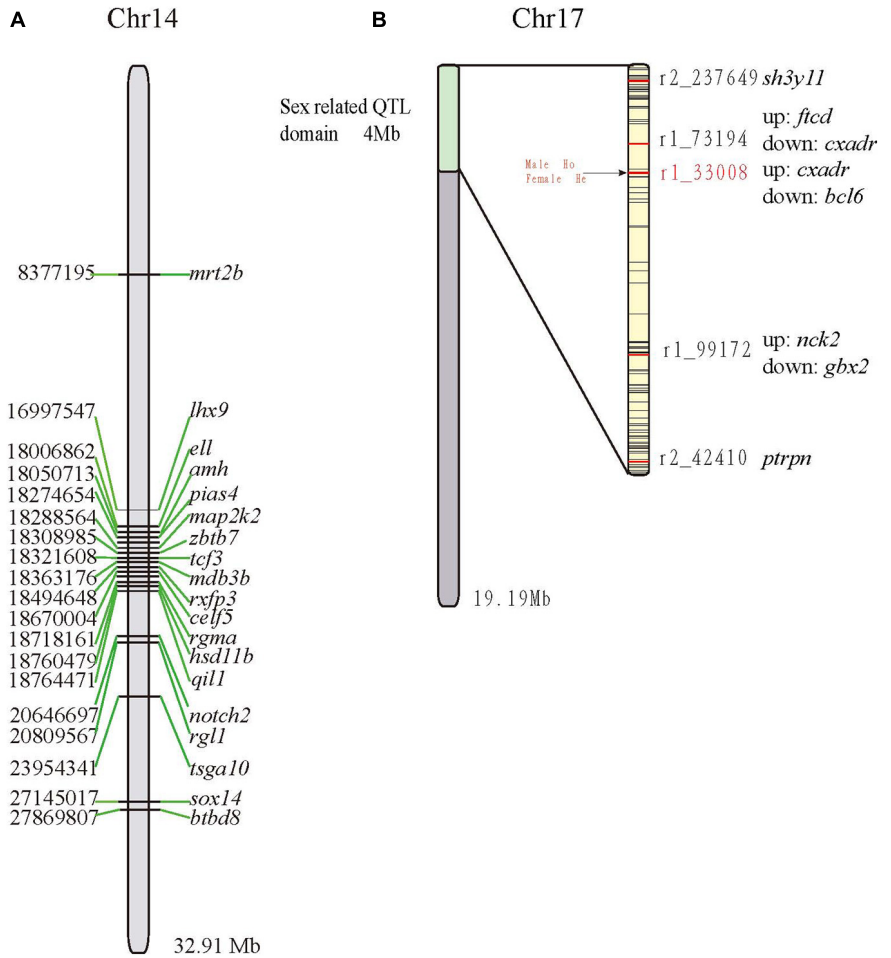

FIGURE 4 | Localization of sex-related genes and QTLs on the Chr14 **(A)** and Chr17 **(B)**, respectively. Abbreviations of genes: *amh*, anti-mullerian hormone; *btbd8*, BTB (POZ) domain containing 8; *celf5*, CUGBP Elav-like family member 5; *dmrt2b*, doublesex and mab-3 related transcription factor 2b; *ell*, RNA polymerase II elongation factor ELL; *hsd11b*, hydroxysteroid 11-beta-dehydrogenase 1-like protein; *lhx9*, LIM homeobox 9; *map2k2*, dual specificity mitogen-activated protein kinase kinase 2; *notch2*, notch homolog 2; *pias4*, Protein inhibitor of activated STAT 4; *qil1*, QIL1; *rgl1*, ral guanine nucleotide dissociation stimulator-like 1; *rxfp3*, relaxin family peptide receptor 3; *tcf3*, E2A-1 transcription factor; sox14, SRY-box transcription factor; *zbtb7*, zinc finger and BTB domain containing 7.

grass carp were not conserved (**Figure 5**). In a previous study (He et al., 2013), a typical *leptin* gene was reported to be composed of three exons and two introns, but the mandarin fish *leptin* gene consisted of two exons and one intron. However, our genomic results confirmed that the *leptin* structure of the mandarin fish is in fact consistent with the other three representative fishes (grouper, large yellow croaker, and grass carp) (**Supplementary Figure 2**). We thereby propose that the errors in the previous study are due to a shortage of whole genome sequence of the mandarin fish at that time.

Neuropeptide Y (*npy*), belonging to the *npy* family, is abundant in the central nervous system (Volkoff, 2006; Holzer et al., 2012). *npy* has been implicated in several centrally mediated physiological functions, such as regulation of body temperature, sexual behavior, energy homeostasis, anxiety, mood, and neuroendocrine secretions (Holzer et al., 2012). Moreover, *npy* is one of the most abundant neuropeptides within the brain and has a major regulatory role in food intake (Yokobori et al., 2012; Zhou Y. et al., 2013). The *npy* gene of the mandarin fish on Chr 2 was comprised of three exons and two introns, which is consistent with other fishes. Mandarin fish compared to six perciforme fishes, NH2-terminal signal peptide (red box, **Supplementary Figure 3**) is variable, mature peptide (Black underline, **Supplementary Figure 3**) is highly conserved. However, its COOH-terminal domain had two significant variant sites (red asterisks in **Supplementary Figure 3**).

Spexin was identified in mammalian adipose tissue. It plays a significant role in the regulation of energy metabolism and food intake (Walewski et al., 2014; Zheng et al., 2017). Its expression is up-regulated in food deprivation and down-regulated in obese rats and humans, suggesting suppression of the orexin in the hypothalamus (Li et al., 2016). In the present study, we performed sequence analysis and found that the *spexin* gene is composed of six exons and five introns, and the amino acids of the mature peptide (spexin-14; **Supplementary Figure 4a**) in the mandarin fish are identical to that of the grouper (Li et al., 2016). Sequence

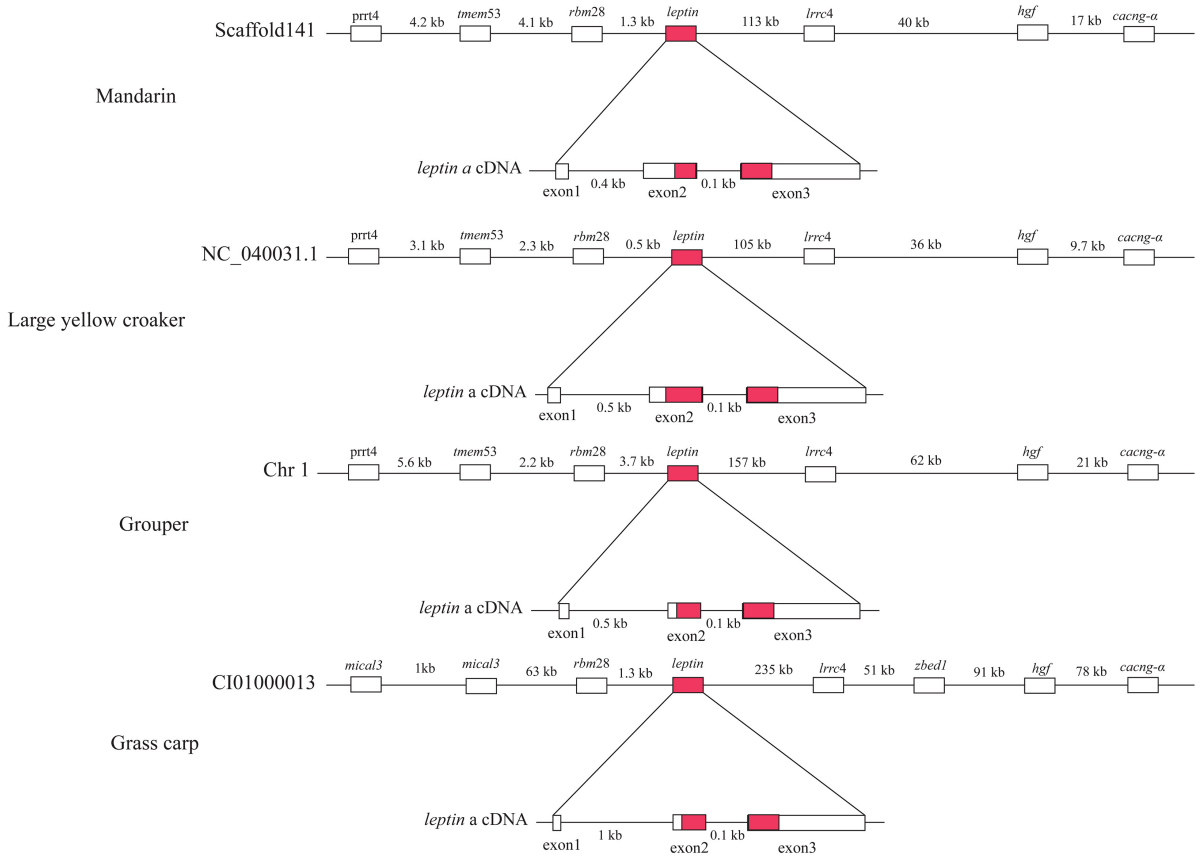

FIGURE 5 | Comparison of *leptin* gene structures among mandarin fish, large yellow croaker, grouper and grass carp. Red marked areas represent open reading frames.

alignment of mandarin fish *spexin* with the other six perciforme fishes reveals that the NH_2-terminal signal peptide is highly variable, and its COOH-terminal domain had two significant variant sites (red asterisks). In contrast, the region covering the *spexin* mature peptide (black box) together with the dibasic processing sites flanking the two ends (RR and GRR, black triangle) is highly conserved (**Supplementary Figure 4b**).

Olfactory Receptor Genes in Teleost Fishes

We identified 133 OR genes in the mandarin fish genome (**Figure 6A**), including 119 functional genes and 13 pseudogenes. The numbers are different from those reported in zebrafish (102 functional genes and 35 pseudogenes) and pufferfish (44 functional genes and 54 pseudogenes) genomes (Niimura and Nei, 2005). We examined zebrafish (Ensembl version: GRCz11) and fugu genomes (Ensembl version: FUGU5) and identified 109 functional genes and 6 pseudogenes in the zebrafish and 72 functional genes and 12 pseudogenes in the fugu, respectively.

We found that the mandarin fish had more OR functional genes than other examined teleosts, except for the spotted gar that was diverged from teleosts before the teleost-specific genome duplication (TGD). Spotted gar has 36 functional genes ascribing to Groups α and γ, which were mostly absent in teleosts. The mandarin fish had the largest numbers of Group β (n = 8) and group δ (n = 69) functional genes among other teleosts. In a previous study (Lv et al., 2019), researchers indicated an expansion of Group β OR genes in the mandarin fish, which was confirmed in our present work (**Figure 6C**). We extracted more Group β OR genes (eight in Chr8) than a previous report (six in Reference (Niimura and Nei, 2005); see **Figures 6B,C**).

Taste Receptor Genes in Teleost Fishes

Taste receptor type 1family (*tas1r*), belonging to the G protein-coupled receptor (*gpcr*), plays a central role in the reception of sweet and umami taste in many vertebrates. A *tas1r2* + 3 heterodimer was identified as the sweet TR (Nelson et al., 2001; Li et al., 2002). Tas1r3 may serve as a receptor for high sucrose concentrations (Zhao et al., 2003). A *tas1r1* + 3 heterodimer and multiple combinations of *tas1r2* with *tas1r3* were identified as a tuned L-amino acid TR in fish (Oike et al., 2007). We extracted three *tas1r* genes in the mandarin fish, Asian arowana and spotted gar (**Table 3** and **Figure 7**). It seems that the gene numbers responding to sweet and umami tastes in the mandarin fish is more primitive since they are much closer to ancient fishes (such as arowana and gars).

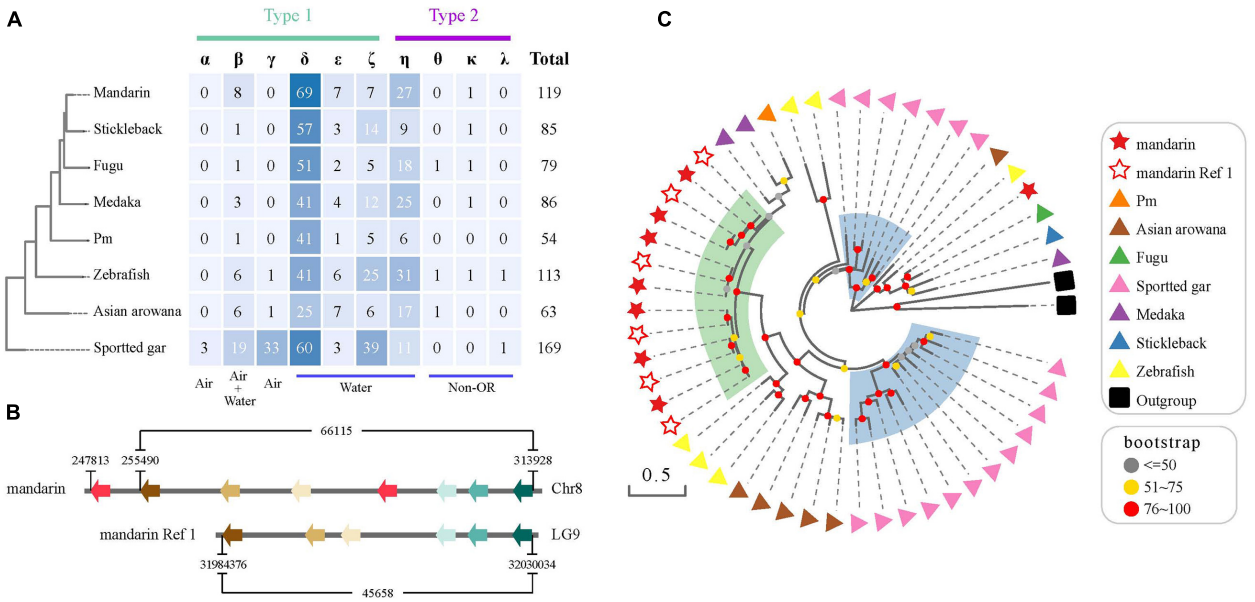

FIGURE 6 | The OR gene family in the mandarin fish genome. **(A)** OR genes in the mandarin fish in comparison with seven other teleost fishes. "Air" and "water" represent the detection of airborne and water-soluble odorants, respectively. **(B)** A comparative genomic analysis of Group β OR functional genes in our genome assembly and the assembly of Ref 1. **(C)** A phylogenetic tree of Group β functional genes in the mandarin fish and seven other teleost fishes. Stars indicate the Group β OR functional genes of the mandarin fish [from our present work and Niimura and Nei (2005)]. Triangles with different colors represent various resources of the other seven teleost fishes. Green transparent rainbow indicates an expansion of Group β OR functional genes in the mandarin fish. Blue transparent rectangles indicate the expansion of Groupe β OR functional genes in spotted gar.

TABLE 3 | Numbers of taste receptor genes in these species.

Group	Mandarin	Zebrafish[I]	Fugu[II]	Stickleback	Medaka	PM	Arowana	Spotted gar
Sour	1	1(1)	1	1	1	1	1(1)	1(2)
Sweet-Umami	3	4	4	9	5	4	3	3
Bitter	1(1)	7	4	3	1(1)	1	5(4)	1(1)
Salty	0	0	0	0	0	0	0	0
Total	5(1)	12(1)	9	13	7(1)	6	9(5)	5(3)

[I]*The version of zebrafish genome is GRCz11.*
[II]*The version of fugu genome is FUGU5.*

Bitter taste preference was likely recognized as a mechanism for avoiding toxic foods. Bitter foods evoke innate aversive behaviors in many animals. Taste receptor type 2 (*tas2r*) family was identified as bitter TRs in mammals. Most vertebrate species have several *tas2r* genes, and their copy numbers varied among various species. In our present study, we identified only one intact *tas2r* gene in the mandarin fish, giant-fin mudskipper, and spotted gar (**Figure 7C**), suggesting that these three species possibly have a low ability to distinguish the bitter foods. However, compared with the other seven teleost fishes, the mandarin fish showed no difference in number of genes for responding to salty and sour tastes (**Table 3**).

Comparison of Mandarin Fish *opsin* Genes With Other Teleost Fish

A total of 50 *opsin* nucleotide sequences, including 14 *RH1* (rhodopsin), 14 *RH2* (green-sensitive), 4 *SWS1* (short wavelength-sensitive 1), 7 *SWS2* (short wavelength-sensitive 2), and 11 *LWS* (long wavelength-sensitive), were successfully derived from eight teleost fishes (**Figure 8A**). To understand the evolutionary relationships among these opsins in teleosts, we constructed a phylogenetic tree (**Figure 8B**) using spotted gar LWS as the outgroup. Obviously, all opsin could be divided into five main clades (RH1, RH2, SWS1, SWS2, and LWS) in the eight examined teleost species.

In this study, we identified six *opsin* genes in the mandarin fish genome (**Figure 8A**), including two *RH1*, one *RH2*, one *SWS2,* and one *LWS*. A previous study (Neafsey and Hartl, 2005) reported loss of *SWS1* genes in fugu and mudskipper (You et al., 2014) genome. We tried to extract *SWS1* sequences in the examined teleosts, but could not find *SWS1* in the mandarin fish, medaka, fugu, and giant-fin mudskipper either. This loss of *SWS1* could be an adaptation to minimize retinal damage from ultraviolet. We identified two *LWS* genes in the mandarin fish, zebrafish and arowana, while only one *LWS* in other

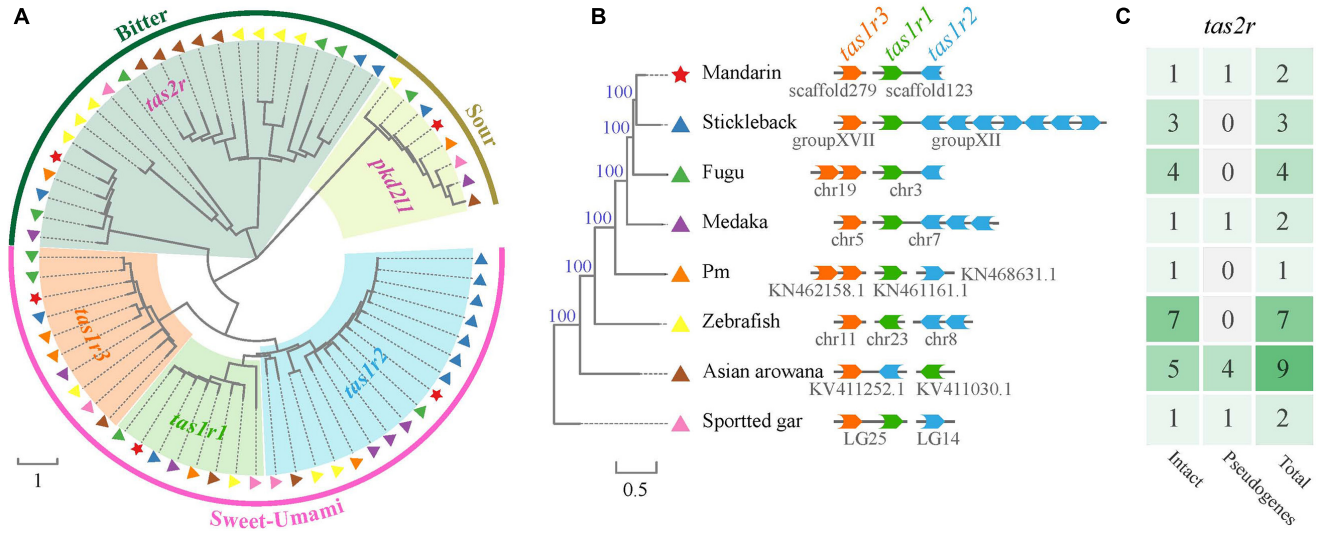

FIGURE 7 | A phylogenetic tree of taste-related genes in the mandarin fish and seven other teleost fishes. **(A)** Comparisons of taste-related genes in the mandarin fish and seven other teleost fishes. Red stars indicate the gens of the mandarin fish. Other symbols indicate the sources of other species. **(B)** Genomic locations of *tas1r1*, *tas1r2*, and *tas1r3* genes in our genome assembly. **(C)** Summary of the copy number of *tas2r* gene in different species [corresponding to those fishes in panel **(B)**].

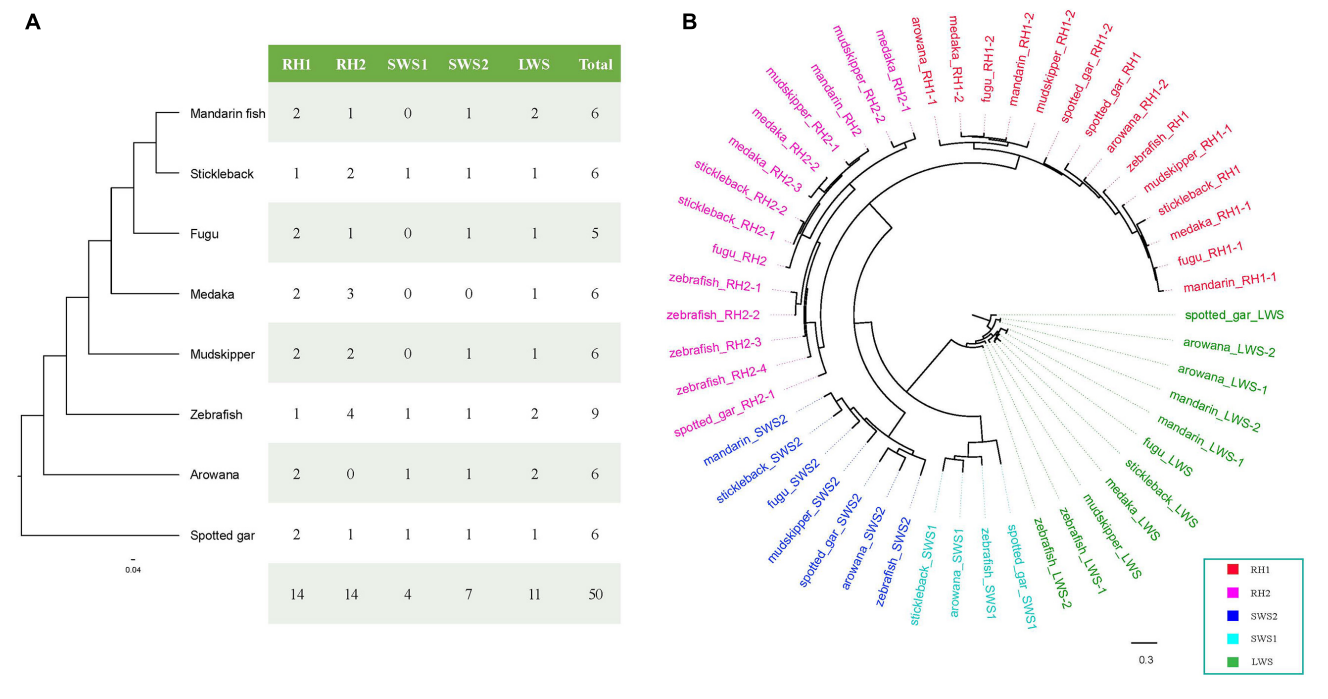

FIGURE 8 | *opsin* genes in the mandarin fish genome. **(A)** *opsin* genes in the mandarin fish and seven other teleost fishes. **(B)** A phylogenetic tree of *opsin* genes in the mandarin fish and seven other teleost fishes. LWS, long wavelength-sensitive; RH1, rhodopsin; RH2, green-sensitive; SWS1, short wavelength-sensitive 1; SWS2, short wavelength-sensitive 2.

fishes. In zebrafish, two *LWS* genes locating in tandem encode different protein sequences (Chinen et al., 2003). However, in the mandarin fish and arowana, two *LWS* genes were also in tandem but encoded the identical protein sequences. According to **Figure 8B** and **Supplementary Figure 4**, m and arin LWS-1 and LWS-2 were completely same in amino acid sequences. And arowana LWS-1 and LWS-2 had same sequences but they were different in zebrafish.

The primary amino acid sequence is very important for *opsin* molecular properties. In this study, we compared their

sequences and identify the differences among the eight examined fishes. We identified some amino acid changes in the mandarin fish that are probably critical for wavelength absorption. Here, we observed five specific sites in the mandarin fish, with significant differences from the other seven fish species (see more details in **Supplementary Figure 5**). We identified the transmembrane domains of *LWS* with TMHMM Server (Version 2.0)[1] (**Supplementary Figure 5**). In the mandarin fish, two specific sites are in the transmembrane domains, L98 and T219, which are potentially important for light absorption. *LWS* is often used for red vision, and shallow water receives more red light. The freshwater fishes have more *LWS* genes than those in seawater (Lin et al., 2017). The more *LWS* genes and several specific sites in transmembrane domains help fishes be more sensitive to light and live prey.

Toxin Genes Were Identified in the Mandarin Fish Genome

Fish toxins have been poorly studied compared to venoms from other animals such as snakes, scorpions, spiders, and cone snails (Utkin, 2015). It is estimated that there are up to 2,900 venomous fishes (Xie et al., 2017) with venom systems convergently evolved 19 times (Harris and Jenner, 2019). Mandarin fish is one of those who can produce toxins in their hard spines to help them defense and prey, and cause pain and swelling at the site of the sting in human as well (Zhang F.-B. et al., 2019). However, apart from several antimicrobial peptides that can be regarded as toxins (Sun et al., 2007), there is no detailed report on venom genes and components of this fish yet.

In this study, a total of 155 toxin proteins were predicted from the mandarin fish genome assembly. They ranged from 87 to 1,895 amino acids (aa), with more than half of them less than 300 aa (**Supplementary Figure 6**). Unlike a vast number (125) of short-length "fragmented" venoms (less than 100 aa) in the Chinese yellow catfish genome (Zhang et al., 2018), there were only two short-length venoms in the mandarin genome, with a length of 87 and 98 aa, respectively. Consistent with the common findings that most toxins are short peptides, the majority (96; 62%) of our predicted toxin proteins had an entire length between 100 and 300 aa.

Among the 155 putative venom proteins, 144 were classified into 37 families, with 11 unclassified toxins (**Supplementary Figure 7**). The top four biggest groups in these toxins included peptidase S1, venom metalloproteinase (M12B), Type-B carboxylesterase, and calmodulin, consisting of 27, 13, 10, and 9 toxin genes respectively. Interestingly, several fish-specific toxins were identified, including SC_GLEAN_10016806 and SC_GLEAN_10016808 that belonged to the stonustoxin (SNTX-a), SC_GLEAN_10016805, and SC_GLEAN_10016807 annotated as SNTX-β. SNTX is a soluble heterodimeric assembly of α and β subunits that share a sequence identity of ∼50% (Ellisdon et al., 2015). It was firstly isolated from the stonefish (Poh et al., 1991) and has been proved to induce platelet aggregation and hemolytic activity (Khoo et al., 1992) and also function as a neurotoxin (Low et al., 1994). The existence of two copies of both SNTX-α and SNTX-β suggesting the probability of the forming of active and functional toxins in the mandarin fish. SNTXs, along with all other toxins identified in this assembled genome, showed the great potential of discovering new drugs.

[1]http://www.cbs.dtu.dk/services/TMHMM/

DISCUSSION

Our high-quality genome assembly of the mandarin fish could provide opportunities to understand SD, special food intake, or other biological processes at the genome level in this economically important fish. The final assembled genome was 758.78 Mb, and approximately 92.8% of the scaffolds were ordered onto 24 chromosomes. The mandarin fish has been widely cultivated in China, with a special feeding habit of accepting only live prey fishes for its delicious meat. However, little is currently known about related genetic mechanisms. Uncovering the molecular mechanisms for regulation of feeding behaviors may not only lead to specific adjustments in fish culture conditions and feeding strategies but also gradually instruct us to develop new technologies to improve feeding, food conversion efficiency and the growth of aquaculture fishes (Volkoff et al., 2010; Zhou Y. et al., 2013).

In fact, feeding is a complex of behaviors, including at least food intake itself and foraging or appetite behavior. Eating is ultimately regulated by the central feeding center in the brain (Keen-Rhinehart et al., 2013; Woods et al., 2014). It also processes information from endocrine signals from the brain and the surrounding environments. These endocrine signals include various hormones. For example, *npy* and *spexin* are two important hormones involved in the regulation of food intake and energy balance (Zheng et al., 2017; Zhou Y. et al., 2013). The latest research suggests that smell might regulate appetite through *npy* in yellowtail (Senzui et al., 2020). *npy* as a neuromodulator in the olfactory epithelium and intensified the activity of OR neurons and olfaction (Negroni et al., 2012; Senzui et al., 2020). In our present study, we observed that the amino acid sequence of *npy* and *spexin* showed a high level of conservation, when compared with the other six examined Perciformes fishes. It seems that *npy* and *spexin* are conserved neuropeptides in fish evolution with important physiological functions. However, the *npy* and *spexin* genes of the mandarin fish had significant variations at the C-termini of the protein sequences (**Supplementary Figures 2, 3**), which may be related to the special diet of the mandarin fish.

Olfaction is also crucial for animals to find foods and to judge whether potential foods are edible or not (Chandrashekar et al., 2000; Nei et al., 2008). It is controlled by a large family of OR genes. Fishes also have this gene family, but the number of genes is much less than mammals (Niimura and Nei, 2006). Previous studies have demonstrated that the beta type OR genes are presented in both aquatic and terrestrial vertebrates, indicating that these receptors detect both water-soluble and airborne odorants (Niimura, 2009); however, delta type OR genes are only in aquatic organisms (You et al., 2014). In the present study, we

determined that the mandarin fish had the largest numbers of Group β (n = 8) and group δ (n = 69) functional genes than the other teleost fishes (**Figure 6A**), which might contribute to its particular carnivorous diet.

Vision is very important for animals because it plays important roles in foraging, mating, information transmission, and escaping from predators (Yokoyama, 2000). Based on their amino acid compositions, *opsin* genes are classified into five common clusters: *RH1* (rhodopsin), *RH2* (rhodopsin-like or the green light-sensitive pigments), *SWS1* (short wavelength−, or the UV or violet light-sensitive pigments), *SWS2* (*SWS1*-like or the blue light-sensitive pigments); *LWS/MWS* (long wavelength- or middle wavelength-sensitive, or the red- and green-sensitive pigments) (Yokoyama, 2000; Shichida and Matsuyama, 2009). *opsin* diversity is usually generated by gene duplication and/or accumulation of mutations. *MWS/LWS* opsins have peak values of light absorption (Terai et al., 2002). The light sensitivity of a visual pigment is determined not only by the chromophore itself, but also by its interaction with the amino acid residues lining the pocket of the opsin (Yokoyama, 1995). In this study, compared with other closely related species, the mandarin fish was identified with more *LWS* genes. *LWS1/2* had five specific sites in the mandarin fish with remarkable differences from the other seven fish species (**Supplementary Figure 5**). Certain mutations of the transmembrane domains, L98 and T219 in the *LWS* genes might be expected to contribute to the special feeding habit of live prey.

Many fish species exhibit sexual dimorphisms, such as Japanese flounder (*Paralichthys olivaceus*) (Shao et al., 2015), half-smooth tongue sole (*Cynoglossus semilaevis*) (Song et al., 2012), displaying significant differences in growth rates or sizes between male and female individuals. Females of mandarin fish present higher growth rates (by 10–20% in body weight) than males (Sun et al., 2017). Therefore, screening of sex-related genes or markers is important for the development of the mandarin industry, which will be helpful for the elucidation of the SD mechanisms in the mandarin fish. Nineteen sex-related genes, localized on the Chr14, were previously reported to be involved in spermatogenesis, SD, and testicular determination. Some studies support that SD is controlled by many major genetic factors that may interact with minor genetic factors, thereby implying that SD should be analyzed as a quantitative trait (Eshel et al., 2012). Five sex-related QTLs in the mandarin fish were previously detected on the Chr17. Therefore, we speculate that both the Chr14 and Chr17 are the potential to be related to SD in the mandarin fish. These results suggest the involvement of multiple chromosomes in sex relation, and provide supportive evidence to the polygenic SD in fishes (Zhang S. et al., 2019). In the coming future, the development of unisex male populations will be necessary for rapid improvement of the quality and quantity of the mandarin fish.

CONCLUSION

In our present study, we generated a chromosomal-level genome assembly for the mandarin fish, which has been an economically important fish in China. Our genome assembly is high in quality, completeness, and accuracy based on multiple evaluations. Gene prediction, functional annotation, and evolutionary analysis provided novel insights into the genomic structure and mechanisms underlying food intake, SD, and prediction of new toxins. Our genome sequences will also offer a valuable genetic resource to support extensive fisheries and artificial breeding programs, and thereby allows for effective disease management, growth improvement, and discovering new drugs in the mandarin fish.

DATA AVAILABILITY STATEMENT

This Whole Genome Shotgun project of mandarin fish has been deposited in CNGBdb with accession number CNA0013732. Raw reads from Illumina sequencing are deposited in the CNGBdb with accession number CNS0204384. The genome assembly of mandarin fish has been deposited in the CNGB Nucleotide Sequence Archive (https://db.cngb.org/cnsa/) under the Project ID CNP0000961.

ETHICS STATEMENT

The animal study was reviewed and approved by the Institutional Review Board on Bioethics and Biosafety of BGI, China (No. FT 18134). Written informed consent was obtained from the owners for the participation of their animals in this study.

AUTHOR CONTRIBUTIONS

XB, XY, and WD conceived and designed the research. WD and XZ performed the genome sequencing. XZ, JL, and YH performed data analyses and wrote the manuscript. WJ, ZC, and MW performed sample preparation. WD, XZ, XB, QS, and XY revised the manuscript. All authors approved submission of the manuscript for publication.

FUNDING

This work was supported by the Jiangsu Agriculture Science and Technology Innovation Fund [No. CX(17)3005]; the Central Public-Interest Scientific Institution Basal Research Fund (2017JBFZ02); the Shenzhen Special Program for Development of Emerging Strategic Industries (No. JSGG20170412153411369); the Ministry of Agriculture and Rural Affairs Project for Conservation of Species Resources (No. 17190412); and the Natural Science Foundation for Fundamental Research in Shenzhen (JCYJ20190812105801661).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.671650/full#supplementary-material

REFERENCES

Abrusan, G., Grundmann, N., Demester, L., and Makalowski, W. (2009). TEclass–a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* 25, 1329–1330. doi: 10.1093/bioinformatics/btp084

Attwood, T. K. (2002). The PRINTS database: a resource for identification of protein families. *Brief Bioinform.* 3, 252–263. doi: 10.1093/bib/3.3.252

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* 6:11.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Bian, C., Hu, Y., Ravi, V., Kuznetsova, I. S., Shen, X., Mu, X., et al. (2016). The Asian arowana (*Scleropages formosus*) genome provides new insights into the evolution of an early lineage of teleosts. *Sci. Rep.* 6:24501.

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. *Genome Res.* 14, 988–995. doi: 10.1101/gr.1865504

Birney, E., and Durbin, R. (2000). Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* 10, 547–548. doi: 10.1101/gr.10.4.547

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31, 365–370. doi: 10.1093/nar/gkg095

Bru, C., Courcelle, E., Carrere, S., Beausse, Y., Dalmar, S., and Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res.* 33, D212–D215.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78–94. doi: 10.1006/jmbi.1997.0951

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinform.* 10:421. doi: 10.1186/1471-2105-10-421

Chandrashekar, J., Mueller, K. L., Hoon, M. A., Adler, E., Feng, L., Guo, W., et al. (2000). T2Rs function as bitter taste receptors. *Cell* 100, 703–711. doi: 10.1016/s0092-8674(00)80706-0

Chiang, L. (1959). On the biology of mandarinfish, *Siniperca chuatsi*, of Liangtze Lake. *Acta Hydrobiol. Sin* 3, 365–385.

Chinen, A., Hamaoka, T., Yamada, Y., and Kawamura, S. (2003). Gene duplication and spectral diversification of cone visual pigments of zebrafish. *Genetics* 163, 663–675. doi: 10.1093/genetics/163.2.663

UniProt Consortium, (2018). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515.

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., et al. (2015). Ensembl 2015. *Nucleic Acids Res.* 43, D662–D669.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Ellisdon, A. M., Reboul, C. F., Panjikar, S., Huynh, K., Oellig, C. A., Winter, K. L., et al. (2015). Stonefish toxin defines an ancient branch of the perforin-like superfamily. *Proc. Natl. Acad. Sci. U.S.A.* 112, 15360–15365. doi: 10.1073/pnas.1507622112

Elsik, C. G., Mackey, A. J., Reese, J. T., Milshina, N. V., Roos, D. S., and Weinstock, G. M. (2007). Creating a honey bee consensus gene set. *Genome Biol.* 8:R13.

Engelmann, J., Hanke, W., Mogdans, J., and Bleckmann, H. (2000). Hydrodynamic stimuli and the fish lateral line. *Nature* 408, 51–52. doi: 10.1038/35040706

Eshel, O., Shirak, A., Dor, L., Band, M., Zak, T., Markovich-Gordon, M., et al. (2014). Identification of male-specific amh duplication, sexually differentially expressed genes and microRNAs at early embryonic development of Nile tilapia (*Oreochromis niloticus*). *BMC Genom.* 15:774. doi: 10.1186/1471-2164-15-774

Eshel, O., Shirak, A., Weller, J. I., Hulata, G., and Ron, M. (2012). Linkage and physical mapping of sex region on LG23 of Nile Tilapia (*Oreochromis niloticus*). *G3* 2, 35–42. doi: 10.1534/g3.111.001545

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230.

Guindon, S., Delsuc, F., Dufayard, J. F., and Gascuel, O. (2009). Estimating maximum likelihood phylogenies with PhyML. *Methods Mol. Biol.* 537, 113–137. doi: 10.1007/978-1-59745-251-9_6

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Guo, W., He, S., Liang, X. T., Changxu, D., Yaqi, L., and Liyuan, M. (2021). A high-density genetic linkage map for Chinese perch (*Siniperca chuatsi*) using 2.3K genotyping-by-sequencing SNPs. *Anim. Genet.* 52, 311–320. doi: 10.1111/age.13046

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., et al. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, D258–D261.

Harris, R. J., and Jenner, R. A. (2019). Evolutionary ecology of fish venom: adaptations and consequences of evolving a venom system. *Toxins* 11:60. doi: 10.3390/toxins11020060

He, S., Liang, X. F., Li, L., Huang, W., Shen, D., and Tao, Y. X. (2013). Gene structure and expression of leptin in Chinese perch. *Gen. Comp. Endocrinol.* 194, 183–188. doi: 10.1016/j.ygcen.2013.09.008

Holzer, P., Reichmann, F., and Farzi, A. (2012). Neuropeptide Y, peptide YY and pancreatic polypeptide in the gut-brain axis. *Neuropeptides* 46, 261–274. doi: 10.1016/j.npep.2012.08.005

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., et al. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Res.* 37, D211–D215.

Jungo, F., Bougueleret, L., Xenarios, I., and Poux, S. (2012). The UniProtKB/Swiss-Prot Tox-Prot program: a central hub of integrated venom protein data. *Toxicon* 60, 551–557. doi: 10.1016/j.toxicon.2012.03.010

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Keen-Rhinehart, E., Ondek, K., and Schneider, J. E. (2013). Neuroendocrine regulation of appetitive ingestive behavior. *Front. Neurosci.* 7:213. doi: 10.3389/fnins.2013.00213

Khoo, H. E., Yuen, R., Poh, C. H., and Tan, C. H. (1992). Biological activities of *Synanceja horrida* (stonefish) venom. *Nat. Toxins* 1, 54–60.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645. doi: 10.1101/gr.092759.109

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Kurokawa, T., Uji, S., and Suzuki, T. (2005). Identification of cDNA coding for a homologue to mammalian leptin from pufferfish, *Takifugu rubripes*. *Peptides* 26, 745–750. doi: 10.1016/j.peptides.2004.12.017

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004). Versatile and open software for comparing large genomes. *Genome Biol.* 5:R12.

Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., et al. (2004). SMART 4.0: towards genomic data integration. *Nucleic Acids Res.* 32, D142–D144.

Li, L., Stoeckert, C. J. Jr., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi: 10.1101/gr.1224503

Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K., et al. (2009). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25, 1966–1967. doi: 10.1093/bioinformatics/btp336

Li, S., Liu, Q., Xiao, L., Chen, H., Li, G., Zhang, Y., et al. (2016). Molecular cloning and functional characterization of spexin in orange-spotted grouper (*Epinephelus coioides*). *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* 196–197, 85–91. doi: 10.1016/j.cbpb.2016.02.009

Li, W., Zhang, T., Ye, S., Liu, J., and Li, Z. (2013). Feeding habits and predator-prey size relationships of mandarin fish *Siniperca chuatsi* (Basilewsky) in a shallow lake, central China. *J. Appl. Ichthyol.* 29, 56–63. doi: 10.1111/j.1439-0426.2012.02044.x

Li, X., Staszewski, L., Xu, H., Durick, K., Zoller, M., and Adler, E. (2002). Human receptors for sweet and umami taste. *Proc. Natl. Acad. Sci. U.S.A.* 99, 4692–4696.

Liang, X., Kiu, J., and Huang, B. (1998). The role of sense organs in the feeding behaviour of Chinese perch. *J. Fish Biol.* 52, 1058–1067. doi: 10.1111/j.1095-8649.1998.tb00603.x

Liang, Y., and Cui, X. (1982). The eco-physiological characteristics of artificial propagation in mandarin fish (*Siniperca chuatsi*). *Acta Hydrobiol. Sin.* 16, 90–92.

Lin, J.-J., Wang, F.-Y., Li, W.-H., and Wang, T.-Y. (2017). The rises and falls of opsin genes in 59 ray-finned fish genomes and their implications for environmental adaptation. *Sci. Rep.* 7:15568.

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., et al. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv preprint*. arXiv:13082012.

Liu, J., Cui, Y., and Liu, J. (1998). Food consumption and growth of two piscivorous fishes, the mandarin fish and the Chinese snakehead. *J. Fish Biol.* 53, 1071–1083. doi: 10.1111/j.1095-8649.1998.tb00464.x

Liu, S., Xu, P., Liu, X., Guo, D., Chen, X., Bi, S., et al. (2021). Production of neo-male mandarin fish *Siniperca chuatsi* by masculinization with orally administered 17 alpha-methyltestosterone. *Aquaculture* 530:680.

Low, K. S., Gwee, M. C., Yuen, R., Gopalakrishnakone, P., and Khoo, H. E. (1994). Stonustoxin: effects on neuromuscular function in vitro and in vivo. *Toxicon* 32, 573–581. doi: 10.1016/0041-0101(94)90205-4

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1:18.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2015). Erratum: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 4:30.

Lv, L.-Y., Liang, X.-F., and He, S. (2019). Genome-wide identification and characterization of olfactory receptor genes in chinese perch, *Siniperca chuatsi*. *Genes* 10:178. doi: 10.3390/genes10020178

Mount, D. W. (2007). Using the basic local alignment search tool (BLAST). *Cold Spring Harb. Protoc.* 2007:pdb. top17.

Neafsey, D. E., and Hartl, D. L. (2005). Convergent loss of an anciently duplicated, functionally divergent RH2 opsin gene in the fugu and *Tetraodon pufferfish* lineages. *Gene* 350, 161–171. doi: 10.1016/j.gene.2005.02.011

Negroni, J., Meunier, N., Monnerie, R., Salesse, R., Baly, C., Caillol, M., et al. (2012). Neuropeptide Y enhances olfactory mucosa responses to odorant in hungry rats. *PLoS One* 7:e45266. doi: 10.1371/journal.pone.0045266

Nei, M., Niimura, Y., and Nozawa, M. (2008). The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. *Nat. Rev. Genet.* 9, 951–963. doi: 10.1038/nrg2480

Nelson, G., Hoon, M. A., Chandrashekar, J., Zhang, Y., Ryba, N. J., and Zuker, C. S. (2001). Mammalian sweet taste receptors. *Cell* 106, 381–390. doi: 10.1016/s0092-8674(01)00451-2

Niimura, Y. (2009). On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 chordate species. *Genome Biol. Evol.* 1, 34–44. doi: 10.1093/gbe/evp003

Niimura, Y., and Nei, M. (2005). Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. *Proc. Natl. Acad. Sci. U.S.A.* 102, 6039–6044. doi: 10.1073/pnas.0501922102

Niimura, Y., and Nei, M. (2006). Evolutionary dynamics of olfactory and other chemosensory receptor genes in vertebrates. *J. Hum. Genet.* 51, 505–517. doi: 10.1007/s10038-006-0391-8

Oike, H., Nagai, T., Furuyama, A., Okada, S., Aihara, Y., Ishimaru, Y., et al. (2007). Characterization of ligands for fish taste receptors. *J. Neurosci.* 27, 5584–5592. doi: 10.1523/jneurosci.0651-07.2007

Pevsner, J. (2005). *Basic Local Alignment Search Tool (BLAST)*. Hoboken, NJ: John Wiley & Sons, Inc.

Poh, C. H., Yuen, R., Khoo, H. E., Chung, M., Gwee, M., and Gopalakrishnakone, P. (1991). Purification and partial characterization of stonustoxin (lethal factor) from *Synanceja horrida* venom. *Comp. Biochem. Physiol. B* 99, 793–798. doi: 10.1016/0305-0491(91)90143-2

Senzui, A., Masumoto, T., and Fukada, H. (2020). Neuropeptide Y expression in response to sensory organ-detected fish meal soluble components and orally fed fish meal-based diet in yellowtail Seriola quinqueradiata. *Aquaculture* 514:734512. doi: 10.1016/j.aquaculture.2019.734512

Shao, C., Niu, Y., Rastas, P., Liu, Y., Xie, Z., Li, H., et al. (2015). Genome-wide SNP identification for the construction of a high-resolution genetic map of Japanese flounder (*Paralichthys olivaceus*): applications to QTL mapping of *Vibrio anguillarum* disease resistance and comparative genomic analysis. *DNA Res.* 22, 161–170. doi: 10.1093/dnares/dsv001

Shichida, Y., and Matsuyama, T. (2009). Evolution of opsins and phototransduction. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 2881–2895. doi: 10.1098/rstb.2009.0051

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Slater, G. S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinform.* 6:31. doi: 10.1186/1471-2105-6-31

Song, W., Li, Y., Zhao, Y., Liu, Y., Niu, Y., Pang, R., et al. (2012). Construction of a High-density microsatellite genetic linkage map and mapping of sexual and growth-related traits in half-smooth tongue sole (*Cynoglossus semilaevis*). *PLoS One* 7:e52097. doi: 10.1371/journal.pone.0052097

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 34, W435–W439.

Sun, B. J., Xie, H. X., Song, Y., and Nie, P. (2007). Gene structure of an antimicrobial peptide from mandarin fish, *Siniperca chuatsi* (Basilewsky), suggests that moronecidins and pleurocidins belong in one family: the piscidins. *J. Fish Dis.* 30, 335–343. doi: 10.1111/j.1365-2761.2007.00789.x

Sun, C., Niu, Y., Ye, X., Dong, J., Hu, W., Zeng, Q., et al. (2017). Construction of a high-density linkage map and mapping of sex determination and growth-related loci in the mandarin fish (*Siniperca chuatsi*). *BMC Genom.* 18:446. doi: 10.1186/s12864-017-3830-3

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* 25, 4.10.11–14.10.14.

Terai, Y., Mayer, W. E., Klein, J., Tichy, H., and Okada, N. (2002). The effect of selection on a long wavelength-sensitive (LWS) opsin gene of Lake Victoria cichlid fishes. *Proc. Natl. Acad. Sci. U.S.A.* 99, 15501–15506. doi: 10.1073/pnas.232561099

Tine, M., Kuhl, H., Gagnaire, P. A., Louro, B., Desmarais, E., Martins, R. S., et al. (2014). European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nat. Commun.* 5:5770.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111. doi: 10.1093/bioinformatics/btp120

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Utkin, Y. N. (2015). Animal venom studies: current benefits and future developments. *World J. Biol. Chem.* 6, 28–33. doi: 10.4331/wjbc.v6.i2.28

Van Ooijen, J. (2006). *JoinMap$4, Software for the Calculation of Genetic Linkage Maps in Experimental Populations*. Wageningen: Kyazma BV.

Volkoff, H. (2006). The role of neuropeptide Y, orexins, cocaine and amphetamine-related transcript, cholecystokinin, amylin and leptin in the regulation of feeding in fish. *Comp. Biochem. Physiol. A Mol. Integr. Physiol.* 144, 325–331. doi: 10.1016/j.cbpa.2005.10.026

Volkoff, H. (2016). The neuroendocrine regulation of food intake in fish: a review of current knowledge. *Front. Neurosci.* 10:540. doi: 10.3389/fnins.2016.00540

Volkoff, H., Hoskins, L. J., and Tuziak, S. M. (2010). Influence of intrinsic signals and environmental cues on the endocrine control of feeding in fish: potential application in aquaculture. *Gen. Comp. Endocrinol.* 167, 352–359. doi: 10.1016/j.ygcen.2009.09.001

Walewski, J. L., Ge, F., Lobdell, H. T., Levin, N., Schwartz, G. J., Vasselli, J. R., et al. (2014). Spexin is a novel human peptide that reduces adipocyte uptake of long chain fatty acids and causes weight loss in rodents with diet-induced obesity. *Obesity* 22, 1643–1652. doi: 10.1002/oby.20725

Woods, I. G., Schoppik, D., Shi, V. J., Zimmerman, S., Coleman, H. A., Greenwood, J., et al. (2014). Neuropeptidergic signaling partitions arousal behaviors in zebrafish. *J. Neurosci.* 34, 3142–3160. doi: 10.1523/jneurosci.3529-13.2014

Wu, Z. (1988). A preliminary ethological analysis on the feeding behavior of mandarin fish. *Freshw. Fish* 5, 18–21.

Xie, B., Huang, Y., Baumann, K., Fry, B. G., and Shi, Q. (2017). From marine venoms to drugs: efficiently supported by a combination of transcriptomics and proteomics. *Mar. Drugs* 15, 1–10. doi: 10.3390/md15040103

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268.

Yokobori, E., Azuma, M., Nishiguchi, R., Kang, K. S., Kamijo, M., Uchiyama, M., et al. (2012). Neuropeptide Y stimulates food intake in the Zebrafish, *Danio rerio*. *J. Neuroendocrinol.* 24, 766–773. doi: 10.1111/j.1365-2826.2012.02281.x

Yokoyama, S. (1995). Amino acid replacements and wavelength absorption of visual pigments in vertebrates. *Mol. Biol. Evol.* 12, 53–61. doi: 10.1093/oxfordjournals.molbev.a040190

Yokoyama, S. (2000). Molecular evolution of vertebrate visual pigments. *Prog. Retin. Eye Res.* 19, 385–419. doi: 10.1016/s1350-9462(00)00002-1

You, X., Bian, C., Zan, Q., Xu, X., Liu, X., Chen, J., et al. (2014). Mudskipper genomes provide insights into the terrestrial adaptation of amphibious fishes. *Nat. Commun.* 5:5594.

Zeng, Q., Liu, S., Yao, J., Zhang, Y., Yuan, Z., Jiang, C., et al. (2016). Transcriptome display during testicular differentiation of channel catfish (*Ictalurus punctatus*) as revealed by RNA-Seq analysis. *Biol. Reprod.* 95:19. doi: 10.1095/biolreprod.116.138818

Zhang, F.-B., Wang, Y., Xiao, J., and Zeng, Y. (2019). Study on distribution of Ichthyotoxic fish in the Jialing River. *Resourc. Environ. Yangtze Basin* 28, 2901–2909.

Zhang, S., Li, J., Qin, Q., Liu, W., Bian, C., Yi, Y., et al. (2018). Whole-genome sequencing of chinese yellow catfish provides a valuable genetic resource for high-throughput identification of toxin genes. *Toxins* 10:488. doi: 10.3390/toxins10120488

Zhang, D.-C., Guo, L., Guo, H.-Y., Zhu, K.-C., Li, S.-Q., Zhang, Y., et al. (2019). Chromosome-level genome assembly of golden pompano (Trachinotus ovatus) in the family Carangidae. *Sci. Data.* 6, 1–11.

Zhang, S., Zhang, X., Chen, X., Xu, T., Wang, M., Qin, Q., et al. (2019). Construction of a High-density linkage map and QTL fine mapping for growth- and sex-related traits in channel catfish (*Ictalurus punctatus*). *Front. Genet.* 10:251. doi: 10.3389/fgene.2019.00251

Zhao, G. Q., Zhang, Y., Hoon, M. A., Chandrashekar, J., Erlenbach, I., Ryba, N. J., et al. (2003). The receptors for mammalian sweet and umami taste. *Cell* 115, 255–266. doi: 10.1016/s0092-8674(03)00844-4

Zheng, B., Li, S., Liu, Y., Li, Y., Chen, H., Tang, H., et al. (2017). Spexin suppress food intake in zebrafish: evidence from gene knockout study. *Sci. Rep.* 7:14643.

Zhou, Y., Liang, X.-F., Yuan, X., Li, J., He, Y., Fang, L., et al. (2013). Neuropeptide Y stimulates food intake and regulates metabolism in grass carp, *Ctenopharyngodon idellus*. *Aquaculture* 380–383, 52–61. doi: 10.1016/j.aquaculture.2012.11.033

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

DATA REPORT
published: 07 July 2021
doi: 10.3389/fgene.2021.695700

Draft Genome of the Mirrorwing Flyingfish (*Hirundichthys speculiger*)

Pengwei Xu[1†], *Chenxi Zhao*[1†], *Xinxin You*[1,2†], *Fan Yang*[3], *Jieming Chen*[1,2], *Zhiqiang Ruan*[1,2], *Ruobo Gu*[2], *Junmin Xu*[2], *Chao Bian*[1,2*] *and Qiong Shi*[1,2*]

[1] College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China, [2] Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, BGI Academy of Marine Sciences, BGI Marine, BGI, Shenzhen, China, [3] Marine Geological Department, Marine Geological Survey Institute of Hainan Province, Haikou, China

Keywords: flying fish, whole genome sequencing, genome assembly, *eevs*, vision-related gene, phylogenetic tree

SUMMARY

Flying fishes are a group of Exocoetidae members with an intriguing epipelagic inhabitant. They have evolved numerous interesting characteristics. Here, we performed whole genome sequencing, *de novo* assembly and annotation of the representative mirrorwing flyingfish (*Hirundichthys speculiger*). We obtained a 1.04-Gb genome assembly using a hybrid approach from 99.21-Gb Illumina and 29.98-Gb PacBio sequencing reads. Its contig N50 and scaffold N50 values reached 992.83 and 1,152.47 kb, respectively. The assembled genome was predicted to possess 23,611 protein-coding genes, of which 23,492 (99.5%) were functionally annotated with public databases. A total of 42.02% genome sequences consisted of repeat elements, among them DNA transposons accounted for the largest proportion (24.38%). A BUSCO (Benchmarking Universal Single Copy Orthologs) evaluation demonstrated that the genome and gene completeness were 94.2% and 95.7%, respectively. Our phylogeny tree revealed that the mirrorwing flyingfish was close to *Oryzias* species with a divergence time of about 85.2 million years ago. Moreover, nine vison-related genes, three melatonin biosynthesis related aanat (aralkylamine *N*-acetyltransferase) genes, and two sunscreen biosynthesis related *eevs* (2-epi-5-epi-valiolone synthase) genes were identified in the assembled genome; however, the loss of *SWS1* (short-wavelength sensitive opsin 1) and aanat1a in amphibious mudskippers was not presented in the mirrorwing flyingfish genome. In summary, we generate a high-quality draft genome assembly for the mirrorwing flyingfish, which provides new insights into physiology-related genes of Exocoetidae. It also serves as a powerful resource for exploring intriguing traits of Exocoetidae at a genomics level.

OPEN ACCESS

Edited by:
Liang Guo,
South China Sea Fisheries Research Institute, China

Reviewed by:
Jitendra Kumar Sundaray,
Central Institute of Freshwater Aquaculture, India
You-Yi Kuang,
Heilongjiang River Fisheries Research Institute, China

***Correspondence:**
Qiong Shi
shiqiong@genomics.cn
Chao Bian
bianchao@genomics.cn

[†] *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *15 April 2021*
Accepted: *03 June 2021*
Published: *07 July 2021*

Citation:
Xu P, Zhao C, You X, Yang F, Chen J, Ruan Z, Gu R, Xu J, Bian C and Shi Q (2021) Draft Genome of the Mirrorwing Flyingfish (Hirundichthys speculiger). Front. Genet. 12:695700. doi: 10.3389/fgene.2021.695700

INTRODUCTION

Flying fishes (Exocoetidae; Beloniformes) have evolved with numerous interesting characteristics, such as gliding over water, marine- to freshwater transition, and unique craniofacial and egg buoyancy. They have been regarded as an extraordinary marine group with enlarged pelvic fins and hypocercal caudal fins, which could help to glide over water to reach a distance up to 400 m (Davenport, 1994). Although the oldest gliding fish fossil (*Potanichthys xingyiensis*) shares certain similar morphology with modern flying fishes, it is not the ancestor of the modern flying fishes, since they are thought to have evolved independently about 65.5 million years ago (Xu et al., 2012). Compared with tetrapod gliders, the gliding behavior of flying fishes could not be considered as an energy-saving strategy for long-distance movement (Rayner, 1986), but it may be just used for escaping from underwater predators [e.g., swordfish, tuna, dolphin, and squid (Kutschera, 2005)].

While the representative mirrorwing flyingfish (*Hirundichthys speculiger*; **Figure 1A**) traverses the air and water interface, it meets a series of challenges [such as relentless sunshine, lack of buoyancy, and high CO_2 accumulation (Wright and Turko, 2016)] as amphibious fishes. The lower refractive index of air usually aggravates this situation, making fishes myopic in air (Baylor and Shaw, 1962). Duplication, loss, differential expression, and crucial tuning of opsin genes could

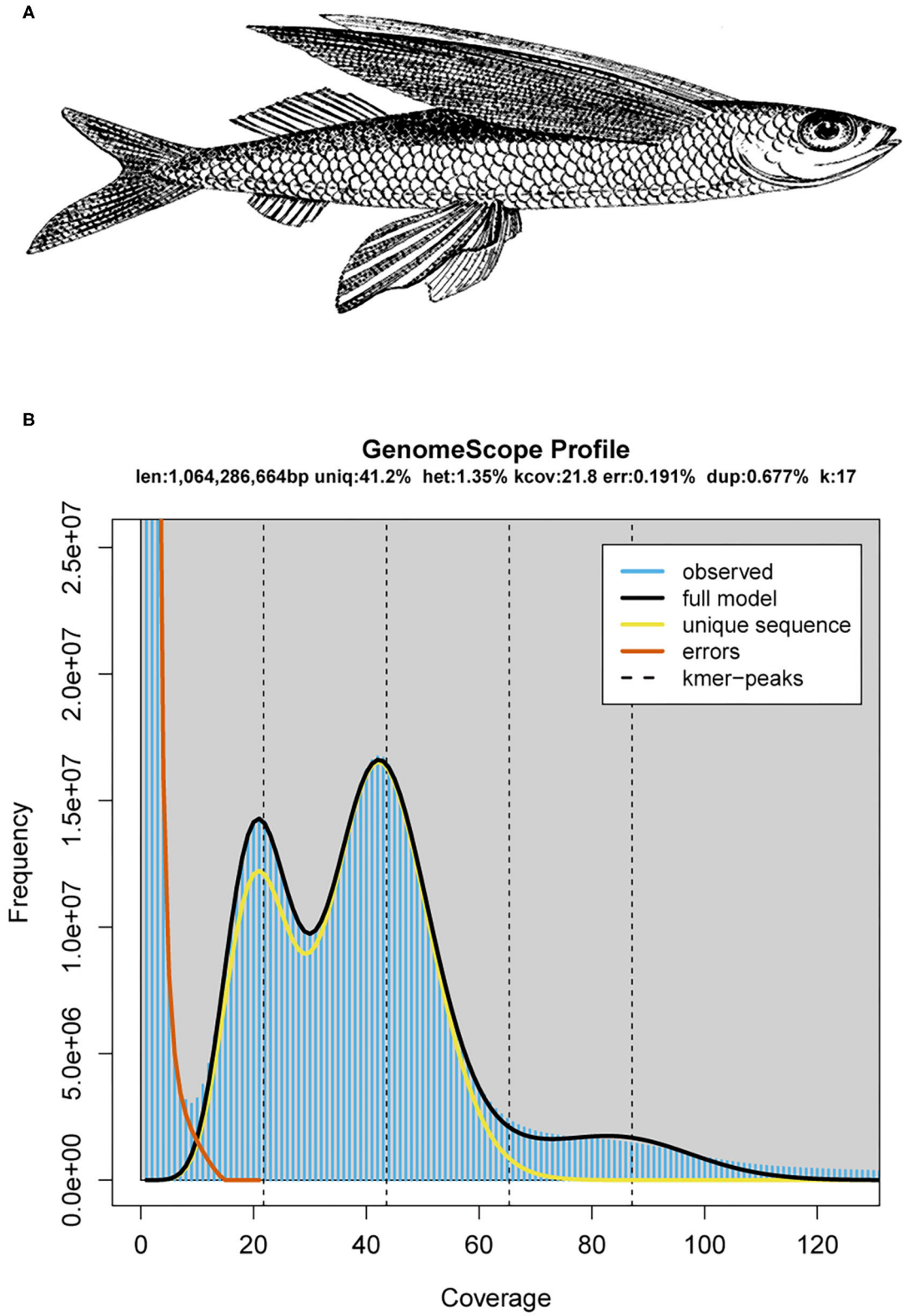

FIGURE 1 | The schematic diagram and genomics feature of the mirrorwing flyingfish. **(A)** A drawing of the mirrorwing flyingfish [adopted from De Bruin et al. (1995)]. **(B)** A *k*-mer analysis of the genome sequencing reads for the mirrorwing flyingfish using GenomeScope.

lead to visual plasticity in vertebrates for adapting to the water-to-air environments (Hauser and Chang, 2017). Five types of opsins, including LWS (red: long wavelength-sensitive), SWS1 (UV: short wavelength-sensitive 1), SWS2 (violet/blue: short wavelength-sensitive 2), RH1 (dim vision: rhodopsin), and RH2 (green: green-sensitive), have been identified in non-mammalian vertebrates (Yokoyama, 2000). Modifications of opsin and melatonin biosynthesis-related arylalkylamine *N*-acetyltransferase (*aanat*) genes could enhance amphibious mudskippers' survival on land (You et al., 2014). When the mirrorwing flyingfish leaps out of water, whether it employs the same mechanisms as mudskippers (including crucial mutation sites of LWS, lack of SWS1, and loss of *aanat1a* in the giant-fin mudskipper; see more details in You et al., 2014) or not is still an open question.

Ultraviolet radiation (UVR: 280–400 nm) often causes DNA damages through oxidative stress, producing a number of disorders (such as sunburn and skin cancer risk) (Kageyama and Waditee-Sirisattha, 2019; Rosic, 2019). UV-absorbing compounds, such as mycosporine-like amino acids (MAAs) and gadusol, are commonly distributed in various marine microorganisms, invertebrates, and algae (Shick and Dunlap, 2002; Miyamoto et al., 2014). The *de novo* synthesis of MAA in invertebrates (such as coral and sea anemone) employed a four-step desmethyl-4-deoxygadusol synthase (DDGS) based pathway as cyanobacteria (Balskus and Walsh, 2010; Rosic and Dove, 2011; Shinzato et al., 2011), while zebrafish (*Danio rerio*) could convert sedoheptulose-7-phosphate (SH7P) to gadusol using 2-epi-5-epi-valiolone synthase (EEVS) and S-adenosyl-L-methionine-dependent methyltransferase [MT-Ox (Osborn et al., 2015)]. The two core genes, *eevs* and *mt-ox*, in zebrafish are flanked by four transcription factor genes [*frmd4B*, *mitf*, *mdfic*, and *foxp1* (Osborn et al., 2015)], which is not consistent with the loss of *mdfic* in Japanese medaka [*Oryzias latipes* (Kim et al., 2018)]. Phylogenetic analysis using mitochondrial genes in Beloniformes had inferred a close relationship between the mirrorwing flyingfish and medaka (Lovejoy et al., 2004; Cui et al., 2018). Whether the mirrorwing flyingfish contains the complete gene cluster as zebrafish or incomplete cluster as medaka is valuable for checking the possible lineage-specific gene rearrangement of *eevs*-like cluster.

Here, we performed whole genome sequencing of the mirrorwing flyingfish and generated a draft assembly with a hybrid method (Ye et al., 2016) for the first time. Our subsequent phylogenetic and comparative genomic analyses between amphibious fishes and ordinary underwater fishes will provide insights into the evolution of vision-related genes, olfactory receptor (OR) genes, and gadusol synthesis-related genes (*eevs*) in the mirrorwing flyingfish. This genome assembly will serve as a valuable resource for the illumination of molecular basis for the special characteristics of flying fishes.

Value of the Data

This is the first genome report of the representative mirrorwing flyingfish. Our final assembly was 1.04 Gb, with a contig N50 of 992.83 kb and a scaffold N50 of 1,152.47 kb.

A phylogeny tree was constructed to demonstrate that the mirrorwing flyingfish was close to *Oryzias* species with a divergence time of about 85.2 Mya. A total of 60.71% of the mirrorwing flyingfish genome region was syntenic with *O. latipes*.

The genome of mirrorwing flyingfish harbored nine vision-related genes, three *aanat* genes, and two *eevs*-like genes. The existence of *SWS1* and *aanat1a* suggests that the mirrorwing flyingfish employs different strategies for visional adaptation in air. A gene cluster of *eevs*-like shared the same synteny as Japanese medaka, implying a uniform gene rearrangement in Beloniformes.

MATERIALS AND METHODS

Fish Sampling and Genome Sequencing

An adult mirrorwing flyingfish was captured by torch fishing in the water area of Iltis Bank, Xisha, China. Genomic DNAs were extracted from muscle tissues and purified and quality checked according to a standard protocol (Sigma-Aldrich, St. Louis, MO, USA).

Subsequently, three paired-end libraries (with insert sizes of 270, 500, and 800 bp, respectively) and three mate-pair libraries (with insert sizes of 2, 5, and 10 kb, respectively) were constructed in accordance with an Illumina standard manual before sequencing on an Illumina X-Ten platform (Illumina Inc., San Diego, CA, USA) with a PE-150 or PE-125 module. Raw reads were then processed using SOAPnuke v1.5.6 (Chen et al., 2018) with optimized parameters ("-n 0.02 -Q 2 -l 15—5 1 -d -I -q 0.4"). An additional SMART Bell library with an insert size of 20 kb was constructed based on a PacBio RS II protocol (Pacific Biosciences, Menlo Park, CA, USA). Six DNA sequencing cells were produced using the P6 polymerase/C4 chemistry (Rhoads and Au, 2015).

Genome Assembly

Distribution of *k-mer* frequency was constructed with jellyfish v2.0 (Marçais and Kingsford, 2011) using clean reads from short-insert libraries (270 and 500 bp). GenomeScope v1.0 (Vurture et al., 2017) was then applied to estimate the genome size and heterozygosity. A routine hybrid pipeline was employed to assemble the high heterozygous flyingfish genome (**Supplementary Figure 1**).

In brief, the Illumina paired-end reads were first assembled using Platanus v1.24 (Kajitani et al., 2014) with optimized parameters (assemble -k 35 -s 5 -u 0.2 -d 0.5). DBG2OLC (Ye et al., 2016) was employed to construct backbone sequences from the best overlaps between the initial contigs and raw PacBio reads. All related PacBio reads were realigned to the backbone with Sparc (Ye and Ma, 2016) to construct the most likely consensus sequences of the genome. All Illumina paired-end reads were aligned to the resulting assembly using BWA-MEM (Li, 2014). The alignments were employed for Pilon v1.24 (Walker et al., 2014) to polish the assembly. All Illumina mate-pair reads were mapped onto the corrected contigs using BWA-MEM (Li, 2014). These alignments were then processed with BESST v2.2.4 (Sahlin et al., 2014) to construct scaffolds. Completeness of the genome assembly was evaluated by BUSCO v3.0 (Simão et al., 2015)

with default parameters "-l actinopterygii_odb9 -m genome -c 3 -sp zebrafish."

Genome Annotation

Transposable elements (TEs) were identified using both homolog-based and *de novo* methods. For the homolog-based method, RepeatMasker v4.06 and ProteinRepeatMasker v4.06 (Chen, 2004) were employed to identify known TEs against the Repbase v21.0 (Jurka et al., 2005). For the *de novo* method, a *de novo* library was constructed using RepeatModeler v2.0 (Flynn et al., 2020) and LTR-FINDER v1.0.6 (Xu and Wang, 2007) firstly. Then, RepeatMasker v4.06 was subjected to identify the *de novo* TEs against the *de novo* library. The tandem repeat sequences were identified using Tandem Repeat Finder (Benson, 1999).

Gene models were also predicted using both homolog-based and *de novo* methods. For the homolog-based methods, protein sequences of zebrafish (*Danio rerio*), three-spined stickleback (*Gasterosteus aculeatus*), human (*Homo sapiens*), Japanese medaka (*O. latipes*), and green spotted pufferfish (*Tetraodon nigroviridis*) were derived from Ensembl-100 and aligned to our flyingfish genome using tBLASTn (Ye et al., 2006) with parameter "-e 1e-5 -m 8 -F." Blasted hits were processed by SOLAR v0.9 (Yu et al., 2006) with parameter "-a prot2 genome2 -z" to determine the potential gene loci. We extracted the candidate gene region with 2-kb flanking sequences and employed Genewise v2.4 (Birney et al., 2004) to determine gene structures. For the *de novo* prediction, we trained the parameters of AUGUSTUS v3.2 (Stanke et al., 2006) using randomly selected 2,000 intact gene models that were derived from the homolog-based method. Then, we used AUGUSTUS to perform *ab initio* prediction on the repeat-masked genome with the trained parameters. Finally, the gene models predicted from both approaches were integrated to form non-redundant gene sets using the similar pipeline as described in a previous study (Xiong et al., 2016). Completeness of the gene sets was evaluated by BUSCO v3.0 (Simão et al., 2015) with parameters "-l actinopterygii_odb9 -m protein -c 3 -sp zebrafish."

Gene function annotation was performed on the basis of sequence and domain similarity. The protein sequences were aligned to Kyoto Encyclopedia of Genes and Genomes (KEGG) v84.0 (Kanehisa et al., 2017), SwissProt, and TrEMBL (Uniprot release 2020-06) (Bairoch et al., 2005) using BLASTP (Ye et al., 2006) with an E-value of 1e−5. InterProScan v5.11-55.0 (Jones et al., 2014) was applied to predict domain information with public databases including Pfam (Bateman et al., 2004), SMART (Letunic et al., 2012), PANTHER (Thomas et al., 2003), PRINTS (Attwood et al., 2000), PROSITE profiles (Sigrist et al., 2010), and ProDom (Servant et al., 2002). Gene Ontology (GO) terms were predicted using the IPR entry list (Burge et al., 2012).

Four types of non-coding RNA were identified in the mirrorwing flyingfish genome. We employed tRNAscan-SE v2.0 (Lowe and Eddy, 1997) to detect transfer RNAs (tRNAs). For microRNAs (miRNAs) and small nuclear RNAs (snRNAs), the Rfam v12.0 (Nawrocki et al., 2015) database was mapped onto the assembled genome, and the matched sequences were delivered into INFERNAL v1.1.4 (Nawrocki and Eddy, 2013) to confirm structures. Ribosomal RNAs (rRNAs) in the genome were searched using animal full-length rRNAs (Quast et al., 2012) as the query.

Gene Family Prediction

To identify gene families in the mirrorwing flyingfish genome, we download protein-coding sequences of 18 representative teleost fishes from the National Center for Biotechnology Information (NCBI) databases (see more details in **Supplementary Table 1**), including *Anabas testudineus* (Ates; climbing perch), *Austrofundulus limnaeus* (annual killifish), *Boleophthalmus pectinirostris* (Bpec; great blue-spotted mudskipper), *Channa argus* (Carg; northern snakehead), *Cyprinodon variegatus* (sheepshead minnow), *D. rerio* (Drer; zebrafish), *Fundulus heteroclitus* (mummichog), *Kryptolebias marmoratus* (Kmar; mangrove rivulus fish), *Monopterus albus* (Asian swamp eel), *Nothobranchius furzeri* (turquoise killifish), *Oreochromis aureus* (Oaur; blue tilapia), *O. niloticus* (Onil; Nile tilapia), *Oryzias latipes* (Olat; Japanese medaka), *O. melastigma* (Omel; marine medaka), *Periophthalmus magnuspinnatus* (Pmag; giant-fin mudskipper), *Poecilia mexicana* (Atlantic molly), *Xiphophorus maculatus* (southern platyfish), and *Maylandia zebra* (Mzeb; *Zebra mbuna*). After removal of alternative splice variants, the protein sequences of the 18 fish species along with the mirrorwing flyingfish (*H. speculiger*; Hspe) were delivered to OrthoFinder v2.3.11 (Emms and Kelly, 2019) with an E-value of 1e−5 to identify orthologous groups.

Protein sequences of single-copy orthologous families were extracted and aligned using MUSCLE v3.8 (Edgar, 2004), and the alignment of protein sequences was converted to codon alignment using PAL2NAL v14 (Suyama et al., 2006). The phase 1 sites of codon aligned were extracted and concentrated to a super gene for each species. PhyML v3.0 (Guindon et al., 2010) and MrBayes v3.2 (Ronquist et al., 2012) were employed to construct a phylogenetic tree. Divergence time of these teleost fishes was estimated using MCMCTREE v4.5 in the PAML v4.5 (Yang, 2007) with five putative calibrations times, which were adapted from TIMETREE (Kumar et al., 2017). We used CAFÉ v3.0 (Han et al., 2013) with optimized parameter (-p 0.05 -t 4 -r 10000 -filter) to assess expansion and contraction of gene families. A branch specific $p < 0.05$ was utilized to define significance in the mirrorwing flyingfish. We employed hypergeometric tests (Falcon and Gentleman, 2008) to investigate pathway enrichments of those significantly expanded gene families, using the whole genome annotation as the background.

Synteny Analysis With Medaka and Zebrafish Genomes

After masking transposon elements of the three genomes, pairwise genome alignment among mirrorwing flyingfish, Japanese medaka, and zebrafish was carried out using LASZT v1.04.03 (Harris, 2007) with optimized parameters ($T = 2$ $C = 2$ $H = 2000$ $Y = 3400$ $L = 6000$ $K = 2200$ –format = axt). The matching length of each pairwise alignment was calculated using an in-house Perl script.

TABLE 1 | Statistics of our genome assembly.

Parameter	Platanus contig		DBG2OLC		Pilon		BESST	
	Size (bp)	Number	Size (bp)	Number	Size (bp)	Number	Size (bp)	Number
N90	131	3,1,85,718	113237	1567	112663	1,567	161399	1,205
N80	161	2,2,24,282	235476	939	233652	939	318262	745
N70	212	1,4,16,513	396517	597	394432	597	513435	485
N60	315	8,49,429	635760	385	630993	385	831356	322
N50	514	4,85,451	998191	257	992826	257	1152470	215
Longest	36570	———	6848566	———	6813063	———	9488118	———
Total Size	1442411998	———	1047997551	———	1042531442	———	1043046751	———
> =100bp	———	4,47,1742	———	3852	———	3,852	———	3,052
> =2kb	———	98,312	———	3849	———	3,849	———	3,049

Platanus: primary contig assembly using Platanus; DBG2OLC: call consensus with blasr and the consensus module (sparc) using the previous result and PacBio subreads; Pilon: polish DBG2OLC result with pair-end reads; BESST: scaffold construct with mate-pair reads.

Identification of Vision-Related Genes

We applied two approaches to obtain the protein sequences of various opsins and *aanat* genes in 12 representative teleost fishes (with abbreviations of Ates, Bpec, Carg, Drer, Kmar, Mzeb, Oaur, Onil, Olat, Omel, Pmag, and Hspe, respectively, in **Supplementary Table 1**). For those with public annotations, gene sequences were directly downloaded from NCBI (**Supplementary Table 2**). For the mirrorwing flyingfish, however, we mapped the protein sequences of blue tilapia, zebrafish, and Japanese medaka to our assembled genome and predicted opsin and *aanat* genes using Exonerate v2.2.0 (Slater and Birney, 2005) with optimized parameters (-model protein2genome –showalignment false –showtargetgff true –bestn 1).

To validate the synteny of opsin genes, we downloaded those genes that have been reported to locate adjacent to an opsin gene (Lin et al., 2017) and obtained the neighboring genes from the genome annotation or using BLAST with an E-value of 1e−5 against the assembled genome. We constructed a rooted neighbor-joining (NJ) tree of opsins, using known opsin from human (ENSP00000358967.4, LWS1; ENSP00000472316.1, MWS; ENSP00000358945.4, MWS2; ENSP00000469970.1, MWS3; ENSP00000296271.3, RH1; ENSP00000249389.2, SWS1) and zebrafish (ENSDARP00000069184.5, OPN3; as the outgroup) by MEGA-X (Kumar et al., 2018) with 1,000 bootstraps.

A phylogenetic tree of *aanat* gene family was also constructed using the NJ method as implemented in the MEGA-X with human AANAT (NP_001079.1) and mouse AANAT (NP_033721.1) as the outgroup (Kumar et al., 2018). We applied Evolview (Subramanian et al., 2019) to edit phylogenetic trees. Five key tuning sites (including 180, 197, 277, 285, and 308) of the LWS opsins had influenced the λ_{max} of vertebrate opsins (Bowmaker, 2008; Yokoyama, 2008). A previous report suggested that a single mutation at S180A, H197Y, Y277F, T285A, A308S, and double mutations S180A/H197Y can lead to a −7, −28, −8, −15, −27, and −11 nm shift, respectively, in the λ_{max} of the pigments (Yokoyama and Radlwimmer, 2001). To investigate classical five key tuning sites of LWS, we obtained the global alignment of LWS in 12 teleost fishes and human being using MUSCLE v3.8 (Edgar, 2004) and highlighted the five crucial sites with Jalview v2.11.1.3 (Waterhouse et al., 2009). F86 of SWS1 opsin is crucial for UV sensing; the mutation of F86V in goldfish led to +1 nm shift in the absorption spectrum of the SWS1 opsins (Tada et al., 2009). The tuning site F86 resulting in the UV perception of SWS1 opsin in vertebrates (Hunt et al., 2007) was also checked in SWS1-containing teleost fishes.

TABLE 2 | Evaluation of the genome and gene completeness with BUSCO.

BUSCO	Genome		Gene	
	Numbers	Percent (%)	Numbers	Percent (%)
Total BUSCOs	4,584			
Complete BUSCOs	4,317	94.2	4,386	95.7
Complete and single-copy BUSCOs	4,074	88.9	4,103	89.5
Complete and duplicated BUSCOs	243	5.3	283	6.2
Fragmented BUSCOs	108	2.4	130	2.8
Missing BUSCOs	159	3.4	68	1.5

Characterization of Gadusol Biosynthesis Genes

To identify gadusol biosynthesis related genes, we extracted the *eevs*-like and *mt-ox* genes and genes adjoined to them in zebrafish, tilapia, and medaka genomes that were collected from the NCBI database (**Supplementary Table 3**) as the references and employed the same method as mentioned for the vision-related genes to predict *eevs*-like and *mt-ox* in in the mirrorwing flyingfish genome. For other 11 selected teleost fishes, we retrieved *eevs*-like and *mt-ox* from the NCBI annotation. We constructed a rooted NJ tree using a dehydroquinate synthase (DHQS-like) derived from cyanobacteria (Balskus and Walsh, 2010) as the outgroup by MEGA-X with 1,000 bootstraps. Conserved domains and motifs of the candidate *eevs*-like genes were predicted using the NCBI Conserved Domain

Database (CDD) (Lu et al., 2020) and MEME website server (Bailey et al., 2006), and then, TBtools suite was applied to illuminate the phylogenetic tree, conserved domains, and motifs (Chen et al., 2020).

Identification of Olfactory Receptor Genes

Reference sequences of olfactory receptor (OR) genes were obtained from a previous paper (Niimura, 2009). The full-length OR protein sequences were aligned to nine teleost fishes (including Ates, Bpec, Pmag, Carg, Kmar, Hspe, Drer, Oaur, and Olat) using tBLASTn (Ye et al., 2006) with an E-value of 1e−5, and the blasted hits were clustered using SOLAR v0.9 (Yu et al., 2006) to define candidate gene loci.

We extracted these candidate gene loci along with 2-kb flank region and employed GeneWise v2.4 (Birney et al., 2004) to predict gene structures. First, the potential OR genes without start/stop codons or with interrupting stop codon(s) or frameshift(s) were excluded. Second, the full-length sequences were inspected using the NCBI non-redundant database (BLASTP with an E-value of 1e−5), but those candidate OR genes with the best hit annotation of non-OR were discarded. Finally, the remaining sequences were further checked using TMHMM v2.0 (Krogh et al., 2001) to identify the putative seven transmembrane domains. We aligned the protein sequences of confirmed OR genes using MUSCLE in the MEGA-X (Kumar et al., 2018) and then constructed a rooted neighbor-joining tree using human G-protein coupled receptor 35 (NP_005292.2) and human G-protein coupled receptor 132 (NP_037477.1) as the outgroup by MEGA-X with the Poisson model and uniform rates.

RESULTS AND DISCUSSION

Summary of the Genome Assembly and Annotation

The Illumina sequencing generated a total of ∼138.13-Gb raw reads, and then, 99.21-Gb clean reads were retained after filtering low-quality sequences (**Supplementary Table 4**). The PacBio sequencing yielded about 29.98-Gb data, consisting of 2,785,344 reads with an N50 length of 16.5 kb (**Supplementary Table 5**).

A *k-mer* analysis predicted that the mirrorwing flyingfish had an estimated genome size of 1.06 Gb and a heterozygosity of 1.35% (**Figure 1B**). After contig building, consensus calling, polishing, and scaffold construction, we generated a final assembly of 1.04 Gb, which is nearly equal to the estimated genome size. The draft assembly consisted of 3,052 scaffolds (> 650 bp in length), and the contig and scaffold N50 values of our final assembly were 992.83 and 1,152.47 kb (**Table 1**).

The BUSCO evaluation indicated that 94.2% of the Actinopterygii gene sets were identified as complete (4,317 out of 4,584, actinopterygii_odb9) in the mirrorwing flyingfish genome (**Table 2**). We also assessed accuracy of the draft assembly by mapping Illumina paired-end reads onto the assembled genome sequences. A total of 94.91% of the Illumina paired-end reads were properly mapped to the assembled genome, with a good coverage of 97.78% (**Supplementary Table 6**). The high completeness of BUSCOs and nucleotide-level accuracy, together with considerable continuity of contig sizes, suggested that our high-quality genome assembly could be qualified for further data analysis.

Repeat content of the mirrorwing flyingfish genome was calculated by combination of both homolog-based and *de novo* methods. We determined that repeat elements occupied 42.02% of the assembled genome, and DNA transposons accounted for the largest proportion (24.38%) of transposable elements (TEs; **Supplementary Table 8**). A total of 8.19% of the mirrorwing flyingfish genome sequences were composed of tandem repeat elements (**Supplementary Table 7**). Divergence rates of the TEs in the mirrorwing flyingfish genome were determined using Repbase and *de novo* libraries, respectively. We observed that 10.72 Mb of identified TEs had a <10% divergence rate from the Repbase consensus; 277.08 Mb of TE sequences (26.56% of the assembly genome) had a <10% divergence rate from the *de novo* library (**Supplementary Figure 2**), which were possible to be active with a recent origin.

We predicted 23,611 protein-coding genes in the mirrorwing flyingfish genome, with an average gene length of 14.35 kb. Moreover, 99.50% of these genes could be functionally annotated by at least one of the four popular databases, with 20,692 KEGG hits, 21,453 SwissProt hits, 23,477 TrEMBL hits, and 21,888 Interpro hits (**Supplementary Table 9**). Additionally, the BUSCO evaluation of genes demonstrated that 95.7% of the Actinopterygii gene sets were predicted as complete (4,386 out of 4,584 actinopterygii_odb9) in the mirrorwing flyingfish gene set (**Table 2**), suggesting high quality of our gene prediction. Furthermore, we identified four types of non-coding RNA, 247 miRNAs, 2,138 tRNAs, 538 rRNAs, and 298 snRNAs in the assembled genome (**Supplementary Table 10**).

Gene Families and Phylogeny

Our gene family data demonstrated that protein-coding sequences in the 19 teleost fishes were clustered into 22,669 gene families, of which 4,632 families were 1:1 single-copy orthologs. A total of 93.5% (22,083 out of 23,611) of the mirrorwing flyingfish protein-coding genes were grouped into 17,352 gene families (**Supplementary Table 11**), defining 7,335 single-copy orthologs and 323 unique paralogs (**Supplementary Figure 3B**).

Using the 4,632 1:1 single-copy orthologous genes, we established a coincident phylogenetic topology with the ML and Bayes methods (**Supplementary Figures 4**, **5**). The divergence tree revealed that the flyingfish was close to the two medaka species with a divergence time of about 85.2 Mya (**Supplementary Figure 6**). A total of 60.71% (633.32 Mb) of the mirrorwing flyingfish genome was syntenic with Japanese medaka, while only 14.66% (152.94 Mb) of the mirrorwing flyingfish genome shared synteny with zebrafish (see more details in **Supplementary Table 12**).

We identified 1,236 expanded gene families and 1,539 contracted gene families in the mirrorwing flyingfish genome (**Supplementary Figure 3A**). Among them, 135 and 131 were significantly expanded and contracted ($p < 0.05$). The KEGG enrichment analysis demonstrated that those genes belonging to the expanded gene families were related to signaling

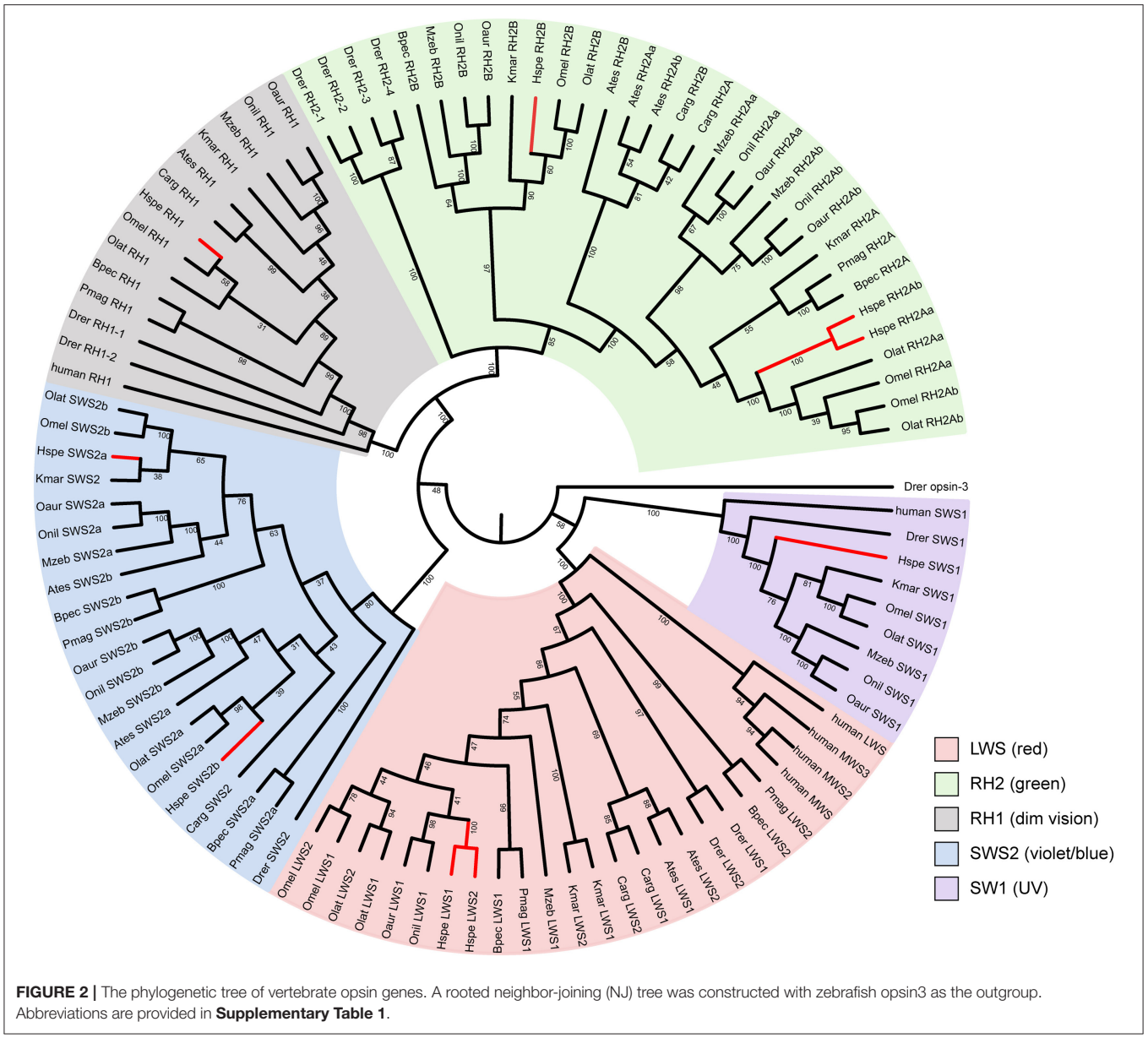

FIGURE 2 | The phylogenetic tree of vertebrate opsin genes. A rooted neighbor-joining (NJ) tree was constructed with zebrafish opsin3 as the outgroup. Abbreviations are provided in **Supplementary Table 1**.

molecules and interaction, nervous system, and immune system (**Supplementary Table 13**, $p < 0.01$).

Various Vision-Related Genes in the Mirrorwing Flyingfish

Vision plays a vital role in animal life, affording an important ability to perceive environmental stimuli. The visual ability of this animal depends on the numbers of opsin proteins (Bowmaker, 2008). Various fishes have accommodated a wide range of habitats (such as freshwater and marine, stagnant and running water, and shallow and deep sea), which provide differential vision adaptation (Hauser and Chang, 2017). We classified 12 teleost fishes into three groups in terms of living habitat, including genuine amphibious inhabitant (Ates, Bpec, Pmag, Carg, Kmar), normal underwater dweller (Drer, Oaur, Onil, Mzeb, Olat, Omel), and temporary water surface traveler (Hspe), for comparison of the variations among opsin proteins.

The mirrorwing flyingfish genome contains five types of opsins, with two LWS, two SWS2, one SWS1, one RH1, and three RH2 (**Figure 2**; **Table 3**). The maximal absorption spectra (λ_{max}) of flyingfish LWS, based on the popular "five-sites" rule (You et al., 2014), are predicted to be 560 nm, which is similar to the parameters in climbing perch, northern snakehead, mangrove rivulus, blue tilapia, Nile tilapia, zebra mbuna, Japanese medaka, and marine medaka (**Supplementary Table 14**). The five crucial sites of LWS in the mirrorwing flying fish are 180S, 197H, 277Y, 285T, and 308A (**Supplementary Figure 7**).

The synteny of opsins in 12 teleost fishes is quite conserved except *SWS1* (**Supplementary Figures 8**, **9**). All amphibious fishes except mangrove rivulus fish have lost *SWS1*

TABLE 3 | Copy number of vison-related genes in the 12 representative teleost fishes.

Species	Common Name	*LWS*	*SWS2*	*SWS1*	*RH1*	*RH2*	Total
A. testudineus	Climbing perch	2	2	0	1	3	8
B. pectinirostris	Blue-spotted mudskipper	2	2	0	1	2	7
P. magnuspinnatus	Giant-fin mudskipper	2	2	0	1	2	7
C. argus	Northern snakehead	2	1	0	1	2	6
H. speculiger	Mirrorwing flyingfish	2	2	1	1	3	9
K. marmoratus	Mangrove rivulus	2	1	1	1	2	7
O. aureus	Blue tilapia	1	2	1	1	3	8
O. niloticus	Nile tilapia	1	2	1	1	3	8
M. zebra	Zebra mbuna	1	2	1	1	3	8
O. latipes	Japanese medaka	2	2	1	1	3	9
O. melastigma	Indian medaka	2	2	1	1	3	9
D. rerio	Zebrafish	2	1	1	2	4	10

TABLE 4 | Copy number of *aanat* genes in the 12 representative teleost fishes.

Species	Common Name	Total Number	*aanat1a*	*aanat1b*	*aanat2*
A. testudineus	Climbing perch	3	1	1	1
B. pectinirostris	Blue-spotted mudskipper	3	1	1	1
P. magnuspinnatus	Giant-fin mudskipper	2	-	1	1
C. argus	Northern snakehead	3	1	1	1
H. speculiger	Mirrorwing flyingfish	3	1	1	1
K. marmoratus	Mangrove rivulus	3	1	1	1
O. aureus	Blue tilapia	3	1	1	1
O. niloticus	Nile tilapia	3	1	1	1
M. zebra	Zebra mbuna	3	1	1	1
O. latipes	Japanese medaka	3	1	1	1
O. melastigma	Indian medaka	3	1	1	1
D. rerio	Zebrafish	2	1	-	1

(**Supplementary Figure 8B**), which is used for UV vision. This *SWS1* missing could be related to the landing activity of these fishes. Since ultraviolet light can cause damages to the retina, the critical mutation of F86V could potentially alter absorption wave of SWS1 opsins toward violet light sensing so as to minimize the UV-induced damages (Cowing et al., 2002). These examined fishes in this study have V (valine) at 86 instead of F (phenylalanine; see **Supplementary Figure 10**), implying that these fishes could be UV sensing. Related amino acid numbering was based on the bovine rhodopsin sequence [GenBank accession no. M21606; (Palczewski et al., 2000)].

The five crucial sites of LWS in the mirrorwing flyingfish showed a narrow range of color sensing, demonstrating the same tendency as some amphibious fishes, such as climbing perch, northern snakehead, and mangrove rivulus fish. When these fishes move out of water, they can keep the same long-wave sensing as that in water. The *SWS1* loss events in the five examined amphibious fishes in our present study may have developed for the water-to-terrestrial adaptation; however, the reservation of *SWS1* in the mirrorwing flyingfish might be due to the short period of gliding in air instead of a real amphibious life (Davenport, 1994).

Low retinal dopamine levels could cause myopia (Feldkaemper and Schaeffel, 2013), and AANAT1a can reduce the dopamine content in the retina via acetylation (Zilberman-Peled et al., 2006). The loss of *aanat1a* in amphibious giant-fin mudskipper could be beneficial for movement in air (You et al., 2014). Interestingly, 12 teleost fishes except for giant-fin mudskipper have one copy of *annat1a* (see more details in **Table 4**; **Figure 3**). A previous study reported that the Atlantic flyingfish (*C. heterurus*) had a pyramidal shape cornea, which could assure both hypermetropic underwater vision and emmetropic vision in air (Baylor, 1967). Since the mirrorwing flyingfish owned three copies of *aanat* (without absence of *aanat1a*), its unique cornea might be responsible for a temporary air vision. Gadusol biosynthesis genes in the mirrorwing flyingfish we identify two copies of *eevs*-like and one copy of *mt-ox* in all the selected 12 fish genomes. Interestingly, the mirrorwing flyingfish has the same gene cluster as medaka, with *mdfic2* missing in the gene cluster "*foxp1b-mdfic2-mt-ox-eevs-a-mitfa-frmd4Ba*" (see more details in **Table 5**). All fishes shared the gene cluster of "*foxp1a-eevs-b-mitfb-frmd4Bb*" except for zebrafish (**Supplementary Figure 11**). Perhaps, the examined zebrafish genome was modified by genetic engineering

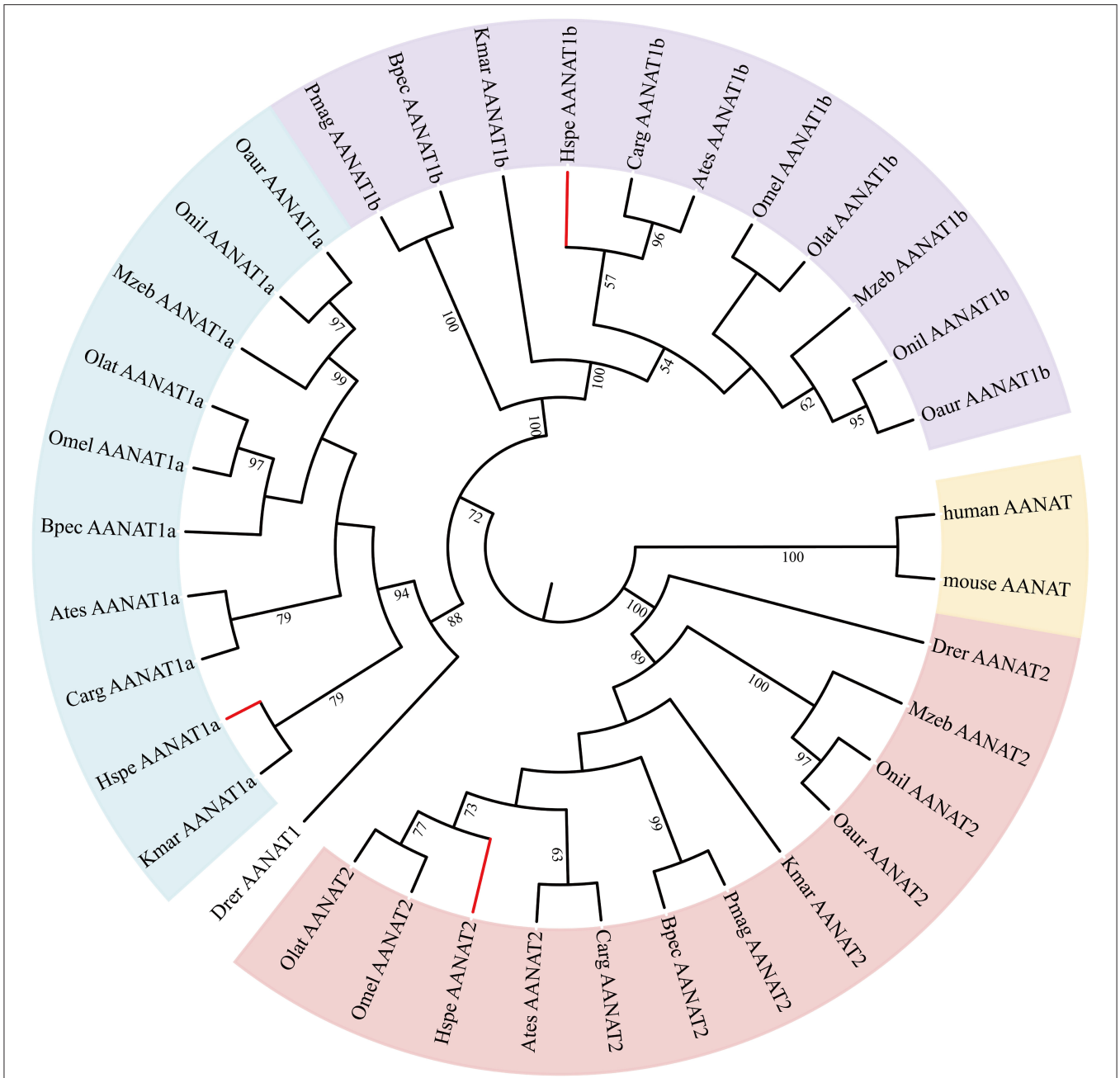

FIGURE 3 | The rooted NJ tree of vertebrate *aanat* genes. It was constructed with human AANAT (NP_001079.1) and mouse AANAT (NP_033721.1) as the outgroup.

(Carpio and Estrada, 2006). The two isotypes of *eevs*-like gene contain five exons, conserved domain CCD, and six conserved motifs (**Figure 4**). It seems that this Beloniformes species had experienced the same gene loss event.

Olfactory Genes in the Mirrorwing Flyingfish

Olfaction is an essential component of the animal sensory system for perceiving water- and air-soluble chemicals that can help to localize food, predators, and spawning migration sites (Hopfield, 1991). We identified 781 intact OR genes in nine representative fishes (**Supplementary Table 15**). These identified ORs could be classified into five subfamilies, including delta, epsilon, zeta, eta, and beta (see more details in **Supplementary Figure 12**).

The mirrorwing flyingfish possessed 50 intact OR genes; among them, the number of air-/waterborne OR genes were much less than climbing perch, northern snakehead, and zebrafish. Surprisedly, we could not find any airborne OR gene in the mirrorwing flyingfish genome. Although this fish could glide a while above water, the detailed classification and copy numbers

TABLE 5 | Genetic analysis of *eevs* and *mt-ox* genes in selected fishes.

Species	Common Name	*foxp1b* *foxp1a*	*mdfic2*	*mt-ox*	*eevsa* *eevsb*	*mitfa* *mitfb*	*frmd4Ba* *frmd4Bb*
A. testudineus	Climbing perch	√√	$\surd_2\times$	$\surd_2\times$	√√	√√	√√
B. pectinirostris	Blue-spotted mudskipper	√√	√×	√×	√√	√√	√√
P. magnuspinnatus	Giant-fin mudskipper	√√	√×	√×	√√	√√	√√
C. argus	Northern snakehead	√√	√×	√×	√√	√√	√√
H. speculiger	Mirrorwing flyingfish	√√	××	√×	√√	√√	√√
K. marmoratus	Mangrove rivulus	√√	√×	√×	×√	√√	√√
O. aureus	Blue tilapia	√√	√×	√×	√√	√√	√√
O. niloticus	Nile tilapia	√√	√×	√×	√√	√√	√√
M. zebra	zebra mbuna	√√	√×	√×	√√	√√	√√
O. latipes	Japanese medaka	√√	××	√×	√√	√√	√√
O. melastigma	Indian medaka	√√	××	√×	√√	√√	√√
D. rerio	Zebrafish	√×	√×	√×	√√	√√	√√

The $\surd_2$ means the climbing perch has two mdfic2 and mt-ox in the gene cluster as follows: foxp1b-mdfic2-mt-ox-mdfic2-mt-ox-eevs-a-mitfa-frmd4Ba; However, zebrafish doesn't have the following gene cluster: foxp1a-eevs-b-mitfb-frmd4Bb. More details are provided in ***Supplementary Figure 11****.*

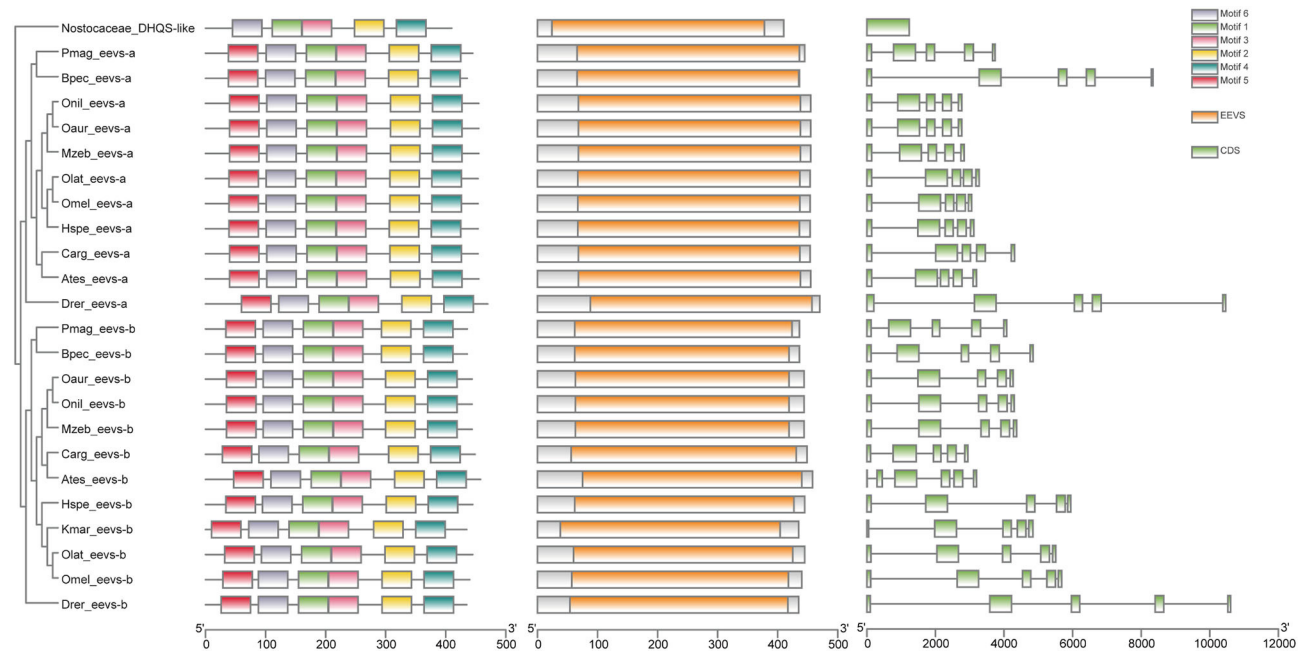

FIGURE 4 | The rooted NJ tree of teleost *eevs*-like genes. It was constructed with cyanobacteria DHQS-like as the outgroup. The first column is the rooted tree, the second column is the six motifs derived from MEME web service, the third column is the conserved domain CDD derived from NCBI, and the four column is the detailed structures of *eevs*-like genes.

of OR genes appear to be the same as those in medaka, while they are different from amphibious fishes (such as mudskippers; see You et al., 2014).

CONCLUSIONS

We obtained a draft genome assembly for the representative mirrorwing flyingfish with a hybrid method after Illumina and PacBio sequencing. We constructed a phylogenetic tree to illuminate the relationship of the mirrorwing flyingfish and other 18 teleost fishes. We also investigated vision-related genes, olfactory receptor genes, and gadusol synthesis-related genes in representative teleost fishes. Since the mirrorwing flyingfish could leave water for a while, it may exhibit similar traits as amphibious fishes. However, our genomic comparisons of vision-related and olfactory receptor genes revealed that the mirrorwing flyingfish potentially shared the same genetic mechanisms as its phylogenetic relatives (medaka species) but different from popular amphibious fishes (such as mudskippers). This high-quality genome assembly provides a valuable genetic resource

for the mirrorwing flyingfish, and it will also facilitate in-depth biomedical studies on various Exocoetoidea fishes.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/genbank/, PRJNA714815; https://figshare.com/, https://doi.org/10.6084/m9.figshare.14600634.v1.

ETHICS STATEMENT

The animal study was reviewed and approved by Animal Care and Use Committee of BGI (approval ID: FT18134).

AUTHOR CONTRIBUTIONS

QS conceived the project. PX, CZ, CB, and XY analyzed the data. XY, JC, ZR, FY, RG, and JX collected samples and assisted data analysis. PX and CZ wrote the manuscript. QS and CB revised the manuscript. All authors approved submission of the final manuscript for publication.

FUNDING

The work was financially supported by Shenzhen Science and Technology Innovation Program for International Cooperation (no. GJHZ20190819152407214) and Shenzhen Dapang Special Program for Industrial Development (no. KJYF202001-17).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.695700/full#supplementary-material

Supplementary Figure 1 | Pipeline of the genome assembly.

Supplementary Figure 2 | Distribution of divergence rates in each type of TEs within the mirrorwing flyingfish genome. The divergence rate was calculated between the identified TEs in the genome and the consensus sequences in the TE library used Repbase **(A)** or *de novo* libraries constructed by RepeatModeler and LTR_Finder **(B)**.

Supplementary Figure 3 | Evolution of 19 representative teleost genomes and their gene families. **(A)** Divergence tree with expanded and contracted gene families. **(B)** Statistics of single-copy orthologs, multiple-copy orthologs, unique paralogs, other orthologs, and unclustered gene numbers in the 19 teleost fishes.

Supplementary Figure 4 | A rooted phylogenetic tree (constructed with Bayes).

Supplementary Figure 5 | A rooted phylogenetic tree (constructed with PhyML).

Supplementary Figure 6 | A fossil-calibrated phylogenetic tree. It was constructed with the following five calibrated times that were adapted from the Timetree: *C. argus-A. testudineus* (66∼78Mya), *H. speculiger-O. latipes* (68∼89 Mya), *O. latipes-O. aureus* (104∼145 Mya), *B. pectinirostris-P. magnuspinnatus* (49∼69 Mya), and *D. rerio-O. aureus* (149∼165 Mya).

Supplementary Figure 7 | The crucial tuning sites in LWS opsins of 12 teleost fishes and human. Five critical sites in the mirrorwing flyingfish include S180A, H197Y, Y277F, T285A, and A308S. Abbreviations of fish species are provided in **Supplementary Table 1**.

Supplementary Figure 8 | The gene architecture of *RH1* **(A)** and *SWS1* **(B)** in 12 representative teleost fishes. Gene names were listed in the first line. Abbreviations of fish species are provided in **Supplementary Table 1**.

Supplementary Figure 9 | The gene architecture of *LWS-SWS2* **(A)** and *RH2* **(B)** in 12 representative teleost fishes. Gene names were listed in the first line. Abbreviations of fish species are provided in **Supplementary Table 1**.

Supplementary Figure 10 | The crucial tuning site of SWS1 (F86V) in 8 selected teleost fishes. The tuning site was marked in light blue. Abbreviations of fish species are provided in **Supplementary Table 1**.

Supplementary Figure 11 | The gene architecture of *eevs*-likes and *mt-ox* in12 representative teleost fishes. Gene names were listed in the first line. Abbreviations of fish species are provided in **Supplementary Table 1**.

Supplementary Figure 12 | A rooted neighbor-joining tree of olfactory receptor genes (ORs) in 9 selected teleost fishes. A total of 787 sequences were collected for construction of the circular cladogram tree. Various OR types were marked in different colors.

Supplementary Table 1 | Information of 19 teleost fishes used in our present study.

Supplementary Table 2 | Accession numbers of known aanat, opsin and neighboring genes.

Supplementary Table 3 | Accession numbers of known adjacent genes of eevs and mt-ox.

Supplementary Table 4 | Libraries and data yields for the whole genome shotgun sequencing.

Supplementary Table 5 | Libraries and data yields for the PacBio sequencing.

Supplementary Table 6 | The alignment result of paired-end reads mapping to H. speculiger genome.

Supplementary Table 7 | Summary of repeat annotations.

Supplementary Table 8 | Classification of repetitive elements.

Supplementary Table 9 | Statistics of function annotation.

Supplementary Table 10 | Statistics of Non-coding RNAs in the genome.

Supplementary Table 11 | Statistics of gene families in 19 teleost fishes.

Supplementary Table 12 | Statistics of pairwise alignment among mirrorwing flyingfish, medaka and zebrafish.

Supplementary Table 13 | KEGG enrichment for genes of expanded gene families of the mirrorwing flyingfish.

Supplementary Table 14 | Estimated maximal absorption spectrum (?max) of LWS.

Supplementary Table 15 | Copy numbers of intact OR genes in each group of fishes and mammals.

REFERENCES

Attwood, T. K., Croning, M. D. R., Flower, D. R., Lewis, A. P., Mabey, J. E., Scordis, P., et al. (2000). PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.* 28, 225–227. doi: 10.1093/nar/28.1.225

Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 34, W369–W373. doi: 10.1093/nar/gkl198

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2005). The universal protein resource (UniProt). *Nucleic Acids Res.* 33, D154–D159. doi: 10.1093/nar/gki070

Balskus, E. P., and Walsh, C. T. (2010). The genetic and molecular basis for sunscreen biosynthesis in

cyanobacteria. *Science* 329, 1653–1656. doi: 10.1126/science.1193637

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., et al. (2004). The Pfam protein families database. *Nucleic Acids Res.* 32, D138–D141. doi: 10.1093/nar/gkh121

Baylor, E. R. (1967). Air and water vision of the Atlantic flying fish, *Cypselurus heterurus*. *Nature* 214, 307–309. doi: 10.1038/214307a0

Baylor, E. R., and Shaw, E. (1962). Refractive error and vision in fishes. *Science* 136, 157–158. doi: 10.1126/science.136.3511.157

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27, 573–580. doi: 10.1093/nar/27.2.573

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and genomewise. *Genome Res.* 14, 988–995. doi: 10.1101/gr.1865504

Bowmaker, J. K. (2008). Evolution of vertebrate visual pigments. *Vision Res.* 48, 2022–2041. doi: 10.1016/j.visres.2008.03.025

Burge, S., Kelly, E., Lonsdale, D., Mutowo-Muellenet, P., Mcanulla, C., Mitchell, A., et al. (2012). Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. *Database* 2012:bar068. doi: 10.1093/database/bar068

Carpio, Y., and Estrada, M. P. (2006). Zebrafish as a genetic model organism. *Biotecnología Aplicada* 23, 265–270.

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Chen, N. (2004). Using Repeat Masker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* 5, 4–10. doi: 10.1002/0471250953.bi0410s05

Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., et al. (2018). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience* 7, 1–6. doi: 10.1093/gigascience/gix120

Cowing, J. A., Poopalasundaram, S., Wilkie, S. E., Robinson, P. R., Bowmaker, J. K., and Hunt, D. M. (2002). The molecular mechanism for the spectral shifts between vertebrate ultraviolet- and violet-sensitive cone visual pigments. *Biochem. J.* 367, 129–135. doi: 10.1042/bj20020483

Cui, L., Dong, Y., Cao, R., Gao, J., Cen, J., Zheng, Z., et al. (2018). Mitochondrial genome of the garfish *Hyporhamphus quoyi* (Beloniformes: Hemiramphidae) and phylogenetic relationships within Beloniformes based on whole mitogenomes. *PLoS ONE* 13, e0205025–e0205025. doi: 10.1371/journal.pone.0205025

Davenport, J. (1994). How and why do flying fish fly? *Rev. Fish Biol. Fish.* 4, 184–214. doi: 10.1007/BF00044128

De Bruin, G. H. P., Russell, B. C., and Bogusch, A. (1995). *FAO Species Identification Field Guide for Fishery Purposes. The Marine Fishery Resources of Sri Lanka.*

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238–238. doi: 10.1186/s13059-019-1832-y

Falcon, S., and Gentleman, R. (2008). "Hypergeometric testing used for gene set enrichment analysis," in *Bioconductor Case Studies*, eds F. Hahne, W. Huber, R. Gentleman, and S. Falcon (New York, NY: Springer), 207–220. doi: 10.1007/978-0-387-77240-0_14

Feldkaemper, M., and Schaeffel, F. (2013). An updated view on the role of dopamine in myopia. *Exp. Eye Res.* 114, 106–119. doi: 10.1016/j.exer.2013.02.007

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., et al. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U.S.A.* 117, 9451–9457. doi: 10.1073/pnas.1921046117

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Han, M. V., Thomas, G. W., Lugo-Martinez, J., and Hahn, M. W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol. Biol. Evol.* 30, 1987–1997. doi: 10.1093/molbev/mst100

Harris, R. S. (2007). *Improved pairwise alignmnet of genomic DNA* (PhD. Thesis). Pennsy lvania State University, Pennsylvania, United States.

Hauser, F. E., and Chang, B. S. W. (2017). Insights into visual pigment adaptation and diversity from model ecological and evolutionary systems. *Curr. Opin. Genet. Dev.*47, 110–120. doi: 10.1016/j.gde.2017.09.005

Hopfield, J. J. (1991). Olfactory computation and object perception. *Proc. Natl. Acad. Sci.* 88, 6462–6466. doi: 10.1073/pnas.88.15.6462

Hunt, D. M., Carvalho, L. S., Cowing, J. A., Parry, J. W. L., Wilkie, S. E., Davies, W. L., et al. (2007). Spectral tuning of shortwave-sensitive visual pigments in vertebrates. *Photochem. Photobiol.* 83, 303–310. doi: 10.1562/2006-06-27-IR-952

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., Mcanulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 110, 462–467. doi: 10.1159/000084979

Kageyama, H., and Waditee-Sirisattha, R. (2019). Antioxidative, anti-inflammatory, and anti-aging properties of mycosporine-like amino acids: molecular and cellular mechanisms in the protection of skin-aging. *Marine Drugs* 17, 222. doi: 10.3390/md17040222

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., et al. (2014). Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* 24, 1384–1395. doi: 10.1101/gr.170720.113

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–d361. doi: 10.1093/nar/gkw1092

Kim, O. T. P., Nguyen, P. T., Shoguchi, E., Hisata, K., Vo, T. T. B., Inoue, J., et al. (2018). A draft genome of the striped catfish, *Pangasianodon hypophthalmus*, for comparative analysis of genes relevant to development and a resource for aquaculture improvement. *BMC Genomics* 19, 733–733. doi: 10.1186/s12864-018-5079-x

Krogh, A., Larsson, B., Von Heijne, G., and Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580. doi: 10.1006/jmbi.2000.4315

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* 34, 1812–1819. doi: 10.1093/molbev/msx116

Kutschera, U. (2005). Predator-driven macroevolution in flyingfishes inferred from behavioural studies: historical controversies and a hypothesis. *Ann. Hist. Phil. Biol.* 10, 59–77.

Letunic, I., Doerks, T., and Bork, P. (2012). SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.* 40, D302–D305. doi: 10.1093/nar/gkr931

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 30, 2843–2851. doi: 10.1093/bioinformatics/btu356

Lin, J. J., Wang, F. Y., Li, W. H., and Wang, T. Y. (2017). The rises and falls of opsin genes in 59 ray-finned fish genomes and their implications for environmental adaptation. *Sci. Rep.* 7:15568. doi: 10.1038/s41598-017-15868-7

Lovejoy, N. R., Iranpour, M., and Collette, B. B. (2004). Phylogeny and jaw ontogeny of beloniform fishes. *Integr. Comp. Biol.* 44, 366–377. doi: 10.1093/icb/44.5.366

Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964. doi: 10.1093/nar/25.5.955

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., et al. (2020). CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* 48, D265–d268. doi: 10.1093/nar/gkz991

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Miyamoto, K. T., Komatsu, M., and Ikeda, H. (2014). Discovery of gene cluster for mycosporine-like amino acid biosynthesis from Actinomycetales microorganisms and production of a novel mycosporine-like amino acid

by heterologous expression. *Appl. Environ. Microbiol.* 80, 5028–5036. doi: 10.1128/AEM.00727-14

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., et al. (2015). Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* 43, D130–D137. doi: 10.1093/nar/gku1063

Nawrocki, E. P., and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29, 2933–2935. doi: 10.1093/bioinformatics/btt509

Niimura, Y. (2009). On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 chordate species. *Genome Biol. Evol.* 1, 34–44. doi: 10.1093/gbe/evp003

Osborn, A. R., Almabruk, K. H., Holzwarth, G., Asamizu, S., Ladu, J., Kean, K. M., et al. (2015). *De novo* synthesis of a sunscreen compound in vertebrates. *Elife* 4:e05919. doi: 10.7554/eLife.05919.028

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., et al. (2000). Crystal structure of rhodopsin: a G protein-coupled receptor. *Science* 289, 739–745. doi: 10.1126/science.289.5480.739

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596. doi: 10.1093/nar/gks1219

Rayner, J. M. V. (1986). Pleuston: animals which move in water and air. *Endeavour* 10, 58–64. doi: 10.1016/0160-9327(86)90131-6

Rhoads, A., and Au, K. F. (2015). PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* 13, 278–289. doi: 10.1016/j.gpb.2015.08.002

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

Rosic, N. N. (2019). Mycosporine-like amino acids: making the foundation for organic personalised sunscreens. *Marine Drugs* 17, 638. doi: 10.3390/md17110638

Rosic, N. N., and Dove, S. (2011). Mycosporine-like amino acids from coral dinoflagellates. *Appl. Environ. Microbiol.* 77, 8478–8486. doi: 10.1128/AEM.05870-11

Sahlin, K., Vezzi, F., Nystedt, B., Lundeberg, J., and Arvestad, L. (2014). BESST–efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics* 15, 281–281. doi: 10.1186/1471-2105-15-281

Servant, F., Bru, C., Carrère, S., Courcelle, E., Gouzy, J., Peyruc, D., et al. (2002). ProDom: automated clustering of homologous domains. *Brief. Bioinform.* 3, 246–251. doi: 10.1093/bib/3.3.246

Shick, J. M., and Dunlap, W. C. (2002). Mycosporine-like amino acids and related Gadusols: biosynthesis, acumulation, and UV-protective functions in aquatic organisms. *Annu. Rev. Physiol.* 64, 223–262. doi: 10.1146/annurev.physiol.64.081501.155802

Shinzato, C., Shoguchi, E., Kawashima, T., Hamada, M., Hisata, K., Tanaka, M., et al. (2011). Using the *Acropora digitifera* genome to understand coral responses to environmental change. *Nature* 476, 320–323. doi: 10.1038/nature10249

Sigrist, C. J., Cerutti, L., De Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., et al. (2010). PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.* 38, D161–166. doi: 10.1093/nar/gkp885

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Slater, G. S. C., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6, 1–11. doi: 10.1186/1471-2105-6-31

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 34, W435–439. doi: 10.1093/nar/gkl200

Subramanian, B., Gao, S., Lercher, M. J., Hu, S., and Chen, W. H. (2019). Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees. *Nucleic Acids Res.* 47, W270–w275. doi: 10.1093/nar/gkz357

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* 34, W609–612. doi: 10.1093/nar/gkl315

Tada, T., Altun, A., and Yokoyama, S. (2009). Evolutionary replacement of UV vision by violet vision in fish. *Proc. Natl. Acad. Sci.* 106, 17457–17462. doi: 10.1073/pnas.0903839106

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., et al. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* 13, 2129–2141. doi: 10.1101/gr.772403

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33, 2202–2204. doi: 10.1093/bioinformatics/btx153

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* 9, e112963–e112963. doi: 10.1371/journal.pone.0112963

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., and Barton, G. J. (2009). Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191. doi: 10.1093/bioinformatics/btp033

Wright, P. A., and Turko, A. J. (2016). Amphibious fishes: evolution and phenotypic plasticity. *J. Exp. Biol.* 219, 2245–2259. doi: 10.1242/jeb.126649

Xiong, Z., Li, F., Li, Q., Zhou, L., Gamble, T., Zheng, J., et al. (2016). Draft genome of the leopard gecko, *Eublepharis macularius*. *Gigascience* 5, s13742–s13016. doi: 10.1186/s13742-016-0151-4

Xu, G. H., Zhao, L. J., Gao, K. Q., and Wu, F. X. (2012). A new stem-neopterygian fish from the middle triassic of China shows the earliest over-water gliding strategy of the vertebrates. *Proc. Biol. Sci.* 280, 20122261–20122261. doi: 10.1098/rspb.2012.2261

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–268. doi: 10.1093/nar/gkm286

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Ye, C., Hill, C. M., Wu, S., Ruan, J., and Ma, Z. S. (2016). DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci. Rep.* 6, 31900–31900. doi: 10.1038/srep31900

Ye, C., and Ma, Z. S. (2016). Sparc: a sparsity-based consensus algorithm for long erroneous sequencing reads. *Peerj* 4, e2016–e2016. doi: 10.7717/peerj.2016

Ye, J., Mcginnis, S., and Madden, T. L. (2006). BLAST: improvements for better sequence analysis. *Nucleic Acids Res.* 34, W6–W9. doi: 10.1093/nar/gkl164

Yokoyama, S. (2000). Molecular evolution of vertebrate visual pigments. *Prog. Retin Eye Res.* 19, 385–419. doi: 10.1016/S1350-9462(00)00002-1

Yokoyama, S. (2008). Evolution of dim-light and color vision pigments. *Annu. Rev. Genomics Hum. Genet.* 9, 259–282. doi: 10.1146/annurev.genom.9.081307.164228

Yokoyama, S., and Radlwimmer, F. B. (2001). The molecular genetics and evolution of red and green color vision in vertebrates. *Genetics* 158, 1697–1710. doi: 10.1093/genetics/158.4.1697

You, X., Bian, C., Zan, Q., Xu, X., Liu, X., Chen, J., et al. (2014). Mudskipper genomes provide insights into the terrestrial adaptation of amphibious fishes. *Nat. Commun.* 5, 1–8. doi: 10.1038/ncomms6594

Yu, X. J., Zheng, H. K., Wang, J., Wang, W., and Su, B. (2006). Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup. *Genomics* 88, 745–751. doi: 10.1016/j.ygeno.2006.05.008

Zilberman-Peled, B., Ron, B., Gross, A., Finberg, J. P., and Gothilf, Y. (2006). A possible new role for fish retinal serotonin-N-acetyltransferase-1 (AANAT1): dopamine metabolism. *Brain Res.* 1073–1074, 220–228. doi: 10.1016/j.brainres.2005.12.028

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

GLOSSARY

Ates,	*Anabas testudineus*
Bpec,	*Boleophthalmus pectinirostris*
Carg,	*Channa argus*
Drer,	*Danio rerio*
Kmar,	*Kryptolebias marmoratus*
Mzeb,	*Maylandia zebra*
Oaur,	*Oreochromis aureus*
Onil,	*Oreochromis niloticus*
Olat,	*Oryzias latipes*
Omel,	*Oryzias melastigma*
Pmag,	*Periophthalmus magnuspinnatus*
Hspe,	*Hirundichthys speculiger*
OR,	olfactory receptor
AANAT,	aralkylamine N-acetyltransferase
TNPO3,	transportin 3
CALUA,	calumenin
SOCS2,	cytochrome c oxidase assembly protein
IRF5,	interferon regulatory factor 5
SWS1,	short wavelength-sensitive 1
HCFC1,	host cell factor C1
LWS,	long wavelength-sensitive
SWS2,	short wavelength-sensitive 2
TFE3b,	transcription factor binding to IGHM enhancer 3
GNL3L,	guanine nucleotide binding protein-like 3-like
SLC6A22.2,	solute carrier family 6 member 22, tandem duplicate 2
RH2,	green-sensitive
SLC6A22.1,	solute carrier family 6 member 22, tandem duplicate 1
SYNPR,	synaptoporin
PRICKLE2,	prickle homolog 2
RH1,	rhodopsin
ADAMTS9,	ADAM metallopeptidase with thrombospondin type 1 motif 9
MAGI1,	membrane-associated guanylate kinase, WW and PDZ domain containing 1
FRMD4B,	FERM domain containing 4B
MDFIC2,	MyoD family inhibitor domain-containing protein 2
FOXP1,	forkhead box P1
MITFA,	melanocyte inducing transcription factor a
MITFB,	melanocyte inducing transcription factor b
EEVS,	2-epi-5-epi-valiolone synthase
MT-Ox,	S-adenosyl-L-methionine-dependent methyltransferase
IRF10,	interferon regulatory factor 10
ATAXIN1,	ataxin-1
RAB32,	Ras-related protein Rab-32
STXBP5B,	syntaxin-binding protein 5b (tomosyn)
SASH1,	SAM and SH3 domain-containing protein 1

frontiers in Genetics

ORIGINAL RESEARCH
published: 04 August 2021
doi: 10.3389/fgene.2021.707228

Phylogenetic Analysis of Core Melanin Synthesis Genes Provides Novel Insights Into the Molecular Basis of Albinism in Fish

Chao Bian[1,2,3], Ruihan Li[2,3], Zhengyong Wen[2,3], Wei Ge[1] and Qiong Shi[2,3]**

[1] *Faculty of Health Sciences, Centre of Reproduction, Development and Aging, University of Macau, Taipa, China,* [2] *Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, Beijing Genomics Institute, BGI Academy of Marine Sciences, BGI Marine, Shenzhen, China,* [3] *College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China*

Melanin is the most prevalent pigment in animals. Its synthesis involves a series of functional genes. Particularly, teleosts have more copies of these genes related to the melanin synthesis than tetrapods. Despite the increasing number of available vertebrate genomes, a few systematically genomic studies were reported to identify and compare these core genes for the melanin synthesis. Here, we performed a comparative genomic analysis on several core genes, including tyrosinase genes (*tyr*, *tyrp1,* and *tyrp2*), premelanosome protein (*pmel*), microphthalmia-associated transcription factor (*mitf*), and solute carrier family 24 member 5 (*slc24a5*), based on 90 representative vertebrate genomes. Gene number and mutation identification suggest that loss-of-function mutations in these core genes may interact to generate an albinism phenotype. We found nonsense mutations in *tyrp1a* and *pmelb* of an albino golden-line barbel fish, in *pmelb* of an albino deep-sea snailfish (*Pseudoliparis swirei*), in *slc24a5* of cave-restricted Mexican tetra (*Astyanax mexicanus*, cavefish population), and in *mitf* of a transparent icefish (*Protosalanx hyalocranius*). Convergent evolution may explain this phenomenon since nonsense mutations in these core genes for melanin synthesis have been identified across diverse albino fishes. These newly identified nonsense mutations and gene loss will provide molecular guidance for ornamental fish breeding, further enhancing our in-depth understanding of human skin coloration.

Keywords: melanin synthesis pathway, core genes for melanin synthesis, albinism phenotype, nonsense mutation, phylogenetic analysis

OPEN ACCESS

Edited by:
Liang Guo,
Chinese Academy of Fishery Sciences (CAFS), China

Reviewed by:
Deshou Wang,
Southwest University, China
Zaijie Dong,
Chinese Academy of Fishery Sciences (CAFS), China

****Correspondence:***
Wei Ge
weige@um.edu.mo
Qiong Shi
shiqiong@genomics.cn

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *09 May 2021*
Accepted: *12 July 2021*
Published: *04 August 2021*

Citation:
Bian C, Li R, Wen Z, Ge W and Shi Q (2021) Phylogenetic Analysis of Core Melanin Synthesis Genes Provides Novel Insights Into the Molecular Basis of Albinism in Fish. Front. Genet. 12:707228. doi: 10.3389/fgene.2021.707228

INTRODUCTION

Pigment patterns and coloration of skin, feathers, hair and scales are among the most variable phenotypes in various vertebrates (Protas and Patel, 2008). Diverse coloration phenotypes present some substantial functions in mate selection, crypsis, aposematism of predators, and species recognition (Protas and Patel, 2008). Melanin is the most prevalent pigment in animals with the primary function of shielding against UV irradiation from sunlight. It is synthesized in pigment cells or chromatophores that are derived from neural crests (Singh and Nüsslein-Volhard, 2015).

Mammals and birds exhibit only one category of pigment cells, also named melanocytes including black and brown types. Reptiles and amphibians possess xanthophores and iridophores. In fishes, there are seven different types of pigment cells, including melanophores, xanthophores, iridophores, erythrophores, leucophores, cyanophores and erythro-iridophores (Nordlund, 1992; Singh and Nüsslein-Volhard, 2015).

The black pigment, also named melanin, is generated from tyrosine in the melanosome (del Marmol and Beermann, 1996). Two types of melanin are present in mammals and birds, including black eumelanin and lighter pheomelanin. In various fishes, however, only eumelanin is found (Burton, 2011). The detailed melanin synthesis pathway is summarized in **Figure 1**. Disruption of the melanogenesis process causes decreased pigmentation, which may lead to complete absence of melanin (Braasch et al., 2007).

The genetic basis of pigment cell development and differentiation is largely conserved between tetrapods and fishes. It is well documented that a big proportion of genes in the melanin synthesis pathway have been duplicated and differentially retained in teleosts (Braasch et al., 2007, 2009a), possibly due to the third round of whole genome duplication (also known as the teleost-specific whole genome duplication, TWGD) approximately 250–350 million years ago (Mya) (Jaillon et al., 2004). Therefore, teleosts usually possess more melanin synthesis genes than tetrapods. Melanin biosynthesis in vertebrates depends on three members of the tyrosinase family, including tyrosinase (*tyr*), and tyrosinase-related protein 1 and 2 (*tyrp1* and *tyrp2*). The *tyrp1* gene might play an additional role in survival and proliferation of melanocytes. The expression of *pmel* (premelanosome protein) and *tyrp1* genes is often regulated by *mitf* (microphthalmia-associated transcription factor a; see **Figure 1**; Braasch et al., 2009a; Bian et al., 2020).

The albinism phenotype (loss of melanin) is a fascinating trait of ornamental fishes, such as glass catfish, white cichlid fish and albino northern snakehead fish. The *casper* zebrafish (named after its ghost like appearance; White et al., 2008) and *casper* stickleback (Hart and Miller, 2017), displaying reduced pigmentation of melanosomes, can also be useful models for tumor engraftment and *in vivo* stem cell analyses with high sensitivity and resolution. Therefore, it is important to understand the detailed molecular mechanism and core genes involved in melanin biosynthesis, which can improve the efficiency of color breeding in ornamental fishes. This can also help to generate a wide variety of species to act as excellent research models. With an increasing number of available vertebrate genome sequences, we can analyze and identify core genes for melanin synthesis in various vertebrates. In this study, we analyzed genes involved in the melanin synthesis pathway in 90 representative vertebrates, including mammals, birds, reptiles, amphibians, Actinopterygii, and Chondrichthyes. For our better understanding of the melanin synthesis pathway in vertebrates, we explored several core genes in this pathway, such as *tyr*, *tyrp1*, *tyrp2*, *pmel*, *mitf,* and *solute carrier family 24 member 5* (*slc24a5*), in various vertebrates.

MATERIALS AND METHODS

Genome Data Collection and Gene Identification

In total, 90 representative vertebrate genome assemblies were selected. They were downloaded from the National Center for Biotechnology Information (NCBI; **Supplementary Table 1**). The scaffold N50 data of these genome assemblies was also shown in **Supplementary Table 1**. Each genome sequence was prepared for construction in a standard BLAST database for subsequent BLAST analyses. The published protein sequences of *tyra*, *tyrb*, *tyrp1*, *tyrp2*, *pmela*, *pmelb*, *mitfa*, *mitfb,* and *slc24a5* from reference genomes (Japanese medaka *Oryzias latipes*, human *Homo sapiens*, chicken *Gallus gallus*, African clawed frog *Xenopus laevis*, and green Anoli lizard *Anolis carolinensiswere*) were downloaded from the public Ensembl database (**Supplementary Table 2**).

In detail, the protein sequences of medaka, human, chicken, frog, and green lizard were used for aligning the assemblies of Actinopterygii, mammals, birds, amphibians, and reptile species by using tBLASTn (version 2.2.28, NCBI, Bethesda, MD, United States) (Mount, 2007) with an E-value of 10^{-5}. These alignments were further filtered and processed by using a Perl script to generate best-hit genomic regions containing putative target genes of each genome assembly with over 50% identity and aligned ratio. GeneWise v2.2.0 (Birney et al., 2004) was employed to predict target gene structures in the best-hit alignments from the 90 representative vertebrate genomes. Most importantly, if this pipeline identified a gene loss in any examined species, we then manually checked the tBLASTn alignments in this species to make sure that the data were indeed true. The gene was examined again in the final gene list from the genome annotation, if it has completed gene structure but its alignment identity and aligned ratio were between 45 and 50%. This lower identity and align ratio could be caused by rapid evolutionary rate of this gene in some species. Meanwhile, we also checked for synteny in genes form up and downstream of these genes as an additional indicator of orthology. Detailed copy numbers of above indicated genes in 90 vertebrate genomes and 9 representative genomes are provided in **Tables 1**, **2** and **Supplementary Tables 3, 4**.

Phylogenetic Analysis

Phylogenetic analysis was performed on the protein sequences. Multiple sequence alignments of amino acid sequences were conducted across various species using the Muscle software (Kumar et al., 2016). The trimAl software (Capella-Gutierrez et al., 2009) with default parameters was employed for alignment trimming. Phylogenetic trees were constructed by PhyML (version 3.1) (Guindon et al., 2010) with the amino acid substitution model of JTT + I + G and the maximum likelihood (ML) method. The protein sequences of ghost shark (*Callorhynchus milii*) were used as the outgroups. A total of 1,000 bootstrap replicates were applied for evaluation of their branch supports. The trees were displayed by FigTree.[1]

[1]http://tree.bio.ed.ac.uk/software/figtree

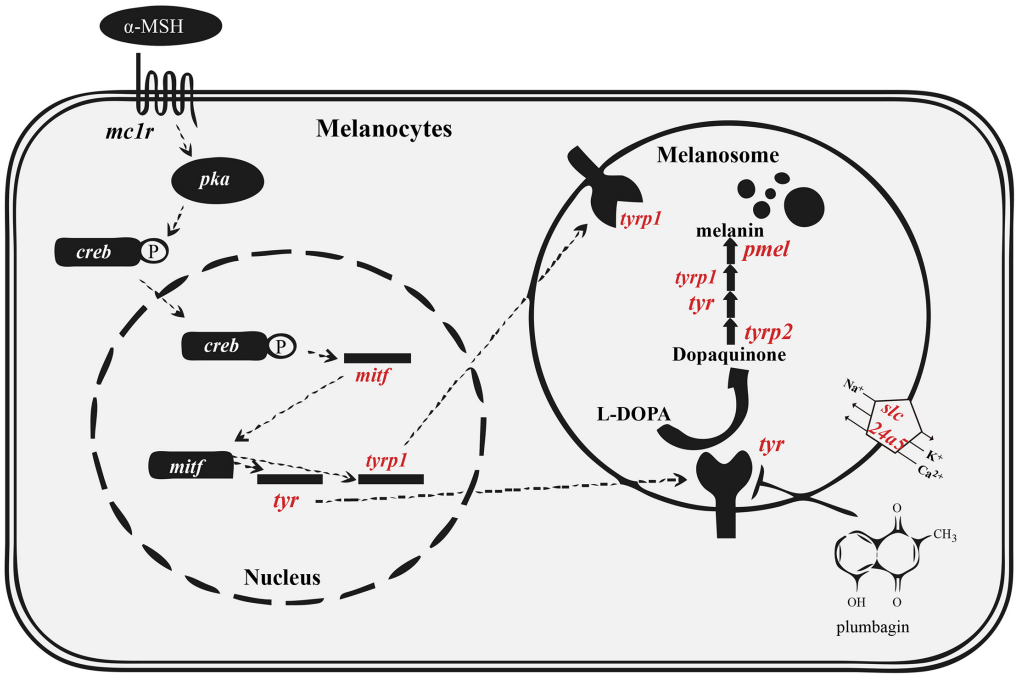

FIGURE 1 | Melanin synthesis pathway in vertebrates. The tyrosinase family (*tyr*, *tyrp1* and *tyrp2*) and *pmel* are essential for melanin synthesis. The melanosomal transporter (*slc24a5*) and regulatory factor *(mitf)* are also important for proper melanin synthesis. Red indicates the genes that we analyzed in detail. This figure was modified from a previous report (Oh et al., 2017). Abbreviations: α-MSH, alpha-melanocyte stimulating hormone; *creb*, cAMP response element binding protein; *mc1r*, melanocortin receptor; *pka*, protein kinase A. See other abbreviations in corresponding text.

TABLE 1 | Copy numbers of representative genes for melanin synthesis in 90 representative vertebrate genomes.

Class	Species number	*tyr*		*tyrp*		*tyrp2*	*pmel*		*mitf*		*slc24a5*
		tyra	*tyrb*	*tyrp1a*	*tyrp1b*		*pmela*	*pmelb*	*mitfa*	*mitfb*	
Actinopterygii	36	37	29	37	36	38 (2)	40 (1)	37 (4)	39 (1)	41	38 (1)
Birds	25	25		25		23 (2)	5		25		25
Amphibians	3	3		5		4	4		4		4
Mammals	13	11		12		12 (1)	13		13		13
Reptiles	12	9		12		13	8		12		12
Chondrichthyes	1	1		1		1	1		1		1

The data in brackets are the predicted copy numbers of pseudogenes.

Pseudogene Identification and Protein Sequence Alignment

We focused on identification of the core genes for melanin synthesis in several representative species with a loss-of-melanin phenotype. These vertebrates included *Astyanax mexicanus*, *Sinocyclocheilus anshuiensis, Protosalanx hyalocranius*, *Pseudoliparis swirei*, and *Danio rerio* (*nacre/casper* mutants). The corresponding nucleotide sequences of three-spined stickleback (*Gasterosteus aculeatus*), *D. rerio* (*Tu* strain), *Oryzias latipes, Anolis carolinensis, Homo sapiens,* and *Gallus gallus* were used as references for comparison. Multiple sequence alignments of the above-mentioned species were produced with the Muscle software in MEGA package (Kumar et al., 2016). Codon-based alignments were utilized to examine irregular shifts in the open reading frame for identification of possible pseudogenes, which were characterized by codon frameshifts or premature stop codon(s).

RESULTS

We screened 90 vertebrate genomes for *tyr*, *tyrp1*, *tyrp2*, *pmel*, *mitf,* and *slc24a5* genes, representing 36 Actinopterygii species, 13 mammals, 25 birds, 12 reptiles, three amphibians, and one chondrichthyan fish. Copy numbers of each gene are summarized in **Table 1**.

Tyrosinase Gene Family

Melanin synthesis in vertebrates involves several important tyrosinase family members, including *tyr*, *tyrp1,* and *tyrp2*

TABLE 2 | Copy numbers of representative genes for melanin synthesis in 9 representative vertebrate species.

Species	*tyr*		*tyrp*		*tyrp2*	*pmel*		*mitf*		
	tyra	*tyrb*	*tyrp1a*	*tyrp1b*		*pmela*	*pmelb*	*mitfa*	*mitfb*	*slc24a5*
A. mexicanus	1	1	1	1	1	1	1	0	0	0 (1)
P. hyalocranius	1	1	1	1	1	1	1	0 (1)	1	1
P. swirei	0	2	1	1	1	1	0 (1)	1	1	1
S. anshuiensis	2	0	0	2	1 (1)	2	0 (2)	0	2	1
D. rerio	1	0	1	1	1	1	1	1	1	1
O. latipes	1	1	1	1	1	1	1	1	1	1
H. sapiens	1		1		1	1		1		1
G. gallus	1		1		1	1		1		1
A. carolinensis	1		1		1	1		1		1
X. tropicalis	1		1		1	1		1		1

Species with melanin loss are marked in red. Predicted copy numbers of pseudogenes are provided in brackets.

(Oetting and Setaluri, 2007). The *tyr* and *tyrp1* genes have been doubled and subsequently preserved in the teleost ancestor after the TWGD. The protein product of *tyr* catalyzes the first two rate-limiting steps of melanin synthesis by converting tyrosine to DOPA, and then catalyzes dopaquinone to melanin (Oetting and Setaluri, 2007). Mutations in the *tyr* gene of the *sandy* zebrafish strain caused melanin loss (Kelsh et al., 1996). Similarly, mutations in *tyr* lead to oculocutaneous albinism type 1 (OCA1) in humans (Dessinioti et al., 2009). The protein product of *tyrp2* catalyzes dopachrome to DHI-2-carboxylic acid (DHICA), while the *tyrp1* could participate not only in stabilizing the *tyr* in melanosome membranes (Kobayashi and Hearing, 2007; Krauss et al., 2014), but also in the formation of indole-5,6-quinone carboxylic acid from DHICA. A mutation in the coding region of *tyrp1a* caused melanophore death in zebrafish (Krauss et al., 2014). The double knock-down experiment of *tyrp1a* and *tyrp1b* in zebrafish led to hypo-pigmented melanophores and brown pigment (Braasch et al., 2009b). The oculocutaneous albinism type 3 (OCA3) of human, also known as Rufous albinism, is caused by mutations (Arg93Cys) in *tyrp1* (Dessinioti et al., 2009). The mutations of *tyrp1* gene also generally caused chocolate or brown coat color in many mammals and birds, like mice (Zdarsky et al., 1990), dog (Hrckova Turnova et al., 2017), chicken (Li et al., 2019), and Japanese quail (Nadeau et al., 2007).

We found that most of the examined teleost fishes possessed two copies of *tyr* (*tyra and tyrb*) and *tyrp1* (*tyrp1a* and *tyrp1b*), but only one copy of *tyrp2*. However, in all examined actinopterygian genomes, only the cave-restricted *S. anshuiensis* lost the *tyrp1a* gene; however, nine fish species have lost the *tyrb* gene (see more details in **Supplementary Table 3**). Both pufferfishes (fugu and Tetraodon) have lost the *tyrp1b* gene, which is consistent with a previous study (Braasch et al., 2007). The three Chinese golden-line barbel fishes (*Sinocyclocheilus anshuiensis*, *S. grahami* and *S. rhinocerous*) have two *tyrp1b* copies, which are consistent with their shared lineage-specific whole genome duplication event. Two forms of *tyrp2* were retained in the three golden-line barbel fishes too; however, one *tyrp2* was a pseudogene with a premature stop codon in *S. anshuiensis* and *S. rhinocerous*. On the other hand, the majority of tetrapods have only one copy of *tyr*, *tyrp1,* and *tyrp2*. According to our reconstructed phylogenetic trees for the *tyr*, *tyrp1,* and *tyrp2* genes (**Supplementary Figures 1–3**), *tyr* genes could be clearly divided into five main groups to represent actinopterygians, amphibians, mammals, reptiles, and birds. Similar to a previous study (Braasch et al., 2007), the *tyr* and *tyrp1* phylogenetic trees (**Supplementary Figures 1, 2**) clearly revealed that *tyr* and *tyrp1* were doubled in teleosts, consistent with the TWGD event.

mitf

The *mitf* gene encodes a vital transcription factor that up-regulates *tyr*, *tyrp1*, *tyrp2,* and *pmel* expression for melanin synthesis (Hsiao and Fisher, 2014). Most of tetrapods possess only one *mitf* gene in their genomes (Steingrimsson et al., 2004). In mice, a representative tetrapod model, *mitf* was determined to be related to development of coat color, eye, osteoclasts, and mast cells (Hodgkinson et al., 1993; Widlund and Fisher, 2003; Steingrimsson et al., 2004).

However, duplicated copies of *mitf* genes (named *mitfa* and *mitfb*) are present in the majority of teleost genomes (Steingrimsson et al., 2004). These *mitf* genes have undergone subfunctionalization after genome duplication at least 100 Mya (Altschmied et al., 2002). It was reported that the *mitfa* gene was involved in melanin synthesis, and the *mitfb* gene coexpressed with *mitfa* in the retinal pigment epithelium at an appropriate time to compensate for loss of *mitfa* function in the *nacre* mutant (Lister et al., 2001). A premature stop codon, identified in the *mitfa* exon of *nacre* zebrafish mutants (see more details in **Figure 2**), caused the loss of melanin pigments in this mutant trunk (Lister et al., 1999; White et al., 2008).

In this study, we identified 135 *mitf* genes in 90 representative vertebrate species (see detailed copy numbers in **Supplementary Table 3**). Among these examined species, the *mitfa* gene was lost in the cavefish *A. mexicanus*, while it harbored a premature stop codon in *P. hyalocranius* (**Figure 2**). The phylogenetic tree of *mitfa* and *mitfb* across 90 vertebrate species (**Supplementary Figure 4**) clearly placed teleosts into one main branch that was split from the common ancestor of teleosts and tetrapods. Subsequently, the teleost branch was divided into *mitfa* and *mitfb* clades, independently.

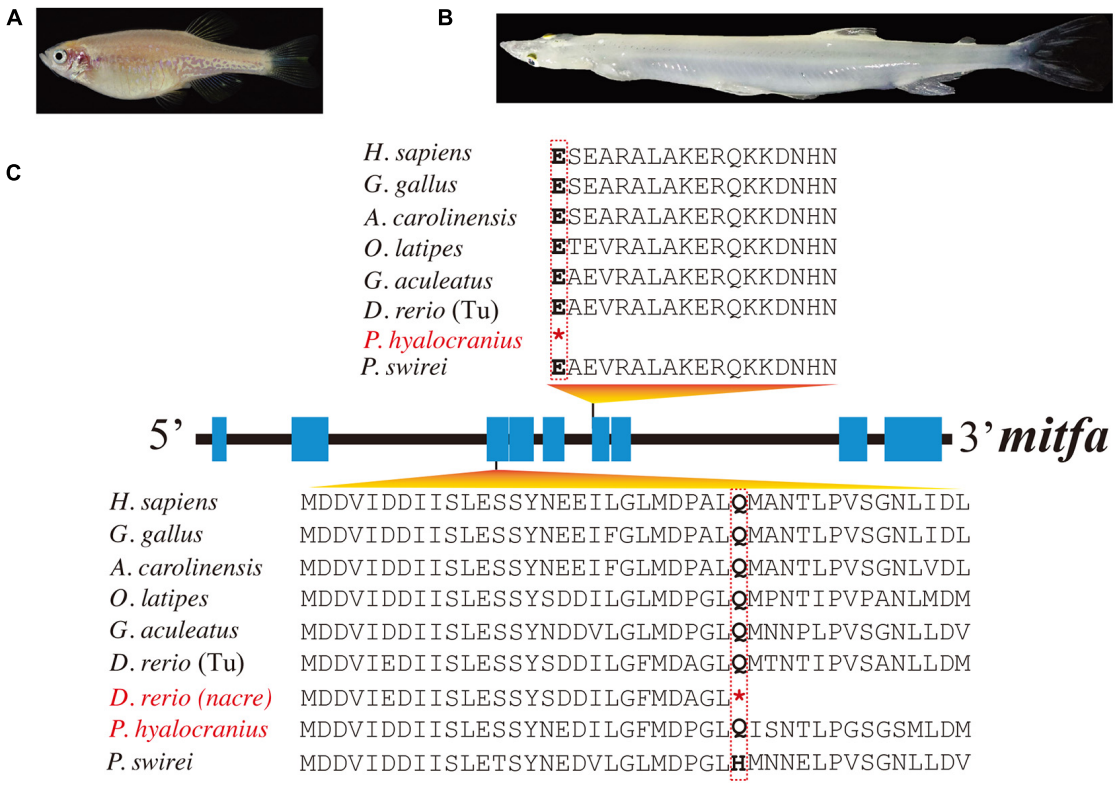

FIGURE 2 | Multiple sequence alignments of *mitfa* of nine representative species. **(A,B)** Photos of transparent *nacre* zebrafish and Chinese clearhead icefish (*P. hyalocranius*). **(C)** Nonsense mutations in the *mitfa* exons of both species. Species with melanin loss are marked in red. *Represents the premature stop codon.

pmel

The *pmel* gene (also known as *silver* locus) encodes a type I integral membrane protein that can catalyze eumelanin production (see **Figure 1**) from indole-5,6-quinone carboxylic acid during melanin synthesis (Chakraborty et al., 1996). Duplication of this gene has been reported in zebrafish (Schonthaler et al., 2005). *pmela* is expressed in melanophores and retinal pigment epithelium, but *pmelb* is expressed exclusively in retinal pigment epithelium (Schonthaler et al., 2005), which is similar to the previously reported subfunctionalization and distributions of the *mitfa* gene and *mitfb* gene in zebrafish (Lister et al., 2001). In mammals, *pmel* transcription is regulated by *mitf* (Du et al., 2003). Several missense mutations in horse *pmel* lead to a phenotype with a characteristic mixture of white and gray hairs (Brunberg et al., 2006). Similarly, a nonsense mutation in the *pmel* gene of Japanese quail caused a completely yellowish plumage phenotype (Ishishita et al., 2018). Furthermore, a 9-bp insertion in the exon 10 of chicken *pmel* led to a *Dominant white* phenotype (Kerje et al., 2004).

In our present study, we found that most teleosts have two *pmel* genes (**Supplementary Table 3**). Related phylogenetic data confirmed the duplication of *pmel* in teleosts after their divergence from tetrapods (**Supplementary Figure 5**). In contrast, *pmel* was lost in the majority of birds. Intriguingly, we observed that the *pmel* genes of *P. swirei* and *S. anshuiensis* commonly harbored premature stop codon mutations. More particularly, the two *pmelb* genes were truncated by nonsense mutations in *S. anshuiensis* (see more details in **Figure 3**).

slc24a5

slc24a5, encoding a transporter protein in the melanosome membrane, is essential for melanin synthesis (**Figure 1**). It has a pigmentation related function that was firstly identified in teleosts and later in mammals. Loss-of-function mutations in this gene led to reduced melanin concentration in zebrafish (Lamason et al., 2005). However, a non-synonymous mutation in the exon 2 of *slc24a5* in a horse generated bright orange-colored eyes (Mack et al., 2017). A recent study also reported that a non-synonymous mutation in *slc24a5* significantly correlated to skin color in African human populations (Crawford et al., 2017).

In the present study, we identified a total of 94 *slc24a5* genes (**Supplementary Tables 3, 4**). It has been clearly shown that most of the examined teleosts only have one copy of the *slc24a5* gene, suggesting that its duplicated paralogous gene was lost in the teleost ancestor after the TWGD. The *A. mexicanus* is a unique species that harbors a premature stop codon mutation in its *slc24a5* coding regions (**Figure 4** and **Table 1**). Interestingly, gene copy number and phylogenetic analyses demonstrated the unique presence of three copies of the *slc24a5* gene in Nile tilapia (*Oreochromis niloticus*, **Supplementary Figure 6**).

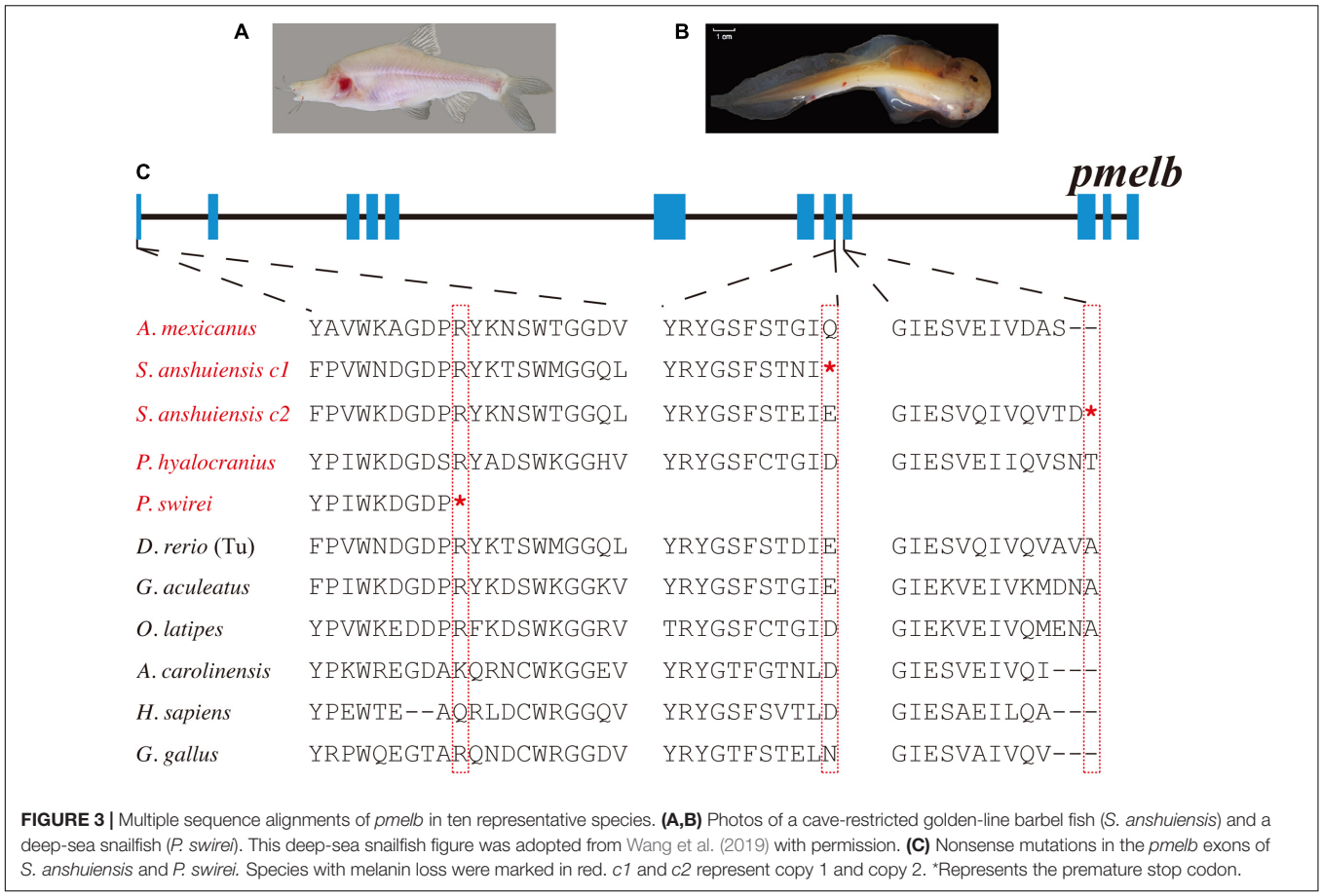

FIGURE 3 | Multiple sequence alignments of *pmelb* in ten representative species. **(A,B)** Photos of a cave-restricted golden-line barbel fish (*S. anshuiensis*) and a deep-sea snailfish (*P. swirei*). This deep-sea snailfish figure was adopted from Wang et al. (2019) with permission. **(C)** Nonsense mutations in the *pmelb* exons of *S. anshuiensis* and *P. swirei*. Species with melanin loss were marked in red. *c1* and *c2* represent copy 1 and copy 2. *Represents the premature stop codon.

DISCUSSION

Possible Molecular Clues Regarding the Melanin Loss Phenotype

Animals that are not exposed to light, such as cavefishes, are often colorless or transparent. The cave ecosystem is associated with several traits that are decreased or degenerated over time (i.e., "regressive" traits) (Protas et al., 2006). Notably, the reduced skin pigmentation in cavefishes is an important adaptation that has independently arisen in diverse species, such as *S. anshuiensis* (Yang et al., 2016) and *A. mexicanus* (Gross, 2012).

We identified premature stop variations and gene loss that are potentially leading to melanin loss in fishes with the albinism phenotype, including Mexican tetra (McGaugh et al., 2014), zebrafish (*Danio rerio, nacre/casper* mutants) (Howe et al., 2013; Bian et al., 2020), Chinese clearhead icefish (Liu et al., 2017), deep-sea Mariana snailfish (*Pseudoliparis swirei*), and cave-restricted golden-line barbel fish (*Sinocyclocheilus anshuiensis*; Yang et al., 2016). These premature stop variants and gene loss could provide novel molecular interpretations for melanin loss and albinism in various vertebrates.

Three high-quality genome assemblies of *Sinocyclocheilus* fishes (Cypriniformes: Cyprinidae), including *S. grahami, S. rhinocerous,* and *S. anshuiensis* have been reported before by us (Yang et al., 2016). *S. grahami* is a surface-dwelling species, *S. rhinocerous* is a semi-cave-dwelling species, while *S. anshuiensis* is a cave-restricted counterpart with an albinism phenotype (**Figure 3A**). These genome data of the three tetraploid fishes were useful for comparative identification of genetic evidence for melanin loss. By comparing these genome assemblies, we found that the *tyrp1a* gene has been lost in the cavefish *S. anshuiensis*. It has been well confirmed that a mis-sense mutation in the *tyrp1a* leads to melanophore death (Krauss et al., 2014). This loss-of-function *tryp1a* could be one potential cue for the melanin loss phenotype in *S. anshuiensis*.

Moreover, both *pmelb* genes in *S. anshuiensis* have a premature stop codon; however, in the other two *Sinocyclocheilus* fish species, *pmelb* genes are commonly normal. In the former case, this may result in functional loss of these genes (to be potential pseudogenes). In fact, many genes for regulating the development of some features may accumulate mutations, ultimately resulting in the loss of functions or even the disappearance of specific traits. The loss of these melanin synthesis genes in *S. anshuiensis* could be important for the white-skin phenotype [15] in this species. A similar nonsense mutation in the *pmelb* coding region is visible in a deep-sea snailfish, *P. swirei*, with a similar melanin loss phenotype (**Figure 3**). This could be a good example of convergent evolution for mutations in the *pmelb* genes of both white-skinned species that exhibit a melanin pigment loss phenotype.

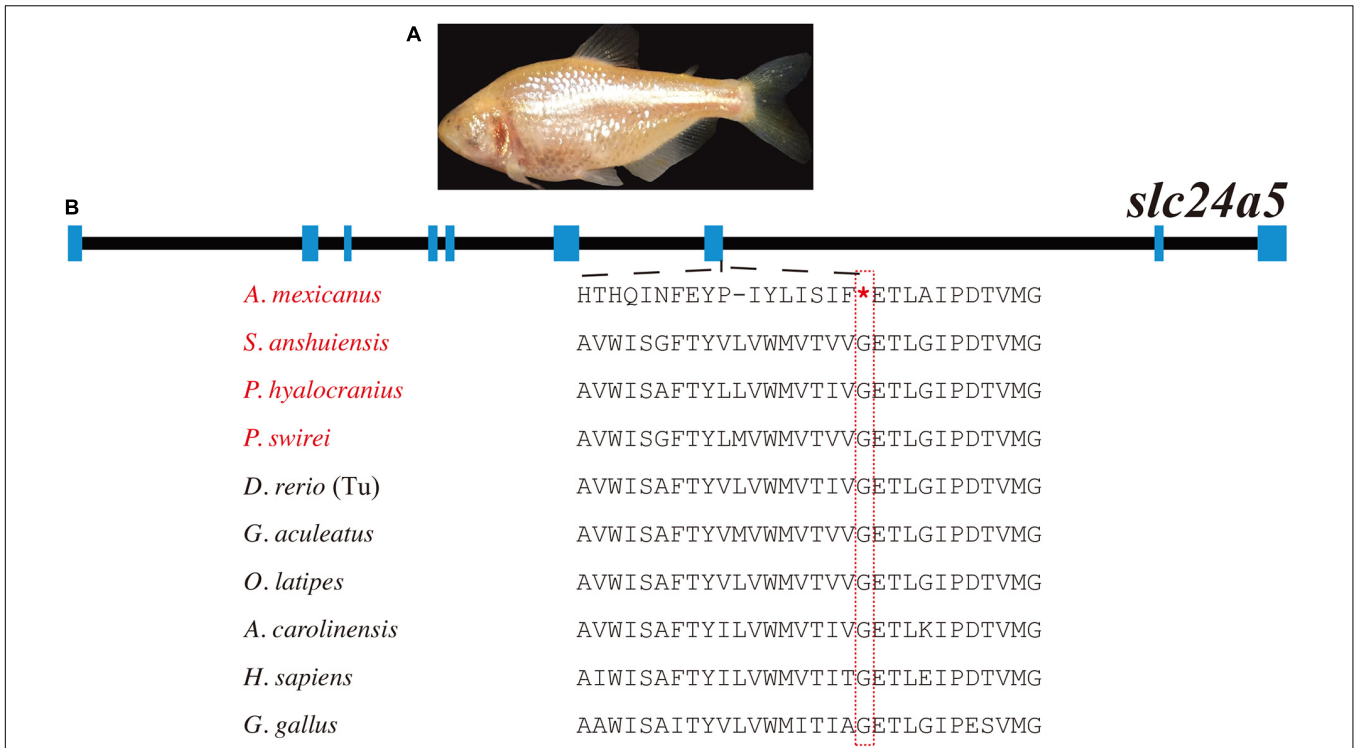

FIGURE 4 | Multiple sequence alignments of *slc24a5* in ten representative species. **(A)** Photo of *A. mexicanus* (cave population), which was downloaded from Wikipedia Commons (NIH Image Gallery, Bethesda, MD, United States). **(B)** A nonsense mutation in the *slc24a5* exon of *A. mexicanus*. Species with melanin loss were marked in red. *Represents the premature stop codon.

A. mexicanus (cave population) is a perfect diploid model for studying the regressive genetics in cavefishes. A previous study has reported that the albino cave population of *A. mexicanus* harbored deletions in its *oca2* gene (Protas et al., 2006) when compared with its surface counterparts. However, roles of these mutations cannot be critically validated by the gene knockout experiments; in fact, the experimental fish will lose the melanin, when each gene encoding the rate-limiting factors like *mitfa* (Koludrovic and Davidson, 2013), *slc24a5* (Lamason et al., 2005), *oca2* (Protas et al., 2006; Klaassen et al., 2018), *mc1r* (Gross et al., 2009), *tyr* (Kobayashi and Hearing, 2007) for melanin synthesis was knocked out or knocked down. Similarly, the gene gain or loss cannot be correctly identified by only using the QTL (quantitative trait locus) method (Protas et al., 2006) in diverse populations of *A. mexicanus*.

By comparing genes across fish species, we observed that the *A. mexicanus* cave population has lost the *mitfa* gene and harbors a nonsense mutation in *slc24a5*. We therefore speculate that the loss of *mitfa* could be a potential reason for melanin loss in this species and this gene loss may be resulted from several changes. The first step involved the accumulation of mutations in one gene, thereby resulting in missense or nonsense mutations. Since this would have caused a loss of function for this gene, and even the gene eventually disappeared from the genome. Therefore, the loss of *mitfa* gene could have appeared earlier than the nonsense mutation in *slc24a5*. A possible hypothesis could be considered as follows: the loss of *mitfa* led to a defect in the melanin synthesis pathway, resulting in the inactivation and gradual accumulation of neutral mutations in the up- and downstream genes of this pathway.

On the other hand, in addition to the black pigmentation defect, the *A. mexicanus* (cave population) exhibits remarkable yellow or golden skin (**Figure 4A**), which is similar to the reported *golden* zebrafish mutant with a mutation in *slc24a5*, thereby leading to more lightly pigmented and "golden" fish (Lamason et al., 2005). This similar phenotype provides additional genetic evidence to support the theory that the mutation of *slc24a5* could be possibly involved in this phenotype of *A. mexicanus* (cave population).

Another fish with an albinism phenotype, *P. hyalocranius*, has been reported in our previous paper (Liu et al., 2017). We also found a nonsense mutation in the sixth exon of *mitfa* (**Figure 2**). It has been well documented that the *mitfa* gene plays a core role in neural crest cell fate specification and melanocyte development (Levy et al., 2006; Wan et al., 2011; Koludrovic and Davidson, 2013). The *nacre/casper* zebrafish strain with a transparent phenotype also harbored a nonsense mutation in its *mitfa* gene (Lister et al., 1999; White et al., 2008). Therefore, the loss-of-function mutation of *mitfa* in *P. hyalocranius* could be the primary molecular mechanism for its melanin loss. Interestingly, the similar mutations in the *mitfa* genes from both *P. hyalocranius* and *nacre/casper* zebrafish were determined, and both fish species present with a similar phenotype of melanin loss in trunk.

CONCLUSION

Large numbers of sequenced vertebrate genomes have provided us a great opportunity to perform comparative genomics studies on some interesting genes. We collected 90 representative vertebrate genome assemblies with high quality to analyze detailed copy numbers and gene structures of core genes for melanin synthesis including *tyr, pmel, mitf,* and *slc24a5*. Phylogenetic analysis and genomic alignments were also performed in this study. We identified some novel genetic evidences that loss or nonsense mutations of these core melanin synthesis genes may contribute to melanin loss in these white-skinned fishes. These genetic resources will help to improve the practical breeding of ornamental fishes and create novel transparent models for theoretical researches. Our interesting findings in this study are also instructive for in-depth investigations of human skin coloration.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

All animal experiments were performed in accordance with the guidelines of the Animal Ethics Committee and were approved by the Institutional Review Board on Bioethics and Biosafety of BGI.

AUTHOR CONTRIBUTIONS

QS and WG conceived the project. CB, RL, and ZW performed the data analysis and figure preparation. CB prepared the manuscript. QS and WG revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was financially supported by the University of Macau (MYRG2016-00072-FHS, MYRG2017-00157-FHS, and CPG2014-00014-FHS) and the Macau Fund for Development of Science and Technology (FDCT089/2014/A2 and FDCT173/2017/A3) to WG, and Shenzhen Grant Plan for Demonstration City Project for Marine Economic Development in Shenzhen (No. 86) to QS.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.707228/full#supplementary-material

REFERENCES

Altschmied, J., Delfgaauw, J., Wilde, B., Duschl, J., Bouneau, L., Volff, J. N., et al. (2002). Subfunctionalization of duplicate mitf genes associated with differential degeneration of alternative exons in fish. *Genetics* 161, 259–267.

Bian, C., Chen, W., Ruan, Z., Hu, Z., Huang, Y., Lv, Y., et al. (2020). Genome and transcriptome sequencing of *casper* and *roy* zebrafish mutants provides novel genetic clues for iridophore loss. *Int. J. Mol. Sci.* 21:2385. doi: 10.3390/ijms21072385

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. *Genome Res.* 14, 988–995. doi: 10.1101/gr.1865504

Braasch, I., Brunet, F., Volff, J. N., and Schartl, M. (2009a). Pigmentation pathway evolution after whole-genome duplication in fish. *Genome Biol. Evol.* 1, 479–493. doi: 10.1093/gbe/evp050

Braasch, I., Liedtke, D., Volff, J. N., and Schartl, M. (2009b). Pigmentary function and evolution of tyrp1 gene duplicates in fish. *Pigment Cell Melanoma Res.* 22, 839–850. doi: 10.1111/j.1755-148X.2009.00614.x

Braasch, I., Schartl, M., and Volff, J. N. (2007). Evolution of pigment synthesis pathways by gene and genome duplication in fish. *BMC Evol. Biol.* 7:74. doi: 10.1186/1471-2148-7-74

Brunberg, E., Andersson, L., Cothran, G., Sandberg, K., Mikko, S., and Lindgren, G. (2006). A missense mutation in PMEL17 is associated with the Silver coat color in the horse. *BMC Genet.* 7:46. doi: 10.1186/1471-2156-7-46

Burton, D. (2011). "THE skin | coloration and chromatophores in fishes," in *Encyclopedia of Fish Physiology: from Genome to Environment*, ed. A. P. Farrell (Amsterdam: Elsevier), 489–496. doi: 10.1016/b978-0-12-374553-8.00041-1

Capella-Gutierrez, S., Silla-Martinez, J. M., and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348

Chakraborty, A. K., Platt, J. T., Kim, K. K., Kwon, B. S., Bennett, D. C., and Pawelek, J. M. (1996). Polymerization of 5,6-dihydroxyindole-2-carboxylic acid to melanin by the pmel 17/silver locus protein. *Eur. J. Biochem.* 236, 180–188. doi: 10.1111/j.1432-1033.1996.t01-1-00180.x

Crawford, N., Kelly, D., Hansen, M., Holsbach Beltrame, M., Fan, S., Bowman, S., et al. (2017). Loci associated with skin pigmentation identified in African populations. *Science* 358:eaan8433. doi: 10.1126/science.aan8433

del Marmol, V., and Beermann, F. (1996). Tyrosinase and related proteins in mammalian pigmentation. *FEBS Lett.* 381, 165–168. doi: 10.1016/0014-5793(96)00109-3

Dessinioti, C., Stratigos, A. J., Rigopoulos, D., and Katsambas, A. D. (2009). A review of genetic disorders of hypopigmentation: lessons learned from the biology of melanocytes. *Exp. Dermatol.* 18, 741–749. doi: 10.1111/j.1600-0625.2009.00896.x

Du, J., Miller, A. J., Widlund, H. R., Horstmann, M. A., Ramaswamy, S., and Fisher, D. E. (2003). MLANA/MART1 and SILV/PMEL17/GP100 are transcriptionally regulated by MITF in melanocytes and melanoma. *Am. J. Pathol.* 163, 333–343. doi: 10.1016/S0002-9440(10)63657-7

Gross, J. B. (2012). The complex origin of *Astyanax* cavefish. *BMC Evol. Biol.* 12:105. doi: 10.1186/1471-2148-12-105

Gross, J. B., Borowsky, R., and Tabin, C. J. (2009). A novel role for Mc1r in the parallel evolution of depigmentation in independent populations of the cavefish *Astyanax mexicanus*. *PLoS Genet.* 5:e1000326. doi: 10.1371/journal.pgen.1000326

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Hart, J. C., and Miller, C. T. (2017). Sequence-based mapping and genome editing reveal mutations in stickleback Hps5 cause oculocutaneous albinism and the *casper* phenotype. *G3* 7, 3123–3131. doi: 10.1534/g3.117.1125

Hodgkinson, C. A., Moore, K. J., Nakayama, A., Steingrimsson, E., Copeland, N. G., Jenkins, N. A., et al. (1993). Mutations at the mouse microphthalmia locus are

associated with defects in a gene encoding a novel basic-helix-loop-helix-zipper protein. *Cell* 74, 395–404. doi: 10.1016/0092-8674(93)90429-t

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., et al. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature* 496, 498–503. doi: 10.1038/nature12111

Hrckova Turnova, E., Majchrakova, Z., Bielikova, M., Soltys, K., Turna, J., and Dudas, A. (2017). A novel mutation in the TYRP1 gene associated with brown coat colour in the Australian shepherd dog breed. *Anim. Genet.* 48:626. doi: 10.1111/age.12563

Hsiao, J. J., and Fisher, D. E. (2014). The roles of microphthalmia-associated transcription factor and pigmentation in melanoma. *Arch. Biochem. Biophys.* 563, 28–34. doi: 10.1016/j.abb.2014.07.019

Ishishita, S., Takahashi, M., Yamaguchi, K., Kinoshita, K., Nakano, M., Nunome, M., et al. (2018). Nonsense mutation in PMEL is associated with yellowish plumage colour phenotype in Japanese quail. *Sci. Rep.* 8:16732. doi: 10.1038/s41598-018-34827-4

Jaillon, O., Aury, J. M., Brunet, F., Petit, J. L., Stange-Thomann, N., Mauceli, E., et al. (2004). Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431, 946–957. doi: 10.1038/nature03025

Kelsh, R. N., Brand, M., Jiang, Y. J., Heisenberg, C. P., Lin, S., Haffter, P., et al. (1996). Zebrafish pigmentation mutations and the processes of neural crest development. *Development* 123, 369–389. doi: 10.1242/dev.123.1.369

Kerje, S., Sharma, P., Gunnarsson, U., Kim, H., Bagchi, S., Fredriksson, R., et al. (2004). The Dominant white, Dun and Smoky color variants in chicken are associated with insertion/deletion polymorphisms in the PMEL17 gene. *Genetics* 168, 1507–1518. doi: 10.1534/genetics.104.027995

Klaassen, H., Wang, Y., Adamski, K., Rohner, N., and Kowalko, J. E. (2018). CRISPR mutagenesis confirms the role of oca2 in melanin pigmentation in *Astyanax mexicanus*. *Dev. Biol.* 441, 313–318. doi: 10.1016/j.ydbio.2018.03.014

Kobayashi, T., and Hearing, V. J. (2007). Direct interaction of tyrosinase with Tyrp1 to form heterodimeric complexes in vivo. *J. Cell Sci.* 120(Pt 24), 4261–4268. doi: 10.1242/jcs.017913

Koludrovic, D., and Davidson, I. (2013). MITF, the Janus transcription factor of melanoma. *Future Oncol.* 9, 235–244. doi: 10.2217/fon.12.177

Krauss, J., Geiger-Rudolph, S., Koch, I., Nusslein-Volhard, C., and Irion, U. (2014). A dominant mutation in tyrp1A leads to melanophore death in zebrafish. *Pigment Cell Melanoma Res.* 27, 827–830. doi: 10.1111/pcmr.12272

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Lamason, R. L., Mohideen, M. A., Mest, J. R., Wong, A. C., Norton, H. L., Aros, M. C., et al. (2005). SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* 310, 1782–1786. doi: 10.1126/science.1116238

Levy, C., Khaled, M., and Fisher, D. E. (2006). MITF: master regulator of melanocyte development and melanoma oncogene. *Trends Mol. Med.* 12, 406–414. doi: 10.1016/j.molmed.2006.07.008

Li, J., Bed'hom, B., Marthey, S., Valade, M., Dureux, A., Moroldo, M., et al. (2019). A missense mutation in TYRP1 causes the chocolate plumage color in chicken and alters melanosome structure. *Pigment Cell Melanoma Res.* 32, 381–390. doi: 10.1111/pcmr.12753

Lister, J. A., Close, J., and Raible, D. W. (2001). Duplicate mitf genes in zebrafish: complementary expression and conservation of melanogenic potential. *Dev. Biol.* 237, 333–344. doi: 10.1006/dbio.2001.0379

Lister, J. A., Robertson, C. P., Lepage, T., Johnson, S. L., and Raible, D. W. (1999). nacre encodes a zebrafish microphthalmia-related protein that regulates neural-crest-derived pigment cell fate. *Development* 126, 3757–3767. doi: 10.1242/dev.126.17.3757

Liu, K., Xu, D., Li, J., Bian, C., Duan, J., Zhou, Y., et al. (2017). Whole genome sequencing of Chinese clearhead icefish, Protosalanx hyalocranius. *Gigascience* 6, 1–6. doi: 10.1093/gigascience/giw012

Mack, M., Kowalski, E., Grahn, R., Bras, D., Penedo, M. C. T., and Bellone, R. (2017). Two variants in SLC24A5 are associated with "Tiger-Eye" Iris pigmentation in Puerto Rican Paso Fino horses. *G3* 7, 2799–2806. doi: 10.1534/g3.117.043786

McGaugh, S., Gross, J., Aken, B., Blin, M., Borowsky, R., Chalopin, D., et al. (2014). The cavefish genome reveals candidate genes for eye loss. *Nat. Comm.* 5:5307. doi: 10.1038/ncomms6307

Mount, D. W. (2007). Using the basic local alignment search tool (BLAST). *CSH Protoc.* 2007:dbto17. doi: 10.1101/pdb.top17

Nadeau, N. J., Mundy, N. I., Gourichon, D., and Minvielle, F. (2007). Association of a single-nucleotide substitution in TYRP1 with roux in Japanese quail (*Coturnix japonica*). *Anim. Genet.* 38, 609–613. doi: 10.1111/j.1365-2052.2007.01667.x

Nordlund, J. J. (1992). The pigmentary system and inflammation. *Pigment Cell Res.* 5(5 Pt 2), 362–365. doi: 10.1111/j.1600-0749.1992.tb00563.x

Oetting, W., and Setaluri, V. (2007). "The tyrosinase gene family," in *The Pigmentary System: Physiology and Pathophysiology*, 2nd Edn, eds J. J. Nordlund, R. E. Boissy, V. J. Hearing, R. A. King, W. S. Oetting, J.-P. Ortonne, et al. (Hoboken, NJ: Blackwell Publishing Ltd), 213–229. doi: 10.1002/9780470987100.ch11

Oh, T. I., Yun, J. M., Park, E. J., Kim, Y. S., Lee, Y. M., and Lim, J. H. (2017). Plumbagin suppresses alpha-MSH-induced melanogenesis in B16F10 mouse melanoma cells by inhibiting tyrosinase activity. *Int. J. Mol. Sci.* 18:320. doi: 10.3390/ijms18020320

Protas, M., and Patel, N. (2008). Evolution of coloration patterns. *Ann. Rev. Cell Dev. Biol.* 24, 425–446. doi: 10.1146/annurev.cellbio.24.110707.175302

Protas, M. E., Hersey, C., Kochanek, D., Zhou, Y., Wilkens, H., Jeffery, W. R., et al. (2006). Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. *Nat. Genet.* 38, 107–111. doi: 10.1038/ng1700

Schonthaler, H. B., Lampert, J. M., von Lintig, J., Schwarz, H., Geisler, R., and Neuhauss, S. C. (2005). A mutation in the silver gene leads to defects in melanosome biogenesis and alterations in the visual system in the zebrafish mutant fading vision. *Dev. Biol.* 284, 421–436. doi: 10.1016/j.ydbio.2005.06.001

Singh, A., and Nüsslein-Volhard, C. (2015). Zebrafish stripes as a model for vertebrate colour pattern formation. *Curr. Biol.* 25, R81–R92. doi: 10.1016/j.cub.2014.11.013

Steingrimsson, E., Copeland, N. G., and Jenkins, N. A. (2004). Melanocytes and the microphthalmia transcription factor network. *Annu. Rev. Genet.* 38, 365–411. doi: 10.1146/annurev.genet.38.072902.092717

Wan, P., Hu, Y., and He, L. (2011). Regulation of melanocyte pivotal transcription factor MITF by some other transcription factors. *Mol. Cell Biochem.* 354, 241–246. doi: 10.1007/s11010-011-0823-4

Wang, K., Shen, Y., Yang, Y., Gan, X., Guichun, L., Hu, K., et al. (2019). Morphology and genome of a snailfish from the Mariana Trench provide insights into deep-sea adaptation. *Nat. Ecol. Evol.* 3, 823–833. doi: 10.1038/s41559-019-0864-8

White, R. M., Sessa, A., Burke, C., Bowman, T., LeBlanc, J., Ceol, C., et al. (2008). Transparent adult zebrafish as a tool for in vivo transplantation analysis. *Cell Stem Cell* 2, 183–189. doi: 10.1016/j.stem.2007.11.002

Widlund, H. R., and Fisher, D. E. (2003). Microphthalamia-associated transcription factor: a critical regulator of pigment cell development and survival. *Oncogene* 22, 3035–3041. doi: 10.1038/sj.onc.1206443

Yang, J., Chen, X., Bai, J., Fang, D., Qiu, Y., Jiang, W., et al. (2016). The Sinocyclocheilus cavefish genome provides insights into cave adaptation. *BMC Biol.* 14:1. doi: 10.1186/s12915-015-0223-4

Zdarsky, E., Favor, J., and Jackson, I. J. (1990). The molecular basis of brown, an old mouse mutation, and of an induced revertant to wild type. *Genetics* 126, 443–449. doi: 10.1093/genetics/126.2.443

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 30 September 2021
doi: 10.3389/fgene.2021.730255

Whole-Genome Sequencing and Genome-Wide Studies of Spiny Head Croaker (*Collichthys lucidus*) Reveals Potential Insights for Well-Developed Otoliths in the Family Sciaenidae

Wu Gan[1,2†], *Chenxi Zhao*[3,4†], *Xinran Liu*[3,4], *Chao Bian*[3,4], *Qiong Shi*[3,4], *Xinxin You*[3,4*] and *Wei Song*[1,3,4*]

[1]East China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Shanghai, China, [2]Key Laboratory of Exploration and Utilization of Aquatic Genetic Resources, Ministry of Education, Shanghai Ocean University, Shanghai, China, [3]BGI Education Center, University of Chinese Academy of Sciences, Shenzhen, China, [4]Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, BGI Academy of Marine Sciences, BGI Marine, BGI, Shenzhen, China

Spiny head croaker (*Collichthys lucidus*), belonging to the family Sciaenidae, is a small economic fish with a main distribution in the coastal waters of Northwestern Pacific. Here, we constructed a nonredundant chromosome-level genome assembly of spiny head croaker and also made genome-wide investigations on genome evolution and gene families related to otolith development. A primary genome assembly of 811.23 Mb, with a contig N50 of 74.92 kb, was generated by a combination of 49.12-Gb Illumina clean reads and 35.24 Gb of PacBio long reads. Contigs of this draft assembly were further anchored into chromosomes by integration with additional 185.33-Gb Hi-C data, resulting in a high-quality chromosome-level genome assembly of 817.24 Mb, with an improved scaffold N50 of 26.58 Mb. Based on our phylogenetic analysis, we observed that *C. lucidus* is much closer to *Larimichthys crocea* than *Miichthys miiuy*. We also predicted that many gene families were significantly expanded (p-value <0.05) in spiny head croaker; among them, some are associated with "calcium signaling pathway" and potential "inner ear functions." In addition, we identified some otolith-related genes (such as otol1a that encodes Otolin-1a) with critical deletions or mutations, suggesting possible molecular mechanisms for well-developed otoliths in the family Sciaenidae.

Keywords: spiny head croaker, whole genome sequencing, chromosome-level genome assembly, otolith development, Sciaenidae

OPEN ACCESS

Edited by:
Liang Guo,
Chinese Academy of Fishery Sciences (CAFS), China

Reviewed by:
Manu Kumar Gundappa,
University of Edinburgh, United Kingdom
Changxu Tian,
Guangdong Ocean University, China
Huaping Zhu,
Chinese Academy of Fishery Sciences, China

***Correspondence:**
Xinxin You
youxinxin@genomics.cn
Wei Song
swift83@sina.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *24 June 2021*
Accepted: *06 September 2021*
Published: *30 September 2021*

Citation:
*Gan W, Zhao C, Liu X, Bian C, Shi Q, You X and Song W (2021) Whole-Genome Sequencing and Genome-Wide Studies of Spiny Head Croaker (*Collichthys lucidus*) Reveals Potential Insights for Well-Developed Otoliths in the Family Sciaenidae. Front. Genet. 12:730255. doi: 10.3389/fgene.2021.730255*

INTRODUCTION

Spiny head croaker (*Collichthys lucidus*), belonging to the family Sciaenidae, is a small economic fish with a main distribution in the coastal waters of Northwestern Pacific (Cheng et al., 2012a), from Philippines, China to Japan. With excellent properties and good meat quality, spiny head croaker has been favored by Chinese consumers with a high market value, whereas it has been overfished in the Yangtze estuary area of China (Hu et al., 2015). Furthermore, Sciaenidae fishes are well known for their well-developed otoliths (Xu et al., 2016), which are acellular crystalline mineral deposits in the inner ears of various teleost fishes (Pracheil et al., 2019).

Otoliths, mainly composed of calcium carbonate and organic matrix, play vital roles in sound sensing, balance, linear acceleration, and gravity in bony fishes (Schulz-Mirbach et al., 2019). Moreover, otoliths are widely applied in fisheries sciences, such as evaluation of fish populations or population migration patterns, and are essential for paleoichthyological and archeological studies (Barrett, 2019; Barnett et al., 2020; Heimbrand et al., 2020). Several critical genes, including Otoconin-90 (Oc90), secreted protein acidic and rich in cysteine (SPARC), SPARC-like1 (SPARCL1), otopetrin-1 (otop1), otolin-1 (otol1), and Otolithmatrixprotein-1 (OMP-1), have been identified to be related to otolith growth and formation in various vertebrates including bony fishes (Hurle et al., 2003; Murayama et al., 2005; Kang et al., 2008; Petko et al., 2008; Xu et al., 2016b; Hołubowicz et al., 2017). Nevertheless, a systematic screening of otolith-related genes in Sciaenidae species has not been reported yet.

With the rapid development of genome sequencing technology and genome-based bioinformatics methods, studies on aquatic genomes and related applications, such as molecular breeding, drug development, new biomaterials, and DNA barcoding technology, have been accumulated (Zhang et al., 2019; Houston et al., 2020; Li et al., 2020; Long et al., 2020). Whole-genome sequencing (WGS) of about 90 fishes has been published around the world by far (Bian et al., 2019; You et al., 2020). At present, genome studies on the Sciaenidae family have focused on the popular large yellow croaker (*Larimichthys crocea*) and its economic traits (Wu et al., 2014; Ao et al., 2015; Gui et al., 2018; Mu et al., 2018; Chen et al., 2019), while there are only few reports on other Sciaenidae species such as spiny head croaker.

Mitochondrial genome maps of spiny head croaker and candidate genes related to its sex determination have been examined by mitochondrial genome sequencing (Cheng et al., 2012b), chromosome assembly (Cai et al., 2019), and RNA sequencing (Song et al., 2020). However, the genetic basis of well-developed otoliths in Sciaenidae species is still unknown. It is therefore necessary to explore genomic resources to gain insights into otolith development mechanisms and to accelerate genome-assisted improvements in biodiversity protection, breeding, and disease prevention of Sciaenidae species.

In the past decades, Illumina short-read sequencing technology has been generally employed to assemble various fish genomes. However, with cost reduction of both short- and long-read sequencing technologies, more and more recent WGS projects have introduced PacBio long-read sequences in order to assemble high repetitive regions and improve assembly quality (You et al., 2020). Here, we produced a nonredundant chromosome-level assembly of the spiny head croaker by combination of Illumina short reads, PacBio Single Molecule Real-Time (SMRT) long reads, high-throughput chromosome conformation capture (Hi-C) data, and transcriptome sequences. Moreover, we performed comparative genomics studies on candidate genes related to otolith development to figure out potential mechanisms for the well-developed otoliths in the family Sciaenidae.

MATERIALS AND METHOD

Sample Collection and Sequencing

We extracted genomic DNAs from pooled muscle tissues of a wild female spiny head croaker and sequenced by using an Illumina HiSeq 2500 sequencing platform (San Diego, CA, United States) and a PacBio Sequel sequencing platform (Menlo Park, CA, United States). The construction of DNA libraries (insert sizes of 500 and 800 bp for Illumina, and 20 kb for PacBio) and subsequent sequencing were performed according to the standard protocols. In total, 52.48 Gb of raw Illumina data and five SMRT cells produced using the P6 polymerase/C4 chemistry, producing 35.24 Gb of PacBio long reads, were generated. After filtering by SOAPnuke (v.1.5.6; Chen et al., 2017), we obtained 49.12 Gb of Illumina clean data and 35.24 Gb of PacBio data for subsequent assembly.

To acquire a chromosome-level assembly of the genome, genomic DNAs were fixed with formaldehyde and were sheared by a restriction enzyme (MboI) to build a Hi-C library, and then sequenced by an Illumina HiSeq X Ten platform. A total of 185.33 Gb of 150 PE Hi-C data were generated. All sequenced data generated in this study were deposited in the CNGB Nucleotide Sequence Archive under Program no. CNP0001197.

We also extracted muscle RNA for transcriptome sequencing by using a HiSeq 2500 platform. Furthermore, to obtain the full-length transcript, the mixed RNA sample from 13 tissues was transcribed to generate full-length cDNA, and the SMRT bell library was constructed using the SMRT bell Template Prep Kit. The libraries were then prepared for sequencing on the PacBio Sequel sequencing platform.

Genome Assembly and Chromosome Assembly

Firstly, we employed Kmerfreq (https://github.com/fanagislab/kmerfreq) to estimate the genome size with 17-bp k-mers and applied GenomeScope (v1.0; Vurture et al., 2017) to estimate genome heterozygosity. Subsequently, a hybrid genome assembly pipeline was employed to obtain genome assembly. Short Illumina reads were first assembled by using Platanus with "-m 300 -k 27 -s 3" (Kajitani et al., 2014), and DBG2OLC (Ye et al., 2016) was performed to combine Platanus-generated contigs with PacBio reads to generate a hybrid contig assembly with default parameters. Pilon (v.1.225; (Walker et al., 2014) was employed to polish the hybrid assembly. After then, redundancies of the primary assembly were removed by Redundans (Pryszcz and Gabaldón, 2016) with "--identity 0.85 --overlap 0.36."

We performed quality control of Hi-C raw reads and obtained valid Hi-C-connected reads by Juicer (v.1.5; Durand et al., 2016). A 3D *de novo* assembly (3D-DNA, v.180922; Dudchenko, et al., 2017) pipeline (Dudchenko et al., 2017) was applied to anchor primary contigs into chromosome-level scaffolds. Completeness of the genome assembly was evaluated using by BUSCO v3.0

(Simão et al., 2015) with "-l actinopterygii_odb9 -m genome -c 3 -sp zebrafish."

De Novo Assembly of Transcriptomes

We *de novo* assembled the RNA-seq reads using the Trinity assembler (v2.9.0; Haas et al., 2013) and TGI clustering tool (TGICL; Pertea et al., 2003). The PacBio ISO-Seq3 pipeline (https://github.com/PacificBiosciences/IsoSeq) was used to obtain full-length non-chimeric (FLNC) transcripts *via* ccs, classify, cluster, and polish stage. FLNC was aligned with genome by minimap2.

Gene Prediction and Annotation

Repetitive elements in the spiny head croaker genome were identified through a combination of homolog-based and *de novo* approaches. For the homolog-based method, RepeatMasker (v.4.0.7) (Smit et al., 2019) and RepeatProteinMask (v.4.0.7; Smit et al., 2019) were used to detect repeats by aligning against the Repbase database (v 21.0; Bao et al., 2015). For the *de novo* method, LTRharvest (Ellinghaus et al., 2008) was applied to predict full long terminal repeat (LTR) retrotransposons. RepeatModeler (v1.0.11; Smit et al., 2019) was employed to build transposable element (TE) consensus sequences as a *de novo* TE library, and TRF (v.4.09; Benson, 1999) was used to obtain tandem repetitive sequences. RepeatMasker was then used to discover and identify repetitive sequences with the combined library of the *de novo* TEs.

Based on the repeat masked genome, we employed *de novo*, homology-based, and transcriptome-based prediction methods to annotate protein-coding genes in the assembled genome. Protein sequences of zebrafish (*Danio rerio*), three-spined stickleback (*Gasterosteus aculeatus*), Atlantic cod (*Gadus morhua*), channel catfish (*Ictalurus punctatus*), spotted gar (*Lepisosteus oculatus*), Nile tilapia (*Oreochromis niloticus*), fugu (*Takifugu rubripes*), downloaded from Ensembl (Hunt et al., 2018), and large yellow croaker (*L. crocea*, GCF_000972845.2) from NCBI were aligned to the spiny head croaker genome by tBLASTn (Kent, 2002) with "-e 1e-5." Subsequently, GeneWise (Birney et al., 2004) was used to predict gene structures from BLAST hits. Augustus (v3.3.1; Stanke et al., 2006) was performed to predict *de novo* genes. We obtain 3,000 intact gene models generated from the homolog-based method randomly to train the parameters of AUGUSTUS then used AUGUSTUS to perform *de novo* prediction based on the repeat-masked genome with the training parameters. These gene sets that were predicted by different methods were integrated into a nonredundant gene set through the pipelines described in previous research (Xiong et al., 2016). After that, the combined gene set was modified with transcriptome data through PASA (v2.3.3; Haas et al., 2003).

Gene functional annotation was performed based on consensus of sequence and domain. The protein sequences were aligned to NCBI Non-Redundant Protein Sequence (NR) databases, Kyoto Encyclopedia of Genes and Genomes (KEGG v89.0; Kanehisa and Goto, 2000), SwissProt, and TrEMBL (Uniprot release 2020-06) (Boeckmann et al., 2003) by BLASTp with "-e 1e-5." The domains were searched and predicted by using InterProScan v5.11–55.0 (Zdobnov and Apweiler, 2001) (Jones et al., 2014) with publicly available databases including PANTHER (Thomas et al., 2003), Pfam (Bateman et al., 2004), PRINTS (Attwood et al., 2000), ProDom (Servant et al., 2002), PROSITE profiles (Sigrist et al., 2010), and SMART (Letunic et al., 2012). Gene ontology (GO) terms (Ashburner et al., 2000) for each gene were predicted from the InterPro descriptions.

Genome Evolution and Gene Family Analysis

In order to identify gene families in spiny head croaker, we collected protein sequences of the same species used for homologous annotation as well as miiuy croaker (*Miichthys miiuy*) and performed the TreeFam methodology (Li et al., 2006) to obtain gene families of these species. We then used RaxML (Stamatakis, 2006) to construct the phylogenetic tree by using the single copy orthologous gene families with the GTRGAMMA model.

To identify the synteny between spiny head croaker and large yellow croaker, BLASTp was used to calculate pairwise similarities (e value < 1e-5), and MCScanX package with default parameters was then used for classification. Then, JCVI was performed to generate visualization.

A MCMCtree program in PAML (v4.9e; Yang, 2007) was performed to estimate the divergence time between various species in the phylogenetic tree with the REV substitution model. Three calibration time points based on the TimeTree database (http://www.timetree.org) were used as references (*T. rubripes-G. aculeatus*: 99–127 MYA; *G. morhua-T. rubripes:* 141–166 MYA; *L. oculatus-D. rerio*: 295–334 MYA), including spotted gar, zebrafish, Atlantic cod, fugu, and three-spined stickleback.

CAFÉ (v3.0; Han et al., 2013) was used to analyze gene family expansion and contraction under a maximum likelihood framework; single-copy orthologous gene families and estimated divergence time between different species were used as input files. To identify possible positive selected genes (PSGs), we first conducted multiple-sequence alignments based on the protein sequences of single-copy gene families by PRANK (Löytynoja, 2014), then the non-synonymous substitution rate (Ka) and synonymous substitution rate (Ks) were calculated by the codeml in PAML (v4.9e) with the branch-site model (cleandata = 1) and spiny head croaker was chosen as foreground species. Only the results with p-values <0.05 and false discovery rate (FDR) < 0.05 were considered as positive selected genes. Based on whole-genome annotation results and the official classification, we use the phyper in R (v3.5.2) to perform KEGG pathway enrichment analysis.

Phylogenetic Analysis of Otolith Related Genes

We downloaded the protein sequences of Otolin-1a, Otolin-1b, Otopetrin-1, Otoconin-90, OMP-1, SPARC, SPARCL1, and SPARCL2 from zebrafish and fugu and whole-genome

TABLE 1 | Statistics of the genome assembly of spiny head croaker.

Parameter	Contig	Scaffold
Total length (bp)	811,227,110	817,240,112
Total number	12,220	10,973
Gap (bp)	0	6,013,002
N50 (bp)	74,891	26,576,940
N90 (bp)	38,245	23,883
Maximum length (bp)	428,343	33,752,147
Hi-C anchored length (bp)	0	643,229,385
Hi-C anchored rate	0	78.71%
GC content	41.60%	41.30%
Evaluation of BUSCO	94.00%	94.20%

sequences of four fishes (*D. rerio*, *L. crocea*, *T. rubripes*, and *Dicentrarchus labrax*) from NCBI and Ensembl. Nucleotide sequences of these genes were aligned from these four species and spiny head croaker (from the present study) by using BLAST, and filtered with identity, then related protein coding sequences were predicted by Exonerate (v.2.2.0) (Slater and Birney, 2005) or GeneWise firstly. Secondly, we converted coding sequences (CDS) to protein sequences and used PRANK (Löytynoja, 2014) to perform multiple-sequence alignments. RaxML (Stamatakis, 2006) was employed to construct a gene family phylogenetic tree with the PROTGAMMAAUTO model, and the genes from spotted gar as outgroup. We also predicted whether these protein-coding genes in Sciaenidae were involved in positive selection by using the codeml in PAML (v4.9e) with the branch-site model (cleandata = 1) and choosing the branch of spiny head croaker and large yellow croaker as foreground species, and searched the domains of these protein sequences by using NCBI Batch CD-Search and generated visualizations *via* EvolView (Subramanian et al., 2019).

According to the results of multiple-sequence alignments and PSG analysis, we selected those amino acid sites with inconsistency between the family Sciaenidae and other species. Potential functional effects of these residual substitutions were evaluated by PolyPhen-2 (Adzhubei et al., 2013; Li et al., 2018) and PROVEAN (Choi and Chan, 2015).

RESULTS

Genome Sequencing and Assembly

We sequenced the genome of a wild female spiny head croaker by using an Illumina HiSeq sequencing platform as well as a PacBio Sequel sequencing platform. After data filtering, we obtained a total of 49.12-Gb Illumina clean reads by SOAPnuke and 35.24-Gb PacBio long reads, representing approximately 60-fold and 43-fold coverage of the spiny head croaker genome, respectively. An entire genome size of 811.25 Mb was estimated by the routine Kmerfreq method (with K = 17; https://github.com/fanagislab/kmerfreq). Employing a hybrid assembly method, we obtained a redundant assembly of 994.29 Mb and then used Redundans (Kajitani et al., 2014; Ye, et al., 2016) to reduce the redundant sequences. About 18.4% sequences in hybrid assembly were removed. We obtained a draft genome of 811.23 Mb with a contig N50 of 74.92 kb (**Table 1**, **Supplementary Table S1**). The mapping ratio with genome sequencing was 97.89% for the chromosome version (**Supplementary Table S2**).

A total of 185.33 Gb Hi-C data were analyzed by Juicer, and contigs in the draft assembly were subsequently anchored into chromosomes by a 3D-DNA pipelin, resulting in a polished genome assembly of 817.24 Mb, with an improved scaffold N50 of 26.58 Mb (**Supplementary Figure S1**, **Table 1**). The final assembly consists of 24 chromosomes (ranging from 24.92 to 33.75 Mb in length) and covers 643.23 Mb which accounts for 78.71% of the whole genome (**Supplementary Table S3**).

We determined that approximately 94.2% of complete reference genes (82.3% single-copy and 11.9% duplicated) were detectable in the final assembly according to BUSCO values (**Supplementary Table S4**).

Gene Prediction and Annotation

In total, approximately 31.69% of the spiny head croaker assembled sequences were annotated as repetitive elements, which is higher than that for large yellow croaker (Ao et al., 2015; Mu et al., 2018). The repetitive sequences include 161.45 Mb of DNA transposons (~19.90%), 76.60 Mb of long interspersed elements (LINEs, ~9.44%), 73.46 Mb of long terminal repeats (LTRs, ~9.06%), and other TEs (**Table 2**, **Figure 1**).

Using the repeat-masked genome assembly, we predicted a total of 29,509 genes after integration of *de novo*, homology-based, and transcriptome-based annotations (**Table 3**). The total annotated gene number is similar to that in a previous published genome of the spiny head croaker (**Figure 1**, **Supplementary Figure S2**). Based on functional annotation, we predicted 29,432 (~99.74%) protein-coding genes with at least one assignment from Swiss-Prot, TrEMBL, InterProScan, Nr database, and KEGGor GO databases.

Genome Evolution and Gene Family Analysis

To determine the phylogenetic relationship of spiny head croaker with other species, we compared its assembly with other nine representative fish genomes. We identified a total of 19,627 gene families (16,005 in spiny head croaker) and 3,955 single-copy orthologues from TreeFam. After construction of a phylogenetic tree by using the single-copy orthologous gene families, we observed that spiny head croaker is much closer to large yellow croaker (**Figure 2**, **Supplementary Figure S3**), and Cichlidae (such as Nile tilapia) is closely related to Sciaenidae (such as spiny head croaker, large yellow croaker, and miiuy croaker). According to the results of MCMCtree, we estimated that spiny head croaker and large yellow croaker diverged around 13 (5.8~28.6) million years ago (Mya), and their ancestor diverged from tilapia around 81 (67.8~97.4) Mya (**Figure 2**).

Based on the phylogenetic tree and species divergence time analysis, we employed CAFÉ to analyze the expansion and contraction of gene families. A total of 1,028 significantly expanded ($p < 0.05$) and 230 significantly contracted

TABLE 2 | Repetitive elements in the assembled genome of spiny head croaker.

Parameter	RepBase TEs		TE proteins		*De novo*		Combined TEs	
	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome
DNA	54,209,863	6.682452	4,882,493	0.601865	138,693,323	17.09673	161,446,484	19.90152
LINE	24,524,858	3.02318	14,324,110	1.765734	51,652,425	6.367197	76,599,297	9.442399
SINE	4,234,508	0.521988	0	0	2,415,507	0.29776	6,410,232	0.79019
LTR	19,821,582	2.443407	9,298,472	1.146223	54,523,316	6.721091	73,457,448	9.055103
Other	9,297	0.001146	0	0	0	0	9,297	0.001146
Unknown	0	0	0	0	40,564,762	5.00042	40,564,762	5.00042
Total	94,358,167	11.631535	28,439,883	3.505786	219,148,467	27.01444	257,046,838	31.68617

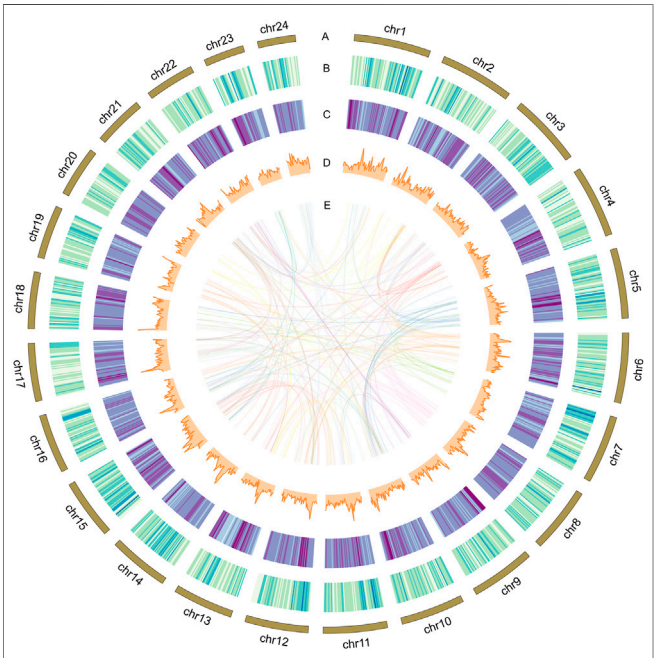

FIGURE 1 | Characterization of the assembled genome for spiny head croaker. From outside to inside: **(A)** chromosomes of spiny head croaker; **(B)** gene density of the genome; **(C)** repeat density of the genome; **(D)** GC content of the genome; **(E)** paralogous genes on different chromosomes. **(B–D)** were drawn in 500-kb sliding windows.

($p < 0.05$) gene families were predicted in spiny head croaker (**Figure 2**). Interestingly, many expanded gene families were enriched in several important KEGG pathways (**Figure 3A**), such as "calcium signaling pathway" (p = 1.30e−39), "circadian entrainment" (p = 1.39e−32), "intestinal immune network for IgA production" (p = 4.40e−47), and "NOD-like receptor signaling pathway" (p = 5.68e−15).

According to the analysis of positive selection with single-copy gene families, 421 positive selected genes (PSGs; $p < 0.05$) were identified in spiny head croaker (**Supplementary Table S5**) by PAML (Yang, 2007). These PSGs were enriched in several interesting KEGG pathways, such as "amino sugar and nucleotide sugar metabolism," "longevity regulating pathway," and "fat digestion and absorption" (**Figure 3B**).

Phylogenetic Analysis of Otolith-Related Genes

We examined several critical otolith-related genes, including otol1a, otol1b, otop1, oc90, omp1, sparc, sparcl1, and sparcl2, to find genetic evidence for the well-developed otoliths in the family Sciaenidae. All sequences were derived from five representative fishes, including zebrafish, fugu, and three Perciformes species (large yellow croaker, spiny head croaker, and European sea bass), and the sequences of zebrafish and fugu (download from NCBI and Ensembl) were used as the queries (**Supplementary Table S6**).

TABLE 3 | Predicted protein-coding genes in the genome of spiny head croaker.

Evidence	Method/species	Numbers	Average gene length (bp)	Average CDS length (bp)	Average exon length (bp)	Average exon per gene	Average intron length (bp)
De novo	AUGUSTUS	40,824	5,815.97	1,092.66	170.20	6.42	871.46
Homolog	*Danio rerio*	26,993	7,176.52	1,404.30	175.12	8.02	822.34
	Gadus morhua	26,917	6,874.98	1,290.37	166.71	7.74	828.53
	Gasterosteus aculeatus	27,805	7,206.74	1,366.09	166.06	8.23	808.25
	Ictalurus punctatus	27,527	7,096.90	1,368.55	175.28	7.81	841.41
	Larimichthys crocea	31,465	7,875.63	1,539.67	173.39	8.88	804.08
	Lepisosteus oculatus	25,501	7,310.53	1,396.47	174.29	8.01	843.40
	Oreochromis niloticus	30,849	7,664.18	1,543.67	180.70	8.54	811.43
	Takifugu rubripes	28,381	7,038.13	1,340.72	172.83	7.76	843.14
Transcriptome	—	16,189	4,348.24	1,295.77	160.29	8.08	2,609.32
Total		29,509	6,425.15	1,336.49	174.59	7.65	764.65

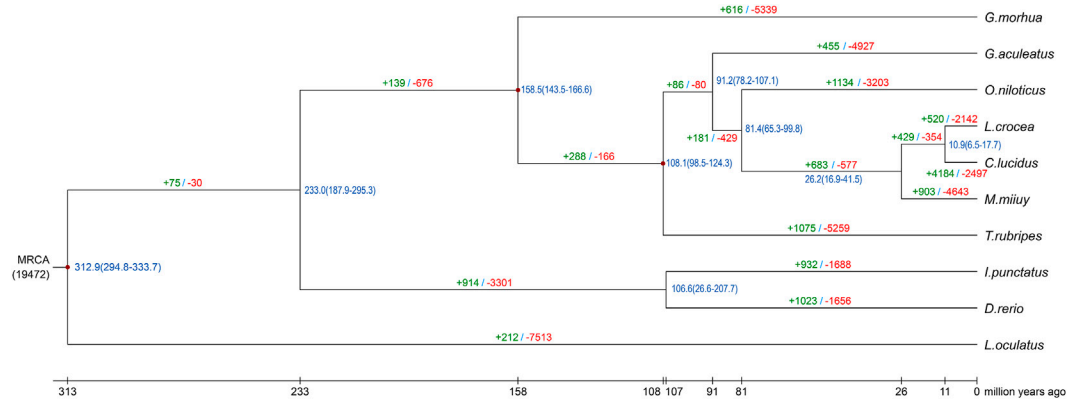

FIGURE 2 | Evolution analysis of spiny head croaker. Green and red numbers on the branches represent the expansion and contraction gene families in each species, blue numbers on the branches show the estimated divergence times in Mya, and dark red points represent the calibration time from TimeTree.

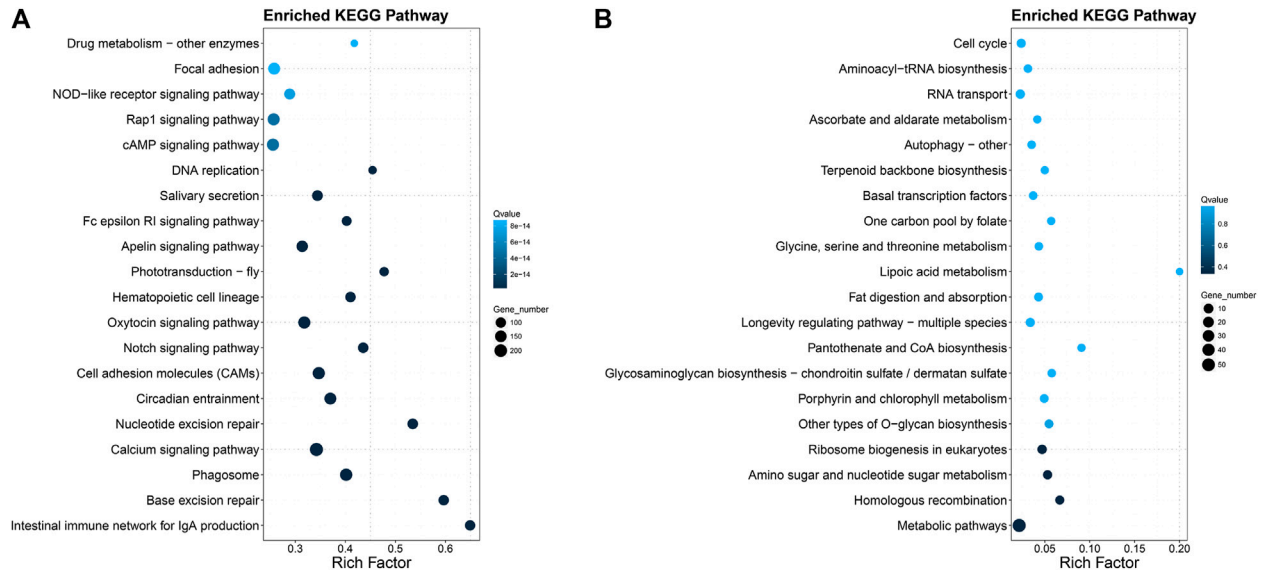

FIGURE 3 | KEGG pathway enrichments in spiny head croaker. **(A)** Pathway enrichments of expanded gene families; **(B)** pathway enrichments of positive selected genes.

Each of these genes was a single copy in these examined species. However, localization and multiple-sequence alignment displayed that the gene previously annotated as otol1 in the family Sciaenidae was more similar to zebrafish otol1b; another gene annotated as inner ear-specific collagen showed a higher sequence similarity to zebrafish otol1a. Domains of otolith-related genes were searched by NCBI Batch CD-Search, and our results proved that the examined domains of these genes were highly conserved in various species (see more details in **Figure 4**). Phylogenetic trees of these otolith-related genes were constructed, and their topological structures were consistent with the species tree. For example, large yellow croaker and spiny head croaker were clustered as sister groups (**Figure 4**), indicating that these genes had a closer relationship in these two croaker species than in other vertebrates.

From the results of multiple-sequence alignment, we found that, in large yellow croaker and spiny head croaker, some nucleotide variances led to amino acid changes in some genes when compared with other fishes (see **Figures 5–8**, **Supplementary Figures S4–S11**). Interestingly, we observed that 16 amino acid residues in SPARC of the family Sciaenidae are different from those in other fishes, although the sequences of the calcium-binding region in SPARC showed high conservation in most species; however, the position 274 of large yellow croaker and spiny head croaker is Asp (D) instead of Gln (Q) or Ser (S) (see more details in **Figure 6**). In Oc90, several residues were changed at PLA2L domains (**Figure 7**).

Based on the best branch-site model, we propose that two critical otolith-related genes (SPARCL1 and OC90) in the family

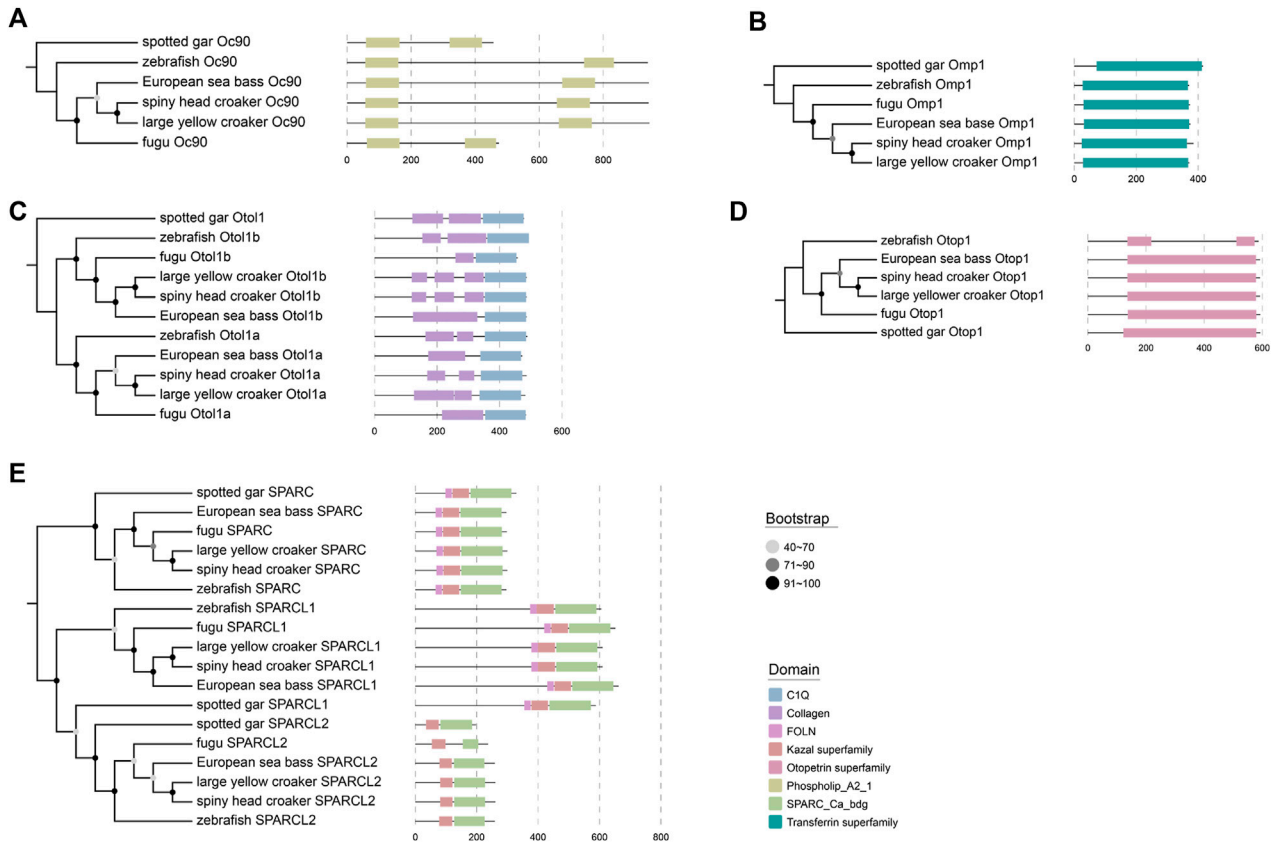

FIGURE 4 | Phylogeny of otolith-related genes among six examined fish species. The detailed phylogenetic topology of Otoconin-90 [Oc90; **(A)**], Otolith matrix protein 1 [Omp-1; **(B)**], Otolin-1a (Otol1a) and Otolin-1b [Otol1b; **(C)**], Otopetrin-1 [Otop1; **(D)**], secreted protein acidic and rich in cysteine (SPARC), SPARC Like 1 (SPARCL1), and SPARC Like 2 [SPARCL2; **(E)**] was constructed individually.

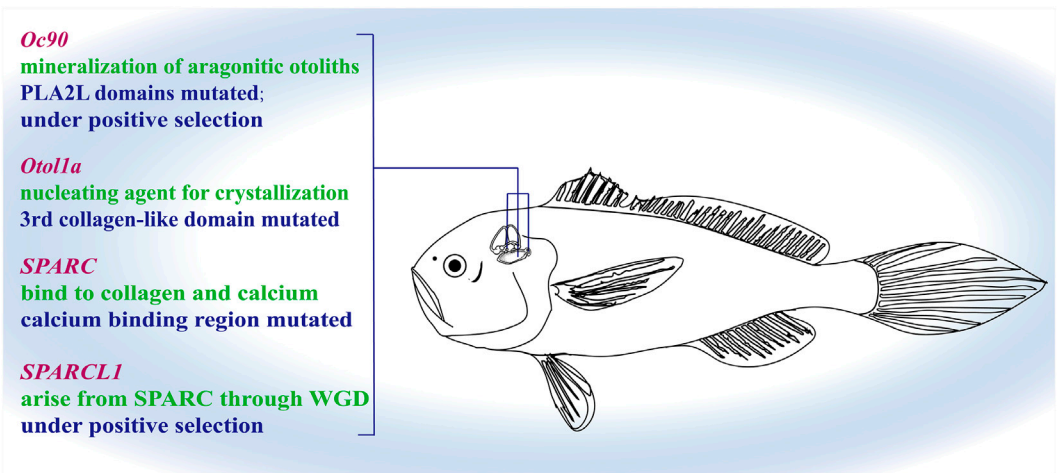

FIGURE 5 | Genetic changes in candidate otolith related genes of the family Sciaenidae. We found that Oc90, Otol1a, and SPARC contained some mutated sites, and Otop1 and SPARCL1 were under positive selection in Sciaenidae.

FIGURE 6 | Multiple alignments of partial SPARC sequences. Red boxes mark the changed amino residues in SPARC from the family Sciaenidae.

Sciaenidae were positively selected (**Supplementary Table S7**). Moreover, based on the multiple alignments, Some of the amino acid substitution sites in these genes were predicted to have a possible effect on the proteins by PolyPhen-2 analysis (Adzhubei et al., 2013) and by PROVEAN analysis (Choi and Chan, 2015), such as position 319 of Otol1a, positions 65 and 204 of SPARC, and the positive selected site in SPARCL1. More importantly, two deletion sites in Otol1a (positions 320 and 321) were predicted to have a possible effect on the protein by PROVEAN (**Figure 8**). Through PAML, we further predicted whether these genes were under positive selection or not.

DISCUSSION

Due to the high heterozygosity rate of some species, heterozygous regions are probably assembled repeatedly, resulting in a redundant genome with a larger size. According to GenomeScope, heterozygosity of spiny head croaker is about 1%. In the present study, we chose a hybrid assembly strategy with a combination of Illumina short reads and PacBio long reads and then employed Redundans (Pryszcz and Gabaldón, 2016) to remove the redundant region so as to obtain a final genome as much as possible to be equal to the estimated haploid genome size. Using this improved version of genome assembly, we reconstructed 24 chromosomes with additional Hi-C reads. The genome size of spiny head croaker (817.24 Mb) is bigger than that of large yellow croaker (708.47 Mb; Chen et al., 2019) and miiuy croaker (636.22 Mb; Xu et al., 2016a). According to our results, especially the annotation of repetitive sequences in this study, we speculate that the higher proportion of repetitive elements may be the major reason for the larger genome of spiny head croaker than its close relatives.

In our present study, a total of 29,509 genes were annotated from the spiny head croaker genome, which are more than those from a previous study (Cai et al., 2019). The number is higher than that of large yellow croaker (23,172 genes; Chen et al., 2019), miiuy croaker (21,960 genes; Xu et al., 2016a), and leopard coral

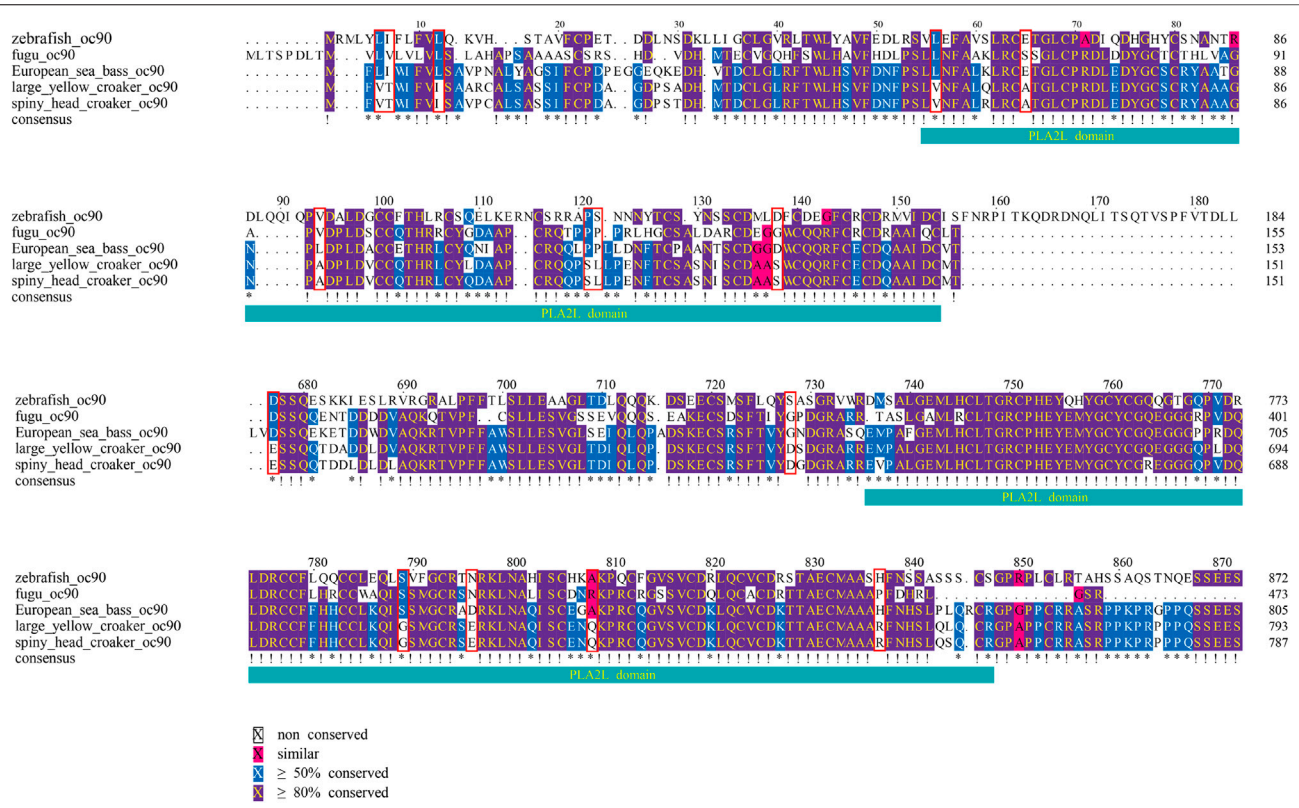

FIGURE 7 | Multiple alignments of partial Oc90 sequences. Red boxes mark the changed amino residues at two PLA2L domains of the Oc90 from the family Sciaenidae.

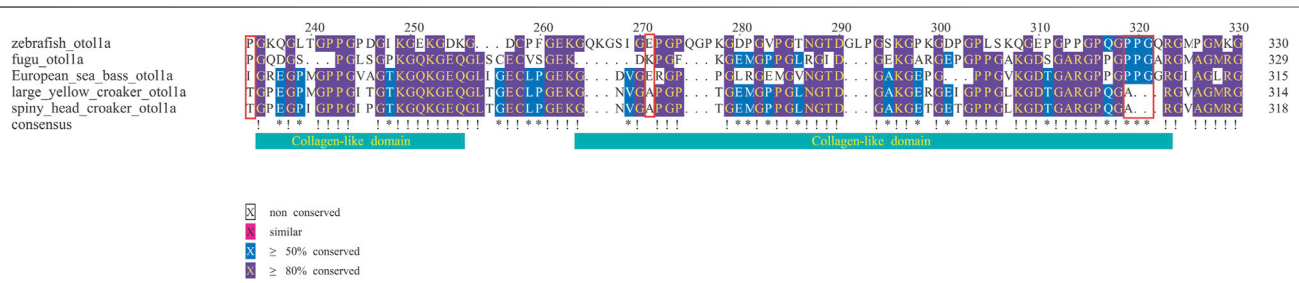

FIGURE 8 | Multiple alignments of partial Otol1a sequences. Red boxes mark the changed amino residues at the collagen-like domain of Otol1a from the family Sciaenidae.

grouper (25,248 genes; Zhou et al., 2020), whereas it is similar to Nile tilapia (29,249 genes; Conte et al., 2017). Moreover, our current work investigated gene families in spiny head croaker based on an integration of PacBio long-read sequencing and Hi-C technology. We identified 3,955 single-copy orthologues and observed that spiny head croaker is much close to large yellow croaker, which is consistent with previous mitochondrial genome studies (Cheng et al., 2012b). Further analysis of gene families showed that some gene families have significantly expanded in the spiny head croaker. Many gene families were enriched in several important pathways, such as "calcium signaling pathway" and "NOD-like receptor signaling pathway," which may be related to some biological characteristics and basic physiological activities of this economically important fish.

Interestingly, in the "calcium signaling pathway," we predicted that the calcium-binding protein (CaBP) gene family and parvalbumin gene family were significantly expanded in the spiny head croaker. Previous studies reported that CaBPs and parvalbumin are early markers of non-mitotic regenerating hair cells in bullfrog vestibular otolith (Steyger et al., 1997). CaBPs, located at the neuroretina, inner ear, and notochord, could modulate calcium levels and distribution, and thereby they were regarded as important regulators of essential

neuronal target proteins (Haynes et al., 2012; Di Donato et al., 2013). Otolith-specific CaBPs were also detected in zebrafish (Söllner et al., 2003) and rainbow trout (Poznar et al., 2017). It was reported that the circadian rhythm of hair cells for secreting these CaBPs is likely to be a vital factor to cause the daily increase of otoliths (Suga and Nakahara, 2012).

Most of the genome-based studies of fishes have focused on growth traits, innate immunity, and/or sex determination (Cai et al., 2019; Zhou et al., 2019). However, no genome study related to the otolith growth and development has been reported yet. One of the main structural proteins in the organic matrix is Otolin-1a, also named as inner ear-specific collagen in some reports, containing calcium-binding sites; its C1q-like domain forms a stable trimer in calcium-containing solutions, suggesting that it participates in the correct arrangement of otolith to the inner ear sensory epithelium and may act as a nucleating agent for crystallization and stabilization of the otolith matrix (Murayama et al., 2005; Hołubowicz et al., 2017). Interestingly, we found that in the third collagen-like domain of Otol1a, the 319th amino acid is substituted by Ala (A), and both positions of 320 and 321 were lost in the family Sciaenidae.

SPARC, a major bone protein with an essential role for fish otolith normal growth and development (Kang et al., 2008), is multifunctional. It is able to bind both collagens and calcium. SPARCL1 and SPARCL2 were derived from SPARC through whole-genome duplication (WGD). When oc90 is absent, both sparc and sparcl1 mRNA levels were significantly upregulated to compensate for the lack of Oc90 and promoted biomineralization of murine otoconia (Xu et al., 2010). In our present study, the sequences of calcium-binding region in SPARC of various fishes showed high conservation in most sites; however, the position 274 of large yellow croaker and spiny head croaker is Asp (D), which might have a higher calcium-binding affinity at a high pH condition (Tang and Skibsted, 2016) and act as a crystal nucleation center whether directly binding with inorganic crystals or interacting with crystal binding proteins; in fact, these two modes are involved in the interaction between bone matrix protein and hydroxyapatite (or apatite; Xu et al., 2010). The PolyPhen-2 and PROVEAN results of amino acid substitution sites 65 and 204 showed possible functional changes of SPARC from the family Sciaenidae. All these data suggest that SPARC in family Sciaenidae plays an important role in calcium and collagen binding capacity, which may relate with formation of well-developed otoliths. While SPARCL1 in the family Sciaenidae was detected as a positive selected gene, whether its function in fish otoliths is similar to that in mice remains unclear. Verification by more studies is required.

Oc90 is a matrix protein of otolith with two PLA2L domains. Although these two domains do not possess enzymatic activity, they contain potential glycosylation sites, retain calcium-binding capacity, and provide a rigid structure for potential CaCO3 deposition (Petko et al., 2008). It is necessary for the early events of otolith biomineralization to play an important role in recruiting other proteins to form the organic matrix (Yang et al., 2011). In the Oc90 of spiny head croaker and large yellow croaker, several sites were substituted at PLA2L domains, which may lead to some changes in the calcium-binding capacity of this protein.

According to the substitutions or deletions of the abovementioned candidate gene sites that were identified at the genomic level, we speculate that these changes may have a relationship with otolith formation, resulting in the interesting status of well-developed otoliths in the family Sciaenidae.

CONCLUSION

In this study, a high-quality chromosome-level genome assembly of spiny head croaker was constructed. Some amino acid substitutions or deletions in several otolith-related genes (such as substitutions in Oc90, Otol1a, SPARC, and deletions in Otol1a) were identified. These changes may be critical for well-developed otoliths in the family Sciaenidae. Our genome resources will provide genetic assistance for in-depth studies on detailed molecular mechanisms of the formation and development of well-developed otoliths in various Sciaenidae species.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available in the CNGB Nucleotide Sequence Archive using accession number CNP0001197.

ETHICS STATEMENT

The animal study was reviewed and approved by All animal study protocols were approved by the Ethics Committee of the Chinese Academy of Fishery Sciences.

AUTHOR CONTRIBUTIONS

WS and XY designed and conceived the study. WG and CZ performed the experiments, wrote the manuscript, and performed the data analyses. XL, CB, and QS revised the manuscript. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by the Basic Research Fund for State-Level Nonprofit Research Institutes of ESCFRI, CAFS (No. Dong2019M02), Key Project of Zhejiang Province of China (2020C02015), Grant Plan for Demonstration City Project for Marine Economic Development in Shenzhen (No. 86), Natural Science Foundation for Fundamental Research in Shenzhen (No. JCYJ20190812105801661), and Shenzhen Science and Technology Program for International Cooperation (No. GJHZ20190819152407214).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.730255/full#supplementary-material

REFERENCES

Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr. Protoc. Hum. Genet.* 76 (1), 7–20. doi:10.1002/0471142905.hg0720s76

Ao, J., Mu, Y., Xiang, L.-X., Fan, D., Feng, M., Zhang, S., et al. (2015). Genome Sequencing of the Perciform Fish Larimichthys Crocea Provides Insights into Molecular and Genetic Mechanisms of Stress Adaptation. *Plos Genet.* 11 (4), e1005118. doi:10.1371/journal.pgen.1005118

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* 25 (1), 25–29. doi:10.1038/75556

Attwood, T. K., Croning, M. D. R., Flower, D. R., Lewis, A. P., Mabey, J. E., Scordis, P., et al. (2000). PRINTS-S: the Database Formerly Known as PRINTS. *Nucleic Acids Res.* 28, 225–227. doi:10.1093/nar/28.1.225

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase Update, a Database of Repetitive Elements in Eukaryotic Genomes. *Mobile Dna* 6 (1), 11. doi:10.1186/s13100-015-0041-9

Barnett, B. K., Chanton, J. P., Ahrens, R., Thornton, L., and Patterson, W. F., 3rd (2020). Life History of Northern Gulf of Mexico Warsaw Grouper Hyporthodus Nigritus Inferred from Otolith Radiocarbon Analysis. *PLoS ONE* 15 (1), e0228254. doi:10.1371/journal.pone.0228254

Barrett, J. H. (2019). An Environmental (Pre)history of European Fishing: Past and Future Archaeological Contributions to Sustainable Fisheries. *J. Fish. Biol.* 94 (6), 1033–1044. doi:10.1111/jfb.13929

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., et al. (2004). The Pfam Protein Families Database. *Nucleic Acids Res.* 32, 138D–141D. doi:10.1093/nar/gkh121

Benson, G. (1999). Tandem Repeats Finder: a Program to Analyze DNA Sequences. *Nucleic Acids Res.* 27 (2), 573–580. doi:10.1093/nar/27.2.573

Bian, C., Huang, Y., Li, J., You, X., Yi, Y., Ge, W., et al. (2019). Divergence, Evolution and Adaptation in ray-finned Fish Genomes. *Sci. China Life Sci.* 62, 1003–1018. doi:10.1007/s11427-018-9499-5

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. *Genome Res.* 14 (5), 988–995. doi:10.1101/gr.1865504

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., et al. (2003). The SWISS-PROT Protein Knowledgebase and its Supplement TrEMBL in 2003. *Nucleic Acids Res.* 31 (1), 365–370. doi:10.1093/nar/gkg095

Cai, M., Zou, Y., Xiao, S., Li, W., Han, Z., Han, F., et al. (2019). Chromosome Assembly of Collichthys Lucidus, a Fish of Sciaenidae with a Multiple Sex Chromosome System. *Sci. Data* 6 (1), 132. doi:10.1038/s41597-019-0139-x

Chen, B., Zhou, Z., Ke, Q., Wu, Y., Bai, H., Pu, F., et al. (2019). The Sequencing and *De Novo* Assembly of the Larimichthys Crocea Genome Using PacBio and Hi-C Technologies. *Sci. Data* 6 (1), 188. doi:10.1038/s41597-019-0194-3

Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., et al. (2017). SOAPnuke: a MapReduce Acceleration-Supported Software for Integrated Quality Control and Preprocessing of High-Throughput Sequencing Data. *Gigascience* 7 (1), gix120. doi:10.1093/gigascience/gix120

Cheng, J., Ma, G.-q., Miao, Z.-q., Shui, B.-n., and Gao, T.-x. (2012a). Complete Mitochondrial Genome Sequence of the Spinyhead Croaker Collichthys Lucidus (Perciformes, Sciaenidae) with Phylogenetic Considerations. *Mol. Biol. Rep.* 39 (4), 4249–4259. doi:10.1007/s11033-011-1211-6

Cheng, J., Ma, G.-q., Song, N., and Gao, T.-x. (2012b). Complete Mitochondrial Genome Sequence of Bighead Croaker Collichthys Niveatus (Perciformes, Sciaenidae): a Mitogenomic Perspective on the Phylogenetic Relationships of Pseudosciaeniae. *Gene* 491 (2), 210–223. doi:10.1016/j.gene.2011.09.020

Choi, Y., and Chan, A. P. (2015). PROVEAN Web Server: a Tool to Predict the Functional Effect of Amino Acid Substitutions and Indels. *Bioinformatics* 31 (16), 2745–2747. doi:10.1093/bioinformatics/btv195

Conte, M. A., Gammerdinger, W. J., Bartie, K. L., Penman, D. J., and Kocher, T. D. (2017). A High Quality Assembly of the Nile Tilapia (Oreochromis niloticus) Genome Reveals the Structure of Two Sex Determination Regions. *BMC genomics* 18 (1), 341. doi:10.1186/s12864-017-3723-5

Di Donato, V., Auer, T. O., Duroure, K., and Del Bene, F. (2013). Characterization of the Calcium Binding Protein Family in Zebrafish. *PLoS One* 8 (1), e53299. doi:10.1371/journal.pone.0053299

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., et al. (2017). De Novo assembly of the Aedes aegypti Genome Using Hi-C Yields Chromosome-Length Scaffolds. *Science* 356 (6333), 92–95. doi:10.1126/science.aal3327

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., et al. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cel Syst.* 3 (1), 95–98. doi:10.1016/j.cels.2016.07.002

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an Efficient and Flexible Software for *De Novo* Detection of LTR Retrotransposons. *BMC bioinformatics* 9 (1), 18. doi:10.1186/1471-2105-9-18

Gui, J. F., Tang, Q. S., Li, Z. J., Liu, J. S., and De Silva, S. S. (2018). *Aquaculture in China: Success Stories and Modern Trends*. John Wiley & Sons. Hoboken, NJ. USA. doi:10.1002/9781119120759

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Jr, Hannick, L. I., et al. (2003). Improving the Arabidopsis Genome Annotation Using Maximal Transcript Alignment Assemblies. *Nucleic Acids Res.* 31 (19), 5654–5666. doi:10.1093/nar/gkg770

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). De Novo transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis. *Nat. Protoc.* 8 (8), 1494–1512. doi:10.1038/nprot.2013.084

Han, M. V., Thomas, G. W. C., Lugo-Martinez, J., and Hahn, M. W. (2013). Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3. *Mol. Biol. Evol.* 30, 1987–1997. doi:10.1093/molbev/mst100

Haynes, L. P., McCue, H. V., and Burgoyne, R. D. (2012). Evolution and Functional Diversity of the Calcium Binding Proteins (CaBPs). *Front. Mol. Neurosci.* 5, 9. doi:10.3389/fnmol.2012.00009

Heimbrand, Y., Limburg, K. E., Hüssy, K., Casini, M., Sjöberg, R., Palmén Bratt, A. M., et al. (2020). Seeking the True Time: Exploring Otolith Chemistry as an Age-determination Tool. *J. Fish. Biol.* 97, 552–565. doi:10.1111/jfb.14422

Houston, R. D., Bean, T. P., Macqueen, D. J., Gundappa, M. K., Jin, Y. H., Jenkins, T. L., et al. (2020). Harnessing Genomics to Fast-Track Genetic Improvement in Aquaculture. *Nat. Rev. Genet.* 21 (7), 389–409. doi:10.1038/s41576-020-0227-y

Hołubowicz, R., Wojtas, M., Taube, M., Kozak, M., Ożyhar, A., and Dobryszycki, P. (2017). Effect of Calcium Ions on Structure and Stability of the C1q-like Domain of Otolin-1 from Human and Zebrafish. *Febs J.* 284 (24), 4278–4297. doi:10.1111/febs.14308

Hu, Y., Zhang, T., Yang, G., Zhao, F., Hou, J. L., Zhang, L. Z., et al. (2015). Assessment of Resource Situation of Collichthys Lucidus in Coastal Waters of the Yangtze Estuary. *Ying Yong Sheng Tai Xue Bao* 26 (9), 2867–2873.

Hunt, S. E., McLaren, W., Gil, L., Thormann, A., Schuilenburg, H., Sheppard, D., et al. (2018). Ensembl Variation Resources). Ensembl Variation Resources. *Database* 2018, bay119. doi:10.1093/database/bay119

Hurle, B., Ignatova, E., Massironi, S. M., Mashimo, T., Rios, X., Thalmann, I., et al. (2003). Non-syndromic Vestibular Disorder with Otoconial Agenesis in Tilted/mergulhador Mice Caused by Mutations in Otopetrin 1. *Hum. Mol. Genet.* 12 (7), 777–789. doi:10.1093/hmg/ddg087

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: Genome-Scale Protein Function Classification. *Bioinform.* 30 (9), 1236–1240. doi:10.1093/bioinformatics/btu031

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., et al. (2014). Efficient *De Novo* Assembly of Highly Heterozygous Genomes from Whole-Genome Shotgun Short Reads. *Genome Res.* 24 (8), 1384–1395. doi:10.1101/gr.170720.113

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28 (1), 27–30. doi:10.1093/nar/28.1.27

Kang, Y.-J., Stevenson, A. K., Yau, P. M., and Kollmar, R. (2008). Sparc Protein Is Required for normal Growth of Zebrafish Otoliths. *J. Assoc. Res. Otolaryngol.* 9 (4), 436–451. doi:10.1007/s10162-008-0137-8

Kent, W. J. (2002). BLAT---The BLAST-like Alignment Tool. *Genome Res.* 12 (4), 656–664. doi:10.1101/gr.22920210.1101/gr.229202

Letunic, I., Doerks, T., and Bork, P. (2012). SMART 7: Recent Updates to the Protein Domain Annotation Resource. *Nucleic Acids Res.* 40, D302–D305. doi:10.1093/nar/gkr931

Li, F., Bian, L., Ge, J., Han, F., Liu, Z., Li, X., et al. (2020). Chromosome-level Genome Assembly of the East Asian Common octopus (Octopus Sinensis)

Using PacBio Sequencing and Hi-C Technology. *Mol. Ecol. Resour.* 20, 1572–1582. doi:10.1111/1755-0998.13216

Li, H., Coghlan, A., Ruan, J., Coin, L. J., Heriche, J.-K., Osmotherly, L., et al. (2006). TreeFam: a Curated Database of Phylogenetic Trees of Animal Gene Families. *Nucleic Acids Res.* 34 (Suppl. l_1), D572–D580. doi:10.1093/nar/gkj118

Li, J.-T., Gao, Y.-D., Xie, L., Deng, C., Shi, P., Guan, M.-L., et al. (2018). Comparative Genomic Investigation of High-Elevation Adaptation in Ectothermic Snakes. *Proc. Natl. Acad. Sci. USA* 115 (33), 8406–8411. doi:10.1073/pnas.1805348115

Long, L., Assaraf, Y. G., Lei, Z.-N., Peng, H., Yang, L., Chen, Z.-S., et al. (2020). Genetic Biomarkers of Drug Resistance: A Compass of Prognosis and Targeted Therapy in Acute Myeloid Leukemia. *Drug Resist. Updates* 52, 100703. doi:10.1016/j.drup.2020.100703

Löytynoja, A. (2014). Phylogeny-aware Alignment with PRANK. *Mult. Seq. alignment Methods*, 1079, 155–170. doi:10.1007/978-1-62703-646-7_10

Mu, Y., Huo, J., Guan, Y., Fan, D., Xiao, X., Wei, J., et al. (2018). An Improved Genome Assembly for Larimichthys Crocea Reveals Hepcidin Gene Expansion with Diversified Regulation and Function. *Commun. Biol.* 1 (1), 1–12. doi:10.1038/s42003-018-0207-3

Murayama, E., Herbomel, P., Kawakami, A., Takeda, H., and Nagasawa, H. (2005). Otolith Matrix Proteins OMP-1 and Otolin-1 Are Necessary for normal Otolith Growth and Their Correct Anchoring onto the Sensory Maculae. *Mech. Dev.* 122 (6), 791–803. doi:10.1016/j.mod.2005.03.002

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., et al. (2003). TIGR Gene Indices Clustering Tools (TGICL): a Software System for Fast Clustering of Large EST Datasets. *Bioinformatics* 19 (5), 651–652. doi:10.1093/bioinformatics/btg034

Petko, J. A., Millimaki, B. B., Canfield, V. A., Riley, B. B., and Levenson, R. (2008). Otoc1: A Novel Otoconin-90 Ortholog Required for Otolith Mineralization in Zebrafish. *Devel Neurobio* 68 (2), 209–222. doi:10.1002/dneu.20587

Poznar, M., Hołubowicz, R., Wojtas, M., Gapiński, J., Banachowicz, E., Patkowski, A., et al. (2017). Structural Properties of the Intrinsically Disordered, Multiple Calcium Ion-Binding Otolith Matrix Macromolecule-64 (OMM-64). *Biochim. Biophys. Acta (Bba) - Proteins Proteomics* 1865 (11 Pt A), 1358–1371. doi:10.1016/j.bbapap.2017.08.019

Pracheil, B. M., George, R., and Chakoumakos, B. C. (2019). Significance of Otolith Calcium Carbonate crystal Structure Diversity to Microchemistry Studies. *Rev. Fish. Biol. Fish.* 29, 569–588. doi:10.1007/s11160-019-09561-3

Pryszcz, L. P., and Gabaldón, T. (2016). Redundans: an Assembly Pipeline for Highly Heterozygous Genomes. *Nucleic Acids Res.* 44 (12), e113. doi:10.1093/nar/gkw294

Schulz-Mirbach, T., Ladich, F., Plath, M., and Hess, M. (2019). Enigmatic Ear Stones: what We Know about the Functional Role and Evolution of Fish Otoliths. *Biol. Rev.* 94 (2), 457–482. doi:10.1111/brv.12463

Servant, F., Bru, C., Carrère, S., Courcelle, E., Gouzy, J., Peyruc, D., et al. (2002). ProDom: Automated Clustering of Homologous Domains. *Brief. Bioinform.* 3, 246–251. doi:10.1093/bib/3.3.246

Sigrist, C. J. A., Cerutti, L., De Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., et al. (2010). PROSITE, a Protein Domain Database for Functional Characterization and Annotation. *Nucleic Acids Res.* 38, D161–D166. doi:10.1093/nar/gkp885

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs. *Bioinformatics* 31 (19), 3210–3212. doi:10.1093/bioinformatics/btv351

Slater, G., and Birney, E. (2005). Automated Generation of Heuristics for Biological Sequence Comparison. *BMC bioinformatics* 6 (1), 31. doi:10.1186/1471-2105-6-31

Smit, A., Hubley, R., and Green, P. (2019). RepeatMasker Open-4.0. 2013-2015, Available at: http://repeatmasker.org/.

Söllner, C., Burghammer, M., Busch-Nentwich, E., Berger, J., Schwarz, H., Riekel, C., et al. (2003). Control of crystal Size and Lattice Formation by Starmaker in Otolith Biomineralization. *Science* 302 (5643), 282–286. doi:10.1126/science.1088443

Song, W., Zhang, Y., Zhang, X., and Gui, J. (2020). De Novo transcriptome Assembly of Four Organs of Collichthys Lucidus and Identification of Genes Involved in Sex Determination and Reproduction. *PLoS ONE* 15 (3), e0230580. doi:10.1371/journal.pone.0230580

Stamatakis, A. (2006). RAxML-VI-HPC: Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. *Bioinformatics* 22 (21), 2688–2690. doi:10.1093/bioinformatics/btl446

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: Ab Initio Prediction of Alternative Transcripts. *Nucleic Acids Res.* 34 (Suppl. l_2), W435–W439. doi:10.1093/nar/gkl200

Steyger, P. S., Burton, M., Hawkins, J. R., Schuff, N. R., and Baird, R. A. (1997). Calbindin and Parvalbumin Are Early Markers of Non-mitotically Regenerating Hair Cells in the Bullfrog Vestibular Otolith Organs. *Int. J. Dev. Neurosci.* 15 (4-5), 417–432. doi:10.1016/s0736-5748(96)00101-3

Subramanian, B., Gao, S., Lercher, M. J., Hu, S., and Chen, W.-H. (2019). Evolview V3: a Webserver for Visualization, Annotation, and Management of Phylogenetic Trees. *Nucleic Acids Res.* 47 (W1), W270–W275. doi:10.1093/nar/gkz357

Suga, S., and Nakahara, H. (2012). *Mechanisms and Phylogeny of Mineralization in Biological Systems: Biomineralization'90.* Springer Science & Business , Media, Berlin, Germany.

Tang, N., and Skibsted, L. H. (2016). Calcium Binding to Amino Acids and Small glycine Peptides in Aqueous Solution: toward Peptide Design for Better Calcium Bioavailability. *J. Agric. Food Chem.* 64 (21), 4376–4389. doi:10.1021/acs.jafc.6b01534

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., et al. (2003). PANTHER: a Library of Protein Families and Subfamilies Indexed by Function. *Genome Res.* 13, 2129–2141. doi:10.1101/gr.772403

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). GenomeScope: Fast Reference-free Genome Profiling from Short Reads. *Bioinformatics* 33 (14), 2202–2204. doi:10.1093/bioinformatics/btx153

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PloS one* 9 (11), e112963. doi:10.1371/journal.pone.0112963

Wu, C., Zhang, D., Kan, M., Lv, Z., Zhu, A., Su, Y., et al. (2014). The Draft Genome of the Large Yellow Croaker Reveals Well-Developed Innate Immunity. *Nat. Commun.* 5 (1), 1–7. doi:10.1038/ncomms6227

Xiong, Z., Li, F., Li, Q., Zhou, L., Gamble, T., Zheng, J., et al. (2016). Draft Genome of the Leopard Gecko, Eublepharis macularius. *GigaSci* 5 (1), 47. doi:10.1186/s13742-016-0151-4

Xu, T., Xu, G., Che, R., Wang, R., Wang, Y., Li, J., et al. (2016a). The Genome of the Miiuy Croaker Reveals Well-Developed Innate Immune and Sensory Systems. *Sci. Rep.* 6, 21902. doi:10.1038/srep21902

Xu, Y., Zhang, H., Yang, H., Zhao, X., Lovas, S., and Lundberg, Y. Y. W. (2010). Expression, Functional, and Structural Analysis of Proteins Critical for Otoconia Development. *Dev. Dyn.* 239 (10), 2659–2673. doi:10.1002/dvdy.22405

Xu, Y., Zhang, Y., and Lundberg, Y. W. (2016b). Spatiotemporal Differences in Otoconial Gene Expression. *Genesis* 54 (12), 613–625. doi:10.1002/dvg.22990

Yang, H., Zhao, X., Xu, Y., Wang, L., He, Q., and Lundberg, Y. W. (2011). Matrix Recruitment and Calcium Sequestration for Spatial Specific Otoconia Development. *PLoS ONE* 6 (5), e20498. doi:10.1371/journal.pone.0020498

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* 24 (8), 1586–1591. doi:10.1093/molbev/msm088

Ye, C., Hill, C. M., Wu, S., Ruan, J., and Ma, Z. (2016). DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. *Sci. Rep.* 6, 31900. doi:10.1038/srep31900

You, X., Shan, X., and Shi, Q. (2020). Research Advances in the Genomics and Applications for Molecular Breeding of Aquaculture Animals, *Aquaculture* 526, 735357. doi:10.1016/j.aquaculture.2020.735357

Zdobnov, E. M., and Apweiler, R. (2001). InterProScan - an Integration Platform for the Signature-Recognition Methods in InterPro. *Bioinformatics* 17 (9), 847–848. doi:10.1093/bioinformatics/17.9.847

Zhang, X., Zhou, L., and Gui, J. (2019). Biotechnological Innovation in Genetic Breeding and Sustainable green Development in Chinese Aquaculture. *Sci. Sin.-Vitae* 49 (11), 1409–1429. doi:10.1360/SSV-2019-0142

Zhou, Q., Gao, H., Zhang, Y., Fan, G., Xu, H., Zhai, J., et al. (2019). A Chromosome-level Genome Assembly of the Giant Grouper (Epinephelus Lanceolatus)

Provides Insights into its Innate Immunity and Rapid Growth. *Mol. Ecol. Resour.* 19 (5), 1322–1332. doi:10.1111/1755-0998.13048

Zhou, Q., Guo, X., Huang, Y., Gao, H., Xu, H., Liu, S., et al. (2020). De Novo sequencing and Chromosomal-scale Genome Assembly of Leopard Coral Grouper, Plectropomus leopardus. *Mol. Ecol. Resour.* 20, 1403–1413. doi:10.1111/1755-0998.13207

Conflict of Interest: Authors CZ, XL, CB, QS, and XY were employed by the company BGI.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

DATA REPORT
published: 05 October 2021
doi: 10.3389/fgene.2021.736500

Whole-Genome Sequencing of *Sinocyclocheilus maitianheensis* Reveals Phylogenetic Evolution and Immunological Variances in Various *Sinocyclocheilus* Fishes

Ruihan Li[1,2†], Xiaoai Wang[3†], Chao Bian[1,2†], Zijian Gao[1,2], Yuanwei Zhang[3], Wansheng Jiang[4], Mo Wang[5], Xinxin You[1,2], Le Cheng[6], Xiaofu Pan[3], Junxing Yang[3] and Qiong Shi[1,2]**

[1]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China, [2]Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, BGI Academy of Marine Sciences, BGI Marine, BGI, Shenzhen, China, [3]State Key Laboratory of Genetic Resources and Evolution, The Innovative Academy of Seed Design, Yunnan Key Laboratory of Plateau Fish Breeding, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China, [4]Hunan Engineering Laboratory for Chinese Giant Salamander's Resource Protection and Comprehensive Utilization, and Key Laboratory of Hunan Forest and Chemical Industry Engineering, Jishou University, Zhangjiajie, China, [5]Key Laboratory for Conserving Wildlife with Small Populations in Yunnan, Faculty of Biodiversity Conservation, Southwest Forestry University, Kunming, China, [6]BGI-Yunnan, Kunming, China

OPEN ACCESS

Edited by:
Roger Huerlimann,
Okinawa Institute of Science and Technology Graduate University, Japan

Reviewed by:
Peng Xu,
Xiamen University, China
Robert Lehmann,
King Abdullah University of Science and Technology, Saudi Arabia

***Correspondence:**
Qiong Shi
shiqiong@genomics.cn
Junxing Yang
yangjx@mail.kiz.ac.cn

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: 05 July 2021
Accepted: 06 September 2021
Published: 05 October 2021

Citation:
Li R, Wang X, Bian C, Gao Z, Zhang Y, Jiang W, Wang M, You X, Cheng L, Pan X, Yang J and Shi Q (2021) Whole-Genome Sequencing of Sinocyclocheilus maitianheensis Reveals Phylogenetic Evolution and Immunological Variances in Various Sinocyclocheilus Fishes. Front. Genet. 12:736500. doi: 10.3389/fgene.2021.736500

An adult *Sinocyclocheilus maitianheensis*, a surface-dwelling golden-line barbel fish, was collected from Maitian river (Kunming City, Yunnan Province, China) for whole-genome sequencing, assembly, and annotation. We obtained a genome assembly of 1.7 Gb with a scaffold N50 of 1.4 Mb and a contig N50 of 24.7 kb. A total of 39,977 protein-coding genes were annotated. Based on a comparative phylogenetic analysis of five *Sinocyclocheilus* species and other five representative vertebrates with published genome sequences, we found that *S. maitianheensis* is close to *Sinocyclocheilus anophthalmus* (a cave-restricted species with similar locality). Moreover, the assembled genomes of *S. maitianheensis* and other four *Sinocyclocheilus* counterparts were used for a fourfold degenerative third-codon transversion (4dTv) analysis. The recent whole-genome duplication (WGD) event was therefore estimated to occur about 18.1 million years ago. Our results also revealed a decreased tendency of copy number in many important genes related to immunity and apoptosis in cave-restricted *Sinocyclocheilus* species. In summary, we report the first genome assembly of *S. maitianheensis*, which provides a valuable genetic resource for comparative studies on cavefish biology, species protection, and practical aquaculture of this potentially economical fish.

Keywords: *Sinocyclocheilus maitianheensis*, whole genome sequencing, assembly, annotation, phylogeny, immunity, cave adaptation

Abbreviations: *ask1*, apoptosis signal-regulating kinase 1, also known as mitogen-activated protein kinase 5 (*map3k5*); *bcl-2*, B-cell lymphoma 2; *bcl2l1*, *bcl-2*-like protein 1; *cd40*, tumor necrosis factor receptor superfamily member 5; *cd401*, tumor necrosis factor ligand superfamily member 5; *daxx*, death domain-associated protein 6; *fadd*, fas-associated protein with death domain; *fas*, tumor necrosis factor receptor superfamily member 6; *fasl*, tumor necrosis factor ligand superfamily member 6; *mkk4*, mitogen-activated protein kinase kinase 4, also named *map2k4*; *mkk6*, mitogen-activated protein kinase kinase 6, also named *map2k6*; *tab1*, mitogen-activated protein kinase 7-interacting protein 1; *tak1*, mitogen-activated protein kinase kinase kinase 7, also named *map3k7*; *tnfa*, Tumor necrosis factor ligand superfamily member 1, also named *tnf*.

INTRODUCTION

Cave-restricted animals live in dark subterranean environments. They have evolved over time to adapt to the cave environments through various trait changes in morphology, behavior, and physiology (Jeffery, 2001). Cavefishes have degraded eyes, less body pigments, lower immune activities, and decrease in circadian rhythms in comparison to surface-dwelling fishes (Jeffery, 2009; Qiu et al., 2016; Yang et al., 2016; Krishnan and Rohner, 2017). As a compensation, the nonvisual sensory system of cavefishes is usually enhanced, such as development of extra taste buds and increased vibration attraction behavior (Yoshizawa et al., 2010; Yang et al., 2016). These facts of cavefishes are extremely interesting and worth to be explored with more investigations.

Previously, we have proved that cavefishes have fewer copies of major histocompatibility complex–related gene families (genes of cell surface proteins essential for acquired immune system) than surface-dwelling and semi–cave-dwelling counterparts, possibly suggesting relatively lower immune activities in cavefishes (Qiu et al., 2016), which may be a specific strategy for cave adaptation. Cave-restricted Mexican tetra (*Astyanax mexicanus*) also shows a big increase in appetite, but its fatty liver did not affect this fish's health (Aspiras et al., 2015), implying that there may be some other immune-related molecular mechanisms for fighting inflammation in cavefishes (Xiong et al., 2019). A recent study (Peuß et al., 2020) proposed that organisms in various environments have developed differential immune strategies with innate immune degradation and T-cell overexpression in cavefishes. However, many of these putative molecular mechanisms are not fully understood.

S. maitianheensis lives originally in the surface of Maitian river in Kunming City, Yunnan Province, China (**Supplementary Figure S1**). Some *Sinocyclocheilus* fishes are also residents in caves. *S. maitianheensis* can therefore be used as a control for comparative studies on cave adaptation. Meanwhile, as a genus of state second-class protection in China, *Sinocyclocheilus* has been propagated with a series of artificial breeding for protection from extinction (Yin et al., 2021). Various *Sinocyclocheilus* species likely shared tetraploid origin and have 96 chromosomes that are twice of most teleosts (Heng et al., 2002). The genome diversity makes fishes in this genus as good models for studying cave adaptation and phylogenetic evolution. Although there are some reports on morphological and mitochondrial genomic evolution of various *Sinocyclocheilus* species (Wu et al., 2010; He et al., 2012; Chen Y.-Y. et al., 2018), whole genome–based comparative studies are rare, except for three representative *Sinocyclocheilus* fishes that we published before (Yang et al., 2016).

Here, we performed whole-genome sequencing, assembly, and annotation of *S. maitianheensis* and subsequently conducted comparative genomic analysis and immune-gene inquiry with four other *Sinocyclocheilus* counterparts (including surface-dwelling *Sinocyclocheilus grahami,* semi–cave-dwelling *Sinocyclocheilus rhinocerous*, and cave-restricted *Sinocyclocheilus anophthalmus* and *Sinocyclocheilus anshuiensis*; their genome sequences are publicly available). Our main purpose is to provide a genetic resource for in-depth studies on cave adaption and cavefish biology. Our study can also contribute to the species protection and exploitation of potentially economical value for *S. maitianheensis*.

MATERIALS AND METHODS

Sampling, Library Constructing, and Genome Sequencing

An adult *S. maitianheensis* was collected from Maitian river in Kunming City, Yunnan Province, China, for genome and transcriptome sequencing. Genomic DNAs were extracted from muscle sample. Seven Illumina paired-end sequencing libraries (with insert sizes of 270 bp, 500 bp, 800 bp, 2 kb, 5 kb, 10 kb, and 20 kb, respectively) were constructed for a routine shotgun whole-genome sequencing in an Illumina HiSeq 2,500 platform (San Diego, CA, United States). SOAPfilter v2.2 (Li R. et al., 2009) (-z -p -g 1 -f -o clean -M 2 -f 0) was used to filter reads. Duplicate reads from polymerase chain reactions, those reads with 10 or more nonsequencing bases (Ns), adapter sequences, and bases with low quality were removed.

Total RNAs were extracted from muscle, skin, eye, liver, heart, and brain for construction of individual cDNA library. cDNAs were then sequenced in an Illumina HiSeq X platform and filtered by SOAPnuke v1.0 (Chen Y. et al., 2018) with optimized parameters [-l 10 (default: 5) -q 0.2 (default: 0.5) -n 0.05 -c 0 -Q 2 (default: 1)]. Reads with nonsequenced (N) base ratio of more than 5% or low-quality base (base quality ≤10) ratio of greater than 20% were discarded to generate a new set of higher-quality reads for subsequent transcriptome-based annotation.

Genome Survey, *de Novo* Assembly, and Assessment

Genome size was estimated via the routine 17-mer frequency distribution analysis with the following formula: genome size = K_{num}/K_{depth}, where k_{num} is the number of k-mers obtained from reads, and K_{depth} is the expected depth of k-mer at a maximum frequency (Song et al., 2016). Two Illumina short-insert libraries (500 and 800 bp) were used for this 17-mer analysis.

The genome assembly strategy includes three steps. First, SOAPdenovo2 v2.04.4 (Luo et al., 2012) was applied to produce primary and final scaffolds with the following parameters: pregraph -K 27 -p 16 -d 1; contig -M 1; scaff -F -b 1.5 -p 16. Contigs and primary scaffolds were generated by using filtered reads from short-insert libraries (200, 500, and 800 bp), and the final scaffolds were constructed by mapping long-insert libraries (2, 5, 10, and 20 kb) onto the primary scaffolds. Second, gaps in scaffolds were then filled in two rounds using paired-end reads from the three short-insert libraries (270, 500, and 800 bp) via GapCloser v1.12 and v1.10 (Li R. et al., 2009) (-t 8 -l 150 and -t 25 -p 25, respectively). Finally, SSPACE V2.0 (Boetzer and Pirovano, 2014) (-k 5 -T 25 -g 2) was used to further extend and fill up both contigs and scaffolds. Completeness assessment of the final genome assembly was

performed by BUSCO v5.2.2 (Simão et al., 2015; Manni et al., 2021) (e-value ≤1e-3) with the popular actinopterygii_odb10 database.

Repeat Annotation

Repeat sequence annotation is composed of three routine methods, including *de novo* annotation, homology prediction, and tandem repeat prediction. First, we used RepeatModeller v1.04 (Chen, 2004) and LTR-FINDER v1.0.6 (Xu and Wang, 2007) to construct a local *de novo* repeat reference, and then our assembled genome was aligned to this reference library by RepeatMasker v4.06 (Chen, 2004). In addition, RepeatMasker v4.06 (Chen, 2004) and RepeatProteinMask v4.06 (Chen, 2004) were applied for homology prediction after identification of transposable elements based on RepBase (Jurka et al., 2005). Moreover, Tandem Repeat Finder v4.09 (Benson, 1999) was separately used to predict comprehensive tandem repeats in our pipeline as previously reported (Liu et al., 2019; Zhao et al., 2021). Finally, these results from the aforementioned three methods are integrated by our in-house perl scripts (https://github.com/liruihanguo/Repeats_integration). These scripts separately classified each type of repeat, integrated all repeats, and then removed those overlaps to obtain a nonredundant repeat set.

Gene Structure and Function Annotations

Two different methods were used for gene annotation to generate a total gene set, including homology annotation and transcriptome-based annotation (Bian et al., 2019). For the homology annotation, we downloaded protein sequences of four representative vertebrates from NCBI (Benson et al., 2006), including zebrafish (*Danio rerio*), Japanese medaka (*Oryzias latipes*), and two *Sinocyclocheilus* fishes (*S. anshuiensis* and *S. rhinocerous*), to align them against our *S. maitianheensis* genome assembly by TBLASTn (e-value ≤1e-5) (Gertz et al., 2006). Each gene structure was predicted by GeneWise v2.4.2 (Birney et al., 2004). For the transcriptome-based annotation, Tophat v2.0.13 (Trapnell et al., 2009) was utilized to obtain potential genes by mapping transcriptome reads onto our assembled genome. Subsequently, Cufflink v2.2.1 (Trapnell et al., 2013) was applied to predict the structures of potential genes on the alignments sorted by samtools v1.1 (Li H. et al., 2009). Lastly, the final consensus gene set was integrated by MAKER v2.31.8 (Cantarel et al., 2008).

All these predicted genes were aligned onto several public databases, including Interpro (Hunter et al., 2009), KEGG (Kanehisa et al., 2017), TrEMBL (Boeckmann et al., 2003), and Swiss-Prot (Boeckmann et al., 2003), using BLASTp (McGinnis and Madden, 2004) (e-value ≤1e-5) to perform function annotation. These results were then assessed by comparing coding sequence (CDS) length, intron length, gene length, exon length, and exon number distributions with the four closely related *Sinocyclocheilus* species, including *S. anshuiensis* (Yang et al., 2016) (SAMN03320099. WGS_v1.1 in NCBI), *S. rhinocerous* (Yang et al., 2016) (SAMN03320098_v1.1), *S. grahami* (Yang et al., 2016) (SAMN03320097. WGS_v1.1), and *S. anophthalmus* [genome assembly was deposited at NCBI under accession no. PRJNA669129 (GCA_018155175.1)]. This unpublished genome of *S. anophthalmus* was sequenced by us on both Illumina Hiseq2500 and PacBio Sequel platforms using muscle genomic DNAs, and the final assembly of 1.9 Gb (with a contig N50 of 229.8 kb, a scaffold N50 of 309.9 kb, and prediction of 49,865 protein-coding genes) was assembled by combining the corrected long PacBio reads and the primary assembly from short Illumina reads by DBG2OLC v1.1 (Ye et al., 2016). We also assessed the completeness of these protein-coding gene sets by BUSCO v5.2.2 (Manni et al., 2021).

Orthogroup Cluster

The protein sequences of *S. maitianheensis* and other ten representative species were used for clustering orthogroups and phylogenetic analyses. These vertebrates include four *Sinocyclocheilus* species (*S. anophthalmus*, *S. grahami*, *S. anshuiensis*, and *S. rhinocerous*), common carp (*Cyprinus carpio*, GCF_000951,615.1 in NCBI), zebrafish (GCF_000002035.6), Japanese medaka (GCF_002234675.1), and Asian arowana (*Scleropages formosus*, GCF_900964775.1), as well as the outgroup of human (*Homo sapiens*, GCF_000001405.39) and mouse (*Mus musculus*, GCF_000001635.27). We used BLASTp (McGinnis and Madden, 2004) (e-value ≤1e-5) to align these protein sequences with each other and OrthoMCL v2.0.92 (Fischer et al., 2011) with default parameters to identify orthologous genes and construct orthogroups.

Phylogenetic and Divergence Time Analyses

Single-copy orthogroups were aligned using MUSCLE v3.8.31 (Edgar, 2004). Subsequently, conserved regions were obtained by Gblocks (Castresana, 2000), and the CDS regions of all single-copy genes from each species were connected to form a supergene for extraction of the 4d sites. We also constructed a phylogenetic tree by using PhyML v3.0 with the maximum likelihood method (Guindon et al., 2010). MCMCtree in the PAML package (Yang and Rannala, 2006) was used to estimate the divergence time of five *Sinocyclocheilus* fishes and other species by three calibration time points of fossil records (Benton and Donoghue, 2007), including 61.5–100.5 Mya for *H. sapiens* and *M. musculus*, 159.9–165.2 Mya for *D. rerio* and *O. latipes*, and 416.1–421.8 Mya for *D. rerio* and *H. sapiens*.

4dTv Analysis to Identify Specific Whole-Genome Duplication in *Sinocyclocheilus* Fishes

To estimate the *Sinocyclocheilus* specific whole-genome duplication (WGD) event, we performed a fourfold degenerative third-codon transversion (4dTv) analysis by comparing five *Sinocyclocheilus* genomes with zebrafish and common carp genome assemblies. The WGD periods were calculated by using the following formula [The recognized

time of 3R WGD (~320 Mya)/the 4dTv peak values of 3R WGD in *Sinocyclocheilus* (0.65–0.75)] * the 4dTv peak values of lineage-specific WGD of *Sinocyclocheilus* (0.04–0.05).

An all-to-all alignments of protein sequences from these seven genomes were applied by using BLASTp (McGinnis and Madden, 2004) with an e-value of 1e-5. Syntenic blocks between species were identified using i-ADHoRe 3.0 (Proost et al., 2012) with default parameters, and then homologous proteins were obtained. Subsequently, homologous pairs were aligned using MUSCLE (Edgar, 2004), after we retrieved these homologous protein sequences and converted them to nucleotide sequences. Lastly, we calculated and corrected the 4dTv values for each gene pair by using the HKY model in PAML package (Yang and Rannala, 2006).

Identification of Immune Genes and Pseudogenes in P38 and Mitochondrial Pathways

Fifteen apoptosis-related genes were identified in five *Sinocyclocheilus* genomes and other five representative vertebrate genomes with high quality to investigate the gene copy number in P38 and mitochondrial pathway. These other five vertebrates include common carp, zebrafish, Japanese medaka, Asian arowana, and Mexican tetra (*A. mexicanus*, GCF_000372,685.1 in NCBI). Each genome sequence was used to construct a standard aligned database in the first place. Protein sequences of *tak1*, *tab1*, *ask1*, *fas*, *fasl*, *fadd*, *tnfa*, *cd40*, *cd40l*, *daxx*, *mkk4a*, *mkk4b*, *mkk6*, *bcl-2a*, and *bcl2l1* of zebrafish were downloaded from public uniport databases (**Supplementary Table S1**) as the queries. These protein sequences were then aligned onto above 10 genomes by TBLASTn (e-value ≤1e-5). The alignments with aligned ratio of less than 0.5, sequences similarity of less than 50%, and redundant data were filtered out to obtain the final hit alignments. Subsequently, target apoptosis-related genes were predicted by GeneWise v2.4.1 (Birney et al., 2004) from these 10 vertebrate genomes.

We performed a multiple-sequence alignment using the Muscle module in MEGA v7.0 (Kumar et al., 2016) to identify pseudogenes in each species indicated above. The whole open reading frames were performed with codon-based alignment to identify potential pseudogenes with irregular shifts of premature stop codon(s), codon frameshifts, or missing exon regions.

RESULTS AND DISCUSSION

Summary of the Genome Assembly and Assessment

We generated a total of 236.7-Gb raw reads, among them approximately 179.5 Gb of clean reads were obtained after removal of low-quality data. For the transcriptome sequencing, a total of 39.3-Gb raw reads were generated. The genome size of *S. maitianheensis* was estimated to be about 1.8 Gb by the routine 17-mer frequency distribution analysis, because the k_{num} is 40,525,178,512, and the K_{depth} is 23 (see **Supplementary Figure S2**).

After *de novo* assembly and gap closing, the final genome assembly was 1.7 Gb in total length, with a scaffold N50 of 1.4 Mb, a contig N50 of 24.7 kb, and the GC content of 37.6% (**Tables 1** and **Supplementary Table S2**). The final assembled genome accounted for 94.4% of the estimated genome size (1.8 Gb). For assessment of our genome assembly, we searched a total of 3,640 BUSCO (Benchmarking Universal Single-Copy Orthologs) groups and determined that 3,536 (97.1%) were complete (with 1,643 single-copy BUSCOs and 1,893 duplicated BUSCOs), suggesting a high completeness of our genome assembly for *S. maitianheensis*.

TABLE 1 | Statistics of the genome assembly for *S. maitianheensis*.

Genome assembly	Parameter
Contig N50 (kb)	24.7
Contig number (>100 bp)	212,978
Scaffold N50 (Mb)	1.4
Scaffold number (>100 bp)	106,737
Total length (Gb)	1.7
Genome coverage (×)	139.2
The longest scaffold (Mb)	12.5
Genome annotation	**Parameter**
Protein-coding gene number	39,977
Mean transcript length (bp)	15,881
Mean exons per gene	8.6
Mean exon length (bp)	188.3
Mean intron length (bp)	1,728

We compared the genome assembly and annotation results for the five examined *Sinocyclocheilus* species and observed that their genome sizes are at a narrow range of 1.7–1.9 Gb (**Supplementary Table S3**). The largest genome is 1.9 Gb for *S. anophthalmus* (GCA_018155175.1 in NCBI) whose assembly was generated with both Illumina (next-generation) and PacBio (third-generation) sequencing reads, whereas others were sequenced only by an Illumina platform with relatively lower scaffold N50 values. GC contents of the five examined *Sinocyclocheilus* species ranged from 37.2 to 38.6% (**Supplementary Table S3**).

Genome Annotation

For repeat annotation, we predicted 664,837,705-bp repeat elements (accounting for 39.4% of the genome assembly; **Supplementary Table S4**). Among them, 455.9 Mb of DNA repeat elements, 135.2 Mb of long terminal repeats (LTRs), 104.9 Mb of long interspersed nuclear elements (LINE), and 9.1 Mb of short interspersed nuclear elements (SINE) were detected (**Supplementary Table S5**). After homology and transcriptome-based annotations, our MAKER results integrated a total of 39,977 protein-coding genes with an average length of 15.9 kb (**Supplementary Table S6**).

For function annotation, 38,677 genes were aligned onto four public databases (Interpro, KEGG, TrEMBL, and Swiss-Prot) with function assignments, which account for 96.8% of the annotated genes (**Supplementary Table S7**). The distributions of CDS, intron, gene, and exon length in *S. maitianheensis* were generally similar to those in other four examined *Sinocyclocheilus* fishes (**Supplementary Figure S3**), which suggested a good reliability of the genome annotation for *S. maitianheensis*. However, there are still some distinctions among them, such as the higher percentages of CDS and genes at length

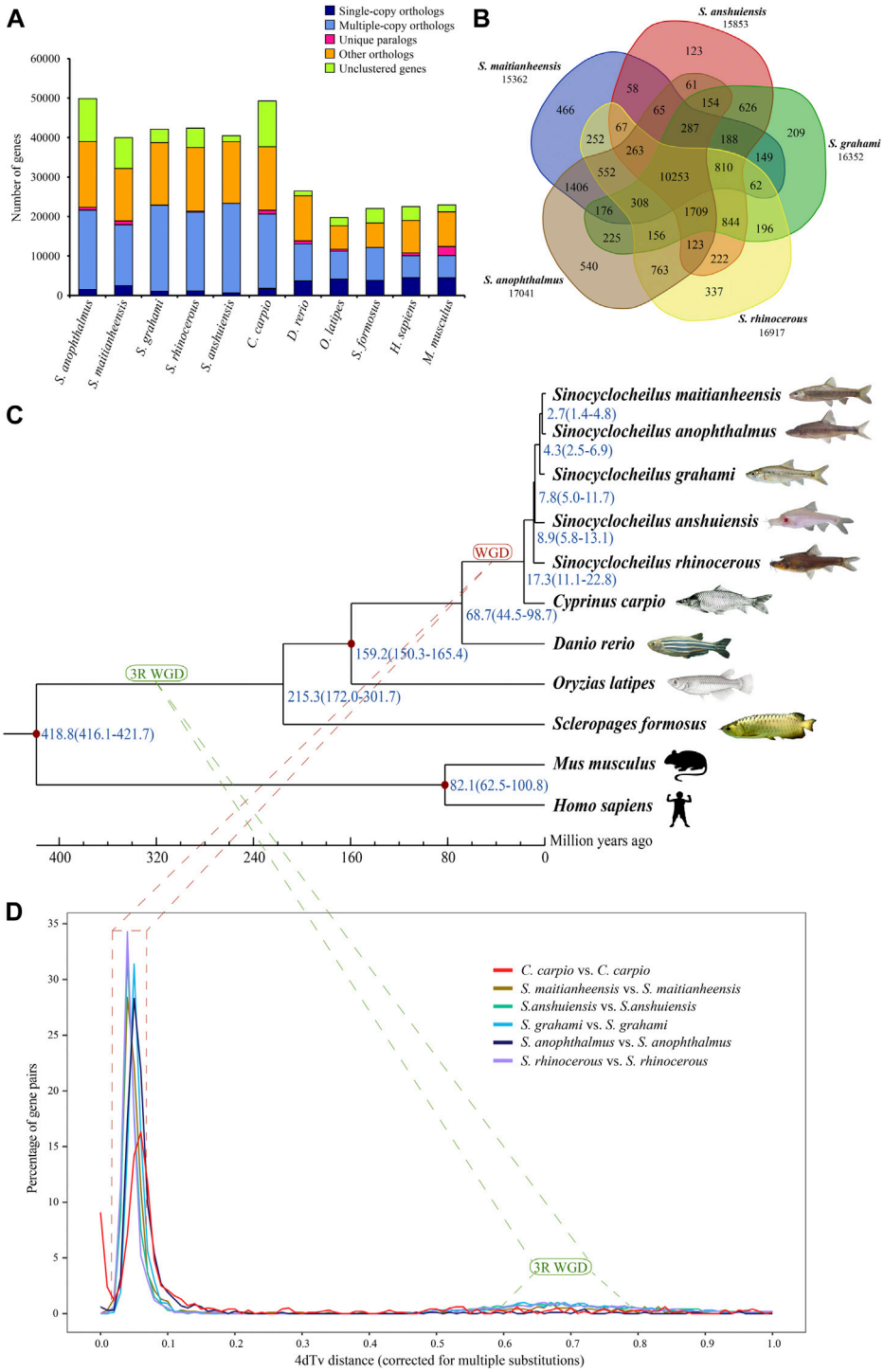

FIGURE 1 | Orthogroups and an evolutionary analysis. **(A)** Distribution of homologous orthogroups among the examined 10 vertebrate species. **(B)** Orthogroup cluster of the five *Sinocyclocheilus* fishes, including *S. maitianheensis*, *S. anophthalmus*, *S. grahami*, *S. anshuiensis*, and *S. rhinocerous*. **(C)** A phylogenetic tree of various representative vertebrates including the five *Sinocyclocheilus* fishes. Diverge time is represented in blue, and the geographic time scale is in million years (for Mya). "WGD" represents the *Sinocyclocheilus*-specific WGD event (red), and "3R WGD" represents the third-round WGD event (green). **(D)** 4dTv distributions of self-alignments in five *Sinocyclocheilus* fishes, and common carp.

approximately 1,000 bp in *S. maitianheensis*. Moreover, our annotation results reveal that *S. maitianheensis* has less protein-coding gene number (39,977) than the other four *Sinocyclocheilus* counterparts (49,865 for *S. anophthalmus*, 42,109 for *S. grahami*, 40,470 for *S. anshuiensis*, and 42,377 for *S. rhinocerous*). The assessment of protein-coding genes among the five fishes showed that there were more missing BUSCOs in *S. maitianheensis* (**Supplementary Table S8**), which may lead to a lower amount of gene number. However, this divergence of gene number may be caused in part by different annotation methods.

Summary of the Gene Orthogroups

A total of 26,875 orthogroups were predicted for the five examined *Sinocyclocheilus* fishes and six other representative vertebrate species. For *S. maitianheensis*, 32,150 protein-coding genes were clustered into 15,617 orthogroups, although 7,827 genes were not clustered; the numbers of multiple-copy, single-copy orthologs, and unique paralogs were 15,400, 2,508 and 981, respectively (**Figure 1A** and **Supplementary Table S9**). We summarized the numbers of orthogroups shared with each other between the five *Sinocyclocheilus* in **Figure 1B**. There are 10,253 common orthogroups among these *Sinocyclocheilus* species.

Phylogenetic Position of *S. maitianheensis*

A total of 191 common single-copy orthogroups among all the examined species were used for construction of a phylogenetic tree. A progressive evolutionary relationship in the five *Sinocyclocheilus* species revealed a new phylogeny based on both genome and transcriptome data (**Figure 1C**). This topology is consistent with the phylogenetic tree based on genomes (Yang et al., 2016) and mitochondrial genes (Zhang and Wang, 2018), but it is different from a previous report of a division into two branches, in which *S. maitianheensis*, *S. anophthalmus*, and *S. anophthalmus* are located in one branch, whereas *S. anshuiensis* and *S. rhinocerous* are located in another branch (Mao et al., 2021) based on the morphological trait of eyes. Therefore, our latest topology provides novel insights into detailed phylogeny for the five *Sinocyclocheilus* species at a genome level.

S. maitianheensis and *S. anophthalmus* diverged at approximately 2.7 (1.4–4.8) Mya, and the divergence time periods with *S. grahami, S. anshuiensis*, and *S. rhinocerous* were 4.3 (2.5–6.9), 7.8 (5.0–11.7), and 8.9 (5.8–13.1) Mya, respectively (**Figure 1C**). Surface-dwelling *S. maitianheensis* has the closest relationship with cave-restricted *S. anophthalmus*; however, there are huge differences in morphological traits between them. Interestingly, the geographical positions of both habitats are close, located in the same Yiliang County (Kunming City, Yunnan Province, China; **Supplementary Figure S1**). *S. anophthalmus* resides in a dark environment of several caves in Jiuxiang Town (Zhao and Zhang, 2009), whereas *S. maitianheensis* lives in the surface of Maitian river. Divergence time of the two *Sinocyclocheilus* species was relatively late, approximately 2.7 (1.4–4.8) Mya. Their separation may be due to a geographical isolation generated by the continuous uplift of the Yunnan–Guizhou Plateau after Himalayan orogeny (50–40 Mya) (Yin and Harrison, 2000), and some of their ancestors swam down along the Maitian river into these surrounding caves.

Estimation of the Lineage-specific WGD in *Sinocyclocheilus*

Cyprinidae experienced a recent genome-wide duplication event (Larhammar and Risinger, 1994; David et al., 2003) after the third-round WGD (3R WGD, also known as teleost-specific WGD) (Jaillon et al., 2004). We performed a 4dTv analysis to estimate the timing of this recent lineage-specific WGD in *Sinocyclocheilus*. Self-alignments with paralogous gene pairs of *S. maitianheensis*, *S. anshuiensis*, *S. rhinocerous*, *S. anophthalmus*, *S. grahami*, and common carp displayed distinct peaks (**Figure 1D**). Their 4dTv distances were calculated to be 0.04, 0.04, 0.04, 0.05, 0.05, and 0.06, respectively. The peak values of the five *Sinocyclocheilus* fishes and common carp are very close (0.04–0.06), implying that these fishes might share the recent genome-wide duplication event.

In order to compare the recent specific WGD between *Sinocyclocheilus* fishes and common carp, we performed a 4dTv analysis on 13,579 orthologous gene pairs between *S. maitianheensis* and common carp genomes (**Supplementary Figure S4**). The peak values in each group of *C. carpio–C. carpio*, *S. maitianheensis–C. carpio*, and *S. maitianheensis–S. maitianheensis* were estimated to be 0.06, 0.04, and 0.04, respectively. The nearly overlapping of peaks indicated that *S. maitianheensis* and common carp might have undergone the recent specific WGD together. Based on the aforementioned 4dTv analyses and the construction of phylogenetic tree, we predicted that the carp-specific WGD occurred ~18.1 Mya, before the evolutionary separation of *Sinocyclocheilus* fishes and common carp (~17.3 Mya; **Figure 1D**). This estimate is little earlier before the specific WGD of common carp based on the divergence time of transposable elements (Xu et al., 2019).

The time estimation of the latest WGD in *Cyprinidae* is contentious, from ~8.2 to ~16 Mya (Larhammar and Risinger, 1994; David et al., 2003; Xu et al., 2014), although these studies almost focused on the common carp. However, a recent study of common carp defined a rough time range (9.7–23 Mya) and further predicted this time point to about 12.4 Mya (Xu et al., 2019). Our result of ~18.4 Mya in the present study is within this time range, and the 4dTv analysis of *Sinocyclocheilus* fishes provides more evidence for the timing of the latest genome duplication in *Cyprinidae*.

Copy Number Variations of Several Immune Genes in P38 and Mitochondrial Pathways

S. anophthalmus has evolved a series of traits to adapt to the caved environment. It had developed huge differences from its sister species (*S. maitianheensis*), such as loss of eyes and the semitransparent body (Zhao and Zhang, 2009). Therefore, compared to *S. maitianheensis*, *S. anophthalmus* is an independent cave species with many valuable traits that are waiting for in-depth explorations.

To investigate potential immunological variances in cave-restricted *Sinocyclocheilus* species, we identified the copy number

of 15 immune genes within P38 and mitochondrial pathways in the five examined *Sinocyclocheilus* genomes (see detailed statistics in **Supplementary Table S10**). These genes include *tak1*, *tab1*, *ask1*, *fas*, *fasl*, *fadd*, *tnfa*, *cd40*, *cd40l*, *daxx*, *mkk4a, mkk4b*, *mkk6*, *bcl-2a*, and *bcl2l1* (**Supplementary Figure S5**). In general, five *Sinocyclocheilus* species and common carp usually own twofold copies compared to the other diploid teleosts. Interestingly, we found copies of some genes in the cave-restricted fishes. For *S. anophthalmus*, a copy of *ask1* (an apoptosis signal-regulating kinase) (Patel et al., 2019) was predicted as a pseudogene; only one copy of *bcl2l1* (encoding apoptotic regulators in BCL-2 family) (Warren et al., 2019) was retained in *S. anophthalmus* compared with *S. maitianheensis* that contained two copies; one more copy of *bcl-2a* and no copy of *mkk4a* were also observed in *S. anophthalmus* genome. In addition, among most of these genes we studied, another cave-restricted fish Mexican tetra (*A. mexicanus*) had fewer gene copy numbers than the other examined fishes. These variances in gene copy number imply that the apoptotic activity might have been decreased in cave-restricted fishes, which is consistent with our previous report of relatively lower immunity in cavefishes (Qiu et al., 2016). However, apoptotic activity is regulated by many factors, and more investigations should be done for in-depth verification.

In addition, we performed high-throughput identification of antimicrobial peptides (AMPs) (Yi et al., 2017; Mwangi et al., 2019). A total of 379, 551, 522, 545, and 552 putative AMP sequences were identified in *S. maitianheensis*, *S. rhinocerous*, *S. anshuiensis*, *S. anophthalmus*, and *S. grahami* genomes, respectively (**Supplementary Table S11**). Thrombin, histone, lectin, chemokine, scolopendin, and ubiquitin are the most abundant AMPs in the five examined *Sinocyclocheilus* species. The lowest number of AMP sequences in *S. maitianheensis* genome may be related to its least protein-coding genes in the five *Sinocyclocheilus* fishes.

CONCLUSION

In summary, we reported the first genome assembly of *S. maitianheensis*, which provides a valuable genetic resource for comparative studies on cavefish biology, species protection and practical aquaculture of this potentially economical teleost fish. This genome assembly also supplies essential genomic data for in-depth genetic analysis. Based on these genomic data, we observed a close relationship between *S. maitianheensis* and *S. anophthalmus*. Some variations of gene copy number in the immune system might indicate the variation in immunity and apoptosis in cave-restricted *Sinocyclocheilus* species.

VALUE OF THE DATA

This is the first draft genome of a representative surface-dwelling Chinese golden-line barbel fish, *Sinocyclocheilus maitianheensis*. The final assembly was 1.7 Gb with a scaffold N50 of 1.4 Mb and a contig N50 of 24.7 kb.

The phylogenetic tree revealed that *S. maitianheensis* is close to *S. anophthalmus* (a cave-restricted species with similar locality). The divergence time between the two relatives is about 2.7 million years ago (Mya).

The 4dTv analysis demonstrated that the recent carp-specific WGD event occurred approximately 18.1 Mya.

A decrease in the copy number of many important immunological genes was observed in cave-restricted *Sinocyclocheilus* species.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

All animal experiments in this study were performed in accordance with the guidelines of the Animal Ethics Committee and were approved by the Institutional Review Board on Bioethics and Biosafety of BGI, China (No. FT18134). Written informed consent was obtained from the owners for the participation of their animals in this study.

AUTHOR CONTRIBUTIONS

Conceptualization, JY and QS; data analysis, RL and ZG; samples collection and assisted data analysis, YZ, WJ, and MW; data curation, XY, LC, and XP; writing-original draft preparation, RL and ZG; writing-review and editing, QS, CB, and XW; supervision, QS and JY; funding acquisition, JY and QS. All authors have read and approved the published version of the manuscript.

FUNDING

This study was supported by Shenzhen Science and Technology Innovation Program for International Cooperation (no. GJHZ20190819152407214), National Natural Science Foundation of China (Nos. 31672282 and U1702233), and Grant Plan for Demonstration City Project for Marine Economic Development in Shenzhen (No. 86).

ACKNOWLEDGMENTS

We appreciate Dr. Yu Huang, a BGI Marine employee, for her editing assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.736500/full#supplementary-material

REFERENCES

Aspiras, A. C., Rohner, N., Martineau, B., Borowsky, R. L., and Tabin, C. J. (2015). Melanocortin 4 Receptor Mutations Contribute to the Adaptation of Cavefish to Nutrient-Poor Conditions. *Proc. Natl. Acad. Sci. USA* 112 (31), 9668–9673. doi:10.1073/pnas.1510802112

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2006). *Genbank. Nucleic Acids Res.* 34, D16–D20. doi:10.1093/nar/gkj157Database issue

Benson, G. (1999). Tandem Repeats Finder: a Program to Analyze DNA Sequences. *Nucleic Acids Res.* 27 (2), 573–580. doi:10.1093/nar/27.2.573

Benton, M. J., and Donoghue, P. C. J. (2007). Paleontological Evidence to Date the Tree of Life. *Mol. Biol. Evol.* 24 (1), 26–53. doi:10.1093/molbev/msl150

Bian, C., Huang, Y., Li, J., You, X., Yi, Y., Ge, W., et al. (2019). Divergence, Evolution and Adaptation in ray-finned Fish Genomes. *Sci. China Life Sci.* 62 (8), 1003–1018. doi:10.1007/s11427-018-9499-5

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. *Genome Res.* 14 (5), 988–995. doi:10.1101/gr.1865504

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., et al. (2003). The SWISS-PROT Protein Knowledgebase and its Supplement TrEMBL in 2003. *Nucleic Acids Res.* 31 (1), 365–370. doi:10.1093/nar/gkg095

Boetzer, M., and Pirovano, W. (2014). SSPACE-LongRead: Scaffolding Bacterial Draft Genomes Using Long Read Sequence Information. *BMC Bioinformatics* 15, 211. doi:10.1186/1471-2105-15-211

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., et al. (2008). MAKER: an Easy-To-Use Annotation Pipeline Designed for Emerging Model Organism Genomes. *Genome Res.* 18 (1), 188–196. doi:10.1101/gr.6743907

Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Mol. Biol. Evol.* 17 (4), 540–552. doi:10.1093/oxfordjournals.molbev.a026334

Chen, N. (2004). Using Repeat Masker to Identify Repetitive Elements in Genomic Sequences. *Curr. Protoc. Bioinformatics* 5. , 2004 Chapter 4Unit 4.10. doi:10.1002/0471250953.bi0410s05

Chen, Y.-Y., Li, R., Li, C.-Q., Li, W.-X., Yang, H.-F., Xiao, H., et al. (2018b). Testing the Validity of Two Putative Sympatric Species from Sinocyclocheilus (Cypriniformes: Cyprinidae) Based on Mitochondrial Cytochrome B Sequences. *Zootaxa* 4476 (1), 130–140. doi:10.11646/zootaxa.4476.1.12

Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., et al. (2018a). SOAPnuke: a MapReduce Acceleration-Supported Software for Integrated Quality Control and Preprocessing of High-Throughput Sequencing Data. *Gigascience* 7 (1), 1–6. doi:10.1093/gigascience/gix120

David, L., Blum, S., Feldman, M. W., Lavi, U., and Hillel, J. (2003). Recent Duplication of the Common Carp (*Cyprinus carpio* L.) Genome as Revealed by Analyses of Microsatellite Loci. *Mol. Biol. Evol.* 20 (9), 1425–1434. doi:10.1093/molbev/msg173

Edgar, R. C. (2004). MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Res.* 32 (5), 1792–1797. doi:10.1093/nar/gkh340

Fischer, S., Brunk, B. P., Chen, F., Gao, X., Harb, O. S., Iodice, J. B., et al. (2011). Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes into New Ortholog Groups. *Curr. Protoc. Bioinformatics Chapter* 6, 11–19. Unit 6.12. doi:10.1002/0471250953.bi0612s35

Gertz, E. M., Yu, Y.-K., Agarwala, R., Schäffer, A. A., and Altschul, S. F. (2006). Composition-based Statistics and Translated Nucleotide Searches: Improving the TBLASTN Module of BLAST. *BMC Biol.* 4, 41. doi:10.1186/1741-7007-4-41

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol.* 59 (3), 307–321. doi:10.1093/sysbio/syq010

He, S., Liang, X.-F., Chu, W.-Y., and Chen, D.-X. (2012). Complete Mitochondrial Genome of the Blind Cave barbelSinocyclocheilus furcodorsalis(Cypriniformes: Cyprinidae). *Mitochondrial DNA* 23 (6), 429–431. doi:10.3109/19401736.2012.710216

Heng, X., Rendong, Z., Jianguo, F., Ming, O., Weixian, L., Shanyuan, C., et al. (2002). Nuclear DNA Content and Ploidy of Seventeen Species of Fishes in Sinocyclocheilus. *Dong Wu Xue Yan jiu= Zoolog. Res.* 23 (3), 195–199.

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., et al. (2009). InterPro: the Integrative Protein Signature Database. *Nucleic Acids Res.* 37 (Database issue), D211–D215. doi:10.1093/nar/gkn785

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., et al. (2004). Genome Duplication in the Teleost Fish *Tetraodon nigroviridis* Reveals the Early Vertebrate Proto-Karyotype. *Nature* 431 (7011), 946–957. doi:10.1038/nature03025

Jeffery, W. R. (2001). Cavefish as a Model System in Evolutionary Developmental Biology. *Dev. Biol.* 231 (1), 1–12. doi:10.1006/dbio.2000.0121

Jeffery, W. R. (2009). Regressive Evolution inAstyanaxCavefish. *Annu. Rev. Genet.* 43, 25–47. doi:10.1146/annurev-genet-102108-134216

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a Database of Eukaryotic Repetitive Elements. *Cytogenet. Genome Res.* 110 (1-4), 462–467. doi:10.1159/000084979

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs. *Nucleic Acids Res.* 45 (D1), D353–d361. doi:10.1093/nar/gkw1092

Krishnan, J., and Rohner, N. (2017). Cavefish and the Basis for Eye Loss. *Phil. Trans. R. Soc. B.* 372 (1713), 20150487. doi:10.1098/rstb.2015.0487

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* 33 (7), 1870–1874. doi:10.1093/molbev/msw054

Larhammar, D., and Risinger, C. (1994). Molecular Genetic Aspects of Tetraploidy in the Common Carp *Cyprinus carpio*. *Mol. Phylogenet. Evol.* 3 (1), 59–68. doi:10.1006/mpev.1994.1007

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009a). The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* 25 (16), 2078–2079. doi:10.1093/bioinformatics/btp352

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., et al. (2009b). SOAP2: an Improved Ultrafast Tool for Short Read Alignment. *Bioinformatics* 25 (15), 1966–1967. doi:10.1093/bioinformatics/btp336

Liu, H.-P., Xiao, S.-J., Wu, N., Wang, D., Liu, Y.-C., Zhou, C.-W., et al. (2019). The Sequence and De Novo Assembly of Oxygymnocypris Stewartii Genome. *Sci. Data* 6 (1), 190009. doi:10.1038/sdata.2019.9

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: an Empirically Improved Memory-Efficient Short-Read De Novo Assembler. *GigaSc.i* 1 (1), 18. doi:10.1186/2047-217x-1-18

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., and Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* 28, msab199. doi:10.1093/molbev/msab199

Mao, T.-R., Liu, Y.-W., Meegaskumbura, M., Yang, J., Ellepola, G., Senevirathne, G., et al. (2021). Evolution in Sinocyclocheilus Cavefish Is Marked by Rate Shifts, Reversals, and Origin of Novel Traits. *BMC Ecol. Evo* 21 (1), 45. doi:10.1186/s12862-021-01776-y

McGinnis, S., and Madden, T. L. (2004). BLAST: at the Core of a Powerful and Diverse Set of Sequence Analysis Tools. *Nucleic Acids Res.* 32, W20–W25. Web Server issue. doi:10.1093/nar/gkh435

Mwangi, J., Hao, X., Lai, R., and Zhang, Z.-Y. (2019). Antimicrobial Peptides: new hope in the War against Multidrug Resistance. *Zool Res.* 40 (6), 488–505. doi:10.24272/j.issn.2095-8137.2019.062

Patel, P., Naik, M. U., Golla, K., Shaik, N. F., and Naik, U. P. (2019). Calcium-induced Dissociation of CIB1 from ASK1 Regulates Agonist-Induced Activation of the P38 MAPK Pathway in Platelets. *Biochem. J.* 476 (19), 2835–2850. doi:10.1042/bcj20190410

Peuß, R., Box, A. C., Chen, S., Wang, Y., Tsuchiya, D., Persons, J. L., et al. (2020). Adaptation to Low Parasite Abundance Affects Immune Investment and Immunopathological Responses of Cavefish. *Nat. Ecol. Evol.* 4 (10), 1416–1430. doi:10.1038/s41559-020-1234-2

Proost, S., Fostier, J., De Witte, D., Dhoedt, B., Demeester, P., Van de Peer, Y., et al. (2012). I-ADHoRe 3.0-fast and Sensitive Detection of Genomic Homology in Extremely Large Data Sets. *Nucleic Acids Res.* 40 (2), e11. doi:10.1093/nar/gkr955

Qiu, Y., Yang, J., Jiang, W., Chen, X., Bian, C., and Shi, Q. (2016). A Genomic Survey on the Immune Differences amongSinocyclocheilusfishes. *Communicative Integr. Biol.* 9 (6), e1255833. doi:10.1080/19420889.2016.1255833

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs. *Bioinformatics* 31 (19), 3210–3212. doi:10.1093/bioinformatics/btv351

Song, L., Bian, C., Luo, Y., Wang, L., You, X., Li, J., et al. (2016). Draft Genome of the Chinese Mitten Crab, Eriocheir Sinensis. *GigaSci*. 5, 5. doi:10.1186/s13742-016-0112-y

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L. (2013). Differential Analysis of Gene Regulation at Transcript Resolution with RNA-Seq. *Nat. Biotechnol.* 31 (1), 46–53. doi:10.1038/nbt.2450

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: Discovering Splice Junctions with RNA-Seq. *Bioinformatics* 25 (9), 1105–1111. doi:10.1093/bioinformatics/btp120

Warren, C. F. A., Wong-Brown, M. W., and Bowden, N. A. (2019). BCL-2 Family Isoforms in Apoptosis and Cancer. *Cell Death Dis.* 10 (3), 177. doi:10.1038/s41419-019-1407-6

Wu, X., Wang, L., Chen, S., Zan, R., Xiao, H., and Zhang, Y.-p. (2010). The Complete Mitochondrial Genomes of Two Species from Sinocyclocheilus (Cypriniformes: Cyprinidae) and a Phylogenetic Analysis within Cyprininae. *Mol. Biol. Rep.* 37 (5), 2163–2171. doi:10.1007/s11033-009-9689-x

Xiong, J. B., Nie, L., and Chen, J. (2019). Current Understanding on the Roles of Gut Microbiota in Fish Disease and Immunity. *Zool. Res.* 40 (2), 70–76. doi:10.24272/j.issn.2095-8137.2018.069

Xu, P., Xu, J., Liu, G., Chen, L., Zhou, Z., Peng, W., et al. (2019). The Allotetraploid Origin and Asymmetrical Genome Evolution of the Common Carp *Cyprinus carpio*. *Nat. Commun.* 10 (1), 4625. doi:10.1038/s41467-019-12644-1

Xu, P., Zhang, X., Wang, X., Li, J., Liu, G., Kuang, Y., et al. (2014). Genome Sequence and Genetic Diversity of the Common Carp, *Cyprinus carpio*. *Nat. Genet.* 46 (11), 1212–1219. doi:10.1038/ng.3098

Xu, Z., and Wang, H. (2007). LTR_FINDER: an Efficient Tool for the Prediction of Full-Length LTR Retrotransposons. *Nucleic Acids Res.* 35, W265–W268. Web Server issue. doi:10.1093/nar/gkm286

Yang, J., Chen, X., Bai, J., Fang, D., Qiu, Y., Jiang, W., et al. (2016). The Sinocyclocheilus Cavefish Genome Provides Insights into Cave Adaptation. *BMC Biol.* 14, 1. doi:10.1186/s12915-015-0223-4

Yang, Z., and Rannala, B. (2006). Bayesian Estimation of Species Divergence Times under a Molecular Clock Using Multiple Fossil Calibrations with Soft Bounds. *Mol. Biol. Evol.* 23 (1), 212–226. doi:10.1093/molbev/msj024

Ye, C., Hill, C. M., Wu, S., Ruan, J., and Ma, Z. (2016). DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. *Sci. Rep.* 6, 31900. doi:10.1038/srep31900

Yi, Y., You, X., Bian, C., Chen, S., Lv, Z., Qiu, L., et al. (2017). High-Throughput Identification of Antimicrobial Peptides from Amphibious Mudskippers. *Mar. Drugs* 15 (11), 364. doi:10.3390/md15110364

Yin, A., and Harrison, T. M. (2000). Geologic Evolution of the Himalayan-Tibetan Orogen. *Annu. Rev. Earth Planet. Sci.* 28 (1), 211–280. doi:10.1146/annurev.earth.28.1.211

Yin, Y. H., Zhang, X. H., Wang, X. A., Li, R. H., Zhang, Y. W., Shan, X. X., et al. (2021). Construction of a Chromosome-Level Genome Assembly for Genome-wide Identification of Growth-Related Quantitative Trait Loci in *Sinocyclocheilus grahami* (Cypriniformes, Cyprinidae). *Zool Res.* 42 (3), 262–266. doi:10.24272/j.issn.2095-8137.2020.321

Yoshizawa, M., Gorički, Š., Soares, D., and Jeffery, W. R. (2010). Evolution of a Behavioral Shift Mediated by Superficial Neuromasts Helps Cavefish Find Food in Darkness. *Curr. Biol.* 20 (18), 1631–1636. doi:10.1016/j.cub.2010.07.017

Zhang, R., and Wang, X. (2018). Characterization and Phylogenetic Analysis of the Complete Mitogenome of a Rare Cavefish, Sinocyclocheilus Multipunctatus (Cypriniformes: Cyprinidae). *Genes Genom* 40 (10), 1033–1040. doi:10.1007/s13258-018-0711-3

Zhao, N., Guo, H., Jia, L., Guo, B., Zheng, D., Liu, S., et al. (2021). Genome Assembly and Annotation at the Chromosomal Level of First Pleuronectidae: Verasper Variegatus Provides a Basis for Phylogenetic Study of Pleuronectiformes. *Genomics* 113 (2), 717–726. doi:10.1016/j.ygeno.2021.01.024

Zhao, Y., and Zhang, C. (2009). Threatened Fishes of the World: Sinocyclocheilus Anophthalmus (Chen and Chu, 1988) (Cyprinidae). *Environ. Biol. Fish.* 86 (1), 163. doi:10.1007/s10641-008-9361-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

DATA REPORT
published: 21 December 2021
doi: 10.3389/fgene.2021.811798

Chromosome-Level Assembly of the Southern Rock Bream (*Oplegnathus fasciatus*) Genome Using PacBio and Hi-C Technologies

*Yulin Bai[1†], Jie Gong[1†], Zhixiong Zhou[1], Bijun Li[1], Ji Zhao[1], Qiaozhen Ke[1,2], Xiaoqing Zou[1], Fei Pu[1], Linni Wu[1], Weiqiang Zheng[2], Tao Zhou[1,3] and Peng Xu[1,3]**

[1]State Key Laboratory of Marine Environmental Science, College of Ocean and Earth Sciences, Xiamen University, Xiamen, China, [2]State Key Laboratory of Large Yellow Croaker Breeding, Ningde Fufa Fisheries Company Limited, Ningde, China, [3]Fujian Key Laboratory of Genetics and Breeding of Marine Organisms, College of Ocean and Earth Sciences, Xiamen University, Xiamen, China

OPEN ACCESS

Edited by:
Roger Huerlimann,
Okinawa Institute of Science and Technology Graduate University, Japan

Reviewed by:
Tianxiang Gao,
Zhejiang Ocean University, China
Syed Farhan Ahmad,
Kasetsart University, Thailand

***Correspondence:**
Peng Xu
xupeng77@xmu.edu.cn

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: 09 November 2021
Accepted: 29 November 2021
Published: 21 December 2021

Citation:
Bai Y, Gong J, Zhou Z, Li B, Zhao J, Ke Q, Zou X, Pu F, Wu L, Zheng W, Zhou T and Xu P (2021) Chromosome-Level Assembly of the Southern Rock Bream (Oplegnathus fasciatus) Genome Using PacBio and Hi-C Technologies. Front. Genet. 12:811798. doi: 10.3389/fgene.2021.811798

The Rock Bream (*Oplegnathus fasciatus*) is an economically important rocky reef fish of the Northwest Pacific Ocean. In recent years, it has been cultivated as an important edible fish in coastal areas of China. Despite its economic importance, genome-wide adaptions of domesticated *O. fasciatus* are largely unknown. Here we report a chromosome-level reference genome of female *O. fasciatus* (from the southern population in the subtropical region) using the PacBio single molecule sequencing technique (SMRT) and High-through chromosome conformation capture (Hi-C) technologies. The genome was assembled into 120 contigs with a total length of 732.95 Mb and a contig N50 length of 27.33 Mb. After chromosome-level scaffolding, 24 chromosomes with a total length of 723.22 Mb were constructed. Moreover, a total of 27,015 protein-coding genes and 5,880 ncRNAs were annotated in the reference genome. This reference genome of *O. fasciatus* will provide an important resource not only for basic ecological and population genetic studies but also for dissect artificial selection mechanisms in marine aquaculture.

Keywords: *Oplegnathus fasciatus*, genomic resources, PacBio, Hi-C, phylogenetic analysis

INTRODUCTION

The Rock Bream (*Oplegnathus fasciatus*), inhabiting in coastal rocky reefs and feeds on invertebrates inhabiting the seabed, is an endemic marine fish in East Asia that belongs to Oplegnathidae in *Perciformes* (**Figure 1A**) (An et al., 2008). In recent years, *O. fasciatus* has become one of the most commercially valuable marine fishery species in China aquaculture (Xiao et al., 2019). As a sedentary

Abbreviations: *B. petinirostris:* Blue Spotted Mudskipper, *Boleophthalmus petinirostris*; *C. semilaevis:* Half-Smooth Tongue Sole, *Cynoglossus semilaevis*; BUSCO, Benchmarking Universal Single Copy Orthologues; CLR, Continuous Long-Read; *D. reio:* Zebrafish, *Danio reio*; GS, Genome selection; *G. aculeatus:* Three-Spined Stickleback, *Gasterosteus aculeatus*; *G. morhu:* Atlantic Cod, *Gadus morhua*; Hi-C, High-through chromosome conformation capture; *L. crocea*: Large Yellow Croaker, *Larimichthys crocea*; *L. japonicus:* Chinese Seabass, *Lateolabrax maculatus*; PacBio SMRT, ME, Minimum Evolution; *N. coriiceps*: Antarctic rock cod, *Notothenia coriiceps*; PacBio Single molecule sequencing technique; *O. fasciatus:* Rock Bream, *O. latipes:* Japanese Rice Fish, *Oryzias latipes*; *Oplegnathus fasciatus*; *P. olivaceus:* Olive Flounder, *Paralichthys olivaceus*; *T. bimaculatus:* Pufferfish, *Takifugu bimaculatus*; TEs, Transposable Elements.

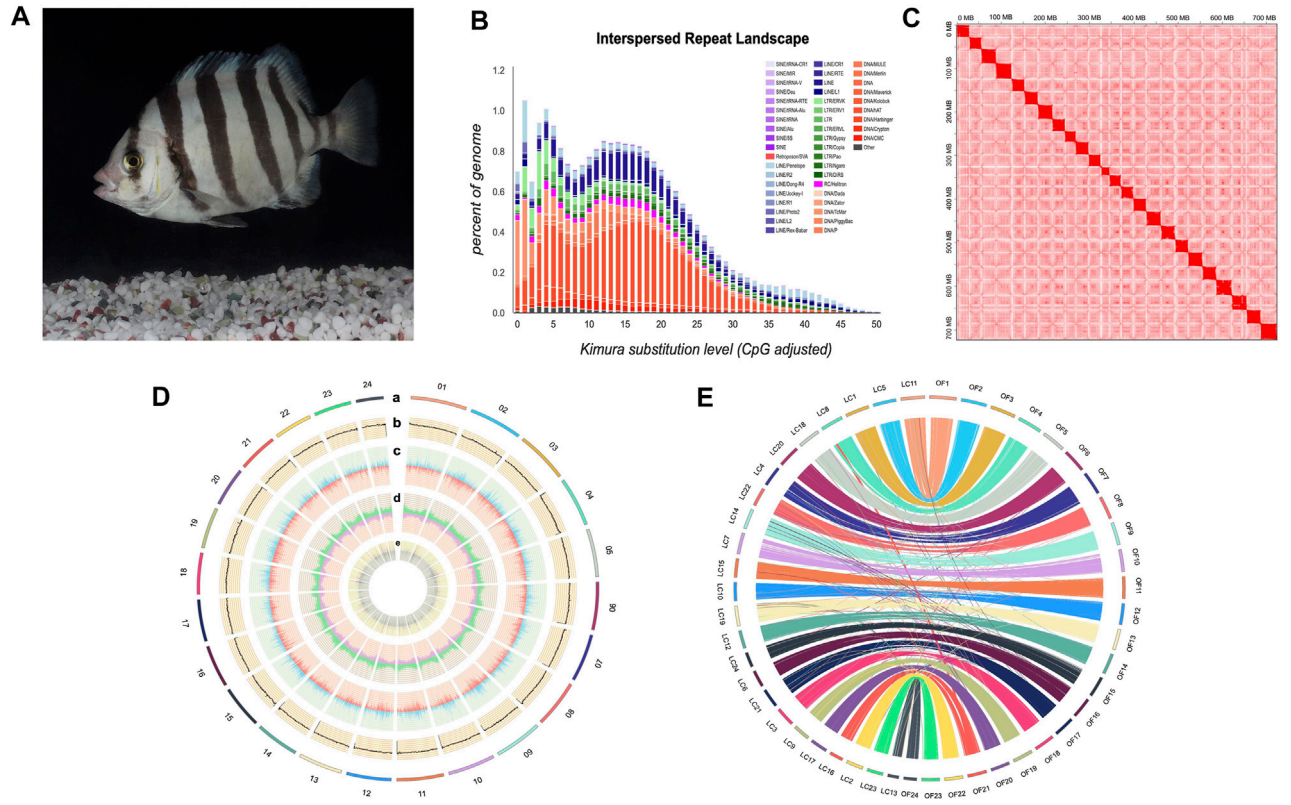

FIGURE 1 | Characteristics of Oplegnathus fasciatus genome assembly. **(A)** A photograph of female *O. fasciatus* (from Zhoushan population). **(B)** Distribution of divergence rate for each type of TE in *O. fasciatus* genome. Percentages of TEs in genome (Y-axis) and Kimura divergence (X-axis). Values from 0-50 reflect older copies (right side) and new copies (left side). **(C)** Heat map of chromosomal interactions in the *O. fasciatus* genome using Hi-C data. **(D)** A circos plot of 24 chromosomes in *O. fasciatus* genome, the tracks from outside to inside are: a. Lines represent *O. fasciatus* chromosomes; b. GC content; c. Gene density; d. Repeat element density; e. ncRNA density. b-e. are drawn in nonoverlapping 100 kb sliding windows. **(E)** Circos diagram between *O. fasciatus* and *L. crocea*. Each coloured arc represents a 1 Kb fragment match between two species. We re-ordered the chromosome numbers of *L. crocea* for better illustration.

species, due to habitat restrictions, *O. fasciatus* is broadly and fragmented distributed in the coastal areas of eastern China (Xiao et al., 2016a). According to the difference of genetic variation, the researchers identified the *O. fasciatus* population as northern (Jiaonan, Qingdao) and southern regions (Zhoushan, Zhejiang) of Chinese coastal waters (Xiao et al., 2016b). The different environments make them divergent in genetic structure and living habits, which provides a suitable fish model for the study of population genetic in marginal sea (Jian et al., 2017; Chen et al., 2019a). The male and female reference genome of *O. fasciatus* derived from the northern population has been reported and revealed its particularity in gender mechanism (Xiao et al., 2019; Xiao et al., 2020). However, a highly accurate reference genome of subtropical *O. fasciatus* species is still lacking, which hinders the progress of genome-scale genetic breeding and genetic studies of its temperature plasticity and adaptation at lower latitudes.

Recently, the genetic breeding of economic fish has attracted much attention, mainly aiming to improve growth rates and disease resistance under aquaculture conditions (Bangera et al., 2017; Gong et al., 2021; Zhao et al., 2021). Genome selection (GS) has become an efficient and popular breeding strategy, which depends on high-quality reference genome (Singh et al., 2018). It is worthwhile to invest time and money to obtain a reference genome more suitable for breeding populations to improve the prediction accuracy (Benevenuto et al., 2019). Our breeding work related to the growth traits of *O. fasciatus* reported recently also supports this conclusion (Gong et al., 2021). In this report, we provided a chromosome-level reference genome of *O. fasciatus* (from the southern population in the subtropical region) using a combination of the PacBio single molecule sequencing technique (SMRT) and high-through chromosome conformation capture (Hi-C) technologies. We assembled the genome sequences into 120 contigs with a total length of 732.95 Mb and a contig N50 length of 27.33 Mb. After chromosome-level scaffolding, 24 scaffolds were constructed corresponding to 24 chromosomes with a total length of 723.22 Mb (98.67% of the total length of all contigs). The availability of data is essential to support the population genetic studies, and will also provide an important resource for the upcoming breeding program of *O. fasciatus*.

MATERIALS AND METHODS

Sample Collection, Library Construction and Sequencing

A healthy female *O. fasciatus* (from Zhoushan population) were obtained from a commercial breeding company, Ningde Fufa Aquatic Breeding (Fujian, China). All samples were collected, snap frozen in the liquid nitrogen and stored at −80°C to maintain nucleic acid integrity.

The muscle tissues were collected for DNA extraction. For the PacBio sequencing project, frozen samples were lysed in SDS digestion buffer with proteinase-K (50 μg/ml). Then, the lysates were purified using AMPure XP beads (Beckman Coulter, High Wycombe, United Kingdom) to obtain High-Molecular-Weight gDNA. Library construction and sequencing were conducted according to the manufacturer's protocol with the PacBio RS-II platform at Novogene (Tianjin). Meanwhile, the Normal-Molecular-Weight gDNA for Illumina sequencing was extracted from the same samples using the PureLink™ Pro 96 Genomic DNA Purification Kit (Invitrogen, Shanghai, China). A pair-end library with 350 bp insert size was constructed and sequenced using the Illumina NovaSeq 6000 platform with a read length of 2 × 150 bp.

For Hi-C sequencing, The DNA was fixed by formaldehyde to maintain the conformation and the restriction enzyme (HindIII) was applied on DNA digestion, followed by repairing 5' overhangs with biotinylated residues. After *in-situ* ligation of these fragments, DNA was reverse-cross linked and purified. Finally, sequencing of the Hi-C library was performed on an Illumina NovaSeq 6000 platform.

Additionally, 11 different tissues (heart, liver, spleen, intestine, kidney, skin, eye, gill, brain, blood, muscle) were collected to extract RNA for RNA sequencing (RNA-seq) following the protocols of the PureLink™ RNA Mini Kit (Invitrogen, Shanghai, China). The library was constructed using the Illumina standard protocol (San Diego, CA, United States) and sequenced on the Illumina HiSeq 6000 platform.

Data Processing and Genome Assembly

Before assembly, the PacBio data was further filtered, and reads with length less than 1500 bp or low-quality were removed. For the Illumina data, adapter sequences and low-quality reads were filtered using fastp (v. 0.23.1) software. The remaining reads were used for further assembly and estimation of genome size using the K-mer analysis of the short reads (see below).

SOAPec (v. 2.01) and GenomeScope (v. 2.0) softwares were used to analyze the K-mer frequencies in the sequencing reads to obtain genome characteristics such as genome size, heterozygosity, and repeatability.

To obtain chromosome-level whole genome assembly for *O. fasciatus*, we utilized a combined approach of Illumina, PacBio and Hi-C technology for the genome assembly and chromosome-level scaffolding. Then, low-quality and duplicated reads were filtered out. The *O. fasciatus* genome was assembled using a hybrid SMRT-Illumina-HiC strategy as follows: 1) Contigs from Continuous Long-Read (CLR) clean reads were assembled using Canu (v. 2.0) (Sergey et al., 2017) with default parameters. The high-fidelity contig sets was produced by using a combination of circular consensus CLR reads, Illumina paired-end reads with sufficient overlap to merge into single extended accurate reads; 2) Purge_Dups (v. 1.2.5) was employed to resolve redundancy in the assembly; 3) The non-redundant contig sets were reordered and scaffolded using 3D-DNA pipeline (Dudchenko et al., 2017); 4) Scaffolds were fine-tuned and discordant contigs were removed from scaffolds using Juicebox (v. 1.5) (Robinson et al., 2018). Finally, we obtained a chromosome level reference genome of *O. fasciatus* containing linkage group information.

Annotation of Genomic Repeats

Repetitive sequences of the *O. fasciatus* genome were annotated using both homology-based search and *de novo* methods. RepeatModeler (v. 2.0.1) and LTR_Finder (v. 1.07) were used to detect repeat sequences in the *O. fasciatus* genome. Combined with Repbase (v. 20,181,026; http://www.girinst.org/repbase), a repeat sequence library was constructed. RepeatMasker (v. 4.1.0) were utilized to search and classify repeats based on this library. TEclass (v. 2.1.3) was used to further annotate unclassified repeats. The built-in script buildSummary.pl from RepeatMasker (v. 4.1.0) was used to summarize Transposable Elements (TEs) annotation results. Then two scripts, calcDivergenceFromalign.pl and createRepeatLandscape.pl, were used to calculate the Kimura divergence value and draw repeated landscapes (**Figure 1B**). The nucleotide distances between all copies of each TE measured using the Kimura two-parameter method were compared to estimate insertion age (Schemberger et al., 2019). Tandem Repeats Finder (v. 4. 09) was used to identify tandem repeats. All repetitive regions except tandem repeats were soft-masked for protein-coding gene annotation.

Protein-Coding Gene Finding and Function Annotation

The coding sequences of genetically close species, including *L. crocea*, *L. maculatus*, *G. aculeatus*, and *P. olivaceus*, were retrieved from Ensembl and NCBI. These coding sequences were provided to the software package of Braker2 (v. 2.1.5) (Bruna et al., 2021), and then the genes in the repeat-masked reference genome were annotated with the close homologous protein model. RNA-seq data were aligned to *O. fasciatus* contigs using Blat (v. 36 × 5) and GMAP (v. 2017-11-15), and a comprehensive transcriptome database was built using PASA (v. 2.4.1) (Haas and Salzberg, 2008). The transdecoder software (v. 5.5.0) was used to predict gene structure based on ESTs evidence, and the credible gene structure annotation information was provided to train parameters for the following *de novo* gene prediction software packages: Augustus (v. 3.4.0) and GeneMark (v. 4.62). Finally, evidence from the gene finders, protein homology searches and ESTs were provided to the EvidenceModeler (v 1.1.1) (Haas and Salzberg, 2008) to obtain a comprehensive and non-redundant gene set.

For gene function annotation, we used Diamond (v. 2.0.6) to search the homologous sequences from the Swiss-Prot (http://www.uniprot.org/), TremBL (http://www.uniprot.org/) and NR

protein databases. We were also subjected to GO annotation and protein family annotation by InterProScan (v. 4.8) (https://www.ebi.ac.uk/interpro/). KO terms for each gene are assigned by an online website (KAAS, https://www.genome.jp/tools/kaas/). The programs tRNAScan-SE (v. 1.3.1) and RNAmmer (v. 1.2) were used to predict tRNA and rRNA, respectively. The other ncRNAs were identified by searching against the Rfam database (http://eggnogdb.embl.de/).

Internal scripts were used to calculate GC content, gene density, repetitive element density, ncRNA density, and gene components distribution based on gff3 format file in the annotation results.

The Completeness Assessment of Assembly and Annotation

Assembly completeness and accuracy were evaluated by multiple methods. First, the Illumina short reads were re-mapped to the genome using BWA (v. 0.7.17), and the mapping ratio was counted by samtools (v. 1.8; with the pattern "flagstat"). Then, we used the Benchmarking Universal Single Copy Orthologues (BUSCO) (v. 5.2.2) (Seppey et al., 2019) to test the integrality of the final assembly and the lineage dataset was actinopterygii_odb10. Similarly, the annotation integrity was evaluated based on the protein sequence sets using the protein pattern built into BUSCO software (v. 5.2.2).

Phylogenetic Analysis

Single-copy genes in *O. fasciatus* and 10 related species (*G. morhua*, *L. maculatus, L. crocea, G. aculeatus*, *N. coriiceps*, *T. bimaculatus*, *O. latipes*, *C. semilaevis*, *B. petinirostris*, and *D. reio*) were identified based on gene families constructed from protein sequences of all species employing OrthoFinder (v. 2.5.4) (Emms and Kelly, 2019) software with default parameters. Single-copy ortholog proteins were aligned by MUSCLE (v. 3.8.31). Subsequently, all obtained alignments were converted to their corresponding coding DNA sequences using an in-house python script. A combined continuous ultra-long sequence was constructed from all the translated coding DNA alignments for minimum evolution (ME) phylogenetic tree construction using MEGA. The divergence time is estimated using MCMCTREE (PAML package) (Yang, 1997) based on the molecular clock data of Timetree database (Hedges et al., 2006).

RESULTS AND DISCUSSIONS

With a 100x sequencing depth sequencing reads from Illumina and Pacbio platforms, we assembled a high-quality chromosome-level reference genome from a female *O. fasciatus*. Based on the 17-kmer analysis and a dominant peak depth of 51.57, the genome size was estimated to be 751.19 Mb with the heterozygosity and repetitive sequence content were approximately 0.28 and 23.09%, respectively (**Supplementary Table S1**). The genome size is slightly smaller than that of the *O. fasciatus* derived from the northern population (previously reported; male ~762 Mb and female ~768.8 Mb) (Xiao et al., 2019; Xiao et al., 2020). A trimodal pattern was observed in the 17-mer frequency distribution analysis. The main peak (the second peak) is much higher than the other two sub peaks, indicating that there is a certain degree of heterozygosity and duplication in *O. fasciatus* genome (**Supplementary Figure S1**) (Manekar and Sathe, 2019).

TABLE 1 | Summary of the *Oplegnathus fasciatus* genome assembly and annotation.

Genome assembly and chromosomes construction	
Contig N50 size (Mbp)	27.33
Contig number (>100 bp)	120
Contig total length (Mbp)	732.95
Scaffold N50 size (Mbp)	31.1
Scaffold number (>100 bp)	153
Scaffold total length (Mbp)	732.99
Number of chromosomes	24
Total length of chromosomes (Mbp)	723.22
Integration efficiency of Hi-C map (%)	98.67
Illumina Short reads re-mapping ratio (%)	97.89
Proportion of BUSCO in genome model (%)	97
Gene Prediction and Annotation	
Protein-coding gene number	27,015
Mean transcript length (bp)	3159.45
Mean exons length (bp)	167.47
Mean exons number per gene	9.27
Proportion of BUSCO in proteins model. (%)	95.6

The PacBio sequencing reads were used for *de novo* assembly of the genome. And the average read length and N50 of read length were 21,083 and 30,815 bp, respectively. The initial assembly yielded a total length of 1.1 Gb, comprising 2,789 contigs with a contig N50 length of 21.91 Mb. The genome assembly was larger than the estimated genome size of 751.19 Mb (see above), because some redundant sequences failed to be merged due to high heterozygosity (Zhang et al., 2012). After eliminating the redundancy, we obtained a final genome assembly of 732.95 Mb for the *O. fasciatus*, which is nearly equal to the estimated genome size (**Table1** and **Supplementary Table S1**). Then, the PacBio draft assembled contigs were anchored and oriented into a chromosomal-scale assembly *via* the Hi-C scaffolding approach (**Figure 1C**). Finally, we generated a chromosome-level genome assembly of 732.99 Mb in length, with a contig N50 and scaffold N50 value of 27.33 and 31.10 Mb, respectively (**Figure 1D** and **Table 1**). The total length of assembly contained 24 chromosomes (the lengths ranged from 18.22 to 37.18 Mb) was 723.22 Mb, and the integration efficiency was 98.67% (**Table 1** and **Supplementary Table S2**). The genome sizes of *O. fasciatus* (733.99 Mb) were similar with two closely related Perciformes species i.e., *L. crocea* (723.86 Mb) (Chen et al., 2019a) and *Oplegnathus punctatus* (718 Mb) (Li et al., 2021). To further verify the integrality of the assembled genome, the reads from the short-insert library were re-mapped onto the assembled genome using BWA (version 0.7.17). A total of 97.89% of the reads mapped to a reference sequence in this genome (**Table 1**). Additionally, we tested completeness by attempting the recovery of conserved single-copy genes from *O. fasciatus*

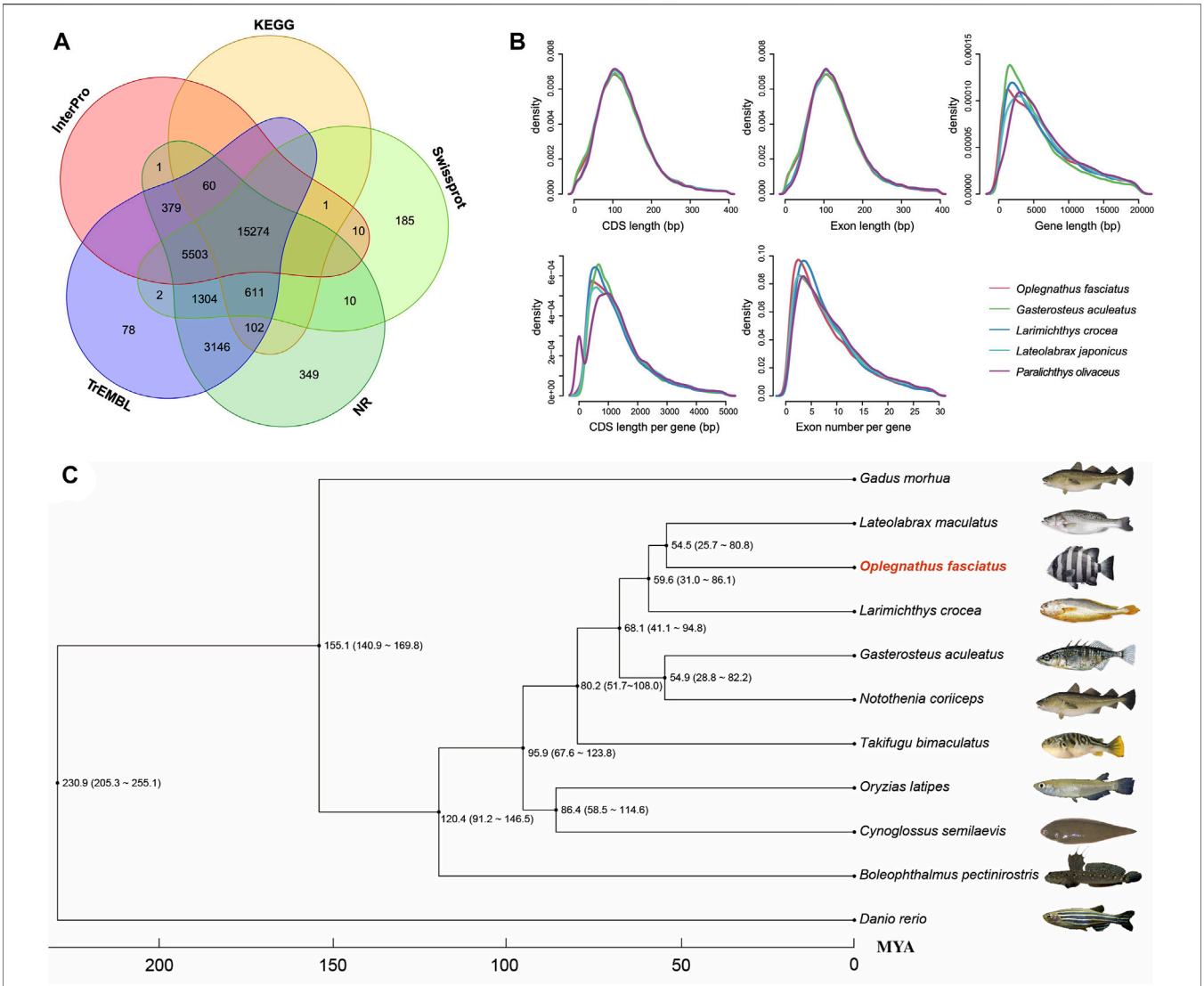

FIGURE 2 | Gene annotations of the Oplegnathus fasciatus genome and phylogenetic relationships among eleven species. **(A)** Venn diagram of the functionally annotated protein-coding genes. **(B)** Gene components distribution patterns among *O. fasciatus* and related species (*G. aculeatus*, *L. crocea*, *L. maculatus*, and *P. olivaceus*). **(C)** Phylogenetic analysis among *O. fasciatus* and representative species (*G. morhua*, *L. maculatus, L. crocea, G. aculeatus*, *N. coriiceps*, *T. bimaculatus*, *O. latipes*, *C. semilaevis*, *B. petinirostris*, and *D. reio*).

genome by BUSCO (v. 5.0.0). Out of a database containing 3,640 single-copy protozoan orthologs, ~97% were fully recovered from the assembly (**Table 1** and **Supplementary Table S3**). The high integration efficiency (~98.67%), mapping ratio (97.89%), and the recognition rate of single copy orthologs (~97%) show that our assembly of *O. fasciatus* is of high quality, which is at the same level as some recently published high-accurate reference genomes of other marine fish (Chen et al., 2019a; Zhou et al., 2019).

There was a total of 224.97 Mb of consensus and nonredundant repetitive sequences obtained by a combination of known, novel and tandem repeats, occupying more than 30.69% of the whole genome assembly (**Figure 1B** and **Supplementary Table S4**). The *de novo* and homologous prediction were utilized to investigate the repeat sequences (**Supplementary Table S5**). TE divergence analysis suggested recent activity of DNA transposons and long terminal repeats in this genome (**Figure 1B**) (Wang et al., 2019). DNA transposons were the most abundant repetitive elements, spanning at least 113.54 Mb, accounting for 15.49% of the whole genome of O. fasciatus. Among them, the repetitive sequences also comprised of long interspersed elements in 44.78 Mb (LINEs; 6.11%), short interspersed nuclear elements in 1.37 Mb (SINEs; 0.19%) and long terminal repeats in 34.02 Mb (LTRs; 4.64%) (**Supplementary Table S5**).

A total of 27,015 nonredundant protein-coding genes were successfully yielded combining *de novo*, homologous searching and transcriptome-assisted predictions (**Table 1**). The statistics of

the predicted gene models were compared to the homologous protein sequences of *G. aculeatus*, *L. crocea*, *L. japonicus* and *P. olivaceus*, which indicated closely distribution patterns in mRNA length, CDS length, exon length and exon number (**Figure 2B** and **Supplementary Table S7**). There were 17,054 complete ORFs, 27,015 complete transcripts and 250,306 exons detected in the *O. fasciatus* genome. The mean lengths of exon were 167.47 bp. The average exon number for each gene was 9.27 and the average length of CDS was 1551.68 bp (**Supplementary Table S7**). We annotated these genes against several public databases, including NR, TrEMBL, Swissprot and InterPro, resulting in 98.98, 97.94, 84.77 and 78.58% of the genes functionally assigned, respectively. Furtherly, we detected protein domains in multiple databases, and 59.76 and 59.40% of the predicted genes were annotated using GO and KEGG database, respectively. Finally, 27,015 genes were successfully functional annotated in at least one of these databases (**Figure 2A** and **Supplementary Table S6**). The number of predicted genes of *O. fasciatus* (27,015) through *de novo* prediction and homologue annotation was slightly more than others in Perciformes, such as *L. crocea* (23,657) (Chen et al., 2019a), *L. maculatus* (22,509) (Chen et al., 2019b), and previous version of *O. fasciatus* (24,003) from the northern population (Xiao et al., 2019). BUSCO analysis with protein pattern suggested that 97.2% (3,537) of the core conserved genes were detected in the *O. fasciatus* gene set, with 3,480 (95.6%) and 57 (1.6%) being identified as complete and fragmented, respectively (**Table 1** and **Supplementary Table S3**). The results indicate that our gene structure annotation is relatively complete (Seppey et al., 2019). To verify the accuracy of the contig arrangement in 24 chromosomes, we aligned 1 K bp small fragments with 50 K bp spacing as anchors of the assembled genome against the published *L. crocea* genome to compare consistency between these two genomes. The 24 chromosomes we identified in the *O. fasciatus* genome aligned exactly against the chromosomes of the *L. crocea*, suggesting high continuity with the *O. fasciatus* genome (**Figure 1E**). Furthermore, four types of non-coding RNAs were identified in *O. fasciatus* genome, including 1,188 miRNA, 1,808 tRNAs, 1,793 rRNAs and 1,091 snRNAs (**Table 1** and **Supplementary Table S8**).

To reveal the phylogenetic relationships among *O. fasciatus* and other species, a total of 2,516 single copy ortholog protein families in a 1:1:1 manner from the 10 related species (as described above) were identified and used for phylogenetic analysis (**Figure 2C**). Previous reports showed that *O. fasciatus* and *L. crocea* had close genetic distance and clustered in Perciformes (Eupercaria) (Xiao et al., 2019). According to our phylogenetic analysis, we further found that the divergence time between *O. fasciatus* and the common ancestor with *L. maculatus* (25.7-80.8 million years ago) was shorter than *L. crocea* (31.0–86.1 million years ago). In addition, the phylogenetic relationship between *O. fasciatus* and other fish is also consistent with previous taxonomic studies (Oh et al., 2007; Betancur-R et al., 2017), indicating that the estimation of divergence time in this study should be reasonable.

Certainly, a high-precision chromosomal genome resource is foremost to identify genetic variation underlying phenotypic traits of economic interest for aquaculture production (Chen et al., 2019a; Chen et al., 2019b; Zhou et al., 2019). The sex determination mechanism of *O. fasciatus* has attracted wide attention of researchers in recent years (Xiao et al., 2020). Previous cytogenetic analysis of *O. fasciatus* has shown that it has morphologically distinguishable sex chromosomes, and there are differences in the number of male (2n = 47; X1X2Y) and female (2n = 48; X1X1X2X2) chromosomes (Xu et al., 2019). Xiao et al. have reported the sex-related comparative genomics study of the northern population of *O. fasciatus* (Xiao et al., 2020), but it is still unclear whether the key genetic locus is restricted by geography. Our report also provides an important resource for further research on the mechanism of gender determination.

CONCLUSION

Here, we report a highly accurate chromosome-level genome assembly of *O. fasciatus* from the southern population in the subtropical region based on PacBio and Hi-C technologies. The genome size (732.99 Mb) is slightly smaller than that of the *O. fasciatus* derived from the northern population. A total of 27,015 gene structures were annotated using the strategy of multi-evidence combination. We found that the divergence time between *O. fasciatus* and the common ancestor with *L. maculatus* was shorter than *L. crocea*. The genome data created in this study will serve as valuable resources for species diversity and population genetic research, and will further promote the progress of genome-scale genetic breeding and genetic studies of its temperature plasticity and adaptation at lower latitudes.

CODE AVAILABILITY

Genome Survey and Assembly

1) SOAPec: version 2.01; -k 17 -l 11; 2) GenomeScope: version 2.0; all parameters were set as default; 3) Canu: version 2.0; all parameters were set as default; 4) Racon: version 1.4.3; all parameters were set as default; 5) NextPolish: version 1.4.0; job_type = local; task = 1212; rewrite = no; rerun = 3; sgs_options = -max_depth 100 -bwa; 6) Purge_Dups: version 1.2.5; all parameters were set as default.

Genome Annotation:

1) RepeatMasker: version 4.1.0; -no_is -nolow -norna -gff -poly -html; 2) RepeatModeler: version 2.0.1; -database genome -engine ncbi; 3) TEclass: version 2.1.3; all parameters were set as default; 4) TRF: version 4.09; 2 7 7 80 10 50, 500 -m -f -d; 5) Braker2: version 2.1.5; -gff3 –species --genome --prot_seq --bam --prg = gth --softmasking 6) Transdecoder: version 5.5.0; -t transcripts.fasta; 7) PASA: version 2.4.1; -c alignAssembly.config -C -R -g genome -t transcripts.fasta.clean -T -u transcripts.fasta --ALIGNERS blat, gmap; 8) Diamond: version 2.0.6; --outfmt 5; 9) EVidenceModeler: version 1.1.1; --gene_predictions --protein_alignments --transcript_alignments --segmentSize 100000 --overlapSize 10000 –weights weights.txt.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, PRJNA778612, https://figshare.com/, https://doi.org/10.6084/m9.figshare.16950832.

ETHICS STATEMENT

The animal study was reviewed and approved by the Animal Care and Use Committee at the College of Ocean and Earth Sciences, Xiamen University.

AUTHOR CONTRIBUTIONS

PX conceived the study. YB, JG, ZZ, BL, and QK performed bioinformatics analysis. YB, JG, JZ, and QK collected samples. YB and JG extracted DNA and RNA. FP, LW, and WZ performed the quality control. YB, TZ, and PX wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This study was supported by National Key R and D Program of China (2019YFE0119000), Special Foundation for Major Research Program of Fujian Province (2020NZ08003), Ningbo Science and Technology Innovation 2025 Major Project (2021Z002), the special project of local science and technology development guided by the central government (2019L3032), China Agriculture Research System (CARS-47) and Zhejiang Provincial Natural Science Foundation of China (LQ20C190008).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.811798/full#supplementary-material

REFERENCES

An, H. S., Kim, M. J., and Hong, S. W. (2008). Genetic Diversity of Rock Bream Oplegnathus fasciatus in Southern Korea. *Gene Genet.* 30, 451–459.

Bangera, R., Correa, K., Lhorente, J. P., Figueroa, R., and Yáñez, J. M. (2017). Genomic Predictions Can Accelerate Selection for Resistance against Piscirickettsia Salmonis in Atlantic salmon (Salmo salar). *Bmc Genomics* 18, 121. doi:10.1186/s12864-017-3487-y

Benevenuto, J., Ferrão, L. F. V., Amadeu, R. R., and Munoz, P. (2019). How Can a High-Quality Genome Assembly Help Plant Breeders? *GigaScience* 8, giz068. doi:10.1093/gigascience/giz068

Betancur-R, R., Wiley, E. O., Arratia, G., Acero, A., Bailly, N., Miya, M., et al. (2017). Phylogenetic Classification of Bony Fishes. *BMC Evol. Biol.* 17, 162. doi:10.1186/s12862-017-0958-3

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., and Borodovsky, M. (2021). BRAKER2: Automatic Eukaryotic Genome Annotation with GeneMark-Ep+ and AUGUSTUS Supported by a Protein Database. *NAR Genom Bioinform* 3, lqaa108. doi:10.1093/nargab/lqaa108

Chen, B., Li, Y., Peng, W., Zhou, Z., and Shi, Y. (2019a). Chromosome-Level Assembly of the Chinese Seabass (Lateolabrax Maculatus) Genome. *Front. Genet.* 10, 275. doi:10.3389/fgene.2019.00275

Chen, B., Zhou, Z., Ke, Q., Wu, Y., Bai, H., Pu, F., et al. (2019b). The Sequencing and De Novo Assembly of the Larimichthys Crocea Genome Using PacBio and Hi-C Technologies. *Sci. Data* 6, 188. doi:10.1038/s41597-019-0194-3

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., et al. (2017). De Novo assembly of the Aedes aegypti Genome Using Hi-C Yields Chromosome-Length Scaffolds. *Science* 356, 92–95. doi:10.1126/science.aal3327

Emms, D. M., and Kelly, S. (2019). OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics. *Genome Biol.* 20, 238. doi:10.1186/s13059-019-1832-y

Gong, J., Zhao, J., Ke, Q., Li, B., Zhou, Z., Wang, J., et al. (2021). First Genomic Prediction and Genome-wide Association for Complex Growth-related Traits in Rock Bream (Oplegnathus fasciatus). *Evol. Appl.* doi:10.1111/eva.13218

Haas, B. J., Salzberg, S. L., Zhu, W., and Pertea, M. (2008). Automated Eukaryotic Gene Structure Annotation Using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* 9, R7. doi:10.1186/gb-2008-9-1-r7

Hedges, S. B., Dudley, J., and Kumar, S (2006). TimeTree: a Public Knowledge-Base of Divergence Times Among Organisms. *Bioinformatics* 22, 2971–2972. doi:10.1093/bioinformatics/btl505

Jian, X., Chao, B., Kunci, C., Liu, G., Jiang, Y., Luo, Q., et al. (2017). Draft Genome of the Northern Snakehead, Channa argus. *GigaScience* 6, 1–5. doi:10.1093/gigascience/gix011

Li, M., Zhang, R., Fan, G., Xu, W., Zhou, Q., Wang, L., et al. (2021). Reconstruction of the Origin of a Neo-Y Sex Chromosome and its Evolution in the Spotted Knifejaw, Oplegnathus Punctatus. *Mol. Biol. Evol.* 38, 2615–2626. doi:10.1093/molbev/msab056

Manekar, S. C., and Sathe, S. R. (2019). Estimating the K-Mer Coverage Frequencies in Genomic Datasets: A Comparative Assessment of the State-Of-The-Art. *Curr. genomics* 20, 2–15. doi:10.2174/1389202919666181026101326

Oh, D. J., Kim, J. Y., Lee, J. A., Yoon, W. J., Park, S. Y., and Jung, Y. H. (2007). Complete Mitochondrial Genome of the Rock Bream Oplegnathus fasciatus (Perciformes, Oplegnathidae) with Phylogenetic Considerations. *Gene* 392, 174–180. doi:10.1016/j.gene.2006.12.007

Robinson, J. T., Turner, D., Durand, N. C., Thorvaldsdóttir, H., Mesirov, J. P., and Aiden, E. L. (2018). Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cel Syst.* 6, 256–258. doi:10.1016/j.cels.2018.01.001

Schemberger, M. O., Nascimento, V. D., Coan, R., Ramos, É., Nogaroto, V., Ziemniczak, K., et al. (2019). DNA Transposon Invasion and Microsatellite Accumulation Guide W Chromosome Differentiation in a Neotropical Fish Genome. *Chromosoma* 128, 547–560. doi:10.1007/s00412-019-00721-9

Seppey, M., Manni, M., and Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol. Biol.* 1962, 227–245. doi:10.1007/978-1-4939-9173-0_14

Sergey, K., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: Scalable and Accurate Long-Read Assembly via Adaptive K-Mer Weighting and Repeat Separation. *Genome Res.* 27, 722–736. doi:10.1101/gr.215087.116

Singh, K., Chhuneja, P., Gupta, O. P., Jindal, S., and Yadav, B. (2018). Shifting the Limits in Wheat Research and Breeding Using a Fully Annotated Refrence Genome. *Science* 361, 1–13. doi:10.1126/science.aar7191

Wang, X., Xu, W., Wei, L., Zhu, C., and Feng, C. (2019). Nanopore Sequencing and De Novo Assembly of a Black-Shelled Pacific Oyster (Crassostrea gigas) Genome. *Front. Genet.* 10, 1211. doi:10.3389/fgene.2019.01211

Xiao, Y., Li, J., Ren, G., Ma, D., Wang, Y., Xiao, Z. Z., et al. (2016a). Pronounced Population Genetic Differentiation in the Rock Bream Oplegnathus fasciatus Inferred from Mitochondrial DNA Sequences. *Mitochondrial Dna* 27, 1–8. doi:10.3109/19401736.2014.982553

Xiao, Y., Ma, D., Dai, M., Liu, Q., Xiao, Z., Li, J., et al. (2016b). The Impact of Yangtze River Discharge on the Genetic Structure of a Population of the Rock

Bream, Oplegnathus fasciatus. *Mar. Biol. Res.* 12, 1–9. doi:10.1080/17451000.2016.1154576

Xiao, Y., Xiao, Z., Ma, D., and Liu, J. (2019). Genome Sequence of the Barred Knifejaw Oplegnathus fasciatus (Temminck & Schlegel, 1844): the First Chromosome-Level Draft Genome in the Family Oplegnathidae. *GigaScience* 8, giz013. doi:10.1093/gigascience/giz013

Xiao, Y., Xiao, Z., Ma, D., Zhao, C., and Herrera-Ulloa, A. (2020). Chromosome-Level Genome Reveals the Origin of Neo-Y Chromosome in the Male Barred Knifejaw Oplegnathus fasciatus. *iScience* 23, 101039. doi:10.1016/j.isci.2020.101039

Xu, D., Sember, A., Zhu, Q., Oliveira, E. A., Liehr, T., Abh, A. R., et al. (2019). Deciphering the Origin and Evolution of the X1X2Y System in Two Closely-Related Oplegnathus Species (Oplegnathidae and Centrarchiformes). *Int. J. Mol. Sci.* 20, 3571. doi:10.3390/ijms20143571

Yang, Z. (1997). PAML: a Program Package for Phylogenetic Analysis by Maximum Likelihood. *Computer Appl. Biosciences Cabios* 13, 555. doi:10.1093/bioinformatics/13.5.555

Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., et al. (2012). The Oyster Genome Reveals Stress Adaptation and Complexity of Shell Formation. *Nature* 490, 49–54. doi:10.1038/nature11413

Zhao, J., Bai, H., Ke, Q., Li, B., Zhou, Z., Wang, H., et al. (2021). Genomic Selection for Parasitic Ciliate Cryptocaryon Irritans Resistance in Large Yellow Croaker. *Aquaculture* 531, 735786. doi:10.1016/j.aquaculture.2020.735786

Zhou, Z., Liu, B., Chen, B., Shi, Y., and Xu, P. (2019). The Sequence and De Novo Assembly of Takifugu Bimaculatus Genome Using PacBio and Hi-C Technologies. *Scientific Data* 6, 187. doi:10.1038/s41597-019-0195-2

Conflict of Interest: WZ and QK were employed by the Ningde Fufa Fisheries Company Limited.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

DATA REPORT
published: 12 January 2022
doi: 10.3389/fgene.2021.808366

The Draft Genome of *Cryptocaryon irritans* Provides Preliminary Insights on the Phylogeny of Ciliates

*Yulin Bai[1,2†], Zhixiong Zhou[2†], Ji Zhao[2], Qiaozhen Ke[1,2], Fei Pu[2], Linni Wu[2], Weiqiang Zheng[2], Hongshu Chi[3], Hui Gong[3], Tao Zhou[1,2,4] and Peng Xu[1,2,4]**

[1]State Key Laboratory of Large Yellow Croaker Breeding, Ningde Fufa Fisheries Company Limited, Ningde, China, [2]State Key Laboratory of Marine Environmental Science, College of Ocean and Earth Sciences, Xiamen University, Xiamen, China, [3]Biotechnology Institute, Fujian Academy of Agricultural Sciences, Fuzhou, China, [4]Fujian Key Laboratory of Genetics and Breeding of Marine Organisms, College of Ocean and Earth Sciences, Xiamen University, Xiamen, China

Keywords: cryptocaryon irritans, oxford nanopore technologies (ONT), illumina, draft genome, phylogenomic analysis

OPEN ACCESS

Edited by:
Liang Guo,
South China Sea Fisheries Research Institute, China

Reviewed by:
Yan He,
Ocean University of China, China
Shaolin Wang,
China Agricultural University, China
Qianqian Zhang,
Yantai Institute of Coastal Zone Research (CAS), China

****Correspondence:***
Peng Xu
xupeng77@xmu.edu.cn

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *03 November 2021*
Accepted: *13 December 2021*
Published: *12 January 2022*

Citation:
Bai Y, Zhou Z, Zhao J, Ke Q, Pu F, Wu L, Zheng W, Chi H, Gong H, Zhou T and Xu P (2022) The Draft Genome of Cryptocaryon irritans *Provides Preliminary Insights on the Phylogeny of Ciliates. Front. Genet. 12:808366. doi: 10.3389/fgene.2021.808366*

INTRODUCTION

Ciliates are one of the most diverse, highly differentiated and ancient groups of microbial eukaryotes (Coyne et al., 2011). The characteristics of nuclear dimorphism (one large macronucleus and one small micronucleus) make ciliates considered to be an excellent model organism in the genetic investigation (Juranek and Lipps, 2007). Some major diseases in marine fish are caused by ciliates, which can cause skin damage, bacterial infection, and even death of the host (Wei et al., 2018; Zhao et al., 2021b). Genomic research on these pathogens is crucial to finding new treatments. It is particularly attractive to identify genes involved in unique metabolic pathways, pathogenicity, and parasite evasion of immune defense mechanisms (Wei et al., 2018).

The ciliated protozoan *Cryptocaryon irritans* is an obligate ectoparasite of marine fish, and its phylogenetic classification has always been controversial (Wright and Colorni, 2002). Due to the high affinity of the morphological characteristics, life cycle and pathogenic mechanism with *I. multifiliis*, *C. irritans* was originally included in the class Oligohymenophorea and also named "Marine Ich" (Corliss, 1979). After comparing the partial 18s rRNA sequence, Wright and Colorni (2002) indicated that *C. irritans* is taxonomically distinct from *I. multifiliis* and justify *C. irritans*' inclusion into the order Prorodontida within the Class Prostomatea (Wright and Colorni, 2002). Parasites usually have relatively complex phylogenetic relationships, and the lack of research on genetic tools such as genomes is the main reason that hinders the development of related biological problems (Ajioka et al., 1998).

Cryptocaryonosis, caused by *C. irritans*, has an extremely wide host range and is responsible for large-scale death of natural populations (Bai et al., 2020), which is one of the most important parasitological problems in marine aquaculture and poses a significant threat to the aquaculture industry (Zhao et al., 2021a). Several strategies for cryptocaryonosis control have been reported, such as antibiotics, vaccines and metal ions, however, they have shown only partial efficacy under field conditions (Yin et al., 2016; Yogeswaran et al., 2010). In addition, the lack of genomic data has hampered the use of molecular tools in developing control strategies for cryptocaryonosis (Kumar et al., 2020). Transcriptome projects have provided partial sequences of many protein-coding genes

Abbreviations: BUSCO, benchmarking universal single-copy orthologues; C. irritans, Cryptocaryon irritans; EST, expressed sequence tag; I. multifiliis, Ichthyophthirius multifiliis; L. crocea, Larimichthys crocea; ONT, oxford nanopore technologies; P. falciparum, Plasmodium falciparum; P. persalinus, Pseudocohnilembus persalinus; P. tetraurelia, Paramecium tetraurelia; RNA-seq, RNA sequencing; S. lemnae, Stylonychia lemnae; T. thermophila, Tetrahymena thermophila; Tris-EDTA, Tris–ethylenediaminetetraacetic acid.

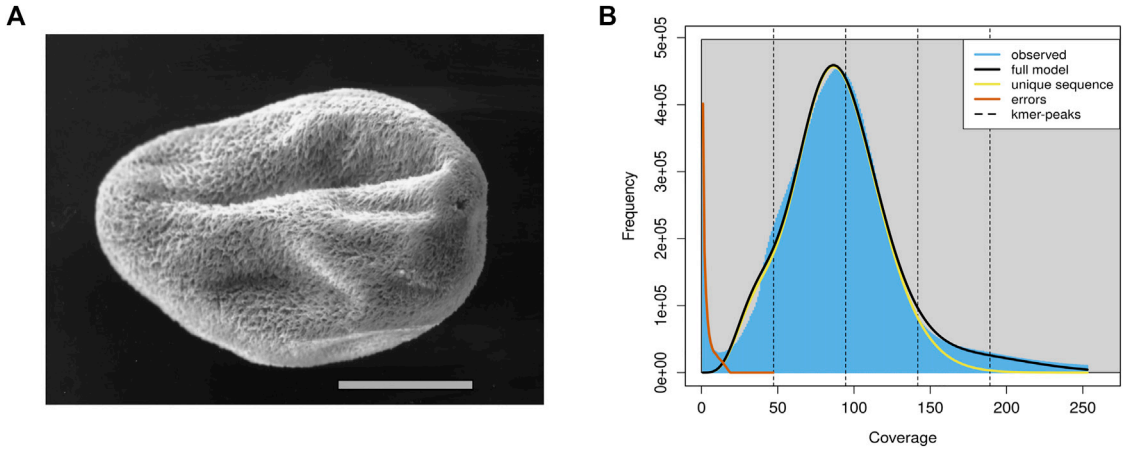

FIGURE 1 | The schematic diagram and genomics feature of the *C. irritans*. **(A)** Pro-tomont of the *C. irritans*. Scale bar 100 μm [adopted from (Bresciani and Buchmann (2001)]. **(B)** A K-mer analysis of the genome sequencing reads for the *C. irritans* using GenomeScope.

(Chi et al., 2020; Yogeswaran et al., 2010), but it is not sufficient to support the physiological metabolism and pathogenic mechanism of *C. irritans*. Therefore, full genome sequence is necessary to perform such analyses.

The parasitic lifestyle, bacterial contamination and other environmental factors always result in a complex sample background, which contributes to contamination in DNA preps (Pan et al., 2019). The limitations on the ability to extract quality DNA with sufficient yields for high-throughput library construction, especially considering the loss of DNA associated extraction and purification step, has likely been the greatest barrier to non-model ciliate genome research (Miller et al., 1999). With the popularization of genome sequencing technology, the genome sequences of the known hosts and closely related co-living species have been sequenced and thoroughly annotated (Chen et al., 2019b; Mao et al., 2013), we have been able to assemble and explore a substantial portion of the genome of *C. irritans*.

In this report, we provided a draft genome of *C. irritans* using Oxford Nanopore Technologies (ONT). We assembled the genome sequences into 2,384 contigs with a total length of 45.61 Mb and a contig N50 length of 21.24 Kb. Furthermore, we identified 4.02 Mb (8.81% of the assembly) of repeat content, 8,729 protein-coding genes and 490 ncRNAs. The first *C. irritans* genome is essential to support the basic genetics and molecular mechanisms studies, and will also provide an important resource for the analysis of host-parasite interaction mechanism.

MATERIALS AND METHODS

Sample Collection and DNA Extraction

C. irritans was isolated from an infected *L. crocea* and propagated by passage on juvenile *L. crocea*. Tomonts (**Figure 1A**) were collected from the bottom of the experimental tank and incubated overnight at room temperature (25–28°C). Theronts from a single beaker were then used to infect a large yellow croaker and offspring from the infection were subsequently maintained by serial passage on fish. Then, all the fish were euthanized by using tricaine methanesulfonate (MS-222; Sigma, St. Louis, MO, United States), and the tomonts were collected, snap frozen in the liquid nitrogen and stored at −80°C.

The harvested cells were washed by low-speed centrifugation through a 0.25 M sucrose solution, then homogenized using a pestle for a 1.5 ml microcentrifuge tube in 0.25 ml of lysis buffer (10 mM Tris; 0.5 M EDTA; 1% SDS; pH 9.5). Proteinase K (0.2 mg/ml) and RNase (40 μg/ml) were then added to digestion. After phenol/chloroform extraction, DNA was dialyzed against Tris–ethylenediaminetetraacetic acid (Tris-EDTA) followed by ethanol precipitation. Nucleic acid concentrations were quantified using a Qubit fluorometer (Thermo Fisher Scientific, Waltham, MA), and then checked by 1.5% agarose gel electrophoresis stained for integrity.

Library Construction and Sequencing

Paired-end sequencing library with a 350 bp insert size was constructed according to the manufacturer's instructions. The library was then sequenced with a paired-end sequencing strategy using the Illumina HiSeq 2,500 platform, and the read length was 2 × 150 bp.

Nanopore sequencing library construction and sequencing were conducted according to the manufacturer's protocol with the Oxford Nanopore MinION platform at Novogene (Tianjin).

Data Filtering and Genome Survey

Before assembly, the Nanopore data was filtered using NanoFilt (v. 2.8.0) and NanoPlot (v. 1.33.0) software, and the reads with length less than 2000bp or mean quality score less than 10 were removed. For the Illumina data, adapter sequences and low-quality reads were filtered using fastp (v. 0.23.1) software. The remaining reads were used for further assembly and estimation of genome size using the K-mer analysis of the short reads.

In order to obtain information such as genome size, heterozygosity, and repeatability, SOAPec (v. 2.01) and

GenomeScope (v. 2.0) softwares were used to analyze the K-mer frequencies in the sequencing reads to efficiently estimate the major genome characteristics.

DNA Contamination Filtration and *de novo* Assembly of the *C. irritans* Genome

To improve continuity and accuracy of protozoan assembly, we filtered the data in multiple ways and performed hybrid assembly (Florencia et al., 2019). First, nanopore data were mapped to the genome sequence including common marine bacteria and *L. crocea* to filter out the DNA contamination of symbiotic bacteria and the host. In order to correct the over-filtered data, Nanofilt (v. 2.8.0) and Minimap (v. 2.17) sorftwares were used to retain the reads containing GC ratios less than 30% or accurately mapped to the genomes of closely homologous species were retained (Coyne et al., 2011). Then, the filtered nanopore data were assembled into contigs by Flye (v. 2.9) with the parameter "--nano-raw". After that, the preliminarily assembled contigs were polished by NextPolish (v. 1.4.0) and Racon (v. 1.4.3) software to correct base errors caused by Nanopore sequencing. Finally, we employed Purge_Dups (v. 1.2.5) to resolve redundancy in the assembly.

The Benchmarking Universal Single-Copy Orthologues (BUSCO) software (v. 5.0.0) was used to evaluate the completeness of assembly with the alveolata_odb10 database.

Annotation of Genomic Repeats

A combination of *de novo* and homology-based predictions were used to identify repeat sequences in the *C. irritans* genome. Firstly, RepeatModeler (v. 2.0.1) and LTR_Finder (v. 1.07) were used to detect repeat sequences in the *C. irritan* genome. Combined with Repbase (20181026), a repeat sequence library was constructed. Then, we used RepeatMasker (v 4.1.0; setting "-nolow -norna -no_is" parameters) to detect and classify repeats based on this library. TEclass (v. 2.1.3) was used to further annotate unclassified repeats. TRF (v. 4.09) was used to identify tandem repeats. Before gene prediction, all regions of repetitive elements were soft-masked with RepeatMasker (v. 4.1.1).

Protein-Coding Gene Finding and Function Annotation

Both homology-based, *de novo* and RNA-seq strategies were used for gene structure prediction of the *C. irritans* genome. As in the case of other ciliates, *C. irritans* translates the TAA and TAG as glutamine instead of termination codons (Hatanaka et al., 2007), so we adjusted some gene structure annotation software parameters. For homology-based annotation, the protein sequences of three closely homologous species (*T. thermophila, I. multifiliis* and *P. tetraurelia*) were downloaded from NCBI and provided to the exonerate software (v. 2.4.0; setting "--geneticcode 6" parameter) to obtain an accurate gene structure for each locus. To train gene finding algorithms, a set of complete gene structures was modeled manually using the Illumina EST alignments to predict genes of *C. irritans* genome. Then, this set was used to train the *de novo* prediction software Augustus (v. 3.4.0) to predict the gene structure based on the repeat-masked genome. The latest RNA-seq data of *C. irritans* were downloaded from NCBI (SRA accession number: PRJNA600221) and mapped to *C. irritans* genome using PASA (v. 2.4.1; setting "--GENETIC_CODE Tetrahymena" parameter) and Stringtie (v. 2.1.4). The transdecoder software (v. 5.5.0; setting "-G Tetrahymena" parameter) was used to predict gene structure based on ESTs evidence. Finally, evidence from the gene finders, protein homology searches and ESTs were used to refine gene models using EvidenceModeler (v 1.1.1; with the "--stop_codons TGA" parameter). After extensive manual annotation of selected genes, a comprehensive and non-redundant gene set was generated.

For gene function annotation, we used Diamond (v. 2.0.6) to align the candidate sequences to the Swiss-Prot, TremBL and NR protein databases with e-values < 1E-5. InterProScan (v. 5.52–86.0) software was used for GO annotation and protein family annotation. KO terms for each gene are assigned by an online website (KAAS, https://www.genome.jp/tools/kaas/).

The programs tRNAScan (v. 2.0) and RNAmmer (v. 1.2) were used to predict tRNA and rRNA, respectively, and other ncRNAs were identified by searching against the Rfam database (http://eggnogdb.embl.de/).

Gene Components Distribution and Phylogenetic Analysis of *C. irritans*

To reveal the phylogenetic relationships and gene components distribution patterns among *C. irritans* and other species, we download the protein sequence of *P. falciparum* (outgroup), *P. persalinus*, and *S. lemnae* in addition to *T. thermophila, I. multifiliis and P. tetraurelia*. These genomes were annotated using the same pipeline, and protein sequences shorter than 50 amino acids were removed. Then, in-house scripts are used to count and plot the gene components. OrthoMCL (v. 6.6) and Diamond (v. 2.0.6) software were used to construct gene families from protein sequences of all species, and single-copy genes are identified based on the gene families. Single-copy ortholog proteins were aligned by MUSCLE (v. 3.8.31). Finally, we combined all the translated coding DNA sequences of each species into a continuous ultra-long sequence and used RAxML (version 8.2.12) software to generate a phylogenetic tree.

RESULTS AND DISCUSSIONS

In total, approximately 8.84 Gb raw illumina data and 16.45 Gb Nanopore reads were generated. After quality filtering, 8.82 Gb clean Illumina data and 12.5 Gb of trimmed Nanopore reads with a mean read length of 8.5 Kb were obtained. For the genome survey analysis, the number of 17-mer was 44,364,461,437, K_depth was estimated as 92.6, GC content and heterozygosity were about 26.86 and 0.5%, respectively. And the corrected genome size is estimated to be 45.67 Mb (**Supplementary Table S1**), which is similar to the *I. multifiliis* genome size

TABLE 1 | Summary statistics of genome assemblies of *C. irritans*.

Summary statistics of genome assembly	
Total length of genome (Mbp)	45.61
Contig N50 size (Kbp)	21.24
Contig N75 size (Kbp)	14.48
Contig L50 size (Kbp)	689
Contig L50 size (Kbp)	1,379
Contig number (>1,000 bp)	2,384
Contig number (>10000 bp)	2,355
Max contig length (Kbp)	748.71
Gene Prediction and annotation	
Protein-coding gene number	8,729
Mean transcript length (bp)	1,635.71
Mean exons length (bp)	269.56
Mean exons number per gene	6.05

(Coyne et al., 2011). A common single-peak pattern was observed in the K-mer frequency distribution analysis, indicating that the genome may have a low level of heterozygosity and repetitive regions (**Figure 1B**).

Affected by contamination and heterozygous region, the initial assembly result is larger than the estimated genome size of 45.67 Mb (see above). Manually adjusting the genome may be the most effective way to eliminate contaminants such as bacterial symbionts (including Pseudomonas and Vibrio) and fish hosts in the current assembly (Coyne et al., 2011). After eliminating the redundancy, we obtained a final genome assembly of 45.61 Mb for the *C. irritans*, which is nearly equal to the estimated genome size (**Table 1**). The draft assembly consisted of 2,384 contigs, and the contig N50 value of our final assembly was 21.24 kb. The summary statistics of *C. irritans* genome assembly compared with other ciliates can be obtained in **Supplementary Table S2**.

Repeat sequences of the *C. irritans* genome were calculated by the combination of both homolog-based and *de novo* methods. There was a total of 4.02 Mb of consensus and nonredundant repetitive sequences obtained by a combination of known, novel and tandem repeats, occupying more than 8.81% of the genome assembly (**Supplementary Table S3**). DNA transposons accounted for the largest proportion of transposable elements,

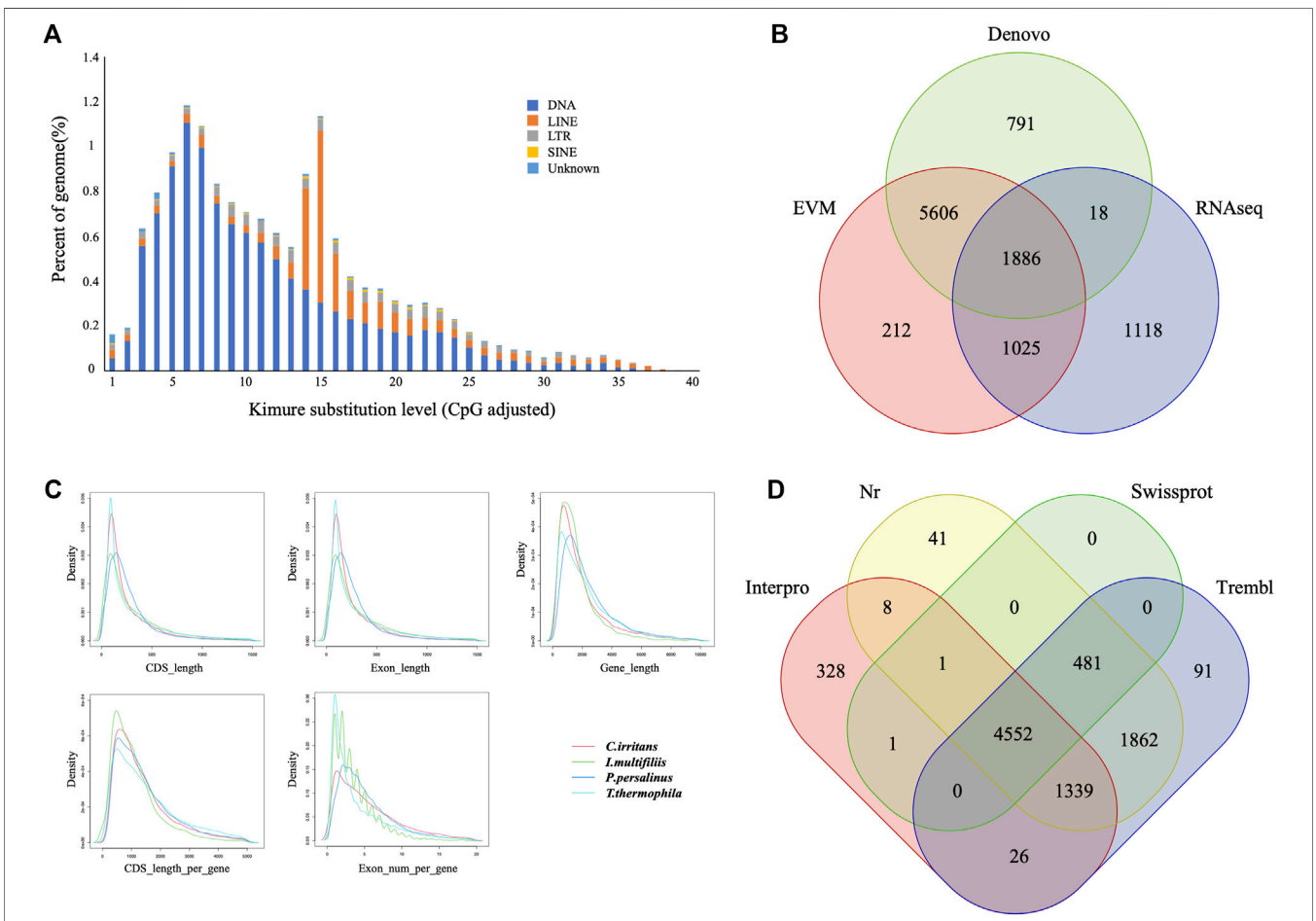

FIGURE 2 | Gene and repetitive element annotations of the *C. irritans* genome. **(A)** Divergence distribution of TEs in the *C. irritans* genome. **(B)** Venn diagram of the number of genes with structure prediction based on different strategies. **(C)** Gene components distribution patterns among *C. irritans* and related species (*I. multifilii*, *P. persalinus*, *T. thermophila*). **(D)** Venn diagram of the number of functionally annotated genes based on different public databases.

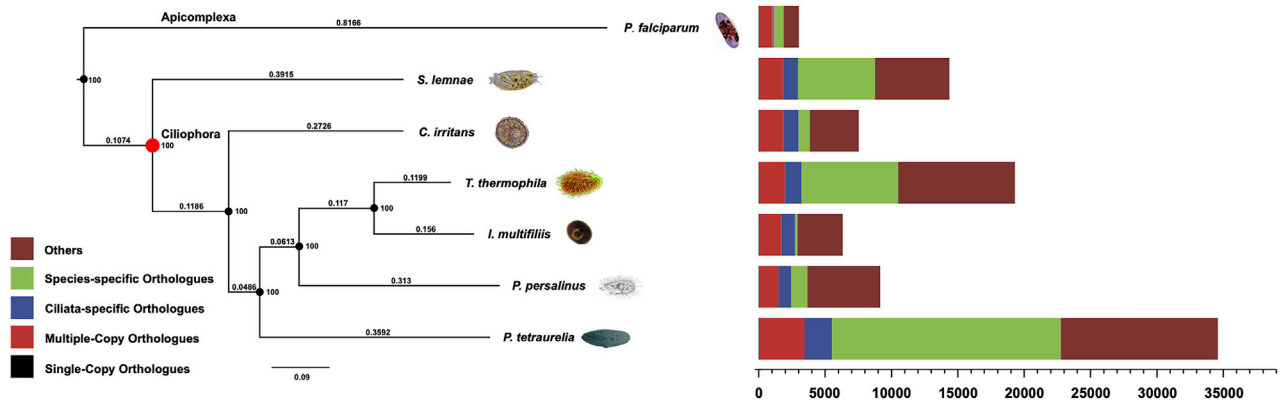

FIGURE 3 | Phylogenetic analysis and distribution of different types of orthologs in representative species (*C. irritans*, *P. falciparum*, *P. persalinus*, *S. lemnae*, *T. thermophila*, *I. multifiliis* and *P. tetraurelia*). The bootstrap value of all branch nodes indicated the robustness of the phylogenetic tree. The numbers above the branches meaning the sequence difference, which were the product of nucleotide mutation rate and divergence time between species.

spanning at least 2.51 Mb accounting for 5.51% of the whole genome. The repetitive sequences are also comprised of long interspersed elements (LINEs; 1.86%), short interspersed nuclear elements (SINEs; 0.10%) and long terminal repeats (LTRs; 0.70%) (**Supplementary Table S4**). TE accumulation analysis suggested long-term TE actives (**Figure 2A**). Furthermore, we identified four types of non-coding RNA, 154 miRNAs, 183 tRNAs, 96 rRNAs, and 57 snRNAs in the assembled genome (**Supplementary Table S5**).

The early-stage gene predictions of the protozoan genome have been largely based on sequence homology (Wang et al., 2003). Gene predictions in newly sequenced species based on the availability of predictions from related species have been used for genome annotation of several protozoa (Mushegian et al., 1998). In order to show sufficiently novel features, new algorithms and strategies need to be developed (Ersfeld, 2003). A total of 8,729 non-redundant protein-coding genes were successfully predicted combining *de novo*, homology searching and transcriptome-assisted predictions in this genome (**Figure 2B**), with an average gene length of 2.21 kb. The statistics of the gene components were compared with closely homologous species, and the result indicated close distribution patterns in mRNA length, CDS length, exon length and exon number (**Figure 2C** and **Supplementary Table S6**). Respectively, 5,034, 8,283, 8,351, and 6,254 genes showed positive hits in Swissprot, NR, TrEMBL and Interpro (**Figure 2D** and **Supplementary Table S7**). In order to verify the integrity of the *C. irritans* genome assembly and annotation, we downloaded the *C. irritans* transcriptome sequence published by Lokanathan et al. (2010). BWA (version 0.7.17) and Samtools (v. 1.8) were used to eliminate host contamination and calculate the mapping ratio. As a result, a total of 88.52% of the reads were mapped to this genome. Additionally, we tested completeness by attempting the recovery of conserved single-copy genes from *C. irritans* genome by BUSCO (v. 5.0.0). Out of a database containing 171 single-copy protozoan orthologs, ~71.4% were fully recovered from the assembly. Similarly, we also tested the published *I. multifiliis* genome (Coyne et al., 2011), and about 82.5% was completely recovered from the assembly (**Supplementary Table S8**). The test of such conserved single-copy genes in protozoa is inconclusive, which might indicate that some genes are not as conserved in ciliates as they are in vertebrates. The percentage might reflect the evolutionary divergence of the ciliate similar to what has been reported for another protozoon (Porcel et al., 2000). It is necessary to develop algorithms and strategies which are more suitable for the evaluation of protozoan genome integrity.

The systematic position of *C. irritan* has long puzzled taxonomists, and their assignment to the class Oligohymenophorea has been controversial (Lasek-Nesselquist and Johnson, 2019). In order to reveal the phylogenetic relationships, the evolutionary position of *C. irritans* was accessed based on single-copy genes of *C. irritans* and six related species (*P. falciparum*, *P. persalinus*, *S. lemnae, T. thermophila, I. multifiliis* and *P. tetraurelia*). As a result, a total of 15,583 Orthogroups were built and 63 single-copy orthologs in a 1:1:1 manner from all seven protozoa species were used for phylogenetic analysis (**Supplementary Table S9**). *C. irritans* shared fewer orthologous genes with *P. falciparum* (1,186) than with *I. multifiliis* (2,131) (**Supplementary Table S10**), which is consistent with its closer morphological resemblance to the latter (Wright and Colorni, 2002). RAxML analyses showed a clear topology in the concatenated tree, that is, with two main evolution nodes are recognizable. *P. falciparum* was used as outgroups, suggesting that it is separated from other species at the class level. The six species, *C. irritans* (Prostomatea), *S. lemnae* (Spirotrichea) and other Oligohymenophorea species which were believed to be members of the phylum Ciliata, were clustered and placed in separate clades (**Figure 3**). *C. irritans* occupied the basal position within the class, indicating that this species might be an ideal representative to demonstrate the ancestral candidate of the Ciliata (Pan et al., 2019). This is consistent with findings of previous studies based on 18s rRNA sequencing which inferred that *C. irritans* is taxonomically different from *I. multifiliis* (Wright and Colorni, 2002). And the similarity in morphology

and development between them may be caused by convergent evolution (Hülsmann and Hausmann, 1994; Zhang et al., 2014).

Inferring phylogenetic relationships based on single genes has certain limitations, and was gradually replaced by other more abundant phylogenetic evidence (Ludwig and Klenk, 2001). With the development of sequencing technology, phylogenetic analysis using whole-genome genetic evidence has been more recognized by researchers (Wu and Eisen, 2008). Similarly, the "Ultra Sequence" was constructed from all the single-copy orthologous genes for phylogenetic tree construction, avoiding many limitations of small data sets (Chen et al., 2019a; Zhou et al., 2019). In our analysis, 63 single-copy orthologous gene sequences of 7 species were used to construct the phylogenomic tree. Therefore, a more accurate estimation of evolutionary relationships could be obtained. However, the current analysis of protozoan genetics is still limited, and expansion of genomic resources is necessary to support future research.

CONCLUSION

Here, we used a combination of Illumina and Oxford Nanopore reads to provide the draft genome assembly of *C. irritans*. A total of 8,729 gene structures were annotated using the strategy of multi-evidence combination. The comparative analysis of the gene components distribution showed that *C. irritans* and other closely homologous species have similar distribution patterns. The phylogenetic tree was constructed to illuminate the relationship of the *C. irritans* and six other protozoa. Meanwhile, we demonstrate that Oxford Nanopore can be a very valuable technology to analyze protozoan genomes. The data generated in this study will contribute to facilitate the enlargement of genomic resources for protozoan pathogens, and provide valuable resources for the study of basic genetics and the pathogenic mechanism of parasites.

CODE AVAILABILITY

The versions, settings and parameters of the software used in this work are as follows:

Genome survey and assembly:

(1) SOAPec: version 2.01; -k 17 -l 52; (2) GenomeScope: version 2.0; all parameters were set as default; (3) NanoFilt: version 2.8.0; -q 9 -l 12000 --headcrop 50 --tailcrop 50; (4) NanoPlot: version 1.33.0; --maxlength 40000 --loglength --plots hex dot pauvre kde; (5) Flye: version 2.9; all parameters were set as default; (6) Racon: version 1.4.3; all parameters were set as default; (7) NextPolish: version 1.4.0; job_type = local; task = 1212; rewrite = no; rerun = 3; sgs_options = -max_depth 100 -bwa; (8) Purge_Dups: version 1.2.5; all parameters were set as default.

Genome annotation:

(1) RepeatMasker: version 4.1.0; -no_is -nolow -norna -gff -poly -html; (2) RepeatModeler: version 2.0.1; -database genome -engine ncbi; (3) TEclass: version 2.1.3; all parameters were set as default; (4) TRF: version 4.09; 2 7 7 80 10 50 500 -m -f -d; (5) Augustus: version 3.4.0; --uniqueGeneId=true --noInFrameStop=true --gff3=on --strand=both; (6) exonerate: version 2.4.0; --model protein2genome --querytype protein --targettype dna --showvulgar no --softmaskquery yes --softmasktarget yes --minintron 20 --maxintron 10000 --showalignment no --showtargetgff yes --showcigar no --geneseed 250 --score 250 --bestn 0 --verbose 0 --geneticcode 6; (7) Transdecoder: version 5.5.0; -t transcripts.fasta -G Tetrahymena; (8) PASA: version 2.4.1; -c alignAssembly.config -C -R -g genome -t transcripts.fasta.clean -T -u transcripts.fasta --ALIGNERS blat,gmap --GENETIC_CODE Tetrahymena; (9) Diamond: version 2.0.6; --query-gencode 6 --outfmt 5; (10) EVidenceModeler: version 1.1.1; --gene_predictions --protein_alignments --transcript_alignments --segmentSize 100000 --overlapSize 10000 –weights weights.txt --stop_codons TGA.

Phylogenetic analysis:

(1) OrthoMCL: version 6.6; all parameters were set as default; (2) MUSCLE: version 3.8.31; parameters: all parameters were set as default; (3) RAxML: version: 8.2.12; parameters: -n sp -m PROTGAMMAAUTO -T 20 -f a.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://figshare.com/, https://doi.org/10.6084/m9.figshare.16922665.v2; https://www.ncbi.nlm.nih.gov/, SRX12890364; https://www.ncbi.nlm.nih.gov/, SRX12890363.

ETHICS STATEMENT

The animal study was reviewed and approved by The Animal Care and Use Committee at the College of Ocean and Earth Sciences, Xiamen University.

AUTHOR CONTRIBUTIONS

PX conceived the study. YB and ZZ performed bioinformatics analysis. YB, JZ, QK, FP, LW, and WZ collected samples. YB, HC, and HG extracted DNA and RNA. YB and TZ performed the quality control. YB and PX wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China (U21A20264); the Special Foundation for Major Research Program of Fujian Province (2020NZ08003); the Industry-University Collaboration Project of Fujian Province (2019N5001); the Open Research Fund Project of

State Key Laboratory of Large Yellow Croaker Breeding (LYC2021RS02); the Local Science and Technology Development Project Guide by The Central Government (2019L3032); the Science and Technology Platform construction of Fujian Province(No.2018N2005); the China Agriculture Research System (CARS-47); the special project of local science and technology development guided by the central government (2019L3032); the Ningbo Science and technology innovation 2025 major special project (2021Z002); the Fundamental Research Funds for the Central Universities (20720200110).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.808366/full#supplementary-material

REFERENCES

Ajioka, J. W., Boothroyd, J. C., Brunk, B. P., Hehl, A., Hillier, L., Manger, I. D., et al. (1998). Gene Discovery by EST Sequencing in Toxoplasma gondiiReveals Sequences Restricted to the Apicomplexa. *Genome Res.* 8, 18–28. doi:10.1101/gr.8.1.18

Bai, H., Zhou, T., Zhao, J., Chen, B., Pu, F., Bai, Y., et al. (2020). Transcriptome Analysis Reveals the Temporal Gene Expression Patterns in Skin of Large Yellow Croaker (Larimichthys Crocea) in Response to Cryptocaryon Irritans Infection. *Fish. Shellfish Immunol.* 99, 462–472. doi:10.1016/j.fsi.2020.02.024

Bresciani, J., and Buchmann, K. (2001). *Parasitic Diseases of Freshwater trout*. Denmark: DSR Publishers Frederiksberg, 1–76.

Chen, B., Li, Y., Peng, W., Zhou, Z., Shi, Y., Pu, F., et al. (2019a). Chromosome-Level Assembly of the Chinese Seabass (Lateolabrax Maculatus) Genome. *Front. Genet.* 10, 275. doi:10.3389/fgene.2019.00275

Chen, B., Zhou, Z., Ke, Q., Wu, Y., Bai, H., Pu, F., et al. (2019b). The Sequencing and De Novo Assembly of the Larimichthys Crocea Genome Using PacBio and Hi-C Technologies. *Sci. Data* 6, 188. doi:10.1038/s41597-019-0194-3

Chi, H., Goldstein, M., Pichardo, A., Wei, Z. H., Chang, W. J., and Gong, H. (2020). Infectivity and Genes Differentially Expressed between Young and Aging Theront Cells of the marine Fish Parasite Cryptocaryon Irritans. *PLoS One* 15, e0238167. doi:10.1371/journal.pone.0238167

Corliss, J. O. (1979). The Ciliated Protozoa: Characterization, Classification, and Guide to the Literature. *Trans. Am. Microsc. Soc.* 98, 03

Coyne, R. S., Hannick, L., Shanmugam, D., Hostetler, J. B., Brami, D., Joardar, V. S., et al. (2011). Comparative Genomics of the Pathogenic Ciliate Ichthyophthirius Multifiliis, its Free-Living Relatives and a Host Species Provide Insights into Adoption of a Parasitic Lifestyle and Prospects for Disease Control. *Genome Biol.* 12, R100. doi:10.1186/gb-2011-12-10-r100

Ersfeld, K. (2003). Genomes and Genome Projects of Protozoan Parasites. *Curr. Issues Mol. Biol.* 5, 61–74.

Florencia, D. V., Sebastián, P., Gonzalo, G., Moreira, D., Gregorio, I., and Carlos, R. (2019). Nanopore Sequencing Significantly Improves Genome Assembly of the Protozoan Parasite Trypanosoma Cruzi. *Genome Biol. Evol.* 11, 1952. doi:10.1093/gbe/evz129

Hatanaka, A., Umeda, N., Yamashita, S., and Hirazawa, N. (2007). Identification and Characterization of a Putative Agglutination/immobilization Antigen on the Surface ofCryptocaryon Irritans. *Parasitology* 134, 1163–1174. doi:10.1017/s003118200700265x

Hülsmann, N., and Hausmann, K. (1994). Towards a New Perspective in Protozoan Evolution. *Eur. J. Protistol.* 30, 365–371. doi:10.1016/s0932-4739(11)80211-7

Juranek, S. A., and Lipps, H. J. (2007). New Insights into the Macronuclear Development in Ciliates. *Int. Rev. Cytol.* 262, 219–251. doi:10.1016/s0074-7696(07)62005-1

Kumar, S., Gupta, S., Mohmad, A., Fular, A., Parthasarathi, B. C., and Chaubey, A. K. (2020). Molecular Tools-Advances, Opportunities and Prospects for the Control of Parasites of Veterinary Importance. *Int. J. Trop. Insect Sci.* 41, 1–10. doi:10.1007/s42690-020-00213-9

Lasek-Nesselquist, E., and Johnson, M. D. (2019). A Phylogenomic Approach to Clarifying the Relationship of Mesodinium within the Ciliophora: A Case Study in the Complexity of Mixed-Species Transcriptome Analyses. *Genome Biol. Evol.* 11, 3218–3232. doi:10.1093/gbe/evz233

Lokanathan, Y., Mohd-Adnan, A., Wan, K.-L., and Nathan, S. (2010). Transcriptome Analysis of the Cryptocaryon Irritans Tomont Stage Identifies Potential Genes for the Detection and Control of Cryptocaryonosis. *Bmc Genomics* 11, 76. doi:10.1186/1471-2164-11-76

Ludwig, W., and Klenk, H. (2001). *Overview: A Phylogenetic Backbone and Taxonomic Framework for Procaryotic Systematics*. New York: Springer.

Mao, Z., Li, M., and Chen, J. (2013). Draft Genome Sequence of Pseudomonas Plecoglossicida Strain NB2011, the Causative Agent of White Nodules in Large Yellow Croaker (Larimichthys Crocea). *Genome Announc* 1, e00586. doi:10.1128/genomeA.00586-13

Miller, D. N., Bryant, J. E., Madsen, E. L., and Ghiorse, W. C. (1999). Evaluation and Optimization of DNA Extraction and Purification Procedures for Soil and Sediment Samples. *Appl. Environ. Microbiol.* 65, 4715–4724. doi:10.1128/aem.65.11.4715-4724.1999

Mushegian, A. R., Garey, J. R., Martin, J., and Liu, L. X. (1998). Large-scale Taxonomic Profiling of Eukaryotic Model Organisms: a Comparison of Orthologous Proteins Encoded by the Human, Fly, Nematode, and Yeast Genomes. *Genome Res.* 8, 590–598. doi:10.1101/gr.8.6.590

Pan, B., Chen, X., Hou, L., Zhang, Q., Qu, Z., Warren, A., et al. (2019). Comparative Genomics Analysis of Ciliates Provides Insights on the Evolutionary History within "Nassophorea-Synhymenia-Phyllopharyngea" Assemblage. *Front. Microbiol.* 10, 2819. doi:10.3389/fmicb.2019.02819

Porcel, B. M., Tran, A.-N., Tammi, M., Nyarady, Z., Rydåker, M., Urmenyi, T. P., et al. (2000). Gene Survey of the Pathogenic Protozoan Trypanosoma Cruzi. *Genome Res.* 10, 1103–1107. doi:10.1101/gr.10.8.1103

Wang, Z., Samuelson, J., Clark, C. G., Eichinger, D., Paul, J., Van Dellen, K., et al. (2003). Gene Discovery in the Entamoeba Invadens Genome. *Mol. Biochem. Parasitol.* 129, 23–31. doi:10.1016/s0166-6851(03)00073-2

Wei, W., Chen, K., Miao, W., Yang, W., and Xiong, J. (2018). Pseudocohnilembus Persalinus Genome Database - the First Genome Database of Facultative Scuticociliatosis Pathogens. *BMC Genomics* 19, 676. doi:10.1186/s12864-018-5046-6

Wright, A.-D. G., and Colorni, A. (2002). Taxonomic Re-assignment of Cryptocaryon Irritans, a marine Fish Parasite. *Eur. J. Protistol.* 37, 375–378. doi:10.1078/0932-4739-00858

Wu, M., and Eisen, J. A. (2008). A Simple, Fast, and Accurate Method of Phylogenomic Inference. *Genome Biol.* 9, R151. doi:10.1186/gb-2008-9-10-r151

Yin, F., Sun, P., Tang, B., Gong, H., Ke, Q., and Li, A. (2016). Anti-parasitic Effects of Leptomycin B Isolated from Streptomyces Sp. CJK17 on marine Fish Ciliate Cryptocaryon Irritans. *Vet. Parasitol.* 217, 89. doi:10.1016/j.vetpar.2015.12.034

Yogeswaran, L., Adura, M. A., Kiew-Lian, W., and Nathan, S. (2010). Transcriptome Analysis of the Cryptocaryon Irritans Tomont Stage Identifies Potential Genes for the Detection and Control of Cryptocaryonosis. *Bmc Genomics* 11, 76. doi:10.1186/1471-2164-11-76

Zhang, Q., Yi, Z., Fan, X., Warren, A., Gong, J., and Song, W. (2014). Further Insights into the Phylogeny of Two Ciliate Classes Nassophorea and Prostomatea (Protista, Ciliophora). *Mol. Phylogenet. Evol.* 70, 162–170. doi:10.1016/j.ympev.2013.09.015

Zhao, J., Bai, H., Ke, Q., Li, B., Zhou, Z., Wang, H., et al. (2021a). Genomic Selection for Parasitic Ciliate Cryptocaryon Irritans Resistance in Large Yellow Croaker. *Aquaculture* 531, 735786. doi:10.1016/j.aquaculture.2020.735786

Zhao, J., Zhou, T., Bai, H., Ke, Q., Li, B., Bai, M., et al. (2021b). Genome-Wide Association Analysis Reveals the Genetic Architecture of Parasite

(Cryptocaryon Irritans) Resistance in Large Yellow Croaker (Larimichthys Crocea). *Mar. Biotechnol. (NY)* 23, 242. doi:10.1007/s10126-021-10019-6

Zhou, Z., Liu, B., Chen, B., Shi, Y., Pu, F., Bai, H., et al. (2019). The Sequence and De Novo Assembly of Takifugu Bimaculatus Genome Using PacBio and Hi-C Technologies. *Sci. Data* 6, 187. doi:10.1038/s41597-019-0195-2

Conflict of Interest: Author QK and WZ were employed by Ningde Fufa Fisheries Company Limited.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 January 2022
doi: 10.3389/fgene.2021.695763

The NGS Magic Pudding: A Nanopore-Led Long-Read Genome Assembly for the Commercial Australian Freshwater Crayfish, *Cherax destructor*

Christopher M. Austin[1,2]*, *Laurence J. Croft*[1,2], *Frederic Grandjean*[3] and *Han Ming Gan*[4]

[1]Deakin Genomics Centre, Deakin University, Geelong, VIC, Australia, [2]Centre for Integrative Ecology, School of Life and Environmental Sciences, Deakin University, Geelong, VIC, Australia, [3]Laboratoire Ecologie et Biologie des Interactions, Equipe Ecologie Evolution Symbiose, Unité Mixte de Recherche 7267 Centre National de la Recherche Scientifique, Université de Poitiers, Poitiers, France, [4]GeneSEQ Sdn Bhd, Rawang, Malaysia

Cherax destructor, the yabby, is an iconic Australian freshwater crayfish species, which, similar to other major invertebrate groups, is grossly under-represented in genomic databases. The yabby is also the principal commercial freshwater crustacean species in Australia subject to explotation via inland fisheries and aquaculture. To address the genomics knowledge gap for this species and explore cost effective and efficient methods for genome assembly, we generated 106.8 gb of Nanopore reads and performed a long-read only assembly of the *Cherax destructor* genome. On a mini-server configured with an ultra-fast swap space, the *de novo* assembly took 131 h (~5.5 days). Genome polishing with 126.3 gb of PCR-Free Illumina reads generated an assembled genome size of 3.3 gb (74.6% BUSCO completeness) with a contig N_{50} of 80,900 bp, making it the most contiguous for freshwater crayfish genome assemblies. We found an unusually large number of cellulase genes within the yabby genome which is relevant to understanding the nutritional biology, commercial feed development, and ecological role of this species and crayfish more generally. These resources will be useful for genomic research on freshwater crayfish and our methods for rapid and super-efficient genome assembly will have wide application.

Keywords: genome, annotation, nanopore, cellulase, aquacultrure, decapoda, Parastacidae

OPEN ACCESS

Edited by:
Ka Yan Ma,
Sun Yat-sen University, China

Reviewed by:
Manu Kumar Gundappa,
University of Edinburgh,
United Kingdom
Jeong-Hyeon Choi,
National Marine Biodiversity Institute of Korea, South Korea

***Correspondence:**
Christopher M. Austin
c.austin@deakin.edu.au

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *15 April 2021*
Accepted: *23 December 2021*
Published: *19 January 2022*

Citation:
Austin CM, Croft LJ, Grandjean F and Gan HM (2022) The NGS Magic Pudding: A Nanopore-Led Long-Read Genome Assembly for the Commercial Australian Freshwater Crayfish, Cherax destructor. *Front. Genet. 12:695763. doi: 10.3389/fgene.2021.695763*

1 INTRODUCTION

Australia's freshwater crayfish are highly diverse and as charismatic as the country's better known avian and mammalian fauna, but far less appreciated and studied. Crayfish are found in a range of freshwater environments, include some exceptionally large species in Australia, and can reach very high densities in both natural and cultured environments (Nyström and Strand, 1996; Whitledge and Rabeni, 1997; Jones and Ruscoe, 2000; Reynolds and Richardson, 2013). As a result, they often represent keystone species and ecosystem engineers in permanent and semi-permanent freshwater systems in many parts of the world. This also means they are an important part of food webs as significant prey items for fish, birds and mammals (Hicks and McCaughan, 1997; Jones and Grey, 2016), and for humans, including indigenous communities (Eyre et al., 1845; Austin, 1998; Kusabs

FIGURE 1 | Adult *Cherax destructor*. Photo provided by Christopher Austin.

and Quinn, 2009). Crayfish also have significant ecological roles within inland aquatic systems as they can consume and process sizeable volumes of a range of organic matter and detritus (Nyström and Strand, 1996; Whitledge and Rabeni, 1997; Reynolds and Richardson, 2013; Jones and Grey, 2016). While crayfish are widely considered as omnivorous and opportunistic feeders their exact ecological role and nutritional biology has been controversial (Momot, 1995) and are assuming greater importance with the frequent translocation of crayfish species and their potential to cause a range of negative ecological impacts both locally and globally (Austin and Ryan, 2002; Lodge et al., 2012; James et al., 2016; Souty-Grosset et al., 2016). Some authors have postulated that freshwater crayfish are primarily carnivorous (Momot, 1995; Weinländer and Füreder, 2012), however molecular and limited NGS-based studied have revealed the presence of cellulase and a diversity of carbohydrate-active related genes supporting an adaption to the processing of plant-based food (Crawford et al., 2005; Tan et al., 2016). The first cellulase reported for freshwater crayfish was from the GH9 family which was found to be especially diverse in *Cherax quadricarinatus* based on a transcriptomic study by Tan et al. (2015) Tan et al. (2016).

To date only one crayfish genome is available for the northern hemisphere species, *Procamabarus virginalis* (Cambaridae) and the southern hemisphere *Cherax quadricarinatus* (Parastacidae). *Cherax destructor*, commonly known as the yabby (**Figure 1**), is an iconic Australian freshwater crayfish species with a wide distribution throughout the river systems, lakes, swamps, and billabongs[1] of inland Australia (Horwitz and Knott, 1995; Nguyen et al., 2004). It is the major commercial freshwater crayfish species in the country (Piper, 2000; Wingfield, 2008) and increasingly scientists are using it or closely related species as a model research species as they are easily maintained and bred in captivity (Mccarthy and Macmillan, 1999; Biro and Sampson, 2015; Beltz and Benton, 2017; Ventura et al., 2019). Despite the decreasing cost of whole-genome sequencing, publicly available whole-genome assemblies for freshwater crayfish species is scarce. Like many decapod crustaceans have large and repetitive genomes (Tan et al., 2020a) so short-read only *de novo* assemblies are memory-intensive and the resulting assemblies are often highly fragmented and difficult to annotate, thereby limiting their utility. While the supplementation of high coverage short-read data sets with low coverage (<10 ×) of long, but less accurate Nanopore or PacBio reads, is increasing the speed and quality of genome assemblies, it is still time-consuming, computationally demanding and challenging (Austin et al., 2017; Tan et al., 2018; Gan et al., 2019).

In this study, we sequence the genome of *Cherax destructor* and demonstrate that by starting with a medium coverage long read data set (~20 × coverage) and similar coverage of Illumina reads for error-correction, the speed at which a quality reference genome can be produced can be greatly increased, even for species with large, and repetitive genomes. We benchmark our assembly against available genome assemblies for decapod crustaceans representing 11 species from a range of infraorders. Given the degree of ongoing interest in the nutritional biology and trophic status of freshwater crayfish, we also examine the diversity of cellulase genes in freshwater crayfish.

2 METHODS

2.1 Genome Sequencing Libraries

A euthanized female crayfish specimen was provided by a local amateur angler in August 2019. The hepatopancreas tissue was dissected and homogenized in DNA/RNAshield (Zymo Research). Then, several gDNA extractions were performed on the homogenized hepatopancreas using the Zymo Quick gDNA kit (Zymo Research). For Nanopore sequencing, 20 µg of gDNA was fragmented to 8–10 kb using Covaris g-tube and 2–4 ug of the fragmented gDNA was subsequently used to construct a Nanopore library with the LSK109 library preparation kit. One-eighth of the eluted library volume was loaded onto an R9.4.1 revD flowcell followed by sequencing. Every 8–16 h, the run was stopped followed by a nuclease flush, library reload, and sequencing. Nanopore sequencing was performed on a total of 12 brand new and eight used (and nuclease flushed) flowcells. Base-calling of the fast5 reads used Guppy v3.3.3 (high accuracy mode). A total of 15,928,097 passed reads were generated totalling to 106.8 gigabases (Mean length: 6,705 bp, Median Length: 5,861 bp and Read Length N_{50}: 8,843 bp, Longest read length: 182,535 bp). For Illumina sequencing, 1 µg of gDNA was fragmented to 350 bp and processed using the TruSeq DNA PCR-Free Kit (Illumina). Sequencing was done on a NovaSeq6000 using a run configuration of 2 × 150 bp. A total of 418,053,185 paired-end reads were generated totaling to 126.3 gigabases.

2.2 Genome Assembly

Whole-genome assembly was performed on an Ubuntu 18.04 mini-server equipped with AMD EPYC 7551P 32-core processor, 256 GB physical RAM, and 750 GB swap space created on a RAID 0 (Redundant Array of Independent Disks) partition comprising

[1]Indigenous Australian name for a stagnant waterhole or river pool accepted into English.

TABLE 1 | Genome assembly and annotation statistics.

Parameter	Details
Organism	Cherax destructor (Australian yabby)
Isolate	CDF2 (female, wild population)
Bioproject	PRJNA588861
Biosample	SAMN13258587
Whole-genome GenBank accession	WNWK00000000
Assembled scaffold/contig length	3,336,744,225 bp/ 3,336,542,896 bp
Scaffold N_{50} (number of sequences)	87,184 bp (98,662)
Contig N_{50} (number of sequences)	80,900 bp (100,635)
GC content	41.43%
BUSCO completeness	74.6% Single-copy, 1.1% Duplicated
Arthropoda odb9 (n = 1,006)	15.1% Fragmented, 9.1% Missing
Number of predicted protein-coding genes	45,673
Number of predicted proteins	47,377
With InterPro signature	21,102 (44.5%)
With gene ontology (GO) term	14,068 (29.7%)

two 1 TB drives. Nanopore reads and intermediate assembly files were all stored on a separate RAID 0 partition comprising four 4 TB hard drives. De novo assembly of the Nanopore reads used wtdbg 2.5 (Ruan and Li, 2019) with the options "-t 60 -p 19 -AS 2 -s 0.05 -L 3000 -g 6G --edge-min 2 --rescue-low-cov-edges". Using this configuration, the *de novo* assembly took 131 h (~5.5 days) to complete with a maximum memory usage of 607 GD. After the wtdbg assembly, one round of polishing with long reads was performed using the wtdbg 2.5 internal polishing tool, wtpoa-cns. For genome polishing with Illumina reads, two rounds of polishing with Pilon v1.22 (Walker et al., 2014) were carried out. The raw paired-end reads were first adapter, quality and poly-G trimmed with fastp v0.20.0 (Chen et al., 2018). For each round of pilon-polishing, the trimmed reads were aligned to the genome using bwa-mem v 0.7.17-r1188 (Li, 2013) followed by correction of individual base errors (SNPs) and small indels using the options "--diploid –fix bases". To overcome memory limitation in Pilon due to large genome size, the genome was split into 10 smaller fasta files, processed with Pilon separately and merged back into a single fasta file. Transcriptome-guided scaffolding of the polished contigs was performed with P_RNA_scaffolder v1 (Zhu et al., 2018) using publicly available transcriptome data (Ali et al., 2015). The genome completeness was assessed using BUSCO v5 (Waterhouse et al., 2017) with the Arthropoda ortholog dataset (Arthropod odb10). Statistics of the resulting assembly were generated using QUAST v5.0.2 (Gurevich et al., 2013) and are presented in **Table 1**. Illumina and Nanopore reads were mapped to the final assembly using bwa-mem (Li, 2013) and minimap2 v2.17 (Li, 2018), respectively. The BAM files were separately processed in Qualimap2 v2.2.1 (Okonechnikov et al., 2016) to generate additional statistics for the genome assembly based on read alignment.

2.3 Repeat Annotation and Protein-Coding Gene Prediction

Repetitive regions were identified using RepeatModeler v1.0.11 (Smit and Hubley, 2010). The *de novo* generated repeat library (Gan et al., 2020) was subsequently used to soft-mask the genome assembly with RepeatMaskerv4.0.7 (Tarailo-Graovac and Chen, 2009) with the options "-no_is –div 40 –xsmall". Using this repeat annotation approach, 61.34% of the genome has been repeat-masked with long interspersed nuclear elements (LINEs) being the most common repeat annotated (31%). For protein-coding gene prediction, BRAKER v2.1.4 (Hoff et al., 2019) was chosen since it can incorporate both RNA-sequencing data and closely related proteins for gene prediction training. Publicly available *Cherax destructor* transcriptome datasets (Ali et al., 2015) were downloaded and aligned to the genome using STAR v2.7.1a (Dobin et al., 2013). To obtain closely related protein sequences, all publicly available *Cherax quadricarinatus* transcriptome data were downloaded from NCBI-SRA as of 2nd December 2019, individually assembled using rnaSPAdes v3.13.0 (Bushmanova et al., 2019) followed by redundancy removal of the concatenated transcripts using EvidentialGene v2013.03.11 (Gilbert, 2019). *Cherax quadricarinatus* translated open reading frames that are larger than 200 amino acid residues and labelled as "complete" e.g., with intact 5′ and 3′ ends, were selected as the protein input (Gan et al., 2020) for training in BRAKER2 using default settings. Using Orthofinder v2.3.8 (Emms and Kelly, 2018), the initial predicted proteins from BRAKER2 were used as the input for orthologous clustering with the available proteomes of the red claw crayfish (*C. quadricarinatus*) (Tan et al., 2020a), pacific white shrimp (*Litopenaeus vannamei*) (Zhang et al., 2019), black tiger prawn (*Penaeus monodon*) (Quyen et al., 2020), marbled crayfish (*Procambarus virginalis*) (Gutekunst et al., 2018), and amphipod (*Parhyale hawaiensis*) (Kao et al., 2016). Then, the predicted *C. destructor* proteins that formed orthologous clusters with at least one of the decapod species were used for subsequent annotation and analysis. Specific comparisons of peptide homology were made with several decpod crustaceans including the recently published clawed lobster genome (clawed lobsters are from the clade most closely related to the freshwater crayfish) (Polinski et al., 2021), the southern hemisphere crayfish (*Cherax quadricarinatus*) using NCBI's *blastx* (evalue $1e^{-10}$). Putative protein functions were inferred using InterproScan v5.35-74.0 (Jones et al., 2014) with the options "—iprlookup –goterms --dp". Identification of Carbohydrate-Active enzymes (CAZy) in the selected crustacean proteomes used dbCAN2 v2.0.0 (Zhang et al., 2018) and the identified GH9 cellulases were further extracted and their diversity explored by phylogenetic analysis. The GH9 cellulases were first aligned with MUSCLE v3.8.31 (Edgar, 2004) followed by trimming in trimal v1.9 (Capella-Gutiérrez et al., 2009) ("-automated1" option) and phylogenetic construction in IqTree v1.6.10 (Nguyen et al., 2014) ("-m TESTNEW –bb 1,000" options). The unrooted IQTree maximum likelihood tree was annotated and visualized in TreeFig v1.4.3 (Rambaut, 2009).

2.4 Data Availability

Raw sequencing libraries have been deposited in NCBI-SRA under the BioProjectPRJNA588861. The genome assembly has been deposited in GenBank under the accession number WNWK000000 (the version described in this paper is

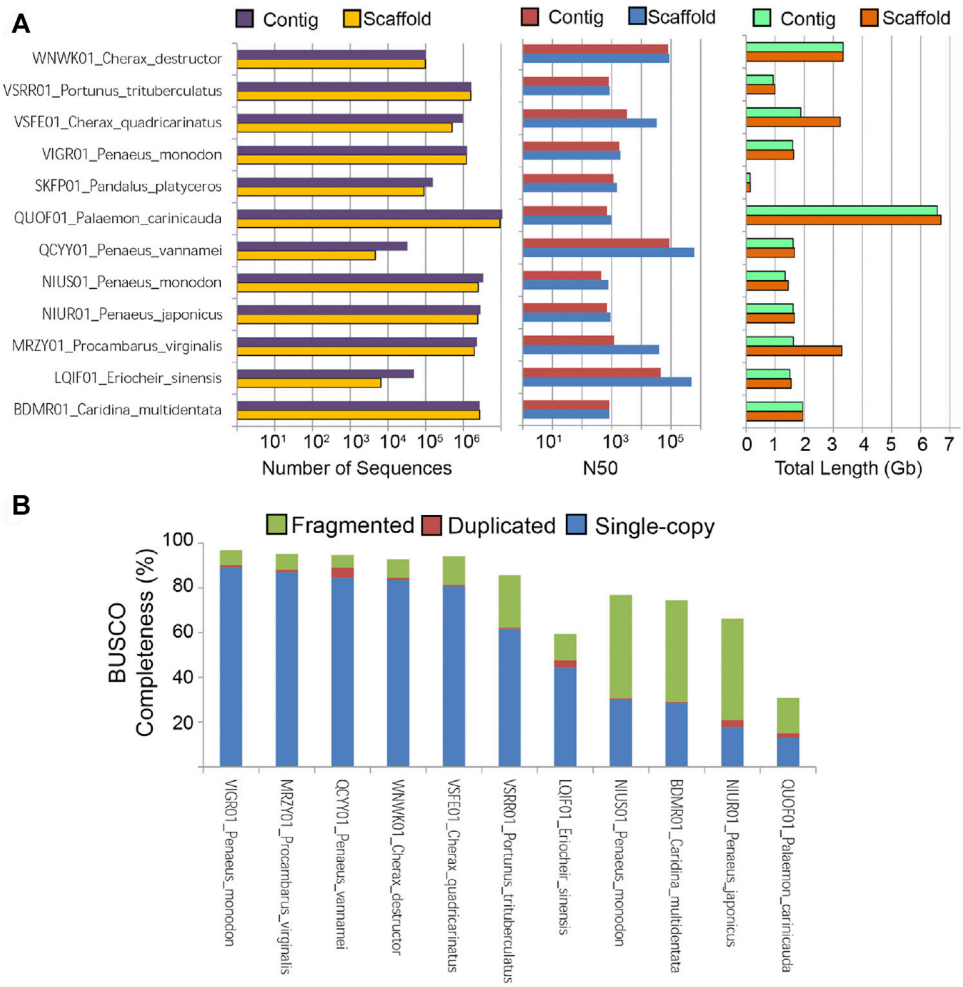

FIGURE 2 | Statistics of publicly available decapod crustacean genome assemblies. **(A)** Number of sequences, Genome N_{50}, and total assembled length **(B)** BUSCO completeness based on the Arthopoda ortholog dataset.

WNWK01000000). The wtdbg2.5 assembly log file, intermediate *C. destructor* genome assemblies, repeat annotation, CAZy annotation, protein-coding gene prediction (GTF format), predicted genes, and proteins have been deposited in the Zenodo repository (Gan et al., 2020). The *C. quadricarinatus* RNASpades transcriptome assemblies, QUAST-generated genome statistics for all Decapod genomes re-analyzed in this paper and their BUSCO calculations are also deposited at Zenodo (Gan et al., 2020).

3 RESULTS AND DISCUSSION

An alignment rate of more than 99.5% was observed for both Illumina and Nanopore reads with the most frequently observed sequencing depth of 29× and 23×, respectively. Assuming the sequencing depth with the highest observed frequency represents the coverage of the single-copy genomic region, the genome size of *Cherax destructor* is estimated to be 4.36–4.64 gb (Total sequencing yield in gigabases divided by single-copy coverage). This is consistent with genome size estimates for the northern hemisphere crayfish *Procambarus virginalis* (~3.5 gb) and *Cherax quadricarinatus* (~5 gb) (Tan et al., 2020a) making Australian crayfish larger than all other crustaceans so far sequenced with the exception of the prawn *Exoplaemon carinicauda* (9.5 gb).

Using 106.8 gb and 126.3 gb of Nanopore and Illumina data, respectively, a 3.3 gb genome assembly was generated with an estimated BUSCO score of 89.7% in less than a week. The assembled genome size was ~27.0% smaller than the genome size estimate. This is quite a common outcome for decapod genome assemblies due to sequencing bias and their repetitive genomes (Tan et al., 2020a; Polinski et al., 2021) and was reflected in the uneven distribution of read depths across scaffolds in our study. Over 3,000 scafolds have over 300x coverage, compared with an average read depth of 111x, consistent with the occurence of a significant proportion of repeat regions and potentially contributing to the discrepency between the assembled genome size and the genome size estimate.

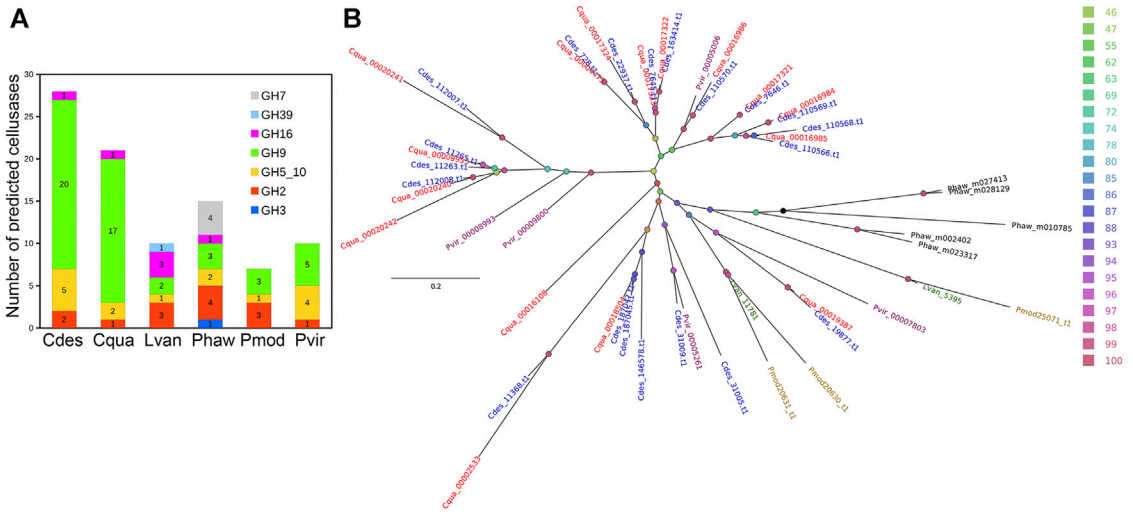

FIGURE 3 | Identification and phylogenetic analysis of cellulases. **(A)** Number of identified cellulases in five decapod crustacean and an amphipod proteomes **(B)** IQTree maximum likelihood tree showing the evolutionary relationships of GH9 cellulases identified from the selected proteomes. The nodes were colored based on ultrafast bootstrap values and the first three letters in each tip label correspond to the species name. Branch lengths indicate number of substitutions per site. Cdes, *Cherax destructor*; Cqua, *Cherax quadricarinatus*; Lvan, *Litopenaeus vannamei*; Phaw, *Parhyale hawaiensis*; Pmod, *Penaeus monodon*; Pvir, *Procambarus virginalis*.

The contig N_{50} of 80,900 bp is the longest to date among currently available for freshwater crayfish genome assemblies. Comparisons with the recently sequenced *Cherax quadricarinatus* genome (Tan et al., 2020a) initially assembled using short reads followed by scaffolding with low coverage Nanopore long reads, show that a long-read led assembly is more efficient though more costly. However, the higher cost of long reads is more than compensated for by increased computational efficiencies due to the availability of speedy and memory-efficient long-read assemblers (wtdbg2) (Ruan and Li, 2019) and lack of reliance on the need to generate large volumes of Illumina reads during the initial assembly stage.

The cumulative scaffold length of *C. destructor* is similar to the *C. quadricarinatus* genome (~3 gb) that was assembled using Illumina reads (191x) followed by scaffolding with low coverage Nanopore reads (x7). In comparison with the other two crayfish assemblies the advantages of a Nanopore-based assembly with an increased volume of long reads can be seen from **Figure 2**, where the difference between the contig and scaffold level assemblies is greatly reduced leading to a less gappy assembly. Also the need for high volumes of short reads is also greatly reduced with only 123.6 gb used in they study compared with used for the assemblies of *C. quadricarinatus* (964 gb) and *Procambarus virginalis* (350 gb).

It is also worth noting that this *C. destructor* genome assembly exhibits a contig N_{50} length of nearly 100 kb which is longest among freshwater crayfish genome assemblies. Recent decapod assemblies increasingly using both short and long reads and Hi-C data which is assisting in more robust decapod crustacean genome assemblies (Zhang et al., 2019; Tang et al., 2020) especiallyfor those species with large repetitive genomes such as *Macrobrachium* shrimps (Jin et al., 2021). The reported *C. destructor* BUSCO genome completeness in this study is also one of the highest to date for freshwater crayfish (**Figure 2B**). A logical next step, given the large and repetitve genomes exhinited by frewshwater crayfish, is to attempt to improve this genome assembly via the inclusion of HiC data (Jin et al., 2021).

An initial 187,638 of putative unigenes were predicted by BRAKER2. The final protein set consisted of 47,377 transcripts (45,673 genes) of which 21,102 and 14,068 were identified with InterPro signature and Gene Ontology term, respectively. The number of predicted proteins with InterPro signatures is very similar to other species of decapod crustaceans. A total of 68.97% of *C. destructor* peptides mapped to the related *C. quadricarinatus* annotation (evalue $1e^{-10}$) (Tan et al., 2016). More specifically, we get 32,677 peptides in common with *Cherax quadricarinatus*, 25,129 with *Procambarus virginalis*, 23,008 with *Penaeus monodon*, 17,159 with *Litopenaeus vannamei*, and 10,318 with *Homarus americanus*. The number of predicted proteins with InterPro signatures is very similar to other species of decapod crustaceans (Tan et al., 2016). While the total number of predicted protein-coding genes is large (45,673) relative to those that have an Interpro signature, this number does not differ greatly from the recently published genome for the clawed lobster, *Hommarus americanus*, which identified 40,732 peptides (Polinski et al., 2021). This high proportion of unique genes is most likely a function of the evolution of a large repetitive genome and the limited genomic data for crayfish and lobsters as pointed out by Polinski et al. (2021) in their recent study of he American lobster (Polinski et al., 2021). Significantly, *Cherax destructor* harbours the highest number of cellulase genes among the currently sequenced decapod crustaceans (**Figure 3A**) with a substantially higher number of GH9 cellulase genes comparable to its close relative, *C. quadricarinatus*, which was previously highlighted in an earlier transcriptomic study (Tan et al., 2016). Phylogenetic analysis of the GH9 cellulases showed a clustering pattern first by the GH9 cellulase variants and then by species relatedness (**Figure 3B**).

Despite the high number of GH9 cellulases identified among the *Cherax* spp., they were generally closely related and localized in a few major clades (**Figure 3B**). Although there were a few that claded with those from the northern hemisphere crayfish *P. virginalis*, indicating a more ancient origin. *Cherax destructor*, is considered to be versatile in its nutrient utilisation based on both dietary and field-based studies (Jones and De Silva, 1997; Beatty, 2006; Giling et al., 2009; Johnston et al., 2011) and is considered an opportunistic omnivorous generalist, that can derive nutrition directly from both animal and plant material and detritus.

A common view is that crayfish, in general, have a trophic role primarily as predators (Momot, 1995) may need to be re-assessed, given the antiquity, and diversity of cellulase and related genes in this group. However there also may be wide variation within and among crayfish species and the diet of particular species can vary in time and space (Beatty, 2006; Giling et al., 2009; Johnston et al., 2011) which has contributed to conflicting views. For example, Johnston et al. (2011) found variation between species from the same crayfish community ranging from primarily herbivorous species to primarily carnivorous species. Other species from this crayfish community, including *C. destructor*, had either mixed diets or switched between plant, and animal diets at different sites. It will, therefore, be of great interest to further examine cellulase diversity and expression in a range of crayfishes species from different environments including under aquaculture conditions and the ability of different crayfish species to utilise plant material in the field and through laboratory trials and how this relates to cellulase gene profiles and their expression.

In general, a significant limitation in further advancing the study of the genomics of non-model organisms is the computational resources and time needed to assemble genomes from predominately short reads, even when aided with long reads for scaffolding (Lewin et al., 2018). This problem is further exacerbated for groups with larger repetitive genomes, which means analyses can take months if not years and still lead to poor quality assemblies. In this study, we demonstrate that a high-quality genome assembly for a decapod crustacean with a large (>3 gb) and repetitive genome can be achieved with modest sequencing volumes, that take advantage of rapid and ongoing developments in third generation sequencing technologies, and can be completed in under 1 week of computation time on a high performance desktop machine.

4 CONCLUSION

This reference genome, along with its annotation, will be useful for future functional, ecological, aquaculture-related and evolutionary genomic studies, and genome-based selection and targeted genetic manipulation of this emerging aquaculture species. Given our finding of an evolutionary proliferation of cellulase genes, we are hoping these data will stimulate new research into the nutritional biology and trophic roles of freshwater crayfish in freshwater ecosystems. We see the continuing advances in Nanopore and other third generation sequencing technologies like the fabled "magic pudding" from a well known Australian children's story (Norman, 1918), it keeps on "giving", similar to the continuing inprovements in efficiency, output volume, and accuracy making the intractable, tractable when it comes to genome sequencing and assembly of non-model species. As a consequence we are able to provide a new model with respect to sequencing platforms, hardware configuration and assembly strategy to enable an ultrafast and efficient genome assembly that can be potentially applied to any species, including those with large and repetitive genomes. We anticipate our strategy and methodology will help elevate the study of interesting and important invertebrate genomes.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/genbank/, PRJNA588861 https://www.ncbi.nlm.nih.gov/genbank/, SAMN13258587 https://www.ncbi.nlm.nih.gov/genbank/, WNWK00000000.1.

AUTHOR CONTRIBUTIONS

HG—Conceived and designed the analysis, collected the data, performed the analysis, and wrote the paper. FG—Conceived and designed the study, contributed to the paper. LC—Contributed to bioinformatics and discussion. CA—Conceived and designed the analysis, collected the data, contributed data, and wrote the paper.

FUNDING

Funding was provided by Deakin University and the University of Poitiers.

ACKNOWLEDGMENTS

We would like to thank Julian Vreugdenburg for the technical support and configuration of the mini-server which enabled the rapid completion of the memory-intensive *de novo* assembly.

REFERENCES

Ali, M. Y., Pavasovic, A., Amin, S., Mather, P. B., and Prentis, P. J. (2015). Comparative Analysis of Gill Transcriptomes of Two Freshwater Crayfish, Cherax Cainii and C. Destructor. *Mar. Genomics* 22, 11–13. doi:10.1016/j.margen.2015.03.004

Austin, C. M. (1998). *Potential for the Commercial Exploitation of Freshwater Crayfish via Aquaculture in the Mt Bosavi Region of Papua New Guinea - a Preliminary Report*. Geelong, Australia: Unpublished Report for the World Wide Fund for Nature (WWF).

Austin, C. M., Tan, M. H., Harrisson, K. A., Lee, Y. P., Croft, L. J., Sunnucks, P., et al. (2017). De Novo genome Assembly and Annotation of Australia's Largest

Freshwater Fish, the Murray Cod (*Maccullochella peelii*), from Illumina and Nanopore Sequencing Read. *GigaScience* 6 (8), 1–6. doi:10.1093/gigascience/gix063

Austin, C. M., and Ryan, S. G. (2002). Allozyme Evidence for a New Species of Freshwater Crayfish of the Genus *Cherax* Erichson (Decapoda : Parastacidae) from the South-West of Western Australia. *Invert. Syst.* 16 (3), 357–367. doi:10.1071/it01010

Beatty, S. J. (2006). The Diet and Trophic Positions of Translocated, Sympatric Populations of *Cherax destructor and Cherax cainii* in the Hutt River, Western Australia: Evidence of Resource Overlap. *Mar. Freshw. Res.* 57 (8), 825–835. doi:10.1071/mf05221

Beltz, B. S., and Benton, J. L. (2017). From Blood to Brain: Adult-Born Neurons in the Crayfish Brain Are the Progeny of Cells Generated by the Immune System. *Front. Neurosci.* 11, 662. doi:10.3389/fnins.2017.00662

Biro, P. A., and Sampson, P. (2015). *Fishing Directly Selects on Growth Rate via Behaviour: Implications of Growth-Selection that Is Independent of Size*. Proc. R. Soc. B, 282. doi:10.1098/rspb.2014.2283

Bushmanova, E., Antipov, D., Lapidus, A., and Prjibelski, A. D. (2019). rnaSPAdes: a De Novo Transcriptome Assembler and its Application to RNA-Seq Data. *GigaScience* 8 (9), giz100. doi:10.1093/gigascience/giz100

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses. *Bioinformatics* 25 (15), 1972–1973. doi:10.1093/bioinformatics/btp348

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). Fastp: an Ultra-fast All-In-One FASTQ Preprocessor. *Bioinformatics* 34 (17), i884–i890. doi:10.1093/bioinformatics/bty560

Crawford, A. C., Richardson, N. R., and Mather, P. B. (2005). A Comparative Study of Cellulase and Xylanase Activity in Freshwater Crayfish and marine Prawns. *Aquac. Res.* 36 (6), 586–592. doi:10.1111/j.1365-2109.2005.01259.x

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: Ultrafast Universal RNA-Seq Aligner. *Bioinformatics* 29 (1), 15–21. doi:10.1093/bioinformatics/bts635

Edgar, R. C. (2004). MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Res.* 32 (5), 1792–1797. doi:10.1093/nar/gkh340

Emms, D. M., and Kelly, S. (2018). OrthoFinder2: Fast and Accurate Phylogenomic Orthology Analysis from Gene Sequences. *BioRxiv*, 1–34. doi:10.1101/466201

Eyre, E. J., and Boone, W. (1845). "*Journals of Expeditions of Discovery into Central Australia, and Overland from Adelaide to King George's Sound, in the Years 1840-1 : Sent by the Colonists of South Australia, with the Sanction and Support of the Government : Including an Account of the Manners and Customs of the Aborigines and the State of Their Relations with Europeans*," London.

Gan, H. M., Falk, S., Moraleś, H. E., Austin, C. M., Sunnucks, P., and Pavlova, A. (2019). Genomic Evidence of Neo-Sex Chromosomes in the Eastern Yellow Robin. *GigaScience* 8 (9), giz111. doi:10.1093/gigascience/giz131

Gan, H. M., Granjean, F., and Austin, C. M. (2020). *Dataset for "Nanopore-Led Long-Read Genome Assembly of the Australian Yabby, Cherax destructor*. Zenodo: Front. Genet.

Gilbert, D. G. (2019). Genes of the Pig, *Sus scrofa*, Reconstructed with EvidentialGene. *PeerJ* 7, e6374. doi:10.7717/peerj.6374

Giling, D., Reich, P., and Thompson, R. M. (2009). Loss of Riparian Vegetation Alters the Ecosystem Role of a Freshwater Crayfish (*Cherax destructor*) in an Australian Intermittent lowland Stream. *J. North Am. Benthological Soc.* 28 (3), 626–637. doi:10.1899/09-015.1

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: Quality Assessment Tool for Genome Assemblies. *Bioinformatics* 29 (8), 1072–1075. doi:10.1093/bioinformatics/btt086

Gutekunst, J., Andriantsoa, R., Falckenhayn, C., Hanna, K., Stein, W., Rasamy, J., et al. (2018). Clonal Genome Evolution and Rapid Invasive Spread of the Marbled Crayfish. *Nat. Ecol. Evol.* 2 (3), 567–573. doi:10.1038/s41559-018-0467-9

Hicks, B. J., and McCaughan, H. M. C. (1997). Land Use, Associated Eel Production, and Abundance of Fish and Crayfish in Streams in Waikato, New Zealand. *New Zealand J. Mar. Freshw. Res.* 31 (5), 635–650. doi:10.1080/00288330.1997.9516795

Hoff, K. J., Lomsadze, A., Borodovsky, M., and Stanke, M. (2019). "Whole-Genome Annotation with BRAKER," in *Gene Prediction* (Berlin/Heidelberg, Germany: Springer), 65–95. doi:10.1007/978-1-4939-9173-0_5

Horwitz, P., and Knott, B. (1995). The Distribution and Spread of the Yabby *Cherax destructor* Comples in Australia: Speculations, Hypotheses and the Need for Research. *Freshw. Crayfish* 10, 11.

James, J., Thomas, J. R., Ellis, A., Young, K. A., England, J., and Cable, J. (2016). Over-invasion in a Freshwater Ecosystem: Newly Introduced Virile Crayfish (*Orconectes virilis*) Outcompete Established Invasive Signal Crayfish (Pacifastacus Leniusculus). *Mar. Freshw. Behav. Physiol.* 49 (1), 9–18. doi:10.1080/10236244.2015.1109181

Jin, S., Bian, C., Jiang, S., Han, K., Xiong, Y., and Zhang, W. (2021). A Chromosome-Level Genome Assembly of the oriental River Prawn, *Macrobrachium nipponense*. *Gigascience* 10 (1). doi:10.1093/gigascience/giaa160

Johnston, K., Robson, B. J., and Fairweather, P. G. G. (2011). Trophic Positions of Omnivores Are Not Always Flexible: Evidence from Four Species of Freshwater Crayfish. *Austral Ecol.* 36 (3), 269–279. doi:10.1111/j.1442-9993.2010.02147.x

Jones, C. M., and Ruscoe, I. M. (2000). Assessment of stocking size and density in the production of redclaw crayfish, *Cherax quadricarinatus* (von Martens)(Decapoda: Parastacidae), cultured under earthen pond conditions. *Aquaculture* 189 (1-2), 63–71. doi:10.1016/s0044-8486(00)00359-8

Jones, E. J., and Grey, J. (2016). in *Environmental Drivers for Population Success: Population Biology, Population and Community Dynamics in Biology and Ecology of Crayfish*. Editor L. M. a. P. Stebbing (Boca Raton: CRC Press), 36.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: Genome-Scale Protein Function Classification. *Bioinformatics* 30 (9), 1236–1240. doi:10.1093/bioinformatics/btu031

Jones, P. L., and De Silva, S. S. (1997). Apparent Nutrient Digestibility of Formulated Diets by the Australian Freshwater Crayfish *Cherax destructor* Clark (Decapoda, Parastacidae). *Aquac. Res.* 28 (11), 881–891. doi:10.1046/j.1365-2109.1997.00913.x

Kao, D., Lai, A. G., Stamataki, E., Rosic, S., Konstantinides, N., Jarvis, E., et al. (2016). The Genome of the Crustacean *Parhyale hawaiensis*, a Model for Animal Development, Regeneration, Immunity and Lignocellulose Digestion. *Elife* 5, e20062. doi:10.7554/eLife.20062

Kusabs, I. A., and Quinn, J. M. (2009). Use of a Traditional Maori Harvesting Method, the Tau Kōura, for Monitoring Kōura (Freshwater *Crayfish, Paranephrops planifions*) in Lake Rotoiti, North Island, New Zealand. *New Zealand J. Mar. Freshw. Res.* 43 (3), 713–722. doi:10.1080/00288330909510036

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., et al. (2018). Earth BioGenome Project: Sequencing Life for the Future of Life. *Proc. Natl. Acad. Sci. USA* 115 (17), 4325–4333. doi:10.1073/pnas.1720115115

Li, H. (2013). Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. arXiv preprint arXiv:1303.3997, Available at: https://www.scienceopen.com/document?vid=e623e045-f570-42c5-80c8-ef0aea06629c

Li, H. (2018). Minimap2: Pairwise Alignment for Nucleotide Sequences. *Bioinformatics* 34 (18), 3094–3100. doi:10.1093/bioinformatics/bty191

Lodge, D. M., Deines, A., Gherardi, F., Yeo, D. C. J., Arcella, T., Baldridge, A. K., et al. (2012). Global Introductions of Crayfishes: Evaluating the Impact of Species Invasions on Ecosystem Services. *Annu. Rev. Ecol. Evol. Syst.* 43, 449–472. doi:10.1146/annurev-ecolsys-111511-103919

Mccarthy, B. J., and Macmillan, D. L. (1999). Control of Abdominal Extension in the Freely Moving Intact Crayfish *Cherax destructor*. I. Activity of the Tonic Stretch Receptor. *J. Exp. Biol.* 202, 11. doi:10.1242/jeb.202.2.171

Momot, W. T. (1995). Redefining the Role of Crayfish in Aquatic Ecosystems. *Rev. Fish. Sci.* 3 (1), 33–63. doi:10.1080/10641269509388566

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2014). IQ-TREE: a Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* 32 (1), 268–274. doi:10.1093/molbev/msu300

Nguyen, T. T. T., Austin, C. M., Meewan, M. M., Schultz, M. B., and Jerry, D. R. (2004). Phylogeography of the Freshwater Crayfish *Cherax destructor* Clark (Parastacidae) in Inland Australia: Historical Fragmentation and Recent Range Expansion. *Biol. J. Linn. Soc.* 83 (4), 539–550. doi:10.1111/j.1095-8312.2004.00410.x

Norman, L. *The Magic Pudding : Being the Adventures of Bunyip Bluegum and His Friends Bill Barnacle and Sam Sawnoff. 1918*. Sydney: Angus & Robertson.

Nyström, P., and Strand, J. (1996). Grazing by a Native and an Exotic Crayfish on Aquatic Macrophytes. *Freshw. Biol.* 36 (3), 673–682.

Okonechnikov, K., Conesa, A., and García-Alcalde, F. (2016). Qualimap 2: Advanced Multi-Sample Quality Control for High-Throughput Sequencing Data. *Bioinformatics* 32 (2), 292–294. doi:10.1093/bioinformatics/btv566

Piper, L. (2000). Potential for Expansion of the Freshwater Crayfish Industry in Australia : a Report for the Rural Industries Research and Development Corporation. Available at: https://www.agrifutures.com.au/wp-content/uploads/publications/00-142.pdf

Polinski, J. M., Zimin, A. V., Clark, K. F., Kohn, A. B., Sadowski, N., Timp, W., et al. (2021). The American Lobster Genome Reveals Insights on Longevity, Neural, and Immune Adaptations. *Sci. Adv.* 7 (26), eabe8290. doi:10.1126/sciadv.abe8290

Quyen, D. V., Gan, H. M., Lee, Y. P., Nguyen, D. D., Tran, X. T., Nguyen, V. S., et al. (2020). Improved Genomic Resources for the Black Tiger Prawn (*Penaeus monodon*). *Marine Genomics* 52, 100751. doi:10.1016/j.margen.2020.100751

Rambaut, A. (2013). FigTree. Available at: http://treebioedacuk/software/figtree/ (Accessed on 9th January 2020).

Reynolds, J., Souty-Grosset, C., and Richardson, A. (2013). Ecological Roles of Crayfish in Freshwater and Terrestrial Habitats. *Freshwater. Crayfish* 19 (2), 197–218. doi:10.5869/fc.2013.v19-2.197

Ruan, J., and Li, H. (2019). Fast and Accurate Long-Read Assembly with Wtdbg2. *Nat. Methods*. 17, doi:10.1038/s41592-019-0669-3

Smit, A. F., and Hubley, R. (2010). RepeatModeler Open-1.0.

Souty-Grosset, C., Anastácio, P. M., Aquiloni, L., Banha, F., Choquer, J., Chucholl, C., et al. (2016). The Red Swamp Crayfish *Procambarus clarkii* in Europe: Impacts on Aquatic Ecosystems and Human Well-Being. *Limnologica* 58, 78–93. doi:10.1016/j.limno.2016.03.003

Tan, M. H., Austin, C. M., Hammer, M. P., Lee, Y. P., Croft, L. J., and Gan, H. M. (2018). Finding Nemo: Hybrid Assembly with Oxford Nanopore and Illumina Reads Greatly Improves the Clownfish (*Amphiprion ocellaris*) Genome Assembly. *GigaScience* 7 (3), 1–6. doi:10.1093/gigascience/gix137

Tan, M. H., Gan, H. M., Gan, H. Y., Lee, Y. P., Croft, L. J., Schultz, M. B., et al. (2016). First Comprehensive Multi-Tissue Transcriptome of *Cherax quadricarinatus* (Decapoda: Parastacidae) Reveals Unexpected Diversity of Endogenous Cellulase. *Org. Divers. Evol.* 16 (1), 185–200. doi:10.1007/s13127-015-0237-3

Tan, M. H., Gan, H. M., Lee, Y. P., Grandjean, F., Croft, L. J., and Austin, C. M. (2020a). A Giant Genome for a Giant Crayfish (*Cherax quadricarinatus*) with Insights into Cox1 Pseudogenes in Decapod Genomes. *Front. Genet.* 11, 201. doi:10.3389/fgene.2020.00201

Tang, B., Zhang, D., Li, H., Jiang, S., Zhang, H., Xuan, F., et al. (2020). Chromosome-level Genome Assembly Reveals the Unique Genome Evolution of the Swimming Crab (Portunus Trituberculatus). *Gigascience* 9 (1), 1–10. doi:10.1093/gigascience/giz161

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Curr. Protoc. Bioinformatics* 25 (1), 4–14. doi:10.1002/0471250953.bi0410s25

Ventura, T., Stewart, M. J., Chandler, J. C., Rotgans, B., Elizur, A., and Hewitt, A. W. (2019). Molecular Aspects of Eye Development and Regeneration in the Australian Redclaw Crayfish, *Cherax quadricarinatus*. *Aquacult. Fish.* 4 (1), 27–36. doi:10.1016/j.aaf.2018.04.001

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PloS one* 9 (11), e112963. doi:10.1371/journal.pone.0112963

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., et al. (2017). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* 35 (3), 543–548. doi:10.1093/molbev/msx319

Weinländer, M., and Füreder, L. (2012). Associations between Stream Habitat Characteristics and Native and Alien Crayfish Occurrence. *Hydrobiologia* 693 (1), 237–249. doi:10.1007/s10750-012-1125-x

Whitledge, G. W., and Rabeni, C. F. (1997). Energy Sources and Ecological Role of Crayfishes in an Ozark Stream: Insights from Stable Isotopes and Gut Analysis. *Can. J. Fish. Aquat. Sci.* 54 (11), 2555–2563. doi:10.1139/f97-173

Wingfield, M. (2008). An Updated Overview of the Australian Freshwater Crayfish Farming Industry. *Freshw. Crayfish* 16, 15–18. doi:10.5869/fc.2008.v16.15

Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., et al. (2018). dbCAN2: a Meta Server for Automated Carbohydrate-Active Enzyme Annotation. *Nucleic Acids Res.* 46 (W1), W95–W101. doi:10.1093/nar/gky418

Zhang, X., Yuan, J., Sun, Y., Li, S., Gao, Y., Yu, Y., et al. (2019). Penaeid Shrimp Genome Provides Insights into Benthic Adaptation and Frequent Molting. *Nat. Commun.* 10 (1), 356. doi:10.1038/s41467-018-08197-4

Zhu, B.-H., Xiao, J., Xue, W., Xu, G.-C., Sun, M.-Y., and Li, J.-T. (2018). P_RNA_scaffolder: a Fast and Accurate Genome Scaffolder Using Paired-End RNA-Sequencing Reads. *BMC genomics* 19 (1), 175. doi:10.1186/s12864-018-4567-3

Conflict of Interest: Author HMG was employed by company GeneSEQ.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

DATA REPORT
published: 19 January 2022
doi: 10.3389/fgene.2021.788547

Chromosome-Level Assembly of the Chinese Hooksnout Carp (*Opsariichthys bidens*) Genome Using PacBio Sequencing and Hi-C Technology

Xiaojun Xu[1†], *Wenzhi Guan*[1†], *Baolong Niu*[1†], *Dandan Guo*[1], *Qing-Ping Xie*[1], *Wei Zhan*[1], *Shaokui Yi*[2*] and *Bao Lou*[1*]

[1]Institute of Hydrobiology, Zhejiang Academy of Agricultural Sciences, Hangzhou, China, [2]School of Life Sciences, Huzhou University, Huzhou, China

OPEN ACCESS

Edited by:
Roger Huerlimann,
Okinawa Institute of Science and Technology Graduate University, Japan

Reviewed by:
Dong-Neng Jiang,
Guangdong Ocean University, China
Tao Zhou,
Xiamen University, China

***Correspondence:**
Shaokui Yi
yishaokui@foxmail.com
Bao Lou
loubao6577@163.com

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: 02 October 2021
Accepted: 23 December 2021
Published: 19 January 2022

Citation:
Xu X, Guan W, Niu B, Guo D, Xie Q-P, Zhan W, Yi S and Lou B (2022) Chromosome-Level Assembly of the Chinese Hooksnout Carp (Opsariichthys bidens) Genome Using PacBio Sequencing and Hi-C Technology. Front. Genet. 12:788547. doi: 10.3389/fgene.2021.788547

Keywords: ***Opsariichthys bidens***, **PacBio sequencing, Hi-C technology, chromosome-level assembly, *Opsariichthyinae***

INTRODUCTION

Chinese hooksnout carp (*Opsariichthys bidens*) is an endemic Cypriniformes minnow in East Asia, and mainly distributed in China. Notably, this common minnow has undergone a long and complex taxonomic history. In 1960s, it was classified in Cyprinidae, Leuciscinae, *Opsariichthys* (Wu, 1964). With the advances on the application of molecular characters for the fish systematics in 1990s, *O. bidens* was assigned into Cyprinidae, Danioninae, *Opsariichthys* (Chen, 1998). Subsequently, its taxonomic status was revised several times (Mayden et al., 2009; Fang et al., 2009; Tang et al., 2010, 2013; Liao et al., 2011; Stout et al., 2016; Huang et al., 2017). According to the latest phylogenetic classification of bony fishes, *O. bidens* has been assigned into Xenocyprididae, Opsariichthyinae, *Opsariichthys* (Betancur-R et al., 2017), which has been adopted by the NCBI database (www.ncbi.nlm.nih.gov/Taxonomy) and FishBase (www.fishbase.org).

For the desirable texture and flavor of the flesh, *O. bidens* has relatively high economic values. Artificial breeding of *O. bidens* began in 2008 (Jing, 2009), and the previous studies focused on the embryonic development (Jin et al., 2017), flesh nutrition content (Zhang Q. K. et al., 2019) and spermatogenesis (Tang et al., 2020) were reported in recent years. Due to the high price, disease resistance, and wide-range temperature adaptation, *O. bidens* (**Figure 1A**) has become an emerging commercial fish species.

Remarkably, *O. bidens* has obvious sex dimorphism (Lian et al., 2017). In aquaculture practice, the adult males are usually twice as large as the female siblings, and have gorgeous nuptial coloration, which brings to high ornamental property as a popular ornamental fish species. Hence, a high-quality genome sequence would facilitate the development of sex-specific markers and sex control breeding.

In this study, the chromosome-level assembly of *O. bidens* was constructed using PacBio sequencing and Hi-C technology. To the best of our knowledge, this is the sequenced genome with the largest chromosome number (2n = 78) in diploid Xenocyprididae (Arai, 2011). The genome resource will facilitate the studies of taxonomy, evolution, and genetic breeding of *O. bidens*.

Data

A total of 135.07 Gb raw data were obtained from the Illumina X Ten platform for genome size estimation. The estimated genome size of *O. bidens* is about 899.69 Mb, and the heterozygous rate of

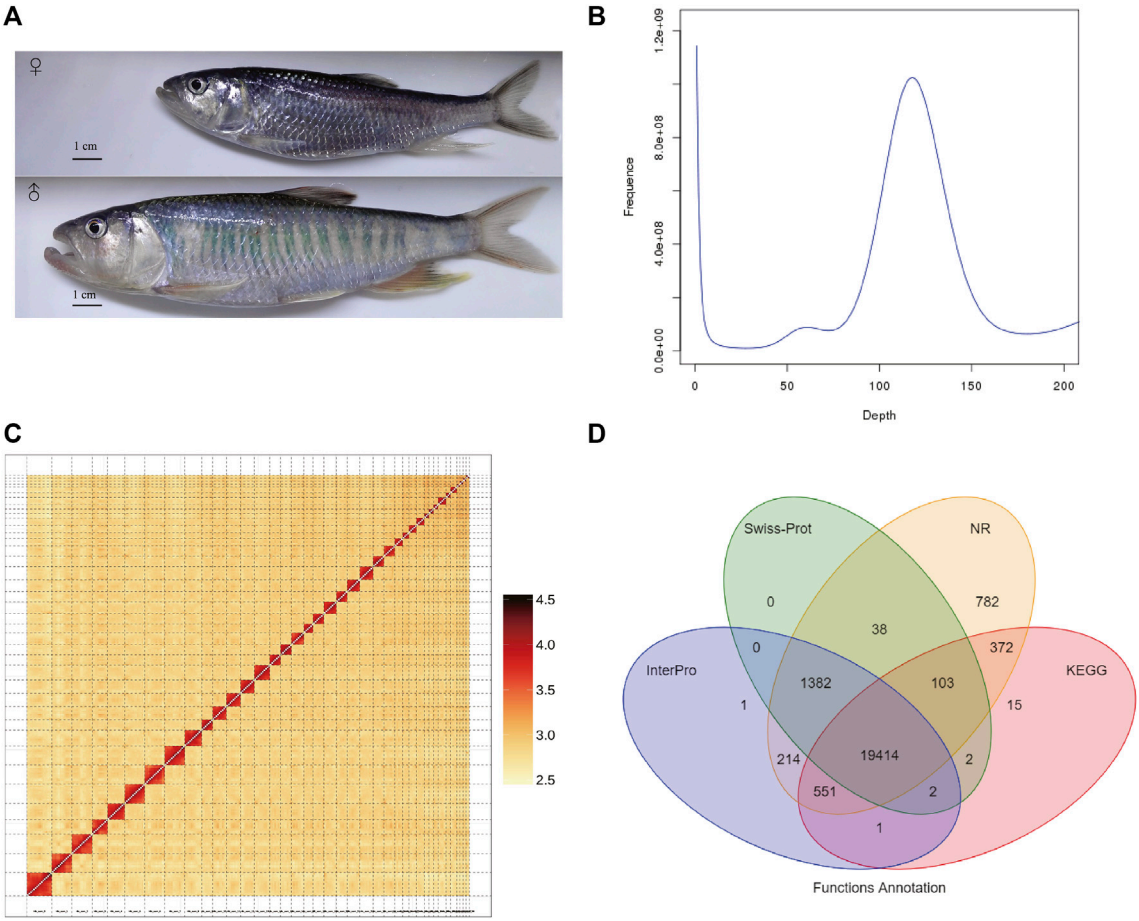

FIGURE 1 | Genome assembly of *Opsariichthys bidens*. **(A)** The photo of female and male *O. bidens*; The body length and weight of female were 112 mm and 13.3 g, respectively; The body length and weight of male were 171 mm and 56.6 g, respectively. **(B)** The Kmer (K = 17) distribution of *O. bidens* genome. **(C)** The Hi-C heatmap used for integrating the scaffolds. **(D)** The Venn graph of the numbers of annotated genes with different databases.

genome was 0.36%. (**Figure 1B**). Meanwhile, 167.17 Gb long reads were generated by PacBio Sequel platform. The average length of long reads was 21,861 bp, and the N50 of long reads was 34,896 bp. The long reads were *de novo* assembled into 403 contigs with total length of 818.75 Mb. The N50 of the assembled contigs was 4.71 Mb and the largest contigs was 22.26 Mb in length.

We used BUSCO analysis to determine the completeness of genome assembly, and the result showed that this assembled genome contained 96.6% complete BUSCOs, including 91.1% complete and single-copy BUSCOs and 5.7% complete duplicated BUSCOs. Meanwhile, the evaluation using CEGMA showed that the completeness of assembly was 97.18%. After polishing with the Illumina short reads using NextPolish (Hu et al., 2020), the total length of assembled contigs was 818.75 Mb. The N50 of these contigs was 4.66 Mb.

Subsequently, 95.64 Gb Hi-C data was generated by Illumina NovaSeq 6000 platform and used for chromosome-level assembly. After quality control of Hi-C reads with HiCUP software (Wingett et al., 2015), a total of 2,210,719 valid pairs were detected and 95.66% unique Di-tags were obtained. With the Hi-C data, 82 contigs were anchored into 39 chromosomes with a total length of 814.71 Mb (**Figure 1C**), and the length of anchored chromosomes ranged from 6.77 to 42.84 Mb. Finally, the *O. bidens* genome was 818.78 Mb in length with N50 value of 25.29 Mb. To further validate the assembly completeness, we mapped the short reads to the final assembly, and the mapping rate was 98.76%.

A total of 42.39% of the genome (347.06 Mb) were identified as repetitive elements. The most abundant transposable elements (TEs) were long terminal repeats (LTRs, 35.12% of the genome), followed by DNA transposons (4.35%) and long interspersed elements (2.19%). Meanwhile, 23,992 protein-coding genes were annotated. The mean gene length was 16,469.11 bp. The average of CDS length was 1,670.54 bp, and the average number of exons per gene was 9.79. The comparative analysis of gene prediction with other fish species was performed (Figure S1). The function annotation of these protein-coding genes showed that 95.4% were annotated by at least one of the public databases (**Figure 1D**). Meanwhile, 1.07 Mb of the genome were annotated as ncRNAs, among which miRNA, tRNA and rRNA accounted for 0.084% of the genome. We performed the BUSCO analysis with the predicted protein-coding genes, and the result showed that a total of 93.5% complete BUSCOs were present with the gene annotation.

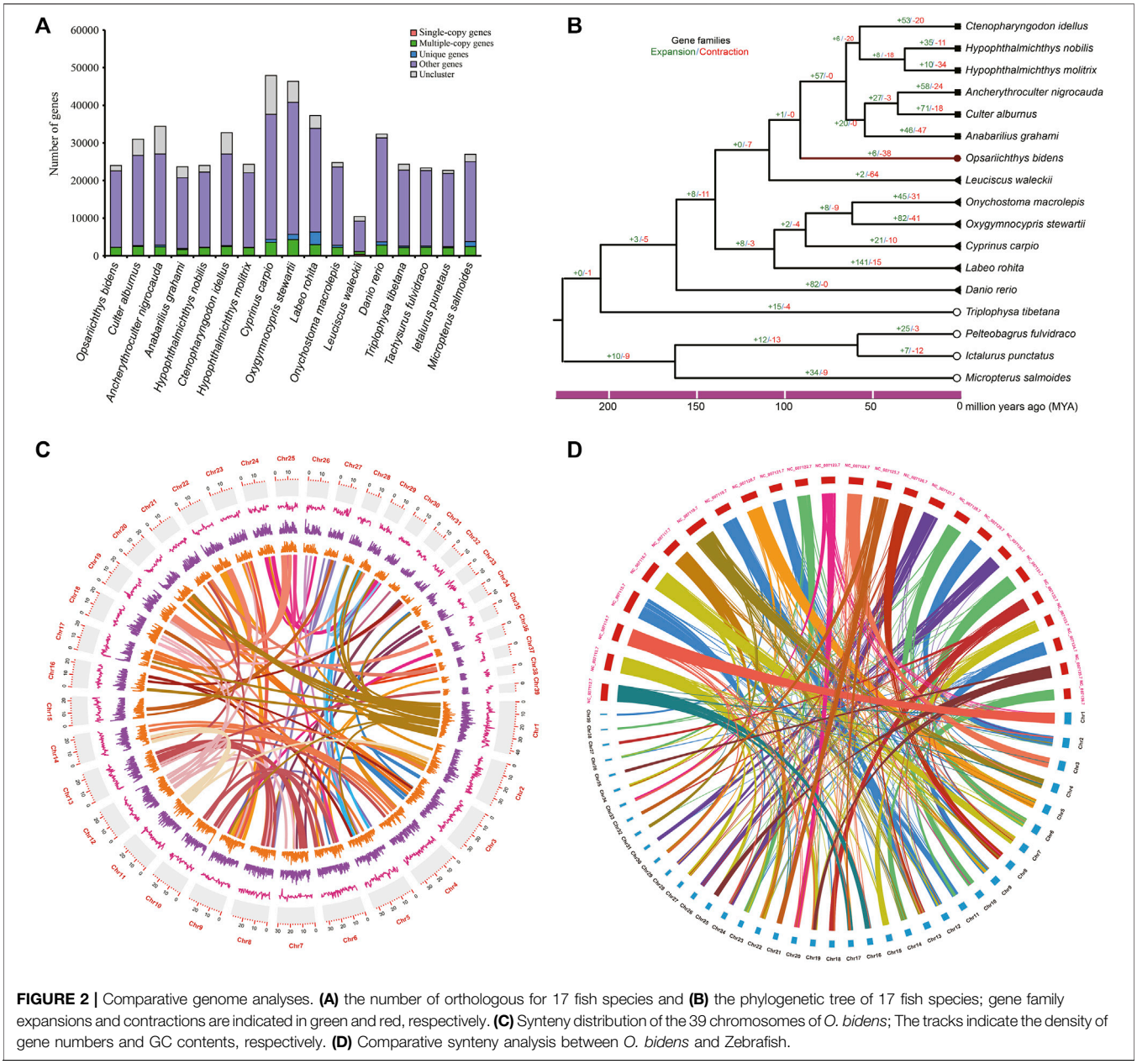

FIGURE 2 | Comparative genome analyses. **(A)** the number of orthologous for 17 fish species and **(B)** the phylogenetic tree of 17 fish species; gene family expansions and contractions are indicated in green and red, respectively. **(C)** Synteny distribution of the 39 chromosomes of *O. bidens*; The tracks indicate the density of gene numbers and GC contents, respectively. **(D)** Comparative synteny analysis between *O. bidens* and Zebrafish.

The single-copy orthologous genes of 17 fish species were identified (**Figure 2A**), and the phylogenetic tree was constructed with 170 single-copy orthologous genes (**Figure 2B**), and the result showed that *O. bidens* was grouped with the species in families of Leuciscinae and Culterinae, indicating a closer relationship with these species. A total of 6 and 38 gene families significantly expanded and contracted in *O. bidens*, respectively. The expanded and contracted gene families contained 43 and 45 genes, respectively.

To further evaluate the quality of genome assembly, we compared *O. bidens* genome with zebrafish genome. The conservation synteny among the 39 chromosomes was shown in **Figure 2C**, and a total of 1,306 blocks were detected among the chromosomes. The gene synteny between *O. bidens* and zebrafish genomes is shown in **Figure 2D**. The chromosomes of *O. bidens* exhibited high homology with the zebrafish chromosomes, and several chromosomes of zebrafish were corresponding to two mini chromosomes of *O. bidens*, indicating that the large number of chromosome of *O. bidens* may originated from the chromosomal break of ancestral chromosomes.

MATERIALS AND METHODS

Sampling, Library Construction, and Sequencing

A healthy female individual was collected from our fish base in Zhejiang Province, China. The muscle, blood, kidney, heart,

brain, liver and ovary tissues were sampled and immediately frozen and stored in liquid nitrogen until extracting the genomic DNA and total RNA. High-quality DNA samples were extracted using the DNA Isolation Reagent Kit (TaKaRa, China) from muscle tissue. DNA quality and integrity was evaluated with 1% agarose gels. Firstly, a DNA sequencing library with insert size 350 bp was constructed following the instructions of Illumina DNA Prep kit. The library was sequenced on the Illumina HiSeq X Ten System using 150 bp paired-end mode in Novogene, Co. Ltd., Beijing. Meanwhile, PacBio SMRT libraries were prepared according to the manufacturer's instructions, and the libraries were sequenced using a PacBio Sequel System. Additionally, total RNAs were extracted from the muscle, kidney, heart, brain, liver and ovary tissues using RNAiso kit (TaKaRa, China). The RNA sequencing library was constructed with the PacBio Iso-Seq Express Template Prep Kit 2.0 (Pacific Biosciences, United States) and sequenced using PacBio Sequel system. The Hi-C library was prepared from muscle tissue of the same individual following the standard protocol described previously (Belton et al., 2012). The constructed Hi-C library was sequenced with Illumina NovaSeq 6000 system.

Genome Size Estimation, Genome Assembly and Polishing

The raw data generated by Illumina platform was filtered with fastp v0.20.0 program (Chen et al., 2018). Frequencies of *K*-mers ($K = 17$) were counted using Jellyfish (Marcais and Kingsford 2012). GenomeScope v1.0 (Vurture et al., 2017) was used to estimate size, repeat content and heterozygosity of the genome with maximum *K*-mer coverage of 10,000. The genome size was calculated as: size = *K*-mer number/peak depth. The genome assembly was performed using the FALCON assembler v2.1.0 (Chin et al., 2016), and the assembled contigs were polished with Illumina reads using NextPolish v1.4.0 software (Hu et al., 2020). The assembly completeness was evaluated by Core Eukaryotic Genes Mapping Approach (CEGMA) (Parra et al., 2007) and Benchmarking Universal Single-Copy Orthologs (BUSCO) v5.2.2 software (Simão et al., 2015) using the Actinopterygii geneset (v10.0). Subsequently, the Hi-C reads were aligned to the assembly using the Juicer v1.6.2 (Durand et al., 2016a). The contigs were ordered and anchored with Hi-C data using the allhic program (Zhang X. et al., 2019), and manually adjusted using the Juicebox Assembly Tools v1.11.08 (Durand et al., 2016b).

Genome Annotation

Repetitive elements in the genome were identified using RepeatMasker (Chen, 2004) and RepeatModeler with default settings. The modeled repeats were classified into their subclasses using the Repbase v20.08 database (http://www.girinst.org/repbase/). Tandem Repeat was extracted using TRF (http://tandem.bu.edu/trf/trf.html) *ab initio* prediction. A custom library generated by a combination of Repbase and the *de novo* TE library which was processed by uclust to yield a non-redundant library was supplied to RepeatMasker for DNA-level repeat identification. Gene prediction was conducted through a combination of homology-based, *ab initio*, and transcript-based prediction methods. The full-length transcripts generated using PacBio Iso-Seq pipeline were used for transcript-based prediction. The transcripts were aligned to the genome using PASA program. Protein sequences of fish species including *Ctenopharyngodon idellus*, *Cyprinus carpio*, *Carassius auratus*, *Danio rerio*, and *Onychostoma macrolepis* were used as queries to search against the genome using tBLASTN. A *de novo* gene prediction was performed with Augustus v3.2.3 (Stanke et al., 2006), GlimmerHMM v3.04 (Majoros et al., 2004) and SNAP (Korf, 2004). The gene model was predicted by combination of three methods with EvidenceModeler v1.1.1 (Haas et al., 2008). Gene functional annotation was performed by aligning predicted protein-coding genes to the public databases using BLASTP and InterProScan70 v5.31 (Mulder and Apweiler, 2007), including NCBI NR, Swiss-prot, Pfam, Gene Ontology (GO), InterPro, and Kyoto Encyclopedia of Genes and Genomes (KEGG).

Phylogenetic Analysis and Species Divergence Time Estimation

To investigate the phylogenetic status of *O. bidens*, we retrieved genome data of 16 fish species, including *Cyprinus carpio* (GenBank: GCA_000951,615.2), *Ictalurus punctatus* (GenBank: GCA_001660625.1), *Danio rerio* (GenBank: GCA_000002035.4), *Anchorythroculter nigrocauda* (NGDC: GWHAAZV00000000), *Micropterus salmoides* (GenBank: GCA_014851395.1), *Pelteobagrus fulvidraco* (GenBank: GCA_003724035.1), *Hypophthalmichthys molitrix* (CNGB：CNP0000974), *Hypophthalmichthys nobilis* (CNGB：CNP0000974), *Culter alburnus* (GenBank: GCA_009869775.1), *Oxygymnocypris stewartii* (GenBank: GCA_003573665.1), *Anabarilius grahami* (GenBank: GCA_003731715.1), *Labeo rohita* (GenBank: GCA_017311145.1), *Onychostoma macrolepis* (GenBank: GCA_012432095.1), *Leuciscus waleckii* (GenBank: GCA_900092035.1), *Triplophysa tibetana* (GenBank: GCA_008369825.1), and *Ctenopharyngodon idellus* (http://www.ncgr.ac.cn/grasscarp/) from public databases. All-to-all BLASTP was employed to identity the similarities among filtered protein sequences in these species with an E-value cutoff of 1e−5. We identified orthologous gene clusters using the OrthoMCL pipeline (Li et al., 2003). Protein sequences from the single-copy gene families were used for phylogenetic tree reconstruction. MUSCLE (Edgar, 2004) was used to generate multiple sequence alignments for protein sequences with default parameters, and the ambiguously aligned positions were trimmed using Gblocks (http://molevol.cmima.csic.es/castresana/Gblocks.html). The alignments of each family were concatenated to a super alignment matrix. The alignment matrix was used for phylogenetic tree reconstruction through maximum likelihood methods. The phylogenetic tree was constructed using RAxML v7.2.9 (Stamatakis, 2014) with 1,000 bootstrap replicates.

Divergence time between species was estimated using MCMCtree with model of JC69 in PAML (Yang, 2007). The divergence time calibration of *Oxygymnocypris stewartii* and *Cyprinus carpio* were obtained from the TimeTree website (http://www.timetree.org/). The likelihood analysis for gene gain and gene loss was identified using CAFE v4.2 (De Bie et al., 2006) with $p < 0.05$.

Synteny Analysis

Synteny analysis of intra-genome was carried out using the MCScanX pipeline (Wang et al., 2012), output were converted to blocks by in-house Perl scripts. Circos (Krzywinski et al., 2009) was used to display the syntenic blocks. We identified syntenic blocks of genes between *O. bidens* and *D. rerio*. For the comparison, we carried out an all-to-all BLAST search of annotated protein sequences and ran MCScanX with the parameters "-s 10 -b 2".

DATA AVAILABILITY STATEMENT

The sequences of genome assembly are available in the National Genomics Data Center (NGDC) with accession number GWHBEIO00000000. The newick file of phylogenetic tree generated by RAxML is available in figShare with doi: https://figshare.com/articles/dataset/phylogenetic_tree_generated_by_RAxML/17085437/1. The karyotype image is available in figShare with doi: https://figshare.com/articles/figure/karyotype_image_of_O_bidens/17161865/1.

ETHICS STATEMENT

The animal study was reviewed and approved by the Animal ethics committee of ZheJiang Academy Of Agricultural Sciences.

AUTHOR CONTRIBUTIONS

XX and BL conceived the study. WG, DG, and WZ collected samples. SY and BN performed the bioinformatics analyses. XX and SY wrote the manuscript. Q-PX revised the manuscript. XX, WG, and BN contributed equally to this work. All authors read and approved the final manuscript.

FUNDING

This study was financially supported by grants of Zhejiang provincial Department of Science and Technology (No.2020C02014).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.788547/full#supplementary-material

Supplementary Figure S1 | The comparative analyses of CDS length, exon length, exon number, gene length and intron length with 6 species. Cau, Cca, Cid, Dre, makouyu and Oma indicate Carassius auratus, Cyprinus carpio, Ctenopharyngodon idellus, Danio rerio, Opsariichthys bidens, and Onychostoma macrolepis, respectively.

REFERENCES

Arai, R. (2011). *Fish Karyotypes: A Check List*. Tokyo: Springer.

Belton, J.-M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: A Comprehensive Technique to Capture the Conformation of Genomes. *Methods* 58, 268–276. doi:10.1016/j.ymeth.2012.05.001

Betancur-R, R., Wiley, E. O., Arratia, G., Acero, A., Bailly, N., Miya, M., et al. (2017). Phylogenetic Classification of Bony Fishes. *BMC Evol. Biol.* 17, 162–201. doi:10.1186/s12862-017-0958-3

Chen, N. (2004). Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Curr. Protoc. Bioinformatics* 5 (4.10), 1–4. doi:10.1002/0471250953.bi0410s05

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). Fastp: An Ultra-fast All-In-One FASTQ Preprocessor. *Bioinformatics* 34, i884–i890. doi:10.1093/bioinformatics/bty560

Chen, Y. Y. (1998). *Fauna Sinica Osteichthys Cypriniformes II*. Beijing: Science Press.

Chin, C. S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., et al. (2016). Phased Diploid Genome Assembly With Single-Molecule Real-Time Sequencing. *Nat. Methods.* 13 (12), 1050–1054. doi:10.1038/nmeth.4035

De Bie, T., Cristianini, N., Demuth, J. P., and Hahn, M. W. (2006). CAFE: a Computational Tool for the Study of Gene Family Evolution. *Bioinformatics* 22, 1269–1271. doi:10.1093/bioinformatics/btl097

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., et al. (2016a). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cel Syst.* 3, 99–101. doi:10.1016/j.cels.2015.07.012

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., et al. (2016b). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cel Syst.* 3, 95–98. doi:10.1016/j.cels.2016.07.002

Edgar, R. C. (2004). MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Res.* 32, 1792–1797. doi:10.2460/ajvr.69.1.8210.1093/nar/gkh340

Fang, F., Norén, M., Liao, T. Y., Källersjö, M., and Kullander, S. O. (2009). Molecular Phylogenetic Interrelationships of the South Asian Cyprinid genera Danio, Devario and Microrasbora (Teleostei, Cyprinidae, Danioninae). *Zool. Scr.* 38, 237–256. doi:10.1111/j.1463-6409.2008.00373

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated Eukaryotic Gene Structure Annotation Using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* 9, R7. doi:10.1186/gb-2008-9-1-r7

Hu, J., Fan, J., Sun, Z., and Liu, S. (2020). NextPolish: A Fast and Efficient Genome Polishing Tool for Long-Read Assembly. *Bioinformatics* 36, 2253–2255. doi:10.1093/bioinformatics/btz891

Huang, S. P., Wang, F. Y., and Wang, T. Y. (2017). Molecular Phylogeny of the opsariichthys Group (Teleostei: Cypriniformes) Based on Complete Mitochondrial Genomes. *Zool. Stud.* 56, e40–52. doi:10.6620/ZS.2017.56-40

Jin, D. L., Zhang, Q. K., Wang, Y. F., Zhu, Y. M., Zhang, Y. M., Wang, J. P., et al. (2017). Observation of Embryonic, Larva and Juvenile Development of *Opsariichthys bidens*. *Oceanol. Limnol. Sin.* 48 (04), 838–847. doi:10.11693/hyhz20170200034

Jing, J. T. (2009). Artificial Breeding and Aquculture experiment of *Opsariichthys bidens* of the Yalu River. *China Fish.* 6, 32–34. doi:10.3969/j.issn.1002-6681.2009.06.017

Korf, I. (2004). Gene Finding in Novel Genomes. *BMC bioinformatics* 5, 59–9. doi:10.1186/1471-2105-5-59

Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an Information Aesthetic for Comparative Genomics. *Genome Res.* 19, 1639–1645. doi:10.1101/gr.092759.109

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* 13, 2178–2189. doi:10.1101/gr.1224503

Lian, Q. P., Mi, G. Q., and Liu, S. L. (2017). Sexual Dimorphism in Morphological Traits of *Opsariichthys bidens*. *XianDai NongYe KeJi* 22, 226–229. doi:10.3969/j.issn.1007-5739.2017.22.126

Liao, T.-Y., Kullander, S. O., and Fang, F. (2011). Phylogenetic Position of Rasborin Cyprinids and Monophyly of Major Lineages Among the Danioninae, Based on Morphological Characters (Cypriniformes: Cyprinidae). *J. Zool. Syst. Evol. Res.* 49, 224–232. doi:10.1111/j.1439-0469.2011.00621.x

Majoros, W. H., Pertea, M., and Salzberg, S. L. (2004). TigrScan and GlimmerHMM: Two Open Source Ab Initio Eukaryotic Gene-Finders. *Bioinformatics* 20, 2878–2879. doi:10.1101/gr.122450310.1093/bioinformatics/bth315

Marcais, G., and Kingsford, C. (2012). Jellyfish: A Fast K-Mer Counter. *Tutorialis e Manuais* 1, 1–8.

Mayden, R. L., Chen, W.-J., Bart, H. L., Doosey, M. H., Simons, A. M., Tang, K. L., et al. (2009). Reconstructing the Phylogenetic Relationships of the Earth's Most Diverse Clade of Freshwater Fishes-Order Cypriniformes (Actinopterygii: Ostariophysi): A Case Study Using Multiple Nuclear Loci and the Mitochondrial Genome. *Mol. Phylogenet. Evol.* 51, 500–514. doi:10.1016/j.ympev.2008.12.015

Mulder, N., and Apweiler, R. (2007). InterPro and InterProScan. *Methods Mol. Biol.* 396, 59–70. doi:10.1007/978-1-59745-515-2_5

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a Pipeline to Accurately Annotate Core Genes in Eukaryotic Genomes. *Bioinformatics* 23, 1061–1067. doi:10.1093/bioinformatics/btm071

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs. *Bioinformatics* 31, 3210–3212. doi:10.1093/bioinformatics/btv351

Stamatakis, A. (2014). RAxML Version 8: a Tool for Phylogenetic Analysis and post-analysis of Large Phylogenies. *Bioinformatics* 30, 1312–1313. doi:10.1093/bioinformatics/btu033

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: Ab Initio Prediction of Alternative Transcripts. *Nucleic Acids Res.* 34, W435–W439. doi:10.1093/nar/gkl200

Stout, C. C., Tan, M., Lemmon, A. R., Lemmon, E. M., and Armbruster, J. W. (2016). Resolving Cypriniformes Relationships Using an Anchored Enrichment Approach. *BMC Evol. Biol.* 16, 244–256. doi:10.1186/s12862-016-0819-5

Tang, D., Gao, X., Lin, C., Feng, B., Hou, C., Zhu, J., et al. (2020). Cytological Features of Spermatogenesis in *Opsariichthys Bidens* (Teleostei, Cyprinidae). *Anim. Reprod. Sci.* 222, 106608. doi:10.1016/j.anireprosci.2020.106608

Tang, K. L., Agnew, M. K., Hirt, M. V., Lumbantobing, D. N., Raley, M. E., Sado, T., et al. (2013). Limits and Phylogenetic Relationships of East Asian Fishes in the Subfamily Oxygastrinae (Teleostei: Cypriniformes: Cyprinidae). *Zootaxa* 3681, 101–135. doi:10.11646/zootaxa.3681.2.1

Tang, K. L., Agnew, M. K., Hirt, M. V., Sado, T., Schneider, L. M., Freyhof, J., et al. (2010). Systematics of the Subfamily Danioninae (Teleostei: Cypriniformes: Cyprinidae). *Mol. Phylogenet. Evol.* 57, 189–214. doi:10.1016/j.ympev.2010.05.021

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). GenomeScope: Fast Reference-free Genome Profiling from Short Reads. *Bioinformatics* 33, 2202–2204. doi:10.1093/bioinformatics/btx153

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a Toolkit for Detection and Evolutionary Analysis of Gene Synteny and Collinearity. *Nucleic Acids Res.* 40, e49. doi:10.1093/nar/gkr1293

Wingett, S. W., Ewels, P., Furlan-Magaril, M., Nagano, T., Schoenfelder, S., Fraser, P., et al. (2015). HiCUP: Pipeline for Mapping and Processing Hi-C Data. *F1000Res* 4, 1310. doi:10.12688/f1000research.7334.1.

Wu, X. W. (1964). *Cyprimid Fishes in China*. Shanghai: Shanghai Science and Technology Press

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi:10.1093/molbev/msm088

Zhang, Q. K., Zheng, X. B., Tang, D. J., Zhu, Y. M., Wang, J. P., and Zhu, J. Q. (2019). Analysis and Evaluation of Nutritional Components in Muscle of Cultured *Opsariichthys bidens*. *J. Ningbo Univ.* 32 (04), 15–19. doi:10.3969/j.issn.1001-5132.2019.04.003

Zhang, X., Zhang, S., Zhao, Q., Ming, R., and Tang, H. (2019). Assembly of Allele-Aware, Chromosomal-Scale Autopolyploid Genomes Based on Hi-C Data. *Nat. Plants* 5, 833–845. doi:10.1038/s41477-019-0487-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

DATA REPORT
published: 19 January 2022
doi: 10.3389/fgene.2021.825742

A Chromosome-Level Genome Assembly of Yellowtail Kingfish (*Seriola lalandi*)

Shuo Li[1,2,3], *Kaiqiang Liu*[1,2], *Aijun Cui*[1,2], *Xiancai Hao*[1], *Bin Wang*[1,2], *Hong-Yan Wang*[1,2], *Yan Jiang*[1,2], *Qian Wang*[1,2], *Bo Feng*[1], *Yongjiang Xu*[1,2]*, *Changwei Shao*[1,2]* and *Xuezhou Liu*[1,2]

[1]*Key Laboratory of Sustainable Development of Marine Fisheries, Ministry of Agriculture and Rural Affairs, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao, China,* [2]*Laboratory for Marine Fisheries Science and Food Production Processes, Qingdao National Laboratory for Marine Science and Technology, Qingdao, China,* [3]*China State Key Laboratory for Managing Biotic and Chemical Threats to the Quality and Safety of Agroproducts, Ningbo University, Ningbo, China*

Yellowtail kingfish (*Seriola lalandi*) is a pelagic marine piscivore with a circumglobal distribution. It is particularly suitable for open ocean aquaculture owing to its large body size, fast swimming, rapid growth, and high economic value. A high-precision genome is of great significance for future genetic breeding research and large-scale aquaculture in the open ocean. PacBio, Illumina, and Hi-C data were combined to assemble chromosome-level reference genome with the size of 648.34 Mb (contig N50: 28.52 Mb). 175 contigs was anchored onto 24 chromosomes with lengths ranging from 12.28 to 34.59 Mb, and 99.79% of the whole genome sequence was covered. The BUSCOs of genome and gene were 94.20 and 95.70%, respectively. Gene families associated with adaptive behaviors, such as olfactory receptors and HSP70 gene families, expanded in the genome of *S. lalandi*. An analysis of selection pressure revealed 652 fast-evolving genes, among which *mkxb*, *popdc2*, *dlx6*, and *ifitm5* may be related to rapid growth traits. The data generated in this study provide a valuable resource for understanding the genetic basis of *S. lalandi* traits.

Keywords: ***Seriola lalandi*, genome, adaptation, rapid growth, aquaculture**

OPEN ACCESS

Edited by:
Roger Huerlimann, Okinawa Institute of Science and Technology Graduate University, Japan

Reviewed by:
Zhenhua Ma, Chinese Academy of Fishery Sciences (CAFS), China
Qiong Shi, Beijing Genomics Institute (BGI), China

***Correspondence:**
Yongjiang Xu
xuyj@ysfri.ac.cn
Changwei Shao
shaocw@ysfri.ac.cn

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *30 November 2021*
Accepted: *22 December 2021*
Published: *19 January 2022*

Citation:
*Li S, Liu K, Cui A, Hao X, Wang B, Wang H-Y Jiang Y, Wang Q, Feng B, Xu Y, Shao C and Liu X (2022) A Chromosome-Level Genome Assembly of Yellowtail Kingfish (*Seriola lalandi*). Front. Genet. 12:825742. doi: 10.3389/fgene.2021.825742*

INTRODUCTION

To develop environmentally friendly and economically sustainable aquaculture, it is necessary to understand the genetic basis of traits that currently limit/enhance development of domestic aquaculture (Rondeau et al., 2013). Genetic resources have been developed and widely used in agriculture and animal husbandry for decades, but only recently have they been used in selected aquaculture species (Ozaki et al., 2013; Dunham et al., 2014). There is still limited information on genetic variation on commercially important traits (Peterson et al., 2020). The methods used to develop these resources offer the best possibilities for genetic improvement or culture practices

Abbreviations: N50, median size; Gb, gigabase pairs; Mb, megabase pairs; kb, kilobase pairs; PacBio, Pacific Biosciences; SMRT, single molecule, real-time; BUSCOs, Benchmarking Universal Single-Copy Orthologs; GC, guanine-cytosine; GO, Gene Ontogeny; KEGG, Kyoto Encyclopedia of Genes and Genomes; Mb, megabase pairs; ML, maximum likelihood; MYA, million years ago; NCBI, national center for biotechnology information; HSP70, heat shock protein 70 family.

(Sodeland et al., 2013). Third-Generation Sequencing (TGS) has improved this area of research through high quality assemblies and decreasing costs, and this has enabled development of genetic resources for a greater number of species (Huete-Pérez and Quezada, 2013; Lee et al., 2016; Lv et al., 2020).

Yellowtail kingfish (*Seriola lalandi*) is an excellent marine economic fish. It has a number of beneficial traits for open ocean aquaculture systems, including large body size, rapid growth, and high-quality flesh (Orellana et al., 2014; Sanchís-Benlloch et al., 2017). Similar in taste to tuna or mackerel, yellowtail kingfish have a large market worldwide and are a popular fish used in sushi (Purcell et al., 2015). These make them a good candidate for aquaculture. Since the 1990s, extensive research in Japan has focused on artificial breeding and breeding technology for *S. lalandi* (Sano, 1998). In China, aquaculture of *S. lalandi* began in 2001 (Jiang et al., 2001), along with biological research, including studies of embryogenesis, seedling cultivation, and effects of salinity stress on growth (Shi et al., 2019; Xu et al., 2019; Liu A et al., 2021).

Here, we report a chromosomal-level genome assembly of *S. lalandi*. Our evolutionary and comparative genomic analysis provide insights into the adaptability of the species to the external environment. Furthermore, the genome analysis provide a valuable resource for further studies of the genetic basis of traits of *S. lalandi*.

Value of the Data

This is the first chromosomal-level genome assembly in *Seriola* genus. It could be a valuable resource to conduct a comparative analysis among the species in the genome of the *Seriola* genus and for further studies of the genetic basis of traits of *S. lalandi*.

MATERIALS AND METHODS

Sampling and Sequencing

Yellowtail kingfish specimens were collected from Dalian Futai Marine Products Farming Co., Ltd. (Dalian, China). Total genomic DNA of a male fish muscle sample was extracted using the QIAamp DNA Mini Kit (QIAGEN, Hilden, Germany) following the manufacturer's protocols. We constructed two paired-end libraries (insert sizes of 200 and 500 bp) following the manufacturer's protocol (Chromium Genome v1, PN120229). The libraries were sequenced on the BGISEQ-500 platform to obtain PE 2 × 150 bp reads. The extracted DNAs were also used to construct a 20 kb library following the PacBio protocol (Pacific Biosciences, Menlo Park, CA, United States). The libraries were then sequenced on the PacBio Sequel platform. We obtained 48.74 and 106.76 Gb of raw sequence data using the BGISEQ-500 and PacBio platforms, respectively (**Supplementary Table 1**).

To construct chromosome-level assemblies, the Hi-C technique was used. A Hi-C library was prepared following the strategy described by Rao et al. (Rao et al., 2014) using blood samples with an ~300 bp insert size. Using the BGISEQ-500 platform to sequence the Hi-C library, we obtained 87.60 Gb of raw Hi-C data (**Supplementary Table 1**).

Four tissues (brain, pituitary, liver, and muscle) were collected for RNA sequencing. RNA from each tissue was extracted and treated with DNase I (TAKARA, Kusatsu, Japan) to remove genomic DNA. For each tissue, a paired-end RNA-sequencing library was constructed with an insert size of 300 bp and then sequenced on the Illumina HiSeq 2,500 platform to generate PE 2 × 150 bp. One muscle specimen was also used to construct an Iso-Seq library and then sequenced on the PacBio Sequel platform. In total, we obtained 307.14 and 26.89 Gb of raw sequence data using the Illumina HiSeq 2,500 and PacBio platforms, respectively (**Supplementary Table 1**).

Genome Assembly, Chromosome Anchoring, and Genome Annotation

Before genome assembly, we estimated the genome by a k-mer analysis using Jellyfish v2.2.6 (Marçais and Kingsford, 2011). For this, a series of k-mers (17, 19, and 21) were extracted from the BGISEQ-500 sequencing data and the frequency of each kmer was calculated. The heterozygosity rate was estimated using 17-mers using GenomeScope v2.0.0 (**Supplementary Figure 1**). Considering the C-value (0.7) from the Animal Genome Size Database, the estimated genome size of *S. lalandi* was 684.60 Mb.

Canu v1.8 was used for the self-correction of long reads sequenced with the PacBio Sequel platform. Then, the corrected reads were assembled using wtdbg2 v2.5 (options: -x rs -g 750 m) (Ruan and Li, 2020). Pilon v1.23 (Walker et al., 2014) was used to polish contigs with short reads by three rounds of alignment. The Hi-C short reads were aligned to the scafiolds using Juicer (Durand et al., 2016) and anchoring was performed using 3D-DNA v180419 (Dudchenko et al., 2017). We finally used Juicebox Assembly Tools v1.9.9 (Durand et al., 2016) to correct the connections. The completeness of the final assembly was assessed using BUSCO v.4.0 (Simão et al., 2015).

Both homology-based and *de novo* predictions were used to annotate repetitive sequences. Transposable elements were identified using RepeatMasker v4.0.7 (http://www.repeatmasker.org) and RepeatProteinMask v1.36 with Repbase v17.01 (Bao et al., 2015). A *de novo* transposable element library was constructed using RepeatModeler v1.0.11 (http://www. repeatmasker.org/RepeatModeler.html) and was then used to predict repeats using RepeatMasker.

To annotate gene structures, we used homology-based prediction, transcriptome-based prediction, and *de novo* prediction. For homology-based annotation, the protein sequences of eight teleost species downloaded from NCBI, including *Seriola lalandi dorsalis*, *Seriola dumerili, Seriola quinqueradiata*, *Seriola rivoliana*, *Echeneis naucrates*, *Oryzias latipes*, *Danio rerio*, and *Takifugu rubripes*, were aligned to the genome assembly by BLAT v3.6 (Kent, 2002), and then GENEWISE v2.4.0 (Birney et al., 2004) was used to predict gene structures. For next-generation

RNA-sequencing annotation, data were aligned to the genome assembly using HISAT2 v2.1.0 (Kim et al., 2015) and the alignments were fed to StringTie v1.3.5 (Pertea et al., 2015) to assemble the transcriptome. TransDecoder v5.0.2 (https://github.com/TransDecoder/TransDecoder/) was used to predicate ORFs and identify candidate gene structures. For third-generation RNA-sequencing annotation, long-read RNA-seq (PacBio Iso-Seq) transcripts were obtained by removing the redundant sequences using cd-hit-est v4.8.1 (Li and Godzik, 2006). Then, the non-redundant transcripts were mapped to the genome by BLAT and assembled using PASA v2.0.2 (https://github.com/PASApipeline/PASApipeline/). For *de novo* prediction, the gene structures were analyzed on the repeat-masked genome assembly using AUGUSTUS v2.5.5 (Stanke et al., 2006), GlimmerHMM v3.0.4 (Allen et al., 2006), and GENSCAN (Burge and Karlin, 1998). Finally, genes predicted from the above methods were merged to obtain a consensus gene set using Evidence Modeler (EVM). For the functional annotation of the gene sets, the protein sequences of these genes were aligned against sequences in public protein databases, including, NR, KEGG, SwissProt, GO, InterPro, and Trembl, to identify homologues using Blastp v2.2.26 with an E-value cutoff of 1e-5.

Phylogenetic Analysis and Gene Family Expansion

To determine single-copy genes of *S. lalandi* and other species (*S. dumerili*, *S. quinqueradiata*, *S. rivoliana*, *E. naucrates*, *O. latipes*, *D. rerio*, *T. rubripes*, *Larimichthys crocea*, *Oreochromis niloticus*, and *Caranx melampygus*), the TreeFam pipeline (Li et al., 2006) was used. Before generating the alignment, the longest transcript of each gene was selected and protein sequences shorter than 50 amino acids were filtered out. Then, Blastp searches were performed for all protein sequences with an E-value cut-off of 1e-5, and fragmented alignments were merged using SOLAR. Hcluster was used to filter segments, group genes, and determine single-copy orthologue families. The phylogenetic tree was inferred using multiple alignments from the single-copy genes using RaxML-ng v0.9.0 (Kozlov et al., 2019) under the site-heterogeneous GTR + G4 model with maximum likelihood estimation (ML).

An ultrametric tree was inferred using r8s v1.71 with fossil records from the TimeTree website (http://www.timetree.org) for calibration. An MCMCTREE analysis implemented in PAML v4.5 (Yang, 1997) was employed to estimate divergence times. CAFÉ v5.0 (De Bie et al., 2006) was used to assess gene family size dynamics, and families with $p < 0.05$ showed significant expansion or contraction. GO and KEGG pathway enrichment analysis were used to analyze the expanded and contracted genes.

Positive Selection Analysis

To identify positively selected genes (PSGs), we re-determined single-copy orthologues shared among five species (*E. naucrates*, *T. rubripes*, *O. latipes*, *D. rerio,* and *S. lalandi*) and constructed a phylogenetic tree. Based on the new phylogenetic tree and single-copy genes, we estimated the rate ratio (ω) of non-synonymous to synonymous nucleotide substitutions using CodeML (PAML package) to examine selective constraint. After obtaining high-quality alignments using prank v.100802 (Löytynoja and Goldman, 2010), Gblocks v0.91b (Castresana, 2000) was used to eliminate poorly aligned positions and divergent regions. Finally, the signature of positive selection ($d_N/d_S > 1$) was identified using the PAML branch site model. GO and KEGG pathway enrichment analysis were used to evaluate PSGs.

RESULTS AND DISCUSSION

Genome Features

To generate a high-quality reference genome, we combined PacBio, Illumina, and Hi-C data (**Supplementary Table 1**). PacBio CLRs with coverage of 165 × were used for genome assembly. The draft assembly was 648.34 Mb, with 277 contigs, a contig N50 of 28.52 Mb, and a GC content of 40.79% (**Supplementary Table 2**). Using ~52 Gb (~87×) of valid Hi-C data, we anchored 175 contigs onto 24 chromosomes (**Figure 1A**, **Supplementary Figure 2**) (Shi et al., 2017). The lengths of the 24 chromosomes ranged from 12.28 to 34.59 Mb, and 99.79% of the whole genome sequence was covered (**Supplementary Table 3**). To evaluate the completeness of the assembly, the BUSCO database (actinopterygii_odb10) and RNA-seq data were used. The genome contained 94.20% complete BUSCOs and the average mapping rate of transcriptome data was 96.30% (**Supplementary Table 4**). The published *Trachinotus ovatus* chromosome-level genome was used to validate the accuracy of the assembly of the chromosomes (Zhang et al., 2019); 567.01 MB synteny blocks (each synteny block > 500 bp) were consistent with the assembled chromosomes (**Figure 1B**).

Repetitive elements comprised 22.46% of the *S. lalandi* genome, similar to the estimate in the *T. ovatus* genome (20.25%, 655 Mb) (Zhang et al., 2019). The most abundant transposable elements (TEs) were DNA transposons (11.51%), followed by long terminal repeats (LTRs, 4.93%) and long interspersed elements (LINEs, 3.85%) (**Supplementary Figure 4**). We integrated *de novo*, homology-based and transcriptome-based methods to predict a protein-coding gene set comprising 22,674 genes (**Supplementary Table 5**), and which 20,568 (90.71%) matched entries in a public database (**Supplementary Table 6**). We identified 95.70% complete BUSCOs from 22,674 protein-coding genes.

Phylogenetic Relationships and Genomic Comparison

We constructed a phylogenetic tree of *S. lalandi* and 10 teleost fish (*S. dumerili*, *S. quinqueradiata*, *S. rivoliana*, *E. naucrates*, *O. latipes*, *D. rerio*, *T. rubripes*, *L. crocea*, *O. niloticus*, and *C. melampygus*) based on 5,067 single-copy genes (**Figure 2A**, **Supplementary Table 7**). According to the phylogeny and the

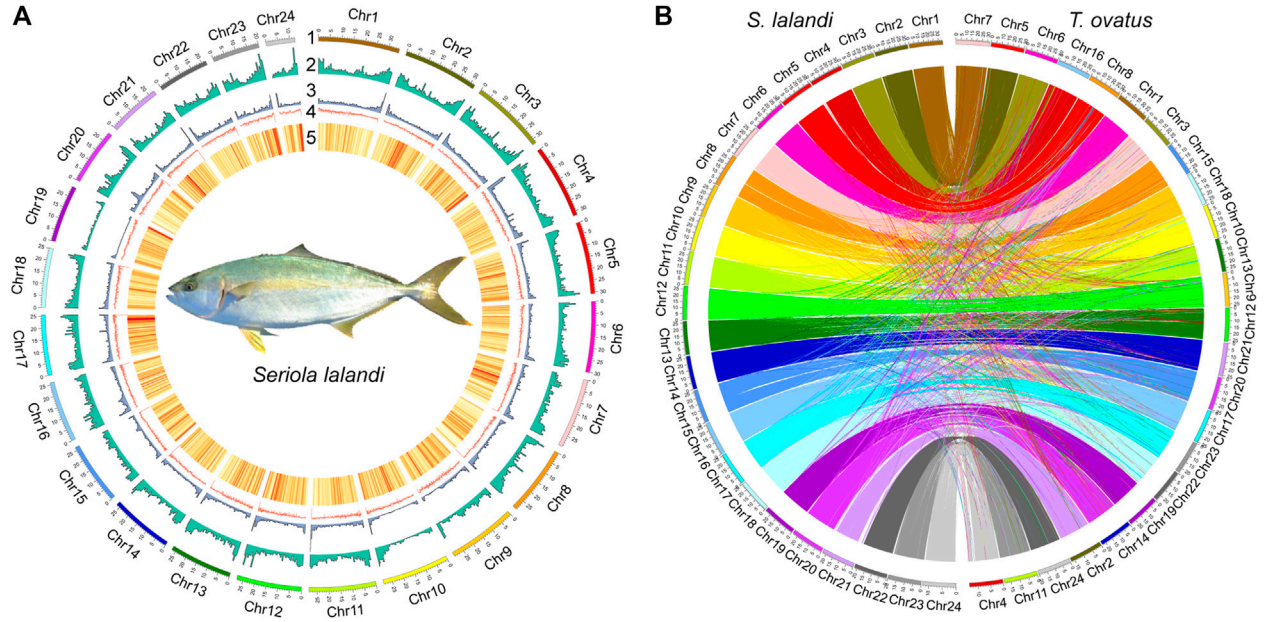

FIGURE 1 | Genome assembly and comparison. **(A)** Circos graph of genome statistics. Genomic features. From outer to inner circles: 1, represents chromosomes; 2, distribution of DNA transposons; 3, distribution of retrotransposons; 4, GC content; 5, gene distribution density; 6, each line joins paralogous genes at different chromosomes. 2–5 are drawn with 500 Kb sliding windows. **(B)** Genome comparison between *S. lalandi* and *T. ovatus*. The *S. lalandi* chromosomes are on the left, and the *T. ovatus* chromosomes are on the left.

fossil record of teleosts, we dated the divergence of *Seriola* from the other teleost species to approximately 72.6 million years ago (**Figure 2A**).

We detected 56 significantly expanded and 1,073 significantly contracted gene families ($p < 0.05$) in *S. lalandi* (**Figure 2A**). Compared with teleost fish except of *Seriola*, the HSP70 family with 19 HSP70 genes was expanded (**Supplementary Table 8**). We found five *hspa12* genes, including *hspa12a*, *hspa12b*, *hspa12l-1*, *hspa12l-2*, and *hspa12l-3*, which was more than observed in *E. naucrates* (2), *O. latipes* (2), *D. rerio* (2), *T. rubripes* (2), *C. melampygus* (3), and *L. crocea* (3) (Liu et al., 2019). In *S. lalandi*, there were three *hspa12l* gene copies. HSP70 is a well-known stress protein (Clark and Peck, 2009), and the expansion of the HSP70 family in *S. lalandi* may contribute to its adaptation to changes in the aquatic environment.

Yellowtail kingfish is a migratory marine fish with high olfactory sensitivity (Martínez-Montaño et al., 2016). We identified 147 olfactory receptor (OR)-like genes from the *S. lalandi* genome, including subfamily "Delta" (68), "Eta" (49), "Zeta" (12), "Epsilon" (9), "Beta" (6), "Thet" (2), and "Kappa" (1) (**Supplementary Table 9**). The expanded subfamilies "Delta" and "Epsilon" are important for the perception of water-soluble odorants (Cong et al., 2019). Most teleosts possess one or two "Beta" OR genes, which are important for detecting both water-soluble and airborne odorants (Liu H et al., 2021). However, subfamily "Beta" of olfactory receptor was expanded in *S. lalandi*. These expansions may contribute to the olfactory detection ability of the species, which could be useful for feeding and migration (Bett and Hinch, 2016).

Fast-Evolving Genes in Yellowtail Kingfish

PSGs are often associated with adaptive evolution and may contribute to new or improved functions. To understand the selective pressure operating on *S. lalandi*, we compared the orthologues of five teleost species (*E. naucrates*, *T. rubripes*, *O. latipes*, *D. rerio*, and *S. lalandi*) and identified 652 fast-evolving genes, including 148 PSGs ($d_N/d_S > 1$) and 504 genes that contain positively selected sites in *S. lalandi* (**Supplementary Table 10**). Consistent with the large body size and fast swimming ability, an enrichment analysis revealed that the PSGs were involved in striated muscle tissue development (GO:0014706), regulation of actin cytoskeleton (dre04810), and fatty acid metabolism (dre01212) (**Figure 2B**).

Muscle tissue development is associated with the growth rate, which is a major economic trait in animal production. Several genes involved in muscle tissue development (*klf2a*, *klhl41b*, *cdk9*, *ndrg4*, *mkxb*, and *popdc2*) showed rapid evolution in *S. lalandi* and likely contribute to the rapid growth of the species (**Supplementary Table 10**). Fast growing muscles also require increased bone support. Two genes, *dlx6* and *ifitm5,* were involved in skeletal system development and promote bone formation to support the large body (**Supplementary Table 10**). Based on the strong muscle and skeletal systems, muscle contraction-related genes (*arhgef12b*, *ramp2*, *tnnt2a*, *tnni1a*, *cald1a*, and *tnnt2a*) with positively selected sites may provide support for fast swimming (**Supplementary Table 10**). Furthermore, fatty acid metabolism-related fast-evolving genes (*hsd17b12b*, *acadm*, *mecr*, *lipg*, and *hao1*) also contributed to energy consumption and growth (**Supplementary Table 10**).

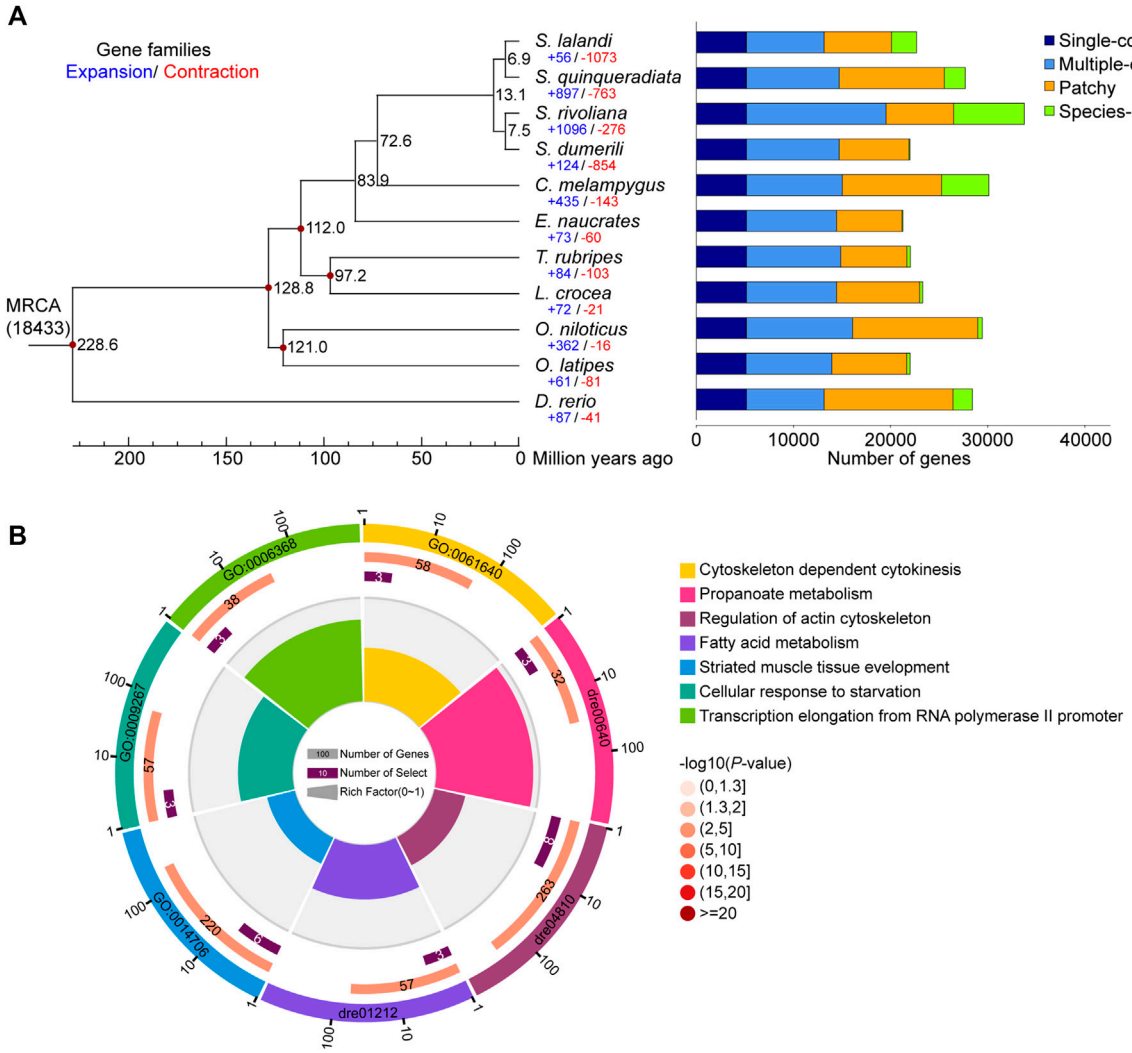

FIGURE 2 | Genome evolution analysis. **(A)** Phylogenetic tree of 11 teleost genomes, which was constructed using 5,067 single copy orthologous genes. The black numbers on the branches indicate the estimated diverge times in millions of years ago, and the blue and red numbers represent the expanded and contracted gene families. The different types of orthologous relationships are shown on the right. **(B)** The enrichment analysis of 148 positively selected genes detected in *S. lalandi* genome.

Moreover, some genes with positively selected sites were associated with growth (*ficn*, *tgfbr3*, *igf1ra*, *gpc1b*, *rnf11*, and *tgfbrap1*) (**Supplementary Table 10**).

We identified other fast-evolving genes, such as the pheromone response gene *ora2*, ear development gene *ddt*, and sensory perception gene *ppef1*, with potential roles in the perception of the external environment (**Supplementary Table 10**). Fast-evolving genes involved in nervous system development and the regulation of neurotransmitter secretion (*rab33a*, *fstl5*, and *cplx2*) provided a tissue basis for sensitive sensory systems (**Supplementary Table 10**). The fast-evolving genes *ldha* and *rnf152* may contribute to adaptation to adverse environmental conditions, such as hypoxia and starvation (**Supplementary Table 10**).

CONCLUSIONS

We sequenced and assembled the genome of *S. lalandi* using Illumina shotgun, PacBio SMRT, and Hi-C data, providing the first chromosome-level genome assembly for the genus *Seriola*. Basing on multiple annotation strategies, we obtained 22,674 protein-coding genes with minimal redundancy. Further genomic analysis revealed gene families associated with expansions of HSP70 and olfactory receptor gene families, and the rapid evolution of muscle and skeletal system development genes, providing insight into the genetic basis underlying the physiological characteristics of *S. lalandi* and its adaptability to the external environment. We believe these

new resources will promote genetic research and accelerate the genetic breeding process for *S. lalandi*.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: GenBank, JAIQDC010000001.1-JAIQDC010000031.1; NCBI BioProject, PRJNA754209.

ETHICS STATEMENT

The animal study was reviewed and approved by Experimental Animal Care, Ethics and Safety Inspection Form Yellow Sea Fisheries Research Institute, CAFS. Written informed consent was obtained from the owners for the participation of their animals in this study.

AUTHOR CONTRIBUTIONS

LX, CS, and XY conceived and designed the experiments. BW, AC, and YJ collected, identified, and photographed the specimens. SL, KL, and XH analyzed the genome and transcriptome data. WH, WQ, and BF performed gene analysis. SL drafted the manuscript. CS, YX, and ZL provided advice on manuscript writing. All authors reviewed the manuscript.

FUNDING

This work was supported the National Key R&D Program of China (2018YFD0900301, 2019YFD0900901, 2018YFD0901204); the Marine S&T Fund of Shandong Province for Pilot National Laboratory for Marine Science and Technology (Qingdao) (2018SDKJ0303-1); the Central Public-interest Scientific Institution Basal Research Fund, CAFS (No.2020TD19, 2020TD47, 2021GH05); the Taishan Scholar Project Fund of Shandong of China; the China Agriculture Research System of MOF and MARA (CARS-47).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.825742/full#supplementary-material

REFERENCES

Allen, J. E., Majoros, W. H., Pertea, M., and Salzberg, S. L. (2006). JIGSAW, GeneZilla, and GlimmerHMM: Puzzling Out the Features of Human Genes in the ENCODE Regions. *Genome Biol.* 7 Suppl 1, S9–S13. doi:10.1186/gb-2006-7-s1-s9

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase Update, a Database of Repetitive Elements in Eukaryotic Genomes. *Mob DNA* 6, 11–16. doi:10.1186/s13100-015-0041-9

Bett, N. N., and Hinch, S. G. (2016). Olfactory Navigation during Spawning Migrations: a Review and Introduction of the Hierarchical Navigation Hypothesis. *Biol. Rev.* 91, 728–759. doi:10.1111/brv.12191

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. *Genome Res.* 14, 988–995. doi:10.1101/gr.1865504

Burge, C. B., and Karlin, S. (1998). Finding the Genes in Genomic DNA. *Curr. Opin. Struct. Biol.* 8, 346–354. doi:10.1016/s0959-440x(98)80069-9

Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Mol. Biol. Evol.* 17, 540–552. doi:10.1093/oxfordjournals.molbev.a026334

Clark, M. S., and Peck, L. S. (2009). Triggers of the HSP70 Stress Response: Environmental Responses and Laboratory Manipulation in an Antarctic marine Invertebrate (*Nacella Concinna*). *Cell Stress and Chaperones* 14, 649–660. doi:10.1007/s12192-009-0117-x

Cong, X., Zheng, Q., Ren, W., Chéron, J.-B., Fiorucci, S., Wen, T., et al. (2019). Zebrafish Olfactory Receptors ORAs Differentially Detect Bile Acids and Bile Salts. *J. Biol. Chem.* 294, 6762–6771. doi:10.1074/jbc.ra118.006483

De Bie, T., Cristianini, N., Demuth, J. P., and Hahn, M. W. (2006). CAFE: a Computational Tool for the Study of Gene Family Evolution. *Bioinformatics* 22, 1269–1271. doi:10.1093/bioinformatics/btl097

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., et al. (2017). *De Novo* assembly of the *Aedes aegypti* Genome Using Hi-C Yields Chromosome-Length Scaffolds. *Science* 356, 92–95. doi:10.1126/science.aal3327

Dunham, R. A., Taylor, J. F., Rise, M. L., and Liu, Z. (2014). Development of Strategies for Integrated Breeding, Genetics and Applied Genomics for Genetic Improvement of Aquatic Organisms. *Aquaculture* 420-421, S121–S123. doi:10.1016/j.aquaculture.2013.10.020

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., et al. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cel Syst.* 3, 95–98. doi:10.1016/j.cels.2016.07.002

Huete-Pérez, J. A., and Quezada, F. (2013). Genomic Approaches in marine Biodiversity and Aquaculture. *Biol. Res.* 46, 353–361. doi:10.4067/S0716-97602013000400007

Jiang, D., Lin, L., and Chen, Y. (2001). Indoor wintering and growth of *Seriola aureorvttata* Temminck et Schegel. *J. Dalian Fish. Univ.* 3, 69–73. doi:10.3969/j.issn.1000-9957.2001.03.012

Kent, W. J. (2002). BLAT-the BLAST-like Alignment Tool. *Genome Res.* 12, 656–664. doi:10.1101/gr.229202

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a Fast Spliced Aligner with Low Memory Requirements. *Nat. Methods* 12, 357–360. doi:10.1038/nmeth.3317

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a Fast, Scalable and User-Friendly Tool for Maximum Likelihood Phylogenetic Inference. *Bioinformatics* 35, 4453–4455. doi:10.1093/bioinformatics/btz305

Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., et al. (2016). Third-generation Sequencing and the Future of Genomics. *BioRxiv*, 048603. doi:10.1101/048603

Li, H., Coghlan, A., Ruan, J., Coin, L. J., Hériché, J.-K., Osmotherly, L., et al. (2006). TreeFam: a Curated Database of Phylogenetic Trees of Animal Gene Families. *Nucleic Acids Res.* 34, D572–D580. doi:10.1093/nar/gkj118

Li, W., and Godzik, A. (2006). Cd-hit: a Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. *Bioinformatics* 22, 1658–1659. doi:10.1093/bioinformatics/btl158

Liu, A., Pirozzi, I., Codabaccus, B. M., Stephens, F., Francis, D. S., Sammut, J., et al. (2021). Effects of Dietary Choline on Liver Lipid Composition, Liver Histology and Plasma Biochemistry of Juvenile Yellowtail Kingfish (*Seriola lalandi*). *Br. J. Nutr.* 125, 1344–1358. doi:10.1017/S0007114520003669

Liu, H., Chen, C., Lv, M., Liu, N., Hu, Y., Zhang, H., et al. (2021). A Chromosome-Level Assembly of Blunt Snout Bream *Megalobrama amblycephala* Genome

Reveals an Expansion of Olfactory Receptor Genes in Freshwater Fish. *Mol. Biol. Evol.* 38, 4238–4251. doi:10.1093/molbev/msab152

Liu, K., Hao, X., Wang, Q., Hou, J., Lai, X., Dong, Z., et al. (2019). Genome-wide Identification and Characterization of Heat Shock Protein Family 70 Provides Insight into its Divergent Functions on Immune Response and Development ofParalichthys Olivaceus. *PeerJ* 7, e7781. doi:10.7717/peerj.7781

Löytynoja, A., and Goldman, N. (2010). webPRANK: a Phylogeny-Aware Multiple Sequence Aligner with Interactive Alignment Browser. *BMC Bioinform* 11, 1–7. doi:10.1186/1471-2105-11-579

Lv, M., Zhang, Y., Liu, K., Li, C., and Wang, J. (2020). A Chromosome-Level Genome Assembly of the Anglerfish *Lophius Litulon*. *Front. Genet.* 11. doi:10.3389/fgene.2020.581161

Martínez-Montaño, E., González-Álvarez, K., Lazo, J. P., Audelo-Naranjo, J. M., and Vélez-Medel, A. (2016). Morphological Development and Allometric Growth of Yellowtail Kingfish *Seriola lalandi* V. Larvae under Culture Conditions. *Aquac. Res.* 47, 1277–1287. doi:10.1111/are.12587

Marçais, G., and Kingsford, C. (2011). A Fast, Lock-free Approach for Efficient Parallel Counting of Occurrences of K-Mers. *Bioinformatics* 27, 764–770. doi:10.1093/bioinformatics/btr011

Orellana, J., Waller, U., and Wecker, B. (2014). Culture of Yellowtail Kingfish (*Seriola lalandi*) in a marine Recirculating Aquaculture System (RAS) with Artificial Seawater. *Aquacultural Eng.* 58, 20–28. doi:10.1016/j.aquaeng.2013.09.004

Ozaki, A., Yoshida, K., Fuji, K., Kubota, S., Kai, W., Aoki, J.-y., et al. (2013). Quantitative Trait Loci (QTL) Associated with Resistance to a Monogenean Parasite (*Benedenia Seriolae*) in Yellowtail (*Seriola quinqueradiata*) through Genome Wide Analysis. *PLOS ONE* 8, e64987. doi:10.1371/journal.pone.0064987

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads. *Nat. Biotechnol.* 33, 290–295. doi:10.1038/nbt.3122

Peterson, B. C., Burr, G. S., Pietrak, M. R., and Proestou, D. A. (2020). Genetic Improvement of North American Atlantic Salmon and the Eastern Oyster *Crassostrea virginica* at the U.S. Department of Agriculture-Agricultural Research Service National Cold Water Marine Aquaculture Center. *North. Am. J. Aquac.* 82, 321–330. doi:10.1002/naaq.10144

Purcell, C. M., Chabot, C. L., Craig, M. T., Martinez-Takeshita, N., Allen, L. G., and Hyde, J. R. (2015). Developing a Genetic Baseline for the Yellowtail Amberjack Species Complex, *Seriola lalandi* Sensu Lato, to Assess and Preserve Variation in Wild Populations of These Globally Important Aquaculture Species. *Conserv Genet.* 16, 1475–1488. doi:10.1007/s10592-015-0755-8

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., et al. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* 159, 1665–1680. doi:10.1016/j.cell.2014.11.021

Rondeau, E. B., Messmer, A. M., Sanderson, D. S., Jantzen, S. G., von Schalburg, K. R., Minkley, D. R., et al. (2013). Genomics of Sablefish (*Anoplopoma fimbria*): Expressed Genes, Mitochondrial Phylogeny, Linkage Map and Identification of a Putative Sex Gene. *BMC Genomics* 14, 1–9. doi:10.1186/1471-2164-14-452

Ruan, J., and Li, H. (2020). Fast and Accurate Long-Read Assembly With Wtdbg2.*Nat. Methods*17, 155–158. doi:10.1038/s41592-019-0669-3

Sanchís-Benlloch, P. J., Nocillado, J., Ladisa, C., Aizen, J., Miller, A., Shpilman, M., et al. (2017). *In-vitro* and *In-Vivo* Biological Activity of Recombinant Yellowtail Kingfish (*Seriola lalandi*) Follicle Stimulating Hormone. *Gen. Comp. Endocrinol.* 241, 41–49. doi:10.1016/j.ygcen.2016.03.001

Sano, T. (1998). Control of Fish Disease, and the Use of Drugs and Vaccines in Japan. *J. Appl. Ichthyol.* 14, 131–137. doi:10.1111/j.1439-0426.1998.tb00630.x

Shi, B., Liu, X., Liu, Y., Zhang, Y., Gao, Q., Xu, Y., et al. (2019). Effects of Gradual Salinity Change on Osmotic Regulation of Juvenile Yellowtail Kingfish (*Seriola Aureovittata*). *Coast Eng.* 38, 63–70.

Shi, B., Liu, Y., Liu, X., Xu, Y., Li, R., Song, X., et al. (2017). Study on the Karyotype of Yellowtail Kingfish (*Seriola Aureovittata*). *PROGRESS FISHERY SCIENCES* 38, 136–141. doi:10.11758/yykxjz.20160816004

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs. *Bioinformatics* 31, 3210–3212. doi:10.1093/bioinformatics/btv351

Sodeland, M., Gaarder, M., Moen, T., Thomassen, M., Kjøglum, S., Kent, M., et al. (2013). Genome-wide Association Testing Reveals Quantitative Trait Loci for Fillet Texture and Fat Content in Atlantic salmon. *Aquaculture* 408-409, 169–174. doi:10.1016/j.aquaculture.2013.05.029

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: Ab Initio Prediction of Alternative Transcripts. *Nucleic Acids Res.* 34, W435–W439. doi:10.1093/nar/gkl200

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE* 9, e112963. doi:10.1371/journal.pone.0112963

Xu, Y., Zhang, Z., Liu, X., Wang, B., Shi, B., Liu, Y., et al. (2019). Morphometric Characteristics of the Embryonic and Postembryonic Development of Yellowtail Kingfish, *Serida Aureovittata*. *J. Fish. Sci. China* 26, 172. doi:10.3724/sp.j.1118.2019.18094

Yang, Z. (1997). PAML: a Program Package for Phylogenetic Analysis by Maximum Likelihood. *Bioinformatics* 13, 555–556. doi:10.1093/bioinformatics/13.5.555

Zhang, D.-C., Guo, L., Guo, H.-Y., Zhu, K.-C., Li, S.-Q., Zhang, Y., et al. (2019). Chromosome-level Genome Assembly of golden Pompano (*Trachinotus Ovatus*) in the Family Carangidae. *Sci. Data* 6, 216. doi:10.1038/s41597-019-0238-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Veterinary Science

ORIGINAL RESEARCH
published: 16 February 2022
doi: 10.3389/fvets.2022.752521

Screening and Validation of p38 MAPK Involved in Ovarian Development of *Brachymystax lenok*

Tianqing Huang[1], Wei Gu[1], Enhui Liu[1], Lanlan Zhang[2], Fulin Dong[1], Xianchen He[3], Wenlong Jiao[4], Chunyu Li[5], Bingqian Wang[1] and Gefeng Xu[1]**

[1] Key Laboratory of Freshwater Aquatic Biotechnology and Breeding, Ministry of Agriculture and Rural Affairs, Heilongjiang River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Harbin, China, [2] Heilongjiang Province General Station of Aquatic Technology Promotion, Harbin, China, [3] Heilongjiang Aquatic Animal Resource Conservation Center, Harbin, China, [4] Gansu Fisheries Research Institute, Lanzhou, China, [5] Xinjiang Tianyun Organic Agriculture Co., Yili Group, Hohhot, China

Brachymystax lenok (lenok) is a rare cold-water fish native to China that is of high meat quality. Its wild population has declined sharply in recent years, and therefore, exploring the molecular mechanisms underlying the development and reproduction of lenoks for the purposes of artificial breeding and genetic improvement is necessary. The lenok comparative transcriptome was analyzed by combining single molecule, real-time, and next generation sequencing (NGS) technology. Differentially expressed genes (DEGs) were identified in five tissues (head kidney, spleen, liver, muscle, and gonad) between immature [300 days post-hatching (dph)] and mature [three years post-hatching (ph)] lenoks. In total, 234,124 and 229,008 full-length non-chimeric reads were obtained from the immature and mature sequencing data, respectively. After NGS correction, 61,405 and 59,372 non-redundant transcripts were obtained for the expression level and pathway enrichment analyses, respectively. Compared with the mature group, 719 genes with significantly increased expression and 1,727 genes with significantly decreased expression in all five tissues were found in the immature group. Furthermore, DEGs and pathways involved in the endocrine system and gonadal development were identified, and p38 mitogen-activated protein kinases (MAPKs) were identified as potentially regulating gonadal development in lenok. Inhibiting the activity of p38 MAPKs resulted in abnormal levels of gonadotropin-releasing hormone, follicle -stimulating hormone, and estradiol, and affected follicular development. The full-length transcriptome data obtained in this study may provide a valuable reference for the study of gene function, gene expression, and evolutionary relationships in *B. lenok* and may illustrate the basic regulatory mechanism of ovarian development in teleosts.

Keywords: full-length transcriptome, p38 MAPK, ovarian development, *Brachymystax lenok*, SMRT and NGS

OPEN ACCESS

Edited by:
Roger Huerlimann, Okinawa Institute of Science and Technology Graduate University, Japan

Reviewed by:
Xiangshan Ji, Shandong Agricultural University, China
Jake Goodall, University of Iceland, Iceland

***Correspondence:**
Bingqian Wang
wangbingqian@hrfri.ac.cn
Gefeng Xu
xugefeng@hrfri.ac.cn

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Veterinary Science

Received: *27 August 2021*
Accepted: *13 January 2022*
Published: *16 February 2022*

Citation:
Huang T, Gu W, Liu E, Zhang L, Dong F, He X, Jiao W, Li C, Wang B and Xu G (2022) Screening and Validation of p38 MAPK Involved in Ovarian Development of Brachymystax lenok. Front. Vet. Sci. 9:752521. doi: 10.3389/fvets.2022.752521

INTRODUCTION

Brachymystax lenok (lenok) is an economically important fish in the Amur Basin with high-quality nutrient-rich meat. As a rare cold water fish native to China, lenok has been listed as a vulnerable species in the Red Book of Endangered Animals in China because of its sharp decline in the wild, and is classified as aquatic wildlife under second-class protection (1). Recent studies

on lenok have mainly focused on resource investigation, nutritional physiology, and artificial reproduction (2–4), and studies on the genetic analysis and reproductive regulation mechanisms of lenok are limited because of sparse genetic information. It is important to analyse the molecular mechanisms that regulate lenok reproduction and development to facilitate better breeding and genetic improvement.

The analysis of full-length transcriptome can demonstrate the type and number of genes at the molecular level and is an effective method for revealing the regulatory mechanisms of different physiological and biochemical processes. Many non-model organisms lack reference genome sequence information, and full-length transcriptome sequencing is a rapid and effective method for investigating gene expression, gene function, and evolutionary analysis in this species (5–7). Single molecule real-time sequencing technology (SMRT) based on the Pacific Biosciences (PacBio) platform has the benefits of long reading fragments and high accuracy. Full-length mRNA can be generated directly without assembly, and this technique has been successfully used for the full-length transcriptome analysis of multiple species, such as cattle (8), rabbits (9), mice (10), and shrimp (11). It has also been widely used in teleosts (12–14).

Next generation sequencing based on the Illumina platform is helpful for characterizing various biological processes and exploiting underlying gene activity. Over the past decade, numerous studies have been conducted on genetic and developmental transcriptomics in fish, including microarray and the serial analysis of gene expression. In zebrafish, transcriptome data at nine different stages of embryonic development were comprehensively analyzed to identify the key roles of pathways and functional genes involved in development (15). In haddock (*Atlantic Haddock*), a genetic network for development has been established by analyzing transcriptome data from the embryo to the early developmental stages of juveniles (16). Numerous studies on gonadal development have been conducted using transcriptome sequencing. For example, genes that are differentially expressed during testis development have been characterized in the channel catfish (*Ictalurus punctatus*) (17). In the spotted knifejaw (*Oplegnathus punctatus*) and fugu (*Takifugu rubripes*), a large number of differential genes were identified in the testes of adult fish compared to the ovaries (18, 19). In gonads from 3–24 weeks after the fertilization (immature to mature) of zebrafish, the dynamic trend of miRNA abundance was characterized by miRNA sequencing (20). However, the number of studies on gonadal development using full length transcriptome methods is limited.

The collection of gonadal tissue samples from lenok could only commence at 300 days post-hatching (dph), and lenok at three years post-hatching (ph) were on the verge of ovulation. These two periods are representative of immature and mature groups for transcriptome sequencing. PacBio and Illumina sequencing were combined to generate two complete full length transcriptomes of immature and mature lenok by analyzing gene expression in five different tissues (liver, muscle, spleen, head kidney, and gonad) and screening for DEGs related to lenok ovarian development. p38 mitogen-activated protein kinases (MAPKs) were defined to express the most significant difference between the immature and mature gonads. MAPKs are a form of conserved serine/threonine protein kinases that participate in the regulation of multiple physiological functions in the form of intracellular signal transmission, and are an important signal transduction system in cells (21). Furthermore, p38 MAPK has been suggested as playing an important role in lenok development and reproduction. As one of the important members of the MAPK family, it is widely expressed in the thyroid, testis, ovary, and pituitary tissues of mammals (22), and plays an important role in the reproductive process (23). However, the role of p38 MAPK in lenok is currently unknown. Therefore, p38 MAPK was selected to demonstrate its role in the balance of reproductive endocrine hormones and the follicular development of lenok. The full-length transcriptome data obtained in this study can provide a valuable reference for further studies on the mechanism of gonadal development and maturation, and make an important contribution to researching the genetic improvement of lenok.

MATERIALS AND METHODS

Ethics Statement

All experiments were performed in accordance with the European Communities Council Directive (86/609/EEC). The experiments were approved by the Animal Husbandry Department of the Heilongjiang Animal Care and Use Committee (202110384464). All fish involved in this research were bred following the guidelines of the Animal Husbandry Department of Heilongjiang, China.

Fish Sampling and RNA Purification

The gonad (G), head kidney (K), liver (L), muscle (M), and spleen (S) were collected from three mature and 12 immature samples of *B. lenok*, which were bred at the Bohai experimental station of the Heilongjiang River Fisheries Research Institute (129° 04′ 64.7753″ E; 44° 14′ 5.983″ N). All samples used in this experiment were obtained from female lenok. The immature group was 300 dph and the mature group was three years ph. Four immature tissue samples were mixed into one RNA sample, which required three repeat RNA samples for sequencing. Therefore, the immature sample size was 12. However, one mature tissue sample can constitute one RNA sample and three duplicate samples require a mature sample size of $N = 3$. Before tissue collection, the fish were euthanized with an overdose of anesthesia in MS-222, as reported previously (2). Each tissue sample was immediately placed into 2 mL sterile tubes and placed in liquid nitrogen. After storing for 1 h, all samples were transferred to a −80 °C refrigerator for further analysis. In addition, gonad samples at 300 dph, 750 dph, and three years ph were used for western blot analysis and were kept at −20 °C. TRIzol reagent (Invitrogen, CA, USA) was used to extract the total RNA, and only RNA samples with a RIN number greater than 7.0 were kept for subsequent experiments. One microgram of each RNA sample was pooled and sequenced using PacBio single-molecule, long-read sequencing (PacBio Sequel, Menlo Park, USA), and Illumina sequencing (Illumina NovaSeq 6000,

California, USA) in parallel. The correlation of each sample was R2 > 0.8 (**Supplementary Figure S1**).

Complementary DNA (CDNA) Library Construction and PacBio Sequencing

The SMARTer PCR cDNA Synthesis Kit (Takara Bio USA, Inc.) was used to prepare the PacBio cDNA library using the following steps: Mix tube 1 (2 μL total RNA, 1 μL 3'SMART CDS Primer II A, and 1.5 μL deionized H_2O) at 72 °C for 3 min, and 42 °C for 2 min. Then, mix tube 2 (2 μl 5xFirst-Strand Buffer, 0.25 μl DTT, 1 μl dNTP Mix, 1 μl SMARTer II A Oligonucleotide, 0.25 μl RNase Inhibitor, and 1 μl SMART Scribe Reverse Transcriptase), and then add into tube 1 at 42°C for 1 h, and 72°C for 10 min. The full-length cDNA was subjected to PCR amplification. The quality and concentration of the cDNA library were determined using a Qubit 2.0 Fluorometer and an Agilent 2100 bioanalyzer (24). The 1–6 KB library was sequenced using PacBio Sequel.

CDNA Library Preparation for Illumina Sequencing

Three replicates for each of the five tissues (gonad, head kidney, liver, muscle and spleen), making a total of 15 RNA samples, were fragmented into small pieces at high temperatures. The mRNA-Seq sample preparation kit (Illumina, San Diego, CA, USA) was used for the reverse transcription of RNA fragments to construct the final cDNA library, and fragments of 250–300 bp were selected using AMPure XP beads (25). The final cDNA library was assessed by PCR and the quality of the cDNA library was determined using an Agilent 2100 Bioanalyzer. Paired-end sequencing was performed on an Illumina NovaSeq 6000, following the recommended protocol (26).

Data Analysis of PacBio Sequencing

Raw reads were processed into error-corrected reads of inserts (ROIs) using the Iso-seq pipeline (27) with min full pass = 3 and min predicted accuracy = 0.9. Full-length, non-chimeric (FLNC) transcripts were determined by searching for the poly-A tail signal and the 5′ and 3′ cDNA primers in the ROIs (28). Iterative clustering for error correction (ICE) was used to obtain consensus isoforms, and FL consensus sequences from the ICE were polished using Quiver (29). High-quality FL transcripts were classified with a post-correction accuracy above 99%. The cluster database with high identity and tolerance (CD-HIT, http://weizhongli-lab.org/cd-hit/) was used for Iso-Seq high-quality FL transcripts to remove redundancy (identity > 0.99).

Data Analysis of Illumina Sequencing

Raw reads in Fastq format were first processed using Perl scripts from our laboratory. In this step, clean reads were obtained by removing reads containing adapters, reads containing poly-N, and low-quality reads from the raw data. The Q20, Q30, GC content, and sequence duplication levels of the clean data were concurrently calculated. All downstream analyses were based on high quality clean data. These clean reads were then mapped to the PacBio reference genome sequence. Only reads with a perfect match or one mismatch were further analyzed and annotated based on the reference genome. Hisat2 (30) tools were used to map the reference genome. Gene expression levels were estimated by fragments per kilobase of transcript per million fragments mapped. Prior to differential gene expression analysis, for each sequenced library, the read counts were adjusted using the edgeR (31) program package through one scaling normalized factor. A differential expression analysis of the two samples was performed using the EBSeq R (32) package. The resulting false discovery rate was adjusted using the posterior probability of DE (PPDE). A false discovery rate < 0.01, and fold change ≥ 2 were set as thresholds for significantly differential expression.

Structure Analysis

Simple sequence repeats (SSRs) in the transcriptome were identified using MISA (http://pgrc.ipk-gatersleben.de/misa/). Candidate coding regions within the transcript sequences were identified using TransDecoder (https://github.com/TransDecoder/TransDecoder/releases). Iso-Seq™ data were directly used to run all-vs-all BLAST (33) with high identity settings. BLAST alignments that met all criteria were considered products of candidate AS events. The coding potential calculator (CPC, http://cpc2.cbi.pku.edu.cn), coding non-coding index (CNCI, https://github.com/www-bioinfo-org/CNCI), coding potential assessment tool (CPAT, http://rna-cpat.sourceforge.net/), and protein family database (Pfam, http://pfam.xfam.org/) were combined to sort non-protein-coding RNA candidates from putative protein-coding RNAs in the transcripts. Transcripts with lengths greater than 200 nt and having more than two exons were selected as long non-coding RNA (lncRNA) candidates.

Gene Functional Annotation and Enrichment Analysis

Gene function was annotated based on the following databases: NCBI non-redundant protein sequences (NR, http://www.ncbi.nlm.nih.gov/), protein family (Pfam, http://pfam.xfam.org/), Clusters of Orthologous Groups of proteins (KOG/COG/eggNOG, http://www.ncbi.nlm.nih.gov/COG/), Swiss-Prot (a manually annotated and reviewed protein sequence database, http://www.uniprot.org/), Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/), and Gene Ontology (GO). The GO enrichment analysis of differentially expressed genes (DEGs) was implemented using the GOseq R (34) packages based on Wallenius non-central hyper-geometric distribution (34), which can adjust for gene length bias in DEGs. KOBAS (35) software was used to test the statistical enrichment of DEGs in KEGG pathways.

Validation of Expressions of DEGs by Real-Time PCR

Real-time PCR was performed to validate the RNAs involved in RNA-Seq. The same RNA samples were used for the deep sequencing. Twelve mRNAs were detected using real-time PCR. A PrimeScript RT Reagent Kit (TaKaRa, Shiga, Japan) was used for reverse transcription. Real-time PCR was performed using FastStart Universal SYBR® Green Master Mix (Roche, Switzerland) and a CFX96 C1000 touch thermal cycler (Bio-Rad, USA). Beta-actin was used as the reference gene. The Ct

values were measured, and the value of the target sequence normalized to the reference sequence was calculated as $2^{-\Delta\Delta}$Ct. The statistical analysis was performed using SPSS (36) version 13.0 for Microsoft Windows.

Western Blot Analysis

Proteins were extracted from gonad samples at 300 dph (immature, $n = 3$) and three years post-hatching (mature, $n = 3$) of female lenoks that were reared under identical conditions to individuals used in transcriptome analyses. To explore the expression profiles of p38 MAPK protein in different tissues, the proteins of the gill, heart, liver, spleen, intestine, skin, and ovary tissues were extracted from female lenok at 750 dph ($n = 3$), which was a time of rapid follicular development. The proteins were boiled for 20 min in SDS-PAGE loading buffer and separated using 12% SDS-PAGE gels. The proteins were transferred to polyvinylidene fluoride membranes (Bio-Rad, USA) and analyzed by western blotting. Anti-p38 MAPK (bs-0637R, Bioss, 1:10000), anti-phospho-p38 MAPK (Thr180) (bs-5476R, Bioss, 1:5000), and beita-actin (ym3121, Immunoway, 1:5000) were used as primary antibodies, and peroxidase-conjugated AffiniPure goat anti-rabbit IgG (1:2000) was used as the secondary antibody (Cell Signaling Technology, USA). After washing with PBST, the protein bands were visualized by infrared fluorescence using the Odyssey Imaging System (LI-COR InC).

SB203580 Inhibitor Injection

SB203580 (Biorbyt) was used to inhibit the phosphorylation of p38 MAPK. Lenoks of approximately 750 dph were randomly divided into three groups ($n = 3$). CK was the no-treatment group, the negative control (DMSO) was injected intraperitoneally with 1 mL DMSO, and the experimental group (DMSO + I) was injected intraperitoneally with 5 mg/kg SB203580 dissolved in 1 ml DMSO. Seven days post-injection, plasma and gonad samples of the three groups were collected for western blot analysis, hormone level determination, and immunohistochemical assays.

Reproductive Hormone Level Assay

Polyclonal antibodies were customized based on partial amino acid sequences of follicle-stimulating hormone (FSH), luteinizing hormone (LH), and gonadotropin-releasing hormone 3 (GnRH3) in the lenok. Enzyme-linked immunosorbent assay (ELISA) kits (MLBIO, China) were coated with specific antibodies and developed to detect FSH, LH, and GnRH3 in the lenok. Referring to D'Cotta's method (37), an ELISA was used to verify the specificity and stability of each kit. Ten positive serum samples of each coated antibody were tested, all of which were positive, and 10 blank controls were tested, all of which were negative. The coefficient of variation between and within batches was less than 15%. Standard curves for each ELISA kit are shown in **Supplementary Figure S2**. The ELISA experiments were performed using a microplate reader at a wavelength of 450 nm. The levels of GnRH3, FSH, LH, and estrogen (E2) were calculated using standard curves. The plasma of three lenoks in CK, DMSO, and DMSO + I groups ($n = 3$) were detected by ELISA, and each sample was assayed three times.

Immunohistochemistry

Gonad samples from the three groups were fixed in 4% paraformaldehyde for 24 h at 4°C, embedded in paraplast, and sectioned at 5 μm thickness. Paraffin sections were incubated with 3% hydrogen peroxide in phosphate-buffered saline for 10 min, and then blocked with 2% bovine serum albumin and 2% goat serum for 1 h. Sections were boiled with 0.01 M citric acid and EDTA solution for 30 min for antigen recovery. The p-p38 antibody (1:300) was applied to the sections as the primary antibody at 4°C overnight. After washing with phosphate-buffered saline, the sections were incubated with biotinylated goat anti-rabbit secondary antibody (1:1000) for 20 min at 25°C. The sections were then treated with peroxidase-conjugated streptavidin, developed with DAB, and counterstained with haematoxylin. For the negative control, sections were not incubated with the primary antibody. For this part, the gonads of three lenoks from CK, DMSO, and DMSO+I groups ($n = 3$) were analyzed.

Statistical Analysis

A one-way analysis of variance was used to assess the differences in p38 protein levels in the gills, heart, liver, spleen, intestine, head, skin and gonad, and in gonad at 300 dph, 750 dph, and three years ph. It was also used to assess the differences in reproductive hormone levels in the CK, DMSO, and DMSO+I groups. Data were shown as the mean ± SEM. The statistical analysis was performed using SPSS version 13.0 for Microsoft Windows, and the statistical significance was set at $p < 0.05$.

RESULTS

Transcriptome Analysis

PacBio Iso-Seq and Bioinformatic Analysis

The PacBio iso-sequencing of lenok was completed, the clean data of mature and immature were obtained by SMRT

TABLE 1 | Statistics of the PacBio Iso-sequencing data.

cDNA library	Mature	Immature
cDNA size	1–6K	1–6K
Date size (GB)	23.21	22.67
CCS number	272,337	271,088
Read bases of CCS	810,860,211	756,750,619
Mean read length of CCS	2,977	2,791
Mean number of passes	37	39
Number of FLNC reads	229,008	234,124
FLNC%	84.09%	86.36%
Number of polished high-quality isoforms	85,270	89,226
Number of polished low-quality isoforms	2,020	2,196
Percent of polished high-quality isoforms	97.25%	97.17%
CD-HIT	61,405	59,372

GB: A unit used to measure the amount of data. KB: One thousand base pairs.

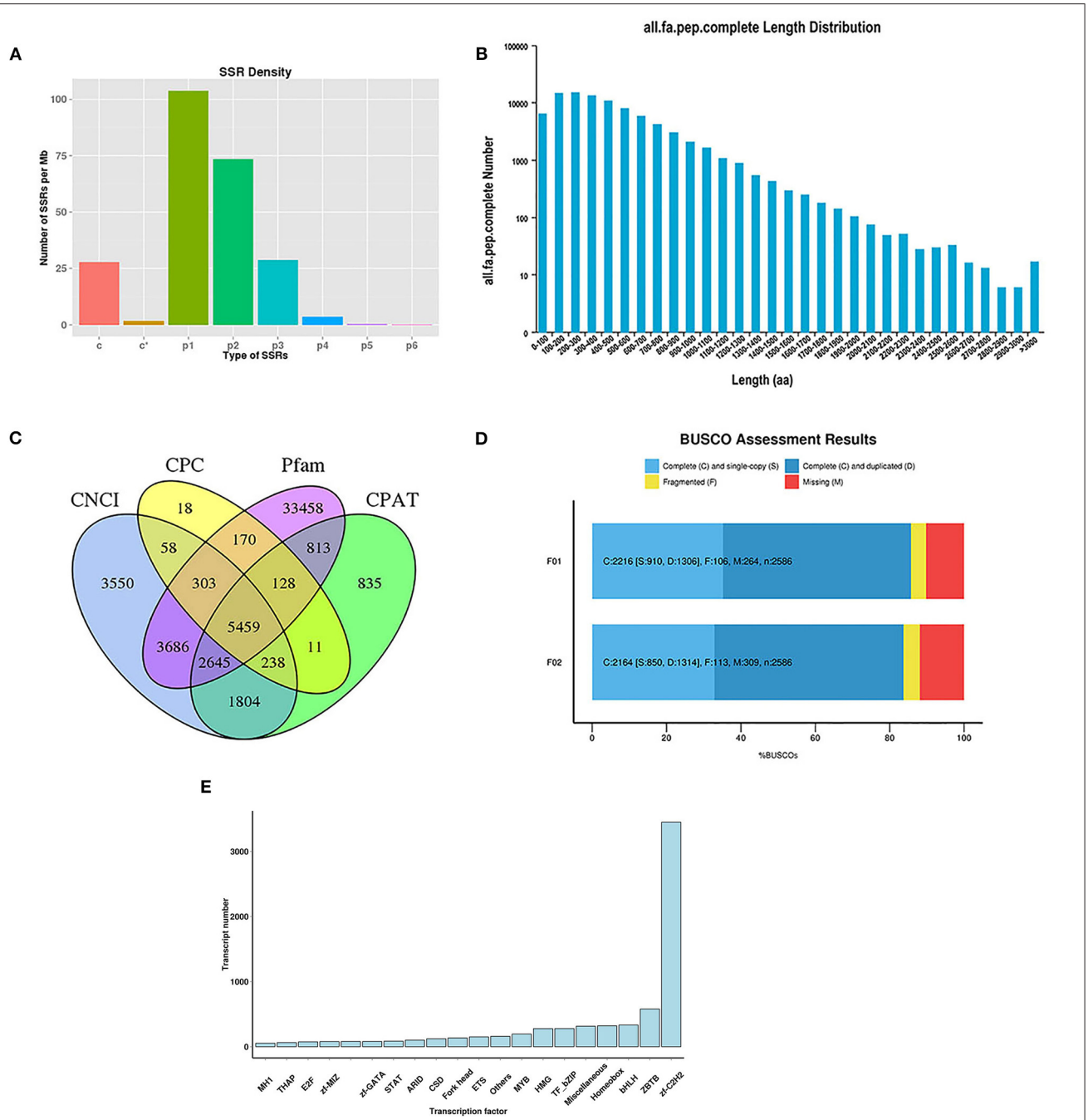

FIGURE 1 | Analysis of SSRs and identification of CDS, TFs, and lncRNAs. **(A)** The density of different type of SSRs; **(B)** The length distribution of predicted CDS; **(C)** Venn diagram for lncRNAs identified by four analysis methods; **(D)** BUSCO assessment results of the completeness of transcripts generated by CD-HIT; **(E)** Distribution of TF families.

sequencing technology, and the data sizes were 23.21 GB and 22.67 GB, respectively. Among these, 272,337 and 271,088 circular consensus sequence (CCS) reads were obtained from mature and immature samples, respectively (**Table 1**), with a mean length of 2,977 and 2,791 bp, respectively, using 37 and 39 passes, respectively, with full passes $\geq$ 3, and a quality consensus accuracy $>$ 0.9. Subsequently, the CCS reads were classified as full-length non-chimeric (FLNC) with 5′ primer, 3′ primer, and poly-A and non-full length reads, with proportions of 84.09 and 86.36%, respectively, for mature and immature samples. As a result, 229,008 and 234,124 high-quality FLNC reads for mature and immature samples, respectively, were obtained through the cluster of FLNC and correction.

We obtained 85,270 and 89,226 high-quality polished consensus sequences (**Table 1**) for the mature and immature samples, respectively. Finally, 61,405 and 59,372 non-redundant transcripts for mature and immature samples, respectively, were obtained by removing redundant transcripts using CD-HIT. By merging the mature and immature data, we obtained 106,647 non-redundant transcripts for *B. lenok* development analysis. The PacBio Iso-Seq raw data were deposited into the SRA-NCBI repository. The BioProject number was PRJNA669274 and the BioProject number of next-generation sequencing was PRJNA669219.

Analysis of Alternative Splicing Events, SSRs and Predictions of Coding Sequences, Transcription Factors, and LncRNAs

In this study, 3,158 and 3,332 alternative splicing (AS) events were analyzed. The absence of a reference genome limited the identification of the AS types. For the analysis of SSRs, 104,373 sequences (344,196,344 bp) were examined, including 103,688 SSRs and 49,478 SSR-containing sequences (**Supplementary Table S1**). There were 23,300 sequences containing more than one SSR, and the number of SSRs present in the compound form was 21,101. Specifically, most sequences were mononucleotides (42,411), dinucleotides (46,227), and trinucleotides (12,599). The number of different types of SSRs is shown in **Figure 1A**.

To identify putative protein-coding sequences, we predicted 100,580 open reading frames (ORFs) using the Trans-Decoder. In total, 88,856 coding sequences (CDSs) were identified with start and stop codons. The distributions of the numbers and lengths of complete CDSs are shown in **Figure 1B**. Among these, 14,579 transcripts (16.41%) were distributed in the 100–200 bp range, 14,893 transcripts (16.76%) in the 200–300 bp range, 13,365 transcripts (15.04%) in the 300–400 bp range, and 10,810 transcripts (12.17%) in the 400–500 bp range.

Here, 5,459 lncRNAs were predicted using a coding potential calculator (CPC), coding-non-coding index (CNCI), Pfam protein structure domain analysis, and coding potential assessment tool (CPAT) (**Figure 1C**) and candidate lncRNAs for future developmental research on lenok were revealed.

The completeness of the transcripts generated by CD-HIT was assessed using benchmarking universal single-copy orthologs (BUSCO, v.2.3). The results showed that 85.69 and 83.68% of the transcripts of mature and immature lenok samples, respectively, were complete (**Figure 1D**). Among the mature group, single-copy and duplicated BUSCOs accounted for 41.06 and 58.94%, respectively. In the immature group, the percentages of complete single-copy and duplicated BUSCOs were 39.28 and 60.72%, respectively. Only 106 and 113 fragmented BUSCOs and 264 and 309 missing BUSCOs were found in our two databases (**Figure 1D**). These results all show that our database is complete and available for subsequent research.

In total, 7,628 putative transcription factors (TFs) were examined by sequencing, and the top 20 families with the highest number of TFs are shown in **Figure 1E**. Most TFs belonged to the zf-C2H2 (3,450), ZBTB (579), bHLH (333), homeobox (361), miscellaneous (315), TF-bZIP (279), and HMG (278) families.

TABLE 2 | Number of annotated transcripts of each database.

Annotated databases	Number of transcripts
COG	33,077
GO	83,729
KEGG	67,965
KOG	74,798
Pfam	88,168
Swiss-Prot	69,821
eggNOG	99,079
NR	102,107
All	102,295

Function Annotation of Transcripts

The annotation information of 102,295 (95.92%) non-redundant transcripts was obtained by blasting eight databases, namely, NR, Swiss-Prot, COG, KOG, Pfam, eggNOG, GO, and KEGG. The number of annotated transcripts in each database is listed in **Table 2**.

By blasting the sequences of homologous species in the NR database, 50.866 (49.83%) annotated transcripts were aligned with *Oncorhynchus mykiss*, followed by *Esox luclus* (26.95%) and *Salmo salar* (13.28%) (**Figure 2A**).

In total, 83,729 transcripts were enriched in the three ontologies based on the GO analysis (**Figure 2B**). In the cellular component (CC), transcripts were mainly enriched in cells (44,531), cell parts (44,364), and organelles (31,422). Most transcripts enriched in molecular processes (MF) were binding (49,208), catalytic activity (29,970), and transporter activity (4,766). For biological processes (BP), transcripts were mainly enriched in cellular processes (45,934), single-organism processes (37,336), and metabolic processes (33,199).

In this study, 99,079 transcripts were annotated by blasting with linear homologous groups in the eggNOG database for function description and classification. The annotation results were classified into 25 categories (**Figure 2C**). The three largest groups were post-translational modification, protein turnover, chaperones (10,389), intracellular trafficking, secretion, vesicular transport (9,355), and signal transduction mechanisms (6,657) followed by transcription (6,166), cytoskeleton (3,078), and translation, ribosomal structure, and biogenesis (2,545).

Based on the KEGG analysis, 67,965 transcripts were enriched in 297 pathways. The five pathways enriched with the most genes were *endocytosis* (2,214), *protein processing in endoplasmic reticulum* (1,658), *regulation of actin cytoskeleton* (1,641), *herpes simplex infection* (1,533), and *MAPK signaling pathway* (1,518) (**Figure 2D**).

Differentially Expressed Genes (DEGs) Analysis

The expression levels of genes in five tissues of immature and mature lenok were investigated, including the gonad (G), head kidney (K), liver (L), muscle (M), and spleen (S). The numbers of upregulated and downregulated DEGs between the immature and mature groups are shown in **Table 3**.

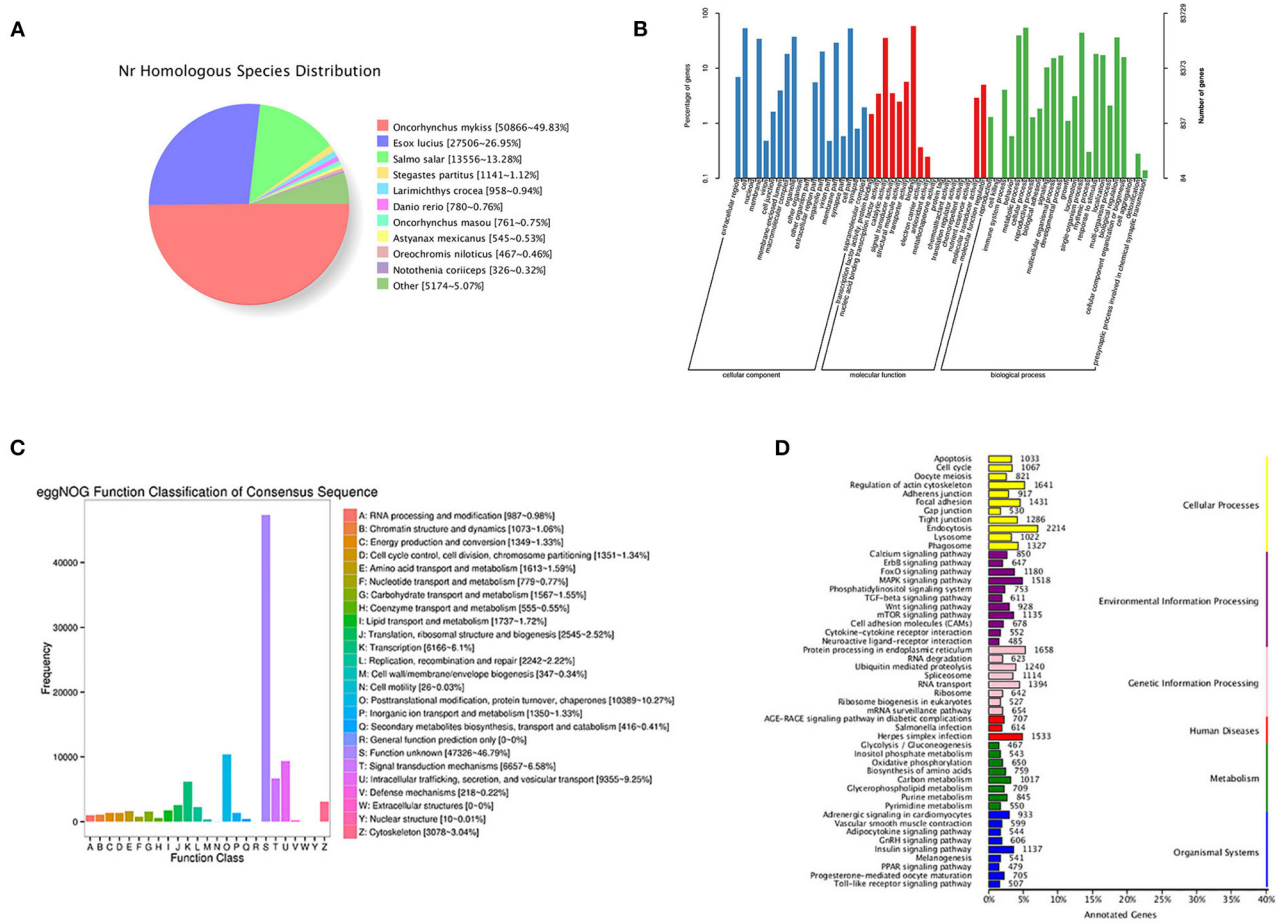

FIGURE 2 | Functional annotation information of non-redundant transcripts. **(A)** Homologous species distribution of transcripts in NR database; **(B)** GO analysis of annotated transcripts; **(C)** eggNOG function classification of transcripts; **(D)** KEGG pathway classified analysis of transcripts.

TABLE 3 | Statistics for DEGs of five tissues between of immature and mature *Brachymystax lenok*.

DEG Set (immature vs. mature)	All DEGs	Up-regulated	Down-regulated
Gonad vs. M gonad	7,728	3,306	4,422
Head kidney vs. M head kidney	10,352	3,840	6,512
Liver vs. M liver	6,690	2,254	4,436
Muscle vs. M muscle	7,968	3,268	4,700
Spleen vs. M spleen	10,837	4,089	6,748

We compared the DEGs between the immature and mature groups in the five tissues. Compared with the mature group, 719 genes with significantly increased expression (**Figure 3A**) and 1,727 genes with significantly decreased expression (**Figure 3B**) in all five tissues were found in the immature group. To investigate the function of DEGs, a KEGG analysis was performed on the overlapping genes. The results showed that 291 of 719 upregulated overlapping DEGs were enriched in the KEGG pathway, among which the significantly enriched pathways were carbon metabolism (23), pyruvate metabolism (11), and protein processing in the endoplasmic reticulum (30) (**Figure 3C**). Among the 1,727 downregulated overlapping DEGs, 693 genes were enriched in the KEGG pathway, among which the enriched pathways included spliceosome (53) and protein processing in the endoplasmic reticulum (65) (**Figure 3D**).

Pathways Related to Endocrine System and Development

The endocrine system and development-related pathways play important roles in the process of gonadal maturation (38). According to the classification information of the KEGG database (https://www.kegg.jp/kegg/pathway.html), the second classifications of endocrine system and development were under the primary classification of organismal systems. We analyzed the enriched DEGs in the pathways related to the endocrine system and development and their expression patterns in different tissues of immature and mature groups. A total of 77 DEGs were screened for the enrichment of these two classifications, and they enriched the *adipocytokine signaling*

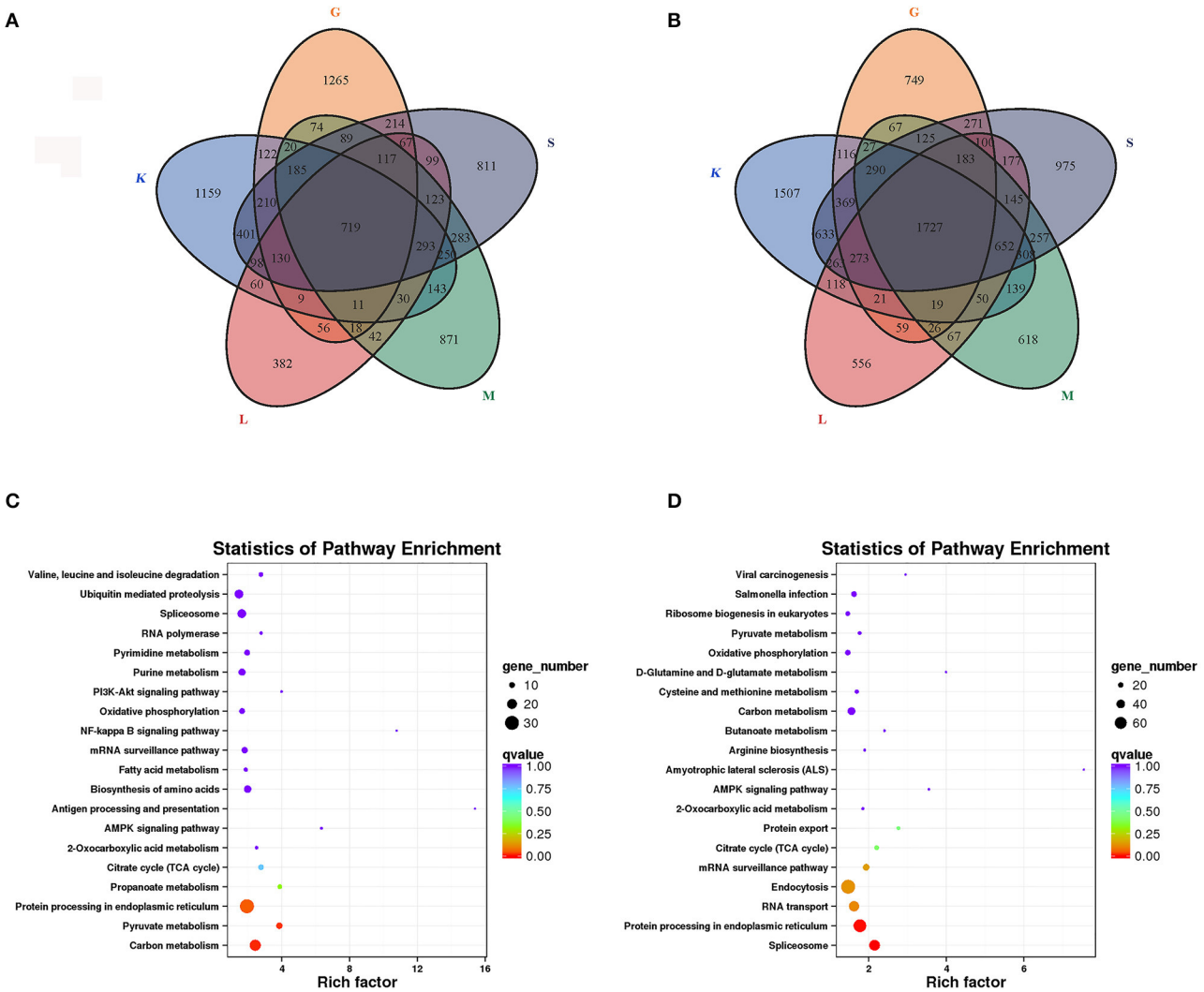

FIGURE 3 | Expression analysis of DEGs in five tissues of lenok. **(A)** Upregulated DEGs; **(B)** Downregulated DEGs; K, head kidney; L, liver; M, muscle; S, spleen; G, gonad; **(C)** KEGG analysis of upregulated DEGs; **(D)** KEGG analysis of downregulated DEGs.

pathway (ko04920), *GnRH signaling pathway* (ko04912), *insulin signaling pathway* (ko04910), *PPAR signaling pathway* (ko03320), *progesterone-mediated oocyte maturation* (ko04914), and *renin secretion* (ko04924). Information regarding the DEGs enriched in these pathways is presented in **Supplementary Table S2**, and their expression profiles in different tissues are shown in **Figure 4**. There were 31 DEGs enriched in the *insulin signaling pathway*, followed by the *adipocytokine signaling pathway* [16], *PPAR signaling pathway* [12], *GnRH signaling pathway* [10], *progesterone-mediated oocyte maturation* [9], *renin secretion* [2], and only one DEG was enriched in the remaining pathways.

Among the 77 DEGs enriched in the above pathways, the expression of 69 DEGs increased in mature lenok, indicating that the expression of genes involved in development and the endocrine system increased with the growth and development of lenok. Among the DEGs enriched in the above pathways, the five genes that expressed the most significant differences in the gonads were F01_transcript_1427 (unknown function), F02_transcript_5540 (p38 mitogen-activated protein kinase), F01_transcript_61649 (unknown function), F01_transcript_83061 (guanine nucleotide-binding protein subunit alpha-11), and F01_transcript_10752 (5-AMP-activated protein kinase) (**Table 4**).

Verification of RNA-seq by qRT-PCR

To verify the accuracy of RNA-seq, 12 DEGs were randomly selected for qRT-PCR verification, and their primer information is presented in **Supplementary Table S3**. The results of RNA-seq and qRT-PCR of these 12 DEGs in the five tissues are shown in **Figure 5**, and the correlation was expressed by Pearson's coefficient ($r^2 = 0.9521$). The results showed consistency and correlation between the results of RNA-seq and qRT-PCR, which proved the effectiveness of RNA-seq.

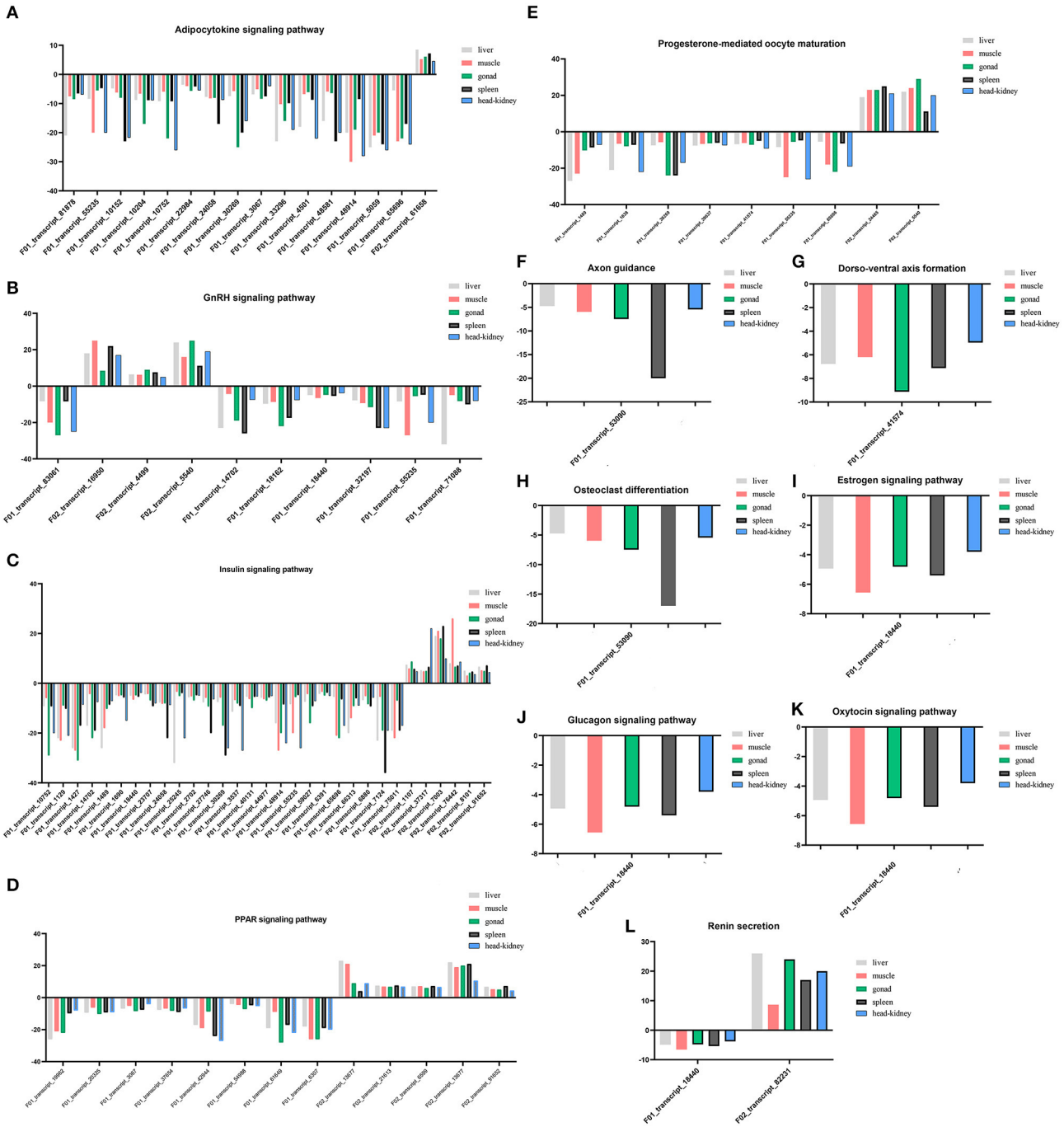

FIGURE 4 | (A–L) Expression profiles of DEGs enriched in pathways involved in endocrine system and development in five tissues.

Validation Analysis

Expression Analysis of p38 MAPK in *B. lenok*

Transcriptome sequencing results showed that p38 MAPK was significantly differentially expressed in mature and immature gonads (**Table 4**). To study the expression pattern of p38 MAPK in lenok, the expression of p38 MAPK protein was detected in eight tissues of lenok, including the gill, heart, liver, spleen, intestine, brain, skin, and gonads (**Figure 6A**), and the results showed that the expression of p38 MAPK was higher in the intestine, heart, and brain tissues, but was not expressed in the gill and spleen. The expression levels of p38 MAPK protein were also detected in the ovaries of lenok at different developmental stages. The results showed that the expression of p38 MAPK at 750 dph was 2.8-fold that at 300 dph, and the expression at three

TABLE 4 | Specifically expressed genes of the top five |log2FC| in gonad.

Gene	|log2FC|	Description
F01_transcript_1427	−31.50	Protein of unknown function
F02_transcript_5540	−29.78	p38 mitogen activated protein kinase
F01_transcript_61649	−28.96	Protein of unknown function
F01_transcript_83061	−26.47	Guanine nucleotide-binding protein subunit alpha-11
F01_transcript_10752	29.37	5'-AMP-activated protein kinase

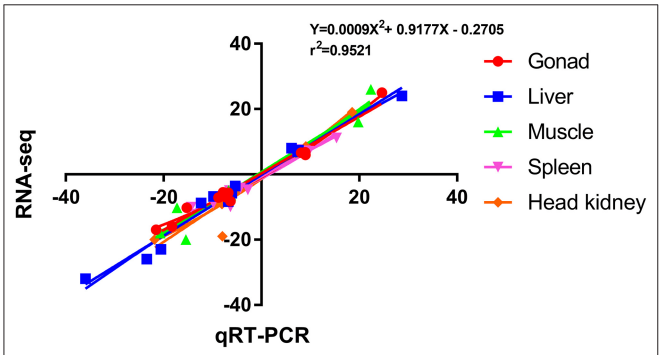

FIGURE 5 | Scatterplot of the correlation between qRT-PCR and RNA-seq. X-axis, qRT-PCR; Y-axis, RNA-seq; r2, Pearson correlation coefficient.

years ph was 2.0-fold of that at 750 dph and 3.1-fold of that at 300 dph (**Figure 6B**). The results showed that the expression of p38 MAPK in the ovary significantly increased with the growth and development of lenok.

Effects of p38 MAPK on Ovarian Development of *B. lenok*

To verify the regulatory effect of p38 MAPK on ovarian development in lenok, an inhibitor of SB203580 was injected intraperitoneally to inhibit p38 MAPK protein phosphorylation. The results showed that the expression levels of p38 MAPK protein in the SB203580 inhibitor, DMSO, and control groups were not significantly different (**Figure 7A**). However, the phosphorylation levels of p38 in the SB203580 inhibitor group were 0.32-fold that of the control group and 0.55 times that of the DMSO group, indicating a significant decreasing trend (**Figure 7B**).

To explore the effect of the p38 MAPK pathway on the balance of hormones in the hypothalamic-pituitary-gonadal (HPG) axis of lenok, gonadotropin-releasing hormone 3 (GnRH3), follicle stimulating hormone (FSH), luteinizing hormone (LH), and oestradiol (E2) hormone levels were detected in the plasma of lenok in the SB203580 inhibitor, DMSO, and control groups, respectively. The values of OD450 in different groups are shown in **Supplementary Table S4**. The repetition rate of the GnRH3 assay was 1.7–7.8%, the FSH was 1.9–9.6%, the LH was 1.3–4.7%, and the E2 was 1.8–7.8%. The variable coefficients of all samples were no greater than 10%, proving that our ELISA results had good repeatability.

The results showed that the level of GnRH3 in the inhibitor group was 1.4-fold higher than that in the control group ($P < 0.05$) and 1.9-fold higher than that in the DMSO group ($P < 0.05$), and there was no significant difference between the control and DMSO groups ($P > 0.05$). The FSH level in the inhibitor group was 0.6-fold lower than that in the control group ($P < 0.05$), and there was no significant difference between the inhibitor and DMSO groups ($P > 0.05$). The LH levels in the three groups were not significantly different ($P > 0.05$). The level of E2 in the inhibitor group was significantly lower than that in the control and DMSO groups, which was 0.4-fold higher than that in the control group ($P < 0.01$) and 0.5-fold higher than that in the DMSO group ($P < 0.05$) (**Figure 7C**).

Immunohistochemical sections labeled with p38 MAPK protein were used to explore the morphological changes in lenok ovaries after p38 MAPK activity was inhibited. There were oocytes at phases II and III in the control group. The cell volume was larger, and the cytoplasm increased with the nucleus inside. A flat layer of nucleolar structure overlapped the oocyte and the theca cells around the follicle were complete (**Figure 7D**). In the DMSO group, the morphology of follicles was almost consistent with that of the control group, and only theca cells appeared irregular (**Figure 7E**), which may be because of the toxic effect of DMSO on the ovary. In the inhibitor group, the structure of follicles was obviously changed, with the rupture of cell nucleus, vacuoles in follicles, and atrophy of theca cells, which showed irregular arrangement (**Figure 7F**).

DISCUSSION

In this study, SMRT technology based on the PacBio platform was used to sequence the mixed tissue samples of immature and mature lenok, including the liver, muscle, head kidney, gonad, and spleen. SMRT technology is superior to next-generation sequencing technologies such as Illumina, which can conduct de novo sequencing of the complete mRNA to obtain the full-length information of the transcripts directly. This method can provide accurate reference sequences, overcome the problems of short transcription splicing and incomplete information of species without reference genomes, and contribute to the screening and identification of functional genes. As a rare cold-water fish native to China, the wild population of *B. lenok* is decreasing annually. The analysis of genetic mechanisms plays an important role in the establishment of artificial breeding technology and genetic improvement of lenok. As a non-model organism, the identification of functional genes in lenok is limited because of

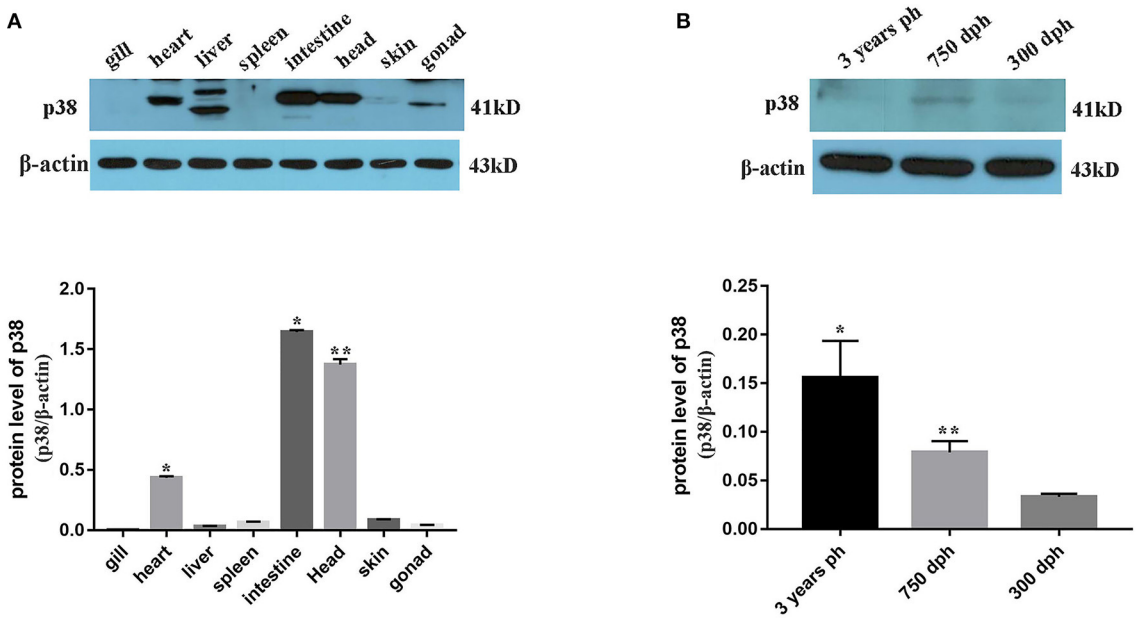

FIGURE 6 | Expression pattern of p38 MAPK protein in *B. lenok*. **(A)** The expression level of p38 MAPK protein in 8 tissues. The significance was expressed "*" as a *P*-value of < 0.05 vs. gill, and "**" as a *P*-value of < 0.01 vs. gill. **(B)** The expression level of p38 MAPK protein in gonad at different developmental stages. The significance was expressed "*" as a *P*-value of < 0.05 vs. 300 dph, and "**" as a *P*-value of < 0.01 vs. 300 dph.

the lack of a reference genome. In this study, key genes and pathways related to the gonadal development of lenok were screened using SMRT sequencing and corrected using Illumina sequencing, and the function of the gene involved in ovarian development of lenok was explored.

The Iso-seq results showed that the lengths of the circular consensus sequence (CCS) reads were 2,791 and 2,977, respectively, and the percentage of FLNC reads with the 5′ end, 3′ end, and poly-A structures were 84.09 and 86.36%, respectively, for mature and immature samples. The length of N50 and proportion of FLNC were better than those of red swamp crayfish (39) and Hong Kong catfish (40). After NGS correction, 61,405 and 59,372 non-redundant transcripts were obtained for the mature and immature groups, respectively, for subsequent expression levels and pathway enrichment analysis. These transcripts were evaluated by BUSCO, and the transcripts that encoded complete proteins accounted for 85.69 and 83.68% in the mature and immature groups, respectively. The proportion of complete transcripts in this study was higher than that obtained from the full-length transcriptome of other aquatic animals, such as Atlantic bluefin tuna (80%) (41) and shrimp (81%) (11). These results proved that our sequencing results were of high quality and were reliable for the analysis of functional gene information.

Eukaryotic transcription factors can specifically bind to the upstream sequence of the 5'-end of a specific gene, thus ensuring that the gene is expressed at a specific intensity at a specific time and space (42). Studies have shown that TFs play important roles in fish morphogenesis (43), growth and development (44), gonadal maturation (45), and immune regulation (46). The sequencing results of the full-length transcriptome in this study showed that the TFs of the zf-C2H2, ZBTB, BHLH, homeobox, miscellaneous, TF-BZIP, and HMG families appeared in large numbers, suggesting that they play pivotal roles in the growth and development of lenok. The C2H2 zinc finger (zf-C2H2) proteins are the most abundant transcriptional regulatory factors in mammals. Most of the zinc finger motifs of zf-C2H2 proteins are not conserved, indicating that they may bind to different DNA sequences to regulate different genes and perform diverse regulatory functions. Most human zf-C2H2 proteins are completely different from those of other species, such as mice, so the function of zf-C2H2 proteins is not consistent between different species (47). In this study, we concluded that the zf-C2H2 protein was the most abundant type of TF in lenok, and its specific function should be further studied. The ZBTB family refers to a class of proteins containing the N-terminal BTB domain and multiple zinc finger domains at the C-terminal (48), and more than 60 types of ZBTB proteins have been identified as being involved in development, differentiation, and tumor formation (49). Recent studies have shown that *zbtb16* can regulate spermatogenesis by controlling the self-renewal and repair of spermatogonia (50, 51). In orange-spotted grouper, *zbtb40* is specifically expressed in male germ cells and regulates spermatogenesis through its interaction with *cyp17a1* (49).

Different tissues play important regulatory roles in the growth and development of fish. Through transcriptome screening, DEGs were identified in the gonad, head kidney, liver, muscle, and spleen tissues between mature and immature lenok groups. Most DEGs were found in the spleen (10,837), followed by the head kidney (10, 352), muscle (7,968), gonad (7,728),

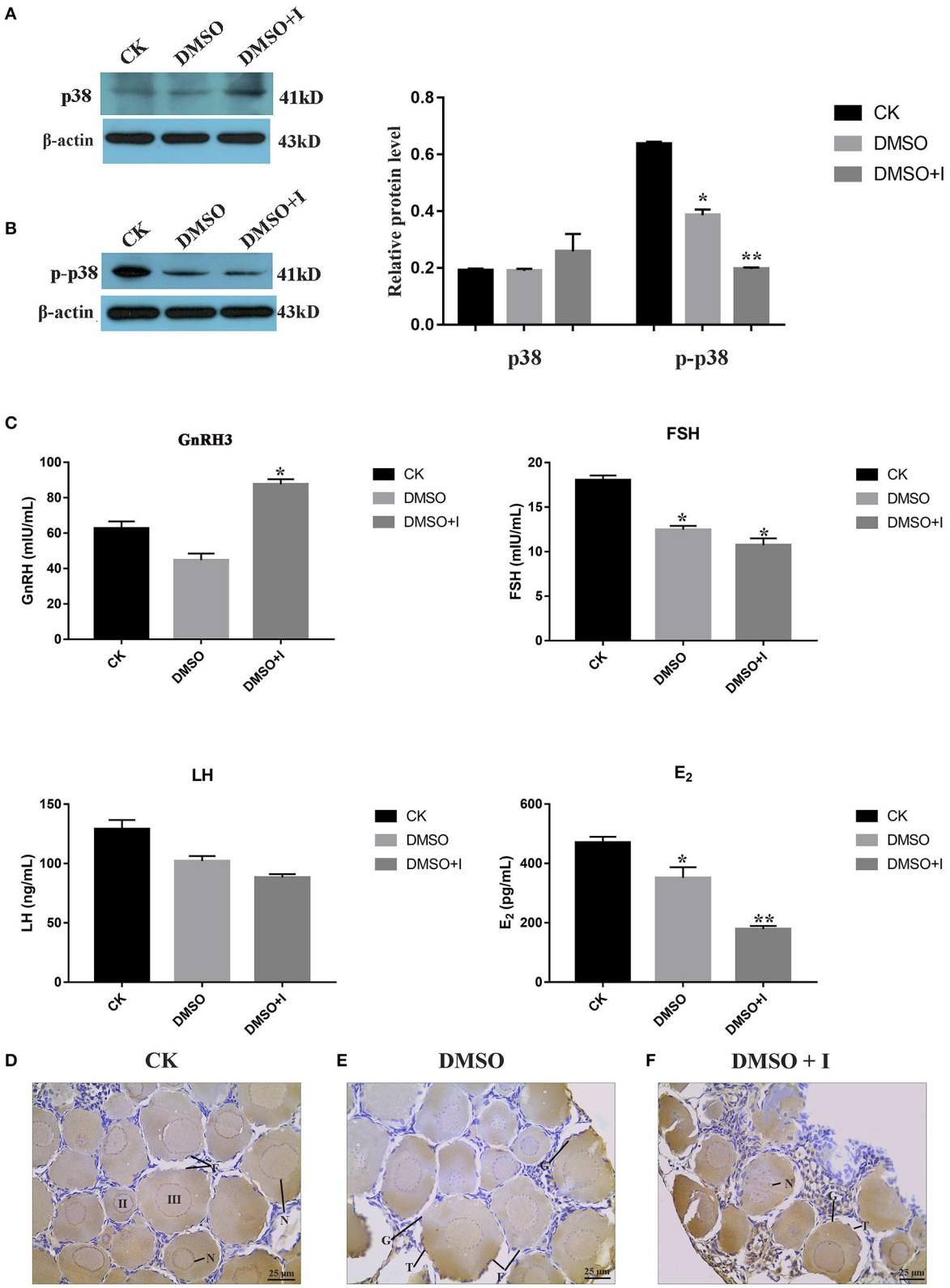

FIGURE 7 | Effects of p38 MAPK on ovarian development of *B. lenok*. **(A)** The expression level of p38 MAPK protein in ovary; **(B)** The phosphorylation level of p38 MAPK protein in ovary of different groups; **(C)** The level of HPG hormones of different groups; The significance was expressed "*" as a *P*-value of < 0.05 vs. CK group, and "**" as a *P*-value of < 0.01 vs. CK group. **(D–F)** Ovarian morphology of different groups observed by immunohistochemistry. II, oocyte of phasell; III, oocyte of phase III; F, follicle; T, theca cell; G, granulosa cell; N, nucleus.

and liver (6,690). The spleen and head kidney are important haematopoietic tissues in fish and play important roles in the generation, storage, and maturation of red blood cells and granulocytes (52, 53). The head kidney also contains phagocytes and B cells, and is an important organ for the production of antibodies (54). In many fish, the ability of specific immunity gradually increases with growth and development (55), and therefore, the number of DEGs in the spleen and head kidney was higher. Compared with the mature group, the number of downregulated DEGs was higher than that of the upregulated DEGs in the immature group, which indicated that the expression of most genes showed a significantly increased trend with the growth and development of lenok.

Due to the rapid decline in natural resources, it is of great practical significance to conduct the artificial reproduction and genetic breeding of lenok. Therefore, it is of great importance to understand the regulatory mechanisms of gonadal development and maturation of lenok. Gonadal developmental processes, such as germline generation, proliferation, and yolk accumulation in fish are regulated by the endocrine system (56). DEGs enriched in pathways involved in the endocrine system were analyzed, and most DEGs were found to be enriched in the *insulin signaling pathway* (ko04910) and the *peroxisome proliferator-activated receptor (PPAR) signaling pathway* (ko03320). Insulin plays an important role in the reproductive process of female animals and has direct and indirect effects on the production of ovarian steroid hormones and the growth of granulosa and theca cells (57). There are many important effectors in the insulin pathway, including insulin receptor (IR), insulin receptor substrate (IRS), phosphatidylinositol 3 kinase (PI3K), and protein kinase B (AKT). IR expression has been detected in granulosa and theca cells and follicles in a variety of animals (58–60) at different stages of development (61). PI3K can regulate Akt to participate in many physiological processes, and the PI3K/Akt pathway plays key regulatory roles in follicular structural differentiation, growth, and development (62). In this study, we preliminarily showed that the insulin signaling pathway plays an essential role in the regulation of the ovarian development in lenok. PPARs are important in the reproductive system, particularly in the HPG axis (63). Some studies have demonstrated that PPARs can regulate ovum proliferation, tissue remodeling, and hormone synthesis (64).

The expression patterns of DEGs enriched in endocrine and development-related pathways were analyzed in the gonads of lenok. The results indicated that p38 MAPK, GNA11, and AMPK were significantly differentially expressed between mature and immature gonads, but the functions of these genes in lenok remain unknown. As a central substance of cell energy metabolism, adenylate activated protein kinase (AMPK) plays a non-negligible role in the process of ovarian development (65). Glucose and adiponectin promote AMPK phosphorylation (66), whereas IGF-1 and FSH inhibit AMPK phosphorylation (67). Phosphorylated AMPK promotes oocyte development, but negatively regulates follicle and granulosa cell development (68). As one of the main branches of the MAPK signaling pathway, p38 MAPK can be activated in response to a variety of environmental stresses or inflammatory stimuli, thereby promoting apoptosis and inhibiting cell growth (22). However, research on lenok remains in its infancy. In this study, the expression of p38 MAPK protein was detected in different tissues of lenok, and the results showed that p38 MAPK plays a role in the regulation of intestinal, heart, and brain tissues. The expression pattern of p38 MAPK protein was simultaneously explored at different developmental stages of the lenok ovary. Under the same artificial feeding conditions, the expression levels of p38 MAPK protein increased gradually with the growth and development of lenok.

SB203580 is a pyridine imidazole derivative that can inhibit the catalytic activity of p38 MAPK by binding competitively to the ATP sites (69), and thus, can specifically inhibit the p38 MAPK signaling pathway (70, 71). The inhibitor was injected intraperitoneally to investigate the effects of p38 MAPK on steroid hormone levels and follicular structures of lenok. The results showed that under the influence of the inhibitor SB203580, the phosphorylation levels of p38 MAPK in the ovary were significantly decreased, indicating that SB203580 blocked the p38 MAPK signaling pathway. We also investigated the influence of p38 MAPK pathway inhibition on the HPG axis of lenok. The synthesis and secretion of GnRH in the hypothalamus promotes LH and FSH synthesis and secretion in the pituitary, thereby stimulating the synthesis and secretion of E2 in the ovary (72). Hormones in the HPG axis maintain normal development of the ovary and oogenesis and regulate the physiological functions of the reproductive system of fish (73). The levels of major reproductive hormones in the plasma were monitored to explore the effects of p38 MAPK signaling pathway inhibition on lenok reproductive hormone balance. Oocytes at 300 dph (immature) ovaries were in the early stage, while a large amount of yolk existed in follicles at three years ph (mature), which would have a substantial influence on histological experiments. However, oocytes of various stages existed in the ovary of lenok at 750 dph, which was more suitable for function verification.

GnRH is synthesized and secreted by GnRH neurons in the medial hypothalamus and maintains hormone balance and homeostasis through autocrine and paracrine mechanisms under the synergistic effect of related hormones (74–76). Compared with the CK and DMSO groups, the GnRH3 level in the inhibitor group increased significantly, indicating that the hormone level of the HPG axis was abnormal, resulting in the continuous secretion of GnRH3 hormone in the hypothalamus. However, no serious tissue damage was caused to the hypothalamus. FSH and LH are secreted by the anterior pituitary gland, and FSH binds to specific receptors and stimulates the formation of LH receptors (77, 78). In addition, FSH acts synergistically with LH to promote follicle maturation and stimulate E2 synthesis and secretion in the ovary (79, 80). In this study, it was found that compared with the CK group, FSH levels were significantly decreased in the DMSO group, which may be caused by toxicity of DMSO. In the inhibitor group, FSH levels were significantly decreased, whereas there was no significant effect on LH levels. Previous studies have shown that inhibition of the p38 MAPK pathway can weaken the stimulatory effect of FSH on E2 and cytochrome P450 aromatase (81). Combined with the results of this study, it was confirmed that the p38 MAPK pathway is involved in the synthesis and function of FSH in lenok.

E2 is a steroid hormone produced by follicular theca cells and is the main estrogen of teleosts (82). It plays important reproductive roles, such as promoting follicular development and regulating the oestrus cycle (83). The results of this experiment showed that the level of E2 in the inhibitor group was significantly lower than that in the CK group. Previous studies have shown that inhibition of p38 MAPK activity can inhibit the generation of E2 (84). In addition, combined with the results of immunohistochemistry in this study, inhibition of p38 MAPK activity could cause atrophy and irregular arrangement of granulosa and theca cells, where E2 was mainly secreted. Thus, inhibition of p38 MAPK activity could result in morphological and functional changes in the granulosa and theca cells of lenok and affect the synthesis and secretion of E2, leading to HPG axis blocking and abnormal follicular development. Therefore, the p38 MAPK pathway plays an important role in maintaining the balance between reproductive hormones and follicle development in lenok.

CONCLUSION

This is the first comparative transcriptome analysis of *B. lenok* combined with SMRT and NGS, and for the first time, the DEGs between immature and mature lenok were analyzed in five tissues. Furthermore, DEGs and pathways involved in the endocrine system and gonadal development were identified, and p38 MAPK was identified to potentially regulate gonadal development in lenok. Inhibiting the activity of p38 MAPK blocked the HPG axis and abnormal follicular development in lenok. Our study illustrates the basic regulatory mechanism of ovarian development and provides a reference for genetic improvement in lenok.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, PRJNA669274 and https://www.ncbi.nlm.nih.gov/, PRJNA669219.

ETHICS STATEMENT

The animal study was reviewed and approved by all experiments were performed according to the European Communities Council Directive (86/609/EEC). The performances were approved by the Animal Husbandry Department of Heilongjiang Animal Care and Use Committee (202110384464).

AUTHOR CONTRIBUTIONS

TH designed and performed the experiments. GX analyzed the data and checked the manuscript. WG, LZ, WJ, and CL cultured and sampled the fish. EL and FD drafted the manuscript. XH and BW reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the China Agriculture Research System of MOF and MARA (CARS-46) and the Central Public-interest Scientific Institution Basal Research Fund, CAFS (No. 2020TD32). The funding agency played no part in the study design, data collection and analysis, decision to publish, or manuscript preparation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fvets.2022.752521/full#supplementary-material

Supplementary Figure S1 | The correlation of samples for next generation sequencing.

Supplementary Figure S2 | The standard curves of Elisa kits for GnRH3, LH, FSH, and E2 in lenok.

Supplementary Table S1 | The analysis of SSR by MISA.

Supplementary Table S2 | The information about the DEGs enriched in pathways involved in endocrine systems and developments.

Supplementary Table S3 | The primer information of DEGs selected for qRT-PCR verification.

Supplementary Table S4 | The results of intra-assay repeat tests of ELISA.

REFERENCES

1. Wang S, Zheng GM, Wang QS. *China Red Data Book of Endangered Animals: Pisces (in Chinese with English Translation)*. Science Press, Beijing. (1998).
2. Chang J, Niu HX, Jia YD, Li SG, Xu GF. Effects of dietary lipid levels on growth, feed utilization, digestive tract enzyme activity and lipid deposition of juvenile Manchurian trout, Brachymystax lenok (Pallas). *Aquaculture Nutrition.* (2017) 24:694–701. doi: 10.1111/anu.12598
3. Yu J, Li S, Chang J, Niu H, Hu Z, Han Y. Effect of variation in the dietary ratio of linseed oil to fish oil on growth, body composition, tissues fatty acid composition, flesh nutritional value and immune indices in Manchurian trout, Brachymystax lenok. *Aquaculture Nutr.* (2019) 25:377–87. doi: 10.1111/anu.12863
4. Olson KW, Jensen OP, Hrabik TR. Feeding ecology and prey resource partitioning of lenok (B rachymystax lenok) and Baikal grayling (T hymallus arcticus baicalensis) in the Eg and Uur rivers. *Mongolia*. (2016) 25:565–576. doi: 10.1111/eff.12234
5. Tomas S, Underwood JG, Tseng E, Holloway AK. Long-read sequencing of chicken transcripts and identifcation of new transcript isoforms. *PLoS ONE.* (2014) 9:e94650. doi: 10.1371/journal.pone.0094650
6. Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, Schilkey F, et al. A survey of the sorghum transcriptome using single-molecule long reads. *Nat Commun.* (2016) 7:11706. doi: 10.1038/ncomms11706
7. Wang B, Tseng E, Regulski M, Clark TA, Hon T, Jiao Y, et al. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat Commun.* (2016) 7:11708. doi: 10.1038/ncomms11708
8. Larsen PA, Smith TP. Application of circular consensus sequencing and network analysis to characterize the bovine Ig G repertoire. *BMC Immunol.* (2012) 13:52. doi: 10.1186/1471-2172-13-52

9. Chen SY, Deng F, Jia X, Li C, Lai SJ. A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. *Sci Reports.* (2017) 7:7648. doi: 10.1038/s41598-017-08138-z
10. Treutlein B, Gokce O, Quake SR, Südhof TC. Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. *Proc Natl A-cad Sci USA*. (2014) 111:1291–9. doi: 10.1073/pnas.1403244111
11. Zeng D, Chen X, Peng J, Yang C, Peng M, Zhu W, et al. Single-molecule long-read sequencing facilitates shrimp transcriptome research. *Sci Reports.* (2018) 8:16920. doi: 10.1038/s41598-018-35066-3
12. Nudelman G, Frasca A, Kent B, Sadler KC, Sealfon SC, Walsh MJ, et al. High resolution annotation of zebrafish transcriptome using longread sequencing. *Genome Res.* (2018) 28:1415–25. doi: 10.1101/gr.223586.117
13. Feng X, Jia Y, Zhu R, Chen K, Chen Y. Characterization and analysis of the transcriptome in Gymnocypris selincuoensis on the Qinghai-Tibetan Plateau using single-molecule long-read sequencing and RNA-seq. *DNA Res.* (2019) 26:353–63. doi: 10.1093/dnares/dsz014
14. Tian Y, Wen H, Qi X, Zhang X, Liu S, Li B, et al. Characterization of full-length transcriptome sequences and splice variants of Lateolabrax maculatus by singlemolecule long-read sequencing and their involvement in salinity regulation. *Front Genet.* (2019) 10:1126–45. doi: 10.3389/fgene.2019.01126
15. Yang HX, Zhou Y, Gu JL, Xie SY, Xu Y, Zhu GF, et al. Deep mRNA sequencing analysis to capture the transcriptomelandscape of zebrafish embryos and larvae. *PLoS ONE.* (2013) 8:64058. doi: 10.1371/journal.pone.0064058
16. Sørhus E, Incardona JP, Furmanek T, Jentoft S, Meier S, Edvardsen RB. Developmental transcriptomics in Atlantic haddock: Illuminating pattern formation and organogenesis in non-model vertebrates. *Dev Biol.* (2016) 411:301–13. doi: 10.1016/j.ydbio.2016.02.012
17. Zeng QF, Liu SK, Yao J, Zhang Y, Yuan ZH, Jiang C, et al. Transcriptome display during testicular differentiation of Channel catfish (Ictalurus punctatus) as revealed by RNA-seq analysis. *Biol Reprod.* (2016) 95:19. doi: 10.1095/biolreprod.116.138818
18. Du X, Wang B, Liu X, Liu X, He Y, Zhang Q, et al. Comparative transcriptome analysis of ovary and testis reveals potential sex-related genes and pathways in spotted knifejaw Oplegnathus punctatus. *Gene.* (2017) 637:203–10. doi: 10.1016/j.gene.2017.09.055
19. Wang Z, Qiu X, Kong D, Zhou X, Guo Z, Gao C, et al. Comparative RNA-Seq analysis of differentially expressed genes in the testis and ovary of Takifugu rubripes. *Comp Biochem Physiol Part D Genomics Proteomics.* (2017) 22:50–7. doi: 10.1016/j.cbd.2017.02.002
20. Presslauer C, Tilahun Bizuayehu T, Kopp M, Fernandes JM, Babiak I. Dynamics of miRNA transcriptome during gonadal development of zebrafish. *Sentific Reports*. (2017) 7:43850. doi: 10.1038/srep43850
21. Widmann C, Gibson S, Jarpe MB, Johnson GL. Mitogen-activated protein kinase: conservation of a three-kinase from yeast to human. *Physiol Rev*. (1999) 79:143–80. doi: 10.1152/physrev.1999.79.1.143
22. Chen S, Yang W, Zhang X, Jin J, Liang C, Wang J, et al. Melamine induces reproductive dysfunction via down-regulated the phosphorylation of p38 and downstream transcription factors Max and Sap1a in mice testes. *Sci Total Environ.* (2021) 770:144727. doi: 10.1016/j.scitotenv.2020.144727
23. Jin J, Ma Y, Tong X, Yang W, Dai Y, Pan Y, et al. Metformin inhibits testosterone-induced endoplasmic reticulum stress in ovarian granulosa cells via inactivation of p38 MAPK. *Hum Reprod*. (2020) 35:1145–58. doi: 10.1093/humrep/deaa077
24. Deng Y, Zheng H, Yan Z, Liao D, Li C, Zhou J, et al. Full-length transcriptome survey and expression analysis of cassia obtusifolia to discover putative genes related to aurantio-obtusin biosynthesis, seed formation and development, and stress response. *Int J Mol Sci.* (2018) 19:2476. doi: 10.3390/ijms19092476
25. Westen AA, van der Gaag KJ, de Knijff P, Sijen T. Improved analysis of long STR amplicons from degraded single source and mixed DNA. *Int J Legal Med.* (2013) 127:741–7. doi: 10.1007/s00414-012-0816-1
26. Modi A, Vai S, Caramelli D, Lari M. The illumina sequencing protocol and the novaseq 6000 system. *Methods Mol Biol.* (2021) 2242:15–42. doi: 10.1007/978-1-0716-1099-2_2
27. Hackl T, Hedrich R, Schultz J, Förster F. Proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics.* (2014) 30:3004–11. doi: 10.1093/bioinformatics/btu392
28. Jia D, Wang Y, Liu Y, Hu J, Guo Y, Gao L, et al. SMRT sequencing of full-length transcriptome of flea beetle Agasicles hygrophila (Selman and Vogt). *Sci Rep.* (2018) 8:2197. doi: 10.1038/s41598-018-20181-y
29. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods.* (2013) 10:563. doi: 10.1038/nmeth.2474
30. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods.* (2015) 12:357–60. doi: 10.1038/nmeth.3317
31. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* (2010) 26:139–40. doi: 10.1093/bioinformatics/btp616
32. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics.* (2013) 29:2073. doi: 10.1093/bioinformatics/btt337
33. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* (1997) 25:3389–402. doi: 10.1093/nar/25.17.3389
34. Young MD, Wakefeld MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* (2010) 11:R14. doi: 10.1186/gb-2010-11-2-r14
35. Wu J, Mao X, Cai T, Luo J, Wei L. KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res.* (2006) 34:W720–4. doi: 10.1093/nar/gkl167
36. Argyrous G. *Statistics for Research: With a Guide to SPSS.* London: SAGE. (2005).
37. D'Cotta H, Fostier A, Guiguen Y, Govoroun M, Baroiller JF. Aromatase plays a key role during normal and temperature-induced sex differentiation of tilapia oreochromis niloticus. *Mol Reprod Dev.* (2001) 59:265–76. doi: 10.1002/mrd.1031
38. Liu YX, Zhang Y, Li YY, Liu XM, Wang XX, Zhang CL, et al. Regulation of follicular development and differentiation by intra-ovarian factors and endocrine hormones. *Front Biosci.* (2019) 24:983–93. doi: 10.2741/4763
39. Shen H, Hu Y, Ma Y, Zhou X, Xu Z, Shui Y, et al. In-depth transcriptome analysis of the red swamp crayfish Procambarus clarkii. *PLoS ONE.* (2014) 9:e110548. doi: 10.1371/journal.pone.0110548
40. Lin X, Zhou D, Zhang X, Li G, Zhang Y, Huang C, et al. A first insight into the gonad transcriptome of Hong Kong Catfish (Clarias fuscus). *Animals*. (2021) 11:1131. doi: 10.3390/ani11041131
41. Marisaldi L, Basili D, Gioacchini G, Canapa A, Carnevali O. De novo transcriptome assembly, functional annotation and characterization of the Atlantic bluefin tuna (Thunnus thynnus) larval stage. *Mar Genomics.* (2021) 58:100834. doi: 10.1016/j.margen.2020.100834
42. GuhaThakurta D. Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Res.* (2006) 34:3585–98. doi: 10.1093/nar/gkl372
43. Marra AN, Cheng CN, Adeeb B, Addiego A, Wesselman HM, Chambers BE, et al. Iroquois transcription factor irx2a is required for multiciliated and transporter cell fate decisions during zebrafish pronephros development. *Sci Rep*. (2019) 9:6454. doi: 10.1038/s41598-019-42943-y
44. Cioffi CC, Pollenz RS, Middleton DL, Wilson MR, Miller NW, William Clem L, et al. Oct2 transcription factor of a teleost fish: activation domains and function from an enhancer. *Arch Biochem Biophys.* (2002) 404:55–61. doi: 10.1016/S0003-9861(02)00227-8
45. Jiang YH, Han KH, Wang SH, Chen Y, Wang YL, Zhang ZP. Identification and expression of transcription factor sox2 in large yellow croaker Larimichthys crocea. *Theriogenology*. (2018) 120:123–37. doi: 10.1016/j.theriogenology.2018.07.025
46. Yao F, Liu Y, Du L, Wang X, Zhang A, Wei H, et al. Molecular identification of transcription factor Runx1 variants in grass carp (Ctenopharyngodon idella) and their responses to immune stimuli. *Vet Immunol Immunopathol.* (2014) 160:201–8. doi: 10.1016/j.vetimm.2014.05.002
47. Schmitges FW, Radovani E, Najafabadi HS, Barazandeh M, Campitelli LF, Yin Y, et al. Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res.* (2016) 26:1742–52. doi: 10.1101/gr.209643.116

48. Lee SU, Maeda T. POK/ZBTB proteins: an emerging family of proteins that regulate lymphoid development and function. *Immunol Rev.* (2012) 247:107–19. doi: 10.1111/j.1600-065X.2012.01116.x
49. Wu X, Yang Y, Zhong C, Guo Y, Li S, Lin H, et al. Transcriptome profiling of laser-captured germ cells and functional characterization of zbtb40 during 17alpha - methyltestosterone - induced spermatogenesis in orange-spotted grouper (Epinephelus coioides). *BMC Genomics.* (2020) 21:1–13. doi: 10.1186/s12864-020-6477-4
50. Buaas FW, Kirsh AL, Sharma M, McLean DJ, Morris JL, Griswold MD, et al. Plzf is required in adult male germ cells for stem cell selfrenewal. *Nat Genet.* (2004) 36:647–52. doi: 10.1038/ng1366
51. Lovelace DL, Gao Z, Mutoji K, Song YC, Ruan J, Hermann BP. The regulatory repertoire of PLZF and SALL4 in undifferentiated spermatogonia. *Development.* (2016) 143:1893–906. doi: 10.1242/dev.132761
52. Van Muiswinkel WB, Lamers CH, Rombout JH. Structural and functional aspects of the spleen in bony fish. *Res Immunol.* (1991) 142:362–6. doi: 10.1016/0923-2494(91)90093-X
53. Bates T, Naumann U, Hoppe B, Englert C. Kidney regeneration in fish. *Int J Dev Biol.* (2018) 62:419–29. doi: 10.1387/ijdb.170335ce
54. Geven EJW, Klaren PHM. The teleost head kidney: integrating thyroid and immune signalling. *Dev Comp Immunol.* (2017) 66:73–83. doi: 10.1016/j.dci.2016.06.025
55. Zhang Z, Chi H, Dalmo RA. Trained innate immunity of fish is a viable approach in larval aquaculture. *Front Immunol.* (2019) 10:42. doi: 10.3389/fimmu.2019.00042
56. Cowan M, Azpeleta C, López-Olmeda JF. Rhythms in the endocrine system of fish: a review. *J Comp Physiol B.* (2017) 187:1057–89. doi: 10.1007/s00360-017-1094-5
57. Reinecke M. Insulin-like growth factors and fish reproduction. *Biol Reprod.* (2010) 82:656–61. doi: 10.1095/biolreprod.109.080093
58. Dupont J, Scaramuzzi RJ. Insulin signalling and glucose transport in the ovary and ovarian function during the ovarian cycle. *Biochem J.* (2016) 473:1483e1501. doi: 10.1042/BCJ20160124
59. Otani T, Maruo T, Yukimura N, Mochizuki M. Effect of insulin on porcine granulosa cells: implications of a possible receptor mediated action. *Acta Endocrinol.* (1985) 108:104–10. doi: 10.1530/acta.0.10 80104
60. Samoto T, Maruo T, Ladines-Llave CA, Matsuo H, Deguchi JU, Barnea ER, et al. Insulin receptor expression in follicular and stromal compartments of the human ovary over the course of follicular growth, regression and atresia. *Endocr J.* (1993) 40:715–26. doi: 10.1507/endocrj.40.715
61. Bossaert P, De Cock H, Leroy JLMR, De Campeneere S, Bols PEJ, Filliers M, et al. Immunohistochemical visualization of insulin receptors in formalin-fixed bovine ovaries post mortem and in granulosa cells collected *in vivo*. *Theriogenology.* (2010) 73:1210–9. doi: 10.1016/j.theriogenology.2010.01.012
62. Zheng W, Nagaraju G, Liu Z, Liu K. Functional roles of the phosphatidylinositol 3-kinases (PI3Ks) signaling in the mammalian ovary. *Mol Cell Endocrinol.* (2012) 356:24–30. doi: 10.1016/j.mce.2011.05.027
63. Bogacka I, Kurzynska A, Bogacki M, Chojnowska K. Peroxisome proliferator-activated receptors in the regulation of female reproductive functions. *Folia Histochemica Cytobiologica.* (2015) 53:189–200. doi: 10.5603/fhc.a2015.0023
64. Barker HM, Brewis ND, Street AJ, Spurr NK, Cohen PT. Three genes for protein phosphatase 1 map to different human chromosomes: sequence, expression and gene localisation of protein serine/threonine phosphatase 1 beta (PPP1CB). *Biochim Biophys Acta.* (1994) 1220:212–8. doi: 10.1016/0167-4889(94)90138-4
65. Downs SM, Hudson ER, Hardie DG. A potential role for AMP-activated protein kinase in meiotic induction in mouse oocytes. *Dev Biol.* (2002) 245:200–12. doi: 10.1006/dbio.2002.0613
66. Dupont J, Reverchon M, Cloix L, Froment P, Rame C. Involvement of adipokines,AMPK, PI3K and the PPAR signaling pathways in ovarian follicle development and cancer. *Int J Develop Biol.* (2012) 56:959–67. doi: 10.1387/ijdb.120134jd
67. Reverchon M, Cornuau M, Rame C, Guerif F, Royere D, Dupont J. Chemerin inhibits IGF-1-induced progesterone and es- tradiol secretion in human granulosa cells. *Human Reproduction.* (2012) 27:1790–800. doi: 10.1093/humrep/des089
68. Chappaz E, Albornoz MS, Campos D, Che L, Palin MF, Murphy BD, et al. Adiponectin enhances in vitro development of swine embryos. *Domest Anim Endocrinol.* (2008) 35:198–207. doi: 10.1016/j.domaniend.2008.05.007
69. Young PR, McLaughlin MM, Kumar S, Kassis S, Doyle ML, McNulty D, et al. Pyridinyl imidazole inhibitors of p38 mitogen-activated protein kinase bind in the ATP site. *J Biol Chem.* (1997) 272:12116–21. doi: 10.1074/jbc.272.18.12116
70. Zhang Y, Zhang K, Zhang Y, Zhou L, Huang H, Wang J. IL-18 Mediates vascular calcification induced by high-fat diet in rats with chronic renal failure. *Front Cardiovasc Med.* (2021) 25:724233. doi: 10.3389/fcvm.2021.724233
71. Lu ZY, Fan J, Yu LH, Ma B, Cheng LM. The Up-regulation of TNF-α maintains trigeminal neuralgia by modulating MAPKs phosphorylation and BKCa channels in trigeminal nucleus caudalis. *Front Cell Neurosci.* (2021) 15:764141. doi: 10.3389/fncel.2021.764141
72. O'Neil MM, Korthanke CM, Scarpa JO, Welsh TH, Cardoso RC, Williams GL. Differential regulation of gonadotropins in response to continuous infusion of native gonadotropin-releasing hormone in the winter anovulatory mare and effects of treatment with estradiol-17β. *J Equine Vet Sci.* (2019) 75:93–103. doi: 10.1016/j.jevs.2019.01.013
73. Scarmuzzi RJ, Brown HM, Dupont J. Nutritional and metabolic mechanisms in the ovary and their role in mediating the effects of diet on folliculogenesis: a perspective. *Reprod Domestic Animals.* (2010) 45:32–41. doi: 10.1111/j.1439-0531.2010.01662.x
74. Ortmann O, Weiss JM, Diedrich K. Gonadotrophin-releasing hormone (GnRH) and GnRH agonists: mechanisms of action. *Reprod Biomed Online.* (2002) 5:1–7. doi: 10.1016/S1472-6483(11)60210-1
75. Breton B, Govoroun M, Mikolajczyk T. GTH I and GTH II secretion profiles during the reproductive cycle in female rainbow trout: relationship with pituitary responsiveness to GnRH-A stimulation. *Gen Comp Endocrinol.* (1998) 111:38–50. doi: 10.1006/gcen.1998.7088
76. Hildahl J, Sandvik GK, Edvardsen RB, Fagernes C, Norberg B, Haug TM, et al. Identification and gene expression analysis of three GnRH genes in female Atlantic cod during puberty provides insight into GnRH variant gene loss in fish. *Gen Comp Endocrinol.* (2011) 172:458–67. doi: 10.1016/j.ygcen.2011.04.010
77. Burow S, Fontaine R, von Krogh K, Mayer I, Nourizadeh-Lillabadi R, Hollander-Cohen L, et al. Medaka follicle-stimulating hormone (Fsh) and luteinizing hormone (Lh): Developmental profiles of pituitary protein and gene expression levels. *Gen Comp Endocrinol.* (2019) 272:93–108. doi: 10.1016/j.ygcen.2018.12.006
78. Hollander-Cohen L, Golan M, Levavi-Sivan B. Differential regulation of gonadotropins as revealed by transcriptomes of distinct LH and FSH cells of fish pituitary. *Int J Mol Sci.* (2021) 22:6478. doi: 10.3390/ijms22126478
79. Waszkiewicz EM, Zmijewska A, Kozlowska W, Franczak A. Effects of LH and FSH on androgen and oestrogen release in the myometrium of pigs during the oestrous cycle and early pregnancy. *Reprod Fertil Dev.* (2020) 32:1200–11. doi: 10.1071/RD20148
80. Nelson LE, Sheridan MA. Regulation of somatostatins and their receptors in fish. *Gen Comp Endocrinol.* (2005) 142:117–33. doi: 10.1016/j.ygcen.2004.12.002
81. Gonzalez-Robayna IJ, Falender AE, Ochsner S, Firestone GL, Richards JS. Follicle- stimulating hormone (FSH) stimulates phosphorylation and activation of protein kinase B (PKB/Akt) and serum and glucocorticoid- induced kinase (Sgk): evidence for a kinase-independent signaling by FSH in granulosa cells. *Molec Endocrinol.* (2000) 14:1283–300. doi: 10.1210/mend.14.8.0500
82. Li M, Sun L, Wang D. Roles of estrogens in fish sexual plasticity and sex differentiation. *Gen Comp Endocrinol.* (2019) 277:9–16. doi: 10.1016/j.ygcen.2018.11.015

83. Ström JO, Theodorsson A, Ingberg E, Isaksson IM, Theodorsson E. Ovariectomy and 17β-estradiol replacement in rats and mice: a visual demonstration. *J Vis Exp.* (2012) 7:e4013. doi: 10.3791/4013
84. Almeida-Pereira G, Vilhena-Franco T, Coletti R, Cognuck SQ, Silva HVP, Elias LLK, et al. 17β-Estradiol attenuates p38MAPK activity but not PKCα induced by angiotensin II in the brain. *J Endocrinol.* (2019) 240:345–60. doi: 10.1530/JOE-18-0095

Conflict of Interest: CL was employed by the company Xinjiang Tianyun Organic Agriculture Co., Yili Group.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

METHODS
published: 18 February 2022
doi: 10.3389/fgene.2022.818344

Integration of Count Difference and Curve Similarity in Negative Regulatory Element Detection

Na He[1,2†], Wenjing Wang[3†], Chao Fang[4], Yongjian Tan[2], Li Li[5,6] and Chunhui Hou[2*]*

[1]Harbin Institute of Technology, Harbin, China, [2]Department of Biology, School of Life Sciences, Southern University of Science and Technology, Shenzhen, China, [3]School of Life Science and State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Hong Kong, Hong Kong SAR, China, [4]Cancer Centre, Faculty of Health Sciences, University of Macau, Macao, China, [5]Department of Bioinformatics, Huazhong Agricultural University, Wuhan, China, [6]Hubei Key Laboratory of Agricultural Bioinformatics, Huazhong Agricultural University, Wuhan, China

Negative regulatory elements (NREs) down-regulate gene expression by inhibiting the activities of promoters or enhancers. The repressing activity of NREs can be measured globally by massively parallel reporter assays (MPRAs). However, most existing algorithms are designed for the statistical detection of positively enriched signals in MPRA datasets. To identify reduced signals in MPRA experiments, we designed a NRE identification program, fast-NR, by integrating the count and graphic features of sequenced reads to detect NREs using datasets generated by experiments of self-transcribing active regulatory region sequencing (STARR-seq). Fast-NR identified hundreds of silencers in human K562 cells that can be validated by independent methods.

Keywords: count difference, curve similarity, silencer, silencer identification, negative regulatory element

OPEN ACCESS

Edited by:
Liang Guo,
Chinese Academy of Fishery Sciences (CAFS), China

Reviewed by:
Sebastiaan Meijsing,
Max Planck Society, Germany
Zhihua Zhang,
Beijing Institute of Genomics (CAS), China

***Correspondence:**
Chunhui Hou
houch@sustech.edu.cn
Na He
11849492@mail.sustech.edu.cn

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *19 November 2021*
Accepted: *20 January 2022*
Published: *18 February 2022*

Citation:
He N, Wang W, Fang C, Tan Y, Li L and Hou C (2022) Integration of Count Difference and Curve Similarity in Negative Regulatory Element Detection. Front. Genet. 13:818344. doi: 10.3389/fgene.2022.818344

INTRODUCTION

Eukaryotic gene expression is tightly controlled by various types of *cis*-regulatory elements (CREs) that are different in regulatory function, genetic, and epigenetic characteristics (Maston et al., 2006). Promoters and enhancers are positive CREs that initiate and enhance transcription, respectively. Enhancers act locally or over long genomic distances through chromatin looping to regulate their target genes (Shlyueva et al., 2014; Haberle and Stark, 2018; Schoenfelder and Fraser, 2019). In contrast, silencers are negative regulatory elements (NRE) that suppress gene expression through mechanisms that are not completely understood (Ogbourne and Antalis, 1998). Mutations in human CREs associate frequently with tumorigenesis, neurodegeneration, and metabolic diseases (Maston et al., 2006), highlighting the functional importance of transcription control in cells.

In eukaryotic genomes, silencers had not been as vigorously investigated as enhancers (Della Rosa and Spivakov, 2020). Most silencers in the database of silencerDB are predicted (Zeng et al., 2021). Potential silencers were also predicted in cell lines (Doni Jayavelu et al., 2020) by gkmSVM which utilizes sequence features of known silencers (Ghandi et al., 2016). Different from enhancers, silencer prediction is currently infeasible because that whether silencers carry ubiquitous epigenetic signatures is unknown. Genome-wide characterization of functional silencers is thus critical to unveil the genetic and epigenetic features of silencers. Genomic sequences of regulatory activity can be systematically assessed by STARR-seq, a widely used MPRA method initially designed for enhancer identification (Melnikov et al., 2012; Arnold et al., 2013; Crocker and Stern, 2013; Gisselbrecht et al., 2013; Mogno et al., 2013; Vanhille et al., 2015; Wang et al., 2018; Sun et al., 2019). Theoretically, STARR-seq measures silencer activity as well. Actually, Doni Jayavelu et al. had

successfully used STARR-seq to measure the transcription-repressing activity of silencers that were predicted by epigenetic features (Doni Jayavelu et al., 2020). Recently, several studies had reported catalogs of silencers that had been predicted or identified by different methods in different model systems at small scales (Petrykowska et al., 2008; Huang et al., 2019; Doni Jayavelu et al., 2020; Pang and Snyder, 2020).

For MPRAs, a statistical method specially designed for silencer identification is needed to facilitate the investigations into silencers' identity and their roles in transcription regulation. To design a silencer identification program, developers need to consider the functional differences between enhancers and silencers (Zhang et al., 2008; Heinz et al., 2010; Lee et al., 2020). Doni Jayavelu et al. measured silencer activities in selected accessible chromatin regions by comparing the sequenced reads of the reporter cDNA to the reads of the input insert DNA using a one-tail *t* test (Doni Jayavelu et al., 2020). While Pang et al. used a model-based method MAGeCK (Li et al., 2014) after counting reads with the method of HTSeq (Anders et al., 2015; Pang and Snyder, 2020). MAGeCK is similar to edgeR (Robinson et al., 2010) and DESeq (Anders and Huber, 2010) in their design strategies, but it is different from the other two methods in its intended usage. MAGeCK is originally used in CRISPR/Cas9 knockout screen assays. Different from these small-scale assays, genome-wide sequenced reads follow a negative binomial distribution. Potentially, methods designed for the detection of differentially methylated regions (DMRs) or differential chromatin modifications (Shen et al., 2013; Lienhard et al., 2014; Zhang et al., 2014; Lun and Smyth, 2016; Gaspar and Hart, 2017) can be used to identify silencers. However, the specificity, robustness, accuracy, and resolution of these programs have not been evaluated for silencer identification. CRADLE, a recently published method, is designed for enhancer identification (Kim et al., 2021). Theoretically, CRADLE can detect silencers as well. Nevertheless, a computational method designed specifically for the identification of silencers has not been reported.

In this research, we provide a program Fast-NR that is designed for the identification of silencers using STARR-seq-generated datasets by integrating the sequenced read count and signal shape features which are considered in the design of many ChIP-seq peak callers including Polyapeak, PICS, and CLC (Thomas et al., 2017; Hower et al., 2011; Cremona et al., 2019; Yan et al., 2020) (Zhang et al., 2011; Wu and Ji, 2014; Strino and Lappe, 2016). Fast-NR is available at https://github.com/Na-He/Fast-NR. We tested this program on simulated and STARR-seq datasets (Johnson et al., 2018; Doni Jayavelu et al., 2020), compared the performance of Fast-NR with several other programs, and show here that Fast-NR can detect NREs under different conditions.

METHODS

Algorithm

DNA fragments of NRE activity reduce their own expression levels in the STARR-seq reporter cDNA library. To identify NREs, we first calculate *p* values for each nucleotide covered by the reporter cDNA and the input insert DNA across the genome. If a *p* value is below an arbitrary threshold, the corresponding genomic region is considered as a potential NRE. We then plot the numbers of the reporter cDNA and input DNA reads as curves and measure the distances between them to determine whether they are similar by using several different methods. For NREs, the similarity scores are supposed to be low. By integrating count number difference and curve similarity features, we designed a computational method, Fast-NR (**Figure 1**), and tested its NRE detection power on simulated and real STARR-seq datasets, respectively. Basically, we first screened nucleotides which had the number of reporter cDNA reads smaller than the input insert reads by at least 12, corresponding to the *p* value threshold (10^{-5}) we set. We calculate *p* values using cumulative density function (CDF) of negative binomial distribution (NBD). Next, we use the single nucleotides that pass the initial screen as anchors and extend the genomic window to upstream and downstream to a total of 601 bp. We further examine the *p* values of each nucleotide in each 601bp window and keep only windows in which 3/4 of all nucleotides are with a *p* value below 10^{-5}. If two windows overlap, we keep the one in the shared region with smaller *p* values. Then, we compare the similarity between the curves of reporter cDNA and the input insert DNA reads, and discard any window with a curve similarity score higher than the arbitrary threshold. Finally, we correct *p* values for each window of identified NRE by Benjamini–Hochberg (BH) test and keep only these with a corrected *p* value smaller than 10^{-5}.

p Value Calculation

We calculate *p* values by cumulative density function (CDF) of the negative binomial distribution for the sequenced reporter cDNA reads. The probability mass function of the number of *k* times failure for a negative binomial distribution is

$$\boldsymbol{CDF}(\boldsymbol{m},\boldsymbol{n},\boldsymbol{p}) = \boldsymbol{P}(\boldsymbol{x}_{\boldsymbol{n}} \le \boldsymbol{m}) = \sum_{\boldsymbol{i}=0}^{\boldsymbol{m}} \left(\frac{\boldsymbol{i}+\boldsymbol{n}-\boldsymbol{1}}{\boldsymbol{n}-\boldsymbol{1}} \right) \boldsymbol{p}^{\boldsymbol{n}} (\boldsymbol{1}-\boldsymbol{p})^{\boldsymbol{i}},$$

where $CDF(m, n, p)$ returns the probability that is fewer than *m* times failure before the *n* th times success, with a single success probability *p*. Here, the *m* is treatCount which comes from a negative binomial distribution, *n* is treatTotal-treatCount, and *p* is (controlTotal-controlCount)/controlTotal. treatTotal and controlTotal are the total fragment numbers of the reporter cDNA and the input insert DNA in the sequenced libraries, respectively. treatCount and controlCount are the count numbers of reporter cDNA and input insert DNA covering each nucleotide, respectively.

Curve Similarity

We compare the shape of the curves of the reporter cDNA and the input insert DNA reads. Cosine, Pearson, Euclidean, and an in-house method gradient (linear slope correlation) are used to calculate the curve similarity in this research.

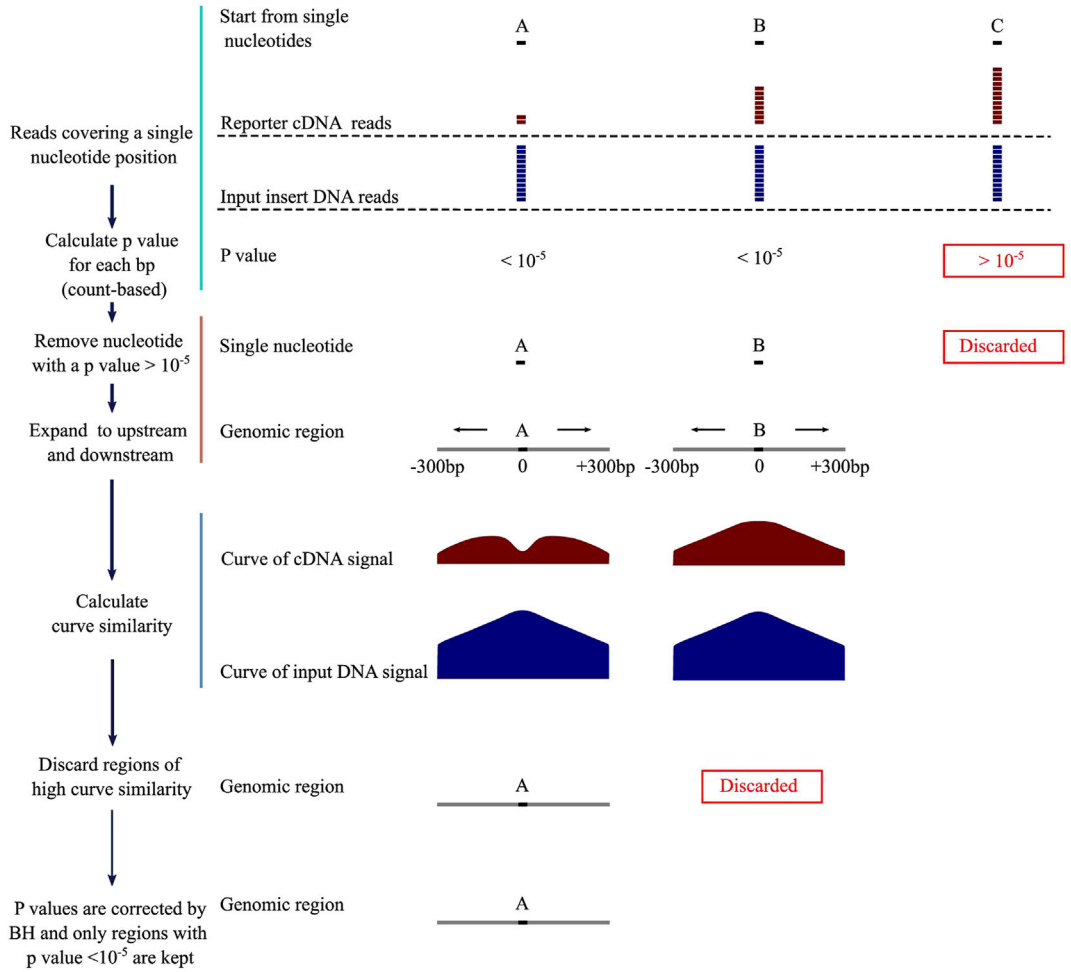

FIGURE 1 | Negative regulatory element identification pipeline. BH, Benjamini–Hochberg correction.

Cosine

The method of cosine computes the cosine distance between the 1-D arrays of u and v:

$$\mathbf{Cos}\ (u,\ v) = 1 - \frac{\sum_{i=0}^{n} u_i v_i}{\sqrt{\sum_{i=0}^{n} u_i^2}\sqrt{\sum_{i=0}^{n} v_i^2}},$$

where u_i and v_i are the reporter cDNA and the input insert DNA read count values in the u and v vectors.

Euclidean

Euclidean method computes the Euclidean distance between the 1-D arrays of u and v:

$$\mathbf{Euclidean\ (dis)} = \sqrt{\sum_{i=0}^{n} (u_i - v_i)^2},$$

where u_i and v_i are the reporter cDNA and the input insert DNA read count values in the u and v vectors.

Pearson

Pearson computes the Pearson correlation coefficient between the 1-D arrays of u and v:

$$cor = \frac{\sum_{i=0}^{n} (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=0}^{n} (u_i - \bar{u})^2}\sqrt{\sum_{i=0}^{n} (v_i - \bar{v})^2}},$$

where u_i and v_i are the reporter cDNA and the input insert DNA read count values in the u and v vectors.

Gradient

This algorithm computes the Pearson correlation coefficients between the gradients of the curves of the reporter cDNA and the input insert DNA reads. The coverage and genomic location values of each silencer form a 2-D array, represented by y and x, respectively. We calculate the gradient between two adjacent points in this array as in the following formula:

$$G(i) = \frac{(y_{i+1} - y_i)}{(x_{i+1} - x_i)},$$

where y_i is the coverage value and x_i is the position value of the i point in the 2D array. Gradient curve similarity index is the Pearson correlation coefficient (as mentioned above) between the cDNA and the input insert DNA $G(i)$ arrays.

p Value Correction

Bonferroni and Benjamini–Hochberg (BH) adjustments are applied to correct p values in our program.

Datasets

To test the performance of Fast-NR, we downloaded STARR-seq datasets for silencer identification in human K562 (GSE142207) (Doni Jayavelu et al., 2020) and for enhancer identification in untreated A549 cells (GSE114063) (Johnson et al., 2018), respectively. We downloaded histone modification datasets (H3K4me1 GSE91306; H3K27ac GSE91337; H3K4me3 GSE91218; H3K9me3 GSE91335) from ENCODE (Consortium, 2012). H3K27me3 (GSE75903) was downloaded from NCBI (Sayers et al., 2022). We mapped reads to human reference genome version hg19 (GRCh37) using bowtie2 (Langmead and Salzberg, 2012) with default parameters except "-p 24 -X 2000 --sensitive," then filtered and kept only reads with a value of MAPQ ≥20 by using samtools (Danecek et al., 2021). We kept only unique reads and discarded duplicates by using picardtools (Broad Institute, 2019) except for K562 STARR-seq datasets.

RESULTS

Performance of Fast-NR and Other Potential NRE Identification Methods Tested on Simulated Data

We simulated a pair of STARR-seq datasets to test the NRE identification powers of Fast-NR and other methods including csaw (Lun and Smyth, 2016), MEDIPS (Lienhard et al., 2014), PePr (Zhang et al., 2014), and CRADLE (Kim et al., 2021) (see **Supplementary Table S1**) using default parameters. We used the input insert reads, which are acquired by sequencing plasmids recovered from the transfected cells, and mapped them to chromosome 22 in the enhancer-screening STARR-seq experiment in A549 cells (GSE114063) as the simulation basis. Chromosome 22 is scanned, binned into 400bp windows and only genomic regions covered by at least 100 sequenced reads are kept. We selected 1,000 regions as simulated silencers (true positive silencers) by removing reads from these regions at four different percentage levels (30, 50, 70, and 90%), thus retained fraction of reads at 70, 50, 30, and 10% in the simulated datasets, respectively. These four datasets are used as the reporter cDNA libraries.

Program csaw detects differentially enriched regions in ChIP-seq dataset by using a sliding window strategy. MEDIPS identifies differentially methylated sites in the dataset generated by methylated DNA immunoprecipitation sequencing (MeDIP-seq). MEDIPS fails to detect any silencer at all from libraries in which 30, 50, and 70% reads are retained for the simulated silencers, while csaw identifies less than 100 silencers independent of what percentages of reads are retained (**Figure 2A**). These results suggest that MEDIPS and csaw may lack NRE detection power.

Program PePr is similar to csaw in their differential signal detection power for ChIP-seq datasets. PePr identifies most simulated silencers when reads are retained at three different levels (10, 30, and 50%) (**Figure 2A**), suggesting PePr could potentially be a usable NRE identification method. Program CRADLE calls both positive and negative regulatory elements for STARR-seq datasets. This program identifies approximately 800 silencers (815, 819, 821, and 812, respectively) independent of the percentages of reads retained (**Figure 2A**). Fast-NR detects 897, 871, 712, and 317 simulated silencers at 10, 30, 50, and 70% retained read levels (**Figure 2A**), suggesting its NRE identification power correlates positively with the read removal percentages. These results together suggest that PePr, CRADLE, and Fast-NR may all be usable NRE identification methods. Also, Fast-NR is more sensitive to signal reduction levels than other programs.

The NRE detection power of PePr, CRADLE, and Fast-NR may change when different p value thresholds are applied. Indeed, all these three methods detect fewer silencers as the p value threshold becomes more stringent (**Figure 2B**). Again, CRADLE is insensitive to the read removal percentage. Interestingly, though Fast-NR detects fewer silencers as p value decreases, it identifies more silencers than CRADLE when 10 and 30% of reads are retained. PePr is also sensitive to the change of p value threshold, especially when the fraction of reads retained is 70% (**Figure 2B**). These results show PePr and Fast-NR are more sensitive to the read retained rates than CRADLE. However, these results do not suggest which program outperforms the others.

Performance of Fast-NR and Other Potential NRE Identification Methods Tested on Real STARR-Seq Datasets

Theoretically, a genomic region of repressing activity is supposed to be transcribed less and underrepresented in the reporter cDNA library of STARR-seq. We downloaded STARR-seq datasets for silencer and enhancer identifications in human K562 (Doni Jayavelu et al., 2020) and A549 (Johnson et al., 2018) cells, respectively (**Supplementary Table S2**). STARR-seq in K562 measures the repressing activity of 7,430 sites in the accessible regions that are predicted as potential silencers based on epigenetic states. Differently, STARR-seq in A549 cells measures enhancer activities genome wide. We tested the NRE detection power of the five programs on the datasets generated by these two STARR-seq experiments (**Figure 3A**). Both Fast-NR and CRADLE identified hundreds and thousands NREs. Program csaw identified 2,399 silencers in K562 and only 31 silencers in A549. PePr identified 359 NREs in K562, but unbelievably, 475,797 NREs in A549. MEDIPS nearly failed to identify any NREs in the two STARR-seq experiments. These results confirm that MEDIPS lacks the NRE detection power for either simulated or real experimental data. Programs csaw and PePr perform differently on the two STARR-seq experiments, and their poor

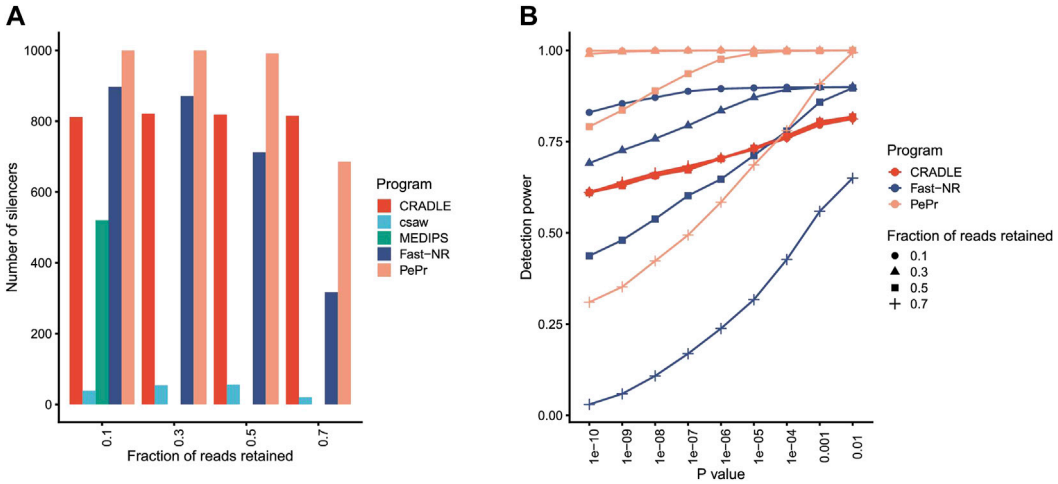

FIGURE 2 | Program performance comparison on simulated datasets. **(A)** The number of silencers identified by different programs. The fraction of reads retained to simulate silencers is shown under *x* axis. *p* value $<10^{-5}$. **(B)** The silencer detection power of different programs at different levels of confidence. Detection power is the ratio between number of identified silencers over total number of simulated silencers.

consistency in performance compromised our confidence to use them for NRE identification. After these comparisons, we kept Fast-NR and CRADLE for more evaluation analyses.

We compared Fast-NR and CRADLE's performance by changing the *p* value threshold for NRE identification. For K562 STARR-seq, Fast-NR identified silencers consistently at high levels and was nearly not affected by the change in the *p* value threshold (**Figure 3B**). In contrast, the number of silencers identified by CRADLE dropped dramatically to only about 10% ($p < 1 \times 10^{-8}$) of these identified at $p < 0.01$. These results provoked us to examine the overlapping rates of silencers identified by Fast-NR and CRADLE in K562. In fact, decent amounts of silencers identified by these two methods overlapped in K562 but not in A549 (**Figure 3C**). Silencers identified were expected to have lower cDNA reads than the input insert DNA reads. We calculated the ratios of (cDNA reads)/(insert reads) for CRADLE-specific and Fast-NR-specific silencers in K562 and A549. The CRADLE-specific silencers showed less reduction in cDNA reads than Fast-NR-specific silencers (**Figures 3D,E**). Over 95% (1,320/1,383) of Fast-NR identified silencers in K562 overlapped with the reported silencers (Doni Jayavelu et al., 2020) (**Figure 3F**). In contrast, only 56% silencers identified by CRADLE overlapped with the reported silencers (**Figure 3F**). These results suggest that many CRADLE-specific silencers were identified because of the heavy correction procedures that are integral to CRADLE (Kim et al., 2021). These CRADLE-specific silencers seemed to be "false-positives" in terms of the reduction rate in the reporter cDNA reads.

To reveal which transcription factors may bind to silencers, we searched through the sequences of silencers and identified a few DNA motifs enriched for transcription factors binding (**Supplementary Table S3**). One of these motifs was the silencing factor REST binding site (Chong et al., 1995), which was particularly enriched in Fast-NR identified silencers. DNA motif for transcription repressor PRDM6 was also enriched (Davis et al., 2006). Histone H4K20 methylation is a mark reported to be associated with silencers (Pang and Snyder, 2020). The binding motif (GC-box sequence) for the transcription factor of Sp1-like factors was also enriched in both Fast-NR and CRADLE silencers. Sp1-like factors activate or repress transcription in response to different physiological and pathological stimuli (Zhao and Meng, 2005). DNA motifs of PAX5 and FOS were enriched at Faste-NR and CATDLE silencers as well. Many transcription factors have dual functional roles in gene regulation, and silencers have been reported to be switchable to enhancers during development and in different cell types (Bessis et al., 1997; Cavalli, 2014; Gisselbrecht et al., 2020). Enrichment of any specific transcription factor's binding motif may not necessarily correlate with the regulatory activity of a CRE in a specific cell type. Nevertheless, silencers are indeed enriched with certain DNA motifs for transcription repressors in our analysis, suggesting that silencers identified by Fast-NR are very likely to be biologically functional.

Curve Similarity Analyses in Fast-NR

Compared to the other four methods, Fast-NR is the only one that takes into account the similarity between the curves of the reporter cDNA and the input insert DNA signals. We examined to what extent the similarity between the curves of the reporter cDNA and the insert DNA reads could affect the NRE identification. We calculated the similarity index values ($-\log_2$CosineDistance) for the NREs identified in A549 and found that the cosine distances between cDNA and plasmid curves are much higher than the random chosen genomic control regions (**Figure 4A**). Interestingly, the similarity index values correlate negatively with the strengths of silencers (**Figure 4B**), suggesting stronger silencers have low curve similarity. We obtained similar results using other curve similarity calculation methods such as Pearson, Euclidean, and gradient (**Supplementary Figure S1A**).

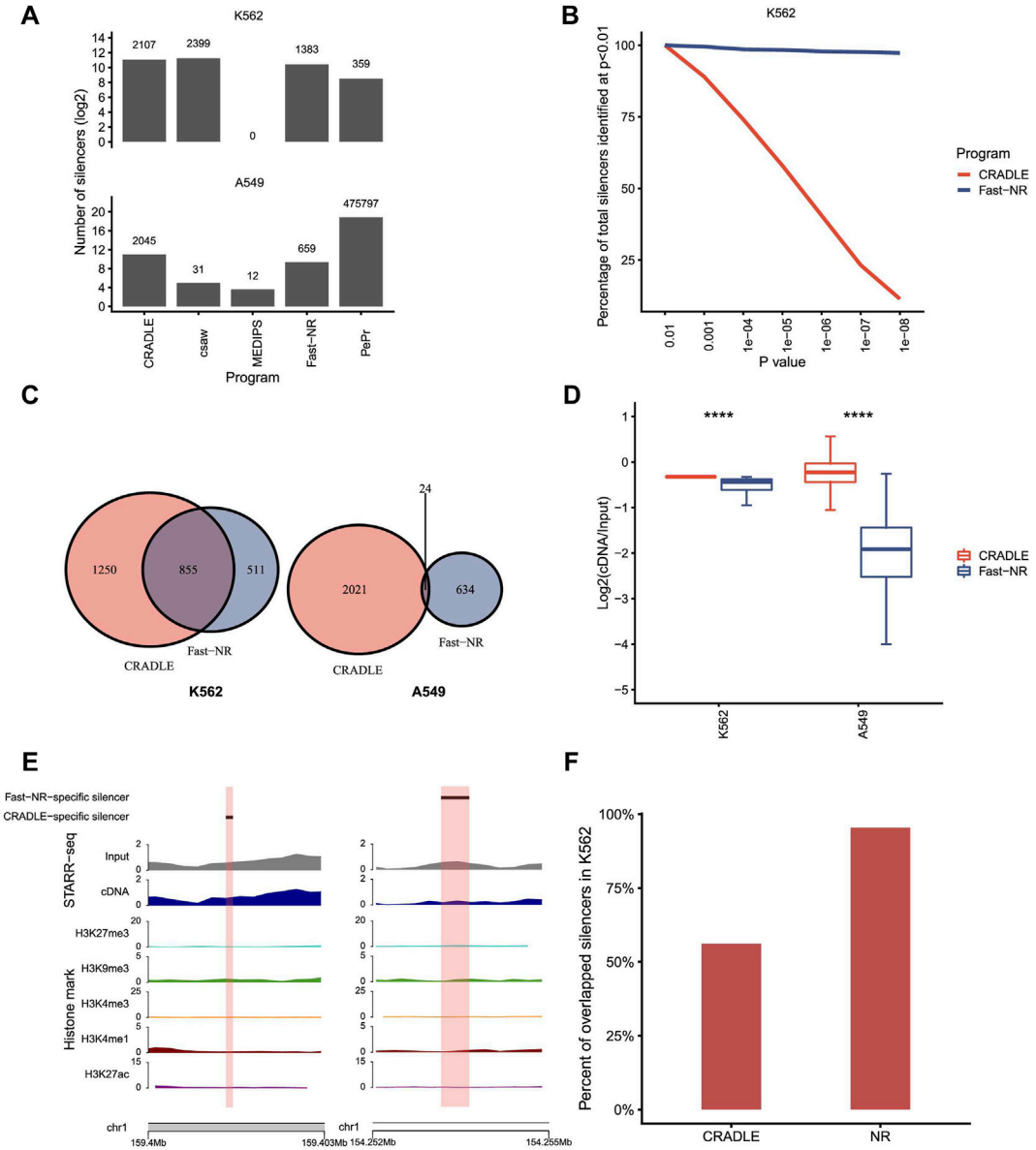

FIGURE 3 | Program performance comparison on STARR-seq datasets. **(A)** The number of silencers identified by different programs. p value $<10^{-5}$. **(B)** The percentage of silencers identified by CRADLE and Fast-NR at different confidence levels. The number of silencers identified at $p < 10^{-2}$ is set at 100%. **(C)** Venn diagrams of silencers identified by Fast-NR and CRADLE in K562 and A549, respectively. p value $<10^{-5}$. **(D)** Reads ratio (reporter cDNA/input inserts) distribution for silencers identified only by CRADLE or Fast-NR in K562 and A549, respectively. **(E)** Exemplary silencers identified only by CRADLE (left) or by Fast-NR (right). **(F)** Percentages of Fast-NR- and CRADLE-identified silencers ($p < 10^{-5}$) reported by Doni Jayavelu et al.

We removed the curve similarity requirement in Fast-NR and identified more silencers (gray dots in **Figure 4B**). These "new" silencers have high curve similarity index values and low silencer strengths compared to the silencers identified with curve similarity considered (blue dots in **Figure 4B**). Interestingly, curve similarity correlated poorly with p values (**Supplementary Figure S1B**), suggesting curve similarity in Fast-NR is a feature independent from the ratio between the reporter cDNA and the input insert DNA reads. The Pearson's correlation coefficients between the curve similarities and the silencer activities could be positive or negative depending on the method used (**Supplementary Figure S1C**). These results together show that curve similarity comparison is also important for the reliable identification of NREs.

DISCUSSION

In this study, we presented a program of Fast-NR, in which both read counts and shape similarity are considered, for the detection

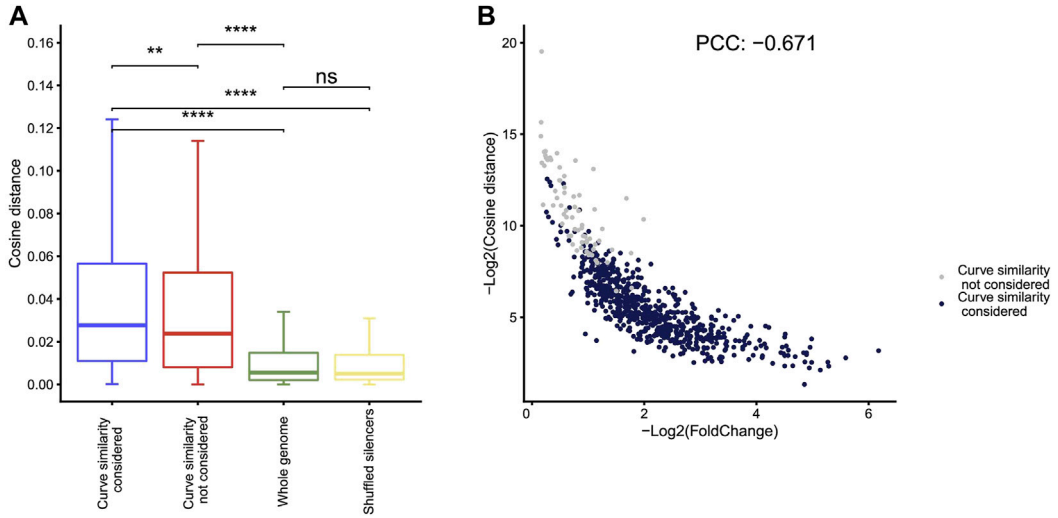

FIGURE 4 | Curve similarity effect on silencer identification. **(A)** The cosine distance distribution for silencers identified by Fast-NR with similarity considered, similarity not considered, and controls of whole genome regions with 400 bp size and shuffled silencer regions. Distance negatively correlates with similarity. $^{**}p < 10^{-3}$, $^{***}p < 10^{-4}$, Wilcoxon rank sum test. **(B)** The correlation between silencer strength and curve similarity. The *X* axis shows the value of $-\log_2$ (cDNA reads/insert DNA reads). The *Y* axis shows the curve similarity index, $-\log_2$ (Cosine distance) of silencers calculated by the method of cosine. Blue dots are silencers that pass curve similarity threshold of 0.9, and gray dots are silencers that do not pass curve similarity threshold.

of NREs using STARR-seq datasets and compare its performance with other four programs of csaw (Lun and Smyth, 2016), MEDIPS (Lienhard et al., 2014), PePr (Zhang et al., 2014), and CRADLE (Kim et al., 2021). Among them, MEDIPS, designed for DNA methylation analysis, shows worst compatibility with silencer identification on either simulated or experimentally generated datasets. Programs csaw and PePr detect significantly differential regions in ChIP-seq data. Neither of them performs consistently when being applied to different types of datasets. Besides methods tested in this research, other methods designed for the identification of differentially enriched signals are not suitable for silencer identification either. For example, DMRfinder (Gaspar and Hart, 2017), DSS (Park and Wu, 2016), and HMST-Seq-Analyzer (Farooq et al., 2020) require specific input data format that are not compatible with NRE analysis.

Fast-NR and CRADLE seem to be good choices for both simulated and experimentally generated datasets. However, many silencers identified by CRADLE showed only small reduction in the reporter cDNA signal than the input insert DNA, and curves of these signals were highly similar. CRADLE uses the GLM approach to correct four types of bias, the DNA structure affecting shear force, Gibbs free-energy affecting PCR efficiency, read sequences mappability, and G-quadruplex affecting DNA polymerase processivity (Kim et al., 2021). CRADLE treats the corrected signals as normal distribution and uses Welch's *t*-test to search for differences. As shown in our analysis, these corrections lead to the detection of "silencers" that cannot be identified based on the differences in the read counts of the reporter cDNA and the input insert DNA.

Being different from methods using sliding window strategy, Fast-NR detects the difference in the number of reporter cDNA and input insert DNA reads at single base-resolution. It is potentially possible to use Fast-NR to reveal the precise locations of regulatory elements and the binding sites of transcription factors.

STARR-seq tests silencers activity in episomal reporter plasmids independent from the endogenous chromatin environment. Ideally, regulatory activities of potential CREs can be tested in their proper chromatin context. Methods for endogenous CREs analysis, such as multiplexed editing regulatory assay (MERA) (Rajagopal et al., 2016) and thousands of reporters integrated in parallel (TRIP) (Akhtar et al., 2013), can be used to measure the regulatory activities of genomic regions in the native cellular context. However, these methods are generally not applicable for unbiased genome-wide analysis of CREs. Nevertheless, the combination of these methods and STARR-seq will help to achieve a global identification, and at the same time, a large scale endogenous validation of CREs.

Another issue we would like to point out is the promoter used in the reporter plasmids. In STARR-seq and related methods, promoter choice could affect the outcome because the promoter used may be irresponsive to some CREs. We speculate that using promoters of house-keeping genes and cell type-specific genes may allow the identification of more CREs that may prefer to regulate different types of promoters. To save the computation time, we filtered potential NREs by applying thresholds on both read counts and *p* values sequentially, which may also, theoretically, reduce false positive rate. However, the thresholds applied could be too strict and exclude some true silencers. We recommend the users to test the threshold effects and choose appropriate thresholds for their own analysis.

Though STARR-seq measures regulatory activity of tested DNA fragment in episomal environment, it provides a

catalogue of CREs that can be further tested at their endogenous loci by alternative methods. We did not take sequencing bias into consideration because experimental data that can be used to determine to what extent biases may affect NRE identification were not available. In summary, by combining read count-based negative binomial test and shape similarity comparison, we have shown that Fast-NR is potentially usable for silencer identification, thus providing a powerful and robust computational method for NRE identification.

CONCLUSION

Silencers are negative regulatory elements that control the precise gene expression during cell proliferation and differentiation. The increasing needs for global silencer characterization require a reliable and user-friendly computational method. Our method Fast-NR integrates single nucleotide read count information and graphic information to detect silencers genome widely. Fast-NR identifies NREs at single base resolution. The wide application of Fast-NR will accelerate the genetic and epigenetic studies of the intriguing functional mechanisms of silencers.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**; further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

NH conceived the study. NH and WW designed the program. NH analyzed the data. CF, YT, and LL participated in the program development. CH supervised the project. NH and CH wrote the article.

FUNDING

We acknowledge financial supports from the National Key R&D Program of China (no. 2018YFC1004500), Shenzhen Science and Technology Innovation Commission (no. 20200925153547003), and support from the Center for Computational Science and Engineering of Southern University of Science and Technology.

ACKNOWLEDGMENTS

We thank Dr. Longjian Niu for technical assistance and Dr. Kunfeng Bai for comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.818344/full#supplementary-material

REFERENCES

Akhtar, W., de Jong, J., Pindyurin, A. V., Pagie, L., Meuleman, W., de Ridder, J., et al. (2013). Chromatin Position Effects Assayed by Thousands of Reporters Integrated in Parallel. *Cell* 154, 914–927. doi:10.1016/j.cell.2013.07.018

Anders, S., and Huber, W. (2010). Differential Expression Analysis for Sequence Count Data. *Genome Biol.* 11, R106. doi:10.1186/gb-2010-11-10-r106

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq--a Python Framework to Work with High-Throughput Sequencing Data. *Bioinformatics* 31, 166–169. doi:10.1093/bioinformatics/btu638

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., and Stark, A. (2013). Genome-wide Quantitative Enhancer Activity Maps Identified by STARR-Seq. *Science* 339, 1074–1077. doi:10.1126/science.1232542

Bessis, A., Champtiaux, N., Chatelin, L., and Changeux, J.-P. (1997). The Neuron-Restrictive Silencer Element: a Dual Enhancer/silencer Crucial for Patterned Expression of a Nicotinic Receptor Gene in the Brain. *Proc. Natl. Acad. Sci.* 94, 5906–5911. doi:10.1073/pnas.94.11.5906

Broad Institute (2019). Available at: http://broadinstitute.github.io/picard/ (Accessed June 24, 2020).

Cavalli, G. (2014). A RING to Rule Them All: RING1 as Silencer and Activator. *Dev. Cel* 28, 1–2. doi:10.1016/j.devcel.2013.12.015

Chong, J. A., Tapia-Ramirez, J., Kim, S., Toledo-Aral, J. J., Zheng, Y., Boutros, M. C., et al. (1995). REST: A Mammalian Silencer Protein that Restricts Sodium Channel Gene Expression to Neurons. *Cell* 80, 949–957. doi:10.1016/0092-8674(95)90298-8

Consortium, E. P. (2012). An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* 489, 57–74. doi:10.1038/nature11247

Cremona, M. A., Xu, H., Makova, K. D., Reimherr, M., Chiaromonte, F., and Madrigal, P. (2019). Functional Data Analysis for Computational Biology. *Bioinformatics* 35, 3211–3213. doi:10.1093/bioinformatics/btz045

Crocker, J., and Stern, D. L. (2013). TALE-mediated Modulation of Transcriptional Enhancers *In Vivo*. *Nat. Methods* 10, 762–767. doi:10.1038/nmeth.2543

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve Years of SAMtools and BCFtools. *Gigascience* 10, giab008. doi:10.1093/gigascience/giab008

Davis, C. A., Haberland, M., Arnold, M. A., Sutherland, L. B., McDonald, O. G., Richardson, J. A., et al. (2006). PRISM/PRDM6, a Transcriptional Repressor that Promotes the Proliferative Gene Program in Smooth Muscle Cells. *Mol. Cel Biol* 26, 2626–2636. doi:10.1128/mcb.26.7.2626-2636.2006

Della Rosa, M., and Spivakov, M. (2020). Silencers in the Spotlight. *Nat. Genet.* 52, 244–245. doi:10.1038/s41588-020-0583-8

Doni Jayavelu, N., Jajodia, A., Mishra, A., and Hawkins, R. D. (2020). Candidate Silencer Elements for the Human and Mouse Genomes. *Nat. Commun.* 11, 1061. doi:10.1038/s41467-020-14853-5

Farooq, A., Grønmyr, S., Ali, O., Rognes, T., Scheffler, K., Bjørås, M., et al. (2020). HMST-Seq-Analyzer: A New python Tool for Differential Methylation and Hydroxymethylation Analysis in Various DNA Methylation Sequencing Data. *Comput. Struct. Biotechnol. J.* 18, 2877–2889. doi:10.1016/j.csbj.2020.09.038

Gaspar, J. M., and Hart, R. P. (2017). DMRfinder: Efficiently Identifying Differentially Methylated Regions from MethylC-Seq Data. *BMC Bioinformatics* 18, 528. doi:10.1186/s12859-017-1909-0

Ghandi, M., Mohammad-Noori, M., Ghareghani, N., Lee, D., Garraway, L., and Beer, M. A. (2016). gkmSVM: an R Package for Gapped-Kmer SVM. *Bioinformatics* 32, 2205–2207. doi:10.1093/bioinformatics/btw203

Gisselbrecht, S. S., Barrera, L. A., Porsch, M., Aboukhalil, A., Estep, P. W., 3rd, Vedenko, A., et al. (2013). Highly Parallel Assays of Tissue-specific Enhancers in Whole Drosophila Embryos. *Nat. Methods* 10, 774–780. doi:10.1038/nmeth.2558

Gisselbrecht, S. S., Palagi, A., Kurland, J. V., Rogers, J. M., Ozadam, H., Zhan, Y., et al. (2020). Transcriptional Silencers in Drosophila Serve a Dual Role as

Transcriptional Enhancers in Alternate Cellular Contexts. *Mol. Cel* 77, 324–337. doi:10.1016/j.molcel.2019.10.004

Haberle, V., and Stark, A. (2018). Eukaryotic Core Promoters and the Functional Basis of Transcription Initiation. *Nat. Rev. Mol. Cel Biol* 19, 621–637. doi:10.1038/s41580-018-0028-8

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cel* 38, 576–589. doi:10.1016/j.molcel.2010.05.004

Hower, V., Evans, S. N., and Pachter, L. (2011). Shape-based Peak Identification for ChIP-Seq. *BMC Bioinformatics* 12, 15. doi:10.1186/1471-2105-12-15

Huang, D., Petrykowska, H. M., Miller, B. F., Elnitski, L., and Ovcharenko, I. (2019). Identification of Human Silencers by Correlating Cross-Tissue Epigenetic Profiles and Gene Expression. *Genome Res.* 29, 657–667. doi:10.1101/gr.247007.118

Johnson, G. D., Barrera, A., McDowell, I. C., D'Ippolito, A. M., Majoros, W. H., Vockley, C. M., et al. (2018). Human Genome-wide Measurement of Drug-Responsive Regulatory Activity. *Nat. Commun.* 9, 5317. doi:10.1038/s41467-018-07607-x

Kim, Y.-S., Johnson, G. D., Seo, J., Barrera, A., Cowart, T. N., Majoros, W. H., et al. (2021). Correcting Signal Biases and Detecting Regulatory Elements in STARR-Seq Data. *Genome Res.* 31, 877–889. doi:10.1101/gr.269209.120

Langmead, B., and Salzberg, S. L. (2012). Fast Gapped-Read Alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi:10.1038/nmeth.1923

Lee, D., Shi, M., Moran, J., Wall, M., Zhang, J., Liu, J., et al. (2020). STARRPeaker: Uniform Processing and Accurate Identification of STARR-Seq Active Regions. *Genome Biol.* 21, 298. doi:10.1186/s13059-020-02194-x

Li, W., Xu, H., Xiao, T., Cong, L., Love, M. I., Zhang, F., et al. (2014). MAGeCK Enables Robust Identification of Essential Genes from Genome-Scale CRISPR/Cas9 Knockout Screens. *Genome Biol.* 15, 554. doi:10.1186/s13059-014-0554-4

Lienhard, M., Grimm, C., Morkel, M., Herwig, R., and Chavez, L. (2014). MEDIPS: Genome-wide Differential Coverage Analysis of Sequencing Data Derived from DNA Enrichment Experiments. *Bioinformatics* 30, 284–286. doi:10.1093/bioinformatics/btt650

Lun, A. T. L., and Smyth, G. K. (2016). Csaw: a Bioconductor Package for Differential Binding Analysis of ChIP-Seq Data Using Sliding Windows. *Nucleic Acids Res.* 44, e45. doi:10.1093/nar/gkv1191

Maston, G. A., Evans, S. K., and Green, M. R. (2006). Transcriptional Regulatory Elements in the Human Genome. *Annu. Rev. Genom. Hum. Genet.* 7, 29–59. doi:10.1146/annurev.genom.7.080505.115623

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., et al. (2012). Systematic Dissection and Optimization of Inducible Enhancers in Human Cells Using a Massively Parallel Reporter Assay. *Nat. Biotechnol.* 30, 271–277. doi:10.1038/nbt.2137

Mogno, I., Kwasnieski, J. C., and Cohen, B. A. (2013). Massively Parallel Synthetic Promoter Assays Reveal the *In Vivo* Effects of Binding Site Variants. *Genome Res.* 23, 1908–1915. doi:10.1101/gr.157891.113

Ogbourne, S., and Antalis, T. M. (1998). Transcriptional Control and the Role of Silencers in Transcriptional Regulation in Eukaryotes. *Biochem. J.* 331, 1–14.

Pang, B., and Snyder, M. P. (2020). Systematic Identification of Silencers in Human Cells. *Nat. Genet.* 52, 254–263. doi:10.1038/s41588-020-0578-5

Park, Y., and Wu, H. (2016). Differential Methylation Analysis for BS-Seq Data under General Experimental Design. *Bioinformatics* 32, 1446–1453. doi:10.1093/bioinformatics/btw026

Petrykowska, H. M., Vockley, C. M., and Elnitski, L. (2008). Detection and Characterization of Silencers and Enhancer-Blockers in the Greater CFTR Locus. *Genome Res.* 18, 1238–1246. doi:10.1101/gr.073817.107

Rajagopal, N., Srinivasan, S., Kooshesh, K., Guo, Y., Edwards, M. D., Banerjee, B., et al. (2016). High-throughput Mapping of Regulatory DNA. *Nat. Biotechnol.* 34, 167–174. doi:10.1038/nbt.3468

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. *Bioinformatics* 26, 139–140. doi:10.1093/bioinformatics/btp616

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., and Comeau, D. C. (2022). Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 50, D20–D26. doi:10.1093/nar/gkab1112

Schoenfelder, S., and Fraser, P. (2019). Long-range Enhancer-Promoter Contacts in Gene Expression Control. *Nat. Rev. Genet.* 20, 437–455. doi:10.1038/s41576-019-0128-0

Shen, L., Shao, N.-Y., Liu, X., Maze, I., Feng, J., and Nestler, E. J. (2013). diffReps: Detecting Differential Chromatin Modification Sites from ChIP-Seq Data with Biological Replicates. *PLoS One* 8, e65598. doi:10.1371/journal.pone.0065598

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional Enhancers: from Properties to Genome-wide Predictions. *Nat. Rev. Genet.* 15, 272–286. doi:10.1038/nrg3682

Strino, F., and Lappe, M. (2016). Identifying Peaks in *-seq Data Using Shape Information. *BMC Bioinformatics* 17 (Suppl. 5), 206. doi:10.1186/s12859-016-1042-5

Sun, J., He, N., Niu, L., Huang, Y., Shen, W., Zhang, Y., et al. (2019). Global Quantitative Mapping of Enhancers in Rice by STARR-Seq. *Genomics, Proteomics & Bioinformatics* 17, 140–153. doi:10.1016/j.gpb.2018.11.003

Thomas, R., Thomas, S., Holloway, A. K., and Pollard, K. S. (2017). Features that Define the Best ChIP-Seq Peak Calling Algorithms. *Brief Bioinform* 18, 441–450. doi:10.1093/bib/bbw035

Vanhille, L., Griffon, A., Maqbool, M. A., Zacarias-Cabeza, J., Dao, L. T. M., Fernandez, N., et al. (2015). High-throughput and Quantitative Assessment of Enhancer Activity in Mammals by CapStarr-Seq. *Nat. Commun.* 6, 6905. doi:10.1038/ncomms7905

Wang, X., He, L., Goggin, S. M., Saadat, A., Wang, L., Sinnott-Armstrong, N., et al. (2018). High-resolution Genome-wide Functional Dissection of Transcriptional Regulatory Regions and Nucleotides in Human. *Nat. Commun.* 9, 5380. doi:10.1038/s41467-018-07746-1

Wu, H., and Ji, H. (2014). PolyaPeak: Detecting Transcription Factor Binding Sites from ChIP-Seq Using Peak Shape Information. *PLoS One* 9, e89694. doi:10.1371/journal.pone.0089694

Yan, F., Powell, D. R., Curtis, D. J., and Wong, N. C. (2020). From Reads to Insight: a Hitchhiker's Guide to ATAC-Seq Data Analysis. *Genome Biol.* 21, 22. doi:10.1186/s13059-020-1929-3

Zeng, W., Chen, S., Cui, X., Chen, X., Gao, Z., and Jiang, R. (2021). SilencerDB: a Comprehensive Database of Silencers. *Nucleic Acids Res.* 49, D221–D228. doi:10.1093/nar/gkaa839

Zhang, X., Robertson, G., Krzywinski, M., Ning, K., Droit, A., Jones, S., et al. (2011). PICS: Probabilistic Inference for ChIP-Seq. *Biometrics* 67, 151–163. doi:10.1111/j.1541-0420.2010.01441.x

Zhang, Y., Lin, Y.-H., Johnson, T. D., Rozek, L. S., and Sartor, M. A. (2014). PePr: a Peak-Calling Prioritization Pipeline to Identify Consistent or Differential Peaks from Replicated ChIP-Seq Data. *Bioinformatics* 30, 2568–2575. doi:10.1093/bioinformatics/btu372

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* 9, R137. doi:10.1186/gb-2008-9-9-r137

Zhao, C., and Meng, A. (2005). Sp1-like Transcription Factors Are Regulators of Embryonic Development in Vertebrates. *Dev. Growth Differ.* 47, 201–211. doi:10.1111/j.1440-169x.2005.00797.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership