

1 **Chromosome-level genome assembly of the greenfin horse-faced filefish (*Thamnaconus***  
2 ***septentrionalis*) using Oxford Nanopore PromethION sequencing and Hi-C technology**

3 Li Bian<sup>1,2†</sup>, Fenghui Li<sup>1,3†</sup>, Jianlong Ge<sup>1,2†</sup>, Pengfei Wang<sup>4,5</sup>, Qing Chang<sup>1,2</sup>, Shengnong Zhang<sup>1,2</sup>, Jie Li<sup>1,2</sup>,  
4 Changlin Liu<sup>1,2</sup>, Kun Liu<sup>6</sup>, Xintian Liu<sup>7</sup>, Xuming Li<sup>8</sup>, Hongju Chen<sup>8</sup>, Siqing Chen<sup>1,2\*</sup>, Changwei Shao<sup>1,2\*</sup>,  
5 Zhishu Lin<sup>9\*</sup>

6 1. Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao, 266071, China

7 2. Laboratory for Marine Fisheries Science and Food Production Processes, Pilot National Laboratory for  
8 Marine Science and Technology (Qingdao), Qingdao, 266237, China

9 3. National Demonstration Center for Experimental Fisheries Science Education, Shanghai Collaborative  
10 Innovation for Aquatic Animal Genetics and Breeding, Shanghai Engineering Research Center of Agriculture,  
11 Shanghai Ocean University, Shanghai, 201306, China

12 4. Key Laboratory of Open-Sea Fishery Development, Ministry of Agriculture, Guangzhou, 510300, China

13 5. Guangdong Provincial Key Laboratory of Fishery Ecology and Environment, South China Sea Fisheries  
14 Research Institute, Chinese Academy of Fisheries Sciences, Guangzhou, 510300, China

15 6. Qingdao Conson Oceantec Valley Development Co., Ltd, Qingdao, 266200, China

16 7. Weihai Fishery Technology Extension Station, Weihai, 264200, China

17 8. Biomarker Technologies Corporation, Beijing, 101300, China

18 9. Qingdao Municipal Ocean Technology Achievement Promotion Center, Qingdao, 266071, China

19 **Abstract**

20 The greenfin horse-faced filefish, *Thamnaconus septentrionalis*, is a valuable commercial  
21 fish species that is widely distributed in the Indo-West Pacific Ocean. It has characteristic blue-

22 green fins, rough skin and spine-like first dorsal fin. *T. septentrionalis* is of a conservation  
23 concern as a result of sharply population decline, and it is an important marine aquaculture fish  
24 species in China. The genomic resources of this filefish are lacking and no reference genome  
25 has been released. In this study, the first chromosome-level genome of *T. septentrionalis* was  
26 constructed using Nanopore sequencing and Hi-C technology. A total of 50.95 Gb polished  
27 Nanopore sequence were generated and were assembled to 474.31 Mb genome, accounting for  
28 96.45% of the estimated genome size of this filefish. The assembled genome contained only  
29 242 contigs, and the achieved contig N50 was 22.46 Mb, reaching a surprising high level among  
30 all the sequenced fish species. Hi-C scaffolding of the genome resulted in 20 pseudo-  
31 chromosomes containing 99.44% of the total assembled sequences. The genome contained  
32 67.35 Mb repeat sequences, accounting for 14.2% of the assembly. A total of 22,067 protein-  
33 coding genes were predicted, of which 94.82% were successfully annotated with putative  
34 functions. Furthermore, a phylogenetic tree was constructed using 1,872 single-copy gene  
35 families and 67 unique gene families were identified in the filefish genome. This high quality  
36 assembled genome will be a valuable genomic resource for understanding the biological  
37 characteristics and for facilitating breeding of *T. septentrionalis*.

### 38 **Key words**

39 Filefish, genome assembly, Oxford Nanopore sequencing, Hi-C

### 40 **1 Introduction**

41 The greenfin horse-faced filefish (*Thamnaconus septentrionalis*; hereafter “filefish”)   
42 belongs to the family Monacanthidae (Tetraodontiformes) and has characteristic blue-green fins,

43 rough skin and spine-like first dorsal fin (Figure 1)(Su & Li, 2002). It is widely distributed in  
44 the Indo-West Pacific Ocean, ranging from the Korean Peninsula, Japan and China Sea to East  
45 Africa. Filefish is a temperate demersal species inhabiting a depth range of 50-120 m, and  
46 feeding on planktons such as copepods, ostracods, and amphipods, as well as mollusks and  
47 benthic organisms(Su & Li, 2002). It goes through annual long-distance seasonal migrations  
48 and has diurnal vertical migration habits during wintering and spawning(Lin, Gan, Zheng, &  
49 Guan, 1984; Su & Li, 2002). Due to a high protein content and good taste, filefish is an  
50 important commercial species in China, Korea and Japan. An interesting feature of filefish is  
51 its rough skin, whose roughness is actually attributed to the covered dense small scales. These  
52 scales are difficult to remove, and people have to peel off the skin before eating. Given this,  
53 filefish is also called “skinned fish” in China.

54 The wild resource of filefish has declined dramatically since 1990 due to overfishing, and  
55 the annual catch in the East China Sea was only 3,842 tons in 1994(Chen, Li, & Hu, 2000).  
56 Since then, researchers have attempted to explore the methods to properly culture filefish.  
57 Several key technologies including fertilized eggs collection, sperm cryopreservation, larval  
58 rearing, tank and cage culturing have been studied, and this species is cultivated commercially  
59 in China, Korea and Japan(Guan et al., 2013; Kang et al., 2004; Li, Jiang, Xu, & Liu, 2002; Liu  
60 et al., 2017; Mizuno, Shimizu-Yamaguchi, Miura, & Miura, 2012). The current main challenge  
61 of filefish cultivation is the high mortality of fish fry during artificial breeding. A better  
62 understanding of the underlying genomic-level characteristics will provide significant  
63 information to break through the bottleneck and benefit the cultivation industry of this filefish.

64 However, the available genetic information of filefish is scarce. At present, only limited genetic  
65 studies regarding microsatellite loci isolation and population structure are available for this  
66 filefish (An et al., 2011; An, Lee, Park, & Jung, 2013; Bian et al., 2018; Xu, Chen, & Tian,  
67 2010; Xu, Tian, Liao, & Chen, 2009).

68 Spectacular improvements in high-throughput sequencing technology, especially the  
69 single-molecule sequencing methods, have remarkably reduced the sequencing costs, making  
70 a genome project affordable for individual labs. Oxford Nanopore sequencing technology is  
71 currently the most powerful method for rapid generation of long-read sequences and has the  
72 potential to offer relatively low-cost genome sequencing of non-model animals. It directly  
73 detects the input DNA without PCR amplification or synthesis, so the length of sequenced DNA  
74 can be very long. The longest read generated by Nanopore sequencing has been up to 2,272,580  
75 bases (Payne, Holmes, Rakyan, & Loose, 2018). Nanopore sequencing has been used in several  
76 fish species to construct high-quality genome assembly or to improve the completeness of  
77 previous genome drafts (Austin et al., 2017; Ge et al., 2019; Jansen et al., 2017; Kadobianskyi,  
78 Schulze, Schuelke, & Judkewitz, 2019; Tan et al., 2018). In the case of red spotted grouper  
79 (*Epinephelus akaara*), a chromosome-level reference genome with a contig N50 length of 5.25  
80 Mb was constructed by taking advantage of Nanopore sequencing and Hi-C technology (Ge et  
81 al., 2019). In clown anemonefish (*Amphiprion ocellaris*), a hybrid Illumina/Nanopore method  
82 generated much longer scaffolds than Illumina-only approach with an 18-fold increase in N50  
83 length and increased the genome completeness by an additional 16% (Tan et al., 2018).

84 In this study, the first chromosome-level genome of filefish was constructed using

85 Nanopore sequencing and Hi-C technology. This genomic data will benefit a comprehensive  
86 conservation study of filefish along the China and Korea coast to implement better protection  
87 of wild populations, and allow us to screen for genetic variations correlated with fast-growth  
88 and disease-resistance traits of filefish in the future.

## 89 **2 Materials and methods**

### 90 **2.1 Sample and DNA extraction**

91 A single female fish (~325 g) was collected on August 2018 from the Tianyuan Fisheries  
92 Co., Ltd (Yantai, China). The muscle tissue below the dorsal fin was taken and stored in the  
93 liquid nitrogen until DNA extraction. Genomic DNA was extracted using CTAB  
94 (Cetyltrimethylammonium bromide) method. The quality and concentration of the extracted  
95 genomic DNA was checked using 1% agarose gel electrophoresis and a Qubit fluorimeter  
96 (Invitrogen, Carlsbad, CA, USA). This high-quality DNA was used for subsequent Nanopore  
97 and Illumina sequencing.

### 98 **2.2 Library construction and genome sequencing**

99 To generate Oxford Nanopore long reads, approximately 15 µg of genomic DNA was size-  
100 selected (30–80 kb) with a BluePippin (Sage Science, Beverly, MA, USA), and processed  
101 according to the Ligation Sequencing Kit 1D (SQK-LSK109) protocol. Briefly, DNA fragments  
102 were repaired using the NEBNext FFPE Repair Mix (New England Biolabs). After end-  
103 reparation and 3'-adenylation with the NEBNext End repair/dA-tailing Module reagents (New  
104 England Biolabs), the Oxford Nanopore sequencing adapters were ligated using NEBNext  
105 Quick Ligation Module (E6056) (New England Biolabs). The final library was sequenced on 3

106 different R9.4 flow cells using the PromethION DNA sequencer (Oxford Nanopore, Oxford,  
107 UK) for 48 hours. The MinKNOW software (version 2.0) was used to conduct base calling of  
108 raw signal data and convert the fast5 files into fastq files. These raw data was then filtered to  
109 remove short reads (<5 kb) and the reads with low-quality bases and adapter sequences.

110 Illumina sequencing libraries were prepared to carry out genome size estimation,  
111 correction of genome assembly, and assembly evaluation. The paired-end (PE) libraries with  
112 insert sizes of 300 bp were constructed according to the Illumina standard protocol (San Diego,  
113 CA, USA) and subjected to PE (2 × 150 bp) sequencing on an Illumina HiSeq X Ten platform  
114 (Illumina, San Diego, CA, USA). After discarding the reads with low-quality bases, adapter  
115 sequences, and duplicated sequences, the clean reads were used for subsequent analysis.

### 116 **2.3 Genome size estimation and genome assembly**

117 A k-mer depth frequency distribution analysis of the Illumina data was conducted to  
118 estimate the genome size, heterozygosity, and content of repetitive sequences of the filefish.  
119 The k-mer analysis was carried out using “kmer freq stat” software (developed by Biomarker  
120 Technologies Corporation, Beijing, China). Genome size (G) was estimated based on the  
121 following formula:  $G = \text{k-mer number} / \text{average k-mer depth}$ , where k-mer number = total k-  
122 mers—abnormal k-mers (with too low or too high frequency).

123 For genome assembly, Canu (version 1.5) (Koren et al., 2017) was conducted for initial  
124 read correction, and the assembly was performed by Wtdbg (<https://github.com/ruanjue/wtdbg>).  
125 The consensus assembly was generated by 2 rounds of Racon (version 1.32) (Vaser, Sović,  
126 Nagarajan, & Šikić, 2017), and 3 rounds of Pilon (version 1.21) (Walker et al., 2014) polishing

127 using the Illumina reads with default settings.

## 128 **2.4 Hi-C library construction and sequencing**

129 For Hi-C sequencing, the muscle tissue of filefish was used for library preparation  
130 according to Rao et al(Rao et al., 2014). Briefly, the tissue cells were fixed with formaldehyde  
131 and restriction endonuclease Hind III was used to digest DNA. The 5' overhang of the fragments  
132 were repaired and labeled using biotinylated nucleotides, followed by ligation in a small volume.  
133 After reversal of crosslinks, ligated DNA was purified and sheared to a length of 300-700 bp.  
134 The DNA fragments with interaction relationship were captured with streptavidin beads and  
135 prepared for Illumina sequencing. The final Hi-C libraries were sequenced on an Illumina  
136 HiSeq X Ten platform (Illumina, San Diego, CA, USA) to obtain  $2 \times 150$  bp paired-end reads.  
137 To assess the quality of Hi-C data, the plot of insert fragments length frequency was first made  
138 to detect the quality of Illumina sequencing. Second, we used BWA-MEM (version 0.7.10-r789)  
139 (Li & Durbin, 2009)to align the PE clean reads to the draft genome assembly. In the end, HiC-  
140 Pro (Servant et al., 2015) (version 2.10.0) was performed to find the valid reads from unique  
141 mapped read pairs.

## 142 **2.5 Chromosomal-level genome assembly using Hi-C data**

143 We first performed a preassembly for error correction of contigs by breaking the contigs  
144 into segments of 500 kb on average and mapping the Hi-C data to these segments using BWA-  
145 MEM (version 0.7.10-r789)(Li & Durbin, 2009). The corrected contigs and valid reads of Hi-  
146 C were used to perform chromosomal-level genome assembly using LACHESIS(Burton et al.,  
147 2013) with the following parameters: CLUSTER\_MIN\_RE\_SITES=22;

148 CLUSTER\_MAX\_LINK\_DENSITY=2; CLUSTER\_NONINFORMATIVE\_RATIO=2;  
149 ORDER\_MIN\_N\_RES\_IN\_TRUNK=10; ORDER\_MIN\_N\_RES\_IN\_SHREDS=10. To  
150 evaluate the quality of the chromosomal-level genome assembly, a genome-wide Hi-C heatmap  
151 was generated by ggplot2 in R package.

## 152 **2.6 Assessment of the genome assemblies**

153 To assess the genome assembly completeness and accuracy, we first aligned the Illumina  
154 reads to the filefish assembly using BWA-MEM (version 0.7.10-r789)(Li & Durbin, 2009).  
155 Furthermore, CEGMA (version 2.5) (Parra, Bradnam, & Korf, 2007) was conducted to find core  
156 eukaryotic genes (CEGs) in the genome with parameter set as identity>70%. Finally, the  
157 completeness of the genome assembly was also evaluated by using BUSCO (version  
158 2.0)(Simao, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) search the genome against  
159 the actinopterygii database, which consisted of 4584 orthologs.

## 160 **2.7 Repeat annotation, gene prediction and gene annotation**

161 We first used MITE-Hunter(Han & Wessler, 2010), LTR-FINDER (version 1.05)(Xu &  
162 Wang, 2007), RepeatScout (version 1.0.5)(Price, Jones, & Pevzner, 2005) and PILER(Edgar &  
163 Myers, 2005) to construct a *de novo* repeat library for filefish with default settings. These  
164 predicted repeats were classified using PASTEClassifier (version 1.0)(Hoede et al., 2014) , and  
165 then integrated with Repbase (19.06)(Bao, Kojima, & Kohany, 2015) to build a new repeat  
166 library for final repeat annotation. In the end, RepeatMasker (version 4.0.6)(Tarailo-Graovac &  
167 Chen, 2009) was performed to detect repetitive sequences in the filefish genome with the  
168 following parameters: “-nolow -no\_is -norna -engine wublast” .



169 *Ab initio*-based, homolog-based, and RNA-sequencing (RNA-seq)-based methods were  
170 conducted in combination to detect the protein-coding genes in filefish genome assembly.  
171 Genscan(Burge & Karlin, 1997), Augustus (version 2.4)(Stanke & Waack, 2003),  
172 GlimmerHMM (version 3.0.4)(Majoros, Pertea, & Salzberg, 2004), GeneID (version  
173 1.4)(Blanco, Parra, & Guigó, 2007), and SNAP (version 2006-07-28)(Korf, 2004) were used  
174 for *ab initio*-based gene prediction in filefish genome assembly. For the homolog-based  
175 method, tiger pufferfish (*Takifugu rubripes*), spotted green pufferfish (*Tetraodon nigroviridis*)  
176 and zebrafish (*Danio rerio*) were chosen to conduct gene annotation using GeMoMa (version  
177 1.3.1)(Keilwagen et al., 2016). For the RNA-seq-based method, a mixture of 10 tissues  
178 (including brain, eye, gill, heart, liver, intestine, spleen, ovary, kidney and muscle) of a  
179 female and the testis of a male filefish was used to construct Illumina sequencing library and  
180 subjected to PE (2 × 150 bp) sequencing on an Illumina HiSeq X Ten platform (Illumina, San  
181 Diego, CA, USA). After discarding the reads with low-quality bases, adapter sequences, and  
182 duplicated sequences, the retained high-quality clean reads were first assembled by Hisat  
183 (version 2.0.4)(Kim, Langmead, & Salzberg, 2015) and Stringtie (version 1.2.3)(Pertea et al.,  
184 2015), and then the gene prediction was performed using TransDecoder  
185 (<http://transdecoder.github.io>) (version 2.0), GeneMarkS-T (version 5.1)(Tang, Lomsadze, &  
186 Borodovsky, 2015), and PASA (version 2.0.2)(Haas et al., 2003). EVM (version 1.1.1)(Haas  
187 et al., 2008) was performed to integrate the prediction results obtained from three methods.  
188 We then added the genes that were supported by homolog and RNA-seq analysis after-manual  
189 evaluation.

190 To functionally annotate the predicted genes, they were aligned to the Non-redundant  
191 protein sequences (NR), eukaryotic orthologous groups of proteins (KOG)(Tatusov et al.,  
192 2003), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto, 2000)and  
193 TrEMBL(Boeckmann et al., 2003) databases using BLAST (version 2.2.31)(Altschul, Gish,  
194 Miller, Myers, & Lipman, 1990) with an e-value cutoff of 1E-5. Gene ontology (GO)  
195 (Consortium, 2004)annotation was performed with Blast2GO (version 4.1)(Conesa et al.,  
196 2005). For non-coding RNA prediction, we first used tRNAscan-SE (version 1.3.1)(Lowe &  
197 Eddy, 1997) to annotate transfer RNAs (tRNAs). Furthermore, Infernal (version 1.1)(Nawrocki  
198 & Eddy, 2013) was conducted to search for ribosomal RNAs (rRNAs) and microRNAs based  
199 on Rfam (version 13.0)(Daub, Eberhardt, Tate, & Burge, 2015) and miRbase (version  
200 21.0)(Griffiths-Jones, Grocock, Van Dongen, Bateman, & Enright, 2006) database.

## 201 **2.8 Comparative genomics**

202 To resolve the phylogenetic position of the filefish, we first used OrthoMCL (version  
203 2.0.9) (Li, Stoeckert, & Roos, 2003) to detect orthologue groups by retrieving the protein data  
204 of eleven teleost species including tiger pufferfish (*Takifugu rubripes*), yellowbelly pufferfish  
205 (*Takifugu flavidus*), spotted green pufferfish (*Tetraodon nigroviridis*), red seabream (*Pagrus*  
206 *major*), medaka (*Oryzias latipes*), large yellow croaker (*Larimichthys crocea*), three-spined  
207 stickleback (*Gasterosteus aculeatus*), nile tilapia (*Oreochromis niloticus*), japanese seabass  
208 (*Lateolabrax maculatus*), spotted gar (*Lepisosteus oculatus*) and zebrafish (*Danio rerio*) . The  
209 single copy orthologous genes shared by all 12 species were further aligned using MUSCLE  
210 (version 3.8.31)(Edgar, 2004) and concatenated to construct a phylogenetic tree with

211 PhyML(Guindon et al., 2010). The divergence time among species was estimated by the  
212 MCMCTree program of the PAML package(Yang, 2007) and CAFÉ(version 4.0) (De Bie,  
213 Cristianini, Demuth, & Hahn, 2006) was used to identified expanded and contracted gene  
214 families.

### 215 **3.Results and discussion**

#### 216 **3.1 Initial characterization of the filefish genome**

217 The k-mer ( $k = 19$  in this case) depth frequency distribution analysis of the 45.97 Gb  
218 clean Illumina data was conducted to estimate the genome size, heterozygosity, and repeat  
219 content of filefish (Table 1). The k-mer depth of 76 was found to be the highest peak in the  
220 plot, and a k-mer number of 37,677,330,713 was used to calculate the genome size of filefish  
221 (Figure S1). The sequences around k-mer depth of 38 were heterozygous sequences, and k-  
222 mer depth more than 153 represented repetitive sequences. The filefish genome size was  
223 estimated to be 491.74 Mb, the heterozygosity was approximately 0.35%, and the content of  
224 repetitive sequences and guanine-cytosine were about 16.62% and 46.05%, respectively.

#### 225 **3.2 Genome assembly**

226 A total of 50.95 Gb high quality clean reads, representing a 104-fold coverage of the  
227 genome, were generated from PromethION DNA sequencer (Table 1, Table S1-2,). These data  
228 was assembled using Wtdbg, followed by Racon and Pilon polishing, which produced a  
229 465.93 Mb genome assembly with a surprising long contig N50 of 22.07 Mb (Table S3). The  
230 length of this assembly was close to the genome size estimated by k-mer analysis (491.74  
231 Mb), indicating an appropriate assembly size was obtained from the Nanopore data. Among

232 the sequenced tetraodontiform species, the genome size of filefish was larger than *Takifugu*  
233 and *Tetraodon* species, but smaller than *Mola mola* (Aparicio et al., 2002; Gao et al., 2014;  
234 Jaillon et al., 2004; Pan et al., 2016) (Table 2).

235 For Hi-C data, overall 39.44 Gb clean reads were obtained and used for subsequent  
236 analysis (Table 1). To assess the quality of Hi-C data, we first made a plot of insert fragments  
237 length frequency, which showed a relatively narrow unimodal length distribution with the  
238 highest peak around 350 bp (Figure S2), indicating efficient purification of streptavidin beads  
239 during library construction. The alignment results revealed that about 89.78% of the Hi-C  
240 read pairs were mapped on the genome, and 78.18% of the read pairs were unique detected on  
241 the assembly (Table S4). Lastly, a total of 47,111,219 valid reads, which accounted for  
242 66.95% of the unique mapped reads, were detected by HiC-Pro in the Hi-C dataset (Table S5).  
243 Taken together, our evaluation suggested an overall high quality of the Hi-C data, and only  
244 the valid read pairs were used for subsequent analysis.

245 Before chromosomal-level genome assembly, an error correction of the initial assembly  
246 was performed by BWA-MEM with Hi-C data. The corrected filefish genome assembly was  
247 approximately 474.30 Mb with only 242 contigs, the contig N50 reached up to 22.46 Mb, and  
248 the longest contig was 32.32 Mb (Table 2, Table S6). The results indicated that high-coverage  
249 Nanopore long read-only assembly, followed by multiple iterations of genome polishing using  
250 Illumina reads is an effective method to generate high-quality genome assemblies.

251 A chromosomal-level genome was then assembled using LACHESIS, the results showed  
252 that overall 147 contigs spanning 471.65 Mb (99.44% of the assembly) were scaffolded into

253 20 pseudo-chromosomes, and 107 contigs spanning 469.46 Mb (98.98% of the assembly)  
254 were successfully ordered and oriented (Table 3). Several of the pseudo-chromosomes were  
255 scaffolded with only 2 or 3 contigs, representing a high contiguity of the genome. The final  
256 assembled genome was 474.31 Mb with a scaffold N50 length of 23.05 Mb and a longest  
257 scaffold of 34.81 Mb (Table 2, Table S6). As far as we know, this assembled genome was one  
258 of the most contiguous fish genome assembly with the highest contig N50 when compared  
259 with other published fish genomes.

260 To further evaluate the quality of the chromosomal-level genome assembly, a genome-  
261 wide Hi-C heatmap was generated. The 20 pseudo-chromosomes could be easily  
262 distinguished and the interaction signal strength around the diagonal was much stronger than  
263 that of other positions within each pseudo-chromosome, which indicated a high quality of this  
264 genome assembly (Figure 2).

### 265 **3.3 Completeness of the assembled genome**

266 Illumina reads were aligned to the filefish assembly, and 97.41% of the clean reads can  
267 be mapped to the contigs (Table S7). Then the CEGMA analysis identified 442 CEGs,  
268 accounting for 96.51% of all 458 CEGs in the program, and 226 CEGs could be detected by  
269 using a highly conserved 248 CEGs dataset (Table S8). Lastly, approximately 94.33%  
270 (4324/4584) of complete BUSCOs were found in the assembly (Table S9). Overall, the  
271 assessment results indicated our filefish genome assembly was complete and of high quality.

### 272 **3.4 Repeat annotation, gene prediction and gene annotation**

273 A total of 67.35 Mb of repeat sequences that accounted for 14.2% of the assembly were

274 found in filefish (Table S10). This repeat content was close to the value (16.62%) obtained  
275 from k-mer analysis. The predominant repeats type were TIRs (4.35%), LINEs (2.40%) and  
276 LARDs (1.65%).

277 The combination of *Ab initio*-based, homolog-based, and RNA-seq-based methods  
278 predicted overall 22,067 protein-coding genes with an average gene length, average exon  
279 length, and average intron length of 11,291bp, 230 bp, and 905 bp, respectively (Table 1,  
280 Table 4). A total of 20,924 genes, which counted for 94.82% of the predicted genes, were  
281 successfully annotated with putative functions (Table 5). The non-coding RNA prediction  
282 identified 1,703 tRNAs, 649 rRNAs and 109 microRNAs, respectively (Table S11).

### 283 **3.5 Comparative genomics**

284 Comparison of the filefish genome assembly with other eleven teleost species genomes  
285 found a total of 22,665 gene families, of which 5,692 were shared among all eleven species,  
286 including 1,872 single-copy orthologous genes (Table S12). Overall 20,261 genes of filefish  
287 can be clustered into 15,433 gene families, including 67 unique gene families containing 193  
288 genes (Table S12). The phylogenetic tree showed that four tetraodontiform species were  
289 clustered together, and the divergence time between filefish and the other three species was  
290 around 124.4 million years ago (Mya) (Figure 3). We also found 59 expanded gene families  
291 and 98 contracted gene families in filefish compared with the other fish species (Figure S3). A  
292 Venn diagram of orthologous gene families among four tetraodontiform species was also  
293 constructed, and 971 unique gene families containing 6485 genes were identified in the  
294 filefish genome (Figure 4).

#### 295 **4. Conclusion**

296 In the present study, we assembled the chromosome-level genome of *T. septentrionalis*, a  
297 first reference genome of the genus *Thamnaconus*. The assembled genome was 474.31 Mb,  
298 which is larger than the sequenced *Takifugu* and *Tetraodon* species, but smaller than *Mola*  
299 *mola*. With the powerful sequencing ability of Oxford Nanopore technology, the contig N50  
300 of the assembled genome achieved 22.46 Mb ,and the longest contig was 32.32 Mb. To the  
301 best of our knowledge, this is the highest contig N50 among all the sequenced fish genomes.  
302 This revealed that a combination of high-coverage Nanopore sequencing and Illumina data  
303 polishing can effectively produce highly contiguous genome assemblies. The contigs were  
304 clustered and ordered onto 20 pseudo-chromosomes with Hi-C data, and several pseudo-  
305 chromosomes were scaffolded with only 2 or 3 contigs. This high-quality genome will lay a  
306 strong foundation for a range of breeding, conservation and phylogenetic studies of filefish in  
307 the future.

308

#### 309 **Acknowledgements**

310 We appreciate the help from Tianyuan Fisheries Co., Ltd (Yantai, China) who provided  
311 the filefish samples. This work was supported by fund of Key Laboratory of Open-Sea  
312 Fishery Development, Ministry of Agriculture, P. R. China (LOF 2017-05), fund of  
313 Guangdong Provincial Key Laboratory of Fishery Ecology and Environment, South China  
314 Sea Fisheries Research Institute, Chinese Academy of Fisheries Sciences, SCSFRI, CAFS  
315 (FEEL-2017-10), Key Research and Development Program of Shandong province,  
316 Department of Science & Technology of Shandong province (2019GHY112073) and Central

317 Public-interest Scientific Institution Basal Research Fund, YSFRI, CAFS (20603022017014).

## 318 **References**

- 319 Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool.  
320 *Journal of molecular biology*, 215(3), 403-410. doi:10.1016/S0022-2836(05)80360-2
- 321 An, H. S., Kim, E. M., Lee, J. W., Dong, C. M., Lee, B. I., & Kim, Y. C. (2011). Novel polymorphic  
322 microsatellite loci for the Korean black scraper (*Thamnaconus modestus*), and their application to  
323 the genetic characterization of wild and farmed populations. *International journal of molecular*  
324 *sciences*, 12(6), 4104-4119. doi:10.3390/ijms12064104
- 325 An, H. S., Lee, J. W., Park, J. Y., & Jung, H. T. (2013). Genetic structure of the Korean black scraper  
326 *Thamnaconus modestus* inferred from microsatellite marker analysis. *Molecular Biology Reports*,  
327 40(5), 3445-3456. doi:10.1007/s11033-012-2044-7
- 328 Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J., Dehal, P., . . . Brenner, S. (2002). Whole-genome  
329 shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science*, 297(5585), 1301-1310.  
330 doi:10.1126/science.1072104
- 331 Austin, C. M., Tan, M. H., Harrisson, K. A., Lee, Y. P., Croft, L. J., Sunnucks, P., . . . Gan, H. M. (2017). *De*  
332 *novo* genome assembly and annotation of Australia's largest freshwater fish, the Murray cod  
333 (*Maccullochella peelii*), from Illumina and Nanopore sequencing read. *GigaScience*, 6(8), gix063.  
334 doi:10.1093/gigascience/gix063
- 335 Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in  
336 eukaryotic genomes. *Mobile Dna*, 6(1), 11. doi:10.1186/s13100-015-0041-9
- 337 Bian, L., Wang, P. F., Chen, S. Q., Li, F. H., Zhang, L. L., Liu, C. L., & Ge, J. L. (2018). Population genetic  
338 structure of *Thamnaconus septentrionalis* in China's coastal waters based on mitochondrial *Cyt b*  
339 sequences. *Journal of Fishery Sciences of China*, 25(4), 827-836. doi:10.3724/SP.J.1118.2018.18072  
340 (in Chinese)
- 341 Blanco, E., Parra, G., & Guigó, R. (2007). Using geneid to identify genes. *Current protocols in bioinformatics*,  
342 18(1), 4.3.1-4.3.28. doi:10.1002/0471250953.bi0403s00
- 343 Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., Estreicher, A., Gasteiger, E., . . . Schneider, M. (2003).  
344 The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids*  
345 *research*, 31(1), 365-370. doi:10.1093/nar/gkg095
- 346 Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of*  
347 *molecular biology*, 268(1), 78-94. doi:10.1006/jmbi.1997.0951
- 348 Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., & Shendure, J. (2013). Chromosome-  
349 scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nature*  
350 *Biotechnology*, 31(12), 1119-1125. doi:10.1038/nbt.2727
- 351 Chen, W. Z., Li, C. S., & Hu, F. (2000). Application and improvement of Virtual Population Analysis (VPA)  
352 in stock assessment of *Thamnaconus septentrionalis*. *Journal of Fisheries of China*, 24(6), 522-526.  
353 (in Chinese)
- 354 Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal  
355 tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*,  
356 21(18), 3674-3676. doi:10.1093/bioinformatics/bti610
- 357 Consortium, G. O. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic acids*



- 358 *research*, 32(Database issue), D258-D261. doi:10.1093/nar/gkh036
- 359 Daub, J., Eberhardt, R. Y., Tate, J. G., & Burge, S. W. (2015). Rfam: annotating families of non-coding RNA  
360 sequences. In Picardi E. (Eds.), *RNA Bioinformatics. Methods in Molecular Biology* (pp. 349-363).  
361 New York, NY: Humana Press.
- 362 De Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: a computational tool for the study  
363 of gene family evolution. *Bioinformatics*, 22(10), 1269-1271. doi:10.1093/bioinformatics/btl097
- 364 Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput.  
365 *Nucleic acids research*, 32(5), 1792-1797. doi:10.1093/nar/gkh340
- 366 Edgar, R. C., & Myers, E. W. (2005). PILER: identification and classification of genomic repeats.  
367 *Bioinformatics*, 21(suppl\_1), i152-i158. doi:10.1093/bioinformatics/bti1003
- 368 Gao, Y., Gao, Q., Zhang, H., Wang, L., Zhang, F., Yang, C., & Song, L. (2014). Draft sequencing and analysis  
369 of the genome of pufferfish *Takifugu flavidus*. *DNA Research*, 21(6), 627-637.  
370 doi:10.1093/dnares/dsu025
- 371 Ge, H., Lin, K., Shen, M., Wu, S., Wang, Y., Zhang, Z., . . . Zheng, L. (in press). *De novo* assembly of a  
372 chromosome-level reference genome of red-spotted grouper (*Epinephelus akaara*) using nanopore  
373 sequencing and Hi-C. *Molecular ecology resources*. doi:10.1111/1755-0998.13064
- 374 Griffiths-Jones, S., Grocock, R. J., Van Dongen, S., Bateman, A., & Enright, A. J. (2006). miRBase:  
375 microRNA sequences, targets and gene nomenclature. *Nucleic acids research*, 34(Database issue),  
376 D140-D144. doi:10.1093/nar/gkj112
- 377 Guan, J., Ma, Z., Zheng, Y., Guan, S., Li, C., & Liu, H. (2013). Breeding and larval rearing of bluefin  
378 leatherjacket, *Thamnaconus modestus* (Gunther, 1877) under commercial scales. *International*  
379 *Journal of Aquaculture*, 3(12), 55-62. doi:10.5376/ija.2013.03.0012
- 380 Guindon, S., Dufayard, J., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms  
381 and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML  
382 3.0. *Systematic biology*, 59(3), 307-321. doi:10.1093/sysbio/syq010
- 383 Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith Jr, R. K., Hannick, L. I., . . . White, O.  
384 (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment  
385 assemblies. *Nucleic acids research*, 31(19), 5654-5666. doi:10.1093/nar/gkg770
- 386 Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., . . . Wortman, J. R. (2008). Automated  
387 eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced  
388 Alignments. *Genome Biology*, 9(1), R7. doi:10.1186/gb-2008-9-1-r7
- 389 Han, Y., & Wessler, S. R. (2010). MITE-Hunter: a program for discovering miniature inverted-repeat  
390 transposable elements from genomic sequences. *Nucleic acids research*, 38(22), e199.  
391 doi:10.1093/nar/gkq862
- 392 Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., & Quesneville, H. (2014).  
393 PASTEC: an automatic transposable element classification tool. *Plos One*, 9(5), e91929.  
394 doi:10.1371/journal.pone.0091929
- 395 Jaillon, O., Aury, J., Brunet, F., Petit, J., Stange-Thomann, N., Mauceli, E., . . . Roest Crollius, H. (2004).  
396 Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-  
397 karyotype. *Nature*, 431(7011), 946-957. doi:10.1038/nature03025
- 398 Jansen, H. J., Liem, M., Jong-Raadsen, S. A., Dufour, S., Weltzien, F. A., Swinkels, W., . . . Henkel, C. V.  
399 (2017). Rapid *de novo* assembly of the European eel genome from nanopore sequencing reads.  
400 *Scientific Reports*, 7(1), 7213. doi:10.1038/s41598-017-07650-6

- 401 Kadobianskyi, M., Schulze, L., Schuelke, M., & Judkewitz, B. (2019). Hybrid genome assembly and  
402 annotation of *Danionella translucida*. *Scientific Data*, 6(1), 156. doi:10.1038/s41597-019-0161-z
- 403 Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids*  
404 *research*, 28(1), 27-30. doi:10.1093/nar/28.1.27
- 405 Kang, K. H., Kho, K. H., Chen, Z. T., Kim, J. M., Kim, Y. H., & Zhang, Z. F. (2004). Cryopreservation of  
406 filefish (*Thamnaconus septentrionalis* Gunther, 1877) sperm. *Aquaculture research*, 35(15), 1429-  
407 1433. doi:10.1111/j.1365-2109.2004.01166.x
- 408 Keilwagen, J., Wenk, M., Erickson, J. L., Schattat, M. H., Grau, J., & Hartung, F. (2016). Using intron  
409 position conservation for homology-based gene prediction. *Nucleic acids research*, 44(9), e89.  
410 doi:10.1093/nar/gkw092
- 411 Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory  
412 requirements. *Nature Methods*, 12(4), 357-360. doi:10.1038/nmeth.3317
- 413 Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable  
414 and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome*  
415 *Research*, 27(5), 722-736. doi:10.1101/gr.215087.116
- 416 Korf, I. (2004). Gene finding in novel genomes. *BMC bioinformatics*, 5(1), 59. doi:10.1186/1471-2105-5-59
- 417 Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform.  
418 *Bioinformatics*, 25(14), 1754–1760. doi:10.1093/bioinformatics/btp324
- 419 Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic  
420 genomes. *Genome Research*, 13(9), 2178-2189. doi:10.1101/gr.1224503
- 421 Li, P. L., Jiang, M. C., Xu, J. B., & Liu, B. (2002). The preliminary net cage culturing experiment of  
422 *Thamnaconus septentrionalis*. *China Fisheries*, 8, 61-62. (in Chinese)
- 423 Lin, X. Z., Gan, J. B., Zheng, Y. J., & Guan, X. D. (1984). The migration research of *Thamnaconus*  
424 *septentrionalis* in China. *Marine Fisheries*, 3, 99-108. (in Chinese)
- 425 Liu, K., Zhang, L. L., Zhang, Q. W., Chen, S. Q., Liu, C. L., & Bian, L. (2017). Study on *Thamnaconus*  
426 *septentrionalis* under industrial aquaculture condition. *Fishery modernization*, 44(3), 35-40.  
427 doi:10.3969/j.issn.1007-9580.2017.03.006 (in Chinese)
- 428 Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes  
429 in genomic sequence. *Nucleic acids research*, 25(5), 955-964. doi:10.1093/nar/25.5.955
- 430 Majoros, W. H., Pertea, M., & Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source *ab initio*  
431 eukaryotic gene-finders. *Bioinformatics*, 20(16), 2878-2879. doi:10.1093/bioinformatics/bth315
- 432 Mizuno, K., Shimizu-Yamaguchi, S., Miura, C., & Miura, T. (2012). Method for efficiently obtaining  
433 fertilized eggs from the black scraper *Thamnaconus modestus* by natural spawning in captivity.  
434 *Fisheries science*, 78(5), 1059-1064. doi:10.1007/s12562-012-0527-z
- 435 Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*,  
436 29(22), 2933-2935. doi:10.1093/bioinformatics/btt509
- 437 Pan, H., Yu, H., Ravi, V., Li, C., Lee, A. P., Lian, M. M., . . . Venkatesh, B. (2016). The genome of the largest  
438 bony fish, ocean sunfish (*Mola mola*), provides insights into its fast growth rate. *GigaScience*, 5(1),  
439 36. doi:10.1186/s13742-016-0144-3
- 440 Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic  
441 genomes. *Bioinformatics*, 23(9), 1061-1067. doi:10.1093/bioinformatics/btm071
- 442 Payne, A., Holmes, N., Rakyen, V., & Loose, M. (2018). BulkVis: a graphical viewer for Oxford nanopore  
443 bulk FAST5 files. *Bioinformatics*, 35(13), 2193-2198. doi:10.1093/bioinformatics/bty841

- 444 Perteua, M., Perteua, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie  
445 enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*,  
446 33(3), 290-295. doi:10.1038/nbt.3122
- 447 Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). *De novo* identification of repeat families in large genomes.  
448 *Bioinformatics*, 21(suppl\_1), i351-i358. doi:10.1093/bioinformatics/bti1018
- 449 Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., . . . Aiden, E.  
450 L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin  
451 looping. *Cell*, 159(7), 1665-1680. doi:10.1016/j.cell.2014.11.021
- 452 Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C. J., Vert, J. P., . . . Barillot, E. (2015). HiC-Pro:  
453 an optimized and flexible pipeline for Hi-C data processing. *Genome Biology*, 16(1), 259.  
454 doi:10.1186/s13059-015-0831-x
- 455 Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO:  
456 assessing genome assembly and annotation completeness with single-copy orthologs.  
457 *Bioinformatics*, 31(19), 3210-3212. doi:10.1093/bioinformatics/btv351
- 458 Stanke, M., & Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel.  
459 *Bioinformatics*, 19(suppl\_2), ii215-ii225. doi:10.1093/bioinformatics/btg1080
- 460 Su, J. X., & Li, C. S. (2002). *Fauna Sinica: Osteichthyes-Tetraodontiformes, Pegasiformes, Gobiesociformes,*  
461 *Lophiiformes*. Beijing: Science Press. (in Chinese)
- 462 Tan, M. H., Austin, C. M., Hammer, M. P., Lee, Y. P., Croft, L. J., & Gan, H. M. (2018). Finding Nemo:  
463 hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish  
464 (*Amphiprion ocellaris*) genome assembly. *GigaScience*, 7(3), gix137.  
465 doi:10.1093/gigascience/gix137
- 466 Tang, S., Lomsadze, A., & Borodovsky, M. (2015). Identification of protein coding regions in RNA  
467 transcripts. *Nucleic acids research*, 43(12), e78. doi:10.1093/nar/gkv227
- 468 Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic  
469 sequences. *Current protocols in bioinformatics*, 5(1), 4.10.11-14.10.14.  
470 doi:10.1002/0471250953.bi0410s25
- 471 Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., . . . Natale, D. A.  
472 (2003). The COG database: an updated version includes eukaryotes. *BMC bioinformatics*, 4(1), 41.  
473 doi:10.1186/1471-2105-4-41
- 474 Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate *de novo* genome assembly from  
475 long uncorrected reads. *Genome Research*, 27(5), 737-746. doi:10.1101/gr.214270.116
- 476 Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., . . . Earl, A. M. (2014). Pilon:  
477 an integrated tool for comprehensive microbial variant detection and genome assembly improvement.  
478 *Plos One*, 9(11), e112963. doi:10.1371/journal.pone.0112963
- 479 Xu, G. B., Chen, S. L., & Tian, Y. S. (2010). New polymorphic microsatellite markers for bluefin  
480 leatherjacket (*Navodon septentrionalis* Gunther, 1877). *Conservation genetics*, 11(3), 1111-1113.  
481 doi:10.1007/s10592-009-9891-3
- 482 Xu, G. B., Tian, Y. S., Liao, X. L., & Chen, S. L. (2009). Isolation and characterization of polymorphic  
483 microsatellite loci from bluefin leatherjacket (*Navodon septentrionalis* Gunther, 1877).  
484 *Conservation genetics*, 10(4), 1181-1184. doi:10.1007/s10592-008-9739-2
- 485 Xu, Z., & Wang, H. (2007). LTR\_FINDER: an efficient tool for the prediction of full-length LTR  
486 retrotransposons. *Nucleic acids research*, 35(Web Server issue), W265-W268.

487 doi:10.1093/nar/gkm286

488 Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology & Evolution*,  
489 24(8), 1586-1591. doi:10.1093/molbev/msm088

490

#### 491 **Data Accessibility**

492 Raw sequencing reads are available on GenBank as BioProject PRJNA565600. Raw

493 sequencing data (Nanopore, Illumina , Hi-C and RNA-seq data) have been deposited in SRA

494 (Sequence Read Archive) database as SRX6875837, SRX6862879, SRX6875660, and

495 SRX6875519.

496

#### 497 **Author Contributions**

498 S.C., C.S. and Z.L. designed and managed the project. L.B., F.L. and J.G. interpreted the data

499 and drafted the manuscript. P.W., S.Z., C.L. and X.L. prepared the materials. Q.C., J.L., K.L.

500 and H.C. performed the DNA extraction, RNA extraction and libraries construction. L.B.,

501 F.L., X.L. and C.S. performed the bioinformatic analysis. All authors contributed to the final

502 manuscript editing.

503

**TABLE 1 Statistics of the sequencing data**

Types	Method	Sequencing platform	Library size (bp)	Clean data (Gb)	Coverage ( $\times$ ) <sup>†</sup>
Genome	Illumina	Illumina HiSeq X	300	45.97	93.48
Genome	Nanopore	PromethION	ultra-long	50.95	103.61
Genome	Hi-C	Illumina HiSeq X	300	39.44	80.20
Transcriptome	Illumina	Illumina HiSeq X	300	11.31	23.00

<sup>†</sup> The coverage was calculated using an estimated genome size of 491.74 Mb.

504

505

506

507

508

**TABLE 2 Assembly statistics of filefish and other tetraodontiform genomes**

Species	<i>T. septentrionalis</i>	<i>Takifugu rubripes</i> <sup>†</sup>	<i>Takifugu flavidus</i>	<i>Tetraodon nigroviridis</i>	<i>Mola mola</i>
Sequencing technology	Oxford Nanopore sequencing	PacBio Sequel	PacBio Sequel	Plasmid library + BAC library sequencing	Illumina Hiseq 2000
Assembly size (Mb)	474.31	384.13	366.29	342.40	639.45
Number of scaffolds	155	128	867	25773	5552
N50 scaffold size (Mb)	23.05	16.71	15.68	0.73	8.77
Number of contigs	242	530	1111	41566	51826
N50 contig length (Mb)	22.46	3.14	4.36	0.03	0.02

509

<sup>†</sup> The assembly statistics of other tetraodontiform genomes were from NCBI assembly database. The GenBank assembly accession numbers were as follows: *Takifugu rubripes* (GCA\_901000725.2), *Takifugu flavidus* (GCA\_003711565.2), *Tetraodon nigroviridis* (GCA\_000180735.1), *Mola mola* (GCA\_001698575.1).

511

512

513

514

515  
516

**TABLE 3 Statistics of the pseudo-chromosome assemblies using Hi-C data**

Group	Contig number	Contig length (bp)
Group 1	3	34,805,468
Group 2	3	34,142,503
Group 3	3	29,239,029
Group 4	13	27,092,115
Group 5	3	24,789,104
Group 6	7	24,144,372
Group 7	10	23,815,151
Group 8	3	23,107,901
Group 9	11	22,985,309
Group 10	5	23,048,615
Group 11	2	22,982,431
Group 12	6	23,025,906
Group 13	3	22,547,364
Group 14	11	22,005,842
Group 15	16	20,921,416
Group 16	3	20,603,809
Group 17	2	19,738,352
Group 18	5	17,694,734
Group 19	13	18,094,054
Group 20	25	16,862,837
Total contigs clustered	147	471,646,312
Total contigs ordered and oriented	107	469,464,378

517  
518  
519

**TABLE 4 Summary of predicted protein-coding genes in the filefish genome**

Method	Software	Species	Number of predicted genes
<i>Ab initio</i>	Genscan		28,628
	Augustus		44,749
	GlimmerHMM		34,576
	GeneID		24,446
	SNAP		58,914
Homology-based		<i>Takifugu rubripes</i>	19,643
	GeMoMa	<i>Tetraodon nigroviridis</i>	21,885
		<i>Danio rerio</i>	19,808
RNA-seq	PASA		30,768
	GeneMarkS-T		47,856
	TransDecoder		78,130
Integration	EVM		22,067

520

521

522

**TABLE 5 Summary of functional annotations for predicted genes**

Annotation database	Annotated number of predicted genes	Percentage (%)
GO	11,257	51.01%
KEGG	13,714	62.15%
KOG	14,760	66.89%
TrEMBL	20,795	94.24%
NR	20,905	94.73%
All Annotated	20,924	94.82%
Predicted Genes	22,067	-

523

524

525

526



527

528

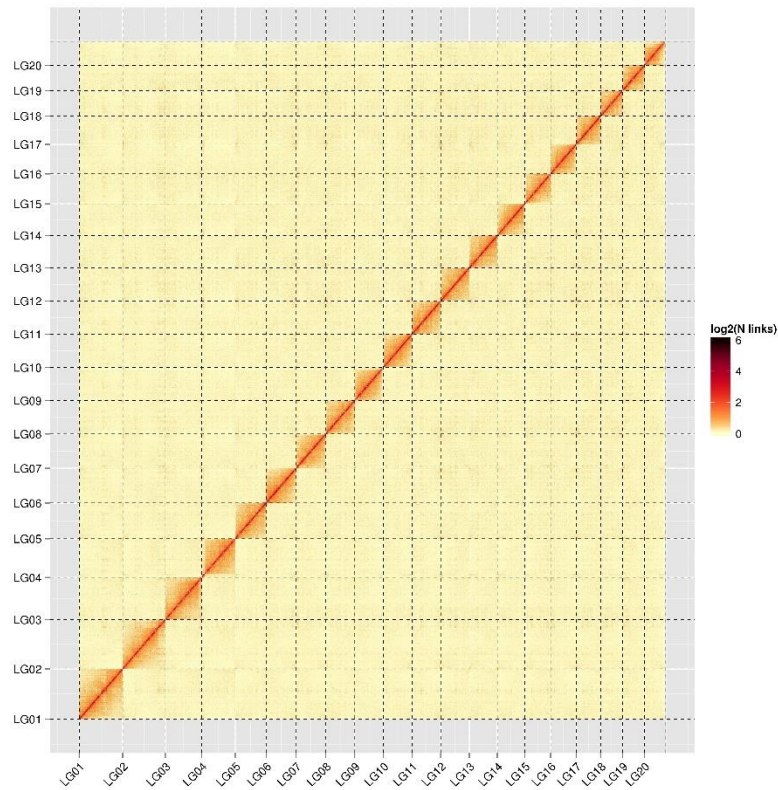
529

530

531

532

**FIGURE 1 The greenfin horse-faced filefish (*Thamnaconus septentrionalis*)**

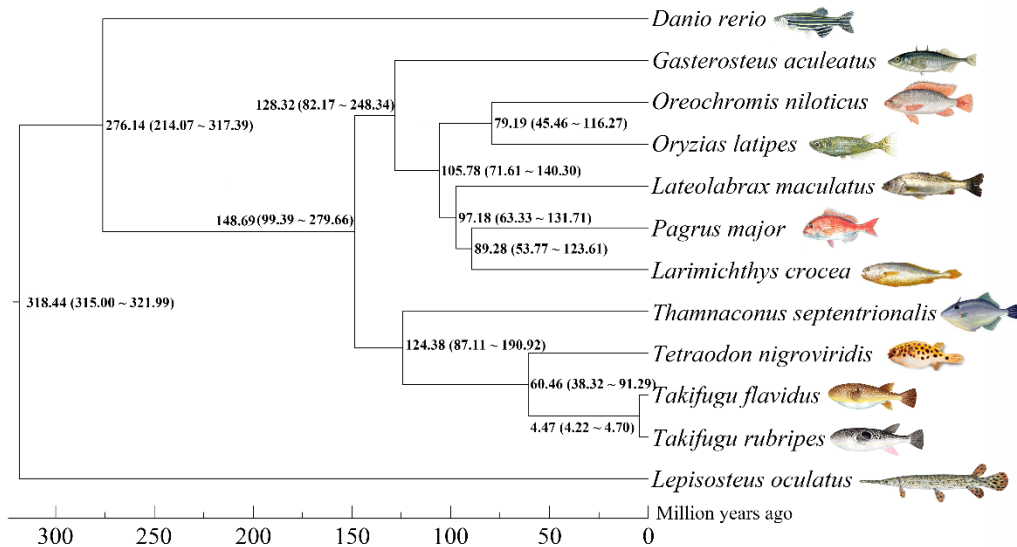


533

534 **FIGURE 2** The genome-wide Hi-C heatmap of the filefish. LG 1-20 are the abbreviations of Lachesis  
 535 **Group 1-20, representing the 20 pseudo-chromosomes.**

536

537

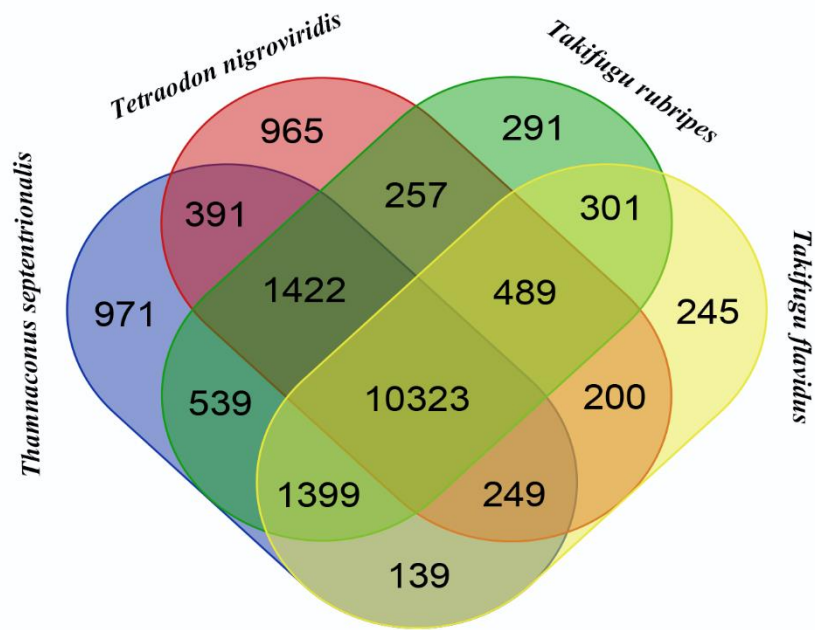


538

539 **FIGURE 3** Phylogenetic analysis of the filefish with other teleost species. *Lepisosteus oculatus* was  
 540 used as the outgroup. The estimated species divergence time (million years ago) and the 95%  
 541 confidential intervals were labeled at each branch site.

542





543

544

545

**FIGURE 4** Venn diagram of orthologous gene families among four tetraodontiform species.