Chromosome-level genome assembly and transcriptome of the tomato hind, Cephalopholis sonnerati (Serranidae, Perciformes)

Zhen-Zhen Xie[1], Cheng Peng[2], Dengdong Wang[2], Qing Wang[2], Shuisheng Li[2], Haoran Lin[2], and Yong Zhang[2]

[1]Jiujiang University
[2]Affiliation not available

March 07, 2024

Abstract

The tomato hind Cephalopholis sonnerati (Valenciennes) (Serranidae), belonging to the genus Cephalopholis, is a bottom dwelling coral reef of 12–120-m depth in the Indo-Pacific and Red Sea. C. sonnerati has also been characterized by complex social structures and behavioural mechanisms. However, due to the lack of genomic resource for C. sonnerati, molecular-genetic studies and genomic breeding remain unexplored in this species. In this study, we reported the chromosome-level genome assembly of C. sonnerati using PacBio sequencing and Hi-C sequencing technologies. We obtained a total length of 1043.66 Mb with an N50 length of 2.49 Mb, containing 795 contigs assembled into 24 chromosomes. Overall 95.8% of the complete BUSCOs were identified in the assembled genome, suggesting the completeness of the genome. Then, we predicted 26,130 protein-coding genes, of which 94.26% were functionally annotated. In addition, C. sonnerati diverged from its common ancestor with E. lanceolatus and E. akaara approximately 41.7 million years ago. Finally, we found tissue-specific expression of 8,108 genes. Functional analyses showed that they mainly consisted of complement and coagulation cascades, DNA replication, synaptic vesicle cycle, long-term potentiation and other glycan degradation. Furthermore, comparative genome analyses indicated that the expanded genes families were highly enriched in the sensory system, which was different from the enrichment analysis of the tissue-specific expression genes. In brief, to our knowledge, we reported the first chromosome-level genome assembly of C. sonnerati, which will provide a valuable genome resource for studies on the genetic conservation, resistance breeding, and evolutionary of C. sonnerati.

Chromosome-level genome assembly and transcriptome of the tomato hind,*Cephalopholis sonnerati* (Serranidae, Perciformes)

Running title: *Cephalopholis sonnerati* genome assembly

Zhenzhen Xie[1,2#], Cheng Peng[3#], Dengdong Wang[2], Qing Wang[4], Shuisheng Li[2], Haoran Lin[2], Yong Zhang[2*],

[1]College of Pharmacy and Life Science, Jiujiang University, Jiujiang 332200, China

[2]State Key Laboratory of Biocontrol, Guangdong Provincial Key Laboratory for Aquatic Economic Animals and Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), School of Life Sciences, Sun Yat-Sen University, Guangzhou, 510275, China.

[3]Guangdong Key Laboratory of Animal Conservation and Resource Utilization, Guangdong Public Laboratory of Wild Animal Conservation and Utilization, Institute of Zoology, Guangdong Academy of Sciences, Guangzhou 510260, China

[4] College of Marine Sciences, South China Agricultural University, Guangzhou 510642, China.

# These authors contributed equally to this paper.

*Corresponding author(s):

Prof. Yong Zhang

Sun Yat-Sen University

No. 135 Xinggang West Road, Guangzhou, Guangdong Province, China

Email: lsszy@mail.sysu.edu.cn

Abstract

The tomato hind*Cephalopholis sonnerati* (Valenciennes) (Serranidae), belonging to the genus*Cephalopholis* , is a bottom dwelling coral reef of 12–120-m depth in the Indo-Pacific and Red Sea. *C. sonnerati* has also been characterized by complex social structures and behavioural mechanisms. However, due to the lack of genomic resource for*C. sonnerati* , molecular-genetic studies and genomic breeding remain unexplored in this species. In this study, we reported the chromosome-level genome assembly of*C.sonnerati* using PacBio sequencing and Hi-C sequencing technologies. We obtained a total length of 1043.66 Mb with an N50 length of 2.49 Mb, containing 795 contigs assembled into 24 chromosomes. Overall 95.8% of the complete BUSCOs were identified in the assembled genome, suggesting the completeness of the genome. Then, we predicted 26,130 protein-coding genes, of which 94.26% were functionally annotated. In addition,*C. sonnerati* diverged from its common ancestor with*E. lanceolatus* and *E. akaara* approximately 41.7 million years ago. Finally, we found tissue-specific expression of 8,108 genes. Functional analyses showed that they mainly consisted of complement and coagulation cascades, DNA replication, synaptic vesicle cycle, long-term potentiation and other glycan degradation. Furthermore, comparative genome analyses indicated that the expanded genes families were highly enriched in the sensory system, which was different from the enrichment analysis of the tissue-specific expression genes. In brief, to our knowledge, we reported the first chromosome-level genome assembly of *C. sonnerati* , which will provide a valuable genome resource for studies on the genetic conservation, resistance breeding, and evolutionary of *C. sonnerati* .

Keywords

Cephalopholis sonnerati , chromosome-level genome assembly, genome annotation, comparative genome analyse

Introduction

Groupers (subfamily Epinephelinae species, Serranidae, Percoidei, Perciformes), the largest subfamily in the Serranidae family, consist of more than 160 species in 16 genera (Zhang et al., 2013). These commercially important fishes possess special characteristics of a long lifespan, large size, slow growth, vulnerability and delayed reproduction (Morris et al., 2000). Moreover, they usually inhabit coral reefs of tropical and subtropical coasts. Of them, the genus*Cephalopholis*is the most abundant serranid in the Gulf of Aqaba (Red Sea) (Shpigel & Fishelson, 2010).

The tomato hind*Cephalopholis sonnerati*(Valenciennes) (Serranidae), belonging to the genus*Cephalopholis* , is a bottom-dwelling coral reef of 12–120-m depth in the Indo-Pacific and Red Sea.*C.sonnerati* are protogynous hermaphrodites in life and feeding on little fish and invertebrates (Shpigel, 1985; Shpigel & Fishelson, 1989a,b; Shpigel & Fishelson, 2010). Furthermore, they are also characterized by complex social structures and behavioural mechanisms. They naturally form social groups, with males and several females occupying individual territories within the male's larger territory (Meyer, 2008; Shpigel & Fishelson, 1989b). However, due to overfishing, anthropogenic activities and water pollution, the natural populations of *C.sonnerati* have directly declined (Hawkins & Roberts, 1994). Previous studies of the genus*Cephalopholi*mainly focused on fishery management, species conservation (Galal-Khallaf et al., 2018), behavior biology (Shpigel & Fishelson, 2010), nutrition biology, and phylogeographic biology (Gaither et al., 2011). Nevertheless, owing to the lack of genomic resources, molecular-genetic studies and genomic breeding remain unexplored in this species.

PacBio (a single-molecule real-time [SMRT] sequencing), a newly third-generation sequencing technology, generates long reads with uniform coverage and high consensus accuracy compared with the second-generation sequencing technology that generates short reads (Rhoads & Au, 2015). Morever, third-generation sequencing technology is less expensive than second-generation sequencing technology and does not depend on amplification for library generation (Ze-Gang & Shao-Wu, 2018). Additionally, Hi-C, a chromosome conformation capture-based method, can convert chromatin interactions, reflecting topological chromatin structures into digital information (Belaghzal et al., 2017). Presently, it has become a mainstream technology in 3D genomics. Despite that more than 270 aquatic organisms' genome sequences have been published (https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/fish), only three genome sequences of grouper species (the giant grouper*Epinephelus lanceolatus* [Zhou et al., 2019], the red-spotted grouper*Epinephelus akaara* [Ge et al., 2019] and the leopard coral grouper, *Plectropomus leopardus* [Zhou et al., 2020] are available. Therefore, it is significantly important to gain more genome sequences of grouper species for the research on the classification, evolutionary, genetics, and biological studies of groupers.

In the present study, we reported the first chromosome-level genome assembly of *C.sonnerati,* which was obtained by using PacBio long-read sequencing and Hi-C sequencing technologies. Our reference genome will lay a solid foundation for studies on the genetics conservation, resistance breeding and evolutionary of *C. sonnerati* .

2. Materials and Methods

2.1 DNA sampling and tissue collections

A female adult (*C. sonnerati*), bred at the farm of Hainan, Dongfang, Gancheng, China, was used for genome sequencing and assembly. The fish was dissected immediately after treatment with 0.2 M eugenol. Genomic DNA of *C. sonnerati* was collected from the caudal vein by a Qiagen Blood & Cell Culture DNA Midi Kit, which was used for genome sequencing. The muscle tissue was used for Hi-C library construction in order to obtain a chromosome-scale genome assembly. Furthermore, tissues from the liver, gill, intestines, kidney, head kidney, brain, pituitary, gonad, heart, skin, and muscle were collected and quickly frozen in liquid nitrogen before RNA sequencing, and then the tissues were kept at –80°C at Sun Yat-Sen University.

2.2 DNA extraction and sequencing

High-quality genomic DNA was extracted from blood samples using a modified CTAB (Hexadecyl Trimethyl Ammonium Bromide) method. The quality and quantity of the extracted DNA were examined using a NanoDrop 2000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA), Qubit ds DNA HS Assay Kit on a Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA), and electrophoresis on a 0.8% agarose gel.

A paired-end sequencing library with an insertion length of 250 bp was constructed using the VAHTS Universal DNA Library Prep Kit for MGI (Vazyme, Nanjing, China). The Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) was used to validated the purity and size distribution. Then, the obtained library was sequenced with the paired-end, 150-bp mode using the MGI-SEQ2000 platform by Frasergen Bioinformatics Co., Ltd. (Wuhan, China).

Ten micrograms (ug) of *C. sonnerati* genomic DNA were used for 20 kb template library preparation using the BluePippin Size Selection system (Sage Science, USA) following the manufacturer's protocol. The library was sequenced on the Pacfic Biosciences Sequel II platform.

2.3 RNA extraction and sequencing

For the gene annotation and the prediction of protein-coding genes, 11 tissues above-mentioned of *C. sonnerati* were used to conduct transcriptome sequencing. Total RNA was extracted with the Trizol Reagent (Invitrogen, USA) according to the manufacturer's instructions. The concentration and integrity of total RNA were estimated using the Agilent 2100 Bioanalyzer (Aglient Technologies, USA) and ethidium bromide staining of 28S and 18S ribosomal bands on a 1% agarose gel, respectively. Equal volumes of RNA samples were pooled together for RNA library construction and sequencing. Briefly, the full-length cDNA was prepared using a SMARTer^{TM} PCR cDNA Synthesis Kit (Takara Biotechnology, Dalian, China). The SMRTbell libraries were constructed with the Pacific Biosciences DNA Template Prep Kit 2.0. Library. Library quantification and size were checked using a Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) and a 2100 Bioanalyzer system (Agilent Technologies, CA, USA), respectively. Subsequently, SMRT sequencing was carried out with a PacBio Sequel II platform by Frasergen Bioinformatics Co., Ltd. (Wuhan, China).

2.4 Genome size estimation

The short-reads from the BGI platform were quality filtered by HTQC v 1.92.310 (Yang. et al.,2013) using the following method. Firstly, the adapters were removed from the sequencing reads. Second, read pairs were excluded if any one end had an average quality lower than 20. Third, the ends of reads were trimmed if the average quality was lower than 20 in the sliding window size of 5 bp. Finally, read pairs with any end shorter than 50 bp were removed. Then, the quality filtered reads were used for genome size estimation. We estimated the genome size of the*C. sonnerati* genome by using the *k-mer* analysis, which was performed with GCE (Liu et al., 2013).

2.5 Genome assembly

The draft assembly of the genome was assembled using mecat2 (Xiao et al., 2017c) with default parameters. To correct errors in the primary assembly, we used gcpp 1.9.0 to polish the genome after the initial assembly of the genome was completed. In addition, we used BGI derived short reads to correct any remaining errors by Pilon 1.22 (Walker et al., 2014). Finally, we used BUSCO v3.0 (Simão et al., 2015) with actinoperygii_odb9 to evaluate the completeness of the assembled genome.

2.6 Chromosome assembly using Hi-C technology

Muscle tissue of *C. sonnerati* was used for Hi-C library construction in our study. The Hi-C experiment included the following steps (Belaghzal, Dekker, & Gibcus, 2017). First, a white muscle sample of *C. sonnerati* was cross-linked using formaldehyde and then lysed. Subsequently, chromatin digestion was carried out with MboI and proximity ligated with T4 DNA ligase. After ligation, cross-linking was reversed by 200 µg/mL proteinase K (Thermo) at 65°C overnight. DNA purification was achieved through the QIAamp DNA Mini Kit(Qiagen) according to the manufacturer's instructions, and the purified DNA was sheared to a length of 300–500 bp. Lastly, the purified DNA was used for Hi-C library construction, and genomic DNA was sequenced on the MGI-SEQ2000 platform in 150PE mode.

The reads from the Hi-C library sequencing were mapped to the polished genome using BWA (bwa 0.7.17) with the default parameters. Paired reads that were mate mapped to different contigs were used to construct the Hi-C associated scaffolding. Lachesis (Burton et al., 2013) was further applied to order and orient the clustered contigs. Then, Jucier (v1.6.2) (Durand et al., 2016) was used to corrected the assembly error in visually.

2.7 Repetitive sequence annotation

Two methods were combined to identify the repeat contents in the genome: homology-based and de novo prediction. For homology-based analysis, we identified the known TEs within the *C. sonnerati* genome using RepeatMasker 4.0.9 (Tarailo-Graovac et al., 2009) to identify with the Repbase TE library (Jurka et al., 2000, 2005). Repeat Protein Mask searches were also conducted using the TE protein database as a query library. For de novo prediction, we constructed a *de novo* repeat library of the *C. sonnerati* genome using RepeatModeler (http://www. org/RepeatModeler/), which can automatically execute two core de novo repeat finding programs, namely, RECON v1.08 (Bao & Eddy, 2002) and RepeatScout (v1.0.5) (Price et al., 2005), to comprehensively conduct, refine and classify consensus models of putative interspersed repeats for the *C. sonnerati* genome. Furthermore, we performed a de novo search for long terminal repeat (LTR) retrotransposons against the *C. sonnerati* genome sequences using LTR_FINDER (v1.0.7) (Xu & Wang, 2007). We also identified tandem repeats using the Tandem Repeat Finder (TRF) package (Benson, 1999) and the non interspersed repeat sequences, including low complexity repeats, satellites and simple repeats, using Repeat Masker. Finally, we merge the library files of the two methods and use repeat maker to identify the repeat contents.

2.8 Gene prediction and annotation

For the prediction of protein-coding genes in the assembled genome of*C. sonnerati* , we used three strategies: homology, *de novo*and transcriptome sequencing. First, protein sequences from*Epinephelus lanceolatus* , *Plectropomus leopardus* ,*Epinephelus akaara* , *Oreochromis niloticus* , *Lates calcarifer* , *Gymnodraco acuticeps* , *Pseudochaenichthys georgianus* and *Cyclopterus lumpus* were downloaded from Ensembl (Flicek et al., 2014) and aligned with *C. sonnerati* for homology annotation. Exonerate (v2.2.0) was used to conduct homology-based gene prediction. Second, we adopted Augustus (v3.3.1) (Stanke et al., 2004) and Genescan (Burge & Karlin, 1997) to perform *de novo* gene prediction. Third, protein-coding gene prediction based on transcriptome sequencing data was carried out using GMAP (version 2018-07-04) (Wu et al., 2005). TransDecoder (3.0.1) (https://github.com/TransDecoder/TransDecoder) was used to form the gene structure. Finally, Maker (v3.00) (Cantarel et al., 2008) was used to integrate the prediction results of the three methods to predict gene models.

Gene functions were inferred according to the best match of the alignments to the non-redundant (NR), TrEMBL (Boeckmann et al., 2003), InterPro (Mitchell et al., 2015), and SwissProt (Boeckmann et al., 2003) protein databases using BLASTP (NCBI blast v2.6.0+) (Altschul et al., 1997; Camacho et al., 2009) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al., 2012) with an e- value threshold of $1e^{-5}$. The protein domains were annotated using PfamScan (pfamscan_version) (Mistry et al., 2007) and InterProScan (v5.35 74.0) (Jones et al., 2014) based on InterPro protein databases. The motifs and domains within gene models were identified by PFAM databases (Finn et al., 2008). Gene Ontology (GO) (Ashburner et al., 2000) IDs for each gene were obtained from Blast2GO (Conesa & Gota, 2008).

In addition, we used tRNAscan SE (v1.3.1) algorithms (Lowe & Eddy, 1997) and tRNAscan with default parameters to identify the genes associated with tRNA. For rRNA identification, we first downloaded the closely related species rRNA sequences from the Ensembl database. Then rRNAs in the database were aligned against our genome using BlastN (Altschul et al., 1997; Camacho et al., 2009) with a cut-off of e-value $<1e^{-5}$, identity of [?]85%, and match length [?] 50bp. MiRNAs and snRNAs were identified by the Infernal (v1.1.2) (Nawrocki et al., 2009) software against the Rfam (v14.1) database (Finn et al., 2008) with default parameters.

2.9 Comparative genomic and phylogenetic analysis

To identify the gene families for phylogenetic tree construction, we compared the genome assembly of *C. sonnerati* with other fish, including *Epinephelus lanceolatus* , *Plectropomus leopardus* ,*Epinephelus akaara* , *Oreochromis niloticus* , *Lates calcarifer* , *Gymnodraco acuticeps* , *Pseudochaenichthys georgianus* , *Cyclopterus lumpus, Danio rerio* , *Salmo salar* , *Monopterus albus* , *Monopterus albus* , *Gadus morhua* , *Oncorhynchus mykiss* , and *Oryzias latipes* .*Latimeria chalumnae* was used as an outgroup. All of the proteins were extracted and aligned to each other using BLASTP (Camacho et al., 2009) programs (NCBI blast v2.6.0) with a maximal e-value of $1e^{-5}$. The OrthoFinder (Emms & Kelly, 2015) method was used to cluster genes from these different species into gene families.

To reveal the phylogenetic relationships among *C. sonnerati* and the aforementioned fishes, protein sequences from 678 single-copy orthologous gene clusters were used for phylogenetic tree reconstruction. The protein sequences of the single-copy orthologous genes were aligned with the MUSCLE (v3.8.31) (Edgar, 2004) program, and the corresponding Coding DNA Sequences (CDS) alignments were generated and concatenated with the guidance of protein alignment. RAxML (v8.2.11) (Stamatakis, 2014) was used to construct the phylogenetic tree with the maximum likelihood method. The phylogenetic relationship of other fish was consistent with previous studies. We used the MCMCTree program of the PAML package (Yang, 2007) to estimate the divergence time among species.

2.10 Gene family expansion and contraction analysis

Based on the identified gene families and the constructed phylogenetic tree with the predicted divergence time of those fish, we used CAFE (Han et al., 2013) to analyze gene family expansion and contraction. In CAFE, a random birth and death model was proposed to study gene gain or loss in gene families across a specified phylogenetic tree. Then, a conditional p-value was calculated for each gene family, and a family with a conditional p-value less than 0.05 was considered to have an accelerated rate for gene gain or loss. These expanded and contracted gene families in R. canadum (p-value [?] 0.05) were mapped to KEGG pathways for functional enrichment analysis, which was conducted using the enrichment methods. This method implemented hypergeometric test algorithms and the Q-value (FDR, False Discovery Rate) was calculated to adjust the p-value using the R package (*https://github.com/StoreyLab/qvalue)*.

2.11 Detection of positive selective genes

Based on the phylogenetic tree, we estimated the rate ratio (ω) of

non synonymous(*Ka*) to synonymous(*Ks*) nucleotide substitutions using the PAML (v4.9e) package (Yang, 2007) to examine the selective constraints on candidate 678 single-copy orthologous genes. After the high-quality alignments of related sequences were obtained as described above, we compared a series of evolutionary models in the likelihood framework using the species trees. A branch site model was used to detect the average ω across the tree (ω0), ω of the appointed branch to test (ω2), and ω of all of the other branches (ω1).

2.12 Identification of differentially expressed genes

To identify the differentially expressed genes in the genome of *C. sonnerati* , 11 tissues (liver, gill, intestines, kidney, head kidney, brain, pituitary, gonad, heart, skin, and muscle) were used to conduct the transcriptome sequencing. For each of the samples, the trimmed short reads were mapped to the genome sequence using Tophat (v2.1.1; https ://ccb.jhu.edu/software/tophat). RSEM (v1.3.0; https://deweylab. github. io/RSEM) was used to calculate isoform level expression in terms of FPKM and TPM (transcripts per million). Differentially expressed genes (DEGs) between sample groups were evaluated by DESeq2 (Love, Huber, & Anders, 2014). The corrected read count data of genes were imported into the R package EdgeR to identify DEGs with the criteria of a fold change of [?] 2.0, a false discovery rate [FDR] and adjusted p value of < 0.05, and expression (FPKM [?] 1) in at least one sample for each comparison.

3. Results and discussion

3.1 Genome assembly

In this study, we generated a high-quality chromosome-level genome assembly of *C. sonnerati* using a combination of PacBio sequencing and Hi-C sequencing technologies. We obtained 56.98 Gb of clean short-read sequencing data from the genome of *C. sonnerati*(Figure 1). Then, the quality clean reads were used for genome size estimation by the k-mer-based methods (Liu. et al., 2013). Accordingly, the genome size of *C. sonnerati* was estimated to be 1015 Mb, with the proportion of repeat sequences and the heterozygosity rate determined to be 0.84% and 42.99%, respectively (Figure 2, Table 1).

With the SMRT cells in the PacBio Sequel platform, we generated ~100X subreads by removing adaptor sequences within sequences. The longest 150X subreads data was used for genome assembly of *C. sonnerati* . Then the draft assembly of the genome was assembled using mecat2 (Xiao et al., 2017) with default parameters. To correct errors in the primary assembly, we used gcpp (v1.9.0) (*https://github.com/PacificBiosciences/gcpp)*to polish the genome after the initial assembly of the genome was completed. In addition, we used Illumina derived short reads to correct any remaining errors by Pilon (v1.22) (Walker et al., 2014). Finally, we produced a total length of about 1043.66 Mb with an N50 length of 2.49 Mb, which accounted for 97.3% of the genome size estimated by k-mer analysis, containing 795 contigs (Table 2). Moreover, the genome of the*C. sonnerati* was longer than that the genome of the leopard coral grouper *Plectropomus leopardus* (881.55 Mb) (Zhou et al., 2020) but shorter than the genome of the red spotted grouper *Epinephelus akaara* (1135 Mb) (Ge et al., 2019). Furthermore, the assembled genome was subjected to BUSCO (Benchmarking Universal Single-Copy Orthologs) v3.0.2 with OrthoDB to evaluate the completeness of the genome. Overall, 95.8% and 95.6% of the complete BUSCOs were identified in the assembled and annotated genome, respectively (Supplementary Table S1). The results validated that the genome assembly was complete.

For anchored contigs, 801,816,224 clean read pairs were generated from the Hi-C library and were mapped to the polished *C. sonnerati*genome using BWA (bwa 0.7.17) with the default parameters. Then, we generated 324,980,877 unique mapped paired-end reads that were used to perform the Hi-C-associated scaffolding. Finally, we successfully clustered 795 contigs into 24 groups with the agglomerative hierarchical clustering method (Burton et al., 2013) in *C. sonnerati* . Subsequently, the genome of *C. sonnerati* was applied to order and orient the clustered contigs. Similarly, there were 767 contigs successfully ordered and oriented with 1.02 Gb. Finally, we obtained the first chromosome-level high-quality assembly, and chromosomal lengths ranged from 2.52 to 44.48 Mb, containing 98.01% of the total sequence (Table 3).

3.2 Genome annotation

Repeat sequences that were 526.92 Mb in length, accounting for 50.47%, were identified in the assembled genome of the *C. sonnerati* . The TEs accounted for 47.23% with 493.11 Mb in length of the assembly genome (Table 4). The percentage was higher than that of*Plectropomus leopardus* (30.74%) (Zhou et al., 2020) and*Epinephelus akaara* (43.02%) (Ge et al., 2019). Among them, DNA transposons, LINEs, and LTRs were the top three categories of repetitive elements, accounting for 24.82, 13.74, and 6.72%, respectively.

We predicted protein-coding genes of the *C. sonnerati* genome by using three methods, including *de novo* , homology-based and transcriptome sequencing-based gene predictions. A total of 26,130 protein-coding genes were generated from the genome of *C. sonnerati* (Supplementary Table S2). Then, the statistics of the predicted gene models were compared with eight closet teleost species (*E. lanceolatus* , *P. leopardus* , *E. akaara* ,*O. niloticus* ,*L. calcarifer* ,*G. acuticeps* ,*P. georgianus* and*C. lumpus*), displaying similar distribution patterns in the exon and intron number, gene and CDS length, exon and intron length, and gene and CDS gene content of *C. sonnerati* (Figure 3). In total, 24,629 genes (approximately 94.26%) were functionally annotated in at least one of the databases (Table 5), which is higher than that of *E. akaara* (23,808) (Ge et al., 2019) and *P. leopardus* (24,364) (Zhou et al., 2020), but lower than that of *E. lanceolatus*(24,794) (Zhou et al., 2019).

For non-coding genes, 373 miRNAs, 2,232 tRNAs, 169 rRNAs and 515 snRNAs were also identified in the genome of *C. sonnerati* (Supplementary Table S3).

3.3 Phylogeny and divergent time

We identified 698 single-copy orthologues by using the sequencing similarities among protein-coding genes between 15 selected species. Additionally, a phylogenetic tree was constructed on the 678 filtered single-copy orthologues from 15 species genomes to reveal the phylogenetic relationship between them. We found that the *C. sonnerati*diverged approximately 41.7 million years ago (mya) from the common ancestor with*E. lanceolatus* and *E. akaara* . In addition,*P.leopardus* was the most closely related ancestor species to the*C. sonnerati* , separating from their common ancestor 66.4 to 75.7 mya (Figure 4).

3.4 Genomic comparison with other species

We conducted functional comparative genomic analyses with the four groupers (*P. leopardus* , *E. akaara* , *E. lanceolatus* and *C. sonnerati*) to reveal the similarities and differences between them by constructing orthologous gene families. The results were demonstrated through the Venn diagram (Supplementary Figure S1). Specifically, the numbers of gene families were highly similar in the four groupers, with 17,125, 16,842, 17,205, and 16,674 in*C. sonnerati* , *E. lanceolatus* , *P. leopardus* and*E. akaara* , respectively. The four groupers shared 14,512 genes, and 406 genes were specific to *C. sonnerati* . We also found that *C. sonnerati* shared 15,818, 15,594, and 15,877 genes with*E. lanceolatus* , *E. akaara* and *P.leopardus* , respectively. In addition, we conducted functional comparative genomic analysis between the genome of *C. sonnerati*and the other three groupers' genomes. We found that the four species had the same karyotype (2n = 48) with a high level of genomic collinearity revealed by the results of chromosome syntenic comparisons between them (Figure 5).

Furthermore, a total of 1224 expanded gene families and 1977 contracted gene families were identified in the *C. sonnerati* genome in comparison to the 15 closet species. Additionally, 112 positive selection genes were found in the *C. sonnerati* genome. Next, there were 18 KEGG pathways and 33 GO terms significantly enriched from the expanded gene families (Table S4). Furthermore, there were 6 KEGG pathways and 22 GO terms significantly enriched from the contracted gene families. The expanded genes families were highly enriched in the sensory system, suggesting that these genes might play an important role in sensory organs development, such as the skin of *C. sonnerati* (Supplementary Figure S2).

3.5 Transcriptome data analysis

We found there were 8,108 tissue-specific expression genes in the*C. sonnerati* genome, based on the detailed analysis of transcriptome data from 11 tissues of *C. sonnerati* (Figure 6). Interestingly, the tissue-specific expression genes were highly enriched in the brain of *C. sonnerati* , which was different from the results of the expanded gene families enrichment analysis in the*C. sonnerati* genome. These tissue-specific expression genes were annotated to the KEGG (Kyoto Encyclopaedia of Genes and Genomes) pathway database for functional analysis. The results showed that these genes mainly consisted of complement and coagulation cascades, DNA replication, synaptic vesicle cycle, long-term potentiation, and other glycan degradation. (Supplementary Figure S3).

4. Conclusion

In this study, we presented the first chromosome-level genome assembly of *C. sonnerati* by combining PacBio long-read sequencing, BGI short-reads sequencing, and Hi-C sequencing technologies. The genome results supplied the first genome from the genus *Cephalopholis* . The genome size was about 1043.66 Mb with an N50 length of 2.49 Mb. In addition, we used Hi-C sequencing technology to scaffold 795 contigs into 24 chromosomes for genome comparison and evolutionary studies between serranid genomes (Kasahara et al., 2007). A total of 26,130 protein-coding genes were predicted in the *C. sonnerati* genome and 24,629 genes (94.26%) were functionally annotated. Interestingly, the enrichment analyses of the expanded gene families suggested a highly enrichment in the sensory system, while the results of the tissue-specific expression genes suggested a highly enrichment in the brain of *C. sonnerati* . These genome resources supply an important reference genome for studies on the genes that influence the sensory system, evolutionary adaption, genetic diversity, and brain development in *C. sonnerati* . Meanwhile, the obtained genome will greatly improve our understanding of the genetic diversity of serranids and promote the development of comparative evolutionary research.

Acknowledgements

We thank Frasergen Bioinformatics Co., Ltd. (Wuhan, China) for completing the bioinformatics analysis of TADs. This work was supported by the Fund of the National Natural Science Foundation of China (41966006 to Z.Z.X.), Science and Technology Research Project of Jiangxi Provincial Department of Education (GJJ170965 to Z.Z.X.), Key Program of Jiangxi Province Science and Technology Plan (20192BBFL60016 to X.G.Z.), Guangdong Provincial Special Fund for Modern Agriculture Industry Technology Innovation Teams (2019KJ143 to Y.Z.), Science and Technology Planning Project of Huizhou (20170214023102296 to Y.Z.).

Author contributions

Z.Z.X., X.G.Z., H.R.L. and Y.Z. designed the research. C.P. and D.D.W. extracted the DNA/RNA and performed the genome sequencing. Q.W collected the sequencing samples. Z.Z.X. wrote the manuscript. D.D.W and S.S.L. analyzed the data.

Conflicts of interests

The authors declare that they have no competing interests.

Data Accessibility Statement

The *C. sonnerati* genome that support the findings of this study have been deposited in the NCBI Sequence Read Archive under the BioProject number PRJNA699118, all raw sequencing data, including BGI and PacBio Sequel II reads, are also deposited under the same BioProject number.

References

Altschul, S. F. et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* , 25, 3389–3402.

Ashburner, M. et al. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics* , 25, 25–29.

Bao, Z. R. & Eddy, S. R. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research* , 12, 1269–1276.

Belaghzal, H., Dekker, J., & Gibcus, J. H. (2017). Hi-C 2.0: An optimized hi-c procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods* , 123, 56-65.

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

Boeckmann, B. et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* , 31, 365–370.

Burge, C. & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of molecular biology* , 268, 78–94.

Burton J N, Adey A, Patwardhan R P, et al. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions[J]. *Nature biotechnology* , 31(12): 1119.

Camacho, C. et al. (2009). BLAST plus: architecture and applications.*BMC Bioinformatics* , 10.

Cantarel, B. L. et al. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* earch, 18, 188–196.

Conesa, A. & Gotz, S. (2008). Blast2GO: a comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics* , 1–12.

Durand, N.C., Shamim, M.S., Machol, I., Rao S.S.P., Huntley, M.H., Lander, E.S., and Aiden E.L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* , 3, 95-98.

Edgar R C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput[J]. *Nucleic acids research* , 32(5): 1792-1797.

Emms D M, Kelly S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy[J]. *Genome biology* , 16 (157):157.

Finn, R. D. et al. (2008). The Pfam protein families database.*Nucleic Acids Research* , 36, D281– D288.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., . . . Fitzgerald, S. (2014). Ensembl 2014. *Nucleic Acids Research* , 42 (Database issue), D749-D755.

Galal-Khallaf, A., Osman, A. G. M., El-Ganainy, A., Farrag, M. M., Mohammed-Abdallah, E., & Moustafa, M. A., et al. (2018). Mitochondrial genetic markers for authentication of major red sea grouper species (Perciformes: Serranidae) in egypt: a tool for enhancing fisheries management and species conservation. *Gene* .

Gaither, M. R., Bowen, B. W., Bordenave, T. R., Rocha, L. A., Newman, S. J., & Gomez, J. A., et al. (2011). Phylogeography of the reef fish*Cephalopholis argus* (Epinephelidae) indicates pleistocene isolation across the indo-pacific barrier with contemporary overlap in the coral triangle. *Bmc Evolutionary Biology,* 11(1), 189-189.

Ge H, Lin K, Shen M, Wu S, Wang Y, Zhang Z, Wang Z, Zhang Y, Huang Z, Zhou C, Lin Q, Wu J, Liu L, Hu J, Huang Z, Zheng L. (2019). De novo assembly of a chromosome-level reference genome of red-spotted grouper (*Epinephelus akaara*) using nanopore sequencing and Hi-C.*Molecular ecology resources* , 19 (6).

Han MV, Thomas GW, Lugo-Martinez J, Hahn MW. (2013). Estimating gene

gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular Biology and Evolution* , 30(8):1987–1997.

Jones, P. et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240.

Jurka, J. (2000). Repbase Update -a database and an electronic journal of repetitive elements. *Trends Genetics* , 16, 418–420.

Jurka, J. et al. (2005). Repbase update, a database of eukaryotic repetitive elements. Cytogenet & Genome Research, 110, 462–467.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research* , 40, D109–D114. Return to ref 41 in article

Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Wei, Q., Ahsan, B., . . . Kasai, Y. (2007). The medaka draft genome and insights into vertebrate genome evolution. *Nature* , 447(7145), 714-719.

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., Li, Z., Chen, Y., Mu, D. and Fan, W. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in *de novo* genome projects. arXiv preprint arXiv:1308.2012.

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research* , 25(5), 955-964.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of

fold change and dispersion for RNA-seq data with DESeq2. *Genome*

Biology , 15(12), 550. https://doi.org/10.1186/s1305 9-014-0550-8

Liu., Shi, Y., Yuan, J., Hu, X., & Wei, F. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *Quantitative Biology* , 35 (s 1–3), 62-67.

Meyer, A. L. (2008). An ecological comparison of *Cephalopholis argus* between native and introduced populations. PhD Thesis, University of Hawaii. Available at http://www.fpir.noaa.gov/Library/HCD/Master%20dissertation%205-31-08.pdf

Mitchell, A. et al. (2015). The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research* , 43, D213–D221.

Mistry, J., Bateman, A. & Finn, R. D. (2007). Predicting active site residue annotations in the Pfam database. *BMC Bioinformatics* , 8, 298. Return

Morris, A. V., Roberts, C. M., & Hawkins, J. P. (2000). The threatened status of groupers (epinephelinae). *Biodiversity & Conservation* , 9 (7), 919-942.

Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics* , 25, 1335–1337.

Rhoads, A., & Au, K. F. (2015). Pacbio sequencing and its applications. *Genomics, Proteomics & Bioinformatics, 13* (5), 278-289.

Roberts, H. C. M. (1994). The growth of coastal tourism in the red sea: present and future effects on coral reefs. Ambio, 23(8), 503-508.

Price, A. L., Jones, N. C. & Pevzner, P. A. (2005). De novo identification of repeat families in large genomes.*Bioinformatics* , 21, I351–I358.

Shpigel, M., & Fishelson, L. (2010). Territoriality and associated behaviour in three species of the genus *Cephalopholis* (Pisces, Serranidae) in the gulf of aqaba, red sea. *Journal of Fish Biology,* 38(6).

Shpigel, M. (1985). Aspects of the biology and ecology of the Red Sea groupers of the genus *Cephalopholis* (Serranidae, Teleostei). PhD Dissertation, Tel Aviv University (in Hebrew, summary in English).

Shpigel, M. & Fishelson, L. (1989a). Food habits and prey selection of three species of groupers from the genus *Cephalopholis*(Serranidae, Teleostei). *Environmental Biology of Fishes*24,67-73.

Shpigel, M. & Fishelson, L. (1989b). Habitat partitioning between species of the genus *Cephalopholis* (Pisces, Serranidae) across the fringing reef of the Gulf of Aqaba (Red Sea). Marine Ecology Progress Series 58, 17–22.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* , 31(19), 3210-3212.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post analysis of large phylogenies. *Bioinformatics* , 30, 1312-1313.

Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res* , 32, W309-W312.

Tarailo-Graovac, M. & Chen, N. (2009). Using Repeat Masker to identify repetitive elements in genomic sequences. *Curr. Protoc* . Bioinformatics Chapter 4, Unit 4.10.

Walker, B. J. et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *Plos One* , 9, e112963.

Wu, T.D. and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* , 21(9), pp.1859-1875.

Xiao, C. et al. (2017). MECAT2: fast mapping, error correction, and de novo assembly for single-moecule sequencing reads. *Nature methods* , 14, 1072.

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

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood.

Molecular Biology and Evolution , 24(8), 1586–1591. https://doi. org/10.1093/molbe v/msm088

Yang., Liu, D., Liu, F., Wu, J., Zou, J., Xiao, X., Zhu, B. (2013). HTQC: a fast quality control toolkit for Illumina sequencing data.*BMC Bioinformatics* , 14(1), 1-4.

Ze-Gang, W., & Shao-Wu, Z.. (2018). Npbss: a new pacbio sequencing simulator for generating the continuous long reads with an empirical model. *Bmc Bioinformatics,* 19 (1), 177.

Zhang, Xuan, Qu, Meng, Zhang, & Xiang, et al. (2013). A Comprehensive Description and Evolutionary Analysis of 22 Grouper (Perciformes, Epinephelidae) Mitochondrial Genomes with Emphasis on Two Novel Genome Organizations. (Doctoral dissertation, PUBLIC LIBRARY SCIENCE).

Zhou Q, Gao H, Zhang Y, Fan G, Xu H, Zhai J, Xu W, Chen Z, Zhang H, Liu S, Niu Y, Li W, Li W, Lin H, Chen S. (2019). A chromosome-level genome assembly of the giant grouper (*Epinephelus lanceolatus*) provides insights into its innate immunity and rapid growth. *Molecular ecology*

resources .

Zhou, Q., Guo, X., Huang, Y., Gao, H., & Chen, S. (2020). *De novo* sequencing and chromosomal-scale genome assembly of leopard coral grouper, *Plectropomus leopardus* . *Molecular Ecology Resources* .

Table 1. Sequencing data for the *C. sonnerati* genome assembly.

Sequencing libraries	Insert size (bp)	Polymerase reads (Gb)	Subreads (Gb)	Mean read length (bp)	Sequence coverage (X)
BGI reads	350	59.70	57.12	148.81	54.70
Pacbio reads	400,00	152.29	152.21	231,09	145.80
Hi-C reads	350	124.03	118.96	148.36	113.94
Total	40,700	336.02	328.29	23,406.17	314.44

Table 2. Statistics of the *C. sonnerati* genome assembly.

Seq type	Total number	Total length (bp)	N50 (bp)	N90 (bp)	Max length(bp)
scaffold	196	1,044,027,303	43,997,100	36,798,269	50,852,404
contig	939	1,043,655,803	2,482,587	683,704	12,435,001

Table 3. Statistics of the repetitive sequences in the *C. sonnerati* genome.

Identification method	Repeat size	% of genome
Trf	43,259,813	4.14
Repeatmasker	155,219,394	14.87
Proteinmask	33,956,519	3.25
De novo	451,894,841	43.28
Total	526,923,565	50.47

Biological classification	Combined TEs Length (bp) % in genome	Combined TEs Length (bp) % in genome
DNA	259,138,824	24.82
LINE	143,448,649	13.74
SINE	20,458,386	1.96
LTR	70,157,749	6.72
Other	11,083	0.00
Unknown	107,454,043	10.29
Total TE	493,109,387	47.23

Table 4. Statistics of gene predictions in the *C. sonnerati*genome.

Gene set	Gene set	Number	Average gene length (bp)	Average CDS length (bp)	Average exon num
De novo	AUGUSTUS	28,361	18,524.94	1458.67	8.34
	Genscan	32,602	23,337.65	1527.94	8.71
Homolog	*O.niloticus*	46,273	11,823.46	1,146.72	5.76
	E.lanceolatus	43,428	12,386.84	1,166.35	6.13
	L.calcarifer	43,682	12,522.15	1,163.49	6.11
	G.acuticeps	42,462	11,486.95	1,128.77	5.79
	C.lumpus	42,018	12,148.39	1,142.28	6.02
	E.akaara	41,589	13,846.16	1,206.10	6.57
	P.georgianus	42,051	12,681.74	1,182.00	5.98
	P.leopardus	53,773	10170.53	904.85	4.78
trans.orf/ISOseq	trans.orf/ISOseq	36,352	20,686.78	1,196.28	9.54
MAKER	MAKER	26,130	20,599.55	1,585.97	9.58

Figure Legends

Figure 1. A picture of *C. sonnerati*

Figure 2. The genome survey of *C. sonnerati* using 17-mer analysis. The peaks of heterozygous, homozygous and repeated 17-mers are highlighted in the plot.

Figure 3. Comparisons of the distribution of gene, CDS, exon and intron length for protein-coding genes between the genomes of*C. sonnerati* and other teleosts.

Figure 4. Divergence time tree constructed using 678 single copy orthologues among *C. sonnerati* and other closely fish species. The estimated divergent time is shown at the branches of the phylogenetic tree, and the confidence intervals are depicted in parentheses.

Figure 5. The whole-genome sequence alignment between*C. sonnerati* and other three groupers.

Figure 6. Heatmap of tissue-specific expression levels from 11 tissues. Genes with a tissue specificity score 1 were considered as showing tissue specificity.

Hosted file

`figure.docx` available at https://authorea.com/users/731293/articles/710407-chromosome-level-genome-assembly-and-transcriptome-of-the-tomato-hind-cephalopholis-sonnerati-serranidae-perciformes