De novo sequencing and chromosomal-scale genome assembly of leopard coral grouper, Plectropomus leopardus

QIAN ZHOU[1], Xinyu Guo[2], Yang Huang[3], Haoyang Gao[2], Hao Xu[1], Shanshan Liu[4], Weiwei Zheng[1], Tianshi Zhang[4], Changxu Tian[3], Chunhua Zhu[3], Haoran Lin[5], and Songlin Chen[6]

[1]Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences
[2]BGI-Qingdao, BGI-Shenzhen
[3]Guangdong Ocean University
[4]Affiliation not available
[5]Sun Yat-Sen University
[6]Yellow Sea Fisheries Research Institute

April 28, 2020

Abstract

The leopard coral grouper, Plectropomus leopardus, belonging to genus Plectropomus, family Epinephelinae, is a carnivorous coral reef fish widely distributing in the tropical and subtropical water of Indo-Pacific Oceans. Due to its appealing body appearance and delicious taste, P. leopardus has become a popular commercial fish for aquaculture in many countries. However, the lack of genomic and molecular resources for P. leopardus hinders its biological studies and genomic breeding programs. Here we report the de novo sequencing and assembly of P. leopardus genome using 10× Genomics and high-throughput chromosome conformation capture (Hi-C) technologies. Using 127.36 Gb 10× Genomics we generated a 902.90 Mb genome assembly with a contig and scaffold N50 of 31.8 Kb and 33.47 Mb, respectively. The scaffolds were clustered and oriented into 24 pseudo-chromosomes with 13.39 Gb valid Hi-C data. BUSCO analysis showed that 95.3% of the conserved single-copy genes were retrieved, indicating a good entirety of the assembly. We predicted 23,234 protein-coding genes, among which 96.5% were functional annotated. The P. leopardus genome provides a valuable genomic resource for genetics, evolutionary and biological studies of this species. Particularly, it is expected to benefit the development of genomic breeding programs in the farming industry.

Abstract

The leopard coral grouper, *Plectropomus leopardus* , belonging to genus Plectropomus, family Epinephelinae, is a carnivorous coral reef fish widely distributing in the tropical and subtropical water of Indo-Pacific Oceans. Due to its appealing body appearance and delicious taste, *P. leopardus* has become a popular commercial fish for aquaculture in many countries. However, the lack of genomic and molecular resource for *P. leopardus* hinders its biological studies and genomic breeding programs. Here we report the *de novo*sequencing and assembly of *P. leopardus* genome using a combination of 10× Genomics, high-throughput chromosome conformation capture (Hi-C) and PacBio long read sequencing technologies. The genome assembly has a total length of 881.55 Mb with a scaffold N50 of 34.15 Mb, consisting of 24 pseudo-chromosomes scaffolds. BUSCO analysis showed that 97.2% of the conserved single-copy genes were retrieved, indicating a good entirety of the assembly. We predicted 25,248 protein-coding genes, among which 96.5% were functional annotated. Comparative genomic analyses revealed that gene family expansions in*P. leopardus* were associated with immune related pathways. In addition, we identified 5,178,453 SNPs based on genome resequencing of 54

individuals. The *P. leopardus* genome and variation data provide valuable genomic resource for genetics, evolutionary and biological studies of the grouper species. Particularly, it is expected to benefit the development of genomic breeding programs in the farming industry.

Keywords : leopard coral grouper, *Plectropomus leopardus* , genome sequencing, chromosomal assembly, genome annotation

Introduction

Groupers (Perciformes, Epinephelinae), acting as the main carnivorous predators, are important biological members in the coral reef ecosystem. The abundance and variety of the groupers have great impact on this fairly complex ecosystem. The family Epinephelidae is comprised of approximately 165 species in 16 genus, classified based on 12S, 16S and histone H3 gene sequences (Craig & Hastings, 2007). However, the taxonomy of these species is still in discussion with a previous classification of family Serranidae (Saad, 2019). Therefore, more genomic evidence are needed to clarify the taxonomy of groupers. However, to date, only two genome sequences of grouper species, the giant grouper (*Epinephelus lanceolatus*) (Zhou et al., 2019) and the red-spotted grouper (*Epinephelus akaara*) (Ge et al., 2019), are available, which significantly hinders the taxonomical, evolutionary and biological studies of the groupers.

The leopard coral grouper, *Plectropomus leopardus* , also namely coral trout or spotted coral grouper, is a representative fish in genus Plectropomus. It naturally inhabits in tropical or subtropical waters of the Indo-Pacific Oceans from southern Japan to Australia and eastwards to the Caroline Islands, Fiji and Tonga (Froese & Pauly, 2019). Like most other groupers, *P. leopardus* is protogynous hermaphroditism, starting out of females and sex-reverse to male later in life (Ferreira, 1995). They are also characterized by complex social structures and a variety of body colour, such as bright red and brown.

Due to low-fat and high-protein flesh, impressive taste and beautiful skin colour, *P. leopardus* has recently become an important commercial species worldwide with a high trading price (Fabinyi, 2012). The increasing demands trigger a rapid development of the aquaculture of*P. leopardus* in many Asian countries and regions, which requires advanced aquaculture technology, such as genomic breeding and metabolic control.

Previous studies in *P. leopardus* mainly focused on species classification (Harrison et al., 2014; Herwerden et al., 2006), reproductive biology (Zeller, 1998), physiological stress responses (Frisch & Anderson, 2005) and behaviour biology (Leis & Carsonewart, 1999; Light & Jones, 1997). Genomic and genetic studies of this species were reported in the development of microsatellite markers (Zhang et al., 2010), the transcriptomic comparison in two colour morphs (Wang et al., 2015), and the muscle metabolic mechanism revealed by gene expression and metabolome analyses (Mekuchi et al., 2017). The insufficient exploitation of genomic resource largely limits the genetic study, conservation and genomic breeding of this species.

Here we report a chromosomal-scale genome assembly and annotation of*P. leopardus* , which was generated using 10× Genomic and Hi-C sequencing technology. The well-annotated genome and the massive sequencing data of leopard coral grouper will provide a crucial resource for genomic, biological and ecological studies of this species, and will efficiently promote its genomic breeding program in aquaculture. In addition, this genome will facilitate future evolutionary, phylogenetic and comparative studies, as well as resource conservation within family Epinephelinae.

Materials and methods

2.1 Sampling and sequencing

Genomic DNA was extracted using a QIAamp DNA purification kit (Qiagen, Germany) according to the manufacturer's instruction. The integrity and quality of the extracted DNA was evaluated using 1% gel electrophoresis. The DNA concentration was assessed using a Pultton DNA/Protein Analyzer (Plextech,

USA). DNA with a total amount [?] 20 μg, 1.8 < $OD_{260/280}$ < 2.0 and a concentration> 12.5 ng/μl were used to construct the sequencing libraries.

A 10× Genomics linked-read library was constructed using the standard protocol (10× Genomics, San Francisco, USA). Raw reads were produced using BGISEQ-500 platform (BGI, Shenzhen, China), with read lengths of 2×100 bp. The raw reads were then filtered with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and QC-Chain (Zhou et al., 2013). The duplicated reads, the adaptor-contaminated reads and the reads having a quality value lower than 20 (representing 1% error rate) were filtered.

To obtain a chromosome-scale genome assembly, we constructed Hi-C library for sequencing. Genomic DNA in blood samples was fixed with formaldehyde in a concentration of 1% and the fixation was terminated using 0.2 M glycine. A Hi-C library was prepared following the Hi-C library protocol (Gong et al., 2018) and then sequenced using a BGISEQ-500 sequencing platform (BGI, Shenzhen, China).

For long-read sequencing, we constructed a SMRTbell library with a fragment size of 20 Kb using the SMRTBell template preparation kit 1.0 (PacBio, USA) following the manufacturer's protocol. The library was sequenced with a PacBio Sequel system, and data from one SMAT cell were generated.

2.2 Tissue collection and RNA sequencing

To facilitate the prediction of protein-coding genes, we extracted total RNA from six tissues of the leopard coral grouper, including gonad, liver, skin, spleen, muscle and fin, using TRIzol reagent (Invitrogen, CA, USA). RNA integrity and quantity was evaluated using the Agilent 2100 Bioanalyzer (Agilent, USA). The six RNA-Seq libraries, which were prepared using the NEBNext Ultra RNA Library Prep Kit (Illumina, USA) following the manufacturer's protocol, was sequenced on a BGISEQ-500 sequencing platform (BGI, Shenzhen, China), producing 70.36 Gb raw data. The quality control of the raw reads were performed with a QC pipeline for RNA-Seq data, RNA-QC-Chain (Zhou et al., 2018), and 69.34 Gb clean RNA-Seq data were remained for further analyses (Table 1).

2.3 Genome assembly and quality assessment

We estimated the genome size and heterozygosity of the *P. leopardus* genome by k-mer analysis, using the quality-filtered reads. The k-mer counts frequencies was computed with Jellyfish (v2.2.10) (Marcais & Kingsford, 2011) using k = 17 and a maximum kmer count of 10,000. The k-mer distribution was measured and plotted using GenomeScope (Vurture et al., 2017). The genome size was calculated with the formula G = $N_{17\text{-mer}}$ /$D_{17\text{-mer}}$, where the $N_{17\text{-mer}}$ is the total number of 17-mers, and $D_{17\text{-mer}}$ denotes the peak frequency of 17-mers.

We *de novo* assembled the 10× Genomics short reads into contigs and scaffolds using Supernova (v1.2) (Weisenfeld et al., 2017) . Gaps in the initial assembly were filled with Gapcloser (v1.12) (Luo et al., 2012) with the parameters of "avg_ins=364, max_ins=500 and min_ins=260". The draft assembly was then anchored and oriented into a chromosomal-scale assembly using the Hi-C scaffolding approach. Firstly, the raw Hi-C reads were filtered with HiC-Pro (v2.8.0) (Servant et al., 2015). Then 3d-dna (v170123) (Dudchenko et al., 2017) with parameters of "-m haploid -s 0 -c 24" was used to anchor the primary contigs and scaffolds into chromosomes. The inter / intra-chromosomal contact maps were built and visualized with Juicebox (Durand et al., 2016).

To further improve the integrity and accuracy of the genome assembly, we employed TGS-GapCloser, which uses low depth ([?] 10x) single molecule sequencing long reads without any error correction to close gaps in the draft assembly (Xu et al., 2019). The long sequences were split into three groups, including total reads (with options –min_idy 0.2, –min_match 200 –r_round 1), reads with length [?] 20 kb (with options –min_idy 0 –min_match 0 –r_round 3) and reads with length in 2-20kb (with options –min_idy 0 –min_match 0 –r_round 3), and each group were used to fill the corresponding aligned gaps.

The completeness of the genome assembly were assessed by Benchmarking Universal Single-Copy Orthologs (BUSCO) (Waterhouse et al., 2017) and GC content analyses. The single copy orthologues of actinopterygii_-

obd9 (BUSCO, v2.0) were searched against the assembled genome using BUSCO tool. The GC content and average sequencing depth across the genome were also measured with 10 Kb non-overlapping sliding windows and the windows harboring more than 50% N's were filtered. No external contamination was found in the genome.

2.4 Repetitive sequence annotation

We annotated the repetitive sequences in the *P. leopardus* genome with both homology searching in known repeat database and *de novo*predictions. Known repeats were identified using RepeatMasker (v3.3.0) (Taraio-Graovac & Chen, 2009) with the RepBase TE library (v14.06) (Bao et al., 2015). RepeatProteinMask (v3.2.2) implemented in RepeatMasker was used to detect the TE-relevant proteins. Novel repeats were predicted using RepeatModeler (http://www.repeatmasker.org) based on the*de novo* repeat library constructed with LTR_Finder (Xu & Wang, 2007) and RepeatScout (Price et al., 2005). In addition, we used Tandem Repeat Finder (TRF, v4.09) (Benson, 1999) to identify the tandem repeats in the genome with parameters of "Match=2, Mismatch=7, Delta=7, PM=80, PI=10, Minscore=50, and MaxPerid=2000".

2.5 Gene prediction and annotation

Based on the repeat masked genome, we employed *de novo* , homology-based and transcriptome-assisted predictions to detect the protein-coding genes. *De novo* gene prediction was performed using Augustus (v2.7) (Stanke et al., 2006) with the *Danio rerio*training set and default settings. For homology-based prediction, protein sequences of *Danio rerio, Takifugu rubripes, Gasterosteus aculeatus, Epinephelus lanceolatus, Epinephelus akaara, Oryzias latipes*and *Cynoglossus semilaevis* were downloaded from NCBI database and aligned to the *P. leopardus* genome using tBLASTn (E-value[?]1e-5). The homologous genome sequences were then aligned against the matching proteins using GeneWise (v2.4.0) (Doerks et al., 2002) for accurate spliced alignments. Transcriptomic data were generated from six RNA-Seq libraries constructed with six tissues, including gonad, liver, skin, spleen, muscle and fin, respectively. A total of 69.34 Gb clean data were aligned to the assembled genome sequences using HISAT2 (v2.0.10) (Pertea et al., 2016) and the putative transcript structures were detected using StringTie (v2.1.1) (Pertea et al., 2016). The candidate protein-coding regions within transcript sequences were then predicted with TransDecoder (v5.5.0) (https://github.com/TransDecoder/TransDecoder/). Finally, genes predicted from the above methods were merged into a consensus gene set using Glean (Elsik et al., 2007).

2.6 Phylogenetic analysis and divergent time estimation

To define the phylogenetic tree, we identified the orthologous gene families by comparing the protein and cDNA sequences among the *P. leopardus* and nine teleosts, including *Takifugu rubripes, Gasterosteus aculeatus, Oreochromis niloticus, Oryzias latipes, Danio rerio, Gadus morhua, Lepisosteus oculatus, Epinephelus lancelatus* and*Epinephelus akaara.* TreeFam (v4.0) (Li et al., 2006) was used to define the orthologous and paralogous relationships among all the organisms. Using the single-copy orthologous genes, a phylogenetic tree was generated with Bayes model using PhyML (v3.0) (Guindon et al., 2010) with 500 bootstrap replications. The MCMCtree program implemented in the PAML package (Yang, 2007) was used to predict the divergence times. The divergent time of *D. rerio* and *T. rubripes* , *G. aculeatus* and *T. rubripes* , and *O. latipes* and*G.morhua* were used as calibration time, which was downloaded from the TimeTree database (http://www.timetree.org/).

Comparative genomic analyses

We compared the genome assembly of *P. leopardus* with the published genomes of grouper species, including *E. lanceolatus*and *E. akaara* . Firstly, to reveal their collinearity relationships, we aligned the chromosomes of the *P. leopardus* to that of the other Epinephelus species using the LASTZ tool (v1.02.00) with default options (Harris, 2007). The chromosomal collinearity were constructed with the mapped regions with lengths >2 Kb for visualizations using Circos (v0.69) (Krzywinski et al., 2009). Secondly, we identified the orthologous groups among these three species, using all-to-all Blast (E-value [?]1e-5, identity [?] 80%) and identified the expanded and contracted gene families using CAFE (De Bie et al., 2006). The enrichment analyses

based on GO and KEGG annotations were performed to identify functional implications of the expanded and contracted genes (Fisher's exact test, adjusted p -value < 0.05).

2.8. Genome-resequencing and SNP calling

For genome resequencing, we sampled a total of 54 individuals of*P. leopardus* from two farming factories in Hainan Province of China. Genomic DNA were extracted from fin tissues of each fish. Pair-ended libraries were constructed according to the standard protocol (Illumina, USA), with an insert size of 300 bp. The sequencing was conducted on the Illumina HiSeq 2000 platform. To avoid the potential influence of low-quality reads in the subsequent analysis, raw reads were checked and filtered using QC-Chain (Zhou et al., 2013), removing reads in the following types: (1) reads containing $> 10\%$ unidentified nucleotides (N's); (2) duplicated reads; (3) reads aligned to adapters, and (4) reads with 10% bases having quality score < 20.

The quality-filtered reads were mapped to the genome assembly using BWA software with default parameters (Li & Durbin, 2009). SNP calling were then performed on a population-scale using GATK (McKenna et al., 2010). The allele frequencies were calculated using VCFtools. We further filtered the SNPs, and only SNPs satisfying the criteria of quality of depth > 2.0, mapping quality > 40, SNP quality > 30, minor allele frequency (MAF) [?] 0.05 and missing rate [?] 0.1 were kept in the final SNP set.

Results and discussion

3.1 Sequencing and genome size estimation

Genomic DNA of a *P. leopardus* individual at around 1 year-old (0.5 kg) **(Figure 1)** , which was provided by Mingbo Aquatic Company (Laizhou, China), was used for genome sequencing. A 10x Genomics linked-read library was constructed and sequenced on BGISEQ-500 platform (BGI, Shenzhen, China), producing a total of 152.61 Gb of raw reads. After quality filtering, we obtained 127.36 Gb clean data (**Table 1**). For the Hi-C sequencing, we obtained 631,842,593 raw read pairs, amounting to 126.37 Gb Hi-C data. Quality control on the Hi-C data finally resulted to 10.60% of the total raw reads as valid Hi-C reads, with two ends mapped to different contigs, which are useful for Hi-C scaffolding (**Table 1**). Single molecule sequencing with PacBio technology generated 1,255,828 reads for 18.05 Gb (**Table 1**).

3.2 Chromosomal-level genome assembly

Based on the clean data, we estimated the genome size to be 871 Mb with the 17-kmer analysis. A dominant peak of the 17 k-mer distribution corresponding to the homozygous peak was demonstrated (**Figure 2a**) and the heterozygosity was estimated to be 0.635%.

The 10x Genomics short reads were *de novo* assembled with Supernova software (v1.2) (Weisenfeld et al., 2017). The contigs and scaffolds in the draft assembly were then anchored and oriented into a chromosomal-scale assembly using the Hi-C scaffolding approach (**Figure 2b**). As a result, we obtained a draft genome assembly of 902.46 megabase (Mb) in length, with a contig of 33.60 Kb. To further improve the completeness and accuracy of the genome assembly, we used PacBio long sequence reads, with a depth of ˜20 x, to close the gaps in the assembly using TAG-Gapcloser. Finally, the total length of the *P. leopardus* genome was 881.55 Mb, with a contig N50 of 855.69 Kb and a scaffold N50 of 34.14 Mb (**Table 2**). The genome assembly had 24 pseudo-chromosomes, with chromosome lengths ranging from 15.72 Mb to 41.71 Mb (**Supplementary TableS1**).

BUSCO analysis showed that the assembly retrieved 97.2% of the conserved single copy orthologue genes, including 94.0% of the complete and 3.2% fragmented genes (**Table 3**). The distribution of GC content and sequencing depth were relatively concentrated, with an average GC content of 39.65%.

3.3 Repetitive sequence annotation

The consensus and non-redundant repetitive sequences were obtained by a combination of known, novel and tandem repeats, generating a total of 298.99 Mb of repetitive sequences, occupying 33.38% of the whole genome assembly (**Table 4**). The repetitive sequences was comprised of DNA transposons in 146.68 Mb

(16.37% of the assembly), long interspersed elements in 48.68 Mb (LINEs; 5.43%), short interspersed nuclear elements in 3.41 Mb (SINEs; 0.38%), long terminal repeats in 38.27 Mb (LTRs; 4.27%), and unknown repeats in 87.71 Mb (9.79%) (**Table 4**).

3.4 Gene prediction and annotation

Combining the results from *de novo* , homology-based and transcriptome-assisted predictions, we successfully generated a non-redundant gene set composing of 25,248 protein-coding genes (**Table 5**). The statistics of the predicted gene models were compared to other teleost species, including *D. rerio* , *O. latipes* and *T. rubripes* , showing similar distribution patterns in mRNA length, CDS length, exon length, intron length and exon number (**Figure 3)** .

We annotated the predicted genes by comparing the protein sequences in several public gene databases, including SwissProt, KEGG and TrEMBL, using BLASTp (E-value[?]1e-5). As a result, 92.3%, 84.6% and 96.4% of the predicted genes got positive hits in SwissProt, KEGG and TrEMBL database, respectively. We also employed InterProScan (v5.0) (Jones et al., 2014) to identify protein domains in multiple protein domain databases of InterPro (ProDom, HAMAP, PANTHER, TIGRFAMs, PRINTS, PIRSF, Gene3D, COILS, PROSITE, Pfam, SMART) (Mitchell et al., 2019) and Gene Ontology (GO), and 88.9% and 70.3% of the predicted genes were annotated in InterPro and GO database, respectively. Finally, a total of 24,364 genes (96.48% out of all predicted genes) were successfully functional annotated in at least one of these databases (**Supplementary Table S2**).

For non-coding genes, 843 tRNAs were identified using tRNAscan-SE (Chan & Lowe, 2019). 1,230 rRNA genes and 324 microRNAs were identified by searching homology against the human rRNA sequence and miRBase (Kozomara & Griffiths-Jones, 2014) database, respectively. Small nuclear RNAs were annotated by the infernal tool (Nawrocki & Eddy, 2013) (http://infernal.janelia.org/) using Rfam database (Kalvari et al., 2018) (**Supplementary Table S3**).

3.5 Phylogeny and divergent time

Using the genomes and genes of 10 selected teleosts, we identified a total of 4,134 were single copy orthologues, based on which a phylogenetic tree was constructed, revealing the phylogenetic relationships of the selected teleost (**Figure 4**). We found that the *P. leopardus* diverged ˜ 57.2 million years ago (mya) from the common ancestor with the linage of*E. lanceolatus* and *E. akaara* . The most closely related species to the grouper linage is G. *aculeatus* , which separated with their common ancestor ˜71.2 mya (**Figure4**).

3.6 Genomic comparison with other groupers

Currently, genome sequences are only available for two groupers (genus Epinephelus), including *E. lanceolatus* and *E.akaara* . To reveal the similarities and differences of the grouper genomes, we conducted functional comparative genomic analyses. The*P. leopardus* and the Epinephelus grouper species have the same karyotype ($2n$ =48), and the chromosome syntenic comparisons showed that they have a high level of genomic collinearity (**Figure 5a**).

Furthermore, gene family evolution was analyzed by constructing orthologous gene families. The numbers of gene families were highly similar in the three groupers, with 18,336, 18,674 and 18,007 in*P. leopardus* , *E. lanceolatus* and *E. akaara* , respectively. A total of 15,497 genes were shared by the three teleost, and 4,427 genes were specific to *P. leopardus* (**Figure 5b**). We also found that *P. leopardus* shared 17,375 genes with*E. lanceolatus* , and 16,929 genes with *E. akaara* , respectively. In *P. leopardus* , a total of 799 gene families were expanded compared to its most recent common ancestor of the Epinephelus linage (**Figure 5c**), among which 126 were significantly expanded ($p < 0.05$). The expanded gene families were significantly enriched in a number of immune systems and immune related pathways (**Figure 5d** , **Supplementary Table S4**), indicating an improved capacity for resistance to diseases in *P. leopardus* . A total of 12 gene families were contracted, however, no enriched KEGG pathway was found.

3.7 Genome re-sequencing for SNP calling

The genome re-sequencing of 54 individuals of *P. leopardus*produced a total of 2,232,057,448 raw sequencing reads. After quality filtering, we obtained a total of 668.90 Gb clean data, with an averagely 12.39 Gb for each fish and a mean sequencing depth of 14.0x (**Supplementary Table S5**). The clean reads were aligned to the assembled genome for each individuals, with 99.63 % of the total reads mapped to the genome. Based on these alignments, we identified a total of 5,178,453 SNPs after quality filtering. The location and effects of these SNPs were also annotated, showing that 132,709 and 57,082 were synonymous and nonsynonymous SNPs, respectively, locating in coding regions (**Table 6**). These SNPs will provide an important genomic resource for the genetic studies, such as population structure analysis, dissection of agronomical traits, identification of selective sweeps, and for genomic selective breeding for superior strains. In the future work, we will use these genomic variations and recorded phenotypes of the corresponding individuals to dissect the genomic associations and to identify key genes playing important roles in the phenotypic differences.

Conclusion

Here we provide a chromosomal-scale genome assembly of the *P. leopardus* by integration of 10x Genomics, Hi-C and PacBio long read sequencing technologies. The genome assembly and annotation supplies the first genome of genus Plectropomus and implement the Epinephelidae genomes, in addition to *E. lanceolatus* and *E. akaara* , thus supplying important genomic data for whole-genome analysis to elucidate the population genetics, evolution and to dissect the genetic diversity underlying their phenotypic traits and adaptions. The genomic variations, together with their functional annotations, will promote accurate genetic analysis and accelerate the genomic breeding programs in aquaculture of the *P. leopardus* .

Acknowledgements

This work was supported by the Fund of Southern Marine Science and Engineering Guangdong Laboratory (Zhanjiang) (ZJW-2019-06), Youth Talent Program Supported by Laboratory for Marine Fisheries Science and Food Production Processes, Pilot National Laboratory for Marine Science and Technology (Qingdao) (2018-MFS-T08), Shandong Superior Variety Project (2016LZGC009),the AoShan Talents Cultivation Program Supported by Pilot National Laboratory for Marine Science and Technology (Qingdao) [2017ASTCP-OS15] and Shandong Taishan Scholar Climbing Project.

Author contributions

S.L.C., H.R.L and C.H.Z. conceived the project. S.L.C. and C.H.Z. managed the project. Q.Z., H.X. and C.X.T. collected the sequencing samples. X.Y.G. extracted the genomic DNA and performed the genome sequencing. Q.Z., G.H and Y. H. analyzed the data. Q.Z. wrote the manuscript. All authors reviewed and approved the final manuscript.

Conflicts of interests

None.

Data Accessibility Statement

The assembled genome has been deposited at DDBJ/ENA/GenBank (VJNF00000000) and NCBI Assembly database with the GenBank accession GCA_008729295.1. Raw sequencing data for *P. leopardus* genome has been deposited in the Sequence Read Archive (SRA) (SRR9330009, SRR11482498). Genome re-sequencing data has been deposited at BioProject PRJNA622646.

References

Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes.*Mobile DNA,* **6** , 11.

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

Chan PP, Lowe TM (2019) tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods in Molecular Biology,* **1962** , 1-14.

Craig MT, Hastings PA (2007) A molecular phylogeny of the groupers of the subfamily Epinephelinae (Serranidae) with a revised classification of the Epinephelini.*Ichthyol Research,* **54** , 1-17.

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

Doerks T, Copley RR, Schultz J, Ponting CP, Bork P (2002) Systematic identification of novel protein domain families associated with nuclear functions. *Genome research,* **12** , 47-56.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP (2017) De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science (New York, N.Y.),***356** , 92.

Durand N, Robinson J, Shamim M, Machol I, Mesirov J, Lander E, Aiden EL (2016) Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom.*Cell systems,* **3** , 99-101.

Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM (2007) Creating a honey bee consensus gene set. *Genome biology,* **8** , R13.

Fabinyi M (2012) Historical, cultural and social perspectives on luxury seafood consumption in China.*Environ Conserv,* **39** , 83-92.

Ferreira BP (1995) Reproduction of the common coral trout Plectropomus leopardus (Serranidae: Epinephelinae) from the central and northern Great Barrier Reef, Australia. *Bull Mar Sci,* **56** , 653-669.

Frisch A, Anderson T (2005) Physiological stress responses of two species of coral trout (Plectropomus leopardus and Plectropomus maculatus). *Comp Biochem Phys A Mol Integrat Phys,* **140** , 317-327.

Froese R, Pauly D (2019) FishBase. World Wide Web electronic publication.

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,* **0** , 1-9.

Gong G, Dan C, Xiao S, Guo W, Huang P, Xiong Y, Wu J, He Y, Zhang J, Li X, Chen N, Gui J-F, Mei J (2018) Chromosomal-level assembly of yellow catfish genome using third-generation DNA sequencing and Hi-C analysis. *GigaScience,***7** .

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology,* **59** , 307-321.

Harris RS (2007) Improved pairwise alignment of genomic DNA. The Pennsylvania State University.

Harrison HB, Feldheim KA, Jones GP, Ma K, Mansour H, Perumal S, Williamson DH, Berumen ML (2014) Validation of microsatellite multiplexes for parentage analysis and species discrimination in two hybridizing species of coral reef fish (Plectropomus spp., Serranidae). *Ecol Evol,* **4** , 2046-2057.

Herwerden LV, Choat JH, Dudgeon CL, Carlos G, Newman SJ, Frisch A, Oppen MV (2006) Contrasting patterns of genetic structure in two species of the coral trout Plectropomus (Serranidae) from east and west Australia: Introgressive hybridisation or ancestral polymorphisms. *Mol Phylogenet Evol,* **41** , 420-435.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics,* **30** , 1236-1240.

Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, Bateman A, Finn RD, Petrov AI (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res,* **46** , D335-d342.

Kozomara A, Griffiths-Jones S (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res,* **42** , D68-d73.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. *Genome research,***19** , 1639-1645.

Leis JM, Carsonewart BM (1999) In situ swimming and settlement behaviour of larvae of an Indo-Pacific coral-reef fish, the coral trout Plectropomus leopardus (Pisces: Serranidae). *Mar Biol,* **134** , 51-64.

Li H, Coghlan A, Ruan J, Coin LJ, Heriche J-K, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, Wong GK-S, Zheng W, Dehal P, Wang J, Durbin R (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic acids research,* **34** , D572-D580.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* **25** , 1754-1760.

Light PR, Jones GP (1997) Habitat preference in newly settled coral trout (Plectropomus leopardus, Serranidae). *Coral Reefs,* **16** , 117-126.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience,* **1** , 18.

Marcais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics,* **27** , 764-770.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research,* **20** , 1297-1303.

Mekuchi M, Sakata K, Yamaguchi T, Koiso M, Kikuchi J (2017) Trans-omics approaches used to characterise fish nutritional biorhythms in leopard coral grouper (Plectropomus leopardus). *Sci Rep,* **7** , 9372.

Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang HY, El-Gebali S, Fraser MI, Gough J, Haft DR, Huang H, Letunic I, Lopez R, Luciani A, Madeira F, Marchler-Bauer A, Mi H, Natale DA, Necci M, Nuka G, Orengo C, Pandurangan AP, Paysan-Lafosse T, Pesseat S, Potter SC, Qureshi MA, Rawlings ND, Redaschi N, Richardson LJ, Rivoire C, Salazar GA, Sangrador-Vegas A, Sigrist CJA, Sillitoe I, Sutton GG, Thanki N, Thomas PD, Tosatto SCE, Yong SY, Finn RD (2019) InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res,* **47** , D351-d360.

Nawrocki EP, Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics,***29** , 2933-2935.

Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc,***11** , 1650.

Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. *Bioinformatics,* **21** , i351-358.

Saad YM (2019) Analysis of 16S mitochondrial ribosomal DNA sequence variations and phylogenetic relations among some Serranidae fishes. *S Afr J Anim Sci,***49** , 80-89.

Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, Barillot E (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing.*Genome biology,* **16** , 259.

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

Taraio-Graovac M, Chen N (2009) Using RepeatMasker to identify repetitive elements in genomic sequences.*Current protocols in bioinformatics* , 4.10.11-14.10.14.

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC (2017) GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics,***33** , 2202-2204.

Wang L, Yu C, Guo L, Lin H, Meng Z (2015) In silico comparative transcriptome analysis of two color morphs of the common coral trout (Plectropomus leopardus). *PloS one,***10** , e0145868.

Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM (2017) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular biology and evolution,* **35** , 543-548.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB (2017) Direct determination of diploid genome sequences. *Genome research,* **27** , 757-767.

Xu M, Guo L, Gu S, Wang O, Zhang R, Fan G, Xu X, Deng L, Liu X (2019) *TGS-GapCloser: fast and accurately passing through the Bermuda in large genome using error-prone third-generation long reads* .

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

Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular biology and evolution,***24** , 1586-1591.

Zeller DC (1998) Spawning aggregations:patterns of movement of the coral trout Plectropomus leopardus (Serranidae) as determined by ultrasonic telemetry. *Mar Ecol Prog,* **162** , 253-263.

Zhang J, Liu H, Yu S (2010) Development and characterization of polymorphic microsatellite loci for a threatened reef fish Plectropomus leopardus. *Conserv Genet Resour,* **2** , 101-103.

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,* **0** , 1-11.

Zhou Q, Su X, Jing G, Chen S, Ning K (2018) RNA-QC-Chain: comprehensive and fast quality control for RNA-Seq data. *BMC Genomics,* **19** , 144.

Zhou Q, Su X, Wang A, Xu J, Ning K (2013) QC-Chain: fast and holistic quality control method for next-generation sequencing data. *PloS one,* **8** , e60234.

Tables

Table 1. Sequencing data for the *P. leopardus* genome assembly.

Sequencing technology	Raw data (Gb)	Clean data (Gb)	Mean read length (bp)
10× Genomics	152.61	127.36	100
Hi-C	126.37	13.39	100
PacBio	180.46	/	14.37k
RNA-Seq	70.36	69.34	100

Table 2. Statistics of the *P. leopardus* genome assembly.

Seq type	Total number	Total length (bp)	N50 (bp)	N90 (bp)	Max length (bp)
scaffold	50,461	881,551,488	34,146,761	8,092	41,705,749
contig	54,036	865,740,848	855,686	7,000	4,608,261

Table 3. BUSCO analysis result of the *P. leopardus* genome.

	Gene number	Percentage
Complete BUSCOs (C)	4,307	94.0
Complete and single-copy BUSCOs (S)	4,167	90.9
Complete and duplicated BUSCOs (D)	140	3.1
Fragmented BUSCOs (F)	147	3.2
Missing BUSCOs (M)	130	2.8
Total BUSCO groups searched	4,584	100

Table 4. Statistics of repetitive sequences in the*P. leopardus* genome.

Identification method	**Repeat size**	**% of genome**
TRF	34,303,179	3.83
RepeatMasker	49,936,126	5.57
ProteinMask	11,710,155	1.31
De novo	280,223,461	31.29
Total	298,989,520	33.38
Biological classification	**Combined TEs Length (bp)**	**Combined TEs % in genome**
DNA	146,681,333	16.37
LINE	48,680,003	5.43
SINE	3,413,395	0.38
LTR	32,129,629	4.27
Other	18,438	0.002
Unknown	87,713,772	9.79
Total	275,295,833	30.74

Table 5. Statistics of gene predictions in the *P. leopardus* genome

	#Gene set	Number of genes	CDS+intron len	CDS len	exon len	intron len	E
Homolog	*Danio rerio*	25,787	23,126.00	1,524.78	181.24	2,913.94	8.
	Takifugu rubripes	21,677	15,087.57	1,424.34	177.75	1,948.16	8.
	Gasterosteus aculeatus	28,700	14,201.45	1,460.43	170.76	1,687.02	8.
	Epinephelus lanceolatus	23,869	14,232.13	1,400.68	169.48	1,766.28	8.
	Cynoglossus semilaevis	33,866	19,789.49	1,521.26	177.79	2,417.60	8.
	Epinephelus akaara	23,142	16,439.61	1,606.08	177.73	1,845.79	9.
	Oryzias latipes	19,672	23,455.32	1,570.11	191.37	3,037.64	8.
De novo	Augustus	24,499	16,070.20	1,435.18	172.04	1,993.26	8.
RNA	Transdecoder	31,207	17,648.02	1,428.34	170.77	1,501.25	8.
Glean	Glean	25,248	15,872.47	1,632.85	181.54	1,781.17	8.

Table 6. Summary of the single nucleotide polymorphisms (SNPs) effects in *P. leopardus*

Type	Number of SNPs
Intergenic	1,740,250
Intron	2,453,403
Upstream	411,242
Downstream	389,553
Splice_site	16,105
Start_lost	82
Codon_change	7
Synonymous_coding	117,662
Non_synonymous_coding	49,633
Stop_gained	438
Stop_lost	78
Total	5,178,453

Figure Legends

Figure 1. A photograph of a red *P. leopardus* (by Qian Zhou).

Figure 2. Genome size estimation and assembly. (a) Graph of$k$ -mer frequency distribution ($k = 17$) generated from 127.36 Gb sequencing data of *P. leopardus* . The highest peak at coverage 66× corresponds to the homozygous peak. The minor peak at coverage 35× corresponds to the heterozygous peak. The minor peak at coverage 138× corresponds to duplications. The illustrations of the lines are marked in the graph. (b) The Hi-C contact map of the *P. leopardus* genome. The color bar shows the contact density from white (low) to red (high).

Figure 3. Comparisons of the predicted gene models between*P. leopardus* genome and other teleost species. (a) mRNA length. (b) CDS length. (c) Exon length. (d) Intron length. (e) Exon number.

Figure 4. Phylogenetic tree constructed using 4134 single copy orthologues among the selected teleost, with 500 bootstraps. The estimated divergent time (Mya, million years ago) and the 95% confidential intervals were labeled at each branch. The red dots indicate the divergent time used for re-calibrations.

Figure 5. Comparative genomic analyses of *P. leopardus* with two grouper species, *E. lanceolatus*and *E. akaara*) . (a) Chromosomal collinearity between*P. leopardus* and *E. lanceolatus*. The colorful bars (Chr1-24) and the green bars (I-IIXIV) indicate each of the 24 chromosomes in *P. leopardus* and *E. lanceolatus* , respectively. (b) Venn diagram of the genes from the three grouper species. (c) Gene family expansion and contraction. (d) The enrichment of KEGG annotations with the expanded gene families in *P. leopardus* ($p < 0.05$).