frontiers in Genetics

DATA REPORT
published: 12 December 2018
doi: 10.3389/fgene.2018.00643

First Draft Genome for Red Sea Bream of Family Sparidae

Ga-Hee Shin[1†], Younhee Shin[1,2†], Myunghee Jung[1], Ji-man Hong[1], Sangmin Lee[1], Sathiyamoorthy Subramaniyam[1], Eun-Soo Noh[3], Eun-Ha Shin[3], Eun-Hee Park[3], Jung Youn Park[3], Young-Ok Kim[3], Kwnag-Min Choi[4], Bo-Hye Nam[3] and Chan-Il Park[4*]*

[1] Research and Development Center, Insilicogen Inc., Yongin-si, South Korea, [2] Department of Biological Sciences, Sungkyunkwan University, Suwon, South Korea, [3] Biotechnology Research Division, National Institute of Fisheries Science, Busan, South Korea, [4] Department of Marine Biology and Aquaculture, College of Marine Science, Gyeongsang National University, Tongyeong, South Korea

Keywords: ***Pagrus major*, genome, PacBio, Sparidae, red sea bream**

OPEN ACCESS

Edited by:
Nguyen Hong Nguyen, University of the Sunshine Coast, Australia

Reviewed by:
Paulino Martínez, University of Santiago de Compostela, Spain
Filippo Biscarini, Italian National Research Council, Italy
James W. Kijas, Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia

***Correspondence:**
Bo-Hye Nam
nambohye@korea.kr
Chan-Il Park
vinus96@hanmail.net

[†] These authors have contributed equally to this work

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

Received: *16 August 2018*
Accepted: *27 November 2018*
Published: *12 December 2018*

Citation:
Shin G-H, Shin Y, Jung M, Hong J, Lee S, Subramaniyam S, Noh E-S, Shin E-H, Park E-H, Park JY, Kim Y-O, Choi K-M, Nam B-H and Park C-I (2018) First Draft Genome for Red Sea Bream of Family Sparidae. Front. Genet. 9:643. doi: 10.3389/fgene.2018.00643

INTRODUCTION

Reference genomes for all organisms on earth are now attainable owing to advances in genome sequencing technologies (Goodwin et al., 2016). Generally, species that contribute considerably to the economy or human welfare are sequenced and are considered more important than others. Furthermore, coastal indigenous people mainly depend on marine species for their food sources, which has resulted in the extinction of several marine species (Cisneros-Montemayor et al., 2016). Of these, an extinction risk assessment of marine fishes, mainly for sea breams (Family: Sparidae), has recently been conducted by way of a global extinction risk assessment from the dataset of the International Union for Conservation of Nature's Red List Process, which mentions that around 25 species are threatened/near-threatened according to their body weight (Comeros-Raynal et al., 2016). Another report clearly showed the benefit of worldwide aquaculture production, which contributed to 47% of total seafood production, and also highlighted the over-fishing of sea breams (FAO, 2018). The Republic of Korea is the fourth largest seafood producer in the world, producing 3.3 million tons in 2015 and exporting seafood worth $1.6 billion in 2016; therefore, aquaculture-associated research is fundamental for Korea. In the present study, the red sea bream (*Pagrus major*), which belongs to the family Sparidae, which comprises 35 genera, 132 species, and 10 subspecies (de la Herran et al., 2001; NCBI, 2018), was assessed. It is widely distributed in the coastal regions of Korea, Japan, China, and Taiwan (Blanco Gonzalez et al., 2015), commonly on rocky substrates, soft sand, and muddy bottoms. Species of this family are hermaphroditic and mature 4 years after birth, surviving for 10 or more years. This group of fishes is an important resource to better understand the genetics of sexual dimorphism. Another major factor affecting this species is microbial infections, which are dominant in the aquaculture industry and account for a considerable decline in aquaculture production (Nam et al., 2016; Sawayama et al., 2017). Few studies have analyzed the molecular markers associated with these problems. Recently, sexual

dimorphism-related genes from the *Sparus aurata* genome have been profiled, including stage-specific expression (Pauletto et al., 2018), and three other studies have assessed molecular markers associated with microbial and environmental toxicity in the red sea bream (Iida et al., 2016; Hano et al., 2017; Sawayama et al., 2017). However, genome-wide molecular marker characterization is needed to conduct genome selection in breeding schemes (López et al., 2014), which is not possible in *P. major*, owing to the absence of a reference genome. To the best of our knowledge, only two draft genomes (*S. aurata* and *Spondyliosoma cantharus*) are available for the entire Sparidae family, which is the largest clade in class Actinopteri (de la Herran et al., 2001), but there is no draft or reference genome sequence for the genus *Pagrus*. Therefore, we constructed a draft genome using contig level assembly, with a size of 829.3 Mb, employing the 90X PacBio sequence alone.

Value of the Data

This draft genome would be considerably useful for detailing the molecular characterization of various breeding-associated problems in species from the family Sparidae as well as other comparative genome mining applications.

MATERIALS AND METHODS

Sample Collection and Genomic DNA Extraction

A single female fish (4.25 kg) was collected on December 2016 from the Jeju Fisheries Research Institute and maintained at 22 ± 0.5°C in aerated seawater (NFRDI-2016-01-2). The abdominal muscle tissues were sampled aseptically and stored in liquid nitrogen for genomic DNA extraction. The complete experimental procedure, from DNA isolation to sequencing, was conducted using DNALink, South Korea (www.dnalink.com), as instructed in the respective product protocols.

Genomic DNA Library Preparation and Sequencing

Highly concentrated genomic DNA (gDNA) (24 μg) from each given sample was prepared using a DNeasy Animal Mini Kit (Qiagen, Hilden, Germany). The complete isolated gDNA was quantified using a ND-1000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA) and Qubit fluorometer. The total gDNA were subjected to other steps i.e., fragmentation with Covaris G-Tube to obtain > 20 KB fragments, filtering of small fragments using 0.45X AMPure®, fragment end repair using ExoVII, ligation of blunt adapters using double standard DNA fragments, attachment of the primer and polymerase to the SMRTbell™ templates (Template Prep Kit 1.0), and the addition of MagBeads. Finally, the impurities were washed out carefully with 1.0X AMPure® and only the double stranded DNA fragments with blunt adapters were subjected to sequencing using C4-chemistry (DNA sequencing Reagent 4.0) in the PacBio (Pacific Biosciences) sequencing platform by capturing a movie for 1 × 240 min of each SMRT cell. Similarly, the isolated gDNAs were also subjected to sequencing library preparation with stranded Illumina paired-end (PE) protocols (Illumina, San Diego, CA, USA). The fragmented libraries were subjected to size selection and sequenced with an Illumina Hiseq 2000 sequencer.

Illumina Pre-process and Genome Size Estimation

Full Illumina DNA sequences were subjected to pre-processing steps, which included adapter trimming, quality trimming (Phred(Q) ≥ 20), and contamination removal. The adapter and quality trims were conducted using Trimmomatic-0.32 functions (Bolger et al., 2014), and the microbial contamination of each sample was removed using CLCMapper v4.2.0 (www.qiagenbioinformatics.com) with an in-house database. Here, the in-house database was constructed from bacterial (ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/prokaryotes.txt), viral (ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/), and marine (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA13694) metagenomes. All the pre-processed sequences from the paired-end library were subjected to genome size estimation using the *k*-mer based method (which was used in the panda genome Li et al., 2009. The *k*-mer frequencies (*k*-mer size = 19) were obtained using the Jellyfish v2.0 method (Marçais and Kingsford, 2011), and the genome size was calculated from the given formulas: Genome Coverage Depth = (*k*-mer Coverage Depth X Average Read Length) / (Average Read Length – *k*-mer size +1) and Genome size = Total Base Number / Genome Coverage Depth. Alternatively, the PacBio sequences were only subjected to error correction using CLCAssemblyCell v4.2.0.

PacBio Error-Correction and *de-novo* Genome Assembly

Complete PacBio sequence reads were processed for error correction (Read Quality ≥ 0.75 and Read Length ≥ 50) with processed Illumina short reads using SMRTAnalysis v2.3 and the error corrected PacBio reads were imported to a diploid-aware hierarchical genome assembler to construct the contigs from the long-sequence PacBio reads, i.e., FALCON (Chin et al., 2016). The assembled contigs were further subjected to sequence polishing using the Quiver consensus method to reduce the base called errors (Chin et al., 2013). Finally, the assembled and polished contigs were assessed to determine genome completeness using BUSCO v3.0 (Simão et al., 2015). The reference BUSCO datasets used were vertebrata_odb9 and actinopterygii_odb9. The quality of the assembly was assessed by short-reads mapping to the draft using CLCMapper v5.0.4.

De novo Repeat Region Prediction and Classification

The repeat regions were predicted using the *de novo* method and classified into repeat subclasses. The *de novo* repeat prediction for *P. major* was conducted using RepeatModeler (www.repeatmasker.org/RepeatModeler/), which includes other methods such as RECON (Bao and Eddy, 2002) (http://eddylab.org/software/recon/), RepeatScout (Price et al., 2005) (https://bix.ucsd.edu/repeatscout/), and TRF (Benson, 1999) (https://tandem.bu.edu/trf/trf.html). The modeled repeats were classified

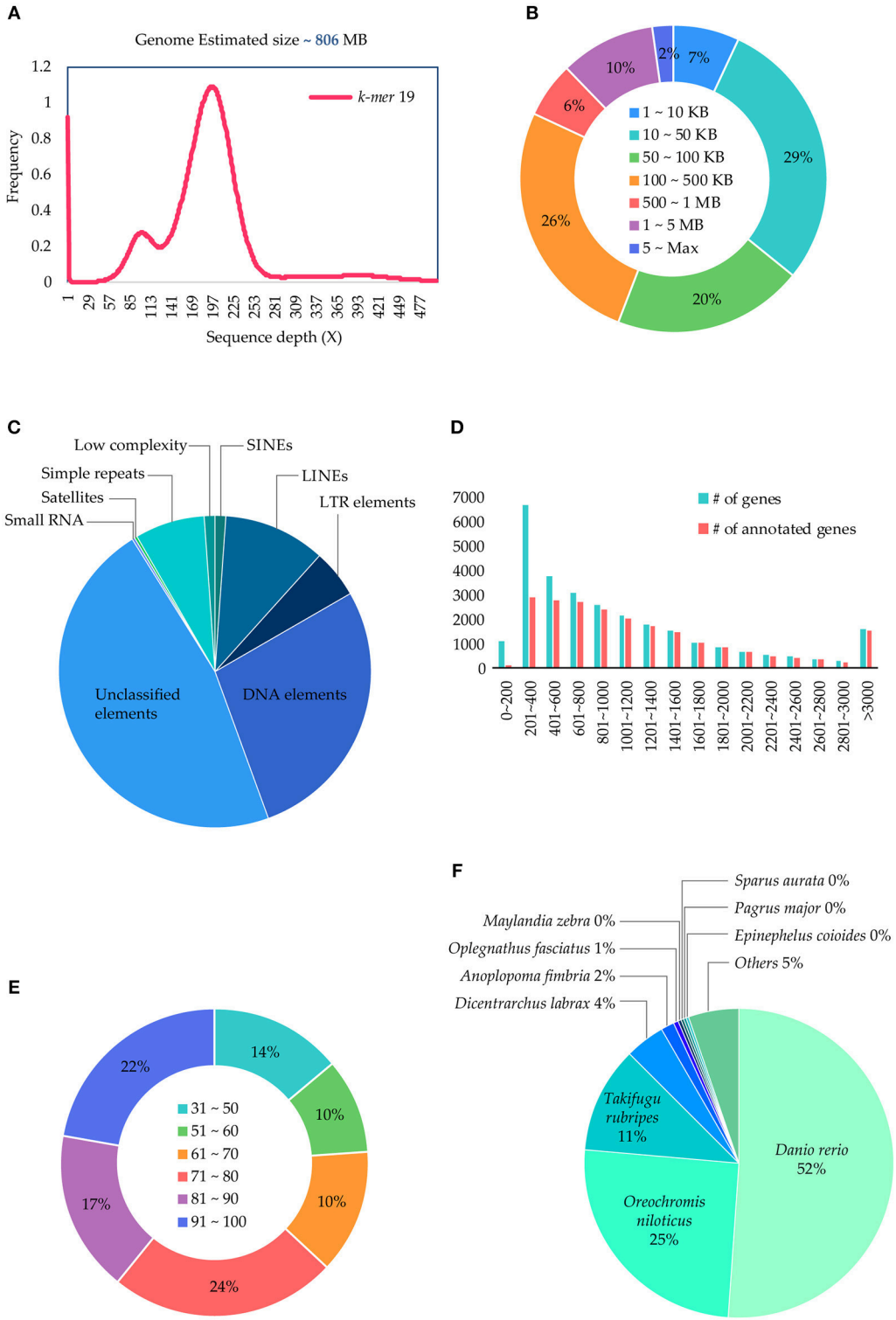

FIGURE 1 | Illustration of genome size, genes, and functional elements of the *Pagrus major* genome. **(A)** *k*-mer based genome size estimation; **(B)** assembled contigs length distributions; **(C)** *de-novo* repeat predictions and sub-class distributions; **(D)** length distribution of the predicted genes with their functional annotation status; **(E)** sequence blast similarity score distribution, and **(F)** species distribution obtained from BLAST.

into their subclasses using the reference Repbase v20.08 database (www.girinst.org/repbase/) (Bao et al., 2015) and these repeats were masked using RepeatMasker v4.0.5 (www.repeatmasker.org) with RMBlastn v2.2.27^{+}.

Gene Prediction and Annotation

The genes from the *P. major* draft were predicted using an in-house gene prediction pipeline, which includes three modules: an evidence-based gene modeler (EVM), an *ab-initio* gene modeler, and a consensus gene modeler. Finally, functional annotation processing was conducted for the consensus genes. Initially, sequenced transcriptomes from two methods [Illumina (186.6 Gb) and IsoSeq (1.2 Gb)] were mapped to the *P. major* repeat masked draft genome using Tophat (Trapnell et al., 2012) and the transcripts/gene structural boundaries were predicted using Cufflink (Trapnell et al., 2012) and PASA (Haas et al., 2003). To train the *ab-initio*, gene modeler and EVM (which includes Exonerate Slater and Birney, 2005, AUGUSTUS Stanke et al., 2006 and GENEID Blanco et al., 2007), as well as several genomes (*Danio rerio, Gasterosteus aculeatus, Tetraodon nigroviridis, Takifugu rubripes, Oryzias latipes, Notothenia coriiceps, Haplochromis burtoni, Stegastes partitus, Sebastes schlegelii, Oplegnathus fasciatus*, and *Homo sapiens*) were used for prediction. Finally, the predicted gene and transcript models from the EVM and *ab-initio* modeler were subjected to the consensus gene modeler (which includes EVidenceModeler, Haas et al., 2008) to produce the final gene and transcript models. Finally, the consensus transcripts were subjected to functional annotation from biological databases (NCBI-NR databases, Swiss-Prot, Gene Ontologies and KEGG pathways) using Blast2GO (Götz et al., 2008).

Preliminary Analysis Report

The *P. major* genome size was estimated as ∼806 Mb (**Figure 1A**) using the *k*-mer method from 190.3 Gb of the short-read sequences (**Table 1**), which were generated using the Illumina sequencer. The 73 Gb long-read sequences, which were generated using the PacBio sequencer, were assembled into 1,657 contigs with a total size of 829.3 Mb and an N50 of 2.8 Mb (**Table 1**), and 92.6% of the paired short-reads were mapped correctly to the assembled contigs, which clearly showed the assembly quality. Particularly, 12% of the contigs were > 1Mb in length (**Figure 1B**) and < 7% of the contigs were < 10 Kb in size (**Figure 1B**). The repeat contents in the genome were 257 Mb (31.1%) bases, which were predicted and classified into their sub-classes (**Figure 1C**). In this genome, 28,343 consensus genes were predicted with an average length of 5,913 bp (**Table 1**, section C) and, among those, 76.2% of the genes obtained annotations from the Uniprot database (**Figure 1E**). Most of the short genes were left unannotated compared to others (**Figure 1D**). Moreover, 52% of the annotated genes obtained annotation from the fish *Danio rerio* (**Figure 1F**). Additionally, BUSCO scores were obtained for the two datasets: 97.8% (2,529/2,586) in vertebrata and 97.1% in actinopterygii (4,447/4,584), which shows the confidence of the completeness of the annotated genes in the assembled genome. Therefore, we propose that this draft version is a near-complete reference genome for *P. major* and, in comparison with 68 other available genome assemblies for the bony fish clade (Percomorphaceae) in the NCBI assembly (lastly accessed: March 2018), this draft is assembled well at the contig level. Moreover, this is the best assembled draft for the genus *Pagrus* and family Sparidae at the contig level and will be good as a base to improve scaffold/chromosomal-level genome assemblies and as a reference for other functional studies.

TABLE 1 | Summary of genome assemblies and gene annotations.

Technology	Illumina	PacBio
A. SEQUENCES		
Raw data in Gb (Coverage)	190.3 (∼240 X)	73.0 (∼90 X)
Pre-processed data in Gb (%)	156.4 (82%)	73.0 (100%)
B. ASSEMBLY		
No of Contigs	1,657	
Total Bases	829,318,935	
Average length	500,494	
Minimum length	153	
Maximum length	12,966,191	
N50	2,896,215	
N (%)	0	
GC (%)	41.23	
C. GENE		
# of genes	28,343 (6.24 exons/ gene)	
Average gene length	5,913 bp	
Average exon length	178 bp	
Repeat elements	31.11%	
Genome coverage (gene region)	20.20%	
D. ANNOTATIONS		
Blast hits	21,605 (76.22%)	
No hits	6,738 (23.77%)	

Deposited Data and Information to the User

The complete sequences, which were used for the genome assemblies and annotations, have been deposited in public data repositories. The DNA libraries used in the current draft genome assembly for *P. major* have been deposited in the NCBI sequence read archive (Project ID: PRJNA480768) and the structural and functional annotation (CDS, gff, repeat regions, and proteins) datasets have been deposited in the figshare repository (doi: 10.6084/m9.figshare.6962867.v1). The format and description of all the deposited datasets are mentioned in the readme file, which have been deposited in the figshare repository.

AUTHOR CONTRIBUTIONS

G-HS, YS, MJ, JH, SL, and SS: genome assembly and annotations. SS and YS: manuscript preparation. E-SN, E-HS, E-HP, JP, Y-OK, and K-MC: sampling and sequencing. B-HN and C-IP: funding and modeled the study.

ACKNOWLEDGMENTS

A grant from the National Institute of Fisheries Science (R2018021) and Collaborative Genome Program (20180430) of the Korea Institute of Marine Science and Technology Promotion (KIMST) funded by the Ministry of Oceans and Fisheries (MOF), Korea supported this research.

REFERENCES

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* 6:11. doi: 10.1186/s13100-015-0041-9

Bao, Z., and Eddy, S. R. (2002). Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res.* 12, 1269–1276. doi: 10.1101/gr.88502

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

Blanco Gonzalez, E., Aritaki, M., Knutsen, H., and Taniguchi, N. (2015). Effects of large-scale releases on the genetic structure of red sea bream (*Pagrus major*, Temminck et Schlegel) populations in Japan. *PLoS ONE* 10:e0125743. doi: 10.1371/journal.pone.0125743

Blanco, E., Parra, G., and Guigó, R. (2007). Using geneid to identify genes. *Curr. Protoc. Bioinform* 18, 4.3.1–4.3.28. doi: 10.1002/0471250953.bi0403s18

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10:563. doi: 10.1038/nmeth.2474

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

Cisneros-Montemayor, A. M., Pauly, D., Weatherdon, L. V., and Ota, Y. (2016). A global estimate of seafood consumption by coastal indigenous peoples. *PLoS ONE* 11:e0166681. doi: 10.1371/journal.pone.0166681

Comeros-Raynal, M. T., Polidoro, B. A., Broatch, J., Mann, B. Q., Gorman, C., Buxton, C. D., et al. (2016). Key predictors of extinction risk in sea breams and porgies (Family: *Sparidae*). *Biol. Conserv.* 202, 88–98. doi: 10.1016/j.biocon.2016.08.027

de la Herran, R., Rejon, C. R., Rejon, M. R., and Garrido-Ramos, M. A. (2001). The molecular phylogeny of the Sparidae (Pisces, Perciformes) based on two satellite DNA families. *Heredity* 87(Pt 6), 691–697.

FAO (2018). *The State of World Fisheries and Aquaculture 2018 - Meeting the Sustainable Development Goals*. Rome: FAO.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333. doi: 10.1038/nrg.2016.49

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., et al. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* 36, 3420–3435. doi: 10.1093/nar/gkn176

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Hannick, L. I., et al. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666. doi: 10.1093/nar/gkg770

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* 9:R7. doi: 10.1186/gb-2008-9-1-r7

Hano, T., Ohkubo, N., and Mochida, K. (2017). A hepatic metabolomics-based diagnostic approach to assess lethal toxicity of dithiocarbamate fungicide polycarbamate in three marine fish species. *Ecotoxicol. Environ. Saf.* 138, 64–70. doi: 10.1016/j.ecoenv.2016.12.019

Iida, M., Fujii, S., Uchida, M., Nakamura, H., Kagami, Y., Agusa, T., et al. (2016). Identification of aryl hydrocarbon receptor signaling pathways altered in TCDD-treated red seabream embryos by transcriptome analysis. *Aquat. Toxicol.* 177, 156–170. doi: 10.1016/j.aquatox.2016.05.014

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., et al. (2009). The sequence and *de novo* assembly of the giant panda genome. *Nature* 463:311. doi: 10.1038/nature08696

López, M. E., Neira, R., and Yáñez, J. M. (2014). Applications in the search for genomic selection signatures in fish. *Front. Genet.* 5:458. doi: 10.3389/fgene.2014.00458

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

Nam, B.-H., Jung, M., Subramaniyam, S., Yoo, S.-,i., Markkandan, K., Moon, J.-Y., et al. (2016). Transcriptome analysis revealed changes of multiple genes involved in haliotis discus hannai innate immunity during *Vibrio parahemolyticus* infection. *PLoS ONE* 11:e0153474. doi: 10.1371/journal.pone.0153474

NCBI (2018). *NCBI Taxonomy Database*. Available online at: https://www.ncbi.nlm.nih.gov/taxonomy/

Pauletto, M., Manousaki, T., Ferraresso, S., Babbucci, M., Tsakogiannis, A., Louro, B., et al. (2018). Genomic analysis of *Sparus aurata* reveals the evolutionary dynamics of sex-biased genes in a sequential hermaphrodite fish. *Commun. Biol.* 1:119. doi: 10.1038/s42003-018-0122-7

Price, A. L., Jones, N. C., and Pevzner, P. A. (2005). *De novo* identification of repeat families in large genomes. *Bioinformatics* 21(Suppl. 1), i351–i358. doi: 10.1093/bioinformatics/bti1018

Sawayama, E., Tanizawa, S., Kitamura, S. I., Nakayama, K., Ohta, K., Ozaki, A., et al. (2017). Identification of quantitative trait loci for resistance to RSIVD in red sea bream (*Pagrus major*). *Mar. Biotechnol.* 19, 601–613. doi: 10.1007/s10126-017-9779-z

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

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

Stanke, M., Schöffmann, O., Morgenstern, B., and Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* 7:62. doi: 10.1186/1471-2105-7-62

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

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