De novo transcriptome assembly and functional annotation of cobia (*Rachycentron canadum*): molecular approach as a base to nutritional studies

David Aciole Barbosa [1 #], Bruno C. Araújo [1 #], Giovana S. Branco [2]; Alexandre W. S. Hilsdorf [1], Luiz R. Nunes [3], Renata G. Moreira [2], Fabiano B. Menegidio [1*]

[1]Núcleo Integrado de Biotecnologia, Universidade de Mogi das Cruzes, Mogi das Cruzes, 08701-970, SP, Brazil

[2]Departamento de Fisiologia, Instituto de Biociências, Universidade de São Paulo, São Paulo, 05508-090, SP, Brazil

[3]Centro de Ciências Naturais e Humanas, Universidade Federal do ABC, São Bernardo do Campo, 09606-045, SP, Brazil

∗ To whom correspondence should be addressed: Fabiano B. Menegidio, PhD, Núcleo Integrado de Biotecnologia, Universidade de Mogi das Cruzes (UMC), Av. Dr. Cândido X. de Almeida e Souza, 200 - Centro Cívico, Mogi das Cruzes - SP, 08780-911, Brazil;

E-mail: fabianomenegidio@umc.br

#. These authors contributed equally to this article

ABSTRACT: Cobia (*Rachycentron canadum*) is a marine teleost species with great productive potential worldwide. However, the genomic information currently available for this species in public databases is limited. This lack of information hinders gene expression assessments, which could bring forward novel insights into the physiology, ecology, evolution, and genetics of this aquaculture species. In this study, we report the first *de novo* transcriptome assembly of cobia liver to improve the availability of gene sequences from this important commercial fish. Thus, Illumina sequencing of liver transcripts generated 1,761,965,794 raw reads, which were filtered into 1,652,319,304 high-quality reads. *De novo* assembly resulted in 101,789 unigenes and 163,096 isoforms, with an average length of 950.61 and 1617.34 nt, respectively. Comparisons against six different databases resulted in functional annotation for 125,993 of these elements (77.3%), providing relevant information regarding the genomic content of *R. canadum.* We trust that the availability of these functionally annotated cobia elements will likely assist future nutrigenomics and breeding programs, involving this important fish farming species.

Keywords: *Rachycentron canadum*; Cobia; Transcriptome; Functional Annotation; Aquaculture; Liver.

1. Introduction

Currently, aquaculture is the most prominent food production industry, with significant growth worldwide. Global aquaculture production increased by 5.3% per year between 2001-2018, with a historical record of 114.5 million tons of farmed species, from which almost 17.7 million tons was due to finfish production. (FAO, 2020). Marine aquaculture plays an essential role in the effort of supplying the increased demand for animal-based protein. Cobia (*Rachycentron canadum*) is a carnivorous

marine fish species of worldwide distribution, and it is the sole representative of the Rachycentridae family. Currently, this species is regarded as the most promising marine fish species in Brazil, mostly due to its fast growth rate (reaching about 4 to 6 kg per year), excellent meat quality (excellent color, texture, and flavor), and high market value (Arnold *et al*., 2002; Benetti *et al*., 2008; Nunes, 2014). However, cobia production is still hindered by lacking of nutrition information, which prevents this species' productivity in industrial aquaculture operation (Fraser and Davies 2009).

Next-generation sequencing (NGS) studies have become an essential molecular tool in aquaculture, assisting in the production of several commercial fish species such as *Sparus aurata* (Calduch-Giner *et al.*, 2013), *Dicentrarchus labrax* (Magnanou *et al.*, 2014), and *Salmo salar* (Glencross *et al.*, 2015; Andrew *et al.*, 2020). *De novo* transcriptome assembly can be used in several different contexts, like genomics/gene expression analyses, which may be applied in many key areas of study, such as conservation genetics, selective breeding, reproductive biology, and nutrition (Leaver *et al.*, 2008; Calduch-Giner *et al.*, 2013; Fox *et al.*, 2014). For example, identifying genes associated with proteins and lipid metabolism in the liver can help the development of specific diets to improve the productive chain of commercial aquaculture species, including cobia. Unfortunately, genomic information in cobia is still scarce, limiting the development of such studies. Thus, to help overcome such limitations, the present manuscript describes an assembled/annotated reference transcriptome of hepatic cells in cobia juveniles. The information contained in this dataset contributes to improve our knowledge regarding the biological and physiological aspects of this fish species. Moreover, it establishes a solid foundation for future studies involving population genomics, breeding programs and nutrigenomics.

2. Data Description

2.1. Sample collection and RNA preparation

Ninety (90) cobia juveniles (128.85 ± 18.43 g) were obtained from a commercial hatchery (Redemar Alevinos, SP, Brazil) and randomly allocated in three 2.000 L tanks. The animals were kept under a mean temperature of 23±1.5 ºC and natural photoperiod throughout the trial, at the Marine Biology Center of the University of São Paulo (CEBIMar). Animals were equally hand-fed twice a day until apparent satiety with a commercial marine fish diet (Guabipirá, Guabi Nutrição e Saúde Animal S.A., SP, Brazil). After six weeks (42 days), fish were anesthetized with benzocaine concentration of 0.4 g □ mL^{-1} and then euthanized by spinal cord section. Hepatic tissue samples from all experimental animals were collected, immediately frozen in liquid nitrogen, and subsequently stored at -80 °C for further analyses. Total RNA from hepatic tissue samples was extracted using PureLink® RNA Mini Kit (Invitrogen 12183018A), following the manufacturer's instructions. RNA samples had their concentration determined with a Nanodrop™ Spectrophotometer (ThermoFisher) and the RNA integrity was determined by the 2100 Bioanalyzer (Agilent Technologies, CA, USA), only samples with RNA integrity number (RIN) above 6.5 were used. This study's experimental procedures were conducted according to the guidelines and approval of the Institutional Animal Care and Use Ethics Committee (#008/2017).

2.2. Library construction and sequencing

The construction of the library and the sequencing of the tissues were carried out at the GENIAL in Center for Research Support Facilities (CEFAP-USP), of the Institute of Biomedical Sciences of the University of São Paulo. RNA extracted from hepatic tissue samples from all animals were equally diluted to 1.000 ng μL^{-1} concentration and pooled for library construction using the TruSeq RNA

Sample Preparation kit, according to the manufacturer's specifications (Illumina Inc., USA). The library fragments were purified with the AM Pure XP system (Beckman Coulter, Beverly, MA, USA) and library quality was quantified by Qubit 2.0 before being subjected to sequencing Finally, paired-end sequencing (2×75 bp) of these cDNA libraries was performed on the NextSeq® equipment (Illumina, Inc., USA) to obtain the sequencing according to the manufacturer's recommendations. MIxS data (Minimum Information about any (x) Sequence**)** for this study is available in **Table 1.**

Information	**Value**
BioProject accession	PRJNA675281
BioProjects name	De novo transcriptome assembly and functional annotation of cobia (*Rachycentron canadum*) liver
BioSamples accession	SAMN16708758, SAMN16708759, SAMN16708760, SAMN16708761, SAMN16708762, SAMN16708763, SAMN16708764, SAMN16708765, SAMN16708766, SAMN16708767, SAMN16708768, SAMN16708769
SRAs accession	SRR13009897, SRR13009896, SRR13009895, SRR13009894, SRR13009893, SRR13009892, SRR13009891, SRR13009890, SRR13009889, SRR13009888, SRR13009887, SRR13009886
TSA accession	GIWT00000000
Sample Name	*Rachycentron canadum* Liver
Organism	*Rachycentron canadum*
Ecotype	"São Sebastião"
Collection date	01/jun/18
Broad-scale environmental context	ENVO_00000447 (marine biome)
Local-scale environmental context	ENVO_00001999 (marine water body)
Environmental medium	ENVO_00002010 (saline water)
Geographic location	Brazil:Sao Sebastiao,Sao Paulo
Latitude and longitude	23.858778 S 45.428833 W
Developmental stage	juvenile
Sequence method	Illumina Nextseq
Assembly method	Trinity
Annotation method	EnTAP, Diamond (RefSeq proteins, Nr proteins, Swiss-Prot, EggNOG)

Table 1. MIxS information for transcriptome assembly of *R. canadum*.

2.3. *De novo* assembly, functional annotation

FASTQ format raw sequencing data was processed in a Public Galaxy Server, available at https://usegalaxy.eu. Initially, the quality of raw sequences was assessed using FastQC (Andrews, 2010) and MultiQC (Ewels *et al*., 2016). Fastp (Chen *et al*., 2018) was then used to remove low-quality reads (Q<30), adapters, and other contaminant sequences. Trinity platform (Haas *et al.*, 2013) was then used for *de novo* transcriptome assembly of filtered reads, and assembly metrics were obtained using the TrinityStats script. Transcriptome completeness was finally assessed using the Benchmarking Universal Single-Copy Orthologs (BUSCO) software (Seppey *et al*., 2019) based on OrthologDB version 9 (Zdobnov *et al*., 2017).

Elements of the assembled transcriptome were functionally annotated by similarity searches (blastn) performed against RefSeq RNA and the "*Lates*/*Seriola* database", a custom RefSeq transcripts database containing sequences from cobia-related species, such as *Lates calcarifer*, *Seriola dumerili*, and *Seriola lalandi dorsalis.* These sequences are also in the Kegg's Organisms Complete Genomes list (available at https://www.genome.jp/kegg/catalog/org_list.html). Additional annotations were obtained with the aid of the Eukaryotic Non-Model Transcriptome Annotation Pipeline (EnTAP) (Hart *et al.*, 2019).

Contigs were also queried (blastx) for homology against the National Center for Biotechnology Information non-redundant protein database (NCBI nr), NCBI proteins reference database (RefSeq), the curated Swiss-Prot database from UniProt Knowledgebase (UniProtKB) (UniProt Consortium, 2019) and the EggNOG proteins database (Huerta-Cepas *et al.*, 2016). The EggNOG hits also helped to assign biological function to the putative cobia genes, identifying GO (Gene Ontology) (Ashburner *et al*., 2000; Gene Ontology Consortium, 2019) and KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa & Goto, 2000; Kanehisa, 2019) terms. The EnTAP functional annotation process was carried

out using a Dugong container environment (Menegidio *et al.*, 2018). Additional transcriptome filtering before NCBI submission was carried out with the aid of seqtk subseq from seqtk toolkit (https://github.com/lh3/seqtk), based on sequence length (>200 nucleotides) and univec (https://ftp.ncbi.nlm.nih.gov/pub/UniVec/).

2.4. Functional and structural characterization of transcriptome

Sequencing of the cDNA libraries derived from *R. canadum* liver material resulted in 1,761,965,794 raw reads. After high-quality reads selection and trimming, we were left with a total of 1,652,319,304 reads (93.77% of raw reads), which were used for *de novo* transcriptome assembly, using Trinity software (Haas *et al.*, 2013). General features of the *R. canadum* liver transcriptome are summarized in **Table 2**, consisting of 101,789 unigenes and 163,096 isoforms, displaying an average length of 1,617.34 and 950.61 nt, respectively. The median length (N50) of these elements was 7,843 nt for unigenes and 2,312 nt for isoforms. A total of 95,075 transcripts (58.29%) presented 500 nt or more. Identification of 83.8% of the complete universal genes (3839 out of the total 4584 genes from Actinopterygii odb9 lineage) supported the high quality and completeness of this transcriptome assembly (**Fig. 1A)**. Among the 3839 conserved BUSCO genes, 38.1% were single copy, while 45.7% were duplicated (**Supplementary Table ST1-S11**).

Illumina sequencing	**raw**	**high-quality**
Number of reads	1,761,965,794	1,652,319,304
Assembly	**unigenes**	**isoforms**
Total sequences	101,789	163,096
N50	7,843	2,312
Average contig length (bp)	1,617.34	950.61
Median contig length (bp)	704	391
Total assembled bases	263,781,171	96,761,286
CG%	44.5	
Functional annotation	**annotated sequences**	
NCBI nr	29,155	
NCBI RefSeq RNA	122,741	
Swiss-Prot	15,507	
Lates/*Seriola*	118,570	
GO-Biological Process	48,550	
GO-Molecular Function	47,766	
GO-Cellular Component	34,492	
Kegg	24,100	
Total	125,993	

Table 2. Summary of the *de novo* transcriptome assembled for *R. canadum*.

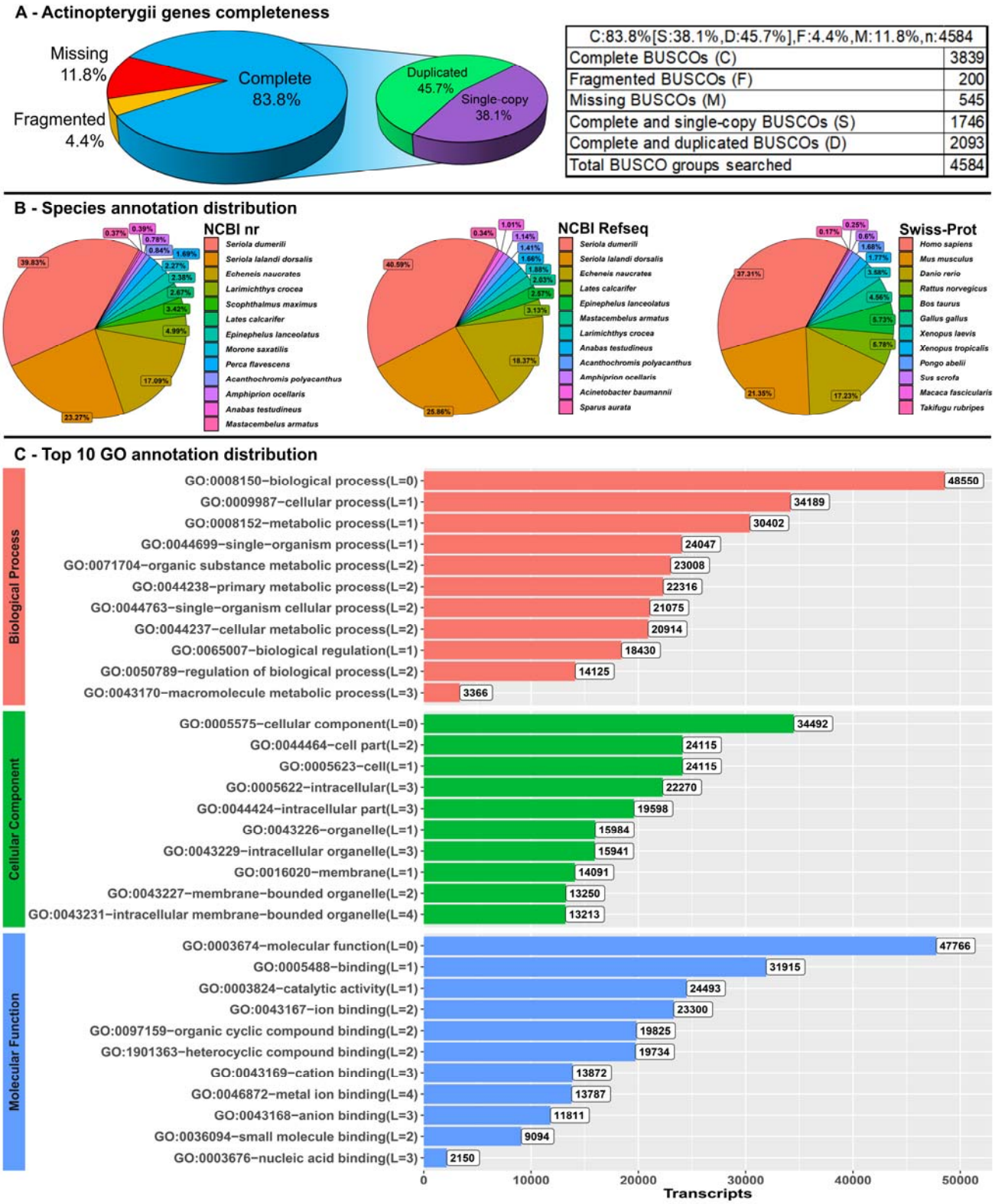


Fig. 1. Completeness and homology search of *R. canadum* liver transcripts. (A) Percentage of completeness on the core set of genes of *R. canadum* based on Actinopterygii database (orthodb9) from BUSCO. (B) Species annotation distribution for the best hits from NCBI nr, NCBI RefSeq and Swiss Prot databases. (C) Gene ontology distribution for the 10 most assigned biological process, cellular component and molecular function of assembled isoforms from *R. canadum* liver transcriptome (Available in 10.6084/m9.figshare.13335698).

Transcriptome annotation against a series of databases (NCBI nr, NCBI RefSeq RNA, SwissProt, GO, KEGG, *Lates*/*Seriola*) resulted in functional assignment for 125,993 transcripts (77.3%). Most sequence homologies were found against NCBI RefSeq RNA (122,741 transcripts, 97.4%), followed by the *Lates*/*Seriola* database (118,570 transcripts, 94.1%), NCBI nr (29,155 transcripts, 23.14%), NCBI RefSeq (25,728, 20.42%) and Swiss-Prot (15,507 transcripts, 12.30%) (**Table1**; **Supplementary Table ST1-S10**). Sequence homologies identified by EnTAP were distributed throughout many bony fish species, of which *Seriola dumerili* was the most frequent (nr = 39.83%, refseq = 40.59%), followed by *Seriola lalandi dorsalis* (nr = 23,27%, refseq = 25.86%), *Echeneis naucrates* (nr 17.9%, refseq = 18.37%) and *Larimichthys crocea* (nr = 4.99%, refseq = 3.13%) (**Fig. 1B**). Most transcripts (75,060) were functionally annotated with GO terms by eggNOG (Huerta-Cepas *et al.*, 2016), which assigned 48,550 transcripts (65%) to biological processes, 34,492 to cellular components (46%) and 47,766 to molecular functions (64%). The ten most representative functional groups within each category are shown in **Fig. 1C**. A total of 24,100 isoforms were annotated into at least one KEGG pathway term (**Table2**; **Supplementary Table ST1-S10**).

In summary, our findings provide an important tool for applied studies related to cobia production. These outcomes will help to improve the knowledge regarding cobia physiology consequently improving aquaculture operations leading to higher productivity.

2.5 Data availability

Sequencing raw data were deposited in the Sequence Read Archive (SRA) repository of the National Center for Biotechnology Information (NCBI), under accession number SRR13009897, SRR13009896, SRR13009895, SRR13009894, SRR13009893, SRR13009892, SRR13009891, SRR13009890, SRR13009889, SRR13009888, SRR13009887 and SRR13009886, associated to the BioProject numbers PRJNA675281 and BioSamples numbers SAMN16708758, SAMN16708759, SAMN16708760, SAMN16708761, SAMN16708762, SAMN16708763, SAMN16708764,

SAMN16708765, SAMN16708766, SAMN16708767, SAMN16708768 and SAMN16708769. The Transcriptome Shotgun Assembly (TSA) project has been deposited at DDBJ/EMBL/GenBank under accession number GIWT00000000. The version described in this paper is the first version, GIWT00000000.1. Supplementary Table S1 is available from the Figshare repository (10.6084/m9.figshare.13242176). Additional data derived from this study (including all intermediate data) are also available from the Open Science Framework (OSF) repository (DOI: 10.17605/OSF.IO/BV3WA).

Authors contributions

B.C.A. sampled the specimens. B.C.A., G.S.B., AWSH and R.G.M. performed molecular analyses and sequencing. D.A.B., L.R.N. and F.B.M. assembled and evaluated the transcriptome. All authors wrote the paper. All authors read and approved the final version of the manuscript.

Declaration of Competing Interest

The authors report no conflicts of interest. The authors alone are responsible for the content and the writing of the paper.

Acknowledgements

This study was financed in part by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (grant # 2019/26018-0) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (grant # 305493/2019-1). D.A.B. and B.C.A. are recipients of scholarship grants from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–Brasil-CAPES

(http://www.capes.gov.br/) (grant # 88882-365823/2019-01) and FAPESP (grant # 2016/12435-0), respectively. A.W.S. Hilsdorf and R.G. Moreira are recipients of CNPq productivity scholarships (grants # 304662/2017-8 and 305493/2019-1, respectively). We are grateful to the Facility GENIAL (Genome Investigation and Analysis Laboratory) - Center for Research Support Facilities (CEFAP-USP), of the Institute of Biomedical Sciences of the University of São Paulo.

Appendix A. Supplementary data

Supplementary data to this article can be found online at Open Science Framework: https://osf.io/bv3wa/.

References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. and Harris, M.A., 2000. Gene ontology: tool for the unification of biology. Nat. Genet 25(1), 25-29. https://doi.org/10.1038/75556.

Andrew, S.C., Primmer, C.R., Debes, P.V., Erkinaro, J. and Verta, J.P., 2020. The Atlantic salmon whole blood transcriptome and how it relates to major locus maturation genotypes and other tissues. Mar. Genom., p.100809. https://doi.org/10.1016/j.margen.2020.100809.

Andrews S., 2010. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Arnold, C.R., Kaiser, J.B., Holt, G.J., 2002. Spawning of cobia *Rachycentron canadum* in captivity. J. World Aquacult. Soc., *33*(2), 205-208. https://doi.org/10.1111/j.1749-7345.2002.tb00496.x.

Benetti, D.D., Orhun, M.R., Sardenberg, B., O'Hanlon, B., Welch, A., Hoenig, R., Zink, I., Rivera, J.A., Denlinger, B., Bacoat, D., Palmer, K., 2008. Advances in hatchery and grow-out technology of cobia *Rachycentron canadum* (Linnaeus). Aquac. Res. 39(7), 701-711. https://doi.org/10.1111/j.1365-2109.2008.01922.x.

Calduch-Giner, J.A., Bermejo-Nogales, A., Benedito-Palos, L., Estensoro, I., Ballester-Lozano, G., Sitjà-Bobadilla, A., Pérez-Sánchez, J., 2013. Deep sequencing for de novo construction of a marine fish (*Sparus aurata*) transcriptome database with a large coverage of protein-coding transcripts. BMC genomics, 14(1),c178. https://doi.org/10.1186/1471-2164-14-178.

Chen, S., Zhou, Y., Chen, Y. and Gu, J., 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 34(17), i884-i890. https://doi.org/10.1093/bioinformatics/bty560

Ewels, P., Magnusson, M., Lundin, S., Käller, M., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics, 32(19), 3047-3048. https://doi.org/10.1093/bioinformatics/btw354.

Fox, S.E., Christie, M.R., Marine, M., Priest, H.D., Mockler, T.C. and Blouin, M.S., 2014. Sequencing and characterization of the anadromous steelhead (Oncorhynchus mykiss) transcriptome. Mar. Genom. 15, 13-15. doi: 10.1016/j.margen.2013.12.001.

FAO. 2020. The State of World Fisheries and Aquaculture 2020. Sustainability in action. Rome. 244 p.

Fraser, T.W. and Davies, S.J., 2009. Nutritional requirements of cobia, *Rachycentron canadum* (Linnaeus): a review. Aquac. Res. 40(11), 1219-1234. https://doi.org/10.1111/j.1365-2109.2009.02215

Gene Ontology Consortium, 2019. The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Res., 47(D1), d330-d338. https://doi.org/10.1093/nar/gky1055.

Glencross, B.D., De Santis, C., Bicskei, B., Taggart, J.B., Bron, J.E., Betancor, M.B. and Tocher, D.R., 2015. A comparative analysis of the response of the hepatic transcriptome to dietary docosahexaenoic acid in Atlantic salmon (*Salmo salar*) post-smolts. BMC genomics, 16(1), 684. https://doi.org/10.1186/s12864-015-1810-z.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc 8(8), 1494-1512. https://doi.org/10.1038/nprot.2013.084.

Hart, A.J., Ginzburg, S., Xu, M., Fisher, C.R., Rahmatpour, N., Mitton, J.B., Paul, R., Wegrzyn, J.L., 2020. EnTAP: bringing faster and smarter functional annotation to non□model eukaryotic transcriptomes. Mol. Ecol. Resour. 20(2), 591-604. https://doi.org/10.1111/1755-0998.13106.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., Rattei, T., Mende, D.R., Sunagawa, S., Kuhn, M., Jensen, L.J., 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 44(D1), D286-D293. https://doi.org/10.1093/nar/gkv1248.

Kanehisa, M. and Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 28(1), 27-30. https://doi.org/10.1093/nar/28.1.27.

Kanehisa, M., 2019. Toward understanding the origin and evolution of cellular organisms. Protein Sci. 28(11), 1947-1951. https://doi.org/10.1002/pro.3715.

Leaver, M.J., Bautista, J.M., Björnsson, B.T., Jönsson, E., Krey, G., Tocher, D.R., Torstensen, B.E., 2008. Towards fish lipid nutrigenomics: current state and prospects for fin-fish aquaculture. Rev Fish Sci. 16, 73-94. https://doi.org/10.1080/10641260802325278.

Magnanou, E., Klopp, C., Noirot, C., Besseau, L., Falcón, J., 2014. Generation and characterization of the sea bass Dicentrarchus labrax brain and liver transcriptomes. Gene, 544(1), 56-66. https://doi.org/10.1016/j.gene.2014.04.032.

Menegidio, F.B., Jabes, D.L., Costa de Oliveira, R. and Nunes, L.R., 2018. Dugong: a Docker image, based on Ubuntu Linux, focused on reproducibility and replicability for bioinformatics analyses. Bioinformatics, 34(3), 514-515. https://doi.org/10.1093/bioinformatics/btx554.

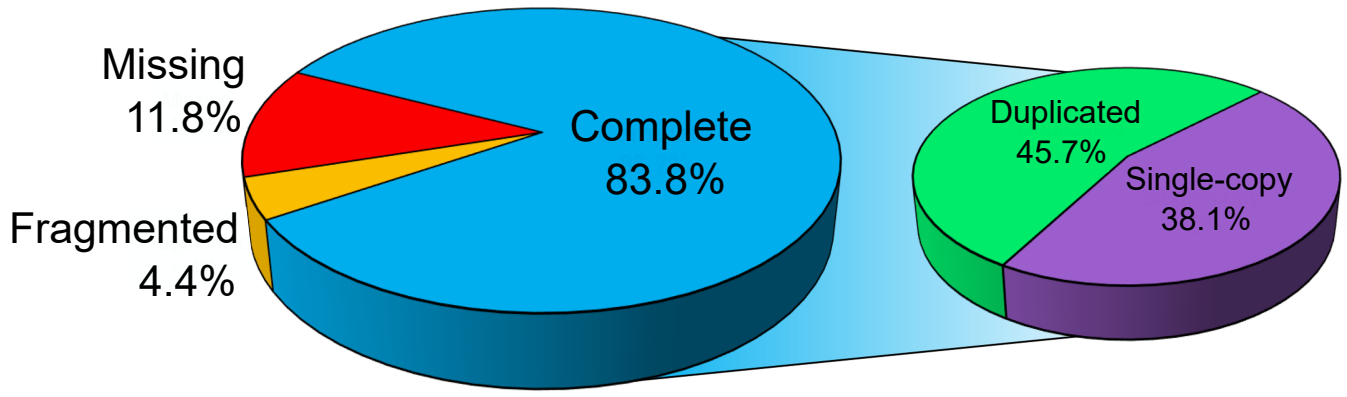
Nunes, A.J.P. 2014. Ensaios com o beijupirá, *Rachycentron canadum*. Fortaleza: Ministério da Pesca e Aquicultura/CNPQ/UFC, 2014. 352 p.

Seppey, M., Manni, M. and Zdobnov, E.M., 2019. BUSCO: assessing genome assembly and annotation completeness. In Gene Prediction (pp. 227-245). Humana, New York, NY. https://doi.org/10.1007/978-1-4939-9173-0_14.

UniProt Consortium, 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 47(D1), D506-D515. https://doi.org/10.1093/nar/gky1049.

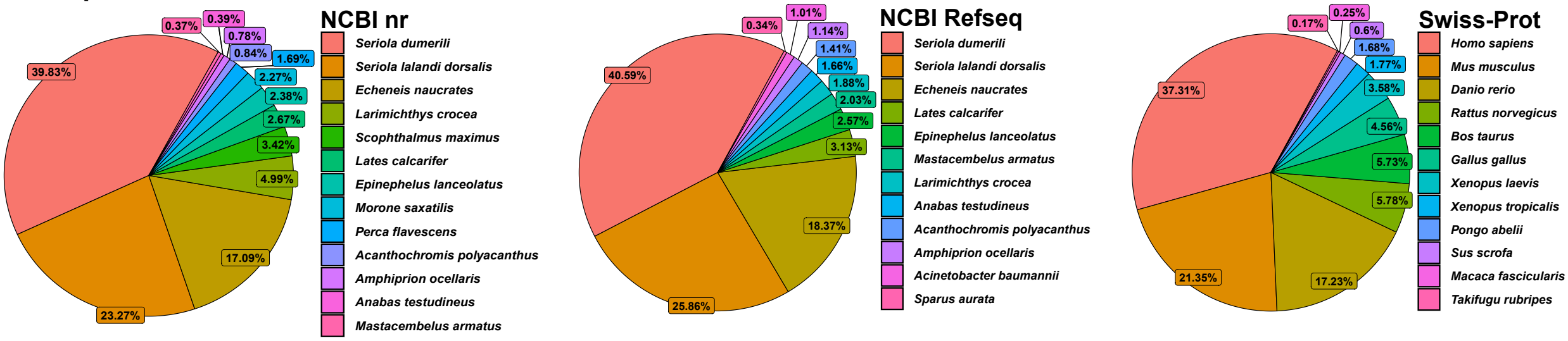
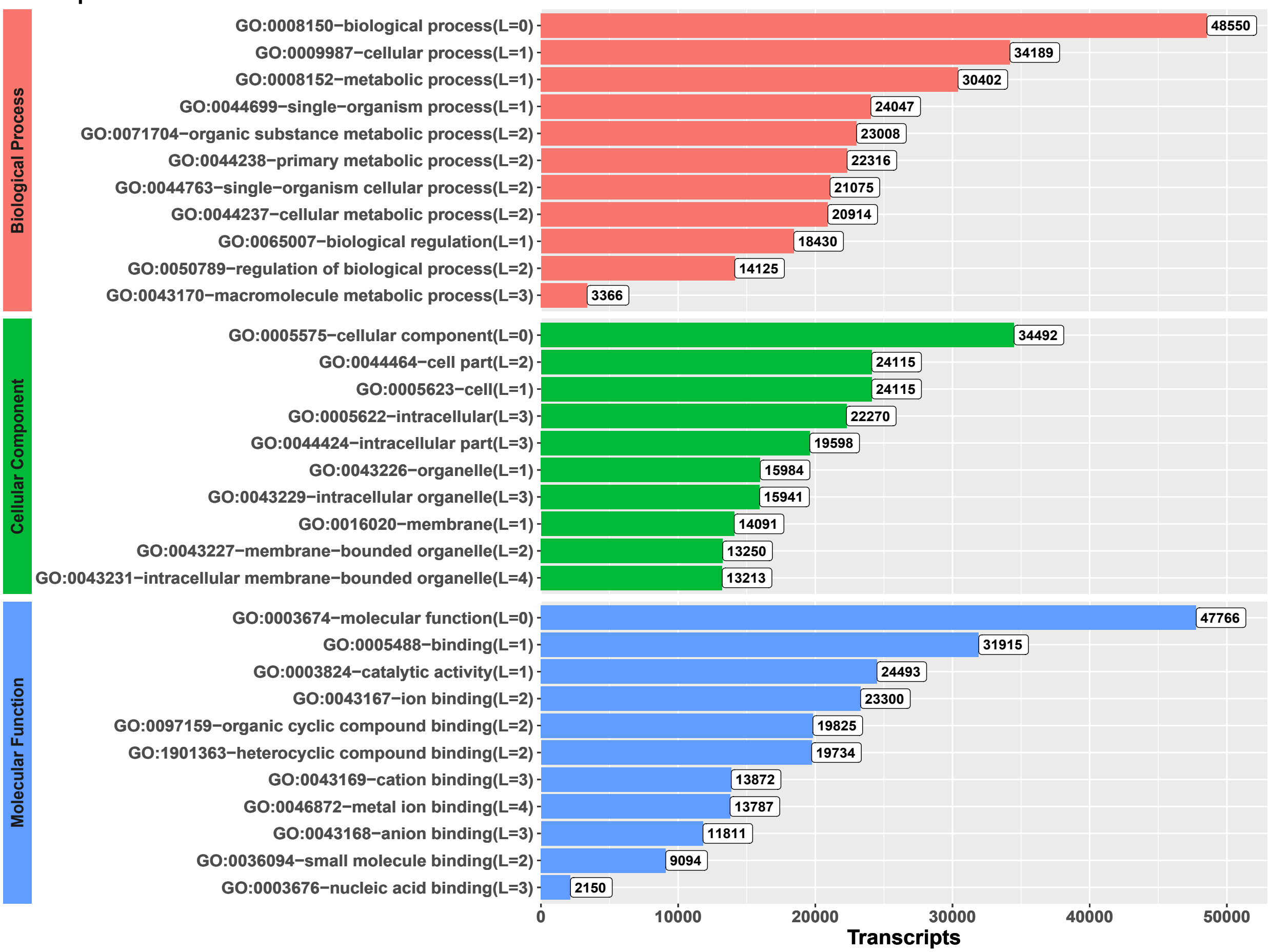
Zdobnov, E.M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R.M., Simao, F.A., Ioannidis, P., Seppey, M., Loetscher, A., Kriventseva, E.V., 2017. OrthoDB v9. 1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Res, 45(D1), D744-D749. https://doi.org/10.1093/nar/gkw1119.

A - Actinopterygii genes completeness



C:83.8%[S:38.1%,D:45.7%],F:4.4%,M:11.8%,n:4584	
Complete BUSCOs (C)	3839
Fragmented BUSCOs (F)	200
Missing BUSCOs (M)	545
Complete and single-copy BUSCOs (S)	1746
Complete and duplicated BUSCOs (D)	2093
Total BUSCO groups searched	4584

B - Species annotation distribution



C - Top 10 GO annotation distribution