Data in brief

Data Article

From the Amazon: A comprehensive liver transcriptome dataset of the teleost fish tambaqui, *Colossoma macropomum*

André M. Machado a,1, Renato Ferraz a,b,1, Ricardo do Amaral Ribeiro c, Rodrigo Ozório a, L. Filipe C. Castro a,d,*

a CIIMAR — Interdisciplinary Centre of Marine and Environmental Research, U. Porto — University of Porto, Porto, Portugal
b ICBAS — Institute of Biomedical Sciences Abel Salazar, U. Porto — University of Porto, Portugal
c Universidade Federal Do Acre, Brazil
d Department of Biology, Faculty of Sciences, U. Porto — University of Porto, Portugal

**ABSTRACT**

The teleost fish tambaqui, *Colossoma macropomum*, is a valuable resource for the Brazilian aquaculture sector, representing more than one-quarter of the total production. In this context, the development of molecular tools is paramount to address and improve productivity, nutrition, and genetic breeding programs. In this study, we applied RNA-seq technology to produce the first comprehensive liver transcriptome in this species. Our analysis generated a gold standard transcriptome with a total of 43,098 transcripts, with an N50 of 1855 bp and the average length of 1312 bp. To functionally annotate the transcripts, the Trinotate pipeline together with several public databases were scrutinized. The blast-x analysis revealed more than 40,000 homologous match hits for each database (NCBI-Nr, Uniref90, Swissprot, Trembl), while the Kaas web
server allowed the mapping of our transcripts to 380 KEGG pathways. The dataset provided in this study entails a comprehensive molecular resource, which will be instrumental to further develop tambaqui aquaculture, specifically in the field of nutrigenomics.

© 2019 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Data

This data article reports the first deep-sequencing of liver RNA of tambaqui (Colossoma macropomum, Cuvier, 1818), which should prove valuable for investigations in other fields such as nutrigenomics [1]. We provide an extended molecular resource that integrates the assembly and annotation of the first liver transcriptome for this species. In addition, we also deliver several analyses of GO terms, KEGG pathways and Clusters of Orthologous Groups focused on the general features of this dataset. Briefly: in Table 1 are described the in silico statistics of the original, filtered and gold standard transcriptome assembly; in Fig. 1e plotted the blast-x analysis of the gold standard transcriptome; in Fig. 2 it is possible to visualize the clusters of orthologous groups analysis; in Fig. 3 is displayed the Functional classification of tambaqui liver transcriptome into three Gene Ontology (GO) categories.

2. Experimental design, materials and methods

2.1. Sampling, RNA isolation and RNA-Seq library preparation

Tambaqui juvenile specimens were obtained from a commercial farm in Rio Branco, Acre, Brazil (10°05’10.6”S 67°32’15.8”W) and transported to Universidade Federal do Acre, Brazil, where they were
anaesthetized. Fresh liver was quickly collected and immediately preserved in an RNA stabilization buffer (RNAlater) and stored at −80 °C prior to RNA extraction. RNA extraction from the liver sample was performed with the kit (illustra RNASpin Mini RNA Isolation Kit, GE Healthcare, UK). The process included an on-column DNase I treatment (provided in the kit). RNA integrity was assessed on a 1% agarose TAE gel stained with GelRed™ nucleic acid stain (Biotium, Hayward, CA, USA). The generated high-quality liver total RNA sample was sequenced in the Illumina Hiseq 4000 platform, using 150bp paired-end sequencing reads by STABVIDA, Lda (Caparica, Portugal).

2.2. Cleaning, de novo assembly, and optimization of the liver transcriptome

The raw reads generated in sequencing were quality-checked with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Trimmomatic [2] was used to trim the first 15 bases of the reads and bases with a quality score below 15 at leading and trailing ends. Reads were then scanned with a 4-base sliding window, cutting when the average quality per base dropped below 20. In the end, only reads with high quality and longer than 50 bases were retained for further analysis. After applying a first technical quality control, the dataset was de novo assembled using the Trinity v2.5.0 software [3], following the Haas and co-workers protocol [4] and specific parameters to our case, strand-specific data and minimum length contig (SS_lib_type RF; min_contig_length 300). From this stage we obtained the Raw transcriptome assembly (Table 1).

The liver transcriptome optimization was done using three independent approaches. Firstly, the TransDecoder (https://transdecoder.github.io/) was used to predict open reading frames (ORFs) with a minimum cut-off of 100 amino acids, with recourse to homology searches (Blast-p against Swissprot Database [5] and Pfam search [6]) as auxiliary. Secondly, all transcripts were blasted against two independent databases, (NR) Non-Redundant of NCBI and Uniref90 of Uniprot [5]. To perform the blast, it was used the blast-x tool of DIAMOND v0.8.36 software [7] and all hits, given a match with Actinopterygii taxon with an E-value cut-off of 1e-5, were retained (Table 1). At this stage, the liver transcriptome filtration was done by overlapping the contigs with ORF to the match hits in Actinopterygii taxon (Filtered de novo transcriptome assembly).

### Table 1

| C. macropomum tissue | Liver |
|-----------------------|-------|
| Number of raw sequencing reads | 52753287 |
| Number of cleaned reads used in the assembly | 43509171 |
| Percentage of reads submitted to assembly (%) | 82.48 |
| Assembly versions | Raw transcriptome assembly | Filtered de novo transcriptome assembly | Gold Standard transcriptome |
| Number of "genes" | 471503 | 30762 | 28638 |
| Number of transcripts | 659372 | 103895 | 43098 |
| N50 transcript length (bp) | 1000 | 2517 | 1855 |
| Median transcript length (bp) | 466 | 1376 | 930 |
| Mean transcript length (bp) | 776 | 1811 | 1312 |
| Length of the smallest transcript | 301 | 301 | 301 |
| Length of the largest transcript | 19846 | 19846 | 18047 |
| Number of transcripts with length > 1k nuc | 122174 | 66468 | 20222 |
| Number of transcripts with length > 10k nuc | 285 | 282 | 36 |
| Total Assembled bases | 511615092 | 188186859 | 56527072 |

*a* Number of groups of transcripts clustered with base on shared sequence content.

*b* N50 is the sequence length of the shortest contig at 50% of the total transcriptome length.

*c* Number of transcripts with length larger than 1000 nucleotides.

*d* Number of transcripts with length larger than 10000 nucleotides.
Thirdly, the tr2aacds pipeline, from the Evidential — Gene package (http://arthropods.eugenes.org/EvidentialGene/), was used as a strategy to handle the redundancy and the number of isoforms per ‘gene’ in this filtered transcriptome. All transcripts classified by the tr2aacds pipeline as ‘primary’ or ‘alternate’ were retained to the next step of the annotation (Gold standard transcriptome assembly). To access the quality of the gold standard transcriptome, the distribution of sequence lengths and the number of isoforms per genes are shown in Fig. 1 a and b and Supplementary Table S1 and S2 in Supplementary File 1. Importantly, all steps of filtering and optimization were supervised with Trinity and Transrate [8] statistics (Table 1). Additionally, we also assessed the completeness of our transcriptome, in terms of gene content, using the eukaryota, and the metazoa lineage-specific profile libraries of Benchmarking Universal Single-Copy Orthologs tool (BUSCO) [9] (Supplementary Table S3 in Supplementary File 1).

2.3. Transcriptome annotation

To perform the transcriptome annotation, the final nucleotide and aminoacid sequences were retrieved from transdecoder.pep and Trinity.fasta initial files with the heads of the gold standard
Fig. 2. Histogram of the clusters of orthologous groups of *C. macropomum* transcriptome (COG).

Fig. 3. Functional classification of *C. macropomum* in three Gene Ontology (GO) categories—biological process (blue), molecular function (grey), cellular component (red).
transcriptome assembly. Subsequently, the sequences were searched against several databases, NR, Uniref90, Trembl, Swissprot, at the local level, using blast-p and blast-x tools of the DIAMOND v0.8.36 software [7], and applying an E-value cut-off of 1e-5. To have a global overview of the orthologs genes contained in our transcriptome assembly, in relation to NR database of NCBI, the top 20 species distribution of blast-x against NR database, the similarity and e-value distributions are plotted in Fig. 1 c and d (Supplementary Tables S4, S5, S6 of Supplementary File 1). The PFAM [6] and HMMER [10] were used to identify protein domains, TMHMM [11] to predict transmembrane regions, GOseq [12] to determine GO and eggNOG v.3.0 [13] to identify clusters of orthologous groups of genes. All the results were integrated into the Trinotate v3.0.1 (http://trinotate.github.io) annotation pipeline and then reported with an E-value cut-off of 1e-5 (Supplementary Table S7 in Supplementary File 1).

2.4. Go terms, clusters of orthologous groups and Kegg patways analyses

The analysis of Go terms and COG's were done using the longest ORF per 'gene', unigenes, of the trinotate report (Annotation statistics to this dataset can be consulted under the sub-title of Final Transcriptome Subset, Supplementary Table S7 in Supplementary File 1). Regarding the search of COG's, it was done through the eggNOG database integrated within the Trinotate pipeline, and briefly, it was checked the percentage of unigenes annotated in at least one of the 25 COG categories contained in COG database. In Fig. 2 and in Supplementary Table S8 of Supplementary File 1 it is possible to access a graphical and tabular distributions of these analyses. The Go terms analyses were based on blast hits of SwissProt, and through this way we assign the unigenes to three main categories: molecular function, cellular component and biological process. The top 15 GO terms with a higher number of mapped unigenes were plotted in Fig. 3. In parallel, to improve the comprehension of functional and metabolic interactions in our transcriptome, the longest protein sequences, also were submitted to KAAS web server for KEGG annotation, using as reference 400,502 teleost fish sequences. This approach allowed to annotate and map a high number of unigenes to several metabolic pathways (Supplementary Table S9 in Supplementary File 1).

Ethics statement

The tissue sampling procedure was approved by the ethical committee of the Federal University of Acre, Brazil, with the project number 23107-009564/2014-29 and protocol number 08/201, under the responsibility of Prof. Anselmo Fortunato Ruiz Rodriguez. The samples were exported from Brazil to Portugal with the approved of the Brazilian Environmental and Renewable Resource Institute (IBAMA) with permit number 16BR021098/DF.

Acknowledgements

This study was supported by CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brazil.

Transparency document

Transparency document associated with this article can be found in the online version at https://doi.org/10.1016/j.dib.2019.103751.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dib.2019.103751.

References

[1] R.B. Ferraz, N. Kabeya, M. Lopes-Marques, A.M. Machado, R.A. Ribeiro, A.L. Salaro, R. Ozório, L.F.C. Castro, Ó. Monroig, A complete enzymatic capacity for long-chain polyunsaturated fatty acid biosynthesis is present in the Amazonian teleost
[1] A.M. Bolger, M. Lohse, B. Usadel, Trimmmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014) 2114–2120, https://doi.org/10.1093/bioinformatics/btu170.

[2] M.G. Grabherr, B.J. Haas, M. Vassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Rachowdury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B.W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nat. Biotechnol. 29 (2011) 644–652, https://doi.org/10.1038/nbt.1883.

[3] B.J. Haas, A. Papanicolaou, M. Vassour, M. Grabherr, P.D. Blood, J. Bowden, M.B. Couger, D. Eccles, B. Li, M. Lieber, M.D. Macmanes, M. Ott, J. Orvis, N. Pochet, E. Strozzi, N. Weeks, R. Westerman, T. Williams, C.N. Dewey, R. Henschel, R.D. Leduc, N. Friedman, A. Regev, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis, Nat. Protoc. 8 (2013) 1494–1512, https://doi.org/10.1038/nprot.2013.084.

[4] A. Bateman, M.J. Martin, C. O’Donovan, M. Magrane, E. Alpi, R. Antunes, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, H. Bye-AJee, A. Cowley, A. Da Silva, M. De Giorgi, T. Dogan, F. di Palma, L. Garmiri, G. Georgiou, D. Gonzalez, E. Hatton-Ellis, W. Li, W. Liu, R. Lopez, J. Luo, Y. Lussi, A. MacDougall, A. Nightingale, B. Palka, K. Pichler, D. Poggiali, S. Purdil, I. Rokhit, G. Qi, S. Rosano, R. Saidi, T. Sawford, A. Speretta, E. Turner, N. Tyagi, V. Volynkin, T. Wardell, K. Warner, X. Watkins, R. Zaru, H. Zeilinger, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimone, G. ArgoudPuy, A. Auchincloss, K. Axelsson, P. Bansal, D. Baratin, M.C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. De Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuerb, M. Gasteiger, S. Gehant, V. Gerekten, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hu, F. Jungo, G. Keller, V. Lara, P. Lemerrier, D. Lieberb, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nospikiel, S. Pao, S. Pedruzzi, S. Pihl, M. Pozzato, M. Prues, C. Rovire, B. Roesch, M. Schneider, C. Sigrist, K. Sonesson, S. Stahl, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A. Vethy, J. Wu, C.H. Arighi, C. Arisp, C. Chen, Y. Chen, J.S. Garavelli, H. Huang, K. Lahk, P. McCarvey, D.A. Natale, K. Ross, C.R. Vinayaka, Q. Wang, Y. Wang, L.S. Yeh, J. Zhang, UniProt: the universal protein knowledgebase, Nucleic Acids Res. 45 (2017) D158–D169, https://doi.org/10.1093/nar/gkw1099.

[5] B. Buchfink, C. Xie, D.H. Huson, Fast and sensitive protein alignment using DIAMOND, Nat. Methods 12 (2014) 59–60, https://doi.org/10.1038/nmeth.3176.

[6] R. Smith-Unna, C. Boursnell, R. Patro, J.M. Hibberd, S. Kelly, TransRate: reference-free quality assessment de novo transcriptome assemblies, Genome Res. 26 (2016) 1134–1144, https://doi.org/10.1101/gr.196499.115.

[7] A.M. Machado et al. / Data in brief 23 (2019) 103751