Latin American Data in Science, v.1, n.2
datainscience.com.br
ISSN: 2763-9290

Functional DNA annotation from a preliminary de novo genome assembly of *Brycon orbignyanus*, an endangered Neotropical migratory fish.

Recebido: 30/05/2021 | Aceito: 21/10/2021 | Publicado: 12/11/2021
https://doi.org/10.53805/lads.v1i2.12

Raissa Cristina D. Graciano [1], Rafael S. Oliveira[2], Isllas Miguel Santos[3], Gabriel M. Yazbeck[1,3*]

ABSTRACT

The predicted sequence for thousands of genes revealed by a preliminary low-coverage genome assembly is presented for *Brycon orbignyanus*, an endangered migratory fish. Neotropical migratory fish stocks have been drastically reduced due to accumulated environmental pressure. *Brycon orbignyanus*, once one of the main fisheries species in the Platine Basin, is now very rare in nature and relies on spawning programs and a few well preserved or still untouched sites. The use of high-throughput DNA sequencing is still untapped regarding the functional genome information from *B. orbignyanus*. In order to help bridging this gap, we present a dataset resulting from the first functional annotation from a *de novo* genome assembly for *B. orbignyanus*, from short reads (90 bp), obtained by the HiSeq 2000 platform (Illumina). The annotation was performed for scaffolds over 10 kb using the Maker pipeline, with reference sequences taken from the NCBI for the Characiformes order. This annotation resulted in the prediction of 12,734 genes, classified with the aid of PANTHER. The data presented here can facilitate the development of basic research in this threatened species, along with practical biotechnological tools for different areas, such as commercial and environmental fish spawning operations (*e.g.* hormonal induction, growth) and human health.

Keywords: Ab initio prediction; Aquaculture; Bioinformatics; Conservation genetics; NGS.

[1] Universidade Federal de São João Del Rei, Programa de Pós-graduação em Biotecnologia, São João Del Rei, Brasil. dna@ufsj.edu.br
[2] Universidade Federal de São João Del Rei, Departamento de Ciência da Computação, São João Del Rei, Brasil.
[3] Universidade Federal de São João Del Rei, Departamento de Zootecnia, São João Del Rei, Brasil.

DATA IMPORTANCE

- *Brycon orbignyanus,* Valenciennes, 1850 (Characiformes: Bryconidae) is an endangered migratory fish, heavily depleted in impounded portions of the Platine Basin in South America, such as the upper Grande River (ROSA; LIMA, 2008; TONELLA *et al.*, 2019). It has been the focus of environmental mitigation initiatives through the practice of hatchery spawning and stocking operations (TONELLA *et al.*, 2019). These programs are still in a relatively early technical-scientific stage, with many unknown variables regarding their real efficiency (*e.g.* do fish used in restocking activities survive? Do they breed? Is artificial stocking changing the genetic makeup of fortuitous remnant demes? Are there local adaptations?). It would be useful to access genomic data from this species to clarify these and other issues.
- This dataset is also aimed for preparing fertile ground for innovation to be seeded, such as in developing supporting tools for hatchery spawning (*e.g.* growth and gonadal maturation hormones), since piracema species typically depend on their migratory struggle for triggering fertility, which is conspicuously missing for tank reared specimens. Thus, efforts in understanding and using genomic information about this species are particularly important in helping environmental mitigation activities and boosting productivity.
- The new dataset presented herein comes from the further exploration of a previously published low-coverage genome assembly for *B. orbignyanus* (YAZBECK *et al*., 2018), as well as new assembly attempts, to perform a genome-wide annotation of its predicted protein coding sequences, for the first time in this species. This approach intends to widen the genetic knowledge horizon for this endangered species and to take the most advantage of the previously available short-reads dataset. We particularly focus on the description of sequences of predicted genes related to gonadal hormones, aiming subsidizing biotechnological innovation in aquaculture of *B. orbignyanus*, and closely related species (LOPERA-BARRERO, 2009).

MATERIALS AND METHODS

For this work we used data available under NCBI's BioProject PRJNA416191, from BioSample SAMN07944233, referring to Voucher 120457 (deposited at LARGE-UFSJ). Raw short-reads can be retrieved through NCBI's SRA database (SRX3350440) and a BAM file is available (SRX3427716), containing these reads' alignment to the preliminary assembly presented in Yazbeck et al. (2018). The genomic data for this project were generated from a single adult female was captured at the Volta Grande Environmental Station Hatchery, Rio Grande, MG, Brazil (-20.026197, -48.220430). Total genomic DNA was extracted from ≈2g of muscle tissue using the Wizard Genomic DNA Purification Kit (Promega, Fitchburg, USA). New genome assemblies were performed with this raw data, by De Bruijn Graph strategy, with SOAPdenovo 2 (LUO et al., 2012). For new assemblies, we varied k-mer sizes, according to estimations performed with KmerGenie (CHIKHI; MEDVEDEV, 2014), assaying the following values of k-mer=33, 47, 49, 53 and 55. The resulting alternative assemblies were evaluated with N_{50} value, allied to a search for conserved core eukaryotic genes, with the aid of BUSCO (SIMÃO et al., 2015), using the *Danio rerio* OrthoDB version 9 database. For the characterization of the functional parts of the genome, we used the MAKER pipeline (CANTAREL et al., 2008) using NCBI's Characiformes database on expressed sequence tags (ESTs) and proteins as training sets. Balancing computational performance, available time, the highly fragmented nature of the published assembly and the average size of the eukaryotic gene (~12 kb) – (COOPER; HAUSMAN, 2016), we restricted this

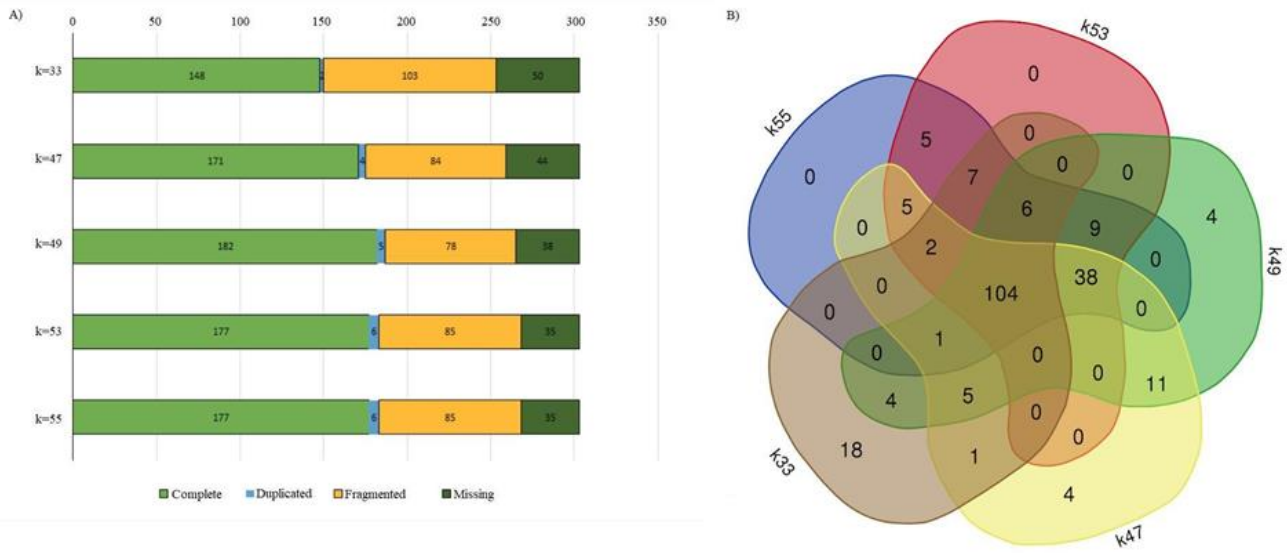
annotation step only for scaffolds over 10 kb (kilobases). For gene classification, we used PANTHER (Protein ANalysis THrough Evolutionary Relationships – (MI et al., 2010) and a local BLAST search against NCBI's Ostariophysi database. All bioinformatic work was performed on the computers at LCC-UFMG, DCOMP-UFSJ and LARGE-UFSJ.

DATA DESCRIPTION

The four new assemblies produced, along with the one presented in Yazbeck *et al*. (2018), had an average total size (including unknown base calls, N) of 1,096,362,179 bp (sd=±11,941,562). The largest assemblies (k=53 e k=55) tie at 1,113,754,917 pb and the shortest (k=49) was 1,086,997,667 pb. N_{50} values for these assemblies varied from 6,729 (k=33) to 8,467 (k=49). These values expose the extensive fragmented nature of obtained assemblies, which reflects the low-coverage nature of the original sequencing experiment (~10-15X). BUSCO analyses results are summarized in Figure. 1A. In average it found 171 (sd=±13.4), out of a list of 303 conserved core eukaryotic genes. Only 104 of these were common to all assemblies (Fig. 1). This illustrates the complexity of *de novo* genome assembly and how different k-mer choices can retrieve different sets of genes, for the same departing short reads dataset.

We selected the assembly resulting from k=55, already published in Yazbeck et al. (2018), to perform the functional annotation, since its N_{50} is only three bases shorter than the one with the largest value, it is some 26 Mb larger than the latter and it showed the smaller number of core eukaryotic genes not detected by BUSCO (Fig. 1B). Also, the former assembly is identical to the one made with k=53.

Figure 1. Summarized results from the evaluation of different *B. orbignyanus* genome assemblies (for different values of k), through BUSCO. A) Number of complete, fragmented, duplicated genes found from a list of 303 Danio rerio genes. B) Venn diagram showing the number of conserved eukaryotic genes shared among alternative assemblies and genes found exclusively in a single assembly.



Dataset

The screening of coding sequences was carried out from scaffolds larger than 10K (totaling just over 500 Mb), from the assembly of the genome k=55 (previously deposited in the Figshare repository, see supplementary materials) led to the annotation of 12,734 predicted genes, for which the characterization of the sequence is revealed herein for the first time (deposited in the Figshare repository, see supplementary materials). The average size of the genes, putative transcripts, exons and introns are detailed in Table 1 and are approximately comparable to those

noted for *Pangasianodon hypophthalmos* (Kim et al., 2018). This result is consolidated in the form of a 2.76 gigabytes General Feature File (GFF3) which can be retrieved from the FigShare repository (see supplementary materials). This file also contains precious data on repeats, transposons and other non-coding DNA elements of *B. orbignyanus*' genome. This information will be useful in different areas, such as taxonomy, ecology, physiology and evolution. The annotation in GFF3 file consists of 9 columns, they are:

- Identity: The name of the scaffold where the annotated sequence is located.
- Method: describes the algorithm or operating procedure that generated the sequence (for example, RepeatMasker or Genbank).
- Type: describes the name of the annotated resource type, such as "exon", "intron" or "Simple repeat".
- Start: Position from which the genomic resource is noted within the scaffold, with a displacement of 1 base.
- End: End of the annotated genomic resource, with an offset of 1 base.
- Score: Value that indicates the source's confidence level in the annotated resource. When it has an indeterminate value, a dot "." Is used.
- Strand: The presence of a single character indicates the strand of the annotated resource: "+" sign (positive or 5'-> 3'), "-", (negative or 3'-> 5') and "." (undetermined).
- Phase: Values 0, 1, 2 are used to indicate the resource phase of the coding sequence (CDS) or "." for other resources.
- Attributes: All other information pertaining to this resource.

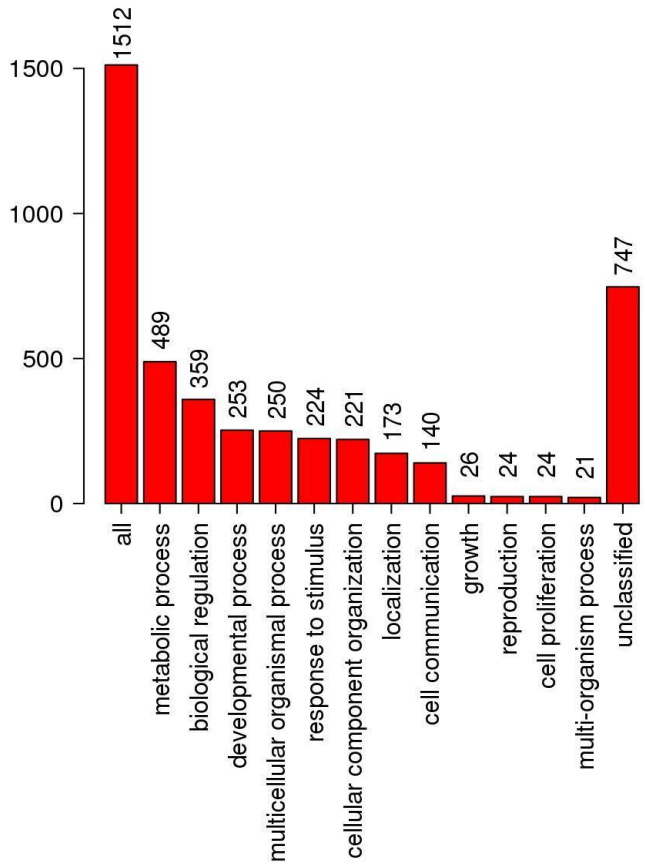
Table 1. Comparison of the current state of *Brycon orbignyanus* genome annotation results with those of three other fish

	Brycon orbignyanus	***Pangasianodon hypophthalmus*[a]**	***Astyanax mexicanus*[b]**	***Danio rerio*[c]**
Number of genes	12,734	28,580	24,690	30,741
Median gene length (bp)	6,170	7,316	11,363	11,007
Median transcripts length (bp)	1,249	978	2,307	1,818
Median exon length (bp)	197	119	137	136
Median intron length (bp)	804	564	891	1,090

[a]Kim et al. (2018); [b]Stockdale et al. (2018); [c]Howe et al. (2013).

Figure 2 summarizes the results obtained from BLAST, according to the Gene Ontology Consortium of 1,512 genes, regarding the classification by biological function, out of 3,808 genes directly identified by this procedure. Roughly, 50% could not be assigned to any know biological process, while 489 genes were found to be related to metabolism, 26 to growth and 24 to reproduction processes.

Figure 2. Gene Ontology classification for 1,512 annotated genes that could be identified through BLAST (against NCBI's Ostariophysi database), regarding known biological function, in *B. orbignyanus*.



Regarding growth hormones systems, two predicted sequences were identified and characterized in the examined assembly for B. orbignyanus, the growth hormone receptor-like gene (GHR) and the insulin like growth factor 2 mRNA binding protein 1 gene (igf2bp1). Growth hormones have been traditionally aimed in hatchery biotechnology, e.g. (MOREAU; CONWAY; FLEMING, 2011), and these data could foster the development of such initiatives for this and closely related species.

Three other predicted genes, related to reproduction were also characterized in the screened portion of the assembly: lutropin-choriogonadotropic hormone receptor (LHCGR); gonadotropin-releasing hormone II receptor (GNRHR2); and gonadotropin subunit beta-like (CGB). These results are presented in Table 2. The use of gonadotropins and gonadotropin releasing hormones has been a viable alternative for fish for a long time now (e.g. (DONALDSON; HUNTER, 1983), although it depends on species-specific hormones for best results. In B. orbignyanus hatcheries, whole carp hypophysis extracts is usually used, which can result in immune reactions and disease transmission (ZOHAR; MYLONAS, 2001). This new functional information could be a first step towards the development new bioproducts and processes useful in B. orbignyanus commercial or environmental rearing operations, since they represent part of the structure of a gonadotropin (CGB) and targeted receptors involved in the sexual maturation process (LHCGR and GNRHR2).

Table 2. Genes related to growth and reproduction predicted from the annotation of *Brycon orbignyanus* genome fragments.

Scaffold ID	Position (start-end)	Description	Nucleotide length (bp)	Amino Acid length (residues)
		Growth related genes		
22651	98-5094	Growth hormone receptor-like (*GHR*)	1,623	540
165032	4481-11259	Insulin like growth factor 2 mRNA binding protein 1 (*igf2bp1*)	1,787	464
		Sexual maturation related genes		
98328	2280-11191	Lutropin-choriogonadotropic hormone receptor (*LHCGR*)	2,015	617
503	6619-12890	Gonadotropin-releasing hormone II receptor (*GNRHR2*)	1,285	381
87608	15135-17881	Gonadotropin subunit beta-like (*CGB*)	676	141

Noteworthy is the mention of the annotation of the full predicted sequence of *B. orbignyanus* LRRC10 protein, at the scaffold 60871 4.4 of the assembly, which has been associated with healthy heart muscle regeneration in *Astyanax mexicanus* (STOCKDALE et al., 2018), and thus potentially holds promising value for human health scientific research.

This dataset present herein considerably broadens the molecular genetics knowledge for this species, by describing almost 13,000 new predicted genes, whereas only 13 mitochondrial gene sequences were previously available. It represents a significant step towards more incisive and dedicated studies into the full genome characterization for *B. orbignyanus*, a *piracema* species with urgent conservation needs and with a potential to diversify fish aquaculture and help in sustainable economic development in South America.

SUPPLEMENTARY MATERIALS

Repository: Figshare

Links to access: https://doi.org/10.6084/m9.figshare.5661802.v1
https://doi.org/10.6084/m9.figshare.16734751.v1
https://doi.org/10.6084/m9.figshare.11793627.v1

ACKNOWLEDGEMENTS

We would also like to thank CNPq, CAPES, FAPEMIG and the CEMIG *Peixe-Vivo* team.

REFERENCES

CANTAREL, B. L. et al. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res, 18:188–196, 2008. DOI: https://doi.org/10.1101/gr.6743907.

COOPER G. M, HAUSMAN, R. E. A Célula: Uma Abordagem Molecular. Artmed Editora: p. 89-138, 2016.

CHIKHI, R.; MEDVEDEV, P. Informed and automated k-mer size selection for genome assembly. Bioinformatics 30:31–37, 2014. DOI: https://doi.org/10.1093/bioinformatics/btt310.

DONALDSON, E. M.; HUNTER, G. A. 7 Induced Final Maturation, Ovulation, and Spermiation in Cultured Fish. In: Hoar WS, Randall DJ, Donaldson EM (eds), Fish Physiology. Academic Press, pp 351–403, 1983.

HOWE, K. et al. The zebrafish reference genome sequence and its relationship to the human genome. Nature, v. 496, n. 7446, p. 498-503, 2013.

KIM, O. T. P. et al. A draft genome of the striped catfish, *Pangasianodon hypophthalmus*, for comparative analysis of genes relevant to development and a resource for aquaculture improvement. BMC Genomics 19:733, 2018. DOI: https://doi.org/10.1186/s12864-018-5079-x.

LOPERA-BARRERO, N. M. Conservation of Brycon orbignyanus natural populations and stocks for their reproductive, genetic, environmental sustainability: A model for species threatened with extinction. Ciencia e investigación agrarian, 36:191–208, 2009. DOI: https://doi.org/10.4067/S0718-16202009000200004.

LUO, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience, 1:18, 2012. DOI: https://doi.org/10.1186/2047-217X-1-18.

MI, H. et al. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Res 38:D204–D210, 2010. DOI: https://doi.org/10.1093/nar/gkp1019.

MOREAU, D. T. R.; CONWAY, C.; FLEMING, I. A. Reproductive performance of alternative male phenotypes of growth hormone transgenic Atlantic salmon (Salmo salar). Evol Appl 4:736–748, 2011. DOI: https://doi.org/10.1111/j.1752-4571.2011.00196.x.

ROSA, R. S.; LIMA, F. C. T. Os peixes brasileiros ameaçados de extinção. In: Livro vermelho da fauna brasileira ameaçada de extinção. Ministério do Meio Ambiente, Brasília, p 278, 2008.

SIMÃO, F. A. et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212, 2015. DOI: https://doi.org/10.1093/bioinformatics/btv351.

STOCKDALE, W. T. et al. Heart Regeneration in the Mexican Cavefish. Cell Rep 25:1997-2007.e7., 2018. DOI: https://doi.org/10.1016/j.celrep.2018.10.072.

TONELLA, L. H. et al. Conservation status and bio-ecology of *Brycon orbignyanus* (Characiformes: Bryconidae), an endemic fish species from the Paraná River basin (Brazil) threatened with extinction. Neotropical Ichthyology, v. 17, 2019. DOI: https://doi.org/10.1590/1982-0224-20190030.

ZOHAR, Y.; MYLONAS, C. C. Endocrine manipulations of spawning in cultured fish: from hormones to genes. Aquaculture 197:99–136; 2001. DOI: tps://doi.org/10.1016/S0044-8486(01)00584-1.

YAZBECK, G. M. et al. A broad genomic panel of microsatellite loci from *Brycon orbignyanus* (Characiformes: Bryconidae) an endangered migratory Neotropical fish. Scientific Reports v. 8, n. 1, p. 1-5, 2018. DOI: https://doi.org/10.1038/s41598-018-26623-x.