Molecular Biology Reports (2018) 45:9–18
https://doi.org/10.1007/s11033-017-4134-z

SHORT COMMUNICATION

Development of microsatellite markers using next-generation sequencing for the fish *Colossoma macropomum*

Raquel B. Ariede[1] · Milena V. Freitas[1] · Milene E. Hata[1] · Vito A. Matrochirico-Filho[1] · Ricardo Utsunomia[2] · Fernando F. Mendonça[3] · Fausto Foresti[2] · Fábio Porto-Foresti[1,4] · Diogo T. Hashimoto[1]

Received: 13 January 2017 / Accepted: 11 December 2017 / Published online: 20 December 2017

Abstract

Tambaqui (*Colossoma macropomum*) is a fish species from the Amazon and Orinoco Rivers, with favorable characteristics to the cultivation system and great market acceptance in South America. However, the construction of a genetic map for the genetic improvement of this species is limited by the low number of molecular markers currently described. Thus, this study aimed to validate gene-associated and anonymous (non-genic) microsatellites obtained by next generation sequencing (RNA-seq and whole genome shotgun—WGS, respectively), for future construction of a genetic map and search for quantitative trait loci (QTL) in this species. In the RNA-seq data, the observed and expected heterozygosity (H_o and H_e) ranged from 0.09 to 0.73, and 0.09 to 0.85, respectively. In the WGS data, H_o and H_e ranged from 0.33 to 0.95, and 0.28 to 0.92, respectively. In general, the evaluation of 200 markers resulted in 45 polymorphic loci, of which 14 were gene-associated (RNA-Seq) and 31 were anonymous (WGS). Moreover, some markers were related to genes of the immune system, biological regulation/control and biogenesis. This study contributes to increase the number of molecular markers available for genetic studies in *C. macropomum*, which will allow the development of breeding programs assisted by molecular markers.

Keywords NGS · Aquaculture · SSR gene-associated · Tambaqui

Introduction

Tambaqui (*Colossoma macropomum*) (Cuvier, 1818) is a migratory fish from the Amazon and Orinoco Rivers basins, belonging to the Characiformes order [1, 2]. This species has a considerable economic importance for aquaculture, mainly due to its high nutritional value of meat, fast growth, captive fitness, resistant and suitable for captive breeding [3]. Furthermore, *C. macropomum* is highly used in the Midwest and Southeast regions of Brazil for crossings with other Serrasalmidae species, resulting in interspecific hybrids [4]. In Brazil, *C. macropomum* and its hybrids tambacu (female *C. macropomum* × male *Piaractus mesopotamicus*) and tambatinga (female *C. macropomum* × male *Piaractus brachypomus*) represent the second most produced fish in aquaculture, representing around 36% of the production (173,301 t) [5]. Moreover, tambaqui has also importance in the aquaculture from others countries in South America [6] and, therefore, it is a target species for breeding programs that will increase the production in aquaculture.

Molecular tools have been used in several fish aquaculture species for *loci* mapping of quantitative traits (QTL), which results in important tools to improve breeding programs through the implementation of marker-assisted selection (MAS). This methodology seeks to locate genomic regions related to genetic variations of economically important phenotypes, i.e., molecular markers linked to a determined trait of interest [7]. However, the discovery and characterization of molecular markers are the first step for the implementation of this approach.

To date, there are 41 anonymous microsatellite markers described for *C. macropomum* [8–10], which were the first studies to contribute with genetic data in this species.

✉ Diogo T. Hashimoto
diogo@caunesp.unesp.br

[1] Centro de Aquicultura da Unesp, Universidade Estadual Paulista, Jaboticabal, SP, Brazil

[2] Departamento de Morfologia, IBB, Universidade Estadual Paulista, Botucatu, SP, Brazil

[3] Instituto do Mar, Universidade Federal de São Paulo, UNIFESP, Santos, SP, Brazil

[4] Departamento de Ciências Biológicas, FC, Universidade Estadual Paulista, Bauru, SP, Brazil

However, this amount is not suitable for QTL characterization, which makes necessary to discover new microsatellite markers in order to apply in genetic breeding programs of *C. macropomum*. In general, hundreds of molecular markers are necessary for QTL detection by genetic mapping in fish species [11, 12].

Historically, the traditional method for microsatellite isolation was expensive and time-consuming task, through the construction of microsatellite-enriched genome libraries, cloning and sequencing by Sanger method [13]. Currently, Next Generation Sequencing (NGS) offers significant advantages in terms of time and cost, facilitating the identification of millions of molecular markers [14, 15]. Whole genomic shotgun (WGS) is one of the NGS strategies used for molecular markers discovery, which consists of the random sequencing of DNA genome, a quick process to locate anonymous markers (mainly not gene-associated) which has been frequently used in fish genomes [16, 17]. Moreover, several studies have been adopting the strategy of transcriptome sequencing (RNA-seq), which consists of mRNA (messenger RNA) sequencing of different tissues for the identification of gene-associated microsatellites in fish [18–21]. In this context, the main purpose of this study was the discovery and characterization of new microsatellites markers in *C. macropomum*, which were obtained by NGS, including the sequencing strategies of WGS and RNA-seq that will be useful for QTL analysis and genetic improvement of *C. macropomum* aquaculture.

Materials and methods

WGS sequencing

For WGS (Whole Genome Shotgun) sequencing, 500 ng of total genomic DNA from an individual of *C. macropomum* (collected at the Centro Nacional de Pesquisa e Conservação de Peixes Continentais, CEPTA, Pirassununga, SP, Brazil) was randomly fragmented by nebulization using compressed nitrogen gas (30 psi). The fragmented DNA was purified using the "MinElute PCR Purification Kit" (Qiagen), according to the manufacturer's instructions. The library construction and sequencing was performed by pyrosequencing on a 454 GS-FLX Titanium® equipment (Roche Diagnostics), in the Instituto Agrobiotecnológico de Rosário, INDEAR, Argentina, following the procedures outlined in Margulies et al. [22].

As we sequenced the *C. macropomum* genome in low coverage by WGS, the prospection of microsatellites was performed directly in the reads, i.e., we did not assembled the sequences because it would result in a low number of contigs. The reads of WGS were analyzed to identify *Simple Sequence Repeats* (SSRs) (minimum of five repeats), using the "Msatcommander" [23] software. Primers design was performed in "Primer3" [24] software. In order to remove any mitochondrial and ribosomal contamination, sequences were assembled against the mitochondrial genome of tambaqui (GenBank accession KP188830) and zebrafish ribosomal RNA RefSeqs (NCBI database) using "CLC Genomics Workbench" (version 10). Moreover, others sequences with homology to transposons and microsatellites already described for *C. macropomum* were identified using the BLASTn and, then, they were manually deleted of the WGS data set. Sequences with SSRs of the WGS that showed homology with genes (mRNA) were included in the RNA-seq database.

RNA-sequencing

To perform the transcriptome sequencing, we collected liver samples from 10 specimens of *C. macropomum*, resulting from five different cultivated stocks in Brazil: n = 3, from CEPTA, Pirassununga, SP; n = 2, from CAUNESP, Jaboticabal, SP; n = 1 from Projeto Surubim, Santa Rita do Tocantins, TO; n = 2 from Fazenda São Paulo, Brejinho de Nazaré, TO; and n = 2 from Fazenda Sambaíba, Porto Nacional, TO. RNA was extracted by the Rneasy Mini Kit (Qiagen). We prepared an equimolar pool of total RNA samples (from 10 individuals) to mRNA enrichment with μMACS mRNA Isolation Kit (Miltenyi Biotec). The library construction and sequencing was performed by pyrosequencing on a 454 GS-FLX Titanium® equipment (Roche Diagnostics), in the HELIXXA company (Campinas, SP, Brazil).

The process of *trimming* (*quality score* Q > 20), assembly, and removal of contamination by mitochondrial DNA and ribosomal RNA was performed using "CLC Genomics Workbench" software. The "CD-HIT" (*Weizhong Li's Group*) software was used to remove sequences smaller than 200 bp and redundancy removal with a 90% identity threshold.

Functional annotation of the unique consensus sequences was performed by homology BLASTx searches against the zebrafish (*Danio rerio*) databank at NCBI (National Center for Biotechnology Information) (cutoff E-value of 1E−3) using BLAST2GO software [25] to obtain the putative gene identity. The gene ontology (GO) terms were assigned to each unique gene based on the GO terms annotated to the corresponding homologs in the NCBI database (e-value cutoff 1e−6). The transcripts were further annotated in InterPro, Enzyme Code (EC) and Kyoto Encyclopedia of Genes and

Genomes (KEGG) metabolic pathways analysis through Bidirectional Best Hit method (BBH).

Microsatellites were identified using the "Msatcommander software" [23] and primers were designed in "Primer3" [24].

Validation of the microsatellites

In the sequences that primers were designed, several filtering steps were adopted for SSRs selection. Initially, sequences containing SSRs were de novo assembled to remove duplicated markers using "CLC Genomics Workbench". We firstly selected tetra and trinucleotide motifs and, then, dinucleotides were included to complete the database for validation. Preferentially, we chosen sequences with high number of repeats. For RNA-seq, we identified the SSR position in the gene by BLASTx, i.e., in coding sequence (cds), 3′ or 5′ untranslated region (UTR). We preferentially selected microsatellites located at the 3′ or 5′ UTR of the genes, as they could be more polymorphic than those positioned in cds.

In total, 100 gene-associated SSRs (RNA-seq) and 100 anonymous SSRs (WGS) were used for validation in a wild population of *C. macropomum* (n = 24), from the Curuá-Uná River (2°26′4.20″S 54°7′43.70″O), Amazon basin. The fin samples were provided by Jonas da Paz Aguiar, from the Universidade Federal do Pará (UFPA), Brazil.

The genomic DNA was extracted from fin samples, following the protocol of the "Wizard Genomic DNA Purification Kit" (Promega). For the polymerase chain reaction (PCR), a final volume of 25 µl were used with: 100 µM of each dNTP, 1.5 mM $MgCl_2$, 1× Taq DNA buffer, 0.1 µM of each primer (F and R), 0.5 units of Taq Polymerase (Invitrogen) and 10–50 ng of genomic DNA. The reactions were performed in a thermocycler (ProFlex™ PCR System, Life Technologies) for 30 cycles under the conditions: 30 s at 95 °C, 30 s at 55–60 °C (adjusted for each primer set), and 20 s at 72 °C. DNA fragments were applied on a 2% agarose gel, stained by Nancy (Sigma), to check the occurrence of polymorphism.

Microsatellites that showed polymorphism in 2% agarose gels were analyzed in the sequencer ABI3730 XL DNA Analyzer (Life Technologies) to get a better accuracy in the alleles determination. The sequencing strategy adopted in this study was according to protocols described by Schuelke [26], using the CAGtag primer (5′-CAGTCGGGCGTC ATCA-3′) labeled with the fluorochromes HEX, FAM or NED. The genotyping PCR was performed with the following reagents: 100 µM of each dNTP, 1.5 mM $MgCl_2$, 1× Taq DNA buffer, 0.1 µM of each primer (F and R), 0.01 µM of the CAGtag primer, 0.5 units of Taq Polymerase (Invitrogen) and 10–50 ng of genomic DNA. PCR conditions were: 9 cycles of 95 °C/30 s, 55–60 °C/30 s (adjusted for each primer set), and 72 °C/20 s; and then 30 cycles of 95 °C/30 s, 50 °C/30 s and 72 °C/20 s. PCR products were analyzed by capillary electrophoresis in the equipment ABI3730 XL, using the DS-30 matrix, with the GeneScan 500 ROX dye Size Standard (Thermo). We used the program GeneMapper 4.0 (Applied Biosystems) to analyze the allele sizes.

The parameters of genetic diversity were estimated by using "GENEPOP" [27] and "ARLEQUIN" 3.5.2.2 [28], including the number of observed alleles (Na), observed heterozygosity (H_o), expected heterozygosity (H_e), Chi square tests for Hardy–Weinberg (HWE), linkage disequilibrium (LD), and inbreeding coefficient (F_{is}) according to the Weir and Cockerham [29] parameters. The levels of significance were adjusted to the multiple tests using a Bonferroni correction [30]. The content of the polymorphic information (PIC) was calculated using the "CERVUS 3.0.7" [31] software. To determine possible genotyping errors and the occurrence of null alleles, the "MicroChecker" [32] software was used with null allele frequency (nf) < 0.1 [33, 34].

Results

WGS sequencing

Whole genome sequencing yielded 42,563 reads, which were deposited in the Short Read Archive (SRA) of NCBI under the accession number SRR5122724. In total, 6105 microsatellites were found in this library (Table 1) and flanking primers were designed for 1255 *loci*. We then selected 100 anonymous microsatellites (non-gene association) for validation, of which 31 were polymorphic when tested in the wild population of *C. macropomum* (GenBank accession number KY379117–KY379147). The genotype analysis

Table 1 Data obtained in the transcriptome (RNA-seq) and WGS sequencing

	RNA-seq	WGS
Sequences	277,245	42,563
Total base pairs (bp)	101,959,245	15,604,478
Average read size (bp)	400.9	366.6
Contigs	7119	–
N50	865	–
SSR	748	6105
SSR with primers	233	1255
Dinucleotides	592	5258
Trinucleotides	138	465
Tetranucleotides	14	339
Pentanucleotides	2	22
Hexanucleotides	–	21

Table 2 Characterization of 31 polymorphic anonymous microsatellites (WGS) in tambaqui (*Colossoma macropomum*)

Loci	*Primer*	*Motif*	Ta (°C)	Size (bp)	HWE	F_{is}	H_o	H_e	A	PIC	nf
r415	F: CAGTCGGGCGTCATCAGTCTCTCAGGCTCATGG R: TGGTTATTGGTCGCTTGTCTC	ACTC (7)	60	250–280	0.463	0.026	0.350	0.359	6	0.337	0.018
r800	F: TTCCTCTTTGATCAGGCGGC R: CAGTCGGGCGTCATCATAAAGGGAGGTGGGTGTGAC	AC (12)	60	240–280	0.069	0.143	0.773	0.899	14	0.867	0.064
r846[a,c]	F: CAGTCGGGCGTCATCACTTAACCCAGCCATGCAG R: GGAAACCATGGGCAGGATTG	AG (10)	60	140–190	0.000	0.604	0.364	0.905	15	0.875	0.294
r872	F: GCAATGTCCAGCTCCCTTTC R: CAGTCGGGCGTCATCAGCTCCATGTCTCAGATTAGCC	AGAT (15)	60	240–320	0.104	0.144	0.773	0.900	13	0.868	0.064
r912[a,c]	F: CAGTCGGGCGTCATCACCACATTGACCACTCTGCTAC R: GCCACTACTGTTTCACTGGG	AC (12)	60	160–210	0.000	0.556	0.381	0.847	7	0.803	0.256
r1163	F: CAGTCGGGCGTCATCAACTGTACATCCAAGCCAGG R: TTATGGGTCTTGAGGCTCCC	AC (9)	60	250–280	0.664	0.083	0.478	0.521	3	0.400	0.034
r1247[a,c]	F: CAGTCGGGCGTCATCAGCAATTAGAGCCTGAGTGTGG R: GGCGAACATGGAACTGCATC	AC (17)	60	210–280	0.000	0.351	0.591	0.903	11	0.872	0.164
r1342	F: ACAGACAAGGAAGGAGAGCG R: CAGTCGGGCGTCATCAGCAGGCACCACACTTTGTTC	AG (10)	60	240–300	0.012	0.050	0.818	0.860	10	0.822	0.013
r1366	F: CAGTCGGGCGTCATCATCTCATAGCGGGTCAGTCTG R: CTGGTCTCTGGTCTCCACTG	ACAT (9)	60	240–290	0.546	0.128	0.750	0.857	9	0.819	0.053
r1935	F: ACACCTGCTGCCATAGACTC R: CAGTCGGGCGTCATCATGGTGAGTGAATTGTGTCGC	AG (15)	60	120–170	0.444	0.080	0.727	0.789	7	0.822	0.022
r1986	F: CAGTCGGGCGTCATCATTCCCATGTGTTTGAGAGCG R: CATGACATCAATGCTTACACGC	AC (13)	60	195–225	0.022	0.167	0.739	0.884	12	0.850	0.080
r2355	F: CAGTCGGGCGTCATCAGTACCGTGAGACCAGATTGC R: ATACACGACGCATGCATTCC	AC (8)	60	230–250	0.760	−0.057	0.619	0.587	5	0.537	−0.028
r2823	F: CAGTCGGGCGTCATCACAAAGAACCCTTCCTGGC R: GCGTACTTACAGCGGAACAC	AC (9)	60	205–225	0.178	0.167	0.565	0.676	6	0.614	0.060
r2899	F: CAGTCGGGCGTCATCATTTCAAACCAGGCGTCTTCC R: GAACGGTTCCTTCGCGAATC	AGC (8)	60	110–130	0.474	0.051	0.545	0.574	4	0.465	0.016
r3061[b]	F: CAGTCGGGCGTCATCACCACTCATGACATTTGACCC R: GTGTTGCTGCTCAGAGTGTG	AGAT (9)	60	280–340	0.022	0.214	0.727	0.921	14	0.891	0.093
r3264	F: CAGTCGGGCGTCATCAGGGAGTAAGTGCAGATCCAG R: GCTGTCATACATAGCGGAAGG	AAT (9)	60	300–315	0.384	0.064	0.522	0.557	4	0.467	0.036
r3429	F: CAGTCGGGCGTCATCAGTCAGTAAAGGCGAGTCTC R: TCTTGTCATGTGTAGTGGTGC	AAT (11)	60	300–330	0.511	0.026	0.708	0.727	5	0.662	0.009
r3620	F: CAGTCGGGCGTCATCACTCCACCCAGCCTTACAGAG R: TCAGCTGTCTTACGCTCTCC	AG (12)	60	115–145	0.012	0.272	0.455	0.621	6	0.574	0.112
r3808	F: CAGTCGGGCGTCATCAGTAATAGAGAGCTGGGCCG R: CGGCAGGTCAGTAACAGGAG	AG (10)	60	130–160	0.727	0.045	0.522	0.546	5	0.463	0.020
r4182[c]	F: CAGTCGGGCGTCATCAGTCCTGTAACGTGTCTCAATG R: TTCTACACTCACGCTGCCTC	AC (18)	60	130–170	0.000	0.120	0.810	0.918	13	0.887	0.049

Table 2 (continued)

Loci	*Primer*	*Motif*	Ta (°C)	Size (bp)	HWE	F_{is}	H_o	H_e	A	PIC	nf
r4481	F: CAGCTTTCACTGCACTGAGG R: CAGTCGGGCGTCATCACTCTCACGGGCGAATCTTTC	AC (10)	60	220–265	0.832	0.005	0.818	0.822	9	0.778	−0.008
r4496	F: CATGTGGTGTTGGGCTGAAC R: CAGTCGGGCGTCATCACTGTTGTCAACCCTCCAC	AC (14)	60	220–250	1.000	−0.063	0.524	0.494	3	0.416	−0.041
r4569	F: TGGGTGGAACTGGAAACGAC R: CAGTCGGGCGTCATCACCAGCCATGAAGATACAGC	AAT (10)	60	160–200	1.000	−0.180	0.333	0.284	2	0.239	−0.183
r4722[b]	F: CAGTCGGGCGTCATCAGCTCAGGATTACAGCAGC R: AGCGCGTTTCTATTCAGCTG	AC (16)	60	110–180	0.459	0.053	0.857	0.904	11	0.871	0.016
r4880[a]	F: ACCAAATCAAACAGCTCCGC R: CAGTCGGGCGTCATCAGTCTGCGGTTGGTCAG	AAT (10)	60	290–320	0.012	0.309	0.609	0.874	10	0.839	0.142
r4985	F: TGCCATCTATCTGTCCTGGTG R: CAGTCGGGCGTCATCACCAATAGACCACACACTGC	AGAT (12)	60	240–285	0.073	0.104	0.783	0.870	9	0.837	0.035
r4990[a]	F: CGTCGCCGCAATAACTACTG R: CAGTCGGGCGTCATCATAGCTGCTGGTCAACAAACG	ATC (11)	60	170–280	0.001	0.417	0.500	0.850	8	0.811	0.191
r4992	F: CAGTCGGGCGTCATCATCGTGCTCACCTGGCAAC R: AGCTCCTAAACACTCCCTCC	AC (9)	60	170–190	0.441	−0.055	0.750	0.712	4	0.637	−0.046
r5455	F: CAGTCGGGCGTCATCATTTCGGTGTACTAGATGGATGG R: CGCTGCAGTATAAGTGGTGC	AGAT (9)	60	140–200	0.228	0.121	0.750	0.851	11	0.816	0.055
r5666	F: TATTGGGCAGCTTCAAAGGC R: CAGTCGGGCGTCATCAGGTTGGAAATGGTCATGTGG	ACAT (7)	60	200–300	0.020	0.060	0.708	0.753	10	0.703	0.014
r6071[b]	F: GGACTTACAGTGGATTTGGGC R: CAGTCGGGCGTCATCATGTTCCCTGTAGATGATGTGC	AGAT (11)	60	250–300	0.298	−0.060	0.958	0.905	11	0.875	−0.044

Ta annealing temperature, *p (HWE)* Hardy–Weinberg equilibrium, F_{is} coefficient of inbreeding, H_o observed heterozygosity, H_e expected heterozygosity, *A* number of alleles, *PIC* polymorphic information content, *nf* null allele frequency

[a] Indicates the presence of null alleles

[b] Indicates linkage disequilibrium

[c] Indicates loci in non-accordance of HWE, after Bonferroni correction (p = 0.0011)

showed the following values: average number of alleles was eight, H_o ranged from 0.33 to 0.95, and H_e ranged from 0.28 to 0.92. The indexes of H_o and H_e had an average of 0.64 and 0.75, respectively (Table 2). Four *loci* demonstrated significant deviation from HWE after Bonferroni correction ($p < 0.0011$), of which three were probably due to presence of null alleles (r912, r846 and r1247). High values of null allele frequency (nf) were detected in five *loci* (r912, r846, r4990, r4880 and r1247) ($nf > 0.1$) (Table 2). We found LD between the microsatellites r6071 and r4722, and r6071 and r3061 ($p = 0.0000$).

RNA-Sequencing

The *trimming* process of RNA-seq resulted in 277,245 reads, which were deposited in the Short Read Archive (SRA) of NCBI under the accession number SRR5122711. These reads were assembled in 7119 contigs (Table 1), of which 4637 contigs were functionally annotated by BLASTx and 4028 sequences (86%) were classified in *Gene Ontology* (GO) terms: biological process (47.7%), molecular function (29.3), and cell component (23%). The main GO terms related to Biological Process were cellular process, metabolic process, and single-organism process; to Mollecular Function were binding, catalytic activity, and transporter activity; and to Cell Component were cell, organelle, and membrane (Fig. 1). We found 3760 annotated sequences with Interpro accession numbers. The transcripts characterization in the KEGG database resulted in 1176 transcripts related to 116 metabolic pathways, with the participation of 328 enzymes.

We found 748 contigs containing microsatellites; however, primers were designed for 233 SSRs *loci*. Functional annotation by BLASTx identified 224 microsatellites with gene identity, of which 100 SSRs were chosen for validation. After genotyping, 14 microsatellites demonstrated polymorphism (GenBank accession number KY379103–KY379116). The average number of alleles was four, the H_o ranged from 0.09 to 0.73, and H_e ranged from 0.09 to 0.85. The mean values of H_o and H_e were of 0.46 and 0.51, respectively (Table 3).

After the Bonferroni correction for multiple *loci* ($p = 0.0011$), all the *loci* were in Hardy–Weinberg Equilibrium. We found LD between the *primers* c3843 and c1842 ($p = 0.0002$). No LD was detected between the microsatellites of the WGS and RNA-seq dataset. We found nf in two *loci* (c3818 and c3843) ($nf > 0.1$) (Table 3).

Discussion

To date, three previous studies were conducted to isolate and characterize 41 microsatellite *loci* in *C. macropomum* [8–10]. Here, using a cost-effective strategy by NGS, we increased the number of microsatellites to be used in genetic management and improvement programs of this species. Most of the motifs found in both WGS and RNA-seq data set were dinucleotides (about 80%), which has been frequently reported in several fish species for anonymous or gene-associated SSRs [35–38]. The levels of polymorphism (allele numbers and heterozygosity) from microsatellites previously described in others studies [8–10] were similar

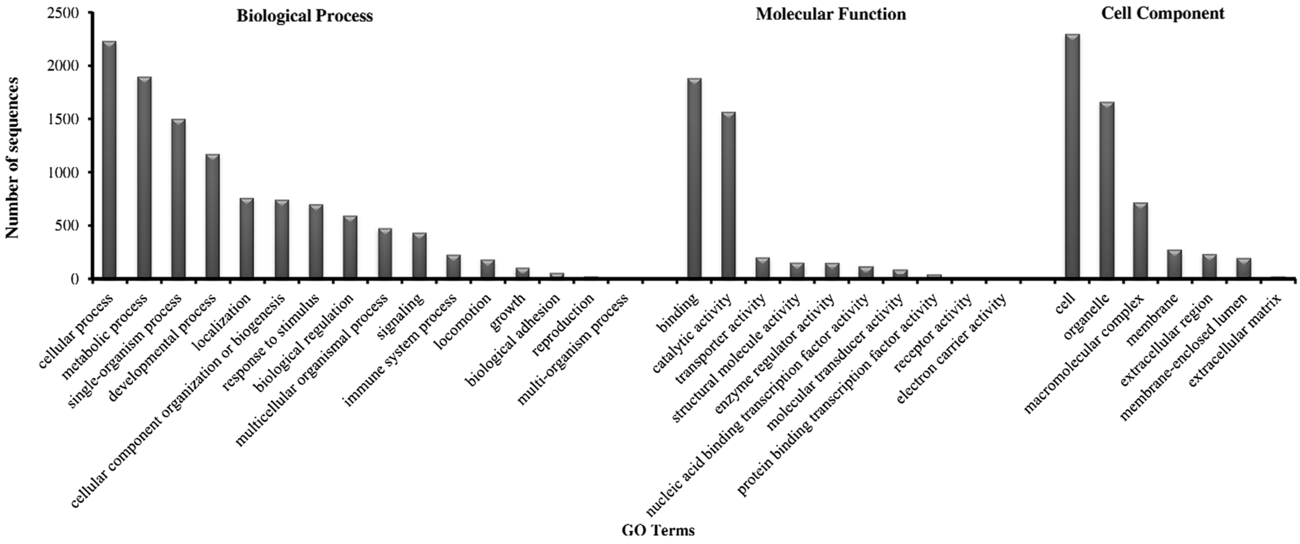

Fig. 1 Gene ontology categories of *Colossoma macropomum* sequences from the RNA-seq data

Table 3 Characterization of 14 polymorphic gene-associated microsatellites (RNA-seq) in tambaqui (*Colossoma macropomum*)

Loci	Gene	Primer	Motif	Ta (°C)	Size (bp)	HWE	F_{is}	H_o	H_e	A	PIC	Gene Position	nf
c841	Uncharacterized protein KIAA0754-like	F: GACGTCAAGAATGCCTGTCG R: CAGTCGGGCGTCATCATTCCAGCGCTGTTCACTCTC	AG(5)	60	210–230	0.559	0.004	0.333	0.334	3	0.303	3′UTR	0.017
c1842[b]	Inositol oxygenase	F: CTACGACACCTGCTGCTTTG R: CAGTCGGGCGTCATCAGCAGAAGGGAGAAAGGTGTG	ATC(7)	60	150–170	1.000	0.007	0.455	0.458	3	0.366	3′UTR	−0.014
c2311	Calnexin precursor	F: CAGTCGGGCGTCATCACAATTCACCGCCTCAGAC R: TCGGCACAGTTAAGGAATGG	ATC(8)	60	150–200	0.357	0.101	0.696	0.772	7	0.719	3′UTR	0.046
c2647	Mannose-1-phosphate guanyltransferase beta	F: CAGTCGGGCGTCATCATCCCAACCCTTACAATACGC R: CTGAGCCTCGCATCATCATG	AC(6)	60	195–215	1.000	−0.026	0.095	0.093	2	0.087	5′UTR	−0.048
c3592	Ranbp1 protein	F: AGGGAGAAGTGTTGCAGGTC R: CAGTCGGGCGTCATCAAAGTGGCGGAGAAACTGG	AG(6)	60	165–185	0.538	0.106	0.273	0.304	2	0.253	5′UTR	0.038
c3818[a]	Ubiquitin carboxyl-terminal hydrolase 44	F: CAGTCGGGCGTCATCAGTCTCTCACGTGCACAC R: TGTAGTAGGGTTCAGCGGTTC	AC(12)	60	200–250	0.003	0.311	0.591	0.851	9	0.813	5′UTR	0.151
c3842	Trinucleotide repeat-containing gene 6B protein-like isoform X5	F: CAGTCGGGCGTCATCAGCTGCTTCCTCTCTGTCC R: GGTCCGACCCATCCACTATC	ACT(5)	60	170–200	0.093	−0.397	0.696	0.502	2	0.371	5′UTR	−0.238
c3843[a,b]	Acyl-CoA-binding domain-containing protein 5A	F: CAGTCGGGCGTCATCATGAACGGGACCTTAGACGG R: TTCACCTGCTCAGCCTCTTC	AGG(6)	60	180–210	0.003	0.483	0.304	0.583	3	0.505	3′UTR	0.242
c3905	Protein asunder homolog	F: AGATGATGTGATGGGCAGGG R: CAGTCGGGCGTCATCACTCTTCCTCTGACAGCCGTG	ATC(5)	58	150–180	1.000	−0.056	0.15	0.142	2	0.129	3′UTR	−0.078
c4296	GTPase IMAP family member 8-like. partial	F: AGCTATTCCTCCTCCCAAACC R: CAGTCGGGCGTCATCACAGCACATATCCAGAGTCC	AG(9)	60	105–125	0.413	0.204	0.583	0.730	5	0.668	5′UTR	0.088
c4604	Arylsulfatase A	F: CAGTCGGGCGTCATCACAGGAGACATGCTAGCTG R: GCATTTCACTCGTACGCTCG	AGC(6)	60	120–160	0.105	0.185	0.542	0.662	5	0.602	3′UTR	0.087
c4706	Beta-1-syntrophin	F: TGAGACGTTCAGCTCCTCAG R: CAGTCGGGCGTCATCATTCCTCTCCGCCAAGATCAG	AGG(5)	60	160–200	0.893	0.068	0.609	0.652	3	0.564	5′UTR	0.023
c5009	Abhydrolase domain-containing protein 3-like	F: CAGTCGGGCGTCATCATCTTCGGTAAACAGGCCATG R: TGGCATTGCGTTAACACACC	AAT(5)	60	120–200	1.000	0.022	0.429	0.438	2	0.336	3′UTR	−0.001

Table 3 (continued)

Loci	Gene	Primer	Motif	Ta (°C)	Size (bp)	HWE	F_{is}	H_o	H_e	A	PIC	Gene Position	nf
c5837	La-related protein 1 isoform X3	F: GTGCTCTGAAGTTCACCGAC R: CAGTCGGGCGTCATCATCTGCCAC TCTGCGTCTTAG	ACTCC(5)	60	160–190	0.511	−0.076	0.739	0.688	6	0.631	3'UTR	−0.036

Ta annealing temperature, *p (HWE)* Hardy–Weinberg equilibrium, F_{is} coefficient of inbreeding, H_o observed heterozygosity, H_e expected heterozygosity, *A* number of alleles, *PIC* polymorphic information content, *nf* null allele frequency, *UTR* untranslated region

[a]Indicates the presence of null alleles

[b]Indicates linkage disequilibrium

to those found in the WGS dataset of the present study. However, the anonymous markers from WGS presented higher rates of polymorphism (31 out of 100 were polymorphic markers), PIC value, allele number, and mean heterozygosity in comparison to the microsatellites from the RNA-seq data (14 polymorphic) (Table 4). In fact, gene-associated microsatellites are more susceptible to selective pressure, which might explain the lower levels of polymorphism in this dataset [39].

Although less polymorphic, transcriptome-derived markers might be used as functional markers in the identification of genes that play important roles in productive traits. For instance, studies of characterization and description of gene-associated microsatellites conducted on barramundi (*Lates calcarifer)*, tilapia (*Oreochromis* sp.), common carp (*Cyprinus carpio*), Atlantic salmon (*Salmo salar*) and Atlantic cod (*Gadus morhua*) [40–44] had already demonstrated the usefulness of these markers in the detection of economically important quantitative traits and its applicability to breeding studies. In the gilthead sea bream (*Sparus aurata*) [45], a dinucleotide microsatellite (alleles 250 and 254) located at the growth hormone (GH) gene was observed in association with fish groups of higher body weight, suggesting that this microsatellite in the promoter region of GH might be considered as a candidate genetic marker for broodstock management and growth selection programs of *Sparus aurata*.

In the present study, we identified different genes related to the immune system, as the *GTPase IMAP* (c4296), which is involved in inflammatory responses caused by the increase of weight and by the consumption of feed containing vegetable products in zebrafish (*Danio rerio*) and Atlantic salmon (*Salmo salar*) [46, 47], respectively; and the *Calnexin precursor* (c2311) gene, which is described as a component of the innate and adaptive immune systems in bony fish [48]. Therefore, future studies concerning these *loci* could improve our knowledge about the association of microsatellites with productive traits. Finally, our study describes 45 novel polymorphic SSRs *loci*, of which 41 (four were in deviations from HWE) are potentially useful for genetic studies of natural and cultivated stocks of *C. macropomum*.

Conclusion

The strategy of NGS allowed the discovery of a high number of microsatellite markers, making available thousands of SSRs for validation in *C. macropomum*. These new microsatellites, added to those already described, will provide important molecular tools to genetic breeding

Table 4 Comparison of our results with others microsatellites previously described in the literature

References	Marker type	Number of markers (motifs)	Allele number	H_e	H_o	PIC
[7]	Anonymous	14 (dinucleotides)	11.5	0.77	0.70	–
[8]	Anonymous	13 (tri and tetranucleotides)	7.0	0.76	0.65	–
[9]	Anonymous	14 (dinucleotides)	11.9	0.85	0.67	0.81
Present study	Anonymous	31 (di, tri and tetranucleotides)	8	0.72	0.65	0.70
Present study	Gene-associated	14 (di, tri and pentanucleotides)	4	0.51	0.46	0.45

H_o observed heterozygosity, H_e expected heterozygosity, *PIC* polymorphic information content

– No data available

programs of *C. macropomum*. In future, these markers could be incorporated in genetic maps to be applied in MAS.

Acknowledgements This work was supported by grants from PROPE/UNESP (07/2015), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq 446779/2014-8 and 305916/2015-7), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP 2014/03772-7), and CAPES. We thanks to Jonas da Paz Aguiar for providing samples of wild individuals from the Curuá-Uná River.

References

1. Eigenmann CH (1915) The Serrasalminae and Mylinae. Ann Carnegie Mus 9:226–271
2. Britski HA (1977) Sobre o gênero *Colossoma* (Pisces, Characidae). Ciência e Cultura 29:810
3. Gomes LC, Simões LN, Araújo-Lima CARM (2010) Tambaqui (*Colossoma macropomum*). In: Baldisserotto B, Gomes LC (eds) Espécies nativas para piscicultura no Brasil, 2nd ed. Editora UFSM, Santa Maria, pp 175–204
4. Hashimoto DT, Senhorini JA, Foresti F, Porto-Foresti F (2012) Interspecific fish hybrids in Brazil: management of genetic resources for sustainable use. Rev Aquacult 4:108–118
5. IBGE (2015) Instituto Brasileiro de Geografia e Estatística. Rio de Janeiro Produção da Pecuária Municipal, vol 43
6. Valladão GMR, Gallani SU, Pilarski F (2016) South American fish for continental aquaculture. Rev Aquacult 0:1–19
7. Souza AP (2001) Biologia molecular aplicada ao melhoramento. In: Nass LL, Valois ACC, Mello IS, Valadares-Inglis MC (eds) Recursos genéticos e melhoramento de plantas. Fundação MT, Rondonópolis, pp 939–965
8. Santos MCF, Hrbek T, Farias IP (2009) Microsatellite markers for the tambaqui (Serrasalmidae, Characiformes), an economically important keystone species of the Amazon River floodplain. Mol Ecol Res 9:874–876
9. Hamoy IG, Cidade FW, Barbosa MS, Gonçalves EC, Santos D (2010) Isolation and characterization of tri and tetranucleotide microsatellite markers for the tambaqui (*Colossoma macropomum*, Serrasalmidae, Characiformes). Conserv Genet Resour 3:33–36
10. Santana GX, Santos CHA, Sousa CFS, Nascimento PRM, Paula-Silva MN, Sousa ACB, Campos T, Almeida-Val VMF (2012) Isolation of novel microsatellite markers for tambaqui (*Colossoma macropomum*, Cuvier 1818), an importante freshwater fish of the Amazon. Conserv Genet Resour 4:197–200
11. Bouza C, Hermida M, Pardo BG et al (2007) A microsatellite genetic map of the turbot (*Scophthalmus maximus*). Genetics 177:2457–2467
12. Ruan X, Wang W, Kong J, Yu F, Huang X (2010) Genetic linkage mapping of turbot (*Scophthalmus maximus* L.) using microsatellite markers and its application in QTL analysis. Aquaculture 308(3):89–100
13. Billotte N, Lagoda PJL, Risterucci AM, Baurens FC (1999) Microsatellite-enriched libraries: applied methodology for the development of SSR markers in tropical crops. Fruits 54:277–288
14. Seeb JE, Carvalho G, Hauser L (2011) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Mol Ecol Res 11:1–8
15. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature 12(7):499–510
16. Carvalho DC, Rodríguez-Zárate CJ, Hammer MP, Beheregaray B (2011) Development of 21 microsatellite markers for the threatened Yarra pygmy perch (*Nannoperca obscura*) through 454 shotgun pyrosequencing. Conserv Genet Resour 3:601–604
17. Carvalho DC, Beheregaray LB (2011) Rapid development of microsatellites for the endangered Neotropical catfish *Conorhynchus conirostris* using a modest amount of 454 shot-gun pyrosequencing. Conserv Genet Resour 3:373–375
18. Liu S, Zhou Z, Lu J (2011) Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array. BMC Genomics 12:53
19. Ji P, Liu G, Xu J, Wang X, Li J, Zhao Z (2012) Characterization of common carp transcriptome: sequencing, de novo assembly, annotation and comparative genomics. PLoS ONE 7(4):e35152
20. Long Y, Li Q, Zhou B (2013) De novo assembly of Mud Loach (*Misgurnus anguillicaudatus*) skin transcriptome to identify putative genes involved in immunity and epidermal mucus secretion. PLoS ONE 8:e56998
21. Gross JB, Furter A, Carlson BM (2013) An integrated transcriptome-wide analysis of cave and surface dwelling *Astyanax mexicanus*. PLoS ONE 8:e55659
22. Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057):376–380
23. Faircloth BC (2008) Msatcommander: detection of microsatellite repeat arrays and automated, locus-specific primer design. Mol Ecol Research 8:92–94
24. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132:365–386
25. Conesa A, Götz S, García-Gómez JM (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676
26. Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. Nat Biotechnol 18:233–234
27. Rousset F (2008) GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. Mol Ecol Resour 8:103–106

28. Excoffier L, Laval G, Schneider S (2005) ARLEQUIN ver. 3.0: an integrated software package for population genetics data analysis. Evol Bioinform Online 1:47–50
29. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370
30. Rice WR (1989) Analyzing tables of statistical tests. Evolution 43:223–225
31. Kalinowsk ST, Taper ML, Marshall TC (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Mol Ecol 16:1006–1099
32. Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) Micro-checker: software for identifying and correcting genotyping errors in microsatellite data. Mol Ecol Notes 4:535–538
33. Chapuis MP, Estoup A (2006) Microsatellite null alleles and estimation of population differentiation. Mol Biol Evol 24(3):621–631
34. Rico C, Cuesta JA, Drake P et al (2017) Null alleles are ubiquitous at microsatellite loci in the Wedge Clam (*Donax trunculus*). PeerJ 5:e3188
35. Carvalho DC, Rodríguez-Zárate CJ, Hammer MP et al (2011) Development of 21 microsatellite markers for the threatened Yarra pygmy perch (*Nannoperca obscura*) through 454 shot-gun pyrosequencing. Conserv Genet Resour 3(4):601–604
36. Chistiakov DA, Hellemans B, Volckaert FA (2006) Microsatellites and their genomic distribution, evolution, function and applications: a review with special reference to fish genetics. Aquaculture 255(1):1–29
37. Mandal S, Jena JK, Singh RK et al (2016) De novo development and characterization of polymorphic microsatellite markers in a schilbid catfish, *Silonia silondia* (Hamilton, 1822) and their validation for population genetic studies. Mol Biol Rep 43(2):91–98
38. Zheng X, Kuang Y, Lü W et al (2014) Transcriptome-derived EST–SSR markers and their correlations with growth traits in crucian carp *Carassius auratus*. Fish Sci 80(5):977–984
39. Papetti C, Harms L, Jürgens J et al (2016) Microsatellite markers for the notothenioid fish *Lepidonotothen nudifrons* and two congeneric species. BMC Res Notes 9:238
40. Yue GH, Li Y, Orban L (2001) Characterization of microsatellites in the IGF-2 and GH genes of Asian seabass (*Lates calcarifer*). Mar Biotechnol 3:1–3
41. Yue GH, Orban L (2002) Microsatellites from genes show polymorphism in two related *Oreochromis* species. Mol Ecol Notes 2:99–100
42. Yue GH, Ho MY, Orban L, Komen J (2003) Microsatellites within genes and ESTs of common carp and their applicability in silver crucian carp. Aquaculture 234:85–98
43. Vasemägi A, Nilsson J, Primmer CR (2005) Expressed sequence tag-linked microsatellite as a source of gene-associated polymorphisms for detecting signatures of divergente selection in Atlantic salmon (*Salmo salar* L.). Mol Biol Evol 22(4):1067–1076
44. Stenvik J, Wesmajervi MS, Fjalestad KT, Damsgard B, Delghandi M (2006) Development of 25 gene-associated microsatellite markers of Atlantic cod (*Gadus morhua* L.). Mol Ecol Notes 6:1105–1107
45. Almuly R, Poleg-Danin Y, Gorshkov S, Rapoport B, Soller M, Kashi Y, Funkenstein B (2005) Characterization of the 5′ flanking region of the growth hormone gene of the marine teleost, gilthead sea bream *Sparus aurata*: analysis of a polymorphic microsatellite in the proximal promoter. Fish Sci 71(3):479–490
46. Drew RE, Settles ML, Churchill EJ, Williams SM, Balli S, Robison BD (2012) Brain transcriptome variation among behaviorally distinct strains of zebrafish (*Danio rerio*). BMC Genomics 13:323
47. Sahlmann C, Sutherland BJG, Kortner TM, Koop BF, Krogdahl A, Bakke AM (2012) Early response of gene expression in the distal intestine of Atlantic salmon (*Salmo salar* L.) during the development of soybean meal induced enteritis. Fish Shellfish Immunol 34:599e609
48. Vasta GR, Nita-Lazar M, Giomarelli B et al (2011) Structural and functional diversity of the lectin repertoire in teleost fish: relevance to innate and adaptive immunity. Dev Comp Immunol 35(12):1388–1399