Title: The liver transcriptome of suckermouth armoured catfish (*Pterygoplichthys anisitsi*, Loricariidae): Identification of expansions in defensome gene families

Authors: Thiago E. Parente[1,2,5*], Daniel A. Moreira[1], Maithê G. P. Magalhães[1], Paula C. C. de Andrade[1], Carolina Furtado[3], Brian J. Haas[4], John J. Stegeman[5], Mark E. Hahn[5]

Addresses: [1] Laboratório de Toxicologia Ambiental, Escola Nacional de Saúde Pública (ENSP), Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, 21040-900, Brasil. [2] Laboratório de Genética Molecular de Microrganismos, Instituto Oswaldo Cruz (IOC), Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, 21040-900, Brasil. [3] Unidade de Genômica, Instituto Nacional do Cancer (INCA), Rio de Janeiro, 20230-130, Brasil. [4] Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA, 02142, USA. [5] Woods Hole Oceanographic Institution (WHOI), Woods Hole, MA, 02543, USA.

Corresponding author: Thiago Estevam Parente

Laboratório de Toxicologia Ambiental, Escola Nacional de Saúde Pública (ENSP), Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, 21040-900, Brasil.

Fax: +55 21 38829113 E-mail: parente@ensp.fiocruz.br & tparente@whoi.edu (TEP)

Abstract

Pterygoplichthys is a genus of related suckermouth armoured catfish native to South America that has invaded tropical and subtropical regions worldwide. Physiological features, including an augmented resistance to organic xenobiotics, may have aided their settlement in foreign habitats. The liver transcriptome of *Pterygoplichthys anisitsi* was sequenced and used to characterize the diversity of mRNAs potentially involved in the responses to natural and anthropogenic chemicals. In total, 66,642 transcripts were assembled. Among the identified defensome genes, cytochromes P450 (CYP) were the most abundant, followed by nuclear receptors (NR), sulfotransferases (SULT) and ATP binding cassette transporters (ABC). A novel expansion in the CYP2Y subfamily was identified, as well as an independent expansion of the CYP2AAs. Two expansions were also observed among SULT1. Thirty-nine transcripts were classified into twelve subfamilies of NR, while 21 encoded ABC transporters. The diversity of defensome transcripts sequenced herein could contribute to this species resistance to organic xenobiotics.

Keywords: RNA-Seq; P450; SULT; ABC transporters; Nuclear Receptors

1 Introduction

Pterygoplichthys is a genus of suckermouth armoured catfish (Siluriformes: Loricariidae) native and abundant in rivers from South America (Lujan et al., 2015). Due to its popularity in the international aquarium trade, followed by intentional or accidental releases, different *Pterygoplichthys* species (e.g.: *P. anisitsi*, *P. pardalis* and *P. disjunctivus*) have established invasive populations in tropical and subtropical regions throughout the globe (Bijukumar et al., 2015; Chavez et al., 2006; Gibbs et al., 2013; Jumawan et al., 2011; Jumawan and Herrera, 2015; Nico et al., 2009). These invasive populations date back to the late 1950's, threaten endangered native species and reach densities two orders of magnitude greater than the native fish biomass (Capps and Flecker, 2013; Courtenay et al., 1974; Nico et al., 2009).

Apart from the lack of natural predators, *Pterygoplichthys* spp. are known to have several distinctive features that might aid their rapid establishment in non-native habitats (Douglas et al., 2002; Ebenstein et al., 2015; Geerinckx et al., 2011; German and Bittong, 2009; Harter et al., 2014; Jumawan and Herrera, 2015; Villalba-Villalba et al., 2013). Among these features, the modified stomach of loricariid catfishes allows the absorption of oxygen through a well documented air-breathing behavior, making these species highly resistant to hypoxia (da Cruz et al., 2013). In fact, according to da Cruz et al. 2013 and CETESB, 2010, *P. anisitsi* is the only fish species able to survive in a river with extremely low O_2 concentration and poor water quality for sustaining aquatic life (CETESB, 2010; da Cruz et al., 2013).

Moreover, *Pterygoplichthys anisitsi* has been shown to be more resistant than other fish species (e.g. Tilapia, *Oreochromis niloticus*) to biodiesel, showing no mortality upon exposure to 6.0 $mL.L^{-1}$ during 15 days (Felício et al., 2015; Nogueira

et al., 2011b). The molecular mechanisms underlying *P. anisitsi* resistance to organic xenobiotics have not been established. However, the cytochrome P450 1A (CYP1A) from *Pterigoplichthys* spp. and some species of the closely related genus *Hypostomus* has been shown to possess amino acid substitutions that alter their substrate specificities, resulting in undetectable or extremely low ethoxyresorufin-*O*-deethylase (EROD) activity in the liver of these fishes (Felício et al., 2015; Nogueira et al., 2011a; Parente et al., 2015, 2014, 2011, 2009). Among Vertebrates, this is an aberrant phenotype of a crucial detoxification enzyme known to take part, for example, in the activation of pre-mutagenic toxins that could potentially be involved in the elevated resistance of *P. anisitsi* to biodiesel.

The aim of this study is to obtain a genome-wide view of the capacity of *P. anisitsi* to handle xenobiotic chemical exposure. This was made possible by the generation of a valuable genetic resource through the sequencing, assembly, and annotation of this species' liver transcriptome. The assembled transcripts were used to infer the mitochondrial genome and the molecular biodiversity of candidate mRNAs for proteins potentially involved in the resistance of this non-model and invasive species to organic toxins.

2 Material and Methods

2.1 Fish sampling

Liver tissue preserved in RNAlater (Life Technologies), from three individuals of suckermouth armoured catfish (*Pterygoplichthys anisitsi*, Loricariidae) collected in the vicinities of Jabuticabal, São Paulo, Brazil, were kindly donated by Prof. Eduardo Almeida from the São Paulo State University (UNESP). Fish handling was carried out in accordance with relevant guidelines and approved by the Ethics Committee, as described elsewhere (Felício et al., 2015). RNA was extracted using either Trizol (Invitrogen) or TRI Reagent (Life technologies) following manufacturer instructions. Briefly, small pieces of tissue were homogenized in Trizol or TRI Reagent using a polytron (T10 Basic ULTRA TURREX, IKA). The homogenate was incubated on ice for 5 min, and then centrifuged for 10 min at 12,000g at 4°C. The supernatant was transferred to another 1.5mL RNase-free and DNase-free plastic tube, in which chloroform was added, vigorously mixed, kept on ice for 15min, and centrifuged for 10 min at 12,000g at 4°C. The aqueous phase was transferred to another tube, mixed with isopropanol, kept on ice for at least 10 min, and centrifuged for 10 min at 12,000g at 4°C. The supernatant was removed and the RNA pellet was washed three times by centrifuging for 2 min at 7500 g at 4°C with 75% ethanol, and dissolved in ultrapure RNase-free water. After extraction, the RNA preparations were quantified using a BioDrop ulite spectrophotometer (Biodrop). RNA quality was evaluated using the kit RNA 6000 Nano for Bioanalyzer (Agilent).

2.2 Library preparation and Illumina sequencing

Libraries of complementary DNA (cDNA) for each individual fish were prepared using 1000 ng of total RNA strictly following the instructions of the TrueSeq RNA Sample kit v2 (Illumina). Each of the three libraries was uniquely identified using specific barcodes. The quality of library preparations was accessed using the DNA 1000 kit for Bioanalyzer (Agilent). Libraries were quantified by qPCR using the Library quantification kit for Illumina GA with revised primers-SYBR fast universal (Kapa Biosystems). The three libraries were clustered, using the TrueSeq PE Cluster kit v3 for cBot (Illumina), in the same lane together with six other samples used in other projects. The 100bp paired-end sequencing reaction was performed in a HiSeq2500 using the TrueSeq SBS kit v3 (Illumina).

2.3 Transcriptome data analyses

Raw Illumina data were demultiplexed using the BCL2FASTQ software (Illumina). Reads were trimmed for Illumina adaptors by Trimmomatic (Bolger et al., 2014) and read quality was evaluated using FastQC (Babraham Bioinformatics). Only reads with PHRED score > 30 were used for the transcriptome assembly. Raw read data of suckermouth catfish liver transcriptome was deposited at the National Center for Biotechnology Information (NCBI) Short Read Archive (SRA) under the accession number of SRR3664326 (single-end) and SRR3664270 (paired-end). This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GETR00000000. The version described in this paper is the first version, GETR01000000.

2.4 Transcriptome assembly and annotation

Cleaned reads from the three individual fish were combined and used for the *de novo* assembly of *Pterygoplichthys anisitsi* transcriptome using the default parameters of Trinity r20131110 (Grabherr et al., 2011; Haas et al., 2013). During the analyses a new Trinity version (2.0.6) was released. The total numbers of assembled transcripts, BLAST hits, and defensome entries were similar using both versions. The Trinotate pipeline was used to achieve a comprehensive functional annotation and analysis (http://trinotate.github.io).

2.5 Mitochondrial genome assembly and annotation

Mitochondrial genome was assembled using the mitochondrial transcripts sequenced in the liver transcriptome, following an approach described elsewhere (Moreira et al., 2015). Briefly, mitochondrial transcripts were retrieved by running a BLASTN search against the mitogenome of the closest related species with a complete mitogenome available, *Pterygoplichthys disjunctivus* (NC_015747.1) (Nakatani et al., 2011). Mitochondrial transcripts were edited according to the information of strand orientation given by the BLASTN result, and aligned with SeaView using the built-in CLUSTAL alignment algorithm and the mitogenome of *P. disjunctivus* (Gouy et al., 2010). The sequence of each CONTIG was manually checked for inconsistencies and gaps. The mitogenome was annotated using the web-based services MitoFish and MITOS (Bernt et al., 2013; Iwasaki et al., 2013). In order to estimate the support of each base in the mitogenome, Bowtie v. 1.0.0 was used to align the reads on the assembled mitogenome. The aligned reads were

viewed using the Integrated Genome Viewer (IGV) or the Tablet (Langmead et al., 2009; Milne et al., 2009; Robinson et al., 2011; Thorvaldsdóttir et al., 2013).

2.6 Defensome gene curation

The assembled transcriptome was subjected to a BLASTX search (E-value < $1e^{-10}$) against two databases, the UniProt entries of humans (*Homo sapiens*), and the Uniprot entries of zebrafish (*Danio rerio*). All the transcrips that had a BLASTX hit with a gene related to the chemical defensome (Goldstone et al., 2006) were retrieved for further analysis.

The retrieved transcripts were aligned to the sequence of their BLASTX top hit with SeaView using the built-in CLUSTAL or MUSCLE alignment algorithm and manually edited to infer the predicted coding sequence (CDS) (Edgar, 2004). The full-length and the partial CDS transcripts that covered >75% of their BLASTX top hit complete CDS were used for phylogenetic analysis. Only defensome gene families with > 15 components were used to build phylogenetic trees. For the construction of phylogenetic trees, the sequences were translated in amino acid, aligned using Muscle and reconstructed using maximum likelihood (PhyML or RAxML algorithm), using the LG model of amino acid substitution optimized for invariable sites and across site rate variation. Branch support was calculated by the approximate likelihood-ratio test (aLRT), using a local computer, and after 1000 bootstrap replicas, using CIPRES Supercomputer (Anisimova and Gascuel, 2006; Felsenstein, 1985; Guindon and Gascuel, 2003). The phylogenetic trees were viewed and edited using FigTree (v1.4.2) (http://tree.bio.ed.ac.uk/software/figtree/).

3 Results and discussion

3.1 Transcriptome assembly and annotation

A total of 60,604,159 100-bp, paired-end reads and 58,617,873 100-bp, single-end reads were generated using Illumina HiSeq2500 technology. After trimming the reads to remove adaptor sequences and after selecting for high quality sequences (Phred score > 30), 177,354,428 reads were used for transcriptome assembly using Trinity **(Table 1)**. In total, 66,642 transcripts were assembled, with a N50 of 1,571bp and an average contig length of 865bp. The BLASTX against *Danio rerio* entries in Uniprot resulted in 28,190 hits (E-value < $1e^{-10}$). The median sequencing depth for the *P. anisitsi* transcripts with BLASTX hit was 13x (average = 646x), and 13% of these transcripts had depth higher than 100x, while 54% were sequenced at a depth higher than 10x **(Table 1, Figure 1)**. The medium ratio between *P. anisitsi* transcript length to the CDS length of its *D. rerio* BLAST top hit (coverage ratio) was 0.9 (average = 1.1), and 47% of these transcripts were longer than their homolog CDS **(Table 1, Figure 1)**. The coverage ratio could often be higher than 1 because for this calculation the entire sequenced transcript was used for *P. anisitsi*, while for *D. rerio* only the complete CDS length was used. Frequently, the *P. anisitsi* transcript include the 5' and the 3'UTR regions, and also unspliced introns.

Figure 1 around here

Transcriptome annotation was also performed using Trinotate, which used the BLASTX algorithm against the general database of Swissprot and Uniref90. *Homo*

sapiens was the most frequent species of the BLASTX top hits, followed by *Mus musculus*, *Rattus norvegicus*, *Danio rerio* and *Pongo abelii* **(Supporting Information Fig. S1)**. Only 73 entries had a Siluriformes fish as the species of the BLASTX top hit. Moreover, 66 of those Siluriformes entries were from a single species, the American channel catfish (*Ictalurus punctatus*). These results contrast with other published transcriptome analysis of fish species that used BLASTX against the NCBI Non-redundant (Nr) database, reflecting the underrepresentation of Siluriformes sequences on more well curated databases (Ali et al., 2014; Zhenzhen et al., 2014).

The transcripts were further classified functionally according their gene ontology (GO) and EggNog IDs. In total, 24,377 Trinity 'genes' had an associated GO term and 12,225 an EggNog term. Of those, there were 10,166 unique GOSlim2 terms and 2,480 unique EggNog terms. Among the transcripts with an assigned GO term, 51% were classified into the Biological Processes category, 29% into Cellular Components, and 20% into Molecular Functions **(Supporting Information Fig. S1)**. The top five GOSlim2 terms for each of the three GO categories were: metabolism (20%), nucleobase, nucleoside, nucleotide and nucleic acid metabolism (8%), biosynthesis (6%), cell organization and biogenesis (6%) and development (6%) for Biological Processes; cell (31%), intracellular (26%), cytoplasm (14%), nucleus (8%) and plasma membrane (3%) for cellular components; and binding (30%), catalytic activity (13%), protein binding (10%), nucleic acid binding (7%), hydrolase activity (5%) for molecular function **(Supporting Information Fig. S1)**. Most of the transcripts with an assigned EggNog term were classified into the 'General function' prediction or into the Function 'unknown classes'. Two other EggNog categories had

> 1,000 entries; Signal transduction mechanisms (1,358 entries) and Posttranslational modification, protein turnover, chaperones (1,078 entries) **(Supporting Information Fig. S1)**.

3.2 Mitochondrial genome assembly and annotation

In order to retrieve mitochondrial transcripts, a BLASTN search of the transcriptome of *P. anisitsi* was performed against the two mitogenomes of Loricariidae species available at the time, *Pterygoplichthys disjunctivus* (NC_015747.1) and *Hypostomus plecostomus* (NC_025584.1) (Liu et al., 2014; Nakatani et al., 2011). In total, 10 transcripts had high score E-values (E-value < $1e^{-10}$). These transcripts were aligned to the *P. disjunctivus* reference mtDNA, covering 96.2% of it, with an average depth of 7,516x, and leaving only six gaps with length varying from 10 to 290 nucleotides, which together sum 632 nucleotides **(Supporting Information Table S2)**. Four of these six gaps had sequences identical between *P. disjunctivus* and *H. plecostomus*, two sister genera, and were most likely to be conserved also in *P. anisitsi*. The sequences of the other two gaps (11,734-11,905 and 15,498-15,788) differed by only a single nucleotide between the mitochondrial genome of *P. disjunctivus* and *H. plecostomus.* The unique features of the mitogenome of *P. disjunctivus* not sequenced in the mitogenome of *P. anisitsi* were six transfer RNAs (tRNAs), tRNA-Val, tRNA-Leu, tRNA-Ser, tRNA-His, tRNA-Pro and tRNA-Thr **(Figure 2).** Among all the 15,889 aligned bases, the mitochondrial genome of *Pterygoplichthys anisitsi* differs from that of *P. disjunctivus* by seven **(Supporting Information Table S2)**, and by 24 nucleotides from the mitogenome of *H. plecostomus*.

This is the sixth published mitogenome of a member of Loricariidae, a family with more than 800 valid species (Lujan et al., 2015). Apart from the two Loricariidae mitogenomes mentioned above, three other (from *Hypoptopoma incognitum*, Accession: KT033767, from *Ancistrus spp*., Accession: KP960569, and from the

endangered species Hypancistrus zebra, Accession: KX611143.1) have recently been published by our group (Magalhães et al., 2016; Moreira et al., 2016, 2015).

Figure 2 around here

3.3 DEFENSOME GENES

The transcripts for which the BLASTX top hit was a gene involved in cellular defense against toxic chemicals were retrieved from the transcriptome. The gene families that comprise the chemical defensome were selected according to the classification of Goldstone *et al.* (Goldstone et al., 2006) and are shown in the **Supporting Information Table S3.** This support table also shows the terms used in the searches, the number of retrieved components of the *P. anisitsi* transcriptome after the BLASTX search against UniProt database for human and zebrafish, as well as the total number of entries for both reference species. The retrieved transcripts were edited and annotated. The coding sequence (CDS) of transcripts encoding for complete CDS of their top BLASTX hits were used as queries for a second round of BLASTX of the *P. anisitsi* transcriptome.

The defensome transcripts retrieved after the second round of BLASTX were manually curated. After this process, 558 components identified in the *P. anisitsi* transcriptome coded for a defensome gene. For some defensome gene families (e.g. Cytochromes P450), this number seems to be higher than expected. However, many of these raw counts are likely to represent fragments of the same transcript and, therefore, could collapse and merge as more genetic information on this species become available. The 183 transcripts coding for a complete coding sequence

(CDS) and at least part of the 5' and 3' untranslated regions (UTRs) should be a better estimate of the real number of defensome genes in the genome of this catfish **(Table 2 and Supporting Information Table S4)**. Cytochromes P450 (CYP), with 43 complete CDS, was the most abundant gene family found in the hepatic transcriptome of *P. anisitsi*, followed by sulfotransferases (SULT) with 33 complete CDS, nuclear receptors (NR) with 32 complete CDS and ATP binding cassette (ABC) transporters with 21 complete CDS. The identification codes for the transcripts covering the complete CDS for all the defensome genes and the ones covering > 75% of the CDS of CYP, NR, SULT and ABC transporters are shown on **Supporting Information Table S5**. Three fragments of AHR gene were identified to align with high percentage identity and E-values to distinct regions of AHR2, with *Danio rerio* and *Carassius auratus* the two most frequent species of the BLASTX hits. Additionally, full-length or nearly full-length transcripts encoding ARNT1, ARNT2, BMAL1 (ARNT-Like 1), and BMAL2 (ARNT-Like 2) were sequenced. Partial sequences encoding NF-E2 and NFE2-Like 1 (NRF1) were also identified, but there were no transcripts identified encoding a NRF2 homolog.

3.4 Cytochromes P450

Cytochromes P450 (CYP) are an ancient superfamily of genes found in all domains of life. CYP genes code for enzymes involved both in the metabolism of endogenous compounds and in the biotransformation of xenobiotics. An astonishing (and still fast growing) diversity of CYP genes has been described (Nelson et al., 2013). Recent analysis of vertebrate genomes reveals that the number of CYP genes in those species range from 50 to more than 100 (Kirischian et al., 2011). In this study, 159 distinct CYP transcripts were detected in the liver transcriptome of *P. anisitsi*, in addition to four cytochrome P450 reductases (POR). Forty-seven of those transcripts contains more than 75% of the coding sequence of a cytochrome P450, several with the adjacent 5' and 3' UTRs. Identical CDSs with distinct UTR regions are shown in **Table 3**.

Transcripts containing > 75% of the complete CDS of a BLASTX top hit were subjected to phylogenetic analyses. CYP51 from human (NM_000786.3), *D. rerio* (NM_001001730.2) and *P. anisitsi* were used to root the trees, resulting in eight well-supported clades **(Figure 3)**. The CYP2 family represented 55% of the CYP transcripts, and was the most abundant family. According to a recent analysis of CYP2 phylogenetic and functional diversity in vertebrates, 14 CYP2 subfamilies have been identified in Actinopterygian species and most vertebrate species are expected to have between 12 and 20 CYP2 genes (Kirischian et al., 2011). We obtained the complete CDS for 25 CYP2 genes in *P. anisitsi*, which were classified into six subfamilies.

Our phylogenetic analysis of CYP2 genes conforms to the one published by Kirischian et al., 2011**,** except for the placement of CYP2AA. CYP2U is a basal

CYP2 subfamily that led to the divergence of two major CYP2 clades (**Figure 3)**. One of these clades is composed of multiple genes in a CYP2Y subfamily, having CYP2M at the base. We detected 12 distinct complete CDS of CYP2Y transcripts in *P. anisitsi*. Differences among these transcripts varied from two amino acids, between CYP2Y#2 and CYP2Y#3, to 90 amino acids, between CYP2Y#3 and CYP2#11. The exact number of correspondent genes cannot be determined, but this might represent a large expansion of this subfamily, which in zebrafish is composed by only one member. Likewise, in this and other similar cases below, numbers were assigned to each transcript only to discriminate them in the context of this manuscript and does not reflect an official nomenclature. As for the CYP2M, its distribution was restricted to a salmonid species (Yang et al., 1998). Recently, however, a CYP2M sequence was reported in *Ictalurus punctatus* (Liu et al., 2012). In this study, we have identified an isoform of CYP2M in *P. anisitsi*. Interestingly, CYP2M was not found in the genome of zebrafish (*Danio rerio*), which is more closely related to the Siluriformes species (superorder Ostariophysi) than to salmonids (superorder Protacanthopterygii). According to (Kirischian et al., 2011), the CYP2M subfamily is a sister group of all mammalian CYP2 genes, except for the CYP2W subfamily.

The other major CYP2 clade shows a second bifurcation. One branch is composed of genes in the CYP2AA subfamily, which was recently described in the zebrafish genome. While zebrafish has 10 CYP2AA genes (Kubota et al., 2013), eight paralogs were sequenced in the liver transcriptome of *P. anisitsi*. The number of amino acid changes among *P. anisitsi* CYP2AAs varied from eight, between CYP2AA#3 and CYP2AA#4, to 201, between CYP2AA#1 and CYP2AA#7.

Interestingly, our phylogenetic analysis suggests that the expansion of this subfamily occurred independently in both fish species. The other branch is a multi-subfamily agglomerate, in which two isoforms of CYP2AD and one of CYP2Z were identified (Figure 3).

Figure 3 around here

3.5 Sulfotransferases (SULT)

The sulfotransferases (SULT) are cytosolic enzymes able to catalyze the sulfonation of a vast array of endogenous and xenobiotic molecules (James and Ambadapadi, 2013). Humans have 13 SULT genes classified into four subfamilies, SULT1, SULT2, SULT4 and SULT6 (Lindsay et al., 2008). The SULT1 subfamily is the most diverse, with eight genes (SULT1A1-4, SULT1B and SULT1C2-4). The other subfamilies of human SULT have just a single gene. In zebrafish (*Danio rerio*), 20 SULT genes have been identified (Kurogi et al., 2013). SULT3 and SULT5 are subfamilies found in zebrafish, but absent in humans. As in humans, zebrafish SULT1 is the most diverse subfamily, with nine genes, followed by SULT3 with five, SULT2 with three and SULT4, SULT5 and SULT6 with one gene each. The zebrafish SULT1 genes follow a distinct nomenclature, ranging from SULT1st1 to SULT1st8. We have identified 67 transcripts that code for SULT enzymes. Among those transcripts, 36 covered > 75% of the sequence of a SULT protein deposited in UniProt database for human or zebrafish. The phylogenetic relationships of *P. anisitsi* SULTs with their homologs from *Homo sapiens*, *Danio rerio*, and *Ictalurus punctatus* were further investigated (**Figure 4)**.

Two clusters of SULT1 genes were observed in the *P. anisitsi* transcriptome, one more closely related to human SULT1As, and another to zebrafish SULT1st6 and *Ictalurus punctatus* SULT1C1 (**Figure 4)**. Three distinct SULT1A CDS from *P. anisitsi* clustered together as a sister group of the clade formed by another *P. anisitsi* SULT1A CDS, one isoform of *I. punctatus*, and most of zebrafish SULT1st transcripts (SULT1st1-4, 7, 8). The three *P. anisitsi* SULT1As were encoded by 12 transcripts, each one having an unique 3'UTR, one or two amino acids differences in

their CDS, and two distinct 5'UTR **(Table 3)**. The more distal *P. anisitsi* SULT1A differed from the others by up to 91 amino acids. The other SULT1 cluster in *P. anisitsi* is formed by six distinct complete CDS (coded by seven transcripts) forming a monophyletic clade with the *I. punctatus* SULT1C1 and the zebrafish SULT1st6 **(Table 3, Figure 4)**. Differences among these transcripts range from a single to 20 amino acids. Basal to this clade is the zebrafish SULT1st5. However, the classification of all the transcripts in this clade as SULT1C is controversial as the three SULT1C transcripts from human form a sister clade to all other SULT1s. Moreover, results indicate the zebrafish SULT1 transcripts are paraphyletic, with SULT1st5 and SULT1st6 being more closely related to the Siluriformes SULT1Cs than to the others zebrafish SULT1st sequences, which in turn are more similar to the SULT1As from human and Siluriformes. In fact, the SULT1st5 from zebrafish is located on chromosome 23 and SULT1st6 is located on chromosome 12, while all others SULT1st are located on chromosome 8 (Kurogi et al., 2013). The chromosomal location of the SULT1st genes from zebrafish corroborates our phylogenetic analysis, which does not support the current nomenclature of SULT1 genes in zebrafish. The fish specific subfamily SULT3 was also expanded in *P. anisitsi*; 14 transcripts were sequenced, 13 different complete CDS, two distinct 5'UTR and seven 3'UTR. *P. anisitsi* SULT3 clustered together with the single isoform known for *I. punctatus* and with the four isoforms of zebrafish.

Figure 4 around here

3.6 Nuclear Receptors (NR)

Nuclear receptors (NR) constitute a superfamily of genes that encode proteins involved in triggering cellular, and ultimately organismal, responses to a diverse range of environmental stimuli. Structurally, NR are composed by two conserved domains: the DNA binding domain (DBD) located at the central part of the protein, and the ligand-binding domain (LBD) at the C-terminal region (Cotnoir-White et al., 2011). The sequence of the DBD is more conserved across the seven NR subfamilies than the sequence of the LBD. Variations inside the LDB are responsible for the specificity of each NR for their ligands, while variations in the DBD distinguish the location where the NR binds to the DNA, triggering distinct responses of gene expression. Among the NR ligands are endogenous compounds (e.g. steroid hormones, vitamin D, retinoic acid and thyroid hormones) and several xenobiotics, as for example: phenobarbital and rifampicin (Pascussi et al., 2008; Xie and Evans, 2001).

Most of the 32 transcripts that code for the complete CDS of nuclear receptors in the transcriptome of *P. anisitsi* have a close homolog in zebrafish. The NR of *P. anisitsi* were classified into twelve subfamilies; NR0B, NR1A, NR1B, NR1F, NR1C, NR1D, NR1H, NR2A, NR2B, NR2F, NR3A and NR5A **(Figure 5)**. Notably, a homolog of NR1I2 (PXR) was sequenced but not included in further analyses as this sequence was only 208 nucleotides long. This fragment, however, shows 66% identity of its inferred amino acid sequence and an E-value of $6e^{-22}$ with the zebrafish homolog. DNA binding domains of *P. anisitsi*'s NR1B1, NR1C3, NR2A1, NR3A1 and NR5A2 are absolutely conserved in comparison to their homolog in zebrafish, while the others have only a few amino acid substitutions. The NR0B transcripts of *P.*

anisitsi, as the NR0B from other species, lack the conventional DBD of nuclear receptors.

Ligand binding domains are slightly different between NRs in *P. anisitsi* and their homologs in other species. Among the most divergent NR sequences are NR2Bs (RXRs). *P. anisitsi* NR2B1 has a 14 amino acid long deletion in the LBD in relation to its zebrafish ortholog (**Supporting Information Fig. S6)**. Three distinct CDS and LBD were found for *P. anisitsi* NR2B2 **(Figure 5)**, each of those coded by two transcripts with different UTR regions. *P. anisitsi* NR2B2#3 differ from its zebrafish ortholog by only four amino acids. However, *P. anisitsi* NR2B2#1 has a 14 amino acid long deletion in the same position as the deletion in NR2B1, while NR2B2#2 has an insertion of 11 amino acids in this region (**Supporting Information Fig. S6)**. Different UTR regions were found for a same CDS of six NR isoforms **(Table 3).**

Figure 5 around here

3.7 ATP Binding Cassette (ABC) transporters

Membrane transporters are crucial to maintain constant over time the electro-chemical gradients across the biological membranes. Active transporters use cellular energy to move molecules in and out of the cell, and through its compartments. ATP Binding Cassette (ABC) transporters hydrolyze ATP to power the transport of ions, nutrients, metabolites and xenobiotics against their concentration gradient (Rees et al., 2009). ABC transporters forms a monophyletic superfamily of genes classified into eight subfamilies according to the sequence similarity at one of its structurally conserved regions, the ATP-binding domain, also known as the nucleotide-binding domains (NBDs) (Dean and Annilo, 2005; Liu et al., 2013).

We have sequenced 113 transcripts for which the top BLAST hit was an ABC transporter. Of those transcripts, 21 had a complete CDS including nucleotides at the 5' and 3' UTR, and 23 coded for > 75% of their BLAST top hit complete CDS (**Table 2)**. A single CDS with two distinct UTR regions was found for ABCB2 and for ABCD3 **(Table 3).** For comparison, 50 ABC transporters genes were recently identified in the genome of another Siluriformes species, *Ictalurus punctatus* (Liu et al., 2013) and also in the marine medaka (Jeong et al., 2015), while zebrafish have 57 (Liu et al., 2013), and humans have 49 ABC transporter genes (Vishwakarma et al., 2014). The 20 *P. anisitsi* unique transcripts coding for > 75% CDS belongs to seven subfamilies; two ABCA isoforms, five ABCB, three ABCC, two ABCD, two ABCE, two ABCF and four ABCG **(Figure 6)**. Our phylogenetic analyses are in accordance with those published before for *I. punctatus* (Liu et al., 2013).

As in other vertebrate species, members of subfamilies ABCD and ABCG code for half transporters, with a single NDB, while ABCB subfamily members have

either half (ABCB3) and full (ABCB11, with two NDBs) transporters, and the other ABC subfamilies code for full transporters. ABCE and ABCF are unique among ABCs as these subfamilies possess two NBD, but no transmembrane domain (TMD) and are, therefore, not functional as transporters proteins (Dean and Annilo, 2005). Similar to other vertebrates, no transmembrane domain was observed in *P. anisitsi* ABCE and ABCF isoforms. In comparison to the sequences in zebrafish, the ABC signature was modified in three transcripts: ABCB3, ABCB3-like and ABCD4. However, when compared to *I. punctatus* the ABC signature was modified only in the sequence of ABCB3, from LS**S**GQ in *I. punctatus* to LS**A**GQ in *P. anisitsi*.

Figure 6 around here

4 Conclusion

The liver transcriptome of *P. anisitsi* was characterized, its mitogenome was assembled and the diversity of a large number of candidate genes involved in this species' resistance to organic contaminants was analyzed. A wide diversity of transcripts encoding enzymes involved in xenobiotic detoxification, especially of CYPs and SULTs, was found at the liver of *P. anisitsi*, which could contribute to this species resistance to organic xenobiotics. Further studies are being conducted to evaluate the modulation of these defensome genes by xenobiotics, and also to characterize the catalytic activity of the encoded proteins toward foreign chemicals. The raw Illumina reads and the assembled transcriptome are available for expanded analyses, and provide a valuable genomic resource for future studies ranging from gene discovery and molecular phylogenetics to control of invasive populations and molecular ecology. Indeed, during the final review of this manuscript a draft genome of the related species *Pterygoplichthys pardalis* was released along with the annotated genome of the channel catfish, *Ictalurus punctatus* (Liu et al., 2016). The genomic data provided by Liu et al., 2016 together with the transcriptomic data provided here can now be used in an iterative process to extend our findings on the diversity of defensome genes in this important group of fish.

Acknowledgments

This study was supported by a PEER grant from USAID (PGA-2000003446 and PGA-2000004790) associated with NSF grant DEB-1120263. T.E.P, D.A.M, and M.G.P.M receives independent fellowships from the Brazilian funding agency CAPES. Preparation of the manuscript was supported also by a Mary Sears Visitor Program Award from WHOI and by the Boston University Superfund Research Program (NIH grant P42ES007381 to M.E.H. and J.J.S.). The authors thank the Program for Technological Development in Tools for Health-PDTIS-FIOCRUZ for use of its facilities. Authors also thank Dr. Eduardo Almeida for donating the fish, and Dr. Mauro Rebelo for suggesting the use of NGS technologies. T.E.P is grateful to José Eduardo Olivé Malhadas for the donation of computer resources.

References

Ali, A., Rexroad, C.E., Thorgaard, G.H., Yao, J., Salem, M., 2014. Characterization of the rainbow trout spleen transcriptome and identification of immune-related genes. Front. Genet. 5, 1–17. doi:10.3389/fgene.2014.00348

Anisimova, M., Gascuel, O., 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst. Biol. 55, 539–52. doi:10.1080/10635150600755453

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Pütz, J., Middendorf, M., Stadler, P.F., 2013. MITOS: improved de novo metazoan mitochondrial genome annotation. Mol. Phylogenet. Evol. 69, 313–319. doi:10.1016/j.ympev.2012.08.023

Bijukumar, A., Smrithy, R., Sureshkumar, U., George, S., 2015. Invasion of South American suckermouth armoured catfishes Pterygoplichthys spp . (Loricariidae) in Kerala , India - a case study. J. Threat. taxa 7, 6987–6995. doi:10.11609/JoTT.o4133.6987-95

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

Capps, K. a, Flecker, A.S., 2013. Invasive aquarium fish transform ecosystem nutrient dynamics. Proc. Biol. Sci. 280, 20131520. doi:10.1098/rspb.2013.1520

CETESB, 2010. Relatório de qualidade das ´guas interiores do estado de São Paulo, São Paulo, Brasil.

Chavez, J.M., De La Paz, R.M., Manohar, S.K., Pagulayan, R.C., Carandang, J.R., 2006. New Philippine record of south american sailfin catfishes (Pisces: Loricariidae). Zootaxa 1109, 57–68. doi:10.11646/zootaxa.1109.1

Cotnoir-White, D., Laperrière, D., Mader, S., 2011. Evolution of the repertoire of nuclear receptor binding sites in genomes. Mol. Cell. Endocrinol. 334, 76–82. doi:10.1016/j.mce.2010.10.021

Courtenay, W.R., Sahlman, H.F., Miley, W.W., Herrema, D.J., 1974. Exotic fishes in fresh and brackish waters of Florida. Biol. Conserv. 6, 292–302. doi:10.1016/0006-3207(74)90008-1

da Cruz, A.L., da Silva, H.R., Lundstedt, L.M., Schwantes, A.R., Moraes, G., Klein, W., Fernandes, M.N., 2013. Air-breathing behavior and physiological responses to hypoxia and air exposure in the air-breathing loricariid fish, Pterygoplichthys anisitsi. Fish Physiol. Biochem. 39, 243–256. doi:10.1007/s10695-012-9695-0

Dean, M., Annilo, T., 2005. Evolution of the ATP-binding cassette (ABC) transporter superfamily in vertebrates. Annu. Rev. Genomics Hum. Genet. 6, 123–142. doi:10.1146/annurev.genom.6.080604.162122

Douglas, R.H., Collin, S.P., Corrigan, J., 2002. The eyes of suckermouth armoured catfish (Loricariidae, subfamily Hypostomus): pupil response, lenticular longitudinal spherical aberration and retinal topography. J. Exp. Biol. 205, 3425–3433.

Ebenstein, D., Calderon, C., Troncoso, O.P., Torres, F.G., 2015. Characterization of dermal plates from armored catfish Pterygoplichthys pardalis reveals sandwich-like nanocomposite structure. J. Mech. Behav. Biomed. Mater. 45, 175–182. doi:10.1016/j.jmbbm.2015.02.002

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. doi:10.1093/nar/gkh340

Felício, A.A., Parente, T.E.M., Maschio, L.R., Nogueira, L., Venancio, L.P.R., Rebelo, M.D.F., Schlenk, D., De Almeida, E.A., 2015. Biochemical responses, morphometric changes, genotoxic effects and CYP1A expression in the armored catfish Pterygoplichthys anisitsi after 15 days of exposure to mineral diesel and biodiesel. Ecotoxicol. Environ. Saf. 115, 26–32. doi:10.1016/j.ecoenv.2015.01.034

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution (N. Y). 39, 783–791.

Geerinckx, T., Herrel, A., Adriaens, D., 2011. Suckermouth armored catfish resolve the paradox of simultaneous respiration and suction attachment: a kinematic study of Pterygoplichthys disjunctivus. J. Exp. Zool. Part A Ecol. Genet. Physiol. 315 A, 121–131. doi:10.1002/jez.656

German, D.P., Bittong, R. a., 2009. Digestive enzyme activities and gastrointestinal fermentation in wood-eating catfishes. J. Comp. Physiol. B Biochem. Syst. Environ. Physiol. 179, 1025–1042. doi:10.1007/s00360-009-0383-z

Gibbs, M. a., Kurth, B.N., Bridges, C.D., 2013. Age and growth of the loricariid catfish Pterygoplichthys disjunctivus in Volusia Blue Spring, Florida. Aquat. Invasions 8, 207–218. doi:10.3391/ai.2013.8.2.08

Goldstone, J. V, Hamdoun, A., Cole, B.J., Howard-Ashby, M., Nebert, D.W., Scally, M., Dean, M., Epel, D., Hahn, M.E., Stegeman, J.J., 2006. The chemical defensome: Environmental sensing and response genes in the Strongylocentrotus purpuratus genome. Dev. Biol. 300, 366–384. doi:http://dx.doi.org/10.1016/j.ydbio.2006.08.066

Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Mol. Biol. Evol. 27, 221–224. doi:10.1093/molbev/msp259

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

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704. doi:10.1080/10635150390235520

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., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N.,

Henschel, R., LeDuc, R.D., Friedman, N., Regev, A., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8, 1494–1512. doi:10.1038/nprot.2013.084

Harter, T.S., Shartau, R.B., Baker, D.W., Jackson, D.C., Val, a. L., Brauner, C.J., 2014. Preferential intracellular pH regulation represents a general pattern of pH homeostasis during acid–base disturbances in the armoured catfish, Pterygoplichthys pardalis. J. Comp. Physiol. B 184, 709–718. doi:10.1007/s00360-014-0838-8

Iwasaki, W., Fukunaga, T., Isagozawa, R., Yamada, K., Maeda, Y., Satoh, T.P., Sado, T., Mabuchi, K., Takeshima, H., Miya, M., Nishida, M., 2013. Mitofish and mitoannotator: A mitochondrial genome database of fish with an accurate and automatic annotation pipeline. Mol. Biol. Evol. 30, 2531–2540. doi:10.1093/molbev/mst141

James, M.O., Ambadapadi, S., 2013. Interactions of cytosolic sulfotransferases with xenobiotics. Drug Metab. Rev. 45, 401–414. doi:10.3109/03602532.2013.835613

Jeong, C.-B., Kim, B.-M., Kang, H.-M., Choi, I.-Y., Rhee, J.-S., Lee, J.-S., 2015. Marine medaka ATP-binding cassette (ABC) superfamily and new insight into teleost Abch nomenclature. Sci. Rep. 5, 15409. doi:10.1038/srep15409

Jumawan, J.C., Herrera, A. a., 2015. Histological and ultrastructural characteristics of the testis of the invasive suckermouth sailfin catfish Pterygoplichthys disjunctivus (Siluriformes: loricariidae) from Marikina River, Philippines. Tissue Cell 47, 17–26. doi:10.1016/j.tice.2014.10.005

Jumawan, J.C., Vallejo, B.M., Herrera, A. a, Buerano, C.C., Fontanilla, I.K.C., 2011. DNA barcodes of the suckermouth sailfin catfish. Philipp. Sci. Lett. 4, 103–113.

Kirischian, N., McArthur, A.G., Jesuthasan, C., Krattenmacher, B., Wilson, J.Y., 2011. Phylogenetic and functional analysis of the vertebrate cytochrome P450 2 family. J. Mol. Evol. 72, 56–71. doi:10.1007/s00239-010-9402-7

Kubota, A., Bainy, A.C.D., Woodin, B.R., Goldstone, J. V, Stegeman, J.J., 2013. The cytochrome P450 2AA gene cluster in zebrafish (Danio rerio): Expression of CYP2AA1 and CYP2AA2 and response to phenobarbital-type inducers. Toxicol. Appl. Pharmacol. 272, 172–179. doi:http://dx.doi.org/10.1016/j.taap.2013.05.017

Kurogi, K., Liu, T.-A., Sakakibara, Y., Suiko, M., Liu, M.-C., 2013. The use of zebrafish as a model system for investigating the role of the SULTs in the metabolism of endogenous compounds and xenobiotics. Drug Metab. Rev. 45, 431–440. doi:10.3109/03602532.2013.835629

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25. doi:10.1186/gb-2009-10-3-r25

Lindsay, J., Wang, L.-L., Li, Y., Zhou, S.-F., 2008. Structure, function and polymorphism of human cytosolic sulfotransferases. Curr. Drug Metab. 9, 99–105. doi:10.2174/138920008783571819

Liu, S., Li, Q., Liu, Z., 2013. Genome-Wide Identification, Characterization and Phylogenetic Analysis of 50 Catfish ATP-Binding Cassette (ABC) Transporter

Genes. PLoS One 8, 1–17. doi:10.1371/journal.pone.0063895

Liu, S., Zhang, J., Yao, J., Liu, Z., 2014. The complete mitochondrial genome of the armored catfish, Hypostomus plecostomus (Siluriformes: Loricariidae). Mitochondrial DNA 1736, 1–2. doi:10.3109/19401736.2014.971281

Liu, S., Zhang, Y., Zhou, Z., Waldbieser, G., Sun, F., Lu, J., Zhang, J., Jiang, Y., Zhang, H., Wang, X., Rajendran, K. V, Khoo, L., Kucuktas, H., Peatman, E., Liu, Z., 2012. Efficient assembly and annotation of the transcriptome of catfish by RNA-Seq analysis of a doubled haploid homozygote. BMC Genomics 13, 595. doi:10.1186/1471-2164-13-595

Liu, Z., Liu, S., Yao, J., Bao, L., Zhang, J., Li, Y., Jiang, C., Sun, L., Wang, R., Zhang, Y., Zhou, T., Zeng, Q., Fu, Q., Gao, S., Li, N., Koren, S., Jiang, Y., Zimin, A., Xu, P., Phillippy, A.M., Geng, X., Song, L., Sun, F., Li, C., Wang, X., Chen, A., Jin, Y., Yuan, Z., Yang, Y., Tan, S., Peatman, E., Lu, J., Qin, Z., Dunham, R., Li, Z., Sonstegard, T., Feng, J., Danzmann, R.G., Schroeder, S., Scheffler, B., Duke, M. V., Ballard, L., Kucuktas, H., Kaltenboeck, L., Liu, H., Armbruster, J., Xie, Y., Kirby, M.L., Tian, Y., Flanagan, M.E., Mu, W., Waldbieser, G.C., 2016. The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. Nat. Commun. 7, 11757. doi:10.1038/ncomms11757

Lujan, N.K., Armbruster, J.W., Lovejoy, N., López-fernández, H., 2015. Multilocus molecular phylogeny of the suckermouth armored catfishes (Siluriformes: Loricariidae) with a focus on subfamily Hypostominae. Mol. Phylogenet. Evol. 82, 269–288. doi:10.1016/j.ympev.2014.08.020

Magalhães, M.G.P., Moreira, D.A., Furtado, C., Parente, T.E., 2016. The mitochondrial genome of Hypancistrus zebra (Isbrücker & Nijssen, 1991) (Siluriformes: Loricariidae), an endangered ornamental fish from the Brazilian Amazon. Conserv. Genet. Resour. 0, 1–6. doi:10.1007/s12686-016-0645-5

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., Marshall, D., 2009. Tablet--next generation sequence assembly visualization. Bioinformatics 26, 401–402. doi:10.1093/bioinformatics/btp666

Moreira, D.A., Furtado, C., Parente, T.E., 2015. The use of transcriptomic next-generation sequencing data to assemble mitochondrial genomes of Ancistrus spp. (Loricariidae). Gene 573, 171–175. doi:10.1016/j.gene.2015.08.059

Moreira, D.A., Magalhaes, M.G.P., de Andrade, P.C.C., Furtado, C., Val, A.L., Parente, T.E., 2016. An RNA-based approach to sequence the mitogenome of Hypoptopoma incognitum (Siluriformes: Loricariidae). Mitochondrial DNA. Part A, DNA mapping, Seq. Anal. 27, 3784–3786. doi:10.3109/19401736.2015.1079903

Nakatani, M., Miya, M., Mabuchi, K., Saitoh, K., Nishida, M., 2011. Evolutionary history of Otophysi (Teleostei), a major clade of the modern freshwater fishes: Pangaean origin and Mesozoic radiation. BMC Evol. Biol. 11, 177. doi:10.1186/1471-2148-11-177

Nelson, D.R., Goldstone, J. V, Stegeman, J.J., 2013. The cytochrome P450 genesis locus : the origin and evolution of animal cytochrome P450s The cytochrome

P450 genesis locus : the origin and evolution of animal cytochrome P450s. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 368, 1612. doi:10.1098/rstb.2012.0474

Nico, L.G., Loftus, W.F., Reid, J.P., 2009. Interactions between non-native armored suckermouth catfish (Loricariidae: Pterygoplichthys) and native Florida manatee (Trichechus manatus latirostris) in artesian springs. Aquat. Invasions 4, 511–519. doi:10.3391/ai.2009.4.3.13

Nogueira, L., Rodrigues, A.C.F., Trídico, C.P., Fossa, C.E., De Almeida, E.A., 2011a. Oxidative stress in Nile tilapia (Oreochromis niloticus) and armored catfish (Pterygoplichthys anisitsi) exposed to diesel oil. Environ. Monit. Assess. 180, 243–255. doi:10.1007/s10661-010-1785-9

Nogueira, L., Sanches, A.L.M., da Silva, D.G.H., Ferrizi, V.C., Moreira, A.B., de Almeida, E.A., 2011b. Biochemical biomarkers in Nile tilapia (Oreochromis niloticus) after short-term exposure to diesel oil, pure biodiesel and biodiesel blends. Chemosphere 85, 97–105. doi:10.1016/j.chemosphere.2011.05.037

Parente, T.E.M., De-Oliveira, a. C. a X., Beghini, D.G., Chapeaurouge, D. a., Perales, J., Paumgartten, F.J.R., 2009. Lack of constitutive and inducible ethoxyresorufin-O-deethylase activity in the liver of suckermouth armored catfish (Hypostomus affinis and Hypostomus auroguttatus, Loricariidae). Comp. Biochem. Physiol. - C Toxicol. Pharmacol. 150, 252–260. doi:10.1016/j.cbpc.2009.05.006

Parente, T.E.M., Rebelo, M.F., da-Silva, M.L., Woodin, B.R., Goldstone, J. V., Bisch, P.M., Paumgartten, F.J.R., Stegeman, J.J., 2011. Structural features of cytochrome P450 1A associated with the absence of EROD activity in liver of the loricariid catfish Pterygoplichthys sp. Gene 489, 111–118. doi:10.1016/j.gene.2011.07.023

Parente, T.E.M., Santos, L.M.F., de Oliveira, A.C.A.X., Torres, J.P. de M., Araújo, F.G., Delgado, I.F., Paumgartten, F.J.R., 2015. The concentrations of heavy metals and the incidence of micronucleated erythrocytes and liver EROD activity in two edible fish from the Paraíba do Sul river basin in Brazil. Vigilância Sanitária em Debate 1, 88–92. doi:10.3395/2317-269x.00278

Parente, T.E.M., Urban, P., Pompon, D., Rebelo, M.F., 2014. Altered substrate specificity of the Pterygoplichthys sp. (Loricariidae) CYP1A enzyme. Aquat. Toxicol. 154, 193–199. doi:10.1016/j.aquatox.2014.05.021

Pascussi, J.-M., Gerbal-Chaloin, S., Duret, C., Daujat-Chavanieu, M., Vilarem, M.-J., Maurel, P., 2008. The tangle of nuclear receptors that controls xenobiotic metabolism and transport: crosstalk and consequences. Annu. Rev. Pharmacol. Toxicol. 48, 1–32. doi:10.1146/annurev.pharmtox.47.120505.105349

Rees, D.C., Johnson, E., Lewinson, O., 2009. ABC transporters: the power to change. Nat. Rev. Mol. Cell Biol. 10, 218–227. doi:10.1038/nrm2646

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P., 2011. Integrative genomics viewer. Nat. Biotechnol. 29, 24–26. doi:10.1038/nbt.1754

Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.

Brief. Bioinform. 14, 178–192. doi:10.1093/bib/bbs017

Villalba-Villalba, A.G., Ramírez-Suárez, J.C., Valenzuela-Soto, E.M., Sánchez, G.G., Ruiz, G.C., Pacheco-Aguilar, R., 2013. Trypsin from viscera of vermiculated sailfin catfish, Pterygoplichthys disjunctivus, Weber, 1991: Its purification and characterization. Food Chem. 141, 940–945. doi:10.1016/j.foodchem.2013.03.078

Vishwakarma, S.K., Ameer, S., Paspala, B., Khan, A.A., 2014. Human ATP Binding Cassette (ABC) Transporters : A Phylogenetic Investigation. Int. J. Sci. Res. 3, 564–571.

Xie, W., Evans, R.M., 2001. Orphan Nuclear Receptors: The Exotics of Xenobiotics. J. Biol. Chem. 276, 37739–37742. doi:10.1074/jbc.R100033200

Yang, Y.H., Wang, J.L., Miranda, C.L., Buhler, D.R., 1998. CYP2M1: cloning, sequencing, and expression of a new cytochrome P450 from rainbow trout liver with fatty acid (omega-6)-hydroxylation activity. Arch. Biochem. Biophys. 352, 271–280. doi:10.1006/abbi.1998.0607

Zhenzhen, X., Ling, X., Dengdong, W., Chao, F., Qiongyu, L., Zihao, L., Xiaochun, L., Yong, Z., Shuisheng, L., Haoran, L., 2014. Transcriptome Analysis of the Trachinotus ovatus: Identification of Reproduction, Growth and Immune-Related Genes and Microsatellite Markers. PLoS One 9, e109419. doi:10.1371/journal.pone.0109419

Data Accessibility

Illumina reads are deposited at the NCBI Sequence Read Archive (SRA) under the accessions: SRR3664270 for paired-end reads, and SRR3664326 for single-end reads. Assembled transcriptome is deposited at the NCBI Transcriptome Shotgun Assembly (TSA) under the accession: GETR00000000. Mitochondrial genomes are deposited at NCBI GenBank under the accession: KT239003, KT239004 and KT239005. NCBI BioProject ID: PRJNA324853. NCBI Biosample: SAMN05216828

Author Contributions

T.E.P designed the work, oversaw sample and library preparation, performed data analyses and wrote the manuscript. D.A.M performed data analyses. M.G.P.M and P.C.C.A prepared sample, libraries and helped in data analyses. C.F performed transcriptome sequencing. B.J.H assisted transcriptome assemble and performed Trinotate analysis. J.J.S and M.E.H oversaw study design and data analyses, and wrote the manuscript. All authors reviewed the manuscript.

Table 1: Summary of *Pterygoplichthys anisitsi* liver transcriptome sequencing and annotation.

Total sequencing reads	**179,826,191**
Reads after QC	**177,354,428**
Transcrispts assembled	**66,642**
Transcript length (bp)	
max	**10,849**
min	**201**
average	**865**
median	**456**
n50	**1,571**
Transcripts with blastx hit	
Uniprot - Zebrafish (*Danio rerio*)	**28,190**
Uniprot - Human (*Homo sapiens*)	**24,498**
eggNOG	**12,225**
GO	**24,377**
Sequencing depth (x)	
average	**646**
median	**13**
Transcripts sequenced at depth ≥ (%)	
10x	**54**
100x	**13**
***Danio rerio* coverage ratio (%)**	
average	**110**
median	**88**
Transcripts with coverage ratio ≥ 1	**47**

Table 2: Number of sequenced components in the hepatic transcriptome of *P. anisitsi* with complete coding sequence (CDS), >75% of the CDS, >50% of the CDS and the total number of contigs for each defensome gene family.

	Coverage			
	Full length	**>75% CDS**	**>50% CDS**	**Total Contigs**
AHR & ARNT	1	3	6	10
Aldo Keto Reductase	5	9	9	10
ATP Binding Cassette (ABC)	21	23	30	113
Basic leucine zipper	2	2	3	3
Catalase	1	1	1	1
Cytochrome P450	43	47	63	159
Epoxide hydroxylase	2	2	2	8
Glucuronosyltransferase	6	8	8	22
Glutatione Peroxidase	5	6	8	11
Glutatione-S-transferase	10	12	12	13
n-acetyl-transferases	10	13	13	15
Nuclear receptor	32	42	67	107
Sulfotransferases (SULT)	33	36	60	67
Superoxide desmutase	2	5	5	5
Thioredoxins (TXN)	3	9	11	14
Total	176	218	298	558

Figure 1: Ratio of *Pterygoplichthys anisitsi* transcript length to *Danio rerio* homologs plotted against *P. anisitsi* transcript sequencing depth. Defensome transcripts are highlighted in red. The total length of each the 28,190 *P. anisitsi* transcripts with a BLASTX hit against in *Danio rerio* Uniprot entries was divided by the length of the CDS of its homolog, and plotted against the average sequencing depth (calculated by dividing the sum of the length of all reads aligned to a transcript by its total length).

Figure 2: Mitochondrial genome of *Pterygoplichthys anisitsi*. Green blocks represent ribosomal RNA, red ones protein coding genes, and the ones in blue tRNAs. Black circles indicate the tRNA not sequenced in *P. anisitsi* mtDNA. Black arrows the approximate region of the six gaps, which coincide to areas with no reads in the log-scale graphic of the reads mapped against the mitogenome showed below. Colored bars indicate heteroplasmic sites.

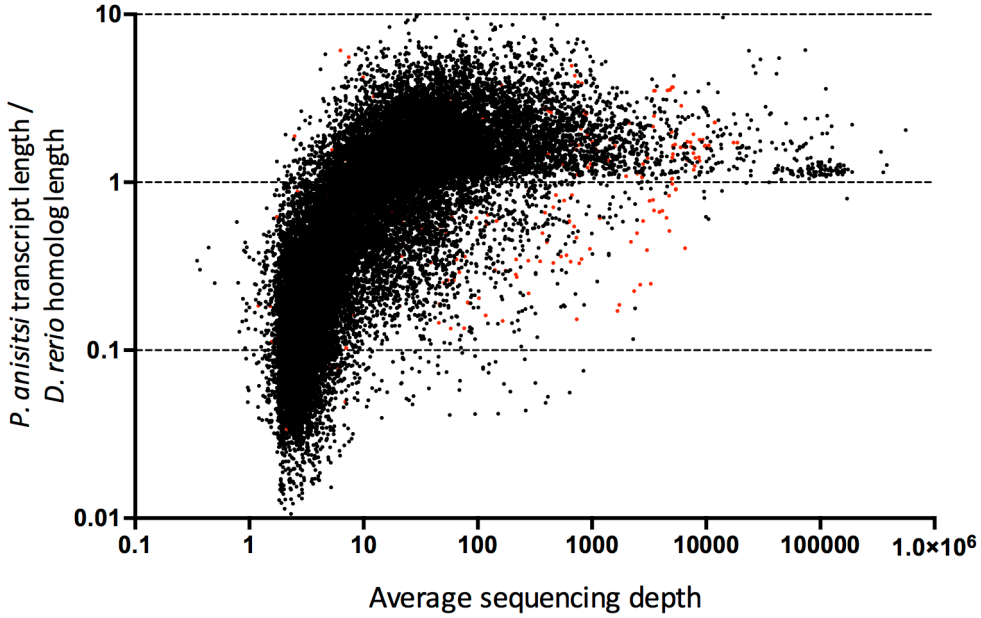
Figure 3: Maximum-likelihood phylogeny of *Pterygoplichthys anisitsi* cytochromes P450 and homologs. The tree is rooted on CYP51. Sequences of *P. anisitsi* are shown in red. Expansions of CYP2Ys and CYP2AAs are highlighted in gray. Bootstrap values are shown on each node (1000 replicates). The translated amino acid sequences were aligned using Muscle and the tree was constructed using RAxML with the LG model for amino acid substitution optimized for invariable sites and across site rate variation. Ps=*Pterygoplichthys anisitsi*; Pte=*Pterygoplichthys* sp.; Anc=*Ancistrus* sp.; Cor=*Corydoras* sp.; Hs=*Homo sapiens*; Dr=*Danio rerio*; Ip=*Ictalurus punctatus*. Gis are shown in Supporting Information Table S5.

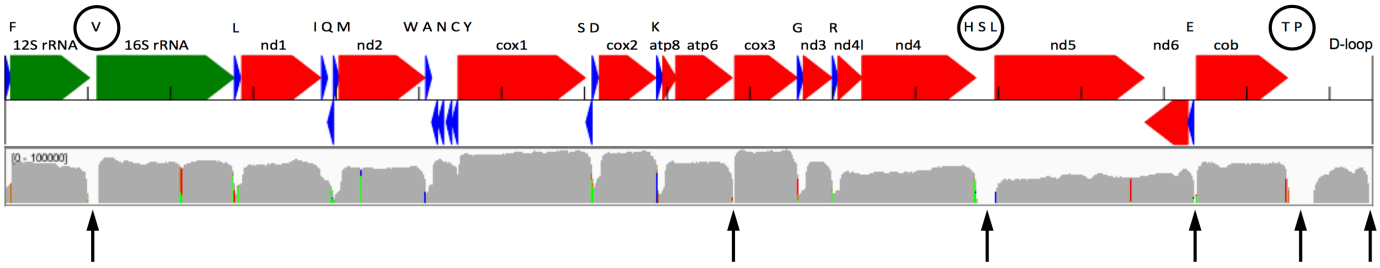
Figure 4: Unrooted maximum-likelihood phylogeny of *Pterygoplichthys anisitsi* sulfotransferases (SULT) and homologs. Sequences of *P. anisitsi* are shown in red. Bootstrap values are shown on each node (1000 replicates). The translated amino acid sequences were aligned using Muscle and the tree was constructed using RAxML with the LG model for amino acid substitution optimized for invariable sites and across site rate variation. Ps=*Pterygoplichthys anisitsi*; Hs=*Homo sapiens*; Dr=*Danio rerio*; Ip=*Ictalurus punctatus*. Gis are shown on Supporting Information Table S5.

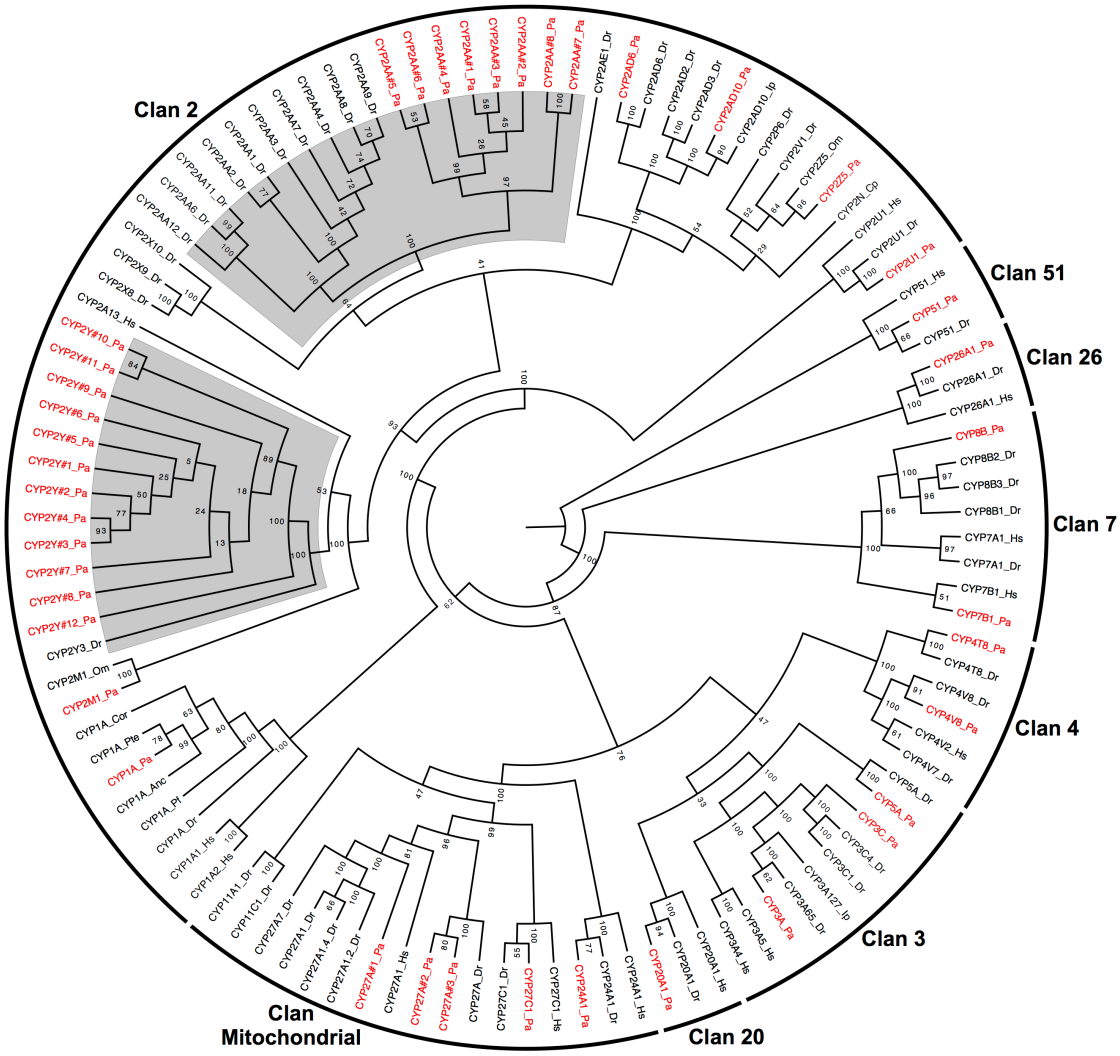
Figure 5: Unrooted maximum-likelihood phylogeny of *Pterygoplichthys anisitsi* nuclear receptors (NR) and homologs. Sequences of *P. anisitsi* are shown in red. Bootstrap values are shown on each node (1000 replicates). The translated amino acid sequences were aligned using Muscle and the tree was constructed using RAxML with the LG model for amino acid substitution optimized for invariable sites and across site rate variation. Ps=*Pterygoplichthys anisitsi*; Hs=*Homo sapiens*; Dr=*Danio rerio*; Ip=*Ictalurus punctatus*. Gis are shown in Supporting Information Table S5.

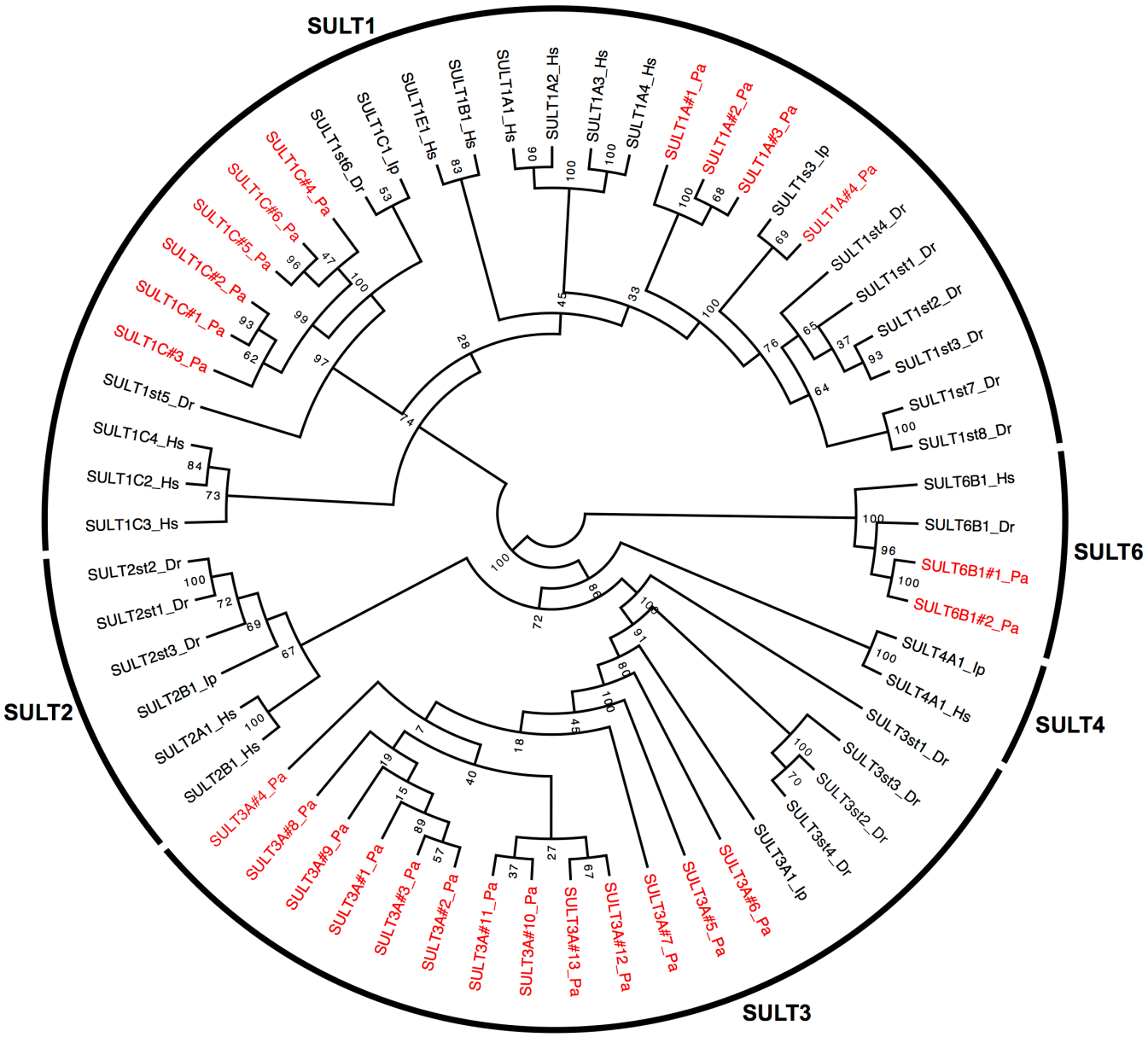
Figure 6: Unrooted maximum-likelihood phylogeny of *Pterygoplichthys anisitsi* ATP Binding Cassete (ABC) transporters and homologs. Sequences of *P. anisitsi* are shown in red. Bootstrap values are shown on each node (1000 replicates). The translated amino acid sequences were aligned using Muscle and the tree was constructed using RAxML with the LG model for amino acid substitution optimized

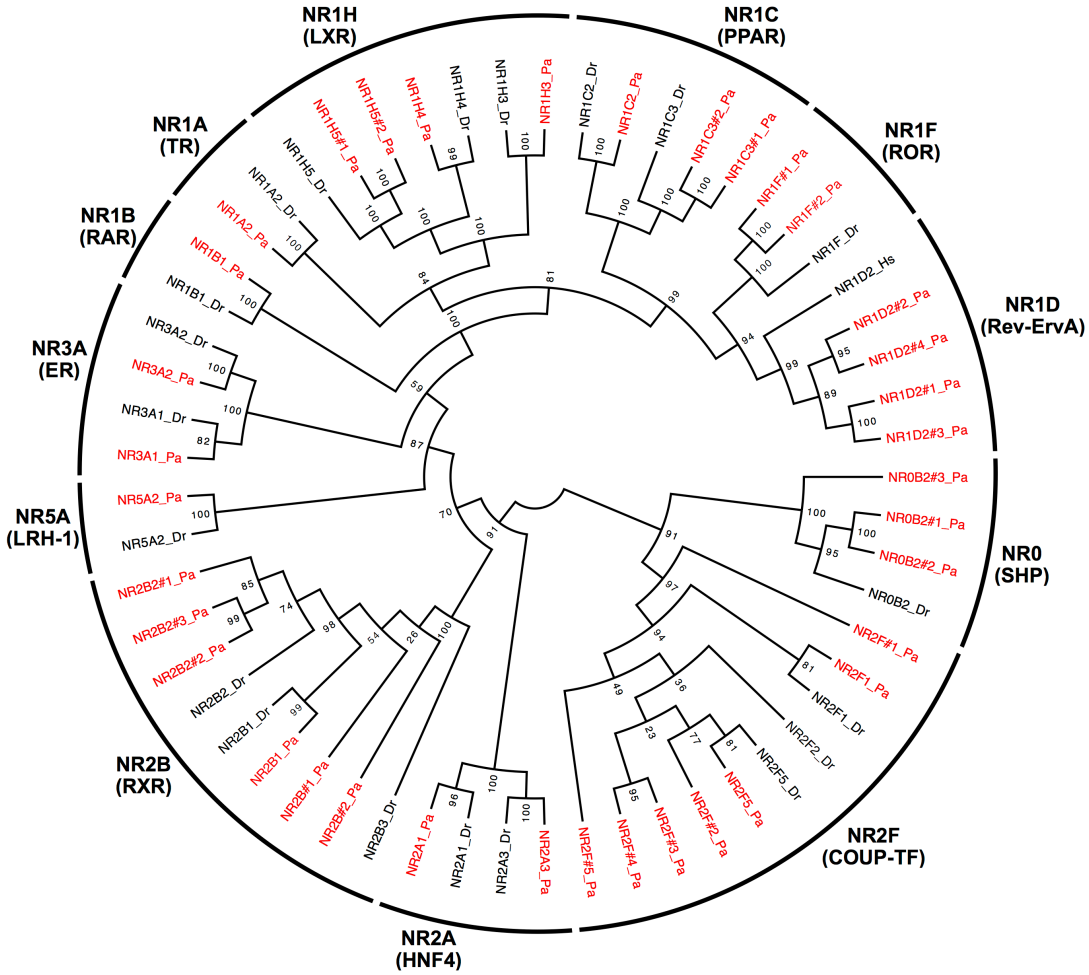
for invariable sites and across site rate variation. GIs are shown on Supporting Information Table S5**.**

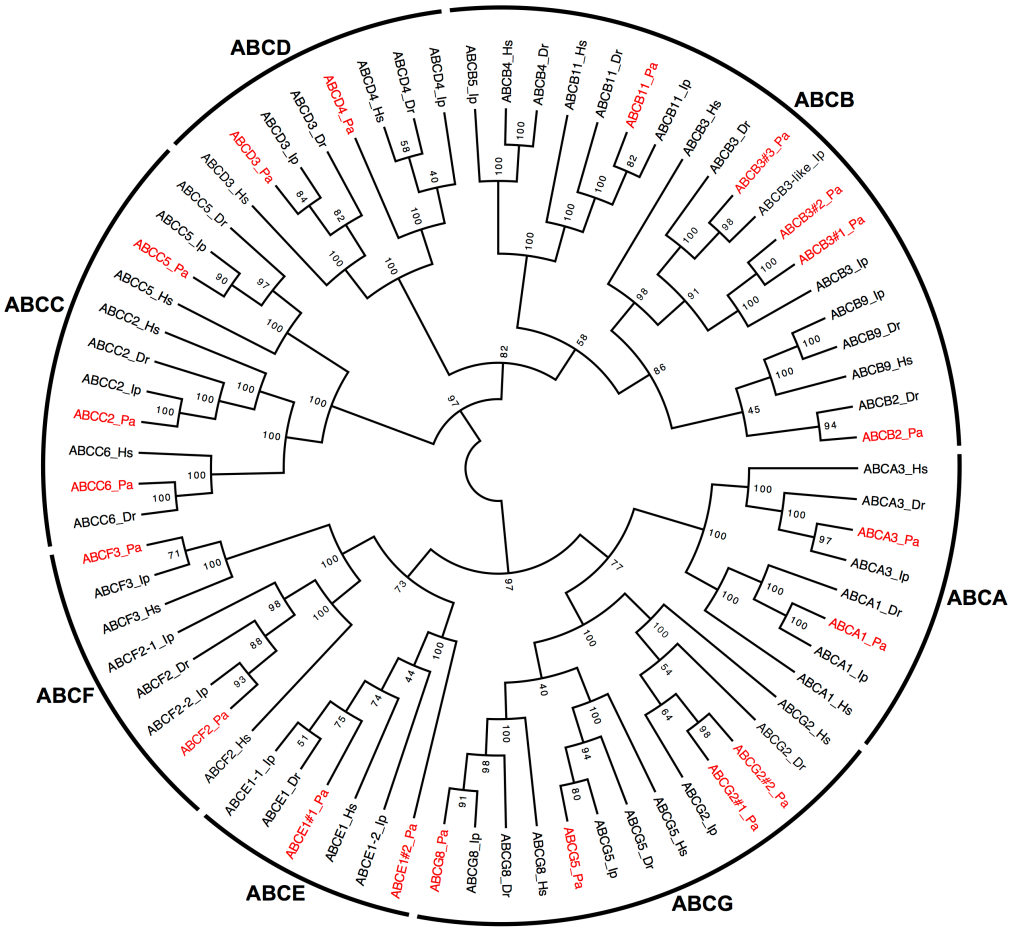

10
1
0.1
0.01
P. anisitsi transcript length /
D. rerio homolog length
0.1
1
10
100
1000
10000
100000
1.0×10^6
Average sequencing depth


F
V
L
I Q M
W A N C Y
S D
K
G
R
H S L
E
T P
12S rRNA
16S rRNA
nd1
nd2
cox1
cox2
atp8
atp6
cox3
nd3
nd4l
nd4
nd5
nd6
cob
D-loop
[0 - 100000]

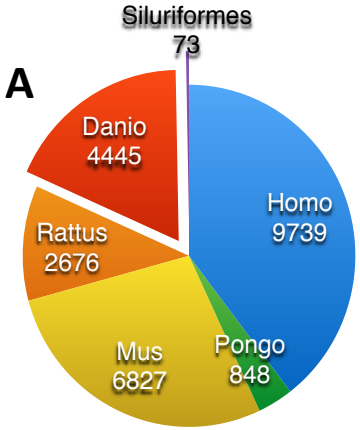

Clan 2
Clan 51
Clan 26
Clan 7
Clan 4
Clan 3
Clan 20
Clan Mitochondrial
CYP2AA#7_Pa
CYP2AA#8_Pa
CYP2AA#2_Pa
CYP2AA#3_Pa
CYP2AA#1_Pa
CYP2AA#4_Pa
CYP2AA#5_Pa
CYP2AA#6_Pa
CYP2AE1_Dr
CYP2AD6_Pa
CYP2AD6_Dr
CYP2AD2_Dr
CYP2AD3_Dr
CYP2AD10_Pa
CYP2AD10_Ip
CYP2P6_Dr
CYP2Y1_Dr
CYP2Z5_Om
CYP2Z5_Pa
CYP2N_Cp
CYP2U1_Hs
CYP2U1_Dr
CYP2U1_Pa
CYP51_Hs
CYP51_Pa
CYP51_Dr
CYP26A1_Pa
CYP26A1_Dr
CYP26A1_Hs
CYP8B_Pa
CYP8B2_Dr
CYP8B3_Dr
CYP8B1_Dr
CYP7A1_Hs
CYP7A1_Dr
CYP7B1_Hs
CYP7B1_Pa
CYP4T8_Pa
CYP4T8_Dr
CYP4V8_Dr
CYP4V8_Pa
CYP4V2_Hs
CYP4V7_Dr
CYP5A_Dr
CYP5A_Pa
CYP3C_Pa
CYP3C4_Dr
CYP3C1_Dr
CYP3A127_Ip
CYP3A65_Dr
CYP3A_Pa
CYP3A5_Hs
CYP3A4_Hs
CYP20A1_Dr
CYP20A1_Pa
CYP24A1_Hs
CYP24A1_Dr
CYP24A1_Pa
CYP27C1_Hs
CYP27C1_Pa
CYP27C1_Dr
CYP27A_Dr
CYP27A#3_Pa
CYP27A#2_Pa
CYP27A#1_Pa
CYP27A1_Hs
CYP27A1-2_Dr
CYP27A1-4_Dr
CYP27A1_Dr
CYP27A7_Dr
CYP11C1_Dr
CYP11A1_Dr
CYP1A2_Hs
CYP1A1_Hs
CYP1A_Dr
CYP1A_Pt
CYP1A_Anc
CYP1A_Pa
CYP1A_Pte
CYP1A_Cor
CYP2M1_Pa
CYP2M1_Om
CYP2Y3_Dr
CYP2Y#12_Pa
CYP2Y#8_Pa
CYP2Y#7_Pa
CYP2Y#3_Pa
CYP2Y#4_Pa
CYP2Y#2_Pa
CYP2Y#1_Pa
CYP2Y#5_Pa
CYP2Y#6_Pa
CYP2Y#9_Pa
CYP2Y#11_Pa
CYP2Y#10_Pa
CYP2A13_Hs
CYP2X8_Dr
CYP2X9_Dr
CYP2X10_Dr
CYP2AA12_Dr
CYP2AA6_Dr
CYP2AA11_Dr
CYP2AA2_Dr
CYP2AA1_Dr
CYP2AA9_Dr
CYP2AA7_Dr
CYP2AA4_Dr
CYP2AA8_Dr
CYP2AA9_Dr

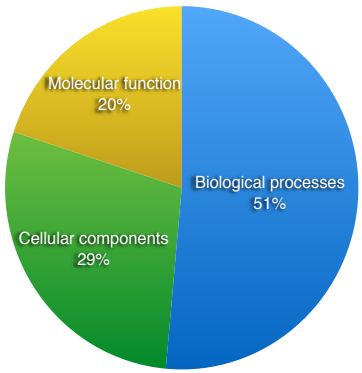

SULT1
SULT2
SULT3
SULT4
SULT6
SULT1A1_Hs
SULT1A2_Hs
SULT1A3_Hs
SULT1A4_Hs
SULT1A#1_Pa
SULT1A#2_Pa
SULT1A#3_Pa
SULT1A#4_Pa
SULT1B1_Hs
SULT1E1_Hs
SULT1C1_Ip
SULT1st6_Dr
SULT1C#4_Pa
SULT1C#6_Pa
SULT1C#5_Pa
SULT1C#2_Pa
SULT1C#1_Pa
SULT1C#3_Pa
SULT1st5_Dr
SULT1s3_Ip
SULT1st4_Dr
SULT1st1_Dr
SULT1st2_Dr
SULT1st3_Dr
SULT1st7_Dr
SULT1st8_Dr
SULT1C4_Hs
SULT1C2_Hs
SULT1C3_Hs
SULT6B1_Hs
SULT6B1_Dr
SULT6B1#1_Pa
SULT6B1#2_Pa
SULT2st2_Dr
SULT2st1_Dr
SULT2st3_Dr
SULT2B1_Ip
SULT2A1_Hs
SULT2B1_Hs
SULT4A1_Ip
SULT4A1_Hs
SULT3st1_Dr
SULT3st3_Dr
SULT3st2_Dr
SULT3st4_Dr
SULT3A1_Ip
SULT3A#6_Pa
SULT3A#5_Pa
SULT3A#7_Pa
SULT3A#12_Pa
SULT3A#13_Pa
SULT3A#10_Pa
SULT3A#11_Pa
SULT3A#2_Pa
SULT3A#3_Pa
SULT3A#1_Pa
SULT3A#9_Pa
SULT3A#8_Pa
SULT3A#4_Pa

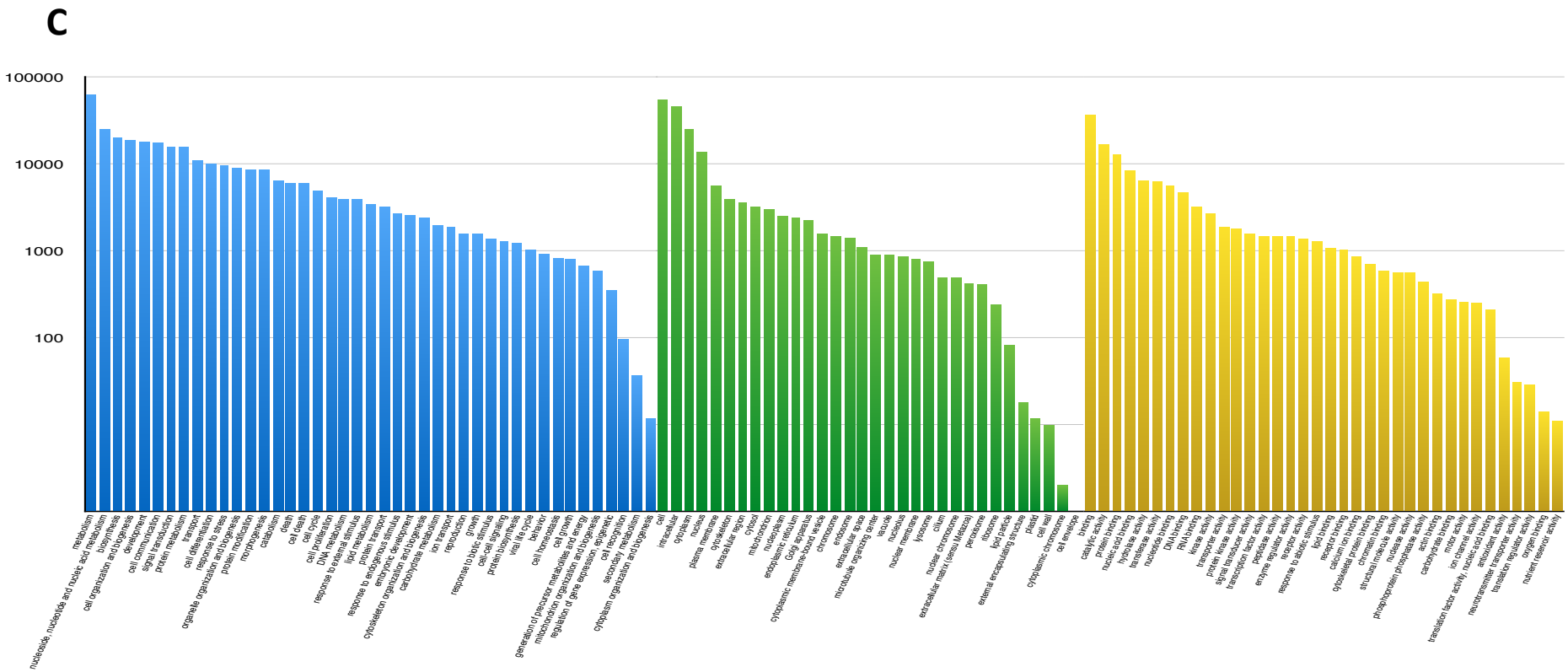

NR1H
(LXR)
NR1C
(PPAR)
NR1A
(TR)
NR1F
(ROR)
NR1B
(RAR)
NR1D
(Rev-ErvA)
NR3A
(ER)
NR5A
(LRH-1)
NR0
(SHP)
NR2B
(RXR)
NR2F
(COUP-TF)
NR2A
(HNF4)
NR1H3_Pa
NR1H3_Dr
NR1H4_Pa
NR1H4_Dr
NR1H5#1_Pa
NR1H5#2_Pa
NR1H5_Dr
NR1C2_Pa
NR1C2_Dr
NR1C3_Dr
NR1C3#2_Pa
NR1C3#1_Pa
NR1F#1_Pa
NR1F#2_Pa
NR1F_Dr
NR1D2_Hs
NR1D2#2_Pa
NR1D2#4_Pa
NR1D2#1_Pa
NR1D2#3_Pa
NR1A2_Dr
NR1A2_Pa
NR1B1_Pa
NR1B1_Dr
NR3A2_Dr
NR3A2_Pa
NR3A1_Dr
NR3A1_Pa
NR5A2_Pa
NR5A2_Dr
NR0B2#3_Pa
NR0B2#1_Pa
NR0B2#2_Pa
NR0B2_Dr
NR2F#1_Pa
NR2F1_Pa
NR2F1_Dr
NR2F2_Dr
NR2F5_Dr
NR2F5_Pa
NR2F#2_Pa
NR2F#3_Pa
NR2F#4_Pa
NR2F#5_Pa
NR2A3_Pa
NR2A3_Dr
NR2A1_Dr
NR2A1_Pa
NR2B3_Dr
NR2B#2_Pa
NR2B#1_Pa
NR2B1_Pa
NR2B1_Dr
NR2B2_Dr
NR2B2#2_Pa
NR2B2#3_Pa
NR2B2#1_Pa


ABCD
ABCB
ABCC
ABCA
ABCF
ABCE
ABCG
ABCD4_Pa
ABCD4_Hs
ABCD4_Dr
ABCD4_Ip
ABCD3_Ip
ABCD3_Dr
ABCD3_Pa
ABCD3_Hs
ABCB4_Hs
ABCB4_Dr
ABCB11_Hs
ABCB11_Dr
ABCB11_Pa
ABCB11_Ip
ABCB3_Hs
ABCB3_Dr
ABCB3#3_Pa
ABCB3-like_Ip
ABCB3#2_Pa
ABCB3#1_Pa
ABCB3_Ip
ABCB9_Ip
ABCB9_Dr
ABCB9_Hs
ABCB2_Dr
ABCB2_Pa
ABCC5_Dr
ABCC5_Ip
ABCC5_Pa
ABCC5_Hs
ABCC2_Hs
ABCC2_Dr
ABCC2_Ip
ABCC2_Pa
ABCC6_Hs
ABCC6_Pa
ABCC6_Dr
ABCA3_Hs
ABCA3_Dr
ABCA3_Pa
ABCA3_Ip
ABCA1_Dr
ABCA1_Pa
ABCA1_Ip
ABCA1_Hs
ABCG2_Hs
ABCG2_Dr
ABCG2#2_Pa
ABCG2#1_Pa
ABCG2_Ip
ABCG5_Hs
ABCG5_Dr
ABCG5_Ip
ABCG5_Pa
ABCG8_Hs
ABCG8_Dr
ABCG8_Ip
ABCG8_Pa
ABCE1#2_Pa
ABCE1-2_Ip
ABCE1_Hs
ABCE1#1_Pa
ABCE1_Dr
ABCE1-1_Ip
ABCF2_Hs
ABCF2-2_Ip
ABCF2_Pa
ABCF2_Dr
ABCF2-1_Ip
ABCF3_Hs
ABCF3_Ip
ABCF3_Pa

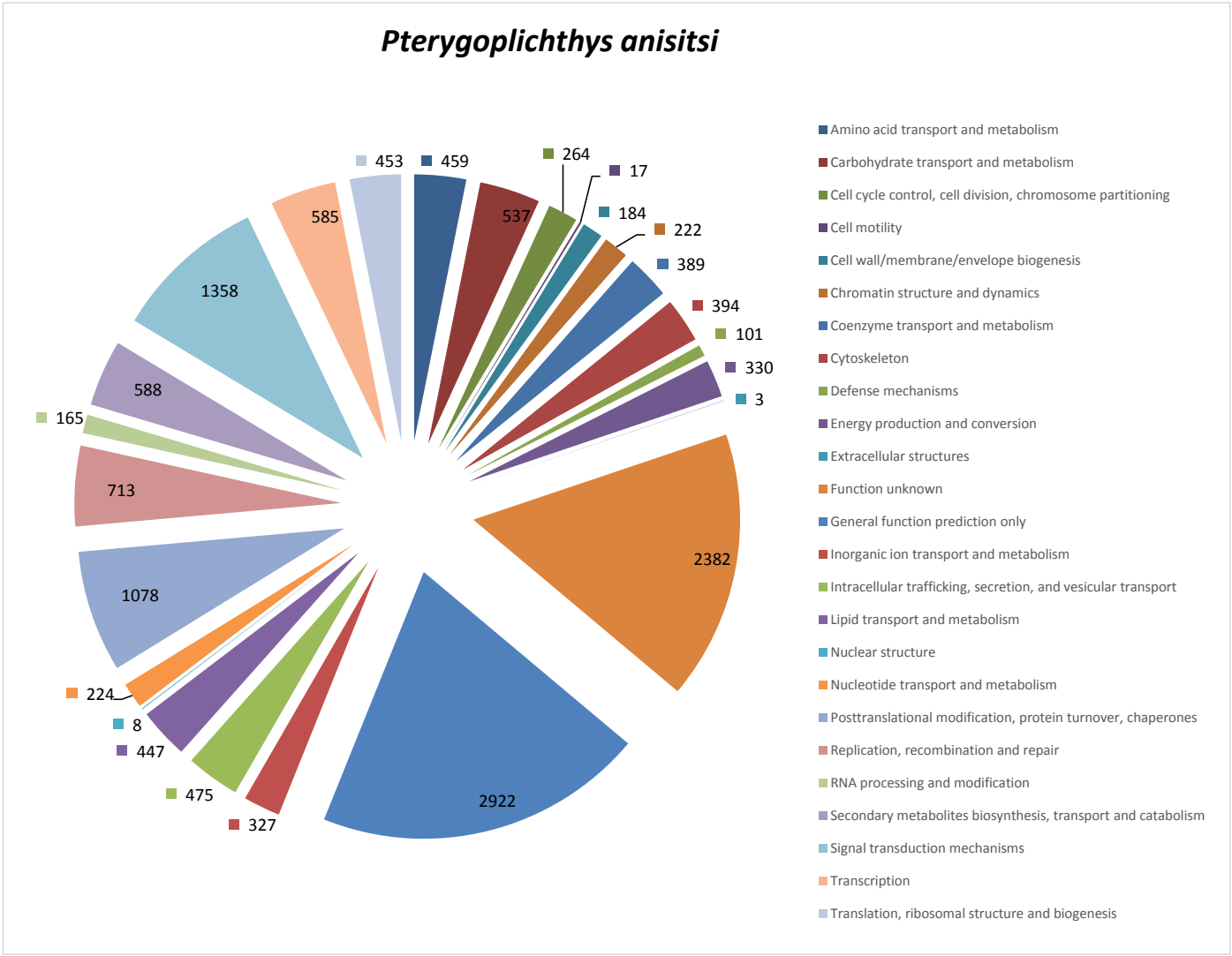
Supplementary Fig. S1 online: Annotation of the liver transcriptome from *P. anisitsi*. A – Species distribution of top blastx top hit. B - Functional classification according to Gene Ontology (GO, first level) terms. C - distribution of GOSlim2 terms over the assembled transcripts.







Supplementary Fig. S1 online: cont.



Supplemental material 3: Nucleotide pairs and gaps between the mitochondrial genome of *Pterygoplichthys anisitsi and P. disjunctivus.*

		P. disjunctivus			
		A	**C**	**G**	**T**
P. anisitsi	**A**	5229	0	4	0
	C	0	4449	0	1
	G	1	0	2412	0
	T	1	0	0	4423
	Gaps	1023-1119; 8789-8799; 11734-11905; 14357-14381; 15498-15788; 16481-16512			

Supplementary Table S3: Retrieved defensome genes in the transcriptome, as well as in the used databases.

Protein Family		Number of hits in *Pterigoplychthys anisitisi* transcriptome against		Number of entries in UniProt database for	
	Term used in the search	Zebrafish DB	Human DB	Zebrafish	Human
subfamily ABCA	abca (ALL)	20	20	37	114
subfamily ABCB	abcb (ALL)	22	17	26	120
subfamily ABCC	abcc (ALL)	14	25	30	75
subfamily ABCD	abcd (ALL)	8	8	9	31
subfamily ABCDE	abce (ALL)	3	3	4	14
subfamily ABCDF	abcf (ALL)	4	4	7	25
subfamily ABCG	abcg (ALL)	8	6	23	40
cytochrome P450	cytochrome p450	26	86	73	315
glutathione-S-transferases	glutathione s-transferase	2	9	16	98
sulfotransferases	sulfotransferase	93	102	109	182
UDP-glucuronosyl transferases	udp*glucuronosyltransferase (*com espaço e com hífen)	30	9	73	122
N-acetyl transferases	n-acetyltransferase	3	12	14	108
aldo-keto reductases	aldo-keto reductase	2	3	2	29
epoxide hydrolases	epoxide hydrolase & EPHX	2	2	1	10
NAD(P)H-quinone oxidoreductases	nad(p)h quinone oxidoreductase	0	0	0	2
superoxide dismutases	superoxide dismutase ; NOT copper chaperone	3	3	9	34
catalases	catalase	1	1	5	5
glutathione peroxidase	glutathione peroxidase	11	11	19	46
thioredoxins	thioredoxin	9	28	16	130
aryl hydrocarbon receptor	aryl hydrocarbon receptor	4	6	15	82
aryl hydrocarbon receptor nuclear translocator	aryl hydrocarbon receptor nuclear translocator	2	4	6	37
nuclear receptor	nuclear receptor	18	68	64	305
pregnane-X-receptor	pregnane x receptor	1	0	3	1
constitutive androstane receptor	constitutive androstane receptor	0	0	0	9
peroxisome proliferator-activated receptors	peroxisome proliferator*activated receptor (espaço e hífen)	6	7	8	32
liver-X-receptor	liver x receptor	0	0	0	4
farnesoid-X- receptor	farnesoid x receptor	0	3	0	1
erythroid-derived 2	erythroid-derived 2 ; erythroid 2-related	3	3	4	7
basic-leucine zipper	basic leucine zipper	3	3	12	51
TOTAL		298	443	585	2029

Supplementary Table S4: Single CDSs with different UTR regions.

	Number of		
	CDS	Sequences	Remarks
CYP2AA	8	8	
CYP2Y	12	13	CDS#1 has two seqs
CYP5A	1	2	
CYP27A	3	7	CDS#2 has four seqs
NR1C2	1	2	
NR1F	2	4	each CDS has two seqs
NR1H3	1	2	
NR2B2	3	6	each CDS has two seqs
NR2F	2	3	CDS#1 has two seqs
NR3A2	1	2	
ABCB2	1	2	
ABCD3	1	3	
SULT1A	4	13	CDS#3 has two seqs; CDS#2 has nine seqs
SULT1C	6	7	CDS#1 has two seqs
SULT3A	13	14	CDS#2 has two seqs

Supplementary Fig. S6: Deletion and insertion in the NR2B from *Pterygoplichthys anisitsi*