MINISTÉRIO DA SAÚDE
FUNDAÇÃO OSWALDO CRUZ
INSTITUTO OSWALDO CRUZ

Doutorado em Biologia Computacional e Sistemas

O QUE DADOS TRANSCRIPTÔMICOS REVELAM SOBRE A BIODIVERSIDADE E EVOLUÇÃO DE LORICARIOIDEI (SILURIFORMES)

DANIEL ANDRADE MOREIRA

Rio de Janeiro
Abril de 2018

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

Daniel Andrade Moreira

O que Dados Transcriptômicos Revelam sobre a Biodiversidade e Evolução de Loricarioidei (Siluriformes)

Tese apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Doutor em Ciências

Orientadores: Dr. Thiago E. Parente
Dra. Renata Schama

RIO DE JANEIRO
Abril de 2018

Moreira, Daniel Andrade .

O que dados transcriptômicos revelam sobre a biodiversidade e evolução de Loricarioidei (Siluriformes) / Daniel Andrade Moreira. - Rio de janeiro, 2018.
xiii, 110 f.; il.

Tese (Doutorado) – Instituto Oswaldo Cruz, Pós-Graduação em Biologia Computacional e Sistemas, 2018.

Orientador: Thiago Estevam Parente.
Co-orientador: Renata Schama.

Bibliografia: Inclui Bibliografias.

1. Sequenciamento de nucleotídeos de alto rendimento. 2. RNA-Seq. 3. Região Neotropical. 4. Loricariidae. 5. Genoma mitocondrial. I. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da Biblioteca de Manguinhos/ICICT com os dados fornecidos pelo(a) autor(a).

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

AUTOR: DANIEL ANDRADE MOREIRA

O QUE DADOS TRANSCRIPTÔMICOS REVELAM SOBRE A BIODIVERSIDADE E EVOLUÇÃO DE LORICARIOIDEI (SILURIFORMES)

ORIENTADORES: Dr. Thiago E. Parente
Dra. Renata Schama

Aprovado em: 26/04/2018

EXAMINADORES:

Dra. Ana Carolina Paulo Vicente - Presidente (IOC/FIOCRUZ)
Dr. Francisco Pereira Lobo (UFMG)
Dr. André Elias Rodrigues Soares (LNCC)
Dr. Gonzalo Bello Bentancor (IOC/FIOCRUZ)
Dr. Francisco José Roma Paumgartten (ENSP/FIOCRUZ)

Rio de Janeiro, 26 de Abril de 2018

Dedico este trabalho
à minha mulher Tatiana Cabrini
por todo o amor e companheirismo
durante essa caminhada.

AGRADECIMENTOS

Aos meus orientadores, Renata Schama e Thiago Parente, por todas as conversas, ensinamentos e pelas contribuições para minha formação acadêmica durante esses 2 anos e 8 meses de doutorado. Em especial para o Thiago, que sempre foi mais que um orientador, foi um amigo que me trouxe de volta para a vida acadêmica, confiou na minha capacidade e sempre me estimulou a dar o meu melhor.

Às “meninas do lab”, Maithê e Paula, pelas conversas científicas e não científicas sempre de maneira descontraída tornando nossa convivência muito agradável.

À todas as pessoas do Laboratório de Toxicologia Ambiental, em especial ao Dr. Francisco Paumgartten por nos acolher em seu laboratório.

Aos membros da banca por aceitarem o convite, é um privilégio ter seus comentários como forma de melhorar meu trabalho e minha formação.

Aos professores do Programa de Pós-Graduação de Biologia Computacional e Sistemas, com os quais aprendi muito durante as diversas disciplinas que cursei. Agradecimento especial à secretária Rose Pani por sempre estar disposta a ajudar.

Aos meus amigos por toda a torcida e apoio, apesar da distância nos impedir de nos encontrarmos com a frequência que gostaríamos, a amizade se fortalece a cada ano que passa.

Aos meus pais, Aldemir e Ely, e minhas irmãs, Myrli e Mayle, por todo o amor e por sempre apoiarem e incentivarem minhas decisões, devo a eles todos os valores e princípios que me formaram. Também agradeço aos meus cunhados pelo carinho e votos de sucesso.

À minha nova família carioca por me acolher com tanto amor, em especial à minha sogra Nailê que me recebeu como um filho.

À minha amada mulher, Tati, por todo o companheirismo, amor, paciência e incentivo dedicados. Sua alegria de viver é contagiante e viver ao seu lado só me engrandece.

Esse trabalho foi finando pela USAID e apoiado com bolsa da CAPES.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

O QUE DADOS TRANSCRIPTÔMICOS REVELAM SOBRE A BIODIVERSIDADE E EVOLUÇÃO DE LORICARIOIDEI (SILURIFORMES)

RESUMO

TESE DE DOUTORADO EM BIOLOGIA COMPUTACIONAL E SISTEMAS

Daniel Andrade Moreira

A biodiversidade é uma entidade multidimensional, que se refere a diferentes elementos e níveis de variabilidade da vida na Terra. As tecnologias de sequenciamento de ácidos nucleicos de alto desempenho, aliadas à biologia computacional, têm permitido uma caracterização mais ampla e profunda da biodiversidade e de suas complexas interações entre a integridade dos ecossistemas, o bem-estar humano e a saúde dos outros seres vivos. De toda a diversidade de vertebrados, aproximadamente 50% das espécies são peixes e destas, 50% são espécies de água doce. A região Neotropical possui a maior riqueza de espécies de peixes do mundo e alto grau de endemismo. Entre as famílias endêmicas, destaca-se a família Loricariidae por ser a quinta família mais diversa entre os vertebrados, com mais de 900 espécies válidas. Os métodos genômicos permitiram que várias espécies de peixes saíssem da ignorância genômica, transformando-as em um novo exército de espécies modelo possibilitando uma análise abrangente das bases genéticas da evolução. Essa tese tem como propósito maior usar a biologia computacional para a integração da ciência por descoberta e da ciência orientada por hipóteses na investigação da biodiversidade da subordem Loricarioidei (Siluriformes), com ênfase na família Loricariidae. Através do uso do sequenciamento de alto desempenho, geramos e analisamos 40 transcriptomas de 34 espécies, incluindo 31 espécies de loricarídeos. A mineração desses transcriptomas possibilitou novos usos para as sequências provenientes de RNA-Seq, como a montagem de genomas mitocondriais, descrita e discutida no Capítulo Um desta tese. Tal abordagem possibilitou ainda a medida dos níveis de expressão de transcritos mitocondriais, o padrão de pontuação da edição pós-transcricional e a detecção de heteroplasmias. Essa metodologia permitiu montar os genomas mitocondriais descritos nos Capítulos Dois e Três. Aliar essa metodologia a uma perspectiva evolutiva possibilitou testar hipóteses filogenéticas e investigar a evolução estrutural desses genomas, como descrito no Capítulo Quatro. No Capítulo Cinco, ampliamos o objeto de pesquisa dos transcritos mitocondriais para todo o transcriptoma da espécie *Pterygoplichthys anisitsi*, onde foi encontrada uma grande diversidade de transcritos que codificam enzimas envolvidas na desintoxicação de xenobióticos, o que pode contribuir para a resistência desta espécie a xenobióticos orgânicos. Esta tese é o primeiro trabalho a fazer uso do sequenciamento de ácidos nucléicos de alto desempenho para o estudo de espécies de Loricarioidei, promovendo a ampliação do conhecimento sobre a diversidade genética, taxonômica, filogenética, estrutural, funcional e fenotípica da fauna de peixes neotropicais.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

WHAT TRANSCRIPTOMIC DATA REVEAL ABOUT THE BIODIVERSITY AND EVOLUTION OF LORICARIOIDEI (SILURIFORMES)

ABSTRACT

PHD THESIS IN COMPUTATIONAL AND SYSTEMS BIOLOGY

Daniel Andrade Moreira

Biodiversity is a multidimensional entity, which refers to different elements and levels of variability of life on Earth. Nucleic acids high-throughput sequencing technologies, coupled with computational biology, have enabled a broader and deeper characterization of biodiversity and its complex interactions among ecosystem integrity, human well-being and the health of other living beings. Of all the vertebrate diversity, approximately 50% of the species are fish, 50% of which are freshwater species. The Neotropical region has the greatest richness of fish species in the world and a high degree of endemism. Among the endemic families, the Loricariidae family stands out as the fifth most diverse family among vertebrates, with more than 900 valid species. Genomic methods allowed several fish species to emerge from genomic ignorance, turning them into a new army of species models making possible a comprehensive analysis of the genetic basis of evolution. This thesis has as its main purpose to use computational biology for the integration of discovery science and hypothesis guided science in the investigation of the biodiversity of the suborder Loricarioidei (Siluriformes), with emphasis on the Loricariidae family. Using high-throughput sequencing, we generated and analyzed 40 transcriptomes of 34 species, including 31 loricariids species. The mining of these transcriptomes made possible new uses for RNA-Seq sequences, such as the assembly of mitochondrial genomes, described and discussed in Chapter One of this thesis. This approach also enabled the measurement of mitochondrial transcripts expression levels, the punctuation pattern of post-transcriptional editing and detection of heteroplasmies. The developed methodology allowed assembling the mitochondrial genomes described in Chapters Two and Three. Combining this methodology to an evolutionary perspective made it possible to test phylogenetic hypotheses and to investigate the structural evolution of these genomes, as described in Chapter Four. In Chapter Five, we extended the research object of mitochondrial transcripts to the whole transcriptome of the species *Pterygoplichthys anisitsi*, where a great diversity of transcripts was found that encode enzymes involved in the detoxification of xenobiotics, which may contribute to the resistance of this species to organic xenobiotics. This thesis is the first work to make use of high-throughput nucleic acid sequencing for the study of Loricarioidei species, increasing the knowledge about the genetic, taxonomic, phylogenetic, structural, functional and phenotypic diversity of Neotropical fish fauna.

ÍNDICE

ÍNDICE DE FIGURAS

INTRODUÇÃO

CAPÍTULO 3.1

CAPÍTULO 3.2

CAPÍTULO 3.3

CAPÍTULO 3.4

CAPÍTULO 3.5

LISTA DE TABELAS

CAPÍTULO 3.1

CAPÍTULO 3.3

CAPÍTULO 3.4

CAPÍTULO 3.5

1 INTRODUÇÃO

Os impactos das atividades humanas sobre o ambiente são responsáveis por mover a Terra para uma nova época geológica, o Antropoceno (1). Essa nova época é caracterizada pela mudança nas relações entre homem e meio ambiente, onde o primeiro já transgrediu muitas fronteiras de sustentabilidade de uso do segundo. Os exemplos são muitos, mas entre os de maior impacto, pode-se citar a alteração da composição química da atmosfera resultando em mudanças climáticas e a rápida diminuição no número de espécies devido à sobre-exploração, destruição de habitats e impactos decorrentes da introdução de espécies invasoras (2,3). Em conjunto, esses impactos nos conduzem ao que já é considerado o sexto evento de extinção em massa (4,5).

Embora a extinção seja um grande impacto em nosso planeta e um poderoso motivador da conservação, a defaunação é muito mais do que a perda de espécies. O termo defaunação é usado para caracterizar tanto a perda de espécies e populações animais, como um declínio na abundância de determinada espécie (6). Nesse contexto de defaunação, a biodiversidade vai muito além da riqueza de espécies, devendo ser vista como uma entidade multidimensional, que se refere aos diferentes elementos e níveis de variabilidade da vida na Terra, seja taxonômico, genético, funcional, filogenético, trófico, espacial, temporal, comportamental e tantas outras dimensões da diversidade da vida em um ecossistema (7,8). A redução da biodiversidade, além de promover mudanças na composição de espécies, também altera funções e serviços ecossistêmicos, como qualidade da água, ciclagem de nutrientes e controle de parasitos, vetores e doenças (9), afetando, dessa forma, o bem-estar humano (2,10).

Essa compreensão de que o bem-estar humano está intrinsicamente associado à biodiversidade e ao meio ambiente contribuiu para a ampliação do próprio conceito de saúde. A abordagem “One Health” estabelece que a saúde dos seres humanos, dos outros seres vivos e do meio ambiente estão conectadas e cada uma delas precisa de igual atenção para assegurar a saúde de todos (11). A exemplo dessa conexão, observa-se que doenças podem ser a causa ou a consequência da perda da biodiversidade (11). Nessa abordagem, existe uma maior demanda pelo conhecimento da saúde dos seres vivos e ambiental, para melhor compreensão da dinâmica de doenças. A malária, as febres hemorrágicas (p.e. dengue, zika, chikungunya e amarela) e a raiva são casos bem conhecidos com fortes ligações à saúde animal e aos fatores ambientais (11).

A redução da taxa de perda de biodiversidade e a promoção do uso sustentável dos ecossistemas são objetivos internacionais desde a ECO-92. Atualmente, esses objetivos integram as metas da Agenda 2030 da Organização das Nações Unidas (ONU), estabelecidas após a incapacidade de atingir as metas para 2010 da Convenção sobre Diversidade Biológica (CBD) (12) e frente ao iminente fracasso para cumprir as Metas de Aichi para 2020 (13). No entanto, novas metas não serão eficazes se não houver melhorias nos sistemas de monitoramento da biodiversidade em todo o mundo, padronizando o uso de indicadores, seu compartilhamento e elevando o nível de compreensão da biodiversidade para uma escala multidimensional (2,14,15). Para a compreensão dos mecanismos que unem o bem-estar humano, a saúde dos seres vivos e a integridade dos ecossistemas, é preciso agregar a maior quantidade de informação de vários níveis e dimensões da biodiversidade. Entretanto, apesar do reconhecimento de que a biodiversidade é multidimensional, os trabalhos na área têm sido predominantemente unidimensionais em sua abordagem, sendo a diversidade taxonômica a dimensão dominante sob investigação (7).

É evidente que uma pesquisa multidimensional da biodiversidade é mais difícil e desafiadora. A busca pela maior densidade de informação envolve a definição de um conjunto de métricas que melhor modelem a biodiversidade (15,16). Para tanto, é necessário o desenvolvimento de novas abordagens metodológicas. Codificados no DNA, diversas dimensões da biodiversidade podem ser reveladas, não somente variações genéticas, mas também inferências taxonômica/filogenéticas e funcionais. Aumentar a capacidade de acessar e interpretar essas informações é o primeiro desafio de novas abordagens em potencial. As tecnologias de sequenciamento de nova geração (sigla do inglês, "NGS"), ou sequenciamento de alto desempenho, aliadas à biologia computacional, evoluíram rapidamente na última década e têm permitido uma caracterização, tanto mais ampla quanto mais profunda, de muitas dimensões da biodiversidade, como variações genéticas adaptativas, identificação taxonômica, relações filogenéticas, inferências demográficas e em diferentes sistemas, como células, tecidos, indivíduos, populações, comunidades e ecossistemas (17).

À medida que essas recentes tecnologias de sequenciamento evoluem, um número crescente de métodos de preparação de amostras, novos equipamentos e, principalmente, ferramentas de análise de dados geram uma imensa diversidade de aplicações científicas. A empresa Illumina, atualmente, oferece o maior rendimento por rodada de sequenciamento e o menor custo por base (18). O seu novo sistema "HiSeq X Ten" tem a capacidade surpreendente de gerar até 1,8Tb de dados brutos por corrida[1]. Essa quantidade de dados é maior do que aquela armazenada até março de 2018 nos bancos de dados do GenBank[2] (desconsiderando os bancos de dados de NGS). Com o desenvolvimento desse sistema, a Illumina afirma ter rompido a barreira de sequenciar um genoma humano pelo custo de US$1.000.

[1] Informações disponíveis em: http://www.illumina.com

[2] Informações disponíveis em: https://www.ncbi.nlm.nih.gov/genbank/statistics/

Essa redução de custo têm permitido que as tecnologias de NGS sejam utilizadas em estudos de genômica em qualquer sistema biológico, sem a necessidade de um organismo modelo estreitamente relacionado ao objeto de estudo e em níveis nunca antes possíveis (19,20).

A geração de informação genética em larga escala está pavimentando o caminho para a expansão da biologia de sistemas, ampliando seu volume de estudo de alguns genes e proteínas para o estudo de centenas de milhares de genes e possibilitando, dessa forma, uma visão mais ampla e capaz de compreender a estrutura e a dinâmica de um sistema biológico como um todo, em suas várias dimensões. Contudo, mesmo o todo sendo mais do que a soma de suas partes, como afirmou Aristóteles em sua obra Metafísica, para todo trabalho de biologia de sistemas são necessários dados sobre três aspectos do sistema: a lista de suas partes, a conectividade entre as partes e o contexto, espacial e temporal, dessas conexões (21,22).

Os objetivos da biologia de sistemas são reunir de forma abrangente as informações de cada um dos distintos níveis dos sistemas biológicos individuais e integrar esses dados para gerar modelos matemáticos preditivos do sistema (23). Entretanto, a descoberta de componentes novos e essenciais para processos biológicos e seu funcionamento dificilmente resultarão de abordagens meramente intuitivas, como abordagens “bottom-up”. Devido à complexidade intrínseca dos sistemas biológicos, uma combinação de abordagens experimentais e, principalmente, novas abordagens computacionais serão necessárias para a interpretação e integração desses dados.

Segundo Kitano (2002b), a biologia computacional tem dois ramos distintos: a descoberta de conhecimento ou mineração de dados, que extrai os padrões ocultos de enormes quantidades de dados experimentais, formando hipóteses como resultado; e a análise baseada em simulação, que testa hipóteses com modelos *in silico*, fornecendo previsões passíveis de serem testadas por estudos experimentais tanto *in vitro*, como *in vivo*. O objetivo da biologia computacional guiada por descoberta é definir todos os elementos em um sistema e criar um banco de dados que contenha essa informação (23). A integração dessas duas abordagens, guiada por dados e orientada por hipóteses, é uma das responsabilidades da biologia de sistemas (23).

Atualmente, a biologia computacional tem um papel fundamental na interpretação da avalanche de novas sequências de ácidos nucleicos geradas. O número de espécies com genomas sequenciados está crescendo continuamente, sem sinais de desaceleração (25,26). Essa explosão no volume de dados genômicos aliados à ciência guiada por descoberta representam uma fonte fundamental de dados relevantes para a pesquisa biológica multidimensional em diversas áreas, como genômica comparativa, filogenia e evolução, biologia da conservação e ciências biomédicas (19,26–29).

Antes do sequenciamento de alto desempenho, uma estratégia comum e ainda em uso da biologia celular e molecular é fazer uso de organismos modelo que servem como representantes para melhor compreensão da biologia humana. Os esforços de sequenciamento dos genomas desses modelos, pelo método de Sanger, são focados em algumas espécies filogeneticamente isoladas, sem levar em conta a grande diversidade de vertebrados, onde aproximadamente 50% dessas espécies são peixes (30,31). Como exemplo, o sequenciamento do genoma do peixe-zebra (*Danio rerio*) mostrou que aproximadamente 70% dos seus genes têm, pelo menos, um ortólogo em humanos, além de facilitar a identificação e a caracterização de mutações causadoras de doenças (32). O conhecimento acumulado em decorrência do sequenciamento do genoma de *D. rerio*, fortaleceu o uso dessa espécie como modelo vertebrado para estudo de desenvolvimento, toxicologia e para diversas doenças humanas (33,34). Esse modelo demonstrou convincentemente a utilidade de se usar uma espécie de peixe para melhorar nossa compreensão dos mecanismos moleculares e celulares que levam a condições patológicas e para o desenvolvimento de novas ferramentas diagnósticas e terapêuticas.

Apesar da utilidade do peixe-zebra na investigação de um amplo espectro de atributos da sua biologia, há evidências que sugerem que outras espécies de peixes podem ser mais adequadas para questões mais específicas (35). Na última década, o crescimento no número de genomas de teleósteos[3] disponíveis (Figura 1) fez emergir um verdadeiro exército de novas espécies modelo para ajudar a responder questões sobre as bases genéticas de importantes adaptações, incluindo modelos de doenças humanas (30,35).

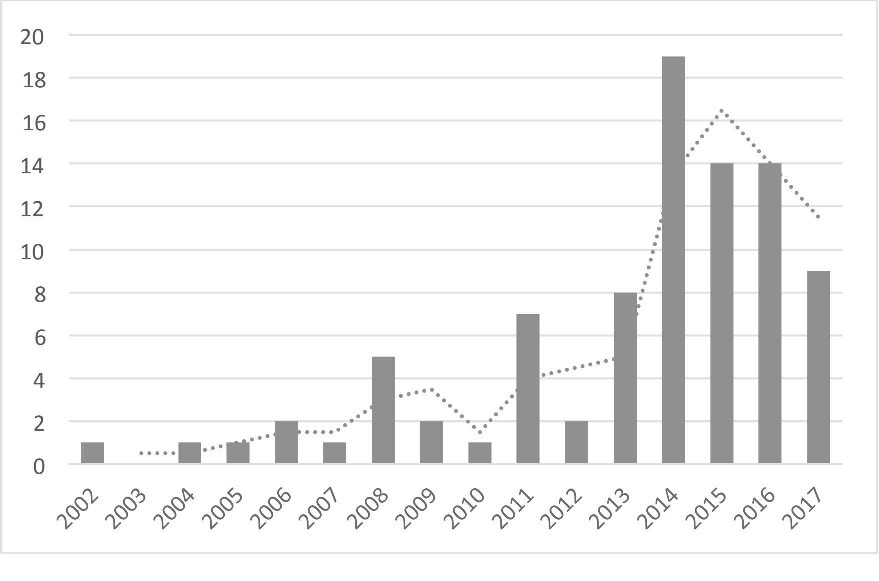

Figura 1: Número de genomas de espécies de peixes teleósteos depositados no GenBank por ano. O número absoluto de genomas de teleósteos é mostrado por ano, desde 2002, em barras sólidas. A linha pontilhada representa a média móvel de dois anos consecutivos (Fonte: NCBI, https://www.ncbi.nlm.nih.gov/).

[3] Teleósteos são organismos pertencentes a infra-classe Teleostei, a mais abundante entre os peixes de nadadeiras raiadas, Actinopterygii.

O conceito de modelos mutantes evolutivos, introduzido por Albertson *et al.* (2009) e revisado por Schartl (2014), se baseia em explorar nos organismos um fenótipo adaptativo, moldado pela evolução, que se assemelhe à doenças humanas. Dessa forma, espera-se obter uma melhor compreensão dos mecanismos de desenvolvimento dessa característica, com o objetivo final de aprimorar o diagnóstico e a terapia da doença humana. Os fenótipos que modelam no animal a doença humana podem ser divididos em duas classes, a primeira onde o estado da doença é adaptativo para a espécie modelo e a segunda onde os animais são acometidos pela mesma doença que os humanos (35). Como exemplo pode-se citar os peixes antárticos, da subordem Notothenioidei, que são modelos de osteoporose e anemia (35). Esses peixes possuem considerável desmineralização dos ossos para aumentar sua flutuabilidade e seu fenótipo é fruto de seleção positiva que confere vantagens adaptativas à espécie. O peixe de caverna *Astyanax mexicanus*, é outro exemplo de espécie modelo (p.ex. degeneração de retina, albinismo e desordens do sono) cujo fenótipo similar a patologia humana é adaptativa para a espécie (35). Nesse caso, todavia, a origem dos fenótipos deu-se por regressão evolutiva. Essa espécie coexiste tanto em locais com luz, como em cavernas e são ótimos alvos para estudos comparativos. Como exemplos de fenótipos que levam os indivíduos a desenvolverem a patologia similar a humana podem ser citados: o gênero *Xiphophorus* e a espécie *Oncorhynchus mykiss*, (truta arco-íris) que são modelos de melanoma e câncer de fígado, respectivamente (35).

Ainda há outros fenótipos onde alguns peixes que são acometidos pelas mesmas condições patológicas que os seres humanos desenvolveram características que os tornam resistentes a tais condições, em particular aquelas causadas por xenobióticos, substâncias não nutrientes estranhas ao organismo provenientes da indústria ou da natureza. Como exemplo, temos o *Fundulus* *heteroclitus*, este peixe tem uma enorme plasticidade, ocupa diversos nichos ecológicos e possui alta tolerância à poluição proveniente de fontes industriais, agrícolas e municipais (35). Essas características tornam o *F. heteroclitus* um modelo mutante evolutivo para estudos toxicogenômicos.

Os métodos genômicos permitiram que vários modelos de peixes saíssem da ignorância genômica, transformando-os em um novo exército de modelos para uma análise abrangente das bases genéticas da evolução (30). O pouco que sabemos sobre as adaptações dessas espécies de peixes deve nos motivar a buscar, em toda a biodiversidade, novos modelos mutantes evolutivos que certamente existem. Todavia, os 87 genomas sequenciados em um universo de mais de 33 mil espécies de peixes ainda representam um número insuficiente de dados para caracterizar a biodiversidade em toda sua multidimensionalidade.

Outra abordagem, utilizada em estudos comparativos, é o sequenciamento e montagem *de novo* de transcriptomas. Essa é uma maneira econômica de obter informações sobre o conteúdo de genes de uma espécie não modelo, especialmente quando o objetivo é comparar múltiplas espécies e quando não se tem um genoma de referência de uma espécie estreitamente relacionada. As abordagens transcriptômicas estão ganhando impulso na biologia (Figura 2), como demonstrado pelo maior crescimento percentual anual no GenBank do “Transcriptome Shotgun Assembly (TSA) Database” (37). Entretanto esse crescimento ainda é focado em espécies isoladas, sem uma perspectiva evolutiva e oportunidades emergentes residem no uso criativo de métodos moleculares e/ou computacionais comparativos para revelar os processos que influenciam a diversidade da vida selvagem. No entanto, antes que espécies com interessantes adaptações possam se tornar modelos e entrar em uma fase de investigações funcionais, é necessária uma caracterização genética, a lista das partes, seguindo os princípios da biologia de sistemas guiados por uma ciência de descoberta.

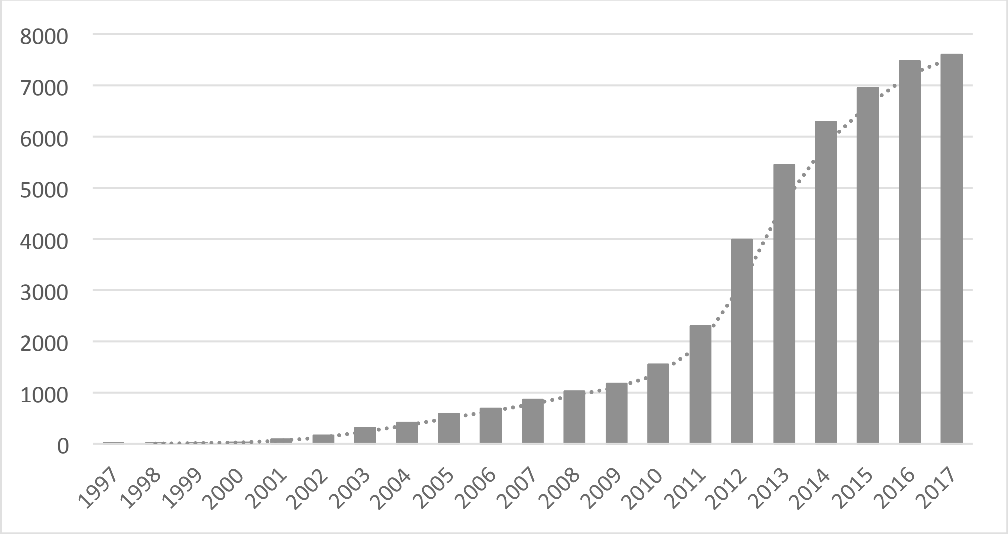

Figura 2: Número de trabalhos indexados no Pubmed na área de transcriptômica. O número total de artigos que contém a palavra "transcriptome" é mostrado por ano, desde 1997, em barras sólidas. A linha pontilhada representa a média móvel de dois anos consecutivos (Fonte: NCBI, https://www.ncbi.nlm.nih.gov/pubmed/?term=transcriptome).

Entre as mais de 33 mil espécies de peixes, aproximadamente 50% são de água doce (38). A região Neotropical possui a maior riqueza de espécies de peixes do mundo (39) e alto grau de endemismo. Entre as famílias endêmicas, destaca-se a Loricariidae por ser a quinta família mais diversa entre os vertebrados, com mais de 900 espécies válidas (40). Contraditoriamente, essa diversidade vem junto com uma grande lacuna sobre sua genética (41) e outras questões básicas, mas relevantes para compreensão de sua biologia e seus papéis na manutenção de um ambiente saudável, como profundas incertezas filogenéticas. O uso recente de alguns marcadores genéticos melhorou a filogenia proposta para esse grupo, mas não conseguiu resolver algumas questões, como por exemplo a respeito da monofilia da subfamília Hypoptopomatinae (42–44). O conhecimento prévio das relações filogenéticas entre os organismos em estudo é essencial para se poder interpretar corretamente a evolução da função dos genes. Essa grande diversidade taxonômica encontrada em Loricariidae também é acompanhada de uma grande diversidade ecológica. Os loricarídeos ocorrem em uma grande variedade de habitats, são predominantemente detritívoros e desempenham importante papel na ciclagem de nutrientes no ambiente aquático (45,46).

A importância desses peixes estende-se através de diferentes áreas do conhecimento, da economia à saúde ambiental e humana, exibindo inúmeras adaptações, como exemplo pode-se mencionar a elevada tolerância a hipóxia e variações de pH das espécies *Pterygoplichthys anisitsi* e *Pterygoplichthys pardalis* que podem ser usados como potenciais modelos mutantes evolutivos (47,48). Além disso, apresentam elevada resistência à toxicidade aguda por biodiesel (49), tornando o *Pterygoplichthys* um apropriado modelo para estudos toxicogenômicos. Além dessas adaptações, outras têm potencial biotecnológico, como é o caso da tripsina de *Pterygoplichthys disjunctivus*, que possui alta atividade em grandes variações de temperatura, pH e salinidade (50) e vem despertando interesse da indústria como possível fonte de uma protease adequada para aplicação em produtos fermentados.

Algumas espécies, principalmente dos gêneros *Pterygoplichthys* e *Hypostomus*, possuem relevância ecológica, já que diversas populações invasoras se estabeleceram em regiões tropicais e subtropicais em todo o mundo e também ameaçam as espécies nativas, alcançando densidades até duas ordens de magnitude maiores do que a biomassa de peixes nativos (51–56).

Na Amazônia, os loricarídeos estão ameaçados pela degradação do habitat e exploração comercial. Muitas espécies de cascudos são vendidas no mercado nacional e internacional de peixes ornamentais (57). O conhecimento das bases genéticas das populações de cascudos tem grande valor conservacionista e econômico, como a preservação do *Hypancistrus zebra*, espécie endêmica de uma região de 100 Km de extensão, chamada de volta grande do rio Xingu, que está ameaçada de extinção e sofre com a implantação da usina de Belo Monte e a sobre-exploração para aquarismo. Essa espécie teve seu genoma mitocondrial sequenciado recentemente, que pode ser usado como importante ferramenta genética no seu monitoramento populacional (58).

Outra importante adaptação, presente em pelo menos dois gêneros de loricarídeos, são substituições de aminoácidos na enzima citocromo P450 1A (CYP1A) que alteram sua especificidade por substrato, resultando na ausência de atividade etoxiresorufina-*O*-desetilase (EROD) (59–61). A atividade de EROD é um importante marcador de poluição ambiental e muito utilizada em biomonitoramento (62). A ausência desse fenótipo é algo extremamente raro, pois o gene CYP1A surgiu antes da radiação dos vertebrados e é altamente conservado entre seus *taxa* (63). Contudo, o significado dessas mudanças funcionais na especificidade do substrato de CYP1A para a fisiologia dos peixes, respostas toxicológicas e adaptação ao meio ambiente ainda não foram elucidados (61). Uma abordagem "top-down", como a genômica, com a caracterização de vários níveis biológicos desse sistema (de genes a famílias gênicas, de espécies a famílias taxonômicas, de um rio a diferentes bacias), permitiria a descoberta de novos componentes relacionados a essa e outras adaptações.

Tendo observado a grande diversidade da família Loricariidae e seu potencial uso em estudos de diversas áreas da biologia, essa tese tem como propósito maior usar a biologia computacional para a integração da ciência por descoberta e da ciência orientada por hipóteses na investigação da biodiversidade da subordem Loricarioidei, com ênfase na família Loricariidae.

O objetivo do Capítulo Um foi desenvolver um método para recuperar genomas mitocondriais a partir de dados de transcriptomas. Esse objetivo foi gerado após uma simples observação dos dados. Quando procurávamos pela sequência que codifica a enzima citocromo c oxidase subunidade I (COX1) encontramos um transcrito com quase um terço do tamanho esperado para toda a sequência do genoma mitocondrial. Essa descoberta nos levou a corroborar a hipótese que seria possível montar e inferir genomas mitocondriais a partir de dados de RNA-Seq. Os Capítulos Dois e Três aplicam a nova abordagem desenvolvida com o objetivo de montar os genomas mitocondriais, reduzindo a lacuna de informação genética desse grupo diverso. O Capítulo Quatro aliou essa metodologia a uma perspectiva evolutiva com objetivo de testar hipóteses filogenéticas e investigar a evolução estrutural desses genomas. No último Capítulo, ampliamos o objeto de pesquisa dos transcritos mitocondriais para todo o transcriptoma em uma determinada espécie, com o objetivo de investigar a evolução de famílias gênicas relacionadas à defesa química do organismo.

O conhecimento adquirido sobre as relações filogenéticas inter e intra-subfamílias foi essencial para darmos continuidade às análises dos dados. No Anexo A ampliamos a análise de todo o transcriptoma de uma para todas as espécies deste trabalho, com o objetivo de investigar a evolução de famílias gênicas e buscar evidências de seleção adaptativa que fomentaram a diversidade genética, fenotípica e ecológica desse rico grupo de organismo.

2 OBJETIVOS

2.1 Objetivo geral

Detectar e descrever a variação biológica acessando a dimensão genética em dados de 40 transcriptomas da subordem Loricarioidei, com ênfase na família Loricariidae.

2.2 Objetivos específicos

- Desenvolver uma abordagem para montar genomas mitocondriais a partir de dados transcriptômicos (Capítulo Um).
- Montar, anotar e analisar a estrutura de genomas mitocondriais (Capítulos Dois e Três).
- Testar hipóteses filogenéticas e investigar a evolução estrutural dos genomas mitocondriais montados (Capítulos Três e Quatro).
- Investigar a evolução de famílias gênicas relacionadas à defesa química do organismo no transcriptoma de *Pterygoplichthys anisitsi* (Capítulo Cinco).
- Testar hipóteses filogenéticas, investigar a evolução de famílias gênicas e buscar evidências de seleção adaptativa usando os transcritos nucleares dos 40 transcriptomas (Anexo A).

3 CAPÍTULOS

3.1 CAPÍTULO UM: O uso de dados transcriptômicos de sequenciamento de nova geração para montar genomas mitocondriais de *Ancistrus* spp. (Loricariidae)

Neste capítulo, o objetivo foi desenvolver uma abordagem inovadora para montar genomas mitocondriais a partir de dados transcriptômicos. Os resultados aqui descritos foram publicados no artigo intitulado “The use of transcriptomic next-generation sequencing data to assemble mitochondrial genomes of *Ancistrus* spp. (Loricariidae)”, na revista Gene, ano 2015; vol.573(1):171–5.

Para alcançar nosso objetivo, utilizamos o RNA total extraído do fígado e a tecnologia Illumina HiSeq. A partir dos dados produzidos, recuperamos os transcritos mitocondriais de três peixes (*Ancistrus* spp.) e montamos os seus mitogenomas. Com base na sequência de DNA de uma espécie proximamente relacionada, estimamos ter sequenciado de 92% a 99% dos mitogenomas desses três indivíduos. Considerando as três sequências juntas, sequenciamos todos os elementos padrões de mitogenomas de vertebrados, 13 genes codificadores de proteínas, dois RNAs ribossômicos, 22 RNAs transportadores e a região controle. O uso de dados transcriptômicos permitiu a observação do padrão de pontuação da maturação do mtRNA, a análise do perfil transcricional e a detecção de sítios heteroplasmáticos. A montagem do mtDNA a partir de dados transcriptômicos supera algumas limitações das estratégias tradicionais para o sequenciamento de mitogenomas. Essa abordagem é, de fato, mais útil para pesquisas que usam dados de RNA-Seq, especialmente aquelas com espécies não modelo, sobre as quais pouco se sabe sobre o organismo.

Gene 573 (2015) 171–175

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Short communication

The use of transcriptomic next-generation sequencing data to assemble mitochondrial genomes of *Ancistrus* spp. (Loricariidae)☆

Daniel A. Moreira [a], Carolina Furtado [b], Thiago E. Parente [a],*

[a] Laboratório de Toxicologia Ambiental, DCB, Escola Nacional de Saúde Pública — ENSP, Fundação Oswaldo Cruz — FIOCRUZ, Rio de Janeiro, Brasil
[b] Divisão de Genética, Instituto Nacional do Cancer — INCA, Rio de Janeiro, Brasil

ARTICLE INFO

Article history:
Received 28 April 2015
Received in revised form 20 August 2015
Accepted 28 August 2015
Available online 3 September 2015

Keywords:
Mitogenome
Mitochondrial transcripts
Transcriptome
RNA
Illumina
Siluriformes
Biodiversity

ABSTRACT

Mitochondrial genes and genomes have long been applied in phylogenetics. Current protocols to sequence mitochondrial genomes rely almost exclusively on long range PCR or on the direct sequencing. While long range PCR includes unnecessary biases, the purification of mtDNA for direct sequencing is not straightforward. We used total RNA extracted from liver and Illumina HiSeq technology to sequence mitochondrial transcripts from three fish (*Ancistrus* spp.) and assemble their mitogenomes. Based on the mtDNA sequence of a close related species, we estimate to have sequenced 92%, 95% and 99% of the mitogenomes. Taken the sequences together, we sequenced all the 13 protein-coding genes, two ribosomal RNAs, 22 tRNAs and the D-loop known in vertebrate mitogenomes. The use of transcriptomic data allowed the observation of the punctuation pattern of mtRNA maturation, to analyze the transcriptional profile, and to detect heteroplasmic sites. The assembly of mtDNA from transcriptomic data is complementary to other approaches and overcomes some limitations of traditional strategies for sequencing mitogenomes. Moreover, this approach is faster than traditional methods and allows a clear identification of genes, in particular for tRNAs and rRNAs.

1. Introduction

The many biological functions of mitochondria and the uses of mitochondrial genes have created a huge interest in sequencing mitogenomes (Huang, 2011). The initial approach to sequence mtDNA was to isolate and fragment it with restriction enzymes, clone the fragments in bacterial plasmids and expand the fragment-plasmid construct in bacteria to obtain enough DNA to be sequenced by the Sanger method (Anderson et al., 1981). Although many different approaches have been developed to isolate mtDNA, it is still a challenge to obtain this molecule with enough quantity, quality and purity for sequencing purposes (Jayaprakash et al., 2015). The sequencing of mitogenomes was made a lot easier by the establishment of the Polymerase Chain Reaction (PCR) as a powerful molecular biology technique. Using long range PCR, overlapping regions of mtDNA could be amplified and sequenced, as before, by the Sanger method (Hu et al., 2002). The complete mtDNA molecule could be assembled by joining the overlapping regions of each PCR amplicon (Hu et al., 2007).

First attempts to use mitochondrial transcripts to sequence mitogenomes were based on Sanger sequencing of expressed sequence tags (ESTs) (Gissi and Pesole, 2003; Samuels et al., 2005). Mitochondrial genomes are transcribed as a single polycistronic transcript, which has the coding sequence of all the 13 intron-less proteins, 22 tRNAs and the two subunits of the ribosomal RNA coded by most metazoan mitogenomes (Bernt et al., 2013a; Friedman and Nunnari, 2014). After transcription, the polycistronic transcript is cleaved on “punctuation marks”, giving origin to mature mitochondrial tRNA, rRNA and mRNA (Ojala et al., 1981). Mature mitochondrial mRNAs are monocistronic or bicistronic and polyadenilated (Bernt et al., 2013a).

More recently, Next-Generation Sequencing (NGS) technologies are being used to study heteroplasmy in humans (Huang, 2011; Jayaprakash et al., 2015) and to sequence mitogenomes from model and non-model organisms (Besnard et al., 2014; Dames et al., 2015; Hahn et al., 2013; Jex et al., 2010), including museum specimens (Fabre et al., 2013). However, most descriptions of mitogenomes by NGS uses mtDNA originated from long range PCR or directly isolated from animal tissues (Dames et al., 2015; Payne et al., 2015; Quispe-Tintaya et al., 2015), and hence, the challenge to isolate this molecule with enough quantity, quality and purity, along with the biases introduced by PCR reactions, remain largely unchanged.

Here, we used mitochondrial RNA (mtRNA), rather than mtDNA, to sequence mitogenomes. The recent advances on RNA-Seq allowed the generation of more than 40 millions 100 bp paired-end Illumina HiSeq2500 reads per sample, using as start material total RNA extracted

Abbreviations: mtDNA, mitochondrial DNA; rRNA, ribosomal RNA; tRNA, transfer RNA; mRNA, messenger RNA; PCR, Polymerase Chain Reaction; ESTs, expressed sequence tags; NGS, Next-generation Sequencing; CDS, coding sequence.

☆ The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

* Corresponding author.
E-mail address: parente@ensp.fiocruz.br (T.E. Parente).

http://dx.doi.org/10.1016/j.gene.2015.08.059
0378-1119/© 2015 Elsevier B.V. All rights reserved.

from liver by the simple phenol:chloroform extraction. Based on transcriptomic NGS data, the mitogenomes of three *Ancistrus* spp. individuals were assembled. The use of mtRNA to sequence mitogenomes is straightforward and overcomes some of the current challenges to isolate mtDNA and the biases introduced by PCR. More importantly, the use of transcriptomic data enables the sequencing of mitogenomes, while still sequence the complete coding region of thousands of nuclear protein-coding genes, and is a valuable approach in face of the fast increasing number of transcriptomic researches using non-model species.

2. Material and methods

2.1. RNA extraction

RNA samples were extracted from the liver of three individual fish of the genus *Ancistrus*, promptly preserved in RNA later and kept at −20°C. The fish were deposited at the Museu Nacional do Rio de Janeiro under the voucher number MNRJ42890. RNA extractions were performed using either TRIzol (Invitrogen) or TRI Reagent (Life technologies) following manufacturer's instructions. 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 sequencing

The complementary DNA (cDNA) libraries 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. 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 (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. A 100 bp single-end and another 100 bp paired-end sequencing reactions were performed in a HiSeq2500 using the TrueSeq SBS kit v3 (Illumina). The library preparation and sequencing reaction were performed at the Divisão de Genética of the Instituto Nacional do Cancer (INCA) in Rio de Janeiro, Brazil.

2.3. Bioinformatic processing

Raw Illumina data were demultiplexed using the BCL2FASTQ software (Illumina). Reads were trimmed for Illumina adaptors by Trimmomatic (Bolger et al., 2014) and its quality were evaluated using FastQC (Babraham Bioinformatics). Only reads with PHRED score > 30 were used for the transcriptome assembly, which was performed using the concatenated reads from the single and paired-ends reactions and the default parameters of Trinity v. 2.0.2 (Grabherr et al., 2011; Haas et al., 2013). Although Trinity was used to assemble the transcriptomes, many *de novo* assemblers are available and should have similar performances.

Each transcriptome was subjected to BLASTX searches against two databases, the Uniprot entries of humans (*Homo sapiens*), and the Uniprot entries of zebrafish (*Danio rerio*) (Altschul et al., 1990; Consortium, 2015). An additional BLASTN was performed against the complete mitogenome of the closest related species, whose mitogenome is publically available, *Pterygoplichthys disjunctivus* (GI: 339506171) (Nakatani et al., 2011). The transcripts aligned with the reference mitogenome were used for mitogenomes assembly. Selected transcripts were edited according to the information of strand orientation given by the BLASTN result, and aligned by SeaView using the built-in CLUSTAL alignment algorithm to the reference mitogenome (Gouy et al., 2010). A CONTIG sequence was generated using the sequence information of just the transcripts of each individual fish. The CONTIG sequence was then manually checked for inconsistencies and gaps. The mitogenomes were annotated using the web-based services MitoFish and MITOS (Bernt et al., 2013b; Iwasaki et al., 2013). In order to estimate the support of each base of the mitogenomes, Bowtie v. 1.0.0 was used to align the reads of each fish on its own assembled mitogenome, and this mapping was 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). Heteroplasmic sites were detected using IGV, setting the software to show positions in which the frequency of the second most frequent base was equal to or higher than 10% and the total reads number were higher than 100.

3. Results

3.1. Library construction, Illumina sequencing and transcriptome assemble

High quality RNA (RIN > 7.0) was extracted from the three fishes (Table 1). The size of the three libraries ranged from 230 to 800 base pairs (bp), and the total number of HiSeq reads from 43.5 to 60.1 million (Table 1). Each of the three transcriptomes is composed by more than 60 thousand transcripts (Table 1). In all cases, around 50% of the assembled transcripts have a BLASTX hit (Table 1).

3.2. Mitogenome assembling and annotation

In order to assemble the mitogenomes, 7 to 13 transcripts of each fish were used (Table 1 and see Table 1 in (Daniel et al., submitted for publication)). In comparison to the complete mitogenome of *Pterygoplichthys disjunctivus*, which is the closest relative of *Ancistrus* spp. with the complete mitogenome available, we sequenced 99.2% of the mtDNA of *Ancistrus* sp. #1, 92.5% of the mtDNA of *Ancistrus* sp. #2a, and 94.7% of the mtDNA of *Ancistrus* sp. #2b (Table 1). The mitogenomes are deposited in GenBank under the GIs: KP960569, KP960568, KP960567, respectively. *Ancistrus* sp. #2a and *Ancistrus* sp. #2b are from the same species, as they share 99.8% identical nucleotides at their mitogenomes. *Ancistrus* sp. #1 is probably a different species, as its mitogenome differs from the other two fish on 6% of the nucleotide positions.

In terms of features number, the mitogenome of *Ancistrus* sp. #2b is the most complete of three *Ancistrus* spp. Having a single gap, which prevented its circularization, we sequenced the 13 protein-coding genes, the two ribosomal RNAs and 21 of the 22 tRNAs in *Ancistrus* sp. #2b (Fig. 1). Its single gap contains the sequence of the displacement loop (D-loop) and the $tRNA_{phe}$. In the mitogenome of *Ancistrus* sp. #1 just three tRNAs are missing: $tRNA_{leu2}$, $tRNA_{his}$, $tRNA_{ser1}$ (Fig. 1). The mtDNA of *Ancistrus* sp. #2a has the lowest coverage, missing the D-loop, and four tRNAs; $tRNA_{phe}$, $tRNA_{leu2}$, $tRNA_{pro}$, $tRNA_{thr}$ (Fig. 1). Supplemental information about the mitogenome features are provided elsewhere (see Table 3 (Daniel et al., submitted for publication)).

Interestingly, the number of supporting reads of each nucleotide varied greatly according to the position in the mitogenome (Fig. 1). More than 84,000 reads were found to support the sequences of

Table 1
Summary of the transcriptome and mitogenome data for the three fish (*Ancistrus* spp.).

	Ancistrus sp. #1	*Ancistrus* sp. #2a	*Ancistrus* sp. #2b
RNA Integrity Number − RIN	8.2	7.4	>7.00
Library insert size (bp)	230–800	268–792	285–370
Reads after QC	43,502,597	53,961,751	60,170,745
Transcripts			
Total	67,098	63,847	67,883
With BLASTX hit	35,710	31,886	33,953
For mitogenome	13	12	7
mtRNA reads (%)	2.6	1.8	0.8
Mitogenome coverage (%)	99.2	92.5	94.7
Heteroplasmic sites	10	8	6

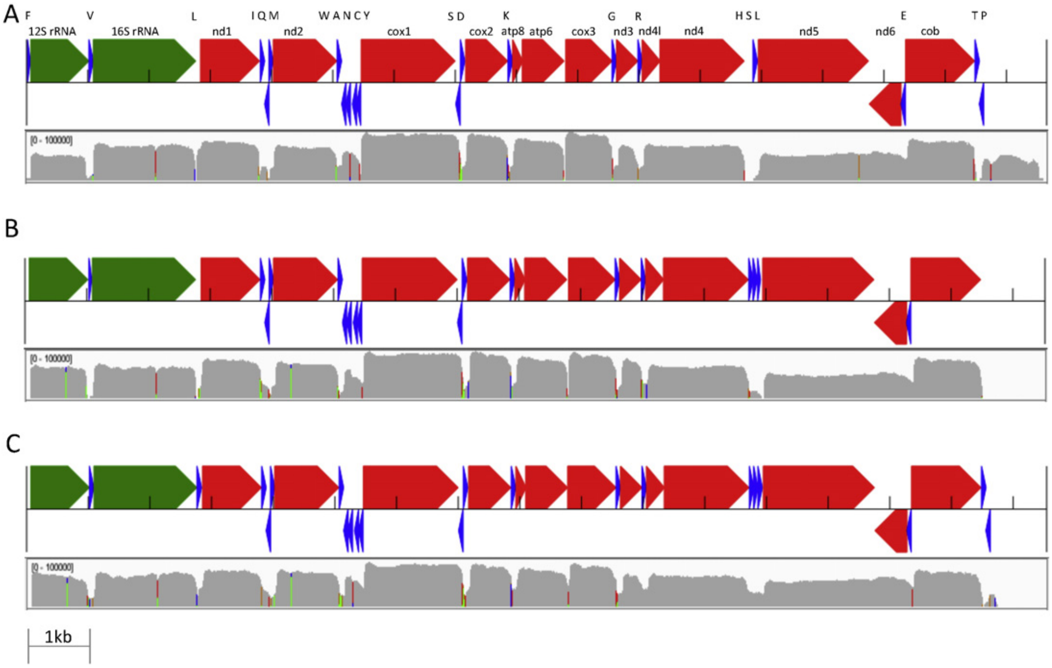

Fig. 1. The assembled mitogenomes of *Ancistrus* sp.#1 (A), *Ancistrus* sp.#2a (B) and *Ancistrus* sp.#2b (C). The number of supporting reads along the sequence is shown in a logarithm scale and below the schematic view of each mitogenome. Heteroplasmic sites are highlighted with different colors on the graphic of supporting reads. Each mitogenomes feature is named at the top of the figure. The tRNAs are named using the one-letter code of the amino acid they transport.

protein-coding genes, especially the Cytochrome Oxidase genes (see Table 1 in (Daniel et al., submitted for publication)). Much less support was found for the sequences of tRNAs. *Ancistrus* sp. #1 was the sample with the highest rate of Illumina reads mapped against its mitogenome, 2.6% of the reads were aligned against its mtDNA, while *Ancistrus* sp. #2b had the lowest, 0.83% (Table 1). This percentage is an approximate value of mitochondrial transcripts representation in the transcriptome (mitochondrial + nuclear).

Finally, using the approach to assemble mitogenomes from transcriptome data, 13 unique heteroplasmic sites were detected in the three mitogenomes (Fig. 1 and see Table 2 in (Daniel et al., submitted for publication)). The heteroplasmies are distributed among intergenic region (one), rRNAs (three), tRNAs (two) and protein-coding(seven) genes. Among the seven heteroplasmic sites in protein-coding genes, four are located in the first codon position, two in the second and one in the third. The eight heteroplasmies found in *Ancistrus* sp. #2a, as well as the six found in *Ancistrus* sp. 2b are shared with at least another fish. Ten heteroplasmatic sites were found in *Ancistrus* sp. #1, six are shared and four are exclusive.

4. Discussion

Reports on the use of transcriptome NGS data to assemble vertebrate mitochondrial genomes are yet very scarce and recent in the literature (Dilly et al., 2015; Fabre et al., 2013; Mercer et al., 2011). Mercer et al. (2011) have studied the human mitochondrial transcriptome using strand specific Illumina GAII NGS platform and were able to assemble 99.9% of the mitochondrial heavy strand, which harbor most of mitochondrial coding sequences (CDS), and 94.6% of the light strand, which encodes ND6 gene and eight tRNAs (Mercer et al., 2011). However, in order to study the human transcriptome specifically from the mitochondria, Mercer et al. used a refined protocol to extract RNA from purified fractions of this organelle derived from a stable cell line. In contrast, we have used total RNA extracted by the straightforward method of phenol:chloroform from liver preserved in RNALater at −20 °C, and yet obtained up to 99.2% of the complete mitogenome.

Our report resembles more the ones from Fabre et al. (2013)) and Dilly et al. (2015). However, while those authors used Roche 454 pyrosequencing, we used the Illumina Hi-Seq2500 platform. Fabre et al. (2013) assembled the mitochondrial protein-coding genes, but were unable to assemble any of the 22 tRNAs. They attributed this inability to the punctuation pattern of mitochondrial gene transcription and the rapid elimination of tRNA sequences; to the differential expression of mitochondrial transcripts; but also to the relative low coverage power of 454 sequencing approach. Indeed, those authors foresaw ultra deep NGS approaches, like the one we used, would better detect transient tRNA molecules.

More recently Dilly et al.(2015) detected some tRNA molecules using 454 platform. However, using the Illumina platform, we were able to sequence more tRNAs; 21 out of the 22 tRNAs present in mitogenomes. In fact, if the three fishes were from the same species, it would have been possible to construct a contig sequence with the complete circular mitogenome, as the tRNA_{phe} and D-loop region missing in *Ancistrus* sp. #2b are present in *Ancistrus* sp. #1. The three specimens used in this work were donated as if they were the same species, but unfortunately, the differences on the mtDNA between *Ancistrus* sp. #1 and the other two fish made evident that they are different species.

Illumina transcriptomic data also made possible the study of transcription patterns and maturation of mitochondrial genes, and the detection of heteroplasmic sites. Similarly to Mercer et al. (2011), the expression of different mitochondrial transcripts varied widely in our analysis. Given their common polycistronic source, our data corroborates the existence of posttranscriptional regulatory mechanisms in fish species. In contrast, however, the mitochondrial contribution to the pool of cellular poly-A RNA was lower in this study than the one reported by Mercer et al. (2011) for human liver. In addition, our findings show the three Cytochrome Oxidases as the most abundant mitochondrial transcripts, rather than the two ribosomal RNAs indicated by Mercer et al. (2011). This can be a species-specific difference, but is most likely due to the different source of extracted RNA; total cellular RNA in our case, and long and short preparations of RNA from isolated mitochondria used by Mercer et al. (2011).

Despite the wide variation of supporting read counts in the assembled mitogenomes, high support was obtained for most regions of protein and rRNA coding genes. The exceptions were usually located at both ends of every transcript, and more notably, at the D-loop region. The D-loop region was sequenced with high number of supporting reads in *Ancistrus* sp. #1, but is missing in *Ancistrus* sp. #2a and #2b. This is more likely to be a species-specific difference, as the D-loop was sequenced in one species of *Ancistrus*, but not in the other. The D-loop region controls both the replication and the transcription of mtDNA (Bernt et al., 2013a) and is generally not polyadenylated or present on mature transcripts. Therefore, it is expected that the D-loop will be among the regions with the lowest numbers of supporting reads, if any, when using transcriptome data to assemble mitogenomes. The D-loop region and other gaps in mitogenomes constructed from transcriptome data can be filled by regular short-range PCR reactions with species-specific primers.

In mammals, the transcription of mitogenomes starts at three sites located in the D-loop region (Bernt et al., 2013a). The primary polycistronic transcripts are matured according to the RNA punctuation model (Ojala et al., 1981). It has been shown that this RNA maturation step cause the isolation into monocistronic mRNA of every protein-coding gene, rRNA and tRNA of the mitogenome, except two bicistronic mRNA coding for nad4/nad4L and atp8/atp6 (Bernt et al., 2013a). The punctuation positions are evident on our assembly of *Ancistrus* spp. mitogenomes from transcriptome data; as the number of supporting reads drops dramatically between almost all genes in the mitogenome, frequently co-localized with the position of the tRNAs. The two exceptions are exactly the two pairs, nad4/nad4L and atp8/atp6, known to code for bicistronic transcripts. The data indicate mitochondrial transcripts maturation in fish species follows a similar pattern from the mammals.

A mitochondrion may harbor many copies of its genome and each cell may possess up to thousands mitochondria. This makes possible the occurrence of many different mtDNA haplotypes in a single specimen and even in a single cell, an event called heteroplasmy (Huang, 2011; Jayaprakash et al., 2015; Sosa et al., 2012). The accurate determination of heteroplasmic sites is challenging because it is very sensitive to the errors introduced by PCR, and to the unpredictable inaccuracies introduced by nuclear mitochondrial DNA (Numts) (Jayaprakash et al., 2015). The use of transcriptome data to assemble mitogenomes detects heteroplasmatic sites in an accurate, reliable and straightforward manner. The number of heteroplasmatic sites found in *Ancistrus* spp. is higher than the 1 to 3 most frequently found in humans (Sosa et al., 2012). It is possible, however, that at least some of the polymorphic sites are due to the presence of functional nuclear copies of mitochondrial DNA (recently integrated in the nuclear genome) or to partial maturation of the transcripts.

The assembly of mtDNA from Illumina transcriptomic data overcomes some limitations of traditional strategies for sequencing mitogenomes and to detect heteroplasmy. Specifically, this strategy use straightforward method of total cellular RNA extraction surpassing the need to obtain high quantity of high quality mtDNA; it does not suffer from the difficulties and biases overlapping long range PCR can represent and introduce; and hardly is compromised by Numts. Illumina platform yields are higher than 454 approach used by other reports and, thus, enabled the sequence of a higher percentage of mitogenomes. Moreover, the use of Illumina transcriptome data also allowed to study the expression and maturation of mitochondrial transcription, and to detect heteroplasmatic sites. Few small gaps in the assembled mtDNA from Illumina transcriptome data are still expected, especially at the D-loop and tRNAs regions. Those gaps can be easily filled by regular PCR reactions with species-specific primers and sequencing by Sanger or NGS techniques. This approach allows a fast and straightforward annotation of mitochondrial genes, especially tRNA and rRNA, which prediction is ineffective in some taxonomic groups (Besnard et al., 2014), while still sequences thousands of nuclear genes. This approach is in fact most useful to transcriptomic-oriented researches, especially those with non-model species.Despite the fast increasing number of transcriptomes of non-model species, almost none has assembled mitogenomes and, therefore, losses the opportunity to molecular characterize the species being used, and more importantly, to contribute to foster barcode initiatives.

Acknowledgments

We are grateful to the Divisão de Genética of the Instituto Nacional do Cancer (INCA) at Rio de Janeiro, Brazil for library preparation and sequencing, namely to Dr. Miguel Moreira and Dr. Héctor Abreu; and to the Setor de Ictiologia of the Museu Nacional, Universidade Federal do Rio de Janeiro, namely to Dr. Paulo A. Buckup and MSc. Emanuel Neuhaus. We acknowledge the use of Woods Hole Oceanographic Institute (WHOI) computer cluster, where TP holds a guest investigator appointment at the Biology Department. This work was supported by the Partnerships for Enhanced Engagement in Science (PEER) initiative from the United States Agency for International Development [grant number: PGA-2000003446] to TP.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Anderson, S., Bankier, a.T., Barrell, B.G., de Bruijn, M.H., Coulson, a.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.a., Sanger, F., et al., 1981. Sequence and organization of the human mitochondrial genome. Nature 290, 457–465.

Bernt, M., Braband, A., Schierwater, B., Stadler, P.F., 2013a. Genetic aspects of mitochondrial genome evolution. Mol. Phylogenet. Evol. 69, 328–338.

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

Besnard, G., Jühling, F., Chapuis, É., Zedane, L., Lhuillier, É., Mateille, T., Bellafiore, S., 2014. Fast assembly of the mitochondrial genome of a plant parasitic nematode (*Meloidogyne graminicola*) using next generation sequencing. C. R. Biol. 337, 295–301.

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

Consortium, T.U., 2015. UniProt: a hub for protein information. Nucleic Acids Res. 43, D204–D212.

Dames, S., Eilbeck, K., Mao, R., 2015. A high-throughput next-generation sequencing assay for the mitochondrial genome. Methods Mol. Biol. 1264, 77–88.

Daniel, A.,.M., Carolina, F., Parente, T.E.M., 2015. Mitochondrial transcripts and associated heteroplasmies of *Ancistrus* spp. (Siluriformes: Loricariidae). Data Br. (submitted for publication).

Dilly, G.F., Gaitán-Espitia, J.D., Hofmann, G.E., 2015. Characterization of the Antarctic sea urchin (*Sterechinus neumayeri*) transcriptome and mitogenome: a molecular resource for phylogenetics, ecophysiology and global change biology. Mol. Ecol. Resour. 15, 425–436.

Fabre, P.H., Jønsson, K.a., Douzery, E.J.P., 2013. Jumping and gliding rodents: mitogenomic affinities of Pedetidae and Anomaluridae deduced from an RNA-Seq approach. Gene 531, 388–397.

Friedman, J.R., Nunnari, J., 2014. Mitochondrial form and function. Nature 505, 335–343.

Gissi, C., Pesole, G., 2003. Transcript mapping and genome annotation of ascidian mtDNA using Est data. Genome Res. 13, 2203–2212.

Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView version 4: amultiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27, 221–224.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.a., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., et al., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8, 1494–1512.

Hahn, C., Bachmann, L., Chevreux, B., 2013. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads — a baiting and iterative mapping approach. Nucleic Acids Res. 41.

Hu, M., Chilton, N.B., Gasser, R.B., 2002. Long PCR-based amplification of the entire mitochondrial genome from single parasitic nematodes. Mol. Cell. Probes 16, 261–267.

Hu, M., Jex, A.R., Campbell, B.E., Gasser, R.B., 2007. Long PCR amplification of the entire mitochondrial genome from individual helminths for direct sequencing. Nat. Protoc. 2, 2339–2344.

Huang, T., 2011. Next generation sequencing to characterize mitochondrial genomic DNA heteroplasmy. Curr. Protoc. Hum. Genet. 19, 8.

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

Jayaprakash, a.D., Benson, E.K., Gone, S., Liang, R., Shim, J., Lambertini, L., Toloue, M.M., Wigler, M., Aaronson, S.a., Sachidanandam, R., 2015. Stable heteroplasmy at the single-cell level is facilitated by intercellular exchange of mtDNA. Nucleic Acids Res. 43, 2177–2187.

Jex, A.R., Littlewood, D.T.J., Gasser, R.B., 2010. Toward next-generation sequencing of mitochondrial genomes — focus on parasitic worms of animals and biotechnological implications. Biotechnol. Adv. 28, 151–159.

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).

Mercer, T.R., Neph, S., Dinger, M.E., Crawford, J., Smith, M.a., Shearwood, A.M.J., Haugen, E., Bracken, C.P., Rackham, O., Stamatoyannopoulos, J.a., et al., 2011. The human mitochondrial transcriptome. Cell 146, 645–658.

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.

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.

Ojala, D., Montoya, J., Attardi, G., 1981. TRNA punctuation model of RNA processing in human mitochondria. Nature 290, 470–474.

Payne, B.A.I., Gardner, K., Coxhead, J., Chinnery, P.F., 2015. Deep resequencing of mitochondrial DNA. Methods Mol. Biol. 1264, 59–66.

Quispe-Tintaya, W., White, R.R., Popov, V.N., Vijg, J., Maslov, A.Y., 2015. Rapid mitochondrial DNA isolation method for direct sequencing. Methods Mol. Biol. 1264, 89–95.

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.

Samuels, A.K., Weisrock, D.W., Smith, J.J., France, K.J., Walker, J.a., Putta, S., Voss, S.R., 2005. Transcriptional and phylogenetic analysis of five complete ambystomatid salamander mitochondrial genomes. Gene 349, 43–53.

Sosa, M.X., Sivakumar, I.K.A., Maragh, S., Veeramachaneni, V., Hariharan, R., Parulekar, M., Fredrikson, K.M., Harkins, T.T., Lin, J., Feldman, A.B., et al., 2012. Next-generation sequencing of human mitochondrial reference genomes uncovers high heteroplasmy frequency. PLoS Comput. Biol. 8.

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.

3.2 CAPÍTULO DOIS: Uma abordagem baseada em RNA para sequenciar o mitogenoma de *Hypoptopoma incognitum* (Siluriformes: Loricariidae)

Neste capítulo utilizamos a técnica descrita no capítulo anterior com o objetivo de montar e anotar o genoma mitocondrial completo de *Hypoptopoma incognitum*. Os resultados aqui descritos foram publicados no artigo intitulado “An RNA-based approach to sequence the mitogenome of *Hypoptopoma incognitum* (Siluriformes: Loricariidae)”, na revista Mitochondrial DNA Part A, ano 2016; vol.27(5):3784–6.

Hypoptopoma incognitum é um peixe da quinta família mais rica de vertebrados e abundante em rios da Amazônia brasileira. Na data da publicação, apenas duas espécies de peixe da família Loricariidae tinham sua sequência do genoma mitocondrial completa depositada no GenBank, mas este foi o primeiro registro cujo espécime foi depositado em coleção ictiológica e coletado em seu habitat nativo. A metodologia baseada em RNA foi utilizada para montar o mitogenoma completo de *H. incognitum* com uma profundidade de sequenciamento média de 5.292 vezes. As características mitocondriais típicas dos vertebrados foram encontradas; 22 genes de tRNA, dois genes de rRNA, 13 genes codificadores de proteínas e uma região controle não codificante. Além disso, o uso dessa abordagem permitiu a medição dos níveis de expressão do mtRNA, o padrão de pontuação da edição pós-transcricional e a detecção de heteroplasmias.

Mitochondrial DNA

http://informahealthcare.com/mdn
ISSN: 1940-1736 (print), 1940-1744 (electronic)

Mitochondrial DNA, Early Online: 1–3
© 2015 Taylor & Francis. DOI: 10.3109/19401736.2015.1079903

Taylor & Francis
Taylor & Francis Group

SHORT COMMUNICATION

An RNA-based approach to sequence the mitogenome of *Hypoptopoma incognitum* (Siluriformes: Loricariidae)

Daniel Andrade Moreira[1], Maithê G. P. Magalhães[1], Paula C. C. de Andrade[1], Carolina Furtado[2], Adalberto L. Val[3], and Thiago Estevam Parente[1]

[1]Laboratório de Toxicologia Ambiental, Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, Brasil, [2]Divisão de Genética, Instituto Nacional do Cancer (INCA), Rio de Janeiro, Brasil, and [3]Laboratório de Ecofisiologia e Evolução Molecular, Instituto Nacional de Pesquisas da Amazônia (INPA), Manaus, Brasil

Abstract

Hypoptopoma incognitum is a fish of the fifth most species-rich family of vertebrates and abundant in rivers from the Brazilian Amazon. Only two species of Loricariidae fish have their complete mitogenomes sequence deposited in the Genbank. An innovative RNA-based approach was used to assemble the complete mitogenome of *H. incognitum* with an average coverage depth of 5292×. The typical vertebrate mitochondrial features were found; 22 tRNA genes, two rRNA genes, 13 protein-coding genes, and a non-coding control region. Moreover, the use of this approach allowed the measurement of mtRNA expression levels, the punctuation pattern of editing, and the detection of heteroplasmies.

Keywords

Fish, heteroplasmy, mtRNA, RNA-Seq, transcriptome

History

Received 12 June 2015
Revised 9 July 2015
Accepted 18 July 2015
Published online 11 September 2015

Introduction

Loricariidae catfishes comprise the fifth most speciose vertebrate family (Lujan et al., 2015). *Hypoptopoma* is a genus within the Hypoptomatinae sub-family of Loricariidae, with 15 valid species distributed over the Neotropics, especially in the Amazon region (Aquino & Schaefer, 2010). Despite the diversity of loricariids, only two complete mitochondrial genomes of this family have been deposited in the Genbank to date, none of those from a Hypoptomatinae specie.

The next-generation sequencing (NGS) technologies promise to be a breakthrough on diverse areas of the biological sciences, generating unprecedented volume of sequence data. NGS techniques are already in use to sequence the mtDNA of several organisms (Hahn et al., 2013), including non-model species (Cui et al., 2009; Hu et al., 2007; Kocher et al., 2014; Jex et al., 2010), and to study heteroplasmy in humans (Dames et al., 2015; Huang, 2011; Seo et al., 2014; Sosa et al., 2012). Nonetheless, the current approaches to sequence mitogenomes by NGS technologies still rely on mtDNA originated from long range PCR or directly isolated from animal tissues (Dames et al., 2015; Payne et al., 2015; Quispe-Tintaya et al., 2015). Here, we present a RNA-based approach to sequence the complete mitochondrial genome of *Hypoptopoma incognitum*.

Methods

A single specimen of *Hypoptopoma incognitum* was collected in Rio Negro around the vicinities of the Instituto Nacional de Pesquisas da Amazônia (INPA) floating base of Catalão (03.16S 59.92W) in Manaus, Brazil. The fish is deposited at the Museu Nacional, UFRJ (MNRJ43412) and the identification was confirmed by ichthyologists. Total RNA was extracted from the liver tissue following conventional phenol–chloroform extraction. RNA quality was accessed using the RNA Nano kit for Bioanalyzer (Agilent, Santa Clara, CA). The cDNA library was constructed using the TruSeq RNA Sample kit v.2 (Illumina, San Diego, CA). Library was accessed for quality (Bioanalyzer, DNA1000 kit, Agilent, Santa Clara, CA) and quantity (Kapa Biosystems, Wilmington, MA). The run was performed in an Illumina HiSeq 2500 using the TrueSeq SBS kit v.3 (Illumina, San Diego, CA). A total of 32 570 486 100 bp paired-end reads were generated. The raw reads were trimmed for adaptor contamination (Trimmomatic) and its quality was evaluated using FastQC (Babraham Bioinformatics, Cambridge, UK). Only reads with Phred score over 30 were used for the transcriptome assembly, which was performed using the default parameters of Trinity (v. 2.0.6) (Haas et al., 2013). The transcriptome was subjected to a BLASTn search against the *P. disjunctivus* (GI: 339506171) and the *H. plecostomus* mitogenomes (GI: 722490383). The sequencing depth of each base was verified using Bowtie v. 1.0.0 to align the reads to the mitogenome. This mapping was visualized using the Integrated Genome Viewer (IGV) (Langmead et al., 2009; Milne et al., 2009; Robinson et al., 2011; Thorvaldsdóttir et al., 2013). Heteroplasmic sites were detected using IGV by setting the software to show positions in which more than one nucleotide was sequenced, and that the frequency of the second most frequent base was equal to or higher than 5%. Mitogenome annotation was performed using MitoFish and MITOS (Bernt et al., 2013; Iwasaki et al., 2013).

Results

In total, five transcripts aligned to the reference mitogenomes and were used to assembly the mitogenome of *H. incognitum* with an average coverage depth of 5292× and a total of 895 189 aligned

Correspondence: Thiago Estevam Parente, Laboratório de Toxicologia Ambiental, Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, Brasil. E-mail: parente@ensp.fiocruz.br

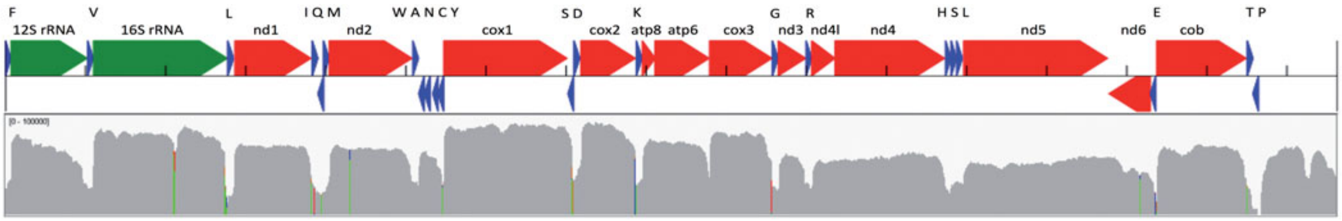

Figure 1. Annotation of the mitochondrial genome of *Hypoptopoma incognitum* (Siluriformes:Loricariidae). Protein-coding genes are colored in red, ribosomal genes in green, and tRNA in blue. The control region is not represented in the upper panel of the figure. The amino acid of each tRNA is shown using the one letter code. The lower panel shows the sequencing depth over the complete mitogenome (log scale). Sharp decreases in read counts represent the punctuation model of mitochondrial transcription. The 33 heteroplasmic sites are colored in the lower panel. Figure was generated by MITOS (upper panel) and by the Integrative Genome Viewer (IGV, lower panel) (Bernt et al., 2013; Thorvaldsdóttir et al., 2013).

reads (Figure 1). The complete mitochondrial genome sequence of *H. incognitum* is 16 630 bp long (GenBank accession no. KT033767), containing the typical vertebrate features: 22 tRNA genes, two rRNA genes, 13 protein-coding genes, and a non-coding control region (D-loop). The majority of genes are encoded on the heavy strand, whereas ND6 and eight tRNAs are found on the light strand. All protein-coding genes used ATG start codon except for COI that used GTG. Seven protein-coding genes are terminated with the complete stop codon, of which four ended with TAA (COI, ATP8, ND4L, and ND5) and three with TAG (ND1, ND2, and ND6). The remaining protein-coding genes are ended by incomplete stop codons, TA (ATP6) or T (COII, COIII, ND3, ND4, and Cytb), which are completed by post-transcriptional polyadenylation. The 12S and the 16S rRNA genes are separated by *tRNA-Val* gene and their lengths are 957 and 1677 bp, respectively. The 22 tRNA genes had sizes ranging from 67 to 75 nucleotides and the control region, located between tRNAPro and tRNAPhe genes, is 985 bp long. The nucleotide composition for the heavy strand was 29.3% A, 23.3%T, 16.3% G, and 31.3% C. The sequencing depth varied greatly along the mitogenome sequence, from as low as of three reads, to as high as 42 195 reads. Cytochrome oxidases I and II were the protein-coding genes with the highest number of reads. Regions with lower sequencing depth tend to code for tRNA. A total of 33 heteroplasmatic sites were detected: five on protein-coding genes, seven on the 16S rRNA, 14 on sequences coding for tRNAs, two in the D-loop region, and five in intergenic sequences (Supplementary material 1).

Discussion

The mitogenome of *H. incognitum* presented here is just the third complete mitochondrial genome of a Loricariidae fish, a family with more than 800 valid species (Lujan et al., 2015). In fact, this is the first mitogenome of loricariid that the fish specimen was sampled in its native habitat and deposited in an ichthyological collection. These are crucial steps when assigning molecular information to Loricariidae species due to their complex taxonomy and still controversial phylogeny. The fish native origin is crucial for its correct identification at the species level and often cryptic species are resolved.

The RNA-based approach to assembly of mtDNA overcomes some limitations of traditional DNA-based strategies for sequencing mitogenomes and to detect heteroplasmy. Specifically, this RNA-based strategy use straightforward method of total cellular RNA extraction that surpass the need to obtain high quantity of high-quality mtDNA; it does not suffer from the difficulties and biases that overlapping long range PCR can represent and introduce; and it is not compromised by Nuclear copies of mitochondrial DNA (Numts). As the number of transcriptomes available in public repositories is growing fast, this approach can be especially advantageous to assembly new mitogenomes from non-model species.

Conclusions

In summary, an innovative RNA-based approach was used to assembly the third complete mitochondrial genome of a species from the speciose family Loricariidae. Moreover, the use of this approach allows the measure of mtRNA expression levels, the punctuation pattern of editing, and the detection of heteroplasmies.

Acknowledgements

The authors are thankful to Dr Jansen Zuanon from INPA for fish identification and to Dr Paulo A. Buckup from the Museu Nacional (MN UFRJ) for fish deposit. Authors acknowledge the use of the computer cluster Titan from the Woods Hole Oceanographic Institution (WHOI), where T. P. holds a guest investigator position.

Declaration of interest

The authors report no conflict of interest. This work was supported by Partnerships for Enhanced Engagement in Research (PEER) grant from USAID (PGA-2000003446) to TP.

References

Aquino AE, Schaefer S. (2010). Systematics of the genus Hypoptopoma Günther, 1868 (Siluriformes, Loricariidae). Bull Am Mus Nat Hist 336:1–110.

Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, Pütz J, et al. (2013). MITOS: Improved de novo metazoan mitochondrial genome annotation. Mol Phylogenet Evol 69:313–19.

Cui Z, Liu Y, Li CP, You F, Chu KH. (2009). The complete mitochondrial genome of the large yellow croaker, *Larimichthys crocea* (Perciformes, Sciaenidae): Unusual features of its control region and the phylogenetic position of the Sciaenidae. Gene 432:33–43.

Dames S, Eilbeck K, Mao R. (2015). A high-throughput next-generation sequencing assay for the mitochondrial genome. Methods Mol Biol 1264:77–88.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494–512.

Hahn C, Bachmann L, Chevreux B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads – A baiting and iterative mapping approach. Nucleic Acids Res 41. doi: 10.1093/nar/gkt371.

Hu M, Jex AR, Campbell BE, Gasser RB. (2007). Long PCR amplification of the entire mitochondrial genome from individual helminths for direct sequencing. Nat Protoc 2:2339–44.

Huang T. (2011). Next generation sequencing to characterize mitochondrial genomic DNA heteroplasmy. Curr Protoc Hum Genet 71:1–12.

Iwasaki W, Fukunaga T, Isagozawa R, Yamada K, Maeda Y, Satoh TP, Sado T, et al. (2013). Mitofish and mitoannotator: A mitochondrial genome database of fish with an accurate and automatic annotation pipeline. Mol Biol Evol 30:2531–40.

Jex AR, Littlewood DTJ, Gasser RB. (2010). Toward next-generation sequencing of mitochondrial genomes – Focus on parasitic worms of animals and biotechnological implications. Biotechnol Adv 28:151–9.

Kocher A, Kamilari M, Lhuillier E, Coissac E, Péneau J, Chave J, Murienne J. (2014). Shotgun assembly of the assassin bug *Brontostoma colossus* mitochondrial genome (Heteroptera, Reduviidae). Gene 552: 184–94.

Langmead B, Trapnell C, Pop M, Salzberg SL. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.

Lujan NK, Armbruster JW, 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 82:269–88.

Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D. (2009). Tablet – Next generation sequence assembly visualization. Bioinformatics 26:401–2.

Payne BAI, Gardner K, Coxhead J, Chinnery PF, . (2015). Deep resequencing of mitochondrial DNA. Methods Mol Biol 1264:59–66.

Quispe-Tintaya W, White RR, Popov VN, Vijg J, Maslov AY. (2015). Rapid mitochondrial DNA isolation method for direct sequencing. Methods Mol Biol 1264:89–95.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. (2011). Integrative genomics viewer. Nat Biotechnol 29:24–6.

Seo S, Zeng X, Assidi M, LaRue B, King J, Sajantila A, Budowle B. (2014). High throughput whole mitochondrial genome sequencing by two platforms of massively parallel sequencing. BMC Genomics 15(Suppl 2):P7.

Sosa MX, Sivakumar IKA, Maragh S, Veeramachaneni V, Hariharan R, Parulekar M, Fredrikson KM, et al. (2012). Next-generation sequencing of human mitochondrial reference genomes uncovers high heteroplasmy frequency. PLoS Comput Biol 8:e1002737.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief Bioinform 14:178–92.

Supplementary material available online

Supplementary Table 1

3.3 CAPÍTULO TRÊS: O genoma mitocondrial completo de *Corydoras nattereri* (Callichthyidae: Corydoradinae)

Neste capítulo aplicamos o método de montagem de genomas mitocondriais em três indivíduos coletados da mesma população. Dessa forma, ampliamos os objetivos da simples descrição da estrutura do mitogenoma de um indivíduo para uma análise da variação genética intra-específica. Os resultados aqui descritos foram publicados no artigo intitulado "The complete mitochondrial genome of *Corydoras nattereri* (Callichthyidae: Corydoradinae)", na revista Neotropical Ichthyology, ano 2016; vol.14(1):e150167.

Corydoras nattereri é uma espécie de bagre encouraçado do sudeste do Brasil. Os mitogenomas produzidos a partir dos transcriptomas hepáticos de três indivíduos resultaram numa sequência de DNA circular de 16.557 nucleotídeos abrangendo 22 genes de tRNA, dois genes de rRNA, 13 genes codificadores de proteínas, a origem de replicação da fita leve (OriL) e a região controle (D-loop). A análise filogenética de sequências proximamente relacionadas da subunidade I do gene citocromo c oxidase (COI) demonstrou a existência de elevada diversidade entre populações morfologicamente similares de *C. nattereri*. *Corydoras nattereri* está inserida num complexo de populações atualmente identificadas como *C. paleatus* e *C. ehrhardti*. A análise da estrutura do mitogenoma demonstra que a inserção de uma sequência de 21 nucleotídeos entre os genes da subunidade 6 da ATPase e do COIII representa um caráter filogeneticamente informativo associado à evolução de Corydoradinae.

Neotropical Ichthyology, 14(1): e150167, 2016
DOI: 10.1590/1982-0224-20150167

Journal homepage: www.scielo.br/ni
Published online: 07 April 2016 (ISSN 1982-0224)

The complete mitochondrial genome of *Corydoras nattereri* (Callichthyidae: Corydoradinae)

Daniel A. Moreira[1], Paulo A. Buckup[2], Marcelo R. Britto[2], Maithê G. P. Magalhães[1], Paula C. C. de Andrade[1], Carolina Furtado[3] and Thiago E. Parente[1]

The complete mitogenome of *Corydoras nattereri*, a species of mailed catfishes from southeastern Brazil, was reconstructed using next-generation sequencing techniques. The mitogenome was assembled using mitochondrial transcripts from the liver transcriptomes of three individuals, and produced a circular DNA sequence of 16,557 nucleotides encoding 22 tRNA genes, two rRNA genes, 13 protein-coding genes and two noncoding control regions (D-loop, OrigL). Phylogeographic analysis of closely related sequences of Cytochrome Oxydase C subunit I (COI) demonstrates high diversity among morphologically similar populations of *C. nattereri*. *Corydoras nattereri* is nested within a complex of populations currently assigned to *C. paleatus* and *C. ehrhardti*. Analysis of mitogenome structure demonstrated that an insertion of 21 nucleotides between the ATPase subunit-6 and COIII genes may represent a phylogenetically informative character associated with the evolution of the Corydoradinae.

O mitogenoma completo de *Corydoras nattereri*, uma espécie de bagres encouraçados do sudeste do Brasil, foi reconstruído através de técnicas de sequencimento de DNA de próxima geração. O mitogenoma foi produzido a partir de produtos de transcrição mitocondrial dos transcriptomas hepáticos de três indivíduos, resultando numa sequência de DNA circular de 16.557 nucleotídeos abrangendo 22 genes de tRNA, dois genes de rRNA, 13 genes codificadores de proteínas e duas regiões de controle não codificadoras (D-loop, OrigL). A análise filogenética de sequências proximamente relacionadas da subunidade I do gene Citocrome Oxidase C (COI) demonstrou a existência de elevada diversidade entre populações morfologicamente similares de *C. nattereri*. *Corydoras nattereri* está inserida num complexo de populações atualmente identificadas como *C. paleatus* e *C. ehrhardti*. A análise da estrutura do mitogenoma demonstra que a inserção de uma sequência de 21 nucleotídeos entre os genes da subunidade 6 da ATPase e do COIII representa um caráter filogeneticamente informativo associado à evolução de Corydoradinae.

Keywords: Barcode, DNA, Mitogenome, Molecular diversity, mtRNA.

Introduction

Corydoradinae are a species-rich group of armored freshwater catfishes that inhabit streams, rivers and floodplains throughout South America (Alexandrou *et al.*, 2011). Together with the Callichthyinae, they comprise the Callichthyidae, a family of catfishes diagnosed by the presence of two series of bony plates on the sides of the body and one pair of barbels at lips junction (Reis, 1998). The Corydoradinae comprises 227 nominal taxa and 188 valid species (Eschmeyer & Fong, 2015), assigned to the *Aspidoras* Ihering, 1907, *Corydoras* Lacépède, 1803, and *Scleromystax* Gunther, 1864 (Britto, 2003). *Corydoras* is the most species-rich genus of catfishes with over 160 described and nearly as many undescribed species (Alexandrou *et al.*, 2011; Eschmeyer & Fong, 2015;). According to Reis (2003) about two new species of *Corydoras* are described each year. While the Callichthyinae is relatively well-known based on morphological (Reis, 1997, 1998, 2003) and molecular studies (Mariguela *et al.*, 2013), the Corydoradinae remains poorly known, despite their great interest to the aquarium hobby. Phylogenetic studies that included species of *Corydoras* have been performed, primarily by Britto (2003) based on morphological characters, and more recently by Alexandrou *et al.* (2011), based on molecular data. In the later study, however, a large proportion of the 52 taxa recognized could not be associated to a valid name, with many species being referenced to informal “C-Numbers” (Fuller & Evers, 2005) available from the aquarium industry or to their geographic origin.

[1] Laboratório de Toxicologia Ambiental, Fundação Oswaldo Cruz (FIOCRUZ), Avenida Brasil, 4036 sala 101, 2104-031 Rio de Janeiro, RJ, Brazil. (DAM) daniel.moreira@ioc.fiocruz.br, (MGPM) magalhaes.maithe@gmail.com, (PCCA) paula.andrade.bio@gmail.com, (TEP) parente@ensp.fiocruz.br

[2] Departamento de Vertebrados, Museu Nacional, Universidade Federal do Rio de Janeiro (UFRJ), Quinta da Boa Vista, 20940-040 Rio de Janeiro, RJ, Brazil. (PAB) buckup@acd.ufrj.br (corresponding author), (MRB) mrbritto2002@yahoo.com.br

[3] Divisão de Genética, Instituto Nacional do Câncer (INCA), André Cavalcanti, 37, 4º andar, 20231-050 Rio de Janeiro, RJ, Brazil. (CF) cfurtado@inca.gov.br

Corydoras nattereri Steindachner, 1876, is a widespread species of *Corydoras* in southeastern Brazil, ranging from rio Mucuri, in Bahia, to the Paranaguá Bay, in Paraná (Britto, 2007; Shimabukuro-Dias *et al*., 2004a). *Corydoras nattereri* and *Scleromystax prionotos* (Nijssen & Isbrücker, 1980) form a pair of color mimics where their distribution overlaps (Alexandrou *et al*., 2011). The geographic range of *C. nattereri* represents a distributional range of about 1,350 km, encompassing numerous isolated coastal river drainages. With such widespread distribution it is not surprising that significant variability has been found among various populations of *C. nattereri*. Oliveira *et al.* (1990) identified three different cytotypes among these populations, that differ in number of chromosomes (2n numbers of 40, 43 and 44), suggesting that more than one species is represented by the taxon. That cytogenetic variation among populations was later correlated with variation in DNA sequence data (Simabukuro-Dias *et al*., 2004b).

Currently partial sequences of the Cytochrome Oxydase C subunit I (COI) of *Corydoras nattereri*, a widely used DNA barcode marker, are available publicly from only two localities (Pereira *et al*., 2011, 2013). Sequences of additional mitochondrial genes have been made available by Alexandrou *et al*. (2011), but only specimens with imprecise locality data have been listed in that study. Within the Corydoradinae, complete mitochondrial sequence data is only available for *C. rabauti* from a specimen without associated locality data (Saitoh *et al*., 2003). Herein, we present the complete mitochondrial genome for *C. nattereri* based on three specimens with precise geographic provenance, thus providing a significant increment in DNA data available for phylogenetic studies of the Corydoradinae.

Material and Methods

Specimens of *Corydoras nattereri* were collected in the rio Suruí (22.600556 S, -43.091667 W) at the Santo Aleixo district, Magé, Rio de Janeiro, Brazil. The fish were deposited at the Museu Nacional, Rio de Janeiro, UFRJ (MNRJ 41520) and tissues from tree individuals (MNTI 8664-8666) were preserved in ethanol and RNALater. Total RNA was extracted from the liver tissue following conventional phenol-chloroform extraction. RNA quality was accessed using the RNA Nano kit for Bioanalyzer (Agilent). Three individual cDNA libraries were constructed using the TruSeq RNA Sample kit v.2 (Illumina). Libraries were accessed for quality (Bioanalyzer, DNA1000 kit, Agilent) and quantity (Kapa Biosystems). Two separated runs (single-end and paired-end) were performed in an Illumina HiSeq 2500 using the TrueSeq SBS kit v.3 (Illumina). Raw Illumina data were demultiplexed using the BCL2FASTQ software (Illumina). Reads were trimmed for Illumina adaptors by Trimmomatic (Bolger *et al*., 2014) and its quality was evaluated using FastQC (Babraham Bioinformatics). Only reads with Phred score over 30 were used for the transcriptome assembly. Cleaned reads from the three individual fish were used for the *de novo* assembly of transcriptomes using the default parameters of Trinity (v. 2.0.2) (Haas *et al*., 2013). Mitochondrial genomes were assembled using the mitochondrial transcripts from the liver transcriptome, following the approach described by Moreira *et al*. (2015a, 2015b). Briefly, mitochondrial transcripts were retrieved running a BLASTN search against the mitogenome of the closest related species with a complete mitogenome available, *Corydoras rabauti* (GI: 29501080) (Saitoh *et al*., 2003). 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 *C. rabauti* (Gouy *et al*., 2010). The sequence of each CONTIG was manually checked for inconsistencies and gaps. Small gaps at mitogenomes, which code for transfer RNAs, were completed with Sanger sequencing data from PCR with specific designed primers. As the identity of the three assembled mitogenomes was higher than 99.8%, the three sequences were concatenated in one consensus sequence. The consensus mitogenome was annotated using the web-based services MitoFish and MITOS (Iwasaki *et al*., 2013; Bernt *et al*., 2013), and the origin of the L-strand replication was identified based on Wong *et al.* (1983). In order to determine sequencing depth 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 Integrated Genome Viewer (IGV), Tablet (Langmead *et al*., 2009; Milne *et al*., 2010; Robinson *et al*., 2011; Thorvaldsdóttir *et al*., 2012), and Geneious version 6 (http://www.geneious.com, Kearse *et al*., 2012).

A Maximum Likelihood (ML) analysis of relationships was performed to position the mitogenomes within a taxonomic and phylogenetic context. All publicly available sequences of COI that exhibited nucleotide identity greater than 90% in a BLAST search of the GenBank Nucleotide Database and Bold Systems were included in the analysis, as well as additional sequences of *Corydoras nattereri*, *Callichthys callichthys*, *Scleromystax barbatus*, and *Aspidoras lakoi* (the latter three used as outgroups) produced in the Museu Nacional (MNLM) laboratory (Table 1) using Sanger sequencing methods, and the corresponding COI sequence extracted from the *C. rabauti* mitochondrial genome (GenBank Accession AB054128). To ensure uniformity of coverage, only nucleotides from positions 58 to 699, and only sequences with full coverage of that segment were included in the analysis, producing a matrix of nucleotide sequences from 35 fish. The ML analysis was performed using Mega 6.06 (Tamura *et al*., 2013) under the Hasegawa-Kishino-Yano model (Hasegawa *et al*., 1985), selected by the corrected Akaike information criterion using jModeltest v.2.1.7 (Darriba *et al*., 2012). Initial tree(s) for the heuristic search were obtained by Neighbor-Join and BioNJ algorithms applied to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach. The discrete Gamma distribution was used to model evolutionary differences among sites, with 5 categories (+G, parameter = 0.1563). Branch support was estimated with the Bootstrap method, using 350 replications. The tree was rooted using *Callichthys callichthys* as outgroup.

Table 1. List of samples used in this study, ordered according to GenBank Accession codes. Voucher museum catalog number is provided only for specimens that are being made available for the first time.

GenBank Accession	BOLD ProcessID	Sample	Voucher	Species	Publication/Source	Latitude	Longitude	Locality
GU701815	FUPR517-09	LBP-32757		*Corydoras ehrhardti*	Pereira *et al.*, 2013	-23,9383	-50,7290	Upper Parana basin, Brazil
GU701816	FUPR516-09	LBP-32756		*Corydoras ehrhardti*	Pereira *et al.*, 2013	-23,9383	-50,7290	Upper Parana basin, Brazil
GU701817	FUPR818-09	LBP-36128		*Corydoras ehrhardti*	Pereira *et al.*, 2013	-23,9383	-50,7290	Upper Parana basin, Brazil
GU701818	FUPR817-09	LBP-36126		*Corydoras ehrhardti*	Pereira *et al.*, 2013	-23,9383	-50,7290	Upper Parana basin, Brazil
GU701819	FUPR816-09	LBP-36124		*Corydoras ehrhardti*	Pereira *et al.*, 2013	-23,9383	-50,7290	Upper Parana basin, Brazil
GU701809	FUPR819-09	LBP-36114		*Corydoras paleatus*	Pereira *et al.*, 2013	-	-	Upper Parana basin, Brazil
GU701810	FUPR822-09	LBP-36119		*Corydoras paleatus*	Pereira *et al.*, 2013	-	-	Upper Parana basin, Brazil
GU701812	FUPR824-09	LBP-36122		*Corydoras paleatus*	Pereira *et al.*, 2013	-	-	Upper Parana basin, Brazil
GU701813	FUPR823-09	LBP-36121		*Corydoras paleatus*	Pereira *et al.*, 2013	-	-	Upper Parana basin, Brazil
GU701814	FUPR515-09	LBP-32755		*Corydoras ehrhardti*	Pereira *et al.*, 2013	-23,9383	-50,7290	Upper Parana basin, Brazil
GU701871	FUPR820-09	LBP-36116		*Corydoras paleatus*	Pereira *et al.*, 2013	-	-	Upper Parana basin, Brazil
GU702213	FPSR082-09	LBP-29094		*Corydoras nattereri*	Pereira *et al.*, 2011	-23,3690	-46,0240	[Rio Paraíba do Sul upstream from Jacareí], SP, Brazil
JN988818	FUPR285-09	LBP-32330		*Corydoras nattereri*	Pereira *et al.*, 2013	-23,5112	-45,8591	[rio Paraitinguinha, Tietê drainage], upper Paraná Basin, SP, Brazil
JN988819	FUPR286-09	LBP-32331		*Corydoras nattereri*	Pereira *et al.*, 2013	-23,5112	-45,8591	[rio Paraitinguinha, Tietê drainage], upper Paraná Basin, SP, Brazil
JN988821	FUPR288-09	LBP-32333		*Corydoras nattereri*	Pereira *et al.*, 2013	-23,5112	-45,8591	[rio Paraitinguinha, Tietê drainage], upper Paraná Basin, SP, Brazil
JN988822	FUPR289-09	LBP-32334		*Corydoras nattereri*	Pereira *et al.*, 2013	-23,5112	-45,8591	[rio Paraitinguinha, Tietê drainage], upper Paraná Basin, SP, Brazil
JX111730	FARGB202-11	UNMDP-T 0370		*Corydoras paleatus*	Rosso *et al.*, 2012	-36,2680	-59,9800	Arroio Tapalque, Buenos Aires, Argentina
JX111731	FARGB238-11	UNMDP-T 0406		*Corydoras paleatus*	Rosso *et al.*, 2012	-38,3350	-61,6014	El Divisorio Stream, Buenos Aires, Argentina
JX111732	FARGB358-11	UNMDP-T 0526		*Corydoras paleatus*	Rosso *et al.*, 2012	-34,9700	-61,0433	Adeg Puente Santa Cruz (Ascencion), Buenos Aires, Argentina
JX111733	FARGB239-11	UNMDP-T 0407		*Corydoras paleatus*	Rosso *et al.*, 2012	-38,3350	-61,6014	El Divisorio Stream, Buenos Aires, Argentina
JX111734	FARGB201-11	UNMDP-T 0369		*Corydoras paleatus*	Rosso *et al.*, 2012	-36,2680	-59,9800	Arroyo Tapalque, Buenos Aires, Argentina
JX111735	FARGB200-11	UNMDP-T 0368		*Corydoras paleatus*	Rosso *et al.*, 2012	-36,2680	-59,9800	Arroyo Tapalque, Buenos Aires, Argentina
JX111736	FARGB184-11	UNMDP-T 0352		*Corydoras paleatus*	Rosso *et al.*, 2012	-35,6056	-59,9914	Chascomus lagoon, Buenos Aires, Argentina
JX111737	FARGB183-11	UNMDP-T 0351		*Corydoras paleatus*	Rosso *et al.*, 2012	-35,6056	-59,9914	Chascomus lagoon, Buenos Aires, Argentina
JX111738	FARGB172-11	UNMDP-T 0340		*Corydoras paleatus*	Rosso *et al.*, 2012	-35,5411	-58,0894	Vitel lagoon, Buenos Aires, Argentina

(**Table 1:** Cont.)

GenBank Accession	BOLD ProcessID	Sample	Voucher	Species	Publication/Source	Latitude	Longitude	Locality
KT874579	MNRJ531-15	MNTI9093/ MNLM5113	MNRJ41901	*Callichthys callichthys*	This study	-19,9753	-40,5478	rio Valssungana Velha, Reis Magos drainage, near bridge of Santa Teresa - Santa Leopoldina road, Santa Teresa, ES, Brazil
KT874580	MNRJ533-15	MNTI554/ MNL1354	MNRJ31639	*Aspidoras lakoi*	This study	-20,7500	-46,2122	Left bank tributary of ribeirão do Turvo, between Fazenda da Serra de Cima and Fazenda Baixadão, Capitólio, MG, Brazil
KT874581	MNRJ538-15	MNTI8665/ MNLM5103	MNRJ41520	*Corydoras nattereri*	This study	-22,6006	-43,0917	Tributary of rio Suruí and rio Suruí downstream of Piabeta - Santo Aleixo road, Magé, RJ, Brazil
KT874582	MNRJ532-15	MNTI7422/ MNLM4992	MNRJ40948	*Scleromystax barbatus*	This study	-26,1972	-49,9219	Rio Lindo (right margin tributary of rio Cubatão), SC-301, bairro Dona Francisca, Joinville, SC, Brazil
KT874583	MNRJ534-15	MNTI3387/ MNLM935	MNRJ37470	*Corydoras nattereri*	This study	-22,4675	-42,2975	Córrego Aldeia Velha, under bridge on road Aldeia Velha, near RPPN Fazenda do Bom Retiro, Casimiro de Abreu, RJ, Brazil
KT874584	MNRJ539-15	MNTI8666/ MNLM5104	MNRJ41520	*Corydoras nattereri*	This study	-22,6006	-43,0917	Tributary of rio Suruí and rio Suruí downstream of Piabeta - Santo Aleixo road, Magé, RJ, Brazil
KT874585	MNRJ537-15	MNTI8664/ MNLM5102	MNRJ41520	*Corydoras nattereri*	This study	-22,6006	-43,0917	Tributary of rio Suruí and rio Suruí downstream of Piabeta - Santo Aleixo road, Magé, RJ, Brazil
KT874586	MNRJ536-15	MNTI7464/ MNLM5000	MNRJ41030	*Corydoras ehrhardti*	This study	-26,2803	-49,3244	rio Vermelho, Itapocu drainage, downstream of dam, upstream of aqueduct, São Bento do Sul, SC, Brazil
KT874587	MNRJ535-15	MNTI3832	MNRJ37693	*Corydoras nattereri*	This study	-18,3106	-40,2411	rio do Sul under bridge of ES-130 between Vinhático and Pinheiros, on municipal border between Montanha and Pinheiros, ES, Brazil

Results

Sequencing depth. The complete mitochondrial genome sequence of *C. nattereri* is 16,557 bp long (GenBank Accessions No. KT239008, KT239009, KT239010, Fig. 1). A total of 152,877,464 100bp reads were used to assemble the transcriptomes. On average, 12 transcripts were aligned to the reference mitogenome. These aligned transcripts were used to assembly the mitogenome of *Corydoras nattereri*, which was sequenced with an average coverage depth of 8,194 and a total of 1,378,370 aligned reads (Fig. 2). The sequencing depth varied greatly along the mitogenome sequence, from as low as of 1 read, to as high as 69,778 reads. Cytochrome oxidase subunits I, II and III were the protein-coding genes with the highest number of reads. Regions with lower sequencing depth tend to code for tRNA. Five small gaps, varying from 21 to 183 nucleotides, were filled by conventional PCR and Sanger sequencing (Table 2).

Table 2. Positions and lengths of the five gaps in the mitogenome assembled using mitochondrial transcripts sequenced using Illumina HiSeq2500. These gaps were filled using conventional PCR and Sanger sequencing with the species-specific primers listed.

Gap position	Length (nucleotides)	Primers	
		Forward	Reverse
1 - 81	81	CAGATTAGGCTCGACCGACG	GGCGTATGACGGCTTGGTAA
995 - 1085	91	CCCAAAACGTCAGGTCGAGG	TTTGCCACAGAGACGGGTTG
7836 - 7910	75	AAATCTGCGGTGCAAACCAC	GGCGGTTATTTTGTCAGCGG
9558 - 9579	22	GAGCCCACCACAGCATCATA	GCAGTCGTGCAGATCCTAGT
11720 - 11903	184	CCCCCGCTACCCAACTTAAT	TGCTTTCTGTTCCCTGGTCT

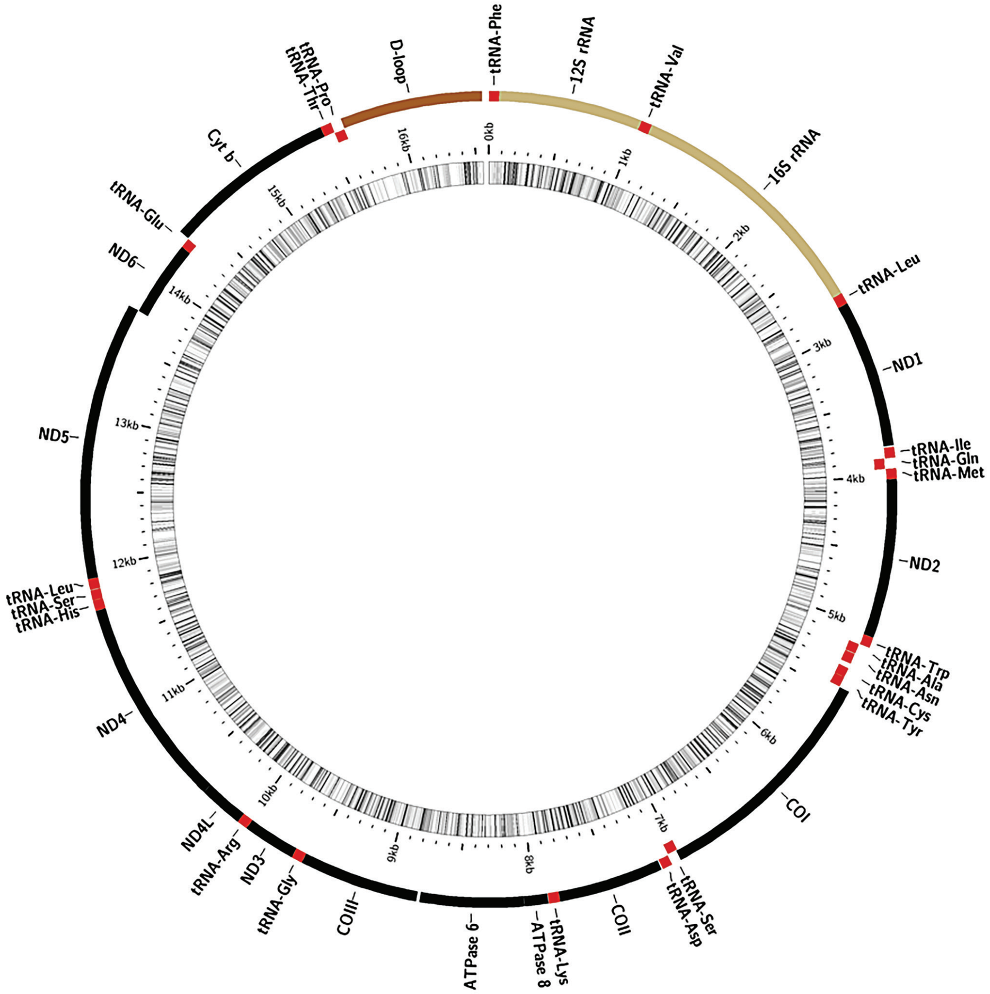

Fig. 1. Circular representation of the mitochondrial genome of *Corydoras nattereri*. Genes encoded in the heavy strand are shown in the outer circle and genes encoded in the light strand are offset inwards. The inner circle represents the CG-content. Figure was generated by the online server MitoFish, http://mitofish.aori.u-tokyo.ac.jp (Iwasaki *et al*., 2012).

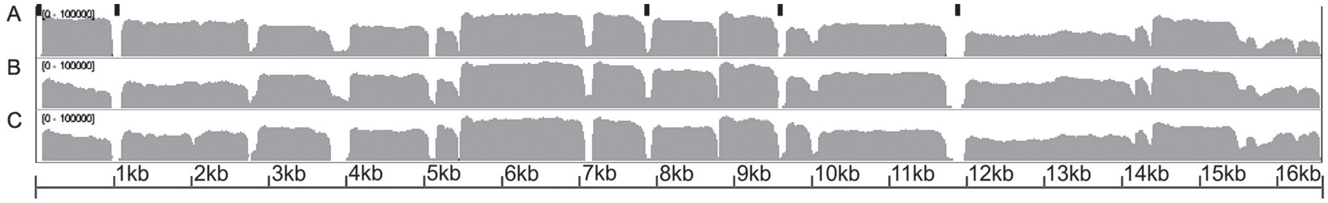

Fig. 2. Sequencing depth over the complete mitogenomes of the three individuals of *Corydoras nattereri*: KT239008 (A), KT239009 (B), and KT239010 (C). Read counts (y-axis) are shown in logarithmic scale and sharp decreases correspond to the punctuation model of mitochondrial transcription (positions correspond to those shown in Fig. 1 and Table 3). Black vertical bars indicate position of gaps that were filled with Sanger sequencing (Table 2). Reads were mapped to the mitogenomes using Bowtie and visualized at the Integrative Genome Viewer (IGV, Bernt *et al*., 2013; Thorvaldsdóttir *et al*., 2013).

Table 3. Positioning of genes in the mitochondrial genome of *Corydoras nattereri*. Negative gap values indicate overlap.

Gene	Start	End	Length	Gap	Direction	Start Codon	Stop Codon
tRNA-Phe	1	68	68	0			
12S rRNA	69	1013	945	0			
tRNA-Val	1014	1085	72	0			
16S rRNA	1086	2752	1667	0			
tRNA-Leu	2753	2827	75	0			
ND1	2828	3799	972	8		ATG	TAG
tRNA-Ile	3808	3879	72	-2			
tRNA-Gln	3878	3948	71	-1	complement		
tRNA-Met	3948	4017	70	0			
ND2	4018	5062	1045	0		ATG	T--
tRNA-Trp	5063	5133	71	1			
tRNA-Ala	5135	5203	69	1	complement		
tRNA-Asn	5205	5277	73	0	complement		
origin-L	5278	5312	35	-3	complement		
tRNA-Cys	5310	5377	68	-1	complement		
tRNA-Tyr	5377	5446	70	1	complement		
COI	5448	7007	1560	-13		GTG	AGG
tRNA-Ser	6995	7065	71	4	complement		
tRNA-Asp	7070	7138	69	6			
COII	7145	7835	691	0		ATG	T--
tRNA-Lys	7836	7909	74	1			
ATPase 8	7911	8078	168	-10		ATG	TAA
ATPase 6	8069	8752	684	21		ATG	TAA
COIII	8774	9557	784	0		ATG	T--
tRNA-Gly	9558	9629	72	0			
ND3	9630	9978	349	0		ATG	T--
tRNA-Arg	9979	10048	70	0			
ND4L	10049	10345	297	-7		ATG	TAA
ND4	10339	11719	1381	0		ATG	T--
tRNA-His	11720	11789	70	0			
tRNA-Ser	11790	11856	67	1			
tRNA-Leu	11858	11930	73	0			
ND5	11931	13757	1827	-4		ATG	TAA
ND6	13754	14269	516	0	complement	ATG	TAA
tRNA-Glu	14270	14337	68	3	complement		
Cyt b	14341	15478	1138	0		ATG	T--
tRNA-Thr	15479	15550	72	-2			
tRNA-Pro	15549	15618	70	0	complement		
D-loop	15619	16557	939				

Genome organization. The complete mitochondrial genome sequence of *Corydoras nattereri* contains the typical vertebrate features: 22 tRNA genes, 2 rRNA genes, 13 protein-coding genes and two noncoding control regions (D-loop, OrigL) (Table 3, Fig. 2). The majority of genes are encoded on the heavy strand, whereas ND6 and eight tRNAs are found on the light strand. All protein-coding genes used ATG start codons, except for COI that used GTG. Seven protein-coding genes are terminated with the complete stop codon, of which five ended with TAA (ATP8, ATP6, ND4L, ND5 and ND6), one with TAG (ND1) and one with AGG (COI). The remaining protein-coding genes are ended by incomplete stop codons, T (COII, COIII, ND2, ND3, ND4 and Cytb), which are completed by post-transcriptional polyadenylation. The 12S and the 16S rRNA genes are separated by tRNA-Val gene and their lengths are 945 and 1,667 bp, respectively. The 22 tRNA genes had sizes ranging from 67 to 75 nucleotides and the control region, located between tRNAPro and tRNAPhe genes, is 939 bp long. A 21-nucleotide insertion sequence between the ATPase subunit-6 and COIII genes was found in *C. nattereri*. The nucleotide composition for the heavy strand was 32.3% A, 25.7%T, 15.1% G, and 27.0% C.

Phylogenetic context. The phylogeographic analysis of publically available COI together with additional COI sequences generated by our research group confirmed our identification of the samples of the rio Suruí as *Corydoras nattereri* (Fig. 3). Our three samples form a monophyletic group with specimens of *C. nattereri* from the Paraíba do Sul and the Paraitinguinha rivers (upper Tietê drainage, upper Paraná basin) (Pereira *et al.*, 2011, 2013). Contrasting with the high similarity of the samples from these three basins, our samples from the rio Aldeia Velha (rio São João coastal basin), and rio Itaúnas are considerably different. The latter are morphologically similar to *C. nattereri*, but their sequences are 2.5%-3.6% divergent in relation to the rio Surui samples. Such high divergence suggests that the populations of *Corydoras* from the São João and Itaúnas river basins represent cryptic species.

Our phylogenetic analysis also demonstrates that the *C. nattereri* clade is nested within a large clade of samples identified in the literature (Pereira *et al.*, 2011, 2013; Rosso *et al.*, 2012) as *C. paleatus* and *C. ehrhardti*. Samples of *C. ehrhardti* (including new sequences produced here) form a monophyletic clade also included among this larger clade, but samples of *Corydoras paleatus* form a complex of non-monophyletic populations. Within this large complex, samples of *C. paleatus* GU701809, GU701810, GU701812, GU701813, and GU701871, from the upper rio Paraná basin, form the monophyletic subunit most closely related to *C. nattereri*, but the bootstrap value for this sister group relationship is low, indicating that further study of *C. paleatus* species complex is still necessary.

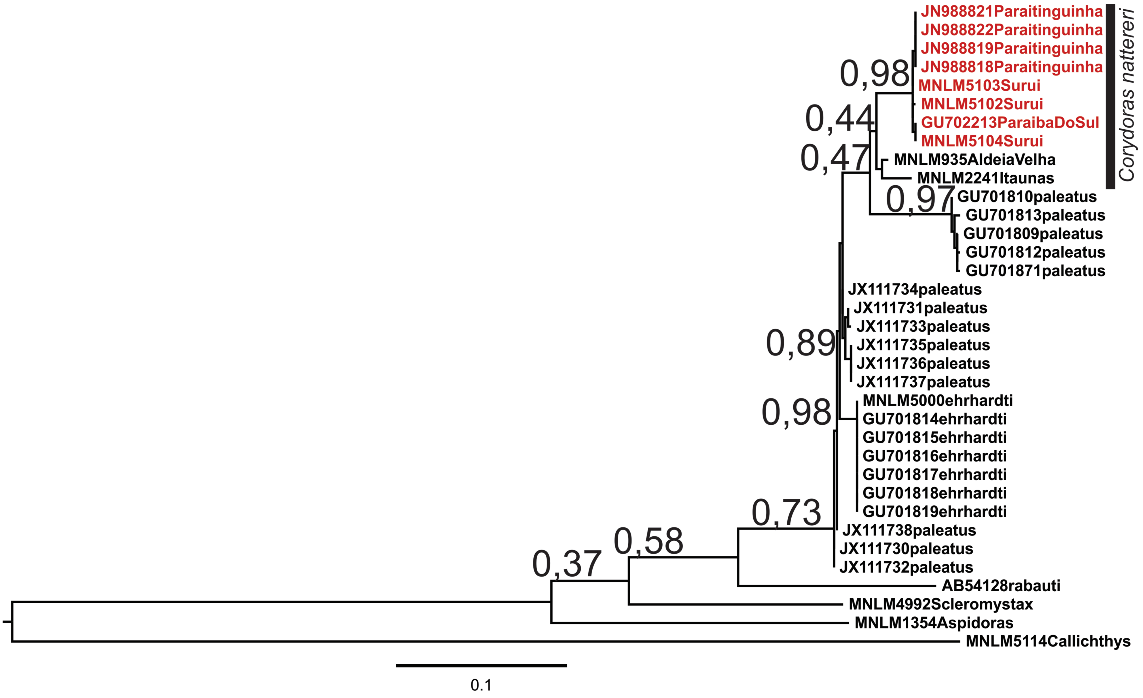

Fig. 3. Maximum likelihood tree (log likelihood = -2410.0522) of *Corydoras* samples with at least 90% similarity to the mitochondrial cytochrome oxidase I sequences of *C. nattereri* from the rio Suruí. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Bootstrap robustness is indicated next to selected branches. Samples of *C. nattereri* have the locality name appended to the sample ID (those for which mitogenomes were produced are from de rio "Surui"), outgroups have the genus name and other samples of *Corydoras* have the species epithet name appended to the ID code (Table 1).

Discussion

Despite the great diversity of *Corydoras,* this is the first mitogenome with voucher specimens sampled in their native habitat and deposited in a permanent collection. A mitogenome of *C. rabauti* has been reported by Saitoh *et al.* (2003), but that study was based on a specimen without locality data obtained from the aquarium trade, and the study does not mention a registration number of a voucher specimen in any biological collection.

Our analysis of COI sequences revealed high levels of genetic divergence and taxonomic complexity among samples closely related to *C. nattereri*. Specimens from the São João and Itaúnas river basins may represent cryptic species, that are currently undistinguishable from topotype specimens of *C. nattereri*. Their level of divergency (2.5% - 3.6%) far exceeds the maximum intraspecific divergence (1.6%) reported among six species of *Corydoras* from the upper Paraná basin (Pereira *et al*., 2013). The analysis also demonstrated the complexity of relationships among populations of *C. nattereri*, *C. paleatus*, and *C. ehrhardti*. Within this context, our newly produced mitogenome *C. nattereri* is likely to provide a solid base to identify additional mitochondrial markers to be used in future studies designed to clarify the relationships among these and other callichthyid taxa.

Comparison of the mitogenome of *Corydoras nattereri* with that of *C. rabauti* (Saitoh *et al*., 2003) reveals features that are likely to be phylogenetically informative and useful in future studies. A 21-nucleotide insertion sequence between the ATPase subunit-6 and COIII genes was found in *C. nattereri*. This insertion corresponds to a 17-nucleotide insertion previously detected in *C. rabauti* (Saitoh *et al*., 2003). Most vertebrate mitochondrial genomes have a head-to-tail junction between the ATPase subunit-6 and COIII genes, and this insertion was considered phylogenetically uninformative in the study of Saitoh *et al*. (2003). Our discovery of an insertion in the homologous position of *C. nattereri* is interpreted as an apomorphic trait shared by the two species. Further investigation about the distribution and length of this insertion among Corydoradinae is likely to yield significant insights about the phylogeny of the group.

Acknowledgements

Authors acknowledge the use of the computer cluster Titan from the Woods Hole Oceanographic Institution (WHOI), where TP holds a guest investigator position. Carla Christie Dijban Quijada, Emanuel Neuhaus, Decio Ferreira de Moraes Junior (UFRJ/MN), and José R. Gomes participated in the field work that yielded the specimens from the rio Surui. William Bryan Jennings (UFRJ/MN) and Gustavo Ferraro (FAPERJ fellowship) provided essential support for laboratory activities at the Laboratório de Pesquisa em Biodiversidade Molecular (UFRJ/MN). This work was supported by Partnerships for Enhanced Engagement in Research (PEER) grant from U.S. Agency for International Development (PGA-2000003446) to TP, and grants from CNPq (307610/2013-6, 564940/2010-0, 476822/2012-2) and FAPERJ (E-26/111.404/2012, E-26/200.697/2014) to PAB.

References

Alexandrou, M. A., C. Oliveira, M. Maillard, R. A. R. McGill, J. Newton, S. Creer & M. I. Taylor. 2011. Competition and phylogeny determine community structure in Müllerian co-mimics. Nature, 469: 84-88.

Bernt, M., A. Donath, F. Jühling, F. Externbrink, C. Florentz, G. Fritzsch, J. Pütz, M. Middendorf & P. F. Stadler. 2013. MITOS: improved *de novo* metazoan mitochondrial genome annotation. Molecular Phylogenetics and Evolution, 69: 313-319.

Bolger, A. M., M. Lohse & B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30: 2114-2120.

Britto, M. R. 2003. Phylogeny of the subfamily Corydoradinae Hoedeman, 1952 (Siluriformes: Callichthyidae), with a definition of its genera. Proceedings of the Academy of Natural Sciences of Philadelphia, 153: 119-154.

Britto, M. R. 2007. Família Callichthyidae. Pp.75-81. In: Buckup, P. A., N. A. Menezes & M. S. Ghazzi (Eds.). Catálogo das espécies de peixes de água doce do Brasil. Rio de Janeiro, Museu Nacional. (Série Livros, 23).

Darriba, D., G. L. Taboada, R. Doallo & D. Posada. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods, 9: 772.

Eschmeyer, W. N. & J. D. Fong. 2015. Catalog of fishes: genera, species, references. Electronic Version. San Francisco, CA, California Academy of Sciences. Available from: http://researcharchive.calacademy.org/research/ichthyology/catalog/fishcatmain.asp (6 August 2015).

Fuller, I. A. M. & H.-G. Evers. 2005. Identifying Corydoradinae catfish. *Aspidoras - Brochis - Corydoras - Scleromystax* & C-numbers. 1st ed. Worcestershire, Ian Fuller Enterprises; Rodgau, A.C.S. GmbH (Aqualog), 384p.

Gouy, M., S. Guindon & O. Gascuel. 2010. SeaView Version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular Biology and Evolution, 27: 221-224.

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. LeDuc, N. Friedman & A. Regev. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols, 8: 1494-1512.

Hasegawa, M., H. Kishino & T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution, 22: 160-174.

Iwasaki, W., T. Fukunaga, R. Isagozawa, K. Yamada, Y. Maeda, T. P. Satoh, T. Sado, K. Mabuchi, H. Takeshima, M. Miya & M. Nishida. 2013. Mitofish and mitoannotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. Molecular Biology and Evolution, 30: 2531-2540.

Kearse, M., R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes & A. Drummond. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics, 28: 1647-1649.

Langmead, B., C. Trapnell, M. Pop & S. L. Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10: R25.

Mariguela, T. C., M. A. Alexandrou, F. Foresti & C. Oliveira. 2013. Historical biogeography and cryptic diversity in the Callichthyinae (Siluriformes, Callichthyidae). Journal of Zoological Systematics and Evolutionary Research, 51: 308-315.

Milne, I., M. Bayer, L. Cardle, P. Shaw, G. Stephen, F. Wright & D. Marshall. 2010. Tablet-next generation sequence assembly visualization. Bioinformatics, 26: 401-402.

Moreira, D. A., C. Furtado & T. E. Parente. 2015a. The use of transcriptomic next-generation sequencing data to assemble mitochondrial genomes of *Ancistrus* spp. (Loricariidae). Gene, 573: 171-175.

Moreira, D. A., M. G. P. Magalhães, P. C. C. Andrade, C. Furtado, A. L. Val & T. E. Parente. 2015b. An RNA-based approach to sequence the mitogenome of *Hypoptopoma incognitum* (Siluriformes: Loricariidae). Mitochondrial DNA: 1-3.

Nijssen, H. & I. J. H. Isbrücker. 1980. On the identity of *Corydoras nattereri* Steindachner, 1877 with the description of a new species, *Corydoras prionotos* (Pisces, Siluriformes, Callichthyidae). Beaufortia, 30: 1-9.

Oliveira, C., L. F. Almeida Toledo & S. A. Toledo Filho. 1990. Comparative cytogenetic analysis in three cytotypes of *Corydoras nattereri* (Pisces, Siluriformes, Callichthyidae). Cytologia, 55: 21-26.

Pereira, L. H. G., G. M. G. Maia, R. Hanner, F. Foresti & C. Oliveira. 2011. DNA barcodes discriminate freshwater fishes from the Paraíba do Sul River Basin, São Paulo, Brazil. Mitochondrial DNA, 22(S1): 71-79.

Pereira, L. H. G., R. Hanner, F. Foresti & C. Oliveira. 2013. Can DNA barcoding accurately discriminate megadiverse Neotropical freshwater fish fauna? BMC Genetics, 14: 20 (1-14).

Reis, R. E. 1997. Revision of the Neotropical catfish genus *Hoplosternum* (Ostariophysi: Siluriformes: Callichthyidae), with the description of two new genera and three new species. Ichthyological Exploration of Freshwaters, 7: 299-326.

Reis, R. E. 1998. Anatomy and phylogenetic analysis of the Neotropical callichthyid catfishes (Ostariophysi, Siluriformes). Zoological Journal of the Linnean Society, 124: 105-168.

Reis, R. E. 2003. Family Callichthyidae (Armored catfishes). Pp. 291-309. In: Reis, R. E., S. O. Kullander & C. J. Ferraris, Jr. (Orgs.). Check list of the freshwater fishes of South and Central America. Porto Alegre, Edipucrs.

Robinson, J. T., H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz & J. P. Mesirov. 2011. Integrative genomics viewer. Nature Biotechnology, 29: 24-26.

Rosso, J. J., E. Mabragaña, M. González Castro & J. M. Díaz de Astarloa. 2012. DNA barcoding Neotropical fishes: recent advances from the Pampa Plain, Argentina. Molecular Ecology Resources, 12: 999-1011.

Saitoh, K., M. Miya, J. G. Inoue, N. B. Ishiguro & M. Nishida, M. 2003. Mitochondrial genomics of ostariophysan fishes: perspectives on phylogeny and biogeography. Journal of Molecular Evolution, 56: 464-472.

Shimabukuro-Dias, C. K., C. Oliveira & F. Foresti. 2004a. Karyotype variability in eleven species of the catfish genus *Corydoras* (Siluriformes: Callichthyidae). Ichthyological Exploration of Freshwaters, 15: 135-146.

Shimabukuro-Dias, C. K., C. Oliveira, R. E. Reis & F. Foresti. 2004b. Molecular phylogeny of the armored catfish family Callichthyidae (Ostariophysi, Siluriformes). Molecular Phylogenetics and Evolution, 32: 152-163.

Tamura, K., G. Stecher, D. Peterson, A. Filipski & S. Kumar. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular Biology and Evolution, 30: 2725-2729.

Thorvaldsdóttir, H., J. T. Robinson & J. P. Mesirov. 2012. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics, 14: 178-192.

Wong, J. F. H., D. P. Ma, R. K. Wilson & B. A. Roe. 1983. DNA sequence of the *Xenopus laevis* mitochondrial heavy and light strand replication origins and flanking tRNA genes. Nucleic Acids Research, 11: 4977-4995.

Submitted October 20, 2015
Accepted January 12, 2016 by Guillermo Ortí

3.4 CAPÍTULO QUATRO: Reduzindo a lacuna de informação sobre a genômica mitocondrial de Loricarioidei (Siluriformes)

Neste capítulo, ampliamos o uso da metodologia desenvolvida para montar genomas mitocondriais de espécies representantes da subordem Loricarioidei, diminuindo a disparidade na quantidade de informação genética disponível para Loricariidae. Usamos essa informação com os objetivos de testar hipóteses filogenéticas e estudar a evolução estrutural desses genomas. Os resultados aqui descritos foram publicados no artigo intitulado “Reducing the information gap on Loricarioidei (Siluriformes) mitochondrial genomics”, na revista BMC Genomics, ano 2017; vol.18(1):345-57.

Conduzimos um estudo abrangente e comparativo entre 31 genomas mitocondriais da subordem Loricarioidei, sendo 26 espécies de Loricariidae e uma de Callichthyidae. As características estruturais foram altamente conservadas. Entretanto, uma deleção parcial no final 3' do Bloco de Sequência Conservada (CSB) D da região controle foi identificada em um clado monofilético. As espécies que compõem esse clado são candidatas a modelos para estudar a replicação e a transcrição do genoma mitocondrial vertebrado. A filogenia recuperada corroborou fortemente a árvore atualmente aceita, embora tenha discutido a proximidade da relação entre duas espécies pertencentes a tribos distintas. O estudo filogenético destacou, também, a baixa variabilidade genética no clado da *Peckoltia*, um grupo eco-morfológico diversificado e taxonomicamente problemático. Os novos recursos genômicos reduzem o hiato da informação sobre a diversidade molecular da fauna de peixes neotropicais, impactando a capacidade de investigar uma variedade de aspectos da ecologia molecular e a evolução desses peixes.

Moreira *et al. BMC Genomics* (2017) 18:345
DOI 10.1186/s12864-017-3709-3

BMC Genomics

RESEARCH ARTICLE Open Access

Reducing the information gap on Loricarioidei (Siluriformes) mitochondrial genomics

Daniel Andrade Moreira[1,2], Paulo Andreas Buckup[3], Carolina Furtado[4], Adalberto Luis Val[5], Renata Schama[2] and Thiago Estevam Parente[1,6*]

Abstract

Background: The genetic diversity of Neotropical fish fauna is underrepresented in public databases. This distortion is evident for the order Siluriformes, in which the suborders Siluroidei and Loricarioidei share equivalent proportion of species, although far less is known about the genetics of the latter clade, endemic to the Neotropical Region. Recently, this information gap was evident in a study about the structural diversity of fish mitochondrial genomes, and hampered a precise chronological resolution of Siluriformes. It has also prevented molecular ecology investigations about these catfishes, their interactions with the environment, responses to anthropogenic changes and potential uses.

Results: Using high-throughput sequencing, we provide the nearly complete mitochondrial genomes for 26 Loricariidae and one Callichthyidae species. Structural features were highly conserved. A notable exception was identified in the monophyletic clade comprising species of the *Hemiancistrus*, Hypostomini and *Peckoltia*-clades, a ~60 nucleotide-long deletion encompassing the seven nucleotides at the 3′ end of the Conserved Sequence Block (CSB) D of the control region. The expression of mitochondrial genes followed the usual punctuation pattern. Heteroplasmic sites were identified in most species. The retrieved phylogeny strongly corroborates the currently accepted tree, although bringing to debate the relationship between *Schizolecis guntheri* and *Pareiorhaphis garbei*, and highlighting the low genetic variability within the *Peckoltia*-clade, an eco-morphologically diverse and taxonomically problematic group.

Conclusions: Herein we have launched the use of high-throughput mitochondrial genomics in the studies of the Loricarioidei species. The new genomic resources reduce the information gap on the molecular diversity of Neotropical fish fauna, impacting the capacity to investigate a variety of aspects of the molecular ecology and evolution of these fishes. Additionally, the species showing the partial CSB-D are candidate models to study the replication and transcription of vertebrate mitochondrial genome.

Keywords: Catfishes, Teleostei, Loricariidae, Biodiversity, Evolution, Neotropical Region, Next-Generation Sequencing

* Correspondence: parente@ioc.fiocruz.br
[1] Laboratório de Toxicologia Ambiental, Escola Nacional de Saúde Pública (ENSP), Fundação Oswaldo Cruz (FIOCRUZ), Av. Brasil, 4036, Rio de Janeiro, Brasil
[6] Laboratório de Genética Molecular de Microrganismos, Instituto Oswaldo Cruz (IOC), Fundação Oswaldo Cruz (FIOCRUZ), Av. Brasil, 4365, Rio de Janeiro, Brasil
Full list of author information is available at the end of the article

Background

Representing around 5% of vertebrate biodiversity [1], catfishes (Siluriformes) are classified in three suborders: Diplomystoidei, with seven valid species distributed in Andean areas of Argentina and Chile; Siluroidei, summing 2250 species distributed worldwide; and Loricarioidei, endemic to the Neotropical Region with 1538 valid species [1–3]. As of March 2017, while one nuclear genome and 102 complete mitochondrial genomes from Siluroidei species were deposited in GenBank, only four Loricarioidei species had their mitochondrial genome sequence publicly released (Table 1). This disproportion is even greater for general nucleotide sequences; for which the number of Siluroidei entries is almost 100 times the amount for Loricarioidei (Table 1).

This information gap on the genetic diversity of Loricarioidei fishes is evident in studies of Siluriformes [3, 4] and Otophysi [5, 6] evolution. Recently, in a time-calibrated mitogenome phylogeny of catfishes, the poor taxonomic representation of the Loricarioidei suborder was hypothesized as the most probable cause for the paraphyletic arrangement of the Callichthyidae and Loricariidae families [1]. According to those authors, although this arrangement "*does not represent a credible topology*", it "*may have implications for the chronology of the basal siluriform nodes*" [1], and therefore, may impact studies on the historical biogeography of catfish dispersal throughout the globe.

The underrepresentation of Loricarioidei genetic information deposited in public databases has also prevented a variety of studies about these species, from molecular ecology, biogeography and phylogeny to potential uses in aquaculture or as sentinels of environmental pollution. For instance, species delimitation and phylogenetic analysis of highly speciose genera such as *Hypostomus*, *Peckoltia* and *Corydoras* could benefit from a larger volume of primary genetic resources. Access to new gene sequences would stimulate work on the biogeography and speciation processes of Loricarioidei fishes with wide distribution (e.g.: *Hoplosternum littorale*, *Callichthys callichthys*). Likewise, the lack of genetic data is a key limitation preventing studies about the molecular ecology of those Neotropical fishes.

Herein, we focused on the Loricariidae, the most species-rich family in the Loricarioidei and the fifth most species-rich family among all vertebrates [2, 7]. The nearly complete mitochondrial genome sequences from 26 species of Loricariidae and one of Callichthyidae were generated, and their structural features were analyzed. Additionally, the sequences of the 13 protein-coding genes and the two ribosomal RNAs were used to test the currently accepted phylogenetic relationships among Loricariidae subfamilies. These resources complement our recent efforts [8–12], augmenting from four to 36 the number of nearly complete mitochondrial genomes available for Loricarioidei fishes.

Results

The mitochondrial genomes of Loricariidae

The mitochondrial genomes from 27 species, representing 21 Loricariidae genera and *Corydoras schwartzi* (Callichthyidae), were sequenced almost to their full length. Detailed information regarding each mitogenome is available in Table 2 and Additional files 1, 2 and 3.

Table 1 Disproportion of genetic information among Siluriformes' suborders. Different types of entries in NCBI database, as well as the number of available, valid and new species described in the last 10 years for Siluriformes according to the Catalog of Fishes database

		Siluriformes		
Database	Parameter	Diplomystoidei	Loricarioidei	Siluroidei
NCBI	Nucleotide	547	8599	855963
	Nucleotide EST	-	-	498206
	Nucleotide GSS	-	-	63406
	Protein	360	6082	76138
	Structure	-	-	7
	Genome	1	4	103
	Popset	25	162	546
	GEO Datasets	-	-	118
	UniGene	-	-	204837
	PubMed Central	17	321	5513
	Gene	13	76	30421
	SRA Experiments	-	9	212
	Probe	-	-	5615
	Assembly	-	-	1
	Bio Project	-	5	86
	Bio Sample	-	9	304
	Clone DB	-	-	12
	PubChem BioAssay	-	-	1
	Taxonomy	13	1433	1766
CAL - Catalog of Fishes	Available	12	1801	3481
	Valid	7	1538	2250
	2008–2017	1	334	284

NCBI: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=7995 and the Catalog of Fishes: http://researcharchive.calacademy.org/research/ichthyology/catalog/SpeciesByFamily.asp, both accessed on March 24, 2017

Coverage, annotation and depth

The new sequences covered from 91.2 to 99.2% the complete *Pterygoplichthys disjunctivus* mitogenome (Table 2), which was chosen as reference as it was one of the four complete mitogenomes of a Loricarioidei species previously available on public databases, not sequenced by our group. The few nucleotide-long gaps verified in the assembled mitochondrial genomes were located at transfer RNA (tRNA) genes. Despite these few gaps, the analysis of the sequenced mitogenomes, along with those available on GenBank, made it possible to infer the mitochondrial gene composition and order for all species, which was found to be identical to the usual pattern for vertebrates (Fig. 1).

Sequences of the two ribosomal RNA (12S and 16S rRNA) and the 13 protein-coding genes were obtained to their full length in most species. Most genes are encoded on the heavy strand, whereas NADH dehydrogenase subunit 6 (nad6) and eight tRNA are found on the light strand (Fig. 1). The majority of the 22 tRNA were completely sequenced in all species, except for *Hypostomus affinis* (Additional file 3). The complete sequence of the mitochondrial termination factor (mTERF) binding site was obtained for 15 species, while partial sequence was obtained for 10 species. The mTERF is located inside the tRNA-Leu2, justifying the highest frequency of partial sequences verified for this tRNA. Nucleotide frequencies among species were similar (Additional file 1). Among the protein-coding genes, cytochrome c oxidase subunit 1 (cox1) had the highest percentage of invariable amino acids (90%) and ATPase subunit 8 (atp8), the lowest (45%).

The origin of L-strand replication (O_L) was sequenced, except for *Loricariichthys castaneus* and for *Hypostomus affinis*, and was conserved in most species. The mitochondrial control region (CR) was partially sequenced in all species, except *Schizolecis guntheri* in which CR was not sequenced (Fig. 1, Additional file 4). The sequence length of the CR varied from zero in *S. guntheri* to 1083 in *K. heylandi*, with a mean length of 792 nucleotides. The three CR domains (I, II and III) were found. The termination associated sequence (TAS) in the Domain I was enriched with adenine (A) and thymine (T) nucleotides, and indels were found among the species. Three conserved sequence blocks (CSB-F, CSB-E and CSB-D) were identified in Domain II.

CSB-F was the most conserved among Loricarioidei species (Additional file 4). CSB-E was conserved among 16 species, but diverged in species of the *Hemiancistrus*, Hypostomini and *Peckoltia*-clades. While the 3′ end of CSB-E is rich in AT in the *Peckoltia*-clade and in A in the *Hemiancistrus* and Hypostomini clades, it is rich in G in the other investigated species (Additional file 4). A deletion of ~60 nucleotides, starting with the seven nucleotides at the 3′ end of CSB-D and spanning over other conserved blocks, follows the same phylogenetic trend (Fig. 2). This deletion is supported by >50 reads in each of the species sequenced herein and by the previously deposited sequence of *Pterygoplichthys disjunctivus* (NC_015747.1). At the same region, a six nucleotide-long insertion was noted in both species of *Rineloricaria* (Fig. 2). In most species, the T-homopolymer region was also found, which was followed by an AT-rich segment that characterizes the third domain. CSB-1, CSB-2 and CSB-3 were fully sequenced and highly conserved among the analyzed genomes.

Sequencing depth is a direct reflection of mitochondrial transcript expression levels. Minimum median sequencing depth was 329 found for *Parotocinclus maculicauda*, while the maximum median depth was 10,105 for *Loricaria cataphracta* (Fig. 1, Additional file 1). The variation among species is regarded as minimal differences in the amount of sample submitted to high-throughput sequencing rather than biological. On the other hand, the variation among regions within a single mitogenome is biologically significant, follows the classical mitochondrial RNA punctuation pattern and indicates the activity of a post-transcriptional gene expression mechanism. Genes coding for the three mitochondrial subunits of cytochrome c oxidase were sequenced at the greatest depth, while short intergenic and tRNA sequences showed the lowest depth (Fig. 1). Heteroplasmic positions were found in all species, except one (Additional file 5). The number of heteroplasmies varied from zero in Neoplecostomini gen. n. (TP065) to 21 in *Loricariichthys castaneus*, with median and average frequencies of 8.0 and 8.9, respectively.

Annotation features

Detailed annotation features of the 13 protein-coding genes of each species are summarized in Additional file 2. The protein-coding genes cytochrome c oxidase subunits 2 and 3 (cox2 and cox3), NADH dehydrogenase subunits 2, 3, 4, 4l and 5 (nad2, nad3, nad4, nad4l, and nad5) were found to be highly conserved, both in total length and nucleotide composition, among species of Loricariidae and Callichthyidae. Cox2, cox3, nad2, nad3, nad4 genes were terminated with an incomplete stop codon (T−), while a complete stop codon (TAA) was found in nad4l and nad5 of all species, except for *Hemipsilichthys nimius*' nad5 which is terminated by a TAG. In *Kronichthys heylandi* nad5 a GTG aligned to the ATG start codon that is found in the other species. However, the codon immediately before this GTG is an ATG. The other six protein-coding mitochondrial genes showed taxon-specific features that are addressed below.

The cytochrome b (cob) gene has 1138 bp and uses an incomplete stop codon (T−) in most species. However, a

Table 2 Geographical coordinates of sampled species and their field and voucher catalog numbers. Vouchers were deposited in the Ichthyological collection of the National Museum belonging to the Federal University of Rio de Janeiro (MNRJ). Quasi-complete mitochondrial genomes were deposited in GenBank and their accession numbers are provided, along with the percentage coverage in comparison to NC015747 for Loricariidae or NC004698 for *Corydoras*

Species	Field no.	Location	Catalog no.	Accession no.	Coverage	Reference
Hemipsilichthys nimius	TP189	23°12′35.2″S 44°47′40.7″W (RJ)	MNRJ43650	KT239011	95.50%	This study
Rineloricaria cf. *lanceolata*	sp16.3	Aquarium specimen (PA)	MNRJ43638	KX087182	97.80%	This study
Rineloricaria sp.	TP144	22°31′06,3″S 42°53′55,5″W (RJ)	MNRJ42544	KX087183	95.40%	This study
Loricariichthys platymetopon	TP179	3°10′50.9″S 59°54′09.3″W (AM)	MNRJ43627	KT239018	95.50%	This study
Loricariichthys castaneus	TP029	21°13′08.7″S 41°18′37.7″W (RJ)	MNRJ41545	KT239015	92.30%	This study
Loricaria cataphracta	TP181	3°10′50.9″S 59°54′09.3″W (AM)	MNRJ43629	KX087174	98.30%	This study
Otocinclus cf. *hoppei*	sp10.7	Aquarium specimen (PA)	MNRJ43634	KX087176	99.00%	This study
Hypoptopoma incognitum	TP171	3°09′36.0″S 59°55′12.0″W (AM)	MNRJ43421	KT033767	100%	Moreira et al. (2016b) [10]
Parotocinclus maculicauda	TP011	22°36′01.6″S 43°05′30.1″W (RJ)	MNRJ41523	KX087179	94.90%	This study
Hisonotus thayeri	TP128	21°32′14.6″S 42°06′54.8″W (RJ)	MNRJ42481	KX087173	96.00%	This study
Kronichthys heylandi	8505	23°12′35.2″S 44°47′40.7″W (RJ)	MNRJ42082	KT239014	99.00%	This study
Neoplecostomini gen. n.	TP065	20°01′35.3″S 40°36′33.3″W (ES)	MNRJ41921	KX087172	95.50%	This study
Neoplecostomus microps	TP088	22°20′01.7″S 44°32′34.3″W (RJ)	MNRJ41752	KX087175	96.40%	This study
Pareiorhaphis garbei	TP009	22°32′03.4″S 43°02′18.7″W (RJ)	MNRJ41511	KX087178	96.80%	This study
Schizolecis guntheri	TP006	22°32′03.4″S 43°02′18.7″W (RJ)	MNRJ41510	KT239017	91.20%	This study
Ancistrus sp. 1	13.3	Aquarium specimen (PA)	MNRJ42890	KP960569	99.20%	Moreira et al. (2015) [8]
Ancistrus sp. 2	13.11	Aquarium specimen (PA)	MNRJ42890	KP960567	94.70%	Moreira et al. (2015) [8]
Ancistrus multispinis	TP003	22°32′03.4″S 43°02′18.7″W (RJ)	MNRJ41509	KT239006	96.30%	This study
Dekeyseria amazonica	TP165	3°09′36.0″S 59°55′12.0″W (AM)	MNRJ43618	KX087168	98.80%	This study
Baryancistrus xanthellus	sp11.19	Aquarium specimen (PA)	Missing	KX087167	99.10%	This study
Pterygoplichthys sp.	sp2	Aquarium specimen (RJ)	MNRJ43652	KX087181	96.80%	This study
Pterygoplichthys pardalis	TP154	3°09′36.0″S 59°55′12.0″W (AM)	MNRJ43607	KT239016	97.10%	This study
Pterygoplichthys disjunctivus		Not informed	Not informed	NC015747	100%	Nakatani et al. (2011) [6]
Hypostomus sp.	sp12.6	Aquarium specimen (PA)	MNRJ43635	KX087171	95.40%	This study
Hypostomus cf. *plecostomus*	TP164	3°09′36.0″S 59°55′12.0″W (AM)	MNRJ43617	KT239012	98.50%	This study
Hypostomus affinis	TP147	22°48′42.6″S 43°37′42.8″W (RJ)	MNRJ43256	KT239013	93.30%	This study
Aphanotolurus emarginatus	TP184	3°10′50.9″S 59°54′09.3″W (AM)	MNRJ43631	KT239019	96.90%	This study
Peckoltia furcata	sp15.2	Aquarium specimen (PA)	MNRJ43637	KX087180	96.30%	This study
Ancistomus snethlageae	sp17.2	Aquarium specimen (PA)	MNRJ43639	KX087166	98.90%	This study
Panaqolus sp.	sp4	Aquarium specimen (RJ)	MNRJ43654	KX087177	99.10%	This study
Corydoras nattereri	TP021	22°36′01.6″S 43°05′30.1″W (RJ)	MNRJ41520	KT239009	100%	Moreira et al. (2016a) [9]
Corydoras schwartzi	TP177	Aquarium specimen (AM)	MNRJ43625	KT239007	98.10%	This study
Corydoras rabauti		Not informed	Not informed	NC004698	100%	Saitoh et al. (2003) [16]

premature stop codon was found in all Loricariinae species, as well as in the four Ancistrinii clade species. This premature stop codon is produced by the change from TGA to either TAA or TAG and aligns with the penultimate codon of the other sequences. Within the context of the phylogenetic tree (see below) the presence of this premature stop codon is most parsimoniously interpreted as independent synapomorphies supporting the monophyly of the Loricariinae and Ancistrini. This finding is supported by other cob sequences from Loricariinae and Ancistrinii species available at GenBank, and can be interpreted as an example of convergent evolution.

The Loricariidae species possess a 1551 bp long cox1 gene, as opposed to Callichthyidae in which this gene consists of 1560 bp. The Loricariidae and Callichthyidae species sequenced herein use GTG as cox1 start codon. All Loricariidae species use TAA as a stop codon for this

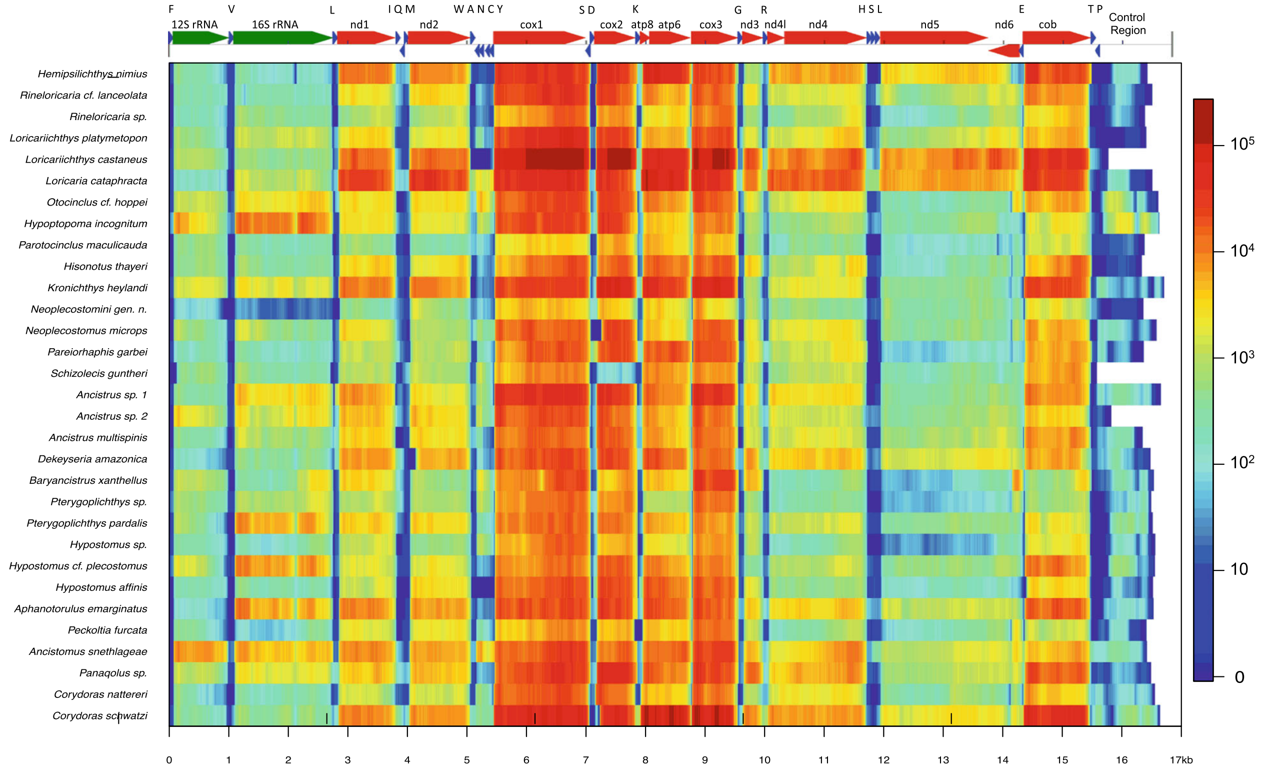

Fig. 1 Annotation and sequencing depth of the 27 newly sequenced mitochondrial genomes. Annotation features are shown on the top of this figure as a linear representation generated by the MITOS WebServer [34]. In the upper panel, protein-coding genes are colored in *red*, ribosomal genes in *green*, and tRNA in *blue*. Each tRNA is identified with the one-letter code for the corresponding amino acid. The sequencing depth of each position of the mitochondrial genomes is indicated below the general annotation. In the lower panel, color gradient corresponds to the number of reads supporting a given nucleotide, according to the legend at the right of the figure. Illumina sequencing reads were aligned to the correspondent mitochondrial genome, using Bowtie 1.0 and counted on the Integrative Genome Viewer (IGV) [35]. Representation of read counts as a color-gradient was generated using R. For comparative purposes, the recently released mitogenomes of *H. incognitum*, *Ancistrus* sp. 1, *Ancistrus* sp.2 and *C. nattereri* were included

gene, except for two species, *Kronichthys heylandi* and *Schizolecis guntheri*, which use an incomplete stop codon (TA-). This exception cannot be confirmed in the literature as the nine cox1 genes sequences publicly available at GenBank and the Barcode of Life Data (BOLD) Systems for *K. heylandi* and the 12 for *S. guntheri* do not include the start or stop codons. Callichthyidae species were found to use AGG stop codon for cox1 termination.

The nad1 gene of the Callichthyidae species and of *Hemipsilichthys nimius* lack an amino acid at the sixth amino acid position, when aligned to the gene sequence of the remaining species. Another amino acid-long gap, at the 257th amino acid position, was found to be exclusive of Loricariinae representatives. In the case of *Loricaria cataphracta*, however, this gap consists of two amino acids. Among the Hypostominae, *Ancistrus sp. 1* and *Ancistrus sp. 2* show an additional amino acid-long gap at the penultimate position. These might be phylogenetically informative characters for further studies of Loricariinae and Callichthyidae.

The nad6 is a small protein-coding gene encoded by 522 bp in the studied Loricariidae, except for the five species of Loricariinae and *Schizolecis guntheri*. In these six exceptions, there is an amino acid-long gap at the 116th position. This gap is shared with the three *Corydoras* species, which also present another gap at the position 142nd.

The gene atp8 is coded by 168 bp in all studied species, except for four of the five Loricariinae species. In the two *Rineloricaria* species and in the two *Loricarichthys* species, an amino acid-long gap was found at the 40th position. As for the ATPase subunit 6 (atp6) gene, *Hemipsilichthys nimius* was the only species with GTG, instead of ATG, as the start codon. All Loricariidae species have the incomplete stop codon TA- to terminate the atp6 gene, while species of Callichthyidae have the complete stop codon TAA. *Corydoras* spp. (Callichthyidae) share an apomorphic insertion between the atp6 and cox3 gene. In the newly sequenced *C. schwartzi* mitogenome, this insertion has 17 nucleotides.

Inferred phylogeny of Loricariidae subfamilies

Sequences from the two ribosomal RNA (12S and 16SrRNA) and the 13 protein-coding mitochondrial

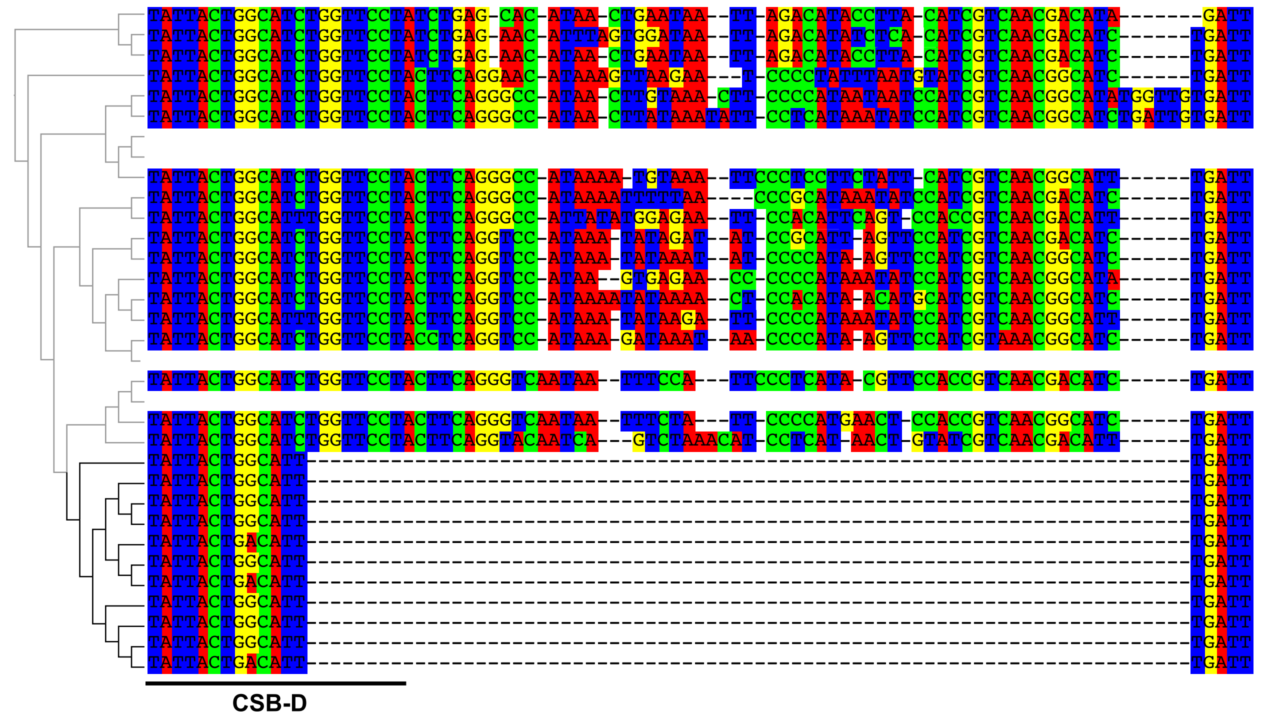

Fig. 2 Long nucleotide deletion at the mitochondrial control region in species belonging to the Hypostomini, *Hemiancistrus* and *Peckoltia*-clades. Insertion/deletion mutations (indels) are shown as hyphen (–). The position of the Conserved Sequence Block D (CSB-D) is indicated at the left bottom. Phylogenetic relationships among sequences are shown on the left. Branches of the clade displaying the deletion are shown in *black*, while other are in *gray*. The names of each taxon are shown in Fig. 3

genes were concatenated generating a super-alignment of 14116 nucleotides, for each of the 30 Loricariidae and three *Corydoras* species, producing a fully resolved maximum likelihood (ML) phylogenetic tree (Fig. 3, Final ML Optimization Likelihood: −144076.08), with branches showing high statistical support. The phylogenetic tree was rooted using sequences from *Corydoras schwartzi* and two additional *Corydoras* species obtained from GenBank (Table 2). *Hemipsilichthys nimius* (Delturinae) appears as the sister taxon of the other Loricariidae sub-families, except for Lithogeninae (not sampled in this study). The remaining loricariids form three large clusters, corresponding to the subfamily Loricariinae and a monophyletic group comprised by two clades: the subfamily Hypostominae and a clade formed by Neoplecostominae surrounded by Hypoptopomatinae.

Notably, the retrieved phylogeny strongly supports the position of Otothyrini, an Hypoptopomatinae tribe, closer to Neoplecostominae species, rather than to Hypoptopomatini species, the other traditional Hypoptopomatinae tribe. Thus, Hypoptopomatinae is considered monophyletic but, instead of two, it is composed by three tribes, Hypoptopomatini, Otothyrini and Neoplescostomini. *Schizolecis guntheri* is currently classified as a member of Otothyrini, but appeared clustered closer to *Pareiorhaphis garbei* and other members of Neoplecostomini than to the other Otothyrini.

It has been shown that for some lineages the phylogenetic tree topology retrieved using data from as few as three mitochondrial genes is the same as the topology retrieved using a concatenated alignment from the 13 protein-coding and the two rRNA genes [13]. In order to evaluate this in Neotropical catfishes, phylogenetic trees were generated using the three longest mitochondrial genes, which also harbor the majority of informative sites (cox1, nad4 and nad5) and the three mitochondrial genes with the highest performance (nad2, nad4 and nad5), as identified by Havird & Santos [13]. Those trees did not recover the same topology as encountered in the tree recovered using information on the 15 mitochondrial genes (Additional file 6). The position of *Loricaria cataphracta* was distinct in the tree based on the three genes with the highest performance, and the position of *Dekeyseria amazonica* was different in the tree based on the three longest genes (Fig. 3, Additional file 6).

Pairwise nucleotide identity (PNI)

Nucleotide identities between pairs of concatenated mitochondrial genes were calculated for every combination of the 33 taxa, arranged according to their phylogenetic relationships and colored as a heat-map for

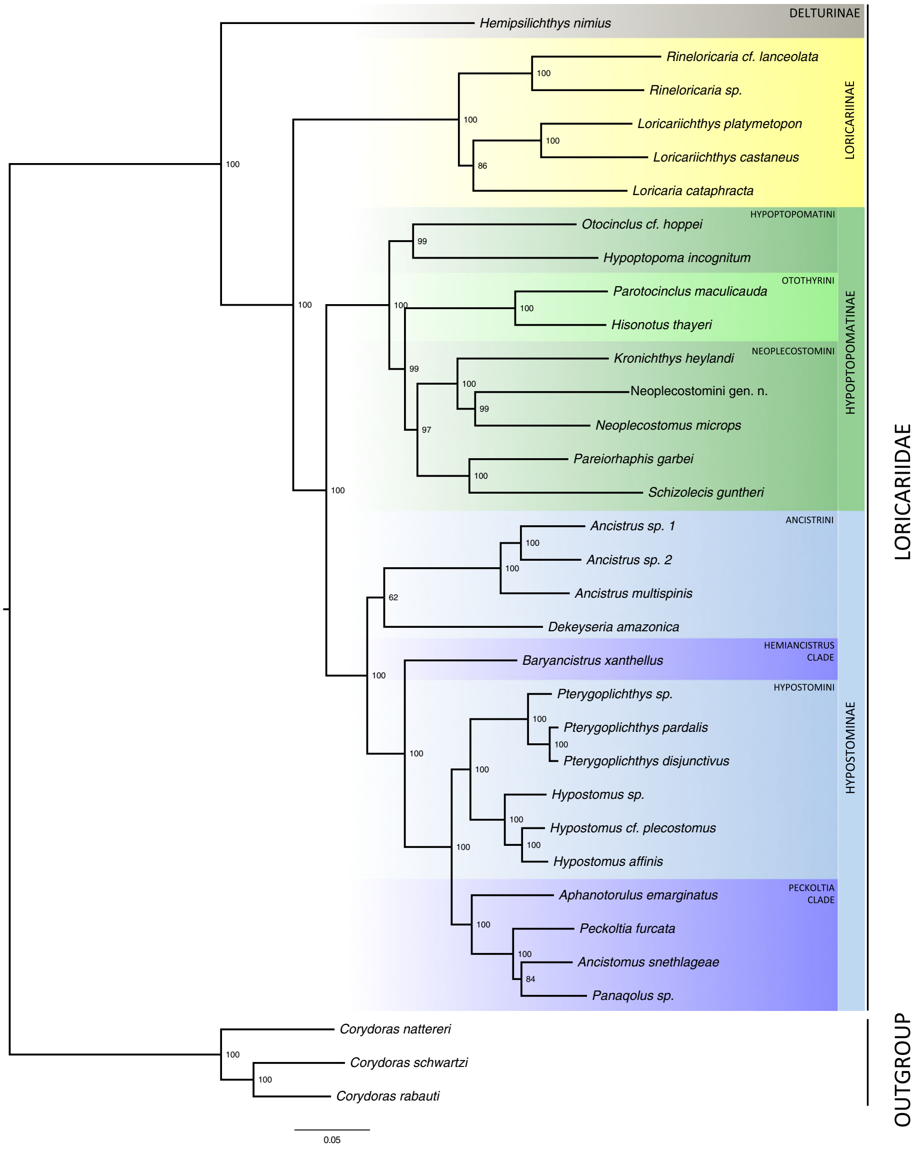

Fig. 3 Maximum likelihood phylogeny of Loricariidae subfamilies. The two ribosomal RNA and 13 protein-coding genes (comprising 14116 nucleotides) were aligned using MUSCLE. Bootstrap support values are shown on each node and are based on 1000 replicates. Callichthyidae species were used to root the tree. Loricariidae subfamilies are highlighted in *gray* (Delturinae), *yellow* (Loricariinae), *green* (Hypoptopomatinae) and *blue* (Hypostominae). The scale bar represents the nucleotide substitution rate, using the GTR + GAMMA + I model

visualization (Fig. 4). Four islands displaying higher pairwise nucleotide identities than their surroundings were identified. Not surprisingly, each of these islands is equivalent to major monophyletic clades. Pairwise nucleotide identities (PNI) among congeneric species displayed a higher degree of similarity, ranging from 89.2% between the two *Rineloricaria* to 95.4% between *Hypostomus* sp. and *H.* cf. *plecostomus*. PNI inside each island were most often above 85%, whereas nucleotide identity outside the islands but contained in the Loricariidae realm was most often below 85%, ranging down to 80%. Outside the Loricariidae realm, nucleotide identity between any Loricariidae species and any *Corydoras* ranged from 78 to 80%.

Discussion

Mitochondrial genome structure

The structure of the newly sequenced mitochondrial genomes follows the general pattern described for ostariophysian fish [14]. Among the identified differences, a ~60 nucleotide-long deletion at the control region shared by *Baryancistrus xanthellus* and species of Hypostomini and *Peckoltia*-clade highlights. In accordance with findings based on 248 ray-finned fish [14], cox1 was the most conserved protein-coding gene in the mitochondrial genomes of Loricarioidei species, while atp8 was the most variable. Corroborating Satoh et al. [14], the mitochondrial termination factor (mTERF) binding site found in the species used in this study was identical to the human sequence, implying functional conservation. The mTERF binding site is located inside the tRNA-Leu2 gene, which was the most frequent tRNA gene with partial sequence, corroborating the functional conservation.

The order of each gene in the mitogenome was inferred as the same as in other ostariophysian fish [14], which is the typical for metazoans [15]. An intriguing difference is the disruption of the classical head-to-tail junction between atp6 and cox3 among *Corydoras* species. A 17 nucleotide-long insertion was initially noted by Saitoh et al. [16] at the mitogenome of *C. rabauti*, but this trait was considered phylogenetically uninformative as it was not shared by any other species evaluated in that study [16]. Subsequently, Moreira et al. [9], observed a homologous 21 nucleotide-long insertion in the mitogenome of *C. nattereri* and hypothesized this apomorphic trait as relevant for the phylogeny of this speciose genus. The 17 nucleotide-long insertion found in *C. schwartzi* corroborates and extends that hypothesis. Although having the same length as in *C. rabauti*, these two 17 nucleotide-long insertions differ by six nucleotide substitutions (one transition and five transversions).

The mitochondrial control region is composed by three domains (I, II and III) containing several conserved sequence blocks (CSB), each showing a taxon-specific distribution in vertebrates [17–20]. Recently, it has been demonstrated that CSB-D and CSB-1 are present in 250 fish species, although their functions are not yet clear [14]. While the newly generated mitochondrial genomes confirmed the presence of CSB-1 within the Loricarioidei, more than one-third of the CSB-D was deleted in species of the *Hemiancistrus*, Hypostomini and *Peckoltia*-clades. A complete CSB-D was found in the other species, including the ones belonging to the subfamily Ancistrini, which is closely related to those clades where the partial CSB-D was identified. There are four monophyletic clades closer to the *Hemiancistrus*, Hypostomini and *Peckoltia*-clades than to Ancistrini [7], suggesting that this partial CSB-D deletion is a synapomorphy for a subset of Hypostominae species. It will be elucidative to test the occurrence of this deletion in species belonging to these four clades. Likewise, comparative studies using species in which CSB-D is partially deleted and their closest relatives with complete CSB-D might provide

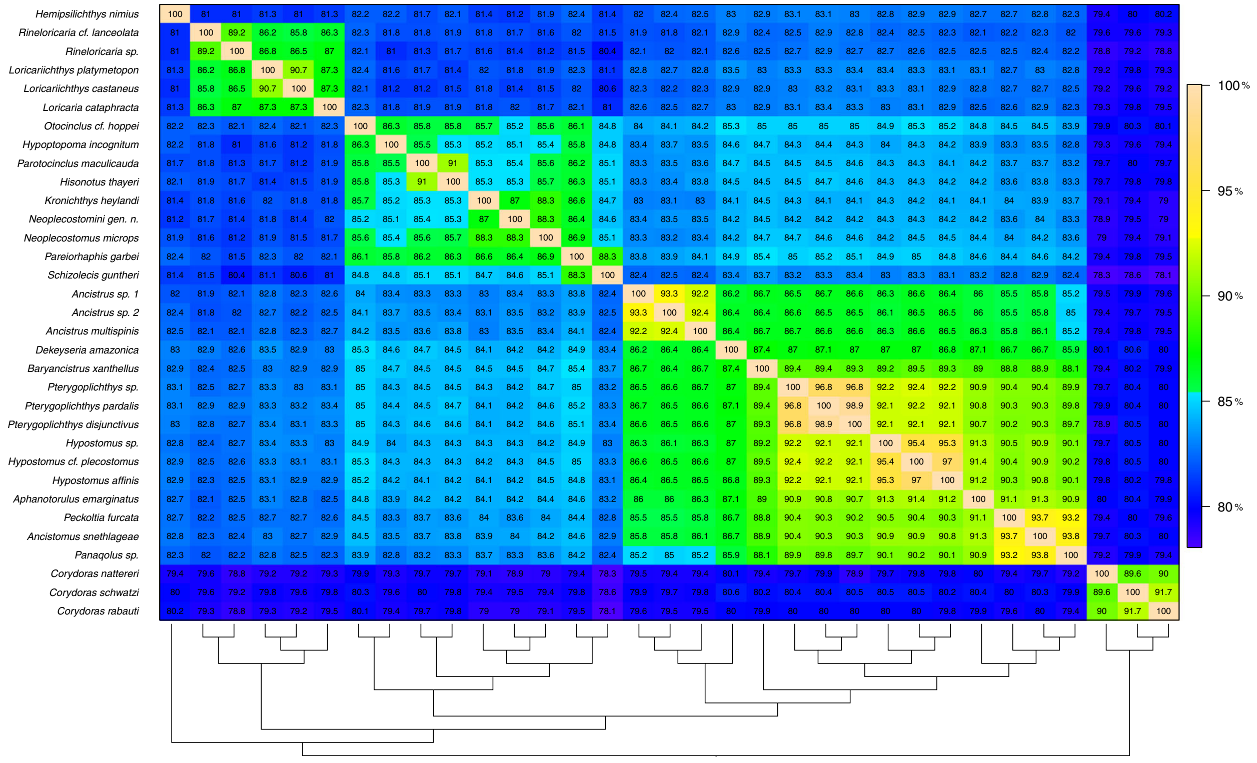

Fig. 4 Pairwise nucleotide identity among sequenced mitochondrial genomes. Species are ordered according to the retrieved phylogeny on both axes. Species names are shown on the *left side*, and their phylogenetic relationships are depicted at the *bottom*. The color gradient represents the percentage of nucleotide identity according to the legend at the *right* of the figure

insights to the role played by this conserved sequence block in replication and transcription of the genome.

Among the 250 fish species studied by Satoh et al. [14], three Siluriformes species were used; two representatives of Siluroidei and *Corydoras rabauti*, a Loricarioidei representative. The sequence of *Pterygoplichthys disjunctivus* (NC_015747.1) corroborates this deletion in the CSB-D, and although it was available since 2011, it was not included in the study of Satoh et al. [14]. Here, the mitochondrial genome of *P. disjunctivus* is used to confirm our findings, as it was obtained using a combination of long and short PCR from DNA and sequenced using the Sanger method [6]. Additionally, this deletion is supported by the recently released mitochondrial genomes of *Hypancistrus zebra* (*Peckoltia*-clade; KX611143.1) [11] and *Pterygoplichthys anisitsi* (Hypostomini; KT239003, KT239004, KT239005) [12].

Mitochondrial gene expression

The variation found in the expression of mitochondrial genes is in accordance with the punctuation pattern, in which the polycistronic RNA is cleaved at the tRNAs genes [21, 22]. This pattern has been observed for other Loricarioidei species [8–11], and suggests the fish mitochondria are transcribed as a polycistronic RNA which is latter edited to monocistronic molecules and to two bicistronic RNAs, coding for nad4/nad4L and atp8/atp6.

The high sequencing depth also allowed for the detection of heteroplasmic sites in all species, except in Neoplecostomini gen. n. (TP065), the species with the lowest count of mapped reads. On the other hand, *Loricariichthys castaneus* was the species with the highest counts of both heteroplasmic sites and mapped reads (Additional file 5). Recently, we reported similar frequencies of heteroplasmic sites found in the mitogenome of other fish species [8, 10, 11]. Mitochondrial heteroplasmies are associated with several health disorders in humans [23–25] and with insecticide resistance in insects [26]. However, the consequences of heteroplasmies on fish physiology, if any, are unknown.

Phylogenetic analysis

Our phylogenetic analysis gives strong statistical support for most of the recently published Loricariidae phylogenies, including branches with poor bootstrap values [7, 27, 28]. Specifically, our data corroborate previous studies on the positioning of Neoplecostomini as a tribe embedded within the subfamily Hypoptopomatinae. The mitochondrial genomic data also supports the positioning of Neoplecostomini closer to Otothyrini, with Hypoptopomatini as the sister taxon of Neoplecostomini and Otothyrini, as proposed by Roxo et al. [27].

Schizolecis guntheri was found to be a member of Neoplecostomini rather than Otothyrini, its current classification based on morphological traits and limited molecular evidence [27, 28]. Indeed, a sister-group relationship between *S. guntheri* and *Pareiorhaphis garbei* was previously recovered by Cramer et al. [28], using a matrix of 4678 nucleotides from partial sequences of one mitochondrial (cox1) and three nuclear genes (recombination activating genes 1 and 2, and F-Reticulon 4). The *S. guntheri* and *P. garbei* specimens used by Cramer et al. [28] were sampled from the same population as in the present study. This finding is corroborated to some extend by the molecular identification of our specimens using the BOLD Systems (Additional file 7). According to the BOLD algorithm, our specimen of *P. garbei* appears in a small cluster composed by a few other *P. garbei* and unidentified *Neoplecostomus* species, in a sister relationship to a larger *S. guntheri* clade, while our *S. guntheri* appears embedded among other *S. guntheri* specimens from other regions in Brazil.

Of note, the pairwise nucleotide identity (PNI) variation encountered among three genera of the *Peckoltia*-clade (*Peckoltia*, *Ancistomus* and *Panaqolus*) is below the range of variation encountered within each of the two genera in its sister clade (Hypostomini), but above the variation between these genera (*Hypostomus* and *Pterygoplichthys*). This is also valid if we include the recently described mitochondrial genome of *Hypancistrus zebra* [11], another member of the *Peckoltia*-clade. However, the morphological diversity within the *Peckoltia*-clade exceeds by far the variation found among Hypostomini. Although the Hypostomini and *Peckoltia*-clade are both species-rich, most Hypostomini species are grouped into two large genera (*Hypostomus* and *Pterygoplichthys*) while the *Peckoltia*-clade is the most genus-rich Loricariidae tribe [7] and also harbors the worst taxonomic problems within this family [29].

The low divergence found among the mitochondrial genomes from the four genera (*H. zebra* included) belonging to the *Peckoltia*-clade is a piece of information that shall assist in the explanation and resolution of these long standing taxonomic problems. In addition, an intriguing question emerges from our results: How is the tremendous eco-morphological diversity characteristic of the Peckoltia-clade achieved with such a low genetic variation?

Conclusions

Herein we have launched the use of high-throughput mitochondrial genomics in the studies of the Loricarioidei species, advancing the knowledge on the genetic diversity of Neotropical fish fauna. Nearly complete mitochondrial genomes were sequenced for 27 species, representing 21 Loricariidae and one Callichthyidae genera.

These new resources greatly reduce the underrepresentation of Loricarioidei among the Siluriformes mitochondrial genomes deposited in GenBank and other public databases. However, the proportion of genetic data available for Siluroidei still outpaces that of Loricarioidei, especially with regard to nuclear genes.

Methods

Taxonomic sampling and RNA extraction

Twenty-seven species from 21 Loricariidae genera (representing six subfamilies) and one Callichthyidae genus, were sampled in the Amazon basin, and in coastal basins of southeastern Brazil, and acquired at local ornamental fish suppliers in the cities of Rio de Janeiro (Rio de Janeiro State) and Belém (Pará State), in Brazil (Table 2). Samples from the Amazon basin were collected near Manaus (Amazonas State); samples from coastal streams were collected in the states of Rio de Janeiro and Espírito Santo. Specimens were initially identified by experienced taxonomists; identifications were then confirmed by similarity searches using the cox1 Folmer region from each specimen against the BOLD Systems database (Additional file 7). Additionally, the complete and nearly complete mitochondrial genomes from Loricarioidei species available on GenBank were downloaded and used for comparative and phylogenetic purposes. GenBank accession numbers of all mitogenomes are displayed in Table 2. All voucher specimens were deposited at a permanent ichthyological collection (Museu Nacional, Universidade Federal do Rio de Janeiro - MNRJ) (Table 2). Fish sampling and handling were authorized by the appropriate Brazilian Government agency (Instituto Chico Mendes de Conservação da Biodiversidade, ICMBio), under the license number 21006-4.

Immediately after sampling, fish were killed, and liver dissected and preserved in RNA Later solution (Thermo Fisher Scientific) at 4 °C until arrival at the laboratory, where samples were stored at −20 °C. Total RNA was extracted from the liver tissue following the phenol-choloroform method, according to manufacture's instructions (TRIzol Reagent, Thermo Fisher Scientific). After extraction, total RNA was initially quantified using a BioDrop DUO (BioDrop) spectrophotometer, and quality and quantity were further evaluated using the Bioanalyzer RNA 6000 nano kit (Agilent). Only RNA preparations with RNA Integrity Number (RIN) above 6.0 were used for the cDNA library preparation.

Library preparation and sequencing

Complementary DNA (cDNA) libraries were prepared using the TruSeq RNA Sample Kit v.2 (Illumina), strictly following the manufacturer's recommendations. The fragmentation time in the thermocycler was adjusted to 45 s to ensure that most fragment lengths were in the range of 300 to 500 base pairs. All libraries were accessed for quality using the Bioanalyzer DNA 1000 kit (Agilent). Library quantification was performed using the Library Quantification Universal Kit for Illumina with Revised Primers-SYBR (Kapa Biosystems) for real time PCR for each library, and once more for each pool of up to nine libraries, individually marked with specific barcodes, destined to clusterization in Illumina HiSeq2500 sequencing lanes. Library pools were clustered using the TrueSeq PE Cluster Kit v3 for cBot (Illumina) into different lanes. Paired-end sequencing of 100 bp were performed on a HiSeq2500 using the TrueSeq SBS Kit v.3 (Illumina). Raw data were demultiplexed using the BCL2FASTQ software (Illumina). Reads were trimmed for adaptors with Trimmomatic [30] and their quality was evaluated using FastQC (Babraham Bioinformatics). Reads with Phred score equal to or higher than 30 were used for *de novo* assembly of transcriptomes using Trinity's default parameters [31, 32].

Mitogenome assembly and annotation

The mitochondrial genomes were assembled using the mitochondrial transcripts retrieved from the liver transcriptome of each fish, following the approach described by Moreira et al. [8]. Briefly, searches using the BLASTn algorithm against the mitogenomes from the closest related species available were performed to capture mitochondrial transcripts among the entire transcriptome. The retrieved mitochondrial transcripts were aligned to the complete mitogenome available from the closest related species (*Pterygoplichthys disjunctivus* NC015747, *Hypoptopoma incognitum* NC028072 or *Ancistrus* sp. KP960567). The assembled mitochondrial genomes were manually edited for removal of poorly aligned bases at both ends. In most cases, some gaps remained in intergenic and tRNA regions due to the punctuation pattern of the mitochondrial transcript expression. These gaps were filled with Ns.

The assembled mitochondrial genomes were annotated using the web-based services MitoFish and MITOS [33, 34], and manually curated for inconsistencies. To estimate the sequencing depth of each base, Bowtie v. 1.0.0 was used to align Illumina reads onto the assembled mitogenomes. Aligned reads were viewed using IGV, Tablet, PysamStats (v. 0.84) and BAMStats (v. 1.25) [35–37]. Heteroplasmic sites were identified using three conditions: first, different nucleotides were sequenced in the same position; second, that position must have more than 100 supporting reads; and third, the frequency of the second most frequent base must be higher than 10%. Using these criteria, every polymorphic nucleotide is supported by at least 10 reads.

Phylogenetic analysis

The sequences for the two rRNA and for the 13 protein-coding genes were recovered in the mitochondrial genomes from all species. The sequences of the 15 mitochondrial genes were aligned using the built-in MUSCLE algorithm from SeaView [38, 39]. Protein-coding gene sequences were aligned by their amino acid residues, and phylogenetic analyses were performed using the nucleotide information. Alignments of each gene were concatenated to create a single super-alignment consisting of 14,116 nucleotides. Pairwise nucleotide identity was calculated using SeaView and tabulated in R (R Core Team [40]).

jModelTest2 (v. 2.1.6) available at the CIPRES Science Gateway were used to select the best-fit model of nucleotide substitution under the Bayesian information criterion (BIC) [41–43]. The ML analysis was performed using RAxML (v. 8.2.4) available at CIPRES under the GTR + GAMMA + I model for all sites [44]. Branch support was evaluated with 1000 bootstrap replicates.

Additional files

Additional file 1: Summary data about mitochondrial genome sequences produced in the study. Mitogenomes were assembled using transcriptomic data. The number of transcripts used to assemble each mitogenome is shown, as well as their lengths without gaps. The 100 bp paired end Illumina Hi-Seq2500 reads were mapped against the assembled mitochondrial genome. The number of total reads mapped, as well as the average and median sequencing depth, which estimates the number each nucleotide was sequenced, are also provided. The nucleotide usage for each mitochondrial genome is shown. (PDF 59 kb)

Additional file 2: Summary data about mitochondrial genome sequences produced in the study. Total length for each mitochondrial gene is given. For the protein coding genes, the used start and stop codons are shown. The complete 12S and 16S ribosomal RNA are highlighted in bold. (PDF 104 kb)

Additional file 3: Completeness of transfer RNAs sequencing. It is shown whether each of the 22 tRNA coded in the mitochondrial genome of 31 Loricarioidei species was sequenced to its complete length (complete), partially sequenced (partial) or not sequenced (not seq.). (PDF 65 kb)

Additional file 4: The mitochondrial control region highlighting the Conserved Sequence Blocks (CSB). Insertion/deletion mutations (indels) are shown as hyphens (–). Background is colored according to the nucleotide at each position, blue for T, red for A, yellow for G and green for C, or in white for indels. The positions of the CSB are delimited at the right bottom of the alignment. (PDF 376 kb)

Additional file 5: Heteroplasmic sites identified on the mitochondrial genomes of Loricarioidei catfishes. This spreadsheet file is sub-divided in 32 independent sheets; one containing the legend, another with a summary and one for each of the 31 Loricarioidei species, identified by their field numbers as detailed on Table 2 in the main text. The sheets for each species show every heteroplasmic position identified, the total number of supporting reads, as well as, for each nucleotide, the absolute count and its proportion. On the Summary sheet, each column contains all the heteroplasmic positions found for a single species (identified on the column head). The first column on the left shows the gene where this position is located. Genes coding for ribosomal RNAs are colored in green, while genes coding for transfer RNAs are colored in blue and protein coding genes are colored in orange. Lines represent homologous positions among the mitochondrial genomes from the different species. Three conditions were used to characterize a heteroplasmic site: first, different nucleotides were sequenced in the same position; second, that position must have more than 100 supporting reads; and third, the frequency of the second most frequent base must be higher than 10%. On the Summary sheet, the positions with more than 1000 supporting reads are highlighted in red. (XLSX 63 kb)

Additional file 6: Phylogenetic tree retrieved using the three longest genes (A) and the three genes with the best performance (B). The three longest genes, cox1, nad4 and nad5, also harbored most informative characters. A - The position of *Dekeyseria amazonica*, highlighted in a red box, changed in relation to the tree using the 15 mitochondrial genes. The three genes with best performance as identified by Havird & Santos [13] were nad2, nad4 and nad5. B - The position of *Loricaria cataphracta*, highlighted in a red box, changed in relation to the tree using the 15 mitochondrial genes. Genes were aligned with built-in MUSCLE using SeaView and used for phylogenetic tree reconstruction using RAxML under the GTR + GAMMA + I model and 1000 bootstrap replica. (PDF 62 kb)

Additional file 7: Confirmation of species identification using cox1 barcode sequence and the BOLD Systems. The Folmer region of Cytochrome c oxidase subunit 1 of each species was used as queries for similarity searches against the BOLD database. The resulting output is presented as a phylogenetic tree generated online. The query species are shown in red and indicated by a red arrow. For each tree, the species name, as well as its field and voucher numbers are shown above the arrow. (PDF 765 kb)

Abbreviations

12S rRNA and 16S rRNA: 12S and 16S ribosomal RNA; AM: State of Amazonas; atp6 and atp8: ATPase subunit 6 and 8; BOLD: Barcode of Life Data; cDNA: Complementary DNA; cob: Cytochrome b; cox1–3: Cytochrome c oxidase subunits 1,2 and 3; CR: Mitochondrial control region; ES: State of Espírito Santo; ICMBio: Instituto Chico Mendes de Conservação da Biodiversidade; ML: Maximum likelihood; MNRJ: Museu Nacional, Universidade Federal do Rio de Janeiro; mTERF: Mitochondrial termination factor; nad1–6 and 4l: NADH dehydrogenase subunit 1–6 and 4l; O_L: Origin of L-strand replication; PA: State of Pará; PNI: Pairwise nucleotide identity; RIN: RNA Integrity Number; RJ: State of Rio de Janeiro; TAS: Termination associated sequence; tRNA: Transfer RNA

Acknowledgments

Authors are thankful to Maithê G.P. Magalhães and Paula C.C. de Andrade for providing essential support for Illumina Libraries preparation at FIOCRUZ, to Emanuel Neuhaus (MN/UFRJ) and Gustavo A. Ferraro (MN/UFRJ, FAPERJ fellow) for providing curatorial support for the processing of voucher specimens at MN/UFRJ, to Dr. Rachel Ann Hauser-Davis for comments on early versions of this MS, and to Dr. Jansen Zuanon (INPA) for valuable assistance in the identification of Amazonian samples.

Funding

Major financial support came from the Partnerships for Enhanced Engagement in Research (PEER) grant from the U.S. Agency for International Development (PGA-2000003446 and PGA-2000004790) to TEP, associated with NSF grant DEB-1120263 to Dr. Mark E. Hahn, from the Woods Hole Oceanographic Institution (WHOI). Authors acknowledge the use of the computer clusters from the IOC/FIOCRUZ and the WHOI. Additional financial support was provided by grants from the following agencies from Brazil: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) e Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM), to PAB (CNPq 307610/2013-6, 564940/2010-0, 476822/2012-2, CAPES/CNPq Protax 440566/2015-0, and FAPERJ E-26/111.404/2012, E-26/200.697/2014), to ALV (CNPq 303930/2014-4, 573976/2008-2 and FAPEAM 0603159/2008), to DAM (CAPES PhD fellowship) and to TEP (CAPES postdoctoral PNPD fellowship). The funding agencies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The voucher specimens for species used herein are deposited in the permanent Ichthyological Collection of the Museu Nacional, Universidade Federal do Rio de Janeiro – MNRJ under the numbers listed in Table 2. The mitochondrial genomes described herein are available on GenBank under the accession numbers listed in Table 2.

Authors' contributions
D.A.M. performed data analyses/interpretation and drafted the manuscript. P.A.B. contributed with sample collection, data interpretation and manuscript preparation. C.F. performed Illumina sequencing and reviewed the manuscript. A.L.V. contributed with sample collection and manuscript preparation. R.S. contributed with data analyses and manuscript preparation. T.E.P. designed the work, contributed with data analyses, performed data interpretation and wrote the manuscript. All authors have read and approved the manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
"Not applicable."

Ethics approval and consent to participate
Fish sampling and handling were authorized by the appropriate Brazilian Government agency (Instituto Chico Mendes de Conservação da Biodiversidade, ICMBio), under the license number 21006–4.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Laboratório de Toxicologia Ambiental, Escola Nacional de Saúde Pública (ENSP), Fundação Oswaldo Cruz (FIOCRUZ), Av. Brasil, 4036, Rio de Janeiro, Brasil. [2]Laboratório de Biologia Computacional e Sistemas, Instituto Oswaldo Cruz (IOC), Fundação Oswaldo Cruz (FIOCRUZ), Av. Brasil, 4365, Rio de Janeiro, Brasil. [3]Departamento de Vertebrados, Museu Nacional, Universidade Federal do Rio de Janeiro (UFRJ), Quinta da Boa Vista, Rio de Janeiro, RJ, Brasil. [4]Unidade de Genômica, Instituto Nacional do Câncer (INCA), Rua André Cavalcanti, 37, Rio de Janeiro, Brasil. [5]Laboratório de Ecofisiologia e Evolução Molecular, Instituto Nacional de Pesquisas da Amazônia (INPA), Av. André Araújo, 2936, Manaus, Brasil. [6]Laboratório de Genética Molecular de Microrganismos, Instituto Oswaldo Cruz (IOC), Fundação Oswaldo Cruz (FIOCRUZ), Av. Brasil, 4365, Rio de Janeiro, Brasil.

Received: 27 January 2017 Accepted: 19 April 2017
Published online: 04 May 2017

References

1. Kappas I, Vittas S, Pantzartzi CN, Drosopoulou E, Scouras ZG. A time-calibrated mitogenome phylogeny of catfish (Teleostei: Siluriformes). PLoS One. 2016;11:e0166988.
2. Eschmeyer WN, Fong JD. Species by family/subfamily. Cat. FISHES. 2016 [cited 2016 Dec 15]. Available from: http://researcharchive.calacademy.org/research/ichthyology/catalog/SpeciesByFamily.asp.
3. Sullivan JP, Lundberg JG, Hardman M. A phylogenetic analysis of the major groups of catfishes (Teleostei: Siluriformes) using rag1 and rag2 nuclear gene sequences. Mol Phylogenet Evol. 2006;41:636–62.
4. Hardman M. The phylogenetic relationships among non-diplomystid catfishes as inferred from mitochondrial cytochrome b sequences; the search for the ictalurid sister taxon (Otophysi: Siluriformes). Mol Phylogenet Evol. 2005;37:700–20.
5. Peng Z, He S, Wang J, Wang W, Diogo R. Mitochondrial molecular clocks and the origin of the major Otocephalan clades (Pisces: Teleostei): a new insight. Gene. 2006;370:113–24.
6. Nakatani M, Miya M, Mabuchi K, Saitoh K, Nishida M. Evolutionary history of Otophysi (Teleostei), a major clade of the modern freshwater fishes: Pangaean origin and Mesozoic radiation. BMC Evol Biol. 2011;11:177.
7. Lujan NK, Armbruster JW, Lovejoy N, López-fernández H. Multilocus molecular phylogeny of the suckermouth armored catfishes (Siluriformes: Loricariidae) with a focus on subfamily Hypostominae. Mol Phylogenet Evol. 2015;82:269–88.
8. Moreira DA, Furtado C, Parente TE. The use of transcriptomic next-generation sequencing data to assemble mitochondrial genomes of *Ancistrus* spp. (Loricariidae). Gene. 2015;573:171–5.
9. Moreira DA, Buckup PA, Andrade PCC, Magalhães MGP, Brito M, Furtado C, et al. The complete mitochondrial genome of *Corydoras nattereri* (Callichthyidae:Corydoradinae). Neotrop Ichthyol. 2016;14:e150167.
10. Moreira DA, Magalhaes MGP, de Andrade PCC, Furtado C, Val AL, Parente TE. An RNA-based approach to sequence the mitogenome of *Hypoptopoma incognitum* (Siluriformes: Loricariidae). Mitochondrial DNA Part A. 2016;27:3784–6.
11. Magalhães MGP, Moreira DA, Furtado C, Parente TE. The mitochondrial genome of *Hypancistrus zebra* (Isbrücker &amp; Nijssen, 1991) (Siluriformes: Loricariidae), an endangered ornamental fish from the Brazilian Amazon. Conserv Genet Resour. 2016;0:1–6.
12. Parente TE, Moreira DA, Magalhães MGP, De Andrade PCC, Furtado C, Haas BJ, et al. The liver transcriptome of suckermouth armoured catfish (*Pterygoplichthys anisitsi*, Loricariidae): identification of expansions in defensome gene families. Mar Pollut Bull. 2017;115:352–61.
13. Havird JC, Santos SR. Performance of single and concatenated sets of mitochondrial genes at inferring metazoan relationships relative to full mitogenome data. PLoS One. 2014;9:1–10.
14. Satoh TP, Miya M, Mabuchi K, Nishida M. Structure and variation of the mitochondrial genome of fishes. BMC Genomics. 2016;17:719.
15. Bernt M, Braband A, Schierwater B, Stadler PF. Genetic aspects of mitochondrial genome evolution. Mol Phylogenet Evol. 2013;69:328–38.
16. Saitoh K, Miya M, Inoue JG, Ishiguro NB, Nishida M. Mitochondrial genomics of ostariophysan fishes: perspectives on phylogeny and biogeography. J Mol Evol. 2003;56:464–72.
17. Lee W-J, Conroy J, Howell WH, Kocher TD. Structure and evolution of Teleost mitochondrial control regions. J Mol Evol. 1995;41:54–66.
18. Cui Z, Liu Y, Li CP, You F, Chu KH. The complete mitochondrial genome of the large yellow croaker, *Larimichthys crocea* (Perciformes, Sciaenidae): unusual features of its control region and the phylogenetic position of the Sciaenidae. Gene. 2009;432:33–43.
19. Nilsson MA. The structure of the Australian and South American marsupial mitochondrial control region. Mitochondrial DNA. 2009;20:126–38.
20. Wang L, Zhou X, Nie L. Organization and variation of mitochondrial DNA control region in pleurodiran turtles. Zoologia. 2011;28:495–504.
21. Ojala D, Montoya J, Attardi G. tRNA punctuation model of RNA processing in human mitochondria. Nature. 1981;290:470–4.
22. Mercer TR, Neph S, Dinger ME, Crawford J, Smith M a, Shearwood AMJ, et al. The human mitochondrial transcriptome. Cell. 2011;146:645–58.
23. Lin Y-F, Schulz AM, Pellegrino MW, Lu Y, Shaham S, Haynes CM. Maintenance and propagation of a deleterious mitochondrial genome by the mitochondrial unfolded protein response. Nature. 2016;533:1–8.
24. Sosa MX, Sivakumar IKA, Maragh S, Veeramachaneni V, Hariharan R, Parulekar M, et al. Next-generation sequencing of human mitochondrial reference genomes uncovers high heteroplasmy frequency. PLoS Comput Biol. 2012;8:e1002737.
25. Huang T. Next generation sequencing to characterize mitochondrial genomic DNA heteroplasmy. Curr Protoc Hum Genet. 2011;19(8):1–12.
26. Van Leeuwen T, Vanholme B, Van Pottelberge S, Van Nieuwenhuyse P, Nauen R, Tirry L, et al. Mitochondrial heteroplasmy and the evolution of insecticide resistance: non-Mendelian inheritance in action. Proc Natl Acad Sci U S A. 2008;105:5980–5.
27. Roxo FF, Albert JS, Silva GSC, Zawadzki CH, Foresti F, Oliveira C. Molecular phylogeny and biogeographic history of the armored neotropical catfish subfamilies Hypoptopomatinae, Neoplecostominae and Otothyri. PLoS One. 2014;9:e105564.
28. Cramer CA, Bonatto SL, Reis RE. Molecular phylogeny of the Neoplecostominae and Hypoptopomatinae (Siluriformes: Loricariidae) using multiple genes. Mol Phylogenet Evol. 2011;59:43–52.
29. Armbruster JW, Werneke DC, Tan M. Three new species of saddled loricariid catfishes, and a review of *Hemiancistrus*, *Peckoltia*, and allied genera (Siluriformes). Zookeys. 2015;123:97–123.
30. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
31. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson D a, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.
32. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512.

33. Iwasaki W, Fukunaga T, Isagozawa R, Yamada K, Maeda Y, Satoh TP, et al. Mitofish and mitoannotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. Mol Biol Evol. 2013;30:2531–40.
34. Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, et al. MITOS: improved de novo metazoan mitochondrial genome annotation. Mol Phylogenet Evol. 2013;69:313–9.
35. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14:178–92.
36. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, et al. Tablet–next generation sequence assembly visualization. Bioinformatics. 2009;26:401–2.
37. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
38. Gouy M, Guindon S, Gascuel O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 2010;27:221–4.
39. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
40. R Development Core Team. R: A Language and Environment for Statistical Computing [Internet]. Team RDC, editor. R Found. Stat. Comput. Vienna: R Foundation for Statistical Computing; 2008. p. 409. Available from: http://www.r-project.org.
41. Lanfear R, Calcott B, Ho SYW, Guindon S. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol. 2012;29:1695–701.
42. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 2012;9:772.
43. Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: Proc. Gatew. Comput. Environ. Work. 2010. p. 1–8.
44. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.

3.5 CAPÍTULO CINCO: O transcriptoma de fígado do cascudo (*Pterygoplichthys anisitsi*, Loricariidae): Identificação de expansões em famílias gênicas do defensoma

Neste capítulo, ampliamos o objeto de pesquisa dos transcritos mitocondriais para todo o transcriptoma em *Pterygoplichthys anisitsi*, com objetivo de investigar a evolução de famílias gênicas relacionadas à defesa do organismo contra ameaças químicas. Os resultados aqui descritos foram publicados no artigo intitulado “The liver transcriptome of suckermouth armoured catfish (*Pterygoplichthys anisitsi*, Loricariidae): Identification of expansions in defensome gene families”, na revista Marine Pollution Bulletin, ano 2017; vol.115(1-2):352–61.

Pterygoplichthys é um gênero de peixes cascudos nativos da América do Sul, que invadiu regiões tropicais e subtropicais em todo o mundo. Suas espécies são conhecidas por terem características distintas que podem ter ajudado no rápido estabelecimento em habitats não-nativos, como uma maior resistência aos xenobióticos orgânicos e uma enzima do sistema de biotransformação com fenótipo aberrante. O transcriptoma de fígado de *Pterygoplichthys anisitsi* foi sequenciado e a diversidade de genes candidatos envolvidos na resistência desta espécie a contaminantes orgânicos foi analisada. No total, 66.642 transcritos foram montados. Uma grande diversidade de transcritos que codificam enzimas envolvidas na desintoxicação de xenobióticos, especialmente de citocromos P450 e sulfotransferases, foi encontrada no fígado de *P. anisitsi*, o que poderia contribuir para a resistência desta espécie a xenobióticos orgânicos.

Marine Pollution Bulletin 115 (2017) 352–361

Contents lists available at ScienceDirect

Marine Pollution Bulletin

journal homepage: www.elsevier.com/locate/marpolbul

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

Thiago E. Parente [a,b,e,*], Daniel A. Moreira [a], Maithê G.P. Magalhães [a], Paula C.C. de Andrade [a], Carolina Furtado [c], Brian J. Haas [d], John J. Stegeman [e], Mark E. Hahn [e]

[a] *Laboratório de Toxicologia Ambiental, Escola Nacional de Saúde Pública (ENSP), Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro 21040-900, Brasil*
[b] *Laboratório de Genética Molecular de Microrganismos, Instituto Oswaldo Cruz (IOC), Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro 21040-900, Brasil*
[c] *Unidade de Genômica, Instituto Nacional do Cancer (INCA), Rio de Janeiro 20230-130, Brasil*
[d] *Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA 02142, USA*
[e] *Woods Hole Oceanographic Institution (WHOI), Woods Hole, MA 02543, USA*

ARTICLE INFO

Article history:
Received 4 November 2016
Received in revised form 2 December 2016
Accepted 5 December 2016
Available online 4 January 2017

Keywords:
RNA-Seq
P450
SULT
ABC transporters
Nuclear receptors

ABSTRACT

Pterygoplichthys is a genus of related suckermouth armoured catfishes native to South America, which have 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 sulfotransferases (SULT), nuclear receptors (NR) 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-two 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.

1. Introduction

Pterygoplichthys is a genus of suckermouth armoured catfish (Siluriformes: Loricariidae) native to 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 $\text{mL}\cdot\text{L}^{-1}$ 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 *Pterygoplichthys* 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., 2009, 2011, 2014, 2015). Among Vertebrates, this is an aberrant phenotype of a crucial detoxification enzyme known to take part, for example, in

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

E-mail addresses: parente@ensp.fiocruz.br, tparente@whoi.edu (T.E. Parente).

http://dx.doi.org/10.1016/j.marpolbul.2016.12.012
0025-326X/© 2016 Elsevier Ltd. All rights reserved.

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 Jaboticabal, 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 12,000*g* at 4 °C. The supernatant was transferred to another 1.5 mL RNase-free and DNase-free plastic tube, in which chloroform was added, vigorously mixed, kept on ice for 15 min, and centrifuged for 10 min at 12,000*g* 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 12,000*g* 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 100 bp 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 transcripts 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

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

Total sequencing reads	179,826,191
Reads after QC	177,354,428
Transcripts 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 ≥ (%)	
10×	54
100×	13
Danio rerio coverage ratio (%)	
Average	110
Median	88
Transcripts with coverage ratio ≥ 1	47

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 1,571 bp and an average contig length of 865 bp. 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 13× (average = 646×), and 13% of these transcripts had depth higher than 100×, while 54% were sequenced at a depth higher than 10× (Table 1, Fig. 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, Fig. 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.

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-

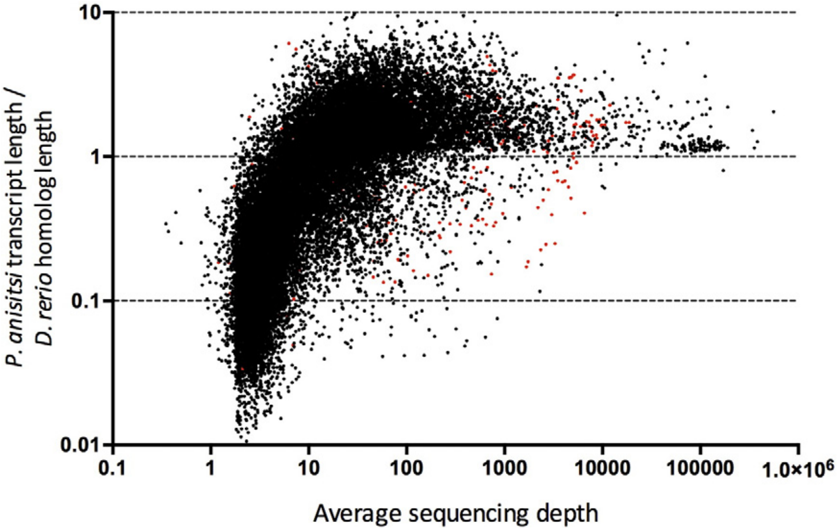

Fig. 1. Ratio of *Pterygoplichthys anisitsi* transcript length to *Danio rerio* homologs plotted against *P. anisitsi* transcript sequencing depth. Defensome transcripts are highlighted in red. The total length of each the 28,190 *P. anisitsi* transcripts with a BLASTX hit against in *Danio rerio* Uniprot 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).

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 7,516×, 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 (Fig. 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 >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 (Moreira et al., 2015, 2016; Magalhães et al., 2016).

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.

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
Glutathione peroxidase	5	6	8	11
Glutathione-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

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 >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 >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 Supporting information Table S4.

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 (Fig. 3). The CYP2 family represented 55% of the CYP transcripts, and was the most abundant family. According to a recent

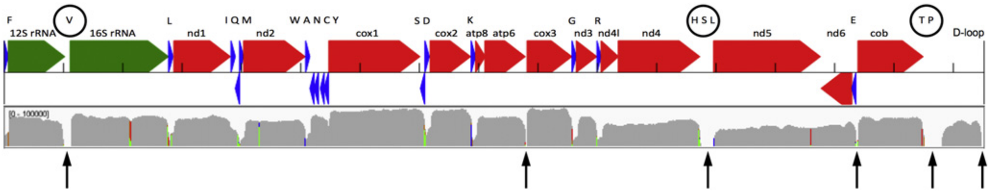

Fig. 2. Mitochondrial genome of *Pterygoplichthys anisitsi*. Green blocks represent ribosomal RNA, red ones protein coding genes, and the ones in blue tRNAs. Black circles indicate the tRNA not sequenced in *P. anisitsi* mtDNA. Black arrows the approximate region of the six gaps, which coincide to areas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

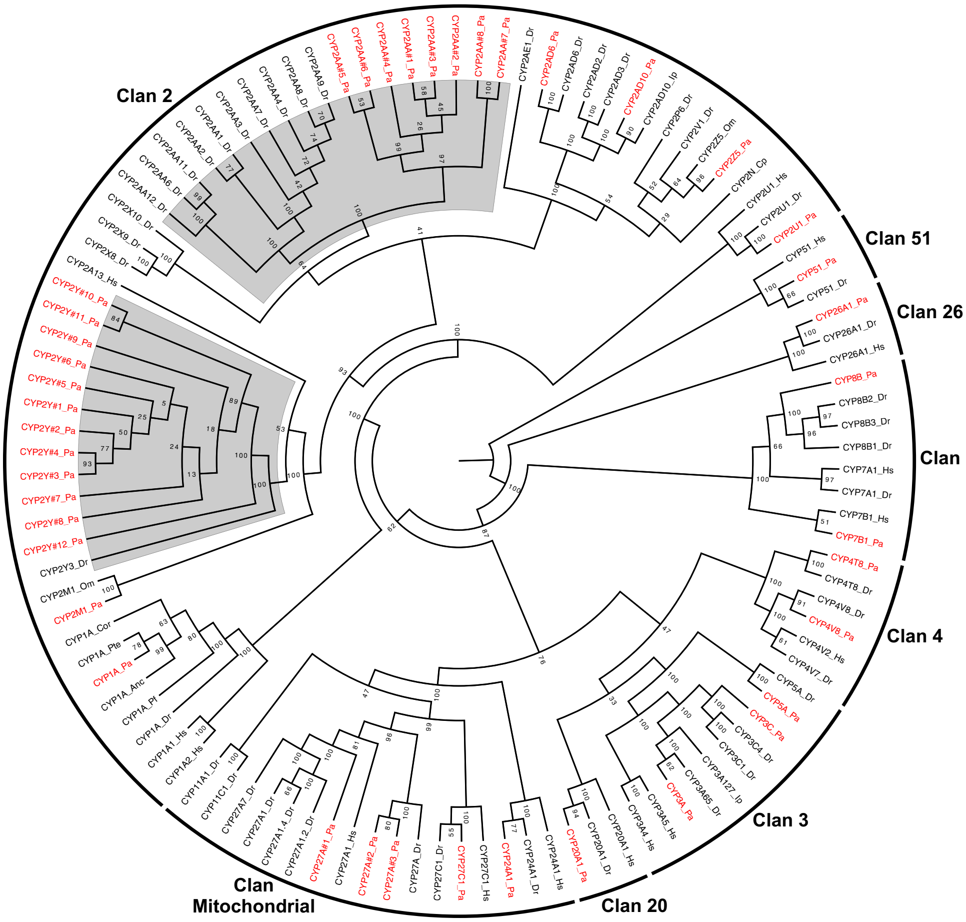

Fig. 3. Maximum-likelihood phylogeny of *Pterygoplichthys anisitsi* cytochromes P450 and homologs. The tree is rooted on CYP51. Sequences of *P. anisitsi* are shown in red. Expansions of CYP2Ys and CYP2AAs are highlighted in gray. Bootstrap values are shown on each node (1000 replicates). The translated 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*; Pf = *Pelteobagrus fulvidraco*. Gis are shown in Supporting Information Table S5.

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 (Fig. 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 (Fig. 3).

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

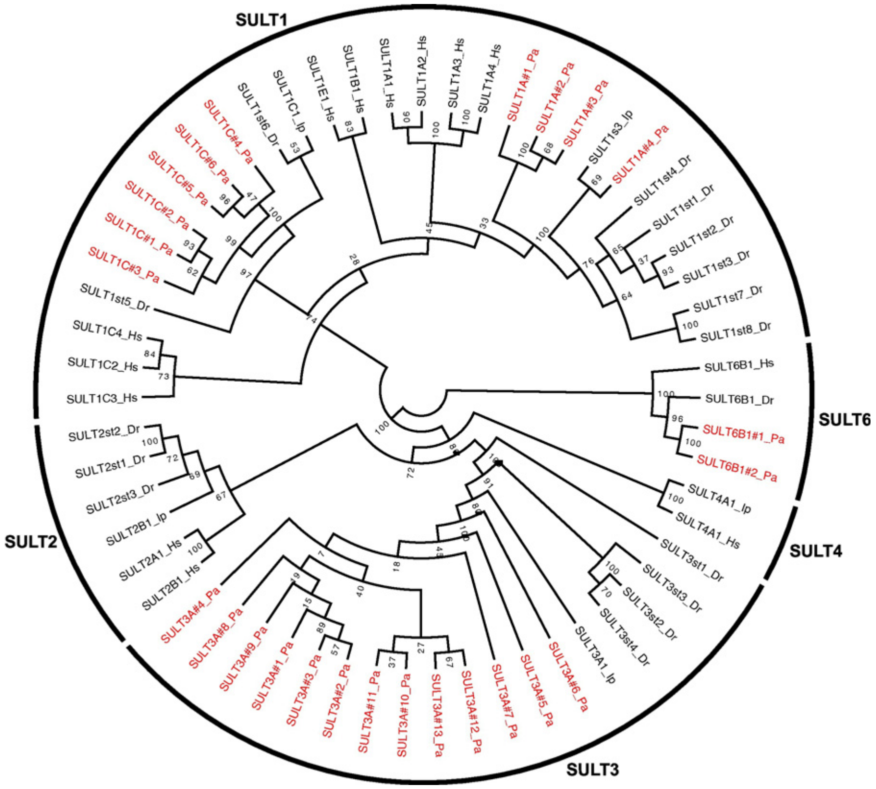

Fig. 4. Unrooted maximum-likelihood phylogeny of *Pterygoplichthys anisitsi* sulfotransferases (SULT) and homologs. Sequences of *P. anisitsi* are shown in red. Bootstrap values are shown on each node (1000 replicates). The translated amino acid sequences were aligned using Muscle and the tree was constructed 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.

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 (Fig. 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 (Fig. 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 (Supporting information Table S4). 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 (Supporting information Table S4, Fig. 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.

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).

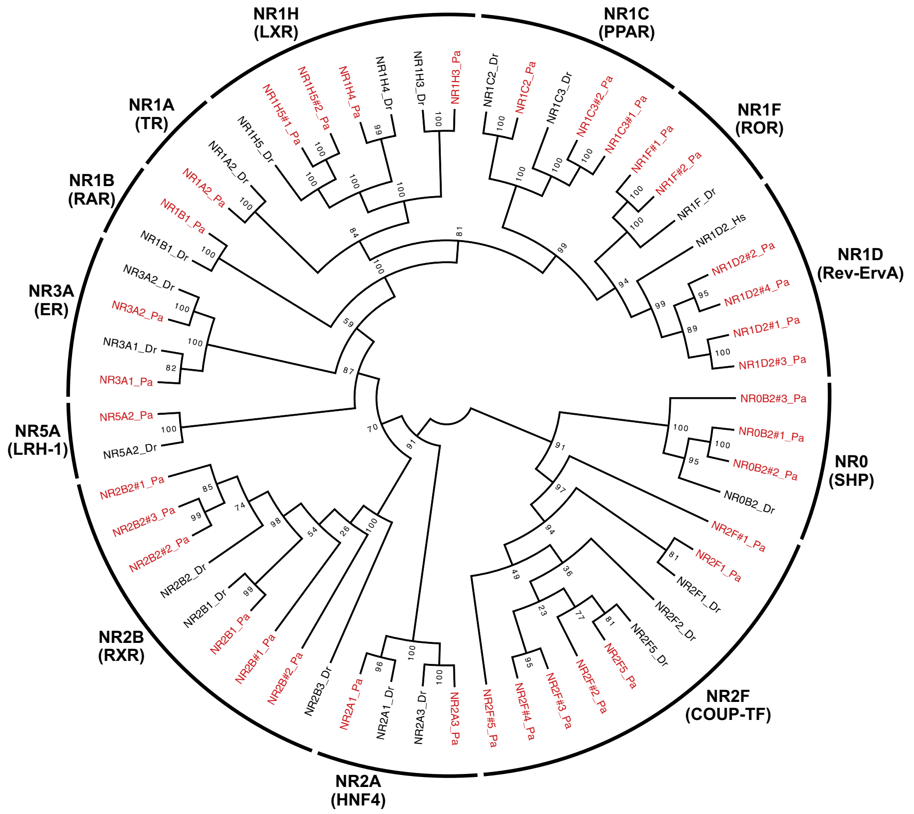

Fig. 5. Unrooted maximum-likelihood phylogeny of *Pterygoplichthys anisitsi* nuclear receptors (NR) and homologs. Sequences of *P. anisitsi* are shown in red. Bootstrap values are shown on each node (1000 replicates). The translated amino acid sequences were aligned using Muscle and the tree was constructed 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.

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 (Fig. 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 (Fig. 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 (Supporting information Table S4).

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 distinct UTR regions was found for ABCB2 and for ABCD3 (Supporting information Table S4). 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 (Fig. 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

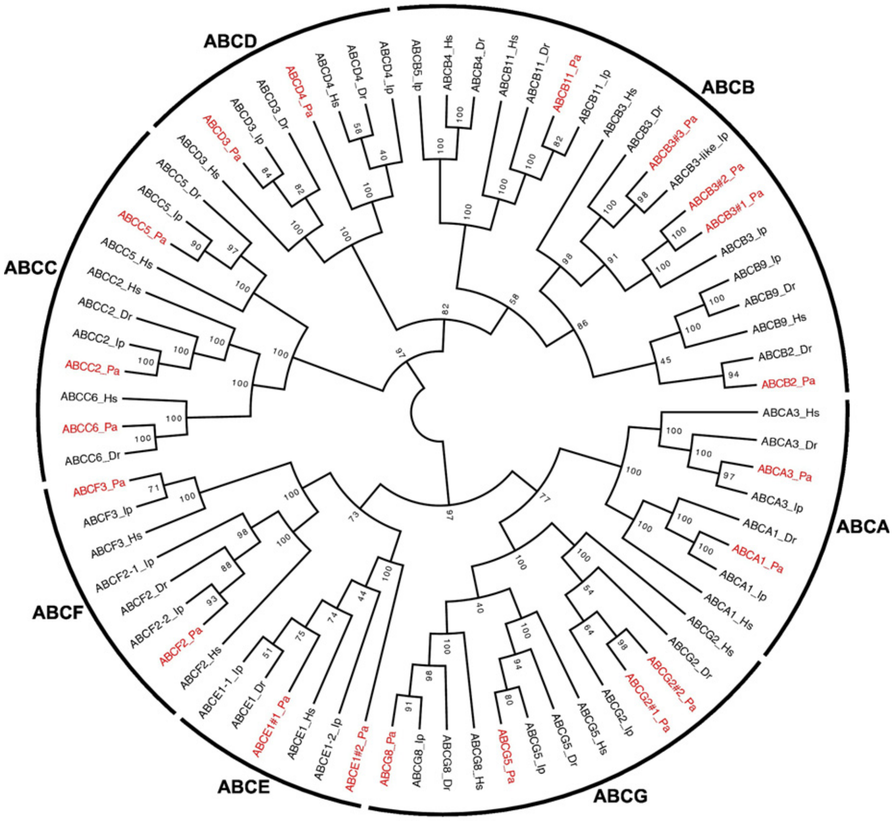

Fig. 6. Unrooted maximum-likelihood phylogeny of *Pterygoplichthys anisitsi* ATP Binding Cassete (ABC) transporters and homologs. Sequences of *P. anisitsi* are shown in red. Bootstrap values are shown on each node (1000 replicates). The translated amino acid sequences were aligned using Muscle and the 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.

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*.

4. Conclusion

The liver transcriptome of *P. anisitsi* was sequenced. From this transcriptome, the *P. anisitsi* mitogenome was assembled, and the diversity 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 fishes.

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 assembly 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.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.marpolbul.2016.12.012.

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. http://dx.doi.org/10.3389/fgene.2014.00348.

Anisimova, M., Gascuel, O., 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst. Biol. 55:539–552. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/10.1093/bioinformatics/btu170.

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

CETESB, 2010. Relatório de qualidade das águas interiores do estado de São Paulo, São Paulo, Brasil. http://aguasinteriores.cetesb.sp.gov.br/publicacoes-e-relatorios/.

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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/10.1093/nar/gkh340.

Felício, A.A., Parente, T.E.M., Maschio, L.R., Nogueira, L., Venancio, L.P.R., Rebelo, M.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. http://dx.doi.org/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. A Ecol. Genet. Physiol. 315A:121–131. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. 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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/10.1093/molbev/mst141.

James, M.O., Ambadapadi, S., 2013. Interactions of cytosolic sulfotransferases with xenobiotics. Drug Metab. Rev. 45:401–414. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. 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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/10.2174/138920008783571819.

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. http://dx.doi.org/10.1186/1471-2164-13-595.

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. http://dx.doi.org/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. http://dx.doi.org/10.3109/19401736.2014.971281.

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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. 368:1612. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/10.1016/j.gene.2011.07.023.

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. http://dx.doi.org/10.1016/j.aquatox.2014.05.021.

Parente, T.E.M., Santos, L.M.F., de Oliveira, A.C.A.X., Torres, J.P.d.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. http://dx.doi.org/10.3395/2317-269x.00278.

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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/10.1371/journal.pone.0109419.

4 DISCUSSÃO

Parafraseando Robbins (2016), a genômica tem se tornado um novo instrumento, o “genomoscópio”, que assim como o telescópio de Galileu, permite olhar para um novo “mundo” ou enxergar com clareza pela primeira vez a dimensão genética da natureza. Isso porque para onde o genomoscópio é “apontado”, novas descobertas surgem. Tais descobertas nem sempre obedecem ao padrão esperado, gerando um desconcerto no *status quo*, estimulando a formulação de perguntas e hipóteses originais e, em última análise, promovendo o avanço da ciência. Entender os novos padrões que emergem das descobertas oriundas do uso do “genomoscópio” exigirá uma mudança de perspectiva sobre alguns dos conceitos e das teorias mais fundamentais da biologia (65).

A obtenção de informações genômicas tornou-se cada vez mais acessível mesmo para organismos não-modelo, entretanto junto com esse avanço tecnológico vem o receio de não conseguir interpretar esse grande volume de informações e transformá-las em conhecimento sobre a biodiversidade e ações práticas (17). De fato, o avanço do sequenciamento de alto desempenho promoveu uma alteração na etapa limitante do processo de geração de conhecimento sobre o patrimônio genético. Antes da disseminação do uso dessas tecnologias, a etapa limitante era a própria produção das sequências, ao passo que hoje o limitante é o processamento dos dados e a interpretação dos resultados. Para lidar com esse desafio, usando a biologia computacional, fomos mudando o foco do nosso “transcriptomoscópio” do genoma mitocondrial de apenas uma para muitas espécies, depois avançamos para os transcritos nucleares de uma espécie para, por fim, analisarmos o todo.

Nesta tese, através do uso do sequenciamento de alto desempenho, geramos e analisamos 40 transcriptomas de 34 espécies da subordem Loricarioidei

(Siluriformes), incluindo 31 espécies da família Loricariidae, a quinta mais diversa entre os vertebrados. O uso do sequenciamento de alto desempenho e de ferramentas computacionais nos permitiu estudar uma maior variedade de informações genéticas, enquanto a aplicação desse método em um espectro de espécies próximas possibilitou a análise e interpretação dos dados obtidos sob a luz da evolução (19).

A mineração dessas informações possibilitou novos usos para as sequências provenientes de RNA-Seq, como a montagem de genomas mitocondriais, descrito e discutido no primeiro capítulo desta tese e aplicado nos três capítulos seguintes. Essa abordagem metodológica inovadora viabilizou a montagem de 31 novos genomas mitocondriais. Contudo, os genomas mitocondriais por si só não trazem muita novidade além da sequência. Seguindo o conselho de Dobzhansky, que disse “nada na biologia faz sentido exceto sob a luz da evolução”, e através da análise comparativa desses genomas mitocondriais, pudemos acessar outras dimensões da biodiversidade de Loricarioidei além da genética, como taxonômica, filogenética e estrutural. Essa visão integrada, por exemplo, possibilitou a sugestão do uso das espécies com a deleção no CSB-D como potenciais modelos para estudos funcionais de replicação e transcrição mitocondrial. Essa característica encontrada exclusivamente em alguns dos peixes estudados nesse trabalho é inédita e exemplifica como nossa perspectiva evolutiva possibilitou a transformação do desconcerto inicial, provocado pela ausência do padrão esperado, em conhecimento. Especula-se que esse padrão tenha sido negligenciado em um estudo abrangente com 248 espécies, compreendendo 42 das 44 ordens de Actinopterygii (66), pois o genoma mitocondrial da espécie *Pterygoplichthys disjunctivus* (Hypostomini:Loricariidae) já estava disponível no banco de dados usado pelos autores, mas não foi considerado no trabalho. É possível que a ausência de genomas mitocondriais de espécies próximas que compartilhem essa deleção, ou seja, a lacuna de informação genética dos loricarídeos, tenha impedido tal achado. Esse fato reforça o potencial do uso de NGS e novas abordagens computacionais para a identificação e descrição de variações genéticas em organismos não modelo, retirando-os da ignorância genômica e pavimentando o caminho para revelar os processos que influenciam a diversidade da vida selvagem.

Os quatro primeiros capítulos foram gerados a partir da mineração dos dados de uma fração, de aproximadamente 2%, das leituras geradas no sequenciamento dos transcriptomas. No entanto, a grande maioria das leituras e diversidade dos transcritos são oriundos do genoma nuclear. Nesses transcritos reside a porção mais valiosa para tentar revelar as forças evolutivas que impulsionaram a diversidade genética, mesmo sem o conhecimento *a priori* dos vínculos entre genótipo e fenótipo. Contudo, para compreender como a seleção age sobre essas sequências é preciso, primeiramente, uma caracterização genética, a lista das partes.

A caracterização genética de transcriptomas de organismos não modelo, em um primeiro momento, é feita através da anotação dos transcritos. Essa anotação geralmente usa métodos baseados em BLAST contra os conjuntos de genes e proteínas de uma espécie modelo como referência para a qual as atribuições de ortologia e função já foram geradas. Quanto mais distante a comparação, menos genes serão anotados, resultando em uma anotação menos informativa (67). Em peixes teleósteos, a detecção de ortólogos é especialmente problemática devido ao “big bang” genômico de sua linhagem após a duplicação genômica no ancestral comum dos teleósteos (a terceira em vertebrados), o que pode conduzir, por exemplo, à confusão entre ortólogos e parálogos, e ainda a inexistência de ortologia devido perda gênica (30). Ter em mãos os 40 transcriptomas é um recurso fundamental para fazer a correta inferência de ortologia, pois as espécies são bem mais próximas entre si do que com os organismos modelos disponíveis. A partir disso, o “transcriptomoscópio” pode ser usado não só para listar as partes de espécies isoladas, mas sob uma ótica evolutiva acessar novas informações sobre as adaptações desses organismos.

A anotação do transcriptoma de *P. anisitsi* foi essencial para descrever as variações encontradas nas famílias gênicas do seu defensoma e providenciar evidências que subsidiam a resistência dessa espécie a contaminantes orgânicos (49) e da sua capacidade de invadir novos ambientes (56). A ausência de atividade EROD em *P. anisitsi* (59) foi só uma das pistas que nos fez explorar sua diversidade genética. Essa gota de informação sobre as adaptações dessa espécie em um oceano de novas possibilidades, nos impulsiona a buscar nessa família e em toda biodiversidade novos candidatos à espécies modelo, que sem dúvida existem. De fato, usando apenas 2% dos dados totais encontramos importantes particularidades na análise comparativa dos genomas mitocondriais, dessa forma, também esperamos descobrir novos padrões ao analisarmos os 98% restantes. Esse grande volume de sequências, até então, desconhecidas representa um grande desafio metodológico para que genes conhecidos possam ser identificados ou até novas funções possam ser descritas. A inferência das funções desses novos transcritos, aumentará o conhecimento da biodiversidade molecular e o repertório de sequências descritas ajudando a desvendar parte da matéria escura biológica (64,68), e possibilitando a formulação de novas hipóteses.

Após a descrição do transcriptoma de *P. anisitsi*, estamos ampliando a análise dos transcriptomas para todas as espécies desse estudo. Usando uma abordagem “top-down”, poderemos descobrir componentes novos e essenciais sobre suas adaptações e responder perguntas, como por exemplo, se as expansões inéditas e independentes de CYPs existem em toda a família, se existe alguma relação de ortologia entre elas e testar hipóteses de seleção positiva nesses ortólogos ao longo de toda a família.

Retirar a quinta família mais diversa entre os vertebrados da ignorância genômica é um passo fundamental para o entendimento da relação entre o genótipo e fenótipo dessas espécies. Apesar da maioria dos fenótipos ser controlada por muitos genes e de muitos fenótipos poderem interferir na dinâmica ecológica, o número de variantes em estudos genômicos é tão grande, que identificar genes que possam ser selecionados e associar esses genes-candidatos a alguns fenótipos é um ponto de partida importante para uma exploração futura (69) e para o melhor entendimento da nossa biodiversidade.

Descrever as partes da biodiversidade brasileira é especialmente relevante no momento atual que entramos no Antropoceno, onde uma das características mais marcantes é a perda de biodiversidade (6). Os resultados aqui gerados aumentam nosso conhecimento sobre a biodiversidade endêmica do nosso país em várias de suas dimensões e, dessa forma, podem fundamentar pesquisas para conservação e controle populacional de espécies ameaçadas de extinção (58) e espécies invasoras (56). Nossos resultados também contribuem para alcançar as metas nacionais relacionadas aos objetivos da Agenda 2030 da ONU, agenda esta que faz parte do plano estratégico da Fiocruz. Por fim, o acesso aos recursos genéticos disponibilizados e analisados nessa tese constitui um passo em direção à melhor compreensão dos mecanismos que unem o bem-estar humano, a saúde dos outros seres vivos e a integridade dos ecossistemas.

5 CONCLUSÕES

Esta tese é o primeiro trabalho a usar sequenciamento de ácidos nucléicos de alto desempenho para o estudo de espécies da subordem Loricarioidei. As análises desses dados transcriptômicos revelaram:

- Genomas mitocondriais completos, os níveis de expressão dos transcritos mitocondriais, o padrão de pontuação da edição pós-transcricional e heteroplasmias.
- Características estruturais apomórficas, como a inserção nucleotídica nos mitogenomas dos calictídeos e a deleção parcial no final 3' do Bloco de Sequência Conservada (CSB) D da região controle em um clado monofilético de loricarídeos.
- Uma filogenia que aumenta a confiança e corrobora em grande parte as relações atualmente aceitas, destacando mudanças na topologia da subfamília Hypoptopomatinae.
- Uma grande diversidade de transcritos que codificam enzimas envolvidas na biotransformação de xenobióticos, especialmente de citocromos P450 e sulfotransferases, encontrada no transcriptoma hepático de *P. anisitsi*.

Os recursos transcriptômicos produzidos por esse estudo irão compor bancos de dados públicos, reduzindo a lacuna de informação genética existente e fornecendo a base para novos estudos sobre a história evolutiva desses peixes e potenciais descobertas de genes e transcritos genuinamente novos.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Zalasiewicz J, Williams M, Haywood A, Ellis M. The Anthropocene: a new epoch of geological time? Philos Trans R Soc A Math Phys Eng Sci. 2011;

2. Pereira HM, Navarro LM, Martins IS. Global Biodiversity Change: The Bad, the Good, and the Unknown. Annu Rev Environ Resour. 2012;37(1):25–50.

3. Waters CN, Zalasiewicz J, Summerhayes C, Barnosky AD, Poirier C, Gałuszka A, et al. The Anthropocene is functionally and stratigraphically distinct from the Holocene. Science. 2016.

4. Steffen W, Grinevald J, Crutzen P, McNeill J. The Anthropocene: conceptual and historical perspectives. Philos Trans R Soc A Math Phys Eng Sci. 2011;

5. Ceballos G, Ehrlich PR, Barnosky AD, Garcia A, Pringle RM, Palmer TM. Accelerated modern human-induced species losses: Entering the sixth mass extinction. Sci Adv. 2015;1(5):e1400253–e1400253.

6. Dirzo R, Young HS, Galetti M, Ceballos G, Isaac NJB, Collen B. Defaunation in the Anthropocene. Science. 2014;345(6195):401–6.

7. Naeem S, Prager C, Weeks B, Varga A, Flynn DFB, Griffin K, et al. Biodiversity as a multidimensional construct: a review, framework and case study of herbivory's impact on plant biodiversity. Proc R Soc B Biol Sci. 2016;283(1844):20153005.

8. Seddon N, Mace GM, Naeem S, Tobias JA, Pigot AL, Cavanagh R, et al. Biodiversity in the Anthropocene : prospects and policy. Proc R Soc B Biol Sci. 2016;1–9.

9. Naeem S, Duffy JE, Zavaleta E. The functions of biological diversity in an age of extinction. Science (80-). 2012;336(6087):1401–6.

10. Naeem S, Chazdon R, Duffy JE, Prager C, Worm B. Biodiversity and human

well-being: an essential link for sustainable development. Proc R Soc B Biol Sci. 2016 Dec 14;283(1844):20162091.

11. Van Helden PD, Van Helden LS, Hoal EG. One world, one health. EMBO Rep. 2013;14(6):497–501.

12. Butchart SHM, Walpole M, Collen B, van Strien A, Scharlemann JPW, Almond REA, et al. Global biodiversity: indicators of recent declines. Science. 2010;328(5982):1164–8.

13. Tittensor DP, Walpole M, Hill SLL, Boyce DG, Britten GL, Burgess ND, et al. A mid-term analysis of progress toward international biodiversity targets. Science. 2014;346(6206):241–4.

14. Feld CK, Martins da Silva P, Paulo Sousa J, de Bello F, Bugter R, Grandin U, et al. Indicators of biodiversity and ecosystem services: a synthesis across ecosystems and spatial scales. Oikos. 2009;118(12):1862–71.

15. Pereira HM, Ferrier S, Walters M, Geller GN, Jongman RHG, Scholes RJ, et al. Essential biodiversity variables. Science. 2013;339(6117):277–8.

16. Lyashevska O, Farnsworth KD. How many dimensions of biodiversity do we need? Ecol Indic. 2012;18:485–92.

17. Shafer ABA, Wolf JBW, Alves PC, Bergström L, Bruford MW, Brännström I, et al. Genomics and the challenging translation into conservation practice. Trends Ecol Evol. 2015;30(2):78–87.

18. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333–51.

19. Stapley J, Reger J, Feulner PGD, Smadja C, Galindo J, Ekblom R, et al. Adaptation genomics : the next generation. Trends Ecol Evol. 2010;25(12):705–12.

20. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-

generation sequencing technology. Trends Genet. 2014;30(9):418–26.

21. Kitano H. Systems Biology: A Brief Overview. Science (80-). 2002;295(5560):1662–4.

22. Raes J, Bork P. Molecular eco-systems biology: towards an understanding of community function. Nat Rev Microbiol. 2008;6.

23. Ideker T, Galitski T, Hood L. A New Approach to Decoding Life: Systems Biology. Annu Rev Genomics Hum Genet. 2001;2(1):343–72.

24. Kitano H. Computational systems biology. Nature. 2002;420(6912):206–10.

25. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, et al. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 2014;42.

26. Ellegren H. Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol. 2014;29(1):51–63.

27. McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol Phylogenet Evol. 2013;66(2):526–38.

28. Grueber CE. Comparative genomics for biodiversity conservation. Comput Struct biotechology J. 2015;13:370–5.

29. Larsen PA, Hayes CE, Williams C V., Junge RE, Razafindramanana J, Mass V, et al. Blood transcriptomes reveal novel parasitic zoonoses circulating in Madagascar’s lemurs. Biol Lett. 2016;12(1).

30. Braasch I, Peterson SM, Desvignes T, McCluskey BM, Batzel P, Postlethwait JH. A new model army: Emerging fish models to study the genomics of vertebrate Evo-Devo. J Exp Zool Part B Mol Dev Evol. 2015;324(4):316–41.

31. The World Conservation Union. Summary Statistics for Globally Threatened Species. Table 1: Numbers of threatened species by major groups of

organisms (1996–2017) [Internet]. IUCN Red List of Threatened Species, 2017.3. 2017 [cited 2018 Jan 15]. Available from: http://cmsdocs.s3.amazonaws.com/summarystats/2017-3_Summary_Stats_Page_Documents/2017_3_RL_Stats_Table_1.pdf

32. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The zebrafish reference genome sequence and its relationship to the human genome. Nature. 2013;496(7446):498–503.

33. Hollert H, Keiter SH. *Danio rerio* as a model in aquatic toxicology and sediment research. Environ Sci Pollut Res. 2015;22(21):16243–6.

34. Lieschke GJ, Currie PD. Animal models of human disease: zebrafish swim into view. Nat Rev Genet. 2007;8(5):353–67.

35. Schartl M. Beyond the zebrafish: diverse fish species for modeling human disease. Dis Model Mech. 2014;7(2):181–92.

36. Albertson RC, Cresko W, Detrich HW, Postlethwait JH. Evolutionary mutant models for human disease. Trends Genet. 2009;25(2):74–81.

37. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res. 2017;45(D1):D37–42.

38. Nelson JS. Fishes of the world. John Wiley; 2006. 601 p.

39. Brosse S, Beauchard O, Blanchet S, Dürr HH, Grenouillet G, Hugueny B, et al. Fish-SPRICH: A database of freshwater fish species richness throughout the World. Hydrobiologia. 2013;700(1):343–9.

40. Eschmeyer WN, Fong JD. SPECIES BY FAMILY/SUBFAMILY. [Internet]. 2018 [cited 2018 Mar 25]. Available from: http://researcharchive.calacademy.org/research/ichthyology/catalog/SpeciesByFamily.asp

41. Nakatani M, Miya M, Mabuchi K, Saitoh K, Nishida M. Evolutionary history of

Otophysi (Teleostei), a major clade of the modern freshwater fishes: Pangaean origin and Mesozoic radiation. BMC Evol Biol. 2011;11(1):177.

42. Cramer CA, Liedke AMR, Bonatto SL, Reis RE. The phylogenetic relationships of the Hypoptopomatinae and Neoplecostominae (Siluriformes: Loricariidae) as inferred from mitochondrial cytochrome c oxidase I sequences. Bull Fish Biol. 2008;9:51–9.

43. Roxo FF, Albert JS, Silva GSC, Zawadzki CH, Foresti F, Oliveira C. Molecular phylogeny and biogeographic history of the armored neotropical catfish subfamilies hypoptopomatinae, neoplecostominae and otothyrinae (siluriformes: loricariidae). PLoS One. 2014;9(8):e105564.

44. Lujan NK, Armbruster JW, Lovejoy N, López-fernández H. Multilocus molecular phylogeny of the suckermouth armored catfishes (Siluriformes: Loricariidae) with a focus on subfamily Hypostominae. Mol Phylogenet Evol. 2015;82:269–88.

45. Lujan NK, Winemiller KO, Armbruster JW. Trophic diversity in the evolution and community assembly of loricariid catfishes Trophic diversity in the evolution and community assembly of loricariid catfishes. BMC Evol Biol. 2012;12(1).

46. Lujan NK, German DP, Winemiller KO. Do wood-grazing fishes partition their niche?: Morphological and isotopic evidence for trophic segregation in Neotropical Loricariidae. Funct Ecol. 2011;25(6):1327–38.

47. da Cruz AL, da Silva HR, Lundstedt LM, Schwantes AR, Moraes G, Klein W, et al. Air-breathing behavior and physiological responses to hypoxia and air exposure in the air-breathing loricariid fish, *Pterygoplichthys anisitsi*. Fish Physiol Biochem. 2013;39(2):243–56.

48. Harter TS, Shartau RB, Baker DW, Jackson DC, Val AL, Brauner CJ. 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. 2014;184(6):709–18.

49. Arantes Felício A, Martins Parente TE, Regina Maschio L, Nogueira L, Rodrigues Venancio LP, de Freitas Rebelo M, et al. 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. 2015;115:26–32.

50. Villalba-Villalba AG, Ramírez-Suárez JC, Valenzuela-Soto EM, Sánchez GG, Ruiz GC, Pacheco-Aguilar R. Trypsin from viscera of vermiculated sailfin catfish, *Pterygoplichthys disjunctivus*, Weber, 1991: Its purification and characterization. Food Chem. 2013;141(2):940–5.

51. Bijukumar A, Smrithy R, Sureshkumar U, George S, George S. Invasion of South American suckermouth armoured catfishes *Pterygoplichthys* spp. (Loricariidae) in Kerala, India - a case study. J Threat Taxa. 2015;7(3):6987–95.

52. Chavez JM, Paz RMD La, Manohar SK, Pagulayan RC, Vi JRC. New Philippine record of south american sailfin catfishes (Pisces: Loricariidae). Zootaxa. 2006;1109(1):57–68.

53. Gibbs MA, Kurth BN, Bridges CD. Age and growth of the loricariid catfish *Pterygoplichthys disjunctivus* in Volusia Blue Spring, Florida. Aquat Invasions. 2013;8(2):207–18.

54. Jumawan JC, Herrera AA. Histological and ultrastructural characteristics of the testis of the invasive suckermouth sailfin catfish *Pterygoplichthys disjunctivus* (Siluriformes: loricariidae) from Marikina River, Philippines. Tissue Cell. 2015;47(1):17–26.

55. Nico LG, Loftus WF, Reid JP. Interactions between non-native armored

suckermouth catfish (Loricariidae: *Pterygoplichthys*) and native Florida manatee (Trichechus manatus latirostris) in artesian springs. Aquat Invasions. 2009;4(3):511–9.

56. Capps KA, Flecker AS. Invasive aquarium fish transform ecosystem nutrient dynamics. Proc R Soc B Biol Sci. 2013;280(1769):20131520–20131520.

57. IBAMA. Diagnóstico Geral das Práticas de Controle Ligadas a Exploração , Captura , Uso De Peixes Para Fins Ornamentais e de Aquariofilia. 2008;

58. Magalhães MGP, Moreira DA, Furtado C, Parente TE. The mitochondrial genome of *Hypancistrus zebra* (Isbrücker &amp; Nijssen, 1991) (Siluriformes: Loricariidae), an endangered ornamental fish from the Brazilian Amazon. Conserv Genet Resour. 2017;9(2):319–24.

59. Parente TEM, De-Oliveira ACAX, Beghini DG, Chapeaurouge DA, Perales J, Paumgartten FJR. 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 Part C Toxicol Pharmacol. 2009;150(2):252–60.

60. Parente TEM, Rebelo MF, Da-Silva ML, Woodin BR, Goldstone J V., Bisch PM, et al. Structural features of cytochrome P450 1A associated with the absence of EROD activity in liver of the loricariid catfish *Pterygoplichthys* sp. Gene. 2011;489(2):111–8.

61. Parente TEM, Urban P, Pompon D, Rebelo MF. Altered substrate specificity of the *Pterygoplichthys* sp. (Loricariidae) CYP1A enzyme. Aquat Toxicol. 2014;154:193–9.

62. Parente TE, Santos LMF Dos, Oliveira ACAX De, Torres JPDM, Araújo FG, Delgado IF, et al. 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. 2015;3(1):88–92.

63. Goldstone HMH, Stegeman JJ. A Revised Evolutionary History of the CYP1A Subfamily: Gene Duplication, Gene Conversion, and Positive Selection. J Mol Evol. 2006;62(6):708–17.

64. Robbins RJ, Krishtalka L. Advances in biodiversity : metagenomics and the unveiling of biological dark matter. Stand Genomic Sci. 2016;1–17.

65. Setubal JC. Desconcertos na ciência. Rev Bras Psicanálise. 2016;50(3):145–52.

66. Satoh TP, Miya M, Mabuchi K, Nishida M. Structure and variation of the mitochondrial genome of fishes. BMC Genomics. 2016;17(1):719.

67. Garcia-Reyero N, Perkins EJ. Systems biology: Leading the revolution in ecotoxicology. Environ Toxicol Chem. 2011;30(2):265–73.

68. Ponting CP, Belgard TG. Transcribed dark matter: meaning or myth? Hum Mol Genet. 2010;19(R2):R162–8.

69. Rudman SM, Barbour MA, Csilléry K, Gienapp P, Guillaume F, Hairston Jr NG, et al. What genomic data can reveal about eco-evolutionary dynamics. Nat Ecol Evol. 2018;2(1):9–15.

ANEXOS

ANEXO A

BASES GENÉTICAS DA DIVERSIDADE, ADAPTAÇÕES E EVOLUÇÃO DE LORICARIIDE REVELADAS PELO SEQUENCIAMENTO DE TRANSCRIPTOMAS

INTRODUÇÃO

As radiações adaptativas são conhecidas pela rápida diversificação morfológica e de espécies em resposta a oportunidades ecológicas (1). Várias teorias evolutivas preveem que as mudanças em características fenotípicas relevantes, do ponto de vista ecológico, sejam positivamente associadas com a taxa na qual novas espécies surgem. A família Loricariidae, com mais de 900 espécies (2), exemplifica uma rica radiação de espécies da região neotropical, sendo a família mais diversa da sub-ordem Loricarioidei e a quinta mais diversa entre os vertebrados. Os integrantes dessa família são facilmente reconhecidos pelo seu corpo coberto por placas dérmicas ossificadas e por sua boca, posicionada na região ventral e adaptada para sucção e raspagem de superfícies submersas. Esses peixes exibem variações extraordinárias na morfologia de suas mandíbulas (3), apesar de apresentarem pouca variação em suas dietas, sendo em sua maioria composta por detritos e algas. Entretanto, Lujan *et al.*, 2012, mostrou que existe um particionamento trófico ao longo de gradientes nutricionais resultando em uma radiação ecológica, consistente com a diversidade morfológica mandibular e a diversificação evolutiva entre os loricarídeos (4).

Nesse sentido, o clado da *Peckoltia* (sensu Lujan, 2015) se destaca pela grande diversidade morfológica, sendo o clado mais rico em gêneros da subfamília Hypostominae. Atualmente, esse clado possui nove gêneros e 65 espécies (5), distribuídas em grande parte do norte da América do Sul, com tamanho corporal variando de 40 a 520 mm (Tabela 1) (6,7). Os organismo deste clado são, em sua maioria, algívoros-detritívoros, contudo existem especializações em seus nichos tróficos, como por exemplo, o gênero *Panaqolus* que é especializado em comer madeira (8) e o gênero *Hypancistrus* que é invertívoro (come invertebrados).

Os avanços recentes em tecnologias de sequenciamento de ácidos nucléicos têm oferecido a oportunidade de gerar grandes quantidades de dados genômicos ou transcriptômicos sobre quase qualquer organismo, mesmo daqueles que atualmente não possuem um genoma de referência (9). Dessa forma, retira-se esses organismos da ignorância genômica, transformando-os em um novo exército de modelos e gerando uma ferramenta inestimável para entender as origens e a manutenção da biodiversidade (10). Essas novas abordagens alteraram a etapa limitante de produção do conhecimento da geração de dados brutos para a análise desses dados, permitindo concentrar os esforços na interpretação de importantes questões da biologia evolutiva, como por exemplo, porque algumas linhagens são mais diversas que outras (11).

Neste estudo realizamos análises comparativas entre 40 transcriptomas de 34 espécies com o objetivo de investigar os mecanismos genéticos da biodiversidade de loricarídeos, como expansões de famílias gênicas, filogenia, taxas evolutivas e seleção positiva que fomentaram a diversidade genética, fenotípica e ecológica desse rico grupo de organismos.

MATERIAL E MÉTODOS

Amostragem taxonômica, extração de RNA e sequenciamento

Além das 29 espécies da família Loricariidae e duas espécies da família Callichthyidae analisadas no Capítulo Quatro, incluímos mais três espécies (dois loricarídeos e um calictídeo) para as análises feitas neste Anexo (Tabela 2), totalizando 40 transcriptomas de 34 espécies de Loricarioidei. A lista com as espécies, os códigos de campo, localização geográfica da coleta ou sua origem e o número de depósito em coleção biológica estão detalhados na Tabela 2.

Após a coleta dos espécimes, o RNA total foi extraído do tecido hepático seguindo o método do fenol/clorofórmio (TRIzol Reagent, Thermo Fisher Scientific). Após a extração, o RNA total foi inicialmente quantificado por espectrofotometria (BioDrop DUO), e sua qualidade e quantidade foram avaliadas adicionalmente com o kit Bioanalyzer RNA 6000 nano (Agilent). Apenas preparações com “RNA Integrity Number” (RIN) acima de 6.0 foram utilizadas para a preparação da biblioteca de DNA complementar (cDNA).

Tabela 1: Diversidade de gêneros e espécies dos clados Hypostomini e da *Peckoltia*. Também são mostradas informações sobre a ocorrência, variação de tamanho e nicho trófico das espécies desses clados usadas neste trabalho.

Tribo	Gênero	Espécie	Número de gêneros	Número de espécies	Ocorrência	Tamanho SL** (mm)	Nicho trófico	Referência
Hypostomini			2	165				(5,12)
	Pterygoplichthys			15				(5)
		Pterygoplichthys sp.						
		Pterygoplichthys pardalis			Bacia do Rio Amazonas	76.1–422.9	Algívoro–detritívoro	(5,12,13)
		*Pterygoplichthys anisitsi**			Bacias dos Rios Paraguai, médio Paraná, Bermejo e Uruguai	164-471	Algívoro–detritívoro	(5,12,14)
	Hypostomus			150				(5)
		Hypostomus sp.						
		Hypostomus affinis			Bacia do Rio Paraíba do Sul	120-433***	Algívoro–detritívoro	(5,12,15)
		Hypostomus cf. *plecostomus*						
Clado da *Peckoltia*			9	65				(5,12)
	Aphanotorulus			6				(5)
		Aphanotorulus emarginatus			Alto Orinoco, Bacia do Rio Essequibo e Bacia do baixo Amazonas	46.5–357.0	Algívoro–detritívoro	(5,7,12)
	Isorineloricaria			4				(5)
	Spectracanthicus			6				(5)
	Peckoltichthys			1				(5)
	Hypancistrus			9				(5)
		Hypancistrus zebra			Bacia do Rio Xingú	39.9-50	Onívoro-invertívoro	(5,6,12)
	Ancistomus			5				(5)
		Ancistomus snethlageae			Bacia do Rio Tapajós	140	Onívoro	(5,12,16)
	Scobinancistrus			2				(5)
	Panaqolus			11				(5)
		Panaqolus sp.					Come madeira	(12)
	Peckoltia			21				(5)
		Peckoltia furcata			Alto Amazonas, Rios Marañon e Ucayali	75.9-153.5	Algívoro–detritívoro	(5,12,17)

* Sinônimo de *Pterygoplichthys ambrosettii*.

** “Standard Length”: A distância da extremidade anterior do peixe até a base da cauda.

*** “Total Length”: A distância da extremidade anterior do peixe até a ponta da cauda.

Tabela 2: Lista das espécies amostradas, seus números de identificação de campo, coordenadas geográficas e números de depósito em coleção biológica. Os vouchers foram depositados nas coleções ictiológicas do Museu Nacional da Universidade Federal do Rio de Janeiro (MNRJ) e do Instituto Nacional de Pesquisas da Amazônia (INPA).

Espécie	ID Campo	Localização	Voucher
Hemipsilichthys nimius	TP189	23°12′35.2″S 44°47′40.7″W (RJ)	MNRJ43650
Rineloricaria cf. *lanceolata*	sp16.3	Espécime de aquário (PA)	MNRJ43638
Rineloricaria sp.	TP144	22°31′06,3″S 42°53′55,5″W (RJ)	MNRJ42544
Loricaria cataphracta	TP181	3°10′50.9″S 59°54′09.3″W (AM)	MNRJ43629
Loricarichthys castaneus	TP029	21°13′08.7″S 41°18′37.7″W (RJ)	MNRJ41545
Loricarichthys platymetopon	TP179	3°10′50.9″S 59°54′09.3″W (AM)	MNRJ43627
Hypoptopoma incognitum	TP171	3°09′36.0″S 59°55′12.0″W (AM)	MNRJ43421
Otocinclus cf. *hoppei*	sp10.7	Espécime de aquário (PA)	MNRJ43634
Pareiorhaphis garbei	TP009	22°32′03.4″S 43°02′18.7″W (RJ)	MNRJ41511
Schizolecis guntheri	TP006	22°32′03.4″S 43°02′18.7″W (RJ)	MNRJ41510
Parotocinclus maculicauda	TP011	22°36′01.6″S 43°05′30.1″W (RJ)	MNRJ41523
Hisonotus thayeri	TP128	21°32′14.6″S 42°06′54.8″W (RJ)	MNRJ42481
Kronichthys heylandi	8505	23°12′35.2″S 44°47′40.7″W (RJ)	MNRJ42082
Neoplecostomini gen. n.	TP065	20°01′35.3″S 40°36′33.3″W (ES)	MNRJ41921
Neoplecostomus microps	TP088	22°20′01.7″S 44°32′34.3″W (RJ)	MNRJ41752
Dekeyseria amazonica	TP165	3°09′36.0″S 59°55′12.0″W (AM)	MNRJ43618
Ancistrus multispinis	TP003	22°32′03.4″S 43°02′18.7″W (RJ)	MNRJ41509
Ancistrus sp. 1	13,3	Espécime de aquário (PA)	MNRJ42890
Ancistrus sp. 2a	13.10	Espécime de aquário (PA)	MNRJ42890
Ancistrus sp. 2b	13,11	Espécime de aquário (PA)	MNRJ42890
Baryancistrus xanthellus	sp11.19	Espécime de aquário (PA)	Ausente
Pterygoplichthys sp.	sp2	Espécime de aquário (RJ)	MNRJ43652
Pterygoplichthys pardalis	TP154	3°09′36.0″S 59°55′12.0″W (AM)	MNRJ43607
Pterygoplichthys anisitsi	A3	Jaboticabal (SP)	Ausente
Pterygoplichthys anisitsi	B3	Jaboticabal (SP)	Ausente
Pterygoplichthys anisitsi	B6	Jaboticabal (SP)	Ausente
Hypostomus sp.	sp12.6	Espécime de aquário (PA)	MNRJ43635
Hypostomus affinis	TP147	22°48′42.6″S 43°37′42.8″W (RJ)	MNRJ43256
Hypostomus cf. *plecostomus*	TP164	3°09′36.0″S 59°55′12.0″W (AM)	MNRJ43617
Aphanotolurus emarginatus	TP184	3°10′50.9″S 59°54′09.3″W (AM)	MNRJ43631
Hypancistrus zebra	TP166	Doação*	INPA 46655
Ancistomus snethlageae	sp17.2	Espécime de aquário (PA)	MNRJ43639
Peckoltia furcata	sp15.2	Espécime de aquário (PA)	MNRJ43637
Panaqolus sp.	sp4	Espécime de aquário (RJ)	MNRJ43654
Corydoras schwartzi	TP177	Espécime de aquário (AM)	MNRJ43625
Corydoras nattereri	TP020	22°36′01.6″S 43°05′30.1″W (RJ)	MNRJ41520
Corydoras nattereri	TP021	22°36′01.6″S 43°05′30.1″W (RJ)	MNRJ41520
Corydoras nattereri	TP022	22°36′01.6″S 43°05′30.1″W (RJ)	MNRJ41520
Hoplosternum littorale	TP149	22°48′42.6″S 43°37′42.8″W (RJ)	MNRJ43258
Hoplosternum littorale	TP156	3°09′36.0″S 59°55′12.0″W (AM)	MNRJ43609

*doado pelo Dr. Jansen Zuanon do Instituto Nacional de Pesquisa da Amazônia (INPA), espécime oriundo de uma apreensão feita pela Polícia Federal Brasileira.

As bibliotecas de cDNA foram preparadas usando o kit TruSeq RNA Sample v.2 (Illumina), seguindo estritamente as recomendações do fabricante. Foi verificada a qualidade de todas as bibliotecas usando o kit Bioanalyzer DNA 1000 (Agilent). A quantificação das bibliotecas foi realizada por PCR quantitativo usando o kit de quantificação de biblioteca para Illumina com Primers Revisados-SYBR (Kapa Biosystems). Cada biblioteca foi identificada utilizando adaptadores específicos e os "pools" de bibliotecas foram agrupados usando o kit TrueSeq PE Cluster v3 para cBot (Illumina). O sequenciamento de 100 pb das duas extremidades de cada fragmento ("paired-end") foi realizado em um sequenciador HiSeq2500 usando o kit True-Seq SBS v.3 (Illumina).

Montagem *de novo* e anotação dos transcriptomas

Os dados brutos foram decompostos usando o software BCL2FASTQ (Illumina). As sequências dos adaptadores e as de baixa qualidade foram removidas das leituras utilizando o Trimmomatic (18) e sua qualidade foi avaliada usando FastQC (Babraham Bioinformatics).

As leituras com valores médios de Phred igual ou superior a 30 foram usadas para montagem *de novo* dos transcriptomas usando os parâmetros padrão do programa Trinity v2.0.6 (19,20). O Transdecoder v3.0.0 (https://github.com/TransDecoder/TransDecoder) foi usado para identificar as regiões codificantes ("open reading frames" - ORFs) candidatas dentro das sequências dos transcritos montados. As sequências proteicas preditas foram anotadas utilizando o algoritmo BLASTP v2.4.0+ (e-value < 1e-10) contra a base de dados UniprotKB de *Danio rerio* e SwissProt de *Homo sapiens*. O melhor alinhamento de cada sequência foi usado para sua anotação.

Identificação de ortólogos

O programa Orthofinder v1.0.3 (21) foi usado para inferir grupos de sequências homólogas. O primeiro passo desse programa usa o método de melhor alinhamento recíproco com o algoritmo BLASTP para computar as pontuações de similaridade entre as sequências de múltiplas espécies. Depois disso, o programa normaliza as pontuações do BLAST para o comprimento da sequência e, em seguida, usa o algoritmo de agrupamento MCL para identificar grupos de sequências altamente similares (ortogrupos) dentro do conjunto de dados.

Foram feitas três rodadas de identificação de ortólogos, que diferem quanto às espécies usadas. A primeira rodada foi feita utilizando os 40 transcriptomas com o objetivo de identificar evidências de potenciais expansões gênicas, de acordo com a hipótese de que a identificação de múltiplos ortogrupos com a mesma anotação seria indício de expansões gênicas. A segunda rodada de identificação de ortólogos foi feita utilizando 34 transcriptomas, um por espécie, com o objetivo de maximizar o número de ortogrupos compostos por uma única sequência por espécie, chamados de ortogrupos de cópia única ("single-copy orthogroups" - SCO), que foram posteriormente utilizados para análises filogenéticas. Para evitar uma falsa atribuição de paralogia entre os transcritos, pois várias isoformas do mesmo gene podem coexistir em um mesmo transcriptoma, o algoritmo "get_longest_isoform_seq_per_trinity _gene.pl" do programa Trinity e o programa CD-HIT-EST v.4.6.6 (identidade >= 95%) foram utilizados para agrupar as sequências de proteínas. A terceira rodada do Orthofinder foi feita utilizando 11 transcriptomas (seis da tribo Hypostomini e cinco do clado *Peckoltia*). Antes da terceira rodada, também foi feito o agrupamento das sequências, como realizado na segunda rodada. O objetivo desta terceira rodada foi maximizar o número de ortogrupos de cópia única, compartilhados entre as espécies das tribos Hypostomini e do clado da *Peckoltia*, para testar se estas sequências apresentam evidências de seleção positiva no clado da *Peckoltia*, seguindo a hipótese de que a grande diversidade morfológica característica destas espécies seja subsidiada por uma taxa evolutiva de seus genes maior do que a do outro clado (Hypostomini) e que seus genes apresentariam sítios de seleção positiva.

Análise filogenética

A análise filogenética foi feita com os SCOs compartilhados entre as 34 espécies deste estudo. As sequências de proteínas dentro de cada SCO foram alinhadas pelo MAFFT v7.215 (22). O programa PAL2NAL v14.0 foi usado para converter o alinhamento de proteínas e as sequências de DNA correspondentes em um alinhamento de códon (23). O programa FASconCAT-G v1.02 (24) foi usado para concatenar os arquivos de alinhamento de sequências de nucleotídeos em uma super-matriz. O Gblocks v0.91b (opção de seleção mais rigorosa), incorporado no SeaView v4.5.3 (25), foi usado para eliminar posições mal alinhadas e regiões divergentes no super-alinhamento. Dessa forma, uma árvore filogenética de máxima verossimilhança foi construída usando o programa RAxML v8.2.9 (26) disponível no portal CIPRES (27) sob o modelo GTR + GAMMA + I para todos os sítios, com 1.000 “bootstraps”. A árvore filogenética foi visualizada e editada usando o programa FigTree (v1.4.2) (http://tree.bio.ed.ac.uk/ software/figtree/).

Análise de evolução molecular

A razão entre as taxas de substituições não sinônimas e sinônimas ($\omega = dN/dS$), é uma estimativa da pressão de seleção à nível de proteína. Se mutações não sinônimas forem fixadas na mesma taxa que as mutações sinônimas, de modo que $dN = dS$ e $\omega = 1$, é esperado que a seleção não tenha efeito no “fitness”. Por outro lado, mutações não as deletérias tendem a não fixação na população, devido a ação da seleção purificadora, de modo que $dN < dS$ e $\omega < 1$. Já mutações não sinônimas que aumentem o “fitness” tendem a ser fixadas pela ação da seleção darwiniana, atingindo uma taxa maior do que mutações sinônimas, resultando em $dN > dS$ e $\omega > 1$. Uma taxa de mutações não sinônimas significativamente maior do que a taxa de mutações sinônimas é, portanto, evidência para a evolução adaptativa das proteínas. A evolução adaptativa raramente ocorre em todas as espécies de uma filogenia, ou em todos os códons de um gene, o que dificulta a detecção dos genes/códons em questão. O cenário mais provável é que a seleção positiva ocorra em alguns ramos da filogenia, ou em alguns sítios específicos do gene ou apenas em sítios específicos de alguns ramos em particular. Cada uma dessas possibilidades é formalizada em um "modelo", para que possíveis processos de evolução possam ser testados explicitamente (28).

O programa CODEML, parte integrante do pacote “Phylogenetic Analysis by Maximum Likelihood” (PAML) (29), foi utilizado para estimar o ω (dN/dS) no conjunto de dados dos SCOs compartilhados entre as 11 espécies da terceira rodada do Orthofinder e para testar hipóteses evolutivas em diferentes modelos.

O primeiro modelo testado, o modelo de ramo (parâmetros: model = 2, NSsites = 0, fix_omega = 0, omega = 0.1, cleandata = 1), permite que um determinado ramo em destaque possua uma razão dN/dS diferente dos ramos de fundo. Já o modelo nulo (parâmetros: model = 0, NSsites = 0, fix_omega = 0, omega = 0.1, cleandata = 1) assume que todos os ramos possuam o mesmo ω. Utilizamos o modelo de ramo para identificar genes que possuam taxas de evolução diferentes entre os clados testados. Para testar a significância dessa diferença, utilizamos o teste de razão de verossimilhança (“likelihood ratio test” - LRT) com um grau de liberdade entre o modelo de ramo e o modelo nulo para cada ortogrupo do conjunto de dados. O valor-p foi calculado a partir da função de distribuição cumulativa da estatística de qui-quadrado e foi corrigido, devido aos múltiplos testes, pelo método da taxa de falsa descoberta (FDR) implementado no R (30). Consideramos que os genes estão evoluindo com uma taxa significativamente mais rápida no ramo em destaque se o valor-p ajustado fosse inferior a 0,05 e um maior valor de ω fosse detectado no ramo em destaque em relação ao ramo de fundo.

No modelo de ramo é improvável que a seleção positiva seja detectada, pois mesmo que alguns sítios da proteína estejam evoluindo rapidamente ao longo do ramo, o ω médio raramente será maior que 1, porque é esperado que a maioria dos sítios da proteína permaneça sob seleção purificadora. O esperado, para a maioria dos genes, é que a seleção positiva afete apenas alguns resíduos de aminoácidos ao longo de linhagens particulares. O modelo de ramo e sítio (31) permite que o ω varie tanto entre os sítios da proteína quanto através dos ramos da árvore e visa detectar a seleção positiva que afeta alguns sítios ao longo de linhagens particulares. Para isso, o modelo de ramo e sítio (parâmetros da hipótese alternativa: model = 2, NSsites = 2, fix_omega = 0, omega = 0.1, cleandata = 1) foi usado para identificar possíveis genes que estejam sob seleção positiva. O teste de razão de verossimilhança foi feito entre o modelo alternativo que permite que os códons estejam sob seleção positiva ($\omega > 1$) no ramo em destaque e o modelo nulo (parâmetros da hipótese nula: model = 2, NSsites = 2, fix_omega = 1, omega = 1, cleandata = 1) que permite duas classes de ω, a seleção neutra ($\omega = 1$) e a seleção de purificação ($\omega < 1$). Os valores-p foram calculados com base na estatística do qui-quadrado e ajustados pelo método FDR. Os genes com valores-p ajustados menores que 0,05 foram tratados como candidatos para seleção positiva.

RESULTADOS E DISCUSSÃO

Montagem *de novo* e anotação dos transcriptomas

Um total de mais de 200 Gigabases foram geradas no sequenciamento dos 40 transcriptomas hepáticos. Após a remoção dos adaptadores e das sequências de baixa qualidade, as leituras foram utilizadas para a montagem *de novo* dos transcriptomas, um por espécime.

As informações sobre o sequenciamento e a montagem, como a média, desvio padrão e total das leituras brutas e utilizadas, número de transcritos montados e N50 estão disponibilizados na Tabela 3, assim como, as informações sobre a predição das ORFs e sua anotação.

Tabela 3: Sumário dos resultados do sequenciamento, montagem *de novo*, predição de ORFs e anotação dos transcriptomas.

	Média	Desv. Padrão	Total
Sequenciamento - Illumina HiSeq2500			
Leituras brutas	51.266.692	15.216.237	2.050.667.660
Leituras utilizadas*	98,7%	0,2%	
Montagem - Trinity			
Transcritos montados	58.720	21.073	2.348.817
N50	1.180	269	
Predição de ORF - Transdecoder			
ORFs (40 spp)	20.751	6.013	830.049
Anotação - BLASTP			
UniProtKB - *Danio rerio*	18.769	5.150	750.754
SwissProt - *Homo sapiens*	17.314	4.650	692.561

*: As leituras utilizadas são mostradas em percentual do total de leituras brutas.

Um maior número de transcritos foi anotado (“E-value” < 1e-10) quando utilizado o banco de dados UniprotKB da espécie *Danio rerio* (Tabela 3), em comparação com a quantidade de transcritos anotados contra o banco de dados SwissProt para a espécie *Homo sapiens*. A fim de melhorar a confiança no resultado do BLASTP, as anotações passaram por um segundo filtro, que selecionou entradas cujo tamanho do alinhamento fosse superior a 60% do tamanho da ORF pesquisada (“query”) e cujo percentual de identidade do alinhamento superior a 40% (filtro 60-40). Devido a distância filogenética, o resultado contra o banco de dados SwissProt de humano foi o que teve o maior número de anotações descartadas (Figura 1B). Entretanto, quando verificamos a especificidade das anotações, observamos que mais de 50% das anotações contra o banco de dados de proteínas de *D. rerio* são descritas como proteínas não caracterizadas (Figura 1C), portanto, não são informativas. Dessa forma, as análises subsequentes levaram em consideração as anotações dos transcritos comparados ao banco de dados SwissProt da espécie *Homo sapiens*.

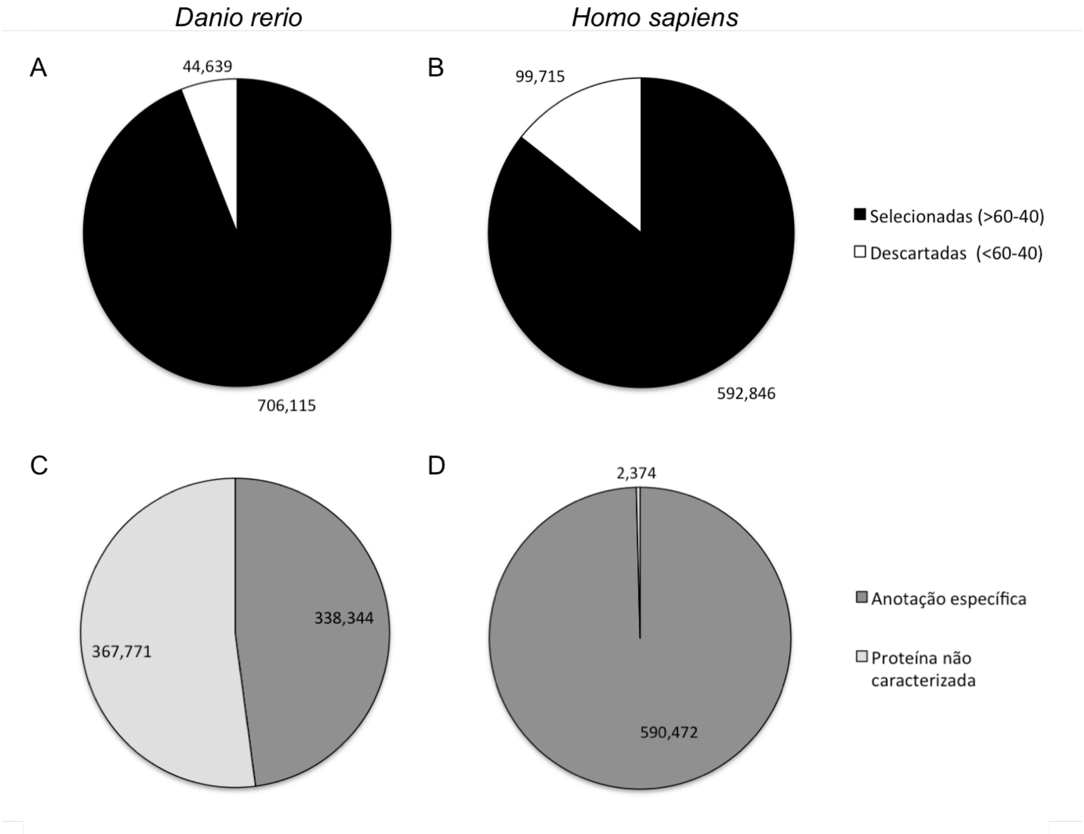

Figura 1: Gráficos de pizza com informações gerais sobre as anotações das ORFs preditas. Quantidade de ORFs anotadas contra os bancos de dados UniProtKB de *D. rerio* (A) e SwissProt de *H. sapiens* (B) que foram superiores ou inferiores ao filtro 60-40. Quantidade de ORFs que foram selecionadas (filtro 60-40) que obtiveram uma anotação específica ou foram anotadas como proteína não caracterizada em relação aos bancos de dados UniProtKB de *D. rerio* (C) e SwissProt de *H. sapiens* (D).

Identificação de ortólogos putativos

O conjunto de dados composto pelas ORFs preditas nos transcriptomas, para cada uma das três rodadas, foi utilizado como entrada para o programa Orthofinder, a fim de inferir grupos de sequências homólogas entre as espécies. A Tabela 4 sumariza os resultados das três rodadas feitas.

Tabela 4: Sumário dos resultados das três rodadas do Orthofinder. O50: O número mínimo de ortogrupos os quais contêm 50% das ORFs. G50: O número de ORFs contidas no menor ortogrupo do conjunto de ortogrupos que contêm 50% das ORFs (O50). Ortogrupo de cópia única: um ortogrupo com exatamente uma ORF (e não mais) de cada espécie.

	Rodadas		
	Primeira	Segunda	Terceira
Número de transcriptomas	40	34	11
Número de ORFs	830.049	586.735	184.926
ORFs em ortogrupos (%)	97,8	97,4	95,9
Número de ortogrupos	22.041	19.728	18.273
Mediana do tamanho dos ortogrupos	18	20	9
O50	3.304	3.320	4.357
G50	71	49	12
Número de ortogrupos com todas as espécies	2.746	2.824	5.265
Número de ortogrupos de cópia única	16	420	2.148

É possível notar que nas três rodadas, para a identificação de grupos de transcritos ortólogos, uma alta percentagem das ORFs (95,9 – 97,8%) encontrou alguma sequência homóloga em outra espécie, sendo agrupadas em ortogrupos. Apesar do percentual de ORFs alocadas em ortogrupos ter sido similar entre as três rodadas de busca por transcritos ortólogos, a mediana de transcritos nos grupos identificados variou de nove, na terceira rodada, a 20, na segunda rodada. À medida que o número de transcriptomas utilizados diminui, a mediana se aproxima da quantidade total de transcriptomas utilizados em dada rodada, isso se deve à dois fatores: primeiro, quanto mais próximo filogeneticamente as espécies analisadas, maiores as chances de homologia entre suas sequências, dessa forma, é maior a possibilidade de todas ou quase todas as espécies estarem em um mesmo ortogrupo; e segundo que de forma diferente do genoma, que é o conteúdo total do material genético herdável, o transcriptoma tem em seu conteúdo apenas os genes expressos em um dado momento da vida do organismo, a presença e detecção de um determinado ortólogo em todas as espécies depende da expressão desse gene, se ele foi sequenciado em uma profundidade suficiente para ser montado e ter sua ORF predita em todas as espécies. Isso também pode ser observado no número de ortogrupos com todas as espécies e no número de ortogrupos de cópia única, os quais aumentam à medida que a quantidade de espécies analisadas diminui.

Análise das expansões gênicas

Com o intuito de identificar possíveis expansões gênicas, resultantes de duplicações, seguimos as premissas: 1 – que transcritos variantes codificados por um mesmo gene e transcritos codificados por genes parálogos identificados em uma única espécie sejam agrupadas em um mesmo ortogrupo; e 2 – que dois ou mais ortogrupos com a mesma anotação contenham transcritos codificados por genes parálogos. Dessa forma, a identificação de múltiplos ortogrupos com a mesma anotação indica um possível processo de expansão gênica nas espécies que compartilham estes ortogrupos, em relação a espécie usada para a anotação. Neste trabalho, a anotação foi feita contra as entradas de duas espécies, *Danio rerio* (espécie modelo) e *Ictalurus punctatus* (espécie filogeneticamente mais próxima de Loricarioidei), encontradas em três bancos de dados (Gene/NCBI, Protein/NCBI e UniProtKB).

Os transcritos que constituem 11.394 dos 22.041 ortogrupos identificados pelo OrthoFinder na primeira rodada (Tabela 4) foram anotados como homólogos contra alguma sequência do banco de dados SwissProt de humano (E-value < 1e-10, identidade > 40%, cobertura da ORF pesquisada > 60%). Dos 11.394 ortogrupos anotados, 7.129 tiveram anotações exclusivas (não compartilhada por nenhum outro ortogrupo) e 4.265 ortogrupos tiveram anotações compartilhadas (não exclusivas), distribuídas em 1.749 anotações. A maior quantidade de ortogrupos com a mesma anotação foi de 12 (Tabela 5) e a menor de 2 ortogrupos.

É observado a limitação do Orthofinder em determinar grupos de ortólogos, devido à algumas variáveis de confusão, como por exemplo, ORFs incompletas. Contudo, essa abordagem de aplicar o Orthofinder em grupo de espécies próximas, sob uma ótica evolutiva, e depois anotar os ortogrupos tem se mostrado eficaz, pois tenta reduzir a confusão entre ortólogos e parálogos comum quando se compara apenas uma espécie contra um banco de dados de uma espécie modelo distante filogeneticamente. A confusão entre parálogos e ortólogos quando comparamos espécies de peixes com banco de dados de humanos se dá principalmente ao evento de duplicação do genoma exclusivo do último ancestral comum dos teleósteos (10,32). Esse evento tem uma importante influência na anotação gênica, particularmente na determinação da ortologia, pois, comparações entre teleósteos e outros vertebrados podem levar a falsas inferências de ortologia. Dessa forma é esperado que um determinado gene, que tenha mantido sua cópia, seja encontrado duplicado quando comparado com humano. A Tabela 5 exibe anotações que se apresentam em mais de sete ortogrupos.

Um caso interessante é o do gene da proteína Complemento C3, que apesar de não ser a anotação mais frequente entre os ortogrupos, é a que tem a maior quantidade de ORFs (911) distribuídas entre os oito ortogrupos com essa anotação. De acordo com o banco dados Genes do NCBI, o complemento C3 é o gene com maior quantidade de entradas no genoma de *D. rerio* e o único gene que tem mais de uma entrada registrada para *I. punctatus* (Tabela 5). Essa expansão gênica de C3 já foi observada em outros peixes, como truta (33), carpa (34) e medaka (35). Em um estudo a respeito dessa expansão de C3 em *D. rerio*, foi descoberto um evento de duplicação comum a todos os peixes teleósteos (36) corroborando o evento de duplicação genômica no ancestral desse grupo. Esses achados em relação ao gene C3 corroboram a hipótese que as premissas usadas nessa etapa deste trabalho sejam verdadeiras.

De forma contrária ao caso do gene C3, outras anotações compartilhadas por múltiplos ortogrupos no nosso conjunto de dados apresentam uma única entrada em *D. rerio* e *I. punctatus* no banco de dados Genes do NCBI, sugerindo expansões gênicas em Loricarioidei não compartilhadas com essas duas espécies usadas como modelo. O primeiro passo dessa análise foi recuperar os ortogrupos com a mesma anotação e comparar essa anotação com a quantidade de registros em bancos de dados de outros peixes, resultando na primeira evidência de uma possível expansão gênica. O próximo passo será recuperar as sequências dos ortogrupos com mesma anotação (candidatos a parálogos), seus homólogos em espécies próximas e fazer uma análise filogenética para confirmar essas possíveis expansões.

Tabela 5: Sumário das anotações compartilhadas por sete ou mais ortogrupos e a quantidade de entradas dessas proteínas nos bancos de dados Gene, Protein e UniProtKB das espécies *Danio rerio* e *Ictalurus punctatus*. Também são apresentados o número mínimo, máximo e total das ORFs nos ortogrupos.

		Danio rerio			*Ictalurus punctatus*			Loricarioidei			
Anotação	Gene	Gene (NCBI)*	Protein (NCBI)*	UniProtKB**	Gene (NCBI)*	Protein (NCBI)*	UniProtKB**	Ortogrupos	ORFs em ortogrupos		
									min.	max.	total
Baculoviral IAP repeat-containing protein 6	BIRC6	1	17	2	1	1	1	8	6	131	249
Brefeldin A-inhibited guanine nucleotide-exchange protein 3	ARFGEF3	1	2	2	1	1	1	7	2	11	41
Centrosomal protein of 290 kDa	CEP290	1	6	3	1	1	1	8	2	28	75
Complement C3	C3	9	16	2	5	6	7	8	26	205	911
Cytoplasmic dynein 1 heavy chain 1	DYNC1H1	1	1	3	1	2	2	7	6	96	251
Cytoplasmic dynein 2 heavy chain 1	DYNC2H1	2	1	0	1	1	1	7	2	83	120
DNA polymerase epsilon catalytic subunit A	POLE	1	1	15	1	1	6	8	2	9	38
DNA-dependent protein kinase catalytic subunit	PRKDC	1	1	1	1	1	1	8	2	98	133
E3 ubiquitin-protein ligase HERC1 Probable	HERC1	1	11	5	1	2	1	8	9	151	408
E3 ubiquitin-protein ligase MYCBP2	MYCBP2	1	19	7	1	18	17	12	5	111	238
E3 ubiquitin-protein ligase UBR4	UBR4	1	28	2	1	9	9	8	7	137	266
Huntingtin	HTT	1	6	3	1	1	2	7	6	109	197
Lysosomal-trafficking regulator	LYST	1	4	6	1	5	5	8	3	108	166
Midasin	MDN1	1	1	2	1	2	2	12	4	86	265
Neurobeachin-like protein 1	NBEAL1	1	2	2	1	1	1	8	7	94	172
Neurofibromin	NF1	2	10	10	1	6	7	7	4	63	113
Nuclear pore membrane glycoprotein 210	NUP210	1	1	1	1	4	2	7	2	5	24
Protein SZT2	SZT2	2	1	1	1	13	13	8	3	54	114
Transformation/transcription domain-associated protein	TRRAP	1	4	2	1	3	2	7	9	102	194
Uncharacterized protein KIAA1109	KIAA1109	1	8	4	1	12	11	8	2	96	227
Vacuolar protein sorting-associated protein 13B	VPS13B	1	0	0	1	12	11	9	6	100	199
Vacuolar protein sorting-associated protein 13D	VPS13D	1	4	3	1	7	7	11	2	83	151

* informação obtida no banco de dados Gene e Protein do NCBI, usando "o ID do gene cada anotação" e "espécie"[porgn:__txid]) (03/2018). *Danio rerio* txid = 7955; *Ictalurus punctatus*; txid = 7998.

** informação obtida no banco de dados UniProt, usando "o ID do gene de cada anotação (GN)" e a espécie (OR) (03/2018).

Análise filogenética

A segunda rodada do OrthoFinder resultou em 420 ortogrupos de cópia única (Tabela 4), esses SCOs foram alinhados e concatenados gerando um super-alinhamento de 415.314 bases, que após o processo de limpeza usando o GBlocks, resultou em 260.817 bases alinhadas nas 34 espécies, produzindo uma árvore filogenética de máxima verossimilhança com ramos com alto suporte estatístico (Figura 2).

A árvore filogenética foi enraizada usando as três espécies da família Callichthyidae (Figura 2 e Tabela 2). A topologia recuperada para as subfamílias é a mesma da árvore gerada com os mitogenomas (Capítulo Quatro). O ramo da subfamília Hypoptopomatinae continua sendo considerado monofilético e a tribo Otothyrini permanece mais próxima das espécies de Neoplecostomini, do que de espécies Hypoptopomatini (37). Contudo, a análise com os ortólogos nucleares não suporta a monofilia da tribo Neoplecostomini, trazendo para debate as divisões taxonômicas recentes em tribos (12) ou subfamílias (37) desse clado.

O resultado da análise filogenética com os ortólogos nucleares corrobora a relação próxima entre as espécies *Schizolecis guntheri* e *Pareiorhaphis garbei* encontrada com o estudo dos genomas mitocondriais. Confirmando que essas espécies devem ser classificadas na mesma tribo, de forma diferente que os recentes trabalhos vêm fazendo (37,38).

As relações entre três das espécies classificadas no clado da *Peckoltia* também foram modificadas, em comparação à árvore filogenética com os genomas mitocondriais apresentada no Capítulo Quatro. Na árvore com base nos dados dos ortólogo nucleares, *Ancistomus snethlageae* formou um grupo externo às espécies *Peckoltia furcata* e *Panaqolus* sp. (Figura 2); enquanto na árvore feita com base nos genomas mitocondriais, *Peckoltia furcata* formou um grupo externo às espécies *Ancistomus snethlageae* e *Panaqolus* sp. (Capítulo 4).

Outra característica importante do clado da *Peckoltia* é a grande proximidade genética entre os gêneros (*Hypancistrus*, *Ancistomus*, *Peckoltia* e *Panaqolus*), observada pelo tamanho dos ramos na Figura 2, quando comparada à distância entre as espécies dos gêneros *Pterygoplichthys* e *Hypostomus* que compõem o clado irmão. Lujan *et al.*, 2015, também encontrou ramos internos curtos entre gêneros do clado da *Peckoltia*. Como o comprimento dos ramos internos foram curtos e o volume da dados usados por Lujan *et al.*, 2015 para computar essas distâncias foi muito menor do que aquele usado neste trabalho, as relações encontrados por aqueles autores foram mal resolvidas ou fracamente suportadas entre os gêneros do grupo *Peckoltia*.

Essa discrepância entre a classificação de espécies geneticamente mais próximas em gêneros diferentes (*Peckoltia*) e de espécies geneticamente mais distantes no mesmo gênero (p. ex. *Hypostomus*) também foi observada e discutida na análise das identidades nucleotídicas entre os genomas mitocondriais no Capítulo Quatro desta tese.

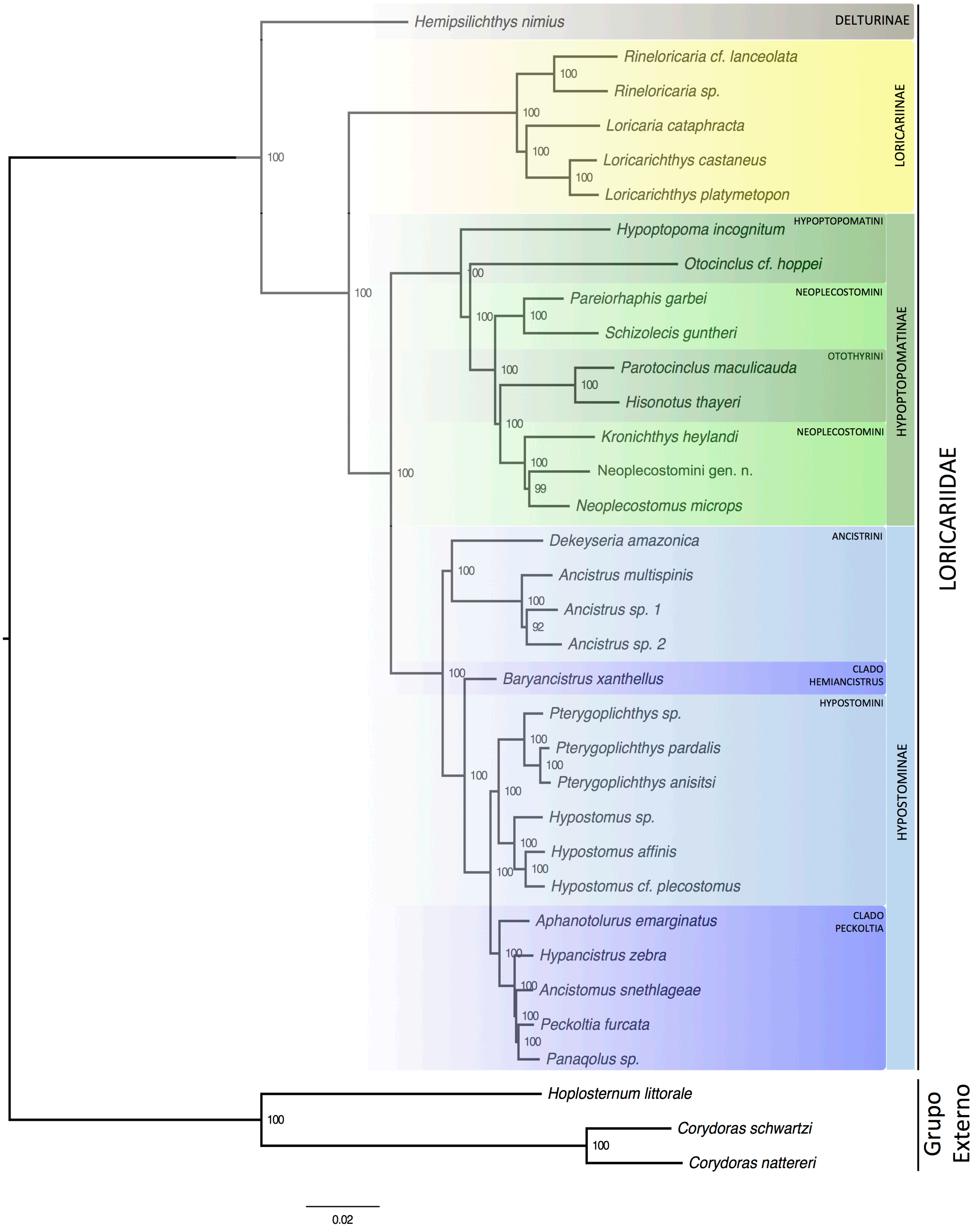

Figura 2: Árvore filogenética de máxima verossimilhança da família Loricariidae. Produzida a partir do super-alinhamento de 420 ortogrupos de cópia única, somando 260.817 bases. Os valores de suporte do Bootstrap são mostrados em cada nó e são baseados em 1000 réplicas. Espécies da família Callichthyidae foram usadas como grupo externo. As subfamílias de Loricariidae são destacadas em cinza (Delturinae), amarela (Loricariinae), verde (Hypoptopomatinae) e azul (Hypostominae). A barra de escala representa a taxa de substituição de nucleotídeos, usando o modelo GTR + GAMMA + I.

Análise de evolução molecular

A terceira rodada do Orthofinder, feita com o conjunto de dados de 11 espécies (seis da tribo Hypostomini e cindo do clado *Peckoltia*), identificou 2.140 SCOs. Em seguida, usamos uma sub-árvore (Figura 3) a partir da árvore de ML gerada neste Anexo em conjunto com um modelo de ramo (PAML) para determinar os valores de dN, dS e ω em todos os 2.140 grupos de ortólogos. Foi utilizado como clado em destaque o ramo composto por quatro espécies do clado da *Peckoltia* (*Hypancistrus zebra*, *Ancistomus snethlageae*, *Peckoltia furcata*, *Panaqolus* sp.), seguindo a hipótese de que a grande diversidade morfológica baseada em uma grande proximidade genética poderia ser sustentada por ω maiores no clado mais diverso. A espécie *Aphanotolurus emarginatus*, pertencente ao clado da *Peckoltia*, não foi colocada no clado em destaque pela grande similaridade morfológica com a tribo Hypostomini, particularmente com o gênero *Hypostomus* e pela maior distância genética em relação às outras quatro espécies de seu clado. O gênero *Aphanotolurus* já foi classificado como *Hypostomus* e têm uma história taxonômica complicada durante a qual vários nomes já foram propostos (7).

O resultado da análise com o modelo de ramo mostrou que o valor médio de ω foi significativamente maior no clado em destaque do que no clado de fundo com valor-p < 0,05 no teste de Mann-Whitney-Wilcoxon (Figura 3), evidenciando uma aceleração da evolução nas espécies destacadas. Contudo, apesar do ω médio para os ortólogos testados ser maior no clado em destaque, o LRT para cada ortogrupo detectou apenas 11 genes que melhor se ajustaram no modelo alternativo, que permite ω diferentes para cada clado, com valor-p ajustado < 0,05 (Tabela 6).

Apesar das expectativas teóricas e empíricas que a taxa de diversidade morfológica está associada a taxa de especiação (1,39–41), a hipótese de que o clado da *Peckoltia* teria uma taxa evolutiva maior que a tribo Hypostomini não pôde ser confirmada pelo modelo de ramo (LRT). De fato, essa associação não é uma regra, um estudo com a família Plethodontidae (salamandras) mostrou que as taxas de diversificação de espécies e evolução morfológica analisadas não foram significativamente correlacionadas, de tal forma que a rápida diversificação pode ocorrer com pouca mudança morfológica, e vice-versa (42).

Não podemos negar que a relação exista, primeiro porque não medimos diretamente a taxa de especiação, e segundo que o poder da nossa análise pode ter sido fraco para detectar essa diferença pela baixa quantidade de *taxa* amostrados.

Entretanto, acreditamos que a causa da baixa frequência de ortogrupos com ω significativamente diferentes (LRT) se deva à problemas na prática taxonômica. A observação de taxas de especiação mais altas nas mesmas linhagens que mostram uma diversidade morfológica (39) pode ser uma consequência de uma maior propensão dos taxonomistas em dividir linhagens morfologicamente variáveis em múltiplas espécies ou gêneros. Várias das espécies do clado da *Peckoltia* são separadas morfologicamente entre si por padrões de coloração (6,7,17), resultando em uma baixa confiança. Idealmente, a delimitação de espécies deve ser feita usando um conjunto de critérios idênticos e quantitativos (43).

A dificuldade de resolução dos ramos internos ao clado da *Peckoltia* existe tanto em nível morfológico (17), quanto molecular (12), assim como a monofilia do gênero *Hypostomus* não pôde ser totalmente suportada em um estudo de filogenia molecular abrangente (12). Apesar da tribo Hypostomini possuir apenas dois gêneros (sensu Lujan, 2015), é a tribo mais rica da subfamília Hypostominae, com 165 espécies (Tabela 1) (5), com ampla distribuição geográfica e também possui espécies especializadas em comer madeira (grupo *Hypostomus cochliodon*, não amostrado nesse estudo).

Nas análises de evolução molecular, além do modelo de ramo, também utilizamos o modelo de ramo e sítio para identificar genes candidatos positivamente selecionados no clado destacado. Entre os 2.140 SCOs, o LRT detectou 73 ortogrupos com valor-p ajustado < 0,05 (tabela 7). Esse resultado é o mais recente entre os obtidos neste trabalho e ainda não foi feita uma análise aprofundada a seu respeito. O próximo passo será verificar quais desses genes podem estar envolvidos na dinâmica ecológica dessas espécies e quais sítios estão sendo positivamente selecionados. Apesar da maioria dos fenótipos ser controlada por muitos genes e de muitos fenótipos interferirem na dinâmica ecológica de um organismo, o número de variantes em uma escala genômica é tão grande que a identificação de genes que possam estar passando por um processo de seleção positiva e a associação desses genes a alguns fenótipos é um ponto de partida importante para a exploração das relações entre genótipos, fenótipos e suas interações com o ambiente (44).

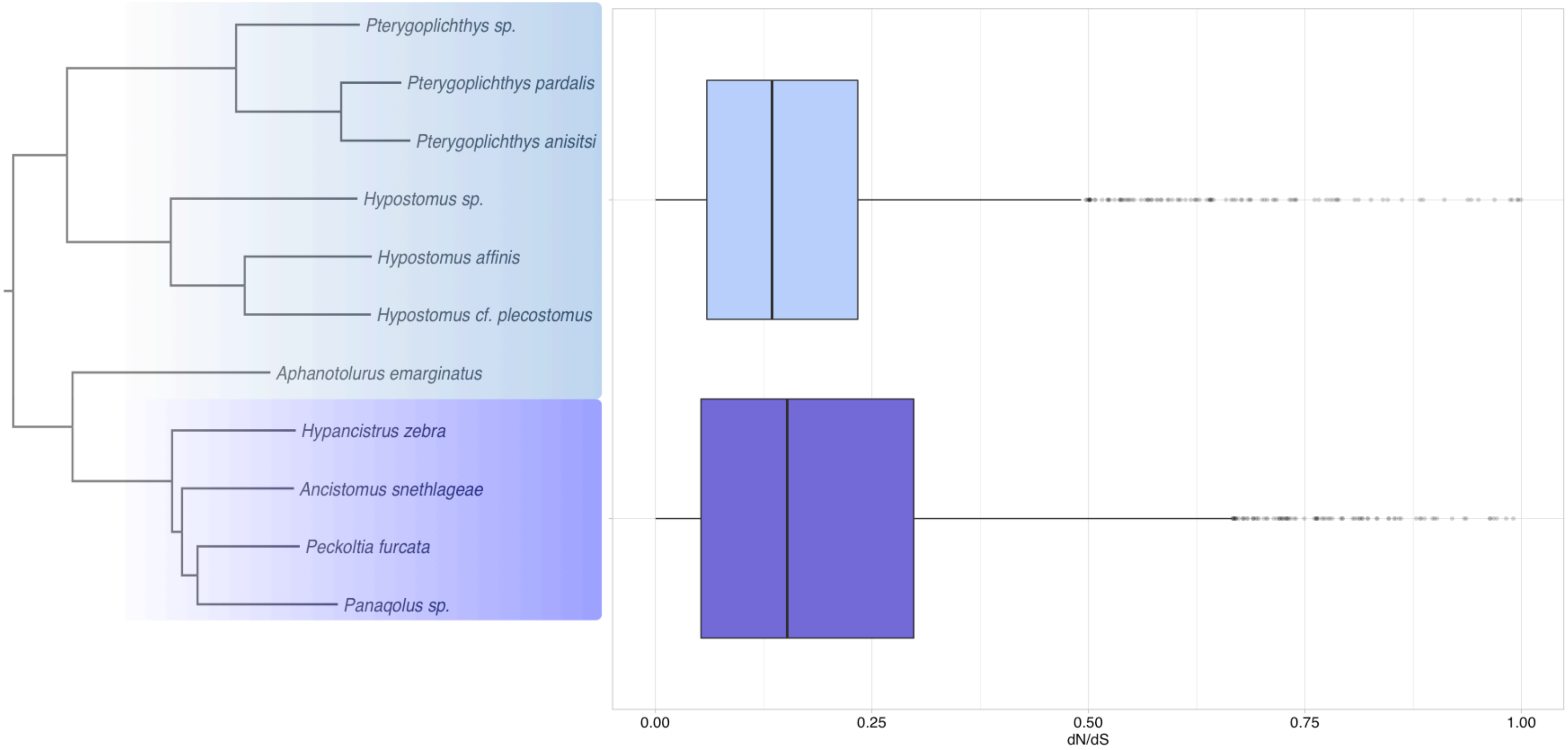

Figura 3: Sub-árvore filogenética usada na análise de evolução molecular e boxplot das razões dN/dS estimadas a partir de cada ortogrupo, no modelo de ramo, para os clados de fundo (azul) e em destaque (roxo).

Tabela 6: Lista dos ortogrupos que melhor se ajustaram ao modelo de ramo e que tiveram razões dN/dS maiores no clado em destaque. São apresentados os "logLikelihood" (lnL) do modelo nulo (H0, ω igual para todos os ramos) e alternativo (H1, ω diferentes para cada clado), o resultado do teste de razão das verossimilhanças (LRT), os valores-p calculados com base na estatística do qui-quadrado e ajustados pelo método FDR, as razões dN/dS para cada clado e a anotação do ortogrupo.

Ortogrupo	lnL - H0	lnL - H1	LRT	Valor-p	dN/dS	dN/dS	Anotação
			(2ΔlnL)	corrigido (fdr)	Clado de Fundo	Clado em Destaque	SwissProt *Homo sapiens*
OG0005553	-1845,8277	-1839,1069	13,4417	0,041	0,1464	0,5620	Unconventional prefoldin RPB5 interactor 1
OG0005863	-1658,4599	-1650,4991	15,9216	0,014	0,1507	0,7913	Neuroguidin
OG0005874	-1743,9179	-1737,0580	13,7197	0,038	0,0726	0,3315	Probable ATP-dependent RNA helicase DDX56
OG0006055	-3221,6315	-3210,5367	22,1897	0,003	0,1845	0,9362	Fatty-acid amide hydrolase 2
OG0006606	-1418,7928	-1410,5983	16,3890	0,014	0,0073	0,1557	Protein FAM199X
OG0006665	-1799,9190	-1791,5291	16,7797	0,014	0,0121	0,2582	BUB3-interacting and GLEBS motif-containing protein ZNF207
OG0006833	-1578,3938	-1567,8760	21,0356	0,003	0,0541	0,6029	3-hydroxyisobutyrate dehydrogenase, mitochondrial
OG0006934	-590,3915	-582,0706	16,6419	0,014	0,1718	0,8222	39S ribosomal protein L35, mitochondrial
OG0007066	-1827,3391	-1815,2799	24,1184	0,002	0,0474	0,4278	Presenilins-associated rhomboid-like protein, mitochondrial
OG0007486	-2271,8490	-2265,4226	12,8528	0,048	0,0509	0,4166	Translation initiation factor eIF-2B subunit gamma
OG0007619	-1013,6153	-1004,3272	18,5763	0,009	0,0001	0,3871	Serrate RNA effector molecule homolog

Tabela 7: Lista dos ortogrupos que possuem sítios positivamente selecionados, ou seja, que melhor se ajustaram no modelo de ramo e sítio com o parâmetro do ω > 1. São apresentados os "logLikelihood" (lnL) do modelo nulo (H0, ω <= 1) e alternativo (H1, ω > 1), o resultado do teste de razão das verossimilhanças (LRT), os valores-p calculados com base na estatística do qui-quadrado e ajustados pelo método FDR e a anotação do ortogrupo.

Ortogrupo	lnL - H0	lnL - H1	LRT	Valor-p	Anotação
			(2ΔlnL)	corrigido (fdr)	SwissProt *Homo sapiens*
OG0005200	-841,2561	-826,7541	29,0039	5,7E-06	Syntaxin-18
OG0005223	-1399,5483	-1390,1262	18,8442	7,3E-04	Mitochondrial dicarboxylate carrier
OG0005253	-548,8007	-540,6056	16,3903	2,3E-03	Secretory carrier-associated membrane protein 1
OG0005382	-1016,0245	-997,1834	37,6823	8,8E-08	Acyl-protein thioesterase 1
OG0005417	-1157,5263	-1134,7920	45,4686	2,4E-09	Cancer-related nucleoside-triphosphatase
OG0005446	-1321,6884	-1315,4948	12,3873	1,5E-02	Non-homologous end-joining factor 1
OG0005450	-1785,8343	-1733,7768	104,1150	0,0E+00	Uncharacterized protein CXorf23
OG0005506	-1233,6739	-1226,5005	14,3467	5,8E-03	28S ribosomal protein S30, mitochondrial
OG0005527	-1147,9195	-1123,8594	48,1203	7,2E-10	Inositol oxygenase
OG0005536	-2617,9484	-2606,4830	22,9310	1,0E-04	no_hit
OG0005542	-1569,3028	-1526,6987	85,2082	0,0E+00	Insulin-like growth factor-binding protein 2
OG0005553	-1817,2688	-1780,4163	73,7049	0,0E+00	Unconventional prefoldin RPB5 interactor 1
OG0005565	-2752,8168	-2746,2565	13,1205	1,0E-02	Eukaryotic peptide chain release factor subunit 1
OG0005608	-2734,1762	-2720,5538	27,2448	1,3E-05	E3 UFM1-protein ligase 1
OG0005645	-860,8561	-855,1801	11,3519	2,4E-02	Protein CCSMST1
OG0005656	-2185,2244	-2175,2880	19,8728	4,4E-04	Pseudokinase FAM20A
OG0005660	-1518,4445	-1510,8386	15,2118	4,1E-03	C-type lectin domain family 4 member G
OG0005665	-928,0316	-918,7503	18,5626	8,0E-04	no_hit
OG0005698	-2603,0766	-2584,1416	37,8701	8,5E-08	Lipase member H
OG0005771	-1244,6444	-1227,3335	34,6217	3,9E-07	Kelch domain-containing protein 3
OG0005779	-868,2252	-862,6942	11,0621	2,7E-02	Centrin-2

Tabela 7: Cont.

Ortogrupo	lnL - H0	lnL - H1	LRT (2ΔlnL)	Valor-p corrigido (fdr)	Anotação SwissProt *Homo sapiens*
OG0005784	-2128,7575	-2122,9836	11,5479	2,2E-02	Cell cycle control protein 50A
OG0005846	-1454,4834	-1445,3005	18,3657	8,7E-04	Lysoplasmalogenase
OG0005863	-1642,3529	-1620,4819	43,7419	5,4E-09	Neuroguidin
OG0005874	-1716,2258	-1692,1230	48,2057	7,2E-10	Probable ATP-dependent RNA helicase DDX56
OG0005886	-1065,1269	-1058,9633	12,3272	1,5E-02	DnaJ homolog subfamily B member 9
OG0005920	-948,8048	-942,7601	12,0895	1,7E-02	Phosducin-like protein 3
OG0005932	-685,5285	-676,2082	18,6407	7,9E-04	Putative peptidyl-tRNA hydrolase PTRHD1
OG0006002	-1205,1759	-1161,6058	87,1403	0,0E+00	Reticulon-3
OG0006052	-404,2630	-392,5137	23,4986	8,1E-05	TCF3 fusion partner
OG0006055	-3202,5923	-3193,6989	17,7868	1,2E-03	Fatty-acid amide hydrolase 2
OG0006211	-1875,0131	-1868,3872	13,2520	9,9E-03	V-type proton ATPase subunit S1
OG0006249	-770,4564	-753,7891	33,3346	6,9E-07	no_hit
OG0006260	-824,5782	-819,6016	9,9532	4,7E-02	Cofilin-2
OG0006382	-740,7149	-726,3893	28,6511	6,6E-06	Ribosomal protein S6 kinase beta-1
OG0006470	-1516,3447	-1497,0827	38,5240	6,5E-08	28S ribosomal protein S22, mitochondrial
OG0006533	-1255,8400	-1247,3275	17,0250	1,7E-03	Steroid receptor RNA activator 1
OG0006535	-2093,1725	-2072,2917	41,7617	1,4E-08	Chitinase domain-containing protein 1
OG0006551	-2158,2640	-2099,6134	117,3012	0,0E+00	Interleukin-10 receptor subunit beta
OG0006553	-2949,5691	-2913,7582	71,6219	0,0E+00	Non-POU domain-containing octamer-binding protein
OG0006603	-2564,3194	-2532,1410	64,3567	2,1E-13	WD and tetratricopeptide repeats protein 1
OG0006606	-1387,4890	-1372,9833	29,0113	5,7E-06	Protein FAM199X
OG0006624	-1657,6375	-1614,6076	86,0598	0,0E+00	Ubiquitin domain-containing protein UBFD1
OG0006635	-2055,7477	-2048,4151	14,6652	5,1E-03	Homocysteine-responsive endoplasmic reticulum-resident ubiquitin-like domain member 2 protein

Tabela 7: Cont.

Ortogrupo	lnL - H0	lnL - H1	LRT (2ΔlnL)	Valor-p corrigido (fdr)	Anotação SwissProt *Homo sapiens*
OG0006652	-1372,8690	-1362,1114	21,5151	2,0E-04	no_hit
OG0006665	-1780,0970	-1767,4109	25,3720	3,3E-05	BUB3-interacting and GLEBS motif-containing protein ZNF207
OG0006723	-1234,2375	-1221,6464	25,1822	3,5E-05	Thyroid transcription factor 1-associated protein 26
OG0006761	-904,7262	-887,9877	33,4771	6,7E-07	Protein-L-isoaspartate(D-aspartate) O-methyltransferase
OG0006770	-979,4372	-958,8669	41,1406	1,8E-08	D-dopachrome decarboxylase
OG0006833	-1548,1223	-1514,5435	67,1577	5,3E-14	3-hydroxyisobutyrate dehydrogenase, mitochondrial
OG0006934	-576,4880	-571,4424	10,0912	4,4E-02	39S ribosomal protein L35, mitochondrial
OG0006971	-377,2381	-365,5092	23,4577	8,1E-05	Protein FAM177A1
OG0006983	-934,4727	-911,0175	46,9104	1,2E-09	Protein Mpv17
OG0007020	-1126,3945	-1116,6171	19,5549	5,1E-04	Ancient ubiquitous protein 1
OG0007066	-1802,9779	-1792,9354	20,0851	4,1E-04	Presenilins-associated rhomboid-like protein, mitochondrial
OG0007070	-1552,9767	-1547,0352	11,8830	1,9E-02	Probable tRNA N6-adenosine threonylcarbamoyltransferase
OG0007084	-1649,3224	-1641,7996	15,0457	4,4E-03	C-type lectin domain family 4 member K
OG0007103	-1644,4888	-1637,1110	14,7556	5,0E-03	NEDD8-activating enzyme E1 regulatory subunit
OG0007112	-1366,0972	-1351,2771	29,6403	4,5E-06	Androgen-induced gene 1 protein
OG0007155	-1738,4892	-1687,7351	101,5082	0,0E+00	Caldesmon
OG0007170	-1518,7181	-1499,9083	37,6197	8,8E-08	Solute carrier family 25 member 33
OG0007205	-1577,7063	-1571,9673	11,4779	2,3E-02	39S ribosomal protein L2, mitochondrial
OG0007252	-1169,6006	-1164,3374	10,5264	3,6E-02	PDZ and LIM domain protein 1
OG0007305	-1116,7128	-1111,1386	11,1484	2,6E-02	Ubl carboxyl-terminal hydrolase 18
OG0007315	-2465,8613	-2455,2636	21,1954	2,3E-04	Saccharopine dehydrogenase-like oxidoreductase
OG0007408	-857,5029	-851,7122	11,5815	2,2E-02	Sodium/bile acid cotransporter 7
OG0007456	-1031,5567	-1024,7952	13,5230	8,9E-03	Regulator of G-protein signaling 4
OG0007459	-620,6622	-613,3754	14,5737	5,3E-03	ATP synthase-coupling factor 6, mitochondrial

Tabela 7: Cont.

Ortogrupo	lnL - H0	lnL - H1	LRT (2ΔlnL)	Valor-p corrigido (fdr)	Anotação SwissProt *Homo sapiens*
OG0007512	-1888,2891	-1880,8425	14,8933	4,7E-03	Mitoferrin-2
OG0007535	-1948,2397	-1940,2205	16,0384	2,7E-03	Cap-specific mRNA (nucleoside-2'-O-)-methyltransferase 1
OG0007619	-993,9903	-980,9624	26,0559	2,4E-05	Serrate RNA effector molecule homolog
OG0007644	-990,9149	-979,6345	22,5609	1,2E-04	28S ribosomal protein S11, mitochondrial
OG0007655	-2265,8969	-2259,2442	13,3055	9,8E-03	tRNA-dihydrouridine(47) synthase [NAD(P)(+)]-like

CONCLUSÕES PARCIAIS

Este é o primeiro trabalho a fazer uso do sequenciamento de ácidos nucléicos de alto desempenho para o estudo desse grupo de espécies, promovendo o conhecimento sobre a diversidade genética, taxonômica, filogenética, funcional e fenotípica da fauna de peixes neotropicais. O sequenciamento, a montagem, a anotação e a análise dos 40 transcriptomas de 34 espécies da subordem Loricarioidei permitiu, até o momento, concluir:

- A filogenia inferida, a partir dos ortólogos nucleares, ressalta a divisão taxonômica controversa da subfamília Hypoptopomatinae e do clado da *Peckoltia*.
- A estratégia para inferência de ortologia desenvolvida neste trabalho foi eficaz para a identificação de expansões gênicas.
- A identificação de apenas 11 ortogrupos (dentre os 2.140 SCOs) que melhor se ajustaram no modelo alternativo de ramo refuta a hipótese que uma alta diversidade morfológica esteja relacionada a uma alta frequência de substituições não sinônimas e destaca os problemas referentes às delimitações taxonômicas.
- Os genes que exibem assinaturas de seleção positiva servirão de base para novas investigações que visam entender a relação desses genes aos fenótipos adaptativos e a sua correlação ecológica.
- Os recursos transcriptômicos produzidos por esse estudo irão compor bancos de dados públicos, reduzindo a lacuna genética existente e fornecendo a base para novos estudos sobre a história evolutiva desses peixes e potenciais descoberta de genes e transcritos genuinamente novos.

REFERÊNCIAS

1. Schluter D. The ecology of adaptive radiation. Oxford University Press; 2000. 288 p.

2. Eschmeyer WN, Fong JD. SPECIES BY FAMILY/SUBFAMILY. [Internet]. 2018 [cited 2018 Mar 25]. Available from: http://researcharchive.calacademy.org/research/ichthyology/catalog/SpeciesByFamily.asp

3. Lujan NK, Armbruster JW. Morphological and functional diversity of the mandible in suckermouth armored catfishes (Siluriformes: Loricariidae). J Morphol. 2012;273(1):24–39.

4. Lujan NK, Winemiller KO, Armbruster JW. Trophic diversity in the evolution and community assembly of loricariid catfishes. BMC Evol Biol. 2012;12(1):124.

5. Eschmeyer WN, Fricke R, Laan R van der. CATALOG OF FISHES: GENERA, SPECIES, REFERENCES [Internet]. 2018 [cited 2018 Mar 25]. Available from: http://researcharchive.calacademy.org/research/ichthyology/catalog/fishcatmain.asp

6. Armbruster JW, Lujan NK, Taphorn DC. Four New Hypancistrus (Siluriformes: Loricariidae) from Amazonas, Venezuela. Copeia. 2007;1:62–79.

7. Ray CK, Armbruster JW. The genera Isorineloricaria and Aphanotorulus (Siluriformes: Loricariidae) with description of a new species. Zootaxa. 2016;4072(5):501–39.

8. Lujan NK, German DP, Winemiller KO. Do wood-grazing fishes partition their niche?: Morphological and isotopic evidence for trophic segregation in Neotropical Loricariidae. Funct Ecol. 2011;25(6):1327–38.

9. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63.

10. Braasch I, Peterson SM, Desvignes T, McCluskey BM, Batzel P, Postlethwait JH. A new model army: Emerging fish models to study the genomics of vertebrate Evo-Devo. J Exp Zool Part B Mol Dev Evol. 2015;324(4):316–41.

11. Stapley J, Reger J, Feulner PGD, Smadja C, Galindo J, Ekblom R, et al. Adaptation genomics : the next generation. Trends Ecol Evol. 2010;25(12):705–12.

12. Lujan NK, Armbruster JW, Lovejoy N, López-fernández H. Multilocus molecular phylogeny of the suckermouth armored catfishes (Siluriformes: Loricariidae) with a focus on subfamily Hypostominae. Mol Phylogenet Evol. 2015;82:269–88.

13. Chavez JM, Paz RMD La, Manohar SK, Pagulayan RC, Vi JRC. New Philippine record of south american sailfin catfishes (Pisces: Loricariidae). Zootaxa. 2006;1109(1):57–68.

14. Marques H, Nobile AB, Dias JHP, Ramos IP. Length-weight and length-length relationships for 23 fish species of Porto Primavera reservoir, Upper Paraná River, Brazil. J Appl Ichthyol. 2016;32(6):1342–6.

15. da Costa MR, Moreti T, Araújo FG. Length-weight relationships of 20 fish species in the Guandu River, Rio de Janeiro State, Southeastern

Brazil. J Appl Ichthyol. 2014;30(1):200–1.

16. PlanetCatfish [Internet]. [cited 2018 Mar 25]. Available from: https://www.planetcatfish.com/common/species.php?species_id=729%0A

17. Armbruster JWW. The genus Peckoltia with the description of two new species and a reanalysis of the phylogeny of the genera of the Hypostominae (Siluriformes: Loricariidae). Zootaxa. 2008;(1822):1–76.

18. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma . 2014 Aug;30(15):2114–20.

19. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson D a, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–52.

20. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013 Aug;8(8):1494–512.

21. Emms DM, Kelly S. OrthoFinder : solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. Genome Biology; 2015;16:1–14.

22. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7 : Improvements in Performance and Usability Article Fast Track. Mol Biol Evol. 2013;30(4):772–80.

23. Suyama M, Torrents D, Bork P. PAL2NAL : robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34:609–12.

24. Kück P, Longo GC. FASconCAT-G : extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Front Zool. 2014;11(81):1–8.

25. Gouy M, Guindon S, Gascuel O. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Mol Biol Evol . 2010 Feb;27(2):221–4.

26. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.

27. Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: Proceedings of the Gateway Computing Environments Workshop (GCE). 2010. p. 1–8.

28. Yang Z. Computational molecular evolution. Oxford University Press; 2006. 357 p.

29. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 2007;24(8):1586–91.

30. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. R Foundation for Statistical Computing. Vienna, Austria: ISBN 3-900051-07-0; 2008. Available from: https://www.r-project.org

31. Zhang J, Nielsen R, Yang Z. Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. Mol Biol Evol. 2005;22(12):2472–9.

32. Meyer A, Schartl M. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. Curr Opin Cell Biol. 1999;11(6):699–704.

33. Zarkadis IK, Sarrias MR, Sfyroera G, Sunyer JO, Lambris JD. Cloning and structure of three rainbow trout C3 molecules: a plausible explanation for their functional diversity. Dev Comp Immunol. 2001;25(1):11–24.

34. Nakao M, Mutsuro J, Obo R, Fujiki K, Nonaka M, Yano T. Molecular cloning and protein analysis of divergent forms of the complement component C3 from a bony fish, the common carp (Cyprinus carpio): presence of variants lacking the catalytic histidine. Eur J Immunol. 2000;30(3):858–66.

35. Kuroda N, Naruse K, Shima A, Nonaka M, Sasaki M. Molecular cloning and linkage analysis of complement C3 and C4 genes of the Japanese medaka fish. Immunogenetics. 2000;51(2):117–28.

36. Forn-Cuní G, Reis ES, Dios S, Posada D, Lambris JD, Figueras A, et al. The evolution and appearance of C3 duplications in fish originate an exclusive teleost c3 gene form with anti-inflammatory activity. PLoS One. 2014;9(6):e99673.

37. Roxo FF, Albert JS, Silva GSC, Zawadzki CH, Foresti F, Oliveira C. Molecular phylogeny and biogeographic history of the armored neotropical catfish subfamilies hypoptopomatinae, neoplecostominae and otothyrinae (siluriformes: loricariidae). PLoS One. 2014;9(8):e105564.

38. Cramer CA, Bonatto SL, Reis RE. Molecular phylogeny of the Neoplecostominae and Hypoptopomatinae (Siluriformes: Loricariidae) using multiple genes. Mol Phylogenet Evol. 2011;59(1):43–52.

39. Rabosky DL, Santini F, Eastman J, Smith SA, Sidlauskas B, Chang J,

et al. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. Nat Commun. 2013;4:1958.

40. Lovette IJ, Bermingham E, Ricklefs RE. Clade-specific morphological diversification and adaptive radiation in Hawaiian songbirds. Proceedings Biol Sci. 2002;269(1486):37–42.

41. Martin CH, Wainwright PC. Trophic Novelty is Linked to Exceptional Rates of Morphological Diversification in Two Adaptive Radiations of Cyprinodon Pupfish. Evolution (N Y). 2011;65(8):2197–212.

42. Adams DC, Berns CM, Kozak KH, Wiens JJ. Are rates of species diversification correlated with rates of morphological evolution? Proc R Soc B Biol Sci. 2009;276(1668):2729–38.

43. Fujita MK, Leaché AD, Burbrink FT, McGuire JA, Moritz C. Coalescent-based species delimitation in an integrative taxonomy. Trends Ecol Evol. 2012;27(9):480–8.

44. Rudman SM, Barbour MA, Csilléry K, Gienapp P, Guillaume F, Hairston Jr NG, et al. What genomic data can reveal about eco-evolutionary dynamics. Nat Ecol Evol. 2018;2(1):9–15.