MINISTÉRIO DA SAÚDE
FUNDAÇÃO OSWALDO CRUZ
INSTITUTO OSWALDO CRUZ

Mestrado em Biologia Computacional e Sistemas

DESCRIÇÃO DA BIODIVERSIDADE MOLECULAR DE *Hypancistrus zebra* (LORICARIIDAE: SILURIFORMES), UMA ESPÉCIE DE PEIXE ORNAMENTAL AMEAÇADA DE EXTINÇÃO

MAITHÊ GASPAR PONTES MAGALHÃES

Rio de Janeiro
Março de 2018

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

‘

INSTITUTO OSWALDO CRUZ
Programa de Pós-graduação em Biologia Computacional e Sistemas

MAITHÊ GASPAR PONTES MAGALHÃES

Descrição da Biodiversidade Molecular de *Hypancistrus zebra* (Loricariidae: Siluriformes), uma Espécie de Peixe Ornamental Ameaçada de Extinção

Dissertação apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Mestre em Biologia Computacional e Sistemas

Orientadores: Prof. Dr. Thiago Estevam Parente
Profa. Dra. Ana Carolina Paulo Vicente

RIO DE JANEIRO
Março de 2018

Magalhães, Maithê Gaspar Pontes .

Descrição da Biodiversidade Molecular de Hypancistrus zebra (Loricariidae: Siluriformes), uma Espécie de Peixe Ornamental Ameaçada de Extinção / Maithê Gaspar Pontes Magalhães. - Rio de janeiro, 2018.
xiv, 57 f.; il.

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

Orientador: Thiago Estevam Parente.
Co-orientadora: Ana Carolina Paulo Vicente.

Bibliografia: Inclui Bibliografias.

1. Transcriptoma. 2. Genética da Conservação. 3. Genoma Mitocondrial. 4. Volta Grande do Xingu. 5. Agenda 2030. 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

AUTORA: Maithê Gaspar Pontes Magalhães

DESCRIÇÃO DA BIODIVERSIDADE MOLECULAR DE *Hypancistrus zebra* (LORICARIIDAE: SILURIFORMES), UMA ESPÉCIE DE PEIXE ORNAMENTAL AMEAÇADA DE EXTINÇÃO

ORIENTADORES: Prof. Dr. Thiago Estevam Parente
Profa. Dra. Ana Carolina Paulo Vicente

Aprovada em: 27/03/2018

EXAMINADORES:

Prof. Dr. Rafael Dias Mesquita **- Presidente** (Universidade Federal do Rio de Janeiro)
Prof. Dr. William Bryan Jennings (Universidade Federal do Rio de Janeiro)
Prof. Dr. André Elias Rodrigues Soares (Laboratório Nacional de Computação Científica)
Profa. Dra. Ana Carolina Guimarães (Fundação Oswaldo Cruz)
Profa. Dra. Nicole de Miranda Scherer (Instituto Nacional do Câncer)

Rio de Janeiro, 27 de março de 2018

Dedico este trabalho aos meus pais, Rosaria e Antônio, pelo apoio incondicional à minha formação acadêmica, à minha irmã, Rafaela, e ao meu noivo, Caio, por estarem sempre dispostos a me ajudar e por zelarem por mim.

AGRADECIMENTOS

Ao meu orientador, Thiago Parente, por todo auxílio, dedicação, paciência, na maioria das vezes, motivação, por se importar de verdade com minha formação acadêmica e por tornar nossa convivência leve, fácil e divertida. Ao meu amigo de laboratório, Daniel Moreira, por toda ajuda, que não foi pouca, pelo carinho e pelos momentos de descontração. À minha amiga de laboratório, graduação, iniciação científica e mestrado, Paula Andrade, por toda torcida, apoio, conversa, entusiamo, animação, mas, principalmente, por todo amor e carinho que demonstra nas pequenas atitudes do nosso dia a dia. Formamos uma excelente equipe.

Aos colegas do Laboratório de Toxicologia Ambiental, em especial ao Francisco Paumgartten e à Ana Cecília de Oliveira, por nos acolherem e por estarem sempre dispostos a ajudar.

Aos meus pais, Antônio e Rosaria, pelo amor e pelo apoio incondicional às minhas escolhas, por me incentivarem, acreditarem em mim e ajudar a tornar tudo possível. Eu sei que posso sempre contar com vocês. À minha irmã, Rafaela, por me ouvir sempre que eu precisava falar e por entender minhas preocupações, mas sempre torcendo pelo melhor. Ao meu noivo, Caio, por compreender minha ausência, impaciência e ansiedade, por me apoiar, por me acalmar, e, principalmente, por sempre ficar sempre ao meu lado. Aos meus avós, Neuza, Carlos e Edite, à minha madrinha, Ana Paula, e ao meu afilhado, Guilherme, pelo incentivo e força que em deram ao longo deste trabalho.

Aos meus amigos biólogos da UFRJ, Deborah, Vanessa e Rafael, por todo suporte, apoio, torcida, por todas as discussões enriquecedoras, principalmente, às que envolviam o evolucionismo, e por permancerem em minha vida.

Ao Instituto Oswaldo Cruz pelo ambiente agradável que proporciona. Aos seus funcionários pela dedicação e pelo excelente trabalho.

FINANCIAMENTO

Bolsa de Mestrado IOC

United States Agency for International Development (USAID, grants PGA-2000003446 e PGA-2000004790).

“I do not know what I may appear to the world, but to myself I seem to have been only like a boy [girl] playing on the seashore, and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me.”

Isaac Newton

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

DESCRIÇÃO DA BIODIVERSIDADE MOLECULAR DE *Hypancistrus zebra* (LORICARIIDAE: SILURIFORMES), UMA ESPÉCIE DE PEIXE ORNAMENTAL AMEAÇADA DE EXTINÇÃO

RESUMO

DISSERTAÇÃO DE MESTRADO EM BIOLOGIA COMPUTACIONAL E SISTEMAS

Maithê Gaspar Pontes Magalhães

Os impactos das atividades humanas na Terra são tão profundos que a proposta de definição de uma nova época geológica, o Antropoceno, tem sido amplamente debatida. Nesta nova época, a perda da biodiversidade destaca-se entre as principais características que afetam a saúde global. A construção de barragens para geração de energia hidrelétrica tem potencial de causar grande impacto na fauna local, especialmente em regiões com elevada biodiversidade e endemismo, como a Amazônia brasileira. O *Hypancistrus zebra* é uma espécie de peixe endêmico da Volta Grande do rio Xingu, na bacia amazônica, ameaçada de extinção devido ao impacto da construção da barragem da Usina de Belo Monte e à captura ilegal para aquariofilia internacional. Apesar disso, até o início desse trabalho, apenas duas sequências de nucleotídeos eram disponíveis no Genbank e no BOLD System, principais bancos de dados públicos desse tipo de informação. Neste trabalho, foram sequenciados sete transcriptomas de diferentes órgãos de *Hypancistrus zebra*. Produzimos mais de 200 milhões de leituras utilizadas para montar mais de meio milhão de transcritos. Neste banco de dados produzido, identificamos mais de 35 mil variantes de nucleotídeo único (SNVs) e quase quatro mil inserções e deleções (indels) distribuídos entre os transcriptomas dos sete órgãos de *H. zebra.* A partir da análise desses dados, desenvolvemos pares de iniciadores para amplificação de indels identificados em seis transcritos e de janelas contendo pelo menos três SNVs identificadas em sete outros transcritos. Sugerimos esse conjunto de transcritos como os mais adequados para aplicação em trabalhos visando a conservação dessa espécie. Foram encontrados elementos genéticos móveis de diversas famílias e expressos nos sete órgãos. A frequência de trancritos com elementos genéticos móveis variou de 12% no transcriptoma do coração a 33% na brânquia. Além disso, montamos o genoma mitocondrial, com 16.330 pb, dessa espécie. As informações e o banco de dados produzidos neste trabalho reduzem a lacuna de conhecimento sobre a diversidade genética do *Hypancistrus zebra* e podem ser usados para estudos de genética de população e da conservação dessa e de outras espécies filogenéticamente próximas. Essas informações, em especial às relacionadas a elementos genéticos móveis, também podem dar apoio à investigação sobre a variação cariotípica encontrada na família Loricariidae, da qual o *H. zebra* faz parte.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

DESCRIPTION OF THE MOLECULAR BIODIVERSITY OF *Hypancistrus zebra* (LORICARIIDAE: SILURIFORMES), AN ENDANGERED ORNAMENTAL FISH

ABSTRACT

MASTER DISSERTATION IN BIOLOGIA COMPUTACIONAL E SISTEMAS

Maithê Gaspar Pontes Magalhães

The impact of human activities on Earth is so deep that a new geological epoch, the Anthropocene, has been widely debated. In this new epoch, biodiversity loss stands out among the key features affecting global health. The construction of dams for hydroelectric power generation has the potential to have a major impact on local fauna, especially in regions with high biodiversity and endemism, such as the Brazilian Amazon. *Hypancistrus zebra* is a species of fish endemic to the Big Bend of the Xingu River, in the Amazon basin, threatened with extinction due to the impact of the Belo Monte Power Plant dam and illegal capture for international fish aquarium. However, until the beginning of this work, only two nucleotide sequences were available from Genbank and the BOLD System, the main public databases of this type of information. In this work, seven transcripts of different organs of *Hypancistrus zebra* were sequenced. We produced more than 200 million readings used to assemble over half a million transcripts. In the generated database, we identified more than 35,000 single nucleotide variants (SNVs) and almost four thousand insertions and deletions (indels) distributed among the seven organs of *H. zebra*. From the analysis of these data, we developed pairs of primers for amplification of indels identified in six transcripts and of frames with at least three SNVs identified in seven other transcripts. We suggest this set of transcripts as the most suitable for application in works aiming the conservation of this species. Mobile genetic elements of several families were found and expressed in the seven organs. The frequency of transcripts with mobile genetic elements varied from 12% in the transcriptome of the heart to 33% in the gill. In addition, we assembled the mitochondrial genome, with 16,330 bp, of this species. The information and the database produced in this work reduces the knowledge gap on the genetic diversity of *Hypancistrus zebra* and can be used for population genetics studies and the conservation of this and other phylogenetically close species. This information, especially those related to genetic genetic elements, can also support research on the karyotype variation found in the Loricariidae family, of which *H. zebra* is a part.

ÍNDICE

RESUMO.....viii
ABSTRACT.....ix
1. INTRODUÇÃO.....1
1.1. Geral.....5
1.2. Específicos.....5
2. MATERIAL E MÉTODOS.....6
2.1. Coleta e sequenciamento.....6
2.2. Montagem e anotação dos transcriptomas.....7
2.3. Montagem do Genoma Mitocondrial.....9
2.4. Desenvolvimento de Marcadores Genéticos.....10
3. RESULTADOS.....12
3.1. Avaliação da qualidade das leituras do Illumina.....12
3.2. Montagem dos sete transcriptomas.....12
3.3. Montagem do genoma mitocondrial.....18
3.4. Desenvolvimento de Marcadores Moleculares para *Hypancistrus zebra*.....24
3.5. Identificação de Elementos Genéticos Móveis.....38
4. DISCUSSÃO.....40
5. CONCLUSÃO.....45
6. REFERÊNCIAS BIBLIOGRÁFICAS.....46
7. ANEXOS.....52
7.1. Anexo 1 - Lista de Comandos.....52
7.2. Anexo 2 - Gráficos dos Valores de “Phred” dos seis órgãos adicionais do *H. zebra*…54
7.3. Anexo 3 - Profundidade média de sequenciamento.....57
7.4. Anexo 4 – Artigo publicado com alguns dos dados desse trabalho.....57

ÍNDICE DE FIGURAS

Figura 1: Modelo de estudo usado neste trabalho, *Hypancistrus zebra*.....3

Figura 2: Etapas da metodologia usada nesta dissertação.....6

Figura 3: Avaliação da qualidade das leituras brutas sequenciadas no transcriptoma da brânquia.....13

Figura 4: Relação entre cobertura e profundidade de sequenciamento para cada órgão.....17

Figura 5: Profundidade de sequenciamento do genoma mitocondrial de *Hypancistrus zebra*.....19

Figura 6: Interseção dos transcritos que contêm pelo menos três SNVs.....26

Figura 7: Interseção dos transcritos que contêm indels.....33

Figura 8: Distribuição de elementos genéticos móveis nos sete órgãos de *Hypancistrus zebra*.....39

LISTA DE TABELAS

Tabela 1: Identificadores de depósitos realizados nesse trabalho.....8

Tabela 2: Informações gerais sobre os transcriptomas montados.....14

Tabela 3: Informações gerais sobre anotações dos transcritos montados.....15

Tabela 4: Sítios heteroplasmáticos no genoma mitocondrial de Hy*pancistrus zebra*.....20

Tabela 5: Total de SNVs e indels encontrados em cada um dos sete órgãos.....24

Tabela 6: Anotação de transcritos selecionados com SNVs.....27

Tabela 7: Variantes de nucleotídeo único (SNVs) selecionados.....28

Tabela 8: Anotação dos transcritos selecionados com indels.....34

Tabela 9: Indels selecionados.....35

Tabela 10: Sugestão de iniciadores para amplificação por PCR das regiões de SNV e Indel.....37

Tabela 11: Proporção do transcriptoma, em pares de base, que codifica elementos genéticos móveis em cada órgão.....38

LISTA DE SIGLAS E ABREVIATURAS

ALT	Alternativo
ANEEL	Agência Nacional de Energia Elétrica
BLAST	“Basic Local Alignment Search Tool”
cDNA	DNA complementar
CDS	“Coding DNA Sequence”
CITES	“Convention on International Trade in Endangered Species of Wild Fauna and Flora”
COI	“Cytochrome c oxidase subunit I”
COII	“Cytochrome c oxidase subunit II”
COIII	“Cytochrome c oxidase subunit III”
CytB	“Cytochrome b”
DNA	“Deoxyribonucleic acid”
EGM	Elementos Genéticos Móveis
Fiocruz	Fundação Oswaldo Cruz
FPKM	“Fragments Per Kilobase Million”
GATK	“Genome Analysis Toolkit”
IBAMA	Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis
ID	Identidade
IGV	Do inglês, “Integrated Genome Viewer”
indel	Inserção ou Deleção
INPA	Instituto Nacional de Pesquisa da Amazônia
MMA	Ministério do Meio Ambiente
Não carac.	Não caracterizada
NCBI	“National Center for Biotechnology Information”
ND5	“Nicotinamide Adenine Dinucleotide Ubiquinone Oxidoreductase Chain 5”
ND6	“Nicotinamide Adenine Dinucleotide Ubiquinone Oxidoreductase Chain 6”
NGS	“Next-Generation Sequencing”
ORFs	“Open Reading Frames”
pb	Pares de base
PCR	“Polymerase Chain Reaction”
POS	Posição
qCov	“Query Coverage”
Quant.	Número total de leituras
REF	Referência

RNA	“Ribonucleic Acid”
rRNA	“Ribosomal RNA”
RSEM	“RNA-Seq by Expectation-Maximization”
SNV	“Single nucleotide variants”
SRA	“Short Read Archive”
TMM	“Trimmed mean of M values”
TPM	“Transcripts Per Kilobase Million”
tRNA	“transfer RNA”
tRNA-Glu	“tRNA-Glutamine”
tRNA-Ser	“tRNA-Serine”

1. INTRODUÇÃO

Os impactos das atividades humanas na Terra são tão profundos que a proposta de definição de uma nova era geológica, o Antropoceno, tem sido amplamente debatida (1,2). O termo Antropoceno é usado para se referir à época atual, em que muitos processos e condições geológicos são profundamente alterados pelas atividades humanas. De acordo com o projeto “Living Planet Index” (3), que fornece uma medida do estado da biodiversidade global com base nas tendências populacionais de espécies de vertebrados, estima-se que houve uma perda de 52% de indivíduos em 40 anos (4). Recordes de extinção para répteis, anfíbios, peixes de água doce têm sido documentados desde o início do século XX (5). Essa perda da biodiversidade é um dos problemas ambientais mais críticos, pois ameaça o fornecimento de serviços ecossistêmicos e o bem estar dos humanos (5,6). Por exemplo, de 23 a 36% dos mamíferos, das aves e dos anfíbios usados para alimentação ou na medicina estão ameaçados de extinção, afetando especialmente populações humanas em locais mais pobres (6). Minimizar esse cenário está entre os objetivos da Agenda 2030 da ONU (7) e em sintonia com a “Estratégia Fiocruz para a Agenda 2030”.

O rápido crescimento da população humana é apontado como uma das principais causas desses impactos. Em apenas 200 anos, de 1800 a 2000, a população mundial cresceu aproximadamente de um bilhão para seis bilhões e meio de pessoas e estima-se que alcance nove bilhões até 2050 (1). O crescimento populacional acelerado está diretamente ligado à Revolução Industrial e, consequentemente, às mudanças no meio ambiente, já que se faz necessário o sustento desse crescente número de pessoas. A Revolução Industrial, nos anos de 1700, e a revolução da termo-indústria, no século XIX, marcou o fim da agricultura como a atividade humana dominante e colocou as espécies em uma trajetória diferente daquela estabelecida durante o Holoceno. A forma como a humanidade afeta o meio ambiente mudou em relação ao tipo de uso dos recursos naturais e em relação à dimensão do consumo dos recusos naturais. A partir de então o crescente gargalo de energia (2) e o impacto negativo sobre a biodiversidade do globo (4) tornaram-se características relevantes para o desenvolvimento da sociedade.

O crescimento da população humana, o desenvolvimento econômico, as mudanças climáticas e a necessidade de fechar a lacuna de acesso à eletricidade estimularam o uso de fontes de energia renovável, em contraposição ao uso de carvão. A energia hidroelétrica foi uma das primeiras alternativas desenvolvidas e amplamente implementadas, especialmente no Brasil, devido a ampla rede hidrográfica naturalmente disponível. Em torno de 3.700 barragens estão planejadas ou em construção, principalmente em países com economias emergentes (8). No Brasil, de acordo com a Agência Nacional de Energia Elétrica (ANEEL) existem 1.308 empreendimentos hidrelétricos (219 Usinas Hidrelétricas, 660 Centrais Geradoras Hidrelétricas e 427 Pequenas Centrais Hidrelétricas) em operação, que juntos são responsáveis por aproximadamente 64% da geração de energia elétrica do país. As construções de novas usinas hidrelétricas têm se concentrado na bacia Amazônica, como as usinas Jirau, Santo Antônio e Belo Monte que foram recentemente construídas (9).

O impacto ambiental causado pela construção de barragens é tão grande que recentemente foi estimado que nenhuma outra atividade antropogênica provocou mais alterações em ecossistemas de água doce (10). O mesmo estudo mostrou que dois terços dos maiores rios do mundo foram divididos por barragens (10). As barragens bloqueiam o fluxo natural do rio, afetando a migração de espécies, a concentração de oxigênio, a temperatura e as condições de sedimentação nos reservatórios e à jusante do rio (9). A Amazônia tem se tornado sinônimo de desenvolvimento de barragens, já que a água que corre nos 6,8 milhões de km² da bacia representa cerca de 18% da descarga de água de todo planeta (10). Essa região, todavia, apresenta uma das maiores taxas de biodiversidade e é caracterizada pelo alto grau de endemismo (11). Um exemplo disso é a Volta Grande do rio Xingu, onde é encontrada a espécie de peixe *Hypancistrus zebra*, pertencente à família Loricariidae.

O *Hypancistrus zebra* é uma espécie de peixe ornamental endêmica de um trecho de aproximadamente 100 km, conhecido como “Volta Grande”, do rio Xingu, no estado do Pará. Seu hábitat está localizado na zona de impacto da quarta maior hidrelétrica do mundo, a Usina de Belo Monte, que foi inaugurada em meados de 2016. Com a construção da barragem da hidrelétrica de Belo Monte, o fluxo de água na Volta Grande chegará a apenas 20% do fluxo natural (10). Quando o reservatório

encher e o nível de água na Volta Grande estiver baixo, acredita-se que os pedrais onde o *Hypancistrus zebra* vive ficarão rasos e quentes demais para sua sobrevivência. Como essa espécie só ocorre nessa região, corre risco de ser extinta. Além dessa ameaça, em fevereiro de 2017, a Secretaria do Meio Ambiente e Sustentabilidade do Estado do Pará concedeu a Licença de Instalação para a empresa Belo Sun Mineração LTDA construir, na Volta Grande do Xingu, a maior mina de ouro a céu aberto do país (12). Outra ameaça para o *Hypancistrus zebra* é sua captura para o mercado clandestino de peixes ornamentais, ainda que a Instrução Normativa No 5, de 21 de maio de 2004 proíba a captura dessa espécie (MMA, 2004). Essas ameaças levaram o *H. zebra* para as listas de espécies ameaçadas de extinção do Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis (IBAMA) e do "Convention on International Trade in Endangered Species of Wild Fauna and Flora" (CITES). Apesar de seu status de ameaçado de extinção, até o início desse projeto, existiam apenas duas sequências de nucleotídeos disponíveis no Genbank e no BOLD System, principais bancos de dados públicos desse tipo de informação. Uma dessas sequências codifica 1.788 pb do gene nuclear F-reticulon (rtn4). A outra sequência contém 2.410 pb do genoma mitocondrial, incluindo parte dos genes ribossomais 12S e 16S e o gene completo de tRNA-Val.

Foto: Thiago Parente

Figura 1: Modelo de estudo usado neste trabalho, *Hypancistrus zebra*. Espécie ameaçada de extinção endêmica da Volta Grande do Rio Xingu, PA, Brasil.

O uso das novas tecnologias de sequenciamento de ácidos nucléicos pode preencher de forma relativamente rápida, simples e com baixo custo essa falta de conhecimento sobre a genética de espécies nativas. As tecnologias de alto desempenho para o sequenciamento de DNA e RNA (NGS) geram um grande volume de informação genética ("Big Data"), possibilitando a observação e a descrição da natureza ao nível molecular. A geração desse grande volume de

dados, possibilitou a ascensão da ciência exploratória (“discovery science”) e consequentemente mais sistemas moleculares foram e estão sendo descritos. Ao contrário da ciência guiada por hipóteses, a ciência exploratória não é direcionada por perguntas específicas (13). A partir da observação da genética da natureza, padrões e quebras desses padrões podem ser descritos e usados para a posterior formulação de novas hipóteses e perguntas. Para possibilitar o processamento e análise desse grande volume de dados, inúmeras ferramentas de biologia computacional são desenvolvidas para visualizar e modelar computacionalmente os dados de NGS para caracterização de sistemas biológicos (13).

A genética da conservação usa marcadores genéticos na preservação da biodiversidade e no manejo de espécies e populações (14). A popularização do uso das tecnologias de sequenciamento de alto desempenho vem possibilitando o desenvolvimento de marcadores genéticos em uma escala genômica. Em animais, o genoma mitocondrial é circular e normalmente contém 37 genes, sendo dois de rRNAs, 22 de tRNAs e 13 codificantes de proteínas. Além disso, contém uma sequência não codificante chamada de região controle, devido ao seu papel na replicação e transcrição de genes mitocondriais. Já o genoma nuclear é maior e mais complexo e, portanto, uma fonte ainda mais rica de marcadores genéticos. Dentre os principais usos desses marcadores estão a estruturação da população, resolução de relações filogenéticas e detecção de caça ou coleta ilegal de animais ameaçados de extinção (15).

Os elementos genéticos móveis (EGM) são os componentes mais abundantes dos genomas de vertebrados, desempenhando papeis na arquitetura e evolução dos genomas (16). A variação cariotípica observada entre muitos grupos de animais vem sendo associada à atividade de alguns desses EGM (17). A variação cariotípica na família Loricariidae é surpreendente, com o número diploide de cromossomos variando de 34 a 96, e EGM já foram associados a rearranjo cromossômico em algumas espécies dessa família (18).

Neste trabalho, dados provenientes do sequenciamento de alto desempenho de RNA foram usados para montar sete transcriptomas de *Hypancistrus zebra*, a partir dos quais montamos o genoma mitocondrial dessa espécie, desenvolvemos marcadores populacionais e identificamos a expressão de elementos genéticos móveis.

OBJETIVOS

1.1. Geral

O principal objetivo desta dissertação é identificar candidatos a marcadores populacionais no banco de dados genéticos de *Hypancistrus zebra* (Loricariidae) produzido neste trabalho.

1.2. Específicos

1. Montar transcriptomas de sete órgãos de *Hypancistrus zebra*;
2. Anotar os sete transcriptomas montados na etapa anterior;
3. Montar o genoma mitocondrial dessa espécie usando os transcritos mitocondriais identificados nos transcriptomas;
4. Identificar transcritos que contenham SNVs e INDELs com maior potencial de aplicação como marcadores genéticos;
5. Analisar o banco de dados genéticos produzido em busca de elementos genéticos móveis.

2. MATERIAL E MÉTODOS

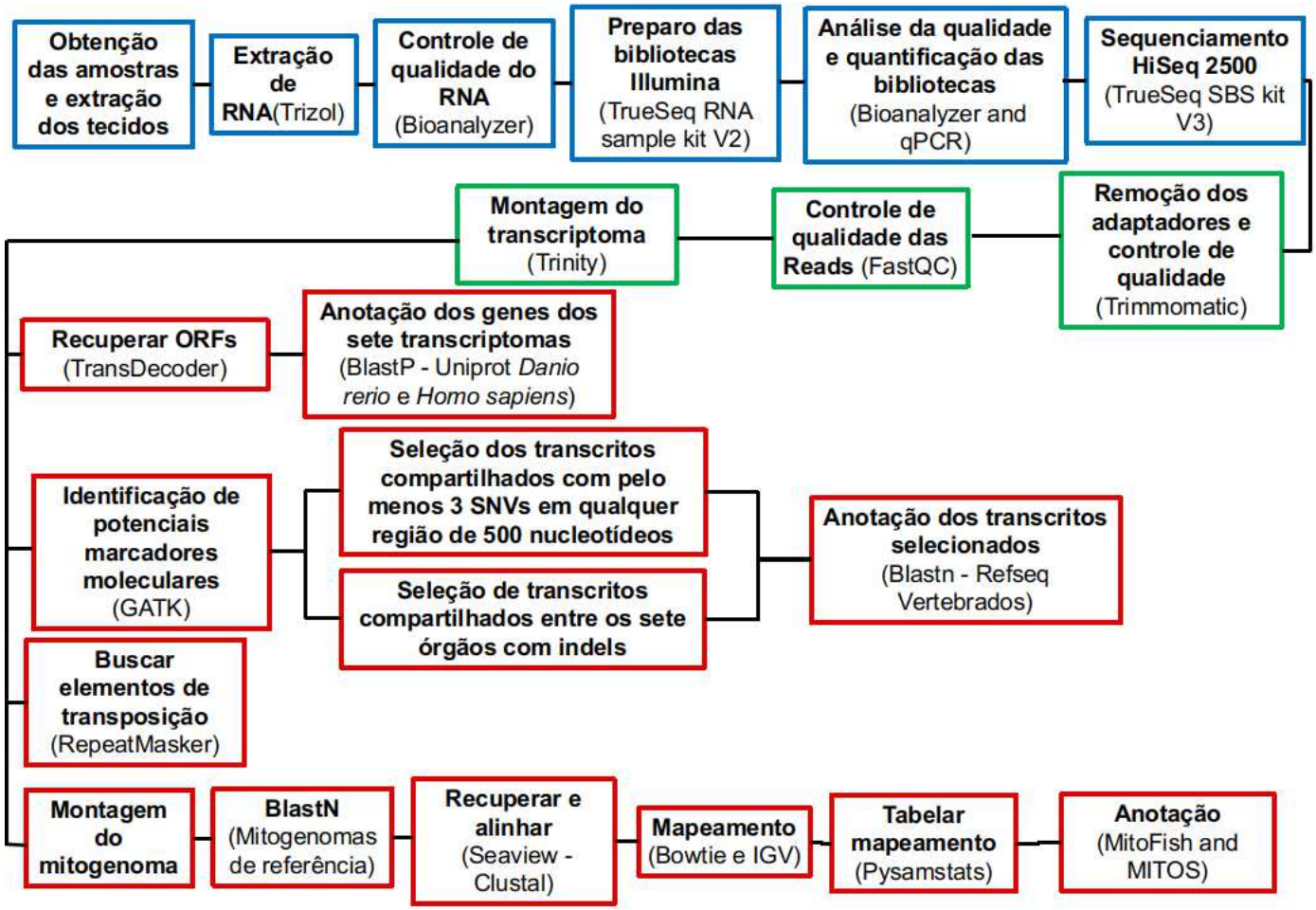


Figura 2: Etapas da metodologia usada nesta dissertação. As molduras em azul representam as etapas de metodologia da bancada molhada; em verde, as etapas de processamento computacional inicial dos dados; e em vermelho, a metodologia de análise computacional dos transcriptomas montados.

2.1. Coleta e sequenciamento

Dois espécimes de *Hypancistrus zebra* foram usados neste trabalho (TP166 e TP167), ambos doados pelo Dr. Jansen Zuanon do Instituto Nacional de Pesquisa da Amazônia (INPA) e originários de uma apreensão feita pela Polícia Federal Brasileira. Os peixes foram depositados na coleção ictiológica do INPA sob número de voucher INPA 46655. Amostras de sete órgãos (fígado, brânquia, intestino, rim, gônada, cérebro e coração) foram extraídas e armazenadas em RNA later (Life Technologies) a -20°C. O RNA total foi extraído de seis órgãos de um espécime (TP166), e de um órgão (coração) do outro espécime (TP167). As extrações de RNA

foram feitas pelo método fenol/clorofórmio usando Trizol (Invitrogen), de acordo com as instruções do fabricante. Em seguida, a quantificação de RNA foi feita usando espectrofotômetro (Biodrop) e sua qualidade foi testada usando o kit de Bioanalyzer RNA 6000 Nano (Agilent).

As bibliotecas de DNA complementar (cDNA) de cada órgão foram preparadas usando 1.000 ng de RNA total, seguindo rigorosamente as instruções do kit TrueSeq RNA Sample v2 (Illumina). Cada biblioteca foi identificada utilizando adaptadores específicos. A qualidade das bibliotecas foi verificada usando o kit DNA 1000 para Bioanalyzer (Agilent). As bibliotecas foram quantificadas por qPCR utilizando o kit de quantificação para Illumina (Kapa Biosystems). As sete bibliotecas foram agrupadas na mesma linha de sequenciamento, utilizando o kit TrueSeq PE Cluster v3 (Illumina). A reação de sequenciamento de 100 pb das duas extremidades de cada fragmento ("paired-end") foi realizada usando o kit TrueSeq SBS v3 (Illumina) no equipamento HiSeq2500 do Instituto Nacional do Câncer (INCA).

2.2. Montagem e anotação dos transcriptomas

Para os dados da biblioteca de cada órgão, as sequências dos adaptadores e as de baixa qualidade foram removidas das leituras brutas utilizando os parâmetros padrões da ferramenta Trimmomatic (19) (Anexo 1), e a qualidade dos conjuntos de leituras aproveitadas foi avaliada utilizando o programa FastQC (Babraham Bioinformatics) (Anexo 1). O conjunto de dados brutos das leituras dos transcriptomas sequenciados de cada órgão foi depositado no banco de dados "Short Read Archive" (SRA), do "National Center for Biotechnology Information" (NCBI).

Tabela 1: Identificadores de depósitos realizados nesse trabalho. São mostrados os códigos de depósito na coleção ictiológica do Instituto Nacional de Pesquisas da Amazônia (INPA), a identificação de campo de cada indivíduo e os identificadores de depósito dos dados brutos no SRA ("Short Read Archive").

Voucher	ID Individual	Órgão	SRA
INPA 46655	**TP166**	**Rim**	SRR6808874
		Brânquia	SRR6808875
		Gônada	SRR6808876
		Cérebro	SRR6808877
		Intestino	SRR6808878
		Fígado	SRR6808879
	TP167	**Coração**	SRR6808880

As leituras editadas de cada órgão foram utilizadas para a montagem *de novo* dos transcriptomas usando o parâmetro padrão do programa de montagem Trinity (v.2.0.6) (20) (Anexo1). N50 é uma medida de qualidade para montagens *de novo*, e foi utilizada neste trabalho a fim de caracterizar os transcriptomas montados. O N50 é o valor referente ao tamanho do transcrito que divide em duas partes iguais a distribuição de transcritos no transcriptoma montado, quando esses estão ordenados pelo comprimento medido em pares de base. Os transcriptomas montados tiveram os quadros abertos de leitura (do inglês, "open reading frames" - ORFs) recuperados com o uso do programa TransDecoder-3.0.0 (21) (Anexo 1).

Os ORFs recuperados foram submetidos a uma busca por similaridade utilizando o algoritmo BLASTP (e-value < 1e-10) (Anexo 1) contra a base de dados Uniprot de *Danio rerio* e Uniprot de *Homo sapiens*. O melhor alinhamento de cada transcrito montado foi usado para sua anotação. A recuperação e a anotação de elementos de transposição nos sete transcriptomas montados foram realizadas usando o programa RepeatMasker 4.0.7 (22), com a ferramenta de busca rmblastn 2.2.28 e o banco de dados RepBase RepeatMasker Edition 2017.01.27.

2.3. Montagem do Genoma Mitocondrial

O genoma mitocondrial foi montado utilizando transcritos de genes mitocondriais sequenciados nos transcriptomas dos dois espécimes, usando uma abordagem desenvolvida pelo nosso grupo (23). Esses transcritos mitocondriais foram recuperados executando-se um BLASTN contra mitogenomas de espécies filogeneticamente próximas e disponíveis em banco de dados públicos. A maioria desses genomas mitocondriais usados como referência foram recentemente sequenciados, montados e depositados pelo nosso grupo (24). Os transcritos mitocondriais recuperados foram editados no SeaView (25) de acordo com a informação de orientação da cadeia fornecida pelo resultado do BLASTN e alinhados contra o mitogenoma de *Pterygoplichthys disjunctivus* (NC_015747.1) usando o algoritmo de alinhamento CLUSTAL implementado no programa SeaView (25). O genoma mitocondrial montado foi depositado no GenBank (KX611143), e anotado usando os programas MitoFish (26) e MITOS (27).

A fim de estimar a profundidade de sequenciamento de cada base no mitogenoma, a ferramenta Bowtie (v.1.0.0) (28) foi usada para alinhar as leituras no mitogenoma montado. As leituras alinhadas foram visualizadas usando o programa Integrated Genome Viewer (IGV) (29) e tabeladas usando o software Pysamstats (30). O critério para classificação de variações na sequência de nucleotídeos como sítios heteroplasmáticos seguiu o estabelecido por MOREIRA et al., 2015. Em síntese, foram consideradas heteroplasmáticas apenas as posições com 100 ou mais leituras mapeadas cuja frequência da segunda base mais abundante foi igual ou maior que 10%. Dessa forma, garante-se um suporte de no mínimo 10 leituras para o nucleotídeo variante.

A fim de estimar os níveis de expressão dos transcritos mitocondriais em cada transcriptoma, o programa RNA-Seq by Expectation-Maximization (RSEM) (31) foi usado para quantificar o número de fragmentos por quilobase de exon por milhão de fragmentos mapeados (FPKM). Esses valores foram normalizados usando o Trimmed mean of M values (TMM) (32).

2.4. Desenvolvimento de Marcadores Genéticos

Para a identificação de marcadores genéticos usamos critérios mais rígidos de processamento inicial das leituras de forma a minimizar as chances de manuteção de eventuais erros de sequenciamento. Dessa forma, ao invés dos parâmetros padrões do software Trimmomatic, a edição das leituras foi feita de forma a aproveitar apenas aquelas que apresentaram uma média de qualidade dos nucleotídeos acima de 28, os parâmetros modificados foram SLIDINGWINDOW:4:28, LEADING:28 e TRAILING:28. De forma similar, para a detecção de marcadores genéticos foi montado um único transcriptoma com as leituras concatenadas dos sete órgãos, mantendo a separação entre leituras senso e anti-senso. A montagem desses transcriptomas também foi realizada usando os parâmetros padrões do Trinity (v.2.0.6).

Os conjuntos de leituras pareadas de cada órgão foram mapeados contra o transcriptoma utilizando o TopHat (v2.0.14) (33) (Anexo 1). Os arquivos bam de saída do TopHat foram preparados para a chamada de variantes utilizando os programas Bowtie2-build (34), SAMtools 1.5 (35) e ferramentas do programa Picard (v.2.10.3-SNAPSHOT) (36). Informações mais detalhadas no Anexo 1.

O programa Genome Analysis Toolkit (GATK v.3.7-0-gcfedb67) (37) foi utilizado para busca de variantes de nucleotídeo único (do inglês, “single nucleotide variants” - SNV) e de inserções ou deleções (indel), de acordo com as recomendações de boas práticas do Genome Analysis Toolkit (GATK) (38,39). Em resumo, os valores de qualidade do mapeamento foram ajustados de acordo com os parâmetros do GATK (Anexo 1), em seguida foi feita a chamada de variantes. Esses variantes foram separados em dois arquivos de acordo com o tipo (SNV ou indel), e por último fizemos uma filtragem desses variantes utilizando o parâmetro padrão para dados de RNA-seq e para as chamadas de variantes sem a utilização do VariantRecalibrator, devido a ausência de uma referência de conjunto de variantes para recalibrar nossos dados (Anexo 1).

Para refinar as listas de SNVs e indels geradas pelo GATK, estabelecemos uma série de critérios com o intuito de aumentar a confiabilidade na identificação de SNVs ou indel e otimizar sua potencial aplicação em trabalhos de genética da conservação. Para o refinamento da lista de SNVs, inicialmente identificamos os

transcritos com pelo menos três SNVs em qualquer região de 500 nucleotídeos. Em seguida, desses transcritos, foram selecionados apenas aqueles cujo conjunto de SNVs foi compartilhado entre os sete órgãos. Como todos os transcritos identificados contendo pelo menos três SNVs em qualquer região de 500 nucleotídeos eram comuns aos sete órgãos tinham anotação, não foi preciso estabelecer um critério específico para ausência ou presença de anotação. Já para o refinamento da lista de indels, selecionamos os transcritos comuns a todos os sete órgãos que apresentaram pelo menos um indel. Esses transcritos comuns aos sete órgãos foram anotados contra o Refseq de vertebrados e aqueles que tiveram anotação com cobertura maior que 70%, E-value igual a 0 e identidade maior que 80% foram selecionados. Desses, foram escolhidos apenas aqueles com indels na mesma posição para os sete órgãos.

A ferramenta Primer-BLAST (40) foi utilizada para buscar candidatos a iniciadores para reação em cadeia da polimerase (PCR) nas sequências dos transcritos identificados na etapa anterior. Os melhores pares de iniciadores foram analisados usando a ferramenta OligoAnalyzer 3.1 (41).

3. RESULTADOS

3.1. Avaliação da qualidade das leituras do Illumina

A qualidade das leituras illumina foi verificada antes e após a etapa de edição, em que as sequências dos adaptadores e as de baixa qualidade foram removidas das leituras brutas. A maioria das leituras brutas teve valor de Phred acima de 30 (Figura 3 e Anexo 1), assim a taxa de erro esperada na chamada do nucleotídeo é de 1 em 1.000, além de demonstrar a alta qualidade do sequenciamento.

3.2. Montagem dos sete transcriptomas

Aproximadamente 98% das leituras sequenciadas dos transcriptomas dos sete órgãos passaram pelos critérios usados na etapa de edição e, portanto, foram usadas para a montagem *de novo* dos transcritos (Tabela 2). Nessa etapa, foi montado um transcriptoma para cada um dos sete órgãos. Apesar do fígado ter o maior número de leituras sequenciadas aproveitadas (50.171.177), não foi o órgão com mais transcritos montados. Os transcriptomas do cérebro e do rim tiveram o número de transcritos montados muito próximo, 101.507 e 110.113 respectivamente, apesar da diferença de mais de 14 milhões de leituras aproveitadas. Da mesma forma, o coração e a gônada tiveram praticamente o mesmo número de transcritos montados, apesar da diferença de quase 30 milhões de leituras aproveitadas (Tabela 2).

O transcriptoma do rim foi o que teve mais transcritos montados (110.113) e anotados tanto contra o banco de dados Uniprot de *Danio rerio* (31.337) quanto contra o banco de dados Uniprot de *Homo sapiens* (29.099). Já o trancriptoma da gônada apresentou o menor número de transcritos montados (48.214), enquanto o do coração teve o menor número de transcritos anotados contra o banco de dados de ambas espécies (Tabela 2 e 3). O número de transcritos anotados contra as entradas de *Danio rerio* no Uniprot foi maior nos sete órgãos em relação ao obtido contra as entradas de *Homo sapiens* no Uniprot. Entretanto, a grande maioria dos transcritos anotados contra as sequências de *D. rerio* foi anotada como proteína não caracterizada, em contraste com o observado na anotação contra *H. sapiens* (Tabela 3).

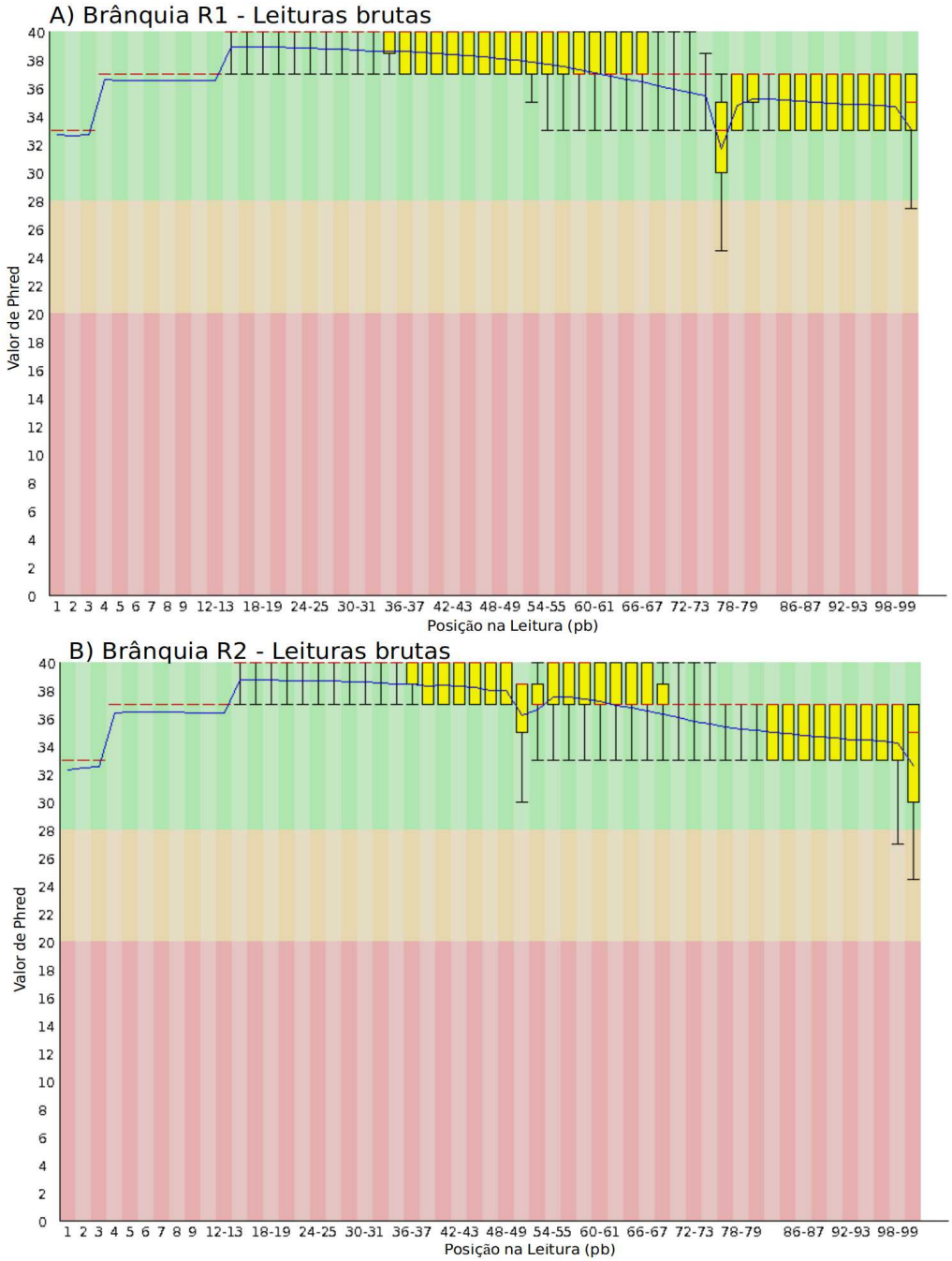


Figura 3: Avaliação da qualidade das leituras brutas sequenciadas no transcriptoma da brânquia. O valor de Phred (eixo y) é classificado em três partes de acordo com a qualidade de sequenciamento de cada nucleotídeo (eixo x). O fundo verde delimita o intervalo do valor de Phred considerado muito bom (28>Phred>40). O valor de Phred é uma medida de qualidade da identificação das bases geradas pelo sequenciamento automatizado. Um valor de Phred igual a 30 indica que a taxa de erro na chamada do nucleotídeo é de 1 em cada 1.000. A avaliação da qualidade das leituras dos demais transcriptomas é mostrada no Anexo 1.

Tabela 2: Informações gerais sobre os transcriptomas montados. Somatório das leituras senso e anti-senso antes e depois das sequências dos adaptadores e as de baixa qualidade serem removidas. Também são mostrados a profundidade média, mediana, mínima e máxima dos transcritos de cada transcriptoma e número de transcritos montados, assim como N50 de cada transcriptoma, uma medida da qualidade da montagem.

	Gônada	Coração	Brânquia	Fígado	Intestino	Cérebro	Rim
Leituras							
brutas	38.911.114	9.527.402	13.040.270	50.737.902	40.047.150	25.669.222	39.672.788
aproveitadas	38.445.295	9.388.445	12.881.333	50.171.177	39.581.476	25.047.519	39.058.249
% aproveitadas	98,8	98,5	98,8	98,9	98,8	97,6	98,5
Profundidade							
média	90,48	19,06	17,49	55,63	39,18	23,35	34,81
mediana	8,44	4,05	4,79	5,80	5,68	5,58	6,62
min	0,133	0,337	0,160	0,100	0,163	0,169	0,171
max	152.710,00	15.848,47	15.063,79	108,445,2	4.2571,13	25.110,92	5.7330,25
Transcritos							
montados	48.214	48.802	74.283	84.407	99.281	101.507	110.113
N50	2037	1016	1434	1662	1766	1535	1931

Tabela 3: Informações gerais sobre anotações dos transcritos montados. Número de transcritos anotados, contra o Uniprot de *Danio rerio* e *Homo sapiens.* A anotação nos bancos de dados é detalhada quanto ao número total de transcritos que não foram anotados, número total de transcritos anotados e, desses, quais foram anotados como proteína não caracterizada.

	Gônada	Coração	Brânquia	Fígado	Intestino	Cérebro	Rim
***Danio rerio* Uniprot**							
Não anotados	26.493	32.153	49.199	57.385	69.696	72.852	78.776
Anotados	21.721	16.649	25.084	27.022	29.585	28.655	31.337
Não caracterizados	12.073	8.503	14.182	15.158	16.979	16.177	17.879
***Homo sapiens* Uniprot**							
Não anotados	27.550	33.361	51.378	59.327	71.895	74.860	81.014
Anotados	20.664	15.441	22.905	25.080	27.386	26.647	29.099
Não caracterizados	230	141	199	211	285	232	303

As medianas da profundidade média de sequenciamento do transcrito (Anexo 3) de cada transcriptoma foram bem próximas, variando de 4,05 a 8,44, ao contrário da média que variou de 17,49 a 90,48 (Tabela 2).

A Figura 4 mostra, para o transcriptoma de cada órgão, como varia a razão do comprimento de cada quadro aberto de leitura identificado pelo comprimento do seu melhor hit em *Danio rerio* em função da profundidade média de sequenciamento (Anexo 3). O ponto destacado em vermelho em cada gráfico corresponde ao seu ponto mediano, ou seja, as medianas tanto da razão de tamanho como de profundidade de sequenciamento. Da análise da Figura 4 depreende-se que os transcritos anotados do transcriptoma da gônada apresentaram maior mediana da razão do tamanho de seus homólogos, além da maior mediana da profundidade média de sequenciamento (Figura 4). A maior frequência de transcritos que cobrem toda a região codificante de seu homólogo (razão transcrito/homólogo > 1) foi encontrada na gônada (32%) e a menor frequência foi encontrada na brânquia (18%). O ponto mediano no gráfico representa a mediana da profundidade média de sequenciamento e a mediana da razão dos comprimentos da CDS do *H. zebra* e da CDS de *D. rerio*. Os pontos medianos dos gráficos dos sete órgãos ficaram bem próximos à 1 em relação à cobertura (Tabela 2).

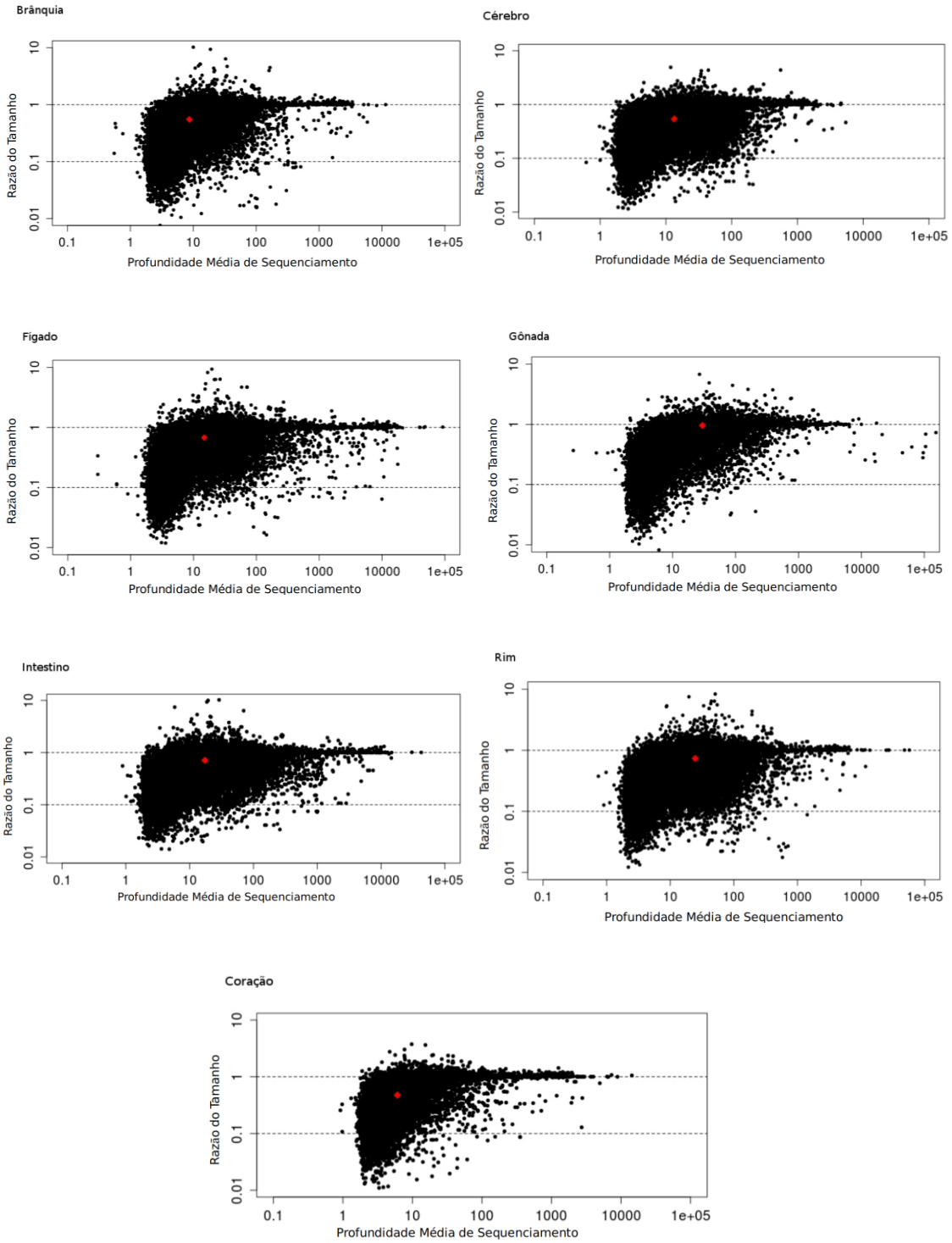


Figura 4: Relação entre cobertura e profundidade de sequenciamento para cada órgão. Em cada gráfico, a profundidade média de sequenciamento de cada transcrito (eixo x) dos transcriptomas dos sete órgãos analisados de *Hypancistrus zebra* é mostrada em relação à razão entre o comprimento de sua sequência codificante (CDS) com o comprimento da CDS do seu homólogo em *Danio rerio* (eixo y). O ponto vermelho representa o ponto mediano do gráfico, que representa a mediana da profundidade média de sequenciamento e a mediana da cobertura.

3.3. Montagem do genoma mitocondrial

Transcritos mitocondriais dos dois espécimes de *H. zebra* foram usados para montar o genoma mitocondrial quase completo, com 16.330 pb de comprimento. A composição do mitogenoma é a usual de vertebrados (2 rRNA, 22 tRNA e 13 genes codificantes de proteína). Todas as sequências de genes codificantes foram recuperadas completas (Figura 5A). Comparando com os genomas mitocondriais disponíveis de alguns outros loricarídeos, estima-se que apenas uma pequena porção de aproximadamente 300 pb da região D-loop esteja faltando.

A profundidade de sequenciamento variou de acordo com a posição no mitogenoma e também entre os diferentes órgãos. Enquanto os genes codificantes de proteínas tiveram maior profundidade de sequenciamento, principalmente COI, COII, COIII e CytB, os genes codificadores de tRNA apresentaram as menores profundidades (Figura 5B).

A expressão de transcritos mitocondriais entre os órgãos, estimada usando FPKM, mostrou que o cérebro teve o maior nível de expressão de transcritos mitocondriais, com FPKM de 15.024.755, seguido do intestino (FPKM = 10.481.360), fígado (FPKM = 10.227.055), brânquias (FPKM = 6.284.735), rim (FPKM = 3.283.662) e gônada (FPKM = 105.136).

Os primeiros nove nucleotídeos do genoma mitocondrial montado neste trabalho e depositado no GenBank do NCBI (KX611143) foram sequenciados apenas em um espécime (TP166). Da mesma forma, os nucleotídeos das posições 2.771 a 2.778 e 16.162 a 16.330 foram obtidos apenas no outro espécime (TP167). A sequência dos dois espécimes diferem entre si em apenas sete posições, são elas: T1004C, A2731G, C3209T, T9728C, G11369A, G11727A e G15771A. A comparação do genoma montado nesse trabalho com a única sequência mitocondrial de *H. zebra* disponível publicamente mostrou diferenças em três posições nucleotídicas (posições 1.658, 2.174, 2.226), compartilhadas com os dois espécimes sequenciados nesse estudo, e uma quarta diferença (posição 1.004) compartilhada com apenas um dos dois espécimes desse trabalho (TP167).

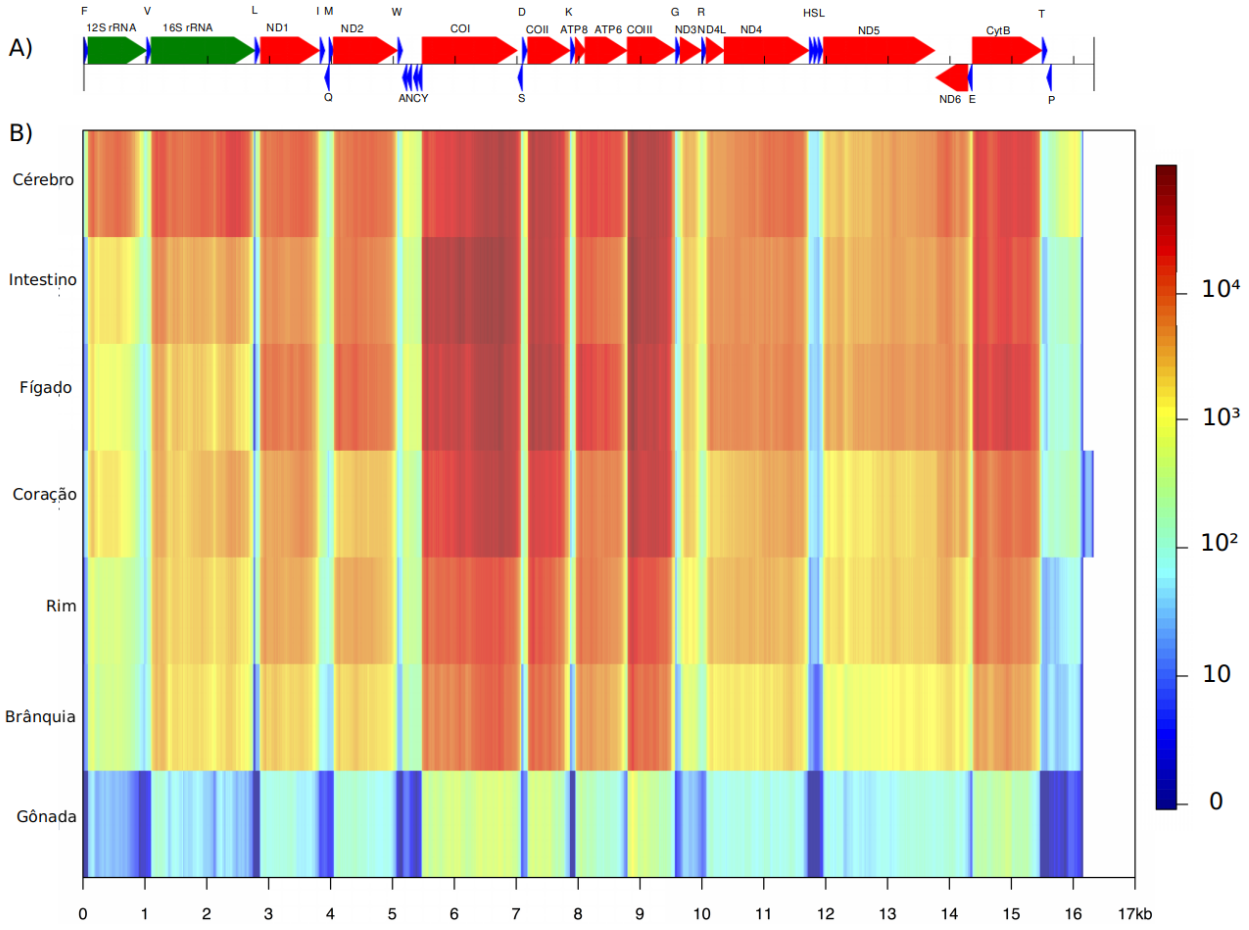


Figura 5: Profundidade de sequenciamento do genoma mitocondrial de *Hypancistrus zebra*. Representação linear (A) e profundidade de sequenciamento por nucleotídeo do mitogenoma de *H. zebra*. Os dois genes de RNAs ribossomais são mostrados em verde, os 13 genes codificantes de proteína em vermelho e os 22 genes de RNAs transportadores em azul. A profundidade de sequenciamento, número absoluto de leituras por posição, é mostrada para cada órgão em um gradiente de cores. A legenda das cores é mostrada ao lado, em escala logarítmica.

No total, 21 sítios heteroplásmicos foram encontrados no mitogenoma do *H. zebra*. As posições 13.592 (ND5) e 13.597 (ND5) foram consideradas heteroplasmáticas nos sete órgãos. As posições 653 (12s rRNA), 656 (12s rRNA), 7.080 (tRNA-Ser), 9.566 (COIII), 14.154 (ND6) e 14.178 (ND6) foram consideradas heteroplasmáticas em seis dos sete órgãos, a exceção foi a gônada, provavelmente devido a baixa proporção de leituras provenientes de transcritos mitocondriais no transcriptoma desse órgão. A posição 998 (12s rRNA) foi considerada heteroplasmática somente no cérebro, enquanto a posição 14.355 (tRNA-Glu) foi

considerada heteroplasmática apenas no coração.

Tabela 4: Sítios heteroplasmáticos no genoma mitocondrial de *Hypancistrus zebra*. Sítios heteroplasmáticos foram definidos como qualquer posição, com pelo menos 100 leituras, na qual a frequência do segundo nucleotídeo mais abundante foi maior ou igual a 10%. A frequência para os quatro nucleotídeos é mostrada para cada sítio heteroplasmático detectado. A posição (Pos) e o gene em que está inserida são mostrados para cada heteroplasmia, além do número total de leituras que alinharam em cada sítio heteroplasmático (Quant.).

Gene	Pos	Coração					Brânquia				
		Frequência %				Quant.	Frequência %				Quant.
		A	C	G	T		A	C	G	T	
12S rRNA	653	87	13	0	0	1042	87	12	0	0	422
12S rRNA	656	77	23	0	0	968	79	21	0	0	371
12S rRNA	998	97	0	3	0	39	100	0	0	0	39
16S rRNA	2129	54	0	1	46	1956	86	0	0	14	887
16S rRNA	2750	12	5	14	69	153	34	6	13	47	53
tRNA Leu	3823	6	1	92	1	558	8	1	90	1	145
ND2	5079	0	99	1	0	149	10	0	0	90	30
tRNA Ser	7080	68	5	20	7	1183	70	4	23	2	246
tRNA Ser	7082	15	1	26	58	183	12	0	15	73	26
COII	7869	23	5	20	52	178	27	7	23	43	44
tRNA Lys	7870	38	37	19	6	108	16	79	5	0	19
ATPase 6	8784	10	2	4	83	181	1	0	0	99	189
COIII	9566	18	10	17	55	656	22	9	13	56	113
COIII	9568	68	5	18	9	230	77	6	13	4	47
tRNA Gly	9571	22	0	0	78	50	10	0	0	90	21
tRNA Gly	9572	7	91	0	2	43	0	100	0	0	19
ND5	13592	16	0	84	0	2558	16	0	84	0	1370
ND5	13597	15	85	0	0	2778	13	87	0	0	1467
ND6	14154	86	14	0	0	3345	84	16	0	0	1275
ND6	14178	83	17	0	0	4117	80	20	0	0	1479
tRNA Glu	14355	85	2	13	0	223	98	0	2	0	206

Tabela 4: cont.

Gene	Pos	Cérebro					Fígado				
		Frequência %				Quant.	Frequência %				Quant.
		A	C	G	T		A	C	G	T	
12S rRNA	653	87	13	0	0	9746	89	11	0	0	758
12S rRNA	656	77	23	0	0	9050	79	21	0	0	657
12S rRNA	998	79	2	16	3	146	100	0	0	0	56
16S rRNA	2129	56	0	0	46	7752	93	0	0	7	1150
16S rRNA	2750	16	7	11	66	655	22	6	29	43	49
tRNA Leu	3823	12	1	84	3	773	6	1	91	1	473
ND2	5079	17	4	15	64	111	13	3	1	83	120
tRNA Ser	7080	74	4	18	4	1709	69	3	24	4	815
tRNA Ser	7082	21	0	19	60	273	8	0	8	83	203
COII	7869	30	8	20	41	299	26	3	19	51	160
tRNA Lys	7870	36	30	26	8	151	18	68	14	0	95
ATPase 6	8784	4	1	3	92	714	4	0	1	95	401
COIII	9566	21	10	21	49	846	19	8	17	56	323
COIII	9568	74	3	16	6	291	82	3	12	3	159
tRNA Gly	9571	29	0	0	71	75	16	0	0	84	77
tRNA Gly	9572	22	78	0	0	68	6	94	0	0	71
ND5	13592	16	0	84	0	6073	17	0	83	0	6079
ND5	13597	15	85	0	0	6565	14	86	0	0	6500
ND6	14154	86	14	0	0	6351	84	16	0	0	4943
ND6	14178	81	19	0	0	7352	82	18	0	0	5949
tRNA Glu	14355	0	100	0	0	753	99	0	1	0	778

Tabela 4: cont.

Gene	Pos	Gônada					Intestino				
		Frequência %				Quant.	Frequência %				Quant.
		A	C	G	T		A	C	G	T	
12S rRNA	653	82	18	0	0	22	86	14	0	0	1454
12S rRNA	656	86	14	0	0	21	78	22	0	0	1300
12S rRNA	998	100	0	0	0	1	97	0	2	1	104
16S rRNA	2129	91	0	0	9	11	68	0	0	32	2314
16S rRNA	2750	0	0	50	50	2	10	6	15	70	124
tRNA Leu	3823	0	0	86	14	7	5	0	94	1	704
ND2	5079	0	0	0	100	2	17	1	4	78	101
tRNA Ser	7080	67	11	11	11	9	71	4	21	4	1419
tRNA Ser	7082	0	0	20	80	5	13	0	17	70	222
COII	7869	0	0	50	50	2	30	6	13	51	240
tRNA Lys	7870	100	0	0	0	1	26	57	14	2	141
ATPase 6	8784	13	0	0	88	8	3	1	1	95	473
COIII	9566	25	0	0	75	4	17	10	21	51	618
COIII	9568	100	0	0	0	3	71	6	16	6	231
tRNA Gly	9571	0	0	0	100	3	40	0	0	60	52
tRNA Gly	9572	0	100	0	0	3	21	79	0	0	39
ND5	13592	14	0	86	0	119	16	0	83	0	4857
ND5	13597	14	86	0	0	125	15	85	0	0	5273
ND6	14154	72	28	0	0	78	84	16	0	0	4228
ND6	14178	65	35	0	0	95	80	19	0	0	4859
tRNA Glu	14355	83	0	17	0	6	96	1	3	0	677

Tabela 4: cont.

Gene	Pos	Rim				
		Frequência %				Quant.
		A	C	G	T	
12S rRNA	653	82	18	0	0	431
12S rRNA	656	73	27	0	0	392
12S rRNA	998	100	0	0	0	37
16S rRNA	2129	70	0	0	30	1293
16S rRNA	2750	19	5	5	71	104
tRNA Leu	3823	10	3	85	2	226
ND2	5079	5	3	13	80	40
tRNA Ser	7080	73	4	20	4	503
tRNA Ser	7082	15	0	17	69	108
COII	7869	20	6	16	58	125
tRNA Lys	7870	15	72	9	4	78
ATPase 6	8784	5	3	1	91	265
COIII	9566	13	10	19	58	253
COIII	9568	78	2	10	10	101
tRNA Gly	9571	20	0	0	80	25
tRNA Gly	9572	13	87	0	0	23
ND5	13592	17	0	83	0	2380
ND5	13597	16	84	0	0	2582
ND6	14154	83	17	0	0	1786
ND6	14178	77	23	0	0	1986
tRNA Glu	14355	97	1	2	0	362

3.4. Desenvolvimento de Marcadores Moleculares para *Hypancistrus zebra*

Um transcriptoma único foi montado usando todos conjuntos de leituras dos dois indivíduos de *H. zebra*. Para busca de SNVs e indels, cada um dos conjuntos de leituras dos sete órgãos foi mapeado contra esse transcriptoma único. O programa GATK foi usado para identificar todos os possíveis SNVs e indels para cada conjunto de leituras mapeadas. O transcriptoma do cérebro foi o que teve maior número de SNVs e de transcritos com SNVs. Já o transcriptoma da gônada foi o que teve menor número de SNVs e transcritos com SNVs (Tabela 5).

Tabela 5: Total de SNVs e indels encontrados em cada um dos sete órgãos. Número de SNVs e indels encontrados em cada órgão e número de transcritos que possuem pelo menos um SNV ou indel.

	SNV	Transcritos com SNV	Indel	Transcritos com indel
Gônada	3.744	2.446	355	329
Coração	4.523	2.768	476	435
Brânquia	4.895	3.195	476	450
Fígado	4.256	2.793	377	353
Intestino	6.970	4.524	730	671
Cérebro	7.315	4.680	834	756
Rim	5.103	3.447	585	543

Os SNVs e indels identificados pelo GATK foram filtrados de acordo com critérios diferentes com o objetivo de refinar a seleção desses marcadores e otimizar seu uso em trabalhos subsequentes de genética de populações e da conservação. Para o refinamento da escolha de SNVs, primeiro foram selecionados os transcritos que apresentaram pelo menos três SNVs em qualquer região de 500 nucleotídeos. Dessa forma, pretende-se otimizar o uso desses marcadores pois será possível obter a informação de três sítios polimórficos em uma mesma reação de PCR e sequenciamento pelo método de Sanger. Em seguida, dentre os transcritos identificados na etapa anterior, foram selecionados aqueles expressos nos sete

órgãos (Figura 6) e cujos SNVs foram identificados na mesma posição (Tabelas 6 e 7). Assim, espera-se ter identificado as regiões dos transcritos com maior potencial de serem efetivamente usados como marcadores populacionais.

Foram encontrados 21 transcritos compartilhados pelos seis órgãos do indivíduo TP166 (Figura 6A) com pelo menos três SNVs em qualquer região de 500 nucleotídeos. Desses, 13 transcritos também são compartilhados com o coração, do indivíduo TP167 (Figura 6B). Dos 13 transcritos com pelo menos três SNVs em qualquer região de 500 nucleotídeos e comuns a todos os órgãos, sete apresentaram SNVs nas mesmas posições. Esses sete transcritos foram anotados contra os bancos de dados Refseq (Tabela 6) e NR. Todos os transcritos apresentaram a mesma anotação em ambos bancos de dados, com cobertura, e-value e identidade muito próximos.

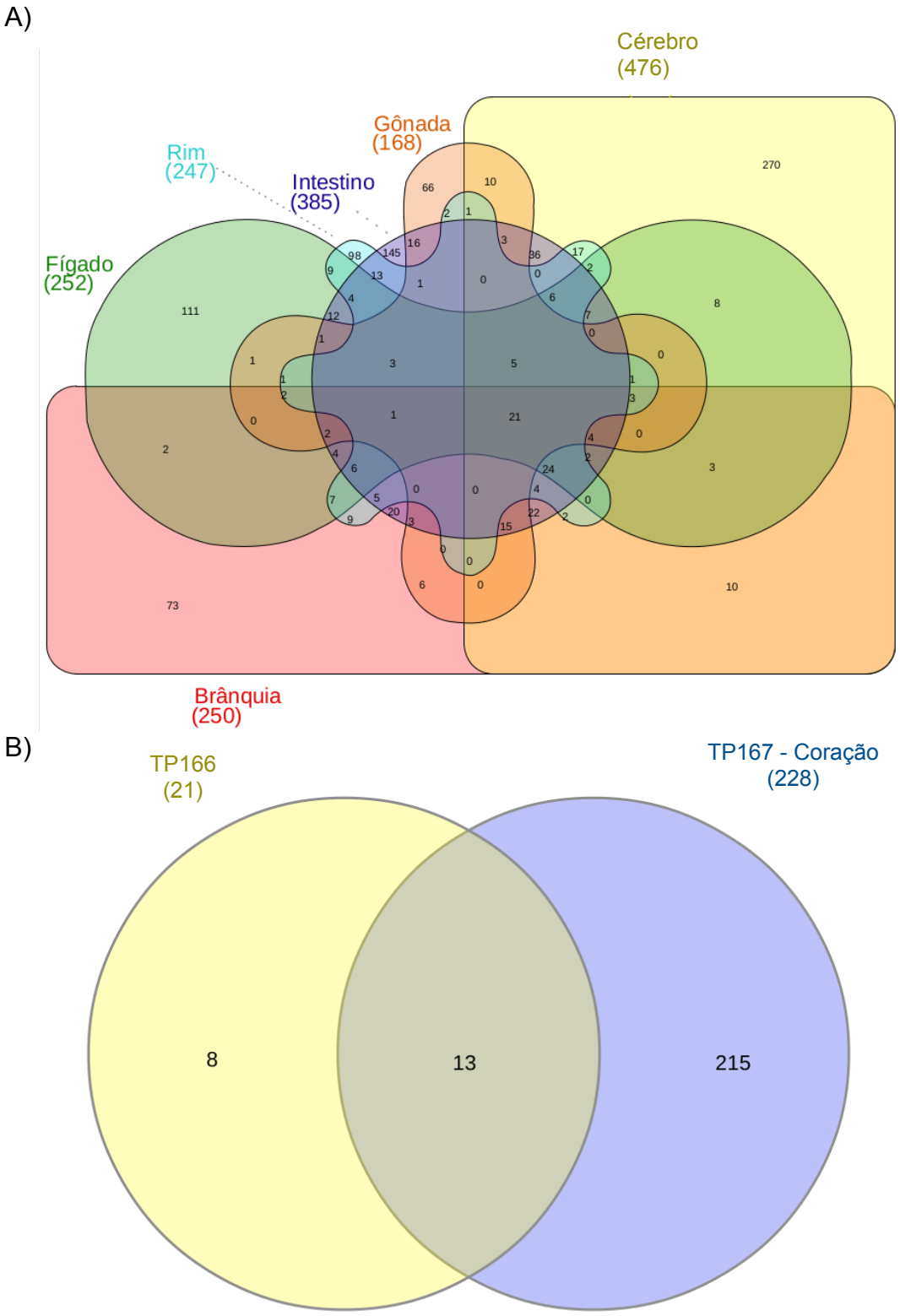


Figura 6: Interseção dos transcritos que contêm pelo menos três SNVs. No painel (A) são mostrados os transcritos expressos com SNVs compartilhados entre os seis órgãos analisados do indivíduo TP166; e no painel (B) são mostrados os compartilhados entre os dois indivíduos amostrados.

Tabela 6: Anotação de transcritos selecionados com SNVs. Anotação contra o banco de dados Refseq dos transcritos que contêm SNVs selecionados. Também são mostrados os percentuais de identidade do transcrito anotado contra a sequência do Refseq (ID%) e cobertura do transcrito anotado contra a sequência do Refseq (qCov%), além do número esperado de anotações ao acaso em uma busca em um banco de dados de um determinado tamanho (E-value).

Transcrito	**Refseq**			
	Anotação	**qCov%**	**E-value**	**ID %**
TR32381|c0_g1_i1	XM_017454953.1 PREDICTED: *Ictalurus punctatus* mitochondrial ribosomal protein L24 (mrpl24), transcript variant X1, mRNA	38	2.98e-152	81,6
TR39489|c0_g1_i1	XM_017451902.1 PREDICTED: *Ictalurus punctatus* golgi phosphoprotein 3 (golph3), mRNA	32	0.0	84,3
TR39678|c0_g1_i1	XM_017723117.1 PREDICTED: *Pygocentrus nattereri* F-box protein 46 (fbxo46), mRNA	80	0.0	80,8
TR37698|c0_g1_i1	XM_017495527.1 PREDICTED: *Ictalurus punctatus* YME1 like 1 ATPase (yme1l1), mRNA	74	0.0	82,3
TR32876|c0_g1_i1	XM_017459237.1 PREDICTED: *Ictalurus punctatus* DnaJ heat shock protein family (Hsp40) member A2 (dnaja2), mRNA	62	0.0	85,4
TR36155|c0_g1_i1	XM_017489735.1 PREDICTED: *Ictalurus punctatus* coxsackie virus and adenovirus receptor (cxadr), mRNA	49	0.0	77,9
TR36776|c0_g1_i1	XM_017476463.1 PREDICTED: *Ictalurus punctatus* glycine N-methyltransferase (gnmt), mRNA	43	0.0	82,5

Tabela 7: Variantes de nucleotídeo único (SNVs) selecionados. Informações sobre SNVs candidatos a serem usados como marcadores populacionais. Região: posição inicial de uma região de 500 nucleotídeos. Posição: posição do SNV. Nº de SNVs: quantidade de SNVs encontrada na região de 500 nucleotídeos. Nucleotídeo: número de leituras que apoiam o nucleotídeo da posição.

Transcrito	Região	Posição	Nº de SNVs	Brânquia				Cérebro			
				Nucleotídeo				Nucleotídeo			
				A	T	C	G	A	T	C	G
TR32381|c0_g1_i1	255	255	3	0	8	8	0	0	35	22	0
		608		0	6	3	0	0	38	46	0
		613		0	6	3	0	0	39	46	0
TR39489|c0_g1_i1	1451	1451	3	12	0	0	17	15	0	0	22
		1460		13	0	16	0	16	0	23	1
		1946		16	0	0	26	14	0	0	16
TR39678|c0_g1_i1	2365	2365	4	0	8	7	0	0	17	13	0
		2368		0	9	7	0	0	17	12	0
		2613		6	0	0	6	19	0	0	19
		2746		4	6	0	0	23	32	0	0
	2746	2746	3	6	0	0	6	19	0	0	19
		2934		4	6	0	0	23	32	0	0
		3213		11	0	0	8	27	0	0	18
TR37698|c0_g1_i1	385	385	3	7	0	0	6	5	0	0	4
		465		4	0	3	0	3	0	3	0
		722		0	7	9	0	0	7	6	0
TR32876|c0_g1_i1	1363	1363	3	0	0	21	33	0	0	24	43
		1457		0	12	0	7	0	26	0	15
		1763		0	55	47	0	0	134	133	0
	1763	1763	3	0	55	47	0	0	134	133	0
		1918		0	34	35	0	0	129	135	0
		2116		24	44	0	0	111	140	1	0
TR36155|c0_g1_i1	1312	1312	3	41	0	0	37	15	0	0	11
		1448		0	26	30	0	0	5	6	0
		1580		24	0	0	28	11	0	0	21
TR36776|c0_g1_i1	1381	1381	3	0	10	14	0	0	10	11	0
		1418		12	8	0	0	7	8	0	1
		1520		0	16	18	0	0	11	13	0

Tabela 7: cont.

Transcrito	Região	Posição	Nº de SNVs	Coração				Fígado			
				Nucleotídeo				Nucleotídeo			
				A	T	C	G	A	T	C	G
TR32381|c0_g1_i1	255	255	3	0	7	28	0	0	26	21	0
		608		0	8	32	0	0	36	35	0
		613		0	8	33	0	0	34	35	0
TR39489|c0_g1_i1	1451	1451	3	7	0	0	3	23	1	0	40
		1460		7	0	3	0	27	0	41	0
		1946		14	0	0	6	35	0	0	29
TR39678|c0_g1_i1	2365	2365	4	0	3	7	0	0	23	18	0
		2368		0	3	7	0	0	21	18	0
		2613		1	0	0	10	12	1	0	12
		2746		4	2	0	0	12	9	0	0
	2746	2746	3	4	2	0	0	12	9	0	0
		2934		5	0	0	8	11	0	0	12
		3213		0	0	3	17	0	0	13	15
TR37698|c0_g1_i1	385	385	3	5	0	0	9	5	0	0	13
		465		1	0	13	0	15	0	3	0
		722		0	10	4	0	0	10	29	0
TR32876|c0_g1_i1	1363	1363	3	0	0	9	8	0	0	53	99
		1457		0	10	0	6	0	38	0	16
		1763		0	5	66	0	0	201	134	0
	1763	1763	3	0	5	66	0	0	201	134	0
		1918		0	39	43	0	0	94	132	0
		2116		25	50	0	0	90	169	0	0
TR36155|c0_g1_i1	1312	1312	3	1	0	0	6	84	0	0	47
		1448		0	0	3	0	0	29	41	0
		1580		0	0	0	7	40	0	0	40
TR36776|c0_g1_i1	1381	1381	3	0	7	11	0	0	892	922	0
		1418		4	10	0	0	785	720	0	5
		1520		0	7	4	0	0	1512	1250	0

Tabela 7: cont.

Transcrito	Região	Posição	Nº de SNVs	Gônada				Intestino			
				Nucleotídeo				Nucleotídeo			
				A	T	C	G	A	T	C	G
TR32381|c0_g1_i1	255	255	3	0	291	244	1	0	30	32	0
		608		0	236	261	0	0	35	34	0
		613		0	234	263	0	0	34	33	0
TR39489|c0_g1_i1	1451	1451	3	83	0	0	119	19	0	0	33
		1460		88	0	122	0	21	0	36	0
		1946		85	0	0	97	32	0	0	31
TR39678|c0_g1_i1	2365	2365	4	0	121	106	0	0	24	21	0
		2368		0	123	106	0	0	24	21	0
		2613		70	1	0	63	38	0	0	26
		2746		66	66	0	0	28	27	0	0
	2746	2746	3	66	66	0	0	28	27	0	0
		2934		122	0	0	100	26	0	0	27
		3213		0	1	51	59	0	0	20	19
TR37698|c0_g1_i1	385	385	3	82	0	0	82	17	0	0	15
		465		63	0	65	0	15	0	8	0
		722		0	119	145	0	0	34	20	0
TR32876|c0_g1_i1	1363	1363	3	0	0	81	145	0	0	49	52
		1457		0	24	0	26	0	28	0	27
		1763		0	343	302	0	0	146	129	0
	1763	1763	3	0	343	302	0	0	146	129	0
		1918		0	180	197	0	0	116	130	0
		2116		148	256	0	0	116	150	0	0
TR36155|c0_g1_i1	1312	1312	3	11	0	0	7	70	0	0	62
		1448		0	5	5	0	0	45	47	0
		1580		7	0	0	7	65	0	0	66
TR36776|c0_g1_i1	1381	1381	3	0	67	77	0	0	13	21	0
		1418		31	38	0	1	13	15	0	1
		1520		0	84	104	0	0	9	19	0

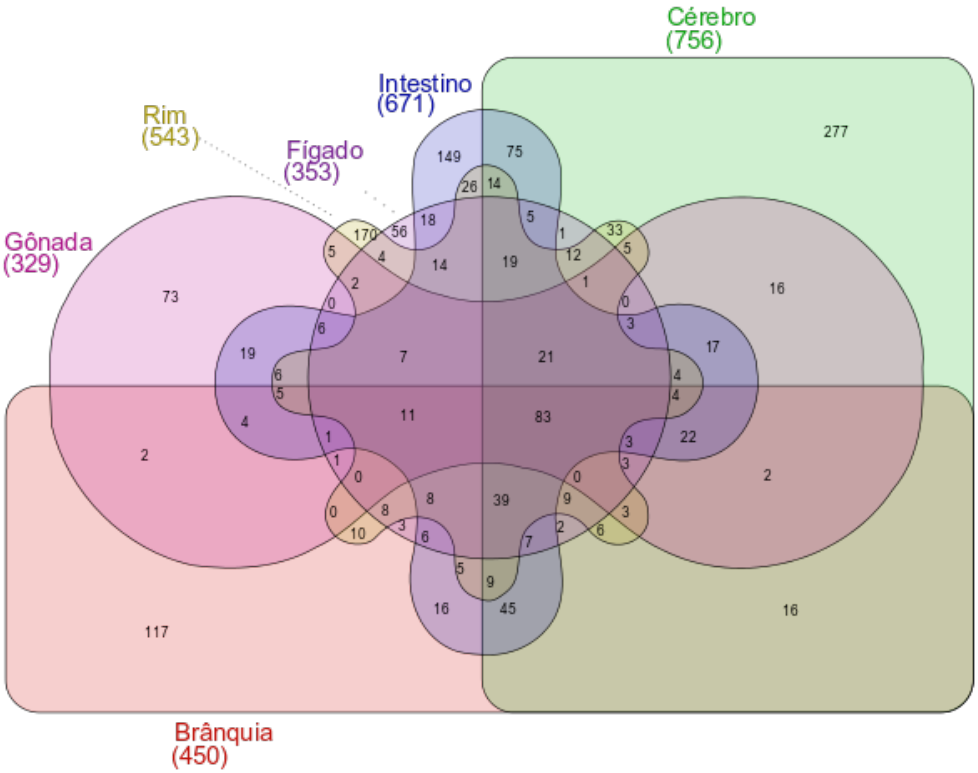
Tabela 7: cont.

Transcrito	Região	Posição	Rim				
			Nº de SNVs	Nucleotídeo			
				A	T	C	G
TR32381|c0_g1_i1	255	255	3	0	17	17	0
		608		0	20	14	0
		613		0	20	14	0
TR39489|c0_g1_i1	1451	1451	3	21	0	0	50
		1460		22	0	50	0
		1946		44	0	0	53
TR39678|c0_g1_i1	2365	2365	4	0	20	31	0
		2368		0	18	33	0
		2613		22	0	0	20
		2746		23	31	0	0
	2746	2746	3	23	31	0	0
		2934		34	0	0	32
		3213		0	0	24	11
TR37698|c0_g1_i1	385	385	3	13	0	0	18
		465		21	0	16	0
		722		0	33	26	0
TR32876|c0_g1_i1	1363	1363	3	0	0	69	82
		1457		0	45	0	39
		1763		0	170	171	1
	1763	1763	3	0	170	171	1
		1918		0	158	204	0
		2116		124	132	0	0
TR36155|c0_g1_i1	1312	1312	3	10	0	0	9
		1448		0	16	12	0
		1580		15	0	0	12
TR36776|c0_g1_i1	1381	1381	3	0	9	16	0
		1418		7	14	0	0
		1520		0	8	11	0

A primeira etapa da seleção de indels foi verificar os transcritos com pelo menos um indel que eram comuns a todos os sete órgãos (Figura 7). Esses transcritos comuns a todos os órgãos foram anotados contra o Refseq de vertebrados e aqueles que tiveram anotação com uma cobertura maior que 70% foram selecionados. Desses, foram aproveitados apenas os indels de mesma posição para os sete órgãos (Tabelas 8 e 9).

Foram encontrados 83 transcritos que contêm indels compartilhados entre os seis órgãos de um dos indivíduos de *H. zebra* (TP166) (Figura 7A). Desses, 56 são compartilhados pelo transcriptoma do coração do outro indivíduo dessa espécie (TP167) (Figura 7B). Dos 56 transcritos mencionados, seis transcritos apresentaram anotação contra os bancos de dados NR e Refseq com cobertura maior que 70%. Todos os transcritos tiveram a mesma anotação em ambos bancos de dados, com cobertura, identidade e E-value muito similares.

A)



B)

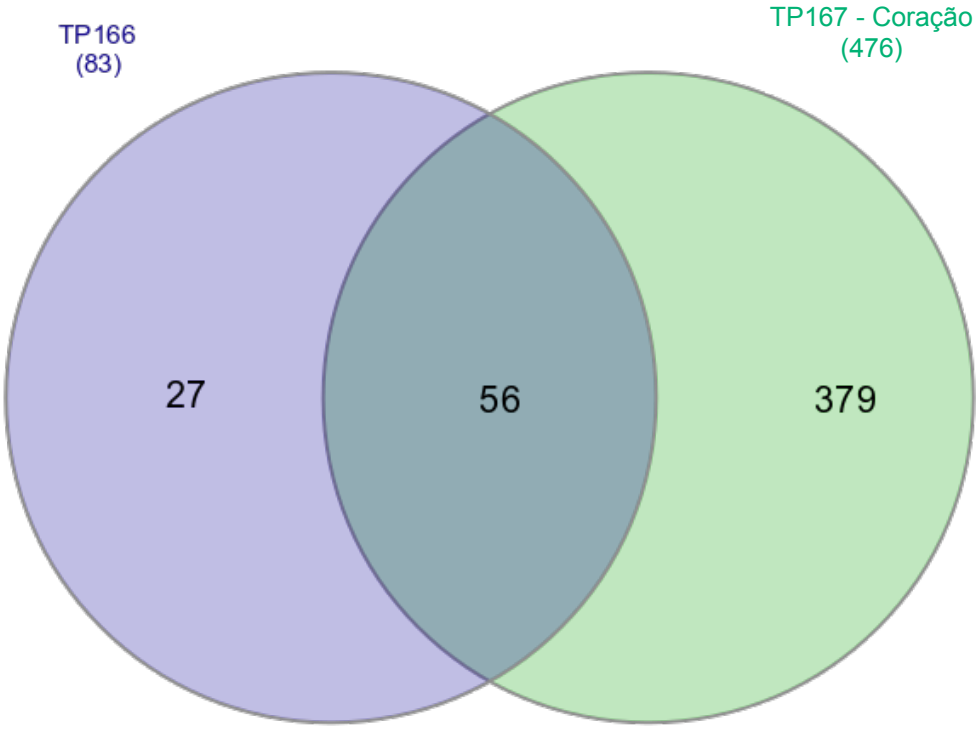


Figura 7: Interseção dos transcritos que contêm indels. No painel (A) são mostrados os transcritos expressos com indels compartilhados entre os seis órgãos analisados do indivíduo TP166; e no painel (B) são mostrados os compartilhados entre os dois indivíduos amostrados.

Tabela 8: Anotação dos transcritos selecionados com indels. Anotação contra o banco de dados Refseq dos transcritos que contêm indels selecionados. ID%: identidade do transcrito anotado contra a sequência do Refseq em porcentagem. qCov%: cobertura do transcrito anotado contra a sequência do Refseq em porcentagem.

Transcrito	**Refseq**			
	Anotação	**qCov %**	**E-value**	**ID %**
TR26307|c0_g1_i1	XM_017479787.1 PREDICTED: *Ictalurus punctatus* DNA damage inducible transcript 3 (ddit3), mRNA	81	0.0	85,8
TR34372|c0_g1_i1	XM_017455682.1 PREDICTED: *Ictalurus punctatus* cleavage and polyadenylation specific factor 2 (cpsf2), mRNA	83	0.0	85,1
TR35455|c0_g1_i1	XM_017471543.1 PREDICTED: *Ictalurus punctatus* fragile X mental retardation syndrome-related protein 1-like (LOC108267464), mRNA	70	0.0	83,3
TR37541|c1_g1_i1	XM_017493024.1 PREDICTED: *Ictalurus punctatus* heat shock 70 kDa protein 4-like (LOC108279070), mRNA	72	0.0	85,6
TR39911|c1_g1_i2	XM_017457414.1 PREDICTED: *Ictalurus punctatus* TATA-box binding protein associated factor 5 (taf5), transcript variant X1, mRNA	93	0.0	83,2
TR40056|c0_g1_i1	XM_017454518.1 PREDICTED: *Ictalurus punctatus* rhomboid 5 homolog 1 (rhbdf1), mRNA	71	0.0	87,7

Tabela 9: Indels selecionados. Informações sobre os indels selecionados. POS: posição do Indel no transcrito. REF: referência. ALT: alternativo. Quant.: número de leituras que suportam aquela referência ou alternativo.

Transcrito	POS	Brânquia				Cérebro			
		REF	Quant.	ALT	Quant.	REF	Quant.	ALT	Quant.
TR26307|c0_g1_i1	1274	CA	13	C	65	CA	41	C	128
TR34372|c0_g1_i1	1497	CA	1	C	5	CA	5	C	10
TR35455|c0_g1_i1	3084	G	3	GA	7	G	18	GA	8
TR37541|c1_g1_i1	2569	GT	19	G	15	GT	40	G	18
TR39911|c1_g1_i2	1458	CA	1	C	5	CA	3	C	11
TR40056|c0_g1_i1	344	TTG	1	T	26	TTG	1	T	14

Tabela 9: cont.

Transcrito	POS	Coração				Fígado			
		REF	Quant.	ALT	Quant.	REF	Quant.	ALT	Quant.
TR26307|c0_g1_i1	1274	CA	3	C	34	CA	36	C	141
TR34372|c0_g1_i1	1497	CA	1	C	7	CA	2	C	33
TR35455|c0_g1_i1	3084	G	16	GA	17	G	8	GA	15
TR37541|c1_g1_i1	2569	GT	32	G	49	GT	49	G	41
TR39911|c1_g1_i2	1458	CA	1	C	3	CA	1	C	2
TR40056|c0_g1_i1	344	TTG	15	T	3	TTG	1	T	36

Tabela 9: cont.

Transcrito	POS	Gônada				Intestino			
		REF	Quant.	ALT	Quant.	REF	Quant.	ALT	Quant.
TR26307|c0_g1_i1	1274	CA	6	C	15	CA	40	C	152
TR34372|c0_g1_i1	1497	CA	18	C	142	CA	5	C	25
TR35455|c0_g1_i1	3084	G	194	GA	143	G	38	GA	33
TR37541|c1_g1_i1	2569	GT	324	G	236	GT	54	G	40
TR39911|c1_g1_i2	1458	CA	14	C	40	CA	2	C	10
TR40056|c0_g1_i1	344	TTG	0	T	39	TTG	2	T	31

Tabela 9: cont.

Transcrito	POS	Rim			
		REF	Quant.	ALT	Quant.
TR26307|c0_g1_i1	1274	CA	65	C	288
TR34372|c0_g1_i1	1497	CA	8	C	38
TR35455|c0_g1_i1	3084	G	30	GA	27
TR37541|c1_g1_i1	2569	GT	37	G	23
TR39911|c1_g1_i2	1458	CA	9	C	26
TR40056|c0_g1_i1	344	TTG	3	T	44

Tabela 10: Sugestão de iniciadores para amplificação por PCR das regiões de SNV e Indel.

Transcrito	Iniciador	Sequência (5'->3')	Tm	Início	Fim
TR26307|c0_g1_i1	**Forward**	GGAGAACGAAAGGAAGGTGC	58.84	998	1017
	Reverse	GAATTATGACAGCTTGGGTGGA	58.38	1429	1408
TR34372|c0_g1_i1	**Forward**	TGGCACGCTACCTCATAGAC	59.26	1370	1389
	Reverse	AGGGGAACATCGGGTACGA	60.00	1679	1661
TR35455|c0_g1_i1	**Forward**	TCTGACGGAACTGAACACAGG	59.93	2777	2797
	Reverse	TCCAAGCCTAACTCTGCCTTT	59.29	3193	3173
TR37541|c1_g1_i1	**Forward**	GGCACTGTCTGTCCACTGATT	60.27	2317	2337
	Reverse	ATACACACAAATCGCAGGCA	58.18	2834	2815
TR39911|c1_g1_i2	**Forward**	GAGGAGGGCAAACCCAAGAA	59.89	1141	1160
	Reverse	TACTGGGTAGTTGTGCCCCT	60.18	1720	1701
TR40056|c0_g1_i1	**Forward**	GGCAGCATTTACCAAGGCACA	61.77	205	225
	Reverse	GTATGACCTGAACGCCCATCT	59.86	662	642
TR32381|c0_g1_i1	**Forward**	CAGGACAGAAAGAGTGTAAGAGA	57.29	223	245
	Reverse	CTCTGTCTCTGTGGGCTTCCT	61.17	646	626
TR39489|c0_g1_i1	**Forward**	GAGGAGCATCTGGCTGTTCTG	60.74	1398	1418
	Reverse	GCTTCCCTTCTGCGTTCTCTC	61.00	1999	1979
TR39678|c0_g1_i1	**Forward**	CCGACTCGCGCTGGAA	59.38	2331	2346
	Reverse	CTTGGAAGCAGACGTGTAATG	57.27	3249	3229
TR37698|c0_g1_i1	**Forward**	TTGACGGATTTGGGAGCGG	60.38	349	367
	Reverse	TTTTCAGGAGCTTATCCAGCGT	60.03	769	748
TR32876|c0_g1_i1	**Forward**	GACGACGAGGGCGGTC	59.47	1312	1327
	Reverse	GCTTTCCACCCACTCGTTTTATT	59.75	2169	2147
TR36155|c0_g1_i1	**Forward**	CGTCGCTCGGTTCCATGTC	61.16	1194	1212
	Reverse	GGAACCTCTTCGCTTTCGGAT	60.41	1736	1716
TR36776|c0_g1_i1	**Forward**	TGCTGGAAATGTCGAACCTGA	59.93	1321	1341
	Reverse	TGAACAAAATGGAATCCCAATCCAG	60.05	1562	1538

3.5. Identificação de Elementos Genéticos Móveis

A representatividade de elementos genéticos móveis (EGM) em cada um dos sete transcriptomas do *Hypancistrus zebra* foi analisada de três formas. Foi observada a proporção, em pares de base, de cada transcriptoma que é composta por EGM, a frequência de transcritos que contêm EGM e a distribuição das classes de EGM em cada órgão. As proporções do transcriptoma que codificam para EGM variaram de 2,61% na gônada a 5,41% no rim (Tabela 11). As frequências de transcritos que contêm EGM encontradas nos sete transcriptomas variaram de 12% a 33%. Os órgãos que apresentaram menor frequência de transcritos com EGM em seu transcriptoma foram o coração (12%), o cérebro (14%) e a gônada (15%). Já a brânquia foi o órgão que apresentou a maior frequência de transcritos com EGM (Figura 8A). A distribuição das classes de EGM foi similar entre os sete órgãos (Figura 8B). A classe de EGM mais abundante é a Tc1-IS630-Pogo, com 66.021 sequências nos transcriptomas dos sete órgãos de *H. zebra*, seguida da hobo-Activator, com 24.959 sequências, e da L2/CR1/Rex, com 12.163 sequências. Já para as classes de elementos genéticos móveis CRE/SLACS, En-Spm e MuDR-IS905 nenhuma sequência representante foi encontrada nos sete transcriptomas.

Tabela 11: Proporção do transcriptoma, em pares de base, que codifica elementos genéticos móveis em cada órgão.

Órgão	Tamanho total do transcriptoma (pb)	Total de elementos genéticos móveis mascarados (pb)	Proporção (%)
Brânquia	62.907.574	2.891.431	4.60
Cérebro	88.881.174	4.575.965	5.15
Coração	34.201.784	1.144.751	3.35
Fígado	77.200.530	3.746.273	4.85
Gônada	54.127.513	1.413.878	2.61
Intestino	94.083.206	4.704.979	5.00
Rim	110.048.620	5.955.650	5.41

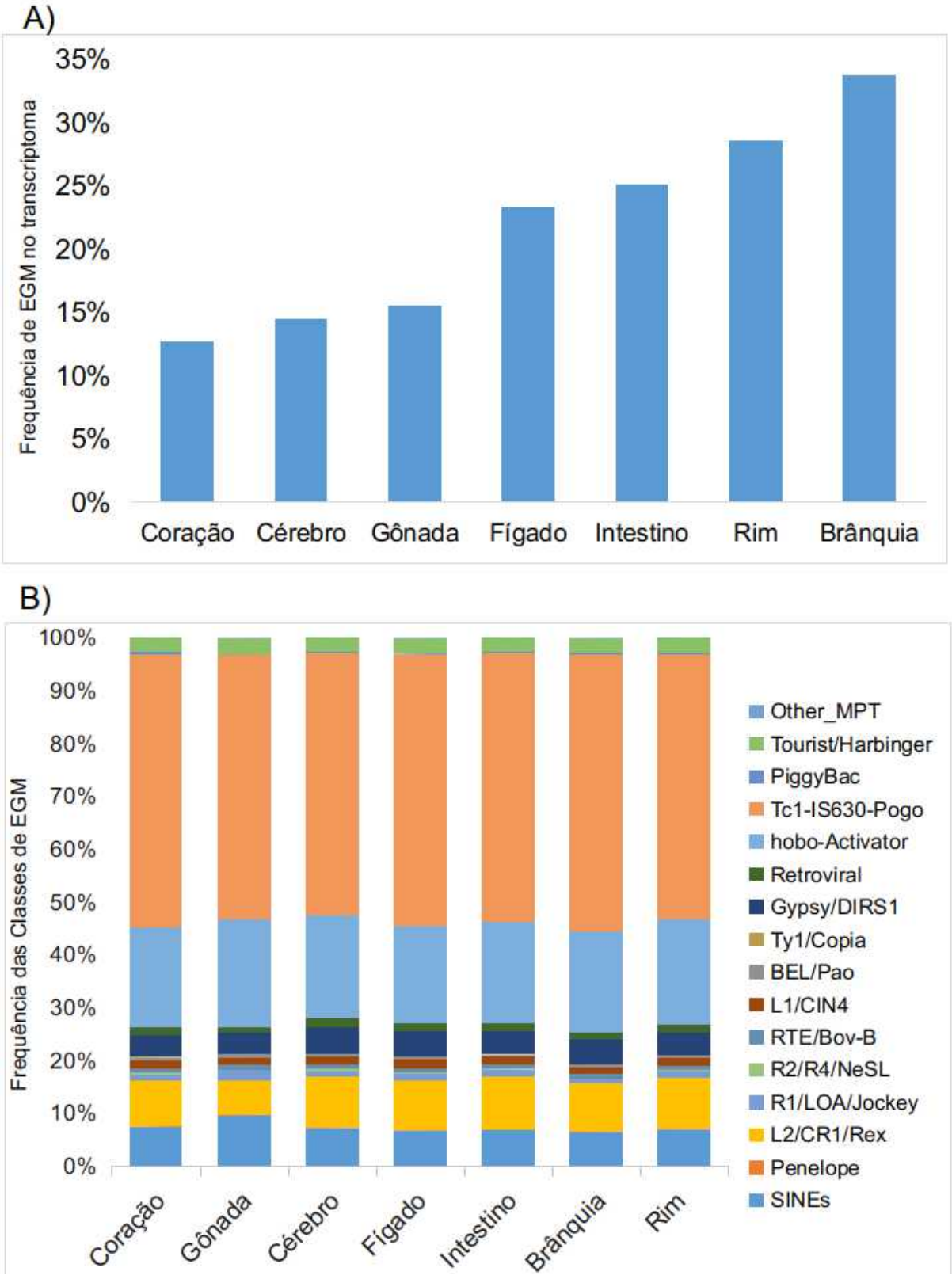


Figura 8: Distribuição de elementos genéticos móveis nos sete órgãos de *Hypancistrus zebra*. Frequência de transcritos que contém elementos genéticos móveis (A) e abundância relativa (B) das famílias de elementos genéticos móveis encontradas em cada um dos sete transcriptomas.

4. DISCUSSÃO

Apesar de ser uma espécie em risco de extinção, até o presente trabalho apenas duas sequências nucleotídicas de *Hypancistrus zebra* estavam disponíveis em bancos de dados públicos, uma de 1.788 nucleotídeos do gene F-reticulon 4 (42) e outra de 2.410 nucleotídeos contendo a sequência parcial de dois genes mitocondriais (12S e 16S) e a sequência completa do tRNA situado entre aqueles genes (tRNA-val) (42). Além dessas sequências de nucleotídeos, apenas uma sequência de aminoácido da proteína F-reticulon 4, já representada no banco de dados de nucleotídeo, era conhecida (42).

A escassez de informações genéticas sobre *H. zebra* pode ser estendida para a maioria das espécies da bacia amazônica. O uso das tecnologias de alto desempenho para sequenciamento de DNA e RNA favorece o preenchimento, de forma relativamente rápida, simples e com baixo custo, dessa falta de conhecimento sobre a genética de espécies nativas. Dessa forma, a geração de dados genéticos cedeu espaço como a etapa limitante do processo de investigação científica para a análise e aplicação desses dados nas diversas áreas das ciências, como por exemplo na genética da conservação (14).

Neste trabalho, produzimos mais de 200 milhões de leituras pareadas de transcriptomas de sete órgãos do *H. zebra*. Essas leituras foram utilizadas para montar mais de meio milhão de transcritos, que compõem os transcriptomas de sete órgãos (brânquia, cérebro, coração, fígado, gônada, intestino e rim) do *H. zebra*. Esse grande volume de dados estará disponível para o público e poderá ser aplicado em estudos de evolução, conservação da espécie e genética de populações.

O comércio de peixes para ornamentação e alimentação é a principal fonte de renda da população ribeirinha do rio Xingu (43,44). O *Hypancistrus zebra* faz parte desse mercado aquariofilista internacional, mesmo sua coleta sendo proibida. Uma medida para preservar essa espécie é o cultivo de *H. zebra* em cativeiro. Contudo, a fiscalização desses cultivos é necessária para previnir a comercialização de peixes coletados na natureza como procedentes de criações. O desenvolvimento de marcadores genéticos é uma medida que possibilitaria a distinção da origem dos peixes comercializados.

Um estudo do transcriptoma de uma espécie de peixe marinho, *Miichthys miiuy*, identificou 8.510 SNVs com a utilização do programa SOAPsnp (45). LIAO et al., 2013 identificaram 5.784 SNVs em uma espécie de carpa, utilizando o programa QualitySNP e critérios menos estringentes do que os utilizados neste trabalho. Já em um trabalho de análise do transcriptoma de uma espécie de Siluriformes, *Pelteobagrus fulvidraco*, foram identificados 26.450 SNVs e 4.145 indels com a utilização do programa samtools 0.1.18 (47). Identificamos mais de 35 mil SNVs e quase quatro mil indels distribuídos entre os sete órgãos de *H. zebra*. Desses, selecionamos aqueles com maior potencial de serem marcadores populacionais e desenhamos iniciadores para os SNVs e indel selecionados. Os transcritos selecionados com SNVs e indels foram anotados contra o banco de dados Refseq. No transcriptoma de outra espécie de Siluriformes, *Clarias batrachus*, também foi encontrado SNV no transcrito anotado contra a proteína chaperona DnaJ Hsp40 (48), assim como em *H. zebra*.

Ao contrário do RNA mensageiro, o DNA possui íntrons. Os iniciadores desenhados para amplificação por PCR das regiões com os SNVs e indels selecionados neste trabalho são baseados em sequências de RNA, sendo assim deve-se levar em consideração que a melhor opção para amplificação dessas regiões é usar amostra de RNA, para que o produto de PCR tenha o tamanho esperado e seja possível sequênciar por Sanger toda região desejada.

Os SNVs e indels selecionados neste trabalho podem ser usados como candidatos prioritários para estudos de genética de populações, visando a conservação dessa espécie. Dessa forma, esses dados podem contribuir para alcançar as metas da Agenda 2030 relacionadas à conservação da biodiversidade e evitar a extinção de espécies ameaçadas.

O genoma mitocondrial quase completo do *Hypancistrus zebra* produzido neste trabalho foi o quinto das mais de 800 espécies da família Loricariidae e o primeiro de qualquer espécie do clado *Peckoltia* disponível em banco de dados públicos. O clado *Peckoltia* é a tribo da família Loricariidae mais rica em gêneros, suas espécies apresentam grande diversidade morfológica, mas pouca divergência genética, o que contribui para sua filogenia ser controversa (49).

À luz da filogenia molecular de LUJAN et al., 2015 uma extensa revisão da taxonomia do clado *Peckoltia* está em andamento, muitas espécies já foram

renomeadas e gêneros sinonimizados (50,51). Apesar da importância da filogenia supracitada, os autores reconhecem que as relações entre os gêneros desse clado estão com uma resolução pobre. A resolução desse clado não pôde ser melhorada devido ao número limitado de *loci* tradicionalmente usado em estudos de filogenia molecular de Loricariidae, assim como à falta de informações genéticas específicas desse clado para desenho de iniciadores para amplificação por PCR de outras regiões. O genoma mitocondrial quase completo do *Hypancistrus zebra* e os marcadores genéticos nucleares encontrados neste trabalho preenchem essa lacuna de conhecimento e podem ajudar tanto na revisão taxonômica em andamento quanto em estudos filogenéticos futuros do clado *Peckoltia*.

Embora o transcriptoma do cérebro tenha metade do número de leituras sequenciadas (~25 milhões) em relação ao do fígado (~50 milhões), foram montados ~17 mil transcritos a mais. De forma similar, o transcriptoma do coração teve o número de transcritos montados muito próximo ao da gônada mesmo tendo ~29 milhões de leituras sequenciadas a menos. A explicação para essa falta de correspondência entre o número de leituras sequenciadas e o número de transcritos montados ainda não é clara, mas pode envolver tanto problemas de montagem, já que é provável que transcritos muito similares sejam montados como um único transcrito (52), como diferenças globais nos níveis de expressão gênica de cada órgão.

Os dois transcriptomas que tiveram o menor número de transcritos montados, gônada e coração, também tiveram menos transcritos anotados nos bancos de dados Uniprot tanto de *Danio rerio* como de *Homo sapiens*. Mais da metade dos transcritos montados não foi anotada contra o banco de dados Uniprot de *Danio rerio* e Uniprot de *Homo sapiens.* A frequência de não anotados variou de ~55% a ~72%, contra as entradas de *Danio* e de ~57% e ~73%, contra as entradas de humano. Resultados similares foram observados em diversos outros trabalhos, inclusive com a utilização de diferentes bancos de dados (53–57). A baixa frequência de transcritos anotados do *H. zebra* contra as entradas de *Danio* e humano podem ser consequência do tamanho relativamente pequeno do banco de dados de *D. rerio* (~59 mil entradas), apesar de ser da espécie de peixe mais estudada. Apesar do banco de dados de *H. sapiens* ter mais entradas (~155 mil) do que aquele de *D. rerio,* o número de transcritos anotados foi menor, em decorrência

da maior distância filogenética entre humanos e *H. zebra*. Entretanto, o banco de dados de humanos é melhor anotado do que o de *D. rerio*. Em reflexo disso, o número de proteínas anotadas como não caracterizadas em *Danio* variou de 8.503 a 17.879, enquanto o número de proteínas anotadas como não caracterizadas contra humanos variou de 141 a 303. Portanto, o banco de dados de *Homo sapiens* é uma melhor opção para caracterização dos transcritos sequenciados, mesmo que *Danio rerio* seja a espécie filogenéticamente mais próxima do modelo deste trabalho.

As sequências sem anotação contra bancos de dados são chamadas de matéria escura biológica, em analogia à matéria escura da cosmologia (58), já que são preditas através do processamento computacional de dados oriundos das tecnologias de sequenciamento de alto desempenho, e, em sua grande maioria, sua existência não foi comprovada de forma experimental. Essas sequências, que neste trabalho representam mais da metade dos transcritos montados, podem ser de transcritos espúrios, que foram montados através da combinação aleatória das leituras produzidas pelos métodos de sequenciamento e que não refletem transcritos efetivamente produzidos pelas células. Um exemplo disso é a montagem de transcritos não biológicos que é influenciada pela complexidade do transcriptoma. A qualidade da montagem *de novo* é inversamente proporcional à complexidade do transcriptoma montado, todos os montadores *de novo* tendem a produzir erros sempre que um gene tem um grande número de eventos de “splicing” alternativo (59). Por outro lado, essas novas sequências até então desconhecidas podem ser moléculas de RNA, codificante e não codificante, que ainda não foram descritas.

Os dados sobre elementos genéticos móveis (EGM) descritos nesse trabalho corroboram alguns dados encontrados no transcriptoma montado de uma espécie de peixe de água doce, *Rutilus rutilus* (60). No trabalho com *Rutilus rutilus*, dois órgãos, fígado e cérebro, foram usados para montar um transcriptoma, que apresentou a frequência de EGM, em relação ao seu tamanho em pares de bases, de 4,71%. Os transcriptomas do fígado e do cérebro de *Hypancistrus zebra* analisados neste trabalho apresentaram resultados similares, com as frequências de 4,85% e 5,15%, respectivamente. Além disso, no transcriptoma de *Rutilus rutilus* as classes de EGM hobo-Activator e Tc1-IS630-Pogo estão entre as classes com maior representatividade, assim como nos sete transcriptomas de *Hypancistrus zebra*. Em contraste, a classe En-Spm está entre as classes com maior representatividade no

Rutilus rutilus enquanto essa classe não foi representada nos sete órgãos de *H. zebra.*

Os elementos genéticos móveis têm papel fundamental na evolução do genoma de eucariotos, estando envolvidos em processos de rearranjo cromossômicos, expressão e regulação de genes, replicação de DNA e diferenciação dos cromossomos sexuais (16,18,61). Sendo assim, especula-se que EGM tenham importante papel na variação cariotípica observada entre muitos grupos (17). Existe uma grande variação cariotípica entre as espécies da família Loricariidae, com o número diplóide de cromossomos de 34 a 96 (18). Dessa forma, preencher as lacunas sobre os EGM em espécies dessa família, pode ajudar a entender essa grande variação cariotípica.

5. CONCLUSÃO

No banco de dados genéticos do *Hypancistrus zebra* produzido neste trabalho foram montados mais de meio milhão de transcritos. Desses, ~ 70% não tiveram anotação contra bancos de dados. Essas sequências podem ser transcritos espúrios ou transcritos biológicos que ainda não foram descritos. Distinguir transcritos espúrios de transcritos biológicos desconhecidos, assim como a inferência das funções desses novos transcritos, pode aumentar o conhecimento da biodiversidade molecular e proporcionar novas oportunidades, pois novos mecanismos de regulação da transcrição e funções protêicas poderiam ser revelados. Do banco de dados produzido, selecionados seis transcritos com indels e sete transcritos com pelo menos três SNVs como sendo os mais adequados para aplicação em trabalhos visando a conservação dessa espécie. O genoma mitocondrial descrito neste trabalho pode auxiliar no desenvolvimento de ferramentas genéticas a serem usadas para a conservação do *Hypancistrus zebra*. Para este fim, as quatro diferenças nucleotídicas entre o mitogenoma depositado e a única outra sequência mitocondrial disponível dessa espécie, assim como as sete diferenças nucleotídicas entre os dois espécimes sequenciados nesse trabalho, e as 21 heteroplasmias detectadas são de grande relevância. Foram encontrados elementos genéticos móveis expressos nos sete transcriptomas, variando em número absoluto de 6.171 a 31.520. Esses dados podem ser usados para investigar a grande variação cariotípica entre as espécies da família Loricariidae, já que os elementos genéticos móveis são responsáveis por uma proporção significante da variação cariotípica observada em muitos grupos. A grande biodiversidade da bacia amazônica está sendo ameaçada pelos profundos impactos causados pelas atividades humanas na Terra. Apesar disso, o conhecimento da genética de espécies nativas ainda é escasso. Nesse sentido, o banco de dados genéticos produzido neste trabalho, por preencher parte dessa lacuna de informação que existe sobre a biodiversidade da bacia amazônica, é de grande importância .

6. REFERÊNCIAS BIBLIOGRÁFICAS

1. Zalasiewicz J, Williams M, Haywood A, Ellis M. The Anthropocene: a new epoch of geological time? Philos Trans R Soc. 2011;835–41.

2. STEFFEN W, GRINEVALD J, CRUTZEN P, MCNEILL J. The Anthropocene: conceptual and historical perspectives. Philos Trans R Soc. 2011;842–67.

3. Freeman R, McRae L, Deinet S, Marconi V, Loh J. Living Planet Index [Internet]. Recuperado de: http://livingplanetindex.org

4. Mcgill BJ, Dornelas M, Gotelli NJ, Magurran AE. Fifteen forms of biodiversity trend in the Anthropocene. Trends Ecol Evol. Elsevier Ltd; 2015;30(2):104–13.

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. 2014;345(6195):401–6.

7. United Nations. Transformando Nosso Mundo: A Agenda 2030 para o Desenvolvimento Sustentável. a/Res/70/1. 2015;1–49.

8. Zarfl C, Lumsdon AE, Berlekamp J, Tydecks L, Tockner K. A global boom in hydropower dam construction. Aquat Sci. 2014;77(1):161–70.

9. de Faria FAM, Jaramillo P. The future of power generation in Brazil: An analysis of alternatives to Amazonian hydropower development. Energy Sustain Dev. International Energy Initiative; 2017;41:24–35.

10. Lees AC, Peres CA, Fearnside PM, Schneider M, Zuanon JAS. Hydropower and the future of Amazonian biodiversity. Biodivers Conserv. 2016;25(3):451–66.

11. Winemiller KO, McIntyre PB, Castello L, Fluet-Chouinard E, Giarrizzo T, Nam S, et al. Balancing hydropower and biodiversity in the Amazon, Congo, and Mekong. Science (80-). 2016;351(6269):128–9.

12. Instituto Humanitas Unisinos [Internet]. Recuperado de: http://www.ihu.unisinos.br/564569

13. Ideker T, Galitski T, Hood L. A N EW A PPROACH TO DECODING L IFE: Systems Biology. Annu Rev Genomics Hum Genet. 2001;2:343–372.

14. 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.

15. Arif IA, Khan HA, Bahkali AH, Al Homaidan AA, Al Farhan AH, Al Sadoon M, et al. DNA marker technology for wildlife conservation. Saudi J Biol Sci. King Saud University; 2011;18(3):219–25.

16. Chalopin D, Naville M, Plard F, Galiana D, Volff JN. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. Genome Biol Evol. 2015;7(2):567–80.

17. Kidwell MG. Navigating wall-sized displays with the gaze: A proposal for cultural heritage. Genetica. 2002;115(June):49–63.

18. Ayres-Alves T, Cardoso AL, Nagamachi CY, de Sousa LM, Pieczarka JC, Noronha RCR. Karyotypic Evolution and Chromosomal Organization of Repetitive DNA Sequences in Species of Panaque , Panaqolus , and Scobinancistrus (Siluriformes and Loricariidae) from the Amazon Basin. Zebrafish. 2017;00(00):zeb.2016.1373.

19. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.

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

21. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Philip D, Bowden J, et al. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Vol. 8, Nat Protocols. 2013. 1-43 p.

22. Smit, AFA, Hubley, R & Green P. RepeatMasker. 2013; Recuperado de: http://www.repeatmasker.org

23. 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, DNA mapping, Seq Anal. 2015;27(5):3784–6.

24. Moreira DA, Buckup PA, Furtado C, Val AL, Schama R, Parente TE. Reducing the information gap on loricarioidei (Siluriformes) mitochondrial genomics. BMC Genomics. BMC Genomics; 2017;18(1):1–13.

25. Gouy M, Guindon S, Gascuel O. Sea view version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 2010;27(2):221–4.

26. 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(11):2531–40.

27. 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. Elsevier Inc.; 2013;69(2):313–9.

28. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3).

29. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief Bioinform. 2013;14(2):178–92.

30. Miles A. Pysamstats [Internet]. Recuperado de: https://github.com/alimanfoo/pysamstats

31. Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12.

32. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3).

33. Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105–11.

34. Langmead. Bowtie2. Nat Methods. 2013;9(4):357–9.

35. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.

36. BROAD INSTITUTE. Picard [Internet]. Recuperado de: https://broadinstitute.github.io/picard/

37. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. Measuring absorptive capacity. Genome Res. 2010;20:1297–303.

38. Depristo MA, Banks E, Poplin R, Garimella K V., Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–501.

39. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Current Protocols in Bioinformatics. 2013. 1-33 p.

40. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden T. Primer-BLAST: A tool to design taget-specific primers for polymerase chain reaction. BMC Bioinformatics. 2012;13:134.

41. Owczarzy R, Tataurov A V., Wu Y, Manthey JA, McQuisten KA, Almabrazi HG, et al. IDT SciTools: a suite for analysis and design of nucleic acid oligomers. Nucleic Acids Res. 2008;36(Web Server issue):163–9.

42. Covain R, Fisch-Muller S, Oliveira C, Mol JH, Montoya-Burgos JI, Dray S. Molecular phylogeny of the highly diversified catfish subfamily Loricariinae (Siluriformes, Loricariidae) reveals incongruences with morphological classification. Mol Phylogenet Evol. 2016;94:492–517.

43. de Araújo JG, dos Santos MAS, Rebello FK, Isaac VJ. Cadeia comercial de peixes ornamentais do Rio Xingu, Pará, Brasil. Bol do Inst Pesca. 2017;43(2):297–307.

44. Carvalho Júnior JR, Carvalho ASDS, Nunes JLG, Camões A, Bezerra MFDC, Santana AR De, et al. Sobre a Pesca De Peixes Ornamentais Por Comunidades Do Rio Xingu, Pará-Brasil: Relato De Caso. Bol do Inst Pesca. 2009;35(3):521–30.

45. Che R, Sun Y, Sun D, Xu T. Characterization of the miiuy croaker (Miichthys miiuy) transcriptome and development of immune-relevant genes and molecular markers. PLoS One. 2014;9(4).

46. Liao X, Cheng L, Xu P, Lu G, Wachholtz M, Sun X, et al. Transcriptome Analysis of Crucian Carp (Carassius auratus), an Important Aquaculture and Hypoxia- Tolerant Species. PLoS One. 2013;8(4):1–11.

47. Chen X, Mei J, Wu J, Jing J, Ma W, Zhang J, et al. A Comprehensive Transcriptome Provides Candidate Genes for Sex Determination/Differentiation and SSR/SNP Markers in Yellow Catfish. Mar Biotechnol. 2015;17(2):190–8.

48. Mohindra V, Singh A, Barman AS, Tripathi R, Sood N, Lal KK. Development of EST derived SSRs and SNPs as a genomic resource in Indian catfish, Clarias batrachus. Mol Biol Rep. 2012;39(5):5921–31.

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

50. 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(480):97–123.

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

52. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. Nature Publishing Group; 2011;12(10):671–82.

53. Yusoff AM, Hari R, Sitam FT, Rovie-ryan JJ, Karuppannan KV. De novo sequencing, assembly and analysis of eight different transcriptomes from the Malayan pangolin. Sci Rep. 2016;(June):1–11.

54. Biscotti MA, Gerdol M, Canapa A, Forconi M, Olmo E, Pallavicini A, et al. The Lungfish Transcriptome: A Glimpse into Molecular Evolution Events at the Transition from Water to Land. Sci Rep. Nature Publishing Group; 2016;6(February):21571.

55. Zhang R, Ludwig A, Zhang C, Tong C, Li G, Tang Y, et al. Local adaptation of Gymnocypris przewalskii (Cyprinidae) on the Tibetan Plateau. Sci Rep. Nature Publishing Group; 2015;5:1–10.

56. Zhu W, Wang L, Dong Z, Chen X, Song F, Liu N, et al. Comparative Transcriptome Analysis Identifies Candidate Genes Related to Skin Color Differentiation in Red Tilapia. Sci Rep. Nature Publishing Group; 2016;6(April):1–11.

57. Brown TM, Hammond SA, Behsaz B, Veldhoen N, Birol I, Helbing CC. De novo assembly of the ringed seal (Pusa hispida) blubber transcriptome: A tool that enables identification of molecular health indicators associated with PCB exposure. Aquat Toxicol. 2017;185:48–57.

58. Robbins RJ, Krishtalka L, Wooley JC. Advances in biodiversity: Metagenomics and the unveiling of biological dark matter. Stand Genomic Sci. Standards in Genomic Sciences; 2016;11(1):1–17.

59. Chang Z, Wang. Z, Li G. The impacts of read length and transcriptome complexity for de ovo assembly: A simulation study. PLoS One. 2014;9(4):1–8.

60. Chi W, Ma X, Niu J, Zou M. Characterizing the transcriptome and molecular marker s information for roach , Rutilus rutilus. J Genet. 2016;95(1):45–51.

61. Favarato RM, Ribeiro LB, Feldberg E, Matoso DA. Chromosomal mapping of transposable elements of the rex family in the bristlenose catfish, ancistrus (siluriformes, loricariidae), from the amazonian region. J Hered. 2017;108(3):254–61.

7. ANEXOS

7.1. Anexo 1 - Lista de Comandos

Lista de comandos com os parâmetros usados para cada programa.

Trimmomatic

$java -jar trimmomatic-0.32.jar PE -threads 20 -phred33 input_R1.fastq input_R2.fastq output_R1_paired.fastq output_R1_unpaired.fastq output_paired.fastq output_R2.unpaired.fastq ILLUMINACLIP:TruSeq2-PE.fa:2:30:10 LEADING:15 TRAILING:15 SLIDINGWINDOW:4:15 MINLEN:36

FastQC

fastqc input1 input2

Trinity

$Trinity --seqType fq --max_memory 100G --left input_R1.fastq --right input_R2.fastq --CPU 30 --min_contig_length 200 –output out_dir/ --full_cleanup

TransDecoder

$TransDecoder.LongOrfs -t ../arquivo_Trinity.fasta
$grep "complete" Trinity.fasta.transdecoder_dir/longest_orfs.pep | sed -e 's/>//g' | cut -d " " -f1 > complete_candidates
$makeblastdb -dbtype prot -in Trinity.fasta.transdecoder_dir/longest_orfs.pep -parse_seqids -out Predicted
$blastdbcmd -db Predicted -entry_batch complete_candidates | sed -e 's/lcl|//g' > complete.fasta
$blastp -db uniprot_Danio_rerio.fasta -query complete.fasta -out complete.blastp -num_threads 10 -evalue 1e-100 -outfmt "6 qacc sacc pident qlen slen length qcovs evalue bitscore" -max_target_seqs 1
$cat complete.blastp | cut -f1 | sort -u > evidences.list
$makeblastdb -dbtype nucl -in Trinity.fasta.transdecoder_dir/longest_orfs.cds -parse_seqids -out cds
$blastdbcmd -db cds -entry_batch evidences.list | sed -e 's/lcl|//g' > train.fasta
$TransDecoder.Predict -t ../Trinity.fasta --cpu 10 --train train.fasta &

BLAST

$blastn/p/x -query input.fasta -db database -num_threads 10 -max_target_seqs 1 -evalue 1e-10 -outfmt "6 std stitle qlen slen qcovhsp frames" > output.txt

TopHat

$ tophat -p 10 <genome_index_base> <reads1_1[,...,readsN_1]> <[reads1_2,...readsN_2]> -g 200 --keep-fasta-order

Picard

MarkDuplicates

$java -jar picard.jar MarkDuplicates I=arquivo.bam O=marked_duplicates.bam M=marked_dup_metrics.txt MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=600

AddOrReplaceReadGroups

$java -jar picard.jar AddOrReplaceReadGroups I=input.bam O=output.bam RGID=4 RGLB=lib1 RGPL=illumina RGPU=unit1 RGSM=20

CreateSequenceDictionary

$java -jar picard.jar CreateSequenceDictionary R=reference.fasta O=reference.dict

ReorderSam

$java -jar picard.jar ReorderSam I=origanal.bam O=reordered.bam R=Reference.fasta CREATE_INDEX=TRUE TMP_DIR=tmp/

GATK

PrintReads : ReassignOneMappingQualityFilter

$java -jar GenomeAnalysisTK.jar -T PrintReads -R reference.fasta -I input.bam -o output.bam -rf ReassignOneMappingQuality -RMQF 255 -RMQT 60 -U ALLOW_N_CIGAR_READS &

HaplotypeCaller

$java -jar GenomeAnalysisTK.jar -R input_transcriptoma.fasta -T HaplotypeCaller -I input.bam -stand_call_conf 30 -mbq 30 --maxReadsInRegionPerSample 100000 -U ALLOW_N_CIGAR_READS -o output.vcf -bamout bamout.bam

VariantFiltration

$java -jar GenomeAnalysisTK.jar -T VariantFiltration -R reference.fa -V raw_snps.vcf -window 35 -cluster 3 --filterExpression "QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0" --filterName "my_snp_filter" -o filtered_snps.vcf

SELEÇÃO SNP

$java -jar GenomeAnalysisTK.jar -T SelectVariants -R reference.fa -V raw_variants.vcf -selectType SNP -o raw_snps.vcf

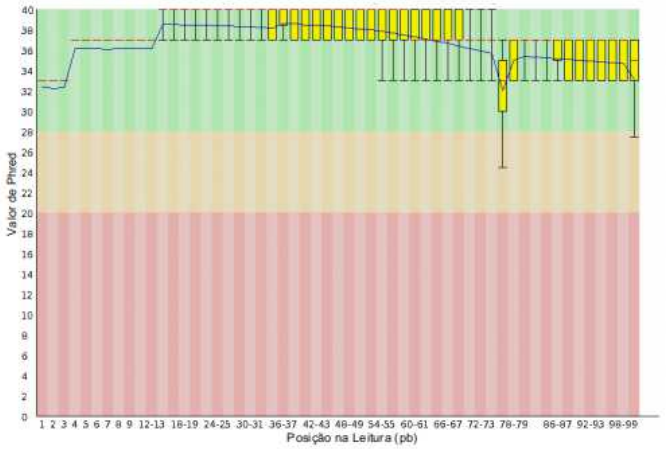
SELEÇÃO INDEL

11)java -jar GenomeAnalysisTK.jar -T SelectVariants -R reference.fa -V raw_variants.vcf -selectType INDEL -o raw_indel.vcf

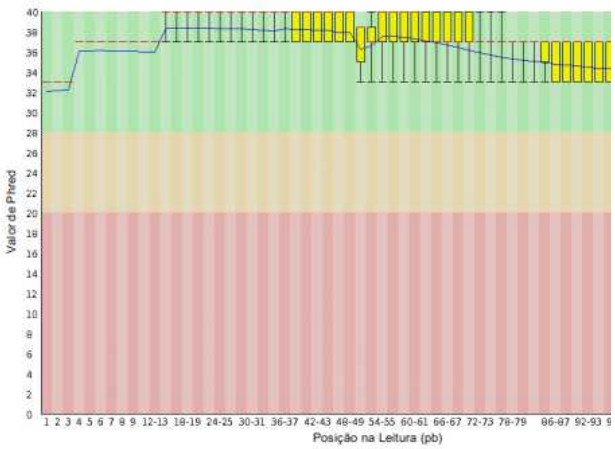
7.2. Anexo 2 - Gráficos dos Valores de “Phred” dos seis órgãos adicionais do *H. zebra*

Avaliação da qualidade das leituras brutas sequenciadas nos transcriptomas dos seis órgãos do *Hypancistrus zebra* (cérebro, coração, fígado, gônada, intestino e rim). O valor de Phred (eixo y) é classificado em três partes de acordo com a qualidade de sequenciamento de cada nucleotídeo (eixo x). O fundo verde delimita o intervalo do valor de Phred considerado muito bom (28>Phred>40). O valor de Phred é uma medida de qualidade da identificação das bases geradas pelo sequenciamento automatizado. Um valor de Phred igual a 30 indica que a taxa de erro na chamada do nucleotídeo é de 1 em cada 1.000.

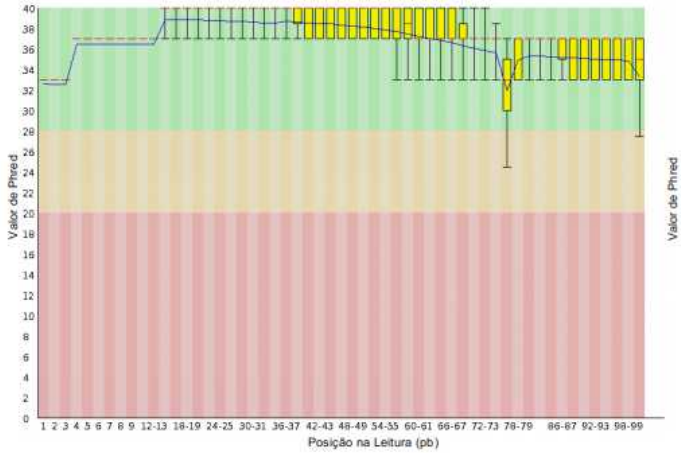
Cérebro R1- Leituras brutas



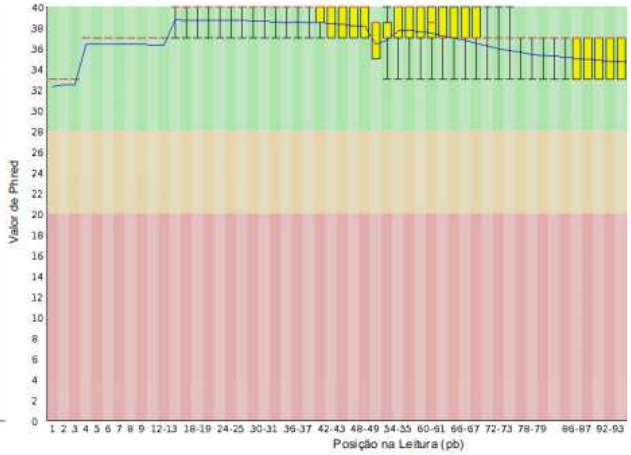
Cérebro R2- Leituras brutas



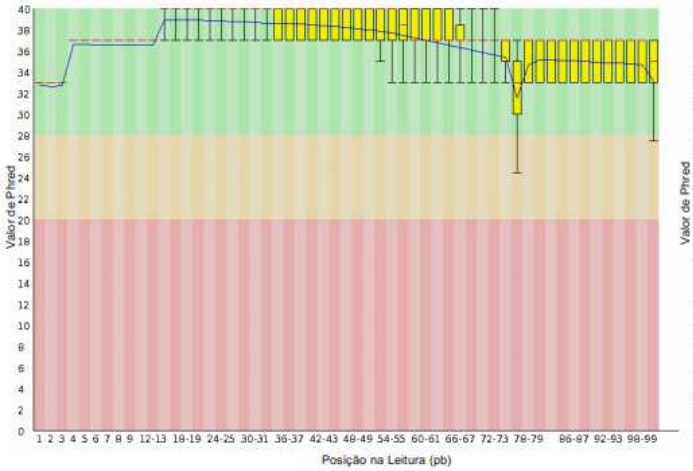
Coração R1- Leituras brutas



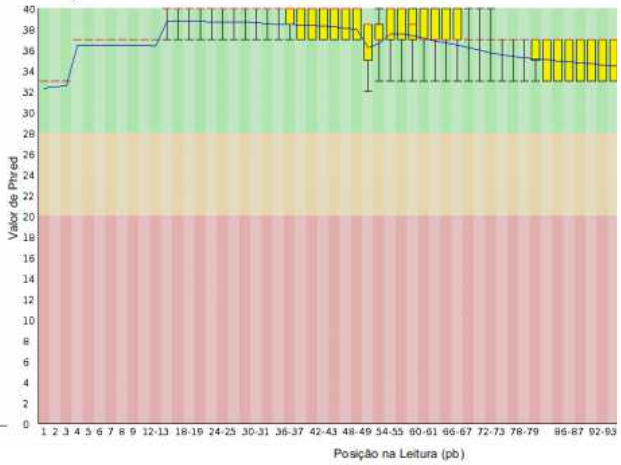
Coração R2- Leituras brutas



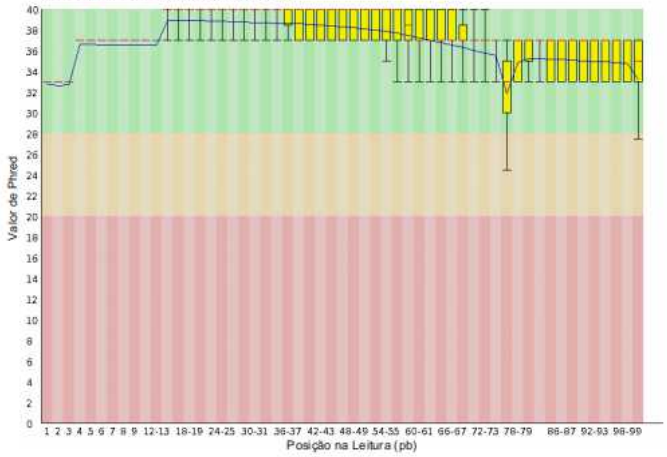
Fígado R1- Leituras brutas



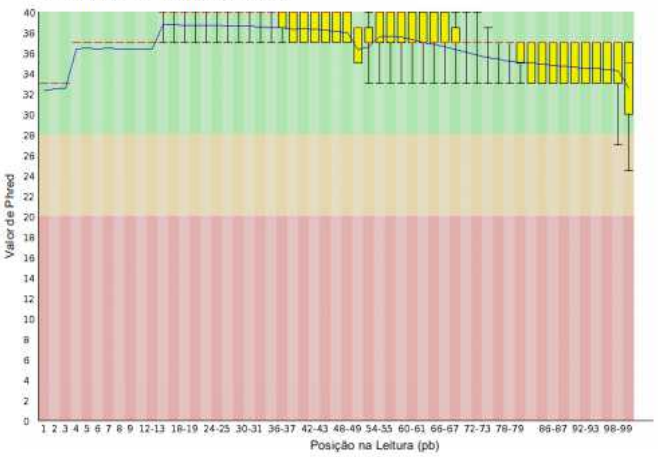
Fígado R2- Leituras brutas



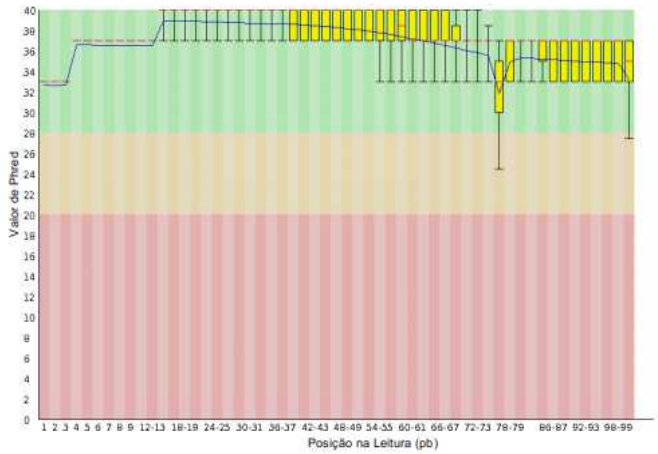
Gônada R1- Leituras brutas



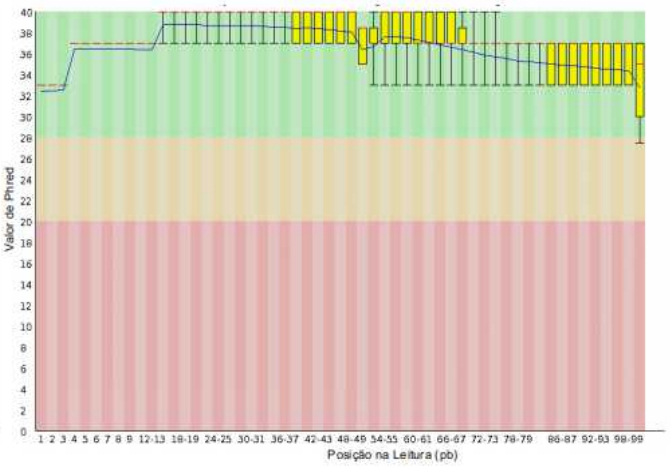
Gônada R2- Leituras brutas



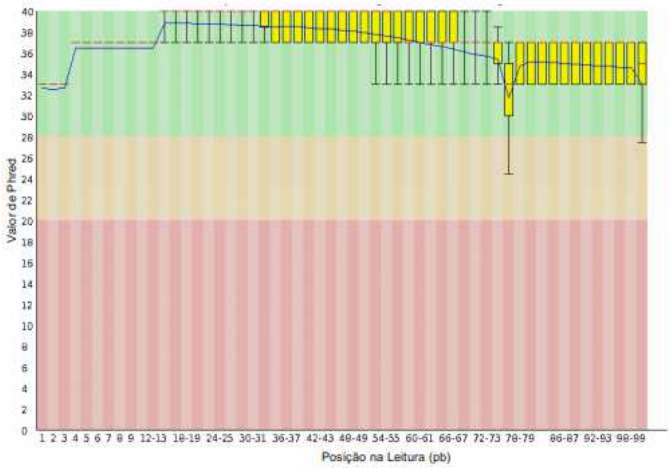
Intestino R1- Leituras brutas



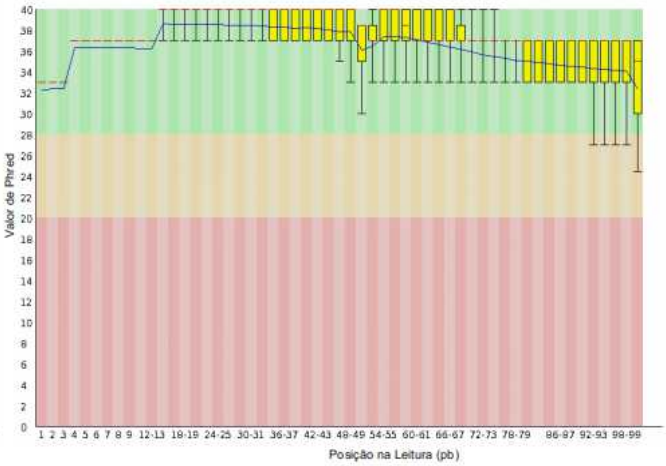
Intestino R2- Leituras brutas



Rim R1- Leituras brutas



Rim R2- Leituras brutas



7.3. Anexo 3 - Profundidade média de sequenciamento

Fórmula da profundidade média de sequenciamento do transcrito.

$$\frac{\text{Número de leituras mapeadas no transcrito} \times \text{Tamanho das leituras mapeadas (pb)}}{\text{Tamanho do transcrito (pb)}}$$

7.4. Anexo 4 – Artigo publicado com alguns dos dados desse trabalho

Artigo publicado durante este trabalho.

Conservation Genet Resour
DOI 10.1007/s12686-016-0645-5

METHODS AND RESOURCES ARTICLE

The mitochondrial genome of *Hypancistrus zebra* (Isbrücker & Nijssen, 1991) (Siluriformes: Loricariidae), an endangered ornamental fish from the Brazilian Amazon

Maithê G. P. Magalhães[1] · Daniel A. Moreira[1] · Carolina Furtado[2] · Thiago E. Parente[1,3]

Received: 3 November 2016 / Accepted: 15 November 2016

Abstract *Hypancistrus zebra* is an ornamental fish endemic to the Xingu river, Amazon basin, which is under the impact zone of the world's fourth-largest hydroelectric dam. Illegal capture is another threat to this species. Despite its critical conservation status only two nucleotide sequences from this fish are publicly available on GenBank and on the BOLD System. Here, the nearly complete mitochondrial genome of *H. zebra* is described; totalizing 16,330 nucleotides, including the complete sequences of the two ribosomal RNA subunits, the 22 transfer RNAs and the 13 protein coding genes. The mitochondrial genome was assembled from transcriptome data, from seven organs, two individual fish, and to an average sequencing depth of 7245x. Seven nucleotide differences were found between the individual fish sequenced here-in, while four were detected among these individuals and the unique mitochondrial sequence publicly available. Additionally, 21 heteroplasmic sites were found among seven organs. This is the first *quasi* complete mitochondrial genome of a species belonging to the *Peckoltia* Clade, a Loricariidae phylogenetically problematic tribe. This genetic resource will be valuable in the efforts to elucidate the phylogenetic relationships among the *Peckoltia* Clade species, as well as it shall subsidize future practices aiming the conservation of this endangered species.

Keywords Aquarium trade · Biodiversity loss · Biopiracy · Conservation genetics · Environmental threats

Introduction

Zebra pleco (*Hypancistrus zebra*, Isbrücker & Nijssen, 1991) is an endangered fish endemic to a ~100 km stretch known as the 'Big Bend' of the Xingu river at the State of Pará on the Brazilian Amazon area (Buckup and Menezes 2007; Rosa and Lima 2008). Its habitat is under the impact zone of Belo Monte hydroelectric dam, which recently started to operate, being the fourth largest in the world (Buckup and Menezes 2007; Rosa and Lima 2008). Another major threat to this species is its illegal capture to supply the international ornamental fish market (Rosa and Lima 2008).

Despite *H. zebra* critical conservation status, only two nucleotide sequences are publicly available. Both sequences were recently submitted to GenBank and integrates a publication showing incongruences between molecular and morphological phylogenies of Loricariidae fish (Covain et al. 2016). One of these sequences is 1788 bp long coding for the partial coding sequence of nuclear F-reticulon (rtn4) gene (KR478530.1), while the other is 2410 bp long containing the partial sequence of the mitochondrial 12S and the 16S ribosomal genes and the complete tRNA-Val gene (KR478209.1).

Here, we assembled a 16,330 bp long quasi-complete mitochondrial genome of *H. zebra*, describing heteroplasmic sites and the relative expression of mitochondrial transcripts among seven organs. The increased genetic

✉ Thiago E. Parente
parente@ensp.fiocruz.br

[1] Laboratório de Toxicologia Ambiental, Escola Nacional de Saúde Pública (ENSP), Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, Brazil

[2] Unidade de Genômica, Instituto Nacional do Câncer (INCA), Rio de Janeiro, Brazil

[3] Laboratório de Genética Molecular de Microrganismos, Instituto Oswaldo Cruz (IOC), Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, Brazil

information will contribute to population genetics studies, which can play a major role in the development of conservation strategies of this species.

Materials and methods

Two specimens of *H. zebra* were used in this study, both provided by Dr. Jansen Zuanon from Instituto Nacional de Pesquisa da Amazônia (INPA), and originated from an apprehension by the Brazilian Federal Police and environmental authorities. Fish were deposited at the Ichthyological collection of the Instituto Nacional de Pesquisas da Amazônia (INPA) under the number INPA 46,655. Tissue samples were dissected and stored in RNA later (Life Technologies) at −20 °C. Total RNA was extracted from six tissues (liver, gill, gut, kidney, gonads, and brain) from one specimen, and from the heart from the other specimen. RNA extractions were performed by phenol/chloroform method using Trizol (Invitrogen), according to the manufacturer's instructions. After extractions, the RNA quantification was performed using a spectrophotometer (Biodrop) and its quality was assessed using the RNA 6000 Nano kit Bioanalyzer (Agilent).

The complementary DNA (cDNA) libraries of each tissue were prepared using 1000 ng of total RNA, following the instructions of TrueSeq RNA Sample v2 kit (Illumina). Each library was identified using specific adapters and quality was determined using Bioanalyzer DNA 1000 Kit (Agilent). The libraries were quantified by qPCR, using the quantification kit for Illumina (kappa Biosystems). The seven libraries were pooled on the same sequencing lane (TrueSeq PE Cluster v3 kit, Illumina). Paired-end sequencing reaction (100pb) was held in a HiSeq2500 using TrueSeq SBS v3 kit (Illumina) at the Instituto Nacional do Cancer (INCA) in Rio de Janeiro, Brazil.

Raw data were demultiplexed using BCL2FASTQ software (Illumina). Adapters sequences and low quality reads were cut off using Trimmomatic and quality was assessed using the program FastQC (v0.11.2). Only reads with Phred score over than 30 were used for the transcriptomes assemblies, which were performed using Trinity 2.0.6. Assemblies statistics were verified by TrinityStats.

The mitochondrial genome was assembled using mitochondrial transcripts from the seven organs (liver, gills, gut, kidney, gonads, brain, heart), following an approach described (Moreira et al. 2015). Mitochondrial transcripts were recovered running BLASTN (v.2.2.31) against the mitogenomes of *Hypoptopoma incognitum*, Aquino & Schaefer, 2010 (NC_028072.1), *Ancistrus* spp., Kner, 1854 (KP960569.1, KP960568.1 and KP960567.1) and *Corydoras nattereri*, Steindachner, 1877 (KT239009.1) (Moreira et al. 2016a, b). Using Seaview and CLUSTAL algorithm for alignments, the recovered mitochondrial transcripts were aligned with *Pterygoplichthys disjunctivus*, Weber, 1991 mitogenome (NC_015747.1), manually edited according to the strand orientation and BLASTN (v.2.2.31) results. The sequence of each contig was manually checked for gaps and inconsistencies. Comparing with others mitogenomes of Loricariidae family, only a small portion of the D-loop region is missing.

The mitogenome annotation was performed using MitoFish (v.3.18) (Iwasaki et al. 2013) and MITOS (revision 656) (Bernt et al. 2013). BAMStats was used to generate mitogenome mapping statistics, such as total reads mapped, average, median and mode. In order to estimate the coverage depth of each base in the sequenced mitogenome, bowtie (v.1.0.0) (Langmead et al. 2009) was used to align the reads to the assembled mitogenome. The aligned reads were visualized using the Integrated Genome Viewer (IGV) (Thorvaldsdóttir et al. 2013) and tabbed using Pysamstats (https://github.com/alimanfoo/pysamstats). Nucleotide positions where the total reads mapped were equal to or greater than 100 and the frequency of the second most abundant base was equal to or greater than 10% were considered heteroplasmic (Moreira et al. 2015).

In order to estimate the expression levels of mitochondrial transcripts, RNA-Seq by Expectation–Maximization (RSEM) was used to quantify the Fragments per kilobase of exon per million fragments mapped (FPKM) (Li and Dewey 2011). FPKM values were normalized using the Trimmed mean of M values (TMM) (Li and Dewey 2011). The expression of the mitochondrial transcripts was compared among the seven organs.

Results

Using mitochondrial transcripts from seven organs of two individual fish a 16,330 bp long sequence, corresponding to the quasi-complete mitochondrial genome of *H. zebra*, was assembled (Fig. 1a). The complete coding sequences of the usual features of Vertebrates mitochondrial genomes (two rRNA, 22 tRNA and 13 protein coding genes) were sequenced. Comparing to the few other Loricariidae mitochondrial genomes available, only a portion of approximately 300 bp at the D-loop region is missing (Fig. 1a).

Sequencing depth varied according to the mitochondrial genome position and among the different organs (Fig. 1b). While protein coding genes presented the highest expression levels, specially COI, COII, COIII and CytB (dark red in Fig. 1b), genes coding for tRNA showed the lowest values (dark blue in Fig. 1b). Taking in account the reads from all the seven tissues, the average sequencing depth was 7245x. The relative expression levels among the seven organs of mitochondrial transcripts were estimated using

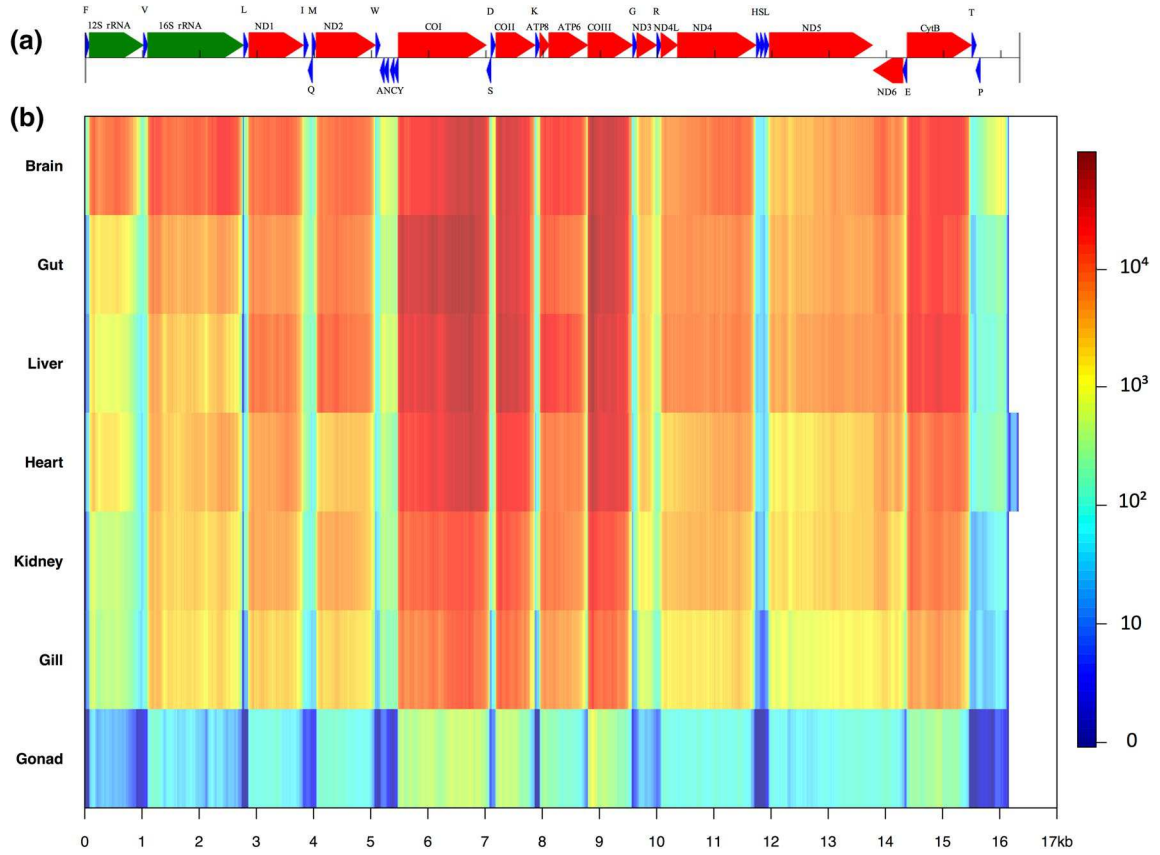


Fig. 1 Linear representation (**a**) and sequencing depth per nucleotide (**b**) for the mitochondrial genome of *H. zebra*. The two rRNA genes are shown green, the 13 protein coding genes in red, and the 22 tRNAs genes in *blue*. Sequencing depth, the absolute number of reads per site, is shown for each tissue as a *color* gradient. The *color* scale is shown in logarithmic scale. Fish used in this study were deposited at Instituto Nacional de Pesquisas da Amazônia under the voucher INPA 46,655. (Color figure online)

FPKM. The brain, with FPKM 15,024,755, had the highest expression of mitochondrial transcripts, followed by intestine (FPKM = 10,481,360), liver (FPKM = 10,227,055), gills (FPKM = 6,284,735), kidney (FPKM = 3,283,662) and gonad (FPKM = 105,136).

The first nine nucleotides of the deposited mitochondrial genome were sequenced only on one specimen. Likewise, a portion in the middle (from 2771 to 2778) and the other end of the sequence (from 16,162 to 16,330) were obtained only on the other specimen. The sequence of both specimens differed by only seven position: T1004C, A2731G, C3209T, T9728C, G11369A, G11727A and G15771A. The comparison with the other mitochondrial sequence public available for *H. zebra* reveals three nucleotide differences (positions 1658, 2174 and 2226) shared with the two specimens sequenced in this study, and a forth difference (position 1004) shared only with one of the two sequences described here-in.

In total, 21 heteroplasmic sites were found in the mitochondrial genome of zebra pleco (Table 1). The positions 13,592 (ND5) and 13,597 (ND5) were considered heteroplasmic in all seven tissues. Positions 653 (12 s rRNA), 656 (12 s rRNA), 7,080 (tRNA-Ser), 9566 (COIII), 14,154 (ND6) and 14,178 (ND6) were considered heteroplasmic in all tissues, except the gonads, what is most probably due to the low sequencing coverage in this tissue. The positions 998 (12 s rRNA) was considered heteroplasmic only in the brain, while position 14,355 (tRNA-Glu) was considered heteroplasmic only in the heart. Interestingly, none of the seven positions by which the two individual zebra pleco differed attended to the criteria to be classified as heteroplasmic.

Discussion

The *quasi* complete mitochondrial genome of *Hypancistrus zebra* described here-in is the first one to be publicly available for a species of the *Peckoltia* (A. Miranda-Ribeiro, 1912) Clade, sensu Lujan et al. 2015, and the fifth from all the >800 species of the Loricariidae family. The *Peckoltia* Clade is a *"historically problematic taxon"*, which was

Table 1 Heteroplasmic sites on the mitochondrial genome of *Hypancistrus zebra*

Gene	Nucl. position	All					Gill				
		A (%)	C (%)	G (%)	T (%)	Counts	A (%)	C (%)	G (%)	T (%)	Counts
12S rRNA	653	86	13	0	0	12,833	87	12	0	0	422
12S rRNA	656	77	22	0	0	11,791	79	21	0	0	371
12S rRNA	998	91	1	7	1	383	100	0	0	0	39
16S rRNA	2129	65	0	0	35	13,407	86	0	0	14	887
16S rRNA	2750	17	6	12	65	987	34	6	13	47	53
tRNA Leu	3823	0	0	100	0	1634	8	1	90	1	145
ND2	5079	14	2	7	77	404	10	0	0	90	30
tRNA Ser	7080	72	4	20	4	4701	70	4	23	2	246
tRNA Ser	7082	14	0	15	70	837	12	0	15	73	26
COII	7869	28	6	17	49	870	27	7	23	43	44
tRNA Lys	7870	26	54	17	4	485	16	79	5	0	19
ATPase 6	8784	0	100	0	0	1699	1	0	0	99	189
COIII	9566	19	10	20	52	2157	22	9	13	56	113
COIII	9568	76	4	14	6	832	77	6	13	4	47
t RNA Gly	9571	25	0	0	75	253	10	0	0	90	21
t RNA Gly	9572	13	87	0	0	223	0	100	0	0	19
ND5	13,592	17	0	83	0	20,878	16	0	84	0	1370
ND5	13,597	15	85	0	0	22,512	13	87	0	0	1467
ND6	14,154	84	16	0	0	18,661	84	16	0	0	1275
ND6	14,178	80	19	0	0	21,720	80	20	0	0	1479
tRNA Glu	14,355	0	0	100	0	2863	98	0	2	0	206
Gene	Nucl. position	Brain					Liver				
		A (%)	C (%)	G (%)	T (%)	Counts	A (%)	C (%)	G (%)	T (%)	Counts
12S rRNA	653	87	13	0	0	9746	89	11	0	0	758
12S rRNA	656	77	23	0	0	9050	79	21	0	0	657
12S rRNA	998	79	2	16	3	146	100	0	0	0	56
16S rRNA	2129	56	0	0	46	7752	93	0	0	7	1150
16S rRNA	2750	16	7	11	66	655	22	6	29	43	49
tRNA Leu	3823	12	1	84	3	773	6	1	91	1	473
ND2	5079	17	4	15	64	111	13	3	1	83	120
tRNA Ser	7080	74	4	18	4	1709	69	3	24	4	815
tRNA Ser	7082	21	0	19	60	273	8	0	8	83	203
COII	7869	30	8	20	41	299	26	3	19	51	160
tRNA Lys	7870	36	30	26	8	151	18	68	14	0	95
ATPase 6	8784	4	1	3	92	714	4	0	1	95	401
COIII	9566	21	10	21	49	846	19	8	17	56	323
COIII	9568	74	3	16	6	291	82	3	12	3	159
t RNA Gly	9571	29	0	0	71	75	16	0	0	84	77
t RNA Gly	9572	22	78	0	0	68	6	94	0	0	71
ND5	13,592	16	0	84	0	6073	17	0	83	0	6079
ND5	13,597	15	85	0	0	6565	14	86	0	0	6500
ND6	14,154	86	14	0	0	6351	84	16	0	0	4943
ND6	14,178	81	19	0	0	7352	82	18	0	0	5949
tRNA Glu	14,355	0	100	0	0	753	99	0	1	0	778

Table 1 (continued)

Gene	Nucl. position	Gonad					Gut				
		A (%)	C (%)	G (%)	T (%)	Counts	A (%)	C (%)	G (%)	T (%)	Counts
12S rRNA	653	82	18	0	0	22	86	14	0	0	1454
12S rRNA	656	86	14	0	0	21	78	22	0	0	1300
12S rRNA	998	100	0	0	0	1	97	0	2	1	104
16S rRNA	2129	91	0	0	9	11	68	0	0	32	2314
16S rRNA	2750	0	0	50	50	2	10	6	15	70	124
tRNA Leu	3823	0	0	86	14	7	5	0	94	1	704
ND2	5079	0	0	0	100	2	17	1	4	78	101
tRNA Ser	7080	67	11	11	11	9	71	4	21	4	1419
tRNA Ser	7082	0	0	20	80	5	13	0	17	70	222
COII	7869	0	0	50	50	2	30	6	13	51	240
tRNA Lys	7870	100	0	0	0	1	26	57	14	2	141
ATPase 6	8784	13	0	0	88	8	3	1	1	95	473
COIII	9566	25	0	0	75	4	17	10	21	51	618
COIII	9568	100	0	0	0	3	71	6	16	6	231
t RNA Gly	9571	0	0	0	100	3	40	0	0	60	52
t RNA Gly	9572	0	100	0	0	3	21	79	0	0	39
ND5	13,592	14	0	86	0	119	16	0	83	0	4857
ND5	13,597	14	86	0	0	125	15	85	0	0	5273
ND6	14,154	72	28	0	0	78	84	16	0	0	4228
ND6	14,178	65	35	0	0	95	80	19	0	0	4859
tRNA Glu	14,355	83	0	17	0	6	96	1	3	0	677
Gene	Nucl. position	Kidney					Heart				
		A (%)	C (%)	G (%)	T (%)	Counts	A (%)	C (%)	G (%)	T (%)	Counts
12S rRNA	653	82	18	0	0	431	87	13	0	0	1042
12S rRNA	656	73	27	0	0	392	77	23	0	0	968
12S rRNA	998	100	0	0	0	37	97	0	3	0	39
16S rRNA	2129	70	0	0	30	1293	54	0	1	46	1956
16S rRNA	2750	19	5	5	71	104	12	5	14	69	153
tRNA Leu	3823	10	3	85	2	226	6	1	92	1	558
ND2	5079	5	3	13	80	40	0	99	1	0	149
tRNA Ser	7080	73	4	20	4	503	68	5	20	7	1183
tRNA Ser	7082	15	0	17	69	108	15	1	26	58	183
COII	7869	20	6	16	58	125	23	5	20	52	178
tRNA Lys	7870	15	72	9	4	78	38	37	19	6	108
ATPase 6	8784	5	3	1	91	265	10	2	4	83	181
COIII	9566	13	10	19	58	253	18	10	17	55	656
COIII	9568	78	2	10	10	101	68	5	18	9	230
t RNA Gly	9571	20	0	0	80	25	22	0	0	78	50
t RNA Gly	9572	13	87	0	0	23	7	91	0	2	43
ND5	13,592	17	0	83	0	2380	16	0	84	0	2558
ND5	13,597	16	84	0	0	2582	15	85	0	0	2778
ND6	14,154	83	17	0	0	1786	86	14	0	0	3345
ND6	14,178	77	23	0	0	1986	83	17	0	0	4117
tRNA Glu	14,355	97	1	2	0	362	85	2	13	0	223

Heteroplasmic sites were defined as any position, with more than 10 reads counts, in which the frequency of the second most abundant nucleotide was equal to or higher than 10%. The frequency for the four nucleotides are shown for each of the heteroplasmic site detected. The exact position and the mitochondrial genome feature in which the polymorphism is located, as well as the total read for each site, are also shown. Fish used in this study were deposited at Instituto Nacional de Pesquisas da Amazônia under the voucher INPA 46,655

recently found to be *"the most genus-rich tribe-level clade"* on the most species rich molecular phylogenetics analysis of Loricariidae fish (Lujan et al. 2015). The phylogenetic relationships recovered by the aforementioned authors for the species on the *Peckoltia* Clade are congruent with the molecular-based phylogeny from Cramer et al., 2011, but only distantly related to previously published phylogenies based on morphological traits (Armbruster 2004, 2008).

In the light of Lujan's molecular phylogeny, a major revision of *Peckoltia* Clade taxonomy is on-going and several species have been already renamed (Armbruster et al. 2015; Ray and Armbruster 2016). Despite the great impact of Lujan's phylogeny on the field, the authors recognized the relationships among genera on the *Peckoltia* Clade as particularly poorly resolved. The resolution of this clade could not be improved due to the limited number of loci traditionally used for molecular phylogeny of Loricariidae fish, as well as to the absence of Clade-specific genetic information to design primers for the amplification of other regions. The release of *H. zebra quasi* complete mitochondrial genome partially cover this knowledge gap and shall assist either the on-going revision and future phylogenetic studies of *Peckoltia* Clade.

Likewise, the mitochondrial genome describe here can assist the development of genetic tools to be used for *H. zebra* conservation. For this purpose, the four nucleotide differences between the *quasi* complete deposited genome and the unique other mitochondrial sequence available for this species, as well as the seven nucleotide differences between the two individuals sequenced in this study, and the 21 heteroplasmies detected are of greater relevance.

Acknowledgements Authors are grateful to Dr. Jansen Zuanon from the Instituto Nacional de Pesquisas da Amazônia (INPA) for the donation of *Hypancistrus zebra* specimens (Voucher INPA 46655) used in this study. T.E.P. thanks Dr. Adalberto L. Val, Nazaré Paula and the entire ADAPTA and LEEM staffs for their hospitality and assistance during his visit to INPA. This work was supported by the U.S. Agency for International Development under Grants PGA-2000003446 and PGA-2000004790; and by the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior (CAPES) from the Brazilian Government for a master fellowship for M. G. P. M., a doctoral fellowship for D. A. M., and a postdoctoral fellowship to T. E. P.

References

Armbruster JW (2004) Phylogenetic relationships of the suckermouth armoured catfishes (Loricariidae) with emphasis on the Hypostominae and the Ancistrinae. Zool J Linn Soc 141:1–80. doi:10.1111/j.1096-3642.2004.00109.x

Armbruster JW (2008) 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 1822:1–76

Armbruster JW, Werneke DC, Tan M (2015) Three new species of saddled loricariid catfishes, and a review of Hemiancistrus, *Peckoltia*, and allied genera (Siluriformes). Zookeys 123:97–123. doi:10.3897/zookeys.480.6540

Bernt M, Donath A, Jühling F et al (2013) MITOS: improved de novo metazoan mitochondrial genome annotation. Mol Phylogenet Evol 69:313–319. doi:10.1016/j.ympev.2012.08.023

Buckup PA, Menezes NA (2007) Catálogo das Espécies de Peixes de Água Doce do Brasil.

Covain R, Fisch-Muller S, Oliveira C et al (2016) Molecular phylogeny of the highly diversified catfish subfamily Loricariinae (Siluriformes, Loricariidae) reveals incongruences with morphological classification. Mol Phylogenet Evol 94:492–517. doi:10.1016/j.ympev.2015.10.018

Cramer CA, Bonatto SL, Reis RE (2011) Molecular phylogeny of the Neoplecostominae and Hypoptopomatinae (Siluriformes: Loricariidae) using multiple genes. Mol Phylogenet Evol 59:43–52. doi:10.1016/j.ympev.2011.01.002

Iwasaki W, Fukunaga T, Isagozawa R 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. doi:10.1093/molbev/mst141

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

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinform 12:323. doi:10.1186/1471-2105-12-323

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 82:269–288. doi:10.1016/j.ympev.2014.08.020

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

Moreira DA, Buckup PA, Andrade PCC et al (2016a) The complete mitochondrial genome of *Corydoras nattereri* (Callichthyidae: Corydoradinae). Neotrop Ichthyol 14:e150167. doi:10.1590/1982-0224-20150167

Moreira DA, Magalhaes MGP, de Andrade PCC et al (2016b) An RNA-based approach to sequence the mitogenome of Hypoptopoma incognitum (Siluriformes: Loricariidae). Mitochondrial DNA Part A, DNA mapping, Seq Anal 27:3784–3786. doi:10.3109/19401736.2015.1079903

Ray CK, Armbruster JW (2016) The genera Isorineloricaria and Aphanotorulus (Siluriformes: Loricariidae) with description of a new species. Zootaxa 4072:501–539. doi:10.11646/zootaxa.4072.5.1

Rosa RS, Lima FCT (2008) Peixes. In: Machado ABM, Drummond GM, Paglia AP (eds) *Livro Vermelho da Fauna Brasileira Ameaçada de Extinção*, 2nd edn. Fundação Biodiversitas, BRASÍLIA, pp 9–278

Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14:178–192. doi:10.1093/bib/bbs017