INSTITUTO NACIONAL DE PESQUISAS DA AMAZÔNIA – INPA
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA CONSERVAÇÃO E BIOLOGIA EVOLUTIVA – PPG-GCBEv

VARIAÇÃO GENÔMICA EM *Paracheirodon axelrodi* (SCHULTZ, 1956) (CHARACIFORMES: CHARACIDAE)

MAYARA KADMAH LAUDELINO DA SILVA

Manaus, Amazonas
Julho, 2024

MAYARA KADMAH LAUDELINO DA SILVA

VARIAÇÃO GENÔMICA EM *Paracheirodon axelrodi* (SCHULTZ, 1956) (CHARACIFORMES: CHARACIDAE)

Orientadora: Izeni Pires Farias
Coorientadora: Cláudia Pereira de Deus

Dissertação apresentada ao Programa de Pós-graduação do Instituto Nacional de Pesquisas da Amazônia – INPA, como parte dos requisitos para obtenção do título de Mestre em Genética, Conservação e Biologia Evolutiva.

Manaus, Amazonas
Julho, 2024

Catalogação na Publicação (CIP-Brasil)

S586v Silva, Mayara Kadmah Laudelino da
Variação genômica em Paracheirodon axelrodi (Schultz, 1956) (Characiformes: Characidae) / Mayara Kadmah Laudelino da Silva; orientadora Izeni Pires Farias; coorientadora Cláudia Pereira de Deus,. - Manaus: [s.l.], 2024.

2,4 MB
58p. : il. color.

Dissertação (Mestrado - Programa de Pós-Graduação em Genética, Conservação e Biologia Evolutiva) - Coordenação do Programa de Pós-Graduação, INPA, 2024.

1. Sequenciamento de nova geração. 2. Montagem de genoma. 3. Genética de populações. I. Farias, Izeni Pires . II. Deus, Cláudia Pereira. III. Título

CDD 570

Sinopse:

Neste trabalho, através do sequenciamento de próxima geração realizado na plataforma Illumina e de soluções em bioinformática, foram analisadas amostras de tetra cardinal *Paracheirodon axelrodi* coletadas em anos anteriores nos municípios de Santa Isabel do Rio Negro e Barcelos (Amazonas, Brasil). Os dados brutos obtidos do sequenciamento foram submetidos à pipelines de bioinformática para (1) a montagem *De Novo* e anotação dos mitogenomas e (2) mapeamento de leituras ao genoma de referência do tetra mexicano *Astyanax mexicanus*. Análises de estrutura populacional realizadas a partir dos SNPs, indicam haver estruturação genética entre as populações das localidades amostradas. A demografia histórica estimada da espécie indica que houve um brusco declínio do tamanho efetivo populacional após o último máximo interglacial, seguida de um constante declínio ao longo do último período glacial, sem haver uma recuperação evidente durante o Holoceno.

Palavras-chave: tetra cardinal, sequenciamento de nova geração, bioinformática, genética de populações.

INPA INSTITUTO NACIONAL DE PESQUISAS DA AMAZÔNIA — UNIDADE DE PESQUISA DO MCTI

PROGRAMA DE PÓS-GRADUAÇÃO PPG GCBEv GENÉTICA, CONSERVAÇÃO E BIOLOGIA EVOLUTIVA

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÕES

GOVERNO FEDERAL BRASIL UNIÃO E RECONSTRUÇÃO

ATA DE DEFESA PÚBLICA DO MESTRADO

PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA, CONSERVAÇÃO E BIOLOGIA EVOLUTIVA

No dia 26 de julho de 2024, às 14h00 (Horário Manaus), reuniu-se a Banca Julgadora da **DEFESA PÚBLICA DE MESTRADO**, composta pelos seguintes Doutores, membros titulares: Jonathan Ready; Fábio de Lima Muniz e Aline Ximenes Mourão, tendo como membros suplentes: Douglas Bastos e Aureo Banhos dos Santos, afim de proceder a arguição pública da **DISSERTAÇÃO** da discente **Mayara Kadmah Laudelino da Silva**, intitulada: **"VARIAÇÃO GENÔMICA EM *Paracheirodon axelrodi* (SCHULTZ, 1956) (CHARACIFORMES: CHARACIDAE)"**. O estudo foi conduzido sob orientação da Profa. Dra. Izeni Pires Farias da UFAM e coorientação da Dra. Cláuia Pereira de Deus do INPA.

Após a exposição da aula, dentro do tempo regulamentar, a discente foi arguida oralmente pelos membros da Banca Julgadora, tendo recebido o conceito final:

(X) Aprovado por unanimidade

( ) Aprovado por maioria

( ) Reprovado

Menção (*se meritório*)

( ) Aprovado com "Distinção" (por maioria)

( ) Aprovado com "Distinção e Louvor" (por unanimidade)

A ATA Foi lavrada e assinada pelos Professores Doutores, membros presentes da Banca Julgadora.

MEMBROS DA BANCA	ASSINATURAS
Jonathan Ready - Univ. Fed. do Pará (UFPA)	gov.br Documento assinado digitalmente JONATHAN STUART READY Data: 26/07/2024 19:00:20-0300 Verifique em https://validar.iti.gov.br
Fábio de Lima Muniz - Univ. Fed. de Rondonópolis (UFR)	gov.br Documento assinado digitalmente FABIO DE LIMA MUNIZ Data: 28/07/2024 23:31:52-0300 Verifique em https://validar.iti.gov.br
Aline Ximenes Mourão - Univ. Fed. do Amazonas (UFAM)	
Douglas Bastos - Inst. Nac. de Pesq. da Amazônia (INPA)	
Aureo Banhos dos Santos - Univ. Fed. do Espírito Santo (UFES)	gov.br AUREO BANHOS DOS SANTOS Data: 29/07/2024 11:38:39-0300 Verifique em https://validar.iti.gov.br

ELIANA FELDBERG, Dra.
Coordenadora do Programa de Pós-Graduação em Genética, Conservação e Biologia Evolutiva - PPG GCBEv.

Esta Ata não tem efeito de conclusão de curso ou diplomação do estudante. Conforme Regulamento PPG GCBEv Art. 62 "Será conferido ao discente otitulo de MESTRE ou DOUTOR em Genética, Conservação e Biologia Evolutiva, desde que cumpridas às exigências das Agências de Fomento, dos regulamentos do PPG-INPA e do PPG GCBEv. Para obtenção do titulo, o estudante deve cumprir, ainda, o exigido nos Arts. 52 ao 55 do Regulamento Geral do INPA e Arts. 61 e 64 do Regulamento PPG GCBEv.

A realização deste estudo foi possível devido:

Ao Instituto Nacional de Pesquisas da Amazônia (INPA) e ao Programa de Pós-graduação em Genética, Conservação e Biologia Evolutiva (PPG GCBEv).

Ao Laboratório de Evolução e Genética Animal (LEGAL) da Universidade Federal do Amazonas (UFAM).

Aos financiamentos fornecidos pelo projeto Dinâmica Populacional e Sustentabilidade do Cardinal Tetra (Paracheirodon axelrodi) no médio Rio Negro da agência de fomento FAPEAM.

À FAPEAM pela concessão da bolsa de Mestrado durante a realização deste estudo.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM) - Programa Institucional de Apoio à Pós-Graduação Stricto Sensu (FAPEAM-POSGRAD).

AGRADECIMENTOS

À minha orientadora, Profª Dra. Izeni Pires Farias, que vem me orientando e abriu as portas do seu grupo de pesquisas para mim lá atrás em 2018, e segue até o presente momento contribuindo para a minha formação e me fazendo agregar conhecimentos e experiências que só pude ter por meio de seu acolhimento, seu conhecimento e sua generosidade em compartilhá-lo tão humildemente com quem a procura.

Ao Dr. Tomas Hrbek, por sua grande colaboração neste estudo ao nos auxiliar na obtenção dos genomas dos cardinais e na realização das análises genômicas; bem como por sua orientação extra oficial ao longo desta dissertação, que foi de suma importância, e a qual sou imensamente grata.

À equipe de cientistas e pesquisadores sem igual do Laboratório de Evolução e Genética Animal (LEGAL) da Universidade Federal do Amazonas (UFAM), que no dia a dia me ensinavam e me auxiliavam dentro do laboratório: Valéria, Nasrah, Aline, Joice, Sandra, Mário, Carlos, Ingrid, Edvaldo, Adriano e Sara.

À Universidade Federal do Amazonas (UFAM), pela estrutura cedida para realização de parte deste estudo e por abrigar o Laboratório de Evolução e Genética Animal (LEGAL), que abriu suas portas para mim e agrega em minha formação desde minha graduação. Ao Instituto Nacional de Pesquisas da Amazônia (INPA) e ao Programa de Pós-graduação em Genética, Conservação e Biologia Evolutiva (PPG GCBEv), por onde tive a oportunidade desenvolver este projeto de mestrado e ao longo destes dois anos me proporcionou experiências memoráveis.

À Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM), pela concessão da bolsa de estudo e por todo o incentivo à pesquisa e formação de pesquisadores no estado do Amazonas que a mesma fornece por meio de seus recursos finanaceiros.

À família em que Deus me permitiu nascer: meu pai, Thiago Coêlho; minha mãe, Crisdeyna Laudelino; minhas irmãs, Stephanie Silva e Giovanna Paiva; minhas avós, Maria Ester e Maria Olga (*in memoriam*). À família que Deus me permitiu construir: meu esposo, Pedro Bittencourt, que além de contribuir neste trabalho, tem sido meu companheiro, meu melhor amigo, meu apoio e a calmaria em que posso repousar quando sinto que tudo parece estar fora dos eixos; e ao meu filho, Heitor Bittencourt, que a cada dia me ensina um pouco mais sobre ser mãe, que me mostra o amor de Deus por mim com um simples sorriso ou um carinho, e que tem me mostrado a real dimensão do amor de uma mãe por um filho. E ao meu bom Deus, que permitiu ter todas essas pessoas maravilhosas na minha vida.

Amo vocês e obrigada por tudo!

EPÍGRAFE

“Tentar não. Faça ou não faça. Tentativa não há”.

Mestre Yoda - Star Wars Episódio IV: Uma Nova Esperança

RESUMO

Dentro da ordem Characiformes, Characidae é a maior e a mais diversa família de peixes da região Neotropical e representa mais de 50% do total da diversidade taxonômica descrita para a ordem. São nos igarapés, igapós e campos interfluviais ao longo das bacias dos rios Negro e Orinoco que são encontradas naturalmente uma das espécies mais emblemáticas do aquarismo mundial: o tetra cardinal *Paracheirodon axelrodi.* Milhões de indivíduos são exportados todos os anos de seus países de origem. Entretanto, as consequências desta extração ainda não são claras do ponto de vista populacional e genético. Além disso, distúrbios climáticos regionais e globais tornam o futuro da espécie incerto. Assim, o objetivo deste trabalho foi gerar um genoma de referência para a espécie e fornecer dados preliminares que possam orientar estratégias de avaliação e manejo das populações naturais e também servir como um *timestamp* do momento atual em que a espécie se encontra para futuras comparações nas próximas décadas. Cinco amostras coletadas em ambientes naturais, georreferenciadas e depositadas em coleções biológicas do Brasil foram sequenciadas através do método PCR Free Whole Genome Sequencing. A partir dos dados brutos do sequenciamento, utilizamos pipelines de bioinformática para (1) a montagem *De Novo* e anotação dos mitogenomas e (2) mapeamento de leituras ao genoma de referência do tetra mexicano *Astyanax mexicanus*. Através da metodologia *De Novo,* obtivemos cinco genomas mitocondriais circulares com tamanho médio de 16.891 ± 92 pb, com características e organização típica de outros mitogenomas de peixes teleósteos. A anotação dos mitogenomas obtidos revelou diferenças em relação aos mitogenomas de *P. axelrodi* já disponíveis no GenBank, como nos tamanhos dos genes COX2 e CYTB e a presença de 3 a 11 elementos repetitivos na extremidade 5’ da região controle do DNA mitocondrial, que podem ser tanto devido a uma variação populacional ou a dificuldade dos algoritmos de bioinformática em lidar com sequências repetitivas. Através do mapeamento ao genoma do tetra mexicano, cerca de 75% das leituras foram mapeadas para as cinco amostras (~ 17,3 Gigabases) e cobrem cerca de 35% da extensão total do genoma do tetra mexicano (1,4 Gigabases). As análises de estrutura populacional a partir dos SNPs indicam haver estrutura entre as localidades amostradas – Santa Isabel do Rio Negro e Barcelos – onde valores de FST são da ordem de 10% e valores de K estiveram entre 4 a 5. A demografia histórica da espécie também foi estimada, indicando um brusco declínio do tamanho efetivo populacional após o último máximo interglacial em seguida de um constante declínio ao longo do último período glacial, sem haver uma recuperação evidente durante o Holoceno.

ABSTRACT

Within the order Characiformes, Characidae is the largest and most diverse family of fish in the Neotropical region and represents more than 50% of the total taxonomic diversity described for the order. It is in the streams, seasonally flooded forests and interfluvial fields along the basins of the Negro and Orinoco rivers that one of the most emblematic species in freshwater fishkeeping is found: the cardinal tetra Paracheirodon axelrodi. Millions of individuals are exported every year from their countries of origin. However, the consequences of this extraction are still unclear from a population and genetic point of view. Furthermore, regional and global climate disturbances make the future of the species uncertain. Thus, the objective of this work was to generate a reference genome for the species and provide preliminary data that can guide assessment and management strategies for natural populations and also serve as a timestamp of the current situation in which the species finds itself for future comparisons in the coming decades. Five samples collected in natural environments, georeferenced and deposited in biological collections in Brazil were sequenced using the PCR Free Whole Genome Sequencing method. From the raw sequencing data, we used bioinformatics pipelines to (1) De Novo assembly and annotate mitogenomes and (2) map reads to the reference genome of the Mexican tetra Astyanax mexicanus. Through De Novo assembly methodology we obtained five circular mitogenomes with an average size of 16,891 ± 92 bp, with characteristics and organization typical of other teleost fish mitogenomes. The annotation of the mitogenomes revealed differences in relation to the mitogenomes of P. axelrodi already available in GenBank, such as the lengths of the COX2 and CYTB genes and the presence of 3 to 11 repetitive elements at the 5' end of the mitochondrial DNA control region, which may be due either to population variation or the difficulty of bioinformatics algorithms to deal with repetitive sequences. Through the reference-guided methodology, about 75% of the five sample's reads (~17.3 Gigabases) were mapped to the Mexican tetra genome, covering about 35% of its total length (1.4 Gigabases). Population structure analyses based on SNPs indicate that there is structure between the sampled locations – Santa Isabel do Rio Negro and Barcelos – where FST values are in the order of 10% and K values were between 4 and 5. The historical demography of the species was also estimated, indicating a sharp decline in effective population size after the Last Interglacial followed by a constant decline throughout the Last Glacial period, without an evident recovery during the Holocene.

LISTA DE TABELAS

Capítulo I

Capítulo II

LISTA DE FIGURAS

Introdução Geral

Capítulo I

Capítulo II

SUMÁRIO

1. INTRODUÇÃO GERAL

1.1 A Espécie Estudada: O Tetra Cardinal *Paracheirodon axelrodi*

Dentro da ordem Characiformes, Characidae é a maior e a mais diversa família de peixes da região Neotropical, com 1.262 espécies válidas, o que representa mais de 50% do total da diversidade taxonômica descrita para a ordem (Fricke et al., 2024). Muitos membros desta família são peixes de tamanho corporal reduzido, com comprimento padrão (CP) < 8 cm, sendo também muito populares entre aquaristas e comumente conhecidos como “tetras” (Mirande, 2019). Diferentemente de outras famílias especiosas de peixes actinopterígeos, os caracídeos apresentam distribuição geográfica apenas na região Neotropical, se estendendo desde o Sudoeste do Texas, México, América Central e América do Sul, até a Patagônia Argentina, sendo especialmente diversos na América do Sul tropical (Van Der Sleen et al, 2018; Mirande, 2019).

A morfologia da família Characidae é altamente conservativa, onde a maior parte de sua variação está relacionada em algum grau com eventos de miniaturização ou hábitos ecológicos, incluindo processos como a reprodução e alimentação (Mirande, 2019). Muitos caracídeos apresentam características que podem ser tanto autapomórficas ou que suportam clados contendo algumas poucas espécies, como o corpo com coloração brilhante e intensa em azul metálico ou faixa lateral azul-verde com densa pigmentação vermelha restrita na região ventral nas espécies de neon tetras do gênero *Paracheirodon* (*P. axelrodi*, *P. innesi*, *P. simulans*), ou como o focinho e cabeça avermelhadas das espécies de rodóstomos do gênero *Petitella* (*Pe. bleheri*, *Pe. georgiae*, *Pe. rhodostoma*), ou a coloração marrom-escura ou preta do lobo inferior da nadadeira caudal, que pode se estender ao pedúnculo caudal ou até mesmo como uma faixa lateral, nos tetras pinguins do gênero *Thayeria* (*T. boehlkei*, *T. ifati*, *T. obliqua*, *T. tapajonica*) (Van Der Sleen et al., 2018).

O tetra cardinal *Paracheirodon axelrodi* (Schultz 1956) (Figura 1) como é comumente chamado, é um peixe endêmico de tributários dos rios Negro, no Brasil, e Orinoco, na Colômbia. De tamanho corporal reduzido (< 35 mm) (Harris & Petry, 2001), existe dimorfismo sexual na espécie, onde as fêmeas apresentam maior tamanho corporal que os machos, principalmente durante a estação reprodutiva, entre os meses de Abril e Maio, durante o período de enchente do rio Negro (Weitzman & Fink, 1983; Rodrigues, 2017). São encontrados em diferentes ambientes aquáticos, como igarapés, igapós e campos interfluviais, onde o acesso a estes ambientes para fins de alimentação, reprodução e sobrevivência depende principalmente do nível da água (Marshall et al., 2008; Marshall et al., 2011). Sua dieta, assim como as demais espécies que compõem este gênero, é composta por microcrustáceos, larvas de mosquitos e algas, estas que acabam sendo uma

grande fonte de energia em períodos onde as águas se encontram em menores níveis (Walker, 2004, Marshall et al., 2008).

Figura 1. *Paracheirodon axelrodi* (tera cardinal). Escala = 3 mm. Foto: Wallice P. Duncan

Fonte: (LINHARES *et al.* Bol. Inst. Pesca 2018, 44(4): e319.)

Por ser uma espécie de coloração brilhante e chamativa, tornou-se um dos peixes de aquário mais populares do mundo e são exportados aos milhões todos os anos. Dados da estatística pesqueira do Estado do Amazonas apontam que entre os anos de 2006 – 2015, mais de 92 milhões de espécimes foram exportados, o que corresponde a mais de 64% do volume total de peixes exportado no período (Tribuzi-Neto, et al. 2020). Entretanto, as consequências da extração de milhões de cardinais anualmente dos seus ambientes naturais ainda não são claras. E com a maior frequência e intensificação de distúrbios climáticos regionais como El Niño e La Niña ou globais como o aquecimento global, o cenário futuro da espécie é incerto. Estudos experimentais demonstram que *P. axelrodi* apresentaram menor tolerância térmica e menores taxas de sobrevivência quando aclimatados artificialmente em ambientes simulando cenários de mudanças climáticas (Campos et al., 2016; Gonçalves et al., 2018).

Assim, o objetivo deste trabalho é gerar um genoma de referência para o tetra cardinal *Paracheirodon axelrodi* e fornecer dados preliminares que possam orientar estratégias de avaliação e manejo das populações naturais e também servir como um *timestamp* do momento atual em que a espécie se encontra para futuras comparações nas próximas décadas.

1.2 Plataformas de Sequenciamento de Nova Geração (NGS)

Muitas metodologias são atualmente implementadas nos estudos genéticos, consequências dos avanços tecnológicos impulsionados pelas plataformas de sequenciamento de nova geração (NGS), e a medida que tal tecnologia vem se desenvolvendo, se desenvolvem também com ela o

número de aplicações correspondentes para a ciência básica e aplicada. Nesses últimos anos, as plataformas de NGSs estão acelerando e impulsionando a pesquisa biológica em diversos campos, como a genômica, transcriptômica, metagenômica, proteogenômica, análise de expressão gênica, descoberta de RNA não codificante, detecção de polimorfismos de nucleotídeo único (do termo em inglês *Single Nucleotide Polymorphism* - SNPs), e a identificação de sítios de ligação de proteínas (El-Metwally et al., 2013). As plataformas de NGS são capazes de produzir um sequenciamento de DNA em grande escala, automatizado, com alto rendimento para muitas amostras a um custo relativamente reduzido (Slatko et al., 2018). Atualmente a plataforma Illumina é a mais utilizada no sequenciamento de DNA, dado a diversidade de ferramentas e protocolos desenvolvidos para obter e tratar os dados gerados por NGS (Kulski, 2016).

Ao utilizar a plataformas de NGS para o sequenciamento genômico, podemos utilizar o grande volume de leituras obtidas para buscar variantes genômicas ao longo dos milhares de fragmentos gerados, como os SNPs. Esta identificação e anotação de variantes pode se dar a partir de uma abordagem comparativa, em que se utiliza um genoma previamente conhecido para servir de referência no mapeamento, ou partindo da abordagem *De Novo*, que não dispõe de um genoma prévio para o mapeamento das leituras. Os SNPs são variações que ocorrem quando um único nucleotídeo é substituído por outro ao longo do genoma, e essas substituições podem ser oriundas de diversos eventos, como mutações pontuais que ocorrem no DNA, como as do tipo transição e transversão (Allendorf et al., 2013). Esses marcadores moleculares são considerados bialélicos e codominantes, são abundantes no genoma e apresentam baixas taxas de mutação; e tendem a ocorrer com maior frequência em regiões do genoma que não codificam proteínas (Turchetto-Zolet et al., 2017).

Realizar o sequenciamento de um genoma inteiro durante uma leitura contínua é impossível atualmente para as tecnologias NGS disponíveis, por conta disso, montar um genoma é um trabalho complexo. O sequenciamento de um genoma pelo método shotgun divide o genoma em milhares de leituras aleatórias e sequencia cada uma das leituras lidas individualmente; reunir essas leituras para reconstruir o genoma inteiro até o nível cromossômico é o que conhecemos e chamamos de montagem do genoma (El-Metwally et al., 2013).

Apesar do seu alto poder e eficiência no sequenciamento genômico, os dados oriundos de NGSs não estão isentos de problemas e erros que precisam ser identificados e corrigidos. Tais erros podem estar relacionados a diversas causas, como a regiões ricas em GC e AT, viés de replicação, erros de substituição, entre outros. Para obter dados de alta qualidade, que sejam imparciais e interpretáveis a partir de NGS, é preciso alcançar profundidade e cobertura de sequência suficientes

para se ter uma confiança estatística; dado que a baixa profundidade de sequenciamento pode contribuir para altas taxas de erro decorrentes de erros de chamada de base e mapeamento, podendo afetar a significância estatística na identificação de genótipos verdadeiros, variantes de nucleotídeos e polimorfismos de nucleotídeo únicos (SNPs) (Kulski, 2016).

Uma ferramenta muito importante que está estritamente conectada às NGSs é a bioinformática, é ela quem vai auxiliar a solucionar os desafios de armazenamento, análise e interpretação de dados NGS. Existem, pelo menos, quatro níveis de análise de sequência de nucleotídeos quando se considera utilizar plataformas de NGS. O primeiro está na utilização de software integrado aos instrumentos de sequenciamento para gerar leituras de sequências; eles convertem os sinais brutos em chamada de base com leituras curtas de sequências de nucleotídeos e índices de qualidade associados. O segundo nível de análise é o alinhamento e montagem de contigs e scaffolds e detecção de variantes. O terceiro é a anotação, integração dos dados e visualização da sequência montada. O último é a fusão de todos os dados gerados pelas plataformas de NGS em um resultado bioinformático único e coerente com ferramentas acessíveis para os interesses biológicos (Kulski, 2016).

Um dos entraves mais comuns quando se lida com dados obtidos a partir de NGS, é que nunca um único pipeline de bioinformática, pacote ou software vai ser capaz de processar todas as informações geradas. Por conta disso, acaba se fazendo necessário o uso e a migração entre diversas plataformas para se conseguir padronizar os dados e gerar informações apropriadas para análise Kulski, 2016). Um conjunto de programas multithread chamado ANGSD desenvolvido por Korneliussen et al. (2014) consegue realizar mapeamento por associação, análises genéticas populacionais como: medida de estrutura populacional, frequência de alelos, testes de mistura e neutralidade; descoberta de SNPs e anotação de variantes usando os dados de sequência bruta. Também é capaz de calcular as probabilidades de genótipos (GL - genotype likelihoods, a probabilidade marginal dos dados de sequenciamento em função de um determinado genótipo), para cada indivíduo em cada sítio, baseado nos *reads* (leituras) alinhados e nos scores de qualidade de sequenciamento e mapeamento. A utilização de GLs é importante quando se lida por exemplo com dados de sequenciamento genômico de baixa cobertura, uma vez que consegue fornecer mais robustez para este tipo de dados e consegue compensar estatisticamente a baixa profundidade das leituras de cada indivíduo analisado (Korneliussen et al., 2014).

Nessa dissertação, foram utilizadas diversas estratégias e soluções em bioinformática para lidar com os dados genômicos obtidos por meio da plataforma Illumina. Soluções estas que se basearam na utilização de *De Novo assembly* para a montagem e anotação do mitogenoma do tetra cardinal e o mapeamento ao genoma de referência do tetra mexicano *Astyanax mexicanus* para a

reconstrução do que vem a ser o primeiro e preliminar genoma de referência para *Paracheirodon axelrodi*.

REFERÊNCIAS BIBLIOGRÁFICAS

Allendorf, F.W.; Luikart, G.H; Aitken, S.N. 2013. *Conservation and the genetics of populations*. John Wiley & Sons.

Campos, D.F.; Jesus, T.F.; Heinrichs-Caldas, W.; Coelho, M.M.; Almeida-Val, V.M.F. 2016. Metabolic rate and thermal tolerance in two congeneric Amazon fishes: Paracheirodon axelrodi Schultz, 1956 and Paracheirodon simulans Géry, 1963 (Characidae). *Hydrobiologia* 789: 133-142. Disponível em: 10.1007/s10750-016-2649-2

El-Metwally, S.; Hamza, T.; Zakaria, M.; Helmy, M. 2013. Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges. *Plos Computational Biology* 9(12) e1003345. Disponível em: https://doi.org/10.1371%2Fjournal.pcbi.1003345

Fricke, R.; Eschmeyer, W.N.; Van Der Slaan, R. 2024. Eschmeyer's catalog of fishes: genera, species, references [Internet]. San Francisco: California Academy of Science. Disponível em: http://researcharchive.calacademy.org/research/ichthyology/catalog/fishcatmain.asp

Gonçalves, L.M.F.; Silva, M.N.P.; Val, A.L.; Almeida-Val, V.M.F. 2018. Differential survivorship of congeneric ornamental fishes under forecasted climate changes are related to anaerobic potential. *Genetics and Molecular Biology* 41(1): 107-118. Disponível em: http://dx.doi.org/10.1590/1678-4685-GMB-2017-0016

Harris, P.M.; Petry,P. 2001. Preliminary report on the genetic population structure and phylogeography of cardinal tetra (Paracheirodon axelrodi) in the rio Negro Basin. p. 205- 225. em: Chao, N.L.; P. Petry; G. Prang; L. Sonneschien & M. Tlusty (Eds.). *Conservation and Management of Ornamental Fish Resources of the Rio Negro Basin, Amazonia, Brazil – Project Piaba*. Editora da Universidade do Amazonas, Manaus, Brazil, p. 303.

Korneliussen, T.S.; Albrechtsen, A; Nielsen, R. 2014. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* 15(356). Disponível em: https://doi.org/10.1186/s12859-014-0356-4

Kulski, J.K. 2016. Next-generation sequencing—an overview of the history, tools, and "Omic" applications. Em: Kulski, J. K. (ed) Next generation sequencing-advances, applications and challenges, *IntechOpen*, London. 10.5772/61964

Linhares, R.M.; Pinagé, C.M.M.F.; Duncan, W.P. 2018. Excessive luminosity fades the skin color of cardinal tetra. *Boletim do Instituto de Pesca* 44(4). Disponível em: DOI: 10.20950/1678-2305.2018.44.4.319

Marshall, B.G; Forsberg, B.R.; Thomé-Souza, M.J.F. 2008. Autotrophic energy sources for Paracheirodon axelrodi (Osteichthyes, Characidae) in the middle Negro River, Central Amazon, Brazil. *Hydrobiologia* (569): 95-103 Disponível em: DOI 10.1007/s10750-007-9060-y

Marshall, B.G.; Forsberg, B.R.; Hess, L.L.; Freitas, C. 2011. Water temperature differences in interfluvial palm swamp habitats of Paracheirodon axelrodi and P. simulans (Osteichthyes: Characidae) in the middle Rio Negro, Brazil. *Ichthyol Explor Freshwaters* 22(4): 377 - 383.

Mirande, J.M. 2019. Morphology, molecules and the phylogeny of Characidae (Teleostei, Characiformes). *Cladistics* 35(2): 282-300. Disponível em: 10.1111/cla.12345

Rodrigues, P.K.S. *Estrutura populacional do tetra cardinal Paracheirodon axelrodi, Schultz 1956 (Characiformes; Characidae) no médio rio Negro, Amazonas – Brasil*. Dissertação (Mestrado em Biologia de Água Doce e Pesca Interior) - Instituto Nacional de Pesquisas da Amazônia, Manaus, 2017.

Slatko, B.E.; Gardner, A.F.; Ausubel, F.M. 2018. Overview of next-generation sequencing technologies. *Current Protocols in Molecular Biology* 122(1):e59. Disponível em: https://doi.org/10.1002/cpmb.59

Tribuzy-Neto, I.A.; Beltrão, H.; Benzaken, Z.S.; Yamamoto, K.C. 2020. Analysis of the ornamental fish exports from the Amazon state, Brazil. *Boletim do Instituto de Pesca* 46(4). Disponível em: https://doi.org/10.20950/1678-2305.2020.46.4.554

Turchetto-Zolet, A.C.; Turchetto, C.; Zanella, C.M.; Passaia, G. 2017. Marcadores moleculares na era genômica: Metodologias e aplicações. *Sociedade Brasileira de Genética*.

Van Der Sleen, P.; Albert, J.S.; Lima, F.C.T.; Netto-Ferreira, A.L.; Mattox, G.M.T; Toledo-Piza, M. 2018. Family Characidae—Tetras and Relatives. In: VAN DER SLEEN, P.; ALBERT, J.S (Eds.). *Field Guide to the Fishes of the Amazon, Orinoco & Guianas*. Princeton University Press, 92-127.

Walker, I. 2004. The food spectrum of the cardinal tetra (Paracheirodon axelrodi, Characidae) in its natural habitat. *Acta Amazonica* 34(1): 69-73.

Weitzman, S.H.; Fink, W.L. 1983. Relationships of the neon tetras, a group of South American freshwater fishes (Teleostei, Characidae), with comments on the phylogeny of New World characiforms. *Bulletin of the Museum of Comparative Zoology at Harvard College* [Internet], 150: 339–95. Disponível em: https://www.biodiversitylibrary.org/part/28701

OBJETIVOS

Capitulo I

Objetivo Geral

Obter o genoma mitocondrial do tetra cardinal *Paracheirodon axelrodi* através da metodologia de montagem *De Novo*.

Objetivos específicos

1. Extrair sequências de DNA mitocondrial a partir de bibliotecas genômicas através de pipelines de bioinformática;
2. Realizar a anotação dos genes mitocondriais através de pipelines de bioinformática;
3. Comparar as anotações obtidas com as já disponíveis em bancos de dados de sequências;
4. Estimar as relações filogenéticas entre a espécie e demais caracídeos utilizando genes codificantes de proteínas (PGCs);
5. Calcular os índices de diversidade nucleotídica entre as populações amostradas.

Capitulo II

Objetivo Geral

Obter um genoma de referência para o tetra cardinal *Paracheirodon axelrodi*.

Objetivos específicos

1. Mapear leituras obtidas a partir de bibliotecas genômicas ao genoma de referência do tetra mexicano *Astyanax mexicanus*;
2. Calcular estatísticas de mapeamento, qualidade e cobertura dos genomas;
3. Realizar a anotação de SNPs variantes em relação ao genoma de referência;
4. Verificar a existência de estrutura populacional entre as localidades amostradas;
5. Estimar o Índice de Fixação FST entre as localidades amostradas;
6. Estimar a demografia histórica da espécie.

CAPÍTULO I

Melhorando a Anotação do Mitogenoma do Tetra Cardinal *Paracheirodon axelrodi* (Schultz, 1956) (Characiformes: Characidae: Stethaprioninae)

Improving the Mitogenome Annotation of the Cardinal Tetra *Paracheirodon axelrodi* (Schultz, 1956) (Characiformes: Characidae: Stethaprioninae)

Mayara Silva[1,2], Pedro Senna Bittencourt[1,3], Cláudia Pereira de Deus[4], Tomas Hrbek[1], Izeni Pires Farias[1].

[1] Laboratório de Evolução e Genética Animal – Legal, Departamento de Genética,Universidade Federal do Amazonas, Av. General Rodrigo Otávio Jordão Ramos, 3000, Coroado, 69077-000, Manaus, Amazonas, Brasil.

[2] Programa de Pós-graduação em Genética, Conservação e Biologia Evolutiva – PPGGCBEv, Instituto Nacional de Pesquisas da Amazônia -INPA, Av. André Araújo, 2936, Aleixo, 69067-375, Manaus, Amazonas, Brasil.

[3] Fundação Espirito-santense de Tecnologia – FEST, Ed. América Centro Empresarial, Av. Fernando Ferrari, Mata da Praia, 29066-380, Vitória, Espírito Santo, Brasil.

[4] Coordenação de Pesquisas em Biodiversidade, Instituto Nacional de Pesquisas da Amazônia - INPA, Av. André Araújo, 2936, Aleixo, 69067-375, Manaus, Amazonas, Brasil.

RESUMO

O sequenciamento completo de genomas através de plataformas de NGS e os avanços da bioinformática para a montagem e anotação de genomas tornaram cada vez mais fáceis a obtenção e análise da estrutura e composição do DNA mitocondrial. Esta crescente geração de dados torna a tarefa de manter bancos de referência curados e anotados por especialistas quase impossível. Tanto GenBank quanto RefSeq contém mitogenomas com erros de anotação dos mais diversos e que podem impactar as análises subsequentes, como análises filogenéticas, DNA *barcoding* e DNA ambiental. O tetra cardinal *Paracheirodon axelrodi* possui dois mitogenomas depositados no GenBank e ambos não possuem informação de procedência ou metodologias utilizadas para montagem e anotação das sequências. A fim de melhor caracterizar o genoma mitocondrial da espécie, este trabalho teve como objetivo realizar a montagem e anotação *De Novo* de cinco mitogenomas de *P. axelrodi* obtidos através de W*hole Genome Sequencing*, todos coletados em ambientes naturais, com amostras georreferenciadas e depositadas em coleções biológicas do Brasil. Com características e organização típica de outros mitogenomas de peixes teleósteos, os mitogenomas gerados apresentam 13 genes codificantes de proteínas (PCGs), 22 RNAs transportadores (tRNAs), 2 RNAs ribossomais (rRNAs) e a região controle, com tamanhos entre 16.799 a 17.039 pb. As principais diferenças em relação aos mitogenomas de *P. axelrodi* disponíveis no GenBank estão no tamanho total do mitogenoma, o códon de início do gene ATP6, os tamanhos do genes COX2 e CYTB e a variação no número de elementos repetitivos na região controle. Apesar de existir variação no códon de início do gene ATP6 entre as espécies de caracídeos analisadas, os tamanhos dos genes COX2 e CYTB foram relativamente conservados, indicando que as anotações de 499/500 pb e 1.077 pb para *P. axelrodi* do GenBank são errôneas. Dada a natureza repetitiva da região D-loop e a diferença de 385 pb observada entre os nossos mitogenomas e os do GenBank, é possível que nossos mitogenomas tenham sido apenas parcialmente recuperados com a metologia *De Novo*.

1. INTRODUÇÃO

O sequenciamento completo de genomas através de plataformas de NGS e os avanços da bioinformática para a montagem e anotação de genomas tornaram cada vez mais fáceis a obtenção e análise da estrutura e composição do DNA mitocondrial. Após a publicação do primeiro mitogenoma de um peixe actinopterígeo (*Formosania lacustris* (Steindacher 1908); Tzeng et al., 1992), o segundo mitogenoma de um peixe actinopterígeo só veio a ser publicado dois anos depois (*Cyprinus carpio* (Linnaeus 1758); Chang, Huang & Lo, 1994). Trinta e dois anos depois, 12.679 mitogenomas completos de actinopterígeos se encontram disponíveis na plataforma NCBI GenBank (ncbi.nlm.nih.gov/genbank/), onde nos últimos cinco anos foram depositados 6.183 mitogenomas e apenas no ano de 2023 foram depositados 1.585 mitogenomas, o que representaria algo em torno de 4,34 mitogenomas/dia.

Esta crescente geração de dados torna a tarefa de manter bancos de referência curados e anotados por especialistas quase impossível. Tanto GenBank quanto RefSeq contém mitogenomas com erros de anotação dos mais diversos e que podem impactar as análises subsequentes, como análises filogenéticas, DNA *barcoding* e DNA ambiental (Bernt et al., 2013; Prada & Boore, 2019). Além disso, muitas destas sequências não possuem informações como local de coleta, número do voucher ou de tombamento em museus ou instituições similares, bem como que tipos de metodologias foram utilizadas em todas as etapas da montagem e anotação destas sequências (Boore, 2006).

O tetra cardinal *Paracheirodon axelrodi* (Schultz 1956) possui dois mitogenomas depositados no GenBank (números de acessos: AB898197 e MH998225) (Zhang et al., 2016; Liu et al., 2019). Ambos os artigos apresentam caracterizações do mitogenoma muito breves, sem informação de procedência ou metodologias utilizadas para montagem e anotação das sequências. O segundo artigo é particularmente curioso, onde fornece coordenadas geográficas para o local de coleta do espécime analisado (15°31'05.0"S 71°45'55.0"W), que fica na Cordilheira dos Andes, no Peru. Além disso, a reconstrução filogenética a partir de Neighbour-Joining dos 12 genes codificantes de proteínas (PCGs) apresentada no trabalho sugere que *P. axelrodi* foi agrupado com *Piaractus brachypomus* (Cuvier 1818) (número de acesso: NC_025315) em 80% das 1.000 réplicas de bootstrap. A fim de melhor caracterizar o genoma mitocondrial da espécie, este trabalho teve como objetivo realizar a montagem, caracterização e anotação de cinco mitogenomas de *P.axelrodi* coletados em ambientes naturais, com amostras georreferenciadas e depositadas em coleções biológicas do Brasil.

2. MATERIAL E MÉTODOS

2.1 Área de Amostragem e Extração de DNA

Amostras de tetra cardinal foram coletadas por colaboradores nos municípios de Barcelos e Santa Isabel do Rio Negro entre os anos de 2012 a 2015, preservadas em tubos com álcool 95% e depositadas na Coleção de Tecidos de Genética Animal (CTGA-UFAM) (Figura 1, Tabela 1). O DNA total foi extraído através da utilização do kit DNeasy® Blood & Tissues Kit (QIAGEN). As amostras foram quantificadas no espectrofotômetro Nanodrop 2000 (Thermo Scientific), a fim de verificar a concentração (ng/µL) e pureza do DNA extraído. Além disso, a integridade do DNA extraído foi visualizada em gel de agarose 1.5% (p/v).

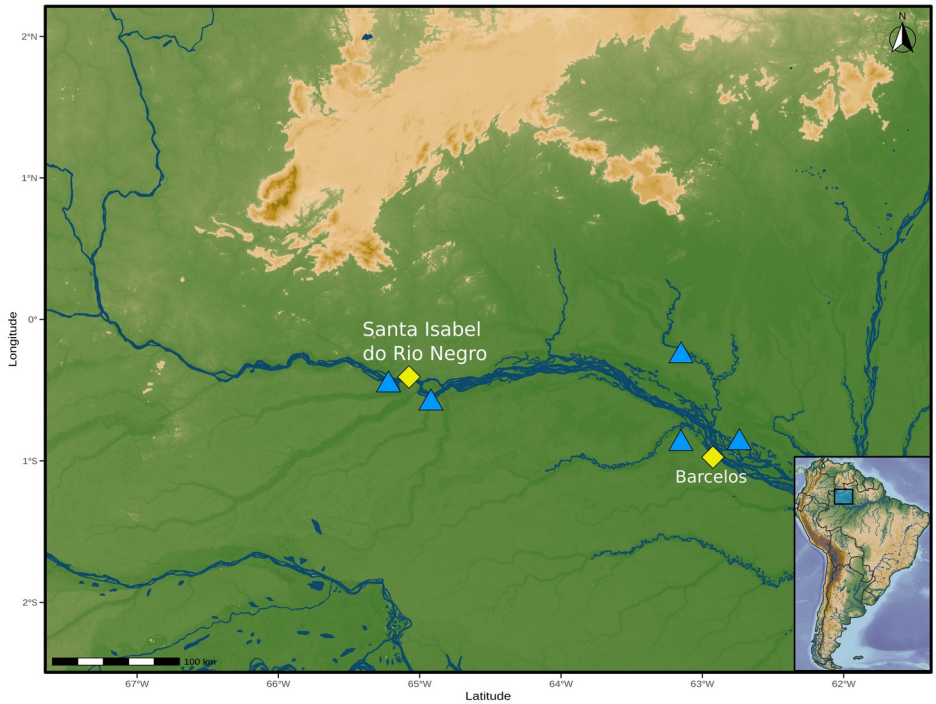

Figura 1. Pontos de coleta das amostras utilizadas neste estudo. Em azul, pontos de coleta; em amarelo, sede dos municípios amostrados.

Tabela 1. Amostragem georreferenciada das localidades utilizadas neste estudo.

Código da amostra	Localidade	Coordenada
CTGA-12368	Santa Isabel do Rio Negro	0°24'55.2"S 65°00'56.1"W
CTGA-105574	Santa Isabel do Rio Negro, rio Aiuanã	00°35'24"S 64°55'10"W
CTGA-15802	Barcelos, rio Negro, Igarapé Puxurituba	00°52.8'25"S 62°40'44.9"W
CTGA-15836	Barcelos, rio Aracá, Igarapé Cutiuaia	00°15'19.9"S 63°3'24.8"W
CTGA-15840	Barcelos, rio Cuiuni, Igarapé Mamulé	00°52'43.9"S 63°13'46.4"W

2.2 Montagem *De Novo* e Anotação do Genoma Mitocondrial

Ao todo, cinco amostras foram selecionadas para serem sequenciadas através do método PCR Free Whole Genome Sequencing na plataforma Illumina NovaSeq 6000. Todas as etapas de preparação das bibliotecas, controle de qualidade e sequenciamento foram feitas pela empresa Novogene (Sacramento, CA). Os dados brutos do sequenciamento foram demultiplexados e tiveram os adaptadores removidos através do programa 'cutadapt 4.4' (Martin, 2011). Em seguida, as leituras pareadas serviram de input para o pipeline de bioinformática 'getOrganelle' (Jin et al., 2020), onde utilizamos a função 'get_organelle_from_reads.py' para realizar montagem do genoma mitocondrial com a utilização de um banco de dados de mitogenomas de referência obtidos do GENBANK RefSeq, com 15 rodadas de extensão (-R 15) e valores de kmer entre 21 e 105 (-k 21,45,65,85,105). Apesar deste programa também fazer a anotação do mitogenoma, nós optamos por utilizar a plataforma MitoAnnotator (Zhu et al., 2023) para a anotação e a plataforma tRNAscan-SE (Chan et al., 2021) foi utilizada para a confirmação dos RNAs transportadores (tRNAs).

2.3 Análises Filogenéticas

Nós extraímos os genes codificantes de proteínas (PCG) dos mitogenomas montados e dos mitogenomas de referência utilizados anteriormente para as análises filogenéticas através do programa Geneious Prime 2024.0.5 (http://www.geneious.com/). Para cada PCG foi criado um arquivo fasta e estes genes foram alinhados separadamente usando MAFFT v 7.490 (Katoh & Standley, 2013) com parâmetros padrão da ferramenta. Nós utilizamos ModelTest-NG (Darriba et al., 2020) para definir modelos de substituição nucleotídica para cada PCG, particionados por posição do códon. Por fim, concatenamos os genes em um único arquivo fasta e estimamos a árvore de máxima verossimilhança a partir de 50 árvores iniciais (25 de parcimônia e 25 randomizadas) e uma lista de modelos de substituição de cada PCG no software RaxML-NG (Kozlov et al., 2019). Valores de suporte dos ramos foram estimados através de 1.000 bootstraps não paramétricos. A árvore foi visualizada através do pacote "ggtree" (Yu et al., 2017), implementado através do software R 4.4.1 (R Core Team, 2024) através da interface gráfica do Rstudio (Posit team, 2023).

2.4 Diversidade Nucleotídica

Para as análises de diversidade nucleotídica, nós utilizamos arquivos FASTA individuais contendo cada um dos 13 PCGs alinhados para estimar esse índice por cada PCG analisado, e um arquivo FASTA com todos os PCGs concatenados para uma análise de sliding window, anotando valores de diversidade nucleotídica em janelas de 300 pb em intervalos de 30 nucleotídeos. Os valores de diversidade nucleotídica foram calculados através da função 'nuc.div' do pacote *pegas*

1.2 (Paradis, 2010) e os intervalos de sliding window foram criados usando a função ‘slidingWindow’ do pacote ‘spider’ 1.5.0 (Brown et al., 2012). Os resultados foram visualizados através dos pacotes ‘ggplot2’ 3.4.4 (Wickham, 2016) e ‘gggenes’ 0.5.1 (Wilkins, 2020).

3. RESULTADOS

Após a extração de DNA das amostras, nós obtivemos concentrações de DNA entre 280 e 699,6 ng/µL e, após sequenciamento das bibliotecas, o número total de leituras brutas (*raw reads*) por amostra variou de 119.998.973 a 181.124.941. Destes, cerca de 0.18% a 0.93% de *reads* foram estimados como correspondentes ao DNA mitocondrial e montados através da metodologia *De Novo Assembly*, gerando genomas circulares com tamanhos entre 16.799 a 17.039 (média= 16.891 ± 92 pb) pares de bases. A proporção média de bases nucleotídicas dos cinco mitogenomas foi de 29,4% para A, 26,0% para C, 15,6% para G e 29,1% para T. O conteúdo médio de G+C (41,6%) foi menor que o conteúdo de A+T (58,4%) (Tabela 2; Figura 2).

Tabela 2. Estatísticas da montagem *De Novo* dos mitogenomas de *P. axelrodi*.

	CTGA-12368	**CTGA-105574**	**CTGA-15802**	**CTGA-15836**	**CTGA-15840**
Raw reads totais (R1+R2)	147.896.134	119.998.973	165.720.559	181.124.941	177.268.147
Tamanho do mitogenoma (pb)	17.039	16.831	16.959	16.830	16.799
Cobertura (X)	517,8	503,3	503,4	497,7	515,5
Proporção do mitogenoma (%)	0.185	0.177	0.937	0.296	0.136
A (%)	29,4	29,3	29,5	29,3	29,3
C (%)	25,9	26,0	26,0	26,0	26,1
G (%)	15,6	15,6	15,5	15,7	15,7
T (%)	29,2	29,1	29,0	29,0	29,0
GC (%)	41,4	41,6	41,5	41,7	41,7

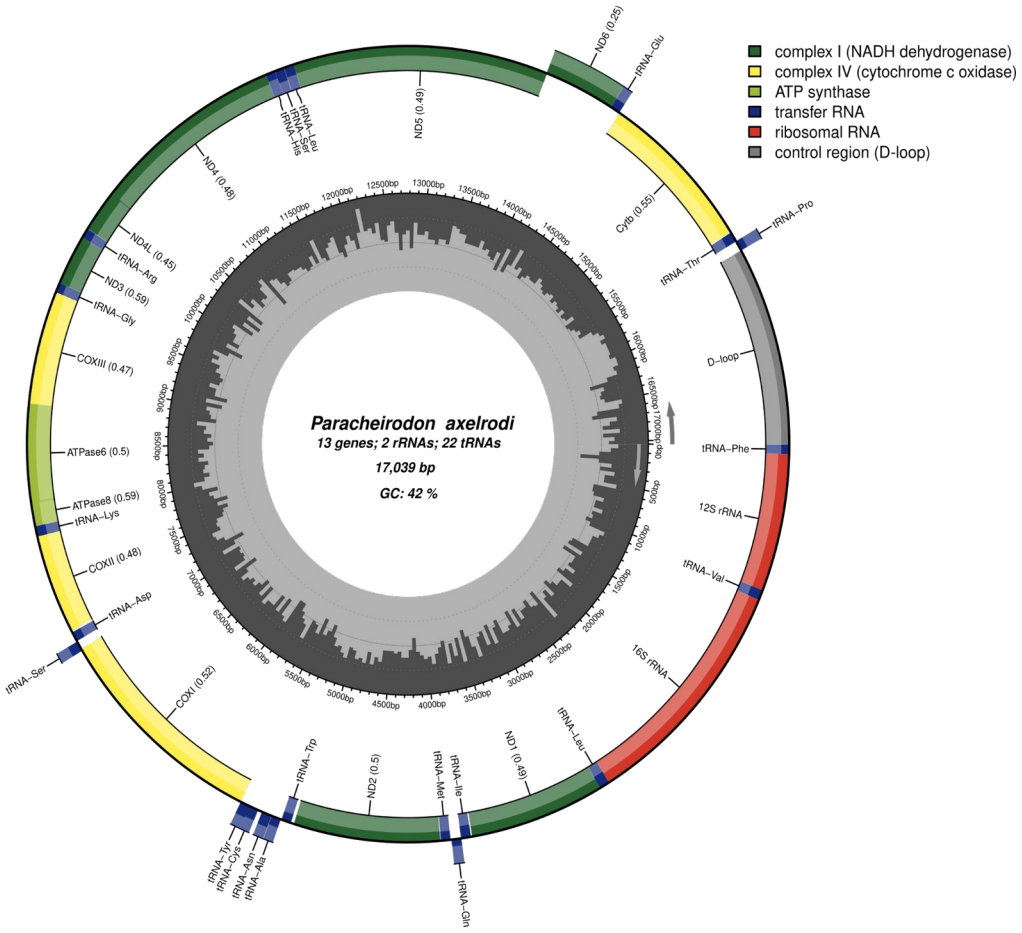

Figura 2. Anotação do mitogenoma de *P. axelrodi* indicando a posição dos genes e a topologia de cada loci: complexo 1 (NADH desidrogenase) em verde escuro; complexo 4 (citocromo c oxidase) em amarelo; ATP sintase em verde claro; RNA transportador em azul; RNA ribossomal em vermelho e região controle - em cinza).

Com características e organização típica de outros mitogenomas de peixes teleósteos, os mitogenomas gerados apresentam 13 genes codificantes de proteínas (PCGs), 22 RNAs transportadores (tRNAs), 2 RNAs ribossomais (rRNAs) e a região controle (Tabela 3). Entre os genes, 28 são codificados na fita pesada (H) e 9 na fita leve (L). Para os PCGs, a média do comprimento total foi de 11.429 pb, onde o menor foi o ATP8 (168 pb) e o maior foi ND5 (1.839 pb). O códon de início de transcrição mais frequente entre os PCGs foi ATG, com exceção apenas do gene ATP6, que apresentou TTG como códon inicial. Entre os códons de parada, TAA foi o mais frequente (6 genes), seguido de T-- (5 genes) e apenas os genes COX1 e ATP8 apresentaram códons de parada AGG e TAG, respectivamente. Entre os rRNAs, a média do comprimento total foi de 2.615 pb, onde o 12S rRNA teve 951 pb e o 16S rRNA variou entre 1.663 a 1.667 pb. Os 22 tRNAs tiveram em média 1.554 pb de comprimento total, variando entre 66 (tRNA-Cys) a 75 (tRNA-Leu)

pares de bases. Finalmente, a região controle do DNA mitocondrial teve um comprimento total médio de 1.229,4 pb ± 103,78, com tamanhos variando entre 1.135 a 1.377 pb. Observamos que a maior parte desta variação se deve à presença de 3 a 9 blocos de DNA repetitivo contendo 35 pb no início da região controle (ACATATAATGCTTAATATTACGCATATGTACTAGT).

Tabela 3. Características do genoma mitocondrial de *P. axelrodi*. Valores em negrito indicam diferenças entre as anotações feitas neste estudo em relação às disponíveis no GenBank.

Nome	Tipo	Fita	Posição	Tamanho	Códon de Início/Parada	Anticódon	Nucleotídeos intergênicos
tRNA-Phe	tRNA	H	1 – 68	68		GAA	0
12S rRNA	rRNA	H	69 – 1.019	951			0
tRNA-Val	tRNA	H	1.020 – 1.091	72		TAC	0
16S rRNA	rRNA	H	1.092 – 2.754	1.663			0
tRNA-Leu2	tRNA	H	2.755 – 2.829	75		TAA	0
ND1	PCG	H	2.830 – 3.801	972	ATG/TAA		+8
tRNA-Ile	tRNA	H	3.810 – 3.881	72		GAT	-2
tRNA-Gln	tRNA	L	3.880 – 3.950	71		TTG	+6
tRNA-Met	tRNA	H	3.957 – 4.026	70		CAT	+3
ND2	PCG	H	4.030 – 5.085	1.056	ATG/TAA		+22
tRNA-Trp	tRNA	H	5.108 – 5.178	71		TCA	+5
tRNA-Ala	tRNA	L	5.184 – 5.252	69		TGC	+1
tRNA-Asn	tRNA	L	5.254 – 5.326	73		GTT	+30
tRNA-Cys	tRNA	L	5.357 – 5.422	66		GCA	-1
tRNA-Tyr	tRNA	L	5.422 – 5.492	71		GTA	+1
COX1	PCG	H	5.494 – 7.053	1.560	ATG/AGG		-13
tRNA-Ser2	tRNA	L	7.041 – 7.112	72		TGA	+3
tRNA-Asp	tRNA	H	7.116 – 7.184	69		GTC	+11
COX2	PCG	H	7.196 – 7.883	**688**	ATG/T--		0
tRNA-Lys	tRNA	H	7.884 – 7.956	73		TTT	+1
ATP8	PCG	H	7.958 – 8.125	168	ATG/TAG		-10
ATP6	PCG	H	8.116 – 8.797	682	**TTG**/T--		0
COX3	PCG	H	8.798 – 9.581	784	ATG/T--		0
tRNA-Gly	tRNA	H	9.582 – 9.651	70		TCC	0
ND3	PCG	H	9.652 – 10.000	349	ATG/T--		0
tRNA-Arg	tRNA	H	10.001 – 10.069	69		TCG	+1
ND4L	PCG	H	10.071 – 10.367	297	ATG/TAA		-7
ND4	PCG	H	10.361 – 11.741	1.381	ATG/T--		0
tRNA-His	tRNA	H	11.742 – 11.810	69		GTG	0
tRNA-Ser1	tRNA	H	11.811 – 11.878	68		GCT	+1
tRNA-Leu1	tRNA	H	11.880 – 11.952	73		TAG	0
ND5	PCG	H	11.953 – 13.791	1.839	ATG/TAA		-4
ND6	PCG	L	13.788 – 14.303	516	ATG/TAA		0
tRNA-Glu	tRNA	L	14.304 – 14.371	68		TTC	+5
CYTB	PCG	H	14.377 – 15.513	**1.137**	ATG/TAA		+5
tRNA-Thr	tRNA	H	15.518 – 15.589	72		TGT	-2
tRNA-Pro	tRNA	L	15.588 – 15.659	72-74		TGG	0
Control region	Control region	H	15.660 – 17.038	1.135 – 1.377			0

Em contraste aos mitogenomas de *P. axelrodi* disponíveis no GenBank, os cinco mitogenomas analisados apresentaram TTG como códon de início do gene ATP6 (vs ATG). Outras diferenças foram nas anotações dos genes COX2 e CYTB, onde obtivemos 688 pb para o gene COX2 (vs 499 pb – AB898197/ 500 pb – MH998225) e 1.137 pb para o gene CYTB (vs 1077 pb – AB898197 e MH998225). A reconstrução filogenética utilizando os 13 PCGs e mitogenomas de referência do RefSeq agruparam as cinco amostras juntamente aos dois mitogenomas de *P. axelrodi* disponíveis e todos estes mais relacionados à sua espécie irmã *P. innesi* do que outras espécies da família Characidae (Figura 3).

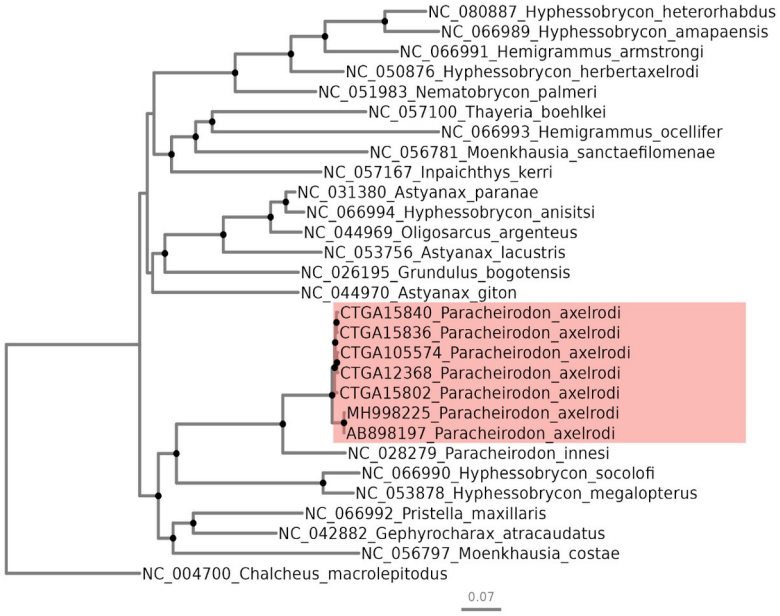

Figura 3. Reconstrução filogenética a partir dos 13 PCGs mitocondriais de *P. axelrodi* e demais espécies de caracídeos disponíveis no RefSeq/GenBank. Nós em preto indicam valores de bootstrap ≥ 75%. Em destaque, amostras de *P. axelrodi* geradas neste estudo e as disponíveis no GenBank.

A diversidade nucleotídica visualizada ao longo dos 11.429 pb dos mitogenomas de *P .axelrodi* através da análise de *sliding window* dos cinco mitogenomas mostra valores de 0 a 0.0173, com valor médio de 0.008 a cada 300 pb analisados em intervalos de 30 pb (Figura 4). Os valores mais altos de diversidade nucleotídica foram observados nos genes ND2, COX1, ND3, ND4L e ND5, enquanto que os valores mais baixos foram observados nos genes COX2, COX3 e ND6. Não foi observada correlação significativa entre os valores de diversidade nucleotídica e conteúdo de GC ao longo dos mitogenomas. Os valores de diversidade nucleotídica variaram entre

0 a 0.014 quando analisados em cada um dos 13 PCGs individualmente, onde, com exceção do gene ATP8, todos foram menores para as amostras de Santa Isabel do Rio Negro quando comparadas com as de Barcelos (Figura 5). Para os genes ND4L e ND4, esta variação foi de uma ordem de magnitude e para os genes COX2, ND3 e ND6 somente os valores para Barcelos foram estimados, dado que não houve variação para estes nas amostras de Santa Isabel do Rio Negro.

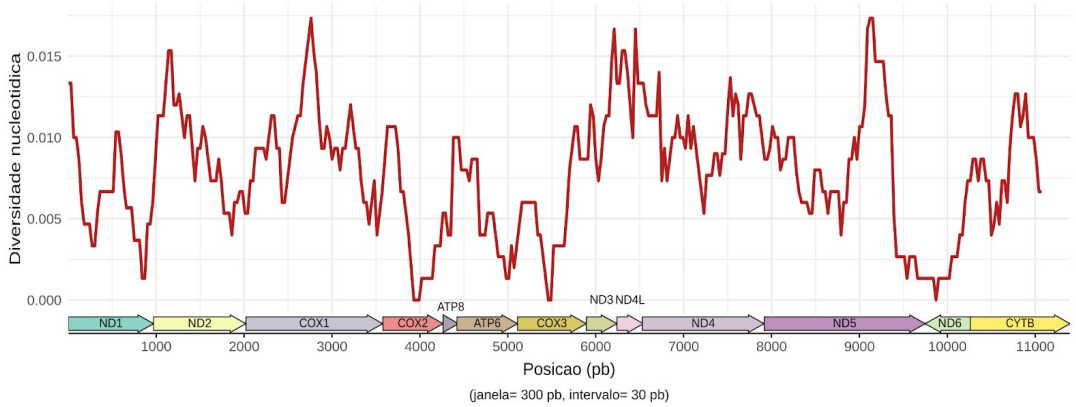

Figura 4. Diversidade nucleotídica observada para os PCGs mitocondriais de *P. axelrodi* através do método de *sliding window* com janela de leitura de 300 pb e intervalo de 30 pb.

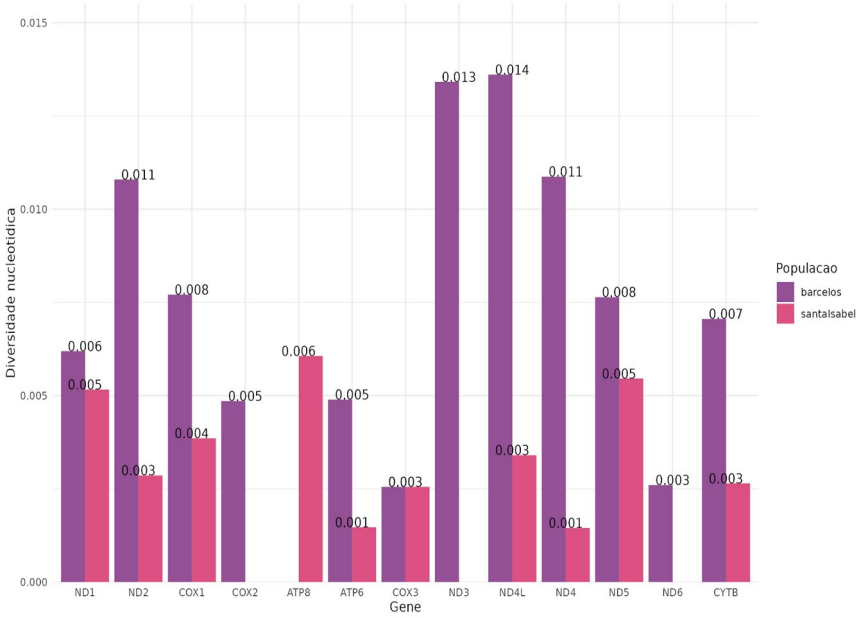

Figura 5. Valores de diversidade nucleotídica observada ao longo dos 13 PCGs mitocondriais entre as duas populações de *P. axelrodi*. Cores representam populações de origem. Barras ausentes indicam diversidade nucleotídica igual a zero.

4. DISCUSSÃO

Os cinco mitogenomas circulares montados e anotados a partir da metodologia *De Novo* apresentaram tamanhos entre 16.799 a 17.039 pb, contendo 13 PCGs, 22 tRNAs, 2 rRNAs e uma região controle, onde apenas o gene ND6 e oito tRNAs (tRNA-Gln, tRNA-Ala, tRNA-Asn, tRNA-

Cys, tRNA-Tyr, tRNA-Ser2, tRNA-Glu, tRNA-Pro) são codificados na fita L enquanto que os demais na fita H, estrutura esta comum em peixes no geral (Satoh et al., 2016). Similar a outros mitogenomas de vertebrados, o conteúdo de A+T foi muito maior que o conteúdo de G+C e a utilização preferencial de códons também foi similar a outros membros da família Characidae (Montag et al., 2023).

Quando comparamos nossos mitogenomas com os disponíveis no GenBank para *P. axelrodi*, as principais diferenças estão no tamanho total do mitogenoma (16.891 ± 92 vs 17.100 pb), o códon de início do gene ATP6 (TTG vs ATG), os tamanhos do genes COX2 (688 vs 499 pb – AB898197/ 500 pb – MH998225) e CYTB (1.137 vs 1.077 pb – AB898197 e MH998225) e a variação no número de elementos repetitivos na região controle (3 vs 11). O códon de início do gene ATP6 observado em nossos mitogenomas (TTG) também foi observado em *P. innesi* e em *Hyphessobrycon heterorhabdus*. Outros códons entre as espécies de caracídeos depositadas no RefSeq foram ATG (*Astyanax giton*, *Gephyrocharax atracaudatus*, *Grundulus bogotensis*, *Hemigrammus armstrongi*, *H. ocellifer*, *Hyphessobrycon amapaensis*, *Hy. anisitsi*, *Hy. herbertaxelrodi*, *Hy. megalopterus*, *Hy. socolofi*, *Inpaichthys kerri*, *Moenkhausia sanctaefilomenae*, *Pristela maxilaris*, *Thayeria boehlkei*), GTG (*Astyanax paranae*, *Astyanax lacustris*, *Oligosarchus argenteus*, *Moenkhausia costae*) e CTG (*Nematobrycon palmeri*). Os tamanhos dos genes COX2 (688 a 691 pb) e CYTB (1.134 a 1.141 pb) também foram relativamente conservados entre as espécies, indicando que as anotações para os dois acessos de *P. axelrodi* depositados no GenBank são errôneas.

A presença/ausência de elementos repetitivos na região controle são conhecidas em várias espécies de caracídeos, onde a variação no número de elementos repetitivos ou sua completa ausência podem ser intrínsecos às populações amostradas ou ao tipo de estratégias de montagem e anotação do mitogenoma (Padhi, 2001; Formenti et al., 2021). Nas amostras que sequenciamos, observamos a presença de três elementos repetitivos na porção 5’ inicial da região controle em quatro amostras e nove elementos repetitivos em um amostra, enquanto que nos mitogenomas disponíveis no GenBank observamos 11 elementos repetitivos – uma diferença de 385 pb entre as amostras analisadas. Esta variação seria suficiente para explicar a diferença de tamanho total dos mitogenomas mencionados anteriormente. Já em *P. innesi*, observamos a presença de elementos repetitivos na extremidade 5’ da região controle com um motivo diferente (ACATACTATGCCTATTACACCTATATGTACTAG) variando de 5 a 11 elementos. Tanto para nossas amostras quanto para *P. innesi* é possível que estas diferenças se devam a particularidades da geração das bibliotecas e da estratégia de montagem e anotação do mitogenoma. Apesar de não

estar descrito em detalhe, os autores do mitogenoma de *P. innesi* (NC_028279) indicam que este foi obtido a partir da plataforma Illumina e montado através da plataforma CLC Genomic WorkBench utilizando o algoritmo IDBA-UD v. 1. Dada a natureza repetitiva da região controle e a diferença de 385 pb observada entre os nossos mitogenomas e os do GenBank, é possível que nossos mitogenomas tenham sido apenas parcialmente recuperados com a metologia *De Novo*.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Bernt, M.; Donath, A.; Juhling, F.; Externbrink, F.; Florentz, C. Fritzsch, G.; et al. 2013. MITOS: Improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution* 69(2): 313-319. Disponível em: https://doi.org/10.1016/j.ympev.2012.08.023

Brown, S.D.J.; Collins, R.P.; Boyer, S.; Lefort, M.C.; Malumbres-Olarte, J.; Vink, C.J.; et al. 2012. SPIDER: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Molecular Ecology Resources* 12(3): 562-565. Disponível em: https://doi.org/10.1111/j.1755-0998.2011.03108.x

Chan, P.P.; Lin, B.Y.; Mak, A.J.; Lowe, T.M. 2021. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research* 49(16): 9077-9096. Disponível em: https://doi.org/10.1093/nar/gkab688

Chang, Y.S.; Huang, F.L.; Lo, T.B. 1994. The complete nucleotide sequence and gene organization of carp (Cyprinus carpio) mitochondrial genome. *Journal of Molecular Evolution* 38: 138-155. Disponível em: https://doi.org/10.1007/BF00166161

Darriba, D.; Posada, D.; Kozlov, A.M.; Stamakis, A.; Morel, B.; Flouri, T. 2020. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Molecular Biology and Evolution* 37(1): 291-294. Disponível em: doi.org/10.1093/molbev/msz189

Formenti, G.; Rhie, A.; Balacco, J.; Haase, B.; Mountcastle, J.; Fedrigo, O.; et al. 2021. Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. *Genome Biology* 22(120). Disponível em: https://doi.org/10.1186/s13059-021-02336-9

Jin, J.J.; Yu, W.; Yang, J.; Song, Y.; Pamphilis, CW.; Yi, T. 2020. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology* 21(241). Disponível em: https://doi.org/10.1186/s13059-020-02154-5

Katoh, K.; Standley, D.M. 2013. MAFFT Multiple Sequence Alignment Software 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* 30(4): 772-780. Disponível em: https://doi.org/10.1093/molbev/mst010

Kozlov, A.M.; Darriba, D.; Flouri, T.; Morel, B.; Stamakis, A. 2019. RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35(21): 4453-4455. Disponível em: doi:10.1093/bioinformatics/btz305

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17(1): 10-12. Disponível em: https://journal.embnet.org/index.php/embnetjournal/article/view/200.doi:https://doi.org10.14806/ej.17.1.200.

Montag, L.F.A.; Koroiva, R.; Santos, A.R.; Magalhães, L.; Cavalcante, G.C.; Silva, C.S.; et al. 2023. The Complete Mitogenome of Amazonian Hyphessobrycon heterorhabdus (Characiformes: Characidae) as a Valuable Resource for Phylogenetic Analyses of Characidae. *Fishes* 8(5). Disponível em: https://doi.org/10.3390/fishes8050233

Padhi, A. 2013. Geographic variation within a tandemly repeated mitochondrial DNA D-Loop region of a North American freshwater fish, Pylodictis olivaris. *Gene* 538(1): 63–68. Disponível em: https://doi.org/10.1016/j.gene.2014.01.020

Paradis, E. 2010. pegas: an R package for population genetics With an integrated-modular approach. *Bioinformatics* 26(3): 419-420. Disponível em: https://doi.org/10.1093/bioinformatics/btp696.

Posit team (2023). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. URL http://www.posit.co/.

Prada, C.F.; Boore, J.L. 2019. Gene annotation errors are common in the mammalian mitochondrial genomes database. *BMC Genomics* 20(73). Disponível em: https://doi.org/10.1186/s12864-019-5447-1

R Core Team (2024). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Satoh, T.P.; Miya, M.; Mabuchi, K.; Nishida, M. 2016. Structure and variation of the mitochondrial genome of fishes. *BMC Genomics* 17(719). Disponível em: https://doi.org/10.1186/s12864-016-3054-y

Tzeng, C.S.; Hui, C.F.; Shen, S.C.; Huang, PC. 1992. The complete nucleotide sequence of the Crossostoma lacustre mitochondrial genome: conservation and variations among vertebrates. *Nucleic Acids Research* 20(18). Disponível em: doi: 10.1093/nar/20.18.4853. PMID: 1408800; PMCID: PMC334242.

Wickham, H. 2016. **g***gplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, ed. 2, New York.

Wilkins, D. (2020). _gggenes: Draw Gene Arrow Maps in 'ggplot2'_. R package version 0.4.1, <https://CRAN.R-project.org/package=gggenes>.

Yu,G.; Smith, D.K.; Zhu, H.; Guan, Y.; Lam, T.T. 2017. GGTREE: An R Packge for Visualization and Annotation of Phylogenetic trees With Their Covariates and Other Associated Data. *Methods in Ecology and Evolution* 8: 28-36. Disponível em: https://doi.org/10.1111/2041-210X.12628

Zhu, T.; Sato, Y.; Sado, T.; Miya, M.; Iwasaki, W. 2023. MitoFish, MitoAnnotator, and MiFish Pipeline: Updates in 10 years. *Molecular Biology and Evolution* 40(3). Disponível em: https://doi.org/10.1093/molbev/msad035

CAPÍTULO II

Em Busca de um Genoma de Referência Completo para um Emblemático Peixe de Aquário: O Tetra Cardinal *Paracheirodon axelrodi* (Schultz, 1956) (Characiformes: Characidae: Stethaprioninae)

Towards a Complete Reference Genome for an Emblematic Aquarium Fish: The Cardinal Tetra *Paracheirodon axelrodi* (Schultz, 1956) (Characiformes: Characidae: Stethaprioninae)

Mayara Silva[1,2], Pedro Senna Bittencourt[1,3], Cláudia Pereira de Deus[4], Tomas Hrbek[1], Izeni Pires Farias[1].

[1] Laboratório de Evolução e Genética Animal – Legal, Departamento de Genética,Universidade Federal do Amazonas, Av. General Rodrigo Otávio Jordão Ramos, 3000, Coroado, 69077-000, Manaus, Amazonas, Brasil.

[2] Programa de Pós-graduação em Genética, Conservação e Biologia Evolutiva – PPGGCBEv, Instituto Nacional de Pesquisas da Amazônia -INPA, Av. André Araújo, 2936, Aleixo, 69067-375, Manaus, Amazonas, Brasil.

[3] Fundação Espirito-santense de Tecnologia – FEST, Ed. América Centro Empresarial, Av. Fernando Ferrari, Mata da Praia, 29066-380, Vitória, Espírito Santo, Brasil.
[4] Coordenação de Pesquisas em Biodiversidade, Instituto Nacional de Pesquisas da Amazônia - INPA, Av. André Araújo, 2936, Aleixo, 69067-375, Manaus, Amazonas, Brasil.

RESUMO

A Bacia Amazônica possui uma riqueza ictiofaunística inigualável quando comparada a outras bacias de drenagem do planeta. São nos igarapés, igapós e campos interfluviais ao longo das bacias dos rios Negro e Orinoco que são encontradas naturalmente uma das espécies mais emblemáticas do aquarismo mundial: o tetra cardinal *Paracheirodon axelrodi*. Por ser uma espécie de coloração brilhante e chamativa, tornou-se um dos peixes de aquário mais populares do mundo e são exportados aos milhões todos os anos. Entretanto, as consequências da extração de milhões de cardinais anualmente dos seus ambientes naturais ainda não são claras. E com a maior frequência e intensificação de distúrbios climáticos regionais e globais, o cenário futuro da espécie é incerto. Assim, o objetivo deste trabalho foi gerar um genoma de referência para o tetra cardinal *P. axelrodi* e fornecer dados que possam orientar estratégias de avaliação e manejo das populações naturais. Cinco amostras coletadas em ambientes naturais, georreferenciadas e depositadas em coleções biológicas do Brasil foram sequenciadas através do método PCR Free Whole Genome Sequencing. Para o mapeamento das bibliotecas, utilizamos o genoma do tetra mexicano *Astyanax mexicanus* como referência e anotamos SNPs autossômicos para as análises subsequentes. Cerca de 73% de todas as leituras geradas (~17,3 Gb) foram mapeadas ao genoma de referência, cobrindo ~35% do genoma de referência (1,4 Gb). O número de SNPs anotados variou de 17.744.913 a 18.698.701 por amostra e o número de SNPs filtrados e não-ligados foi de 303.412 para todas as amostras. A análise de PCA indica a separação entre as amostras provenientes de Barcelos e Santa Isabel do Rio Negro, com valores de K entre 4 a 5 indicados como melhores valores, a depender do algoritmo utilizado. Valores de FST global entre as localidades amostradas foram de 0.146 (total) e de 0.105 (corrigido). Já os valores de FST usando o método de *sliding-window* através dos 25 cromossomos mapeados resultou em valores entre 0,04 e 0,05. Valores de $FST \geq 0.427$ estiveram no percentil 99% da distribuição dos dados, com valores até 10 vezes maiores que as médias por cada cromossomo. A demografia histórica da espécie sugere uma queda brusca do tamanho efetivo populacional da espécie após o último período interglacial seguido de um declínio contínuo durante todo o último período glacial, até o Holoceno. Mudanças nas condições ambientais às quais os

organismos se encontram adaptados desafiam seus limites fisiológicos de maneira direta e a velocidade com que as mudanças climáticas ocorrem aumentam o risco de extinção das espécies. Diante disso, o sequenciamento do genoma completo do tetra cardinal e o monitoramento genético podem vir a detectar estas mudanças em curto prazo e servir de guia para esforços de conservação e que avaliem o risco de extinção da espécie frente às mudanças climáticas.

1. INTRODUÇÃO

A Bacia Amazônica possui uma riqueza ictiofaunística inigualável quando comparada a outras bacias de drenagem do planeta, contando com 2.716 espécies distribuídas entre 529 gêneros, 60 famílias e 18 ordens (Dagosta & De Pinna, 2019). O rio Negro, um dos tributários principais da bacia Amazônica, comporta 1.165 espécies distribuídas entre 389 gêneros, 56 famílias e 17 ordens. Na grande maioria dos ambientes aquáticos disponíveis, a maior composição das assembleias de peixes são da ordem Characiformes, principalmente nos lagos, tributários, praias sazonais e igarapés (Beltrão et al., 2019). A ordem inclui várias espécies de importância econômica, como o tambaqui (*Colossoma macropomum*), dourado (*Salminus brasiliensis*), matrinxã (*Brycon amazonicus*), jaraquis (*Semaprochilodus* spp.), pacus e piranhas (Serrasalmidae) e espécies de aquário comumente conhecidas como tetras (Characidae) (Toledo-Piza et al., 2023).

São nos igarapés, igapós e campos interfluviais ao longo das bacias dos rios Negro e Orinoco que são encontradas naturalmente uma das espécies mais emblemáticas do aquarismo mundial: o tetra cardinal *Paracheirodon axelrodi* (Marshall et al., 2008; Marshall et al., 2011). Descrito em 1956 por dois grupos de pesquisadores independentes (Schultz, 1956; Myers & Weitzmann, 1956) com apenas um dia de diferença entre as publicações, os boatos de que um “novo neon” havia sido descoberto na bacia Amazônica eram verdadeiros. A espécie é facilmente distinguível de quaisquer outros caracídeos devido a sua coloração corporal brilhante e intensa, com uma faixa lateral azul-metálica ou azul-esverdeada que se estende desde a cabeça até a nadadeira adiposa e com uma densa pigmentação vermelha restrita a região ventral que se estende até o pedúnculo caudal (Van Der Sleen & Lima, 2018).

De tamanho corporal reduzido (< 35 mm) (Harris & Petry, 2001), existe dimorfismo sexual na espécie, onde as fêmeas apresentam maior tamanho corporal que os machos, principalmente durante a estação reprodutiva, entre os meses de Abril e Maio, durante o período de enchente do rio Negro (Weitzman & Fink, 1983; Rodrigues, 2017). Sua dieta é composta por microcrustáceos, larvas de mosquitos e algas, estas que acabam sendo uma grande fonte de

energia em períodos onde as águas se encontram em menores níveis (Walker, 2004, Marshall et al., 2008; Marshall et al., 2011).

Por ser uma espécie de coloração brilhante e chamativa, tornou-se um dos peixes de aquário mais populares do mundo e são exportados aos milhões todos os anos. Dados da estatística pesqueira do Estado do Amazonas apontam que entre os anos de 2006 – 2015, mais de 92 milhões de espécimes foram exportados, o que corresponde a mais de 64% do volume total de peixes exportado no período (Tribuzi-Neto, et al. 2020). Entretanto, as consequências da extração de milhões de cardinais anualmente dos seus ambientes naturais ainda não são claras. E com a maior frequência e intensificação de distúrbios climáticos regionais como El Niño e La Niña ou globais como o aquecimento global, o cenário futuro da espécie é incerto. Estudos experimentais demonstram que *P. axelrodi* apresentaram menor tolerância térmica e menores taxas de sobrevivência quando aclimatados artificialmente em ambientes simulando cenários de mudanças climáticas (Campos et al., 2016; Gonçalves et al., 2018).

Assim, o objetivo deste trabalho foi gerar um genoma de referência para o tetra cardinal *Paracheirodon axelrodi* e fornecer dados preliminares que possam orientar estratégias de avaliação e manejo das populações naturais e também servir como um *timestamp* do momento atual em que a espécie se encontra para futuras comparações nas próximas décadas.

2. MATERIAL E MÉTODOS

2.1 Área de Amostragem e Extração de DNA

Amostras de tetra cardinal foram coletadas por colaboradores nos municípios de Barcelos e Santa Isabel do Rio Negro entre os anos de 2012 a 2015 e preservadas em tubos com álcool 95% (Figura 1, Tabela 1). O DNA total foi extraído através da utilização do kit Qiagen DNeasy Blood & Tissues Kit. As amostras foram quantificadas no espectrofotômetro Nanodrop 2000 (Thermo Scientific), a fim de verificar a concentração (ng/µL) e pureza do DNA extraído. Além disso, a integridade do DNA extraído foi visualizada em gel de agarose 1.5% (p/v).

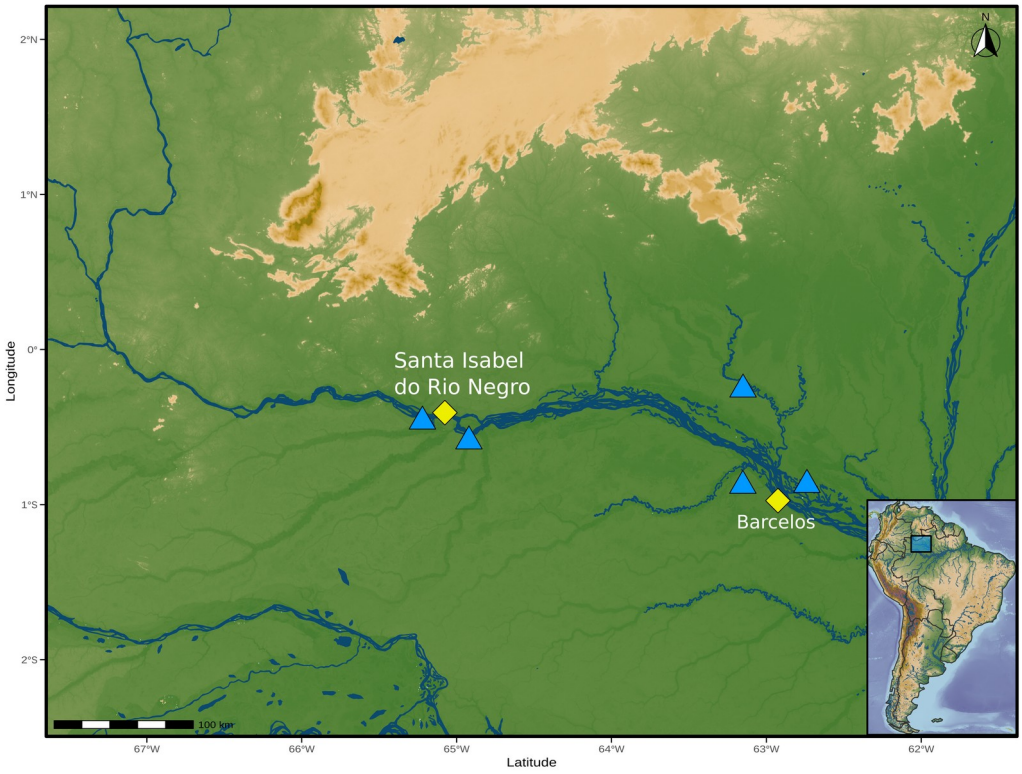

Figura 1. Pontos de coleta das amostras utilizadas neste estudo. Em azul, pontos de coleta; em amarelo, sede dos municípios amostrados.

Tabela 1. Amostragem georreferenciada das localidades utilizadas neste estudo.

Código da amostra	Localidade	Coordenada
CTGA-12368	Santa Isabel do Rio Negro	0°24'55.2"S 65°00'56.1"W
CTGA-105574	Santa Isabel do Rio Negro, rio Aiuanã	00°35'24"S 64°55'10"W
CTGA-15802	Barcelos, rio Negro, Igarapé Puxurituba	00°52.8'25"S 62°40'44.9"W
CTGA-15836	Barcelos, rio Aracá, Igarapé Cutiuaia	00°15'19.9"S 63°3'24.8"W
CTGA-15840	Barcelos, rio Cuiuni, Igarapé Mamulé	00°52'43.9"S 63°13'46.4"W

Ao todo, cinco amostras foram selecionadas para serem sequenciadas através do método PCR Free Whole Genome Sequencing na plataforma Illumina NovaSeq 6000. Para cada amostra, foram geradas bibliotecas contendo leituras pareadas com tamanho máximo de 150 pares de bases. Todas as etapas de preparação das bibliotecas, controle de qualidade e sequenciamento foram feitas pela empresa Novogene (Sacramento, CA).

2.2 Mapeamento e Anotação de Variantes

Para o mapeamento das bibliotecas, utilizamos o genoma do tetra mexicano *Astyanax mexicanus* (25 cromossomos, 84 scaffolds e cobertura 35X; RefSeq GCF_023375975.1) (Warren et al., 2021) como genoma de referência, por ser até o presente momento o único genoma de caracídeo disponível em bancos de dados de sequência como GenBank. Após mapear, ordenar e marcar leituras duplicadas, os arquivos BAM mapeados ao genoma do tetra mexicano serviram de input para a anotação dos polimorfismos de nucleotídeo único (do termo em inglês *Single Nucleotide Polymorphism* – SNPs), que foram realizados de duas maneiras: a primeira através dos softwares GATK 4.5.2.0 (McKenna et al., 2010) e BCFtools 1.51.1, onde foram gerados arquivos VCF (Variant Call Format) contendo os nucleotídeos variantes, sua posição e metadados associados; e a segunda através do software ANGSD 0.941-21-g8a58e17 (Korneliussen et al., 2014), onde foram estimadas as probabilidades de genótipo (Genotype Likelihoods - GL) a partir das frequências alélicas de cada SNP anotado. Em ambas foram aplicados filtros de qualidade restritivos, a fim de selecionar corretamente e reduzir a incerteza na identificação dos SNPs analisados, gerando assim inputs nos formatos VCF e Beagle para as análises subsequentes. Todas as estatísticas de mapeamento das leituras e dos SNPs foram geradas através dos programas SAMtools 1.19 (Danecek et al., 2021), BCFtools 1.51.1 e assembly-stats (https://github.com/sanger-pathogens/assembly-stats). Detalhes das etapas de mapeamento, controle de qualidade e anotação de variantes se encontram no Material Suplementar 1.

2.3 Análise de Componentes Principais (PCA)

Para avaliar a existência de estrutura populacional entre as localidades amostradas, um arquivo Beagle contendo GLs para cada SNP por amostra foi criado e filtrado com os argumentos “-minInd 5 -minMapQ 20 -minQ 20 -remove_bads 1 -C 50 -uniqueOnly 1 -only_proper_pairs 1 -setMinDepth 5 -setMaxDepth 100 -minMaf 0.05 -SNP_pval 1e-6 -skipTriallelic 1”. O arquivo criado foi em seguida filtrado pela posição de SNPs fisicamente próximos (distância menor ou igual a 150 pb) através de um script customizado na linguagem R (R Core Team, 2024) utilizando os pacotes ‘arrow’ (Richardson et al., 2024) e ‘tidyverse’ (Wickham et al., 2019) para leitura e filtragem de dados massivos. Este arquivo filtrado serviu de input para análise de desequilíbrio de ligação no software ngsLD (Fox et al., 2019). Nesta etapa, foram feitas comparações par-a-par para cada SNP anotado e foram mantidos SNPs (1) cuja distância (pb) fosse maior ou igual a 2000 e (2) cujos valores de R^2 fossem menores ou iguais a 0.5, onde 0 indica SNPs não ligados (unlinked) e 1 indica SNPs completamente ligados, também através de um script customizado em R. Os SNPs filtrados e não-ligados remanescentes serviram de entrada para a Análise de Componentes

Principais (PCA) utilizando o software PCAngsd (Meisner & Albrechtsen, 2018). A partir da matriz de covariância gerada pelo PCAngsd, foram extraídos posteriormente os autovalores (eigenvalues) e autovetores (eigenvectors) através da função ‘eigen’ do R base e os resultados da PCA foram visualizados graficamente através dos pacotes ‘ggplot2’ (Wickham, 2016) e ‘patchwork’ (Pedersen, 2024).

2.4 Análises de Admixture

Para verificar a existência de mistura genética entre as populações, realizamos duas abordagens, onde a primeira foi derivando os valores a partir dos resultados da PCA usando PCAngsd com o argumento “--admix” e a segunda foi através do programa NGSadmix (Skotte et al., 2013). Para as análises usando NGSadmix, valores de K entre 1 a 5 foram estimados através de 20 réplicas para cada valor de K. Os resultados gerados no NGSadmix foram clusterizados através da plataforma online ‘Clumpak’ (Kopelman et al., 2015). Para a determinação do melhor valor de K, os valores de probabilidade posterior (LnP(D)) e os de Delta K estimados através do método de Evanno (Evanno et al., 2005) foram visualizados graficamente no R.

2.5 Índice de Fixação (Fst)

As cinco amostras foram organizadas de acordo com suas localidades de origem e foram calculados os seus respectivos GLs. Em seguida, as probabilidades de frequências alélicas das amostras (SAFs) foram estimadas para cada população e entre populações (Joint SFS) através do software ANGSD/realSFS (Mas-Sandoval et al., 2022). Os arquivos resultantes foram usados para estimar o valor de FST global entre as duas populações. Além disso, os valores de Fst foram estimados em uma análise de *sliding window*, anotando valores de Fst a cada 10 megabases (Mb) com intervalos de 5 Mb. Os resultados foram visualizados graficamente no R.

2.6 Demografia Histórica

Para estimar a demografia histórica da espécie, realizamos a análise de PSMC (*Pairwise Sequentially Markovian Coalescent*) (Liu & Hansen, 2017) através do pacote ‘psmcr’ (github.com/emmanuelparadis/psmcr), utilizando como input o genoma de referência do tetra mexicano anotado com SNPs de cardinal através do comando ‘bcftools consensus’. O genoma e o arquivo VCF de cada amostra foram lidos através da função ‘VCF2DNABIN’ e foi intercalado em intervalos a cada 100 bases através da função ‘seqBinning’, sendo em seguida lido pela função ‘psmc’, com os parâmetros ‘parapattern = "4+10*1+20*2+4+6", trratio = 5, niters= 30, B=100’ que já foram utilizados para outras espécies de peixes actinopterígeos (Li et al., 2021). Os resultados

foram visualizados graficamente no R, onde os eixos x e y da figura foram escalados com o tempo de geração de 1.5 ano e taxa de mutação de 5.97e-9 (Bergeron et al., 2023).

3. RESULTADOS

Após o sequenciamento das bibliotecas, o número total de leituras brutas (*raw reads*) por amostra variou de 119.998.973 a 181.124.941. Após a etapa de mapeamento, o número de leituras mapeadas ao genoma de referência do tetra mexicano *A. mexicanus* esteve entre 71,73 a 74,46% do total de leituras geradas. Destes, entre 40,03 a 44,65% das leituras pareadas foram propriamente mapeadas, ou seja, leituras se encontram dentro de uma distância aceitável de seus respectivos pares. A cobertura vertical média variou entre 10,63 a 21,46X e a cobertura horizontal média variou entre 33,89 a 35,48%, cobrindo cerca de um terço do genoma de referência do tetra mexicano (1,4 Gigabases). O número de SNPs anotados variou de 17.744.913 a 18.698.701 (Tabela 2).

Tabela 2. Estatísticas do mapeamento do genoma de *P. axelrodi*.

	CTGA 12368	**CTGA 105574**	**CTGA 15802**	**CTGA 15836**	**CTGA 15840**
Raw reads	147.896.134	119.998.973	165.720.559	181.124.941	177.268.147
Reads mapeados	109.596.184	89.357.097	119.009.751	129.928.183	129.332.140
Reads mapeados (%)	74,10	74,46	71,81	71,73	72,96
Reads propriamente mapeados	65.767.672	53.576.688	66.535.528	72.508.786	74.617.114
Reads propriamente mapeados (%)	44,47	44,65	40,15	40,03	42,09
Singletons	7.877.634	6.503.586	9.920.268	10.918.543	10.200.339
Singletons (%)	5,33	5,42	5,99	6,03	5,75
Bases mapeadas	16.312.936.338	13.304.321.642	17.752.170.105	19.381.083.757	19.263.555.431
Taxa de erro	0,074	0,073	0,074	0,074	0,074
SNPs anotados	18.314.055	17.744.913	18.542.561	18.653.065	18.698.701

Indels anotados	5.368.935	5.059.129	5.524.752	5.583.070	5.579.542
SNPs multialélicos	64.234	57.255	66.660	65.785	66.096
Consenso N50	51.779.647	51.783.652	51.774.855	51.774.719	51.776.334
Consenso L50	11	11	11	11	11
Cobertura horizontal (%)	34,68	33,89	35,14	35,48	35,29
Cobertura vertical (média)	12,49	10,63	15,70	21,46	20,90
BaseQ (média)	35,46	35,46	35,48	35,34	35,33
MapQ (média)	8,32	8,52	8,46	8,53	8,46

Um total de 571.708 SNPs foram mantidos em todas as cinco bibliotecas após a leitura e filtragem dos arquivos BAM processados pelo ANGSD para as análises de estrutura populacional. Após os filtros de distância entre SNPs (dist ≤ 150 pb) e distância e R^2 (dist ≥ 2000 & $R^2 \leq 0.5$), o número de SNPs filtrados e não-ligados foi de 303.412. A partir deste banco de dados, os resultados da PCA indicam a separação das amostras de Barcelos e Santa Isabel do Rio Negro nas comparações entre os eixos PC1-PC2 e PC1-PC3, enquanto que nos eixos PC2-PC3 esta diferenciação não é evidente. Em todos os eixos visualizados, as amostras CTGA105574 e CTGA12368 foram proximamente relacionadas, enquanto que as amostras de CTGA15802 e CTGA15836 são próximas apenas no PC1-PC2. (Figura 2).

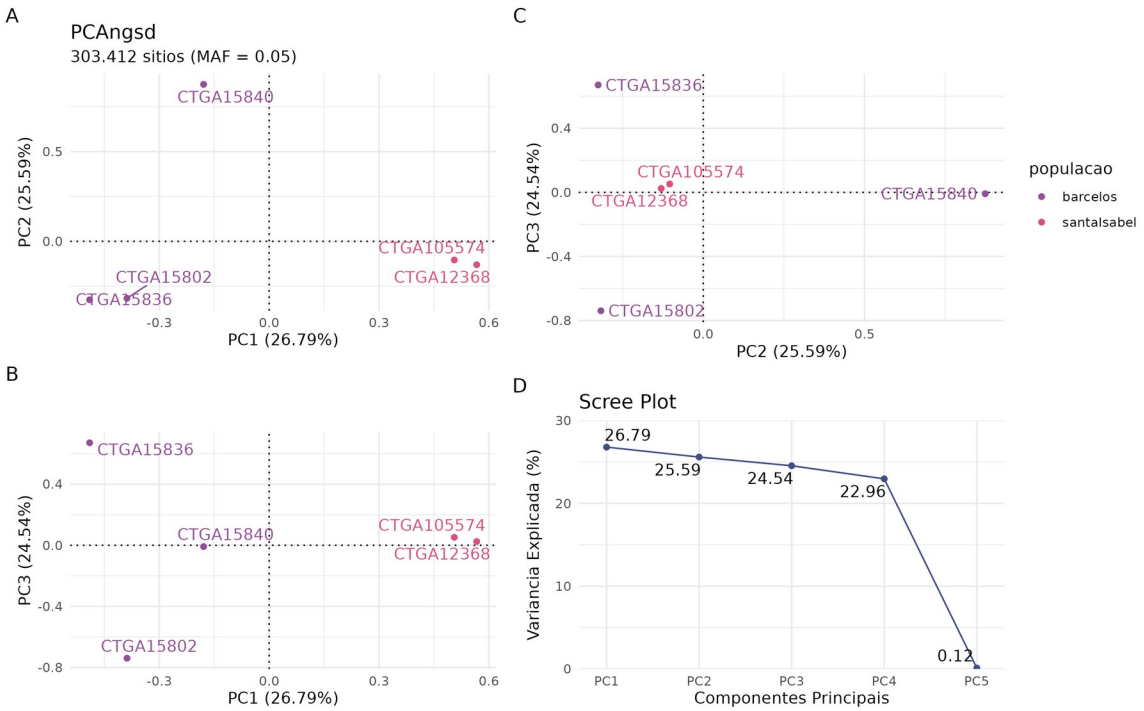

Figura 2. Análise de Componentes Principais (PCA) obtida através do programa ANGSD. Cores indicam as populações de origem. A) Componentes Principais 1 e 2 (PC1-P2); B) Componentes Principais 1 e 3 (PC1-PC3); C) Componentes Principais 2 e 3 (PC2-PC3); D) Variância explicada por cada Componente Principal.

A análise de admixture a partir dos resultados da PCA indica um valor de K=4, onde as amostras CTGA105574 e CTGA12368 possuem uma pequena proporção de mistura, entre 1 a 2% (Figura 3). Já as análises realizadas com NGSadmix, tanto as médias de probabilidade posterior quanto valores de delta K após 20 réplicas indicam que K=5 seria a melhor partição do banco de dados, o que sugere que cada amostra faz parte de uma população individual e sem compartilhamento de alelos (Figura 4).

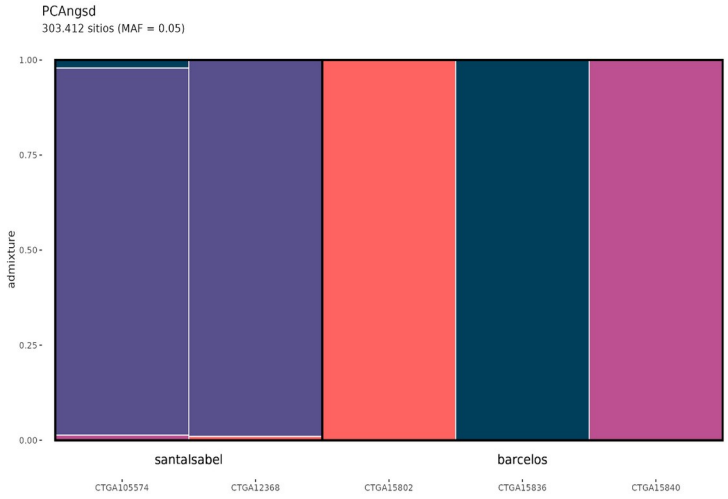

Figura 3. Análise de admixture a partir da PCA do PCAngsd. Cores indicam a atribuição populacional inferida pela análise (K= 4).

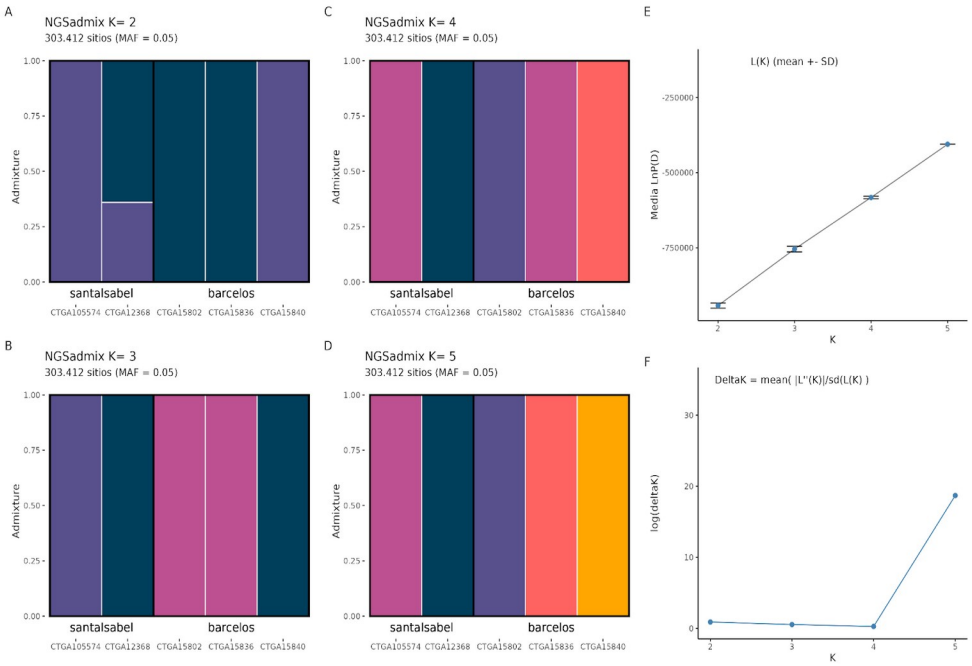

Figura 4. Análise de admixture através do programa NGSadmix. Cores indicam a atribuição populacional inferida pela análise para cada valor de K. A) Admixture para K=2 populações; B) Admixture para K= 3 populações; C) Admixture para K= 4 populações; D) Admixture para K= 5 populaçõe; E) Valores de probabilidade posterior para cada valor de K estimado e sumarizado após 20 réplicas F) Valores de delta K para cada valor de K estimado e sumarizado após 20 réplicas.

A análise de FST global realizada a partir dos SAFs de cada amostra/população teve um valor total de 0.146 e um valor corrigido de 0.105, valores estes que indicam a existência de estrutura populacional entre as localidades amostradas. Já os valores de FST através do método *sliding window* com janela de leitura de 10 kb e intervalo de 5 kb evidencia a existência de posições com altos valores de FST, onde valores de FST ≥ 0.427 se encontram no percentil de 99% da distribuição (Figura 5).

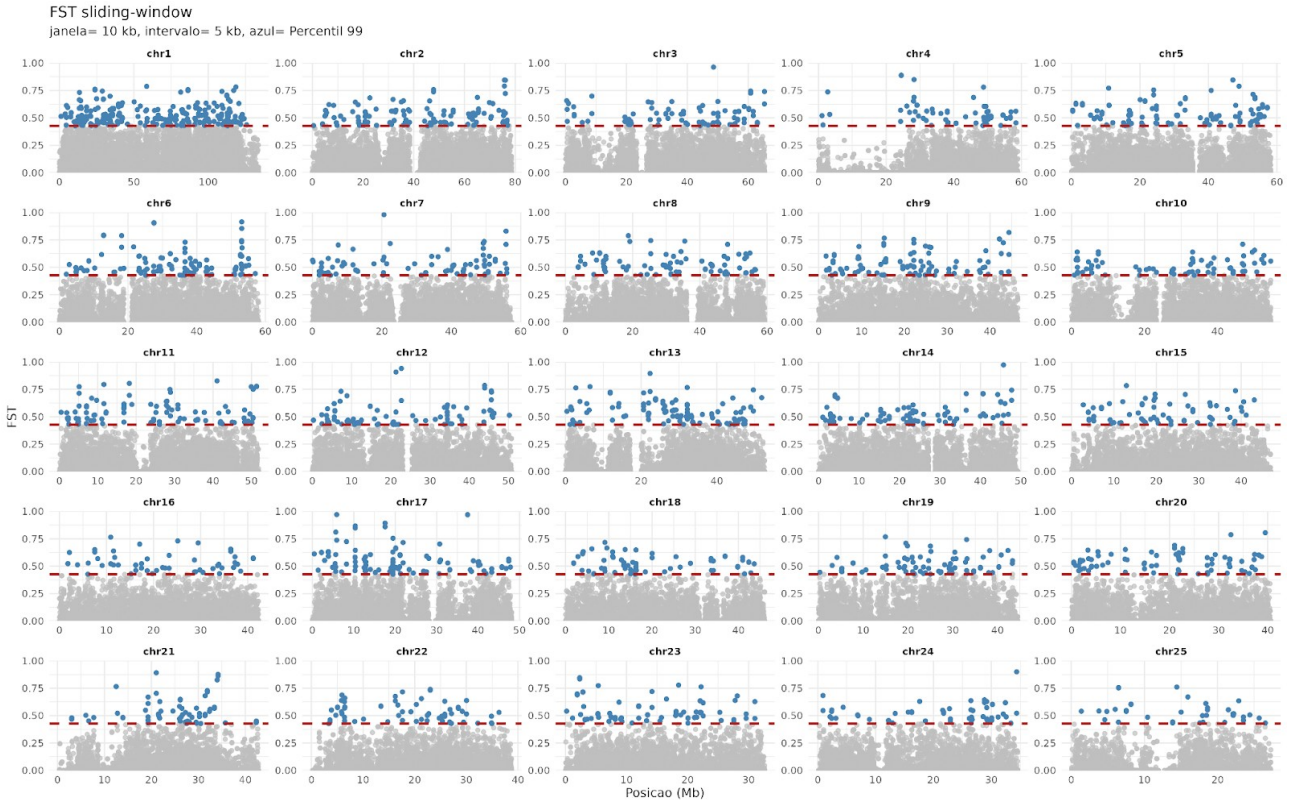

Figura 5. Análise de FST através do método *sliding window* com janela de leitura de 10 kb e intervalo de 5 kb a partir das probabilidades de frequências alélicas estimadas pelo programa realSFS e mapeadas aos 25 cromossomos do tetra mexicano *Astyanax mexicanus*. Em azul, loci cujos valores de Fst se encontram no percentil de 99% da distribuição de Fst (Fst ≥ 0.427).

Quando sumarizamos os índices de FST para as leituras mapeadas a cada um dos 25 cromossomos de *A. mexicanus*, os valores de FST considerando todas as janelas de leitura (213.558) apresentaram valores entre 0,04 a 0,05. Já para os valores de FST que se encontram no percentil de 99% da distribuição, seus valores variam entre 0,51 a 0,56, valores estes cerca de 10 vezes maiores que a média por cromossomo quando considerando todas as posições. A proporção de janelas de leitura com valores de FST acima do percentil de 99% apresentou valores entre 0,66 a 1,39%, totalizando 2.154 janelas (Tabela 3).

Tabela 3. Índice de Fst por cromossomo através do método *sliding window* com janela de leitura de 10 kb e intervalo de 5kb. N= número de janelas de leitura; Fst total= média de Fst por cromossomo; N99%= Número de janelas de leitura no percentil de 99% da distribuição dos valores de Fst; Fst 99%= média de Fst para as janelas de leitura no percentil 99%; N99/N= proporção de janelas de leitura no percentil de 99% da distribuição dos valores de Fst.

Cromossomo	N	Fst total	N99%	Fst 99%	N99/N (%)
chr1	21840	0,051	247	0,526	1,13
chr2	13135	0,047	120	0,524	0,91
chr3	10131	0,046	84	0,531	0,82
chr4	6386	0,043	62	0,546	0,97
chr5	9865	0,046	110	0,538	1,11
chr6	9681	0,053	109	0,538	1,12
chr7	9830	0,047	80	0,53	0,81
chr8	8744	0,046	79	0,534	0,90
chr9	8340	0,049	92	0,532	1,10
chr10	8939	0,05	69	0,515	0,77
chr11	8745	0,052	99	0,548	1,13
chr12	8607	0,047	85	0,528	0,99
chr13	7812	0,047	109	0,545	1,39
chr14	8550	0,048	85	0,532	0,99
chr15	7920	0,044	65	0,541	0,82
chr16	7786	0,045	52	0,534	0,66
chr17	7778	0,05	104	0,56	1,33
chr18	7474	0,042	61	0,528	0,81
chr19	7151	0,046	74	0,52	1,03
chr20	7132	0,045	78	0,544	1,09
chr21	5573	0,045	60	0,548	1,07
chr22	6295	0,045	64	0,533	1,02
chr23	5758	0,049	72	0,547	1,25
chr24	6002	0,045	57	0,526	0,95
chr25	4084	0,045	37	0,536	0,90

A demografia histórica estimada através do método PSMC apresentou um mesmo padrão para todas as amostras analisadas, onde foi observado um declínio brusco do tamanho efetivo populacional da espécie no intervalo entre 110 a 100 mil anos atrás, após o último período interglacial, e continuou em declínio através de todo o último período glacial, sem haver uma recuperação significativa do tamanho efetivo populacional durante o Holoceno (Figura 6).

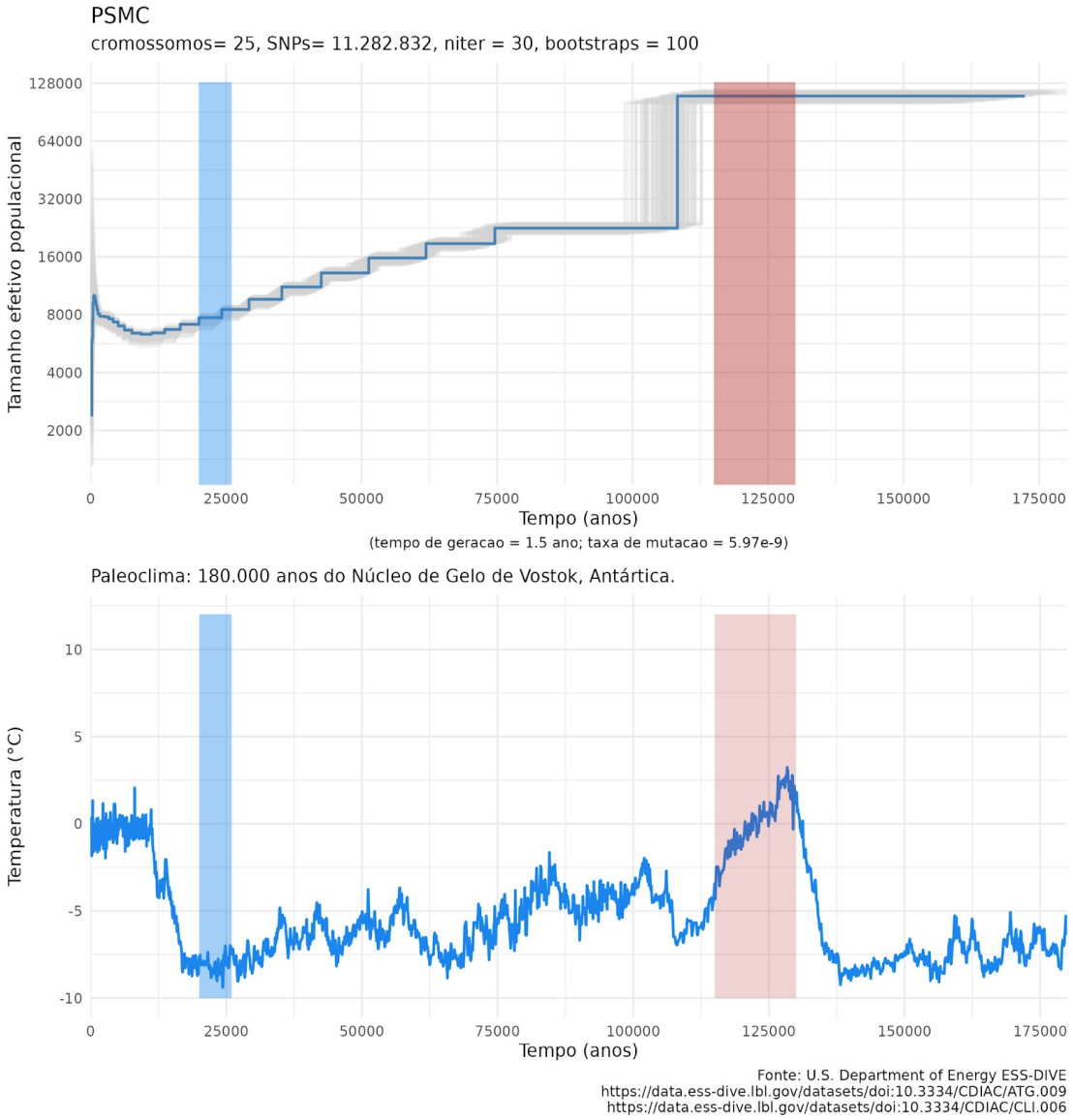

Figura 6. Análise de estimativa da demografia histórica de *P. axelrodi* pelo método PSMC. No gráfico superior se observa a variação e o drástico declínio no tamanho efetivo populacional (Ne) em razão do tempo (anos). No gráfico inferior se observa a variação na temperatura em razão do tempo (anos).

4. DISCUSSÃO

Durante a montagem de um genoma, duas abordagens podem vir a ser implementadas: a abordagem *De Novo* e a abordagem baseada em mapeamento. Apesar de a abordagem *De Novo* ser conceitualmente ideal, recriando a sequência original do genoma através de leituras contendo sobreposições, esta abordagem é altamente dependente de leituras que sejam longas, possuam alta cobertura e qualidade para que seja possível recuperar integralmente a informação contida no genoma, o que aumenta os custos do sequenciamento (Liao et al., 2019). A abordagem baseada em mapeamento mapeia as leituras obtidas a um genoma de referência de alta qualidade gerado anteriormente ou disponível em bancos de dados públicos como Ensembl ou GenBank, simplificando a montagem do genoma e exige relativamente menos recursos financeiros e computacionais, além de tornar possível o sequenciamento de genomas de espécies não-modelo e abordagens populacionais como o sequenciamento do genoma completo de baixa cobertura (lcWGS) (Lou et al., 2021; Prasad et al., 2022).

Neste trabalho, mapeamos bibliotecas genômicas do tetra cardinal *Paracheirodon axelrodi* ao genoma do tetra mexicano *Astyanax mexicanus*, único genoma de caracídeo disponível até o momento. Verificamos que ~73% de todas as leituras geradas para as cinco amostras sequenciadas (aproximadamente 17,3 Gb) foram mapeadas entre os 109 scaffolds (25 cromossomos + 84 scaffolds), cobrindo ~35% do genoma de referência. Com base no Genome Sequencing Coverage Calculator (https://support.illumina.com/downloads/sequencing_coverage_calculator.html), seriam necessários 50 Gb de dados para obter o genoma completo de uma amostra nas mesmas especificações do genoma do tetra mexicano (1,4 Gb, 35X). Ainda que nossas leituras representem apenas um terço do total desejado, estes dados nos permitiram vislumbrar aspectos da estrutura populacional e da demografia histórica da espécie.

A análise de PCA indica a separação entre as amostras provenientes de Barcelos e Santa Isabel do Rio Negro, onde na análise de admixture a partir da PCA indica um valor de K= 4 com a possibilidade de admixture entre amostras destas localidades. Entretanto, as análises usando NGSadmix com valores de K variando de 2 a 5 indicam K= 5 como melhor valor de K, onde cada amostra representaria um cluster populacional independente. Valores de FST global entre as localidades tiveram valores de 0.146 (global) e de 0.105 (corrigido). A estimativa de FST usando o método de *sliding-window* através dos 25 cromossomos mapeados resultou em valores de FST por cromossomo entre 0,04 e 0,05 quando considerando todas as janelas. Considerando os resultados da PCA e os valores de FST, indicando em ambos a existência de estrutura populacional entre as localidades amostradas, é possível que o baixo número amostral, o grande número de SNPs e loci

com altos valores de heterozigosidade possam de alguma forma afetar a performance dos métodos de admixture.

Sanchez-Bernal et al. (2023), investigando a filogeografia do tetra cardinal ao longo de sua distribuição geográfica nos rios Orinoco e Negro através de marcadores mitocondriais e nucleares, observaram estrutura populacional entre estas bacias de drenagem, com K= 2 como o melhor valor para o número de populações a partir dos dados microssatélites analisados, separando pontos de coleta localizados nas bacias do rio Orinoco e na bacia do rio Negro em dois clusters populacionais distintos, com uma zona de transição na região do alto rio Negro, em Cucuí e São Gabriel da Cachoeira. Especificamente para as localidades de Barcelos e Santa Isabel do Rio Negro, apenas com K= 6 de Sanchez-Bernal et al. (2023) que é possível visualizar estrutura populacional entre estas localidades. O valor de FST estimado a partir dos dados microssatélites foi de 0.051, similar aos valores observados a partir do FST por cromossomo, porém inferior aos valores de FST globais observados com os dados genômicos.

A demografia histórica da espécie a partir da análise de PSMC sugere uma queda brusca do tamanho efetivo populacional da espécie após o último período interglacial (~130 a 115 mil anos atrás), entre 110 a 100 mil anos atrás, com um declínio contínuo durante todo o último período glacial (~115 a 11,7 mil anos atrás), até o Holoceno. Li et al. (2021) observaram padrões similares de queda do tamanho efetivo populacional que poderiam estar associados ao último período glacial para 10 das 12 espécies de peixes teleósteos analisados a partir de genomas de alta qualidade. Na Amazônia, padrões similares de declínio também foram observados para aves do gênero *Willisornis* (Dalapicolla et al., 2024) e para primatas Neotropicais (Kuderna et al., 2023). Ainda que os genomas obtidos neste estudo sejam preliminares e parciais, estudos previamente publicados e simulações demonstram que o padrão demográfico geral pode ser recuperado mesmo com genomas de baixa cobertura, onde as maiores fontes de cautela seriam nos valores estimados dos tamanhos efetivos populacionais e menores intervalos de tempo estimados para o passado (Prasad et al., 2022; Dalapicolla et al., 2024). Ou seja, a geração de um genoma de referência de alta cobertura só melhoraria as estimativas já obtidas neste estudo.

Perspectivas futuras

Mudanças nas condições ambientais às quais os organismos se encontram adaptados desafiam seus limites fisiológicos de maneira direta e a velocidade com que as mudanças climáticas ocorrem aumentam o risco de extinção das espécies (Bernatchez et al., 2024). Bittencourt et al. (2017) correlacionaram mudanças nas frequências alélicas de uma única população de cardinal tetra ao longo dos anos de 2007-2010 a eventos de El Ñino e La Ñina que ocorreram durante o

monitoramento da espécie. Estudos experimentais também demonstram que *P. axelrodi* apresentaram menor tolerância térmica e menores taxas de sobrevivência quando aclimatados artificialmente em ambientes simulando cenários de mudanças climáticas (Campos et al., 2016; Gonçalves et al., 2018).

Eventos naturais, como secas ou cheias severas, ou atividades de origem antrópica, como a sobrepesca, podem impactar na diversidade genética das espécies, provocando modificações na estrutura populacional das espécies afetadas diretamente e causando rápidas mudanças na composição de uma comunidade. Os tetras do gênero *Paracheirodon* são espécies com ciclo de vida curto (12-16 meses) e dependentes do ciclo hidrológico anual para sobrevivência e manutenção das populações. Assim, as populações das espécies deste gênero podem ser fortemente influenciadas por eventos ambientais estocásticos que afetem sua diversidade genética. Diante disso, o sequenciamento do genoma completo do tetra cardinal e o monitoramento genético podem vir a detectar estas mudanças em curto prazo e servir de guia para esforços de conservação que, entre outros, avaliem o risco de extinção da espécie frente às mudanças climática e em que grau eventos climáticos regionais como El Niño e La Niña afetam as populações da espécie.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Bittencourt, P.S.; Marshall, B.; Hrbek, T.; Farias, I.P. 2017. Life after the drought: temporal genetic structure of Paracheirodon axelrodi Schultz, 1956 (Characiformes: Characidae) in the middle Negro River. *Pan-American Journal of Aquatic Sciences* 12(3): 184-193.

Beltrão, H.; Zuanon, J.; Ferreira, E. 2019. Checklist of the ichthyofauna of the Rio Negro basin in Brazilian Amazon. *ZooKeys* 881:53-89. Disponível em: https://doi.org/10.3897/zookeys.881.32055

Bergeron, L. A.; Besenbacher, S.; Zheng, J.; Bertelsen, M.F.; Quintard, B.; Hoffman, J.I.; et al. 2023. Evolution of the germline mutation rate across vertebrates. *Nature* 615: 285– 291. Disponível em: https://doi.org/10.1038/s41586-023-05752-y

Bernatchez, L.; Ferchaud, A.L.; Berger, CS.; Venney, C.J.; Xuereb, A. 2024. Genomics for monitoring and understanding species responses to global climate change. *Nature Reviews Genetics* 25: 165–183. Disponível em: https://doi.org/10.1038/s41576-023- 00657-y

Campos, D.F.; Jesus, T.F.; Heinrichs-Caldas, W.; Coelho, M.M.; Almeida-Val, V.M.F. 2016. Metabolic rate and thermal tolerance in two congeneric Amazon fishes: Paracheirodon

axelrodi Schultz, 1956 and Paracheirodon simulans Géry, 1963 (Characidae). *Hydrobiologia* 789: 133-142. Disponível em: 10.1007/s10750-016-2649-2

Dagosta, F.C.P.; De Pinna, M. 2019. The Fishes of the Amazon: Distribution and Biogeographical Patterns, with a Comprehensive List of Species. *Bulletin of the American Museum of Natural History* 431: 1-163. Disponível em: https://doi.org/10.1206/0003-0090.431.1.1

Dalapicolla, J.; Weir, J.T.; Vilaça, S.T.; Quaresma, T.F.; Schneider, M.P.C.; Vasconcelos, A.T.R.; et al. 2024. Whole genomes show contrasting trends of population size changes and genomic diversity for an Amazonian endemic passerine over the late quaternary. *Ecology and Evolution* 14(4). Disponível em: https://doi.org/10.1002/ ece3.11250

Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J. Ohan, V.; Pollard, M.O.; et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* 10(2). Disponível em: https://doi.org/10.1093/gigascience/giab008

Evanno, G.; Regnaut, S.; Goudet, J. 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* 14: 2611-2620. Disponível em: https://doi.org/10.1111/j.1365-294X.2005.02553.x

Fox, E.A.; Wright, A.E.; Fumagalli, M.; Vieira, F.G. 2019. ngsLD: evaluating linkage disequilibrium using genotype likelihoods, *Bioinformatics* 35(19):3855–3856. Disponível em: https://doi.org/10.1093/bioinformatics/btz200

Gonçalves, L.M.F.; Silva, M.N.P.; Val, A.L.; Almeida-Val, V.M.F. 2018. Differential survivorship of congeneric ornamental fishes under forecasted climate changes are related to anaerobic potential. *Genetics and Molecular Biology* 41(1): 107-118. Disponível em: http://dx.doi.org/10.1590/1678-4685-GMB-2017-0016

Harris, P.M.; Petry,P. 2001. Preliminary report on the genetic population structure and phylogeography of cardinal tetra (Paracheirodon axelrodi) in the rio Negro Basin. p. 205- 225. em: Chao, N.L.; P. Petry; G. Prang; L. Sonneschien & M. Tlusty (Eds.). *Conservation and Management of Ornamental Fish Resources of the Rio Negro Basin, Amazonia, Brazil – Project Piaba*. Editora da Universidade do Amazonas, Manaus, Brazil, 303 p.

Korneliussen, T.S.; Albrechtsen, A; Nielsen, R. 2014. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* 15(356). Disponível em: https://doi.org/10.1186/s12859-014-0356-4

Kopelman, N.M.; Mayzel, J.; Jakobson, M.; Rosenberg, N.A.; Mayrose, I. 2015. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources* 15: 1179-1191. Disponível em: https://doi.org/10.1111/1755-0998.12387

Kuderna, L.F.K.; Gao, H.; Janiak, M.C.; Kuhlwilm, M.; Orkin, J.D.; Bataillon, T.; et al. 2023. A global catalog of whole-genome diversity from 233 primate species. *Science* 380(6648): 906-913. Disponível em: DOI:10.1126/science.abn7829

Li, J.; Bian, C.; Yi, Y.; Yu, H.; You, X.; Shi, Q. 2021. Temporal dynamics of teleost populations during the Pleistocene: a report from publicly available genome data. *BMC Genomics* 22(490). Disponível em: https://doi.org/10.1186/s12864-021-07816-7

Liao, X.; Li, M.; Zou, Y.; Wu, F.; Wang, J. 2019. Current challenges and solutions of de novo assembly. *Quantitative Biology* 7: 90-109. Disponível em: https://doi.org/10.1007/s40484-019-0166-9

Liu, S.; Hansen, M.M. 2017. PSMC (pairwise sequentially Markovian coalescent) analysis of RAD (restriction site associated DNA) sequencing data. *Molecular Ecology Resources* 17(4): 631-641. Disponível em: doi: 10.1111/1755-0998.12606.

Lou, R.N.; Jacobs, A.; Wilder, A.P.; Therkildsen, N.O. 2021. A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology* 30(23): 5966-5993. Disponível em: doi: 10.1111/mec.16077

Marshall, B.G; Forsberg, B.R.; Thomé-Souza, M.J.F. 2008. Autotrophic energy sources for Paracheirodon axelrodi (Osteichthyes, Characidae) in the middle Negro River, Central Amazon, Brazil. *Hydrobiologia* (569): 95-103 Disponível em: DOI 10.1007/s10750-007-9060-y

Marshall, B.G.; Forsberg, B.R.; Hess, L.L.; Freitas, C. 2011. Water temperature differences in interfluvial palm swamp habitats of Paracheirodon axelrodi and P. simulans (Osteichthyes: Characidae) in the middle Rio Negro, Brazil. *Ichthyol Explor Freshwaters* 22(4): 377 - 383.

Mas-Sandoval, A.; Pope, N.S.; Nielsen, K.N.; Altinkaya, I.; Fumagalli, M.; Korneliussen, T.S. 2022. Fast and accurate estimation of multidimensional site frequency spectra from low- coverage high-throughput sequencing data. *Gigascience* 11. Disponível em: doi: 10.1093/gigascience/giac032.

Mckenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A; et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Resources* 20:1297-303. Disponível em: DOI: 10.1101/gr.107524.110. Disponível em: DOI: 10.1101/gr.107524.110.

Meisner, J.; Albrechtsen, A. 2018. Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. *Genetics* 210(2): 719-731. Disponível em: doi: 10.1534/genetics.118.301336.

Myers, G.S.; Weitzman, S.H. 1956. Two new Brazilian fresh water fishes. *Stanford Ichthyological Bulletin*, 7(1): 1-4.

Pedersen, T. (2024). *patchwork: The Composer of Plots. R package version 1.2.0.9000*, https://github.com/thomasp85/patchwork, https://patchwork.data-imaginist.com.

Prasad, A.; Lorenzen, E.D.; WestburyY, M.V. 2022. Evaluating the role of reference-genome phylogenetic distance on evolutionary inference. *Molecular Ecology Resources* 22: 45-55. Disponível em: https://doi.org/10.1111/1755-0998.13457

R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Rodrigues, P.K.S. *Estrutura populacional do tetra cardinal Paracheirodon axelrodi, Schultz 1956* (*Characiformes; Characidae) no médio rio Negro, Amazonas – Brasil.* 2017. Dissertação (Mestrado em Biologia de Água Doce e Pesca Interior) - Instituto Nacional de Pesquisas da Amazônia, Manaus, 2017.

Richardson, N. et al. Apache Arrow (2024). _arrow: Integration to 'Apache' 'Arrow'_. R package version 15.0.1, <https://CRAN.R-project.org/package=arrow>.

Sanchez-Bernal, D.; Martinez, J.G.; Farias, I.P.; Hrbek, T.; Caballero, S. 2023. Phylogeography and population genetic structure of the cardinal tetra (Paracheirodon axelrodi) in the Orinoco basin and Negro River (Amazon basin): evaluating connectivity and historical patterns of diversification. *PeerJ* 11:e15117. Disponível em: https://doi.org/10.7717/peerj.15117

Schultz, L.P. 1956. The amazing new fish called the scarlet characin. *Tropical Fish Hobbyist* 4 (4):41-43.

Skotte, L.; Korneliussen, T. S; Albrechtsen, A. 2013. Estimating individual admixture proportions from next generation sequencing data. *Genetics* 195(3): 693-702. Disponível em: doi: 10.1534/genetics.113.154138.

Toledo-Piza, M.; Baena, E.G.; Dagosta, F.C.P.; Menezes, N.A.; Ândrade, M.; Benine, R.C; et al. 2024. Checklist of the species of the Order Characiformes (Teleostei: Ostariophysi). *Neotropical Ichthyology* 22(1): e230086. Disponível em: https://doi.org/10.1590/1982-0224-2023-0086

Tribuzy-Neto, I.A.; Beltrão, H.; Benzaken, Z.S.; Yamamoto, K.C. 2020. Analysis of the ornamental fish exports from the Amazon state, Brazil. *Boletim do Instituto de Pesca* 46(4). Disponível em: https://doi.org/10.20950/1678-2305.2020.46.4.554

Warren, W.C.; Boggs, T.E.; Borowsky, R.; Carlson, B.M.; Ferrufino, E.; Gross, J.B; et al. 2021. A chromosome-level genome of Astyanax mexicanus surface fish for comparing population-specific genetic differences contributing to trait evolution. *Nature Communications* 12(1447). Disponível em: https://doi.org/10.1038/s41467-021-21733-z

Weitzman, S.H.; Fink, W.L. 1983. Relationships of the neon tetras, a group of South American freshwater fishes (Teleostei, Characidae), with comments on the phylogeny of New World characiforms. *Bulletin of the Museum of Comparative Zoology at Harvard College* [Internet] 150: 339–95. Disponível em: https://www.biodiversitylibrary.org/part/28701

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L.D.; François, R; et al. 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4(43). Disponível em: https://doi.org/10.21105/joss.01686

MATERIAL SUPLEMENTAR

Os dados brutos do sequenciamento foram demultiplexados e foram processados através de um pipeline de bioinformática customizado escrito na linguagem Bash e executado através da interface de linha de comando (Terminal) do sistema operacional Linux Ubuntu 22.04 LTS. Este pipeline consiste em uma série de scripts e loops que executam várias tarefas de forma automatizada e sequencial, onde o resultado de uma operação serve de input para a operação seguinte até que se chegue ao final do pipeline.

Entre as operações executadas pelo pipeline estão: a ordenação das leituras R1 e R2 de maneira intercalada (‘interleaved’) através da função ‘mergepe’ do programa seqtk 1.3-r117-dirty (https://github.com/lh3/seqtk), seguido da remoção de adaptadores através do programa ‘cutadapt 4.4’ (Martin, 2011) e o mapeamento das leituras pareadas ao genoma de referência de *Astyanax mexicanus* (RefSeq GCF_023375975.1) foi realizado através do programa bwa-mem2 (Vasimuddin et al., 2019). As leituras mapeadas foram ordenadas, receberam adição de mate score tags e foram marcadas para duplicadas através das funções ‘sort’, ‘fixmate’ e ‘markdup’ do software SAMtools 1.51.1 (Danecek et al., 2021), sendo em seguidas convertidas para o formato .bed através da função ‘bamtobed’ do programa BEDtools (Quinlan & Hall, 2010). Os arquivos resultantes tiveram seus read groups modificados através da função ‘AddOrReplaceReadGroups’ do software Picard (Picard Toolkit, 2019), indexadas com SAMtools ‘index’ e a anotação de SNPs e indels foi feita através da função ‘HaplotypeCaller’ e a genotipagem através da função ‘GenotypeGVCFs’ do software GATK 4.2.5.0 (McKenna et al, 2010). Os arquivos .vcf foram transformados em .bcf através da função ‘view’ e filtrados através da função ‘filter’ do software BCFtools 1.51.1 (Danecek et al., 2021), onde foram usados os argumentos (QD=2, MQRank=-12.5, FS=60, SOR=3, ReadPos_SNP=-8, ReadPos_ind=-20, snpgap=10, indelgap=5, pvalue= 0.000001, MQ=20, QUAL=20, MAF=0.05, mindepth= 0.8, maxdepth=300, missingdataSITE=0.9, MIN_HET_AD=0.25, MIN_COV=2, MAX_COV=100). O arquivo .bcf filtrado foi então convertido para .vcf através da função ‘bcftools view’ e este foi utilizado nas análises subsequentes.

REFERÊNCIAS BIBLIOGRÁFICAS

Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J. Ohan, V.; Pollard, M.O.; *et al*. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* 10(2). Disponível em: https://doi.org/10.1093/gigascience/giab008

Mckenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A; *et al*. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Resources* 20:1297-303. Disponível em: DOI: 10.1101/gr.107524.110.

“Picard Toolkit.” 2019. *Broad Institute*, GitHub Repository. Disponível em: https://broadinstitute.github.io/picard/

Vasimuddin, M.; Misra, S.; Li, H.; Aluru, S. 2019. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. *IEEE Parallel and Distributed Processing Symposium (IPDPS)* 314-324. Disponível em: https://doi.org/10.1109/IPDPS.2019.00041

Quinlan, A.R.; Hall, I.M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics* 26(6): 841–842. Disponível em: https://doi.org/10.1093/bioinformatics/btq033

CONCLUSÃO GERAL

Plataformas de NGS nos permitiram ter acesso a um volume de dados genômicos de espécies modelo e não modelo sem precedentes na história da biologia molecular, abrindo caminho para novas possibilidades de estudo, tanto na ciência básica quanto aplicada. No entanto, ainda existem grandes desafios inerentes a montagem e anotação de genomas, sejam estes devido a natureza do genoma analisado ou a capacidades computacionais e de bioinformática.

Neste estudo, através de duas estratégias diferentes de montagem de genomas, foi possível mapear leituras geradas na plataforma Illumina para o tetra cardinal *Paracheirodon axelrodi* – um importante e emblemático peixe no aquarismo mundial e o peixe mais exportado do Amazonas para fins de aquarismo. Como resultados, obtivemos cinco mitogenomas circulares através da montagem *De Novo* e geramos anotações mais precisas do que as atuais que se encontram disponíveis no GenBank/RefSeq através de pipelines de bioinformática, e geramos um rascunho do genoma nuclear de cinco amostras após o mapeamento contra o genoma de referência do tetra mexicano *Astyanax mexicanus*, onde cerca de 75% das leituras foram mapeadas e cobrem cerca de 35% da extensão total do genoma do tetra mexicano.

As análises de estrutura populacional a partir dos SNPs indicam haver estrutura entre as localidades amostradas – Santa Isabel do Rio Negro e Barcelos – onde valores de FST são da ordem de 10% e valores de K estiveram entre 4 a 5. A demografia histórica da espécie também foi estimada, indicando um brusco declínio do tamanho efetivo populacional após o último máximo glacial em seguida de um constante declínio ao longo do último período glacial, sem haver uma recuperação evidente durante o Holoceno. Ainda que os genomas nucleares obtidos neste estudo sejam preliminares e parciais, eles nos permitiram acessar importantes aspectos das populações amostradas e forneceram dados preliminares que podem orientar estratégias de avaliação e manejo das populações naturais e servirão como um *timestamp* do momento atual em que a espécie se encontra e para futuras comparações nos próximos anos e décadas, principalmente devido ao futuro incerto da espécie, que pode ser sensivelmente afetada pelas mudanças climáticas regionais e globais que virão no futuro próximo.