UNIVERSIDADE FEDERAL DO PARÁ
NÚCLEO DE ECOLOGIA AQUÁTICA E PESCA DA AMAZÔNIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ECOLOGIA AQUÁTICA E PESCA

LUIZ FILIPE BRITO DE OLIVEIRA

SEQUENCIAMENTO E AVALIAÇÃO DA MONTAGEM DE GENOMAS MITOCONDRIAIS VIA *LONG READS* PARA TRÊS ESPÉCIES DE PEIXES DE ÁGUA DOCE DA AMAZÔNIA

Belém - PA
2022

LUIZ FILIPE BRITO DE OLIVEIRA

SEQUENCIAMENTO E AVALIAÇÃO DA MONTAGEM DE GENOMAS MITOCONDRIAIS VIA *LONG READS* PARA TRÊS ESPÉCIES DE PEIXES DE ÁGUA DOCE DA AMAZÔNIA

Dissertação apresentada ao Programa de Pós-Graduação em Ecologia Aquática e pesca da Universidade Federal do Pará, como requisito parcial para obtenção do título de Mestre em Ecologia Aquática e Pesca.

Orientador: Prof. Dr. Jonathan Stuart Ready

Belém – Pa
2022

LUIZ FILIPE BRITO DE OLIVEIRA

SEQUENCIAMENTO E AVALIAÇÃO DA MONTAGEM DE GENOMAS MITOCONDRIAIS VIA *LONG READS* PARA TRÊS ESPÉCIES DE PEIXES DE ÁGUA DOCE DA AMAZÔNIA

Dissertação apresentada ao Programa de Pós-Graduação em Ecologia Aquática e pesca da Universidade Federal doPará, como requisito parcial para obtenção do título de Mestre em Ecologia Aquática e Pesca.

Data: 30/08/2022

Orientador:

_____
Prof. Dr. Jonathan Stuart Ready
(UFPA – Instituto de Ciências Biológicas)

Banca examinadora:

_____
Prof. Dr. Marcelo Nazareno Vallinoto De Souza
(UFPA – Instituto de Estudos Costeiros)

_____
Prof. Dr. Leonardo dos Santos Sena
(UFPA – Instituto de Ciências Biológicas)

_____
Prof. Dr. Santelmo Selmo de Vasconcelos Júnior
(ITV-DS – Instituto Tecnológico Vale)

Suplentes:

_____
Prof. Dr. Tibério Cesar Tortola Burlamaqui
(CEABIO – Centro de Estudos Avançados da Biodiversidade)

_____
Prof. Dr. João Bráullio de Luna Sales
(UFPA – Instituto de Ciências Biológicas)

AGRADECIMENTOS

A Universidade Federal do Pará e a Pós-Graduação em Ecologia Aquática e Pesca pelo financiamento de minha pesquisa ao longo do mestrado.

Aos meus pais Luiz e Valdilene por todo amor e carinho concedido a mim, pelo esforço sem medidas que me possibilitaram ter ótimas oportunidades de estudar e aprender, e principalmente por terem me permitido ser livre nas escolhas e objetivos que decidi seguir. A minha família por todo seu apoio e carinho, e que sempre torceu pelo meu desempenho.

Aos meus avôs e avós, ainda que de origens muito humildes e sem quase nenhuma base escolar, me ensinam lições valiosas com sua experiência. A todos os meus tios e tias, em especial aos meus primos e primas, por todo companheirismo e estarem presentes nos momentos de maior alegria durante a minha infância.

À minha namorada Ariel, por permitir compartilhar parte da sua vida comigo, onde sempre foi preenchido por amor e afeto do mais genuíno possível, sendo meu porto seguro para todos os momentos que mais necessitei de alguém. Além da minha sogra, Leide, que sempre me trouxe ótima companhia e acolhimento em todas as vezes que estive na sua casa.

Aos meus melhores amigos que fiz na biologia, Kevin, Pedro, Sissa, Gio, Talita, que sem dúvida alguma foram imprescindíveis para que eu me encontrasse aqui hoje, não somente pelos inúmeros momentos de felicidade, mas também pelas valiosas conversas de quando precisei que alguém me escutasse.

Aos meus melhores amigos de graduação Lucas, Gabriel, Túlio, Anny, Emerson, Jéssica e Luyann, embora não tenhamos mais o contato de antes, afirmo com toda a certeza que as presenças de vocês naquele tempo ainda se mantiveram essenciais até aqui, sou grato por todas as maravilhosas experiências e bons momentos que tivemos e que espero ter no futuro.

Ao Prof. Dr. Jonathan Stuart Ready que me proporcionou a oportunidade de participar do seu grupo de pesquisa desde a primeira vez que pus os pés na UFPA. A todos do GIBI, em especial aos meus companheiros de laboratório Fabrício, Alan, Cíntia, Karol, Suellen, Bianca, Igor, Rasna, Orlando, Ilzane, Aisha, Elena, Naelma, Yany, Silvia, Yasmim, Márcia, Fernanda, as Jéssicas, ao Tibério que esteve mais tempo ao meu lado durante o desenvolvimento desse trabalho, e a copa do CEABIO, por proporcionar o lugar de melhor refúgio durante o tempo de trabalho no laboratório, com café e conversas das mais banais às mais complexas, e principalmente fofoca que sempre adoramos.

A todos, os meus mais genuínos agradecimentos.

RESUMO

A bacia amazônica contém a fauna de peixes de água doce mais diversa do planeta. A utilização de marcadores genéticos mitocondriais incluindo o COI tem possibilitado que identificações baseadas em ferramentas moleculares atuem na discriminação da riqueza desse grupo. O monitoramento desta diversidade requer novas ferramentas como o metabarcoding, porém, dificuldades com bancos de dados referenciais incompletos ainda tornam sua aplicação um desafio. A implantação de diversas tecnologias de sequenciamento permite a obtenção de todos os possíveis marcadores mais abundantes, por exemplo, por meio da montagem de genomas mitocondriais completos utilizando abordagens de bom custo-benefício como o *genome skimming*. Atualmente, isto vem sendo realizado principalmente com técnicas tradicionais de sequenciamento ou usando *short reads*, reduzindo o potencial reconhecimento de possíveis regiões problemáticas dentro do genoma. Foi testada a produção de mitogenomas de três espécies usando sequenciamento via *long reads*, para verificar o potencial ganho de informação em comparação com sequências referências já existentes em diferentes condições de ordem taxonômica. Foi possível recuperar o mitogenoma completo de *Gymnorhamphichthys rondoni*, e quase completo para *Carnegiella strigata* e *Potamorrhaphis guianensis*. Além disso, a montagem obtida possibilitou confirmar que a região controle de *C. strigata* possui tamanho maior em relação à referência, contendo sequências com diferentes padrões de repetições *in tandem*. Os resultados mostraram a conservação geral da estrutura e organização de mitogenomas para estas espécies. Em conclusão, as *long reads* foram capazes de auxiliar na obtenção de mitogenomas completos, bem como elucidar regiões ainda problemáticas e pouco exploradas. Atualmente, a combinação de métodos de sequenciamento por *short* e *long reads* para a realização de montagens híbridas se faz necessária para gerar sequências com maior confiabilidade, visto que este último apresenta taxas de erro levemente maiores do que o primeiro. O contínuo avanço de tecnologias para sequenciamento e redução de erros de leituras longas deve torná-lo a ser um dos principais métodos a serem adotados no futuro.

Palavras-chave: Peixes, mitogenoma, *long reads*, região controle, repetições.

ABSTRACT

The Amazon basin contains the most diverse freshwater fish fauna in the planet. The use of mitochondrial genetic markers including the COI made it possible for identifications based on molecular tools to act in the discrimination of the richness for this group. Monitoring this diversity requires new tools such as metabarcoding, however, difficulties with incomplete referential databases still a challenge. The implementation of several sequencing technologies allows obtaining all the possible most abundant markers, for example, by assembling complete mitochondrial genomes using cost-effective approaches such as genome skimming. Currently, this is being done mainly with traditional techniques or using short reads, reducing the potential recognition of possible problematic regions within the genome. The production of mitogenomes of three species was tested using sequencing through long reads to verify the potential gain of information in comparison with existing reference sequences under different conditions of taxonomic order. It was possible to recover the complete mitogenome for *Gymnorhamphichthys rondoni*, almost complete for *Carnegiella strigata* and *Potamorrhaphis guianensis*. In addition, the assembly obtained made it possible to confirm that the control region of *C. strigata* has a larger size compared to the reference, containing sequences with different patterns of *in tandem* repeats. The results showed the general conservation of mitogenome structure and organization for these species. In conclusion, the long reads were able to assist the obtation of the complete mitogenomes, as well as elucidate regions that are still problematic and little explored. Currently, the combination of sequencing methods by short and long reads to perform hybrid assemblies is very necessary to generate sequences with greater reliability, since the latter has slightly higher error rates than the former. The continuous advancement of technologies for sequencing and error reduction of long reads should be one of the main methods to be adopt in the future.

Keywords: Fish, mitogenome, long reads, control region, repeats.

SUMÁRIO

1 INTRODUÇÃO

1.1 APLICAÇÃO DE ESTUDOS MOLECULARES EM PEIXES NEOTROPICAIS.

A Amazônia é constituída por grandes bacias hidrográficas que drenam rios de diferentes ordens, os quais concentram a maior diversidade de peixes de água doce da região neotropical (Albert et al., 2020). Em especial, a drenagem da bacia amazônica acolhe aproximadamente 15% dessa fauna em ambientes de água doce no mundo (Tedesco et al., 2017). O registro de catalogação desse grupo já ultrapassa mais de 2,400 espécies no banco de dados AmazonFish (https://www.amazon-fish.com/). Esse inventário online compila a informação mais atualizada da distribuição de espécies de peixes de água doce, e os seus dados são oriundos de diferentes recursos, como literatura científica, coleções em museus, banco de dados online e registros de expedições de campo (Jézéquel et al., 2020).

Na região neotropical, essa grande diversidade está distribuída de forma bastante desigual entre os táxons de níveis mais altos, visto que os Siluriformes, Gymnotiformes, Characiformes, Cyprinodontiformes e Cichliformes, são as cinco ordens mais representativas (Malabarba et al., 2020). Perciformes foi considerada por muito tempo como a maior ordem de vertebrados existente no mundo, apresentando-se como uma "cesta de lixo taxonômica polifilética", que incluía diversas famílias sem sinapomorfias definidas (Betancur-R et al., 2017). Diante de estudos utilizando dados moleculares mais robustos, uma definição monofilética tem sido proposta para essa ordem, resultando em alterações filogenéticas como, a separação de membros anteriormente pertencentes aos Perciformes para serem incluídos agora em Cichliformes (Nelson et al., 2016).

O conhecimento acerca da diversidade da fauna de peixes de água doce neotropicais vem crescendo nos últimos anos (Reis et al., 2016). O desafio de documentar essa riqueza deve continuar para que seja possível a proteção dessa fauna diante das ações antropogênicas (Birindelli & Sidlauskas, 2018). A manutenção de populações estáveis e espécies que habitam regiões de cabeceira é particularmente difícil (Albert et al., 2020), uma vez que características como dispersão limitada e distribuições restritas podem resultar em problemas de sobrevivência no ambiente (Nogueira et al., 2010). Eventos de especiação alopátrica são comuns nestes táxons, e espécies ainda a serem descobertas podem não ser conhecidas em tempo hábil frente às ameaças antropogênicas, incluindo a fragmentação de habitats (Dias et al., 2013).

A riqueza de espécies (configurada como linhagens evolutivas independentes) é considerada importante para o gerenciamento de conservação da biodiversidade (Magurran

2013). A contribuição de abordagens baseadas em marcadores genéticos como o DNA *barcoding*, que requer a obtenção do espécime, tem acelerado a compreensão dessa medida de diversidade na região neotropical (Carvalho et al., 2011; Pereira et al., 2013; Guimarães et al., 2018). Além disso, abordagens atuais e menos invasivas como o *metabarcoding* têm sido cada vez mais utilizados em trabalhos focados em peixes (Milan et al., 2020; Jackman et al., 2021). No entanto, embora esta última seja caracterizada como uma alternativa promissora em comparação à mensuração tradicional de riqueza, sua aplicabilidade depende de banco de dados de referência da comunidade taxonômica esperada, a fim de que ocorra a realização da interpretação dos dados moleculares em espécies biológicas (Schenekar et al., 2020).

A realização de inventários da fauna na região neotropical baseada em amostragem tradicional reflete lacunas ainda a serem preenchidas (Frota et al., 2016). Nos últimos 15 anos, o aprimoramento de métodos que as abordagens já mencionadas utilizam, como o sequenciamento de DNA, possibilita cada vez mais a produção de dados de origem genética e em escala de informação cada vez maior (e.g. Mitogenomas) (Miya & Nishida, 2015). Portanto, potencializar a produção desses dados para a ictiofauna de água doce implicaria em um suporte para estudos filogenéticos de determinado táxon, ou na descrição da riqueza de comunidades de peixes da Amazônia.

1.2 TECNOLOGIAS DE SEQUENCIAMENTO DE ALTO RENDIMENTO (High Throughput Sequencing – HTS).

O desenvolvimento de tecnologias que possibilitaram desvendar o código genético dos seres vivos foi inicialmente concebido pelos métodos de Maxam e Gilbert (1977) e Sanger & Coulson (1975). Este último tornou-se o padrão utilizado de modo mais extensivo no sequenciamento de primeira geração. Além disso, a difusão dessa técnica de forma mais promissora ocorreu no campo clínico, principalmente em decorrência da criação do “Projeto Genoma Humano”, com a liberação do primeiro rascunho do genoma completo em 2004 (Levy & Myers, 2016).

Fatores como redução de custos por base, rapidez e aumento na geração de dados genômicos, representaram um aperfeiçoamento em ritmo veloz das técnicas de sequenciamento de DNA (Dijk et al., 2014). A maior quantidade de dados gerados é exponencial comparado com a técnica de Sanger; pois, ocorre pela capacidade de gerar milhões de sequências de nucleotídeos em paralelo em um tempo relativamente curto (Kulski, 2016). Esse aprimoramento fez com que as diversas tecnologias de sequenciamento de DNA, que

emergiram no mercado nas últimas duas décadas, fossem chamadas de tecnologias de sequenciamento de alto rendimento (High Throughput Sequencing - HTS) (Reuter et al., 2015).

A abordagem HTS é um termo mais geral para se referir principalmente às tecnologias modernas de sequenciamento de segunda (Next Generation Sequencing – NGS) e terceira gerações (Pradhan et al., 2019). Embora existam diversas empresas voltadas para o desenvolvimento de plataformas NGS, a Illumina detém mais de 70% de dominância nesse mercado, sendo atualmente o sistema de sequenciamento mais bem sucedido (Kulski, 2016). Ao avançar no campo de sequenciamento de terceira geração, a MinION é uma plataforma de sequenciamento desenvolvida pela Oxford Nanopore Technologies, que vem recebendo destaque nos últimos anos por apresentar portabilidade de uso e velocidade na geração de seus dados (Lu et al., 2016). As duas plataformas citadas se diferenciam bastante pela tecnologia e química implementada, mas também, pelo tipo de sequências produzidas.

1.3 SEQUENCIAMENTO DE SEGUNDA GERAÇÃO (NGS) – Illumina.

Em tecnologias de NGS é possível conceituar e classificar diferentes abordagens pelo princípio de sequenciamento de cada base, e a química envolvida (Ambardar et al., 2016). O método *sequencing-by-sinthesis* (SBS; sequenciamento por síntese) adotado pela Illumina, se baseia em repetidas rodadas de inserção de nucleotídeos com a dependência da DNA polimerase (Fan et al., 2006). Incluído ainda no SBS está a abordagem *cyclic reversible termination* (CRT), que usa terminadores de cadeia marcados com fluorescência removível (Metzker, 2010). Ainda que diferentes tecnologias sejam comercializadas, todas seguem um fluxo de trabalho de sequenciamento bastante similar (Shendure & Ji, 2008), que será abordado de forma mais específica a seguir com foco na Illumina.

A primeira etapa é chamada de preparação de biblioteca, e consiste na fragmentação aleatória do DNA genômico em diversos tamanhos e a sua ligação com adaptadores. Após isso, é realizado um processo de “PCR em ponte” que amplifica milhões de fragmentos em uma célula de fluxo, gerando agrupamentos separados com mesma sequência que irão produzir um forte sinal óptico para a próxima etapa. E por fim, o sequenciamento, que acontece por: 1) ciclos de adição de nucleotídeos modificados com extremidade 3’ – OH bloqueadas no DNA molde, 2) detecção do sinal fluorescente de cada base que será interpretada pelo sequenciador e 3) remoção química da fluorescência tornando a extremidade 3’ – OH livre para a continuação da polimerização da fita molde.

A saída de sequências brutas dos sequenciadores Illumina gera fragmentos de DNA de 40 – 300 pb, sendo classificadas como *short reads*. Vantagens como custo-benefício, baixas taxas de erro, grande quantidade de dados gerados e diversos protocolos de fragmentação do DNA, tornam esta abordagem de sequenciamento bastante versátil e atrativa (Kumar et al., 2016; Kchouk et al., 2017). Embora o uso de sequenciadores de NGS possibilite uma série de finalidades, existem algumas desvantagens inerentes à utilização de *short reads*, como problemas na identificação de regiões repetitivas e duplicadas, bem como variantes estruturais e o acesso de regiões ricas em conteúdo GC; isto torna processos como a montagem de genomas uma tarefa desafiadora (Pollard et al., 2018; Ho et al., 2020).

1.4 SEQUENCIAMENTO DE TERCEIRA GERAÇÃO (TGS) – MinION.

Os vieses de amplificação contidos na utilização de *short reads* podem resultar em montagens *de novo* super fragmentadas, que tendem a gerar problemas de resolução de sequências repetitivas de genomas grandes (Schatz et al., 2010). As tecnologias de sequenciamento de terceira geração (TGS), surgiram da necessidade de contornar as dificuldades enfrentadas pelo NGS através do incremento do tamanho das *reads* geradas (Pollard et al., 2018; Lu et al., 2016). O sequenciamento de *long reads*, como assim são chamadas as sequências produzidas por TGS, é capaz de gerar dados que excedem quilobases e abrangem regiões consideravelmente maiores de um genoma em um único fragmento gerado (Goodwin et al., 2016).

Em 2014 a Oxford Nanopore Technologies (ONT) foi responsável pelo lançamento da plataforma MinION, que se tornou um dos sequenciadores de TGS mais recentes e bem-sucedidos do mercado (Mikheyev & Tin, 2014; Deamer et al., 2016). Esse dispositivo apresenta características interessantes por ser um equipamento compacto, portátil e que funciona conectado a um computador via porta USB 3.0; isto possibilita vantagens como a exibição em tempo real dos dados gerados ainda que a corrida de sequenciamento não tenha sido finalizada (Kulski, 2016; Kchouk et al., 2017).

O método de sequenciamento desenvolvido pela ONT é baseado na identificação do efeito da corrente iônica durante a passagem de nucleotídeos através de nanoporos (Bayley, 2015). Embora não haja a necessidade de amplificação por PCR, uma etapa de preparação de biblioteca assim como no NGS é necessária. Inicialmente, dois tipos de adaptadores são utilizados, um oligonucleotídeo chamado de “líder”, e outro constituído por um oligonucleotídeo simples

chamado de “grampo” (Ip et al., 2015). Ambos os adaptadores são ligados em qualquer lado do DNA e possuem a função de guiá-lo para próximo da membrana onde os poros estão inseridos (Goodwin et al., 2016).

O sequenciamento inicia quando o “líder” alcança a proteína-motor presente no poro que descompacta o DNA, permitindo que a fita simples seja deslocada através do nanoporo até alcançar o “grampo”, o qual possibilita a passagem da fita complementar da mesma maneira (Lu et al., 2016). Durante essa passagem, mudanças na corrente iônica são observadas diante da ocupação dos nucleotídeos no poro. A magnitude e o tempo de duração desses eventos de mudança são detectados por sensores, e por meio de modelos gráficos gerados pelo sequenciador, são interpretados em “k-mers”, que representam um conjunto de 3–6 nucleotídeos usados para a determinação da sequência (Jain et al., 2016).

A plataforma MinION se configura como uma abordagem bastante atrativa, mas que possui limitações inerentes à sua tecnologia. Fatores como a velocidade não homogênea de deslocamento dos nucleotídeos no interior dos canais, nucleotídeos de estrutura similar e ausência de mudança de sinal na passagem de homopolímeros, podem se configurar como complicações químicas que causam uma baixa precisão na chamada de bases durante o sequenciamento (Rang et al., 2018). Em um estudo de sequenciamento utilizando a plataforma da ONT, foram encontrados valores considerados altos para taxas de erros em inserções, deleções e substituições de 4,9%, 7,8% e 5,1%, respectivamente (Jain et al., 2015). No entanto, a taxa de erro finalmente reduziu bastante no lançamento da versão de *flow cells* R10, permitindo uma confiabilidade maior nas sequências produzidas com uma cobertura menor de *reads* (Fig. 1). Dessa forma, uma a avaliação do uso da plataforma MinION para a obtenção de genomas completos se torna necessária diante dos vieses de sequenciamento da ONT em relação ao implementado pela Illumina.

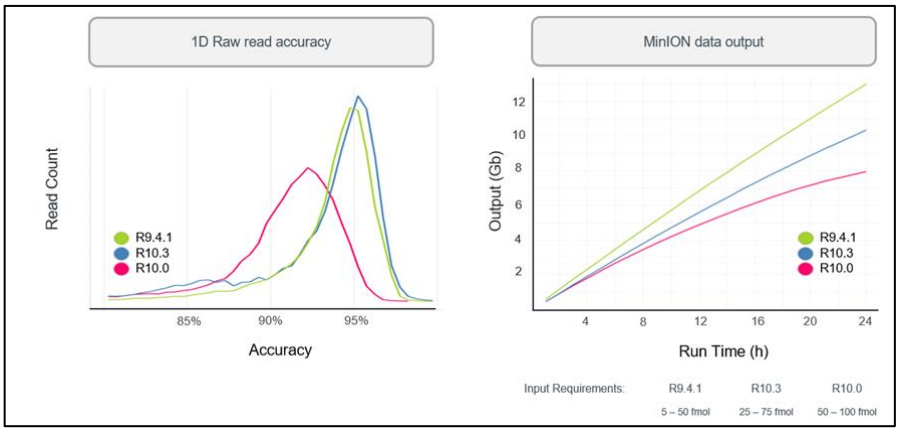


Figura 1: Qualidade de reads de versões de *flow cells* da Oxford Nanopore (Fonte: https://nanoporetech.com/about-us/news/r103-newest-nanopore-high-accuracy-nanopore-sequencing-now-available-store).

1.5 GENOME SKIMMING.

Devido ao padrão constante de redução de custos nas últimas décadas em relação às tecnologias de HTS, programas de pesquisa envolvendo DNA *barcoding* tradicional migraram gradativamente para trabalhos com escala maior, referindo-se à quantidade de espécimes e loci gênicos utilizados, representando uma abordagem mais poderosa (Coissac *et al*., 2016). Essa mudança de paradigma para a era genômica, influencia diretamente na aplicação de diferentes abordagens de sistemática molecular mais completas em virtude do aumento substancial de dados. A adição de outros loci informativos além do DNA *barcode* padrão do COI vêm garantindo resoluções filogenéticas mais robustas para um determinado grupo taxonômico, o que implica em um melhor entendimento de história evolutiva (Malé et al., 2014; Tan et al., 2022).

Um método que vem possibilitando a realização desse desenvolvimento é o chamado *genome skimming*, termo que foi primeiramente utilizado por Straub et al. (2012) em uma análise de táxons vegetais, propondo basicamente uma abordagem de sequenciamento de baixa cobertura para regiões com altas quantidades de cópias no genoma, como o DNA organelar de plastídios e mitocôndrias, além de regiões repetitivas do genoma nuclear. Em animais, essa abordagem está sendo essencialmente associada a abordagens de bioinformática (p. ex. montagem *de novo*), tornando possível reconstruir a sequência de um genoma mitocondrial completo para diferentes espécies de forma eficiente e rápida, e com um bom custo-benefício (Grandjean *et al*., 2017). Além disso, estudos recentes têm encontrado que esse método aparece

como uma ótima alternativa para recuperação sequências gênicas a partir de amostras de DNA degradado (Trevisan et al., 2019; Nevill et al., 2020).

Diante da performance de aplicação do *metabarcoding* para a discriminação das comunidades naturais, sua confiabilidade para aplicações de conservação da biodiversidade vem se tornando cada vez maior, considerando que as potenciais limitações associadas ao método têm sido superadas com o avançar do desenvolvimento das tecnologias de sequenciamento (Taberlet et al., 2012). Assim como o *metabarcoding* proporcionou uma perspectiva de aumento de dados em caráter taxonômico em estudos ambientais, a abordagem agora chamada de *metagenome skimming* ou *mitochondrial metagenomics*, tornou possível a execução dessa tarefa acrescentando uma maior quantidade de dados genéticos, não mais especializada em loci específicos para o DNA *barcode* tradicional, mas em escala de genomas inteiros (Papadopoulou et al., 2015).

Diversos trabalhos que investigam processos ecológicos e evolutivos em larga escala utilizando genomas mitocondriais completos, estão sendo realizados empregando a abordagem citada acima para diferentes tipos de amostras, como por exemplo, utilizando amostras compostas de espécimes, ou até mesmo amostras ambientais (Campton - Platt et al., 2016). Além disso, essa abordagem pode ser estendida para inferir melhor a resolução filogenética e estruturação de comunidades de grupos taxonômicos, como no caso de besouros (Coleóptera), que têm sido o principal alvo desses estudos em diferentes regiões do planeta (Andújar et al., 2015; Linard et al., 2015).

Os resultados obtidos por esses estudos mostram evidências de superação dos problemas enfrentados em trabalhos de *metabarcoding* tradicional utilizando amostras de DNA ambiental, por exemplo, mostrando uma riqueza de espécies subestimada por conta do uso de primers universais que acabam não amplificando determinada região em alguns táxons (Collins et al., 2019). Situações como essa criam a necessidade do desenho de primers específicos de acordo com a resolução taxonômica de determinados loci em bancos referenciais existentes (Elbrecht & Leese, 2015). Há também critérios de opções de loci por preferência de pesquisadores em estudos baseados na identificação molecular de espécies ou de *metabarcoding*, o que seria contornado uma vez que fosse obtido todos os loci oriundos de um genoma organelar completo (Coissac et al., 2016).

1.6 GENOMA MITOCONDRIAL.

Mitogenomas são representados por uma molécula circular simples, fechada e com conformação em dupla-fita; sendo encontrados em múltiplas cópias dentro de uma única célula na maioria dos vertebrados (Shadel & Clayton 1997). Comumente apresentam uma herança predominantemente unissexual (linha materna) mas com exceções conhecidas (Ladoukakis & Zouros, 2017). A estrutura e composição do DNA mitocondrial (mtDNA) entre a maioria dos vertebrados se mostra de modo bastante conservado, com um tamanho médio entre 16 a 17 kbp e composto por 37 genes, sendo 22 de tRNAs, dois de rRNAs e 13 genes codificadores de proteína (GCPs) (Boore, 1999).

A dupla fita de seu material genético é categorizada pela sua composição nucleotídica, contendo uma fita pesada (H – *heavy-strand*) ou leve (L - *light-strand*). Metazoários apresentam uma tendência de riqueza no conteúdo de bases nucleotídica adenina e guanina (A + G) na terminação de códons dos genes codificados pela fita H, havendo, por outro lado, uma maior proporção de citosina e timina (C + T) na fita L (Asakawa et al., 1991; Bernt et al., 2013). A fita H apresenta a maior parte da informação genética, geralmente contendo genes para os dois rRNAs, 14tRNAs e 12 regiões codificadoras de proteína, enquanto a fita L abriga oito tRNAs e um GCP (Taanman 1998; Broughton et al., 2001).

Entre os 13 genes codificadores de proteínas pertencentes aos complexos enzimáticos da fosforilação oxidativa, estão o citocromo b (CYTB), subunidades I-III da citocromo c oxidase (COI, COII e COIII), ATP sintase (ATP6 e ATP8) e sete genes que codificam subunidades da cadeia respiratória NADH dehidrogenase (NAD1, NAD2, NAD3, NAD4, NAD4L, NAD5 e NAD6) (Ingman, 2006). Cada GCP é separado por pelo menos por um tRNA (Ojala et al., 1981). Além disso, existe uma região controle (RC) não codificadora chamada de D-Loop, que contém sequências de sinais regulatórios responsáveis pela replicação e transcrição das fitas do mtDNA (Saccone et al., 1987).

Cópias inativas de regiões do mitogenoma presentes no genoma nuclear são classificadas como pseudogenes (NUMTs) (López et al., 1994), os quais têm sido amplamente documentados no DNA nuclear de eucariotos (Schiavo et al., 2017; Grau et al., 2020; Wang et al., 2020). Sua provável origem remonta da transferência de regiões da mitocôndria para o núcleo após o evento simbiótico, e se estabeleceu ao longo do processo evolutivo dos eucariotos (Timmis et al., 2004). A detecção desses pseudogenes interfere diretamente nas análises

trabalhos que visam a identificação de espécies usando o DNA mitocondrial (e.g. DNA *barcoding* ou *metabarcoding*), uma vez que a sua presença pode superestimar a riqueza real de espécies pela coamplificação de NUMTs (Song et al., 2008; Andújar et al., 2020).

Embora genomas nucleares de peixes e vertebrados, de modo geral, sejam conhecidos por não apresentar pseudogenes, estes foram reportados no mitogenoma de três espécies de peixes em um estudo realizado por Antunes et al. (2005). Características conservadas do mtDNA como a ordenação dos genes, apresentaram diferentes padrões que foram compartilhados polifiléticamente entre espécies de peixes distantemente relacionadas (Satoh et al., 2016). Além disso, pseudogenes podem ser representados em forma de rearranjos na estrutura do DNA mitocondrial, como no caso de duplicações em regiões de tRNA que foram encontradas retidas no genoma de *Chlorurus sordidus* (Mabuchi et al., 2004).

Os indícios das existências de NUMTs no genoma mitocondrial em geral estão associados a códons de parada, inserções ou duplicações (Song et al., 2008). No entanto, outros indicativos podem ser observados, como regiões altamente mapeadas que apresentem variações nucleotídicas durante um sequenciamento (Fig. 2) (Pereira et al., 2020). Dessa forma, a obtenção de genomas inteiros por si pode auxiliar na função de identificação com maior precisão regiões que podem conter NUMTs, ou proporcionar outras opções de utilização de marcadores para abordagens de identificação molecular de espécies. Além disso, vieses como regiões duplicadas oriundas do genoma nuclear ou originalmente pertencentes ao DNA mitocondrial, seriam detectados com maior precisão utilizando sequenciamento de *long reads* (Fig. 3) (Gan et al., 2019).

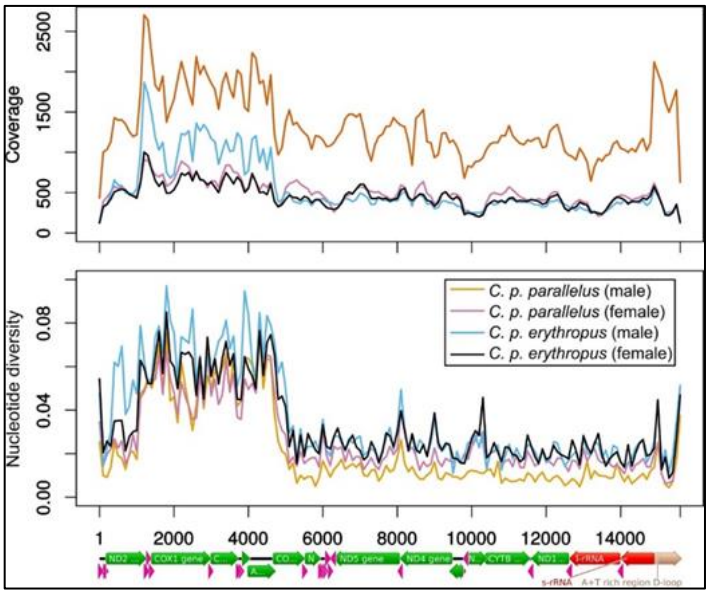


Figura 2: Gráfico apresentando uma maior cobertura de sequenciamento e variação nucleotídica no COI e outros genes que podem indicar a influência de *NUMTs* nesta região do genoma mitocondrial (Fonte: Pereira et al., 2020).

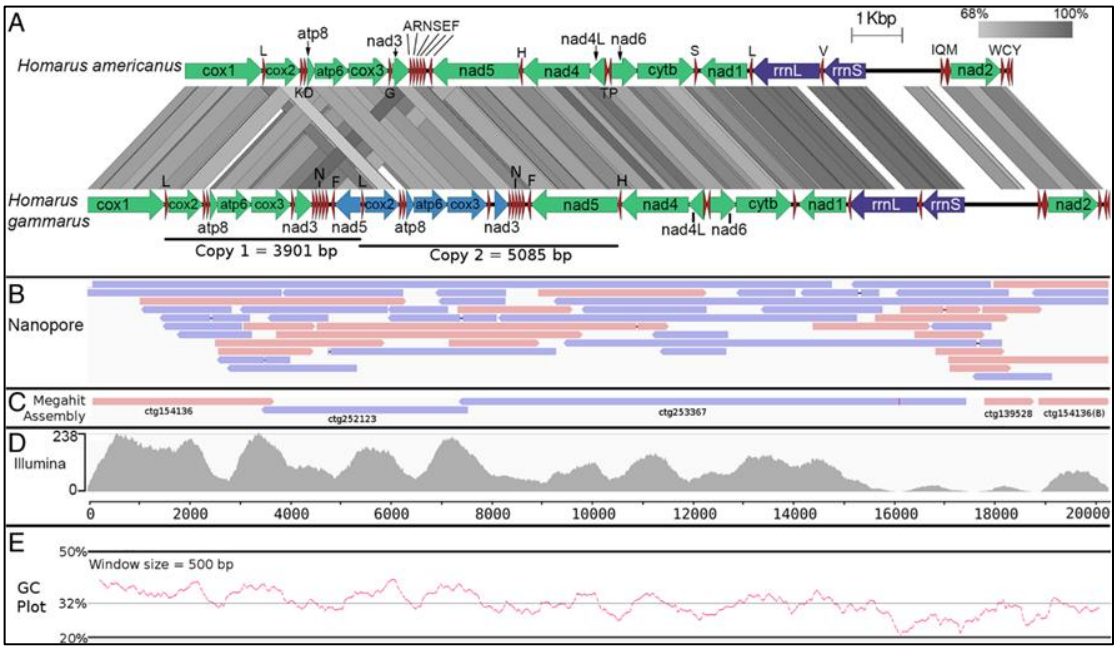


Figura 3: Contigs gerados pelo sequenciador MinION abrangendo regiões verdadeiramente duplicadas do genoma mitocondrial da lagosta *Homarus gammarus* indicando pseudogenização (Fonte: Gan et al., 2019).

O uso de ferramentas moleculares em estudos focando a discriminação da ictiofauna que habita os ambientes de água doce da Amazônia ainda é incipiente. Muitos trabalhos utilizando métodos tradicionais de amostragem e de identificação de espécies foram realizados na região, destacando a necessidade da caracterização dessa riqueza para uma melhor compreensão dos processos ecológicos e evolutivos que garantem a viabilização de medidas de

conservação de populações naturais. A popularização de diversas plataformas de tecnologias de sequenciamento configura uma excelente alternativa para impulsionar o conhecimento da biodiversidade. No entanto, a utilização ideal desses sistemas depende não somente da ampliação do banco de dados de referência, mas também na produção de sequências genômicas de qualidade que possam garantir a estruturação taxonômica mais próxima da realidade. Portanto, o objetivo deste estudo foi avaliar a qualidade de mitogenomas produzidos utilizando sequências *long reads*, unido a uma abordagem de sequenciamento eficiente e com um bom custo-benefício como o *genome skimming*, além de incrementar a produção de genomas mitocondriais de peixes amazônicos de água doce.

2 OBJETIVOS

2.1 OBJETIVO GERAL:

Gerar e avaliar genomas mitocondriais completos para três espécies de peixe da Amazônia utilizando sequências *long reads* seguindo *genome skimming* como abordagem de sequenciamento.

2.2 OBJETIVOS ESPECÍFICOS:

- Aferir a eficiência da montagem dos mitogenomas usando *long reads* para as três espécies de peixe;
- Investigar o valor de dados long *reads* para evidenciar a presença de potenciais *numts* existentes no genoma de peixes;
- Avaliar a confiabilidade e qualidade dos mitogenomas montados para a abordagem de sequenciamento utilizada.

3 MATERIAL E MÉTODOS

3.1 AMOSTRAGEM.

Foi utilizado uma amostra de tecido de um espécime de *Carnegiella strigata* (Fig. 4), *Gymnorhamphichthys rondoni* (Fig. 5) e *Potamorrhaphis guianensis* (Fig. 6) presentes na coleção de tecidos do grupo de pesquisa GIBI (Grupo de Investigação Biológica Integrada). Para a seleção das espécies, foi investigado suas potenciais referências no GenBank com base na distância taxonômica e utilização de métodos de sequenciamento diferente. Assim para *C. strigata*, foi escolhida uma sequência da própria espécie (AP011983.1), para *G. rondoni* uma referência do mesmo gênero foi optada (*Gymnorhamphichthys sp*. - AP011980.1), enquanto

para *P. guianensis*, o mitogenoma de *Ablennes hians* (AP006774.1) foi selecionado por ser membro da mesma família.

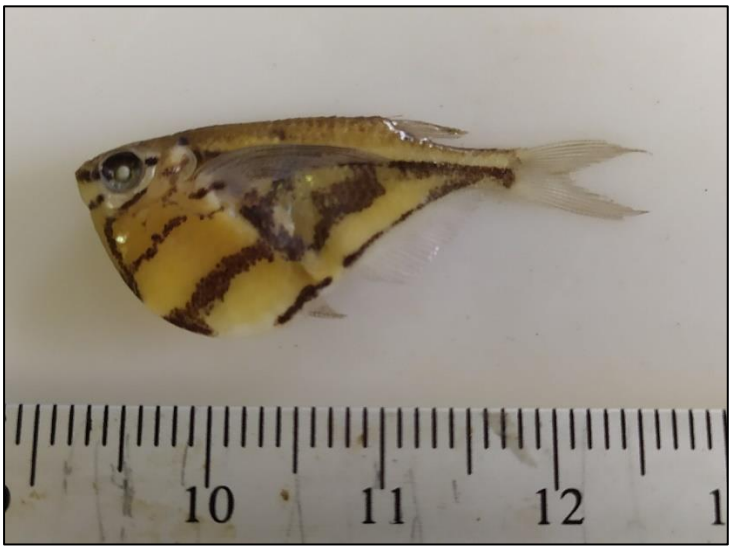


Figura 4: Espécime utilizado de *Carnegiella strigata*.

Figura 5: Espécime utilizado de Gymnorhamphichthys *rondoni*.

Figura 6: Espécime utilizado de *Potamorrhaphis guianensis*.

3.2 EXTRAÇÃO DE DNA E SEQUENCIAMENTO DE *LONG READS*.

O DNA genômico foi extraído das amostras de tecido utilizando o protocolo CTAB (Doyle & Doyle, 1987) a fim garantir altas quantidades de DNA genômico. O produto extraído de todas as amostras foi quantificado no Qubit® Fluorometer (Thermo Fisher) e diluído ou concentrado a fim de obter a concentração apropriada para seguir de acordo com as demandas de cada plataforma de sequenciamento. A preparação da biblioteca para o DNA extraído foi processada pelo SQK-LSK110 *Ligation sequencing* kit e após isso foram submetidos ao sequenciador MinION R10 *flow cell* seguindo as instruções do fabricante.

3.3 MONTAGEM DE GENOMA MITOCONDRIAL.

Para a realização da montagem do mitogenoma de cada espécie, as *long reads* brutas foram mapeadas contra a referência escolhida (Ver seção 3.1) usando o Minimap2 (Li, 2018). Posteriormente as *reads* potencialmente mitocondriais foram extraídas utilizando o SAMtools (Danecek et al. 2018), e assim, as sequências restantes foram utilizadas para a realização da montagem *de novo* utilizando o Flye (Kolmogorov et al. 2019). Os contigs gerados sucederam a uma etapa de polimento para correção de erros e *indels* por meio do software Medaka (Wick et al. 2019).

3.4 ANOTAÇÃO E ANÁLISE DO GENOMA MITOCONDRIAL.

Os contigs mitocondriais foram anotados usando a plataforma MITOS 2 Web Server (Bernt et al. 2013) e GeSeq (Tillich et al. 2017). Para os genes que apresentaram complicações ao realizar sua anotação, uma anotação manual foi conduzida por meio da busca por ORFs (Open Reading Frames) utilizando o programa Geneious Prime 2022.1.1. A visualização do genoma e contigs mitocondriais foi efetuada no OrganellarGenomeDRAW (Greiner et al. 2019).

A tradução das sequências, apuração e correção de códons de cada GCP foi aplicada por meio das ferramentas online Web Expasy (Gasteiger et al. 2013) e Codon Plot (Stothard, 200). A cobertura de cada contig montado foi calculado pelo Pysamstats (https://github.com/alimanfoo/pysamstats), e a busca por repetições *in tandem* foi explorada pelo Tandem Repeats Finder (Benson, 2000), com a predição da possível formação de estruturas secundárias em repetições sendo executada no Mfold (Zuker, 2003).

4 RESULTADOS E DISCUSSÃO

4.1 *Carnegiella strigata*.

4.1.1 SEQUENCIAMENTO, MONTAGEM E COBERTURA.

O sequenciamento bruto gerou um total de 139.398 *reads*. Para a seleção das *long reads* mitocondriais, foi escolhido o mitogenoma de referência da mesma espécie disponível no GenBank (AP011983.1). O mapeamento final resultou em 514 reads que foram utilizadas para a etapa de montagem do genoma mitocondrial de *C. strigata* (Tabela 1).

Tabela 1: Número de reads brutas, mapeadas e o código de acesso da referência utilizada para cada espécie.

Espécie	Reads brutas	Reads mapeadas	Referência GenBank
Carnegiella strigata	139.398	514	*Carnegiella strigata*/ AP011983.1
Gymnorhamphichthys rondoni	64.459	866	*Gymnorhamphichthys sp.*/ AP011980.1
Potamorrhaphis guianensis	3116007	574	*Ablennes hians*/ AP006774.1

O resultado da montagem gerou quatro contigs que não foram capazes de abranger o conteúdo completo do mitogenoma da espécie. O tamanho de cada contig foi calculado: contig 1 (3557 pb), contig 2 (7073 pb), contig 3 (2078 pb) e contig Região Controle (2486 pb). A cobertura do sequenciamento de cada contig pode ser observada na Figura 7.

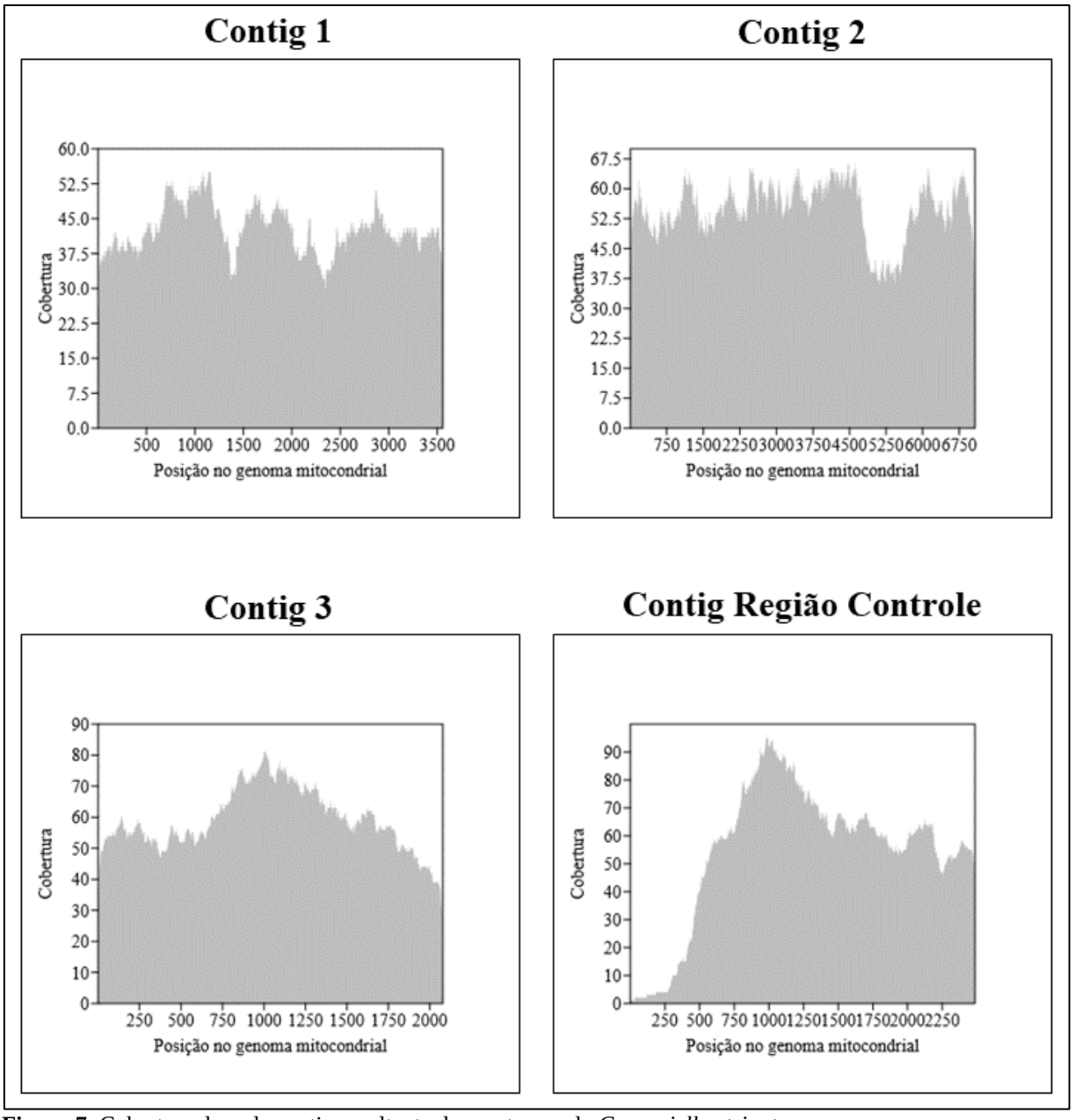


Figura 7: Cobertura de cada contig resultante da montagem de *Carnegiella strigata*.

4.1.2 ESTRUTURA E ORGANIZAÇÃO DO MITOGENOMA.

Por não ter sido possível a recuperação de um contig circular, uma anotação com configuração linear foi apresentada (Figura 8). Um total de 28 genes foram anotados além da região não codificadora configurada como a RC, dos quais 25 foram recuperados de forma completa e três de forma parcial (NAD1, COI e NAD5), com todos os GCPs sendo encontrados na fita pesada exceto NAD6. Apesar disso, todos os genes recuperados nas montagem seguiram o padrão geral de ordenação encontrado em Characiformes (Xu et al. 2021).

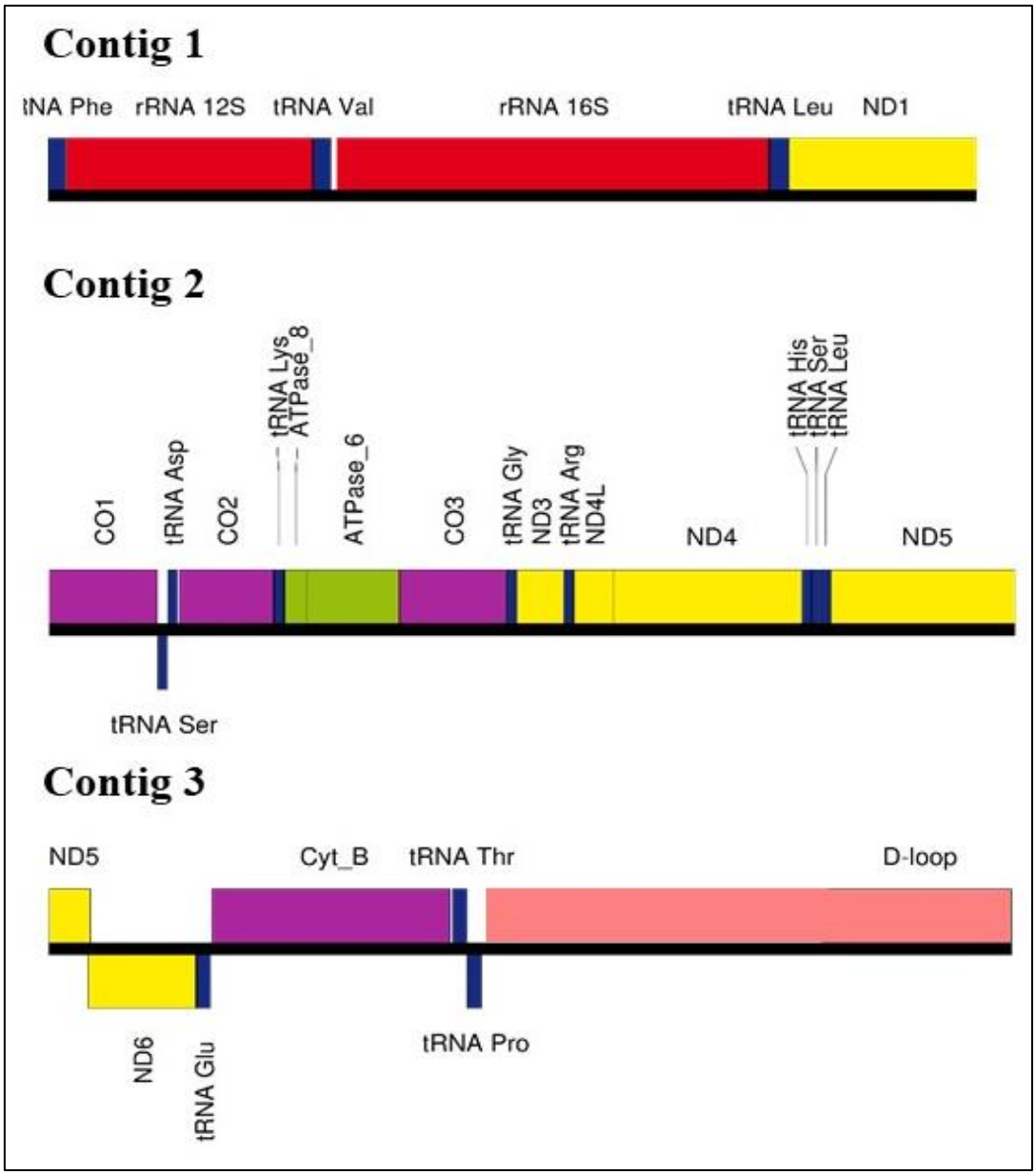


Figura 8: Anotação gênica para cada contig gerado da montagem de *Carnegiella strigata*, com o contig 3 unido ao contig da região controle para melhor visualização.

O códon de início ATG foi o único representante para todos os genes codificadores de proteína preditos (NAD1, COII, ATP8, ATP6, COIII, NAD3, NAD4L, NAD4, NAD5, NAD6 e CYTB), ocorrendo da mesma forma que a sequência referência de *C. strigata* para estes mesmos genes. Considerando a terminação gênica, diferentes códons ocorreram, sendo TAA o predominante (COI, NAD4L e ATP8). No entanto, a presença de códons incompletos também se fez bastante presente, sendo T para os genes COII, COIII e NAD3, bem como o códon TA para o CYTB, e o códon não convencional AGA para o gene NAD5.

Realizando a mesma comparação feita anteriormente, mas para os códons de início, somente os genes COII, ATP8, COIII, NAD4L, NAD4, NAD5 e CYTB, apresentaram o mesmo padrão

de códon de terminação que a sequência referência, exceto o gene COI, que apresentou TAA, diferentemente do mitogenoma de referência de *C. strigata* que exibiu o códon TAG. Essa informação é suportada pelo valor de cobertura estável (Cobertura do códon TAA = 53) para a posição desse códon, o que confere um grau de confiabilidade maior quanto a confirmação da troca da base A por G.

A configuração estrutural do contig produzido em comparação com a referência também sustenta essa modificação, em que a existência de um espaço intergênico de mesmo tamanho (1pb) entre o COI e o tRNA Ser1, também é encontrado na sequência de referência (Figura 9). Estudos focados em populações de *C. strigata* na Amazônia utilizando fragmentos de genes mitocondriais estão conferindo um caráter críptico para esta espécie, em que distâncias genéticas consideráveis utilizando GCPs foram encontradas. Logo, as mudanças observadas no uso de códons de terminação no gene COI, desempenhando a mesma função, podem estar reforçando tal situação (Schneider et al. 2018).

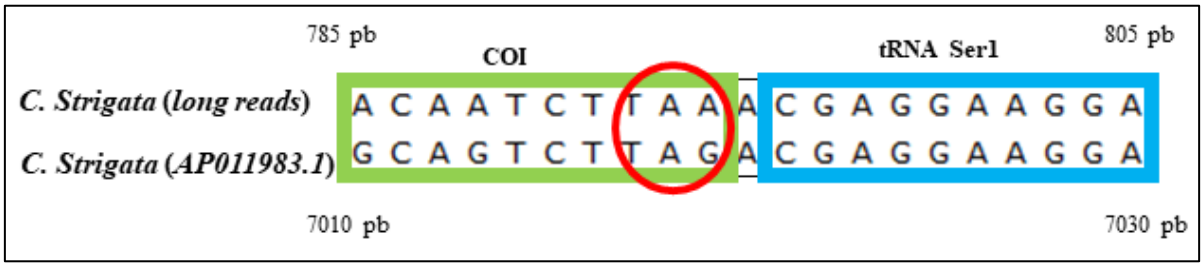


Figura 9: Alinhamento entre *C. strigata long reads* e referência destacando a mudança de base A por G no códon de parada do gene COI.

Não foi possível definir corretamente os *stop codons* para os genes ATP6, NAD4 e NAD6, uma vez que terminaram suas sequências com A, C, e ATA, respectivamente. Isto ocorreu por não ter sido possível identificar corretamente suas sequências como ORFs (Open Reading Frames) mesmo após a curadoria manual. Portanto, foi decidido manter a posição inicial e final desses genes no mitogenoma pelos métodos de predição gênica adotados neste trabalho.

A causa disso possivelmente está atrelada a *indels* por efeito da natureza do sequenciamento de leituras longas, as quais dificultam a validação dos códons verdadeiros no genoma, e assim impedindo a determinação mais precisa da abrangência de cada gene. Mas também, é possível que este problema esteja indicando traços de pseudogenes (*Numts*) que apresentam sequências similares às *reads* mitocondriais. É provável que o sequenciamento e mapeamento destas *numts* tenham influenciado diretamente o resultado da montagem, o que

pode explicar a presença de diversos stop codons presentes na sequência dos genes ATP6, NAD4 e NAD6 (Figura 10) (Flynn et al. 2015).

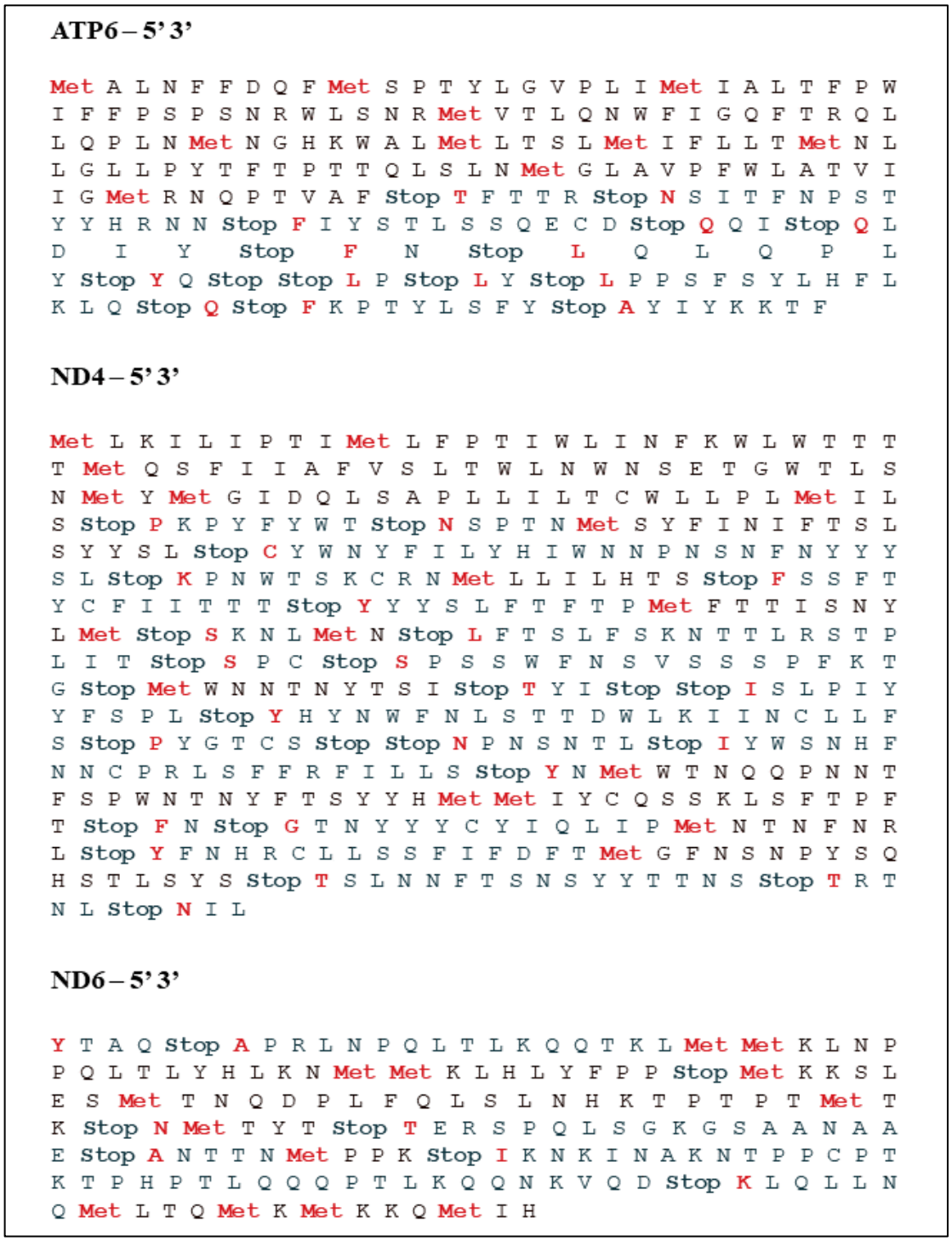


Figura 10: Tradução das sequências dos genes ATP6, NAD4 e NAD6, com destaque para a detecção de stop codons internos em *C. strigata*.

Foram identificados 14 tRNAs, que variaram de tamanho em 68 pb (tRNA Phe) a 75 pb (tRNA Leu1), estando 11 presentes na fita pesada e apenas três na fila leve. Os rRNAs foram

recuperados com tamanhos de 1013 pb e 2761 pb para o rRNA 12S e rRNA 16S, respectivamente, e com sua localização típica entre os tRNA Phe, Val e Leu1. Em relação aos espaços intergênicos e sobreposições, a primeira culminou em um total de 51 pb distribuídos oito vezes no genoma mitocondrial de *C. strigata*, sendo o maior com 20 pb localizado entre tRNA Val e o rRNA 16S. As sobreposições se apresentaram em menor número, aparecendo somente cinco vezes somando 35 pb: rRNA 12S e tRNA Val (1 pb), ATP8 e ATP6 (10 pb), NAD4L e NAD4 (7 pb), NAD5 e NAD6 (15 pb), tRNA Thr e tRNA Pro (2 pb). As características dos genes anotados e como estão organizados em cada contig estão descritas na Tabela 3.

Tabela 2: Características do mitogenoma de *Carnegiella strigata*, incluindo a localização, tamanho, espaços intergênicos e códons de cada gene encontrado por contig produzido, asteriscos identificam códons provavelmente incorretos.

Gene	Localização		Tamanho	Espaço Intergênico/ Sobreposição (pb)	Códon		Fita
	Início	Fim			Início	Parada	
Contig 1							
tRNA Phe	1	68	68				H
rRNA 12S	69	1013	945	0			H
tRNA Val	1013	1084	72	-1			H
rRNA 16S	1105	2761	1657	20			H
tRNA Leu1	2762	2836	75	0			H
NAD1	2837	3557	721	0	ATG	-	H
Contig 2							
COI	1	794	794	-	-	TAA	H
tRNA Ser1	796	866	71	1			L
tRNA Asp	871	941	71	4			H
COII	956	1646	691	14	ATG	T	H
tRNA Lys	1646	1721	75	0			H
ATP8	1723	1890	168	1	ATG	TAA	H
ATP6	1881	2565	685	-10	ATG	A*	H
COIII	2566	3349	784	0	ATG	T	H
tRNA Gly	3350	3421	72	0			H

NAD3	3422	3770	349	0	ATG	T	H
tRNA Arg	3771	3840	70	0			H
NAD4L	3841	4137	297	0	ATG	TAA	H
NAD4	4131	5511	1381	-7	ATG	C*	H
tRNA His	5513	5581	69	1			H
tRNA Ser2	5582	5649	68	0			H
tRNA Leu2	4651	5723	73	1			H
NAD5	5724	7073	1350	0	ATG	-	H
Contig 3							
NAD5	1	200	200	0	-	AGA	H
NAD6	186	698	513	-15	ATG	ATA*	L
tRNA Glu	699	767	69	0			L
CYTB	773	1902	5	0	ATG	TA	H
tRNA Thr	1912	1983	72	9			H
tRNA Pro	1982	2051	70	-2			L
Contig Região Controle							
D-loop	1	2486	2486				

4.1.3 IDENTIFICAÇÃO DE REPETIÇÕES *IN TANDEM* PRESENTES NA REGIÃO CONTROLE DE *C. strigata*.

A região controle foi identificada com um comprimento de 2486 pb, sendo um tamanho não apenas diferente, porém maior do que é encontrado na referência de *C. strigata* no GenBank. Na referência, é possível observar um trecho de 1304 pb retratado no GenBank como região desconhecida, e logo em seguida, o fragmento D-loop de 898 pb, totalizando 2202 pb de uma região não codificadora (Figura 11).

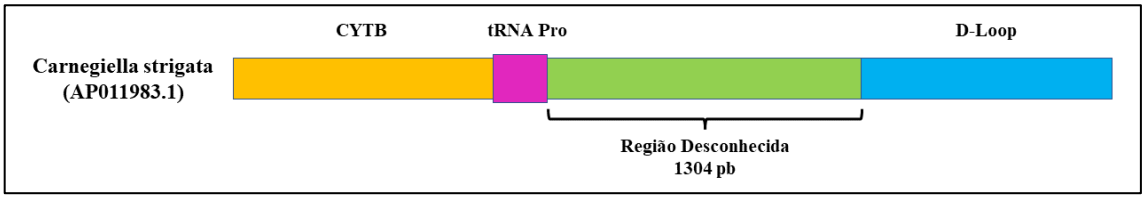


Figura 11: Representação da porção final do mitogenoma de *C. strigata* disponibilizada no GenBank.

Ao fazer o mapeamento da região controle produzida com *long reads* contra a referência (Figura 12), embora seja possível observar diferentes pontos de divergência e inserção nucleotídica, foi constatado que a região desconhecida e D-loop da referência coincidiu com a montagem produzida por *long reads*, além de um trecho de 348 pb não mapeado, o que pode indicar uma região não codificadora ainda maior do que os dados disponibilizados no GenBank.

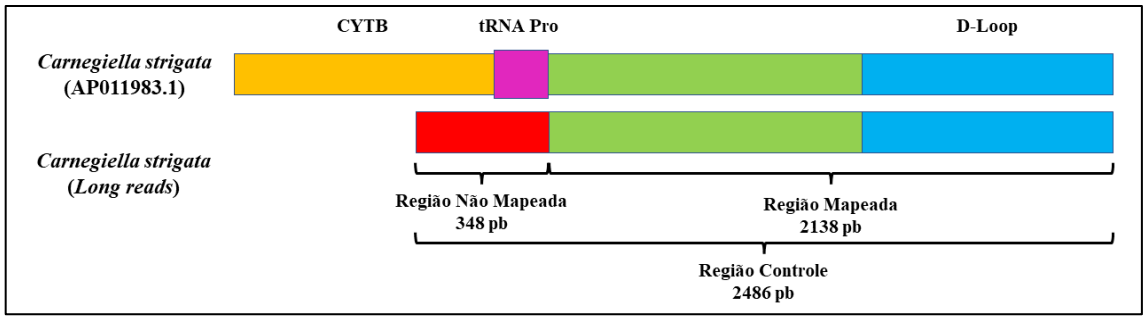


Figura 12: Representação do mapeamento da região controle gerada com sequências *long reads* contra o mitogenoma de referência de *C. strigata*.

Tendo em vista o resultado do mapeamento acima, três sequências foram utilizadas para a realização da busca por repetições *in tandem*, sendo uma a região não codificadora da referência de 2202 pb (Região desconhecida + D-loop) (Figura 11), e outras duas derivadas da região controle *long reads*, uma com um total de 2138 pb (Região controle mapeada), e outra com 2468 pb comportando a anterior mais a região não mapeada (Região controle completa) (Figura 12).

Foram encontrados sete diferentes padrões de repetições *in tandem* para as três sequências utilizadas de *Carnegiella strigata* (Tabela 4). A região não codificadora do mitogenoma de referência exibiu um único padrão de repetições a partir da posição 18 a 1196, compreendendo quase que todo o trecho descrito como região desconhecida, a sequência ATAATATTACATATGTACTAGTACATATTATGC se manteve como o consenso de cada repetição com um tamanho final de 33 pb e ocorrendo 35,4 vezes.

A sequência acima foi encontrada como um padrão de repetição para as duas regiões controle produzidas por *long reads*, em que somente na região mapeada dispôs de 37,4 número de cópias, começando na posição 1 a 1236, enquanto na região controle completa apareceu 48,4 vezes, tendo início na posição 1 a 1548. Além disso, foram identificados um total de dois padrões de repetições para a RC mapeada, e quatro para a RC completa. Estes padrões basicamente consistiam na quebra ou junção da sequência consenso acima. Os outros padrões

e a forma como estão apresentados dentro genoma mitocondrial pode ser consultado no apêndice A - G.

Repetições *in tandem* presentes na RC já vêm sendo relatados no genoma mitocondrial de peixes e outros vertebrados, inclusive na ordem Characiformes (Xu et al. 2021). Assim, é assumido que as regiões que incluem as repetições façam parte da região controle total de *C. strigata*, evidenciando uma região controle acima da média encontrada de 1100 pb para peixes (Terencio et al. 2012, Satoh et al. 2016). Além disso, é sugerido que o comprimento desta região seja maior do que é observado na referência (2202 pb), em comparação com os dados sequenciados com a abordagem utilizada neste trabalho (2486 pb). Estas observações reforçam e validam a eficiência desta tecnologia na resolução de regiões contendo estas características no mitogenoma (Formenti et al. 2021; Kinkar et al. 2021).

No entanto, esta variação de tamanho da RC de populações naturais de peixes e outros organismos, mesmo nas moléculas biológicas de um único indivíduo, já é um fenômeno conhecido como heteroplasmia (Árnason e Rand, 1992; Bentzen et al. 1998). Este evento está estritamente relacionado à presença de sequências repetitivas na RC, onde a formação de potenciais estruturas secundárias nessas sequências levam a desempenhar um papel importante no término ou na continuação da síntese do DNA mitocondrial durante a replicação (Kornienko et al. 2018). Este fenômeno ocorre porque as repetições contendo múltiplas sequências associadas ao término (TAS), comumente TACAT e o seu complementar, tendem a formar estruturas espaciais secundárias que atuam como uma espécie de barreira para a DNA polimerase (Bernacki e Kilpatrick, 2020). As TAS foram identificadas para a maioria dos padrões de repetição e com a formação de estruturas secundárias, podendo ser consultadas no apêndice H–M.

Tabela 3*:* Características das repetições *in tandem* encontradas nas sequências pertencentes a região controle de *C. strigata*..

Região desconhecida + D-loop (AP011983.1)									
Índice	**Cópias**	**Tamanho**	**Match (%)**	**Indels (%)**	**Score**	**A**	**C**	**G**	**T**
18 - 1196	35.4	33	98	1	2247	39	11	9	39
Região controle mapeada (*Long reads*)									
Índice	**Cópias**	**Tamanho**	**Match (%)**	**Indels (%)**	**Score**	**A**	**C**	**G**	**T**

1 - 1236	37.4	33	99	0	2440	39	12	8	39
22 - 1235	111.6	10	60	25	106	39	12	8	39

Região controle completa (*Long reads*)

Índice	**Cópias**	**Tamanho**	**Match (%)**	**Indels (%)**	**Score**	**A**	**C**	**G**	**T**
1 - 1584	48.4	33	91	4	1794	39	12	8	39
1 - 1584	23.7	66	93	3	2765	39	12	8	39
1 - 1584	15.8	99	96	2	2932	39	12	8	39
248 - 1583	122.5	10	60	25	122	39	12	8	39

4.2 *Gymnorhamphichthys rondoni.*

4.2.1 SEQUENCIAMENTO, MONTAGEM E COBERTURA.

O sequenciamento do espécime de *Gymnorhamphichthys rondoni* gerou um total de 64.459 reads. Para a seleção apenas das sequências de origem mitocondrial, o mitogenoma de *Gymnorhamphichthys sp*. (AP011980.1) disponível no GenBank foi utilizado como referência para a realização do mapeamento, o que resultou num total de 866 reads que prosseguiram para a montagem do genoma mitocondrial, como mostra a Tabela 1.

O tamanho total do genoma encontrado para *G. rondoni* foi de 16.566 pb. A cobertura de sequenciamento se manteve em um valor estável e pode ser visualizada na Figura 13. A composição nucleotídica foi: T, 25,8%; C, 30,4%, A, 27,5% e G, 16.3%, sendo o conteúdo GC total de 46,7%, estando um pouco abaixo do conteúdo AT comumente observado para outras espécies de peixes teleósteos (Prosdocimi et al. 2011).

4.2.2 ESTRUTURA E ORGANIZAÇÃO DO MITOGENOMA.

As *long read*s mitocondriais utilizadas foram suficientes para a confecção de um único contig circular que abrange o genoma mitocondrial completo da espécie (Figura 14). Após a anotação automática e manual, foram recuperados todos os 37 genes mitocondriais, 13 GCPs 22 tRNAs, dois rRNAs e uma região controle D-loop de 955 pb. Apenas o gene NAD6 foi esteve presente na *light strand*, enquanto a *heavy strand* exibiu os demais GCPa. A ordenação e organização identificada de todos os genes foi congruente com a conformação de outros mitogenomas de Gymnotiformes (Aguilar et al. 2019). Apesar do número ainda baixo de genomas mitocondriais reportados nos bancos de dados públicos para a ordem, o grupo vem

sendo um alvo em trabalhos de anúncios de mitogenomas (Elbassiouny et al. 2016, Sandoval et al. 2018).

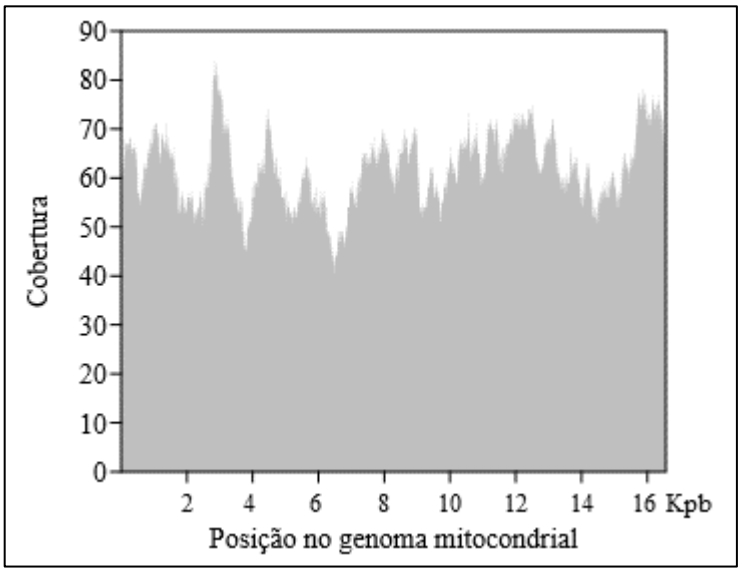


Figura 13: Cobertura da montagem final de *Gymnorhamphichthys rondoni*.

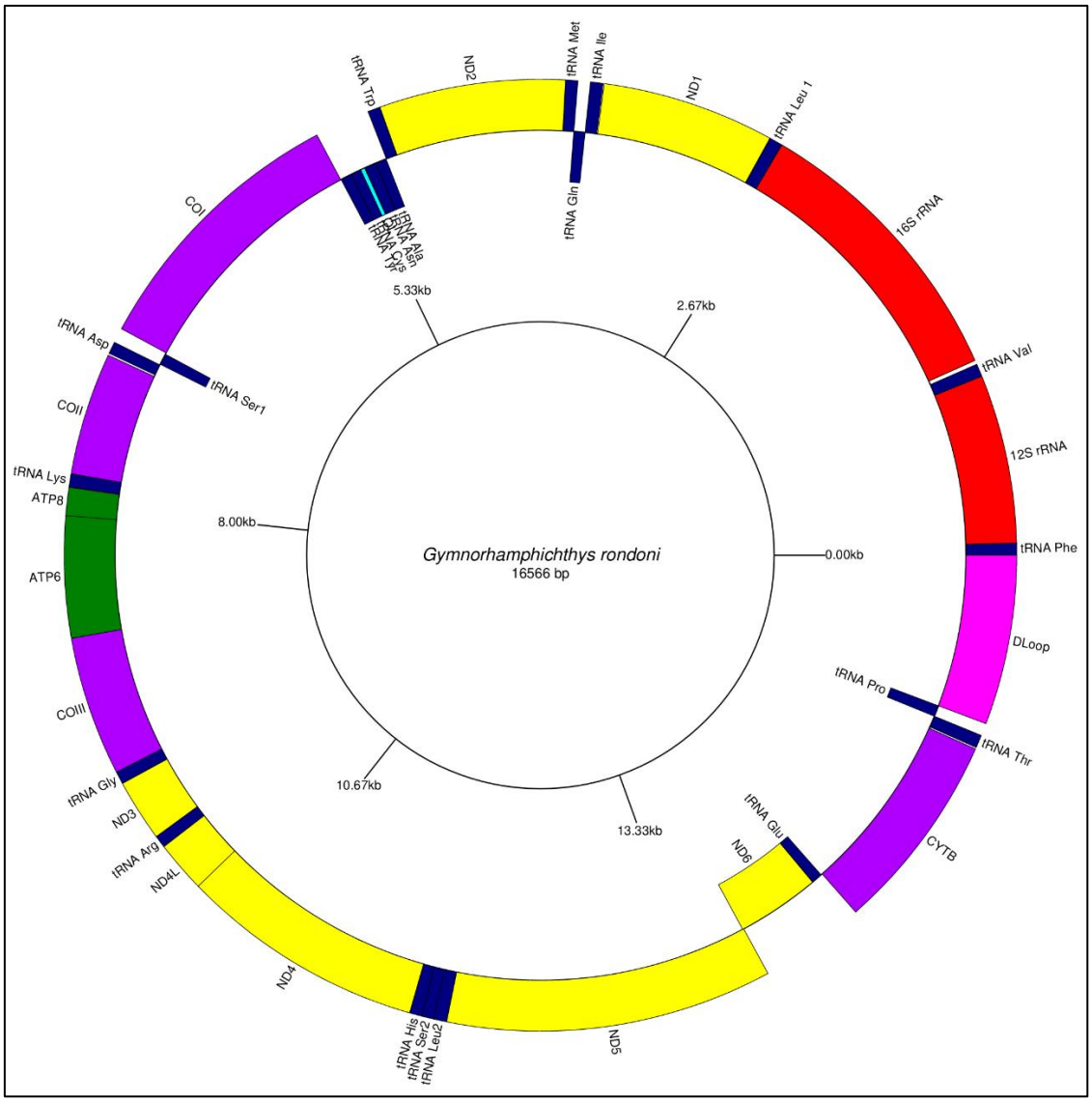


Figura 14: Genoma mitocondrial circular de *Gymnorhamphichthys rondoni*.

O códon de início ATG foi o mais presente para os genes codificadores de proteína, sendo representado dez vezes para os seguintes genes NAD2, COII, ATP8, ATP6, COIII, NAD3, NAD4L, NAD4, NAD5 e CYTB, enquanto o códon de início GTG ocorreu nos três últimos genes restantes (NAD1, COI e NAD6). Dez desses mesmos genes (NAD1, NAD2, COI, ATP8, ATP6, COII, NAD4L, NAD5, NAD6 e CYTB) terminaram com o códon TAA, e apenas o gene NAD3 terminou com o códon alternativo TAG, enquanto COII e NAD4 terminaram sua sequência com o códon incompleto T, sendo uma característica bastante encontrada no DNA mitocondrial de peixes, em que apesar de incompletos, são gerados após a etapa de transcrição por meio da poliadenilação do mRNA (Ojala et al. 1981; Sun et al. 2021).

Foram identificados todos os 22 tRNAs presentes no mitogenoma da espécie do presente estudo, 14 localizados na *heavy strand* e oito na *light strand*, dos quais variaram em

comprimento de 66 pb (tRNA Cys) a 74 pb (tRNA Lys). Os tRNAs Leu e Ser tiveram suas duas formas (1 e 2) encontradas dentro do genoma mitocondrial. Enquanto os rRNAs ribossomais 12S e 16S se mantiveram localizados entre os tRNAs Phe e Val, e Val e Leu, e com tamanho de 948 e 1629 pb, respectivamente.

Algumas sobreposições e espaços intergênicos foram observados, seus tamanhos e o local onde ocorreram no mitogenoma podem ser visualizados na Tabela 4. Um total de 99 pb de espaços intergênicos foram encontrados, estando distribuídos 13 vezes ao longo do genoma mitocondrial, com destaque para o maior espaço intergênico de 23 pb encontrado entre o rRNA 16S e o tRNA Leu. Além da região intergênica de 29 pb localizada entre o tRNA Asn e Cys, a qual está associada a origem de replicação OL da *light strand*, bem como uma região de 955 pb representando a região controle D-loop. As sobreposições entre genes totalizaram 31 pb, estando presente 10 vezes nas junções: tRNA Ile e tRNA Gln (1 pb), tRNA Gln e tRNA Met (1 pb), NAD2 e tRNA Trp (2 pb), ATP6 e ATP8 (7 pb), ATP6 e CO33 (1 pb), COIII e tRNA Gly (1 pb), NAD3 e tRNA Arg (2 pb), NAD4L e NAD4 (7 pb), NAD5 e NAD6 (8 pb) e tRNA Thr e tRNA Pro (1 pb). O tamanho e posição inicial e final de cada gene está representado na Tabela 2.

Tabela 4: Características do mitogenoma de *Gymnorhamphichthys rondoni*, incluindo a localização, tamanho, espaços intergênicos e códons para cada gene encontrado.

Gene	Localização		Tamanho	Espaço Intergênico/ Sobreposição (pb)	Códon		Fita
	Início	Fim			Início	Parada	
tRNA Phe	1	68	68				H
12S rRNA	68	1016	948	0			H
tRNA Val	1017	1088	72	0			H
16S rRNA	1112	2740	1629	23			H
tRNA Leu 1	2741	2815	75	0			H
NAD1	2816	3787	972	0	GTG	TAA	H
tRNA Ile	3793	3864	72	5			H
tRNA Gln	3864	3934	71	-1			L
tRNA Met	3934	4002	69	-1			H
NAD2	4003	5049	1047	0	ATG	TAA	H
tRNA Trp	5048	5118	71	-2			H

tRNA Ala	5120	5188	69	1			L
tRNA Asn	5190	5262	73	1			L
OL	5265	5292	29	0			L
tRNA Cys	5292	5357	66	0			L
tRNA Tyr	5358	5427	70	0			L
COI	5429	6976	1548	1	GTG	TAA	H
tRNA Ser1	6985	7055	71	8			L
tRNA Asp	7060	7128	69	4			H
COII	7139	7829	691	10	ATG	T	H
tRNA Lys	7830	7903	74	0			H
ATP8	7905	8069	165	1	ATG	TAA	H
ATP6	8063	8746	684	-7	ATG	TAA	H
COIII	8746	9531	786	-1	ATG	TAA	H
tRNA Gly	9531	9601	71	-1			H
NAD3	9602	9952	351	0	ATG	TAG	H
tRNA Arg	9951	10020	70	-2			H
NAD4L	10021	10317	297	0	ATG	TAA	H
NAD4	10311	11691	1381	-7	ATG	T	H
tRNA His	11692	11760	69	0			H
tRNA Ser2	11761	11827	67	0			H
tRNA Leu2	11827	11901	73	1			H
NAD5	11902	13737	1836	0	ATG	TAA	H
NAD6	13730	14254	525	-8	GTG	TAA	L
tRNA Glu	14255	14323	69	0			L
CYTB	14331	15461	11311	7	ATG	TAA	H
tRNA Thr	15470	15542	73	8			H
tRNA Pro	15542	15611	70	-1			L
D-Loop	15612	16566	955				

4.3 *Potamorrhaphis guianensis.*

4.3.1 SEQUENCIAMENTO, MONTAGEM E COBERTURA.

Um total de 3116007 reads foram obtidas com o sequenciamento do espécime de *Potamorrhaphis guianensis*. Após a realização do mapeamento, 584 reads potencialmente

mitocondriais permaneceram para a etapa de montagem, isto tendo como referência o genoma mitocondrial de *Ablennes hians*, membro da família Belonidae, assim como *P. guianensis* (Tabela 1).

O resultado após a montagem proporcionou dois contigs, um primeiro de 7930 pb e um segundo de 5444 pb. Embora ambos tenham apresentado uma cobertura consideravelmente mais alta (>80) do que as duas espécies apresentadas neste estudo (Figura 15), os contigs não tiveram abrangência e sobreposição suficiente para contemplar o mitogenoma completo de *P. guianensis*.

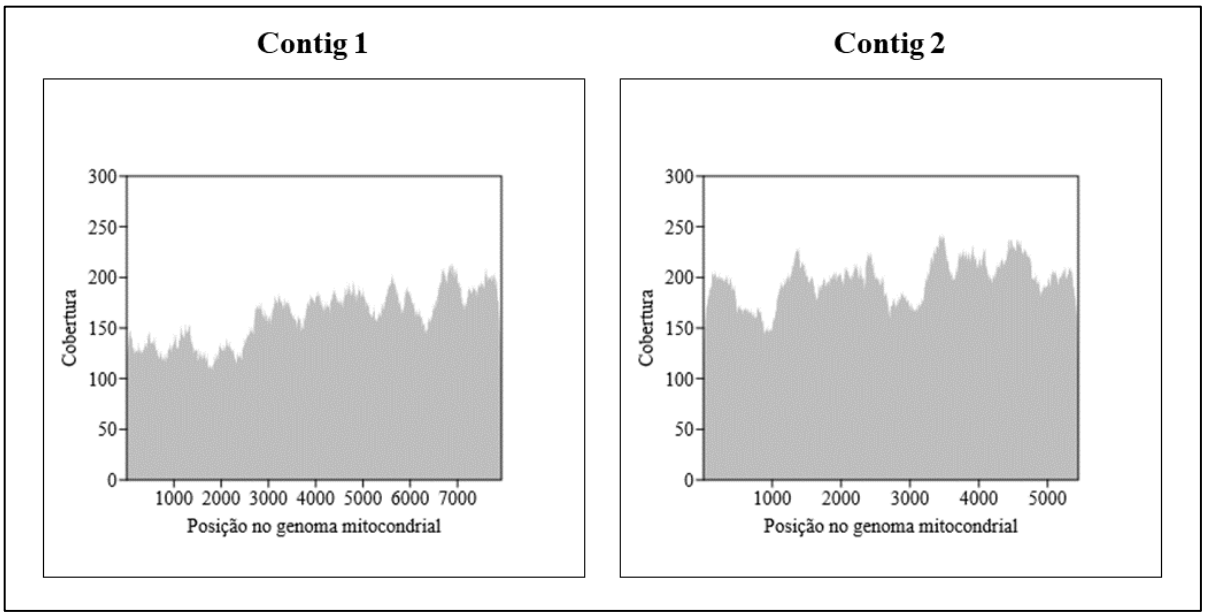


Figura 15: Cobertura de cada contig resultante da montagem de *Potamorrhaphis guianensis*.

4.3.2 ESTRUTURA E ORGANIZAÇÃO DO MITOGENOMA.

A predição gênica de cada contig está representada linearmente na Figura 16. O primeiro apresentou a porção inicial convencional do genoma mitocondrial, iniciando com o tRNA Phe ao tRNA Lys, enquanto o segundo contig partiu do fragmento do gene ATP6 ao fragmento do gene NAD5. Um total de 31 genes foram recuperados para o mitogenoma de *P. guianensis* usando sequências *long reads*, sendo 19 tRNAs, 2 rRNAs, e 9 GCPs. Como já foi mencionado anteriormente, os genes ATP6 e NAD5 foram recuperados de forma parcial, ao passo que os 29 genes restantes foram anotados completamente. No entanto, a ordenação dos genes se mostrou semelhante com o típico padrão encontrado em Beloniformes e outros peixes (Cui et al. 2018).

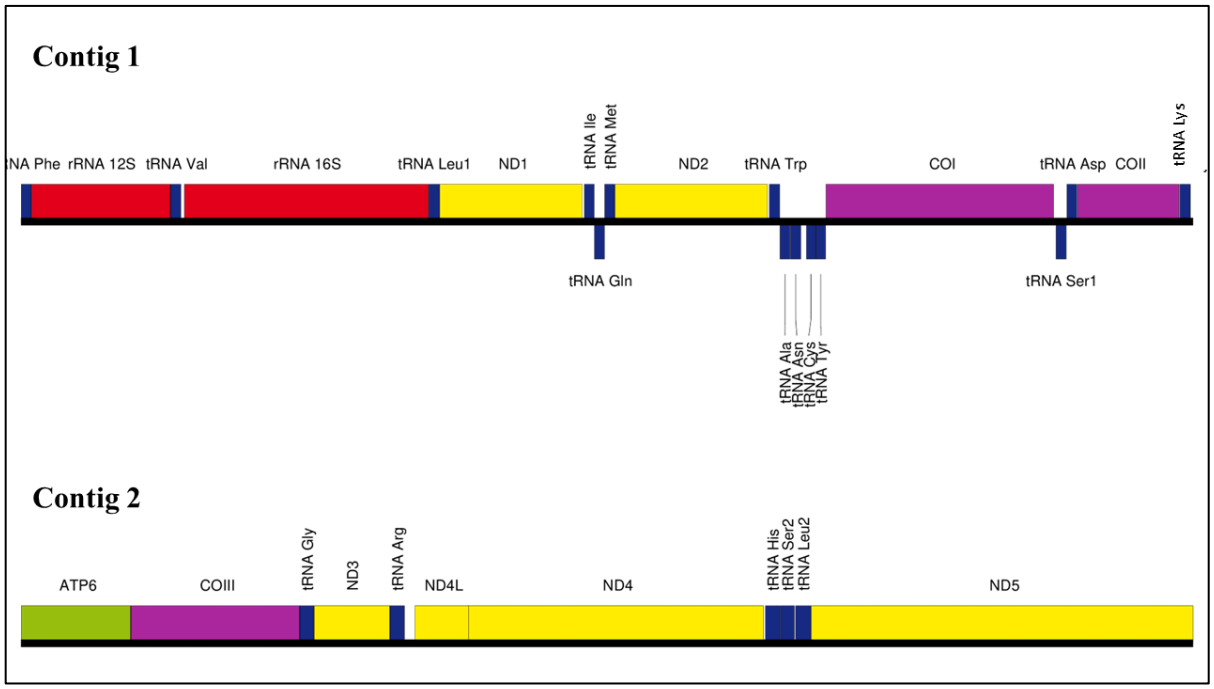


Figura 16: Anotação gênica para cada contig gerado da montagem de *Potamorrhaphis guianensis*.

A localização, tamanho e códons utilizados para todos os genes recuperados estão mostrados na Tabela 5. Mais uma vez, a maioria dos GCPs (NAD1, NAD2, COII, COIII, NAD4 e NAD5) apresentaram ATG como principal *start codon*, com a exceção do gene COI que exibiu GTG, além de NAD3 e NAD4L com ATA, sendo o primeiro registro deste último códon em peixes da família Belonidae. Não foi possível registrar o códon de início de ATP6 por não ter sido possível a recuperação da porção inicial do gene em nenhum contig.

Os PCGs em grande maioria apresentaram *stop codons* incompletos: NAD1 e COII (T); ATP6 e COIII (TA). Os genes NAD3 e NAD4L foram os únicos a apresentar os códons de parada mais convencionais TAG e TAA, respectivamente. Além disso, mesmo após anotação manual por meio da procura de ORFs, não foi possível determinar o códon de parada para NAD2, COI e NAD4, que terminaram sua sequência com TT, CAA e GA, respectivamente possivelmente pelas mesmas razões discutidas anteriormente para alguns genes que apresentaram a mesma dificuldade em *C. strigata* (Ver seção 4.2.2), sendo possível visualizar os *stop codons* internos após a realização da tradução do conteúdo gênico destes genes (Figura 17).

ND2 – 5' 3'

Met N P Y I T P Met Met L L S Met S L G Y V I T L T S S H W F L A
W Met G L E I S T L A I L P L Met A G L H H P R A I E A T T K Y F L
T Q A A A A A Met I L F A T T S N A W I T G Q W D I Y Q L S H P L P
L T Met I T I A L A L K I G L A P L H T W L P E V L Q G L N L S T G
L I L S T W Q K I A P F S L L I Q I Q P Y N T F P L I I L G V I S T
L V G G W G G L N Q T Q L R K I L A Y S S T A H L G W A V L A Met Q
F Y P L L P L L Q F L F Met L Y P H H Q H F Stop H S K P T S P Q T
L I H F P P P E P K P L S Stop H P Stop F H Stop F S S H L E D S L
P L Q A S Y Q N D Stop F Y K N S L N K T S Q S L Q H Stop P L S P L
F S A F T S T S E F H T P Stop S L Q Y H Q Met T Y Q A Stop H P D A
S L F T K I H F Y Stop L H Y Q P Q L F A S Y P Stop P P P S L H

COI - 5' 3'

V T F V R W Met F S T N H K D I G T L Y L I F G A W A G Met V G T A
L S L L I R A E L S Q P G S L L G D D Q I Y N V I V T A H A F
V Met I F F Met V Met P V Met I G G F G N W Met I P L Met I G A P
D Met A F P R Met N N Met S F W L L P P S F L L L L A S S G I E A G
A G T G W T V Y P P L A S N L A H A G A S V D L T I F S L H L A G V
S S I L G A I N F I T T I I N Met K P P T I S Q Y Q T P L F V W A V
L I T A V L L L L F T T R V S C Stop N Y N T S Y Stop P K S K H H I
L W P C R Stop Stop Stop P Y P L P A P I L I L W T P Stop S L Y S N
S P Stop I W F N L S H C S F L L Stop Q K Stop T F W L Y G Y
S Met S N N S N W S P W L Y R L S S S H I Y
S Stop Y Stop R Stop H T C L L Y I R Y H N Y C H P N Stop C K S I Q
L A C D I A W Stop S N Q Met Stop N P S P V S T R F Y F F I Y S
W Stop F N W Y Y F S Stop L F F Stop Y Y S P W H L L R C S T L P L
C F I Y G S C F C N H Stop S I R P L I P F I Y Stop L Y F T R H L N
K N P L Y Y Y I L W G Stop S H F F P S T F L Stop P C Stop Y T P T
L F Stop L P H A Y A L W N T I S S Met G S F Met S L V A V I L F L
F I I W E A F T S K R E Met L S I E Met T S Met N V E W L H G Y P P
L L H T F E E P A F V Q T Q

ND4 - 5' 3'

Met L K I L A P T T L L L L T T W L L P P K K L Met I R N S F Met Q
L T H C L Stop Q P F Met N Stop I T H G N Stop Met N T P K P L H G
N Stop S T L N P F T Y S H L L T L T L N N Y S K P K P Y I A R T Y
Y S T T I I Y F P S Y I T S N L P Y P F L Q R H Stop N N Y V L Y H
I W S Y P H P N F I Y Y Y P L Stop Q P N Stop T T Q C Stop N Met F
P F L Y T S R L T S T T N R P P P S T K H N W N P V S S Y P
S Met Y Stop S P T L N S F C R Q N L Met S Stop Met P S S V P R K N
T P L T E P I Y D S Q K P T Stop K P Q L Q D L Stop F L P Q F
Y Stop N Stop E A T A W Y E F Y Q H Stop S P Stop P K N Stop V T P
L L S Stop P Y E V Stop Stop Stop Q A L S A S D K Q T Stop N P L L
L T H L L V T Stop A Stop L L L V F L S K H H E A L P D
L Stop F L Stop S L Met D Stop H H P P Y S V Stop P T P T T N E H T
A E Q Stop F L Y E A F K Stop P S P L Stop L H D D
F Stop L A Stop P I S P S P L S P I S Stop E N Stop Stop L L P L Y
L T D P L E P L F L L A L E P Stop L Q Q A T H S L Y F Y Q L N K A
P S P T Met L L L F I Q H T P E N T Y Stop F F F I Stop P P F Y Y S
S S N Q N Stop S W A

Figura 17: Tradução das sequências dos genes NAD2, COI e NAD4, com destaque para a detecção de stop *codons* internos em *P. guianensis*..

Foram identificados 19 tRNAs, que variam de tamanho entre 65 pb (tRNA Cys) a 74 pb (tRNA Lys), 13 localizados na *heavy strand* e seis localizados na *light strand*. Os rRNAs também tiveram sua localização típica entre o tRNA Phe, tRNA Val e tRNA Leu1, para rRNA 12S e rRNA 16S, com 942 pb e 1653 pb, respectivamente. Os espaços intergênicos apareceram 13 vezes somando os dois contigs e totalizaram 170 pb, sendo o valor mais alto apresentado para as espécies deste estudo, tendo tRNA Arg e NAD4L apresentando o maior valor (46 pb), além do espaço intergênico de 35 pb configurado como a origem de replicação da *light strand* localizada entre o tRNA Asn e Cys. Já as sobreposições, ocorreram sete vezes, num total de apenas 16 pb: rRNA 12S e tRNA Val (1 pb); tRNA Ile e tRNA Gln (1 pb); tRNA Gln e tRNA Met (1 pb); COIII e tRNA Gly (1 pb); tRNA Gly e NAD3 (3 pb); NAD3 e tRNA Arg (2 pb) e NAD4L e NAD4 (7 pb). Todas estas informações estão contidas na Tabela 5 abaixo.

Tabela 5: Características do mitogenoma de *Potamorrhaphis guianensis*, incluindo a localização, tamanho, espaços intergênicos e códons para cada gene encontrado por contig produzido, asteriscos identificam códons provavelmente incorretos.

Gene	Localização		Tamanho	Espaço Intergênico/ Sobreposição (pb)	Códon		Fita
	Início	Final			Início	Parada	
Contig 1							
tRNA Phe	1	69	69	-			H
12S rRNA	70	1011	942	0			H
tRNA Val	1011	1082	72	-1			H
16S rRNA	1105	2757	1653	22			H
tRNA Leu 1	2759	2831	73	1			H
NAD1	2832	3798	967	0	ATG	T	H
tRNA Ile	3811	3879	69	12			H
tRNA Gln	3879	3949	71	-1			L
tRNA Met	3949	4017	69	-1			H
NAD2	4018	5051	1034	0	ATG	TT*	H
tRNA Trp	5063	5134	72	11			H
tRNA Ala	5135	5203	69	0			L
tRNA Asn	5205	5277	73	1			L
OL	5278	5312	35	0			L

tRNA Cys	5313	5377	65	0			L
tRNA Tyr	5378	5444	67	0			L
COI	5446	6987	1542	1	GTG	CAA*	H
tRNA Ser1	7002	7072	71	14			L
tRNA Asp	7076	7145	70	3			H
COII	7148	7835	688	2	ATG	T	H
tRNA Lys	7839	7912	74	3			H
Contig 2							
ATP6	1	509	509	-	-	TA	H
COIII	510	1294	785	0	ATG	TA	H
tRNA Gly	1294	1362	69	-1			H
NAD3	1360	1713	354	-3	ATA	TAG	H
tRNA Arg	1712	1782	71	-2			H
NAD4L	1829	2080	252	46	ATA	TAA	H
NAD4	2074	3449	1376	-7	ATG	GA*	H
tRNA His	3457	3525	69	7			H
tRNA Ser2	3526	3593	3593	0			H
tRNA Leu2	3597	3669	73	2			H
NAD5	3670	5444	1775	-	ATG	-	H

5 CONCLUSÃO

As investigações nesta área ainda estão em andamento, porém, os resultados aqui obtidos parecem confirmar indícios quanto a hipótese da ocorrência de *numts* no genoma de peixes amazônicos, como *Carnegiella strigata* e *Potamorrhaphis guianensis*. Isto demonstra a necessidade de cautela na construção bancos referenciais de marcadores genéticos, e utilização em trabalhos que dependem destes dados para descrição de diversidade taxonômica. Além disso, a exploração da região controle de *C. strigata* revelou um padrão incomum quanto ao tamanho encontrado em peixes e de dados disponíveis da mesma espécie, apontando fortes evidências de heteroplasmia, suportado ainda pela detecção de estruturas secundárias em numerosas repetições *in tandem* presentes nesta região. Apesar desses achados serem congruentes com o que é encontrado na literatura, trabalhos futuros deverão se concentrar nas questões levantadas quanto ao entendimento maior da importância e função deste tipo de estruturação na região controle de algumas espécies. Os resultados provenientes deste trabalho suportam a ideia de que a abordagem de sequenciamento *genome skimming*, associada a tecnologia de produção de sequências do tipo *long reads*, foram efetivas na recuperação da maior parte do conteúdo do genoma mitocondrial completo para as três estudadas. Mais além, apresentamos o mitogenoma circular e a sua anotação funcional de *Gymnorhamphichthys rondoni* que futuramente será disponibilizado em bancos de dados públicos, isto poderá eventualmente conduzir a construção de filogenias mais robustas e completas dentro da ordem Gymnotiformes. Nosso estudo nos levou a concluir que pesquisas futuras voltadas para mitogenomas, devem aspirar a aplicação de montagens híbridas em conjunto com dados do tipo *short reads*, a fim de garantir um produto de maior confiabilidade. Em suma, nossa abordagem escolhida se mostrou satisfatória quanto a tentativa de obtenção de genomas mitocondriais completo, bem como a apuração de regiões ainda a serem mais bem compreendidas.

REFERÊNCIAS

AGUILAR, Celestino et al. Mitogenomics of Central American weakly-electric fishes. Gene, v. 686, p. 164-170, 2019.

ALBERT, James S.; TAGLIACOLLO, Victor A.; DAGOSTA, Fernando. Diversification of Neotropical freshwater fishes. Annual Review of Ecology, Evolution, and Systematics, v. 51, p. 27-53, 2020.

AMBARDAR, Sheetal et al. High throughput sequencing: an overview of sequencing chemistry. Indian journal of microbiology, v. 56, n. 4, p. 394-404, 2016.

ANDUJAR, Carmelo et al. Phylogenetic community ecology of soil biodiversity using mitochondrial metagenomics. Molecular Ecology, v. 24, n. 14, p. 3603-3617, 2015.

ANDUJAR, Carmelo et al. Validated removal of nuclear pseudogenes and sequencing artefacts from mitochondrial metabarcode data. BioRxiv, 2020.

ANTUNES, Agostinho; RAMOS, Maria João. Discovery of a large number of previously unrecognized mitochondrial pseudogenes in fish genomes. Genomics, v. 86, n. 6, p. 708-717, 2005.

ARNASON, E.; RAND, D. M. Heteroplasmy of short tandem repeats in mitochondrial DNA of Atlantic cod, Gadus morhua. Genetics, v. 132, n. 1, p. 211-220, 1992.

ASAKAWA, Shuichi et al. Strand-specific nucleotide composition bias in echinoderm and vertebrate mitochondrial genomes. Journal of molecular evolution, v. 32, n. 6, p. 511-520, 1991.

BAYLEY, Hagan. Nanopore sequencing: from imagination to reality. Clinical chemistry, v. 61, n. 1, p. 25-31, 2015.

BENSON, Gary. Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research, v. 27, n. 2, p. 573-580, 1999.

BENTZEN, Paul et al. Tandem repeat polymorphism and heteroplasmy in the mitochondrial control region of redfishes (Sebastes: Scorpaenidae). Journal of Heredity, v. 89, n. 1, p. 1-7, 1998.

BERNACKI, Lucas E.; KILPATRICK, C. William. Structural variation of the turtle mitochondrial control region. Journal of Molecular Evolution, v. 88, n. 7, p. 618-640, 2020.

BERNT, Matthias et al. Genetic aspects of mitochondrial genome evolution. Molecular phylogenetics and evolution, v. 69, n. 2, p. 328-338, 2013.

BERNT, Matthias et al. MITOS: improved de novo metazoan mitochondrial genome annotation. Molecular phylogenetics and evolution, v. 69, n. 2, p. 313-319, 2013.

BETANCUR-R, Ricardo et al. Phylogenetic classification of bony fishes. BMC evolutionary biology, v. 17, n. 1, p. 1-40, 2017.

BIRINDELLI, José LO; SIDLAUSKAS, Brian L. Preface: How far has Neotropical Ichthyology progressed in twenty years?. Neotropical Ichthyology, v. 16, n. 3, 2018.

BOORE, Jeffrey L. Animal mitochondrial genomes. Nucleic acids research, v. 27, n. 8, p. 1767-1780, 1999.

BROUGHTON, Richard E.; MILAM, Jami E.; ROE, Bruce A. The complete sequence of the zebrafish (Danio rerio) mitochondrial genome and evolutionary patterns in vertebrate mitochondrial DNA. Genome research, v. 11, n. 11, p. 1958-1967, 2001.

COISSAC, Eric et al. From barcodes to genomes: extending the concept of DNA barcoding. Molecular ecology, v. 25, n. 7, p. 1423-1428, 2016.

COLLINS, Rupert A. et al. Non-specific amplification compromises environmental DNA metabarcoding with COI. Methods in Ecology and Evolution, v. 10, n. 11, p. 1985-2001, 2019.

CRAMPTON-PLATT, Alex et al. Mitochondrial metagenomics: letting the genes out of the bottle. GigaScience, v. 5, n. 1, p. s13742-016-0120-y, 2016.

DANECEK, Petr et al. Twelve years of SAMtools and BCFtools. Gigascience, v. 10, n. 2, p. giab008, 2021.

DE CARVALHO, Daniel C. et al. Deep barcode divergence in Brazilian freshwater fishes: the case of the São Francisco River basin. Mitochondrial Dna, v. 22, n. sup1, p. 80-86, 2011.

DEAMER, David; AKESON, Mark; BRANTON, Daniel. Three decades of nanopore sequencing. Nature biotechnology, v. 34, n. 5, p. 518-524, 2016.

DIAS, Murilo S. et al. Natural fragmentation in river networks as a driver of speciation for freshwater fishes. Ecography, v. 36, n. 6, p. 683-689, 2013.

ELBASSIOUNY, Ahmed A. et al. Mitochondrial genomes of the South American electric knifefishes (Order Gymnotiformes). Mitochondrial DNA Part B, v. 1, n. 1, p. 401-403, 2016.

ELBRECHT, Vasco; LEESE, Florian. Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass—sequence relationships with an innovative metabarcoding protocol. PloS one, v. 10, n. 7, p. e0130324, 2015.

FAN, Jian-Bing; CHEE, Mark S.; GUNDERSON, Kevin L. Highly parallel genomic assays. Nature Reviews Genetics, v. 7, n. 8, p. 632-644, 2006.

FLYNN, Tanya et al. Mitochondrial genome diversity among six laboratory zebrafish (Danio rerio) strains. Mitochondrial DNA Part A, v. 27, n. 6, p. 4364-4371, 2016.

FORMENTI, Giulio et al. Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. Genome biology, v. 22, n. 1, p. 1-22, 2021.

FROTA, Augusto et al. Inventory of the fish fauna from Ivaí River basin, Paraná State, Brazil. Biota Neotropica, v. 16, n. 3, 2016.

GAN, Han Ming; LINTON, Stuart M.; AUSTIN, Christopher M. Two reads to rule them all: Nanopore long read-guided assembly of the iconic Christmas Island red crab, Gecarcoidea natalis (Pocock, 1888), mitochondrial genome and the challenges of AT-rich mitogenomes. Marine genomics, v. 45, p. 64-71, 2019.

GASTEIGER, Elisabeth et al. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic acids research, v. 31, n. 13, p. 3784-3788, 2003.

GOODWIN, Sara; MCPHERSON, John D.; MCCOMBIE, W. Richard. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics, v. 17, n. 6, p. 333, 2016.

GRANDJEAN, Frederic et al. Rapid recovery of nuclear and mitochondrial genes by genome skimming from Northern Hemisphere freshwater crayfish. Zoologica Scripta, v. 46, n. 6, p. 718-728, 2017.

GRAU, Erwin Tramontin et al. Survey of mitochondrial sequences integrated into the bovine nuclear genome. Scientific reports, v. 10, n. 1, p. 1-11, 2020.

GREINER, Stephan; LEHWARK, Pascal; BOCK, Ralph. OrganellarGenomeDRAW (OGDRAW) version 1.3. 1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic acids research, v. 47, n. W1, p. W59-W64, 2019.

GUIMARÃES, Karen Larissa Auzier et al. DNA barcoding of fish fauna from low order streams of Tapajós River basin. PloS one, v. 13, n. 12, p. e0209430, 2018.

GUREVICH, Alexey et al. QUAST: quality assessment tool for genome assemblies. Bioinformatics, v. 29, n. 8, p. 1072-1075, 2013.

HEBERT, Paul DN et al. Biological identifications through DNA barcodes. Proceedings of the Royal Society of London. Series B: Biological Sciences, v. 270, n. 1512, p. 313-321, 2003.

HO, Steve S.; URBAN, Alexander E.; MILLS, Ryan E. Structural variation in the sequencing era. Nature Reviews Genetics, v. 21, n. 3, p. 171-189, 2020.

INGMAN, Max; GYLLENSTEN, Ulf. Vertebrate mitochondrial DNA. Reviews in Cell Biology and Molecular Medicine, 2006.

IP, Camilla LC et al. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. F1000Research, v. 4, 2015.

JACKMAN, Jake M. et al. eDNA in a bottleneck: obstacles to fish metabarcoding studies in megadiverse freshwater systems. bioRxiv, p. 2021.01. 05.425493, 2021.

JAIN, Miten et al. Improved data analysis for the MinION nanopore sequencer. Nature methods, v. 12, n. 4, p. 351-356, 2015.

JAIN, Miten et al. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome biology, v. 17, n. 1, p. 239, 2016.

JÉZÉQUEL, Céline et al. A database of freshwater fish species of the Amazon Basin. Scientific data, v. 7, n. 1, p. 1-9, 2020.

KCHOUK, Mehdi; GIBRAT, Jean-Francois; ELLOUMI, Mourad. Generations of sequencing technologies: from first to next generation. Biology and Medicine, v. 9, n. 3, 2017.

KINKAR, Liina et al. Nanopore sequencing resolves elusive long tandem-repeat regions in mitochondrial genomes. International journal of molecular sciences, v. 22, n. 4, p. 1811, 2021.

KOLMOGOROV, Mikhail et al. Assembly of long, error-prone reads using repeat graphs. Nature biotechnology, v. 37, n. 5, p. 540-546, 2019.

KORNIENKO, I. V. et al. Termination of replication and mechanisms of heteroplasmy in sturgeon mitochondrial DNA. Molecular Biology, v. 53, n. 1, p. 107-117, 2019.

KULSKI, Jerzy K. Next-generation sequencing—an overview of the history, tools, and "omic" applications. Next generation sequencing-advances, applications and challenges, p. 3-60, 2016.

KUMAR, Kishore R.; COWLEY, Mark J.; DAVIS, Ryan L. Next-generation sequencing and emerging technologies. In: Seminars in thrombosis and hemostasis. Thieme Medical Publishers, 2019. p. 661-673.

LADOUKAKIS, Emmanuel D.; ZOUROS, Eleftherios. Evolution and inheritance of animal mitochondrial DNA: rules and exceptions. Journal of Biological Research-Thessaloniki, v. 24, n. 1, p. 1-7, 2017.

LEVY, Shawn E.; MYERS, Richard M. Advancements in next-generation sequencing. Annual review of genomics and human genetics, v. 17, p. 95-115, 2016.

LI, Heng. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, v. 34, n. 18, p. 3094-3100, 2018.

LINARD, Benjamin et al. Metagenome skimming of insect specimen pools: potential for comparative genomics. Genome biology and evolution, v. 7, n. 6, p. 1474-1489, 2015.

LOPEZ, Jose V. et al. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. Journal of molecular evolution, v. 39, n. 2, p. 174-190, 1994.

LU, Hengyun; GIORDANO, Francesca; NING, Zemin. Oxford Nanopore MinION sequencing and genome assembly. Genomics, proteomics & bioinformatics, v. 14, n. 5, p. 265-279, 2016.

MABUCHI, Kohji et al. Gene rearrangements and evolution of tRNA pseudogenes in the mitochondrial genome of the parrotfish (Teleostei: Perciformes: Scaridae). Journal of molecular evolution, v. 59, n. 3, p. 287-297, 2004.

MAGURRAN, Anne E. Measuring biological diversity. John Wiley & Sons, 2013.

MALABARBA, Luiz Roberto; MALABARBA, Maria Claudia. Phylogeny and classification of Neotropical fish. In: Biology and Physiology of Freshwater Neotropical Fish. Academic Press, 2020. p. 1-19.

MALÉ, Pierre-Jean G. et al. Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. Molecular ecology resources, v. 14, n. 5, p. 966-975, 2014.

MAXAM, Allan M.; GILBERT, Walter. A new method for sequencing DNA. Proceedings of the National Academy of Sciences, v. 74, n. 2, p. 560-564, 1977.

METZKER, Michael L. Sequencing technologies—the next generation. Nature reviews genetics, v. 11, n. 1, p. 31-46, 2010.

MIKHEYEV, Alexander S.; TIN, Mandy MY. A first look at the Oxford Nanopore MinION sequencer. Molecular ecology resources, v. 14, n. 6, p. 1097-1102, 2014.

MILAN, David T. et al. New 12S metabarcoding primers for enhanced Neotropical freshwater fish biodiversity assessment. Scientific reports, v. 10, n. 1, p. 1-12, 2020.

MIYA, Masaki; NISHIDA, Mutsumi. The mitogenomic contributions to molecular phylogenetics and evolution of fishes: a 15-year retrospect. Ichthyological Research, v. 62, n. 1, p. 29-71, 2015.

NELSON, Joseph S.; GRANDE, Terry C.; WILSON, Mark VH. Fishes of the World. John Wiley & Sons, 2016.

NEVILL, Paul G. et al. Large scale genome skimming from herbarium material for accurate plant identification and phylogenomics. Plant methods, v. 16, n. 1, p. 1-8, 2020.

NOGUEIRA, Cristiano et al. Restricted-range fishes and the conservation of Brazilian freshwaters. PloS one, v. 5, n. 6, p. e11390, 2010.

OJALA, Deanna; MONTOYA, Julio; ATTARDI, Giuseppe. tRNA punctuation model of RNA processing in human mitochondria. Nature, v. 290, n. 5806, p. 470-474, 1981.

OJALA, Deanna; MONTOYA, Julio; ATTARDI, Giuseppe. tRNA punctuation model of RNA processing in human mitochondria. Nature, v. 290, n. 5806, p. 470-474, 1981.

PAPADOPOULOU, Anna; TABERLET, Pierre; ZINGER, Lucie. Metagenome skimming for phylogenetic community ecology: a new era in biodiversity research. Molecular Ecology, v. 24, n. 14, p. 3515-3517, 2015.

PEREIRA, Luiz HG et al. Can DNA barcoding accurately discriminate megadiverse Neotropical freshwater fish fauna? BMC genetics, v. 14, n. 1, p. 1-14, 2013.

PEREIRA, Ricardo J. et al. Mind the numt: Finding informative mitochondrial markers in a giant grasshopper genome. Journal of Zoological Systematics and Evolutionary Research, 2020.

POLLARD, Martin O. et al. *LONG READS*: their purpose and place. Human molecular genetics, v. 27, n. R2, p. R234-R241, 2018.

PRADHAN, Dibyabhaba et al. High-throughput sequencing. In: Data Processing Handbook for Complex Biological Data Sources. Academic Press, 2019. p. 39-52.

RANG, Franka J.; KLOOSTERMAN, Wigard P.; DE RIDDER, Jeroen. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome biology, v. 19, n. 1, p. 90, 2018.

REIS, Roberto E. et al. Fish biodiversity and conservation in South America. Journal of fish biology, v. 89, n. 1, p. 12-47, 2016.

REUTER, Jason A.; SPACEK, Damek V.; SNYDER, Michael P. High-throughput sequencing technologies. Molecular cell, v. 58, n. 4, p. 586-597, 2015.

RINCÓN-SANDOVAL, Melissa; BETANCUR-R, Ricardo; MALDONADO-OCAMPO, Javier A. Mitochondrial genomes of the South American electric knifefishes Eigenmannia humboldtii (Steindachner 1878), Eigenmannia limbata (Schreiner and Miranda Ribeiro 1903), Sternopygus aequilabiatus (Humboldt 1805) and Sternopygus macrurus (Bloch and Schneider 1801), (Gymnotiformes, Sternopygidae). Mitochondrial DNA Part B, v. 3, n. 2, p. 572-574, 2018.

SACCONE, C.; ATTIMONELLI, M.; SBISA, E. Structural elements highly preserved during the evolution of the D-loop-containing region in vertebrate mitochondrial DNA. Journal of molecular evolution, v. 26, n. 3, p. 205-211, 1987.

SANGER, Fred; COULSON, Alan R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of molecular biology, v. 94, n. 3, p. 441-448, 1975.

SATOH, Takashi P. et al. Structure and variation of the mitochondrial genome of fishes. BMC genomics, v. 17, n. 1, p. 1-20, 2016.

SCHATZ, Michael C.; DELCHER, Arthur L.; SALZBERG, Steven L. Assembly of large genomes using second-generation sequencing. Genome research, v. 20, n. 9, p. 1165-1173, 2010.

SCHENEKAR, Tamara et al. Reference databases, primer choice, and assay sensitivity for environmental metabarcoding: Lessons learnt from a re-evaluation of an eDNA fish assessment in the Volga headwaters. River Research and Applications, v. 36, n. 7, p. 1004-1013, 2020.

SCHIAVO, Giuseppina et al. A genomic landscape of mitochondrial DNA insertions in the pig nuclear genome provides evolutionary signatures of interspecies admixture. DNA Research, v. 24, n. 5, p. 487-498, 2017.

SCHNEIDER, Carlos Henrique et al. Cryptic diversity in the mtDNA of the ornamental fish Carnegiella strigata. Journal of fish biology, v. 81, n. 4, p. 1210-1224, 2012.

SHADEL, Gerald S.; CLAYTON, David A. Mitochondrial DNA maintenance in vertebrates. Annual review of biochemistry, v. 66, n. 1, p. 409-435, 1997.

SHENDURE, Jay; JI, Hanlee. Next-generation DNA sequencing. Nature biotechnology, v. 26, n. 10, p. 1135-1145, 2008.

SONG, Hojun et al. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proceedings of the national academy of sciences, v. 105, n. 36, p. 13486-13491, 2008.

STOTHARD, Paul. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques, v. 28, n. 6, p. 1102-1104, 2000.

STRAUB, Shannon CK et al. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. American Journal of Botany, v. 99, n. 2, p. 349-364, 2012.

SUN, Cheng-He et al. Mitochondrial genome structures and phylogenetic analyses of two tropical Characidae fishes. Frontiers in genetics, v. 12, p. 627402, 2021.

TAANMAN, Jan-Willem. The mitochondrial genome: structure, transcription, translation and replication. Biochimica et Biophysica Acta (BBA)-Bioenergetics, v. 1410, n. 2, p. 103-123, 1999.

TABERLET, Pierre et al. Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular ecology, v. 21, n. 8, p. 2045-2050, 2012.

TAN, Edwin YW et al. Genome skimming resolves the giant clam (Bivalvia: Cardiidae: Tridacninae) tree of life. Coral Reefs, v. 41, n. 3, p. 497-510, 2022.

TEDESCO, Pablo A. et al. A global database on freshwater fish species occurrence in drainage basins. Scientific data, v. 4, p. 170141, 2017.

TILLICH, Michael et al. GeSeq–versatile and accurate annotation of organelle genomes. Nucleic acids research, v. 45, n. W1, p. W6-W11, 2017.

TIMMIS, Jeremy N. et al. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nature Reviews Genetics, v. 5, n. 2, p. 123-135, 2004.

TREVISAN, Bruna et al. Genome skimming is a low-cost and robust strategy to assemble complete mitochondrial genomes from ethanol preserved specimens in biodiversity studies. PeerJ, v. 7, p. e7543, 2019.

VAN DIJK, Erwin L. et al. Ten years of next-generation sequencing technology. Trends in genetics, v. 30, n. 9, p. 418-426, 2014.

WANG, Jian-Xia et al. Tracking the Distribution and Burst of Nuclear Mitochondrial DNA Sequences (NUMTs) in Fig Wasp Genomes. Insects, v. 11, n. 10, p. 680, 2020.

WICK, Ryan R.; JUDD, Louise M.; HOLT, Kathryn E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome biology, v. 20, n. 1, p. 1-10, 2019.

XU, Wei; LIN, Shupeng; LIU, Hongyi. Mitochondrial genomes of five Hyphessobrycon tetras and their phylogenetic implications. Ecology and evolution, v. 11, n. 18, p. 12754-12764, 2021.

ZUKER, Michael. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic acids research, v. 31, n. 13, p. 3406-3415, 2003.

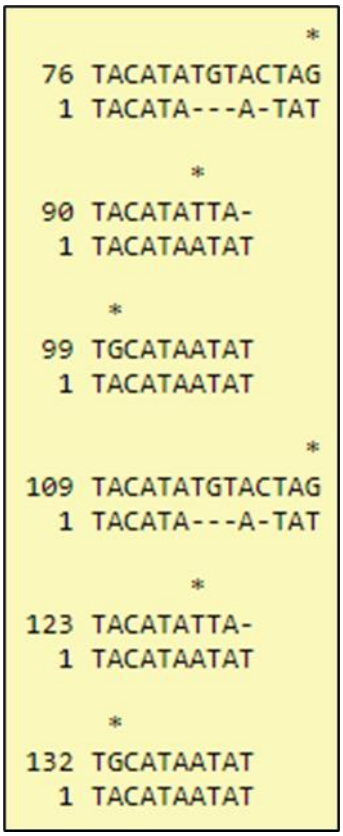
APÊNDICE A – Trecho de repetições *in tandem* da região desconhecida + d-loop (AP011983.1) no índice 18 – 1196.

```
 84 ATAATATTACATATGTACTAGTACATATTATGC
  1 ATAATATTACATATGTACTAGTACATATTATGC


117 ATAATATTACATATGTACTAGTACATATTATGC
  1 ATAATATTACATATGTACTAGTACATATTATGC


150 ATAATATTACATATGTACTAGTACATATTATGC
  1 ATAATATTACATATGTACTAGTACATATTATGC


183 ATAATATTACATATGTACTAGTACATATTATGC
  1 ATAATATTACATATGTACTAGTACATATTATGC
```

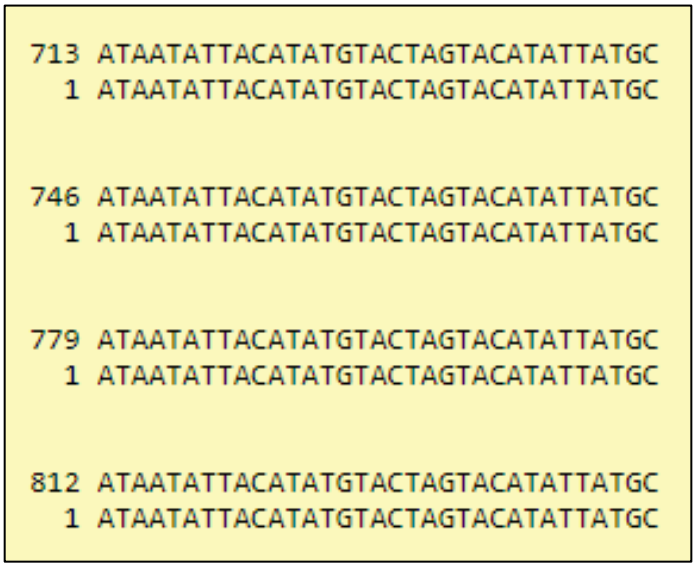
APÊNDICE B – Trecho de repetições *in tandem* da região controle mapeada (*long reads*) no índice 1 – 1236.

```
 69 ATAATATTACATATGTACTAGTACATATTATGC
  1 ATAATATTACATATGTACTAGTACATATTATGC


102 ATAATATTACATATGTACTAGTACATATTATGC
  1 ATAATATTACATATGTACTAGTACATATTATGC


135 ATAATATTACATATGTACTAGTACATATTATGC
  1 ATAATATTACATATGTACTAGTACATATTATGC


168 ATAATATTACATATGTACTAGTACATATTATGC
  1 ATAATATTACATATGTACTAGTACATATTATGC
```

APÊNDICE C – Trecho de repetições *in tandem* da região controle mapeada (*long reads*) no índice 22 – 1235.



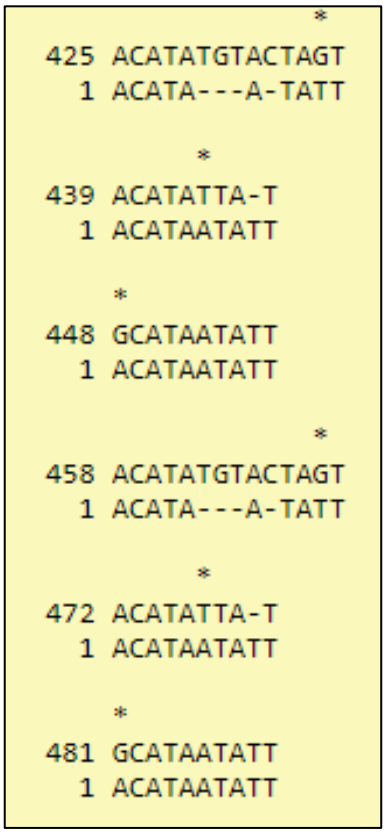
APÊNDICE D – Trecho de repetições *in tandem* da região controle completa (*long reads*) no índice 1 – 1584 (33 pb).



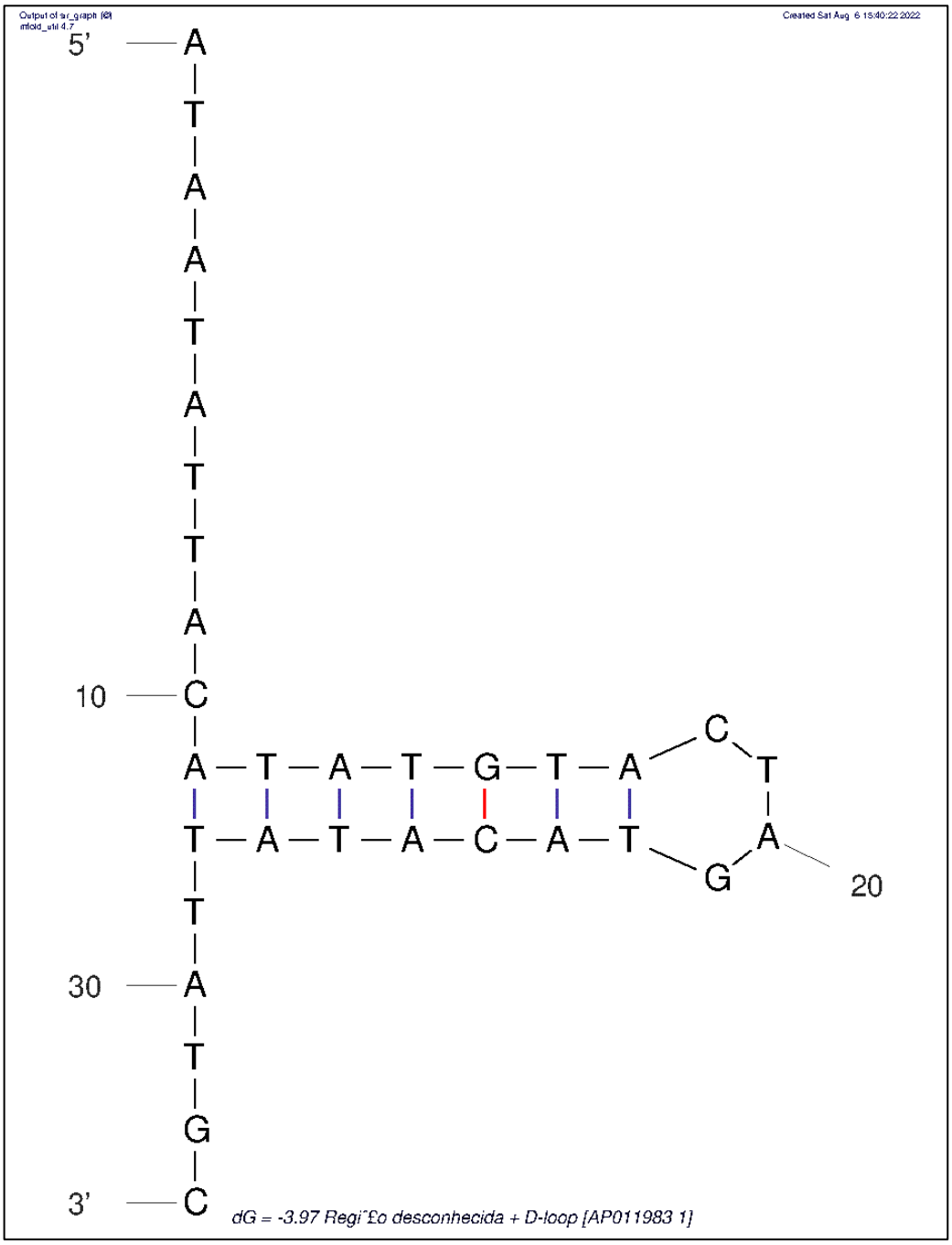
APÊNDICE E – Trecho de repetições *in tandem* da região controle completa (*long reads*) no índice 1 – 1584 (66 pb).

APÊNDICE F – Trecho de repetições *in tandem* da região controle completa (*long reads*) no índice 1 – 1584 (99 pb).

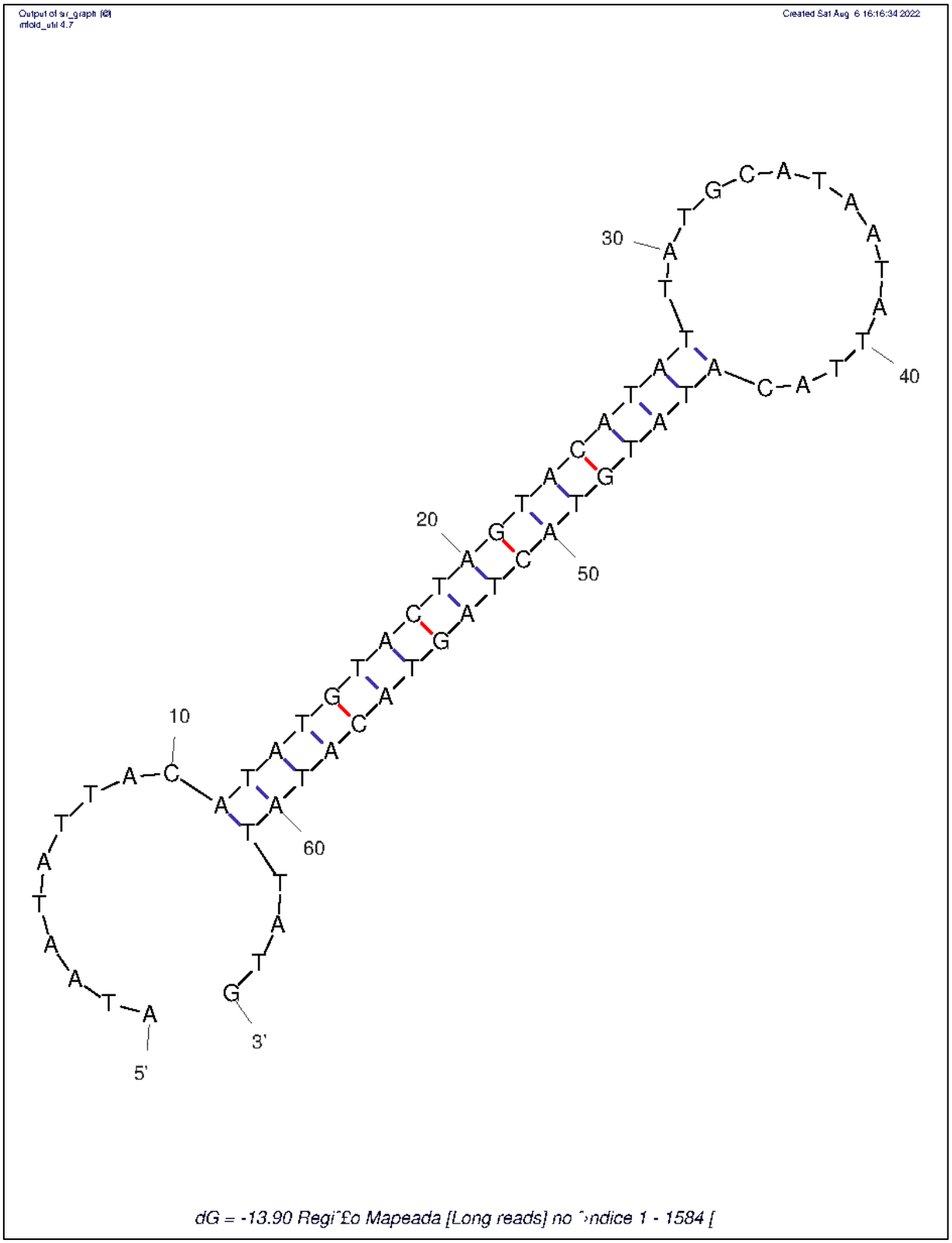
APÊNDICE G – Trecho de repetições *in tandem* da região controle completa (*long reads*) no índice 248 – 1583.



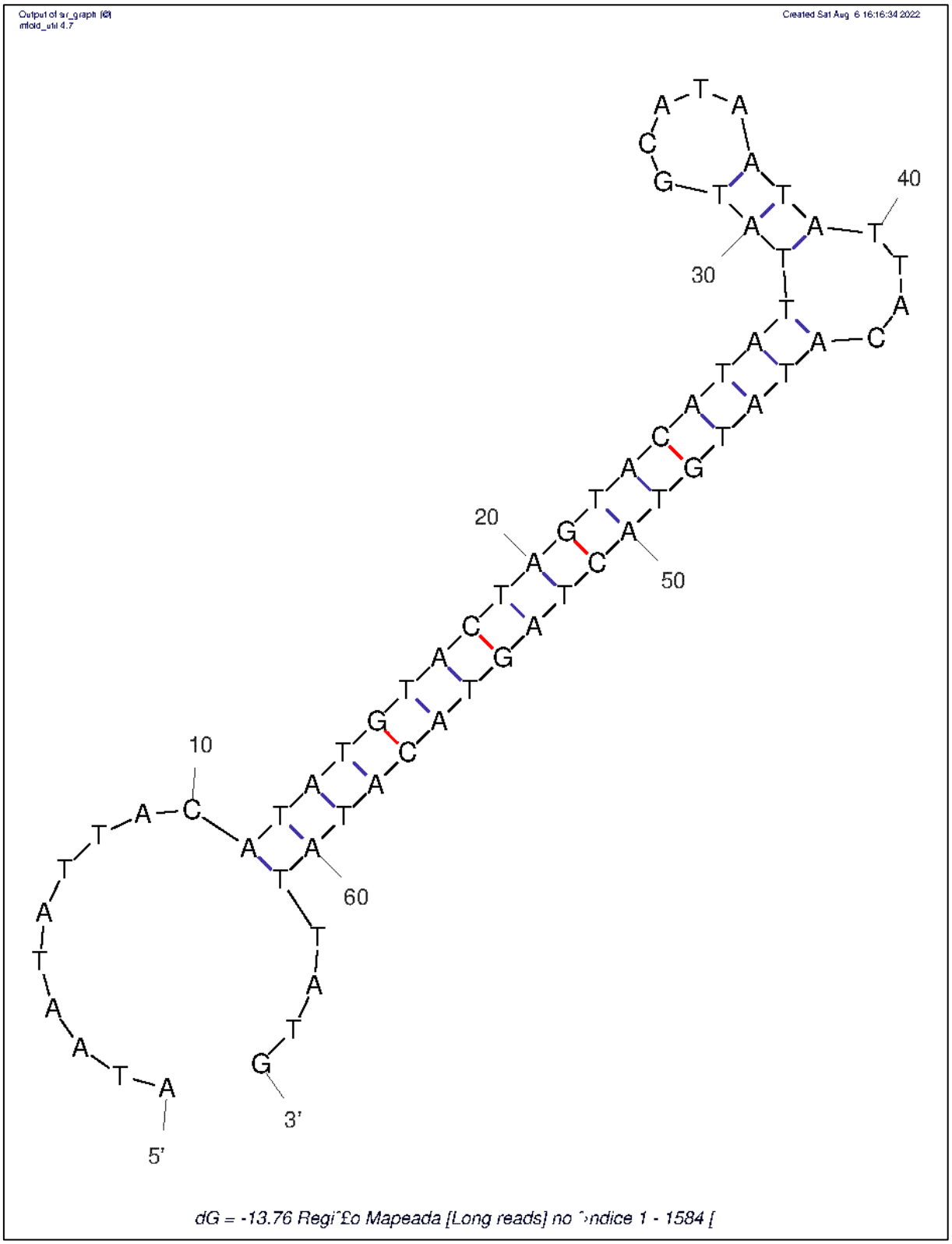
APÊNDICE H – Estrutura secundária (ΔG = -3.97 kcal/mol) da repetição *in tandem* consenso da região desconhecida + d-loop (AP011983.1) no índice 18 – 1196, região controle mapeada (*Long reads*) no índice 1 – 1236 e região controle completa (*Long reads*) no índice 1 – 1584 (33 pb).



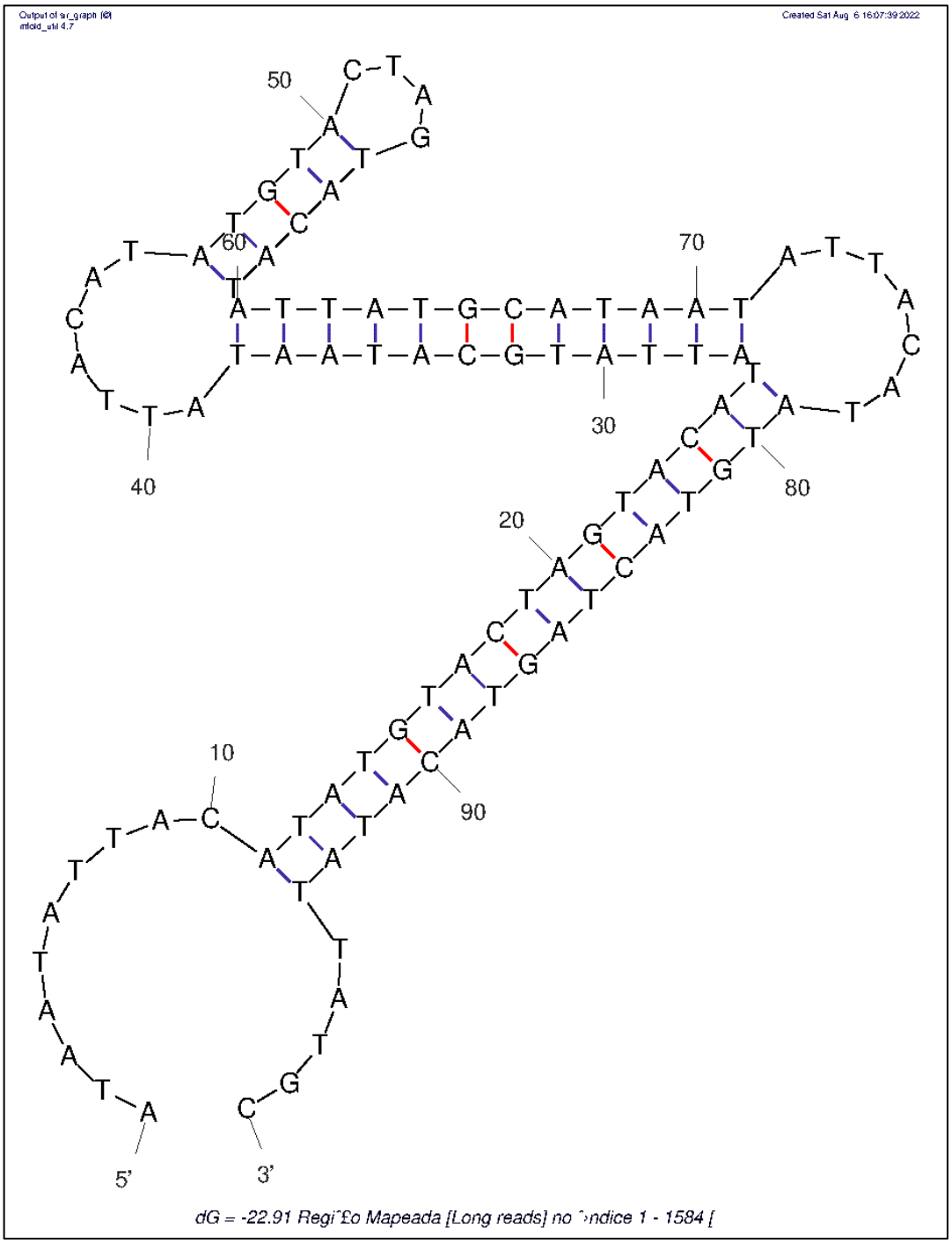
APÊNDICE I – Estrutura secundária 1 (ΔG = -13.90 kcal/mol) da repetição *in tandem* consenso da região controle completa (*Long reads*) no índice 1 – 1584 (66 pb).



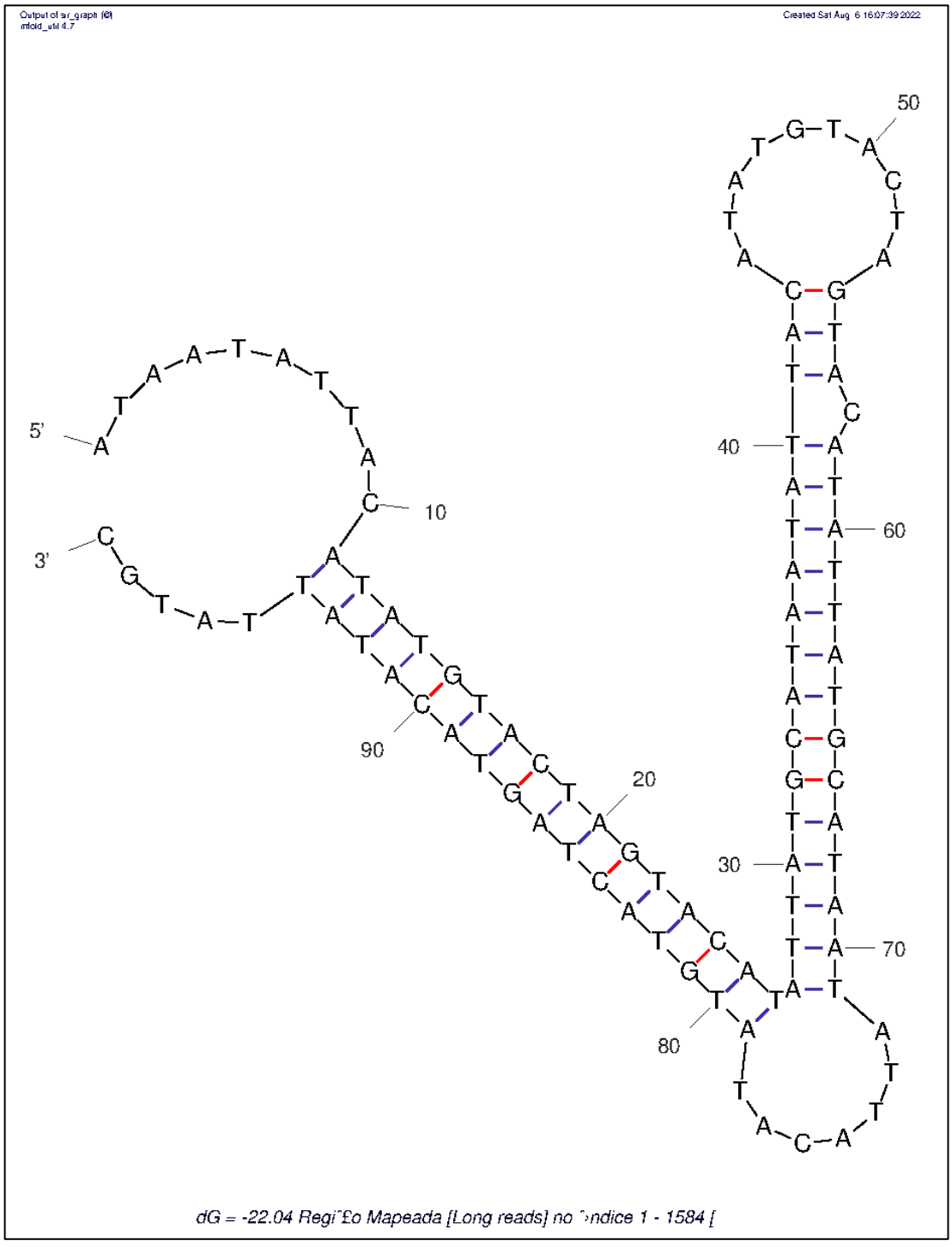
APÊNDICE J – Estrutura secundária 2 (ΔG = -13.76 kcal/mol) da repetição *in tandem* consenso da região controle completa (*Long reads*) no índice 1 – 1584 (66 pb).



APÊNDICE K – Estrutura secundária 1 (ΔG = -22.91 kcal/mol) da repetição *in tandem* consenso da região controle completa (*Long reads*) no índice 1 – 1584 (99 pb).



APÊNDICE L – Estrutura secundária 3 (ΔG = -22.04 kcal/mol) da repetição *in tandem* consenso da região controle completa (*Long reads*) no índice 1 – 1584 (99 pb).



APÊNDICE M – Estrutura secundária 3 (ΔG = -22.04 kcal/mol) da repetição *in tandem* consenso da região controle completa (*Long reads*) no índice 1 – 1584 (99 pb).