Universidade Federal
de São João del-Rei

Departamento de Ciências Naturais

Programa de Pós-Graduação em Ecologia

Desenvolvimento bioinformático de marcadores moleculares para a conservação de *Brycon orbignyanus* (Characiformes: Bryconidae), uma espécie de importância socioambiental

Rosiane de Paula Santos

São João del-Rei

2018

Rosiane de Paula Santos

Desenvolvimento bioinformático de marcadores moleculares para a conservação de *Brycon orbignyanus* (Characiformes: Bryconidae), uma espécie de importância socioambiental

Orientador: Dr. Gabriel de Menezes Yazbeck

Co-orientador: Dr. Rafael Sachetto Oliveira

Dissertação apresentada ao Programa de Pós- graduação em Ecologia da Universidade Federal de São João del-Rei, como requisito parcial à obtenção do título de mestre.

São João del-Rei

2018

Ficha catalográfica elaborada pela Divisão de Biblioteca (DIBIB)
e Núcleo de Tecnologia da Informação (NTINF) da UFSJ,
com os dados fornecidos pelo(a) autor(a)

S237d

Santos, Rosiane de Paula.
Desenvolvimento bioinformático de marcadores moleculares para a conservação de Brycon orbignyanus (Characiformes: Bryconidae), uma espécie de importância socioambiental / Rosiane de Paula Santos ; orientadora Gabriel de Menezes Yazbeck; coorientador Rafael Sachetto Oliveira. -- São João del-Rei, 2018.
57 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ecologia) -- Universidade Federal de São João del Rei, 2018.

1. recursos genéticos. 2. recursos pesqueiros. 3. genética da conservação. 4. genética de populações. 5. aquicultura. I. Yazbeck, Gabriel de Menezes, orient. II. Oliveira, Rafael Sachetto, co-orient. III. Título.

Nome: Rosiane de Paula Santos

Título: Desenvolvimento bioinformático de marcadores moleculares para a conservação de *Brycon orbignyanus* (Characiformes: Bryconidae), uma espécie de importância socioambiental

Dissertação apresentada ao Programa de Pós-graduação em Ecologia da Universidade Federal de São João del-Rei, como requisito parcial à obtenção do título de mestre.

Aprovada em 18 de julho de 2018

Banca Examinadora

Prof. Dr. Gabriel de Menezes Yazbeck (Orientador)
Universidade Federal de São João del Rei

Prof. Dr. Juliano de Carvalho Cury (membro titular)
Universidade Federal de São João del Rei

Prof. Dr. Leandro Márcio Moreira (membro titular)
Universidade Federal de Ouro Preto

Universidade Federal de São João del-Rei
Departamento de Ciências Naturais
Programa de Pós-Graduação em Ecologia (PGE)

ATA DE DEFESA

Aos dezoito dias do mês de julho do ano de 2018, às 14:00h na sala 3.08 B, no terceiro andar da biblioteca NEAD/UFSJ, realizou-se a defesa de dissertação intitulada "Desenvolvimento bioinformático de marcadores moleculares para a conservação de Brycon orbignyanus (Characiformes: Bryconidae), uma espécie de importância socioambiental", de autoria da candidata **Rosiane de Paula Santos**, aluna regular do Programa de Pós-Graduação em Ecologia, em nível de Mestrado. A Comissão Examinadora foi constituída pelos professores: Gabriel de Menezes Yazbeck (Presidente), Leandro Márcio Moreira (Membro Titular) e Juliano de Carvalho Cury (Membro Titular). Concluídos os trabalhos de apresentação e arguição, a candidata foi Aprovada pela Comissão Examinadora. Foi concedido um prazo máximo de 30 dias para a candidata efetuar as correções sugeridas pela Comissão Examinadora e apresentar o trabalho em sua redação definitiva, sob pena de não expedição do Diploma. E, para constar, foi lavrada a presente ata, que vai assinada pelos membros da Comissão.

São João del-Rei, 18/07/2018

Orientador-Presidente

1° Examinador/Membro Banca

2° Examinador/Membro Banca

( ) **Vide verso**: Em caso de alteração do título pela Comissão Examinadora

À PROPE

Certifico que o candidato cumpriu com as exigências da Comissão Examinadora e do Regimento Interno do PGE

Em ____/____/____

Juliano de Carvalho Cury
Coordenador do Programa de Pós-Graduação em Ecologia da UFSJ

Ao Setor de Expedição de Diplomas

Certifico que o candidato cumpriu com as exigências para emissão do Diploma.

Em ___/___/___

Juliano de Carvalho Cury
Coordenador do Programa de Pós-Graduação em Ecologia da UFSJ

Por sugestão da Comissão Examinadora, o novo título passa a ser:

_____

_____

_____

_____

_____

_____

_____

Financiamentos

Programa de Pós-Graduação em Ecologia, Universidade Federal de São João del-Rei, LARGE-UFSJ

Apoio e Colaborações

LCC-CENAPAD, CNPq, CEMIG, FAPEMIG, Programa de Pós Graduação em Ciência da Computação, Universidade Federal de São João del-Rei

DEDICATÓRIA

Dedico esse trabalho aos meus pais João e Tarcísia e aos meus irmãos Robson e Rosana, pelo amor incondicional e pelo apoio em todas as coisas em que realizo.

AGRADECIMENTOS

À Deus por ser a razão de tudo e pela força e saúde que permitiram a concretização deste trabalho. Em especial a Jesus no Santíssimo Sacramento;

Aos meus pais João e Tarcísia, meus irmãos e cunhada Robson, Rosana e Rosy, por todo amor, carinho, compreensão e apoio em tudo que realizo;

Aos meus padrinhos Wlad e Bia pelo incentivo, apoio e carinho. E as pequenas Maria Gabriela e Maria Cecília, por serem verdadeiros anjos em minha vida;

A todos os professores que tive desde o início de minha formação, por todo o conhecimento que me trouxe até aqui;

À professora e amiga Carla Brighenti e toda sua família, pelo importante incentivo que foi decisivo na continuação deste mestrado e por todo apoio, amizade e companheirismo;

A Wanderson e Thayane, pela amizade, apoio espiritual, pelas horas de conversa, conselhos e por toda a ternura que muito me ajuda.

À Comunidade Ponte de Misericórdia e todos os amigos que lá fiz, pela formação e apoio espiritual que tem transformado a minha vida;

Aos amigos do Movimento Shalom da Arquidiocese de Mariana, pelo apoio e pelos momentos que juntos vivemos;

Ao prof. Daniel Carvalho pelo apoio, colaboração e compreensão face a essa situação compartilhada em que vivemos;

Ao prof. Dominique Lavenier pela disponibilidade e ajuda inicial deste trabalho;

Aos amigos da república S.N, com que posso partilhar e vivenciar bons momentos;

Aos meus colegas do Laboratório de Recursos Genéticos, pela parceria, trabalho e aprendizado em conjunto, em especial a José Mauro e Raissa Dias pela disponibilidade, ajuda e bons momentos no laboratório;

A Carol Yazbeck, Melissa e Agata pela paciência e compreensão;

A Universidade Federal de São João del-Rei pela estrutura e formação;

A Gleison, Rita e Júlia, pela amizade, cuidado e acolhida em suas vidas, que fez toda a diferença na fase final desse trabalho e por que têm feito diferença na minha vida!

A todos que contribuem para o meu viver, acompanhando, rezando e partilhando momentos.

Ao meu co-orientador Rafael Sachetto, pela disponibilidade, paciência, ajuda e por todo o aprendizado a mim possibilitado;

E por fim de maneira muito especial, àquele que desde o início de minha vida acadêmica tem sido muito mais que apenas professor ou orientador: Prof. Gabriel Yazbeck, obrigada por toda a paciência, compreensão, ajuda, incentivo. Obrigada pela amizade, pelo direcionamento e orientação que me possibilitou ser quem eu sou hoje!

“Não importa o que a vida fez de você, importa o que
você fez com o que a vida fez de você!”

Jean Paul Sartre

LISTA DE ABREVIATURAS E SIGLAS

AFLP: *Amplified Fragment Length Polymorphism*

ANL: *Anonymous nuclear loci*

API: *Application programming interface*

CEMIG: Companhia Energética de Minas Gerais

CENAPAD: Centro Nacional de Processamento de Alto Desempenho

CNPq: Conselho Nacional de Desenvolvimento Científico e Tecnológico

ddRAD-seq: *Double digest RAD sequencing*

DNA: Ácido Desoxirribonucleico

EST: Marcadores de sequência expressa

FAPEMIG: Fundação de Amparo à Pesquisa de Minas Gerais

GATB: *Genome Assembly and Analysis Tool Box*

Gb: Gigabases

GBS: *Genotyping-by-sequencing*

INDEL: *Insertions/deletions*

LARGE: Laboratório de Recursos Genéticos

LCC: Laboratório de Computação Científica

mtDNA: DNA mitocondrial

MTP: Mecanismos de transposição de peixes

NGS: Sequenciamento de Nova Geração

pb: pares de bases

PCR: Reação em cadeia da Polimerase

RADSeq: *Restriction site Associated DNA Sequencing*

RFLP: *Restriction Fragment Length Polymorphism*

SNP: *Single Nucleotide Polymorphism*

QTL: *Quantitative Trait Locus*

UFMG: Universidade Federal de Minas Gerais

UFSJ: Universidade Federal de São João del-Rei

SUMÁRIO

RESUMO

A exploração intensa de recursos aquáticos tem causado diversos impactos sobre as populações naturais, especialmente em peixes. *Brycon orbignyanus* é uma espécie de piracema, praticamente extinta da natureza. Portanto, é necessária a obtenção de dados sobre sua diversidade genética, para fins de conservação. Marcadores moleculares são ferramentas que permitem a observação direta de polimorfismos de DNA e, portanto, a determinação da distribuição espacial da variação genética. Um tipo de marcador que vem sendo amplamente utilizado é o SNP, *Single Nucleotide Polymorphism*, *i*.e. uma variação populacional em uma única base nitrogenada de uma sequência de DNA, em uma posição específica do genoma. Os chamados loci nucleares anônimos ou ANL, são marcadores moleculares (tipo SNP ou variação de inserção/deleção ou INDEL) de regiões não codificantes do genoma, portanto, *a priori,* sem efeitos de seleção, o que os caracteriza como ferramentas úteis para aplicações na sistemática molecular e genética conservacionista. Este trabalho visa a identificação bioinformática de um amplo conjunto de sequências candidatas a SNP, INDEL e ANL, a partir de dados do sequenciamento do genoma de *B. orbignyanus*, realizado na plataforma Illumina HiSeq 2000. A garimpagem de SNPs foi realizada a partir dos dados com a utilização da interface de programação de aplicações GATB - *Genome Assembly and Analysis Tool Box*, utilizando-se os métodos implementados no programa DiscoSNP. Foi obtida, então, uma lista de candidatos contendo SNPs e INDELs. Posteriormente, foi criado um *script* bioinformático para filtragem e seleção dos melhores candidatos. Os SNPs foram selecionados com base no tamanho de fragmento e tamanho de borda em relação a um rascunho do genoma montado. A seleção de marcadores tipo INDELs foi feita permitindo variação na diferença de tamanho entre 1 e 100 pb. A caracterização dos ANL foi feita utilizando-se uma concatenação de passos automatizados (*pipeline*) previamente descritos, para remoção de sequências repetitivas ou altamente conservadas e regiões codificadoras. Foram produzidos um total de 17.481 possíveis candidatos a marcadores moleculares, dos quais 14.441 são do tipo SNPs e 3.040 são do tipo INDEL. Depois de filtragem usando o *script* criado aqui foram caracterizados um total de 8.631 candidatos a SNP, sendo que 377 são ANL. Foram propostos pares de oligonucleotídeos (*primers*) para a reação em cadeia da polimerase (PCR), para testes *in vitro* de amplificação da sequência contendo a variação. Esses resultados fornecem um novo e amplo recurso para o desenvolvimento rápido e econômico de dezenas de novos marcadores moleculares para *B. orbigny*anus. Estes resultados certamente irão contribuir para futuras práticas de manejo e conservação desta espécie.

Palavras-chave: recursos genéticos, recursos pesqueiros, genética da conservação, genética de populações, aquicultura

ABSTRACT

Intensive exploitation of aquatic resources has been causing many impacts on natural populations, especially on fish. *Brycon orbignyanus* is a *piracema* species, practically extinct in nature. Thus, it is important to gather data on genetic diversity, for conservation purposes in this species. Molecular markers are tools that allow the direct observation of DNA polymorphisms and, therefore, the determination of the spatial distribution of genetic variation. One type of molecular marker that has been widely used is the SNP, single-nucleotide polymorphism, *i.e*. a variation on a single DNA base in a sequence occurring at a specific position in the genome. Anonymous nuclear loci (ANL) are molecular markers (either SNPs or insertion/deletion variants or INDEL) at non-coding regions of the genome, therefore free of selection effects, which renders it as useful tools for potential applications in molecular systematics and conservation genetics. This work aims the bioinformatic identification of a wide set of SNP, INDEL and ANL candidates, from the sequencing of the *B. orbignyanus* genome, through the HiSeq 2000 platform. The mining of SNPs from the data was performed using the GATB-Genome Assembly and Analysis Tool Box API, through the methods implemented in the DiscoSNP program. A list of candidates containing SNPs and INDELs was generated. Then, a shell script was created to select the best candidates from this list. SNPs were selected based on fragment size and flanking region size relative to the assembled genome. The selection of INDELs was made allowing variation between 1 and 100 pb. The characterization of the ANL was done using a previously described pipeline for the removal of highly conserved repetitive sequences and coding sequences. A total of 17,481 possible candidate molecular markers is produced, of which 14,441 are SNPs and 3,040 are INDELs. After filtering using a script developed herein, a total of 8,631 SNPs and 377 ANL candidates were characterized. Pairs of oligonucleotides (primers) were proposed for the polymerase chain reaction (PCR) for in vitro amplification of the sequence containing the variation. These results provide a new and broad resource for the rapid and economic development of dozens of new molecular markers for *B. orbignyanus*. These results will certainly contribute to management and conservation practices in this species.

Keywords: genetic resources, fisheries resources, conservation genetics, population genetics, aquaculture

INTRODUÇÃO

1.1 Ecologia Molecular

Ecologia é a ciência que trata das relações dos organismos, uns com os outros e com o ambiente em que vivem (Haeckel, 1866) ou, mais precisamente, o estudo da distribuição e abundância dos organismos, das interações e processos que as determinam (Begon *et al.,* 2009). A ecologia consiste de uma teoria geral que contém o domínio e a descrição de um conjunto de princípios fundamentais, entre os quais está o padrão espacial e temporal dessa distribuição e abundância de organismos, incluindo suas causas e consequências (Scheiner e Willig, 2008). Essa ciência é importante para a sociedade atual em um contexto em que cada vez é necessário uma readequação da postura do ser humano para com o ambiente, tendo em vista os grandes impactos da atuação do homem sobre os ecossistemas naturais (Simpson e Christensen, 2012). Através de estudos nessa área são desenvolvidas estratégias de conservação e manejo que podem ser adotados como procedimentos racionais para a utilização de recursos naturais e manutenção da biodiversidade (Chu, 2003).

Para a manutenção da biodiversidade são necessários estudos e conhecimentos gerados a partir de áreas distintas da biologia que visam produzir informações úteis em um contexto mais amplo (Jeltsch *et al.,* 2013). Algumas disciplinas com essas características vem se popularizando, uma delas é a ecologia molecular, formada pela interdisciplinaridade entre genética, biologia molecular e ecologia (Freeland *et al.*, 2011) e a outra é a genética da conservação (Frakham *et al.*, 2010). Esses campos de estudos advêm da utilização de diferentes conceitos e metodologias para o avanço do conhecimento e a resolução de problemas ecológicos. Para tal, um passo inicial importante é o desenvolvimento de marcadores moleculares para aplicação na avaliação genética de populações, resolução de incertezas taxonômicas e avaliação das relações filogenéticas entre espécies (Avise *et al*., 2012).

1.2 Marcadores Moleculares

Marcadores moleculares são quaisquer segmentos específicos de DNA identificáveis, cuja sequência básica difere em genótipos distintos, permitindo diferenciá-los (Malone e Zimmer, 2005). São ferramentas que permitem a observação direta de polimorfismos de DNA e, portanto, a determinação da forma de distribuição espacial da variação molecular ou estrutura genética populacional. Essa informação pode ser útil na compreensão da biologia básica de espécies, delimitação de populações e avaliação de conectividade entre populações fragmentadas (Sunnucks, 2000; Schlötterer, 2004).

O desenvolvimento das técnicas de Biologia Molecular e de obtenção de informações genômicas possibilitou o surgimento de diversas ferramentas para avaliação direta da variabilidade genética (Koboldt *et al*., 2013). Dentre essas ferramentas são encontrados diversos tipos de marcadores moleculares de DNA, destacando-se o RFLP *(Restriction Fragment Length Polymorphism)*, AFLP (*Amplified Fragment Length Polymorphism*), microssatélites, SNPs (*Single Nucleotide Polymorphism)*, ANL *(Anonymous Nuclear Loci)* e marcadores mitocondriais que se baseiam na variação natural da sequência de base do DNA (Karl e Avise, 1993; Ferreira e Grattapaglia, 1996; Schlötterer, 2004; Avise, 2012) .

Os marcadores moleculares são ferramentas muito efetivas para os estudos de fluxo gênico e viabilidade de populações e ainda para quantificar os efeitos da fragmentação de habitats e da redução no tamanho populacional devido a impactos sobre as populações naturais, auxiliando assim na proposição e avaliação da eficiência de estratégias de conservação (Parker *et al.,* 1998; Selkoe e Toonen, 2006; Attard *et al.,* 2016).

Uma classe de marcadores viabilizada com técnicas de reação em cadeia da polimerase (PCR) e sequenciamento de DNA é o SNP, ou *Single Nucleotide Polymorphism*, *i.e.* uma variação em uma única base nitrogenada (Adenina, Citosina, Timina e Guanina), em uma posição específica no genoma, que varia entre indivíduos (Garvin *et al.,* 2010). Esta substituição pode ser de dois tipos, sendo que os mais comuns são as chamadas transições, onde acontece a troca de uma purina por outra purina (A por G ou vice-versa) ou de uma pirimidina por outra pirimidina (C por T ou vice-versa). Com frequência menor, ocorrem as transversões ou seja, troca de uma purina por uma pirimidina, ou vice-versa (Saenger *et al.,* 2013).

Normalmente, esses marcadores são bi-alélicos, ou seja, geralmente são encontrados apenas duas variantes em uma espécie. Esses marcadores são abundantes, podendo ser encontrados em regiões codificadoras ou com função regulatória, porém, constantemente são encontrados em espaços intergênicos, sem função determinada (Mammadov *et al.,* 2012). Além disso, possuem uma capacidade de detecção de polimorfismo em praticamente todo o genoma, isso os faz uma excelente opção para estudos de diversidade genética (Vignal *et al.,* 2012). E ainda, cada variação está presente em algum grau apreciável dentro de uma população, o que torna esses marcadores úteis em estudos de genética, ecologia e evolução (Garvin *et al.,* 2010).

Outra classe de marcadores moleculares são conhecidos como *Anonymous Nuclear Loci* (ANL). Esses são marcadores de regiões não codificantes do genoma que não estão *a priori* sob efeitos de seleção. Isso os qualifica como uma ferramenta útil para aplicações potenciais na sistemática molecular e genética conservacionista (Karl e Avise, 1993). Os ANL são encontrados em regiões "anônimas" do DNA e podem ser isolados e usados para construir oligonucleotídeos iniciadores de PCR (ou *primers*) para serem utilizados e amplificar-se regiões homólogas (Silva *et al.,* 2011).

Esses marcadores são encontrados em grande quantidade em animais, dado que uma representativa porcentagem do genoma de eucariotos compreende regiões não codificadoras (Thomson *et al.,* 2010). Por estarem nessas regiões, esses marcadores são suscetíveis a estarem sob efeitos de evolução de forma independentes, sendo então passíveis de acumularem um elevado número de mutações e proporcionarem, portanto, marcadores de taxa de mutação rápida. Estes marcadores são úteis em estudos de genética populacional e, ainda, filogenia e filogeografia (Bertozzi *et al.,* 2012; Dowie *et al.,* 2017).

Outra fonte abundante de marcadores genéticos que estão amplamente espalhados pelo genoma, embora não tão comuns quanto os SNPs, são conhecidos como INDELs (Vali *et al.,* 2008). Estes são marcadores polimórficos, amplamente espalhados pelo genoma, baseados em fragmentos de inserção e deleção de sequências do DNA (Bhattramakki *et al.,* 2002). INDELs oferecem marcadores moleculares cujo resultado pode ser revelado em procedimentos simples baseados em separação de fragmentos, como a eletroforese e, diferentemente dos SNPs, não é necessária a realização de reação de sequenciamento de DNA (Pereira *et al,* 2010).

INDELs são úteis para uma variedade de fins, como exemplo, abordagem da estrutura genética de populações humanas (Tishkoff *et al.,* 2009), inferindo proporções de ancestralidade de indivíduos e de populações (Yang *et al.,* 2005) e na identificação de espécies (Mahadani e Ghosh, 2014; Pereira *et al.*, 2010).

1.3 Sequenciamento de Nova Geração (NGS)

O sequenciamento de DNA, metodologia que determina a sequência exata de bases em um fragmento do genoma, foi tradicionalmente realizado pelo chamado método de Sanger. Essa foi a opção utilizada para sequenciamento de ácidos nucleicos por três décadas consecutivas desde sua invenção em 1977 (Heather e Chain, 2016). Este período serviu de base para a era genômica, caracterizada por avanços técnicos que permitiram análises de segmentos de DNA e posteriormente o sequenciamento de genomas completos de diversos organismos (Behjati e Tarpey, 2013).

O conjunto de metodologias conhecidas como Sequenciamento de Nova Geração (NGS) surgiu com novas estratégias de sequenciamento e desde então têm promovido uma verdadeira revolução na genética (Mardis, 2008; Metzker, 2010). Diferentes plataformas estão hoje disponíveis como Ion Torrent, Pacific Biosystems e a antiga Solexa, que hoje foi incorporada na marca Illumina (Liu *et al.,* 2012). Essas novas plataformas de sequenciamento estão se tornando amplamente acessíveis, assim reduzindo o custo do sequenciamento que agora pode ser feito com maior eficiência e rapidez (Liu *et al.,* 2012).

Estes avanços nas ferramentas de sequenciamento de DNA permitiram melhorar a eficiência e a velocidade de produção de dados genômicos (Liu *et al.,* 2012) ao mesmo tempo em que os custos do processo de sequenciamento foram dramaticamente reduzidos (Buermans e Dunnen., 2014, Van Nimwegen *et al.*, 2016). O NGS trouxe aos pesquisadores uma ampla variedade de aplicações e estudos sobre sistemas biológicos e de questões complexas que exigem uma profundidade de informação, além da capacidade das tecnologias tradicionais de sequenciamento de DNA (Ekblom e Galindo, 2011). O desenvolvimento dessas técnicas permitiu a criação de uma enorme quantidade de dados e possibilitou o avanço do conhecimento de genomas, promovendo uma verdadeira revolução nos estudos de genética e, consequentemente, novas aplicabilidades em conservação (Ekblom e

Galindo, 2011). Essas técnicas aumentam consideravelmente a produção de informações em diversos estudos com organismos não-modelos.

Anteriormente, estudos genéticos de população em escala de genoma só eram acessíveis a organismos modelos, por serem bem financiados e compreendidos. No entanto, o sequenciamento de DNA associado à restrição, como o RADSeq, permitiu a descoberta e genotipagem a custos razoáveis de milhares de marcadores genéticos para qualquer espécie, incluindo organismos não-modelo (Davey e Blaxter, 2010; Andrews *et al.,* 2016). Uma técnica relacionada, conhecida como ddRAD-seq (Peterson *et al.,* 2012), sequenciamento de DNA de dupla restrição, também permite uma abordagem com relativo baixo custo para o desenvolvimento/análise populacional acoplada de numerosos marcadores SNPs. Outros avanços baseados em dados de NGS são as abordagens de *genotyping-by-sequencing* (GBS), que permitem que uma fração direcionada do genoma seja sequenciada por NGS em múltiplos indivíduos simultaneamente, mesmo em espécies com pouca ou nenhuma informação genômica prévia e grandes genomas com a utilização de enzimas de restrição ou sondas de captura ou sequenciando o transcriptoma (Davey *et al.,* 2011).

É possível obter resultados preliminares para o desenvolvimento rápido e econômico de diversos tipos de marcadores moleculares, aproveitando-se os dados genômicos obtidos em um experimento de NGS (Ekblom e Galindo, 2011). A utilização desses marcadores permite análises genéticas em larga escala (Poke *et al,* 2005), visto que revelam a variação entre os indivíduos através da detecção de polimorfismos em várias posições distintas do genoma, permitindo estudos da diversidade genética e análises de diferenciação populacional.

Apesar de toda facilidade criada pelo NGS, a análise eficiente de dados de DNA de natureza massiva (*e.g.* bilhões de bases ou mesmo bilhões de sequências em um único conjunto de dados) depende da aplicação de técnicas informáticas para armazenamento, transferência, manipulação e análise, o que geralmente depende de avançadas plataformas de computação científica de alto desempenho, como *clusters* e *workstations*.

1.4 Métodos Bioinformáticos

A bioinformática teve sua ascensão marcada com avanços na área de biologia molecular, advindo do desenvolvimento das técnicas do processo de sequenciamento de ácidos nucleicos, especialmente dos projetos genomas, que geram volumes intensos de informações obtidas a partir do sequenciamento de fragmentos (Polanski e Kimmel, 2007).

A bioinformática pode ser definida como sendo a aplicação de soluções computacionais a problemas biológicos (Hogweg, 2011). Com a rápida expansão e desenvolvimento das técnicas de sequenciamento de DNA, o volume de sequências biológicas aumentou consideravelmente (Chaitankar *et al.*, 2016). Isso evidenciou uma demanda por recursos computacionais cada vez mais poderosos para a manipulação e processamento desse elevado volume de dados gerados, dando origem à bioinformática contemporânea (Dai *et al.,* 2012). Isso envolve o desenvolvimento de bancos de dados e de algoritmos utilizados para o processamento deste conjunto de informações, produzindo conhecimento e informação útil para a pesquisa biológica (Lee *et al.*, 2012).

A API ou interface de programação de aplicação GATB - *Genome Assembly and Analysis Tool Box* (disponível em http://gatb.inria.fr) (Drezen *et al.,* 2014) é uma suíte de módulos, bibliotecas e programas para a personalização de conjuntos concatenados de etapas (*pipeline* informático) e de programas para análise de dados NGS. Essa plataforma armazena uma biblioteca otimizada de algoritmos de código aberto para análise de dados NGS em computadores comuns, com pouca memória disponível (um dos maiores gargalos da análise NGS).

O programa DiscoSNP, incluído no GATB, usa dados brutos de sequenciamento NGS, geralmente caracterizados por sequências relativamente curtas, armazenadas no formato FASTQ (que alia leituras curtas ou *reads* da sequência com um valor de qualidade/confiabilidade na escala PHRED) para encontrar os SNPs sem uma montagem genômica *de novo* completa ou sem um genoma de referência. O DiscoSNP visa localizar possíveis SNPs que estejam isolados de outras possíveis variantes, por um mínimo de k nucleotídeos. (Uricaru *et al.,* 2014). Uma abordagem de micro montagem, utilizando o programa Minia, (programa encontrado na API do GATB para montagem de leituras pareadas) gera um arquivo fasta contendo cada SNP identificado ao contig (*i.e.* sequência local contígua montada) que ele pertence, representado por um par de sequências que

diferem apenas no sítio polimórfico. O rótulo da sequência fornece informações sobre cobertura e qualidade média da leitura (Uricaru *et al.,* 2014).

O algoritmo implementa grafos de *de Bruijn* (sistemas representados por vértices e arestas direcionadas - Compeau *et al.,* 2011), a partir de um ou mais conjuntos de dados para localizar padrões específicos de SNP esperados nos grafos como as chamadas "bolhas" (Uricaru *et al.*, 2014). No caso de um conjunto de dados, oriundo de um único indivíduo diplóide (*i.e.* com dois complementos homólogos do genoma), os candidatos a SNPs são garimpados a partir do encontro fortuito de posições em heterozigose no espécime sequenciado. Isso significa que os mesmos devem ser posteriormente testados empiricamente em uma amostra populacional para validação e quantificação do polimorfismo, que geralmente respeita valores limites de frequência do alelo de menor ocorrência (*e.g.* ≧ 5%) para serem considerados marcadores SNPs.

1.5 Ecologia de peixes de piracema

A região neotropical possui a maior diversidade de peixes de água doce do mundo (Albert e Reis, 2011), sendo que a América do Sul possui uma rica diversidade de peixes de água doce e marinhos com uma riqueza estimada em mais de 9.100 espécies (Reis *et al.*, 2016).

Um número representativo de peixes neotropicais são de espécies migratórias, denominadas de peixes de piracema. Tais peixes percorrem longas distâncias com finalidade reprodutiva, ao longo do seu ciclo de vida (Zaniboni-Filho *et al.,* 2017). Essa migração é de extrema importância para o sucesso reprodutivo dessas espécies, pois proporciona a reunião de um grande número de indivíduos aptos para reprodução em um local ideal para oviposição e posterior sobrevivência dos alevinos (alto nível de oxigenação, grande disponibilidade de recursos alimentares e baixo risco de predação devido à maior turgidez da água durante inundações) (Carolsfeld, 2004).

O sucesso reprodutivo é dependente da variação do regime hidrológico iniciada por alguns fatores tais como: cheias prolongadas acompanhadas da subida do nível da água e elevação nas temperaturas (Neiff, 1990; Agostinho *et al.,* 2002). A migração de peixes de água doce geralmente ocorre da seguinte forma: durante a estação chuvosa, os indivíduos adultos movem-se em direção a montante do rio e após o período de reprodução os peixes adultos retornam a jusante (Petrere, 1985).

Ovos fecundados darão origem a larvas que serão conduzidos passivamente para as planícies de inundação, que se formam ao longo das margens dos rios e são consideradas “berçários” para espécies migratórias, pois apresentam condições extremamente favoráveis para o desenvolvimento das larvas, tais como: alto nível de oxigenação, grande disponibilidade de recursos naturais e baixo risco de predação devido a turbidez da água (Lowe-Mcconnell., 1999).

No entanto, esse processo pode ser diretamente afetado devido a construção de usinas hidrelétricas, pois muitos rios vêm sendo represados transformando-se, em uma sucessão de reservatórios e causando o bloqueio de rotas migratórias de peixes, o que impacta diretamente peixes de piracema (Agostinho *et al.,* 2008). Uma das consequências do barramento é o isolamento de populações naturais, o que pode levar a efeitos como perda de variabilidade genética e, consequentemente, a uma redução do valor adaptativo médio populacional, aumentando o risco de extinção local de espécies de peixes (Barletta, 2010). Para evitar isso, estruturas e mecanismos de transposição de peixes (MTPs) têm às vezes sido criados em locais onde há represamentos, com intuito de viabilizar a ocorrência do processo reprodutivo, facilitando a passagem de indivíduos ou cardumes. Em teoria, estes MTPs ajudam a reverter impactos ambientais causados pelas barragens. No entanto, discussão corrente na literatura desafia essa noção, argumentando que a falta de habitats apropriados para reprodução à montante pode, na realidade, constituir uma armadilha ecológica para peixes migratórios (Pelicice e Agostinho, 2008; Kochalski, *et al., 2*018).

Outra forma proposta para mitigação de impactos da geração hidrelétrica são as estações de piscicultura para a produção de estoques de espécies nativas para a realização de atividades de estocagem chamadas peixamento (Sugunan, 1997). A princípio, os programas de peixamento eram realizados sem estudo prévio sobre a ictiofauna nativa da região, o que culminou na liberação de espécies não-nativas ou exóticas. Estes programas só se modificaram a partir da década de 80, período no qual a liberação de espécies nativas para manutenção das populações selvagens se tornou objetivo principal destes programas (Agostinho *et al*., 2008; Lopes e Bedore, 2008). Esta tem sido uma das estratégias ambientais mais usadas por hidrelétricas (Agostinho *et al.,* 2007). Esta atividade quando realizada de forma devida, é capaz de auxiliar no restabelecimento de populações selvagens ameaçadas. No entanto, tal prática ainda é motivo de constantes debates, devido a inúmeras tentativas mal

sucedidas e impactos negativos decorrentes desta atividade (Agostinho *et al.,* 2007; Agostinho *et al.,* 2010).

Esforços de conservação de peixes e atividades mitigadoras de impactos ambientais são iniciativas de interesse social, para fins de sustentabilidade das atividades de pesca e produção de peixes. Isso é particularmente importante, visto o alto potencial de geração de trabalho e renda, que podem contribuir como importantes mecanismos de desenvolvimento econômico-social de segmentos da população de baixa renda, que formam a maior parcela entre os trabalhadores dessas atividades (Begossi, 1998).

A aquicultura de espécies ameaçadas pode contribuir com a diminuição da pressão de pesca em remanescentes naturais. Além disso, atividades de pesca e produção de peixes ainda contribuem fortemente para fixação das populações humanas no campo, a disseminação da consciência ambiental, além de sua integração com medidas de preservação de bacias hidrográficas e seus recursos naturais através de atividades de recuperação e proteção ambiental (Pauly *et al.,* 2002; Craig, 2016).

1.6 *Brycon orbignyanus* (Valenciennes, 1850) - Bryconidae

A ordem Characiformes contém cerca de 2.000 espécies, distribuídas em 23 famílias, sendo 19 exclusivamente neotropicais (Oliveira *et al.,* 2011; Eschmeyer & Fong, 2017). A família Bryconidae, constituída das subfamília Salminae e Bryconinae, compreende um grupo de peixes de grande importância para a pesca e para a aquacultura (Reis *et al.,* 2003). O gênero *Brycon*, incluído dentro da subfamília Bryconinae, é um dos gêneros mais numerosos com 42 espécies descritas (Eschmeyer e Fong, 2017). Espécies desse gênero distribuem-se desde o sul do México até o Panamá, ao longo das bacias hidrográficas da América do Sul trans-Andina, nas principais bacias hidrográficas da América do Sul cis-Andina e na maioria dos sistemas costeiros do Caribe e Atlântico (Lima, 2003; Abe *et al.,* 2014).

O Brasil possui uma relevante diversidade de peixes de água doce devido à presença de um clima favorável e da existência de diversos sistemas hidrográficos (Nogueira *et al.,* 2010). Apesar disso, muitas espécies vêm sofrendo ameaça devido a diferentes fatores que causam alterações no ecossistema. Uma delas é *Brycon orbignyanus*, popularmente conhecida como piracanjuba, que ocorre na bacia hidrográfica do Prata (bacias dos rios Paraná, Paraguai e Uruguai, a segunda maior

da região Neotropical, e compreende um grupo de peixes de notável relevância para a pesca e para a aquacultura (Reis *et al.,* 2003; Oliveira *et al.,* 2017).

Brycon orbignyanus (Valenciennes 1850) (Figura 1.6) é descrita taxonomicamente como sendo da classe Actinopterigii, Infraclasse Teleostei, Ordem Characiformes, família Bryconidae, subfamília Bryconinae e gênero Brycon (Lauder e Lien 1983, Britski *et al.,* 1988, Oliveira et al., 2011, Abe *et al.,* 2014)

Fig 1.6: *Brycon orbignyanus* (Valenciennes, 1850). **Fonte:** Oliveira *et al., 2017*

Diversos são os fatores que causam a mudança no status de conservação de uma espécie de peixe, tais como: a perda e degradação do habitat causados principalmente por mudanças no uso da terra, degradação de matas ciliares, poluição, represamento hidrelétrico, introdução de espécies, transposição da água para irrigação, urbanização, sedimentação e sobrepesca (Reis *et al.,* 2016). *B. orbignyanus* anteriormente era um dos peixes mais abundantes para pesca e hoje se encontra com uma distribuição distinta do que era décadas atrás (Figura 1.6.2). Essa espécie consta no livro vermelho da fauna ameaçada de extinção (Rosa e Lima, 2008). Atualmente, as populações nativas desta espécie encontram-se restritas a poucos rios e pequenos afluentes onde as condições ambientais permanecem preservadas (Oliveira *et al.,* 2017).

Essa espécie de piracema possui uma estratégia reprodutiva periódica que ocorre apenas durante a cheia após realizar a migração, além disso, realiza desova total e sazonal, primeira maturação tardia, alta fecundidade, ovos pequenos e ausência de cuidado parental (Rosa e Lima, 2008). Todos esses fatores, considerados em conjunto, implicam em uma situação de ameaça ainda mais alta a essa espécie, sendo imprescindível o desenvolvimento de estudos aprofundados e ações para a conservação dos estoques destes peixes.

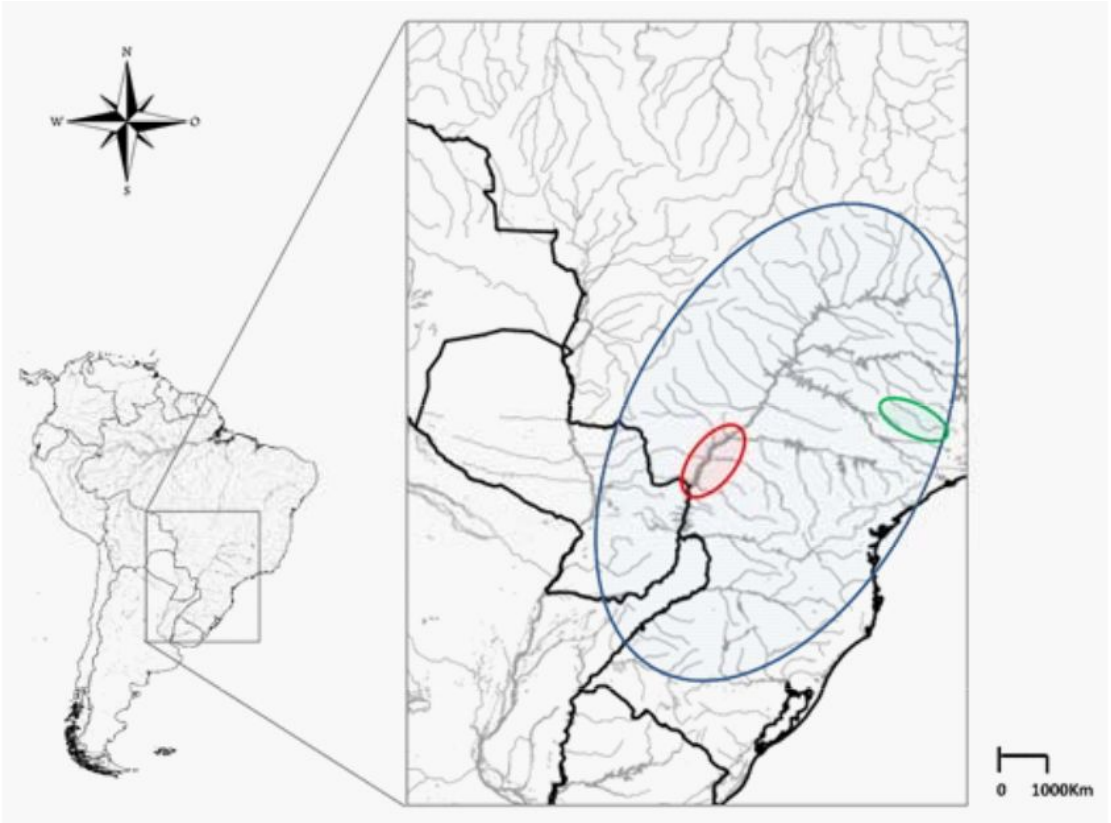

Fig 1.6.2: Mapa exibindo a distribuição de *Brycon orbignyanys* na bacia do Prata. O círculo vermelho mostra o trecho livre do rio Paraná. O círculo verde mostra o rio Mogi-Guaçu, onde a ocorrência desta espécie é atualmente rara. **Fonte:** Oliveira *et al., 2017*

1.7 Desenvolvimento de marcadores moleculares específicos para *B. orbignyanus*

Diante do status de conservação dessa espécie, se torna necessário obter dados sobre a diversidade genética, em diferentes classes de marcadores moleculares, nos estoques naturais e domesticados, que possam auxiliar gestores, profissionais e técnicos sobre procedimentos de manejo e reprodução de indivíduos e populações de interesse. Para isso, devem ser realizados estudos que permitam acesso direto ao genótipo de indivíduos de peixes, como o desenvolvimento de marcadores moleculares de DNA (Choupina e Martins, 2014; Pradhan *et al.*, 2016).

Apesar da importância de marcadores moleculares para a conservação, apenas recentemente marcadores específicos foram descritos para essa espécie: uma classe de marcadores amplamente utilizados em estudos de diversidade genética são marcadores do tipo microssatélites e, até então, a utilização desses marcadores era baseada apenas em marcadores heterólogos (isolados em espécies próximas), utilizando-se loci originalmente descritos em espécies como *B. opalinus*

(Barroso *et al.,* 2003) e *B. hilarii* (Sanches e Galetti Jr., 2006). Um exemplo foi a avaliação da estrutura genética das populações de vida livre na qual foi constatado, através da análise de quatro microssatélites, o estado crítico dessa espécie e a necessidade de criação de zonas prioritárias de conservação (Ashikaga *et al.*, 2015). Apenas em 2016 foi descrito o primeiro conjunto de 29 marcadores de microssatélites polimórficos para essa espécie (Arias *et al.,* 2016), a partir dos mesmos dados NGS brutos que subsidiam o presente trabalho. Nesse mesmo ano foi também publicado o genoma mitocondrial completo para essa espécie (Siqueira *et al.,* 2016). Recentemente, foram publicados um painel genômico de potenciais microssatélites visando o desenvolvimento rápido e barato destes marcadores, juntamente com um rascunho montado de seu genoma (Yazbeck *et al.*, 2018), além de cinco novos marcadores polimórficos validados de técnicas clássicas de clonagem molecular (Souza *et al.*, 2018).

Tradicionalmente, marcadores moleculares eram obtidos por meio destas técnicas de clonagem molecular, que demandam muito tempo e produzem um número reduzido de marcadores (Zane *et al.,* 2002). No entanto, com o advento de NGS e sua produção volumosa, o foco das técnicas em desenvolvimento de marcadores moleculares se moveu para a etapa de análise, que é manipulada com auxílio de técnicas e métodos da bioinformática, área emergente e em constante desenvolvimento (Ekblom e Galindo, 2011).

JUSTIFICATIVA

Fatores, tais como, destruição das florestas ciliares, represamentos, poluição e introdução de espécies, têm contribuído para o status de conservação da piracanjuba, hoje como sendo ameaçada de extinção e apenas com pequenas populações nativas em rios e afluentes distantes de centros urbanos (Oliveira *et al.,* 2017). Estudos de genética populacional para esta espécie, ainda que realizado com poucos marcadores heterólogos, mostraram que a população encontra-se estruturada em diferentes subpopulações na bacia do rio da Prata, sendo que existem alguns estoques naturais preservados com condições ambientais adequadas e que ainda retêm alguma variabilidade genética, sendo essas áreas sugeridas como unidades de manejo independentes (Ashikaga *et al.,* 2015).

Estudos de esforços comparáveis para a descrição de marcadores microssatélites a partir de dados NGS nas espécies de *B. orbignyanus e S. brasiliensis* sugerem baixa variabilidade genética, uma vez que para *S. brasiliensis*

foram validados 47 marcadores microssatélites polimórficos (Cao *et al.,* 2016), enquanto que para *B. orbignyanus* foram validados apenas 29 (Arias *et al.,* 2016).

Além disso, comparações realizadas entre o mitogenoma completo obtido a partir do conjunto de dados que subsidiaram este trabalho (GenBank KY825192.1), com o primeiro mtDNA publicado para essa espécie (Siqueira *et al.,* 2016) mostra virtualmente pouca variação, enquanto que a mesma comparação direta para *S. brasiliensis* foi possível encontrar um número de diferenças muito maior (KY825190.1 em Brandão-Dia *et al.,* 2016).

Finalmente, análises visando a otimização de amplificações heterólogas com marcadores microssatélites em diferentes plantéis de *B. orbignyanus* (e.g. Carmo *et al.*, 2015 e Lopera-Barrero *et al.,* 2010*)* mostraram diferentes combinações entre marcadores monomórficos/polimórficos, o que sugere fixação de alelos em loci alternativos em diferentes estoques. Isso pode ser reflexo de repetidos eventos de gargalo genético populacional ou efeito fundador na formação de estoques de procriação.

Tudo isso, em conjunto, aponta para uma provável depauperação genética em *B. orbignyanus*, o que evidencia a importância de esforços intensivos de pesquisa genética nesta espécie em particular. Sendo assim, é de grande relevância o uso de novas técnicas para o desenvolvimento de marcadores genéticos inéditos que visam subsidiar estudos genéticos nesta espécie, que possam auxiliar no planejamento e avaliação de eficiência de atividades de mitigação de impactos ambientais e em programas de conservação e produção dessa espécie.

O presente trabalho visa o aproveitamento de um conjunto de dados NGS, gerados no Laboratório de Recursos Genéticos (LARGE-UFSJ) para a espécies *B. orbignyanus,* para o desenvolvimento bioinformático de uma ampla coleção de candidatos a marcadores moleculares inéditos, para alavancar os estudos e práticas de manejo e aquicultura para essa importante espécie de interesse socioambiental.

OBJETIVOS

Geral

Caracterizar regiões genômicas candidatas ao desenvolvimento de marcadores moleculares inéditos para *Brycon orbignyanus*, úteis para aplicação em estudos, manejo ambiental e aquicultura.

Específicos

- Identificar potenciais marcadores moleculares tipo polimorfismos de um único nucleotídeo, SNP;
- Identificar potenciais marcadores moleculares tipo inserção/deleção, INDEL;
- Identificar dentre estes, potenciais marcadores moleculares tipo loci nucleares anônimos, ANL;
- Propor listas de *primers* para ensaios empíricos de reação em cadeia da polimerase (PCR), com os potenciais candidatos;
- Disponibilização pública de listas de candidatos a marcadores moleculares em base de dados científicas, para o desenvolvimento rápido e econômico de novos marcadores moleculares em *B. orbignyanus* e espécies próximas.

METODOLOGIA

3.1 Descrição dos dados NGS

As análises realizadas neste trabalho foram feitas a partir do sequenciamento do genoma de um espécime de *Brycon orbignyanus* efetuado utilizando-se a tecnologia NGS HiSeq 2000 (Illumina, San Diego, EUA), com aproximadamente 10x de cobertura, *i.e,* a média de quantas vezes cada base de DNA foi sequenciada. Um total de 16.04 Gb de dados filtrados foram obtidos, representando cerca de 178.2 milhões de leituras pareadas (com comprimento de 90 bases), através do sequenciamento a partir de uma biblioteca genômica de fragmentos de 500 bp. O conjunto de dados inicial e o rascunho do genoma montado utilizados nesta proposta podem ser recuperados em https://www.ncbi.nlm.nih.gov/sra/SRX335044 e doi:10.6084/m9.figshare.5661802, respectivamente. Ambos conjuntos de dados encontram-se detalhadamente descritos em Yazbeck *et al.* (2018).

3.2 Caracterização de SNPs e INDELS

As análises foram executadas em computador com processador Intel core i5, com 8 gigabytes de memória de acesso aleatório (RAM) e sistema operacional Linux Ubuntu (versão 16.04 LTS). O desenvolvimento de candidatos SNPs foi realizado com a utilização da plataforma GATB (disponível em http://gatb.inria.fr) (Drezen *et al.*, 2014). Foi utilizado o pacote DiscoSNP com parâmetros *default* a partir das leituras pareadas em formato FASTQ, produzidas pelo processo de sequenciamento. A seleção de INDELs foi feita com o mesmo *pipeline* DiscoSNP, permitindo-se inserções e deleções entre 1 e 100 pb. Foi gerada então uma lista de candidatos contendo SNPs e INDELs. A figura 3.2 delineia os passos básicos de funcionamento do DiscoSNP.

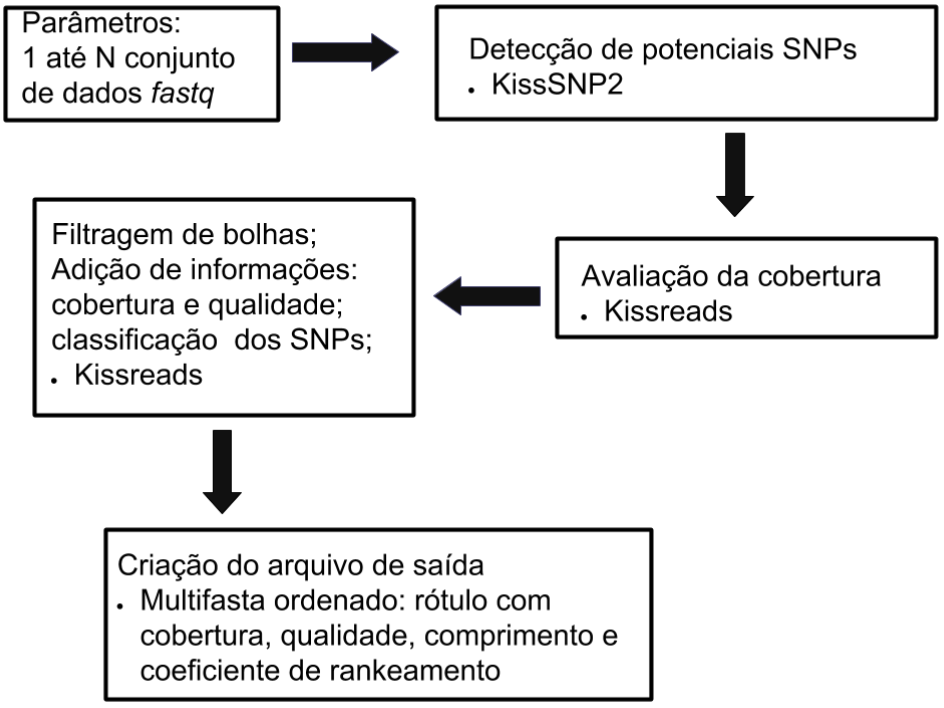

Fig 3.2: Fluxograma de funcionamento do DiscoSNP.

3.3 Caracterização de ANL

As análises foram realizadas remotamente no *cluster* de computadores do LCC-CENAPAD da Universidade Federal de Minas Gerais-UFMG a partir do rascunho do genoma montado para *B. orbignyanus* (Yazbeck *et al.*, 2018). Foi

realizada a filtragem e remoção das sequências não-alvo, tais como: sequências repetitivas, altamente conservadas e as sequências codificadoras. O software RepeatScout (Price *et al.*, 2005) foi utilizado para a criação de uma biblioteca contendo sequências não-alvo, incluindo o genoma mitocondrial disponível da espécie. Esta biblioteca foi utilizada pelo software RepeatMasker (Smit *et al.,* 2010) para mascarar essas sequências altamente repetitivas no arquivo FASTA original para que estas pudessem ser removidas com a utilização do *script* SeqClean-PERL (disponível em: https://sourceforge.net/projects/seqclean/). As sequências resultantes foram confrontadas com o GenBank através do BLASTN (Boratyn *et al.,* 2013). Para isso, o programa makeblastdb (parte do pacote BLAST) foi utilizado para a criação de um banco de dados local contendo sequências altamente conservadas, regiões codificadoras já definidas, RNA mensageiro e marcadores de sequência expressa (EST). Na execução deste *pipeline*, sequências com valor esperado (*e-value*) $<10^{-4}$ foram removidas do arquivo FASTA. Foi utilizado um *script* PERL adaptado (http://www.samuseum.sa.gov.au/about/staff/dr-terry-bertozzi) para automatizar a análise (Bertozzi *et al.,* 2012).

Após a preparação desta biblioteca, a lista de candidatos a SNPs gerada pelo DiscoSNP foi confrontada com este resultado para a caracterização dos ANL. Para isso, com o *script* search_scaffs.sh (Apêndice B) foi então criada uma lista com todos os contigs/scaffolds que possuem candidatos a ANL. Posteriormente, com o *script* filters_anl.sh (Apêndice E) foram identificados e listados os contigs/scaffolds com SNP que possuem candidatos únicos a ANL. O fluxograma de funcionamento do *pipeline* pode ser visto na figura 3.3.

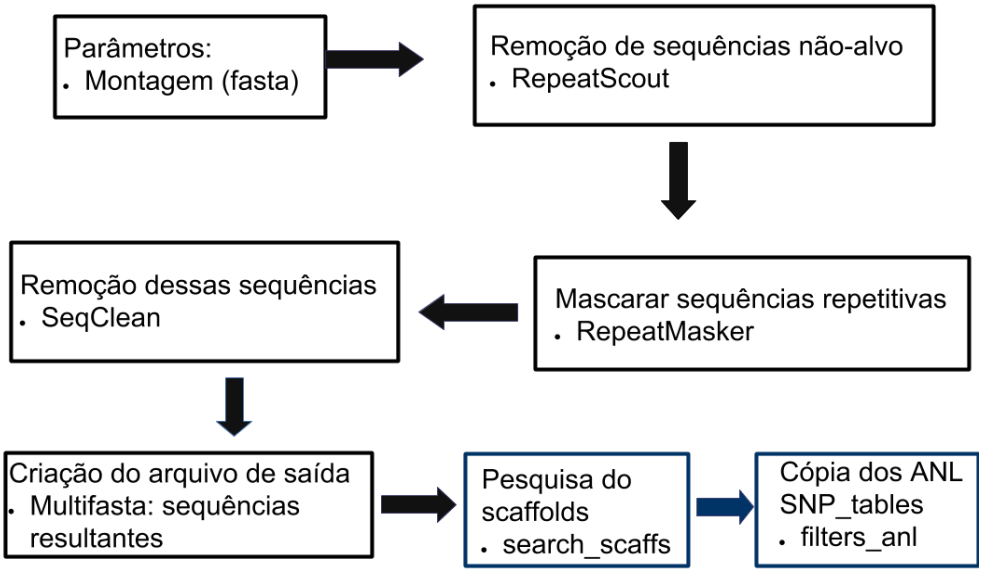

Figura 3.3: Fluxograma de funcionamento do *pipeline* para geração dos ANL. Em preto os passo realizados pelo *pipeline* anomaker.pl (Bertozzi *et al.,* 2012) e em azul os passos adicionados por este trabalho.

3.4 Filtragem dos dados

Inicialmente foi feita a inspeção manual de uma amostra dos resultados obtidos. Para isso, foi feita a caracterização dos 20 primeiros potenciais SNPs da lista gerada pelo DiscoSNP. Cada possível SNP foi buscado no banco de dados criado com o programa makeblastdb e, após encontrar sua localização no contig/scaffold, este foi copiado para um arquivo de texto (FASTA) juntamente com as duas sequências alternativas apresentando a posição variante do candidato a SNP. Esse arquivo foi aberto na suíte bioinformática UGENE (Okonechnikov *et al.,* 2012) para a realização do alinhamento com o programa Muscle (Edgar, 2004). Em seguida, foi feita a busca e inspeção visual da profundidade de cobertura das regiões de interesse, no contig/scaffold, no arquivo binário de mapeamento de alinhamento de sequência, formato BAM (disponível em: https://www.ncbi.nlm.nih.gov/sra/SRX3427716) do rascunho do genoma montado em Yazbeck et al. (2018), com auxílio do programa Tablet (Milne *et al.*, 2012), para verificação da posição do SNP.

Posteriormente, foi criado um script em *shell* para selecionar os melhores candidatos dessa lista, considerando apenas os marcadores SNPs e ignorando os marcadores INDELs. O fluxograma de funcionamento do script pode ser visto na Figura 3.1.

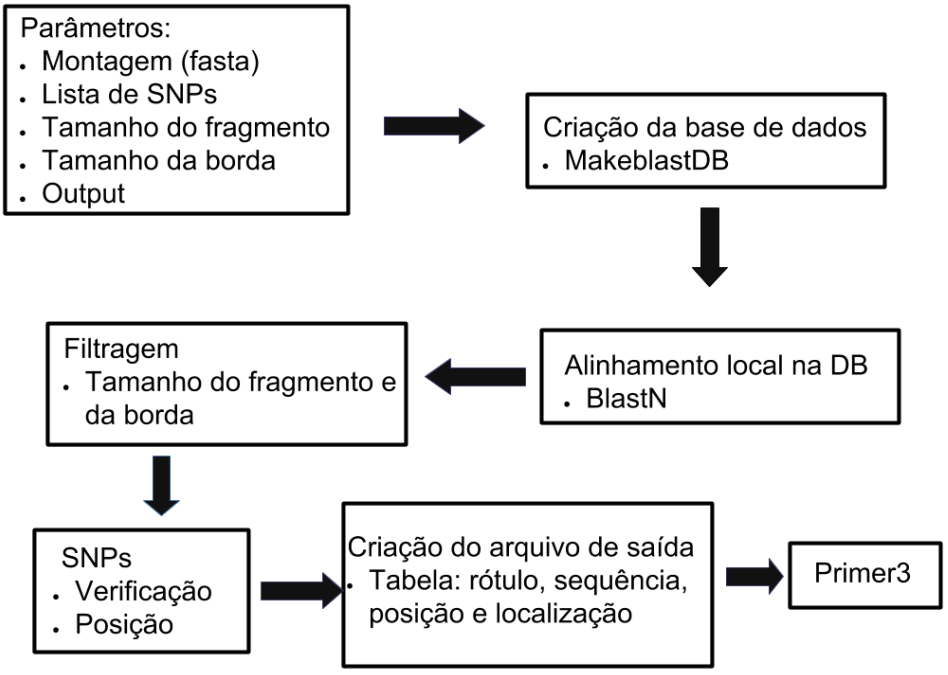

Figura 3.4: Fluxograma de funcionamento do *script* para filtragem dos SNPs.

Como parâmetros o *script* recebe o rascunho do genoma montado para ser utilizado como referência para localização dos candidatos a SNPs. Além disso, são passados como parâmetros, a lista de candidatos a SNPs, o tamanho do fragmento e borda a serem considerados. Esses dois últimos parâmetros são utilizados para que seja possível realizar o desenho dos *primers* para um fragmento viável de ser amplificado. O primeiro passo é a criação de uma base de dados utilizando o rascunho do genoma, realizado com a execução do programa makeblastdb. Em seguida, para cada candidato a SNP, foi feito um alinhamento local, com blastn, com o rascunho do genoma montado. O resultado deste alinhamento foi salvo em um arquivo separado que foi utilizado para a realização da filtragem do tamanho do fragmento e da borda, verificação da existência de um SNP e em caso positivo, é feita a verificação da posição deste. Todas essas informações são escritas em um arquivo e para cada sequência considerada aceita é gerado um arquivo de configuração que será utilizado para o desenho dos *primers.* O desenho de *primers* foi realizado com a utilização do programa 'Primer3' (Untergasser *et al.,* 2012), especificando o alvo de amplificação como sendo a possível posição SNP e mais duas bases (uma de cada lado), o tamanho do produto entre 100 e 400 pb e demais

configurações *default.* Após o processo de geração dos *primers,* foi utilizado o *script* filter_primers.sh (Apêndice C) para selecionar o primeiro par de primers dentre os possíveis pares sugeridos pelo Primer3. Em seguida foi utilizado o *script* complete_table.py (Apêndice D) que completa uma tabela incluindo as sequências dos *primers forward* e *reverse* e a classificação como sendo SNP ou ANL.

RESULTADOS

4.1 *Scripts* bioinformáticos

O desenvolvimento de *scripts* informáticos para a efetivação dos objetivos deste trabalho são considerados resultados em si. Foram produzidos 5 *scripts* para o refinamento da lista gerada pelo DiscoSNP. O *script* execute_pipeline (Apêndice A-10.6084/m9.figshare.6813053) realiza a filtragem dos SNPs e produz uma tabela com as informações do SNP em relação ao rascunho genoma montado. O *script* search_scaffs é utilizado para confrontar o resultado produzido pelo DiscoSNP com aquele obtido após a execução do *pipeline* anonmaker.pl (Bertozzi *et al.,* 2012). O *script* filters_primers.sh, é utilizado para selecionar apenas um par de *primers* do resultados produzidos pelo Primer3. O *script* complete_table.py foi utilizado para completar a tabela gerada após a filtragem, adicionando os pares de *primers* e a classificação de SNPs que se enquadram como ANL. Todos *scripts* desenvolvidos para esta pesquisa estão delineados nos Apêndices de A a E e o modo de execução do trabalho com a utilização de cada um dos *scripts* está delineado no arquivo readme, Apêndice F.

4.2 Candidatos a marcadores caracterizados

O método implementado no programa DiscoSNP, baseado em grafos de *de bruijn*, profundidade de cobertura e qualidade das leituras, foi capaz de apontar um total de 17.481 possíveis candidatos a marcadores moleculares (10.6084/m9.figshare.6813038), dos quais 14.441 foram identificados como SNPs e 3.040 como INDELs. A cobertura média dos SNPs foi 17 ± 117 e o valor de qualidade média PHRED foi 100 ± 6.

A inspeção manual de uma amostra dos resultados permitiu a verificação dos candidatos a SNPs. Um exemplo de verificação encontra-se ilustrado na figura 4.2,

onde o resultado da saída do DiscoSNP (duas sequências variantes) foram alinhadas com o contig que contém essa região.

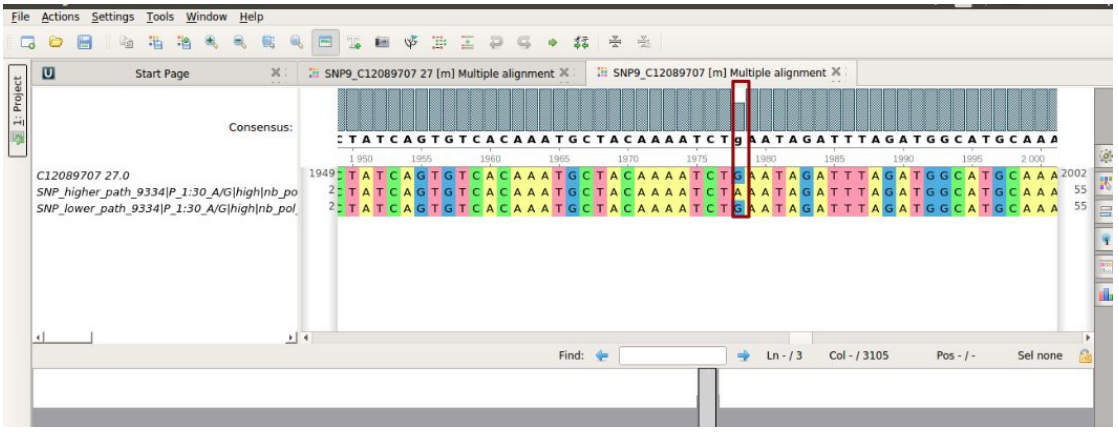

Fig. 4.2: Exemplo de visualização do alinhamento de um SNP com o contig pelo programa UGENE. Repare na presença do candidato à SNP (G/A).

A seguir, para o processo de inspeção visual de uma amostra de resultados apontados, o contig/scaffold e a possível variável em questão foram analisados no arquivo BAM, que mapeia as leituras curtas NGS na montagem do rascunho do genoma. O número de confirmações independentes de cada posição da sequência dá a profundidade de cada região, e a região candidata pode ser confirmada como contendo leituras com bases alternativas para aquele potencial SNP. Um exemplo deste procedimento pode ser visto na figura 4.2.2.

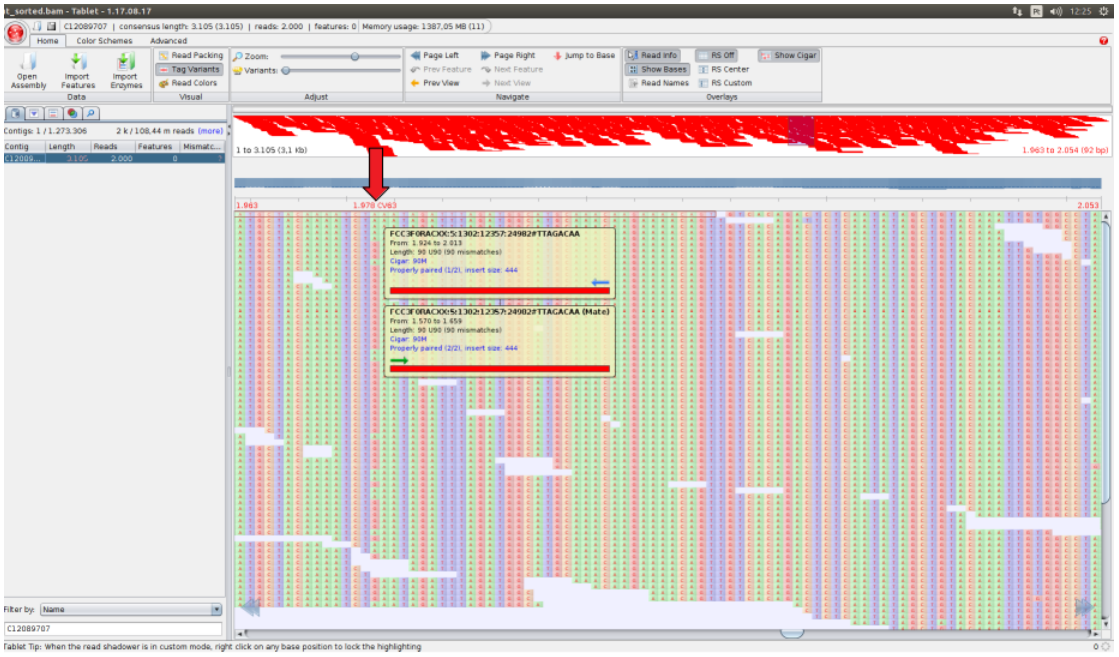

Fig 4.2.2: Exemplo de visualização de possível SNP na interface gráfica do programa Tablet. As leituras independentes são alinhadas a uma montagem e a seta vermelha aponta a posição-alvo. O diagrama em vermelho, acima, representa a profundidade em cada região específica desta sequência.

O tratamento da lista gerada pelo DiscoSNP com o *script* execute_pipeline.sh (Apêndice A) desenvolvido neste trabalho, que considera alvos com tamanho de fragmento de no mínimo 100 pb e tamanho de borda disponível 60 pb, encontrou um total de 8.631 pares de sequências (sequência 1 e 2 variam pela diferença em uma única base) como candidatos SNPs (10.6084/m9.figshare.6813104). Após o exame da lista de SNPs, dentro dos resultados obtidos na caracterização de ANL foram encontrados 377 contigs/scaffolds, sendo que desse número 171 são candidatos únicos a ANL, *i.e.* potenciais marcadores exclusivos em um contig ou scaffold, enquanto que os 206 restantes são possíveis marcadores em contigs/scaffolds com mais de um candidato a marcador presente. Este resultados estão sumarizados na Figura 4.2.3.

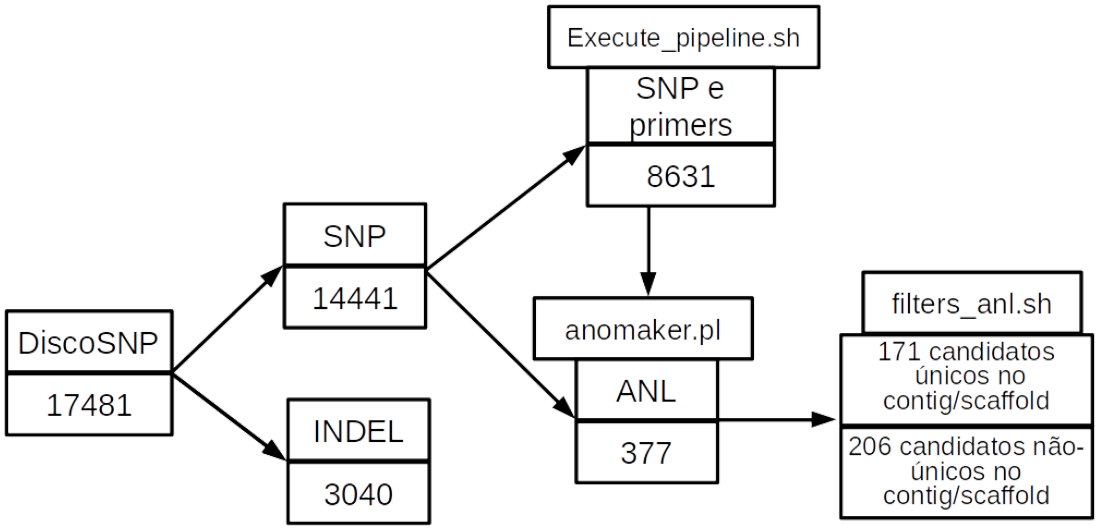

Figura 4.2.3: Sumário dos resultados obtidos com os *pipelines* empregados neste estudo

Foi realizada a análise dos tipos de mutações ocorridas nos candidatos a SNPs. Os resultados podem ser visualizados na Tabela 4.1.

Tabela 4.1: Frequência de substituições observadas nos candidatos a SNPs.

Tipo de mutação	Bases	Número	Total
Transições	A <---> G	2680	5403
	T <---> C	2723	
Transversões	A <---> C	877	3228
	A <---> T	1011	
	G <---> C	548	
	G <---> T	792	

Para 34 SNPs não foi possível realizar o desenho dos *primers* devido a sua localização em contigs contendo muitas lacunas de sequências desconhecidas (*gaps*).

Foi gerada uma lista contendo 3.040 candidatos a INDELs. O resultado encontra-se disponível *online* no repositório Figshare, de acesso público (10.6084/m9.figshare.6813107). Para os INDELs o tamanho médio encontrado foi 6,97 ± 11.2, sendo que a diferença mais constante de INDELs foi 1pb.

DISCUSSÃO

A exploração intensa de recursos aquáticos tem causado diversos impactos sobre as populações naturais, especialmente em peixes. *Brycon orbignyanus* é uma espécie de piracema, quase extinta da natureza e com indícios de depauperação genética, isso faz com que seja necessário desenvolver recursos que possam subsidiar estudos relacionados à genética dessa espécie. Para isso, torna-se importante o uso de novas metodologias e aplicações.

A bioinformática envolve o desenvolvimento de banco de dados e de algoritmos para a geração de informação útil e conhecimento para a pesquisa biológica (Pongor e Landsman, 1999; Lapatas *et al.,* 2015). É indispensável a utilização de computadores no manuseio e análise dos dados biológicos devido à velocidade de processamento, poder de resolução de problemas e pela necessidade da comunicação dos dados através da internet (Khan *et al.,* 20). Particularmente, a bioinformática é imprescindível para tratar o massivo volume de dados gerados por NGS. Neste trabalho, um conjunto de dados de mais de 16 bilhões de bases de DNA foi analisado.

Os novos *scripts* aqui gerados foram considerados eficientes para a automatização de tarefas nestes *pipelines* empregados e serão úteis para caracterização dos marcadores moleculares em outros conjuntos de dados NGS, para outras espécies. A disponibilização destes contribuirá para o processo de reprodutibilidade da metodologia em novos conjuntos de dados, algo que vem sendo cada vez mais discutido e indispensável no método científico (Piccolo e Frampton, 2016).

O software DiscoSNP oferece a vantagem de encontrar SNPs robustos sem a necessidade de executar a montagem do genoma nuclear. Portanto, ele pode ser aplicado com sucesso a genomas de organismos não-modelo (Quillery *et al*., 2014) em computadores com recursos limitados (*e.g*. 4 GB de memória ram) A utilização do DiscoSNP mostrou-se eficiente para a identificação de candidatos a marcadores a partir da análise de posições em heterozigose, em um único indivíduo. Esta espécie ainda não possui marcadores SNPs, ANL ou INDELS disponíveis na literatura. Os métodos bioinformáticos foram capazes de elucidar milhares de candidatos marcadores moleculares gerados a partir dos dados de NGS de *B. orbignyanus*, resultando na produção de um novo e amplo recurso genético para a espécie.

Considerando este conjunto de dados NGS com aproximadamente 16 Gb, foram encontrados cerca de 17 mil candidatos a marcadores moleculares. Utilizando a mesma ferramenta para a busca de SNPs em ervilhas, utilizando um conjunto de dados de quatro variedades, totalizando em 1.32 bilhões de leituras curtas, foram encontrados 419.024 SNPs (Boutet *et al.,* 2016). Naquele trabalho foi feita validação 64.754 SNPs, por meio de GBS para a construção de um mapa genético. Outro estudo, utilizando um conjunto de dados de 1,4 milhões de reads gerados pela plataforma 454, encontrou 321.088 possíveis candidatos a SNPs para uma espécie de carrapato, *Ixodes ricinus* (Quillery *et al.,* 2013), sendo validados empiricamente 384 SNPs por meio de genotipagem por meio de tecnologia nano-fluídica (Fluidigm). Em outro estudo, utilizando um conjunto de dados de 10 bibliotecas de Illumina em 360 milhões de bases foram encontrados 643.647 candidatos a SNPs (Thongthawee e Volkaert, 2014). Estes trabalhos utilizaram múltiplos indivíduos diferentes para a preparação das bibliotecas NGS. Possivelmente, o número de SNPs caracterizados é proporcional ao número de indivíduos utilizados para a criação da biblioteca, uma vez que a análise simultânea de mais de um indivíduo permite o encontro direto de regiões potencialmente polimórficas. Neste trabalho, foi utilizado apenas um indivíduo. Ainda assim, um grande número de candidatos a SNPs foram passíveis de detecção.

A relativa profundidade e a alta qualidade dos dados NGS aplicados, quando comparado com dados típicos de resultados de clonagem molecular e sequenciamento de DNA de Sanger, implica em alta confiabilidade das sequências estudadas e da natureza precisa e específica dos *primers* propostos. Isso leva à expectativa que, apesar da limitação desta pesquisa em não transpor a barreira *in silico* (com validação *in vitro*), futuros trabalhos de teste empírico destes candidatos a marcadores têm grandes chances de serem frutíferos e de alta produtividade, exemplo de Arias et al. (2016) que, então consistiu no maior número de marcadores empiricamente validados para a piracanjuba (29), logo suplantados pelos 47 de Cao *et al.* (2016) em dourado- *Salminus brasiliensis*, ambos com dados oriundos de NGS. Essa eventual confirmação empírica dos resultados aqui obtidos também será uma validação metodológica dos *pipelines* aqui aproveitados e desenvolvidos.

As análises do número de substituições nos candidatos a SNP/ANL demonstram que transições correspondem a 62,6% das mutações e transversões a 37,4%. O possível polimorfismo dos candidatos a SNPs mostra que a frequência de transições foi maior que de transversões, isso porque é mais frequente a ocorrência

de mutações entre as bases de mesma categoria, i.e, pirimidina substitui outra pirimidina, ou uma purina substitui outra purina, do que entre quando uma purina substitui uma pirimidina, ou vice-versa. A relação transições/transversões (1,67) vista aqui foi semelhante à relatada em outras espécies de peixes, como salmão - *Salmo solar* (Hayes *et al.,* 2007), dourada - *Sparus aurata* (Cenadelli *et al.,* 2007), paulistinha - *Danio rerio* (Stickney *et al.,* 2002) e pregado - *Scophthalmus maximus* (Vera *et al.,* 2011).

Diferentes ANL (ou mesmo diferentes SNPs) em um mesmo contig, indicam potenciais marcadores em desequilíbrio de ligação, por não segregarem de forma independente. A futura anotação gênica (determinação da função) de SNPs contidos em sequências retidas durante a filtragem para caracterização de ANL pode elucidar fortuitamente variantes de regiões codificadoras do genoma de piracanjuba.

Com o uso do DiscoSNP foi caracterizado um número representativo de INDELs. Foi observado que a maioria dos INDELs variam em apenas 1 pb de tamanho. Ainda assim, não foi possível fazer o desenho dos *primers* para o conjunto total de candidatos a INDELs de forma automatizada, com a metodologia aplicada ao processamento dos SNPs, devido a limitação do BLAST em realizar alinhamentos com mais de 10 pb de discrepância. A metodologia implementada pelo programa SWIPE (Rognes, 2011) poderia ser um substituto potencial ao BLAST, mas requer muito mais poder computacional e tempo de análise do que disponível quando da execução deste projeto. Sendo a lista de sequências contendo INDELs foi disponibilizada sem a lista de *primers*, para futuro desenvolvimento.

Marcadores de DNA tipo INDELS serão muito convenientes para análise de polimorfismos genéticos nesta espécie, uma vez que um simples ensaio de PCR, seguido por eletroforese em agarose ou acrilamida pode ser suficiente para revelar diferenças de tamanho de fragmento, com comportamento co-dominante, sendo mais econômicos do que marcadores SNPs e, portanto, facilmente aplicáveis a laboratórios limitados em operações de produção, como pisciculturas ambientais ou comerciais.

A disponibilização pública de painéis de potenciais marcadores de diferentes classes para esta espécie, como o delineado em Yazbeck *et al.* (2018) e no presente trabalho, inaugura uma nova era de pesquisa genética populacional para *B. orbignyanus* visto que, agora, toda a comunidade científica tem à disposição sequências de DNA e *primers* já caracterizados para serem testados empiricamente, potencialmente capazes de levar ao desenvolvimento de centenas de novos

marcadores baseados em PCR nos próximos anos. O acúmulo de marcadores irá não só guiar as iniciativas de procriação e repovoamento da espécie, mas permitirão trabalhos efetivos de melhoramento genético nos estoques de piscicultura ambiental e comercial, devido à sua alta densidade por cromossomo. Isso viabiliza os chamados estudos de associação, pela caracterização de marcadores de loci de características fenotípicas de variação contínua ou QTLs (*e.g*. Boulding *et al.,* 2008; Moen *et al.,* 2009; Gonen *et al.,* 2015), úteis para estudos ecológicos (por envolverem genes potencialmente sujeitos a pressões seletivas) e estudar os efeitos da seleção em espécimes de cativeiro, usados nos programas de peixamento.

A execução desse trabalho gerou milhares de novos candidatos a marcadores para a espécie *B. orbignyanus.* Esses possíveis marcadores inéditos irão contribuir para aplicação no manejo e conservação desta espécie, visto que constituem instrumentos úteis na caracterização genética de indivíduos e populações, na delimitação genética de estoques, no planejamento e avaliação de eficiência de peixamentos, na seleção de matrizes e sistemas de procriação em estações ambientais, na avaliação da necessidade e eficiências de MTPs e medidas de manejo ambiental diante de desastres e mortandades.

CONCLUSÃO

Esta pesquisa caracterizou o primeiro conjunto de sequências genômicas e *primers* candidatos a revelarem marcadores moleculares tipo SNP e ANL para *Brycon orbignyanus* por meio de abordagem de mineração de um vasto conjunto de dados do genoma desta espécie, previamente disponível. Foram disponibilizados publicamente painéis de potenciais marcadores moleculares deste peixe em bases de dados, juntamente com *scripts* que poderão ser aplicados ao desenvolvimento de marcadores em outros conjuntos de dados NGS, de outras espécies. Isto permitirá a reprodutibilidade do processo e servirá de base para novos estudos e aplicações. Os resultados aqui gerados fornecem um novo e amplo recurso para o desenvolvimento rápido e econômico de novas classes de marcadores moleculares, ainda inexistentes nesta espécie, para uso em pesquisas, programas de conservação e no manejo ambiental de pisciculturas para *B. orbignyanus*.

REFERÊNCIAS BIBLIOGRÁFICAS

Agostinho, Aa., Gomes, Lc. And Pelicice Fm., (2007). Ecologia e manejo de recursos pesqueiros em reservatórios do Brasil. Maringá: Eduem. 501 p.

Agostinho, A. A., Pelicice, F. M., & Gomes, L. C. (2008). Dams and the fish fauna of the Neotropical region: impacts and management related to diversity and fisheries. *Brazilian Journal of Biology*, *68*(4), 1119–1132. https://doi.org/10.1590/S1519-69842008000500019

Agostinho, Angelo Antonio, Gomes, L. C., Fernandez, D. R., & Suzuki, H. I. (n.d.). Efficiency of fish ladders for neotropical ichthyofauna. *River Research and Applications*, *18*(3), 299–306. https://doi.org/10.1002/rra.674

Agostinho, Angelo Antonio, Pelicice, F. M., Gomes, L. C., & Junior, H. F. J. (2010). Reservoir Fish Stocking: When One Plus One May Be Less Than Two. https://doi.org/10.4322/natcon.00802001

Albert, J. S., & Reis, R. E. (2011). *Historical Biogeography of Neotropical Freshwater Fishes*. University of California Press.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*(2), 81–92. https://doi.org/10.1038/nrg.2015.28

Arias, M. C., Aulagnier, S., Baerwald, E. F., Barclay, R. M. R., Batista, J. S., Beasley, R. R., … Zou, S. (2016). Microsatellite records for volume 8, issue 1. *Conservation Genetics Resources*, *8*(1), 43–81. https://doi.org/10.1007/s12686-016-0522-2

Ashikaga, F. Y., Orsi, M. L., Oliveira, C., Senhorini, J. A., & Foresti, F. (2015). The endangered species *Brycon orbignyanus*: genetic analysis and definition of priority areas for conservation. *Environmental Biology of Fishes*, *98*(7), 1845–1855. https://doi.org/10.1007/s10641-015-0402-8

Attard, C. R. M., Beheregaray, L. B., & Möller, L. M. (2016). Towards population-level conservation in the critically endangered Antarctic blue whale: the number and distribution of their populations. *Scientific Reports*, *6*, srep22291. https://doi.org/10.1038/srep22291

Avise, J. C. (2012). *Molecular Markers, Natural History and Evolution*. Springer Science & Business Media.

Barletta, M., Jaureguizar, A. J., Baigun, C., Fontoura, N. F., Agostinho, A. A., Almeida-Val, V. M. F., … Corrêa, M. F. M. (2010). Fish and aquatic habitat

conservation in South America: a continental overview with emphasis on neotropical systems. *Journal of Fish Biology*, *76*(9), 2118–2176. https://doi.org/10.1111/j.1095-8649.2010.02684.x

Barroso, R. M., Hilsdorf, A. W. S., Moreira, H. L. M., Mello, A. M., Guimarães, S. E. F., Cabello, P. H., & Traub-Cseko, Y. M. (2003). Identification and characterization of microsatellites loci in *Brycon opalinus* (Cuvier, 1819) (Characiforme, Characidae, Bryconiae). *Molecular Ecology Notes*, *3*(2), 297–298. https://doi.org/10.1046/j.1471-8286.2003.00435.x

Begon, M., Townsend, C. R., & Harper, J. L. (2009). Ecology: From Individuals to Ecosystems.

Begossi, A. (1998). Linking Social and Ecological Systems: Management Practices and Social Mechanisms for Building Resilience. Cambridge, UK: Cambridge University Press.

Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Archives of Disease in Childhood. Education and Practice Edition*, *98*(6), 236–238. https://doi.org/10.1136/archdischild-2013-304340

Bertozzi, T., Sanders, K. L., Sistrom, M. J., & Gardner, M. G. (2012). Anonymous nuclear loci in non-model organisms: making the most of high-throughput genome surveys. *Bioinformatics (Oxford, England)*, *28*(14), 1807–1810. https://doi.org/10.1093/bioinformatics/bts284

Bhattramakki, D., Dolan, M., Hanafey, M., Wineland, R., Vaske, D., Register, J. C., … Rafalski, A. (2002). Insertion-deletion polymorphisms in 3′ regions of maize genes occur frequently and can be used as highly informative genetic markers. *Plant Molecular Biology*, *48*(5–6), 539–547. https://doi.org/10.1023/A:1014841612043

Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., … Zaretskaya, I. (2013a). BLAST: a more efficient report with usability improvements. *Nucleic Acids Research*, *41*(Web Server issue), W29-33. https://doi.org/10.1093/nar/gkt282

Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., … Zaretskaya, I. (2013b). BLAST: a more efficient report with usability improvements. *Nucleic Acids Research*, *41*(Web Server issue), W29–W33. https://doi.org/10.1093/nar/gkt282

Boutet, G., Carvalho, S. A., Falque, M., Peterlongo, P., Lhuillier, E., Bouchez, O., Baranger, A. (2016). SNP discovery and genetic mapping using genotyping

by sequencing of whole genome genomic DNA from a pea RIL population. *BMC Genomics*, *17*(1), 121. https://doi.org/10.1186/s12864-016-2447-2

Brandão-Dias, P. F. P., Carmo, A. O. do, Martins, A. P. V., Pimenta, R. J. G., Alves, C. B. M., & Kalapothakis, E. (2016). Complete mitochondrial genome of *Salminus brasiliensis* (Characiformes, Characidae). *Mitochondrial DNA Part A*, *27*(3), 1577–1578. https://doi.org/10.3109/19401736.2014.958676

Boulding, E. G., Culling, M., Glebe, B., Berg, P. R., Lien, S., & Moen, T. (2008). Conservation genomics of Atlantic salmon: SNPs associated with QTLs for adaptive traits in parr from four trans-Atlantic backcrosses. *Heredity*, 101(4), 381–391. https://doi.org/10.1038/hdy.2008.67

Buermans, H. P. J., & den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica Et Biophysica Acta*, *1842*(10), 1932–1941. https://doi.org/10.1016/j.bbadis.2014.06.015

Cao, Y.-L., Caputo, L. I., Cheng, H., Carmo, F. M. da S., Carvalho, L. C. de, Yazbeck, G. de M., … Zhou, C.-H. (2016). Microsatellite records for volume 8, issue 3. *Conservation Genetics Resources*, *8*(3), 359–370. https://doi.org/10.1007/s12686-016-0581-4

Carmo, F. M. da S., Polo, É. M., Silva, M. A. da, Yazbeck, G. de M., Carmo, F. M. da S., Polo, É. M., … Yazbeck, G. de M. (2015). Optimization of heterologous microsatellites in Piracanjuba. *Pesquisa Agropecuária Brasileira*, *50*(12), 1236–1239. https://doi.org/10.1590/S0100-204X2015001200015

Carolsfeld, J., Centre (Canada), I. D. R., Bank, W., & Trust, W. F. (2003). *Migratory Fishes of South America: Biology, Fisheries and Conservation Status*. IDRC.

Craig, J. F. (2016). Freshwater Fisheries Ecology. John Wiley & Sons.

Cenadelli, S., Maran, V., Bongioni, G., Fusetti, L., Parma, P., & Aleandri, R. (n.d.). Identification of nuclear SNPs in gilthead seabream. *Journal of Fish Biology*, *70*(sc), 399–405. https://doi.org/10.1111/j.1095-8649.2007.01454.x

Chaitankar, V., Karakülah, G., Ratnapriya, R., Giuste, F. O., Brooks, M. J., & Swaroop, A. (2016). Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research. *Progress in Retinal and Eye Research*, *55*, 1–31. https://doi.org/10.1016/j.preteyeres.2016.06.001

Chauhan, T., & Rajiv, K. (2010). Molecular markers and their applications in fisheries and aquaculture. *Advances in Bioscience and Biotechnology*, *01*(04), 281. https://doi.org/10.4236/abb.2010.14037

Choupina, A. B., Martins, I. M., Choupina, A. B., & Martins, I. M. (2014). Molecular markers for genetic diversity, gene flow and genetic population structure of freshwater mussel species. *Brazilian Journal of Biology*, *74*(3), S167–S170. https://doi.org/10.1590/1519-6984.25112

Chu, E. W. (2003). Essentials of Conservation Biology. Third Edition. By Richard B Primack. *The Quarterly Review of Biology*, *78*(2), 253–254. https://doi.org/10.1086/378013

Compeau, P. E. C., Pevzner, P. A., & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, *29*(11), 987–991. https://doi.org/10.1038/nbt.2023

Dai, L., Gao, X., Guo, Y., Xiao, J., & Zhang, Z. (2012). Bioinformatics clouds for big data manipulation. *Biology Direct*, *7*, 43. https://doi.org/10.1186/1745-6150-7-43

Davey, J. W., & Blaxter, M. L. (2010). RADSeq: next-generation population genetics. *Briefings in Functional Genomics*, *9*(5–6), 416–423. https://doi.org/10.1093/bfgp/elq031

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, *12*(7), 499–510. https://doi.org/10.1038/nrg3012

Dowie, N. J., Grubisha, L. C., Burton, B. A., Klooster, M. R., & Miller, S. L. (2017). Development of Anonymous Nuclear Loci for *Pterospora andromedea* (Monotropoideae) Using Illumina and Ion Torrent Sequencing Data. *Conservation Genetics Resources*, *9*(3), 371–373. https://doi.org/10.1007/s12686-017-0686-4

Dowie, Nicholas J., Grubisha, L. C., Burton, B. A., Klooster, M. R., & Miller, S. L. (2017). Increased phylogenetic resolution within the ecologically important Rhizopogon subgenus Amylopogon using 10 anonymous nuclear loci. *Mycologia*, *109*(1), 35–45. https://doi.org/10.1080/00275514.2017.1285165

Drezen, E., Rizk, G., Chikhi, R., Deltel, C., Lemaitre, C., Peterlongo, P., & Lavenier, D. (2014). GATB: Genome Assembly & Analysis Tool Box.

Bioinformatics, *30*(20), 2959–2961. https://doi.org/10.1093/bioinformatics/btu406

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Ekblom, R., & Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, *107*(1), 1–15. https://doi.org/10.1038/hdy.2010.152

Ferreira, M. E.; Grattapaglia, D. (1996) Introdução ao uso de marcadores moleculares em analise genetica. 2.ed. Brasilia: Embrapa-Cenargen.

Frankham, R., Ballou, J. D., & Briscoe, D. A. (2010). *Introduction to Conservation Genetics* (Edição: 2). Cambridge: Cambridge University Press.

Freeland, J. R., Kirk, H., & Petersen, S. (2011). Molecular Genetics in Ecology. In *Molecular Ecology* (pp. 1–34). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470979365.ch1

Garvin, M. R., Saitoh, K., & Gharrett, A. J. (2010). Application of single nucleotide polymorphisms to non-model species: a technical review. *Molecular Ecology Resources*, *10*(6), 915–934. https://doi.org/10.1111/j.1755-0998.2010.02891.x

Gonen, S., Baranski, M., Thorland, I., Norris, A., Grove, H., Arnesen, P., … Houston, R. D. (2015). Mapping and validation of a major QTL affecting resistance to pancreas disease (salmonid alphavirus) in Atlantic salmon (*Salmo salar*). *Heredity,* 115(5), 405–414. https://doi.org/10.1038/hdy.2015.37

Haeckel, E. H. P. A. (1866). *Generelle morphologie der organismen. Allgemeine grundzüge der organischen formen-wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte descendenztheorie.* Berlin, G. Reimer. Retrieved from http://archive.org/details/generellemorphol01haec

Harvey, B., & Baer, A. (2004). *Migratory Fishes of South America: Biology, Fisheries and Conservation Status*. (J. Carolsfeld & C. Ross, Eds.). Washington DC: IDRC.

Hayes, B., Laerdahl, J. K., Lien, S., Moen, T., Berg, P., Hindar, K., … Høyheim, B. (2007). An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed

sequences. *Aquaculture*, *265*(1), 82–90. https://doi.org/10.1016/j.aquaculture.2007.01.037

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Hogeweg, P. (2011). The Roots of Bioinformatics in Theoretical Biology. *PLoS Computational Biology*, *7*(3). https://doi.org/10.1371/journal.pcbi.1002021

Jeltsch, F., Bonte, D., Pe'er, G., Reineking, B., Leimgruber, P., Balkenhol, N., … Bauer, S. (2013). Integrating movement ecology with biodiversity research - exploring new avenues to address spatiotemporal biodiversity dynamics. *Movement Ecology*, *1*, 6. https://doi.org/10.1186/2051-3933-1-6

Karl, S. A., & Avise, J. C. (1993). PCR-based assays of mendelian polymorphisms from anonymous single-copy nuclear DNA: techniques and applications for population genetics. *Molecular Biology and Evolution*, *10*(2), 342–361. https://doi.org/10.1093/oxfordjournals.molbev.a040002

Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Mahmoud Ali, W. K., Alam, M., Gani, A. (2014). Big Data: Survey, Technologies, Opportunities, and Challenges. *The Scientific World Journal*, *2014*. https://doi.org/10.1155/2014/712826

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, *155*(1), 27–38. https://doi.org/10.1016/j.cell.2013.09.006

Lapatas, V., Stefanidakis, M., Jimenez, R. C., Via, A., & Schneider, M. V. (2015). Data integration in biological research: an overview. *Journal of Biological Research*, *22*(1). https://doi.org/10.1186/s40709-015-0032-5

Lee, H. C., Lai, K., Lorenc, M. T., Imelfort, M., Duran, C., & Edwards, D. (2012). Bioinformatics tools and databases for analysis of next-generation sequence data. *Briefings in Functional Genomics*, *11*(1), 12–24. https://doi.org/10.1093/bfgp/elr037

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., … Law, M. (2012). Comparison of Next-Generation Sequencing Systems [Research article]. https://doi.org/10.1155/2012/251364

Lopera-Barrero, N. M., Vargas, L., Nardez-Sirol, R., Pereira-Ribeiro, R., Aparecido-Povh, J., Jr, S., … Cristina-Gomes, P. (2010). Diversidad genética y contribución reproductiva de una progenie de Brycon

orbignyanus en el sistema reproductivo seminatural, usando marcadores microsatélites. *Agrociencia*, *44*(2), 171–181.

Lopes J.M., Bedore A.G. (2008). Peixamento como medida ambiental mitigadora do impacto na ictiofauna. Ação Ambiental (UFV), 39:28-34.

Lowe-McConnell, R. H (1999). Estudos ecológicos de comunidades de peixes tropicais. São Paulo, EDUSP,

Luz-Agostinho, K. D. G., Agostinho, A. A., Gomes, L. C., Júlio-Jr., H. F., & Fugi, R. (2009). Effects of flooding regime on the feeding activity and body condition of piscivorous fish in the Upper Paraná River floodplain. *Brazilian Journal of Biology*, *69*(2), 481–490. https://doi.org/10.1590/S1519-69842009000300004

Malone, G. & P.D. Zimmer. (2005). Marcadores Moleculares. Vermelho (Oryza sativa L.). p.77-113. In P.D. Zimmer, A.C. Oliveira & G. Malone. *Ferramentas da biotecnologia no melhoramento genético vegetal*. Universitária/ UFPel, Pelotas. 158 p.

Mahadani, P., & Ghosh, S. K. (2014). Utility of indels for species-level identification of a biologically complex plant group: a study with intergenic spacer in Citrus. *Molecular Biology Reports*, *41*(11), 7217–7222. https://doi.org/10.1007/s11033-014-3606-7

Mammadov, J., Aggarwal, R., Buyyarapu, R., & Kumpatla, S. (2012). SNP Markers and Their Impact on Plant Breeding [Research article]. https://doi.org/10.1155/2012/728398

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, *9*, 387–402. https://doi.org/10.1146/annurev.genom.9.081307.164359

McConnell, R. (1999). *Estudos Ecológicos de Comunidades de Peixes Tropicais Vol. 03*. Edusp.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews. Genetics*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626

Milne, I., Stephen, G., Bayer, M., Cock, P. J. A., Pritchard, L., Cardle, L., … Marshall, D. (2013). Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics*, *14*(2), 193–202. https://doi.org/10.1093/bib/bbs012

Moen, T., Baranski, M., Sonesson, A. K., & Kjøglum, S. (2009). Confirmation and fine-mapping of a major QTL for resistance to infectious pancreatic necrosis

in Atlantic salmon (*Salmo salar*): population-level associations between markers and trait. *BMC Genomics,* 10, 368. https://doi.org/10.1186/1471-2164-10-368

Neiff, J. J. (1990). Ideas para la interpretación ecológica del Paraná. Interciência, 15: 424-441.

Okonechnikov, K., Golosova, O., Fursov, M., & UGENE team. (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics (Oxford, England)*, *28*(8), 1166–1167. https://doi.org/10.1093/bioinformatics/bts091

Oliveira, D. J. de, Ashikaga, F. Y., Foresti, F., & Senhorini, J. A. (2017). Conservation Status of the "Piracanjuba" *Brycon orbignyanus* (Valenciennes, 1850) (Characiformes, Bryconidae): Basis for Management Programs. *Biodiversidade Brasileira*, *7*(1), 18–33.

Parker, P. G., Snow, A. A., Schug, M. D., Booton, G. C., & Fuerst, P. A. (1998). What Molecules Can Tell Us about Populations: Choosing and Using a Molecular Marker. *Ecology*, *79*(2), 361–382. https://doi.org/10.2307/176939

Pauly, D., Christensen, V., Guénette, S., Pitcher, T. J., Sumaila, U. R., Walters, C. J., Zeller, D. (2002). Towards sustainability in world fisheries. *Nature* https://doi.org/10.1038/nature01017

Pelicice, F. M., & Agostinho, A. A. (2008). Fish-passage facilities as ecological traps in large neotropical rivers. *Conservation Biology: The Journal of the Society for Conservation Biology*, *22*(1), 180–188. https://doi.org/10.1111/j.1523-1739.2007.00849.x

Pereira, F., Carneiro, J., Matthiesen, R., van Asch, B., Pinto, N., Gusmão, L., & Amorim, A. (2010). Identification of species by multiplex analysis of variable-length sequences. *Nucleic Acids Research*, *38*(22), e203. https://doi.org/10.1093/nar/gkq865

Petrere Junior, M., 1985. Migraciones de peces de agua dulce en America Latina: algunos comentarios. Copescal Doc.Ocas., (1):17 p.

Petrere Junior, M. (n.d.). River fisheries in Brazil: A review. *Regulated Rivers: Research & Management*, *4*(1), 1–16. https://doi.org/10.1002/rrr.3450040102

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLOS ONE*, *7*(5), e37135. https://doi.org/10.1371/journal.pone.0037135

Piccolo, S. R., & Frampton, M. B. (2016). Tools and techniques for computational reproducibility. *GigaScience*, *5*(1), 30. https://doi.org/10.1186/s13742-016-0135-4

Poke, F. S., Vaillancourt, R. E., Potts, B. M., & Reid, J. B. (2005). Genomic research in Eucalyptus. *Genetica*, *125*(1), 79–101. https://doi.org/10.1007/s10709-005-5082-4

Polanski, A., & Kimmel, M. (2007). *Bioinformatics*. Berlin, Heidelberg: Springer-Verlag.

Pongor, S., & Landsman, D. (1999). Bioinformatics and the developing world. *Biotechnology and Development Monitor*, *40*, 10–13.

Pradhan, S. K., Barik, S. R., Sahoo, A., Mohapatra, S., Nayak, D. K., Mahender, A., … Pandit, E. (2016). Population Structure, Genetic Diversity and Molecular Marker-Trait Association Analysis for High Temperature Stress Tolerance in Rice. *PLoS ONE*, *11*(8). https://doi.org/10.1371/journal.pone.0160027

Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics (Oxford, England)*, *21 Suppl 1*, i351-358. https://doi.org/10.1093/bioinformatics/bti1018

Reis, R. E., Albert, J. S., Di Dario, F., Mincarone, M. M., Petry, P., & Rocha, L. A. (2016). Fish biodiversity and conservation in South America. *Journal of Fish Biology*, *89*(1), 12–47. https://doi.org/10.1111/jfb.13016

Rognes, T. (2011). Faster Smith-Waterman database searches with inter-sequence SIMD parallelisation. *BMC Bioinformatics*, 12, 221. https://doi.org/10.1186/1471-2105-12-221

Rosa, R. S.; Lima, F. C. T. (2008.). Os Peixes Brasileiros Ameaçados de Extinção. In: *Livro vermelho da fauna brasileira ameaçada de extinção.* Brasília: Ministério do Meio Ambiente.

Saenger, W. (2013). *Principles of Nucleic Acid Structure*. Springer Science & Business Media.

Sanches, A., & Galetti, P. M. (2006). Microsatellites loci isolated in the freshwater fish *Brycon hilarii*. *Molecular Ecology Notes*, *6*(4), 1045–1046. https://doi.org/10.1111/j.1471-8286.2006.01427.x

Scheiner, S. M., & Willig, M. R. (2008). A general theory of ecology. *Theoretical Ecology*, *1*(1), 21–28. https://doi.org/10.1007/s12080-007-0002-0

Schlötterer, C. (2004). The evolution of molecular markers — just a matter of fashion? *Nature Reviews Genetics*, *5*(1), 63–69. https://doi.org/10.1038/nrg1249

Selkoe, K. A., & Toonen, R. J. (2006). Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters*, *9*(5), 615–629. https://doi.org/10.1111/j.1461-0248.2006.00889.x

Silva, M. C., Duarte, M. A., & Coelho, M. M. (2011). Anonymous Nuclear Loci in the White-Faced Storm-Petrel Pelagodroma marina and Their Applicability to Other Procellariiform Seabirds. *Journal of Heredity*, *102*(3), 362–365. https://doi.org/10.1093/jhered/esr016

Simpson, R. D., & Christensen, N. L. (2012). *Ecosystem Function & Human Activities: Reconciling Economics and Ecology*. Springer Science & Business Media.

Siqueira, F. de F., Carmo, A. O. do, Pimentel, J. da S. M., & Kalapothakis, E. (2016). Complete mitochondrial genome sequence of *Brycon orbignyanus* (Characiformes, Bryconidae). *Mitochondrial DNA. Part A, DNA Mapping, Sequencing, and Analysis*, *27*(3), 1942–1943. https://doi.org/10.3109/19401736.2014.971298

Smit, AFA, Hubley, R & Green, P (2013-2015). *RepeatMasker Open-4.0*. <http://www.repeatmasker.org>.

Souza, F. P. de, Urrea-Rojas, A. M., Ruas, C. de F., Povh, J. A., Ribeiro, R. P., Ruas, E. A., … Lopera-Barrero, N. M. (2018). Novel microsatellite markers for the endangered neotropical fish *Brycon orbignyanus* and cross-amplification in related species. *Italian Journal of Animal Science*, *0*(0), 1–5. https://doi.org/10.1080/1828051X.2018.1436008

Stickney, H. L., Schmutz, J., Woods, I. G., Holtzer, C. C., Dickson, M. C., Kelly, P. D., … Talbot, W. S. (2002). Rapid Mapping of Zebrafish Mutations With SNPs and Oligonucleotide Microarrays. *Genome Research*, *12*(12), 1929–1934. https://doi.org/10.1101/gr.777302

Sugunan, V. V. (1997). Fisheries management of small water bodies in seven countries in Africa, Asia and Latin America. *FAO Fisheries Circular*, (No. 933). Retrieved from https://www.cabdirect.org/cabdirect/abstract/19981802595

Sunnucks, null. (2000). Efficient genetic markers for population biology. *Trends in Ecology & Evolution*, *15*(5), 199–203.

Thongthawee, S and Volkaert, H, (2014) Analysis of teak *(Tectona grandis)* genome and its diversity, TBS

Thomson, R. C., Wang, I. J., & Johnson, J. R. (2010). Genome-enabled development of DNA markers for ecology, evolution and conservation. *Molecular Ecology*, *19*(11), 2184–2195. https://doi.org/10.1111/j.1365-294X.2010.04650.x

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., … Williams, S. M. (2009). The Genetic Structure and History of Africans and African Americans. *Science (New York, N.Y.)*, *324*(5930), 1035–1044. https://doi.org/10.1126/science.1172257

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3--new capabilities and interfaces. *Nucleic Acids Research*, *40*(15), e115. https://doi.org/10.1093/nar/gks596

Uricaru, R., Rizk, G., Lacroix, V., Quillery, E., Plantard, O., Chikhi, R., … Peterlongo, P. (2015). Reference-free detection of isolated SNPs. *Nucleic Acids Research*, *43*(2), e11. https://doi.org/10.1093/nar/gku1187

Quillery, E., Quenez, O., Peterlongo, P., & Plantard, O. (2014). Development of genomic resources for the tick *Ixodes ricinus*: isolation and characterization of single nucleotide polymorphisms. Molecular Ecology Resources, 14(2), 393–400. https://doi.org/10.1111/1755-0998.12179

Väli, Ü., Brandström, M., Johansson, M., & Ellegren, H. (2008). Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. *BMC Genetics*, *9*, 8. https://doi.org/10.1186/1471-2156-9-8

van Nimwegen, K. J. M., van Soest, R. A., Veltman, J. A., Nelen, M. R., van der Wilt, G. J., Vissers, L. E. L. M., & Grutters, J. P. C. (2016). Is the $1000 Genome as Near as We Think? A Cost Analysis of Next-Generation Sequencing. *Clinical Chemistry*, *62*(11), 1458–1464. https://doi.org/10.1373/clinchem.2016.258632

Vera, M., Álvarez-Dios, J. A., Millán, A., Pardo, B. G., Bouza, C., Hermida, M., … Martínez, P. (2011). Validation of single nucleotide polymorphism (SNP) markers from an immune Expressed Sequence Tag (EST) turbot, Scophthalmus maximus, database. *Aquaculture*, *313*(1), 31–41. https://doi.org/10.1016/j.aquaculture.2011.01.038

Vignal, A., Milan, D., SanCristobal, M., & Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics.

Genetics, Selection, Evolution: GSE, *34*(3), 275–305. https://doi.org/10.1051/gse:2002009

Yang, N., Li, H., Criswell, L. A., Gregersen, P. K., Alarcon-Riquelme, M. E., Kittles, R., … Seldin, M. F. (2005). Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine. *Human Genetics*, *118*(3–4), 382–392. https://doi.org/10.1007/s00439-005-0012-1

Yazbeck, G. M., Oliveira, R. S., Ribeiro, J. M., Graciano, R. D., Santos, R. P., Carmo, F. M. S., & Lavenier, D. (2018). A broad genomic panel of microsatellite loci from *Brycon orbignyanus* (Characiformes: Bryconidae) an endangered migratory Neotropical fish. *Scientific Reports*, *8*(1), 8511. https://doi.org/10.1038/s41598-018-26623-x

Zaniboni-Filho, E., Ribolli, J., Hermes-Silva, S., Nuñer, A. P. O., Zaniboni-Filho, E., Ribolli, J., … Nuñer, A. P. O. (2017). Wide reproductive period of a long-distance migratory fish in a subtropical river, Brazil. *Neotropical Ichthyology*, *15*(1). https://doi.org/10.1590/1982-0224-20160135

APÊNDICES

Apêndice A: *Script* execute_pipeline.sh, este realiza a filtragem a partir do resultado do DiscoSNP, gerando uma lista com os melhores candidatos.

```
################################################################################
#Comment lines below to skip filter phase
if [ -f $output_file ]; then
       rm $output_file
fi
echo "Filtering SNPs to match filter parameters"
for match_file in `ls -v $matches_dir/*.out`; do
       text=`grep -A 9 '>' $match_file`
       text=`echo "$text" | sed s/^--$/\$/g`
       IFS=$'$' y=($text)
       line_number=`echo "$match_file" | awk -F '/' '{print $2}' | awk -F '_'
'{print $2}' | awk -F '.' '{print $1}'`
       snp_seq=`awk "NR==$line_number" $snps_file`
       line_number=`expr $line_number - "1"`
       snp_name=`awk "NR==$line_number" $snps_file`
       if [[ $snp_name = *"INDEL"* ]]; then
       echo "Discarding $snp_name. We do not handle INDELs in this script"
       else
       query_length=`cat "$match_file" | grep -m 1 'Length=' | awk -F '=' '{print
$2}'`;
       for item in ${y[@]}; do
              scaff=`echo $item |  grep ">"`
              scaff_length=`echo $item |  grep "Length=" | awk -F '=' '{print
$2}'`
         blast_query_length=`echo $item |  grep "Identities =" | awk -F '='
'{print $2}' | awk '{print $1}' | awk -F '/' '{print $2}'`
              blast_match_length=`echo $item |  grep "Identities =" | awk -F '='
'{print $2}' | awk '{print $1}' | awk -F '/' '{print $1}'`
              if [  "$scaff_length" -lt "$filter_lenght" ]; then
              echo "Discarding scaffold ${scaff}. The size to discard is
$filter_lenght, and the scaffold size is $scaff_length"
              else
              snp_pos=`echo $item | grep '|'`
              snp_pos_trimmed=`echo $snp_pos | sed -e 's/^[[:space:]]*//'`
              before_space=`echo $snp_pos_trimmed | awk '{print $1}'`
              len=${#before_space}
              border_right=`echo $item | grep Sbjct | awk '{print $4}'`
              border_left=`echo $item | grep Sbjct | awk '{print $2}'`
              if [  "$border_right" -lt "$border_left" ]; then
                     aux=$border_right
                     border_right=$border_left
                     border_left=$aux
              fi
              border_left_size=`expr $border_left - "1"`
              border_right_size=`expr $scaff_length - $border_right`
              snp_position=`expr $border_left + $len`
              #find scaff seq to generate primer config
              scaffseq=`grep -A1 $scaff $assembly_file | grep -v $scaff`
              if [ "$blast_query_length" -ne "$query_length" ]; then
                     echo "Discarding scaffold ${scaff}. The blast query length
($blast_query_length) is different from the original query length
($query_length)"
              elif [  "$border_left_size" -lt "$border_min_size" -o
"$border_right_size" -lt "$border_min_size" ]; then
                     echo "Discarding scaffold ${scaff}. The size to discard is
Left border ($border_left_size) or Right border ($border_right_size) is less than
$border_min_size"
              else
```

```
                        index_lower=`echo "$match_file" | awk -F '/' '{print $2}' |
awk -F '_' '{print $2}' | awk -F '.' '{print $1}'`;
                        index_higher=`expr $index_lower - "2"`;
                        higher_seq=`sed -n "${index_higher}p" $snps_file`;
                        lower_seq=`sed -n "${index_lower}p" $snps_file`;

                        if [ "$blast_match_length" -eq "$query_length" ]; then
                        echo $higher_seq > /tmp/higher_seq_tmp111
                        echo $lower_seq > /tmp/lower_seq_tmp111
                        new_len_txt=`cmp -l   /tmp/higher_seq_tmp111
/tmp/lower_seq_tmp111`
                        new_len=`echo $new_len_txt | awk '{print $1}'`
                        snp_position=`expr $border_left + $new_len`
                        snp_position=`expr $snp_position - 1`
                        echo $snp_name, $snp_seq, $snp_position, $scaff >>
$output_file
                        generate_primer $scaffseq $snp_position tmp_primer3.primer3
$(basename "$match_file" ".out").$scaff.primer3_result
                        else
                        match_diff=`expr $query_length - $blast_match_length`
                        if [ "$match_diff" -ne "1" ]; then
                                echo "Discarding scaffold ${scaff}. Scaffold and SNP
differs by more than 1 nucleotide"
                        else
                                nucleotide_higher=${higher_seq:$len:1}
                                nucleotide_lower=${lower_seq:$len:1}

                                if [ "$nucleotide_higher" != "$nucleotide_lower" ];
then
                                echo $snp_name, $snp_seq, $snp_position, $scaff >>
$output_file
                                generate_primer $scaffseq $snp_position
tmp_primer3.primer3 $(basename "$match_file" ".out").$scaff.primer3_result
                                else
                                echo "Discarding scaffold ${scaff}. Scaffold and SNP
differs by more than 1 nucleotide"
                                fi


                        fi

                        fi

                fi

                fi

        done

        IFS=$'\n'

        fi
done
#################################################################################
```

Apêndice B: *script* search_scaffs.sh, faz a seleção dos contigs/scaffolds contendo ANL.

```
#Parameters: scaffold's list and DiscoSnp output
while read p; do
        grep "$p" $2
done <$1
```

Apêndice C: *script* filter_primers.sh, seleciona apenas um par de *primers* a partir do resultado gerado pelo primer3.

```
#!/bin/bash

primer3_output_folder=$1
for primer_file in $primer3_output_folder/*; do

        primer_left=`grep PRIMER_LEFT_0_SEQUENCE "$primer_file"`;
        primer_right=`grep PRIMER_RIGHT_0_SEQUENCE "$primer_file"`;

        echo $primer_left > "$primer_file".primer_seq
        echo $primer_right >> "$primer_file".primer_seq

done
```

Apêndice D: *script* complete_table.py, completa a tabela adicionando os pares de *primers* e a classificação dos SNPs.

```
from sys import argv
import glob
table_file = open(argv[1], 'r')
original_file = open(argv[2], 'r')
primer_folder = argv[3]
matched_dict = {}
count = 0
for line in original_file:
	if count % 2 == 0:
	matched_dict[line[:-1]] = str(count+2)
	count += 1
for line in table_file:
	label = line.split()[0][:-1].strip()
	full_scaff = line.split(',')[-1].strip()
	scaff = line.split(',')[-1].split()[0]
	seq   = line.split()[1][:-1]
	pos   = line.split()[2][:-1]
	primer_file_name =
primer_folder+'/filtered/'+'match_'+matched_dict[label]+'.'+scaff+"*"
	possible_file = glob.glob(primer_file_name)
	if len(possible_file) == 0:
	#aqui temos que procurar dentro da pasta de anls
	primer_file_name =
primer_folder+'/anls/filtered/'+'match_'+matched_dict[label]+'.'+scaff+"*"
	possible_file = glob.glob(primer_file_name)
	primer_file = open(possible_file[0])
	lines = primer_file.readlines()
	try:
		l = lines[0].split('=')[1].strip()
		r = lines[1].split('=')[1].strip()
	except:
		l = " "
		r = " "
	print label, ";", seq, ";", pos, ";", full_scaff, ";", l, ";", r, ";",
"ANL"
	else:
	primer_file = open(possible_file[0])
	lines = primer_file.readlines()
	try:
		l = lines[0].split('=')[1].strip()
		r = lines[1].split('=')[1].strip()
	except:
		l = " "
		r = " "
	print label, ";", seq, ";", pos, ";", full_scaff, ";", l, ";", r, ";",
"SNP"
```

Apêndice E: *script* filter_anl.sh, seleciona os ANLs que são único em um contig/scaffold

```
#!/bin/bash

while IFS='' read -r line || [[ -n "$line" ]]; do
    echo "ls $2 | grep $line"
done < "$1"
```

Apêndice F: readme file de execução do trabalho

```
This document describes how to generate a list containing putative SNPs, INDELs
and ANL from raw set(s) of reads, followed by filtering of the results for the
best candidates and primers design for in vitro PCR-based amplification.

# To generate a list of candidate SNPs and INDELs
       #Run DiscoSNP, available in https://gatb.inria.fr/software/discosnp/
       ./run_discoSnp++.sh -r fof.txt

# Perform data filtering
       #Run execute_pipeline
       usage:      sh      execute_pipeline.sh      assembly_file      snps_file
min_scaffold_size_filter min_border_size_filter output_file_name
       example:    sh    execute_pipeline.sh    brycon_prefix_oneline_k55.scafSeq
O_discoRes_k_31_c_auto_D_100_P_1_b_0_coherent_Rosiane.fa 500 60 snp_tables.txt

# To characterize anonymous nuclear loci
       # Use of the AnonMarker PERL script
       (http://www.samuseum.sa.gov.au/about/staff/dr-terry-bertozzi).
       perl AnonMarker.pl <assembly_file>

# To select the contigs/scaffolds containing ANL.
       #run script search_scaffs.sh
       sh script search_scaffs.sh <output_AnonMaker>

# To selects only one pair of primers from the results generated by primer3.
       # run shell script filter_primers.sh on the folder primer3_results
       sh filter_primers.sh

# To complete the table by adding the pairs of primers and the classification of
the SNPs.
       #run python script complete_table.py
       python2 complete_table.py

# To select ANLs that are single in a contig/scaffold
       run shell script filter_anl.sh
       sh script filter_anl.sh
```