Genômica da Biodiversidade Brasileira

PROD. TEC. ITV DS – N024/2023
DOI 10.29223/PROD.TEC.ITV.DS.2023.24.Oliveira

RELATÓRIO TÉCNICO ITV DS

MANUAL PARA O MONITORAMENTO DE BIODIVERSIDADE POR MEIO DE DNA AMBIENTAL (eDNA *metabarcoding*)

Renato R. M. Oliveira[1] • Vitória Catarina Cardoso Martins[1]
Paulo Henrique de O. Costa[1] • Izabela Santos Mendes[1 2]
Silvia Britto Barreto[1 2] • José Bitencourt[1] • Santelmo Vasconcelos[1]
Gisele Nunes[1] • Guilherme Oliveira[1]
Alexandre Aleixo[1]

Belém / PA

Novembro / 2023

Parceria entre:
INSTITUTO TECNOLÓGICO VALE
ICMBio

Genômica da Biodiversidade Brasileira

Título: Manual para o monitoramento de Biodiversidade por meio de DNA ambiental (eDNA metabarcoding)	
PROD. TEC. ITV DS N024/2023	Revisão
Classificação: ( ) Confidencial ( ) Restrita ( ) Uso Interno (x) Pública	00

Informações Confidenciais - Informações estratégicas para o ITV e ICMBio. Seu manuseio é restrito a usuários previamente autorizados pelos Gestores da Informação.
Informações Restritas - Informação cujo conhecimento, manuseio e controle de acesso devem estar limitados a um grupo restrito que necessitam utilizá-la para exercer suas atividades profissionais.
Informações de Uso Interno - São informações destinadas à utilização interna por empregados e prestadores de serviço
Informações Públicas - Informações que podem ser distribuídas ao público externo, o que, usualmente, é feito através dos canais corporativos apropriados

Nota de capa

1 Instituto Tecnológico Vale, Belém/PA;

2 Instituto Chico Mendes de Conservação da Biodiversidade.

Citar como
OLIVEIRA, R. R. M. *et al.* **Manual para o monitoramento de Biodiversidade por meio de DNA ambiental (eDNA metabarcoding).** Belém: 2023. (Relatório Técnico N024/2023) DOI 10.29223/PROD.TEC.ITV.DS.2023.24.Oliveira

Dados Internacionais de Catalogação na Publicação (CIP)

O48 Oliveira, Renato Renison Moreira.
Manual para o monitoramento de Biodiversidade por meio de DNA ambiental (eDNA metabarcoding). / Renato Renison Moreira Oliveira ... [*et al.*] - Belém: 2023.
46 p. : il.

Relatório Técnico (Instituto Tecnológico Vale) – 2023
PROD.TEC.ITV.DS – N024/2023
DOI 10.29223/PROD.TEC.ITV.DS.2023.24.Oliveira

1. Conservação. 2. Biodiversidade. 3. DNA ambiental. 4. Metabarcoding. I. Martins, Vitória Catarina Cardoso. II. Costa, Paulo Henrique de Oliveira. III. Mendes, Izabela Santos. IV. Barreto, Silvia Britto. V. Bitencourt, José Augusto Pires. VI. Vasconcelos Júnior, Santelmo Selmo. VII. Nunes, Gisele Lopes. VIII. Oliveira, Guilherme Corrêa de. IX. Aleixo, Alexandre Luis Padovan. X. Título

Bibliotecária responsável: Nisa Gonçalves / CRB 2 – 525

Parceria entre:
INSTITUTO TECNOLÓGICO VALE
ICMBio

Genômica da Biodiversidade Brasileira

RESUMO EXECUTIVO

Este documento é resultado da integração e atualização de dois manuais previamente desenvolvidos para estudos de monitoramento da ictiofauna e flora e apresenta o eDNA metabarcoding como uma solução para identificar múltiplas espécies através do DNA ambiental. Esta técnica depende de bancos de dados de códigos de barra de DNA e está sendo muito importante para o monitoramento ambiental e a conservação da biodiversidade, oferecendo resultados rápidos e precisos. Além disso, é essencial para estratégias de manejo e atendimento a propósitos regulatórios e de preservação. Os métodos de monitoramento de flora e ictiofauna são aqui usados como estudos de caso, exemplificando a aplicação prática do eDNA metabarcoding no Projeto Genômica da Biodiversidade Brasileira (GBB), uma iniciativa crucial para o conhecimento da biodiversidade brasileira e para orientar esforços de conservação.

Parceria entre:
INSTITUTO TECNOLÓGICO VALE
ICMBio

RESUMO

O Brasil, reconhecido por sua rica biodiversidade, enfrenta desafios significativos para mapear e conservar sua vasta fauna e flora, devido à grandeza e diversidade de seu território. A escassez de taxonomistas agrava a situação, tornando os levantamentos de biodiversidade mais raros e custosos. Nesse contexto, o DNA *metabarcoding* surge como uma solução inovadora, permitindo a identificação simultânea de múltiplas espécies através da análise do DNA ambiental encontrado em amostras de solo, água, sedimentos e/ou tecidos biológicos. A eficiência da técnica se depende de bancos de dados de códigos de barra de DNA, gerados a partir do sequenciamento de marcadores moleculares específicos de espécimes previamente identificados por especialistas, garantindo assim a confiabilidade dos resultados. Esta abordagem vem revolucionando os projetos de monitoramento ambiental e conservação da biodiversidade, oferecendo resultados rápidos e precisos. Além disso, a técnica tem se mostrado essencial para estratégias de manejo eficazes e para atender a diversos propósitos regulatórios e de preservação. Este manual integra métodos de monitoramento de flora e ictiofauna, propondo-se a exemplificar a aplicação prática da técnica de DNA *metabarcoding*, visando sua adoção no Projeto Genômica da Biodiversidade Brasileira (GBB). A iniciativa representa um passo crucial para gerar conhecimento aprofundado da biodiversidade brasileira, essencial para minimizar impactos ambientais e orientar esforços de conservação.

Palavras-chave: conservação; biodiversidade; *metabarcoding;* DNA ambiental.

Genômica da Biodiversidade Brasileira

ABSTRACT

Brazil, known for its rich biodiversity, faces significant challenges in mapping and conserving its vast fauna and flora, due to the size and diversity of its territory. The shortage of taxonomists exacerbates the situation, making biodiversity surveys rare and more costly. In this context, DNA metabarcoding emerges as an innovative solution, allowing for the simultaneous identification of multiple species through the analysis of environmental DNA found in soil, water, sediments, and/or biological tissue samples. The technique relies on DNA barcode databases, compiled from sequencing specific molecular markers of specimens previously identified by experts, thus ensuring the reliability of the results. This approach revolutionizes environmental monitoring and biodiversity conservation projects, offering fast and accurate results. Moreover, the technique is proving essential for effective management strategies and meeting various regulatory and preservation purposes. This manual integrates monitoring methods for flora and ichthyofauna, aiming to demonstrate the practical application of the DNA metabarcoding technique, seeking its adoption in the Genomics of Brazilian Biodiversity (GBB) Project. This initiative represents a crucial step in generating in-depth knowledge of Brazilian biodiversity, which is essential for minimizing environmental impacts and guiding conservation efforts.

Keywords: conservation; biodiversity; metabarcoding; environmental DNA.

LISTA DE ILUSTRAÇÕES

Genômica da Biodiversidade Brasileira

SUMÁRIO

1 INTRODUÇÃO

O Brasil é considerado o país com a maior biodiversidade do mundo, abrigando 15-20% da biodiversidade do planeta (ELLWANGER; NOBRE; CHIES, 2022), abrigando ecossistemas únicos. Conhecer essa biodiversidade é o ponto básico para a sua conservação. Porém, realizar o levantamento da fauna e flora no Brasil é um desafio, dada a extensão e ampla diversidade de um país continental. Adicionalmente, a escassez de financiamento para a investigação taxonômica básica levou ao chamado “impedimento taxonômico”, reduzindo o número de taxonomistas qualificados e, consequentemente, limitando os avanços na avaliação e descrição da biodiversidade (VOGEL ELY *et al.,* 2017). Dessa forma, os estudos da fauna e flora de ambientes complexos se tornam cada vez mais raros e dispendiosos.

Neste sentido, tecnologias inovadoras baseadas em DNA, que visam a caracterização de espécies em larga escala, vêm sendo amplamente utilizadas. Uma dessas técnicas é a análise de DNA de amostras ambientais (e.g., água, solo, sedimento, ar etc.), denominada DNA ambiental (eDNA) (FICETOLA *et al.,* 2008). Essas amostras de eDNA são amplificadas utilizando iniciadores conservados para regiões previamente definidas, informativas para os táxons de interesse, e são submetidas ao sequenciamento de alto rendimento (*High Throughput Sequencing* – HTS. Esse procedimento permite a identificação taxonômica simultânea de múltiplas espécies a partir de uma única amostra, sem a necessidade de isolamento ou captura dos organismos alvo, metodologia conhecida como eDNA metabarcoding (CRISTESCU, 2014; DEINER *et al.,* 2017; FAHNER *et al*., 2016).

Para a identificação taxonômica, a abordagem de DNA *metabarcoding* necessita de bancos de dados de códigos de barras de DNA (i.e., barcodes),

construídos a partir de sequências de DNA de espécimes previamente identificadas por especialistas (FAZEKAS *et al.,* 2009; KRESS; ERICKSON, 2012). A construção desses bancos de referência é um pré-requisito para a obtenção de resultados confiáveis de forma robusta, idealmente obtendo-se identificações até o menor nível taxonômica possível (i.e., espécies) (HEBERT *et al.,* 2003).

O eDNA *metabarcoding* é uma ferramenta poderisa, que tem contribuído positivamente para o monitoramento da biodiversidade, gerando conhecimento necessário para minimizar o impacto ao meio ambiente. Esse método tem sido aplicado em projetos de monitoramento ambiental, permitindo avaliar a eficácia das estratégias de manejo, assim como em estudos de conservação da biodiversidade, fornecendo um resultado rápido e confiável para uma ampla gama de propósitos regulatórios e de conservação. Levantamentos da fauna e flora brasileira visam gerar o conhecimento necessário afim de minimizar o impacto no meio ambiente. Entretanto, para avaliar a biodiversidade e a composição da comunidade de organismos usando essa metodologia são necessários protocolos eficazes em cada uma de suas etapas.

Esse manual resulta da combinação de dois manuais elaborados anteriormente para o monitoramento da flora (MARTINS et al., 2021) e da ictiofauna (COSTA et al., 2021). A versão atual é uma demanda do Projeto Genômica da Biodiversidade Brasileira (GBB), do eixo DNA *metabarcoding*, trazendo os dois manuais previamente citados como estudos de caso para demonstrar a aplicação da metodologia de DNA *metabarcoding* nas atividades de monitoramento, além de atualizações metodológicas.

Nas próximas seções serão abordados os materiais e métodos necessários para o monitoramento da flora e ictiofauna nas etapas de coleta das amostras, extração de DNA, amplificação das regiões de interesse,

preparo das bibliotecas, sequenciamento dos amplicons e análises de bioinformática.

2 COLETA DE AMOSTRAS, EXTRAÇÃO DE DNA, AMPLIFICAÇÃO E PREPARO DA BIBLIOTECA PARA ESTUDOS DA FLORA

2.1 COLETA DE AMOSTRAS DE SOLO PARA ESTUDOS DA FLORA

Um dos principais pontos de atenção que se deve ter durante a coleta de material para estudos genômicos está relacionado aos procedimentos para evitar a contaminação da amostra coletada, bem como a contaminação cruzada (entre amostras). Logo, todos os procedimentos de amostragem e filtragem devem ser conduzidos com luvas de acetonitrila descartáveis. Além disso, um par de luvas novas deve ser utilizado para cada local amostrado. Caso a luva toque uma superfície que esteja contaminada ela deve ser trocada.

Para a coleta de solo, espátulas de metal limpas e autoclavadas são utilizadas (um número suficiente de espátulas deve ser considerado por coleta) (**Figura 1A**). A coleta deve ser realizada em triplicata. Em campo, a assepsia das espátulas deve ser feita utilizando álcool 70% e, quando possível, flambadas ao fogo, tomando sempre cuidado para não provocar danos ao ambiente ou ao coletor. Também pode ser utilizada solução de hipoclorito de sódio a 2% seguida de água ultrapura para a remoção de possíveis resíduos.

O solo deve ser armazenado em contêineres estéreis, sendo a recomendação usar sacos de poliuretano de alta densidade (HDPE), selados e esterilizados por radiação gama, como os do tipo *whirl-pak* (**Figura 1B**). A identificação deve ser feita previamente e com números únicos, para evitar troca de amostras. O identificador da amostra pode ser escrito ou colado no saco com etiquetas próprias para este fim. Caso opte pelas etiquetas, imprimir previamente cópias em triplicata para colar no saco de amostragem e levar caderno de anotações e uma etiqueta extra para quaisquer eventualidades.

Independente da escolha, os números de identificação devem ser anotados/colados no caderno de anotações juntamente com os metadados obtidos na coleta.

O transporte das amostras até o laboratório deverá ser feito o mais rápido possível em um contêiner termoestável ou isopor. Recomenda-se a temperatura de -20ºC. Entretanto, devido às condições de logística e as adversidades do campo, pode-se seguir as recomendações do Guia Nacional de Coleta e Preservação de Amostras[1], que preconiza o transporte a 4ºC. Seguir as devidas recomendações são importantes para manutenção da qualidade da amostra, visando reduzir a degradação do material genético alvo.

No campo, uma alternativa é usar sacolas térmicas com gelo químico para o armazenamento da amostra. O uso de gelo deve ser feito com cuidado, devido a uma possível contaminação das amostras pelo degelo. Ao chegar no laboratório, o material deve ser imediatamente armazenado em ultra freezer a -80ºC.

[1] Disponível em: http://arquivos.ana.gov.br/institucional/sge/CEDOC/Catalogo/2012/GuiaNacionalDeColeta.pdf

Figura 1- Amostragem em campo (A); Saco *whirl-pak* (B).

Fonte: Martins *et al.*, 2021

Os metadados são todas as informações extras referentes a cada amostra. O seu preenchimento é uma etapa obrigatória, uma vez que os parâmetros de amostragem são de suma importância tanto para a identificação das amostras quanto para análises posteriores. Em campo, os metadados obrigatórios são basicamente: coordenadas geográficas ou pontos de amostragem, nome descritivo, data da coleta, hora da coleta e coletor. Outros fatores podem ser considerados (sazonalidade, dados físico-químicos, profundidade, quantidade, replicatas, método de coleta etc.), dependendo do tipo de material a ser coletado. É preciso tomar o devido cuidado para anotar quais metadados estão associados a quais amostras.

Fotos do local de coleta também são desejáveis, sendo que a mesma câmera utilizada para as fotos pode ser utilizada para tirar fotos das anotações.

Isso facilita a conexão das imagens aos dados da amostra. A planilha com os metadados deve ser digitalizada preferencialmente no mesmo dia da coleta.

2.2 EXTRAÇÃO DE DNA DE SOLO

A etapa de extração de eDNA deve ser feita em laboratório, devidamente limpo e equipado, seguindo o fluxo de trabalho (**Figura 2**). O kit sugerido para extração de DNA de solos é o Kit DNeasy PowerSoil Kit (Apêndice) (QIAGEN, Hilden, Germany). Este kit reduz a ação de inibidores naturais do solo que podem causar prejuízos à qualidade do DNA a ser extraído, resultando então em um DNA de alta qualidade.

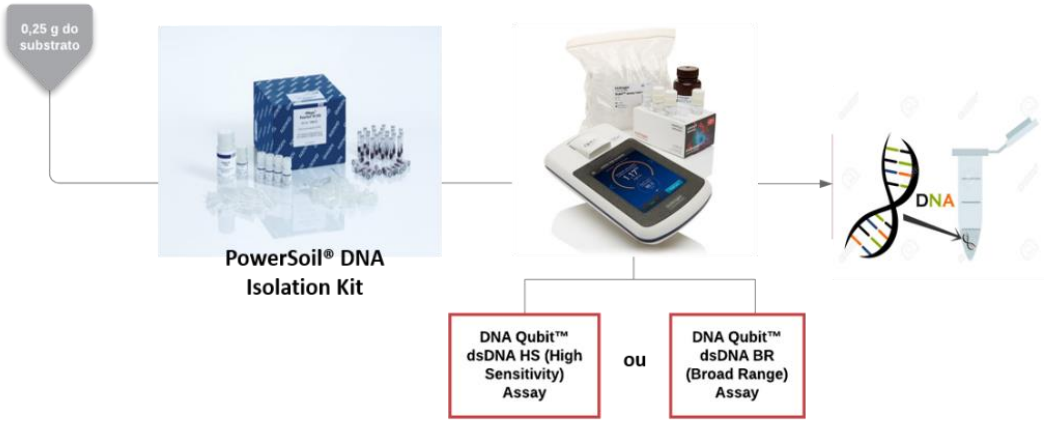
Para evitar contaminação, todo o processo de extração deve ser executado em capela de fluxo laminar, previamente limpa com Hipoclorito de sódio 2% e Etanol 70% e posteriormente, esterilizada com luz UV por 30 minutos. Além disso, controles negativos devem ser preparados a cada nova extração. Os controles negativos são feitos utilizando água no lugar do eDNA, seguindo os mesmos procedimentos e kits utilizados para as amostras de eDNA. Recomenda-se fazer extrações de DNA em triplicatas, mas isso depende da questão a ser respondida e da necessidade do projeto.

A extração do eDNA deverá ser feita usando o Kit DNeasy PowerSoil Kit (QIAGEN), conforme as instruções abaixo:

1. Identificar os tubos com o código das amostras;
2. Pesar 0,25 g de solo de cada amostra e adicionar nos tubos identificados;
3. Adicionar 60 µL de solução C1 e inverter os tubos manualmente por 10 vezes;

4. Vortexar por 10 minutos a 3.000 rpm. Recomenda-se adaptar o mix para vortexar horizontalmente;
5. Posteriormente, os tubos devem ser centrifugados a 14.000 rpm por 30 segundos. Transferir um volume de 500 µL do sobrenadante para um novo microtubo, fornecido pelo kit;
6. Adicionar 250 µL de solução C2 e agitar o tubo no vortex (Mix Mate) por 5 segundos. Imediatamente incubar a 4°C por 5 minutos;
7. Centrifugar por 1 minuto a 14.000 rpm e transferir até 600 µL de sobrenadante (evitando o pellet) para um novo microtubo fornecido pelo kit;
8. Adicionar 200 µL de solução C3 e agitar no vortex (Mix Mate) por 5 segundos e incubar novamente a 4°C por 5 minutos. Os microtubos são centrifugados a 14.000 rpm por um minuto e, em seguida, até 750 µL de sobrenadante é transferido para um novo microtubo (evitando o pellet), no qual é adicionado 1.200 µL da solução C4. Agitar no vortex (Mix Mate) por 5 segundos;
9. Adicionar 675 µL da solução em um microtubo que contém uma coluna de sílica acoplada e centrifugar por 1 minuto a 14.000 rpm. Esse passo é repetido três vezes e, logo após, é adicionado 500 µL de solução C5, e o tubo é centrifugado por 30 segundos a 14.000rpm;
10. Descartar o conteúdo do tubo coletor e centrifugar o microtubo vazio por 30 segundos a 14.000 rpm para remover qualquer solução C5 residual. Em seguida, a coluna de sílica é transferida para um novo microtubo e deve ser adicionado 100 µL de solução C6 no centro da coluna para fazer a eluição do DNA. Centrifugar o microtubo por 30 seg a 14.000 rpm;
11. A coluna de sílica é descartada e o DNA é armazenado a -20°C.

Figura 2 - Fluxograma simplificado mostrando o kit recomendado de extração do DNA total das amostras de solo e o método de quantificação do DNA.



Fonte: Martins *et al*, 2021

Ao final, as amostras devem ser eluídas em 100 µL de tampão C6 fornecido pelo próprio kit PowerSoil (10 mM Tris- HCl pH 8.5) e quantificadas no fluorômetro Qubit ® (Thermo Fisher Scientific, Apêndice) com o kit Qubit™ dsDNA HS (High Sensitivity) Assay (Thermo Fisher Scientific, Apêndice), caso a concentração de DNA esteja entre 10pg/µL e 100ng/µL. Caso a concentração esteja entre 100pg/µL e 1,000ng/µL utilizar o kit Qubit™ dsDNA BR (Broad Range) Assay (Thermo Fisher Scientific, Apêndice). O produto da extração deve ser colocado em microtubos Eppendorf Safe-Lock 1,5 µL e armazenado a -20ºC.

2.3 AMPLIFICAÇÃO DE MARCADORES GENÉTICOS A PARTIR DO DNA TOTAL EXTRAÍDO DE AMOSTRAS DE SOLO

Pares de *primers* (fita simples e curta utilizada para dar início a síntese de DNA) direcionados para a amplificação das regiões de ITS (espaçador interno transcrito do rDNA nuclear) foram testados para a adequação ao protocolo de sequenciamento nas plataformas de sequenciamento de nova geração (NGS), como o Ion Torrent e o Illumina MiSeq) conforme indicados na **Tabela 1**. O ITS é o principal marcador usado para identificar rastros de DNA de plantas em amostras de solo (ZHAO; WANG; HUA, 2018). Após a verificação de eficiência de diferentes combinações de *primers*, o par ITS2F-ITS3R foi selecionado para as etapas posteriores devido à sua alta eficiência de amplificação para as amostras de DNA ambiental.

Tabela 1 - Lista de *primers* testados para amplificação e sequenciamento nas plataformas de NGS.

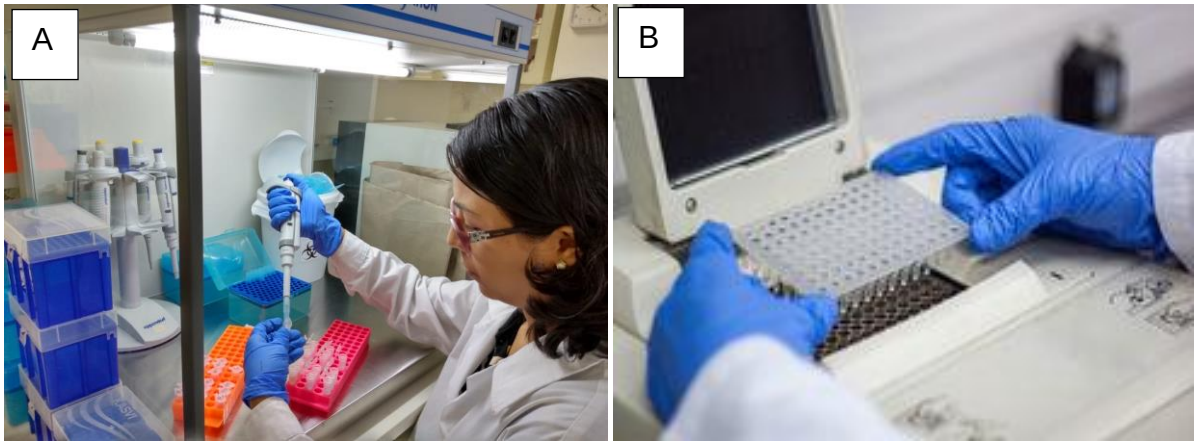
Região	Primer	Sequência (5'-3')
ITS2	ITS2F	ATGCGATACTTGGTGTGAAT
ITS2	ITS3R	GACGCTTCTCCAGACTACAAT
ITS	ITS1	TCCGTAGGTGAACCTGCGG
ITS	ITS1F	CTTGGTCATTTAGAGGAAGTAA
ITS	ITS4	TCCTCCGCTTATTGATATGC
ITS	ITS4-R	CAGGAAACAGCTATGACTCCTCCGCTTATTGATATGC

Como sugestão, os *primers* podem ser sintetizados pela empresa IDT (*Integrated DNA Technologies*) em escala de síntese padrão. Antes de preparar a solução estoque deve-se centrifugar o tubo do primer liofilizado por 30 segundos. Para preparar a solução estoque, a quantidade em nmols é multiplicada por 10. Isso vai gerar a quantidade em µL de água ultrapura que precisa ser adicionada ao primer. Recomenda-se vortexar por 30 segundos a 1 minuto para ter certeza de que o primer está bem dissolvido. Para preparar a

solução de uso que vai ser usada na reação em cadeia da polimerase (PCR) é necessário diluir para 10 µM. Para tanto, utiliza-se 10 µL da solução estoque e 90 µL de água ultrapura.

As reações de amplificação devem ser realizadas em triplicata para cada amostra, e controles negativos devem ser utilizados toda vez que for realizada uma PCR (**Figura 3B)**. Outra recomendação importante é não utilizar a pipeta multicanal para pipetar o DNA.

Figura 3 - Extração do DNA (A); Amostras sendo colocadas no termociclador (B)



Fonte: Martins *et al*, 2021

A preparação da reação de sequenciamento, chamada de Mix, que contém todos os reagentes necessários à síntese de novas cópias de DNA, deve ser sintetizada de acordo com a Tabela 2. Não é recomendado utilizar o 5x *Green Buffer* que vem com o tampão de carregamento. O reagente dimethyl sulphoxide (DMSO), adquirido pela empresa HIMEDIA, deve ser usado em alíquotas para evitar contaminação. O reagente TBT-PAR (solução 5x) é preparado no laboratório conforme o protocolo do Apêndice.

Tabela 2 - Componentes e volumes necessários para a preparação do Mix.

Composto	**µL**
5x Buffer	2,5
25 mM $MgCl_2$	1,2
2 mM dNTP	1
TBT	2,5
DMSO	1
10pmol ITS 2F	0,25
10pmol ITS 3R	0,25
Taq	0,1
DNA	2
H_2O ultrapura	1,7
Total	12,5
Volume para 1 amostra	

Fonte: autores (2023).

Para cada amostra, recomenda-se que as PCRs sejam feitas em triplicatas para posterior corrida no termociclador, como o modelo Veriti 96-Well Thermal Cycler (Thermo Fisher), usado no ITV. As condições da PCR dependem do conjunto de primer utilizado. Para o conjunto ITS2F-ITS3R, as seguintes condições foram utilizadas: desnaturação inicial a 94°C por 3 minutos, seguida por 30 ciclos de amplificação com 1 minuto a 94°C, 1 minuto a 54°C, em seguida 1 minuto a 72ºC e extensão final por 7 minutos a 72ºC.

2.4 PREPARO DA BIBLIOTECA PARA O SEQUENCIAMENTO

Esta etapa é importante para a inclusão de adaptadores de sequências únicas (indexes) Illumina para o sequenciamento. Recomendamos a

construção das bibliotecas utilizando o protocolo 16S Metagenomic Sequencing Library Preparation da Illumina (Apêndice), que apesar de mostrar a amplificação de outras regiões de interesse, é um protocolo prático e objetivo.

Inicialmente, todas as triplicatas geradas pela PCR anterior devem ser unidas antes da primeira purificação. Os procedimentos da primeira purificação do DNA devem ser feitos a partir do produto da PCR (*amplicons*) utilizando o kit Agencourt Ampure XP beads (Beckman Coulter), seguindo as instruções do protocolo anteriormente mencionado. Para avaliar a qualidade dos *amplicons*, as amostras devem ser submetidas à eletroforese em gel de agarose 1%, seguindo o protocolo de preparo no laboratório ou por meio de eletroforese capilar, utilizando o Bioanalyzer Agilent Technology 2100 (PANARO et al., 2000).

Posteriormente, os *indexes* devem ser adicionados a cada amostra através da etapa de *Index* utilizando o kit Nextera XT Library prep (Illumina, San Diego, CA, USA), conforme instruções do protocolo (16S *Metagenomic Sequencing Library Preparation* da Illumina). Pares de *indexes* utilizados na montagem da biblioteca não devem ser repetidos nas amostras que serão sequenciadas em uma mesma corrida. Em seguida, uma nova PCR com os *indexes* que foram adicionados a cada amostra deve ser feita utilizando 5 µL de DNA, 5µ L de Nextera XT *index* Primer1, 5 µL de Nextera XT *index* Primer2, 25 µL de enzima 2x Kapa Hifi HotStart Ready mix e 10 µL de água ultrapura, para um volume final de 50 µL. A etapa de *Index* PCR deve ser feita de acordo com as seguintes condições: desnaturação inicial a 95°C por 3 minutos, seguida por 8 ciclos de amplificação com 30 segundos a 95°C, 30 segundos a 55°C, em seguida 30 segundos a 72ºC e extensão final por 5 minutos a 72ºC.

O produto da etapa de Index PCR deve ser purificado novamente utilizando o kit *Agencourt Ampure XP beads* conforme instruções do protocolo.

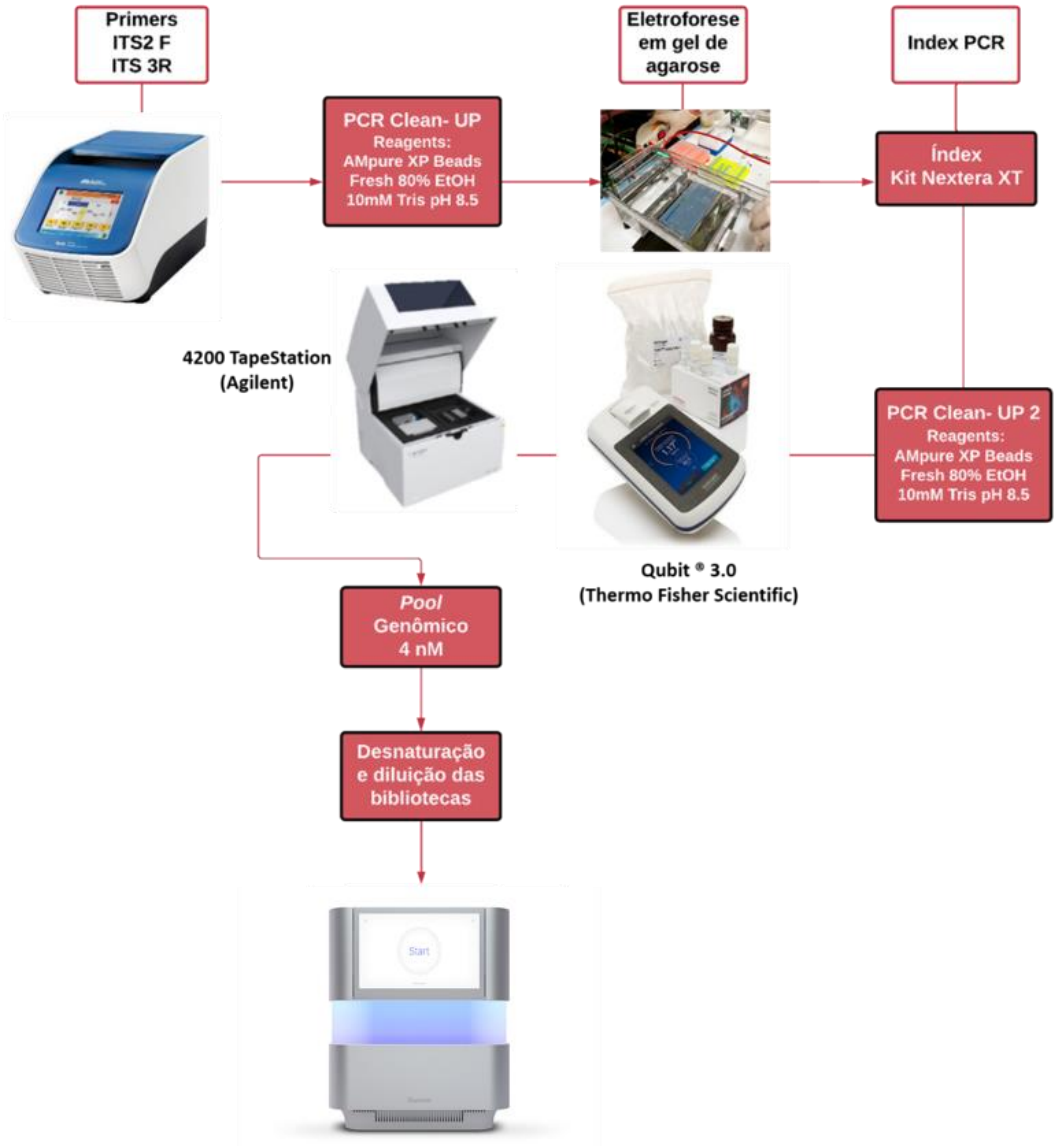
As bibliotecas devem ser quantificadas no fluorômetro Qubit ® (Thermo Fisher Scientific) e o tamanho do *amplicon* também deve ser analisado no Tapestation System 4200 (Agilent). Para a escolha do kit (d1000, d5000 ou High Sensitive) deve-se primeiro verificar o tamanho do fragmento para depois seguir o protocolo do fabricante. O *pool* genômico (produto do *Index* PCR purificado) deve ser padronizado para a concentração de 2 nM utilizando a Equação 1, que deve ser preenchida de acordo com a concentração de DNA quantificado e o tamanho médio do *amplicon* obtido.

Equação 1:

$$\frac{(concentração\ em\ ng/\mu l)}{(660\ g/mol \times tamanho\ médio\ da\ biblioteca)} \times 10^6 = concentração\ em\ nM$$

A corrida de sequenciamento deve ser realizada utilizando os kits de corrida NextSeq 2000 P1 (60 Gb) ou P2 (180 Gb) (2 x 300bp) (**Figura 4)**.

Figura 4 - Fluxograma simplificado mostrando as etapas de construção de bibliotecas para o sequenciamento no NextSeq 2000 Illumina.



Fonte: adaptado de Martins *et al*., 2021.

3 COLETA DE AMOSTRAS, EXTRAÇÃO DE DNA, AMPLIFICAÇÃO E PREPARO DA BIBLIOTECA PARA ESTUDOS DA ICTIOFAUNA

3.1 COLETA DE AMOSTRAS DE ÁGUA PARA ESTUDOS DA ICTIOFAUNA

As amostras de água são coletadas com o uso de tubos com filtros de 0,22 □m tipo Sterivex™ com conexão do tipo "*Luer Outlet*" (**Figura 5A**). O pacote do filtro deve ser aberto no topo e guardado. No filtro é então colada uma etiqueta com um número identificador único. A água deve ser coletada com uma seringa de 50 ou 60 mL que tenha um bico do tipo "*Luer Lock*" (**Figura 5B**). Após a coleta da água, a seringa deve ser acoplada ao Sterivex™ pelo bico "*Luer Lock*" e apertada manualmente para o líquido ser filtrado (**Figura 5C**). Após a filtragem de todo o líquido, a seringa é desacoplada e o processo repetido. O filtro deve ser colocado dentro do pacote em que estava após ser desacoplado para evitar contaminação. O processo é repetido pelo número de vezes necessário para que o filtro sature e não seja mais possível forçar a filtragem. O número de vezes em que o processo é repetido depende da quantidade de partículas e de matéria orgânica suspensa na água.

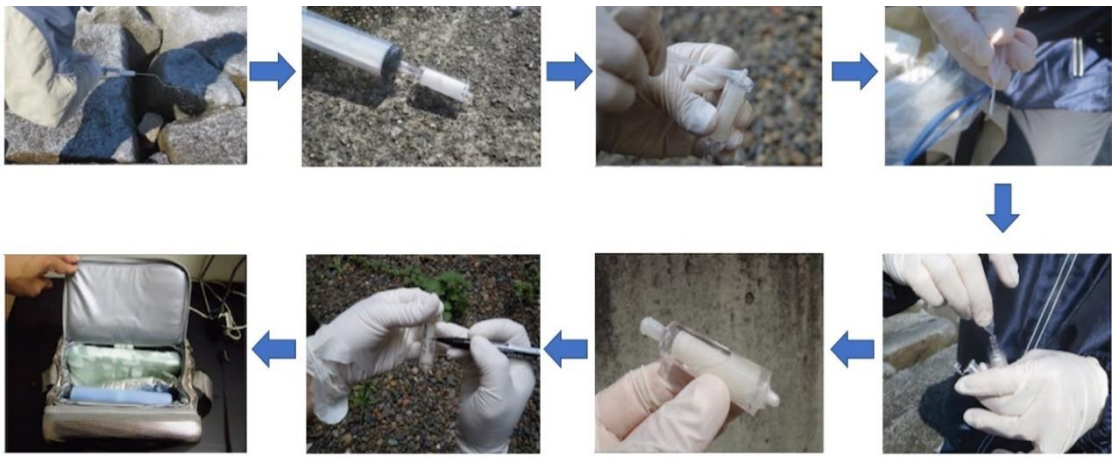
Figura 5 - Filtro Sterivex™ com conexão "*Luer Outlet*" mostrada na parte inferior do filtro em A; Seringa com bico "*Luer Lock*" (B) deve ser acoplada ao filtro como mostrado em C.

A B C

Fonte: Merck Millipore (202-?).

Filtros devem ser coletados em triplicata e a cada filtro deve ser dado um número identificador único. Escrever em um saco *whirl-pak* novo informações sobre o local da coleta. Os três filtros do mesmo local devem ser colocados em um saco whirl-pak que é em seguida fechado. O saco é então colocado em uma sacola térmica contendo gelo químico que foi congelado de um dia para o outro a -20ºC (**Figura 6**).

Figura 6 - Etapas para coleta de amostra com filtro Sterivex™



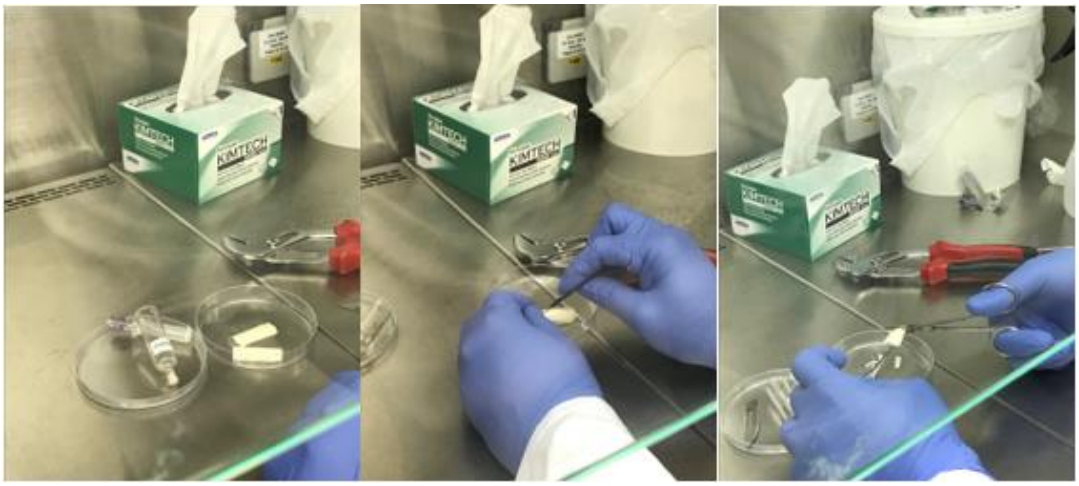
Fonte: THE EDNA SOCIETY, 2019

3.2 EXTRAÇÃO DE DNA

O processo de extração de DNA deve ser realizado em capela de fluxo laminar, previamente limpa e esterilizada com UV para evitar contaminação. Além disso, controles negativos devem ser utilizados. A extração de DNA deve ocorrer para cada tubo Sterivex (Millipore) que é aberto utilizando um alicate esterilizado. O filtro é retirado do tubo e cortado em pequenas seções, como

demostrado na (**Figura 7**). Em seguida, devem ser realizados os passos 3 a 11 da seção 2.2, com a diferença de adição de 50 µL de solução C6, no passo 10.

Figura 7 - Processo de remoção do filtro do Sterivex™ (Millipore) para extração de DNA



Fonte: Costa *et al.,* 2021

3.3 AMPLIFICAÇÃO DE MARCADORES GENÉTICOS EM DNA TOTAL EXTRAÍDOS DE AMOSTRAS DE ÁGUA

Para a realização da amplificação da região de interesse pela PCR, são utilizados os pares de *primers* L1091 (AAAAAGCTTCAAACTGGGATTAGATACCCCACTAT) e H1478 (TGACTGCAGAGGGTGACGGGCGGTGTGT) (YANG et al., 2014), com adaptadores para a plataforma Illumina MiSeq ou NextSeq 2000. Esse conjunto de *primer* permite amplificar fragmentos de DNA pertencentes ao gene mitocondrial 12S DNAr. As amplificações devem ser realizadas em triplicatas para cada amostra e combinadas na etapa da primeira purificação. Controles

negativos devem ser feitos para cada PCR realizada. Para a primeira PCR, são utilizados 2,5 µL de 5x Buffer, 1,2 µL de 25 mM $MgCl_2$, 1 µL de 2 mM dNTP, 0,25 µL do primer L1091 10 pmol, 0,25 µL do primer H1478 10 pmol, 0,1 µL de Taq, 4,2 µL de H_2O ultrapura e 3 de DNA, totalizando um volume de 12,5 µL para cada amostra. As PCRs são então colocadas no termociclador (e.g, Veriti 96) com o seguinte programa: desnaturação inicial a 94°C por 3 minutos, seguida por 30 ciclos de amplificação com 1 minuto a 94°C, temperatura de anelamento de 57°C a 1 minuto, em seguida 1 minuto a 72ºC e extensão final por 2 minutos a 72ºC.

Assim como descrito na seção 2.3, antes de preparar a solução estoque deve-se centrifugar o tubo do *primer* liofilizado por 30 segundos. A quantidade em nmols é então multiplicada por 10 e o valor da multiplicação é a quantidade em µL de água ultrapura que precisa ser adicionada ao *primer* liofilizado. Recomenda-se vortexar por 30 segundos a 1 minuto para ter certeza de que o *primer* esteja bem dissolvido. Para preparar a solução que vai ser adicionada na PCR, deve-se fazer a diluição do primer estoquer para 10 µM, utilizando 10 µL da solução estoque e 90 µL de água ultrapura.

3.4 PREPARO DA BIBLIOTECA PARA O SEQUENCIAMENTO

Recomendamos para a construção da biblioteca utilizar o protocolo 16S Metagenomic Sequencing Library Preparation da Illumina (Illumina, San Diego, CA, USA), com modificação somente dos *primers* para a construção da biblioteca referente ao gene mitocondrial 12S DNAr. Os *primers* modificados (L1091 e H1478) contêm sequências nucleotídicas chamadas *overhang*, que permitem a adição dos *indexes* nas etapas posteriores. Para as próximas etapas, devem ser seguidos os memos procedimentos descritos na seção 2.4.

4 SEQUENCIAMENTO ILLUMINA

Após o a preparação do *pool* genômico, conforme descrito na seção 3.3, o mesmo deve ser unido ao pool de PhiX (biblioteca padrão da Illumina utilizada para aumentar a diversidade de fragmentos nas corridas de sequenciamento). Para sequenciamento de eDNA *metabarcoding*, a concentração de PhiX utilizada é de 30%. Posteriormente, é utilizado 9 µL do *pool* e 15 µL de RSB (*Resuspension Buffer*) para a diluição na concentração apropriada (750 pM). Por fim, é realizado o sequenciamento de *amplicons* na plataforma Illumina NextSeq 2000 (**Figura 8**), demonstrado na, utilizando os kits de corrida NextSeq 2000 P1 (60 Gb) ou P2 (180 Gb) (2 x 300bp).

Figura 8 - Plataforma NextSeq 2000 System Illumina

Fonte: Illumina (2023)

A seguir, é descrito o passo-a-passo de execução do carregamento do pool genômico para sequenciamento de acordo com o protocolo Guia do Sistema NextSeq 2000 (Apêndice A):

- ✓ **1° Passo**: Preparar o cartucho de reagente pré-carregado para uso, agitando-o cuidadosamente, invertendo-o 10 vezes para garantir uma mistura adequada dos reagentes.
- ✓ **2° Passo**: Se for necessário, dilua a biblioteca para 2 nM. Com uma pipeta P1000, perfure o reservatório da biblioteca e pressione as extremidades do fecho metálico para ampliar o orifício. Em seguida, coloque 20 µL da mistura da biblioteca no reservatório designado do cartucho de reagente. Descarte a ponta da pipeta para prevenir contaminação.
- ✓ **3° Passo**: Limpe e seque cuidadosamente a lâmina de fluxo (*flow cell*). Insira a mesma no compartimento frontal do cartucho, pressionando firmemente para assegurar uma inserção adequada.
- ✓ **4° Passo**: Na interface do *software*, selecione *Load* (Carregar). O *software* de controle do NextSeq 2000 abrirá o visor e ejetará a bandeja. Insira o cartucho de reagente na bandeja do NextSeq 2000 com a etiqueta virada para cima e a lâmina de fluxo devidamente posicionada dentro do cartucho. Empurre o cartucho até que ele se encaixe de maneira adequada no local designado.
- ✓ **5° Passo**: Na interface do *software*, selecione *Close* (Fechar) para recolher o cartucho e fechar o visor. O software do NextSeq 2000 mostrará informações sobre os suprimentos digitalizados após dois minutos.
- ✓ **6° Passo** [opcional]: Na interface do *software*, selecione *Eject Cartridge* (Ejetar Cartucho) para retirar o cartucho. O visor se abrirá automaticamente após um minuto e ejetará o cartucho.

- ✓ **7° Passo**: Na interface do *software*, selecione *Settings* (Configurações) para iniciar as etapas de configuração da execução ou conecte-se ao BaseSpace ou ao BaseSpace Onsite.
- ✓ **8° Passo**: Revise os parâmetros da execução e os resultados da verificação pré-execução e, em seguida, selecione *Sequence* (Sequenciar) na interface do software.

5 ANÁLISES DE BIOINFORMÁTICA

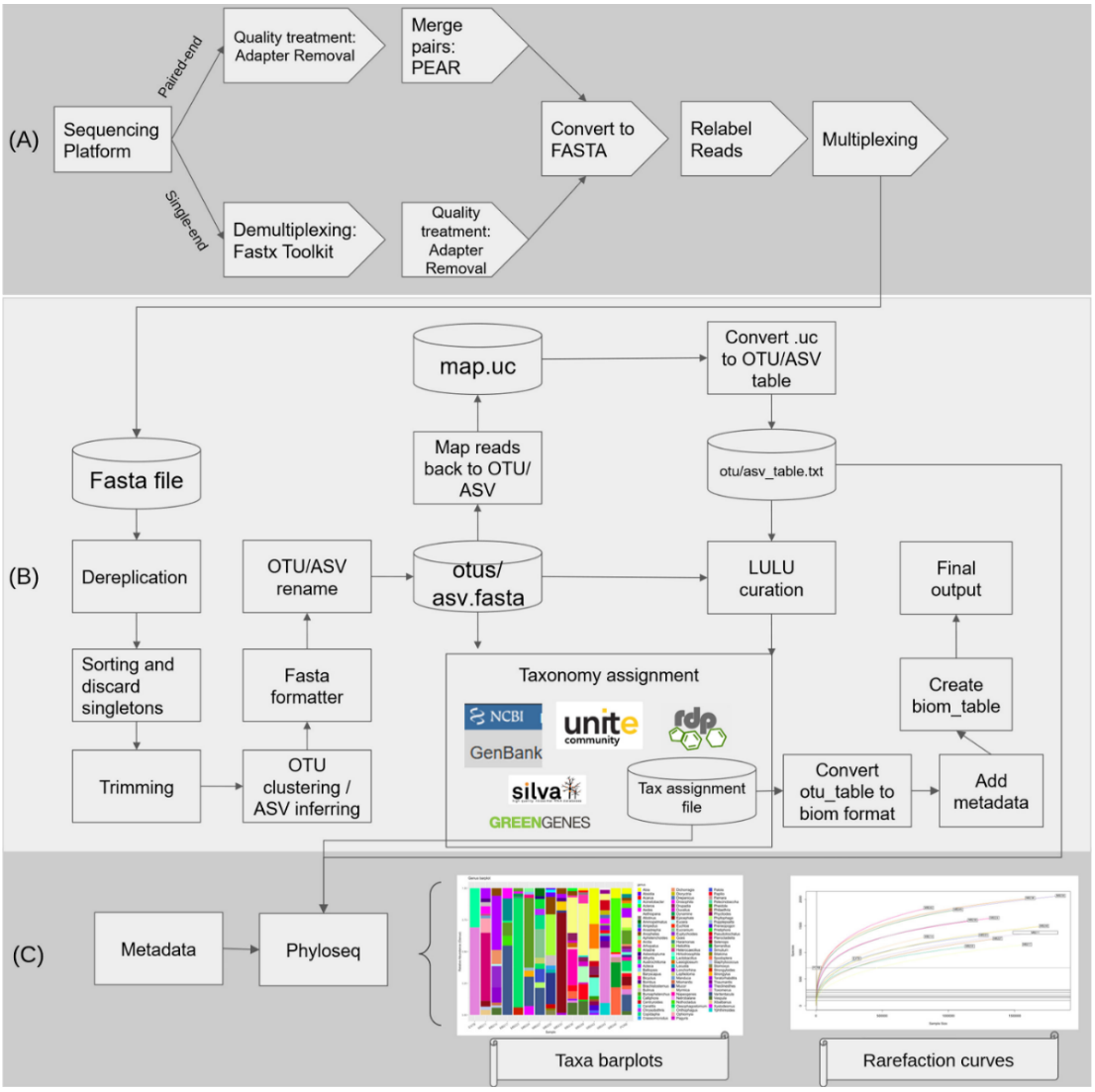
Recomenda-se que as sequências brutas geradas pelo sequenciamento, também chamadas de leituras ou *reads*, sejam submetidas ao PIMBA (*A Pipeline for MetaBarcoding Analysis*) (OLIVEIRA et al., 2021), um *pipeline* desenvolvido para análise de dados de eDNA *metabarcoding*.

O PIMBA foi desenvolvido pelo grupo de bioinformática do Instituto Tecnológico Vale – Desenvolvimento Sustentável e tem como base o QIIME (*Quantitative Insights Into Microbial Ecology*) (BOLYEN et al., 2019; CAPORASO et al., 2010), um pipeline referência desenvolvido para estudos de *metabarcoding*. O diferencial é que o PIMBA permite a utilização de bancos de dados taxonômicos distintos, desde os mais tradicionais, como o SILVA (QUAST et al., 2013), RDP (COLE et al., 2014) , Greengens (DESANTIS et al., 2006), UNITE (ABARENKOV et al., 2010), etc; até bancos próprios, personalizados ou baixados, como BOLD (RATNASINGHAM; HEBERT, 2007) e Genbank (BENSON et al., 2013). A escolha do banco de dados é uma etapa muito importante na análise de eDNA *metabarcoding* pois depende do táxon alvo (plantas, animais, microrganismos etc.) e do marcador utilizado (ITS, COI, 12S DNAr, 16S DNAr etc.).

O PIMBA é composto por várias etapas: (1) tratamento de qualidade das sequências, (2) classificação taxonômica, (3) reconstrução filogenética e (4) análise de diversidade. Essas etapas foram adaptadas para que possam ser concluídas em apenas três comandos de linha, referentes a três módulos (**pimba_prepare.sh**, **pimba_run.sh** e **pimba_plot.sh**), cujos parâmetros podem ser ajustados conforme a necessidade do usuário (**Figura 9**). O download deste pipeline e seu manual de uso estão disponíveis no GitHub (https://github.com/reinator/pimba, Apêndice A), com instruções para executá-lo

nos sistemas operacionais baseados em LINUX, como Ubuntu, MacOS ou Windows com subsistema Linux ativado (WSL2).

Figura 9 - Fluxograma simplificado mostrando as etapas de análise do pipeline PIMBA: (A) pimba_prepare, etapa de tratamento de qualidade e trimagem das sequências. (B) pimba_run, envolve a remoção de reads duplicadas, agrupamento por OTUs ou ASVs, remoção de quimeras e classificação taxonômica. (C) pimba_plot, com análises estatísticas básicas realizadas pelo pacote phyloseq do R.



Fonte: Oliveira *et al.,* 2021.

Resumidamente, o PIMBA fará a etapa de trimagem (remoção de adaptadores) e filtragem por qualidade (Phred >20, *default*) utilizando a ferramenta Adapter Removal (SCHUBERT; LINDGREEN; ORLANDO, 2016). Posteriormente, somente sequências *forward* e *reverse* de alta qualidade serão montadas através do montador Pear (ZHANG et al., 2014). Após a montagem, ocorre a etapa de derreplicação para a remoção das sequências duplicadas e em seguida a exclusão de sequências com cópia única. Com o intuito de melhorar a qualidade dos dados, as sequências menores que 100 pares de base (pb) (*default*) são descartadas (ajustado de acordo com o marcador molecular utilizado).

Em seguida, o usuário pode optar em seguir com a abordagem onde as sequências são agrupadas em Unidades Taxonômicas Operacionais (OTUs) utilizando o VSEARCH (ROGNES et al., 2016) ou são inferidas as *Amplicon Sequence Variants* (ASVs) usando a ferramenta SWARM (MAHÉ et al., 2015). O agrupamento das sequências em OTUs é usado para classificar grupos de indivíduos com similaridade > 97% (essa porcentagem depende do grupo de organismo a ser trabalhado ou marcador molecular utilizado). Caso se opte por utilizar ASVs, o usuário deve informar o número máximo de nucleotídeos que podem ser diferentes (o *default* é de apenas uma base nucleotídica). Finalmente, as OTUs ou ASVs são comparadas com sequências de referência disponíveis em bancos de dados públicos como o NCBI (https://www.ncbi.nlm.nih.gov), ou bancos específicos como o ITVBiobase, um banco de dados que contêm sequências de plantas (*mat*K, *rbc*L, *rpo*B, and *rpo*C1, *atp*F-*atp*H, *psb*K-*psb*I and *trn*H-*psb*A e ITS2) (VASCONCELOS et al., 2021), invertebrados (COI), morcegos (COI) e peixes (12S e 16S DNAr) que

ocorrem em Carajás. Esse banco foi devidamente construído e curado por especialistas taxonomistas com foco na área de atuação da Vale.

Por fim, o usuário pode realizar o posicionamento filogenético com as OTUs ou ASVs encontradas, caso possua uma árvore de constrição de interesse. Todas as etapas do pipeline serão explicadas com mais detalhes nas seções seguintes.

5.1 PASSO A PASSO CONTENDO TODAS AS ETAPAS E COMANDOS A SEREM UTILIZADOS NO PIMBA

Para habilitar o WSL2 no windows, siga as instruções do seguinte link: https://marcelo-albuquerque.medium.com/como-instalar-o-wsl-2-no-windows-10-3e26d99d7161.

A única exigência para poder executar o PIMBA é ter o Docker instalado. Para isso, abra o terminal no seu computador e rode o seguinte comando:

```
sudo apt-get install docker.io
```

Após isso, você precisa baixar os scripts do PIMBA (**pimba_prepare.sh, pimba_run.sh, pimba_plot.sh, pimba_tax.sh, pimba_place.sh, adapters.txt** e **databases.txt**) na página https://github.com/reinator/pimba. (configurar o arquivo adapters.txt de acordo com os adaptadores utilizados na etapa de sequenciamento, vide https://github.com/reinator/pimba, Apêndice)

1. Criar uma pasta do projeto;

```
mkdir nome_projeto
```

2. Criar uma pasta chamada *rawdata* para alocar os dados brutos;

```
mkdir rawdata_dir
```

3. Dar permissão de acesso à pasta

```
chmod -R 777 rawdata_dir
```

Assim que os dados do sequenciamento estiverem disponíveis na pasta *rawdata* execute os comandos dos três passos abaixo. Lembrar que para executar o comando, o usuário deve estar dentro da pasta **nome_projeto**

5.1.1 Passo 1: Preparar os dados (pimba_prepare.sh)

Esta é a etapa para executar os *scripts* de tratamento de qualidade nos dados sequenciados pela plataforma Illumina, usando o seguinte comando:

```
./pimba_prepare.sh illumina <rawdata_dir> <output_reads> <num_threads>
<adapters.txt> <min_length> <min_phred>
```

Parâmetros:

<rawdata_dir> = caminho dos arquivos R1 e R2 *reads*;

<output_reads> = nome da pasta dos resultados;

<num_threads> = número de *threads*;

<adapters.txt> = arquivo contendo os adaptadores usados no sequenciamento;

<min_length> = tamanho mínimo da *read* após o tratamento de qualidade;

<min_phred> = menor qualidade PHRED da *read* após o tratamento de qualidade.

Exemplo:

```
./pimba_prepare.sh illumina rawdata_illumina/ AllSamples 24
adapters.txt
```

5.1.2 Passo 2: Rodar a análise de clusterização (OTU ou ASV) e de classificação taxonômica (pimba_run.sh)

O arquivo *fasta* que vai ser gerado na etapa anterior é mandatório nesta etapa 2. Portanto, é importante sempre verificar se o arquivo foi gerado.

O comando padrão para rodar a análise de clusterização com OTUs ou ASVs é o seguinte:

```
./pimba_run.sh  -i <input_reads> -o <output_dir> -w <approach> -s
<otu_similarity> -a <assign_similarity> -c <coverage> -l <otu_length>
-h <hits_per_subject> -g <marker_gene> -t <num_threads> -e <E-value> -
d <databases.txt> -x <run_lulu>
```

Parâmetros

-i <input_reads> = arquivo fasta contendo as *reads* geradas pelo pimba_prepare.sh;

-o <output_dir> = diretório com os resultados;

-w = estratégia a ser utilizada. Pode ser ‘otu’ ou ‘asv’. Se for ‘otu’, o PIMBA usa o programa Vsearch. Se for ‘asv’, o PIMBA usa o programa swarm. *Default*: ‘otu’;

-s <otu_similarity> = porcentagem de similaridade utilizada para clusterização das OTUs. *Default* é 0.97;

-a <assign_similarity> = porcentagem de similaridade utilizada para a atribuição taxonômica. *Default* é 0.9;

-c <coverage> = cobertura mínima para alinhamento. *Default* é 0.9;

-l <otu_length> = comprimento da trimagem das reads. Se 0, as *reads* não serão trimadas;

-h <hits_per_subject> = se 1, escolhe o melhor *best hit*. Se > 1, escolhe pela maioria. *Default* é 1;

-g <marker_gene> = gene marcador e *database*. Pode ser: (16S-SILVA, 16S-GREENGENES, 16S-RDP, 16S-NCBI, ITS-FUNGI-NCBI, ITS-FUNGI-UNITE, ITS-PLANTS-NCBI, COI-NCBI, COI-BOLD 12S-BIOBASE).

-t <num_threads> =número de *threads* usado para o *blast*. *Default* é 1;

-e = *e-value* usado para o *blast*. *Dafault* é 0.0001;

-d <databases_file.txt> = arquivo contendo o caminho da base de dados;

-x = se o parâmetro for 'lulu', o PIMBA vai descartar OTUs ou ASVs errôneos usando o software LULU. O *Default* é não usar o LULU.

Exemplo:

```
./pimba_run.sh -i AllSamples.fasta -o 12Sbiobase -w otu -s 0.97 -a 0.7
-c 0.8 -l 100 -h 1 -g 12S-BIOBASE  -t 24 -e 0.1 -d databases.txt -x
lulu
```

5.1.3 Passo 3: Gráficos de diversidade alfa e beta (pimba_plot.sh)

Quando terminar o pimba_run.sh, você será capaz de gerar alguns gráficos básicos para seus resultados, como PCoA, curvas de rarefação e gráficos de diversidade alfa e beta. Tudo o que você precisa são dois arquivos que o pimba_run.sh irá gerar (otu_table.txt e tax_assignment) e um arquivo de metadados (tabela em formato csv) que deverá ser fornecido (**Figura 10**)**.** Em posse desses arquivos, o seguinte comando poderá ser executado:

```
./pimba_plot.sh -t <otu_table> -a <tax_assignment> -m <metadata> -g
<group_by>
```

Parâmetros:

-t <otu_table> = tabela OTU gerada por pimba_run;

-a <tax_assignment> = arquivo de atribuição taxonômica gerado por pimba_run;

-m = arquivo CSV com colunas "sample_id" (primeira coluna contendo os Ids das amostras) e outros atributos relacionados a cada amostra (local, estação do ano etc.);

-g = indica uma coluna da tabela de metadados onde se pode agrupar as informações a fim de gerar os gráficos (por exemplo, local de coleta). Se não quiser agrupar os resultados, não o especifique.

Exemplo:

```
./pimba_plot.sh -t AllSamples_otu_table.txt -a
AllSamples_otus_tax_assignments.txt -m mapping_file.csv -g local
```

Figura 10 - Exemplos de arquivos finais gerados pelas etapas de análise do pipeline PIMBA. Esses são os arquivos que podem ser utilizados em análises estatísticas posteriores.

Example of OTU table (otu_table.txt) generated by pimba_run.sh:

OTUId	EXTB	MB211	MB214	MB217	MB223	MB224	MB227	MB230	MB233	MB236	MB239	MB242	MB245	MB248	PCRB
OTU955365479	74371	116	24	52	236	98	147	19	45	1039		2355		60	16	3	20
OTU812100800	155496	225	32	88	86	152	525	17	147	329	126	308	73	9	35		
OTU789743117	54074	21	15	22	20	48	191	8	11	329	14	2	71	4	14		
OTU645601205	8938	4	2	7	3	5	27	2	4	10	6	1	8	0	1		
OTU212911189	2037	2	0	0	0	0	2	0	0	1	0	1	0	0	0		
OTU964442511	140	76	82	227	64	128	291	135	94	373557	265	167	957	113	76		
OTU533344192	164	133	62	265895	97	67	246	43	55	444	106	31	124	51	92		
OTU103992458	724	0	2	1	0	1	2	0	6	1	1	9	0	0	0		
OTU524853145	530	389	212	799483	347	3301		722	142	183	1150		284	118	377	165	236
OTU844428705	310	310	180	390	320	253	350	518	214	353599	610124	64796		39526		1131	212
OTU293268522	270	37	28	92	1007		120	121096	48	14	51	33	71	71	76	18	
OTU884692304	253	0	0	1	0	0	3	0	0	0	0	0	0	0	0		
OTU986112541	434	0	0	0	0	0	1	0	0	0	2	0	0	0	0		
OTU176656043	97	0	0	0	0	0	2	0	0	0	0	0	0	0	0		
OTU689340459	3	0	0	0	0	0	2	1	0	296	0	1	3	0	1		
OTU886162264	411	2	0	3	2	1	0	0	1	0	1	2	3	0	1		
OTU771688223	499	0	0	0	1	0	1	0	2	1	1	0	0	0	0		
OTU484930833	8	10	6	4	14	3	1	11940		272	3	4	5	2	9	0	
OTU467557698	1	0	0	0	0	0	0	0	0	202	5	243	0	0	0		

Example of OTU Tax assignment (tax_assignment.txt):

OTU698595363	k__Metazoa; p__Arthropoda; c__Insecta; o__Lepidoptera; f__Eriocraniidae; g__Eriocrania; s__sp.ER06	0.96154
OTU766111137	k__Metazoa; p__Arthropoda; c__Insecta; o__Diptera; f__Calliphoridae; g__Huascaromusca; s__sp.1MATM-2015	0.94574
OTU670530351	k__Metazoa; p__Arthropoda; c__Insecta; o__Lepidoptera; f__Nymphalidae; g__Hypothyris; s__moebiusimoebiusi	0.96154
OTU120497499	k__Metazoa; p__Arthropoda; c__Insecta; o__Diptera; f__Culicidae; g__Anopheles; s__koliensis	0.84733
OTU977420622	k__Metazoa; p__Arthropoda; c__Insecta; o__Lepidoptera; f__Nymphalidae; g__Ideopsis; s__gaura	0.95385
OTU374660010	k__Metazoa; p__Arthropoda; c__Arachnida; o__Scorpiones; f__Buthidae; g__Isometrus; s__maculatus	0.87692
OTU8660394	k__Metazoa; p__Arthropoda; c__Insecta; o__Diptera; f__Empididae; g__Rhamphomyia; s__sp.1BKC-2015	0.96899
OTU261821968	k__Metazoa; p__Arthropoda; c__Insecta; o__Diptera; f__Syrphidae; g__Toxomerus; s__apegiensis	0.95349
OTU302289343	k__Metazoa; p__Arthropoda; c__Insecta; o__Lepidoptera; f__Lycaenidae; g__Allotinus; s__davidis	0.84733
OTU412412883	k__Metazoa; p__Arthropoda; c__Insecta; o__Diptera; f__Drosophilidae; g__Zygothrica; s__ptilialis	0.94615
OTU285620827	k__; p__; c__Florideophyceae; o__Gigartinales; f__Peyssonneliaceae; g__Riquetophycus; s__sp.HSY-2014a	0.91473
OTU288350188	k__Metazoa; p__Arthropoda; c__Insecta; o__Lepidoptera; f__Nymphalidae; g__Hypomenitis; s__alphesiboea	0.96124
OTU425378105	k__Metazoa; p__Nematoda; c__Chromadorea; o__Tylenchida; f__Aphelenchoididae; g__Bursaphelenchus; s__paraluxuriosae	0.89552
OTU733572620	k__Metazoa; p__Arthropoda; c__Insecta; o__Diptera; f__Syrphidae; g__Toxomerus; s__difficilis	0.96923
OTU240332622	k__Metazoa; p__Arthropoda; c__Insecta; o__Diptera; f__Culicidae; g__Anopheles; s__nuneztovari	0.93846

Example of Metadata file (metadata.csv):

```
SampleID,SampleName,BarcodeSequence,LinkerPrimer,Description
EXTB,EXTB,CGTACTAG+GCGTTAGA,CTGTCTCTTATACACATCT,Branco
PCRB,PCRB,CGTACTAG+TCTTACGC,CTGTCTCTTATACACATCT,Branco
MB211,MB211,AGGCAGAA+TCTTACGC,CTGTCTCTTATACACATCT,Cav1
MB214,MB214,TCCTGAGC+TCTTACGC,CTGTCTCTTATACACATCT,Cav1
MB217,MB217,GGACTCCT+TCTTACGC,CTGTCTCTTATACACATCT,Cav1
MB223,MB223,TAGGCATG+TCTTACGC,CTGTCTCTTATACACATCT,Cav2
MB224,MB224,CTCTCTAC+TCTTACGC,CTGTCTCTTATACACATCT,Cav2
MB227,MB227,CAGAGAGG+TCTTACGC,CTGTCTCTTATACACATCT,Cav2
MB230,MB230,GCTACGCT+TCTTACGC,CTGTCTCTTATACACATCT,Cav3
MB233,MB233,CGAGGCTG+TCTTACGC,CTGTCTCTTATACACATCT,Cav3
MB236,MB236,AAGAGGCA+TCTTACGC,CTGTCTCTTATACACATCT,Cav3
MB239,MB239,GTAGAGGA+TCTTACGC,CTGTCTCTTATACACATCT,Cav4
MB242,MB242,TAAGGCGA+ATAGAGAG,CTGTCTCTTATACACATCT,Cav4
MB245,MB245,CGTACTAG+ATAGAGAG,CTGTCTCTTATACACATCT,Cav4
MB248,MB248,AGGCAGAA+ATAGAGAG,CTGTCTCTTATACACATCT,Cav4
```

Fonte: Oliveira *et al*., 2021

Esses são os três principais passos para se ter resultados ao analisar dados de DNA *metabarcoding* com o PIMBA. Existem outros dois módulos opcionais que o usuário pode executar e que serão descritos a seguir.

5.1.4 Passo opcional 1: Realizar apenas atribuição taxonômica (pimba_tax.sh)

Caso o usuário já possua um arquivo fasta contendo as OTUs ou ASVs geradas pelo **pimba_run.sh**, assim como a sua respectiva OTU/ASV table, ele poderá obter diferentes atribuições taxonômicas, variando apenas o banco de dados de referência a ser utilizado. O comando a ser utilizado é mostrado abaixo:

```
./pimba_tax.sh -i <otus_fasta> -u <otu_table> -o <output_dir> -a
<assign_similarity> -c <coverage> -h <hits_per_subject> -g
<marker_gene> -t <num_threads> -e <E_value> -d <databases_file.txt>
```

Parâmetros:

-i <otus_fasta> = arquivo fasta contendo as *OTUs* ou ASVs geradas pelo pimba_run.sh;

-u <otu_table> = arquivo contendo a OTU/ASV table;

-o <output_dir> = diretório com os resultados;

-a <assign_similarity> = porcentagem de similaridade utilizada para atribuição taxonômica. *Default* é 0.9;

-c <coverage> = cobertura mínima para alinhamento. *Default* é 0.9;

-h <hits_per_subject> = se 1, escolhe o melhor *best hit*. Se > 1, escolhe pela maioria. *Default* é 1;

-g <marker_gene> = gene marcador e *database*. Pode ser: (16S-SILVA, 16S-GREENGENES, 16S-RDP, 16S-NCBI, ITS-FUNGI-NCBI, ITS-FUNGI-UNITE, ITS-PLANTS-NCBI, COI-NCBI, COI-BOLD, 12S-BIOBASE);

-t <num_threads> = número de *threads* usado para o *blast*. *Default* é 1;

-e = e-value usado para o *blast*. *Dafault* é 0.0001;

-d <databases_file.txt> = arquivo contendo o caminho da base de dados.

Exemplo:

```
./pimba_tax.sh -i otus.fasta -u otu_table.txt -o
AllSamplesCOI_98clust90assign -a 0.9 -c 0.9 -h 1 -g 16S-GREENGENES -t
24 -e 0.1 -d databases.txt
```

5.1.5 Passo opcional 2: Realizar posicionamento filogenético (pimba_place.sh)

Adicionalmente, o módulo **pimba_place.sh** permite realizar análise de posicionamento filogenético com as OTUs ou ASVs geradas pelo **pimba_run.sh**. A partir de uma árvore de constrição e seus respectivos arquivos de alinhamento, é possível gerar um arquivo JPLACE contendo o posicionamento filogenético das OTUs/ASVs encontradas, com o seguinte comando:

```
./pimba_place.sh -i <input_name> -c <constraint_tree_newick>  -a
<constraint_alignment> -x <constraint_tree_taxonomy> -t <num_threads>
-d <sequence_type> -o <output_dir>
```

Parâmetros:

-i <input_name> = arquivo fasta com as sequências a serem posicionadas na árvore de constrição;

-c <constraint_tree> = arvore de constrição no formato Newick;

-a <constraint_alignment> = arquivo fasta com as sequências alinhadas da árvore de contrição;

-x <constraint_tree_taxonomy> = arquivo separado por tabulação contendo os identificadores das sequências presentes na árvore de constrição e sua respectiva taxonomia;

-t <num_threads> = número de threads a serem utilizadas;

-d <sequence_type> = tipo de sequências a serem analisadas. Pode ser 'nt' or 'aa';

-o <output_dir> = diretório onde serão armazenados os resultados gerados.

Exemplo:

```
./pimba_place.sh -i data/AllSamples_otus_plants.fasta -c
data/rbcL_ITS2_alignment.newick  -a data/rbcL_ITS2_alignment.fasta -x
data/taxonpath.tsv -t 5 -d nt -o output
```

O arquivo JPLACE gerado será salvo no diretório <output_dir>/no_clustering/placed/<input_name>.jplace

REFERÊNCIAS

ABARENKOV, K. *et al.* **The UNITE database for molecular identification of fungi – recent updates and future perspectives**. **The New Phytologist**WileyNew Phytologist Trust, 2010. Disponível em: https://www.jstor.org/stable/27797548. Acesso em: 22 ago. 2020

BENSON, D. A. *et al.* GenBank. **Nucleic Acids Research**, v. 41, n. D1, p. D36–D42, 1 jan. 2013.

BOLYEN, E. *et al*. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. **Nature Biotechnology**, v. 37, n. 8, p. 852–857, 24 jul. 2019.

CAPORASO, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. **Nature Methods**, v. 7, n. 5, p. 335–336, 11 maio 2010.

CATARINA, V. *et al.* Monitoramento da biodiversidade da flora de canga, serra dos carajás, Pará, através de DNA metabarcoding. Belém: 2021. (Relatório Técnico N003/2021) DOI 10.29223/PROD.TEC.ITV.DS.2021.03.Martins Acesso em: 1 jan. 2023.

COLE, J. R. *et al.* Ribosomal Database Project: data and tools for high throughput rRNA analysis. **Nucleic Acids Research**, v. 42, n. D1, p. D633–D642, 1 jan. 2014.

COSTA, P. *et al.* Manual para o monitoramento da ictiofauna por meio de dna ambiental (eDNA). Belém: 2021. (Relatório Técnico N024/2020) DOI 10.29223/PROD.TEC.ITV.DS.2020.24.Costa Acesso em: 1 jan. 2023.

CRISTESCU, M. E. From barcoding single individuals to metabarcoding biological communities: Towards an integrative approach to the study of global biodiversity. **Trends in Ecology and Evolution,** v. 29, n. 10, p. 566-571, out. 2014. Acesso em: 3 mar. 2021

DEINER, K. *et al*. Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. **Molecular Ecology**, v. 26, n. 21, p. 5872–5895, 1 nov. 2017.

DESANTIS, T. Z. *et al*. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. **Applied and Environmental Microbiology**, v. 72, n. 7, p. 5069–5072, 1 jul. 2006.

ELLWANGER, J. H.; NOBRE, C. A.; CHIES, J. A. B. Brazilian Biodiversity as a Source of Power and Sustainable Development: A Neglected Opportunity. **Sustainability 2023, Vol. 15, Page 482**, v. 15, n. 1, p. 482, 28 dez. 2022.

FAHNER, N. A. *et al.* Large-Scale Monitoring of Plants through Environmental DNA Metabarcoding of Soil: Recovery, Resolution, and Annotation of Four DNA Markers. **PLOS ONE**, v. 11, n. 6, p. e0157505, 1 jun. 2016.

FAZEKAS, A. J. *et al.* Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? **Molecular Ecology Resources**, v. 9, n. SUPPL. 1, p. 130–139, maio 2009.

FICETOLA, G. F. *et al.* Species detection using environmental DNA from water samples. **Biology Letters**, v. 4, n. 4, p. 423–425, 23 ago. 2008.

HEBERT, P. D. N. *et al.* Biological identifications through DNA barcodes. **Proceedings of the Royal Society of London B: Biological Sciences**, v. 270, n. 1512, 2003.

KRESS, W. J.; ERICKSON, D. L. DNA barcodes: Methods and protocols. **Methods in Molecular Biology**, v. 858, p. 3–8, 2012.

MAHÉ, F. et al. Swarm v2: highly-scalable and high-resolution amplicon clustering. **PeerJ**, v. 3, n. 12, p. e1420, 10 dez. 2015.

OLIVEIRA, R. R. M. *et al.* PIMBA: A PIpeline for MetaBarcoding Analysis. p. 106–116, 2021.

PANARO, N. J. *et al.* Evaluation of DNA Fragment Sizing and Quantification by the Agilent 2100 Bioanalyzer. **Clinical Chemistry**, v. 46, n. 11, p. 1851–1853, 1 nov. 2000.

QUAST, C. *et al.* The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. **Nucleic Acids Research**, v. 41, n. D1, p. D590–D596, 1 jan. 2013.

RATNASINGHAM, S.; HEBERT, P. D. N. BARCODING: bold: The Barcode of Life Data System (http://www.barcodinglife.org). **Molecular Ecology Notes**, v. 7, n. 3, p. 355–364, 24 jan. 2007.

ROGNES, T. *et al.* VSEARCH: a versatile open source tool for metagenomics. **PeerJ**, v. 4, n. 10, p. e2584, 18 out. 2016.

SCHUBERT, M.; LINDGREEN, S.; ORLANDO, L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. **BMC Research Notes**, v. 9, n. 1, p. 1–7, 12 fev. 2016.

THE EDNA SOCIETY. Environmental DNA Sampling and Experiment Manual. 2019.

VASCONCELOS, S. *et al*. Unraveling the plant diversity of the Amazonian canga through DNA barcoding. **Ecology and Evolution**, v. 11, n. 19, p. 13348–13362, 1 out. 2021.

VOGEL ELY, C. *et al*. Implications of poor taxonomy in conservation. **Journal for Nature Conservation**, v. 36, p. 10–13, 1 abr. 2017.

YANG, L. *et al*. Species identification through mitochondrial rRNA genetic analysis. **Scientific Reports**, v. 4, n. 1, p. 1–11, 13 fev. 2014.

ZHANG, J. *et al.* PEAR: a fast and accurate Illumina Paired-End reAd mergeR. **Bioinformatics**, v. 30, n. 5, p. 614–620, 1 mar. 2014.

ZHAO, N.; WANG, Y.; HUA, J. The Roles of Mitochondrion in Intergenomic Gene Transfer in Plants: A Source and a Pool. **International Journal of Molecular Sciences,** v. 19, n. 2, p. 547, 11 fev. 2018.

APÊNDICE A

Kit de extração de DNA DNeasy PowerSoil
O manual pode ser obtido neste link:
https://www.qiagen.com/us/resources/resourcedetail?id=9bb59b74-e493-4aeb-b6c1-f660852e8d97&lang=en

Qubit™ 4 Fluorometer
O manual do Qubit™ 4 Fluorometer pode ser obtido neste link:
https://www.thermofisher.com/document-connect/document-connect.html?url=https://assets.thermofisher.com/TFS-Assets%2FLSG%2Fmanuals%2FMAN0017209_Qubit_4_Fluorometer_UG.pdf

Qubit® dsDNA HS Assay Kit
O manual do Qubit® dsDNA HS and BR Assay Kits pode ser obtido neste link:
https://www.thermofisher.com/document-connect/document-connect.html?url=https://assets.thermofisher.com/TFS-Assets%2FLSG%2Fmanuals%2FQubit_dsDNA_HS_Assay_UG.pdf

16S Metagenomic Sequencing Library Preparation
O manual do procedimento para a construção das bibliotecas pode ser obtido neste
link:
https://support.illumina.com/content/dam/illuminasupport/documents/documentation/chemistry_documentation/16s/16s-metagenomiclibrary-prep-guide-15044223-b.pdf

NextSeq 2000 Sequencing System Guide
O manual do NextSeq 2000 System pode ser obtido neste link:
https://www.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/nextseq-1000-2000-spec-sheet-m-na-00008/nextseq-1000-2000-spec-sheet-m-na-00008.pdf

Guia de instalação e uso do PIMBA:
https://github.com/reinator/pimba