UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR

CONSTRUÇÃO DE UM PIPELINE PARA ANÁLISE METAGENÔMICA

ADAN RODRIGUES DE OLIVEIRA

Belém
NOVEMBRO 2022

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR

CONSTRUÇÃO DE UM PIPELINE PARA ANÁLISE METAGENÔMICA

ADAN RODRIGUES DE OLIVEIRA

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Genética e Biologia Molecular da Universidade Federal do Pará como requisito parcial para obtenção de grau de Mestre em Genética com ênfase em Bioinformática.

Orientador: Prof. Dr. Diego Assis das Graças.

Belém
NOVEMBRO 2022

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

O48c Oliveira, Adan Rodrigues de.
CONSTRUÇÃO DE UM PIPELINE PARA ANÁLISE METAGENÔMICA / Adan Rodrigues de Oliveira. — 2022.
74 f. : il. color.

Orientador(a): Prof. Dr. Diego Assis das Graças
Dissertação (Mestrado) - Universidade Federal do Pará, Instituto de Ciências Biológicas, Programa de Pós-Graduação em Genética e Biologia Molecular, Belém, 2022.

1. Metagenômica. 2. Pipeline. 3. Usearch. 4. Python. 5. Tkinter. I. Título.

CDD 579.135

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico através da concessão de bolsa de mestrado. O projeto foi desenvolvido no Laboratório De Engenharia Biológica (ENGBIO) e no Centro De Genômica E Biologia De Sistemas (CCGBS), ambos nas instalações da Universidade Federal do Pará (UFPA).

AGRADECIMENTOS

A todos que estiveram presentes no decorrer deste trabalho, de qualquer forma ajudando e apoiando.

Ao professor e doutor Diego Assis Das Graças, por sempre estar orientando o melhor caminho para esta pesquisa.

A minha família, por estar sempre apoiando e motivando para que este trabalho fosse concluído com excelência.

A minha companheira e agora Mestre, Cynthia Letícia Serra Cabeça, por compreender e apoiar sempre que necessário.

Ao CNPQ por ter fornecido bolsa de Mestrado para ajudar nos custeios da pesquisa e permanência nas instalações dos laboratórios.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS 6
LISTA DE FIGURAS E TABELAS 10
RESUMO 11
ABSTRACT 12
1. INTRODUÇÃO 13
1.1 Metagenômica 15
1.2 Análise de Riqueza e Diversidade 24
1.2.1 Análise de CHAO1 26
1.2.2 Análise de SHANNON 27
1.2.3 Análise de ACE 27
1.2.4 Análise de Simpson e Simpson inverso 28
1.3 Linguagens De Programação e Pipelines em Bioinformática 28
1.4 Ferramentas de processamento de dados metagenômicos 31
2. OBJETIVOS 33
2.1 Objetivo Geral 33
2.2 Objetivos Específicos 33
3. MATERIAL E MÉTODOS 34
3.1 Considerações iniciais 34
3.3 Funcionamento Do Pipeline 35
4. RESULTADOS E DISCUSSÃO 36
4.1 Discussão 42
4.2 Perspectivas futuras 45
CONCLUSÃO 47
REFERÊNCIAS 48
ANEXOS 56

LISTA DE ABREVIATURAS E SIGLAS

16S	RNA Ribosomal Da Subunidade Menor Do Ribossomo.
antiSMASH	Programa Para Análise De Metabólitos Secundários Em Microrganismos.
API's	Application Programming Interface.
ASV	Amplicon Sequence Variant.
BLAST	Dispositivo De Alinhamento De Sequências.
BOWTIE2	Dispositivo De Alinhamento Para Sequências De Referência Longas.
bwa-mem	Algoritmo Para Alinhar Leituras De Extremidade Única Ou Emparelhada.
CLUSTAW	Clustaw É Um Programa De Alinhamento Múltiplo Para DNA ou Proteínas.
COGs	Clusters Of Orthologous Groups.
DADA2	Algoritmo Para Inferência De Amostras De Amplicon De Nucleotídeo Único.
DIAMOND	Algoritmo Para Alinhamento De Proteínas De Alto Rendimento.
eggNOG	Banco de dados para anotação funcional, de ontologia, e histórias de genes.
eggNOG-mapper	Ferramenta para anotação funcional de novas sequências.
EMBL-EBI	Instituto Europeu de Bioinformática.
ENA	European Nucleotide Archive.
FASTA	Formato De Arquivo Utilizado Para Armazenamento De Sequências.
FASTQ	Formato Para Armazenar Sequências E Suas Pontuações De Qualidade.
FragGeneScan	Aplicativo Para Encontrar Genes Em Leituras Curtas.
gene3D	Algoritmo Para Anotação Estrutural E Funcional De Famílias De Proteínas.
GO	Gene Ontology.
GO slim	Método Para Criação De Lista De Genes A Partir De Termos Selecionados.
GP	Utilitários Para Manipular Sequências De DNA/RNA/Proteína.

InterProScan	Algoritmo De Classificação De Família De Proteínas.
ITS	Internally Transcribed Spacer.
ITSoneDB	Banco de dados de ITS.
KALIGN	Algoritmo De Alinhamento De Sequência Múltipla.
KEGG	Kyoto Encyclopedia Of Genes And Genomes.
MAFFT	Software De Alinhamento Múltiplo De Aminoácidos ou Nucleotídeos.
MAGs	Metagenome-Assembled Genome.
MAPSeq	Algoritmo Para Classificação De Sequências Metagenômicas.
metaSPAdes	Algoritmo Para Montagem De Leituras Metagenômicas.
MGnify	Instituto Europeu De Bioinformática.
MOTHUR	Pacote De Software Para Processamento De Dados De Bioinformática.
mOTUs2	Ferramenta Para Perfil De OTU Baseado Em Gene Marcador.
MUSCLE	Multiple Sequence Comparison By Log- Expectation.
NCBI	National Center For Biotechnology Information.
NCBI Refseq DB	Banco De Dados De Sequência De Referência NCBI.
NGS	New Generation Sequencing.
OS	Operational System.
OTUs	Operational Taxonomic Unit.
PERL-CGI	Protocolo Para Execução De Scripts Por Meio De Solicitações Da Web.
Pfam	Base De Dados De Famílias De Proteínas.
PH	Escala Para Especificar A Acidez Ou Basicidade De Uma Solução.
Phred 33	O Índice de Qualidade De Uma Base.
PIP	Administrador De Instalação De Pacotes Da Linguagem Python.
Pipeline	Mapa Das Etapas Que Compõem Um Processo.

PLOTS	Produção De Uma Imagem Por Meio De Desenho De Linhas.
PRINTS	Compêndio De Impressões Digitais De Proteínas.
Prodigal	Software De Previsão De Genes.
PROSITE	Banco De Dados De Proteínas.
Qiime2-Deblur	Plugin (QIIME2) Que Realiza O Controle De Qualidade De Sequências.
QIIME-uclust	Algoritmo de interação com OTU's, presente no QIIME2.
R	Linguagem Voltada Para Processamento De Dados.
RAPSearch	Ferramenta Para Procura De Similaridade De Proteínas.
REST	Representational State Transfer.
Rfam	Banco De Informações Sobre Elementos De RNA Estruturado.
rRNA	RNA Ribossomal.
SeqPrep	Ferramenta De Tratamento De Sequências De Leituras.
Shell	Linguagem Presente Na Maioria Dos Sistemas Operacionais.
SOAP	Sistema Que Permite Comunicação De Processos Em OS Diferentes.
SPAdes	Algoritmo De Montagem De Genoma.
SSU e LSU	Regiões de rDNA.
T-COFFE	Software De Alinhamento Múltiplo De Sequências.
TIGRFAMs	Database Of Protein Families.
Tkinter	Pacote De Interface Para Linguagem De Programação Python.
Trimmomatic	Uma Ferramenta De Corte De Leitura Flexível Para Dados Illumina.
UBLAST	Algoritmo Para Pesquisa De Banco De Dados Para Alinhamentos.
UCLUST	Algoritmo que Agrupa Sequências De Nucleotídeos ou Aminoácidos.
UniRef90	Algoritmo Para Clusterização De Sequências.
UNITE	Ambiente de gerenciamento de banco de dados.

USEARCH	Ferramenta Para Análise De Sequências.
USEARCH-UNOISE3	Algoritmo Para Remoção De Erros Em Leituras De Amplicon.
USEARCH-UPARSE	Algoritmo Para Clusterização De OTU's.
ViPR/IRD DB	Virus Pathogen Database And Analysis Resource.
XML	Extensible Markup Language.

LISTA DE FIGURAS E TABELAS

Figura 1 - Segmentos da Bioinformática 14
Figura 2 - Locais de Recuperação dos MAG's para produção do catálogo "GEM". 19
Figura 3 - Incidência/Origem de microrganismos presentes em vagões e estações de trem. ... 20
Figura 4 - As 20 espécies mais abundantes em vagões de trem e em "Touch Screens". 21
Figura 5 - Esquema representando o esquema de automação para o pipeline final. 36
Figura 6 - Funcionamento do Metadoon Pipeline. 37
Figura 7 - Entrada de dados. 38
Figura 8 - Mesclando pares. 38
Figura 9 - Filtro de qualidade. 38
Figura 10 - Buscando leituras únicas. 39
Figura 11 - Organizando os grupos em OTU's. 39
Figura 12 - Montando tabelas de OTU's. 39
Figura 13 - Alinhando as sequências com banco de dados. 40
Figura 14 - Opções de bancos de dados. 40
Figura 15 - Opções de saída dos dados. 41

RESUMO

A metagenômica é normalmente aplicada para compreensão da análise de diversidade em um ambiente. O desenvolvimento da metagenômica deve-se principalmente ao crescimento do campo computacional, que consegue auxiliar o desenrolar desse estudo de maneira direta. A utilização e automação de rotinas computacionais se torna uma tendência nos laboratórios de estudos genéticos e metagenômicos tanto quanto a produção dos mesmos. Dessa forma, o presente trabalho visou a produção de um pipeline que execute análises metagenômicas (Bar-coding) de rRNA 16S, enfatizando estatísticas de diversidade e de abundancia. Através da integração de Python, R e Shell, e da interface gráfica "Tkinter", presente nos repositórios Python, os processos de analise foram desenhados. As rotinas automatizadas nesta ferramenta começam utilizando o Usearch para a preparação dos dados, mesclando os arquivos de entrada "fastq", aplicando o filtro de qualidade, buscando por leituras únicas, agrupando as sequências em OTUs, criando as tabelas de OTUs e a tabela de dados taxonômicos. Em seguida, executa análises de ACE, CHAO1, Shannon, Simpson e Simpson Inverso, utilizando o software R e, por último, o pacote "canvas" do python agrupa os resultados, plotando e montando um report final que possa auxiliar nas análises de resultados e na discussão dos mesmos. A produção do Metadoon pipeline consegue interligar as partes de uma pesquisa bioinformata-metagenômica (como; normalização e organização dos dados, pre processar, gerar gráficos) e ao mesmo tempo gerar reportes que auxiliam na análise do resultado.

Palavras – Chave: Metagenômica, Pipeline, Usearch, Python, Tkinter.

ABSTRACT

Metagenomics is normally applied to understand the analysis of diversity in an environment. The development of metagenomics is mainly due to the growth of the computational field, which manages to help the development of this study in a direct way. The use and automation of computational routines becomes a trend in laboratories for genetic and metagenomic studies as much as their production. Thus, the present work aimed to produce a pipeline that performs metagenomic analyzes (Bar-coding) of 16S rRNA, focusing on diversity and abundance statistics. Through the integration of Python, R and Shell, and the "Tkinter" graphical interface, present in the Python repositories, the analysis processes were designed. The automated routines in this tool start using Usearch for data preparation, merging the “fastq” input files, applying the quality filter, searching for unique reads, grouping the sequences into OTUs, creating the OTU tables and the taxonomic data table. Then, it performs ACE, CHAO1, Shannon, Simpson and Inverse Simpson analysis, using the R software and, finally, the python “canvas” package groups the results, plotting and assembling a final report that can assist in the analysis of results and in their discussion. The production of the Metadoon-Pipeline manages to interconnect the parts of a bioinformatics and metagenomics research (such as; normalization and organization of data, pre-processing, generating graphs) and at the same time generate reports that help in the analysis of the result.

Keywords: Metagenomics, Pipeline, Usearch, Python, Tkinter.

1. INTRODUÇÃO

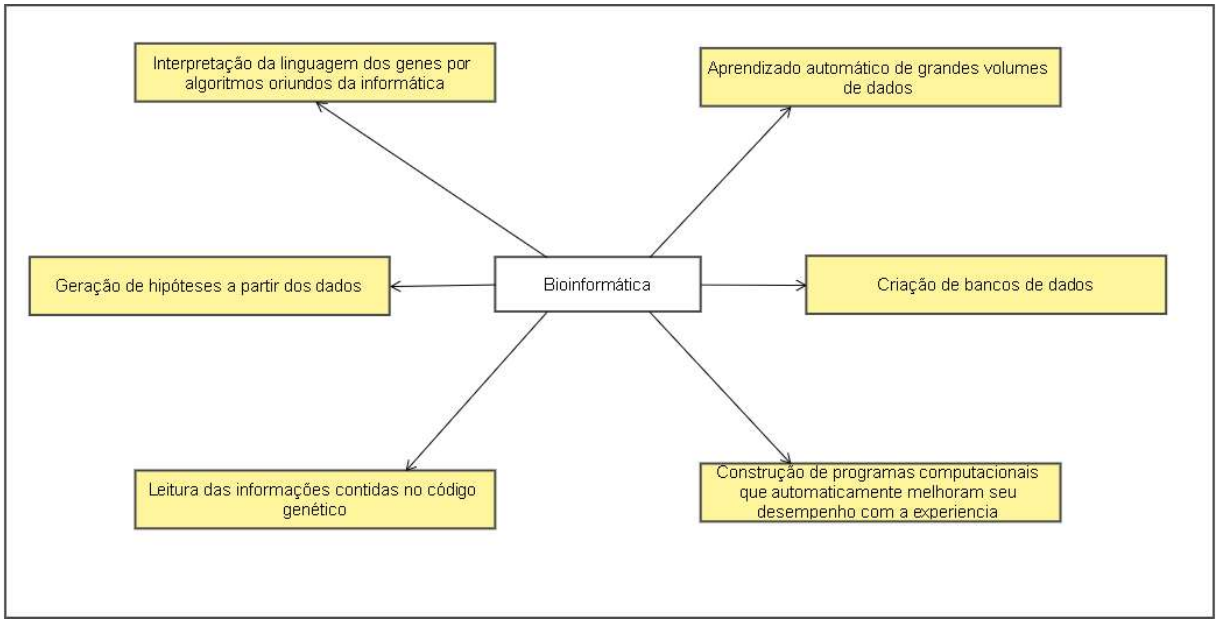
O marco dos anos 2000 para o meio genômico foi o sequenciamento do genoma humano, que marcou o avanço na era pós genômica. Este avanço se caracterizou pela coleta massiva de dados e o empenho frenético e colaborativo em fomentar um imenso banco de dados que pudesse apoiar, de modo geral, o meio científico. O resultado deste imenso estudo, apesar de não totalmente satisfatório, foi muito esclarecedor e proporcionou avanços importantes, como a descoberta de variantes para diversas mutações, fator que mais tarde auxiliaria a medicina moderna no combate a doenças mais graves (HENG, *et al*., 2018).

Os achados permitiram inovar áreas do conhecimento e atrelar a outras ciências já existentes, aperfeiçoando a precisão do saber humano. A dúvida sempre fora em como o fruto deste sequenciamento iria contribuir para a ciência, atualmente, entretanto, sabe-se que a medicina e as ciências forenses, dentre outras áreas do saber, se apropriam de alguma forma dos dados gerados pela pesquisa sobre o genoma humano. Hoje é possível diagnosticar com mais precisão certas mutações, descobrindo a predisposição a desenvolver certas doenças genéticas bem como calcular a probabilidade de desenvolver a mesma no decorrer da vida (ZATZ, 2002).

Acreditou-se, a partir de um certo momento, que o gene passaria a ter o seu devido destaque. Os estudos gênicos são interpretados através das impressões dadas pelas ciências ômicas (transcriptômica, metabolômica e proteômica), porém grande parte desse processo, que está sob as vistas da bioinformática, é estatístico, protocolar e demorado, se desenvolvendo de acordo com o avanço tecnológico da época. Juntamente com o avanço desse nicho científico existe avanço das dificuldades que cada vez se tornam mais complexas. (HENG, *et al*., 2018).

Com a consolidação robótica no século XX, novas tecnologias computacionais foram surgindo, fato que possibilitou a incorporação tecnológica nas ciências genéticas e consequentemente o avanço nos estudos e agrupamento dos resultados dos mesmos. É nesse cenário que, por volta de 1980, a bioinformática ou biocomputação surge com o objetivo de inovar as abordagens genéticas a fim de torná-las mais completas e melhorar os resultados em análises biológicas (ARAÚJO, et al., 2008).

Figura 1 - Segmentos da Bioinformática.



Adaptado de ARAÚJO e colaboradores (2008)

Acima é possível observar os principais segmentos que a bioinformática trabalha no ramo das ciências.

O instituto europeu de bioinformática (EMBL-EBI) fornece desde 1998 várias aplicações e soluções para utilização na análise de sequências, incluindo análise de similaridade de sequências, tais como: “BLAST”, “FASTA”, Interproscan e o Alinhador de sequências múltiplas (CLUSTAW, T-COFFE, MUSCLE, KALIGN, MAFFT). Todos esses

serviços são providos pelo framework "PERL-CGI". O processo de manutenção e de acesso desses serviços tem gerado um número considerável de empregos desde os anos de 2009.

Para melhoria do serviço oferecido aos pesquisadores e usuários de todo esse aparato, a instituição criou um módulo chamado JDispatcher e começou a trabalhar nele. JDispatcher é voltado tanto para usuários iniciantes quanto para usuários avançados e apresenta novidades na amostragem da anotação e visualização. Essa ferramenta está disponível na web através das interfaces SOAP e REST (GOUJON et.al., 2010).

1.1 Metagenômica

A biologia molecular divide-se em outras ciências correlatas, são exemplos dessas: Genômica, Proteômica e Metabolômica. Além disso, outras ciências ainda mais especializadas surgem no intuito de se complementarem e realizar uma boa compreensão sobre o conteúdo gênico: a Genômica Estrutural, que preocupa-se em determinar a estrutura do DNA em seus nucleotídeos ordenados; a Genômica Funcional, que se preocupa em analisar a informação proveniente dos ácidos nucleicos; a Proteômica, que identifica e caracteriza as proteínas codificadas pelo genoma e a Metabolômica que estabelece o status dos organismos ao analisar a as micromoléculas metabólicas que provém dos processos biológicos e por sua vez se associam ao conteúdo gênico, os dados provenientes destes estudos citados tendem a se tornar massivos (SOARES e FRANCO, 2007).

Com o avanço nos estudos da microbiologia ambiental através de métodos independentes de cultivo, surgiu uma nova abordagem quanto ao estudo de diversidade de grupos biológicos tidos como "não cultiváveis". Esse estudo foi acompanhado das tendências dos estudos genéticos, como a aplicação do método de clonagem para obtenção de um produto químico de preferência, processo que normalmente é precedido de um estudo metagenômico.

O termo metagenômica foi primeiro descrito por Jo Handelsman, da Universidade de Wisconsin, nos Estados Unidos, partindo da ideia de vários protocolos para acessar informações sobre microrganismos do solo. Seu interesse era sobre os produtos naturais gerados pelos resíduos presentes em tais organismos, visto que nunca teriam sidos isolados do solo e cultivados em laboratório. Sugeriu -se que tal ambiente seria uma grande fonte de recursos químicos intocados pelos estudos químicos humanos sobre produtos naturais. De maneira geral, as análises metagenômicas são feitas a partir do isolamento do DNA proveniente de uma amostra do ambiente alvo, fragmentação do mesmo e inserção de pedaços de interesse em vetores que possam replicar tais fragmentos funcionais no âmbito da produção em larga escala. Dessa forma podemos separar dois tipos de abordagens metagenômicas: screenings baseados em análise funcional e screenings baseados em sequenciamento. (FILHO, 2010).

A primeira clonagem através de amostras ambientais diretas foi feita a cerca de 30 anos por Pace e colaboradores (1976), utilizando o método de profiling usando o marcador 16S rRNA das amostras coletadas e purificadas. O DNA foi originado de uma população mista de microrganismos e o método permitiu a recuperação dos genes individuais e seu sequenciamento. Dessa forma foi possível, avaliar sequências completas ou parciais de 16S rRNA de cada grupo da população das bactérias achadas. Eventos seguintes, como a descoberta da importância dos microrganismos presentes nos solos para as plantas, abriram horizontes no que diz respeito ao número de fatores químicos e biológicos presentes no solo e desde então passou a se destacar ainda mais a importância das ciências ômicas no meio científico. O sequenciamento do gene 16S rRNA facilita a identificação de grupos experimentais e isso se deve ao fato de que o gene de rRNA 16S sofre poucas mutações, apresentando diversidade para identificação muito grande já que, em sua estrutura, apresenta

9 regiões que se mostram de maneiras diferentes entre as espécies de microrganismos. O sequenciamento do 16S se tornou uma tendência no meio da biologia computacional pelo seu preço e facilidade, além do seu potencial para identificação de OTUs (ALVES, *et al.*, 2018).

O uso do gene de RNA ribossomal 16S trouxe mudanças positivas para o estudo da ecologia microbiana pois, através dele, é possível entender sobre as posições filogenéticas de comunidades bacterianas de meio ambiente (HENTSCHEL, et al.,2002). Um dos primeiros estudos com o gene 16S rRNA foram feitos por Carl Woese que, em seu estudo, mostrou que esta molécula era um excelente marcador molecular (ATLAS & BARTHA, 1998).

A utilização da técnica de clonagem molecular e a produção de bibliotecas metagenômicas permitem com certa precisão realizar comparações entre comunidades bacterianas em duas ou mais áreas que apresentem diferentes características geográficas. Diferenças de diversidade entre conjuntos de amostras de microrganismos podem ser avaliadas através de uma região conservada do genoma bacteriano denominada 16S rDNA (gene 16S rRNA), diminuindo as dificuldades presentes durante as técnicas de cultivo e aprimorando o entendimento sobre a diversidade dos microrganismos e os biomas em que eles vivem. O RNA ribossomal 16S é um gene que codifica para a subunidade ribossômica menor, que é parte do sítio de ocorrência na síntese proteica, isto é, está presente em todas as bactérias. Nessa estrutura gênica é possível observar características que foram muito bem conservadas ao longo da evolução, dessa forma, servem como indicador de como os microrganismos podem estar intimamente relacionados durante a evolução em níveis de milhões de anos (SILVEIRA, 2004).

A metagenômica barcoding é uma abordagem crescente para a caracterização de comunidades microbianas em diferentes ambientes, incluindo o solo, a água, o ar e o corpo humano. Essa abordagem combina técnicas de amplificação gênica seletiva e sequenciamento

de nova geração para produzir dados de amplicon de alta qualidade, que são utilizados para avaliar a diversidade genética e funcional dessas comunidades (LIU et al., 2016).

Nos últimos anos, houve uma série de projetos e trabalhos de destaque na área de metagenômica barcoding, destacando a importância e a eficácia dessa abordagem para a caracterização de comunidades microbianas. Um dos maiores projetos na área de metagenômica barcoding foi o Earth Microbiome Project (EMP), que teve como objetivo produzir um banco de dados de referência para as comunidades microbianas do planeta. O EMP utilizou a metagenômica barcoding para avaliar a diversidade taxonômica e funcional de diferentes ambientes, incluindo o solo, a água e o ar (LIU et al., 2016).

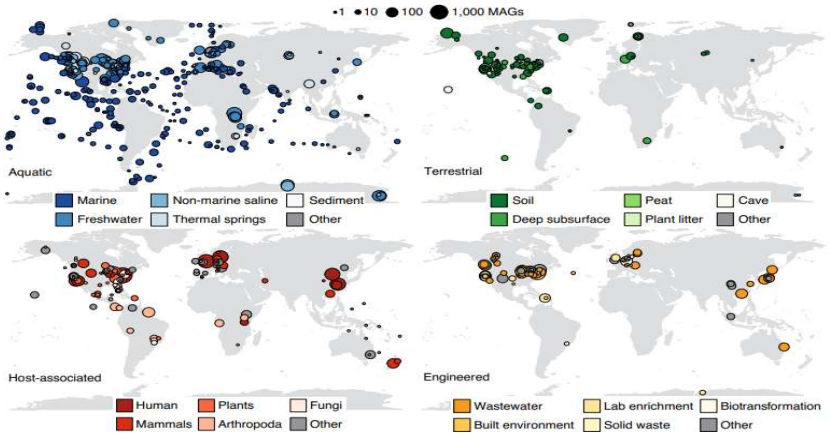
Outro trabalho de destaque na área de metagenômica barcoding foi a avaliação da composição da microbiota intestinal humana. Esse estudo utilizou a metagenômica barcoding para produzir dados de amplicon de alta qualidade que foram utilizados para avaliar a diversidade taxonômica e funcional da microbiota intestinal de diferentes indivíduos (KLINDWORTH et al., 2013).

Além desses projetos, houve uma série de trabalhos na área de metagenômica barcoding que investigaram a composição da microbiota em diferentes ambientes, incluindo o solo, a água e o corpo humano. Esses estudos forneceram importantes insights sobre a diversidade e a função das comunidades microbianas em diferentes ambientes. Em resumo, os maiores projetos e trabalhos na área de metagenômica barcoding destacam a importância e a eficácia dessa abordagem para a caracterização de comunidades microbianas em diferentes ambientes (LIU et al., 2016).

Grandes projetos metagenômicos têm sido produzidos com o passar do tempo, alguns com abordagens mais complexas, outros com abordagens mais simples e significativas. Neste ponto é possível citar o trabalho de Nayfach e colaboradores (2020), que objetivou a produção

de um catálogo genômico abrangente de bactérias não cultivadas e *archaea.* A produção deste catálogo proporciona uma oportunidade de comparação em larga escala genômica, principalmente para a mineração de genes e funções de interesse e construção de modelos de vias metabólicas, facilitando as abordagens computacionais de biologia de sistemas (NAYFACH, ROUX, et al., 2020). Este projeto recuperou cerca de 52.515 metagenomas de genomas montados (MAGs). O catálogo "GEM" foi construído com 10.450 metagenomas de amostras extraídas de diversos habitats microbianos e localizações geográficas;

Figura 2 - Locais de Recuperação dos MAG's para produção do catálogo "GEM".



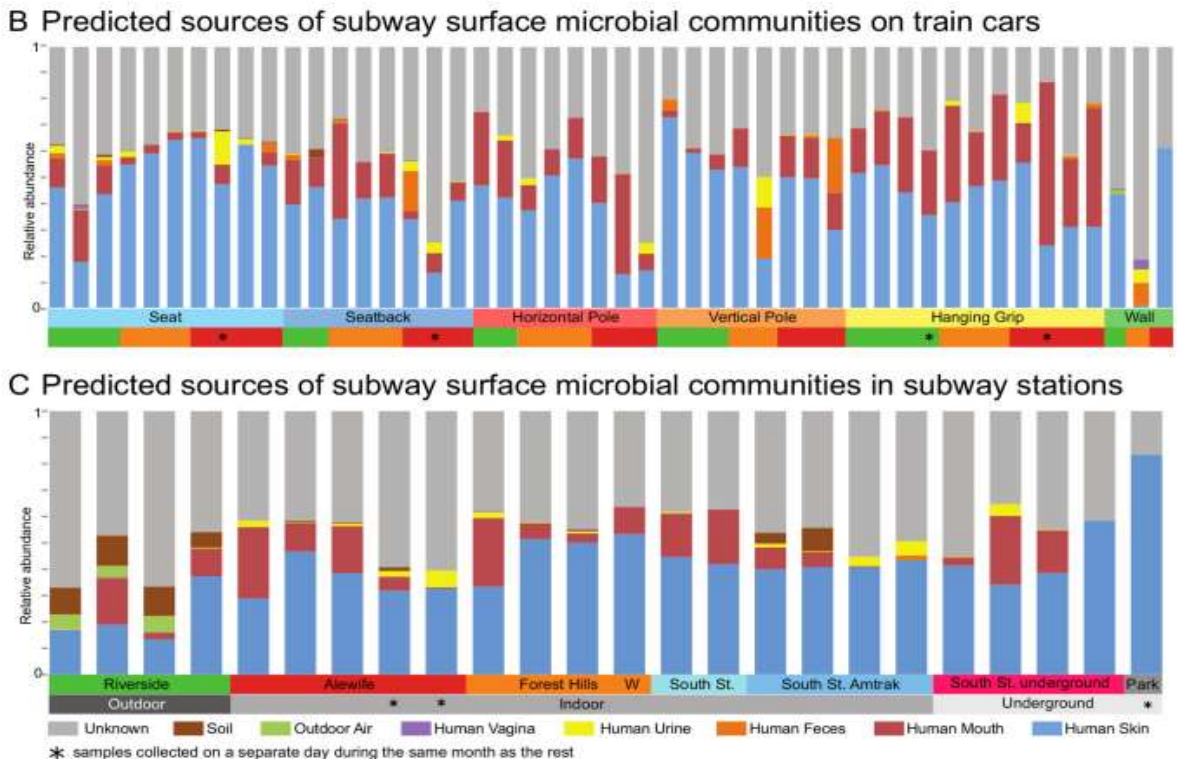
(NAYFACH, ROUX, et al., 2020)

Esses genomas representam cerca de 12.556 novas OTUs em nível de espécies candidatas, representando um recurso que captura uma ampla e diversa quantidade funcional de bactérias não cultivadas e archaea. A produção destes dados foi feita através da modelagem metabólica e o estudo da predição sobre a ligação host-virus. Como resultado, obteve-se vários dados sobre potenciais de vias e sobre "host connection" de vírus, entre outras descobertas. A utilização das ferramentas de interpretação metagenômica para um fim como o do trabalho de NAYFACH e colaboradores, representa o maior esforço até hoje para capturar a amplitude de bactérias e diversidade genômica archaeal em todos os biomas da

Terra, destacando dessa forma a importância do estudo de metagenomas para a ciência como um todo (NAYFACH, et al., 2020).

Um outro projeto de suma relevância para a metagenômica foi o produzido por HSU e colaboradores (2016). O projeto teve como objetivo utilizar do sequenciamento de amplicon 16S e "shotgun" para perfilar as comunidades microbianas em locais de trânsito público (Linhas de ônibus e trem) em Boston. As coletas foram feitas nas superfícies dos transportes públicos, especificamente locais onde o contato humano é mais presente: assentos, alças e barras de apoio, telas de equipamentos eletrônicos fixos e até mesmo panfletos grudados nas paredes. Esse estudo conseguiu reunir dados sobre a riqueza de espécies dentro das estações de acordo com seu local de coleta e com as origens (Habitat normal) dos microrganismos ali encontrados,

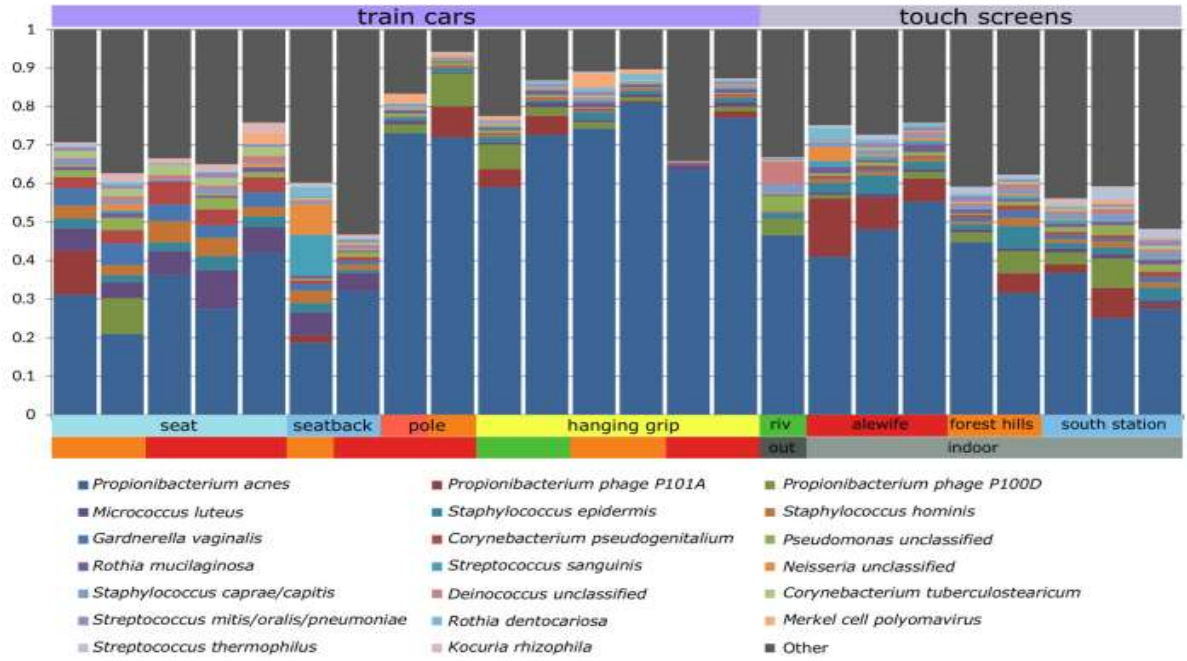
Figura 3 - Incidência/Origem de microrganismos presentes em vagões e estações de trem.



(HSU, et al., 2016)

bem como dados sobre a diversidade de 20 das espécies presentes nas amostras extraídas durante o estudo.

Figura 4 - As 20 espécies mais abundantes em vagões de trem e em "Touch Screens".



(HSU, et al., 2016)

Em metagenômica, além da abordagem com OTUs que leva em consideração as unidades taxonômicas funcionais, uma nova abordagem vem se popularizando no meio científico. Essa abordagem é denominada de ASV. A abordagem pelo algoritmo ASV permite, segundo pesquisas mais recentes, lidar melhor com erros de sequenciamento. O algoritmo ASV tende a modelar o erro do sequenciador e clusterizar suas leituras de forma que suas distribuições dentro dos clusters sejam consistentes com o modelo de erro. A abordagem evita fazer suposições sobre a variação dentro de um grupo taxonômico, supostamente uma fraqueza dos métodos OTU. Ao considerar tanto a semelhança de sequência quanto a sua abundância no modelo, os métodos ASV levam em conta o perfil de erro que resulta de experimentos de NGS, que podem produzir dezenas de milhares de leituras para uma única sequência de modelo de gene 16S rRNA (CARUSO, 2019).

A abordagem ASV tenta determinar quais sequências exatamente foram lidas e quantas vezes cada sequência foi lida. Os dados gerados serão incorporados em um modelo de erro para a execução de sequenciamento, permitindo a comparação de leituras semelhantes para determinar a probabilidade de que uma dada leitura em uma determinada frequência não seja devido a um erro do sequenciado. Isso cria, em tese, um valor "p" para cada sequência, considerando a hipótese nula como sendo essa sequência, sendo uma consequência do erro de sequenciamento. Em seguida, as sequências são filtradas de acordo com algum valor limite de confiança, deixando para trás um conjunto de sequências com uma confiança estatística definida. Por serem sequências exatas, geradas sem clustering ou utilização de bancos de dados de referência, os resultados com ASV podem ser facilmente comparados entre estudos usando a mesma região alvo. Além disso, pelo fato de uma dada sequência alvo sempre gerar o mesmo ASV e um determinado ASV, sendo uma sequência exata, pode ser usada para ser comparada a um banco de dados de referência em escala muito maior, permitindo maior precisão na identificação, até o nível de espécie e potencialmente além (CALLAHAN BJ, et al., 2019).

A metagenômica é normalmente aplicada para compreensão da análise de diversidade em um ambiente, essa aplicação pode ser utilizada na procura de novas espécies, patógenos ou genes de interesse dentro de um organismo em questão. A metagenômica mais antiga acopla, em seu modelo, métodos mais tradicionais, tanto no sequenciamento quanto no tratamento de dados. As plataformas NGS mais recentes tem como grande característica a sua rápida utilização de pipelines específicos, que facilitam o desenvolvimento do workflow planejado e utilizado no decorrer dos trabalhos. (LI, *et al.*, 2016).

A Unidade Taxonômica Operacional (OTU) é uma abordagem que permite a classificação de micro-organismos a partir de sua genética, morfologia e biologia. Essa

metodologia foi introduzida como uma alternativa ao sistema tradicional de classificação baseado em características físicas e morfológicas, que apresenta algumas limitações. A OTU é baseada na ideia de que uma unidade taxonômica é uma comunidade de micro-organismos com características genéticas e funcionais similares. Ela é definida como a menor unidade taxonômica que pode ser detectada e é diferenciada de outras unidades com base em critérios objetivos (VANDAMME et al., 1996).

Os critérios utilizados para a definição de OTUs incluem a divergência genética, a similaridade fenotípica e o comportamento fisiológico. A análise da sequência de DNA é a principal ferramenta utilizada para determinar a divergência genética entre as UTOs. A OTU é considerada uma abordagem mais precisa e abrangente na classificação de micro-organismos, uma vez que leva em consideração aspectos genéticos e funcionais além de características morfológicas. Além disso, a OTU permite a identificação de novos tipos de micro-organismos e a avaliação da diversidade microbiana em diferentes ambientes (VANDAMME et al., 1996).

1.1.1 DNA Barcoding

A técnica de DNA barcoding foi proposta pela primeira vez por Paul Hebert e sua equipe em 2003, como uma forma de identificar espécies de insetos (HEBERT et al. 2003). Desde então, tem sido aplicado em uma ampla variedade de grupos taxonômicos, incluindo plantas, animais e fungos (ADAMOWICZ SJ et al. 2018).

A eficácia do DNA barcoding foi comprovada em vários estudos (HOLLINGSWORTH et al. 2008), mostrando que é uma ferramenta valiosa para a identificação de espécies. Além disso, o uso de DNA barcoding tem se mostrado particularmente útil em grupos taxonômicos com características morfológicas similares, como insetos e peixes (WARD et al. 2005).

1.2 Análise de Riqueza e Diversidade

A Riqueza é uma medida que mostra o número dos diferentes tipos de espécies presentes na amostra, podendo levar em conta as relações filogenéticas entre os indivíduos distribuídos entre os tipos dentro das amostras, como divergência, riqueza ou uniformidade (TUCKER, et al., 2017). Os índices de diversidade em ecologia são usados para mensurar os tipos de interesse que geralmente são espécies, mas também podem ser outras categorias como gêneros, famílias, tipos funcionais ou haplótipos. As entidades geralmente são plantas ou animais e a medida de abundância pode ser, por exemplo, número de indivíduos, biomassa ou cobertura perante o total (HILL, 1973).

A maioria dos índices mensuram apenas a diversidade categórica entre os sujeitos. Esses índices, no entanto, não levam em conta a variação total quanto a diversidade que pode ser mantida entre os grupos, o que pode ocorrer apenas quando são calculadas tanto a diversidade categórica quanto a qualitativa (TUOMISTO, 2010). A diversidade "verdadeira", ou o número efetivo de tipos, refere-se ao número dos grupos proporcionalmente abundantes necessários para que a abundância proporcional média dos tipos seja igual à observada no conjunto de dados da amostra (TUOMISTO, 2010).

A riqueza(R) é uma medida que quantifica quantos tipos o conjunto de dados de interesse contém. A riqueza de espécies (geralmente indicada como S) de um conjunto de dados é o número de espécies na lista de espécies desse mesmo conjunto. A riqueza é simplesmente uma medida, por isso tem sido um índice de diversidade popular em ecologia, onde os dados de abundância muitas vezes não estão disponíveis para os conjuntos de dados finais (JOST, 2006). A riqueza não leva em conta a abundância dos tipos, diferente da diversidade, que leva em conta a abundância. Se a verdadeira diversidade for calculada com

q = 0, é possível concluir que o número efetivo de tipos (0D) é igual ao número real de tipos, que é semelhante à Riqueza (R) (TUOMISTO, 2010).

O estudo de diversidade em macroorganismos proporcionou abordagens de análise em padrões de diversidade no ambiente tal que se tornassem passíveis de aplicação em micro-organismos (COLWELL e CODDINGTON, 1994).

Na amostra de uma comunidade, os tipos de organismos em números observados aumentam de acordo com o número de amostras. Esse dado revela informações importantes quanto a diversidade total da comunidade estudada. O padrão descrito pode ser mostrado em uma curva de abundância, por exemplo, ou até mesmo em uma curva cumulativa. Curva cumulativa é o nome dado ao gráfico que mostra os tipos observados na amostra pelos números de indivíduos amostrados. Visto que as comunidades têm um número finito de espécies, se as coletas continuassem a ocorrer, logo a curva alcançaria o número real de riqueza da comunidade, dessa forma essas curvas podem demonstrar fatores importantes como a qualidade da amostragem. Quanto mais horizontal uma curva se torna, mais qualidade tem sua amostragem (HUGHES, HELLMANN, *et al.*, 2001).

Para estimar a riqueza de uma comunidade a partir de uma amostra ou entre diferentes amostras, é necessário utilizar um estimador de diversidade. A partir de então se utilizam de dois tipos de estimadores: “estimadores paramétricos” e “estimadores não paramétricos”. Os métodos paramétricos utilizam-se de modelos de abundância relativa de espécies e encaixam os resultados em algum modelo adequado e, dentre esses modelos, tem-se o “log normal”. Já os métodos não-paramétricos estipulam a diversidade sem que se suponha um modelo específico de abundância para os dados, muitos desses métodos estatísticos são adaptados de estatística de marca-recaptura (mark-release-recapture-MRR). Essas abordagens consideram a possibilidade da quantidade de OTUs recapturadas em relação a aquelas contadas apenas

uma vez. Em uma comunidade consideravelmente grande, as chances de uma espécie ser recapturada mais de uma vez é muito baixa, ou seja, um indivíduo irá representar sua respectiva espécie. Em comunidades menores, a probabilidade de uma espécie ser recapturada é maior (PEIXOTO, 2013).

Para caracterizar a taxonomia da comunidade de indivíduos, os espécimes devem ter seus genes agrupados em unidades taxonômicas operacionais (OTUs) através do nível de similaridade. As OTUs acabam por auxiliar no processo de caracterização e descrição de comunidades e agrupamentos bacterianos (LAMBAIS et al., 2005). Grande parte dos estudos com OTUs utilizam grau de similaridade em níveis de 97%.

1.2.1 Análise de CHAO1

O índice de diversidade CHAO1 foi proposto por Anne Chao em 1987, com o objetivo de estimar o tamanho da população para dados de (re)captura com capacidade de captura desigual. Os dados utilizados por Anne foram cedidos por Edwards e Eberhardt (1967, Journal of Wildlife Management 31, 87-96) e Carothers (1973, Journal of Animal Ecology 42, 125-146). Utiliza o método de quantificar a captura de uma dada sequência na amostra. Chao1 é um estimador de riqueza baseado na abundância de espécies, diferente de Chao1 por exemplo, que é um outro estimador de riqueza, porém que se baseia em incidência (DIAS, 2004). Chao1 calcula o número de capturas de uma determinada espécie dentro de um campo amostral (ROBERT K. COLWELL & JONATHAN A. CODDINGTON, 1994), e pode ser calculado por:

$$S_1^a = S_{obs} + (a^2/2b)$$

Onde “S” é o número de espécies observadas na amostra, “a” é o número de espécies observadas que pode ser representada por apenas um único indivíduo nessa amostra e “b” é o número de espécies observadas que podem ser representadas por exatos dois únicos indivíduos na amostra (ROBERT K. COLWELL & JONATHAN A. CODDINGTON, 1994).

Chao1 pode ser calculado no software “R” pela utilização de alguns pacotes como “FOSSIL” ao instalar e importar o mesmo no script de análise (VAVREK, MATTHEW J. 2011).

1.2.2 Análise de SHANNON

O índice de Shannon é um índice ecológico bastante popular na literatura, também chamado de índice de diversidade de Shannon (SPELLERBERG &FEDOR, 2003). A medida foi originalmente proposta por Claude Shannon para quantificar a entropia em conteúdo de strings (SHANNON, 1948). O princípio deste índice é de que, quanto mais letras houver, e quanto mais próximas forem suas abundâncias proporcionais na sequência de interesse, mais difícil será prever corretamente qual letra será a próxima na sequência. O índice de Shannon quantifica a incerteza (entropia ou grau de surpresa) associada a esta previsão da sequência seguinte esperada (TUOMISTO, 2010). Para realizar o cálculo de Shannon é preciso obter o número da abundância crua das espécies ($Pi = X / \Sigma_x$), onde X é o número de espécies observadas e Σ_x é a soma dos valores de X, além disso, considera-se apenas valores maiores que zero para X. A partir de então o índice de Shannon pode ser calculado: $(-\Sigma_{(Pi)} * \mathrm{Log}_{(Pi)})$, isto é; a soma negativa da abundância crua das espécies multiplicado pelo log da abundância crua das espécies (TUOMISTO, 2010).

1.2.3 Análise de ACE

ACE é um índice característico da ecologia por calcular a riqueza extrapolada de espécies de uma população usando os Estimadores de cobertura baseados em “Abundância”. A análise de ace pode ser executada dentro de um script R pelo pacote “Fóssil”. A função executada para análise de ace no R aceita um vetor, matriz ou quadro de dados de qualquer tamanho composto de inteiros positivos e zeros. As matrizes são organizadas de tal forma que cada linha seja um táxon (OTU) diferente e cada coluna seja uma amostra ou localidade. É importante ressaltar que o ACE se destina apenas ao uso de dados de abundância, e não com dados de ausência ou presença. A função de ACE pode retornar “NaN” ou “Inf” como um valor, nesse caso, Chao1 é usado em seu lugar (VAVREK, MATTHEW J. 2011).

1.2.4 Análise de Simpson e Simpson inverso

O índice de Simpson foi introduzido em 1949 por Edward Simpson. Foi formulado para mensurar um dado grau de concentração quando indivíduos são classificados em tipos. Existem algumas variações e aplicabilidades desse índice e abaixo destaca-se uma delas:

$$\ell = \frac{\sum^{R}_{I=L} N_I(N_I-1)}{N(N-1)}$$

Onde “n_i“ é o número de entidades pertencentes ao tipo “$_i$” (que se multiplica com a soma dos indivíduos contados no dataset) e “N” é o número total de entidades no conjunto de dados. Outro índice que pode ser usado em adição ao de Simpson é o índice de Simpson inverso, onde captura-se o valor do índice de Simpson e o inverte da seguinte forma: “1 / Simpson” (SIMPSON, 1949).

1.3 Linguagens De Programação e Pipelines em Bioinformática

Através de cálculos matemáticos, os computadores podem fazer diversos trabalhos e até auxiliam em tarefas humanas. Estão presentes hoje em dia em diversos usos humanos e no âmbito científico, as primeiras linguagens de programação utilizadas foram COBOL, FORTRAN, entre outros. As linguagens a nível científico levam em consideração uma mesma estrutura de dados, no entanto a sintaxe se difere de acordo a necessidade da aplicação e hoje em dia as linguagens evoluíram e surgiram novas aplicações. No âmbito de aplicações, a automatização de processos e criação de inteligências artificiais tem se tornado tendências pelo fato de promoverem facilidade (SEBESTA, 2018). Ferramentas de desenvolvimento estão ao lado do bioinformata. A aplicação de SOAP's e API's são definitivamente parte do trabalho de quem trabalha com biologia computacional. Tais métodos de aplicações podem ser integrados localmente e aplicados em workflows e pipelines para a análise desejada. A utilização de API's pode ser aplicada ao desenvolvimento de aplicações robustas em várias linguagens de programação, tais como: C/C++, C#, Java, Pearl, PHP, Python e Ruby (GOUJON et.al., 2010).

De certa forma as linguagens são lançadas e pacotes específicos também são liberados para nichos específicos de usuários, em python por exemplo, existe o pacote Biopython que auxilia na execução de tarefas e tratamento de dados de anotação e sequenciamento. A vantagem de utilizar pacotes como esse, vem da facilidade proveniente dos projetos "open-source", onde a produção, utilização e atualização dos pacotes são feitas, de maneira comunitária, pelos usuários. O python é uma linguagem com desenvolvimento em script, orientada a objetos, que se mostra adequada para processar arquivos de textos e automatizar tarefas comuns. Pelo seu grande potencial com relação ao paradigma de orientação a objetos, Python pode ser utilizado para projetos em larga escala, oferecendo uma grande diversidade de aplicações. Dentre eles, existe o tratamento de dados brutos de texto, além da classificação

por métodos de Data mining, como por exemplo, máquinas de vetorização, naive bayes, entre outros (CHAMPMAN & CHANG, 2000).

O powershell também é uma linguagem com grande potencial para aplicação em tarefas de biologia computacional. O powershell é uma linguagem bastante recursiva e apresenta características notáveis como uma linha de comando interativa orientada a objetos, navegação fluente, boa manipulação de arquivos XML, capacidade de executar tarefas na web através de via de linha de comando, sintaxe e gramáticas consistentes, expressões de regulagem ricas e formatação de saída de dados avançada.

Pipelines para metagenômica de barcoding são ferramentas essenciais para o processamento e análise de dados metagenômicos. Eles permitem a automação de várias etapas, incluindo a qualidade do controle, a filtragem de baixa qualidade, a assemblagem de sequências e a identificação de marcadores taxonômicos. Um exemplo de pipeline amplamente utilizado é o MetaPhlAn. Este pipeline utiliza técnicas de barcoding para identificar a presença de bactérias, vírus, fungos e outros microrganismos a partir de sequências metagenômicas. Além disso, ele também pode ser usado para estimar a abundância relativa de diferentes taxas em uma amostra (SEGATA et al., 2011).

Outro pipeline popular é o Kraken (WOOD et al., 2014). Este pipeline usa um banco de dados taxonômico pré-construído para classificar sequências de DNA em diferentes taxas. Ele é especialmente eficiente em sequências de baixa complexidade e pode ser aplicado a grandes conjuntos de dados de forma rápida e precisa. Além dos pipelines mencionados, há também muitos outros disponíveis para metagenômica de barcoding, cada um com suas próprias vantagens e desvantagens. É importante lembrar que a escolha do pipeline certo depende de vários fatores, incluindo o tamanho e a complexidade da amostra, o objetivo da análise e as preferências pessoais do usuário.

1.4 Ferramentas de processamento de dados metagenômicos

O programa USEARCH é uma ferramenta de bioinformática amplamente utilizada para a análise de sequências de DNA e proteínas. De acordo com Edgar (2010), USEARCH é conhecido por sua eficiência e rapidez na realização de tarefas de alinhamento de sequências, clustering e classificação taxonômica.

A capacidade do programa de processar grandes quantidades de dados em questão de minutos é possível graças ao uso de algoritmos avançados de busca de sequências, como o algoritmo de busca de similaridade global (EDGAR, 2010). Além disso, o programa possui uma interface de usuário intuitiva e fácil de usar, o que facilita a realização de análises complexas por parte de usuários sem conhecimento avançado em bioinformática (EDGAR, 2010).

O programa USEARCH também é amplamente utilizado em estudos de diversidade microbiana, onde é possível identificar diferentes tipos de bactérias e fungos em amostras de solo, água e ar, entre outros (KLINDWORTH et al., 2013). Além disso, o programa pode ser usado para realizar análises filogenéticas, onde é possível identificar relações evolutivas entre espécies (KLINDWORTH et al., 2013).

O QIIME 1 e QIIME 2 são dois programas de análise de dados de sequências bacterianas amplicon-baseadas. Ambos são utilizados para investigar a composição taxonômica e funcional dos micróbios presentes em amostras ambientais, clínicas e agrícolas.

O QIIME 1 é uma ferramenta baseada em linha de comando, que foi publicada pela primeira vez em 2010 (CAPORASO et al., 2010). Desde então, tornou-se um dos principais programas de análise de dados de sequências bacterianas amplicon-baseadas. O QIIME 1 oferece uma ampla gama de métodos de análise, incluindo filtragem, clusterização, classificação e visualização de dados (CAPORASO et al., 2011).

O QIIME 2, por sua vez, é uma versão totalmente nova e atualizada do QIIME 1, que foi lançada em 2017 (BOLYEN et al., 2019). O QIIME 2 foi projetado para ser mais flexível e fácil de usar, oferecendo uma ampla gama de plugins que permitem aos usuários personalizar suas análises. Além disso, o QIIME 2 oferece uma interface baseada em nuvem, que permite aos usuários acessar e analisar seus dados de qualquer lugar com uma conexão à internet. Em resumo, o QIIME 1 e o QIIME 2 são duas ferramentas valiosas para a análise de dados de sequências bacterianas amplicon-baseadas. Enquanto o QIIME 1 é uma ferramenta baseada em linha de comando, o QIIME 2 é uma ferramenta mais flexível e fácil de usar, com uma interface baseada em nuvem.

O programa UQIIME (Unified Quality-Filtered Informed Metagenome Assembler) é uma ferramenta open-source que permite a análise e a visualização de dados metagenômicos (sequências de DNA obtidas de ambientes não cultiváveis).

Desenvolvido por Edgar et al. (2010), o UQIIME é amplamente utilizado por pesquisadores em biologia e ecologia microbiana para obter uma compreensão mais profunda da biodiversidade e da dinâmica dos ecossistemas microbianos. Com o UQIIME, é possível realizar tarefas como a classificação taxonômica de sequências, a construção de árvores filogenéticas, a identificação de padrões de associação funcional e a análise de sobreposição de comunidades. Além disso, o programa também suporta a integração de dados metagenômicos com outros tipos de dados, como dados de transcriptômica e proteômica, para uma visão mais completa da atividade microbiana.

O MOTHUR é um software de bioinformática amplamente utilizado para análise de sequências de amplicon de DNA. Desenvolvido por Patrick Schloss, o MOTHUR é uma ferramenta de análise de amplicon de alta qualidade que permite aos usuários realizar uma

série de tarefas, desde a qualidade do controle até a classificação taxonômica (SCHLOSS et al., 2009).

Uma das principais vantagens do MOTHUR é sua capacidade de lidar com grandes quantidades de dados de sequência, tornando-o uma escolha popular para estudos de biologia molecular. Além disso, o MOTHUR oferece uma ampla gama de recursos, incluindo a classificação taxonômica baseada em referência, a normalização de biblioteca, a clusterização por similaridade e a construção de árvores filogenéticas (SCHLOSS et al., 2009).

O MOTHUR também é conhecido por sua capacidade de lidar com dados de alta complexidade, tornando-o uma escolha popular para estudos envolvendo comunidades microbianas. Além disso, o MOTHUR é altamente customizável, permitindo que os usuários ajustem seus processos de análise de acordo com suas necessidades específicas (SCHLOSS et al., 2009).

Em resumo, o MOTHUR é uma ferramenta valiosa para a comunidade de bioinformática, oferecendo recursos avançados e alta capacidade de processamento para ajudar os usuários a realizar análises de amplicon de DNA de alta qualidade (SCHLOSS et al., 2009).

2. OBJETIVOS

2.1 Objetivo Geral

Construir um pipeline que automatize o processo de análise metagenômica dos dados de sequenciamento e gere relatórios de fácil entendimento.

2.2 Objetivos Específicos

- Produção de um pipeline para análises de diversidade e abundância dentro de amostras de metagenomas (Bar-coding);
- Facilitar e padronizar análises de diversidade bruta dentro do campo da bioinformática;
- Diminuir o tempo de filas de espera ao utilizar dispositivos em web- servers para tratar de dados de sequenciamento;
- Facilitar o acesso e execução de tarefas bioinformatas para pesquisadores não muito familiarizados com pipelines tradicionais.

3. MATERIAL E MÉTODOS

3.1 Considerações iniciais

A produção do pipeline foi feita em ambiente Linux, na distribuição Ubuntu em sua versão 22.04 (Jammy Jellyfish). Os scripts foram produzidos em Python, aproveitando a biblioteca "OS", para importar comandos shell dentro do script de interface. A interface utilizada foi a "Tkinter" (LUNDH, 1999), disponível na linguagem python, e automaticamente presente na plataforma Windows e disponível para download na plataforma Linux via pip (instalador python).

A interface automatiza comandos shell presentes no programa Usearch, que é uma ferramenta muito utilizada para execução de tarefas como: mesclar pares em arquivos fastq; fazer filtro de qualidade; formar uma tabela com OTUS e executar blast com uma base de dados específica, entre outros comandos. A saída dos dados leva os dados tratados para um script em R que faz análises e plota os dados para que, no final, seja gerado um report final com os dados (PLOTS e análises) para registro dos resultados.

O script automatizado em python com a interface tkinter, tem a intenção de executar tarefas importantes na metagenômica como: mesclagem dos pares de sequências nos arquivos "fastq"; filtro de qualidade; pesquisa de sequências únicas dentro das sequências filtradas; clusterização das sequências em OTUS; criação da tabela de OTUS; comparação das OTUS com uma base de dados opcional para o usuário; análise de diversidade e análise funcional; criação de um report final com os plots dos dados e gráficos para discussão dos resultados.

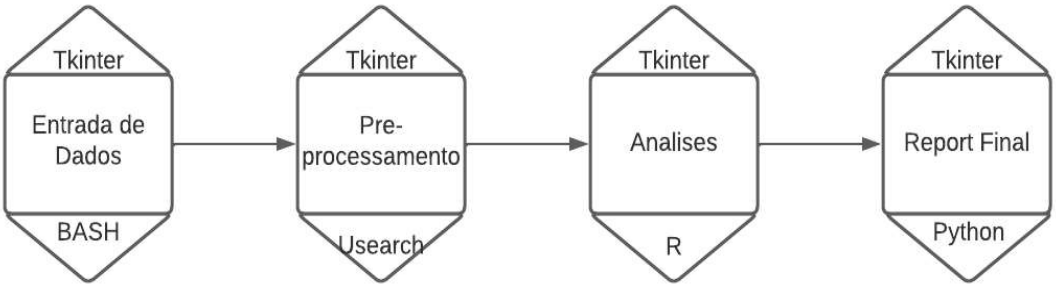
Para a instalação do pipeline, foi produzido um script de instalação em bash, que instala e atualiza as bibliotecas que o pipeline precisa para executar as suas partes, como a instalação da biblioteca Tkinter, a atualização da linguagem python no ambiente, a instalação do R, e alocação dos arquivos do pipeline nas pastas do sistema.

3.3 Funcionamento Do Pipeline

A intenção é que o pipeline execute tarefas básicas como: receber dados do usuário; pré-processar esses dados; executar análises pré dispostas neste trabalho e gerar um relatório final com o intuito de auxiliar o pesquisador na interpretação desses dados e na montagem de sua pesquisa. A biblioteca Tkinter do python deve servir para melhorar a interação do pipeline com o usuário, no intuito de facilitar o acesso e execução dos comandos através de botões, caixas de textos, combo boxes etc.

A intenção é que o funcionamento do pipeline deve seguir o fluxograma na figura 5;

Figura 5 - Esquema representando o esquema de automação para o pipeline final.



Acima, a imagem mostra a intenção de funcionamento geral do pipeline. Cada passo terá uma interação com o usuário dentro da interface Tkinter. A entrada de dados é gerada através de comandos bash, onde se copia os arquivos brutos do usuário para uma pasta do pipeline, onde serão feitos os novos passos. O pré processamento que é feito pelo Usearch gera dados tabelados para execução de análises que é feita no R e, em seguida, ocorre uma releitura desses dados no python com o intuito de gerar o arquivo de relatório final.

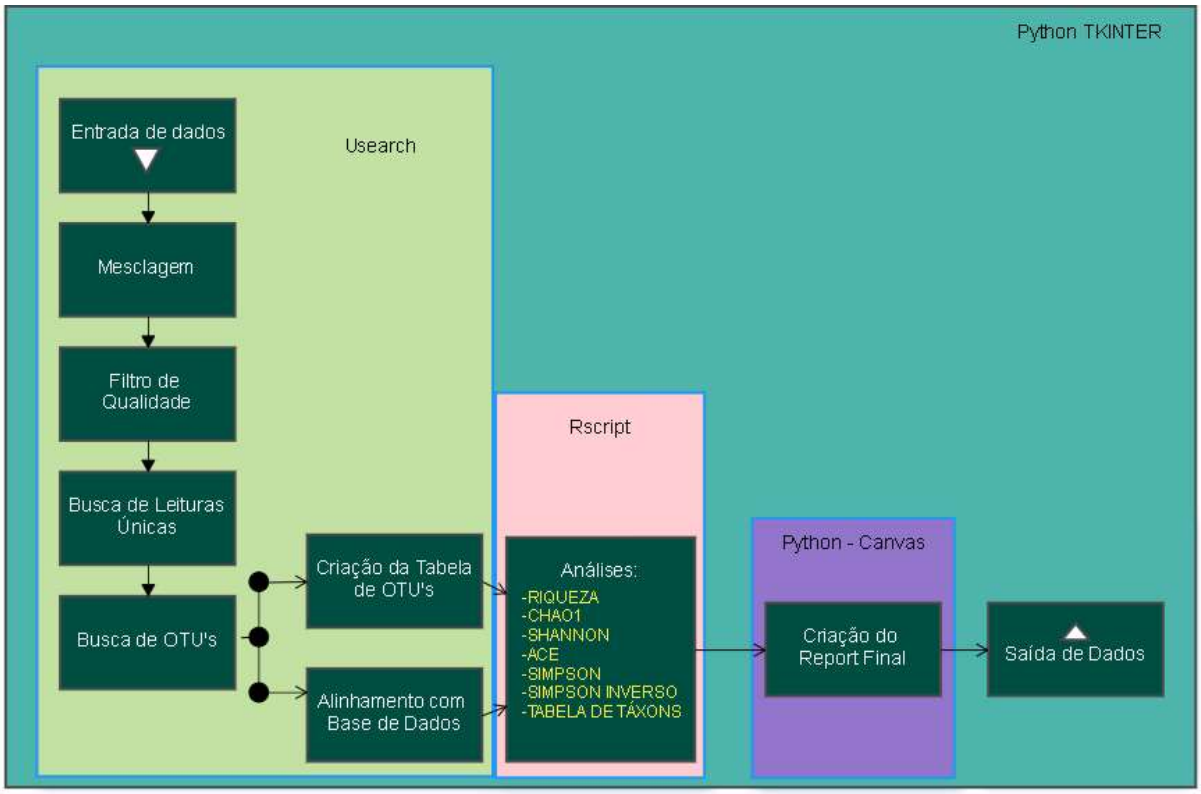
Os dados utilizados para testes são amostras retiradas de microrganismos de manguezais.

4. RESULTADOS E DISCUSSÃO

Embora a proposta deste trabalho tenha sido simples, a demanda de tempo acabou precisando ser maior devido aos fatos gerados pela pandemia de covid 19, decorrida desde o ano de 2020 até os dias atuais onde seus resquícios ainda permanecem relutantes.

O funcionamento geral do pipeline segue o fluxo abaixo, tendo como principal fonte de execução de análise a ferramenta Usearch (Versão 11).

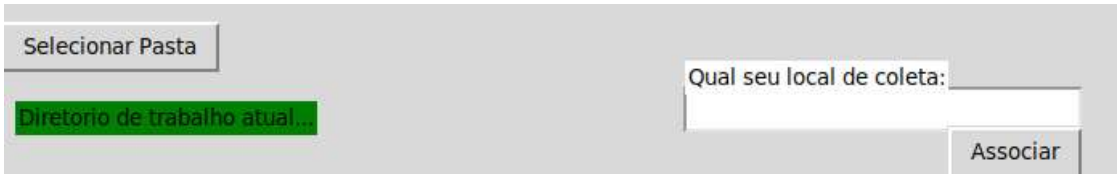
Figura 6 - Funcionamento do Metadoon Pipeline.



O pipeline traz em si um script principal em python, no qual será apresentada a interface “Tkinter” principal. Dentro desse script estarão incorporadas linhas de comandos para a execução dos processos descritos acima.

As primeiras entradas de dados do pipeline pedem ao usuário a pasta onde suas amostras estão direcionadas e o ambiente em que elas foram encontradas. O nome do ambiente onde elas foram encontradas acompanha todos os principais arquivos de saída que forem gerados ao longo do processo, isto é, se seus dados foram coletados no manguezal, você deve digitar no campo “Qual seu local de coleta:” o nome “manguezal”, dessa forma todos os principais arquivos gerados serão nomeados com o prefixo manguezal, à exemplo disso se tem o report final que será gerado com o nome que você deu a sua “pesquisa”.

Figura 7 - Entrada de dados.

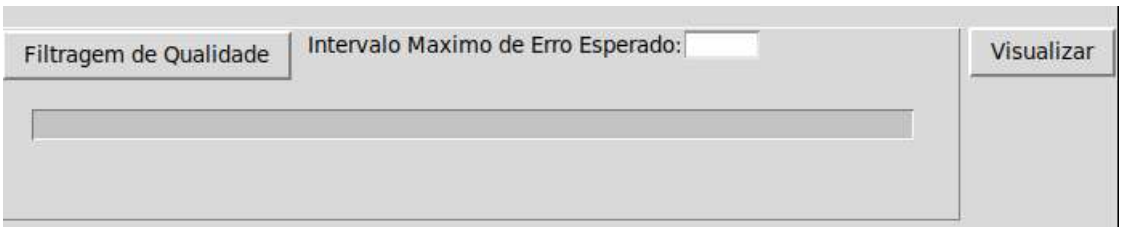


Em seguida os dados das amostras são mesclados, com o comando "fastq_mergepairs", isto é, concatenados em um só arquivo:

Figura 8 - Mesclando pares.

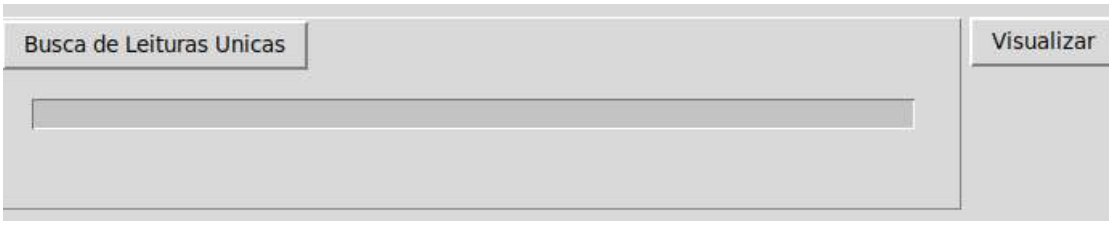
Depois desse passo, os arquivos mesclados são filtrados, com o comando "fastq_filter", onde o Usearch irá analisar a qualidade das bases mescladas:

Figura 9 - Filtro de qualidade.



Após esse passo será rodado o comando "fastx_uniques" que busca nos campos das amostras, por sequências únicas(singletoons):

Figura 10 - Buscando leituras únicas.

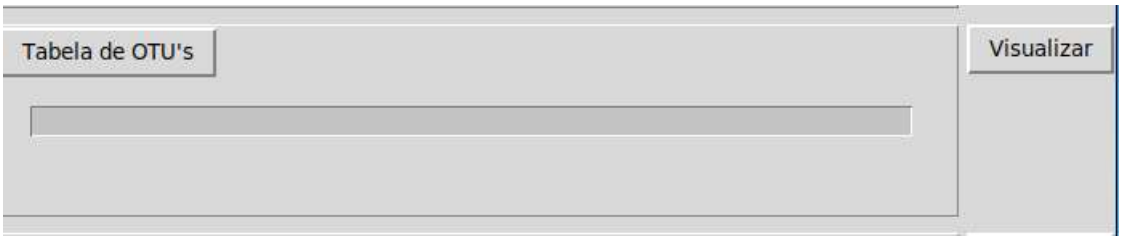


Após isso, o pipeline pode executar o comando “cluster_OTUs” que agrupa as sequências das amostras em OTUs por nível de semelhança:

Figura 11 - Organizando os grupos em OTU's.

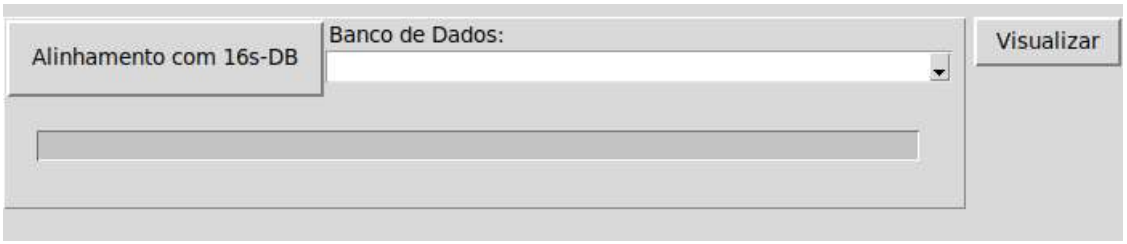
Logo em seguida o comando “usearch_global” é utilizado para pegar os dados de OTUs e organizá-los em uma tabela que é utilizada para fazer os cálculos dos índices de riqueza e diversidade:

Figura 12 - Montando tabelas de OTU's.



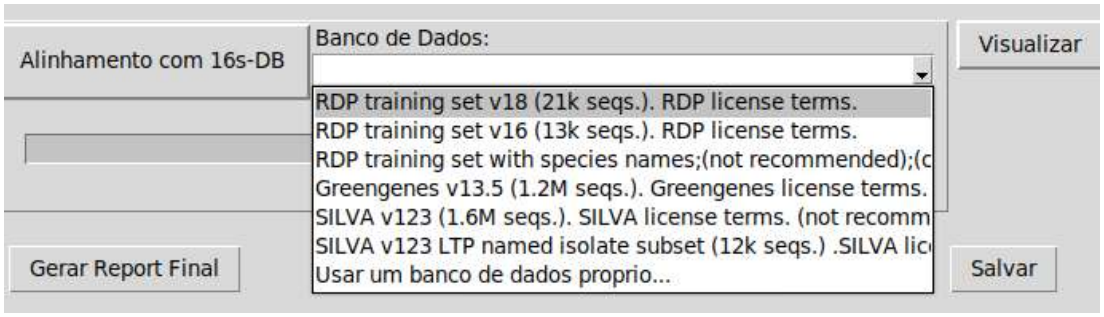
Ao final da preparação desses dados, é executado o comando “nbc_tax” que compara as sequências com uma base de dados de escolha do usuário:

Figura 13 - Alinhando as sequências com banco de dados.



Nesse passo, o pipeline oferece a possibilidade de que o usuário possa selecionar uma base de dados descrita como opção, ou adicionar uma base de dados própria para executar esse ponto do processo:

Figura 14 - Opções de bancos de dados.



Após o processo de preparação das amostras, ocorre a execução das análises de diversidade, tais como: análises de Shannon; ace; chao1; Simpson; Simpson inverso e análise de riqueza. Ocorre também a formação e plotagem de uma tabela com os filos presentes nas amostras e suas respectivas frequências no total de amostras. Por fim, os dados gerados são reunidos e agrupados em um documento no formato PDF, que deve servir para ajudar o usuário a entender e discutir seus dados com relação às principais descrições estatísticas observadas neste mesmo documento. Este relatório final traz uma

breve introdução sobre o que cada análise significa para aquele conjunto de dados, e em seguida os gráficos gerados pelo pipeline. Ao todo, 3 documentos são gerados, além dos arquivos pré-tratados e agrupados na pasta final, são eles: o relatório final em inglês; o relatório final em português e o gráfico de frequência dos filos por OTUs.

A geração desses processos é feita ao apertar o botão “Gerar Report Final”, onde os dados condensados são salvos em uma pasta dentro dos arquivos do pipeline.

Figura 15 - Opções de saída dos dados.

Depois de produzido, o arquivo do relatório final pode ser visualizado ao apertar o botão “Visualizar”. Além disso, todos os arquivos finais podem ser salvos em uma pasta fora dos arquivos do pipeline, determinada pelo usuário. Os passos descritos nas imagens acima são executados de maneira independente quanto a sua ordem de execução. O único requisito claro é a execução dos passos de forma sequencial, isto é, executar um passo de cada vez. O pipeline pode ser rodado em um kernel Linux apenas, embora não tenha sido testado para outras plataformas além de Linux e Windows.

A execução em ambiente Windows acabou não sendo possível de ser finalizada no tempo determinado devido a um conjunto de bugs, a maioria deles não foram possíveis de serem solucionados.

O pipeline criado utiliza a versão 32 bits do Usearch, por tanto suas análises restringem o tamanho dos arquivos de acordo com o sistema operacional utilizado e a arquitetura do programa, dessa forma as análises estão limitadas a arquivos com que possam utilizar apenas 4Gb (Linux) e 2Gb (Windows). Isso se deve ao fato de o licenciamento dado a versão de 64

bits ser paga. O Pipeline foi chamado até o momento de Metadoon, de certa forma acaba distribuindo a versão gratuita do Usearch, mas automatizando passos que seriam normalmente feitos por um pesquisador da área, além da plotagem dos dados e condensação dos mesmos em algo mais fácil de compreender.

Durante o processo de produção, foi possível perceber que a utilização do pipeline segue uma rotina que poderia ser um tanto difícil de entender em um primeiro contato e o que facilitaria o entendimento e importância das ordens de execução seria basicamente a disposição dos botões e os textos que os acompanham.

4.1 Discussão

Como visto e revisado no trabalho de ARAÚJO e colaboradores (2008), uma parte da aplicação da bioinformática ao campo acadêmico está na análise de rotinas e produção de maneiras de automatizar as mesmas com o intuito de aprimorar o desempenho da pesquisa com base na experiência. A produção do metadoon pipeline tenta trabalhar nesse segmento, produzindo desta forma uma ferramenta que aprimore o desempenho da pesquisa com base não só na experiência como no aprendizado.

Além de ressaltar a importância dada à bioinformática na educação curricular, WILLIAMS e colaboradores (2019) utilizam de questionários para entender as barreiras na integração da bioinformática, em ensino de graduação, através das respostas de 11.000 endereços em uma lista de correspondência de professores de biologia nos EUA (compradas em http://schooldata.com) e para outros membros de redes de professores com interesses na educação em ciências da vida. O questionário fez várias perguntas livres que buscaram entender as barreiras para o fator integração. As respostas foram analisadas por meio de uma abordagem de métodos mistos. A barreira mais comumente relatada pelos 1.260 entrevistados foi a falta de conhecimento/treinamento do corpo docente e, além disso, outros impedimentos

foram relatados, como: falta de interesse do aluno; currículos excessivamente cheios e falta de preparação dos alunos. Aplicações mais profundas da bioinformática foram feitas ainda por WILLIAMS campo educacional (como o "DNA-SUBWAY"), com o intuito de melhorar e facilitar o aprendizado que permite o acesso ao conhecimento da biologia computacional.

O DNA - SUBWAY é uma plataforma de código aberto para o armazenamento, análise e visualização de dados de sequenciamento de próxima geração e outros dados de biologia. Ele foi desenvolvido pelo Center for Plant Science Innovation e empacota ferramentas de bioinformática de nível de alto desempenho em pesquisa, computação em bancos de dados e fluxos de trabalho em uma interface fácil de usar (TAWDE & WILLIAMS et al., 2014).

Durante a sua pesquisa, ATHER e colaboradores (2020) produziram um pipeline para fins educacionais escrito em linguagem markdown. O nome dado ao pipeline foi "SeqAcademy". O pipeline enfatiza o uso de dados RNA-Seq e ChIP-Seq disponíveis publicamente e reúnem ferramentas populares que preenchem as lacunas entre as leituras cruas de sequenciamento e os insights biológicos. Nessa produção, os pesquisadores fizeram um conjunto de considerações conceituais para vários RNA-Seq e ChIP-Seq além disso apresentaram uma aplicação prática com etapas de análise com um caso de uso biológico em leveduras (modelo de texto e demonstração). O trabalho (Pipeline) de ATHER e colaboradores complementa os pipelines já existentes e mais utilizados para tratar dados de RNA-Seq e ChIP-Seq, introduzindo sutilmente a partir daí componentes críticos de análises desses dados a novos bioinformatas, utilizando facilidade.

Dessa forma é importante destacar que o Metadoon pipeline, tenta melhorar as primeiras experiências de usuário com as mecânicas e dispositivos bioinformatas, automatizando uma parte usual de programas comuns em workflows dentro da metagenômica como o Usearch. Além disso, com a produção do Metadoon, se cria uma maneira mais tátil de

entender e usar os padrões de produção presentes em pesquisas com metagenômica e rRNA 16S.

No mesmo segmento de educação sobre bioinformática, existem outros trabalhos também que conseguem descrever e produzir sobre. YI ZHANG e colaboradores (2018), ao estudarem os processos de avanço das ferramentas NGS em bioinformática, entenderam que o aprendizado para pesquisadores e estudantes da área se torna um desafio gradualmente ligado ao desenvolvimento dessas novas tecnologias. Com isso, desenvolveram “SequencEnG” , que é um recurso interativo da web para aprendizado ativo ou ensinamento de técnicas de NGS, através de protocolos detalhados graficamente e pipelines de análises detalhados de maneira descritiva. No “SequenceEng”, o usuário tem a oportunidade de interagir com uma árvore de conhecimento NGS para aprender sobre 66 técnicas NGS descritas e também explorar os pipelines de análise de dados NGS para reunir rapidamente uma lista de ferramentas populares e úteis no campo da bioinformática, dessa forma favorecendo contato e aprendizado para estudantes e pesquisadores da área.

Ao entender os trabalhos que se desenvolvem e baseiam-se em cima da educação de novos pesquisadores, especialmente bioinformatas, entendemos que este é um campo de aplicação com extrema importância. A necessidade de aplicação nessa área educacional da bioinformática deve-se ao constante avanço que ocorre no passo do desenvolvimento da tecnologia, como citado acima no trabalho de YI ZHANG e colaboradores (2018), portanto pensar mais em desenvolver facilidades e um melhor acesso a novas tecnologias é pensar no futuro da própria pesquisa, entendendo que quem as exerce precisa entender antes, passos graduais dos novos métodos e das novas aplicações.

É possível perceber que a melhoria no desempenho dos pesquisadores e estudantes de bioinformática, através do uso de dispositivos e técnicas computacionais é algo que se torna

popular entre as aplicações de pesquisa, justamente pela necessidade ressaltada na maioria dos trabalhos de adaptação às novas tecnologias, que surgem com o avanço das técnicas computacionais. Dessa forma é possível olhar para Metadoon pipeline como uma possível aplicação no campo educacional por conter comandos e procedimentos básicos de tratamento de workflows em metagenômica, por conseguir gerar dados que tanto ajudam a entender as amostras quanto ajudam a interpretar os próprios dados gerados e por utilizar de uma interface gráfica para ajudar a fazer procedimentos característicos da metagenômica de maneira mais fácil.

4.2 Perspectivas futuras

Neste tópico será abordado as perspectivas futuras pensadas sobre este trabalho, no sentido de torná-lo sempre melhor e necessário ao atender novas demandas em um possível novo tempo. O primeiro aspecto da existência do "Metadoon Pipeline" foi considerar em atender as necessidades educacionais para novos bioinformatas, assim como em trabalhos citados durante esta escrita. O segundo aspecto, foi considerar as necessidades em pesquisas com a metagenômica, associadas a geração facilidade de resultados para análise e a falta de interface gráfica para acompanhar esse aspecto.

Ao utilizar das primeiras impressões sobre o "Metadoon Pipeline", pretende-se acatar as mesmas com base na experiência do usuário. A interface gráfica será melhorada, ao passo que os usuários denunciarem defeitos ou possíveis melhorias para o pipeline em questão. Em um segundo passo, pretende-se que, em um futuro não tão distante, o Metadoon Pipeline rode tanto dados em campos de OTU's como em ASV; nesse mesmo sentido, que não rode apenas Usearch, mas que possa rodar também outros algoritmos em suas sequências de execução, permitindo ao usuário uma maior diversidade de resultados.

Entende-se que a publicação deste trabalho é importante para que o próprio produto gerado nesta pesquisa possa evoluir, sendo, portanto, necessário também divulgar este estudo por meios científicos. Dessa forma declara-se aqui como a publicação sendo uma das perspectivas futuras.

CONCLUSÃO

Com a produção e finalização deste trabalho, é possível concluir que a ideia de facilitar, divulgar e produzir ferramentas que possam atrelar pontos importantes de uma pesquisa, deve se tornar uma tendencia. A produção do metadoon pipeline consegue interligar as partes de uma pesquisa bioinformata-metagenômica (como; normalização e organização dos dados, pre processar, gerar gráficos) e ao mesmo tempo gerar reportes que auxiliam na análise do resultado.

A produção do pipeline foi feita com todas, se não a maioria das ferramentas descritas. A interface que facilita familiarização do pipeline para bioinformatas não habituados com linhas de comando, os resultados gerados que tanto facilitam o tempo de processo para conclusão de uma pesquisa quanto reduzem o tempo normalmente utilizado para gerar primeiros resultados sobre uma dada pesquisa. Dessa forma é possível descrever que os objetivos da produção do metadoon pipeline tiveram êxito; tendo para cada um deles uma solução funcional aplicada no produto final desta dissertação.

REFERÊNCIAS

ADAMOWICZ, S. J. et al. Trends in DNA barcoding and metabarcoding. Genome, v. 62, n. 3, p. v–viii, mar. 2019.

ALVES, L. DE F. et al. Metagenomic Approaches for Understanding New Concepts in Microbial Science. International Journal of Genomics, v. 2018, p. 1–15, 23 ago. 2018.

ARAÚJO, N. D. DE et al. A ERA DA BIOINFORMÁTICA: SEU POTENCIAL E SUAS IMPLICAÇÕES PARA AS CIÊNCIAS DA SAÚDE. Estudos de Biologia, v. 30, n. 70/72, 27 nov. 2008.

ATHER, S. H. et al. SeqAcademy: an Educational Pipeline for RNA-Seq and ChIP-Seq Analysis. F1000Research, v. 7, p. 628, 22 maio 2018.

ATLAS, R. M.; BARTHA, R.; ATLAS, D. Microbial Ecology: Fundamentals and Applications. [s.l.] Benjamin/Cummings, 1998. v. 4

BENDIA, A. G. et al. A Mosaic of Geothermal and Marine Features Shapes Microbial Community Structure on Deception Island Volcano, Antarctica. Frontiers in Microbiology, v. 9, 7 maio 2018.

BOLYEN, E., RIDEOUT, J.R., DILLON, M.R., BOKULICH, N.A., ABNET, C.C., AL-GHALITH, G.A. ET AL. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol, 37(8), 852–857.

CALLAHAN, B. J. et al. High-throughput Amplicon Sequencing of the full-length 16S rRNA Gene with single-nucleotide Resolution. Nucleic Acids Research, v. 47, n. 18, p. e103–e103, 3 jul. 2019.

CAPORASO, J.G., KUCZYNSKI, J., STOMBAUGH, J., BITTINGER, K., BUSHMAN, F.D., COSTELLO, E.K. ET AL. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat Methods, 7(5), 335–336.

CAPORASO, J.G., KUCZYNSKI, J., STOMBAUGH, J., BITTINGER, K., BUSHMAN, F.D., COSTELLO, E.K. ET AL. (2011). QIIME: quantitative insights into microbial ecology. Nat Methods, 7(5), 335–336.

CARUSO, V. et al. Performance of Microbiome Sequence Inference Methods in Environments with Varying Biomass. MSystems, v. 4, n. 1, 26 fev. 2019.

CHAPMAN, B.; CHANG, J. Biopython. ACM SIGBIO Newsletter, v. 20, n. 2, p. 15–19, ago. 2000.

COLWELL, R. K.; CODDINGTON, J. A. Estimating Terrestrial Biodiversity through Extrapolation. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, v. 345, n. 1311, p. 101–118, 29 jul. 1994.

DOS SANTOS, S. D. S. et al. Haemophilus influenzaeType b Immunization in Adults Infected with the Human Immunodeficiency Virus. AIDS Research and Human Retroviruses, v. 20, n. 5, p. 493–496, maio 2004.

EDGAR, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26(19), 2460-2461.

EDGAR, R. C. Search and Clustering Orders of Magnitude Faster than BLAST. Bioinformatics, v. 26, n. 19, p. 2460–2461, 12 ago. 2010.

EDGAR, R. C., HAAS, B. J., CLEMENTE, J. C., QUINCE, C. & KNIGHT, R. (2010). UClust-Unified Clustering Methods for Quantitative Microbial Community Analysis. The National Academy of Sciences, 107(27), 12014–12019. https://doi.org/10.1073/pnas.1002611107

EMBL-EBI. MGnify - EBI. Disponível em: <https://www.ebi.ac.uk/metagenomics/pipelines>. Acesso em: 22 out. 2022.

FILHO, M. A. C. DE P. P. F. (ED.). Metagenômica E Sua Aplicação No Estudo De Diversidade E Função De Microrganismos De Solos Do Cerrado. - Portal Embrapa. Disponível em: <https://www.embrapa.br/busca-de-publicacoes/-/publicacao/884750/metagenomica-e-sua-aplicacao-no-estudo-de-diversidade-e-funcao-de-microrganismos-de-solos-do-cerrado>. Acesso em: 22 out. 2022.

GOUJON, M. et al. A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Research, v. 38, n. Web Server, p. W695–W699, 3 maio 2010.

HEBERT, P. D. N. et al. Biological identifications through DNA barcodes. Proceedings. Biological sciences, v. 270, n. 1512, p. 313–21, 2003.

HENG, H. H. et al. A Postgenomic Perspective on Molecular Cytogenetics. Current Genomics, v. 19, n. 3, p. 227–239, 20 fev. 2018.

HENTSCHEL, U. et al. Molecular Evidence for a Uniform Microbial Community in Sponges from Different Oceans. Applied and Environmental Microbiology, v. 68, n. 9, p. 4431–4440, set. 2002.

HILL, M. O. Diversity and Evenness: a Unifying Notation and Its Consequences. Ecology, v. 54, n. 2, p. 427–432, mar. 1973.

HOLLINGSWORTH, P. M. et al. A DNA barcode for land plants. Proceedings of the National Academy of Sciences, v. 106, n. 31, p. 12794–12797, 30 jul. 2009.

HSU, T. et al. Urban Transit System Microbial Communities Differ by Surface Type and Interaction with Humans and the Environment. MSystems, v. 1, n. 3, 28 jun. 2016.

HUGHES, J. B. et al. Counting the Uncountable: Statistical Approaches to Estimating Microbial Diversity. Applied and Environmental Microbiology, v. 68, n. 1, p. 448–448, jan. 2002.

KLINDWORTH, A. et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Research, v. 41, n. 1, p. e1–e1, 28 ago. 2012.

KLINDWORTH, A., PRUESSE, E., SCHWEER, T., PEPLIES, J., QUAST, C., HORN, M., ... & GLOCKNER, F. O. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic acids research, 41(1), e1-e11.

LAMBAIS, M. R. et al. Diversidade Microbiana Nos solos: Definindo Novos Paradigmas. Tópicos Em Ciência Do solo, v. 4, 2005.

LI, Y. et al. VIP: an Integrated Pipeline for Metagenomics of Virus Identification and Discovery. Scientific Reports, v. 6, n. 1, 30 mar. 2016.

LIU, J.; YAN, H.-F.; GE, X.-J. The Use of DNA Barcoding on Recently Diverged Species in the Genus Gentiana (Gentianaceae) in China. PLOS ONE, v. 11, n. 4, p. e0153008, 6 abr. 2016.

Lundh, Fredrik. An Introduction to Tkinter. (1999).

MELO, A. T. O.; BARTAULA, R.; HALE, I. GBS-SNP-CROP: a reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. BMC Bioinformatics, v. 17, n. 1, 12 jan. 2016.

MELO, H. V. F. Desenvolvimento De Um Pipeline Para Análise Genômica E Transcriptomica Com Base Em Web Services. Dissertação De Mestrado—UNIVERSIDADE FEDERAL DE SÃO CARLOS; PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA: [s.n.].

MITCHELL, A. L. et al. MGnify: the Microbiome Analysis Resource in 2020. Nucleic Acids Research, 7 nov. 2019.

MOHAMMED, Y.; PALMBLAD, M. Simple Proteomics Data Analysis in the Object-Oriented PowerShell. Mass Spectrometry Data Analysis in Proteomics, p. 379–391, 2013.

NAYFACH, S. et al. A Genomic Catalog of Earth's Microbiomes. Nature Biotechnology, v. 39, n. 4, p. 499–509, 9 nov. 2020.

PEIXOTO, B. M. Classificação De Sequências E Análise De Diversidade Em Metagenômica. Dissertação De Mestrado—Universidade Estadual de Campinas: Instituto de Computação: [s.n.].

PRODAN, A. et al. Comparing Bioinformatic Pipelines for Microbial 16S rRNA Amplicon Sequencing. PLOS ONE, v. 15, n. 1, p. e0227434, 16 jan. 2020.

RODRIGUES, J. Introdução À Bioinformática. Disponível em: <https://www.revista-programar.info/artigos/introducao-a-bioinformatica/>. Acesso em: 14 ago. 2021.

SCHLOSS, P. D., WESTCOTT, S. L., RYABIN, T., HALL, J. R., HARTMANN, M., HOLLISTER, E. B., LESNIEWSKI, R. A., OAKLEY, B. B., PARKS, D. H., ROBINSON, C.

J., ET AL. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol., 75(23), 7537-7541.

SEBESTA, R. W. Conceitos De Linguagens De Programação - 11.ed. [s.l.] Bookman Editora, 2018.

SEGATA, N., IZARD, J., WALDRON, L., GEVERS, D., MIROPOLSKY, L., GARRETT, W. S., & HUTTENHOWER, C. (2011). Metagenomic microbial community profiling using unique clade-specific marker genes. Nature methods, 8(9), 811-814.

SHANNON, C. E. A Mathematical Theory of Communication. Bell System Technical Journal, v. 27, n. 3, p. 379–423, 1948.

SILVEIRA, É. L. DA. Identificação De Comunidades Bacterianas De Solo Por Sequenciamento Do Gene 16S rRNA. Dissertação De Mestrado—Repositório Institucional UNESP: [s.n.].

SIMPSON, E. H. Measurement of Diversity. Nature, v. 163, n. 4148, p. 688–688, abr. 1949.

SOARES, M. F.; FRANCO, M. Aplicações Da Biologia Molecular À Patologia Renal: Revisão Da Literatura. J. Bras. Nefrol., v. 29, n. 2, p. 95–102, 2 abr. 2007a.

SOARES, M.; FRANCO, M. Aplicações Da Biologia Molecular À Patologia Renal: Revisão Da Literatura Application of Molecular Biology to Renal Pathology: Review of the Literature Artigo De Revisão. [s.l: s.n.]. Disponível em: <https://bjnephrology.org/wp-content/uploads/2019/08/jbn_v29n2a7.pdf>. Acesso em: 22 out. 2022.

SPELLERBERG, I. F.; FEDOR, P. J. A Tribute to Claude Shannon (1916-2001) and a Plea for More Rigorous Use of Species richness, Species Diversity and the “Shannon-Wiener” Index. Global Ecology and Biogeography, v. 12, n. 3, p. 177–179, 11 abr. 2003.

TUCKER, C. M. et al. A Guide to Phylogenetic Metrics for conservation, Community Ecology and Macroecology. Biological Reviews, v. 92, n. 2, p. 698–715, 20 jan. 2016.

TUOMISTO, H. A Consistent Terminology for Quantifying Species diversity? Yes, It Does Exist. O Ecologia, v. 164, n. 4, p. 853–860, 27 out. 2010b.

TUOMISTO, H. A Diversity of Beta diversities: Straightening up a Concept Gone awry. Part 1. Defining Beta Diversity as a Function of Alpha and Gamma Diversity. Ecography, v. 33, n. 1, p. 2–22, fev. 2010a.

VANDAMME, P. et al. Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiological reviews, v. 60, n. 2, p. 407–438, 1996.

Vavrek MJ (2011). “Fossil: palaeoecological and palaeogeographical analysis tools.” Palaeontologia Electronica, 14(1), 1T. R package version 0.4.0.

WARD, R. D. et al. DNA barcoding Australia’s fish species. Philosophical Transactions of the Royal Society B: Biological Sciences, v. 360, n. 1462, p. 1847–1857, 15 set. 2005.

WILLIAMS, J. et al. DNA Subway -An Educational Bioinformatics Platform for Gene and Genome Analysis: DNA Barcoding, and RNA-Seq. [s.l: s.n.]. Disponível em: <https://www.asas.org/docs/default-source/wcgalp-proceedings-oral/227_paper_9798_manuscript_948_0.pdf?sfvrsn=2>. Acesso em: 27 maio. 2021.

WILLIAMS, J. J. et al. Barriers to Integration of Bioinformatics into Undergraduate Life Sciences education: a National Study of US Life Sciences Faculty Uncover Significant Barriers to Integrating Bioinformatics into Undergraduate Instruction. PLOS ONE, v. 14, n. 11, p. e0224288, 18 nov. 2019.

WOOD, D. E., SALZBERG, S. L., & HAFT, D. H. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology, 15(3), R46

ZATZ, M. A biologia molecular contribuindo para a compreensão e a prevenção das doenças hereditárias. Ciência & Saúde Coletiva, v. 7, n. 1, p. 85–99, 2002.

ZHANG, Y. et al. SequencEnG: an Interactive Knowledge Base of Sequencing Techniques. Bioinformatics, v. 35, n. 8, p. 1438–1440, 8 set. 2018.

ANEXOS

Anexo I - Manual de Uso Para o Metadoon Pipeline

Instalação:

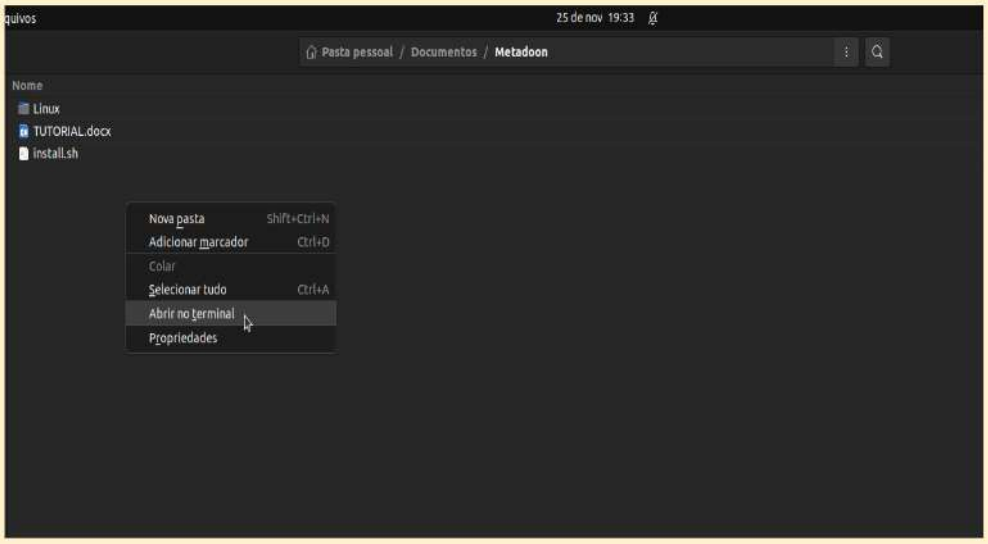
1-PASSOS INICIAIS

OBS: Somente execute os passos a seguir quando tiver certeza de que baixou TODOS os arquivos para instalação:

Arquivos disponíveis em : https://mega.nz/folder/67xwSZRb#E6DdY7j01xbEda7Do5gHOQ

- Metadoon-Pipeline apresenta um script de instalação denominado “install.sh”;
- O script normalmente está ao lado da pasta “Linux”;
- Você deve instalar o pipeline e suas dependências pelo script em questão;
- Para isso dê as permissões de execução para esse script pelo terminal.

~Abra a pasta que contém a pasta “Linux” e o arquivo de instalação “install.sh” com o terminal.



Em seguida digite;

$chmod 777 install.sh

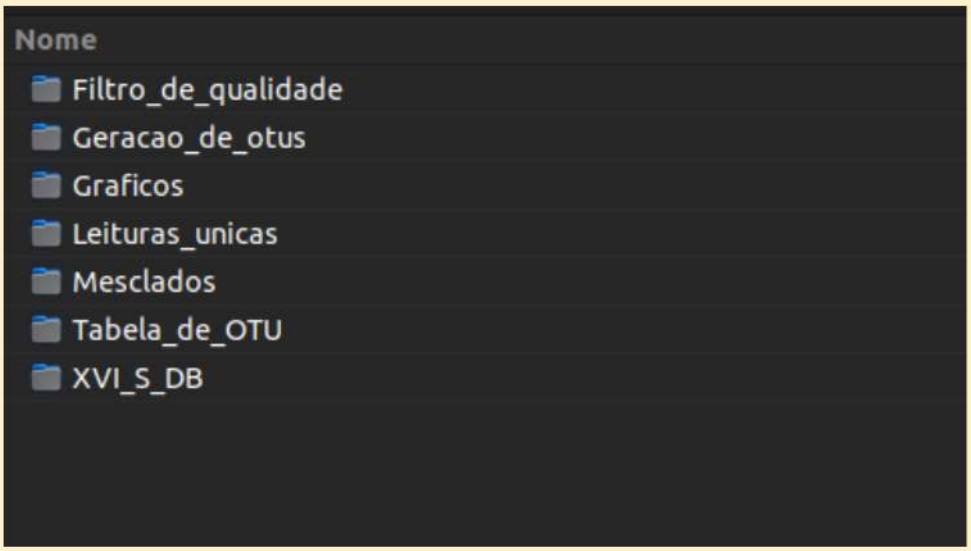
Para dar permissão de execução ao script de instalação. Em seguida digite;

$./install.sh

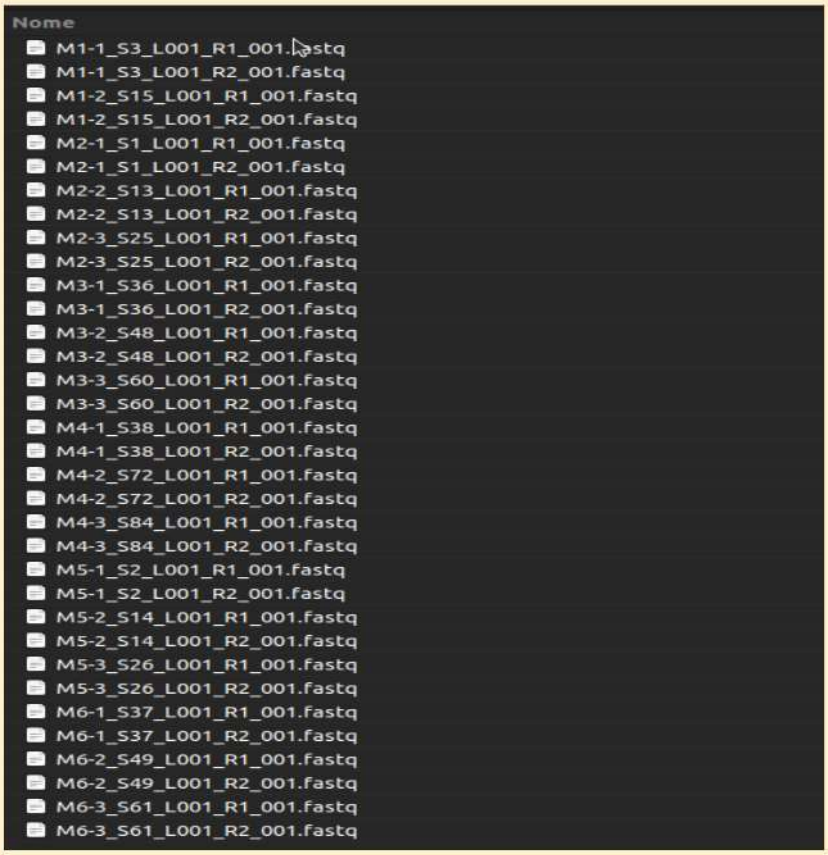
Para rodar o script no terminal Linux.

Sobre as Pastas:

- A pasta “Metarquivos” é onde o Metadoon organiza seus dados pré-processados.
    - Dentro dessa pasta são armazenados os arquivos pré-processados com Usearch.



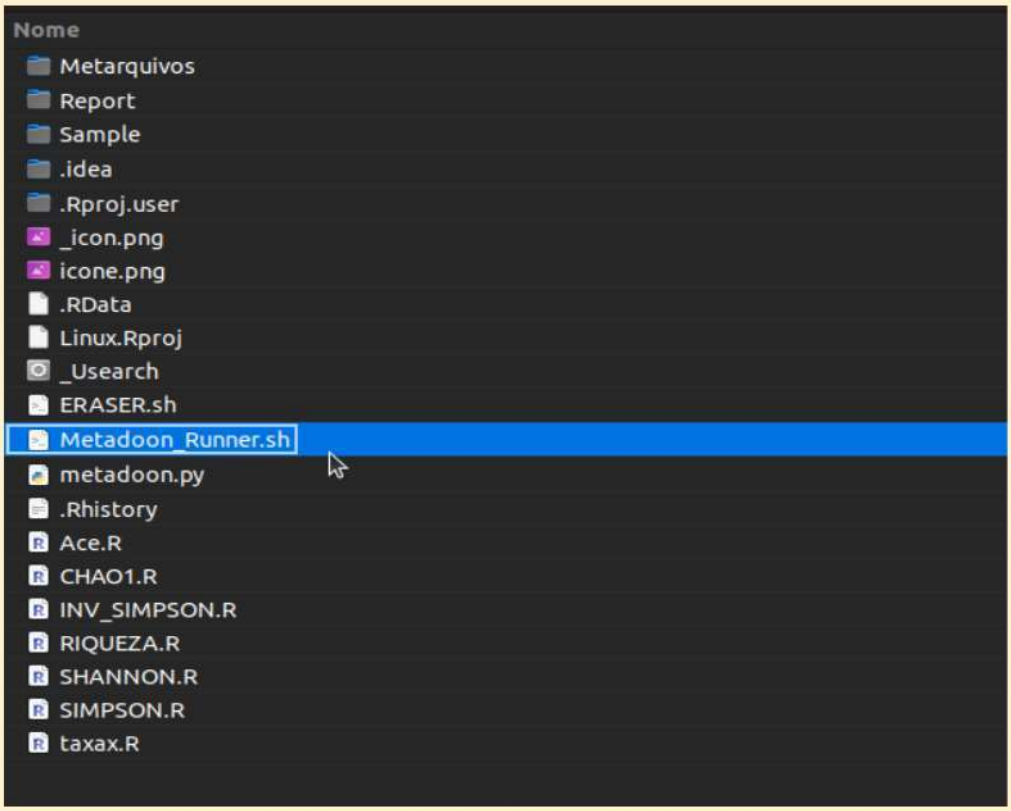
- A pasta Report é onde seus reports finais são salvos.
- A pasta “Sample” é onde seus dados “fastq” são trazidos para raiz do pipeline, para servirem de entrada de dados ao início do pré processamento com USEARCH.



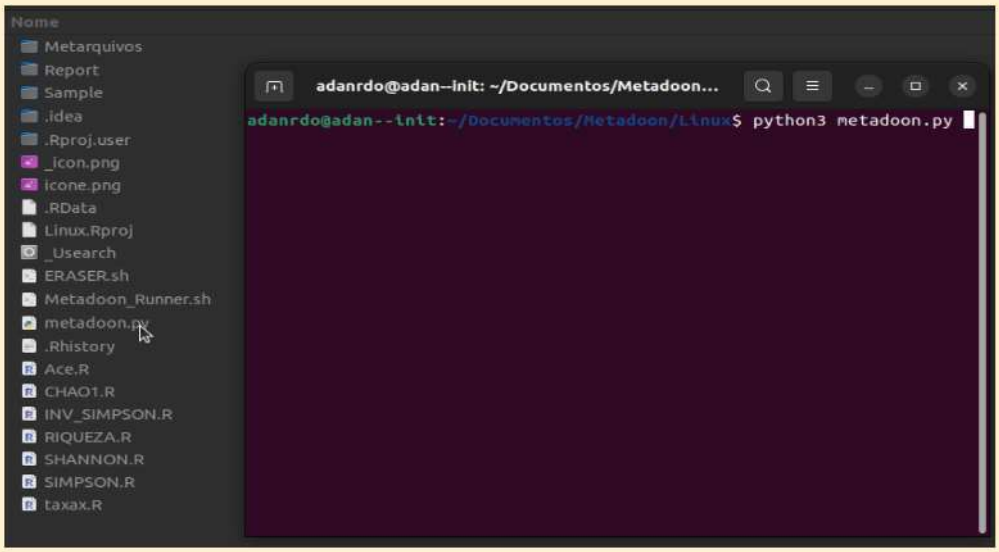
- O conteúdo na pasta Sample é uma cópia do diretório que você deve selecionar (Diretório onde estão seus arquivos “fastq”) ao início do pipeline.

Utilização:

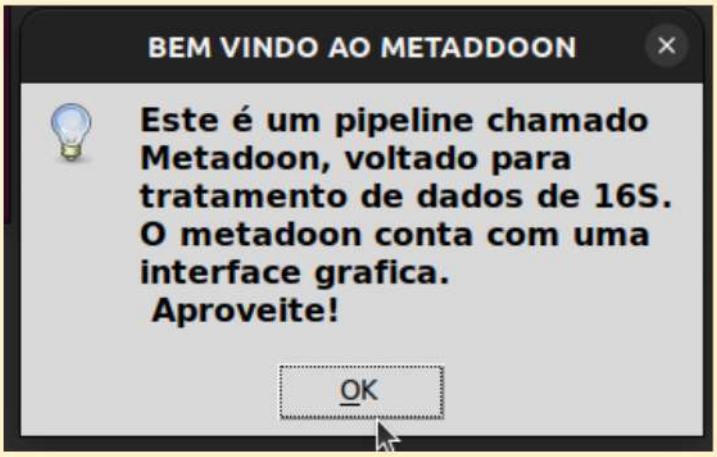
- Depois de instalado, você pode abrir o Metadoon Pipeline executando o “Metadoon_Runner.sh”;



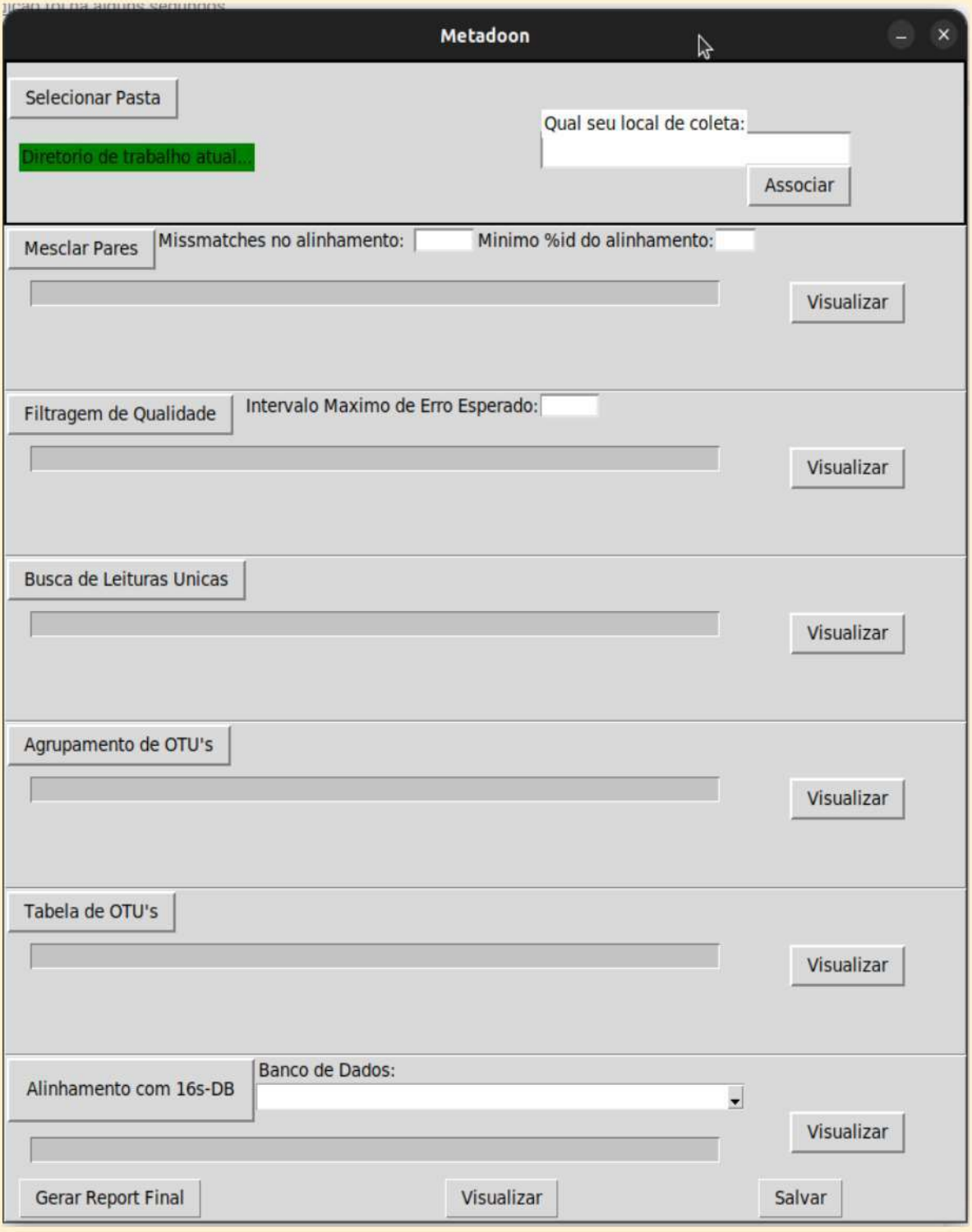
Ou executando o script Python “metadoon.py” ultilizando o python3 pelo terminal;



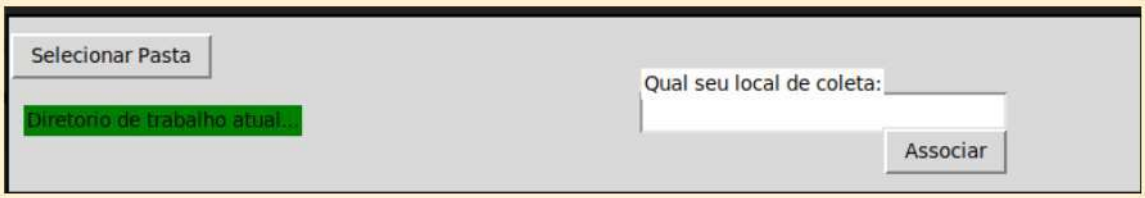
- Ao executar o Pipeline, você verá essa mensagem:



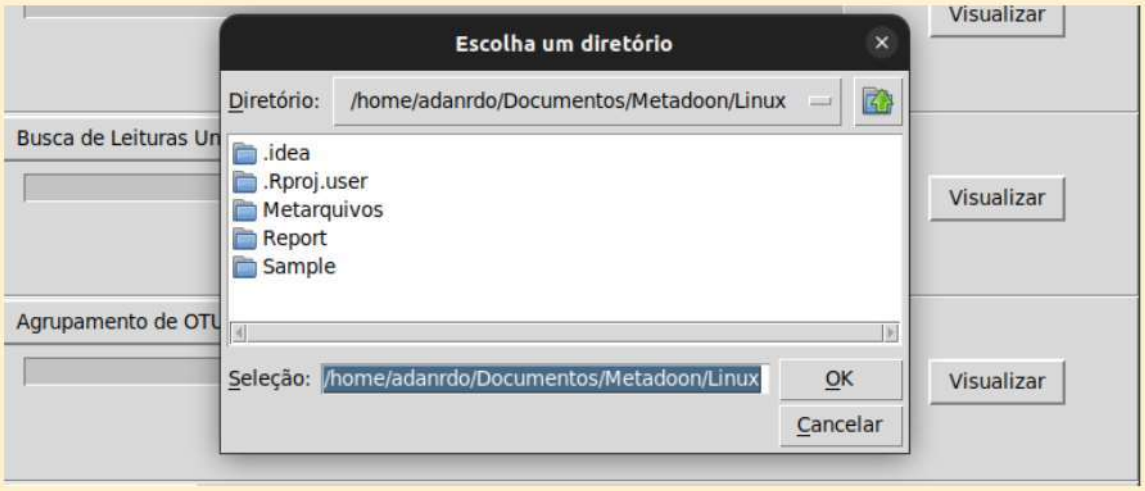
- Ao clicar em “ok”, ou fechar esta mensagem, você verá a face inicial do pipeline:


Metadoon
Selecionar Pasta
Diretorio de trabalho atual...
Qual seu local de coleta:
Associar
Mesclar Pares
Missmatches no alinhamento:
Minimo %id do alinhamento:
Visualizar
Filtragem de Qualidade
Intervalo Maximo de Erro Esperado:
Visualizar
Busca de Leituras Unicas
Visualizar
Agrupamento de OTU's
Visualizar
Tabela de OTU's
Visualizar
Alinhamento com 16s-DB
Banco de Dados:
Visualizar
Gerar Report Final
Visualizar
Salvar

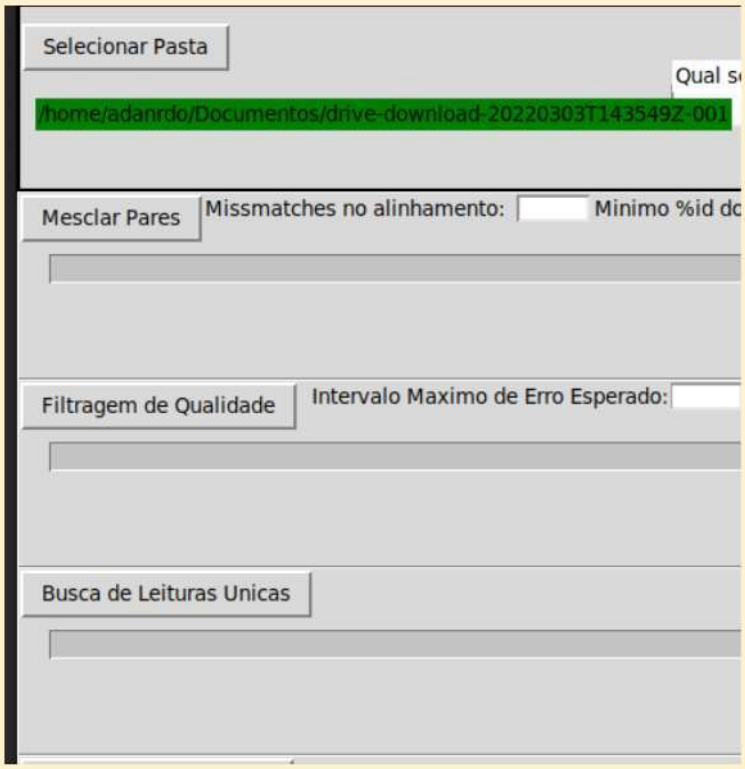
2- ENTENDENDO O PIPELINE

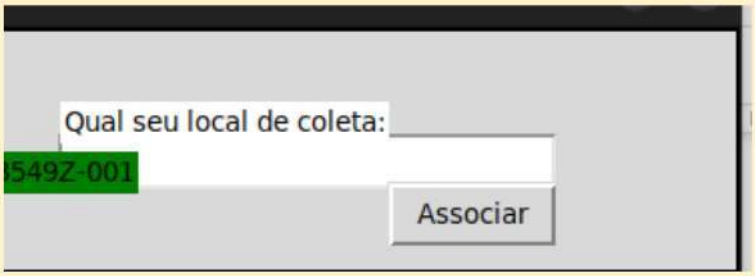


- O botão "Selecionar pasta", irá abrir uma caixa de seleção que pedirá ao usuário o caminho da origem dos arquivos "FASTQ" de seus dados;



- Em seguida o diretório selecionado deverá aparecer na região marcada em verde.





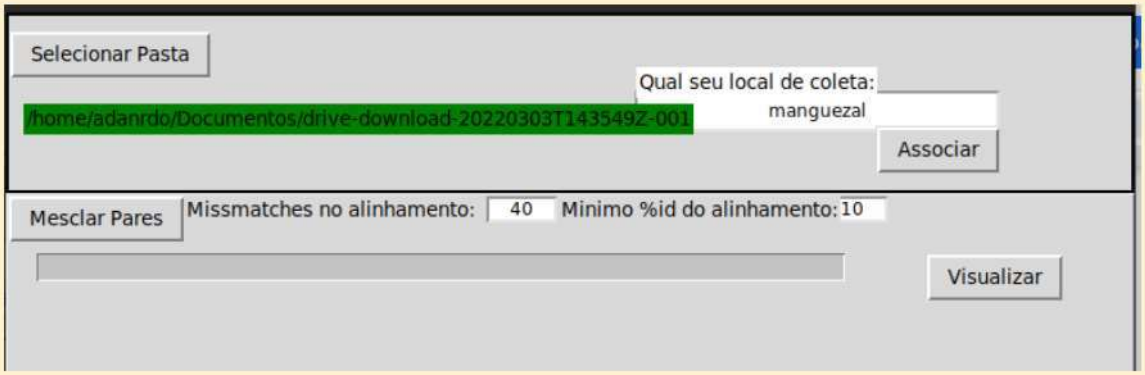
- Nessa área você deve identificar seus arquivos finais pelo nome que você deseja, a sugestão é que seja pelo local ou ecossistema em que realizou sua coleta.
- Os botões “Mesclar Pares”, “Filtragem de Qualidade”, “Busca de Leituras Únicas", “Agrupamento de OTU’s” e “Tabela de OTU’s”, são botões que executam comandos do USEARCH.
- O botão “Alinhamento com 16s-DB” além de executar um comando do Usearch, também joga os resultados para Rscript, a fim de proceder com as análises.

O esquema abaixo ilustra e relaciona os botões aos comandos Usearch:

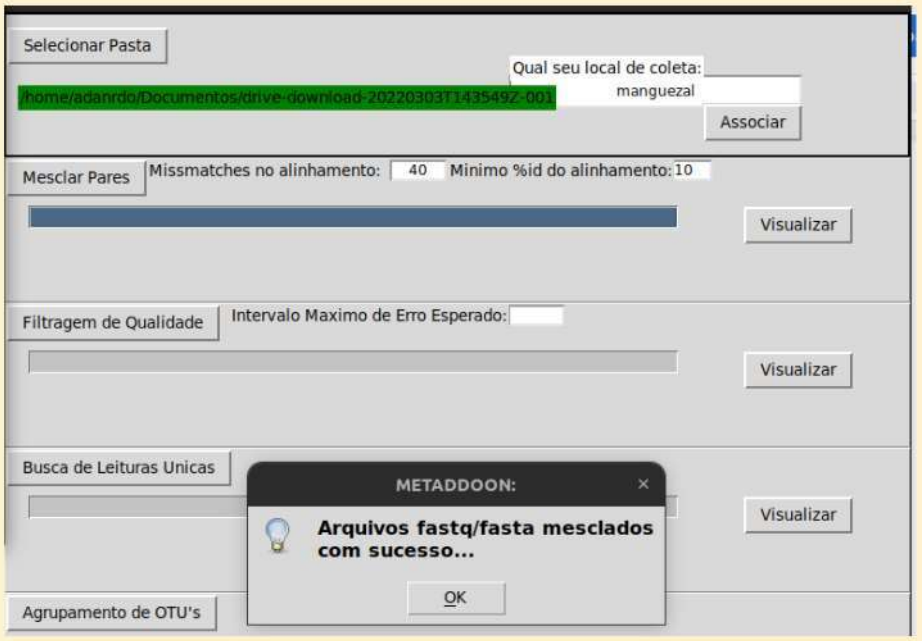
NOME DOS BOTÕES	COMANDOS USEARCH
Mesclar Pares	“fastq_mergepairs”;
Filtragem De Qualidade	“fastq_filter”;
Busca De Leituras Unicas	“fastx_uniques”;
Agrupamento De Otu’s	“cluster_OTUs”;
Tabela De Otu’s	“usearch_global”;
Alinhamento Com 16s-Db	“nbc_tax”.

2.1 MESCLANDO PARES

- Antes de mesclar os pares você deve preencher o número máximo de incompatibilidades (mismatches) no alinhamento (Padrão 50, mas Considere aumentar se você tiver sobreposições longas).
- Antes de mesclar os pares você deve preencher também o mínimo de identidade das suas amostras no alinhamento (Minimo %id do alinhamento), Padrão 90. Considere diminuir se você tiver sobreposições longas.



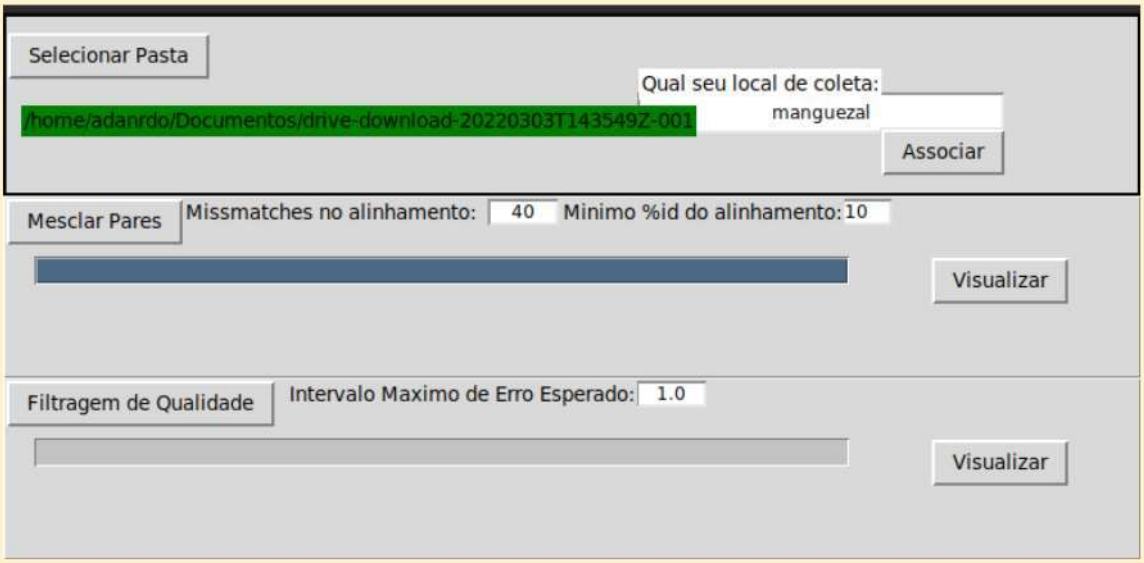
- Depois clique em “Mesclar Pares”, e espere a barra inferior ser preenchida e/ou o aviso aparecer na tela;



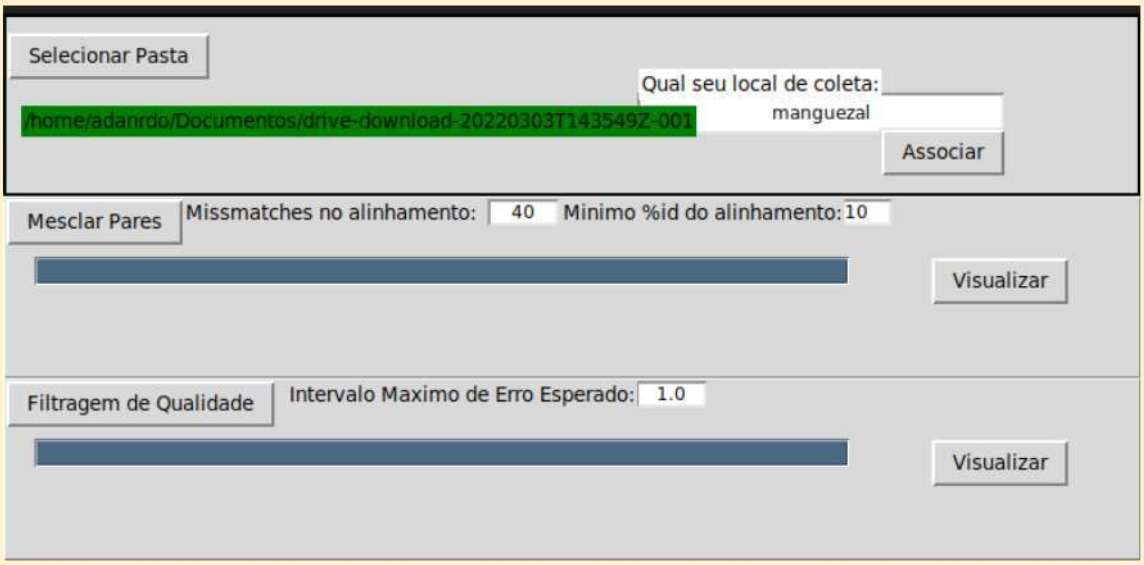
- A mesclagem dos seus dados foi feita:✅

2.2 FILTRO DE QUALIDADE

- Antes de executar a filtragem de qualidade, você deve estabelecer um intervalo de erro esperado.
- É recomendado usar a configuração do limite máximo de erro esperado para 1,0



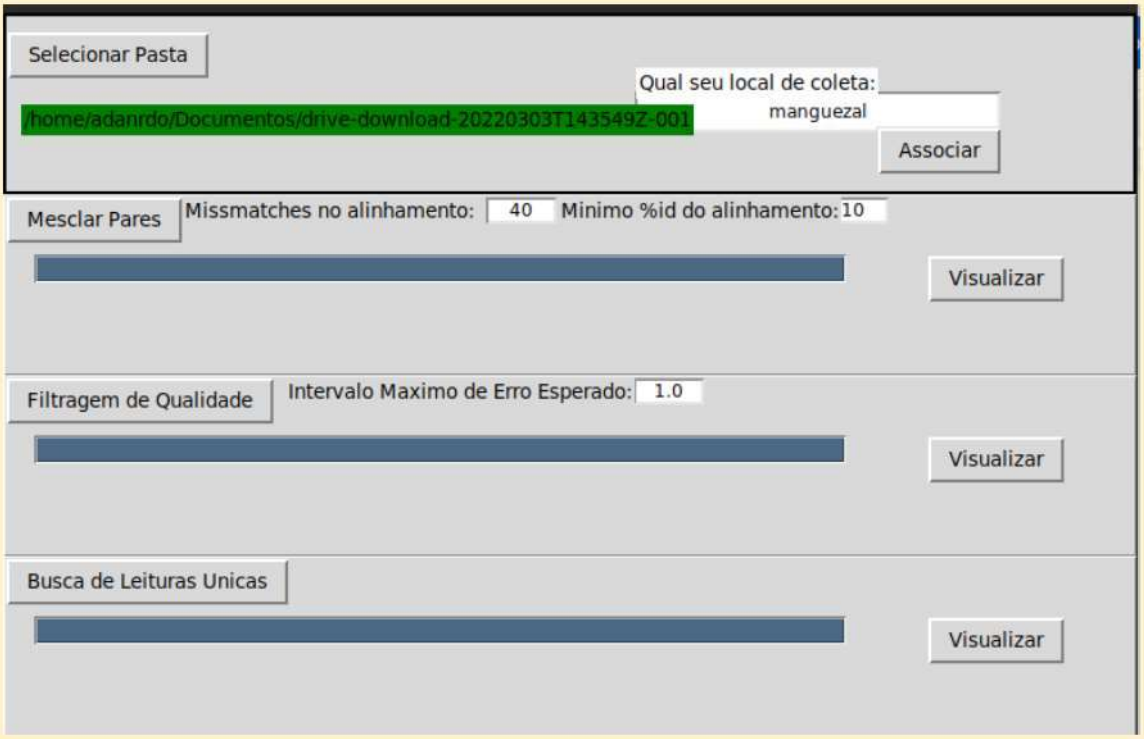
- Depois clique em "Filtragem de qualidade", e espere a barra inferior ser preenchida e/ou o aviso aparecer na tela;



- A filtragem dos seus dados foi feita:✅

2.3 BUSCANDO POR LEITURAS ÚNICAS

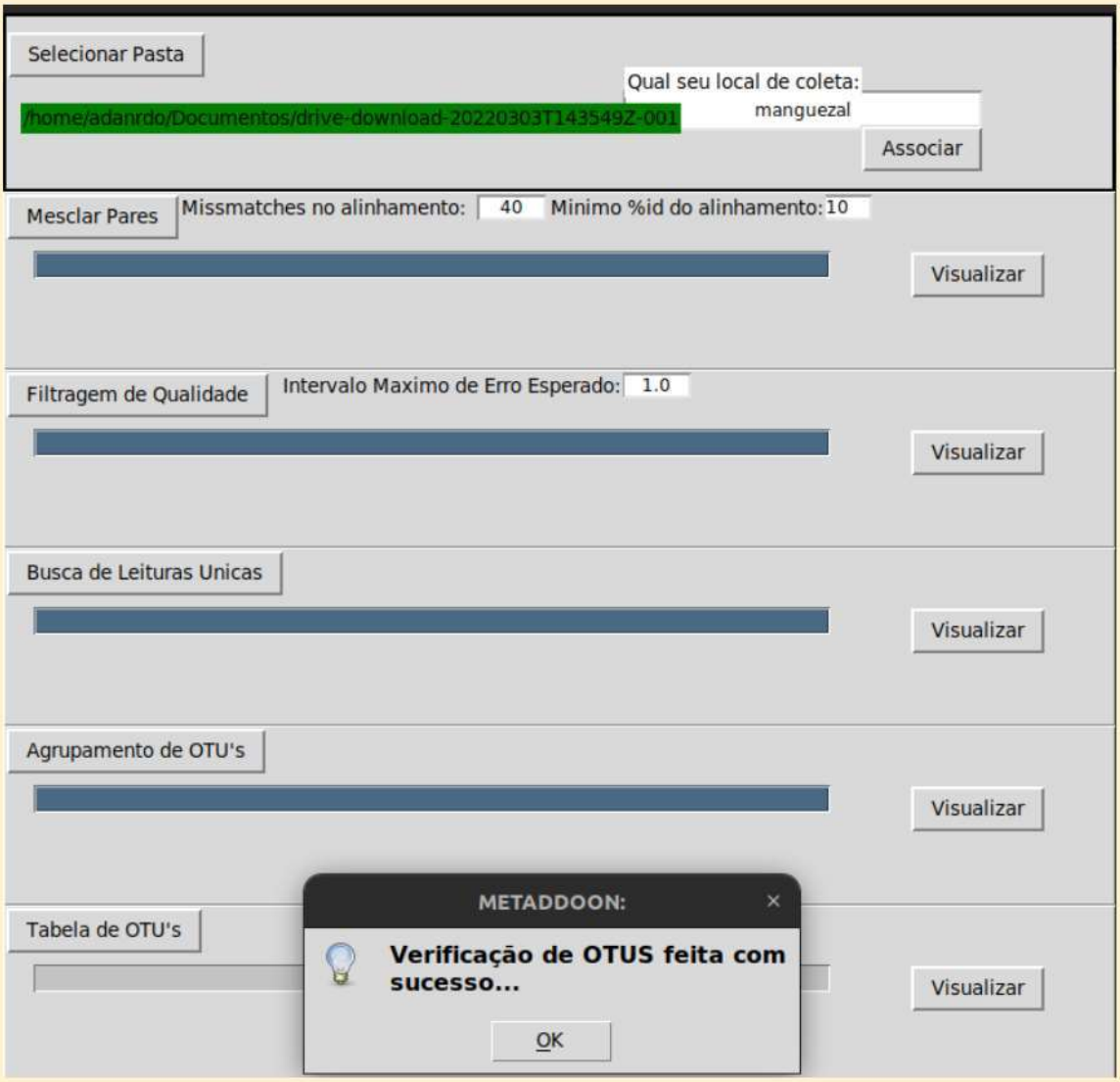
- Nesse passo, após, mesclar e filtrar seus dados, você só precisa apertar em “Buscar Leituras Únicas" e esperar a barra inferior ser preenchida;



- As leituras únicas dos seus dados foi armazenada:✅

2.4 AGRUPAMENTO DE OTU’S

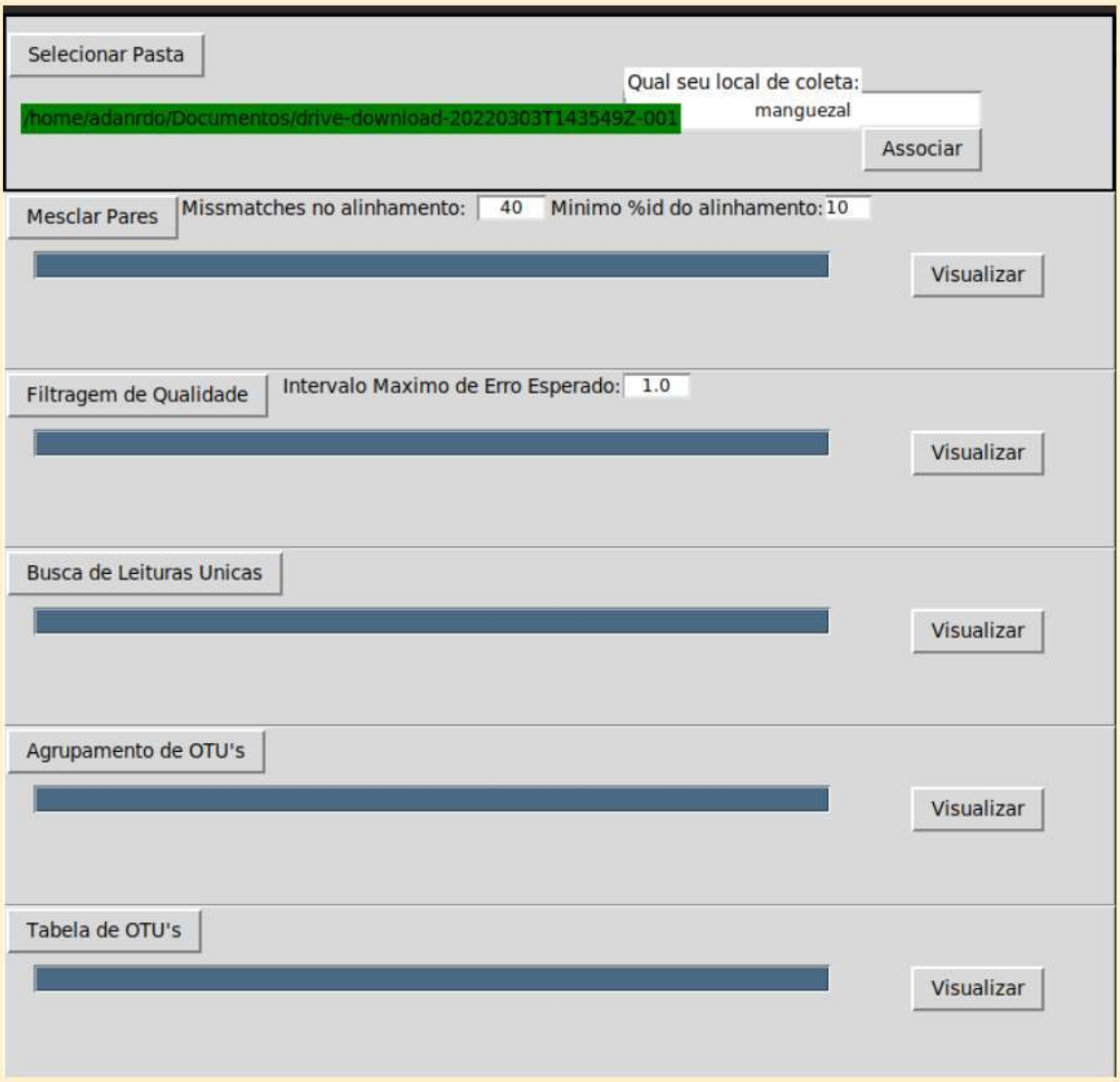
- Nesse passo, após, mesclar, filtrar e agrupar as leituras únicas dos seus dados, você só precisa apertar em “Agrupamento de OTU’s" e esperar a barra inferior ser preenchida;



- Os OTU’s foram agrupados com sucesso:✅

2.5 CRIANDO TABELA DE OTU’S

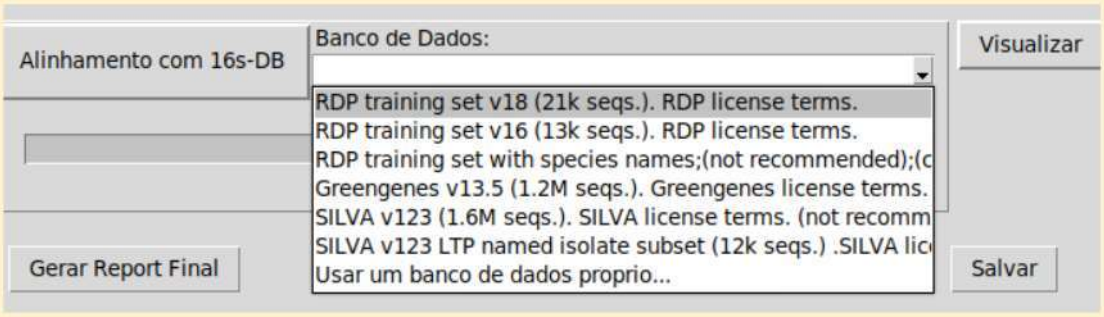
- Nesse passo, após, mesclar, filtrar, agrupar as leituras únicas e agrupar os OTU’s dos seus dados dos seus dados você só precisa apertar em “Tabela de OTU’s" e esperar a barra inferior ser preenchida;



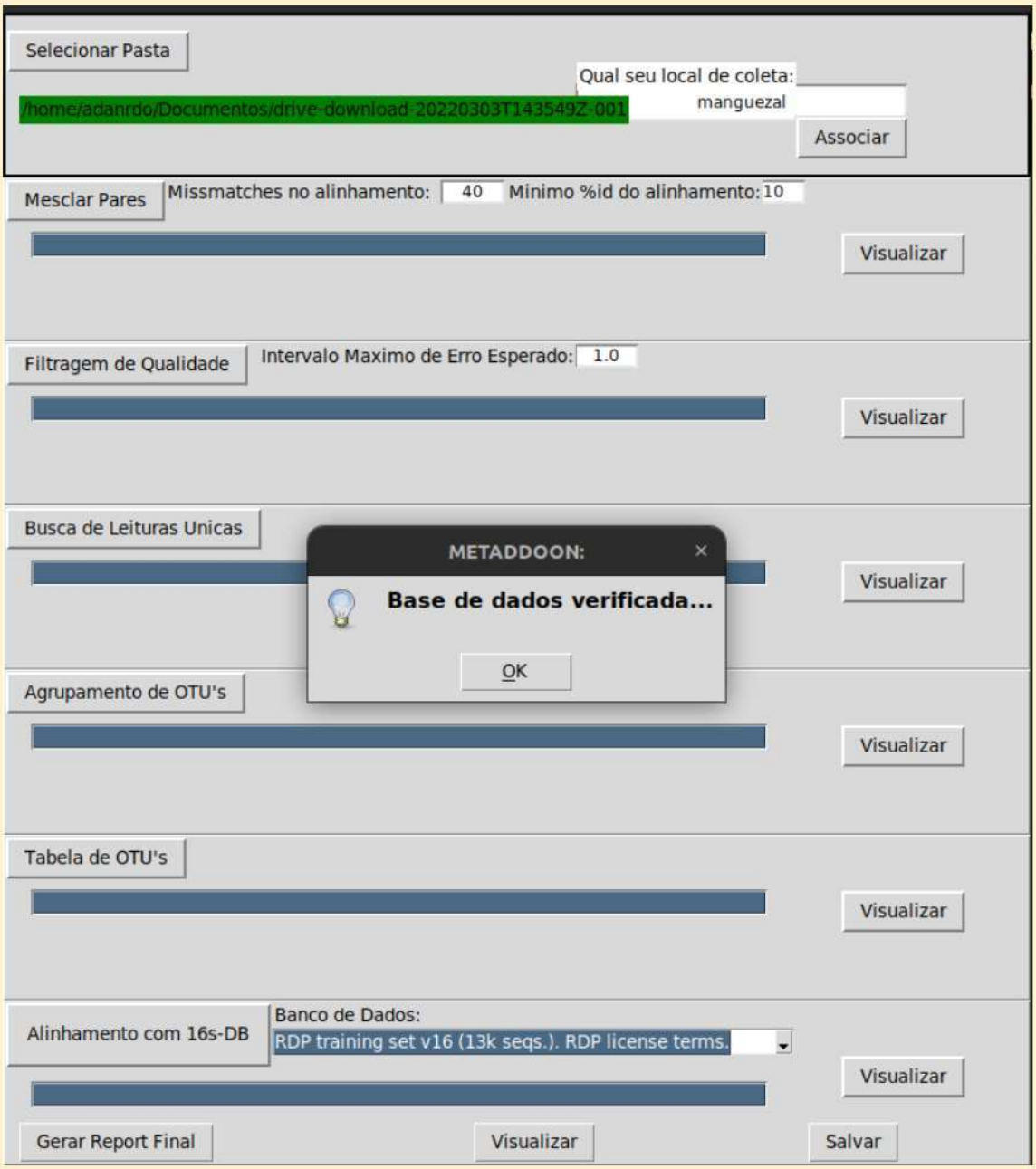
- A tabela de OTU's foi gerada com sucesso:✅

2.6 CRIANDO ARQUIVO DE TAXONS

- Nesse passo, você deve selecionar na selectbox qual banco de dados vc quer usar para comparar aos seus dados:



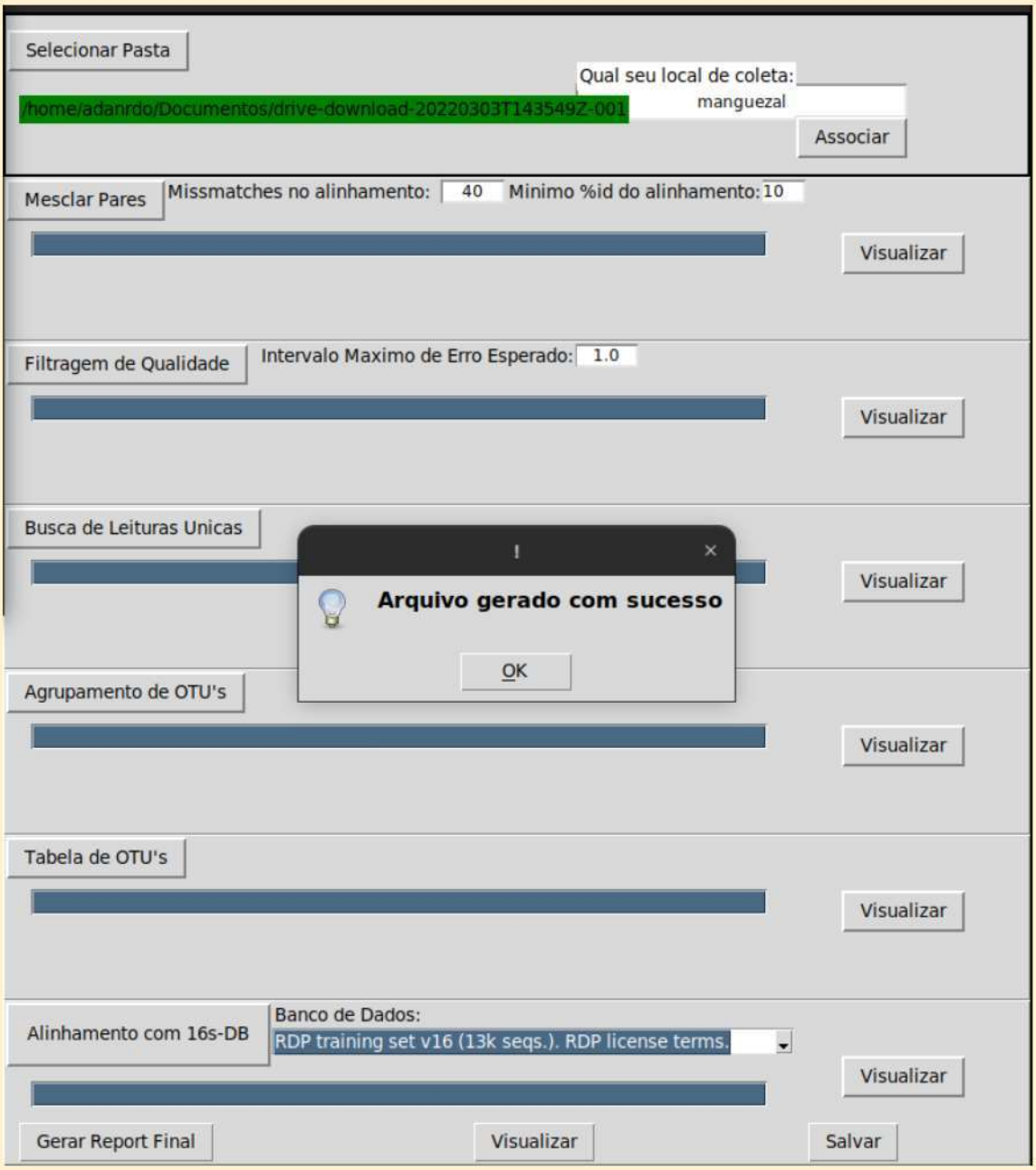
- Caso não deseje nem um dos listados, clique na ultima opção "Usar um banco de dados proprio..."



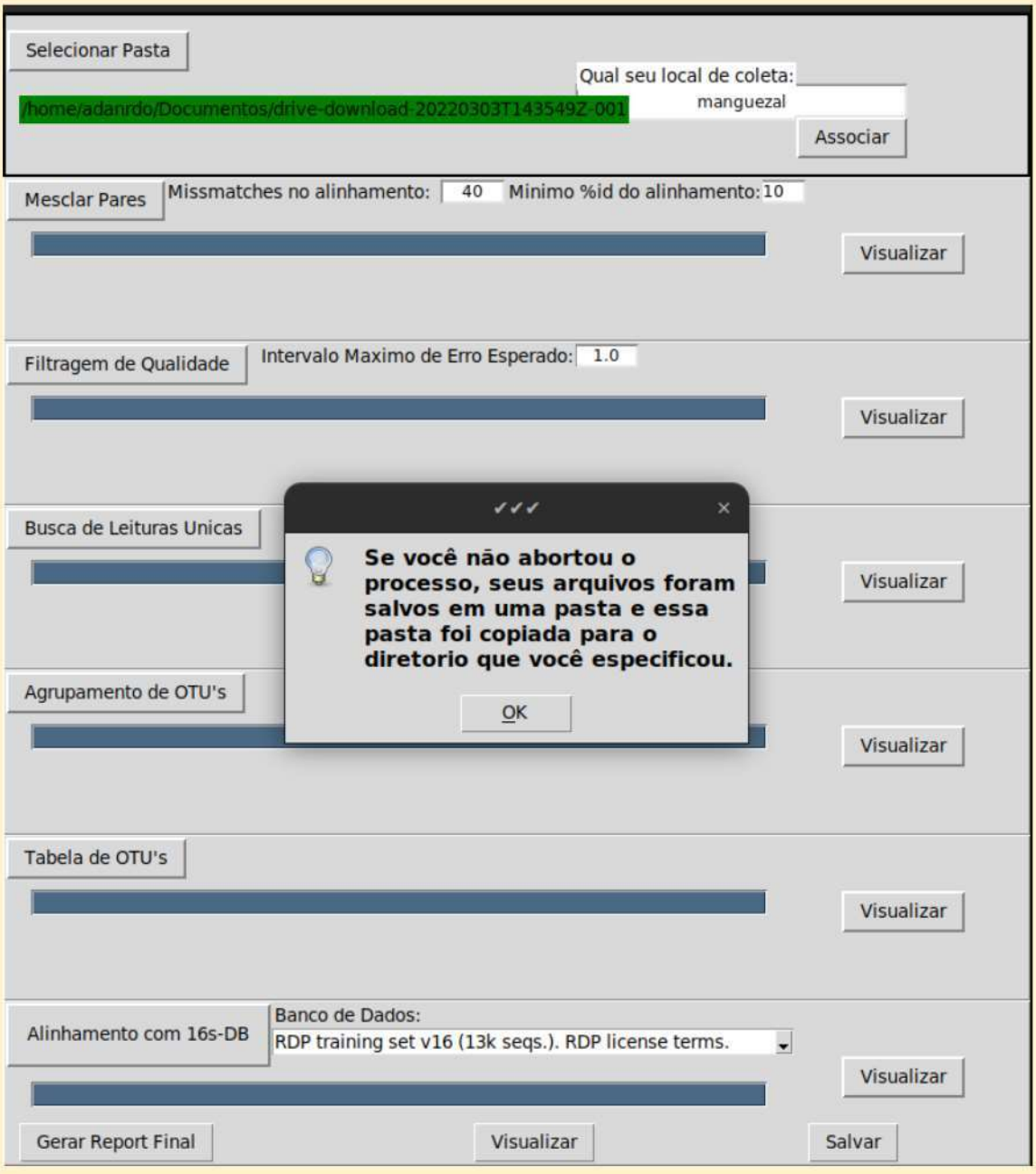
- A base de dados foi verificada com sucesso:✅

3-REPORT FINAL

- Para gerar o report final vc deve clicar em “Gerar Report final”;

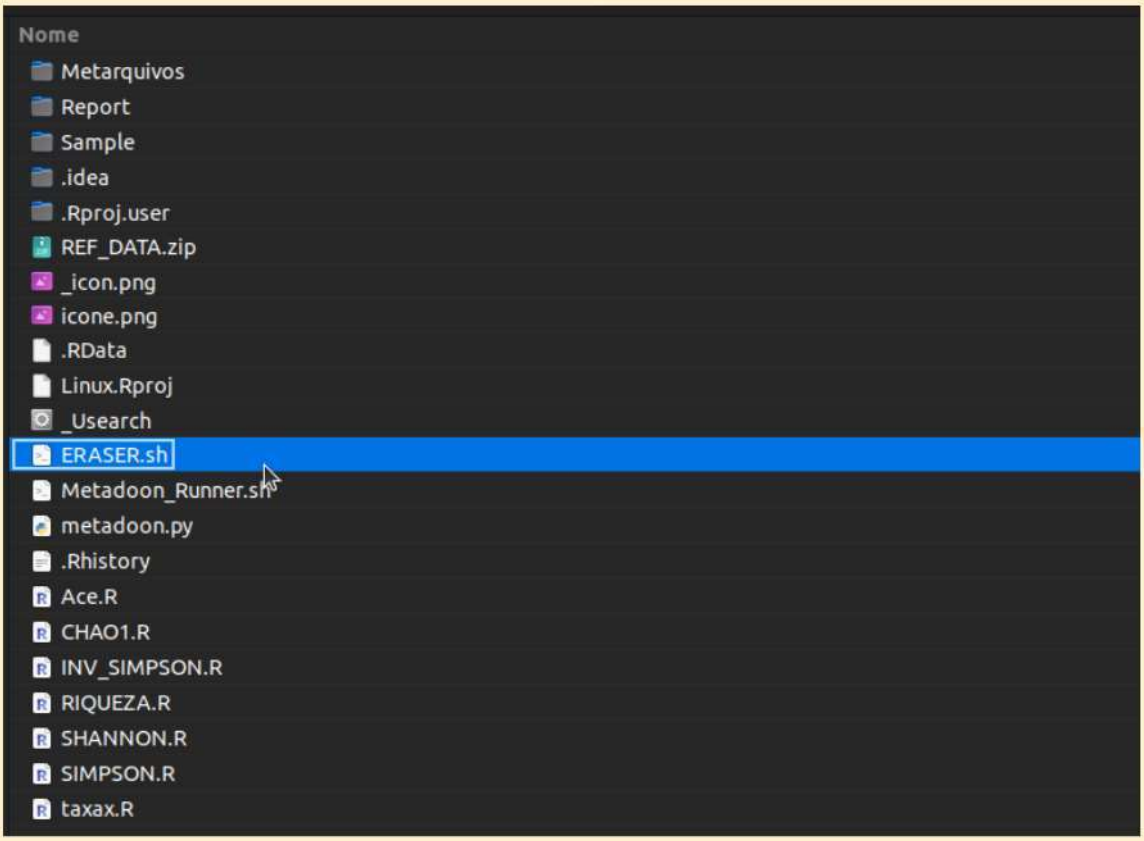
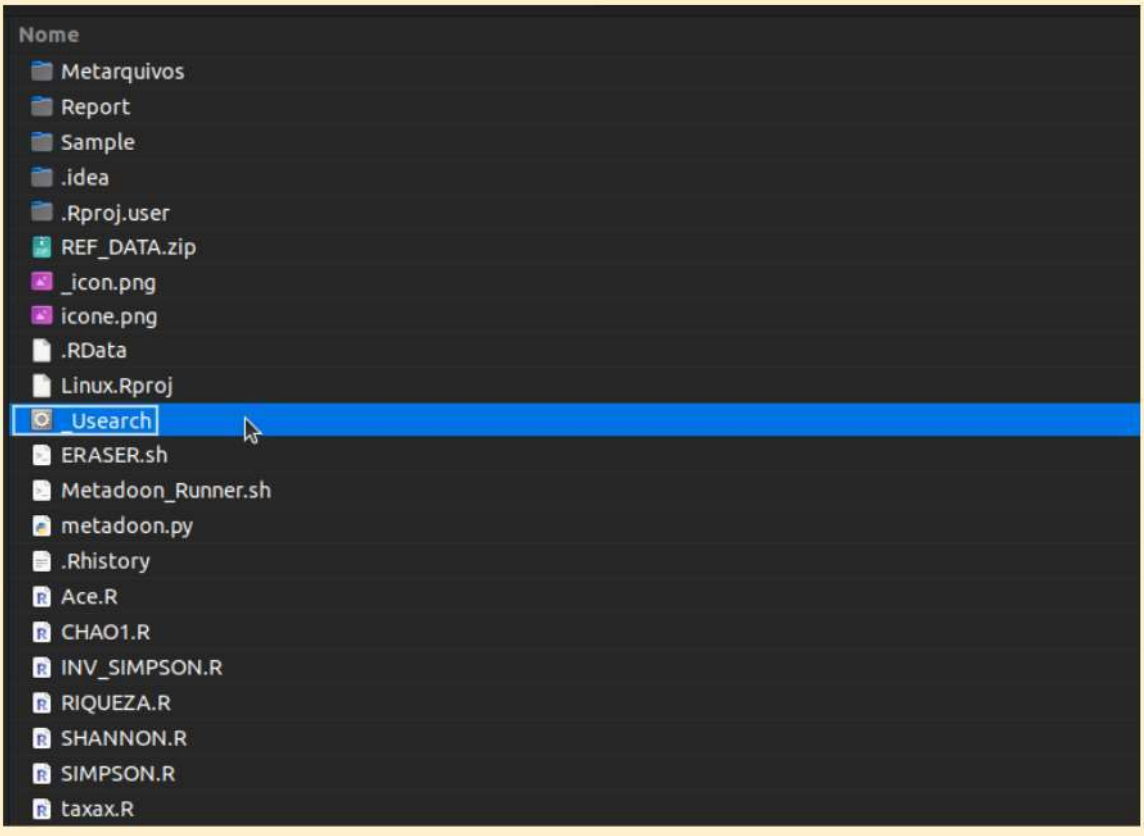


- O report final foi gerado com sucesso:✅
- O botão de visualizar, abre os dados finais gerados.
- O botão salvar abre uma caixa de seleção que permite o usuário salvar seus dados em uma pasta separada.
-



Observações:

- AO FINAL DO PROCESSO VC DEVE RODAR O SCRIPT “ERASER.sh” PARA LIMPAR OS DADOS PARA UM PRÓXIMO USO.



- Se você deseja usar a versão 64bits do Usearch neste pipeline, basta obter licença, e substituir o arquivo 32bits do Usearch pelo arquivo 64bits.