INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
DE MINAS GERAIS (IFMG)
CAMPUS BAMBUÍ
LICENCIATURA EM CIÊNCIAS BIOLÓGICAS

Willian Mendes Gonçalves

MONTAGEM DE AMBIENTE DE ANÁLISES ECOLÓGICAS DE MICROBIOMAS USANDO A PLATAFORMA QIIME2

Bambuí – MG
2024

WILLIAN MENDES GONÇALVES

MONTAGEM DE AMBIENTE DE ANÁLISES ECOLÓGICAS DE MICROBIOMAS USANDO A PLATAFORMA QIIME2

Trabalho de conclusão de curso apresentado ao Curso de Licenciatura em Ciências Biológicas do Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG) – *Campus* Bambuí para obtenção do grau de Licenciatura em Ciências Biológicas.

Orientador: Gustavo Augusto Lacorte

Bambuí – MG
2024

Catalogação na Fonte Biblioteca IFMG - Campus Bambuí

G635e Gonçalves, Willian Mendes.
Montagem de ambiente de análises ecológicas de microbiomas usando a Plataforma Qiime2. / Willian Mendes Gonçalves. – Bambuí, 2024.
80 f.: il.; color.

Orientador: Gustavo Augusto Lacorte.
Trabalho de Conclusão de Curso (graduação) - Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais – Campus Bambuí, MG, Curso Licenciatura em Ciências Biológicas, 2024.

1. Bioinformática. 2. QIIME 2. 3. Microbioma. I. Lacorte, Gustavo Augusto. II. Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais – Campus Bambuí, MG. III. Título.

CDD 631.46

Elaborada por Douglas Bernardes de Castro- CRB-6/2802

Willian Mendes Gonçalves

MONTAGEM DE AMBIENTE DE ANÁLISES ECOLÓGICAS DE MICROBIOMAS USANDO A PLATAFORMA QIIME2

Trabalho de conclusão de curso apresentado ao Curso de Licenciatura em Ciências Biológicas do Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG) – *Campus* Bambuí para obtenção do grau de Licenciatura em Ciências Biológicas.

Aprovado em 30 de Agosto de 2024 pela banca examinadora:

Gustavo Augusto Lacorte – IFMG – *Campus* Bambuí – (Orientador)
Ludimilla Portela Z. L. Suzuki – IFMG – *Campus* Bambuí
Marcos Roberto Ribeiro – IFMG – *Campus* Bambuí

Dedico este trabalho, primeiramente, a Deus, por guiar meus passos e iluminar meu caminho em cada desafio que enfrentei. À minha família, por seu amor incondicional, apoio constante e por acreditar em mim, mesmo nos momentos mais difíceis. Aos meus amigos, por sua amizade sincera, pelas palavras de incentivo e por estarem ao meu lado durante toda essa jornada. Sem vocês, percorrer essa jornada não seria possível.

"Se você tem medo não o faça, se você o está fazendo não tenha medo! Genghis Khan"

AGRADECIMENTOS

E pensar que já estou me formando! Bem, conheci muitas pessoas nesses últimos anos: fiz muitos amigos; tive bons momentos; outros, nem tanto assim. Aprendi coisas, vivenciei coisas, algumas que nunca imaginei que viveria para vivenciar, mas isso é assunto para outra ocasião.

Hoje venho agradecer a todos que, de alguma maneira, contribuíram para a minha formação.

Em primeiro lugar gostaria de agradecer meus pais, que sempre me apoiaram e me incentivaram a continuar estudando, afinal, como bons pais que são, sempre desejaram o melhor para meus irmãos e para mi. Aos bons amigos, que sempre estiveram comigo, alguns de longa data (da época que nem barba eu tinha) e outros que vieram depois, mas, mesmo assim, tornaram-se bons amigos. Aos meus professores, que sempre se esforçaram ao máximo para nos proporcionar boas aulas e boas experiencias em sala. Para não me delongar muito, agradeço de coração a todos os que fizeram parte dessa minha caminhada, sei que, sem muitos de vocês, a mesma teria sido bem mais árdua e dura do que foi.

Então, meu muito obrigado a todos!

RESUMO

A Bioinformática é uma área interdisciplinar que integra conhecimentos de diversas áreas, como Ciências Biológicas, Ciência da Computação, Estatística, entre outras. Há um debate em torno da necessidade de biólogos tornarem-se proficientes em programação, levando em consideração a complexidade e o tempo necessário para especializar-se nessa área. Embora essa exigência seja questionável, o conhecimento básico na área pode facilitar a comunicação entre biólogos e especialistas em computação, especialmente em situações em que programas prontos não atendem a necessidades específicas. Diante desse cenário, este Trabalho de Conclusão de Curso visou realizar uma avaliação de diferentes estratégias de montagem de ambiente de uso da plataforma Qiime 2, para análises metataxonômicas. Para validar o ambiente e a *pipeline*, foram analisados dados de sequenciamento de *amplicons*, provenientes do Parque Nacional da Serra da Canastra, focando em áreas impactadas e não impactadas pelo fogo, durante o um Manejo do Fogo Integrado. Ao final das analises, foi possível identificar como a ação do fogo mudou as comunidades microbiológicas do solo. A análise comparativa de máquinas para bioinformática revelou que, embora computadores com menor disponibilidade de recursos sejam capazes de executar processos básicos, tarefas mais complexas, especialmente em estudos de microbiomas, exigem máquinas com maior disponibilidade de recursos. Nos testes realizados, a máquina com 64 GB de RAM e 12 vCPUs demonstrou ser a mais adequada para lidar com grandes volumes de dados e processos intensivos. No entanto, para pesquisas com restrições orçamentárias, uma abordagem mais econômica e eficiente pode ser a combinação estratégica de diferentes máquinas, permitindo a otimização de custos sem comprometer a qualidade dos resultados.

Palavras-chave: Bioinformática, QIIME 2, microbioma, microbiota, microbiologia.

ABSTRACT

Bioinformatics is an interdisciplinary field that integrates knowledge from several areas, such as Biological Sciences, Computer Science, Statistics, among others. There is a debate about the need for biologists to become proficient in programming, taking into account the complexity and time required to specialize in this area. Although this requirement is questionable, basic knowledge in the area can facilitate communication between biologists and computer specialists, especially in situations where ready-made programs do not meet specific needs. Given this scenario, this Final Course Work aimed to evaluate different strategies for setting up an environment for using the Qiime 2 platform for metataxonomic analyses. To validate the environment and the pipeline, amplicon sequencing data from the Serra da Canastra National Park were analyzed, focusing on areas impacted and not impacted by fire during the 1st Integrated Fire Management. At the end of the analyses, it was possible to identify how the action of fire changed the soil microbiological communities. The comparative analysis of bioinformatics machines revealed that, although computers with lower resource availability are capable of performing basic processes, more complex tasks, especially in microbiome studies, require machines with higher resource availability. In the tests carried out, the machine with 64 GB of RAM and 12 vCPUs proved to be the most suitable for handling large volumes of data and intensive processes. However, for research with budget constraints, a more economical and efficient approach may be the strategic combination of different machines, allowing cost optimization without compromising the quality of the results.

Keywords: Bioinformatics, QIIME 2, microbiome, microbiota, microbiology.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

IFMG	– Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais
PNSC	– Parque Nacional da Serra da Canastra
QIIME 2	– Quantitative Insights Into Microbial Ecology 2
LAB.BIOMOL	– Laboratório de Biologia Molecular
PCR	– Reação em Cadeia da Polimerase
PB	– Pares de Bases

SUMÁRIO

1 INTRODUÇÃO

A Bioinformática possui uma natureza interdisciplinar, que abrange conhecimentos das mais diversas áreas, como: Ciências Biológicas, Ciência da Com- putação, Estatística, Química, Farmácia, Matemática, entre outras (ATTWOOD *et al.*, 2019; MELO-MINARDI *et al.*, 2013; BAYAT, 2002). Por possuir essa grande interdisciplinaridade e uma vasta gama de aplicações, o uso da Bioinformática tornou-se uma tendência global na atualidade. Existe um debate, no qual se questiona se todos os Biólogos devem ou não se tornar proficientes em linguagens de programação e em desenvolvimento de software. Graças à grande quantidade de tempo demandado para se especializar em Bioinformática e devido ao fato de que parte dos conhecimentos necessários para essa especialização são desenvolvidos em cursos da área de exatas, essa ideia é questionável (FERRARI; GASTALDI, 2018; LEHUGEUR; MELO, 2018).

Entretanto, haverá situações em que não existirão programas prontos para realizar uma análise específica desejada. Mesmo quando tais programas estão disponíveis, pode ser necessário configurá-los adequadamente. Nessas ocasiões, faz-se necessária a interação entre Biólogos e Bioinformatas ou Cientistas da Computação, os quais serão capazes de lidar com o lado computacional da análise. Em situações como essas, o conhecimento básico em Bioinformática faz-se vantajoso, para promover um diálogo eficaz entre o Biólogo e o Bioinformáta/Cientista da Computação. Mesmo que a aprender uma linguagem de programação seja considerado um desafio para muitos, ainda é uma habilidade que tem sido cada vez mais apreciada nas mais diversas áreas (DAMASCENO, 2023; FERRARI; GASTALDI, 2018).

Segundo Abreu *et al.* (2004), o fogo atua como um agente transformador, com grande potencial de alterar o ambiente. Ele pode interferir nas trajetórias sucessionais das comunidades vegetais e impactar a microbiota do solo, além de modificar as propriedades físico-químicas e estruturais do solo.

A análise de dados biológicos é essencial para a compreensão e intendimento dos mais diversos processos biológicos. Entender o funcionamento dos processos biológicos e sua importância é algo fundamental para a preservação ambiental, principalmente tratando-se de ciclos biogeoquímicos, executados por microorganismos (PORTO *et al.*, 2013).

Os microorganismos do solo são responsáveis por funções essenciais, como a decomposição da matéria orgânica, ciclagem de nutrientes, fixação de nitrogênio e solubilização de compostos como o fósforo. Eles também contribuem para a agregação do solo, biodegradação de poluentes, controle biológico de pragas e doenças. A atividade microbiana está diretamente ligada à sustentabilidade dos ecossistemas, destacando a importância de conhecer e preservar os microorganismos sel-

vagens, para a aplicação em tecnologias de recuperação ambiental e tratamento de resíduos (PORTO *et al.*, 2013; NAIR; NGOUAJIO, 2012; GOI; SOUZA, 2006; ALFONSO; LEYVA; HERNÁNDEZ, 2005).

Graças à importância dos processos realizados pelos microorganismos do solo, faz-se necessário o entendimento de como o fogo afeta a microbiota do solo, uma vez que esse agente possui capacidade transformadora suficiente para modificar as trajetórias sucessionais das comunidades de microorganismos, presentes no solo. Diante do exposto, o presente trabalho teve como objetivo desenvolver um ambiente para análise de microbiomas, assim como construir uma *pipeline* customizada para analisar como o fogo afeta a composição da microbiota presente no solo, durante o primeiro manejo integrado do fogo no Parque Nacional da Serra da Canastra.

1.1 Objetivo Geral

O presente trabalho teve como principal objetivo avaliar as diferentes estratégias de montagem de ambiente de uso da plataforma Qiime 2 para análises metataxonômicas.

1.2 Objetivos Específicos

Para atingir o objetivo proposto, foram definidos os seguintes objetivos específicos.

- Definir configurações de máquinas apropriadas para a realização de análises de dados de sequenciamento de *amplicons* utilizando a plataforma QIIME 2 para estudos metataxonômicos.
- Validar as configurações de máquinas analisadas, com o intuito de identificar seu desempenho e possíveis gargalos em alguns processos da *pipeline* utilizada.
- Comparar o custo de processamento de cada máquina observada com o intuito de identificar qual a melhor escolha de máquina para a realização de pesquisas universitárias.
- Analisar um conjunto de dados de sequenciamento de *amplicons* disponíveis no Laboratório de Biologia Molecular do *campus*, provenientes do 1° Manejo do Fogo no Parque Nacional da Serra da Canastra, com o objetivo de identificar como o fogo modifica a microbiota do solo.

2 REFERENCIAL TEÓRICO

Este capítulo apresenta os fundamentos essenciais para uma melhor compreensão do estudo realizado no presente trabalho.

2.1 Bioinformática

O termo Bioinformática foi cunhado na década de 1970, por Ben Hesper e Paulien Hogeweg, uma vez que os pesquisadores acreditavam que era importante distinguir a Bioinformática como um campo de pesquisa. Desse modo, a definição do novo campo do conhecimento seria necessária. Assim, o termo cunhado referia-se ao "estudo dos processos informáticos em sistemas bióticos" (HOGEWEG, 2011).

Antes mesmo do termo Bioinformática ser cunhado por Hesper e Hogeweg, algorítmos, com o intuito de resolver problemas biológicos, já haviam sido desenvolvidos, na década de 1960, quando a Bioinformática ainda estava sob o nome de Biologia Computacional. Exemplares de algoritmos para o alinhamento de sequências, como algoritmos para o estudo da filogenia, foram desenvolvidos nessa época (CORTES, 2019; MELO-MINARDI *et al.*, 2013; OTTO *et al.*, 2007).

Segundo Cortes (2019), foi nos anos 1970, com o surgimento de computadores com maior poder de processamento, que estudos utilizando *software* de simulações se tornaram possíveis, como é o caso do estudo realizado por Michael Levitt e Arieh Warshel, denominado *Computer simulation of protein folding* de 1975. Do mesmo modo, foi durante as décadas de 1970 e 1980, que foram estabelecidos os bancos de dados de sequências de proteínas e ácidos nucleicos (BISHOP *et al.*, 2015).

Tais bancos de dados tornaram-se mais acessíveis aos pesquisadores mundo afora, graças à difusão do acesso à Internet e, consequentemente, aos *websites* e serviços de e-mail. Essas ferramentas possibilitaram uma maior e mais rápida troca de informações entre os pesquisadores (OUZOUNIS; VALENCIA, 2003).

Graças à maior conectividade e troca de informações, houve um crescimento na quantidade de sequências de ácidos desoxirribonucléicos (DNA), provenientes do sequenciamento de genomas inteiros. Outro fator que corroborou para alavancar o desenvolvimento da tecnologia de sequenciamento genômico, assim como o número de sequenciamento genômico completo, de diversas espécies, foi o anúncio do Projeto Genoma Humano (PGH), o qual pode ser concluído com dois anos de antecedência, graças ao avanço das técnicas de sequenciamento (CORTES, 2019; LANDER, 2011; FLÓRIA-SANTOS; NASCIMENTO, 2006).

É sabido que a Bioinformática é uma área do conhecimento essencialmente interdisciplinar, uma vez que ela abrange conhecimentos que pertencem às mais diver-

sas áreas, como Ciências Biológicas, Ciência da Computação, Estatística, Química, Farmácia, Matemática, entre outras (ATTWOOD *et al.*, 2019; MELO-MINARDI *et al.*, 2013; BAYAT, 2002). Segundo Verli *et al.* (2014), as pesquisas em Bioinformática podem ser divididas em duas vertentes: a da Bioinformática Tradicional ou Clássica (pela primazia do nome bioinformática), responsável, principalmente, por estudar e resolver problemas relacionados ao sequenciamento de nucleotídeos e aminoácidos. Já a segunda, é a vertente da Bioinformática Estrutural, responsável por abordar questões biológicas, sob uma ótica tridimensional, abarcando grande parte das técnicas compreendidas pela Química Computacional ou Modelagem Molecular.

Os estudos em Bioinformática também são responsáveis pelos avanços nos métodos de armazenamento e recuperação de dados biológicos, bem como na construção de modelos e algorítmos, que são capazes de identificar sequências de genes, assim como inibidores de enzimas; realizar a predição 3D de proteínas; organizar e relacionar informações biológicas; agrupar proteínas homólogas; estabelecer árvores filogenéticas e analisar experimentos de expressão gênica (OLIVEIRA, 2023), (MAGANA *et al.*, 2017; VERLI *et al.*, 2014; ARAÚJO *et al.*, 2008; CATTLEY; ARTHUR, 2007) .

2.1.1 Linguagem

Segundo Damasceno (2023), a escolha da linguagem de programação para escrever um *software* deve levar em consideração algumas tecnicidades, como seu nível de complexidade (baixo ou alto), facilidade de escrita (sintaxes) e seu nível de portabilidade. Outro fator a ser levado em consideração é o desempenho do programa, determinando se ele atende ou não o desempenho esperado. No entanto, é muito frequente que a escolha da linguagem para construção do *software* seja definida pelos gostos e afinidades dos programadores.

As linguagens de programação mais utilizadas nos estudos genômicos são as linguagens R, Perl e Python. A escolha dessa três linguagens deve-se aos grandes acervos de ferramentas para o manuseio de dados biológicos, que existem e já são amplamente reconhecidas, como é o caso do Bioconductor, BioPerl, BioPython, Blast, Galaxy, CLC Genomics, Qiime 2, entre outros. Esses acervos são aglomerados de módulos e bibliotecas, que são constantemente atualizados por suas respectivas comunidades (COCK; ANTAO *et al.*, 2009; GENTLEMAN *et al.*, 2004; STAJICH *et al.*, 2002).

A facilidade com que cada linguagem utilizada é outro fator a ser observado. Nesse tópico, Python e R saem na frete da linguagem Perl. O que torna a linguagem Perl um pouco mais complexa do que as outras duas em questão é a sua versatilidade. Uma vez que com ela é possível resolver o mesmo problema de diferentes formas, isso

acaba por dificultar o entendimento de um programa feito por outra pessoa. Mesmo assim, a sintaxe dessa linguagem pode ser facilmente entendida por alguém que está começando a aprender a programação (FERRARI; GASTALDI, 2018; STAJICH *et al.*, 2002).

Assim como a linguagem Perl, a linguagem R é considerada por alguns uma linguagem complicada no início, no entanto, ela também permite realizar uma vasta gama de atividades sem muitas dificuldades. Em contrapartida, a linguagem Python busca formas mais diretas de resolver os problemas propostos, a sintaxe é mais simples e há uma vasta gama de recursos para quem esta iniciando na progra- mação. Atualmente, R e Python são as linguagens mais recorrentes quando o assunto é Bioinformática, haja visto o seu grande potencial em *Machine Learning e Deep Learning* (FERRARI; GASTALDI, 2018; MCKINNEY, 2018; GENTLEMAN *et al.*, 2004).

2.1.2 Ambiente

As principais ferramentas dos bioinformatas são os computadores, comumente chamado de máquinas, os quais são responsáveis por realizar o trabalho de processamento bruto, juntamente às mais diversas análises (FERRARI; GASTALDI, 2018).

Quanto ao Sistema Operacional (SO) a ser utilizado pelos bioinformatas, há uma certa preferência aos Sistemas da Distribuição Linux. Isso ocorre devido a alguns fatores que favorecem essa escolha. O primeiro fator é que a Distribuição Linux é um *software* livre, o que permite os usuários criarem uma distribuição totalmente personalizada ao uso desejável. Isso não seria diferente para a Bioinformática, uma vez que uma distribuição específica para análises de dados biológicos foi implementada e distribuída para a comunidade denominada Bio-Linux (VAUGHAN-NICHOLS, 2017; FILGUEIRAS *et al.*, 2016).

Outro fator que favorece a utilização das Distribuições Linux são os melhores resultados no gerenciamento de recursos. Esse fator não se limita só ao contexto da Bioinformática, ele reflete em boa parte da área da computação, já que, desde de novembro de 2017, todos os 500 mais potentes supercomputadores do mundo utilizam o Linux como sistema operacional. No entanto, a utilização de Linux não é uma restrição para a análise de dados e os sistemas Windows e MacOS também podem ser utilizados para esse fim (VAUGHAN-NICHOLS, 2017; HIRATA, 2011).

É importante ressaltar que as configurações de máquina, necessárias para a realização de análises de dados, podem variar com base nas particularidades de cada projeto, assim como na escala dos dados a serem analisados. No entanto, é essencial ter em mente que uma configuração mínima é necessária para que a análise possa ser realizada. Também é recomendável ter como opção a possibilidade de

aumentar a capacidade de processamento, memória e armazenamento, conforme a complexidade e demanda dos projetos (ZACARIA, 2018).

2.1.3 Computação em Nuvem (Cloud Computing)

Segundo Rodrigues (2019), computação em nuvem, do termo em inglês Cloud Computing, é o nome utilizado para referir-se a serviços computacionais contratados e gerenciados por meio da *Internet*.

Esses serviços em nuvem surgiram devido à subutilização dos recursos computacionais em empresas, especialmente as de Tecnologia da Informação (TI). Combinando a capacidade computacional e a experiência em gerenciamento de TI, essas empresas criaram um novo mercado computacional, que oferece serviços, como aluguel de infraestruturas, armazenamento e backup de arquivos pessoais, além de plataformas para o desenvolvimento e implantação de aplicações (RODRIGUES, 2019; BUYYA; BROBERG; GOSCINSKI, 2011).

Ainda segundo Rodrigues (2019), sistemas de computação em nuvem ou servidores virtuais oferecem uma ampla gama de serviços, para atender às mais diversas necessidades. Para usuários comuns, estão disponíveis serviços que abrangem desde soluções de software até armazenamento e backup de arquivos pessoais. Por outro lado, no âmbito empresarial, a computação em nuvem disponibiliza plataformas destinadas ao desenvolvimento e implementação de aplicações, além de servidores virtuais, que podem ser configurados de forma flexível para atender a demandas específicas.

Para Nhacutouo (2020) e Ogunyemi e Johnston (2017), a virtualização de servidores proporciona benefícios, que giram em torno da redução do custo total de propriedade, melhor aproveitamento do hardware, redução do espaço físico refrigerado necessário, redução no consumo de energia, flexibilidade na criação de novas máquinas virtuais, assim como o fácil redimensionamento dos recursos necessários para as máquinas já existentes.

2.2 Plataformas de Análise de Dados Biológicos

Essa sessão será responsável por apresentar algumas plataformas para análise de dados biológicos, disponíveis no mercado.

2.2.1 Galaxy

Segundo Neto e Cintra (2016) e Aguiar (2011), o Galaxy é um *framework* voltado para *Web* de código aberto, desenvolvida em linguagem de programação

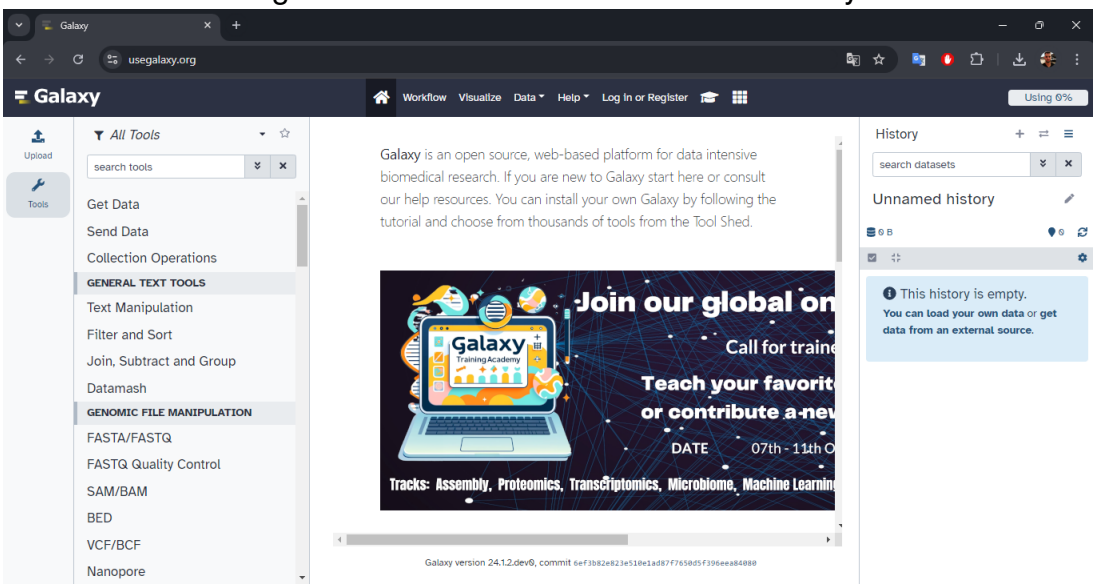
Python, muito utilizada na análise dos mais diversos tipos de dados biológicos. Essa ferramenta pode ser utilizada, através de um servidor publico (https://usegalaxy.org) ou pode ser utilizada em uma instância local (https://galaxyproject.org). O Galaxy possui como propósito facilitar, por meio de um ambiente *Web*, a análise de grandes volumes de dados, assim como facilitar a reprodutividade, transparência e compartilhamento de dados.

Graças à uma próspera e ativa comunidade, que continua a usar e contribuir com a manutenção e crescimento do projeto, é possível oferecer serviços de análises e treinamento virtual aos interessados, de forma gratuita. A Galaxy Training Network oferece suporte e treinamento virtual autodirigido e gratuito, contando com mais de 230 tutoriais integrados (COMMUNITY, 2022).

A plataforma Galaxy proporciona, aos usuários, acesso a soluções poderosas e práticas, quando o assunto é análise de dados, fornecendo acesso a hardware, ferramentas e dados abrangentes, que podem ser utilizados com mínimo de treinamento possível. Graças às mais diversas pesquisas, mais de 8.000 pacotes de *software* populares foram integrados ao Galaxy. A diversidade crescente de ferramentas disponíveis no Galaxy permite os mais diversos tipos de análise e como todas as manipulações de dados são realizadas por meio de ferramentas, a reprodutibilidade é garantida (COMMUNITY, 2022).

A figura 1, demonstra a tela inicial da plataforma de análise Galaxy, disponível em (https://usegalaxy.org/).

Figura 1 – Tela inicial da Plataforma Galaxy



Fonte: (COMMUNITY, 2022)

2.2.2 CLC Genomics Workbench

De acordo com Liu e Di (2020), o CLC Genomics Workbench é um *software* de bioinformática, que oferece uma gama de funcionalidades essenciais para análise de dados genômicos. Este *softwar*e possibilita a realização de análises detalhadas de ressequenciamento direcionado, assim como a montagem completa de genomas e transcriptomas. No entanto, para utilizar esse software é necessário comprar uma licença para utilizá-lo.

Sua arquitetura é baseada na linguagem de programação Java. Esse fato permitiu a CLC bio, em 2006, proporcionar aos conhecedores da linguagem Java a possibilidade de desenvolver seus próprios *plugins*, os quais são integrados ao CLC Workbenches e ao Viewer e fornecem uma maneira fácil de personalizar e ampliar suas funcionalidades CLC (2023) e Qiagen (2023).

Diferentemente do Galaxy, o qual já fornece acesso a servidores, que já possuem as ferramentas de análise de dados instaladas e com recursos como memória, armazenamento e poder de processamento, os quais não são informados aos usuários, o CLC Genomics necessita de ser instalado em maquinas físicas ou servidores. Por esse motivo, a empresa responsável pelo *software* sugere a quantidade de recursos mínimos para garantir o funcionamento correto da ferramenta (QIAGEN, 2023).

Sendo assim, a quantidade miníma de memória RAM indicada, para cada análise realizada pelo *software*, pode variar entre 8 GB e 64 GB de memória RAM. Isso será definido mediante o tamanho do banco de dados de referência utilizado, juntamente ao tipo de análise a ser realizada. Por exemplo, a execução de uma análise, cujo o banco de dados possua 14 GB de informações, requer, pelo menos, 16 GB de RAM. Fato interessante a respeito do *software* CLC é que, ao se criar um banco de dados de referência com a ferramenta *Download Microbial Reference Database tool*, uma mensagem, contendo os requisitos de memória necessários para executar o *Taxonomic Profiler* com este banco de dados, é apresentada ao usuário (QIAGEN, 2023).

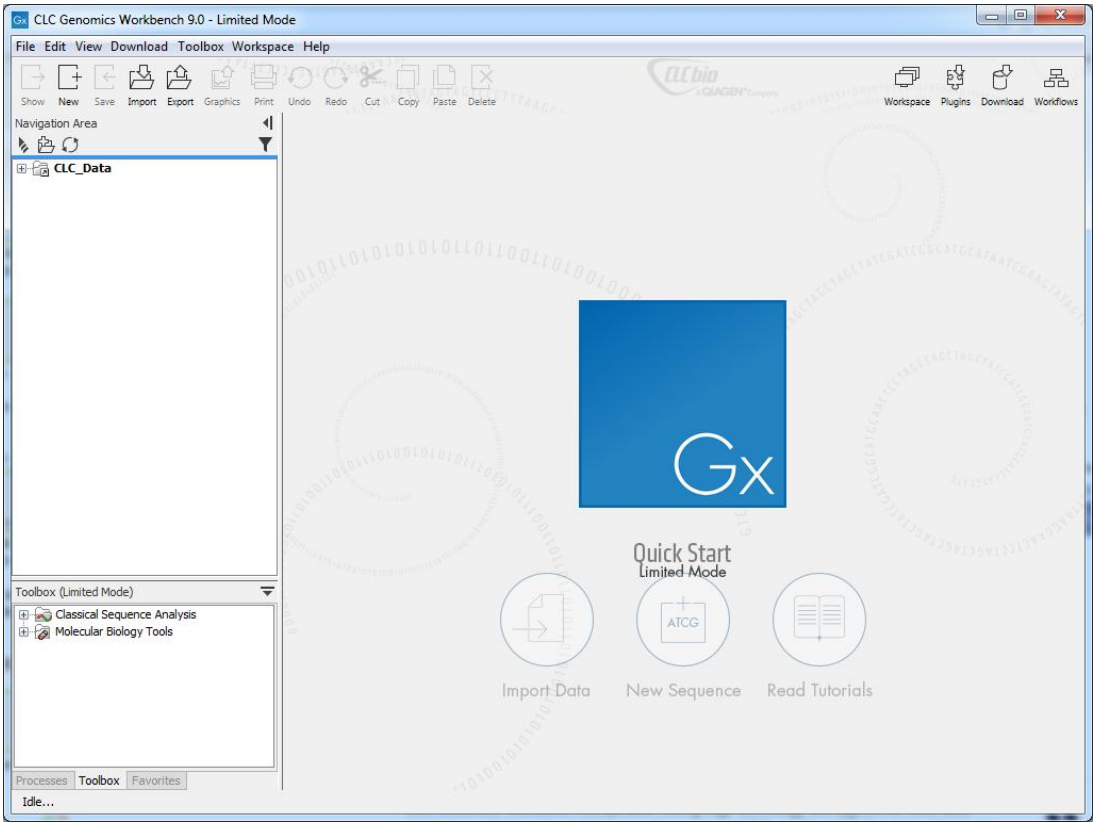
Ainda de acordo com Qiagen (2023),aumentar o número de CPUs pode diminuir o tempo necessário para realizar o mapeamento de leituras, contudo acredita-se que o ganho de desempenho seja limitado, acima de aproximadamente 40 *threads*. Em relação à quantidade de armazenamento, assim como a memória RAM, é variável, tendo que se levar em consideração o tamanho dos bancos de dados referenciais, o tamanho dos dados analisados e a quantidade de análises realizadas.

Graças a sua arquitetura o *software* é compatível com diversos Sistemas Operacionais: Windows, MacOS e Linux (QIAGEN, 2023).

A figura 2, demonstra como é a tela inicial da Plataforma CLC Genomics

Workbench.

Figura 2 – Tela inicial da Plataforma CLC Genomics Workbench



Fonte: (QIAGEN, 2023)

2.2.3 QIIME 2

O QIIME 2 é um pacote de *software* de análise de microbioma de código aberto, que converte dados brutos de sequenciamento em visualizações interpretáveis e resultados estatísticos (FUNG *et al.*, 2021).

Segundo Estaki *et al.* (2020) e Bolyen *et al.* (2019), a ferramenta QIIME 2 é a reformulação da plataforma QIIME 1, que foi responsável por apoiar diversos estudos de microbiomas e ganhou uma ampla comunidade de usuários e desenvolvedores. Sua arquitetura é baseada em um sistema de *plugins*, desenvolvidos com o auxílio da linguagem Python.

Os *plugins* são componentes ou módulos de software, que podem ser adicionados a um programa ou sistema existente para estender suas funcionalidades ou adicionar recursos específicos (OLIVEIRA, L. N.; LUIZ SCHIAVONI, 2018).

O QIIME 2 oferece novas ferramentas de visualização interativa, que simplificam a realização de análises exploratórias e a criação de relatórios de resultados. Um exemplo de ferramenta para a visualização de análises, com a qual o QIIME 2 conta é o QIIME 2 View, uma ferramenta de fácil manuseio, que pode ser acessada

de forma remota, através da *web*. Essa ferramenta de visualização é um novo serviço exclusivo, que permite que os usuários possam compartilhar e interagir de forma segura com os resultados provenientes do QIIME 2, sem a necessidade de instalar a ferramenta (BOLYEN *et al.*, 2019).

Ainda segundo Bolyen *et al.* (2019), os princípios que nortearam o design do QIIME 2, foram os princípios de reprodutibilidade, a transparência e a clareza da ciência de dados do microbioma. Com base nesses princípios, o QIIME 2 possui um sistema descentralizado de rastreamento de proveniência de dados, o que permite realizar uma retrospectiva de qualquer resultado que foi gerado com o auxilio da ferramenta. Isso é possível, uma vez que todos os detalhes das etapas da análise, incluindo referências a dados intermediários, são automaticamente registrados nos resultados. Outra funcionalidade importante do QIIME 2 é a detecção de resultados corrompidos, indicando a falta de confiabilidade dos mesmos.

Vale ressaltar que, com exceção de metadados, todos os arquivos gerados pelo QIIME 2, existem como artefatos QIIME 2 e usam a extensão de arquivo .QZA (QIIME Zipped Artifact), para arquivos manipuláveis, e .QZV, para arquivos de visualização. Artefatos são geralmente arquivos compactados em formato ZIP, que possuem, em sua composição, dados em formato FASTA ou FASTQ (ESTAKI *et al.*, 2020).

Arquivos em formato FASTA, são aqueles constituídos por textos simples, em que cada sequência é representada por uma única linha descritiva, que se inicia com o caractere “>” seguido por uma ou mais linhas de sequência. É importante destacar que arquivos no formato FASTA não possuem índices de qualidade referente às suas sequências. No entanto, arquivos no formato FASTQ, que também são arquivos simples de texto, como os arquivos FASTA, têm a capacidade de armazenar um índice de qualidade numérico associado a cada nucleotídeo, em uma sequência (COCK; FIELDS *et al.*, 2009).

Para reduzir a ocorrência de erros, os resultados gerados pelo QIIME 2 são estruturados semanticamente, com diretrizes claras sobre os tipos de entrada aceitáveis. Dessa forma, é possível garantir a integridade e a confiabilidade dos dados, promovendo uma análise mais precisa e confiável (BOLYEN *et al.*, 2019).

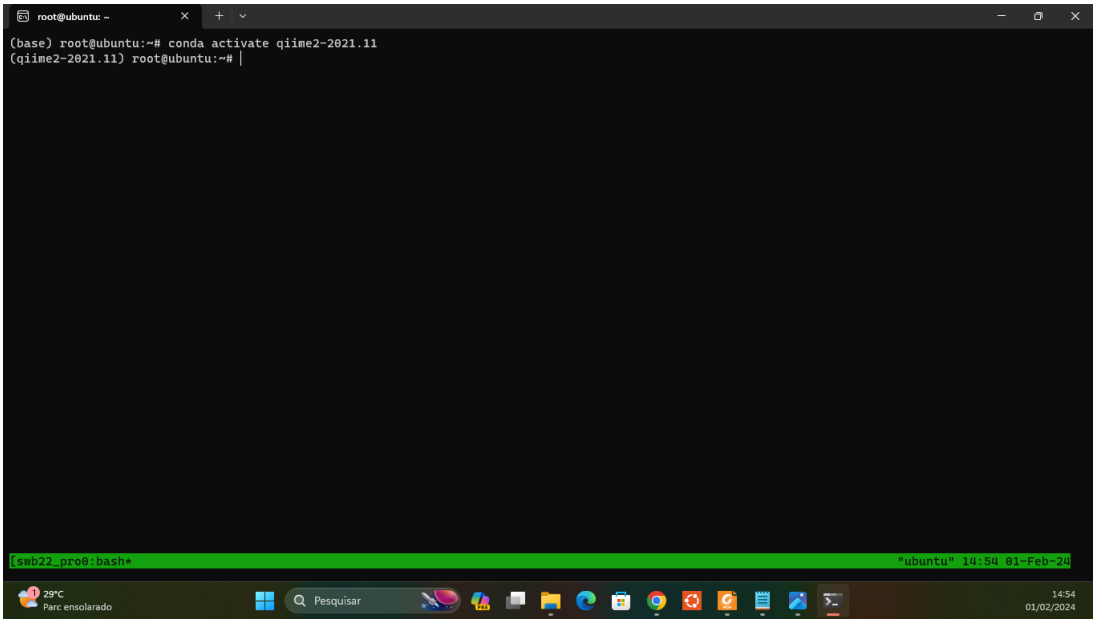
Atualmente, o QIIME 2 pode ser manuseado de duas formas. A primeira forma do QIIME 2 é voltada para um público que não é familiarizado com terminais de linhas de comando e que prefere ferramentas que contam com interfaces gráficas de fácil manuseio. Para esse público, o QIIME 2 conta com sua versão gráfica, denominada QIIME 2 Studio, uma interface gráfica projetada para biólogos, médicos e formuladores de políticas.

Em contrapartida, para aqueles com conhecimentos mais avançados em programação e análise de dados, o QIIME 2 oferece várias opções de interface. Isso

inclui a interface de programação de aplicativos QIIME 2, ideal para cientistas de dados, que desejam automatizar fluxos de trabalho ou trabalhar interativamente no Jupyter Notebooks. Além disso, o QIIME 2 disponibiliza o q2cli e o q2cwl, que fornecem uma interface de linha de comando e *wrappers* CWL, para atender às necessidades de especialistas em computação de alto desempenho (BOLYEN *et al.*, 2019).

A figura 3, demonstra a tela de linha de comando do QIIME 2 referente a sua versão q2cli.

Figura 3 – Painel de linha de comando referente a versão q2cli do QIIME 2



Fonte: Gerado pelo autor, 2024

2.2.4 Possibilidades de análises

A plataforma QIIME 2 permite que dados de microbiomas sejam todos tratados em um mesmo lugar, realizando desde os primeiros processos da análise, como a importação de dados brutos de sequenciamento, até análises de diversidade e taxonomia (SILVA, L. R. G. d., 2022).

Atualmente, a ferramenta conta com *plugins* específicos, voltados para diferentes objetivos de pesquisa. Dentre as análises que podem ser realizadas pelo QIIME 2, podemos citar a análise de metagenômica, a análise de diversidade, a análise de composição taxonômica, a análise de abundância, a análise de alfa e beta diversidade, a criação de árvores filogenéticas, a classificação taxonômica, entre vários outros estudos, relacionados aos estudos gênicos (POOR, 2022; RIBEIRO, J. C. M., 2022; SILVA, L. R. G. d., 2022), (FERRERO, 2021; VAZ, 2021; BOKULICH *et al.*, 2018).

2.2.5 Parâmetros de Máquina e Modo de Uso

Nessa seção, serão apresentados conceitos referentes aos parâmetros necessários para a execução do software e seu modo de uso.

2.2.5.1 Parâmetros de Máquina - *Hardware*

Devido à sua alta compatibilidade, o QIIME 2 pode ser instalado em Sistemas Windowns, linux e MacOS. Atualmente, para a sua instalação, são recomendadas máquinas que possuam, no mínimo, 7 GB de armazenamento livre em disco e, pelo menos, 4 GB de memória RAM, como ponto de partida para a análise de pequenos conjuntos de dados.

A quantidade de memória RAM disponível, assim como o poder de processamento e espaço livre em disco, podem variar de acordo com o projeto e análises desejadas (ESTAKI *et al.*, 2020).

2.2.5.2 *Pipeline*

Uma vez tendo acesso a um ambiente devidamente configurado para análises, utilizando o QIIME, há uma série de etapas de transformações de arquivos a serem seguidas. Essa sequência de etapas são caracterizadas por gerarem saídas, que alimentam as entradas das transformações seguintes. Esse conjunto de processos é denominado de *pipelines* (TEIXEIRA; LEAL, s.d.).

Segundo Moreira (2019), uma *pipeline* para análise de dados de microbiomas pode ser resumida em três etapas principais, sendo elas: o pré-processamento dos dados, a limpeza e remoção de ruídos e, por último, a análise dos dados.

Vale ressaltar que, antes mesmo do pré-processamento dos dados, é necessário realizar a preparação das pastas e arquivos desejados, nomeando-os de acordo com convenção de nomenclatura Illumina (regra de nomenclatura para os arquivos de entrada utilizados no QIIME 2). Esse processo faz-se necessário, devido às estruturas semânticas de entrada, permitidas pela ferramenta (BOLYEN *et al.*, 2019).

Em seu trabalho, Moreira (2019) fala que o pré-processamento dos dados é composto por duas etapas fundamentais, sendo a primeira delas o agrupamento dos dados por amostra de origem, utilizando os *barcodes* (códigos de barras) como referência. A segunda etapa é o processo de de multiplexagem, que consiste na remoção dos *barcodes*, adicionados durante a Reação em Cadeia da Polimerase (PCR).

É importante salientar que existem formatos, nos quais as bibliotecas de sequenciamento não contam com arquivos de *barcodes*, como é o caso dos dados em formato *Casava 1.8 paired-end demultiplexed fastq*. Nesse formato, existem dois

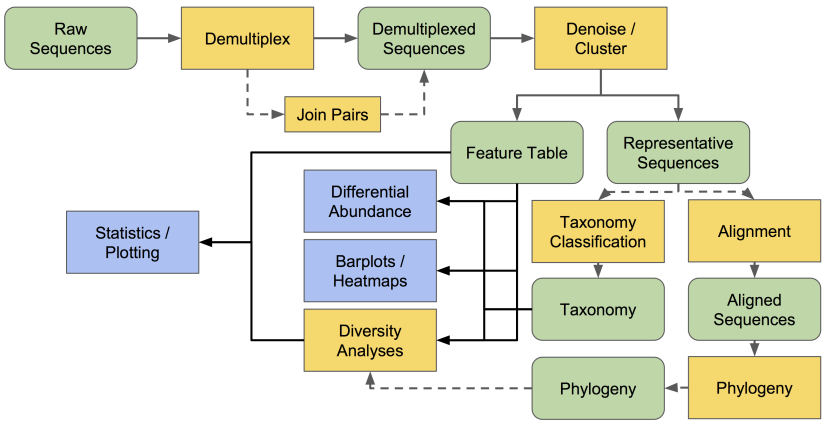
arquivos de dados referentes a uma mesma amostra, um arquivo contendo as leituras diretas e outro, contendo as sequências reversas. Dados nesse formato, por não possuirem marcadores moleculares, dispensam a etapa de retirada dos mesmo, exigindo na etapa de pré-processamento apenas o emparelhamento das leituras (BOLYEN *et al.*, 2019).

A etapa de limpeza e eliminação de ruídos consiste em remover *reads*, contendo nucleotídeos ambíguos, homopolímeros; *reads*, contendo menos de 10 pares de bases (bp) e *reads* com baixa ocorrência, sequências de ocorrência única (*singletons*), sequências quiméricas, causadas por erro na PCR, e sequências idênticas (deduplicação, mantendo apenas uma cópia de cada sequência) (THOMPSON *et al.*, 2022; MOREIRA, 2019).

A análise dos dados consiste no agrupamento taxonômico, com identidade mínima de 97% em OTUs e identificação dos OTUs, através da comparação com banco de dados consolidados (JULI, 2017). É importante ressaltar que as *pipelines* não são uma sequência estática de processos a serem seguidos, elas podem ser criadas e personalizadas, de acordo com o objetivo do pesquisador (ESTAKI *et al.*, 2020).

A Figura 4 demonstra um exemplo de *pipeline* para a análise de DNA de microorganismos.

Figura 4 – Exemplo de uma *Pipeline* para analise de dados biológicos



Fonte: (BOLYEN *et al.*, 2019)

2.2.5.3 *Pipeline* - Importação, Filtragem e Criação da árvore filogenética

É notório dizer que, para se dar início à *pipeline*, primeiramente, é necessário obter as sequências desejadas para a realização da análise. Essas sequências podem ser provenientes de experimentos do próprio pesquisador ou cedidas por terceiros.

Após a obtenção das sequências e o preparo das mesmas, é necessária a importação desses dados para um artefato do QIIME, para isso, é utilizado o comando *qiime tools import*, juntamente com o parâmetro *type*, acompanhado pelo tipo semântico do artefato desejado. Outros parâmetros ainda são esperados pela ferramenta, como o parâmetro de entrada (*input*), o qual é responsável por identificar quais serão os arquivos a serem importados, e o parâmetro de saída (*output*), que será responsável por nomear o arquivo de saída, com o nome escolhido, e direcioná-lo ao diretório desejado (BOLYEN *et al.*, 2019).

Em seguida, para importação dos dados para um artefato QIIME, é necessário realizar o controle de qualidade dos mesmos. Nessa etapa da *pipeline* o controle de qualidade dos dados podem ser realizados por vários plugins dentro da ferramenta. Alguns exemplos são: DADA2, *Deblur* e *basic quality-score-based filtering*. Esses *plugins* são responsáveis por gerarem o mesmo resultado dentro da *pipeline*, portanto podem substituir um ao outro sem problemas (BOLYEN *et al.*, 2019).

Ambos os métodos resultaram em dois artefatos QIIME 2, um *FeatureTable[Frequency]*, que contém a frequência de cada sequência, única em cada amostra do conjunto de dados, e um *FeatureData[Sequence]*, que mapeia os identificadores de recursos nas sequências representativas, contidas na *FeatureTable* (BOLYEN *et al.*, 2019).

Ao concluir a etapa de filtragem, para explorar os dados resultantes, é necessário criar arquivos de visualização, os arquivos .QZV. Os arquivos .QZV gerados conterão informações sobre a quantidade de sequências, que estão associadas a cada amostra e a cada recurso, além de histogramas referentes à distribuição das sequências analisadas e algumas estatísticas de resumo relacionadas. Além disso, o comando fornecerá um mapeamento de IDs de recursos para sequências e fornecerá links para facilitar as buscas BLAST de cada sequência, no banco de dados NCBI nt (BOLYEN *et al.*, 2019).

Com esses dados em mãos, torna-se possível a criação de uma árvore filogenética, que é responsável por estabelecer as relações entre os dados de contagem de características encontrados na *FeatureTable[Frequency]*. Para estabelecer as relações entre essas características, é essencial criar uma árvore filogenética enraizada. Esse processo é realizado com a ajuda do *plugin* q2-phylogeny, usando o comando *align-to-tree-mafft-fasttree* Bolyen *et al.* (2019).

Ainda segundo Bolyen *et al.* (2019), nesse ponto da *pipeline*, é utilizado o programa *mafft*, que realiza o alinhamento múltiplo das sequências do *FeatureData[Sequence]* criando um artefato *FeatureData[AlignedSequence]*. Em seguida, o software executa uma etapa de filtragem ou máscara para eliminar posições que exibem alta variabilidade. Após os dados passarem pela máscara, o *FastTree*, gera-se uma árvore filogenética, com base no alinhamento que foi previamente mascarado. É

importante destacar que o *FastTree* gera inicialmente uma árvore sem raiz. Portanto, na etapa final deste processo, é realizada a técnica de enraizamento no ponto médio, que posiciona a raiz da árvore no ponto médio da maior distância entre as extremidades da árvore sem raiz. Esse último passo é essencial para garantir que a árvore filogenética esteja devidamente enraizada e pronta para análises subsequentes, principalmente na avaliação da diversidade filogenética.

Com a árvore gerada, torna-se possível a realização das análises de alfa e beta diversidade dos dados estudados. Essas análises podem ser realizadas dentro do QIIME, através do *plugin q2-diversity*, que é responsável por realizar o cálculo de métricas de diversidade alfa e beta, aplicando testes estatísticos relacionados e gerando visualizações interativas (BOLYEN *et al.*, 2019).

2.2.5.4 *Pipeline* - beta e alfa diversidade

Para o início das análises de beta diversidade, é utilizado o método *core-metrics-phylogenetic*, acompanhado pelo parâmetro *–p-sampling-depth*,que é responsável por setar a profundidade de amostragem uniforme (rarefação), bem como outros parâmetros de entrada, que indicam quais arquivos serão utilizados nessa análise e o parâmetro de saída. O método *core-metrics-phylogenetic* rarefaz a *FeatureTable[Frequency]* a uma profundidade pré-estabelecida pelo usuário, realiza o cálculo de várias métricas de diversidade alfa e beta e gera gráficos de análise de coordenadas principais (PCoA), usando o EMPeror para cada uma das métricas de diversidade beta (BOLYEN *et al.*, 2019; VÁZQUEZ-BAEZA *et al.*, 2013).

Dentro da análise de diversidade, é possível calcular a composição microbiana da amostra, em alfa e beta diversidade, em que, para análise de composição alfa, se utiliza o comando *qiime diversity alpha-group-significance*, acompanhado dos parâmetros de entrada e saída duas vezes. Na primeira vez, é setado o parâmetro de entrada, utilizando o arquivo: *--i-alpha-diversity core-metrics-results/faith_pd_vector.qza"* e a segunda *--i-alpha-diversity core-metrics-results/evenness_vector.qza"*, (BOLYEN *et al.*, 2019).

Já a análise de composição beta, utiliza o comando *qiime diversity beta-group-significance*, acompanhado dos parâmetros de entrada e saída duas vezes, assim como na alfa. O que difere a análise beta da alfa são suas entradas, uma vez que se utilizam o arquivo *"unweighted_unifrac_distance_matrix.qza"* em suas duas entradas, mas variam o parâmetro (–m-metadata-column) de acordo com as colunas do metadata utilizado (BOLYEN *et al.*, 2019).

Para a exploração da diversidade alfa, utiliza-se o comando *qiime diversity alpha-rarefaction*, acompanhado dos parâmetros de entrada e saída. Esse comando permite calcular várias métricas de diversidade alfa em diferentes profundidades de

amostragem. Para cada profundidade de amostragem, são criadas 10 tabelas rarefeitas e, em seguida, as métricas de diversidade são calculadas para todas as amostras contidas nessas tabelas. Como resultados, os valores médios da diversidade são plotados para cada amostra, em cada profundidade de amostragem. Além disso, é possível agrupar as amostras com base em metadados fornecidos, se estes estiverem disponíveis através do parâmetro correspondente, na visualização resultante (BOLYEN *et al.*, 2019).

Como resultado desse processo, será gerado um arquivo de visualização, contendo dois gráficos. O gráfico superior é um gráfico de rarefação alfa e é usado, principalmente, para determinar se a riqueza das amostras foi totalmente observada ou sequenciada. Já o gráfico inferior, nesta visualização, desempenha um papel crucial ao agrupar as amostras com base em metadados. Ele mostra o número de amostras, que permanecem em cada grupo, quando a tabela de características é submetida à rarefação em diferentes profundidades de amostragem (BOLYEN *et al.*, 2019).

2.2.5.5 *Pipeline* - Atribuição Taxonômica

Outra opção de análise existente no QIIME é a análise taxonômica, que consiste em um processo de atribuição taxonômica para as sequências, contidas no artefato *FeatureData[Sequence]*, através de comparações com banco de dados de referência. Para realizar essa atribuição, são necessários um classificador taxonômico pré-treinado, que já está embutido no comando *plugin q2-feature-classifier*, e um banco de dados conceituado, como referência (BOLYEN *et al.*, 2019).

Classificadores pré-treinados são softwares que foram treinados sob a ótica de aprendizado de máquina, utilizando uma grande base de dados. Esses classificadores foram condicionados a reconhecer padrões, com base na vasta gama de dados de referência, e sua acurácia de reconhecimento apresentou resultados aceitáveis. Desse modo, são considerados apropriados para a classificação de novos dados (RAMOS *et al.*, 2018).

Como resultado desse processo, é gerado um arquivo de visualização, no formato de tabela, no qual são apresentadas três colunas: a primeira contém o ID, correspondente à amostra; a segunda contém informações taxonômicas relacionadas a ela e a terceira contém o percentual de confiabilidade da taxonomia, variando de 0 a 1 (BOLYEN *et al.*, 2019).

Ainda dentro da classificação taxonômica, é possível criar gráficos interativos, referentes às amostras, utilizando o comando *qiime taxa barplot*, juntamente aos parâmetros de entrada e saída desejados (BOLYEN *et al.*, 2019).

2.3 Estudos de microbiomas

Nessa seção, serão abordados conceitos importantes a respeito do que são microbiomas.

2.3.1 Microbiomas: Conceitos

O termo “microbioma” foi citado, pela primeira vez, por Joshua Lederberg e Alexa T. McCray, em seu trabalho, no qual definiram microbioma como “a comunidade ecológica de microrganismos comensalistas, simbiontes ou patogênicos, que literalmente ocupam o espaço de nosso corpo”. Essa definição, inicialmente, referia-se ao microbioma humano. No entanto, para abranger os microrganismos em qualquer ambiente, o termo refere-se ao conjunto de microrganismos que habitam um determinado hospedeiro ou que coexistem em um ambiente específico (CARDOSO; ANDREOTE, 2016; BOON *et al.*, 2014; ANDREOTE, 2014).

Segundo o Ribeiro (s.d.), diversidade significa característica ou estado do que é diverso, diferente, diversificado. É notório dizer que exista uma diversidade muito grande, referente aos seres vivos do planeta Terra. Para se referir a essa diversidade, foram cunhados dois termos, os quais são intercambiáveis entre si, mas que se referem a essa grande variedade de seres vivos, “Biodiversidade” e “Diversidade Biológica”.

Estes dois termos vieram com o intuito de abranger temas fundamentais da ecologia e da biologia evolutiva, referentes à diversidade de espécies e os ambientes que os sustentam (FRANCO, 2013). A biodiversidade ou diversidade biológica corresponde ao somatório de toda a variação biótica existente, partindo desde genes a ecossistemas (ROCHA; VALE, 2017; PURVIS; HECTOR, 2000).

2.3.2 Análises de microbiomas

As análises de microbiomas referem-se ao estudo dos microrganismos presentes em um determinado ambiente. Essas análises são cruciais para compreender a diversidade, a função e a dinâmica dessas comunidades microbianas, em relação ao ambiente em que vivem (SALGADO, 2021; KLEIN, 2021; RIBEIRO *et al.*, 2014).

Segundo Galloway-Peña e Hanson (2020), ao se estudar microbiomas, as diferenças entre amostras são frequentemente avaliadas, usando métricas de diversidade alfa e beta.

A diversidade alfa é responsável por calcular a abundância de indivíduos e a riqueza de espécies, em um determinado local ou comunidade. Para o cálculo dessa diversidade, existem métricas amplamente utilizadas pela comunidade, como os índi-

ces de Shannon-Wiener, Simpson, estimador de abundância Chao1, diversidade filogenética de Faith, além de várias outras métricas (BENÍCIO *et al.*, 2023; CARVALHO, F. A.; FELFILI, 2007).

Diversidade beta é um cálculo realizado sobre a variação na composição das espécies, entre diferentes unidades amostrais. Uma abordagem convencional para calcular a diversidade beta é por meio da dissimilaridade de Bray-Curtis, uma medida quantitativa, que leva em consideração a abundância de táxons ao comparar duas comunidades. Além de considerar as abundâncias de táxons, a distância Unifrac ponderada também incorpora o parentesco filogenético ao avaliar as diferenças entre duas comunidades. Por outro lado, a distância Unifrac não ponderada é uma medida qualitativa, que leva em conta apenas a presença e ausência de táxons. Ambas as medições do Unifrac exigem o uso de uma árvore filogenética, já que as pontuações são derivadas do cálculo das distâncias totais dos ramos entre as bactérias compartilhadas e não compartilhadas, em uma árvore filogenética (GALLOWAY-PEÑA; HANSON, 2020).

Existem vários outros métodos para o cálculos de diversidade beta, como, por exemplo, pelos índices de diversidade Whittaker, Sørensen, Jaccard. O índice de Jaccard, também conhecido como coeficiente de similaridade, é uma medida qualitativa, que não leva em consideração as abundâncias relativas, focando apenas na presença ou ausência de elementos. Ao avaliar a presença, ausência e abundância dos táxons, é possível examinar a extensão das diferenças nas composições das comunidades entre as amostras. Quando possível, a inclusão do parentesco filogenético permite uma avaliação da divergência evolutiva (KNIGHT *et al.*, 2018). Esses índices fornecem medidas quantitativas da diversidade beta e podem ser utilizados para a comparação de diferentes habitats e locais, em relação a suas espécies (MOURA, 2020; MAGURRAN, 1988).

2.3.3 Correlação com metadados

Segundo Madigan, Martinko e Bender (2016), os microorganismos estão presentes em todos os lugares que ofereçam suporte à vida, em nosso planeta. Isso engloba desde de habitats familiares a nós, como solo, água e ar, assim como ambientes extremos para a sobrevivência de macroorganismos. Exemplos desses ambientes extremos são os vulcões, lagos congelados ou águas extremamente salinas.

Além de ser conhecido como o substrato responsável pela fixação e desenvolvimento das plantas, o solo também deve ser considerado um “ente vivo”, pois, em sua composição, estão presentes inúmeros animais e microorganismos (PARRON *et al.*, 2015). Segundo Bald *et al.* (2021), devido à sua heterogeneidade constituinte, o solo apresenta maior diversidade bacteriana que a água e o ar. Os microorganismos

do solo, comumente denominados de microbiota, são representados por cinco grandes grupos, possuindo diversidade e quantidade elevadas: bactérias, actinomicetos, fungos, cianobactérias e protozoários.

Ainda segundo Bald *et al.* (2021), os microorganismos do solo correspondem a uma porcentagem entre 1 e 4% do carbono total contido no solo e ocupam cerca de 5% do espaço poroso desse substrato. No entanto, graças às suas diversas características, o solo é um ambiente estressante, cuja principal causa é a limitação por recursos, desse modo, o número de bactérias e fungos, em estado ativo, fica abaixo dos 30%.

Em relação à diversidade genética, o solo destaca-se como um ambiente com inúmeras e variadas populações de microrganismos, tornando-se o reservatório final para a maioria deles. Além disso, o solo abriga bilhões de organismos, os quais possuem funções e nichos ecológicos específicos. Nesse intricado ecossistema, cada organismo contribui para uma variedade de processos, que ocorrem no solo, promovendo uma interconexão vital para a saúde e a fertilidade desse ambiente (MATTOS, 2015).

Em seu trabalho, Cotta (2016) fala que a diversidade microbiana está intimamente ligada a uma série de fatores, tanto abióticos (como atmosfera, temperatura, água, pH, potencial redox, fontes nutricionais, entre outros) quanto bióticos (incluindo genética microbiana e interações entre microrganismos), que desempenham papéis essenciais no desenvolvimento microbiano e na organização da comunidade viva presente nos solos. A interação entre esses fatores exerce uma influência direta sobre a ecologia, a atividade e a dinâmica populacional dos microrganismos presentes no solo (KUBIAK, 2017).

2.4 Tipos de Sequenciamento

O sequenciamento do ácido desoxirribonucléico (DNA) envolve uma série de procedimentos bioquímicos, que visam determinar a sequência das bases nitrogenadas que compõem o código genético de cada ser vivo. As bases nitrogenadas constituintes do DNA são: adenina (A), guanina (G), citosina (C) e timina (T) (SANTOS *et al.*, 2013; WATSON; CRICK, 1953)

Conforme destacado por Silva, Lima e Souza (2022), os sequenciadores de DNA, atualmente, são categorizados em quatro gerações distintas. A primeira geração é representada pelos métodos Maxam-Gilbert e Sanger, em que o primeiro método de sequenciamento de DNA, proposto por Allan Maxame e Walter Gilbert, no início dos anos 1970, se baseia na clivagem química da molécula de DNA. Em 1977, baseado nos estudos de Maxam-Gilbert, Frederick Sanger aperfeiçoou e popularizou o método. O método de sequenciamento Sanger é realizado a partir de uma cadeia “simples

de DNA", obtida pela desnaturação da molécula nativa por meio de calor. Esta fita "simples de DNA" atuará como molde para gerar a metade complementar da dupla hélice desejada. Essa técnica também conta com a marcação por radioatividade, que identifica fragmentos de DNA e a síntese desses fragmentos, a partir da fita molde (SILVA; LIMA; SOUZA, 2022; HAWKINS, 2017; MARTINS, 2013; SANTOS *et al.*, 2013).

O sequenciamento Sanger, assim como vários outros sequenciadores, é otimizado através da reação em cadeia da polimerase (PCR), um processo que gera pequenos fragmentos homogêneos de DNA. Esses fragmentos, especialmente aqueles com tamanhos na faixa de 200 a 1000 pares de bases (pb), são particularmente vantajosos para o sequenciamento Sanger. A PCR simplifica a obtenção desses fragmentos, contribuindo para a eficiência e precisão da técnica de sequenciamento ao produzir materiais genéticos ideais para análise (HAWKINS, 2017).

Já na segunda geração, o Método de Pirosequenciamento, método desenvolvido por Mostafa Ronaghi e Pål Nyrén, no Instituto Real de Tecnologia, em Estocolmo, no ano de 1996 (CARVALHO, M.; SILVA, D., 2010). O Pirosequenciamento é um método de sequenciamento, no qual se detecta a atividade da DNA polimerase, ao se medir o pirofosfato (PPi) liberado pela adição de um dNTP à extremidade 3' de um primer. Essa metodologia possibilita a identificação da sequência de uma única fita de DNA, através da síntese de sua fita complementar, adicionando um par de bases por vez e identificando a base adicionada em cada etapa (DIAS, 2014; MELLO *et al.*, 2012; PYROSEQUENCING. . . , s.d.).

A terceira geração foi marcada a partir de 2011, com o lançamento da Plataforma Ion Torrent PGM™, pelas empresas Life Technologies e Thermo Scientific. A nova forma de sequenciamento foi apresentada à comunidade como uma ferramenta de sequenciamento de bancada (*benchtop*), que se dedica principalmente ao sequenciamento de pequenos genomas ou grupos específicos de genes (SILVA; LIMA; SOUZA, 2022).

Esse novo método, para o sequenciamento de nucleotídeos, foi o primeiro a ser capaz de realizar o sequenciamento do DNA sem a utilização de moléculas pré-modificadas, como é o caso dos nucleotídeos marcados com fluorescência. Graças a isso, o método Ion Torrent tornou os equipamentos e reagentes para o sequenciamento mais baratos e mais acessíveis (RIBEIRO, I., 2020).

De acordo com Oliveira (2015), o fluxo de trabalho do método Ion Torrent possui três etapas, sendo a primeira a construção da biblioteca, a segunda, o preparo do *template* e, por fim, o sequenciamento em chip semicondutor. O processo de construção da biblioteca desejada tem início com a fragmentação do DNA de interesse, sendo essa fragmentação resultante de uma via enzimática ou mecânica. Em seguida, os fragmentos recebem, em suas extremidades, dois adaptadores de DNA,

cujas sequências são denominadas de A e P1.

Um diferencial das plataformas de terceira geração é o fato de possibilitarem o sequenciamento de múltiplas amostras, em uma única corrida de sequenciamento. Para que isso ocorra, são anexados, junto aos adaptadores A e P1, sequências denominadas *barcodes*, as quais são específicas para cada amostra. Isso permite que, durante a análise, cada amostra seja corretamente identificada. Esse processo de sequenciamento múltiplo é denominado *multiplex* (LOBATO, 2011), (YAMAGUTI, 2009).

Segundo Silva, Lima e Souza (2022), a quarta geração de sequenciamento genômico foi marcada pelo Método Oxford Nanopore, que foi apresentado em 2012, pela empresa Oxford Nanopore Technologies, o qual consiste na incorporação de um poro de proteína a uma membrana de polímero sintético, dentro do microchip. Essa tecnologia baseia-se em nanoporos individuais, inseridos em uma membrana sintética. À medida que cadeias específicas de DNA atravessam esses nanoporos, cada nucleotídeo provoca alterações específicas em uma corrente elétrica, sendo essa informação traduzida em dados de sequenciamento (EISENSTEIN, 2012). O método de sequenciamento da Oxford Nanopore dispensa o uso de marcação fluorescente, resultando em uma redução de custos e um aumento significativo na velocidade do processo (COSTA, 2020).

No cenário atual, a crescente geração de dados genéticos, como o sequenciamento de genomas e transcritomas, está se tornando cada vez mais acessível, à medida que os custos continuam a diminuir. Isso ocorre devido aos progressos nas técnicas de sequenciamento, juntamente ao aumento na capacidade de processamento e armazenamento de computadores, permitindo a produção extensiva de dados relacionados à abundância de proteínas e metabólitos (SILVA, 2022; OLIVEIRA, V. G. P. d., 2022; MELO, 2015; SILVA, P. I. T., 2012).

2.5 Aplicação do Sequenciamento

As técnicas de sequenciamento de DNA desempenham um papel fundamental em estudos dentro da biologia molecular, genômica, medicina, entre várias outras áreas. Essas técnicas possibilitam, aos pesquisadores, analisar as informações genéticas contidas nas amostras, fornecendo informações cruciais sobre a composição genômica, variações genéticas e a expressão de genes (CAMPOS, 2021; MARCO, 2019; STRÖHER, 2018).

Devido às limitações dos métodos tradicionais para a identificação de microorganismos e ao crescente interesse nas comunidades microbianas, é possível, por meio das tecnologias de sequenciamento de DNA, explorar não apenas os microorganismos cultiváveis em laboratório, mas também aqueles que não podem ser

cultivados em bancada, uma vez postulado que, cerca de 95-99% das espécies microbianas presentes em determinados nichos ecológicos, não são cultiváveis através dos métodos clássicos (FRANCO, 2022; NISHIYAMA, 2021; GOI; SOUZA, 2006).

Graças à inviabilidade do cultivo de vários microorganismos de interesse, os microbiologistas aderiram à prática de abordagens moleculares, independentes de cultivo para a realização de seus estudos, referentes à composição e distribuição das comunidades microbianas. Optaram principalmente pela análise de genes do RNA, amplificado diretamente do DNA obtido através da amostra em estudo e submetida à clonagem molecular (FRANCO, 2022).

Segundo Paranhos (2019) e Ronconi (2008), o sequenciamento do gene rrs, popularmente conhecido por 16S, é uma ferramenta essencial na taxonomia bacteriana, permitindo a identificação das espécies. Isso ocorre devido ao RNA ribossomal (rRNA) ser uma molécula altamente conservada, tanto funcional como sequencialmente, com regiões moderadamente variáveis.

De acordo com Vogel *et al.* (2009), os microrganismos presentes no “solo vivo” desempenham um papel fundamental para toda a vida superior em nosso planeta, sendo responsáveis por processos terrestres, que impactam diretamente nossa qualidade de vida. Esses microrganismos desempenham funções essenciais na fertilidade do solo, na ciclagem do carbono e na ciclagem de nutrientes, contribuindo de maneira crucial para os ecossistemas terrestres e para a sustentabilidade ambiental.

Embora os microrganismos sejam os responsáveis por funções essenciais nos solos, devido a limitações para o cultivo desses seres em laboratório, pouco se sabe sobre suas populações e funções desempenhadas. Nesse contexto, a metataxonômica apresenta-se como uma poderosa ferramenta de análise da diversidade microbiana do solo (SANTI, 2019).

Em seu trabalho, Marchesi e Ravel (2015) fala que a metataxonomia é o termo utilizado para denominar o processo de alto rendimento, usado para caracterizar toda a microbiota e criar uma árvore metataxonômica, a qual relaciona todas as sequências obtidas.

Graças ao conhecimento aprofundado dos microorganismos e seus processos biológicos, torna-se possível desenvolver estratégias mais eficientes para diversas aplicações, como a promoção da saúde humana, a otimização de processos industriais e agrícolas e a preservação ambiental, explorando da melhor forma possível os benefícios proporcionados pelos microorganismos em vários setores (KOIDE, 2022; SANTI, 2019).

2.6 Caracterização do Parque Nacional da Serra da Canastra (PNSC)

Situado na região sudeste do estado de Minas Gerais - Brasil, abrangendo 11 municípios mineiros, o Parque Nacional da Serra da Canastra (PNSC) foi criado em 3 de abril de 1972, pelo decreto n°70.355, o qual estabeleceu uma área de extensão, para o mesmo, de 197.787ha. Dessa área total estabelecida, atualmente, cerca de 82.000ha são regularizados (MESSIAS; FERREIRA, 2016).

O relevo da região oscila entre altitudes, que variam entre 900 a 1.496 metros, alternado entre seus tipos de solos, latossolos vermelho-amarelo, distróficos de textura argilosa. O clima da região exibe um padrão sazonal, com grande incidência de chuvas, no período de verão, e estiagem durante o inverno. A temperatura média, durante o mês mais frio, fica a abaixo de 18 °C, enquanto a do mês mais quente excede 22 °C (SILVA, A. A. C., 2019).

A região do PNSC está situada no bioma cerrado e apresenta uma diversidade de fitofisionomias, incluindo florestas mesófilas de encosta, capões, cerradão, cerrado sensu stricto, campo cerrado, campo limpo e campo rupestre em sua composição vegetal (EDMUNDO, 2014).

Segundo ICMBio (2016), o PNSC é lar de uma notável diversidade de fauna e flora. No PNSC, encontram-se numerosas espécies raras, endêmicas e/ou ameaçadas, conferindo-lhe status de área prioritária para conservação na região Neotropical, especialmente devido à riqueza de sua avifauna. Além disso, é considerado um abrigo de grande relevância para aves, designado como Área Importante para a Conservação das Aves (IBA- *Important Bird Are*), no domínio do Cerrado. No que diz respeito à flora, muitas espécies são exclusivas do parque e algumas, até recentemente, eram desconhecidas pela comunidade científica.

2.6.1 Causa e impactos das queimadas e incêndios florestais

Os raios representam uma das principais origens de ignição em savanas, onde a estação seca prolongada torna o componente herbáceo suscetível a incêndios. As queimadas resultantes de descargas elétricas são consideradas mais como um distúrbio, capaz de contribuir para a conservação da biodiversidade natural em ecossistemas propensos ao fogo (MAGALHÃES; LIMA; RIBEIRO, G. A., 2012).

Diante dos benefícios decorrentes da influência do fogo na vegetação dessas regiões, a abordagem para combater os incêndios de origem natural no PNSC envolve apenas intervenção se a chuva não for suficiente para extinguir os focos. O combate direto é implementado apenas quando os incêndios atingem proporções consideráveis, ameaçando as áreas de proteção que abrigam fitofisionomias florestais, locais de pesquisa e refúgios de espécies (ICMBIO, 2016).

No Brasil, os incêndios florestais nas Unidades de conservação são provocados, em sua grande maioria, pela utilização errada do fogo para: (a) a renovação de pastagens, melhorando a qualidade do alimento, (b) limpeza de áreas, tanto agrícola quanto florestal, (c) como ferramenta para abertura de novas fronteiras agrícolas, (d) controle de pragas e doenças, (e) melhorar o manejo de pré-colheita em diversas culturas (REDIN *et al.*, 2011; MEDEIROS; FIEDLER, 2004).

A utilização do fogo, normalmente, é realizada sem a construção de aceiros para contenção das chamas, assim, como as condições climáticas não são verificadas, o período da queima é inadequado e há, por parte dos praticantes, o desconhecimento sobre equipamentos para o controle do fogo e os impactos provocado por ele no ambiente (MEDEIROS; FIEDLER, 2004).

O fogo, como agente transformador, possui um grande potencial de modificação do ambiente, uma vez que o mesmo pode ser responsável por interferir nas trajetórias sucessionais das comunidades vegetais, assim como na microbiota do solo; além de modificar as propriedades físico-químicas e estruturais do mesmo (ABREU *et al.*, 2004). Várias dessas alterações possuem caráter sinérgico, dos quais, muitas vezes, são irreversíveis e podem vir a impactar o clima e o processo hidrológico do local. É necessário salientar que os efeitos microbianos do solo, assim como os efeitos do carbono no solo, são de longo prazo, podendo durar até mesmo décadas (MASSMAN; FRANK; MOONEY, 2010).

2.6.2 O Plano de Manejo Integrado do Fogo do Parque Nacional da Serra da Canastra (PNSC)

Segundo Souza *et al.* (2016), o Parque Nacional da Serra da Canastra (PNSC) encontra-se entre as unidades de conservação (UC) brasileiras que mais sofrem com a alta frequência de incêndios. Os incêndios criminosos são identificados como as principais causas das extensas queimadas que afetam a unidade. Apesar dos esforços contínuos de prevenção e combate aos incêndios, ao longo dos anos no parque, os resultados obtidos ainda não são completamente satisfatórios.

Nesse contexto, de acordo com o INPE (2016), a implementação do sistema de monitoramento por imagens de satélite, resultante da parceria entre INPE, IBAMA e ICMBIO, a partir de 2010, teve como objetivo registrar a ocorrência de incêndios florestais e avaliar a extensão das áreas afetadas.

Essa colaboração proporcionou informações precisas, possibilitando a elaboração de Relatórios de Ocorrência de Incêndios, enriquecidos por mapas detalhados das regiões atingidas pelo fogo. Com base nessas informações, observou-se que as ações de prevenção e combate a incêndios florestais não têm alcançado os resultados desejados, especialmente em relação aos incêndios tardios que ocorrem no final

da estação seca. Esses eventos têm causado consideráveis danos à biodiversidade. No entanto, esses relatórios têm desempenhado um papel crucial, como norteadores no planejamento do manejo integrado do fogo (SILVA, A. A. C., 2019; ICMBIO, 2016).

Conforme descrito por Schmidt *et al.* (2016), o MIF é uma abordagem que leva em consideração aspectos ecológicos, culturais e de manejo. Essa abordagem propõe a utilização de queimadas controladas, juntamente às medidas de prevenção e combate a incêndios, com o objetivo de assegurar a conservação e o uso sustentável dos ecossistemas.

Segundo Souza (2017), em 2017, foi implantado o Programa de Manejo Integrado do Fogo no PNSC. O MIF almeja, como resultados práticos finais, a busca pela criação ou promoção dos cenários mais favoráveis para todos os participantes e elementos envolvidos na questão do fogo, abrangendo a biodiversidade, os recursos naturais, as comunidades humanas tradicionais que utilizam o fogo, assim como aquelas que não fazem uso direto, mas que podem ser impactadas em diferentes escalas e intensidades por seus efeitos diretos e indiretos. Esses objetivos estão alinhados com os princípios da biologia da conservação, enquanto disciplina científica, e com as práticas de manejo de UC, conforme estipulado pela legislação brasileira, por meio da Lei 9985/2000 ou em outros países.

Além disso, foi implementada no PNSC a criação de aceiros, com o intuito de gerenciar e, consequentemente, reduzir os impactos de incêndios controlados e não controlados (SILVA, A. A. C., 2019).

De acordo com o ICMBio (2016), as queimadas devem ser realizadas de forma controlada, em setores alternados de modo que: (i) a área queimada só se torna sujeita a uma nova queima dois anos após a ocorrência da queimada inicial; (ii) a queima deve ser realizada durante os meses de março, abril e maio, conforme o regime climático, evitando períodos de estiagem; (iii) os setores destinados à queima serão definidos pelos aceiros, adotando a prática de não queimar a totalidade de um talhão, simultaneamente. O processo inicia-se em uma lateral, avançando para outros setores. Após um intervalo de tempo adequado (quanto mais extenso, melhor), uma nova frente de fogo será iniciada em outra lateral, seguindo esse ciclo, sucessivamente, até que todos os limites sejam queimados.

O procedimento de queima proposto possibilitará que os animais não se vejam ameaçados pelo fogo, ao contrário do que ocorre em queimadas convencionais, onde os quatro lados são queimados simultaneamente. Além disso, o intervalo de retorno ao talhão favorece o brotamento da vegetação, proporcionando alimento e atraindo a fauna para a área queimada, minimizando o risco de mortes da fauna local por falta de alimento (SILVA, A. A. C., 2019; SOUZA, 2017).

2.6.3 Microrganismos do Solo e Seu Papel Ecológico

Os microrganismos desempenham um papel crucial na biodiversidade terrestre, contribuindo de maneira significativa para os ciclos biogeoquímicos e influenciando no funcionamento dos ecossistemas (SILVA, F. C. d. *et al.*, 2016; RODRIGUES, 2013; BELL *et al.*, 2005; TORSVIK; ØVREÅS; THINGSTAD, 2002).

Segundo Pedrosa *et al.* (2015), a maior concentração de atividade biológica encontrada no solo estão principalmente situadas em suas primeiras camadas. Nessas camadas, o componente biológico corresponde a uma parcela de 0,5% do volume total do solo e corresponde a menos de 10% de sua matéria orgânica.

Os microrganismos encontrados no solo podem habitar tanto na camada de serapilheira, onde contribuem para a decomposição do material orgânico na superfície do solo, quanto na rizosfera, região ao redor do conjunto de raízes, que constitui uma zona de intensa atividade microbiana. Essa maior atividade de microrganismos na rizosfera deve-se ao fato das raízes secretarem muitos tipos de açúcares, aminoácidos e vitaminas, que estimulam o desenvolvimento microbiano (MANHAES; FRANCELINO, 2013).

Dentre essas funções desempenhadas pelos microrganismos do solo, destacam-se a capacidade de fixar nitrogênio, a decomposição de resíduos orgânicos, a desintoxicação de pesticidas, o fornecimento de nutrientes para o solo e a produção de compostos bioativos, como vitaminas e hormônios que auxiliam no crescimento das plantas (NAIR; NGOUAJIO, 2012; GOI; SOUZA, 2006; ALFONSO; LEYVA; HERNÁNDEZ, 2005).

A Fixação Biológica de Nitrogênio (FBN) desempenha um papel essencial para a manutenção da vida em nosso planeta. Embora o nitrogênio constitua aproximadamente 80% da atmosfera terrestre, animais e plantas não têm a capacidade de absorvê-lo, diretamente, em sua forma gasosa (LESSA, 2007). Para atender às suas necessidades nutricionais relacionadas a esse elemento, os seres vivos dependem da absorção de suas formas nitrogenadas, como amônia, nitratos ou compostos orgânicos. Essas formas nitrogenadas são metabolizadas para a construção de biomassa, uma vez que o nitrogênio, como macronutriente essencial, está presente em aminoácidos, proteínas, DNA, entre outras estruturas celulares (GOMES *et al.*, 2020; MARCHETTI; BARP, 2015).

A FBN é realizada por microrganismos procarióticos, conhecidos como diazotróficos. Estes microrganismos podem existir de maneira independente, associar-se a espécies vegetais ou estabelecer simbiose com leguminosas. Os diazotróficos desempenham um papel significativo no crescimento vegetal, tanto por meio da FBN, quanto pela produção de substâncias que favorecem o desenvolvimento radicular, como o ácido indol acético, entre outros. Portanto, as bactérias diazotróficas associati-

vas são reconhecidas como rizobactérias promotoras do crescimento vegetal (RPCV), desempenhando um papel crucial na interação com as raízes das plantas e na ciclagem de nutrientes (SANTOS *et al.*, 2017; MOREIRA *et al.*, 2010).

Como já dito por Alfonso, Leyva e Hernández (2005), uma das funções desempenhadas pelos microrganismos é a decomposição e liberação de nutrientes no solo. Segundo Tauk (1990), a decomposição de resíduos orgânicos no solo é um processo complexo e diversificado. Inicialmente, ocorre uma rápida decomposição de materiais facilmente degradáveis, seguido por um processo mais lento para materiais mais resistentes à degradação. Essa desaceleração pode ser explicada pela adsorção, estabilização de metabólitos e redução da taxa de biomassa no solo.

Conforme descrito por Machado *et al.* (2012), a decomposição dos resíduos no solo contribui para o aumento da disponibilidade de nutrientes, promovendo maiores concentrações de nitrogênio (N) e fósforo (P), além de ampliar a disponibilidade de outros nutrientes nas camadas superficiais do solo.

Segundo Gavrilescu (2005), corroborado por Vacondio (2014), os microrganismos, por meio da produção de suas enzimas, têm a capacidade de transformar pesticidas e outros compostos químicos em fonte de alimento e energia. Essas enzimas provocam alterações na estrutura química, propriedades toxicológicas e no potencial poluente desses compostos (MATTOS, 2015).

Ainda de acordo com Vacondio (2014), a aplicação de microrganismos como agentes de degradação de várias substâncias, como corantes, cosméticos, detergentes, medicamentos e produtos químicos agrícolas, é reconhecida como um método eficaz para mitigar os efeitos adversos desses contaminantes no ambiente. A degradação biológica, promovida por atividades microbianas, é considerada o principal processo para a dissipação de contaminantes no solo.

O solo constitui um ambiente naturalmente diverso em sua composição, abrigando inúmeras e variadas populações de microrganismos de todos os tipos. Com bilhões de indivíduos, cada um desempenhando funções específicas e ocupando nichos ecológicos distintos, o solo contribui para a manutenção da vida em todo o planeta. Os principais conjuntos de organismos presentes no solo englobam bactérias, vírus, fungos, algas, e uma macrofauna composta por artrópodes e protozoários (LEAL *et al.*, 2021; MATTOS, 2015).

3 MATERIAIS E MÉTODOS

Neste capítulo, serão abordados os materiais e métodos adotados para a execução deste trabalho.

Devido ao seu acesso gratuito e à ampla gama de análises que permite realizar em um único ambiente, a Plataforma QIIME 2 foi escolhida como a ferramenta de análise de dados biológicos.

A metodologia utilizada consistiu em pesquisar, na literatura e em fóruns especializados na ferramenta Qiime 2, informações sobre os recursos computacionais necessários para a execução de processos de mineração de dados biológicos. As informações encontradas foram utilizadas para nortear a escolha das máquinas a serem testadas quanto ao seu desempenho, em processos de bioinformática. A pesquisa analisou o desempenho de cinco configurações distintas de máquina.

Devido ao alto custo para a montagem de todas as configurações de máquinas analisadas, aliado à facilidade de criação e gerenciamento de máquinas virtuais, foi contratado um serviço de virtualização de servidores. Isso possibilitou maior economia para a pesquisa, além de oferecer maior variedade de opções de configurações de máquinas a serem testadas.

Para validar os ambientes configurados e avaliar o desempenho de cada máquina, foi realizada a análise de um conjunto de dados biológicos, utilizando o Qiime 2. O desempenho de cada máquina foi então avaliado com base na conclusão da pipeline desenvolvida e no tempo necessário para completar o processo.

3.1 Descrição da *pipeline* Qiime 2

A *pipeline* utilizada foi constituída por quatro etapas: pré-processamento dos dados, limpeza e remoção de ruídos, mineração dos dados e, por último, análise dos dados. A primeira etapa envolveu a importação das 96 bibliotecas para um artefato Qiime 2. Após a criação desse artefato, foi gerado um arquivo de visualização correspondente. Ambos os artefatos foram utilizados na etapa de limpeza dos dados. O artefato contendo os dados importados serviu como *input* no processo de limpeza, enquanto o artefato de visualização foi utilizado para determinar os pontos de corte e truncamento a serem aplicados.

A etapa de limpeza e eliminação de ruídos consistiu em remover todas as sequências com baixa qualidade de sequenciamento, sequências únicas, quiméricas e duplicadas. Vale ressaltar que, nessa etapa, o corte e truncamento das sequências foram realizados manualmente, com base na análise dos gráficos de qualidade, gerados a partir do artefato de visualização criado anteriormente.

A etapa de mineração dos dados foi composta por processos de manipu-

lação do arquivo resultante da etapa de limpeza. Nessa fase, os dados foram submetidos a processos de classificação taxonômica, criação de uma árvore filogenética, análise de alfa e beta diversidade e identificação de vias metabólicas.

O processo de classificação taxonômica utilizou um classificador taxonômico pré-treinado e um banco de dados de referência, reconhecido pela comunidade científica para atribuir a taxonomia às sequências de DNA analisadas.

Já para a criação da árvore filogenética, foi utilizado o comando *qiime phylogeny align-to-tree-mafft-fasttree*, tendo como entrada o artefato *representative_sequences.qza* resultante do processo realizado anteriormente. Essa árvore, inicialmente, foi gerada sem raiz e, posteriormente, foi enraizada no ponto médio da maior distância entre as extremidades da árvore sem raiz.

Para a análise de alfa e beta diversidade, foram utilizados os comandos *qiime diversity alpha* e *qiime diversity core-metrics-phylogenetic*, respectivamente. Cada comando foi definido com os parâmetros necessários para a realização de cada uma das métricas.

Por fim, temos a identificação das vias metabólicas, que seguem a mesma sequência de processos já descrita. No entanto, ela é desenvolvida dentro de um *plugin*, adicionado ao Qiime 2, denominado Picrust 2. Vale ressaltar que não é necessário importar e filtrar os dados novamente, podem-se iniciar os processos utilizando o artefato resultante do processo de filtragem já realizado.

Após a conclusão das etapas anteriores, a análise dos dados envolve a interpretação das informações, geradas ao final da mineração, e a aplicação dessas informações aos objetivos propostos.

3.2 Diferentes estratégias de delineamento de máquina para execução da *pipeline*

Com base nas informações encontradas na literatura e em fóruns especializados na ferramenta Qiime 2, foi possível delinear as configurações necessárias para a execução das análises de dados, utilizando o Qiime 2. Para as análises, foram configuradas cinco máquinas distintas para os testes, em que, cada uma, apresentou um aumento significativo de recursos em relação à máquina anterior. As configurações analisadas variaram desde máquinas comuns, encontradas em computadores pessoais, até configurações mais potentes, geralmente utilizadas em pequenos servidores. A primeira máquina configurada foi a máquina 01, que possuía a menor disponibilidade de recursos, com 2 vCPUs e 4 GB de RAM. Essa configuração mais simples foi utilizada nos testes para avaliar se máquinas, com recursos limitados, são capazes de realizar processos de bioinformática. Da mesma forma, as máquinas 02 e 03, com 4 vCPUs e 8 GB de RAM e 6 vCPUs e 16 GB de RAM, respectivamente,também foram

testadas para verificar se computadores pessoais, com diferentes níveis de capacidade, podem executar processos de mineração de dados biológicos de forma eficaz.

As máquinas 04 e 05 forão as que apresentaram disponibilidade de recursos não encontrados com frequência, em computadores pessoais, apresentando 8 vCPUs e 32 GB de RAM e 16 vCPUs e 64 GB de RAM, respectivamente. A máquina 04 foi descrita, com base nas informações coletadas, como aquela com a configuração necessária para a realização de pesquisas de bioinformática, em ambientes de pesquisa reais. Já a máquina 05, que apresenta a maior disponibilidade de recursos, foi testada para verificar se o aumento na quantidade de recursos resulta em melhor desempenho no processamento de análises de bioinformática.

Devido à grande quantidade de recursos demandados para a realização da pesquisa, a contratação de um serviço de virtualização de servidores fez-se necessária. Isso ocorreu porque a máquina 05 possui recursos não encontrados no cotidiano. Graças aos diferentes níveis de processamento das máquinas analisadas,

foi possível avaliar as estratégias para o delineamento de recursos, nas análises de bioinformática. Para a avaliação das máquinas, todas foram submetidas às mesmas baterias de teste, nas quais todas seguiram os mesmos processos descritos, na *pipeline* específica.

Os resultados observados, em cada um dos testes, foram computados em tabelas e posteriormente analisados, levando-se em consideração a conclusão dos testes e qual o tempo demandado para o mesmo. Para evitar erros na cronometragem do tempo demandado na execução de cada teste, foi inserido o parâmetro *time* no código de cada processo executado. Esse parâmetro registrava a informação sobre quanto tempo o processo levou para ser executado.

3.3 Validação da *pipeline* com os dados da microbiota do solo do PNSC

Para a validação dos ambientes configurados e da *pipeline* desenvolvida, foram utilizados dados da microbiologia do solo, presentes no PNSC. Essa validação consistiu em executar toda a *pipeline* desenvolvida, a partir da leitura de outras pipelines, presentes na literatura, e analisar os dados provenientes de seus processos.

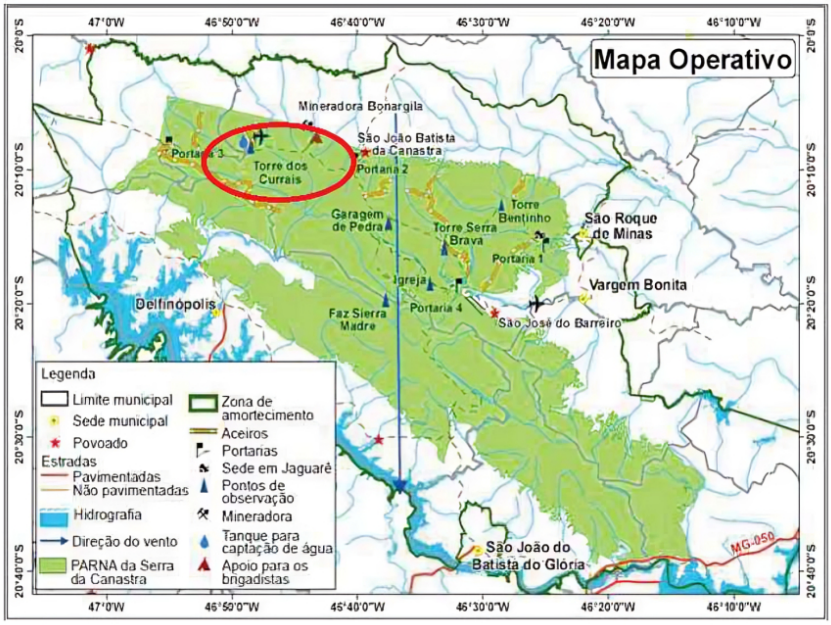
A validação foi principalmente baseada na interpretação dos resultados provenientes da análise de beta diversidade dos dados utilizados. Os resultados desse processo foram úteis para analisar como as chamas modificam a microbiologia do solo.

3.4 Dados utilizados

Os dados utilizados nesses trabalhos são provenientes do banco de dados do Laboratório de Biologia Molecular (LAB.BIOMOL), do Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais – *Campus* Bambuí. Os dados selecionados são provenientes de pesquisas conduzidas no LAB.BIOMOL, oriundos de outros projetos, porém, até o momento, não foram objeto de estudo específico.

O conjunto de dados, destacados do acervo do LAB.BIOMOL, pertence a uma pesquisa de mestrado, intitulada “Levantamento da Comunidade Bacteriana em Diferentes Tipos de Solo Impactados e Não Impactados Por Queimadas No Parque Nacional da Serra da Canastra – Minas Gerais”, no qual a autora da pesquisa realizou o levantamento de dados durante o plano de manejo integrado do fogo do Parque Nacional da Serra da Canastra (PNSC), que ocorreu entre 20 e 23 de maio, de 2017, pelo Instituto Chico Mendes de Conservação da Biodiversidade (ICMBio), em colaboração com uma equipe de brigadistas do ICMBio e outra equipe de brigadistas do PRÓ-FOGO Brasília. A Figura 5 demonstra a área na qual foi implantado o MIF no PNSC.

Figura 5 – Mapa operativo do 1º MIF no PNSC.



Fonte: (SILVA, A. A. C., 2019)

Os dados foram coletados de maneira sistemática, envolvendo a amostragem de oito áreas distintas, cada uma caracterizada por suas particularidades. A primeira área apresenta o maior histórico de queimadas; a segunda, o menor histórico de queimadas; a terceira, queimada no último ano; a quarta, queimada há dois anos; a quinta, foi coletada antes da queimada (Queimada há dois anos - MIF); a sexta, após a queimada (a favor do vento – coletada, aproximadamente, 30-40 minutos após a quei-

mada -MIF); a sétima, antes da queimada (queimada há dois anos - MIF) e a oitava, após a queimada (contra o vento – coletada imediatamente após a queimada - MIF). A Tabela 1, demonstra, de forma mais simples, a distribuição que foram coletadas as amostras.

Tabela 1 – Disposição das amostras coletadas nas áreas atingidas pelo manejo do fogo no PNSC

Áreas de coleta	Grupos amostrais	
	Rizosfera da Planta (RP)	Solo (S)
1. Maior histórico de queimadas	T1RP	T1S
2. Menor histórico de queimadas	T2RP	T2S
3. Queimada no último ano	T3RP	T3S
4. Queimada há dois anos	T4RP	T4S
5. Antes da queimada -MIF	T5RP	T5S
6. Após a queimada - MIF (a favor do vento)	T6RP	T6S
7. Antes da queimada - MIF	T7RP	T7S
8. Imediatamente após a queimada - MIF (contra o vento)	T8RP	T8S

Fonte: Elaborado pelo autor, 2024.

Para cada ponto de coleta definido, foram abertas covas, utilizando ferramentas, como enxada e sacho. Em cada local de amostragem, foram coletadas três amostras distintas, em duas profundidades diferentes (de 0-10cm, abrangia a região da rizosfera da planta e de 10-20cm, compreendia uma porção do solo e a terceira amostra seria uma amostra composta de 0-20cm de profundidade), desse modo totalizando um montante de 48 amostras (8 pontos amostrais x 2 profundidades x 3 repetições).

A partir das paredes de cada cova, utilizando uma faca previamente desinfetada com álcool 70% e posteriormente flambada, foram coletadas amostras de aproximadamente 50 gramas de solo. Essas amostras foram transferidas para sacos plásticos esterilizados e acondicionadas em uma caixa de isopor, à temperatura ambiente. Posteriormente, foram transportadas para o laboratório de Biologia Molecular do IFMG - Campus Bambuí e armazenadas em geladeira, mantendo uma temperatura aproximada de 10℃ (SILVA, A. A. C., 2019).

A extração do DNA total das amostras de solo foi conduzida por meio do Kit

FastDNA® SPIN Kit for Soil (BIO 101 - MP Biomedicals), com o número de catálogo #6560-200, seguindo o protocolo fornecido pelo fabricante do produto. O material extraído foi posteriormente empregado na construção de bibliotecas de rDNA 16S, para análise de diversidade e outros estudos (SILVA, A. A. C., 2019).

As bibliotecas de sequenciamento foram preparadas, utilizando o TruSeq® DNA PCR-Free Sample Preparation Kit (Illumina, EUA), seguindo as orientações do fabricante, e códigos de índice foram incorporados. A qualidade da biblioteca foi avaliada no sistema de Fluorômetro Qubit 2.0 (Thermo Scientific) e no Agilent Bioanalyzer 2100. Finalmente, a biblioteca foi sequenciada em uma plataforma Illumina HiSeq 2500, gerando leituras pareadas de 250 pares de bases (SILVA, A. A. C., 2019).

3.5 Plataforma de *Cloud Computing*-Absam

Os serviços de computação em nuvem, utilizados neste trabalho, foram contratados na plataforma Absam, uma *startup* brasileira que busca oferecer soluções tecnológicas vantajosas para empresas de todos os portes, incluindo pequenas, médias e grandes organizações. A plataforma oferece serviços de hospedagem de sites, sistemas desktop, bancos de dados, cargas de trabalho de IA e *Machine Learning*, *game servers*, VPN ou ambientes de CI / CD e *Cloud Computing Server*(ABSAM, 2024).

O serviço contratado foi o de *Cloud Computing Server*, que possibilitou a criação de máquinas virtuais e a configuração dos ambientes para a realização dos testes e análises propostos. Todas as máquinas virtuais foram estabelecidas, com base no mesmo sistema operacional (SO): Ubuntu 20.04 (LTS) X64. O armazenamento de cada máquina virtual consistiu no armazenamento padrão oferecido pela plataforma, que foi de 50GB.

As máquinas virtuais variavam entre si apenas em relação à quantidade de núcleos de processamento e à capacidade de memória RAM. A configuração das máquinas foi a seguinte: a máquina mais simples possuía 2 vCPUs e 4 GB de RAM; a segunda máquina, 4 vCPUs e 8 GB de RAM; a terceira, 6 vCPUs e 16 GB de RAM, a quarta e principal máquina do experimento apresentava 8 vCPUs e 32 GB de RAM, e a quinta, a mais robusta, contava com 16 vCPUs e 64 GB de RAM. A Tabela 2 demonstra de forma mais simples como os recursos foram distribuidos entre as máquinas.

Todas as máquinas foram configuradas com a instalação do Miniconda, que foi responsável por criar o ambiente, no qual a ferramenta QIIME 2 e seus *plugins* serão instalados, ou seja, a ferramenta QIIME e seus *plugins* serão utilizados dentro de um ambiente próprio e exclusivo dentro da máquina.

O Miniconda é um instalador mínimo e gratuito, projetado para o conda. Ele representa uma versão enxuta e inicial do Anaconda, contendo apenas o conda,

Tabela 2 – Tabela de configurações de máquinas

Máquinas	Armazenamento (GB)	vCPUs	RAM (GB)	Custo por hora (R$)	Custo por minuto (R$)
Máquina 01	50	2	4	0,0875	0,00146
Máquina 02	50	4	8	0,175	0,00292
Máquina 03	50	6	16	0,35	0,00583
Máquina 04	50	8	32	0,70	0,0117
Máquina 05	50	16	64	1,40	0,02333

Fonte: Elaborado pelo autor, 2024.

o Python, os pacotes fundamentais necessários para ambos e um conjunto restrito de outros pacotes úteis, como pip, zlib, entre outros (ANACONDA, INC., 2024).

Com o ambiente conda devidamente instalado e iniciado, o próximo passo foi a instalação do QIIME 2, para isso, foi necessário acessar o *site* da plataforma, escolher qual versão deseja instalar e seguir os tutoriais de instalação. Para esse trabalho, a versão escolhida foi a q2cli 2021.11 (uma interface de linha de comando). A instalação padrão do QIIME fornece um conjunto de *plugins* básicos para a realização de diversas análises, caso haja a necessidade da instalação de outros *plugins* específicos, esses poderão ser adicionados de forma fácil e simples pelo usuário.

Um exemplo de *plugin*, que foi utilizado neste trabalho e não estava incluído no conjunto padrão de *plugins* do QIIME, foi o PICRUST2. Trata-se de um software utilizado para prever abundâncias funcionais, com base apenas em sequências de genes marcadores. Sua instalação foi realizada conforme as instruções fornecidas no tutorial, disponível na página do *GitHub* (DOUGLAS, 2024).

3.6 Análises de bioinformática

Nessa sessão, serão apresentados os conceitos importantes a respeito da análise de bioinformática.

3.6.1 Preparação dos dados

Para validar o ambiente configurado, como proposto anteriormente, foram realizadas análises de microbiomas. Assim, com o ambiente configurado e os dados selecionados, as análises foram conduzidas, seguindo os procedimentos descritos dentro da *pipeline* customizada, que foi desenvolvida de acordo com as *pipelines* descritas nos tutoriais, fornecidos pela plataforma QIIME. Essa *pipeline* inclui várias etapas, começando pela importação das bibliotecas.

Antes da importação dos dados, uma etapa de pré-processamento dos mesmos fez-se necessária. Essa etapa de pré-processamento dos dados consistiu em renomear os arquivos e diretórios, de forma que atendessem os padrões de no-

menclatura Ilumina e as diretrizes de entrada da ferramenta. Com essa pré-etapa concluída, a etapa de importação pode ser realizada.

Outro pré-tratamento necessário, que se deve realizar antes do início da análise, é a criação de três arquivos *mapping file*, os quais consistem em arquivos de extensão (.TSV). O primeiro arquivo *mapping file* que foi criado, denominado de (*mappingfile_wil_full.tsv*), possui a relação de todas as amostras estudadas e seus classificadores.

O Segundo arquivo *mapping file* criado, denominado de (*mappingfile_beta_rp_will_full.tsv*), diferente do primeiro, possui apenas a relação das amostras classificadas como (RP) e por último, o terceiro *mapping file* denominado (*mappingfile_beta_s_will_full.tsv*), ao contrario do segundo possui a relação das amostras classificadas como (S).

3.6.2 Importação das bibliotecas

As bibliotecas fornecidas pelo LAB.BIOMOL, consistiam em dados no formato (.fastq.qz), que possui índices de qualidade associados às sequências nucleotídeos, contidos nos arquivos. Esses dados tratavam-se também de dados de multiplexados, ou seja, as leituras foram separadas e organizadas em dois arquivos distintos, referentes a uma mesma amostra. Nesse contexto, há dois arquivos fastq.gz para cada amostra, cada um contendo as leituras direta ou reversa de cada amostra sequenciada, sendo assim, as 48 amostras sequenciados geraram um total de 96 bibliotecas fastq. É importante salientar que essas bibliotecas foram fornecidas sem a presença de *barcodes* inseridos durante o processo de sequenciamento, o que dispensa o processo para retirá-los.

Com todas essas informações, é possível identificar que o método de importação das bibliotecas a ser utilizado é o método "*Casava 1.8 paired-end demultiplexed fastq*". Esse processo de importação recebe as 96 bibliotecas, que estão todas dentro de um mesmo diretório, como entrada e converte-as em um único artefato de saída QIIME 2 (.qza). Após a importação, é necessária a criação de um arquivo de visualização (.qzv), o qual permite observar a quantidade de sequências obtidas por amostra e também gerar resumos de qualidade, em cada posição dos dados sequenciados (BOLYEN *et al.*, 2019).

Os gráficos gerados no arquivo de visualização são resultados de uma amostragem aleatória de 10.000 amostras, de um montante de 3.844.404 sequências, sem reposição. No gráfico referente às leituras diretas, o comprimento mínimo de sequência, identificado durante a subamostragem, foi de 217 bases. Já no gráfico referente às leituras reversas, o comprimento mínimo de sequência, identificado durante a subamostragem, foi de 213 bases.

De acordo com os dados observados nos gráficos de qualidade, os pontos de truncamento escolhidos, para serem utilizados no controle de qualidade, foram os pontos 200 e 186 para as sequências diretas e reversas, respectivamente. Como a qualidade do sequenciamento das bases iniciais apresentavam alta qualidade, não foi necessário o corte das sequências iniciais.

3.6.3 Controle de qualidade das sequências

Após analisar os gráficos de qualidade resultantes no processo anterior, foi realizado o controle de qualidade das sequências e a construção da tabela de características. Para esse processo, foram utilizados os índices de qualidade identificados nos gráficos de qualidade, juntamente à *pipeline* DADA2, uma *pipeline* para identificar e corrigir, quando possível, dados de sequência de *amplicon* da Illumina. Este processo de controle de qualidade, implementado no *plugin* q2-dada2, é responsável por filtrar e remover sequências quiméricas (BOLYEN *et al.*, 2019). O controle de qualidade resultará em três arquivos artefatos, denominados, por padrão, como *denoising_stats.qza, representative_sequences.qza e table.qza*, os quais serão os arquivos de entrada nos próximos processos.

3.6.4 Criação de um arquivo de visualização relacionado a um arquivo mapping file

Baseado no arquivo artefato *table.qza*, juntamente com o arquivo *mapping-file_wil_full.tsv*, criado anteriormente, é possível gerar um arquivo de visualização interativo, através do *plugin qiime feature-table summarize*, no qual é possível filtrar as amostras pela quantidade de sequências contidas nelas.

3.6.5 Atribuição taxonômica e criação da árvore filogenética

A atribuição de taxonomia foi realizada com o auxílio de um banco de dados referência, o banco de dados referência SILVA, e um classificador pré-treinado, já embutido em um *plugin* do QIIME, o *q2-feature-classifier*. Esse processo confronta as sequências estudadas com o banco referência e determina a qual classificação taxonômica cada sequência estudada pertence. O resultado desse processo é um artefato QIIME, que pode ser convertido em um arquivo de visualização. Nesse arquivo, é possível observar uma tabela de três colunas, onde a primeira coluna contém as sequências analisadas; a segunda coluna, a taxonomia associada e a terceira coluna apresenta o grau de confiabilidade dessa atribuição.

Utilizando o comando *qiime phylogeny align-to-tree-mafft-fasttree*, tendo

como entrada o artefato r*epresentative_sequences.qza*, foi possível gerar uma árvore filogenética, baseadas nas sequências contidas no arquivo de entrada. Essa árvore, inicialmente, é gerada sem raiz e, posteriormente, é enraizada no ponto médio da maior distância entre as extremidades da árvore sem raiz.

3.6.6 Análise de diversidade com todas as amostras estudadas

O primeiro passo no teste de hipóteses, em ecologia microbiana, geralmente, envolve a avaliação da diversidade dentro (alfa) e entre amostras (beta). Uma abordagem comum para isso é utilizar o método *qiime diversity core-metrics-phylogenetic*. Este método rarifica a tabela de recursos de entrada, calcula diversas métricas de diversidade alfa e beta e gera visualizações de análise de coordenadas principais (PCoA) no Emperor. As métricas calculadas para Diversidade Alfa são: Índice de diversidade de Shannon, Recursos observados, Diversidade de Faith e Equidade de Pielou. Já para Diversidade Beta, são calculadas: Distância Jaccard, Distância Bray-Curtis, Distância UniFrac não ponderada e Distância UniFrac ponderada.

Nesse processo, é importante setar a profundidade em que as amostras serão rarefeitas, para isso, é utilizado o *script —p-sampling-depth*.Este *script* realizará uma subamostragem aleatória de cada amostra, ajustando a profundidade de amostragem para o valor fornecido. Vale salientar que, se as contagens de qualquer amostra for menor que o valor estipulado, essa amostra será retirada da análise de diversidade.

Para garantir que todas as amostras estivessem dentro da análise de diversidade, foi estipulada uma rarefação com uma contagem total igual a 2000. Ao final desse processo, foi gerado automaticamente um diretório, contendo o resultado de todos os cálculos referentes às métricas mencionadas anteriormente, assim como alguns arquivos de visualização, os quais possibilitam a observação de como os grupos bacterianos estão distribuídos geneticamente.

Outros dois testes foram realizados nessa análise, o teste de *Chao1* e o teste de PERMANOVA. O teste de *Chao1* busca estimar a riqueza total de espécies em uma comunidade biológica, com base nas observações de espécies comuns e raras. Para isso, utiliza-se o artefato *table.qza*, gerado na etapa de controle de qualidade, juntamente ao *plugin qiime diversity alpha* e o parâmetro *—p-metric chao1*. Esse processo resultará em um artefato QIIME, que pode ser transformado em um arquivo de visualização posteriormente.

O teste de PERMANOVA é uma análise estatística, baseada em permutações, que tem como objetivo avaliar a composição das amostras, comparando-as para determinar o grau de similaridade entre elas. Esse processo é realizado pelo *plugin qiime diversity beta-group-significance* e recebe, como entrada, o ar-

quivo *weighted_unifrac_distance_matrix.qza*, gerado pelo *plugin qiime diversity core-metrics-phylogenetic*, já executado anteriormente, junto ao *mappingfile_wil_full.tsv* e o parâmetro de *–m-metadata-column*. É setado com o título da coluna que especifica qual ponto de coleta cada amostra foi retirada.

3.6.7 Identificação de vias Metabólicas

Para a identificação das vias metabólicas presentes nas amostras, o *plugin* PICRUST2 foi utilizado, tendo como entrada os artefatos *table.qza* e *representative_sequences.qza*, juntamente com os parâmetros *–p-placement-tool epa-gn*, *–p-hsp-method mp* e *–p-max-nsti 2*, em que o primeiro parâmetro indica qual ferramenta de posicionamento (epa-ng ou sepp) deve ser usada, para adicionar as sequências de consulta na árvore de referência. O segundo parâmetro indica o método de previsão de estado oculto a ser usado e, por fim, o terceiro parâmetro especifica o quão distante uma sequência precisa estar na filogenia de referência, antes de ser excluída, o limite padrão é 2.

O resultado desse processo são três arquivos artefatos (*ko_metagenome.qza, ec_metagenome.qza e pathway_abundance.qza*), que serão utilizados como arquivos de entrada para outros processos. Com esses arquivos disponíveis, é possível transformar o arquivo de saída *pathway_abundance.qza* em um arquivo de visualização, o qual, após análise, apresentará dados importantes para os próximos processos, como a contagem de recursos de cada amostra e as vias metabólicas observadas.

As previsões do metagenoma podem ser integradas em uma série de análises do QIIME 2, então é possível usar essas tabelas resultantes do PICRUST2, como arquivos de entrada para o QIIME 2 e analisar a diversidade de vias metabólicas pre sentes nas amostras.

3.6.8 Análise de diversidade em relação a profundidade das amostras

Para realizar uma análise mais detalhada, foram realizadas três baterias de análises: uma com todas as amostras, que já foi descrita anteriormente, e outras duas análises, que visavam repetir os processos, mas com grupos amostrais menores. Para essas análises, as amostras foram divididas e analisadas, separadamente, em dois grupos, referentes à profundidade em que cada amostra foi coletada (RP e S).

Para a realização dessa bateria de análises, primeiramente, foram criados os artefatos *table_RS.qza* e *table_S.qza*, com o auxílio do *plugin qiime feature-table filter-samples*. A entrada para esse processo incluiu o artefato *table.qza*, o arquivo *metadata* associado a ele (*mappingfile_wil_full.tsv*), e o parâmetro de filtragem com

base no qual o novo artefato deveria ser gerado (RS e S).

Com esses arquivos prontos, o próximo passo é a análise de alfa e beta diversidade, semelhante ao que foi feito na primeira bateria de análises. No entanto, em vez de utilizar o artefato *table.qza* e o arquivo de *metadata mappingfile_wil_full.tsv*, serão utilizados os artefatos *table_RS.qza* e *table_S.qza* e os arquivos de metadados *mappingfile_beta_rp_wil_full.tsv* e *mappingfile_beta_s_wil_full.tsv*, respectivamente, para as análises desses dois grupos de amostras.

3.7 Metodologia de comparação entre máquinas

Para a comparação entre as máquinas, foram escolhidos quatro processos dentro da *pipeline* para comporem a baterias de testes, sendo eles: o processo de importação das amostras, controle de qualidade, classificação taxonômica e, por fim, o processo de identificação de vias metabólicas.

A bateria de testes foi composta por dez repetições de cada processo, em cada uma das máquinas analisadas. Esse número de repetições foi escolhido para verificar o tempo médio, que cada máquina demandou para concluir cada um dos processos. O tempo gasto foi mensurado em minutos, graças ao parâmetro *time*, adicionado ao comando de execução de cada processo. Ao fim de cada repetição, o tempo de execução de cada máquina foi anotado em uma tabela, contendo a identificação da máquina, o processo executado e o valor gasto para realização do teste.

4 RESULTADOS E DISCUSSÃO

Nessa sessão, serão apresentados os resultados obtidos através do presente trabalho e as discussões pertinentes, a respeito dos resultados.

4.1 Configuração do ambiente de análise

Como primeiro resultado, temos a criação de máquinas virtuais, seguindo a parametrização de recursos, descrita na metodologia. Essas máquinas foram utilizadas como base para a configuração do ambiente de análise de bioinformática, baseado na ferramenta QIIME 2.

O que diferencia cada uma das máquinas, utilizadas neste estudo, é a quantidade de recursos de memória e processamento disponíveis em cada uma delas. Uma vez que as máquinas sejam criadas e devidamente configuradas com ambiente Conda e a ferramenta QIIME devidamente instalada, o ambiente, ao ser iniciado, está pronto para ser utilizado.

4.2 Análise de desempenho entre as máquinas

Visando atender o objetivo específico que pretende validar as configurações de máquinas analisadas, com o intuito de identificar seu desempenho e possíveis gargalos em alguns processos da *pipeline* utilizada, foram criadas, além da máquina principal, outras quatro máquinas teste, com o objetivo de identificar os possíveis gargalos que surgiriam durante a execução da pipeline. Essas máquinas foram configuradas, de forma que cada uma possuísse o dobro da quantidade de memória RAM e 50% a mais de poder de processamento do que a maquina anterior, com exceção da máquina 02, que apresenta o dobro de poder de processamento e de memória RAM que a máquina 01.

Ao final de cada bateria de teste, foi calculado o tempo médio necessário para execução de cada um dos processos analisados, vale ressaltar que todos os resultados foram expressados em uma escala de minutos. A Tabela 3, demonstra a relação entre o tempo médio demandado por cada uma das máquinas para o primeiro processo analisado (*Tools Import*).

Como observado na Tabela 3, a máquina 01, que possui a menor quantidade de recursos, apresenta a maior média de tempo gasto para realizar o processo analisado, seguida pela máquina 02, que possui o dobro de recursos de memória e processamento em comparação a ela. Já as máquinas 03 e 04, além de apresentarem resultados idênticos nessa bateria de testes, também apresentaram os menores tempos médios, necessários para a execução do processo de importação de dados.

Tabela 3 – Relação entre os tempos demandados no processo *Tools Import*

Máquina	Concluído	Tempo Médio (min)	Desvio Padrão
Máquina 02	SIM	5,882	1,603
Máquina 03	SIM	4,430	0,351
Máquina 04	SIM	4,480	0,310
Máquina 05	SIM	5,526	1,133

Fonte: Elaborado pelo autor, 2024.

No entanto, o que mais chama a atenção é o fato da máquina 05 apresentar a terceira maior média necessária para realizar o processo analisado.

Como o comando utilizado para a execução do processo não define a quantidade de recursos a serem utilizados para o mesmo, por padrão, a ferramenta aloca o máximo de recursos disponíveis para execução do processo. Contudo percebe-se, ao analisar o resultado da bateria de testes, que o *plugin Tools Import* não faz o melhor aproveitamento dos recursos disponíveis, não explorando ao máximo os recursos para uma melhor e mais rápida execução do processo de importação de dados.

Do mesmo modo, percebe-se que a máquina 04, mesmo possuindo uma maior quantidade de recursos em comparação à máquina 03, apresenta o mesmo desempenho que esta, o que implica em uma utilização muito parecida dos recursos disponíveis para realização do processo de importação.

Outro fato interessante, observado nesta bateria de testes, é que a máquina 05 apresenta o segundo maior desvio padrão entre as máquinas observadas, sendo precedida pela máquina 02, que possui um quarto dos recursos disponíveis se comparada com a máquina 05. Esse desvio padrão maior indica que os tempos de execução do *plugin Tools Import* são mais variáveis, ou seja, há uma maior discrepância nos tempos registrados. Isso sugere que o desempenho dessas máquinas é menos consistente, com tempos de execução que podem variar, significativamente, de uma execução para outra.

Por outro lado, as máquinas 03 e 04 destacam-se novamente, apresentando o menor desvio padrão, indicando que os tempos de execução do *plugin* analisado são menos variáveis, ou seja, diferente da máquina 05, a discrepância entre os tempos observados é menor. Sugerindo, então, um desempenho mais consistente dessas duas máquinas.

Por fim, e não menos importante, a máquina 01 apresentou o terceiro menor desvio padrão, demonstrando que, mesmo com a menor disponibilidade de recursos, ainda assim, apresentou um desempenho mais consistente do que máquinas mais poderosas, como as máquinas 02 e 05.

Esse resultado preliminar mostra que a máquina 04, a máquina principal dos testes, cujas especificações foram escolhidas com base em fóruns da ferramenta QIIME 2, correspondeu às expectativas, apresentando um resultado satisfatório no teste analisado.

Para a segunda bateria de testes, foram observados os resultados referentes ao processo de filtragem (dada 2), como observado na Tabela 4.

Tabela 4 – Relação entre os tempos demandados no processo *Dada2*

Máquina	Concluído	Tempo Médio (min)	Desvio Padrão
Máquina 01	SIM	968,725	148,553
Máquina 02	SIM	471,413	43,413
Máquina 03	SIM	377,246	18,758
Máquina 04	SIM	323,316	28,849
Máquina 05	SIM	307,205	17,689

Fonte: Elaborado pelo autor, 2024.

Como demonstrado na Tabela 4, a máquina 01 apresenta a maior média necessária para a execução do processo de filtragem, como já esperado. Nota-se também que, à medida que a quantidade de recursos aumenta, o tempo necessário para a execução do processo diminui. Com um aumento de 100% nos recursos de processamento e memória RAM da máquina 02, em comparação com a Máquina 01, percebe-se que o tempo gasto no processo de filtragem reduziu para menos da metade, comparado-se à máquina 01.

No entanto, essa proporção não se mantém em relação às próximas máquinas, em que o tempo médio gasto para execução do processo diminui significativamente, em relação às máquinas 02, 03, 04 e 05. Como observado na Tabela 4, a diferença entre as médias de execução torna-se cada vez menor, à medida que as máquinas se tornam mais potentes. Um exemplo claro desse fato, é quando comparamos as máquinas 04 e 05, que apresentam uma diferença em seus tempos médios, inferior a vinte minutos.

Ao analisarmos o desvio padrão das cinco máquinas observadas, nesta sequência de testes, e compará-los ao desvio padrão da série de testes anterior, percebemos que as cinco máquinas apresentaram resultados diferentes nesses dois testes. Nesta série de testes, a máquina 05 apresentou o menor desvio padrão observado, diferente de seu desempenho anterior, em que apresentou o segundo maior desvio padrão computado.

Já a máquina 04 apresentou o terceiro menor desvio padrão, diferentemente do teste anterior, em que apresentou o menor desvio observado. A máquina

03, no entanto, apresentou o segundo menor desvio padrão, em ambos os testes realizados. As máquinas 01 e 02 apresentaram, no teste anterior, o terceiro e quinto maior desvio padrão observado, respectivamente. No teste atual, seus respectivos desvios inverteram-se, de modo que a máquina 01 apresentou o quinto maior desvio padrão, enquanto a máquina 02 apresentou o terceiro maior desvio padrão.

Para a terceira bateria de testes, foram observados os resultados referentes ao processo de Classificação Taxonômica (*Feature-Classifier*), como observado na Tabela 5.

Tabela 5 – Relação entre os tempos demandados no processo *Feature-Classifier*

Máquina	Concluído	Tempo Médio (min)	Desvio Padrão
Máquina 01	NÃO	2,830	0,356
Máquina 02	NÃO	1,687	1,266
Máquina 03	NÃO	1,259	0,086
Máquina 04	NÃO	6,768	9,599
Máquina 05	SIM	348,18	20,262

Fonte: Elaborado pelo autor, 2024.

Nesta bateria de testes, algo inesperado ocorreu em relação às expectativas sobre quais máquinas seriam capazes de realizar os processos analisados. Uma vez que apesar da máquina 04 ter sido configurada para ser capaz de realizar todos os testes propostos neste trabalho, ela não conseguiu concluir a série de testes com êxito. Assim como as máquinas 01, 02 e 03, devido à falta de recursos de memória RAM.

Como observado na Tabela 5, a máquina com maior disponibilidade de recursos (máquina 05), foi a única a concluir o *plugin* de classificação, demandando um tempo médio de quase seis horas de execução. Isso nos mostra a complexidade do processo e o quanto a disponibilidade de recursos é algo importante para as análises de bioinformática, uma vez que, sem a quantidade de recursos mínimos necessários para a execução dos processos, os mesmos não puderam ser realizados, inviabilizando a conclusão da pesquisa.

Para a quarta e última bateria de testes, foram observados os resultados referentes ao processo de Identificação de Vias Metabólicas através do (*plug Picrust 2*), como observado na Tabela 6.

Como esperado, nessa quarta bateria de testes, as máquinas 04 e 05 foram as únicas a concluírem com êxito a execução do *plugin Picrust 2*. No entanto, a Máquina 03 não concluiu a bateria de testes, por não possuir recursos de memória RAM suficiente para concluir a parte final do processo de identificação de vias meta-

Tabela 6 – Relação entre os tempos demandados no processo *Picrust 2*

Máquina	Concluído	Tempo Médio (min)	Desvio Padrão
Máquina 01	NÃO	54,252	0,910
Máquina 02	NÃO	54,295	0,617
Máquina 03	NÃO	103,38	2,778
Máquina 04	SIM	157,337	3,439
Máquina 05	SIM	173,067	13,107

Fonte: Elaborado pelo autor, 2024.

bólicas. Por isso, ela demandou tanto tempo para falhar no teste, uma vez que possuía recursos suficientes para a execução das primeiras etapas do processo.

As máquinas 01 e 02 apresentaram resultados idênticos, demandando o mesmo tempo médio para falharem. Mas o que mais chama atenção nessa bateria de testes é o fato da máquina 04 apresentar um resultado ligeiramente melhor que a máquina 05, o que nos leva a pensar que o *plugin Picrust 2*, assim como o *Tools Import*, não é capaz de aproveitar totalmente os recursos disponíveis ou, a partir de certo ponto, uma quantidade maior de recursos torna-se desnecessária, uma vez que a ferramenta não apresenta um melhor desempenho, após atingir um determinado nível de recursos.

Outra informação interessante, extraída deste teste, é que a máquina 04 apresentou um desvio padrão menor, em comparação ao desvio padrão da máquina 05. Como discutido anteriormente, um desvio padrão menor, indica uma maior consistência na execução do *plugin* analisado, o que sugere menor variação nos tempos necessários para a execução do processo. Isso significa que os resultados são mais previsíveis e estáveis, com menos discrepâncias em relação ao tempo de execução.

4.3 Comparação custo-benefício de máquinas virtuais no contexto de pesquisas universitárias

Para a comparação custo-benefício das máquinas, em um contexto universitário, as mesmas foram analisadas referentes a seu desempenho em cada processo observado anteriormente, nos quais as máquinas foram avaliadas somente se tivessem conseguido concluir, com êxito, as baterias de teste.

A Tabela 7 demonstra essa relação entre as máquinas e o custo médio do processo *Tools Import*, executado por cada uma delas.

Graças ao baixo custo de processamento das máquinas, para a execução do *plugin* de importação, como observado na Tabela 7, a escolha de qual máquina

Tabela 7 – Relação custo-benefício *Tools Import*

Máquinas	Tempo Médio Total (min)	Valor min (R$)	Custo do processo (R$)
Máquina 01	9,274	0,00146	0,0135
Máquina 02	5,882	0,00292	0,0172
Máquina 03	4,430	0,00583	0,0258
Máquina 04	4,480	0,0117	0,0523
Máquina 05	5,526	0,02333	0,1289

Fonte: Elaborado pelo autor, 2024.

escolher para realizar a importação dos dados, fica a critério do tempo gasto para executar o processo. No entanto, para esse processo em questão, a diferença de tempo demandado para a conclusão do *plugin* não é grande o suficiente para determinar a escolha de uma máquina mais potente. Uma vez que o maior tempo médio, observado para a execução da importação dos dados, é inferior a 10 minutos.

Ao analisar a Tabela 7, percebemos que todas as máquinas concluíram o processo de importação dos dados, apresentando um custo médio total inferior a 15 centavos. De modo geral, a máquina que apresenta o melhor desempenho, ao compararmos tempos de execução e custo de processamento, é a máquina 03, que combina o menor tempo gasto na execução do processo, com o terceiro menor custo de execução.

A Tabela 8 demonstra a relação entre as máquinas e o custo médio do processo Dada 2, executado por cada uma delas. Como observado na tabela 8, devido à grande quantidade de tempo demandado para a execução do processo de filtragem dos dados, percebemos que o custo de execução do processo nas máquinas, torna-se discrepante o suficiente para tendenciar a escolha de uma máquina ou outra.

Tabela 8 – Relação custo-benefício dada2

Máquinas	Tempo Médio Total (min)	Valor min (R$)	Custo do processo (R$)
Máquina 01	968,725	0,00146	1,413
Máquina 02	471,413	0,00292	1,375
Máquina 03	377,246	0,00583	2,201
Máquina 04	323,316	0,0117	3,772
Máquina 05	307,205	0,02333	7,168

Fonte: Elaborado pelo autor, 2024.

A Tabela 8, mostra que a máquina 02 apresenta o menor custo para a execução do *plugin* dada2, no entanto, apresenta o quarto maior tempo de execução, que corresponde a quase oito horas de processamento. O segundo menor custo de execução é apresentado pela máquina 01, que corresponde a um tempo médio superior a 16 horas.

Essas duas máquinas apresentam os menores valores médios a serem pagos para a execução do *plugin* dada2. No entanto, apresentam tempos médios para a conclusão do dada2, superiores a sete horas. Isso difere das outras máquinas, que apresentam tempos médios para a execução do *plugin* analisado inferiores a sete horas.

A máquina 03 apresenta tanto o valor quanto o tempo mediano das cinco máquinas observadas, demandando pouco mais de seis horas para completar a execução do *plugin* dada2. Ao observarmos a máquina 04, percebemos que ela apresenta o segundo maior custo e o segundo menor tempo para a execução do dada2, assim como a máquina 05, que apresenta o maior custo e o menor tempo de execução para a conclusão do processo.

Ao compararmos as cinco máquinas, percebemos que as máquinas 01 e 02, mesmo apresentando os menores custos, demandam os maiores tempos para a conclusão do processo, o que acaba por tornar o desenrolar da pesquisa mais lento. Isso difere das máquinas restantes, principalmente das máquinas 04 e 05, que proporcionam um custo maior para a execução do processo, ao mesmo tempo que reduz -se o tempo necessário para conclusão do mesmo.

Desse modo, a escolha da melhor máquina para a execução do processo observado torna-se subjetiva ao pesquisador, que decidirá entre um maior custo de execução e um menor tempo necessário, ou o inverso, baseado no cenário vivenciado por ele.

Já a terceira bateria de teste não apresentou dados suficientes para a comparação custo-beneficio entre as máquinas, devido ao fato da máquina 05 ter sido a única máquina a concluir, com êxito, o *plugin* de classificação.

Por fim, a quarta bateria de teste apresentou dois resultados, como apresentado na Tabela 9, que demonstra a relação entre as máquinas e o custo médio do processo *Picrust 2*, executado por cada uma delas.

Tabela 9 – Relação custo-benefício *Picrust2*

Máquinas	Tempo Médio Total (min)	Valor min (R$)	Custo do processo (R$)
Máquina 04	157,337	0,0117	1,836
Máquina 05	173,067	0,02333	4,038

Fonte: Elaborado pelo autor, 2024.

Analisando o custo de processamento e o tempo demandado, por cada uma das máquinas, para a conclusão do *plugin* Picrust 2, percebe-se que a máquina 04 apresenta o melhor custo-benefício entre as duas máquinas, devido ao fato de apresentar o menor custo e o menor tempo necessário para a execução do processo.

4.4 Validação do ambiente de análise

Nessa seção, serão apresentados os resultados e discussão a respeito da validação do ambiente de análise.

4.4.1 Análises de bioinformática

A importação das 96 bibliotecas, referentes às 48 amostras utilizadas nesse estudo, geraram um total de 3.844.404 sequências pareadas. Após o processo de filtragem e rarefação, baseados nos dados provenientes dos gráficos de qualidade de sequenciamento de sequências diretas e reversas, restaram 1.206.859 sequências pareadas, que equivalem a 31,39% do conjunto de dados inicial, com uma rarefação de 28.421 sequências.

4.4.2 Beta diversidade

Ao executar todas as etapas da *pipeline*, obteve-se os resultados do teste de PERMANOVA, referentes às comparações entre os oito grupos amostrais estudados. Com esses resultados em mãos, é possível observar as características das comunidades microbianas presentes no solo, em diferentes pontos, no decorrer dos ciclos de queimadas do PNSC. .

Analisando os resultados dos testes de comparação, executados pelo QIIME 2, podemos perceber a diferenciação das comunidades microbianas em cada área, baseado no *p-value*, como demonstrado na Tabela 10.

Tabela 10 – Matriz de comparação entre áreas levando em consideração o índice de similaridade entre as amostras

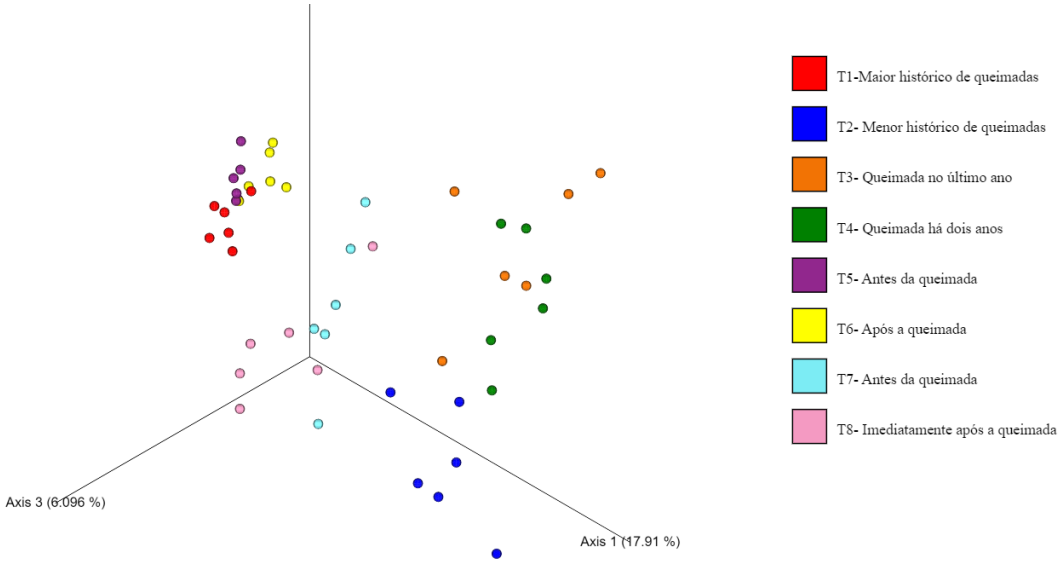
Áreas	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**
T1		*0.004*	*0.001*	*0.002*	*0.184*	*0.079*	*0.003*	*0.003*
T2			*0.002*	*0.002*	*0.003*	*0.002*	*0.029*	*0.041*
T3				*0.218*	*0.003*	*0.003*	*0.100*	*0.010*
T4					*0.003*	*0.005*	*0.026*	*0.003*
T5						*0.321*	*0.003*	*0.003*
T6							*0.005*	*0.003*
T7								*0.474*

Fonte: Elaborado pelo autor, 2024

Para melhor compreensão dos dados apresentados na Tabela 10, a Figura

6 demonstra o mesmo resultado, na forma de um gráfico tridimensional, o qual foi gerado a partir dos resultados provenientes da execução do *plugin qiime diversity core-metrics-phylogenetic*.

Figura 6 – Gráfico tridimensional da disposição das comunidades microbianas



Fonte: Elaborado pelo autor, 2024

Para fins de comparação de beta diversidade, analisaremos as amostras T1 (com a área de maior incidência de queimadas), T2 (Menor histórico de queimadas), T3 (a área queimada no último ano) e T4 (área queimada há dois anos) e adotarems a amostra T2 como grupo controle para as comparações posteriores.

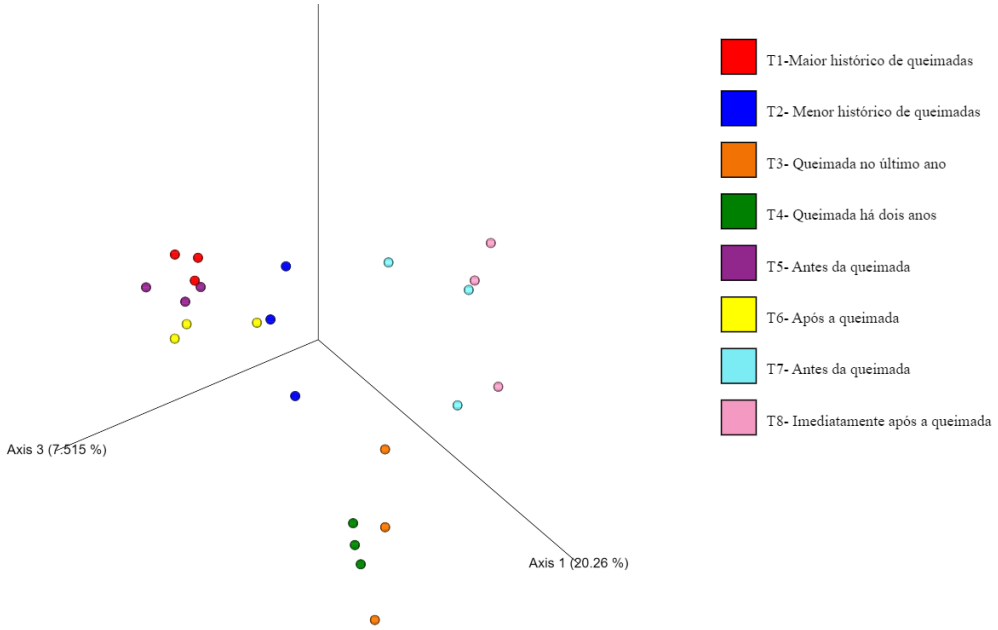
Como observado na Tabela 10, quando comparamos o grupo controle T2, com a áreas T1, T3 e T4, percebemos que todas as áreas comparadas, ao grupo controle, apresentam comunidades distintas a ele. Isso reforça a hipótese que a cada nova ação do fogo no solo, modifica as comunidades microbiológicas presentes nesse habitat.

No entanto, essa alteração nas comunidades não se dá de forma imediata, uma vez que, de acordo com a Tabela 10, as áreas T3 e T4 (queimada a um ano e a dois anos, respectivamente), apresentam uma probabilidade de serem semelhantes, superior a 20%. Isso demonstra que, apesar de ter a capacidade de alterar a composição da microbiota do solo, a ação do fogo não provoca mudanças imediatas, reforçando a ideia de que os microorganismos presentes nesse habitat possuem características evolutivas, que lhes conferem certa resistência aos efeitos das chamas.

Quando se compara as áreas analisadas em uma mesma profundidade de amostra, observamos uma diferenciação mais nítida entre as áreas, como ilustrado nas Figuras 7 e 8. A figura 7 demonstra os resultados referentes a região da rizosfera

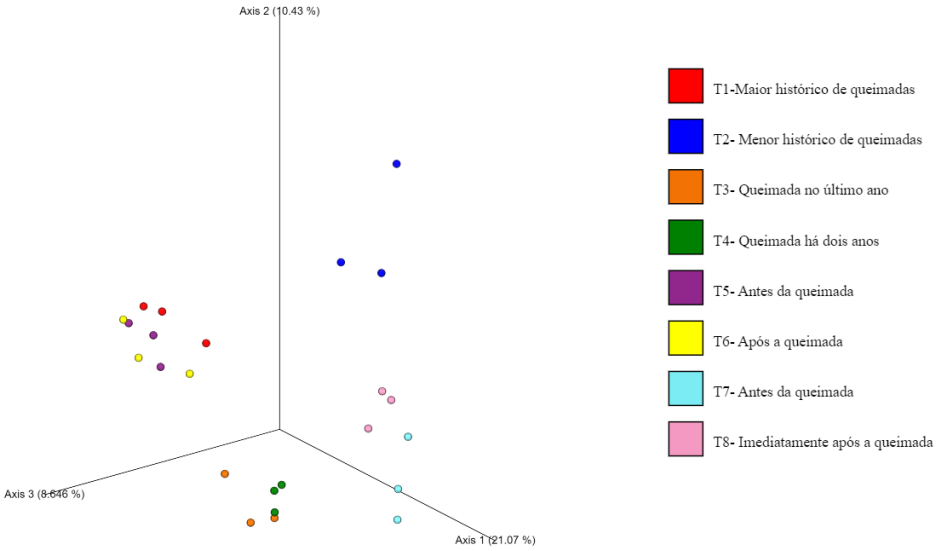
da planta, enquanto a Figura 8 apresenta os resultados das comparações entre as amostras de solo.

Figura 7 – Gráfico tridimensional da disposição das comunidades microbianas da região da rizosfera da planta.



Fonte: Elaborado pelo autor, 2024.

Figura 8 – Gráfico tridimensional da disposição das comunidades microbianas presentes no solo.



Fonte: Elaborado pelo autor, 2024.

Analisando a Figura 8, referente ao solo, nota-se que o grupo controle está distante das demais amostras, mantendo-se isolado, apesar de apresentar uma certa

divisão interna entre seus constituintes. Já em relação às outras três amostras analisadas, observa-se que a amostra T1 se mantém distante das demais, mesmo apresentando uma certa discrepância em relação aos seus constituintes, ao contrário das amostras T3 e T4, que exibem certa proximidade entre si.

Vale ressaltar que a amostra T3 apresenta um certo distanciamento entre seus três constituintes, ao contrário da amostra T4. Como demonstrado na Figura 8, a amostra T4 apresenta dois de seus constituintes bem próximos e apenas um constituinte discrepante, o qual se aproxima de dois grupos constituintes da amostra T3. Evidenciando a similaridade entre essas duas amostras, que sofreram com o efeito do fogo, em dois diferentes momentos.

Já quando se analisa a Figura 7, referente à região da rizosfera da planta, percebemos que o grupo controle apresenta a mesma conformação do grupo controle, referente ao solo. Mantendo-se isolado das demais amostras analisadas, ao mesmo tempo em que apresenta um certo distanciamento entre seus grupos constituintes.

Ao dar atenção aos demais grupos observados, nota-se algumas diferenças em relação à análise anterior, uma vez que, na análise das comunidades presentes no solo a amostra T1, mesmo permanecendo isolada das demais, apresentou um certo distanciamento entre seus constituintes. Já em relação à microbiota presente na região da rizosfera da planta, esse distanciamento entre os constituintes da amostra T1 diminui, diferentemente da amostra T3, na qual percebe-se um maior distanciamento entre os constituintes, quando comparado à análise anterior.

Esse maior distanciamento entre seus constituintes demonstra que, mesmo pertencendo à mesma amostra observada, os grupos constituintes da amostra T3 tornaram-se mais distantes, evolutivamente, uns dos outros, em comparação com a análise anterior. Da mesma forma, os grupos constituintes da amostra T4 também apresentaram um maior distanciamento evolutivo, em relação à análise anterior.

4.4.3 Análise de Beta diversidade entre a mesma área antes e depois da queima

Para fins de comparação de beta diversidade em uma mesma área, foram analisadas dois pontos dentro das áreas a serem manejadas no MIF, sendo esses pontos representados pelos grupos amostrais T5 e T7. Ambos representam amostras microbiológicas do solo, antes do manejo do fogo.

Como observado na Tabela 10, quando compara-se o grupo amostral T5 (antes da queimada - MIF) com o grupo amostral T6 (após a queimada, aproximadamente 40 minutos, a favor do vento), nota-se que, mesmo após a ação do fogo, a probabilidade dos grupos de microorganismos, presentes no solo, serem semelhantes ao grupo controle é superior a 30%. Demonstrando que, mesmo a ação das chamas sendo devastadora à microbiota do solo, apresenta certa resistência a esse fenômeno.

Ao observarmos a Tabela 8, que reflete a distribuição dos grupos amostrais analisados, referentes à microbiota do solo, percebemos que, apesar das alterações causadas pelas chamas, o grupo T6 apresenta uma certa proximidade ao grupo controle T5.

Como observado, dois dos subgrupos constituintes da amostra T6 estão afastados dos subgrupos presentes no grupo controle T5, enquanto o subgrupo restante está localizado muito próximo à amostra T5, correspondendo a mais de 30% de chances de a amostra T6 ser semelhante à amostra T5, conforme indicado na tabela 10.

Ao observarmos a imagem 7, referente à análise da região da rizosfera, notamos uma conformação diferente entre as amostras analisadas. O grupo controle T5 apresenta dois de seus subgrupos constituintes bem próximos, enquanto o terceiro subgrupo mantém-se isolado dos demais subgrupos da amostra T5 e das demais amostras observadas.

Já a amostra T6 apresenta uma conformação similar à amostra T5. No entanto, o que difere a amostra T6 do grupo controle é que os subgrupos constituintes da amostra T6 estão divididos em dois grupos: o primeiro é composto por dois constituintes da amostra T6, dos quais um se aproxima do grupo controle T5, enquanto o outro se mantém afastado; o segundo grupo é formado pelo constituinte restante, que se distancia de todos os subgrupos ao seu redor.

Algo interessante a se observar da comparação entre as amostras T5 e T6, é o fato de que o fogo modificou as comunidades microbiológicas do solo de formas distintas. Como observado na comparação entre as Figuras 7 e 8, percebemos que a microbiota presente na região da rizosfera sofreu uma alteração maior do que a microbiota do solo.

Essa alteração indica que a maior exposição às chamas, sofrida pela microbiota presente na rizosfera da planta, justifica essa maior alteração em sua comunidade ou que os microrganismos presentes, no solo da amostra T5, possuem maior resistência aos efeitos do fogo. Essa maior resistência às queimadas dá-se, provavelmente, através do processo evolutivo, o qual seleciona os indivíduos com as características mais adaptadas ao ambiente, para sobreviverem e reproduzirem-se, transmitindo essas características favoráveis para as próximas gerações.

Ao analisar as amostras T7 (antes da queimada - MIF) e T8 (após a queima, contra o vento) referentes à Figura 8, pode-se perceber que ambas as amostras apresentam a mesma conformação, em que dois de seus constituintes estão próximos, enquanto o terceiro constituinte encontra-se afastado dos demais. No entanto, o que mais chama a atenção em relação a essas duas amostras é o fato que a amostra T8 se distanciou muito do grupo controle T7, perdendo as características que remetiam ao grupo controle, caracterizando, assim, o grupo T8 como um grupo distinto do grupo

T7.

Ao voltar nossa atenção para a Figura 7, notamos que as amostras T7 e T8 apresentam a mesma conformação, com dois de seus constituintes localizados próximos, enquanto o terceiro constituinte de cada amostra distancia-se dos demais. No entanto, diferente do analisado na Figura 8, as amostras T7 e T8 apresentam proximidade, se comparadas. Essa proximidade mostra que a comunidade microbiológica, presente na região da rizosfera, sofreu menor influência das chamas que do que a microbiota presente no solo.

Comparando-se as amostras T7 e T8, em diferentes profundidade de amostras, percebe-se que cada uma das regiões do solo sofreram alterações distintas, em relação à ação das chamas. No entanto, ao se analisar a Tabela 10, nota-se que, graças à maior resistência aos efeitos do fogo presente na microbiota da região da rizosfera, as duas amostras apresentam uma probabilidade de serem parecidas, superior a 45%, o que evidencia que, mesmo a microbiota do solo sofrendo uma mudança significativa, a ponto de caracterizá-la como um grupo distinto do grupo controle, a microbiota da região da rizosfera pouco mudou, conferindo à amostra amostra T8 essa alta probabilidade de ser similar à comunidade microbiológica presente no grupo controle.

Outra observação interessante a se fazer, a respeito da comparação entre as amostras T7 e T8, é como a sucessão ecológica ocorre após a ação do fogo. Grupos de microorganismos presentes na superfície do solo, que apresentavam populações estáticas, devido a fatores limitantes como pH, alimento, espaço, entre outros, podem sofrer mudanças expressivas em suas populações, após um fenômeno transformador como as queimadas, como já discutido por Evangelista *et al.* (2013), em seu trabalho.

Como demonstrado na Figura 8, após o efeito das chamas, a comunidade presente no solo emergiu em uma conformação totalmente distinta do grupo controle, mostrando como o fogo possui um efeito modificador não apenas no ambiente, mas também nas comunidades que vivem nos habitats atingidos por ele.

Um paralelo interessante a se fazer entre essas duas comparações (T5 com T6 e T7 com T8) é o fato de que o fogo modificou a microbiologia do solo de formas distintas nessas duas áreas. Como apresentado anteriormente, a microbiota presente na rizosfera, das amostras T5 e T6, sofreu uma modificação maior, devido aos impactos do fogo, do que a microbiota do solo.

Ao contrário da comparação entre as amostras T7 e T8, em que a microbiota presente no solo sofreu uma alteração maior do a microbiota presente na rizosfera. Isso indica que os microorganismos, em diferentes locais do PNSC, evoluem de formas distintas, adaptando-se da melhor forma possível ao ambiente em que vivem.

4.5 Conclusão

Nessa sessão, serão apresentadas as conclusões feitas a partir dos testes realizados no presente trabalho.

4.5.1 Analise de microbioma

Ao analisar os resultados apresentados, é possível notar que o fogo altera todo o ambiente por onde passa, seja essa alteração em uma escala macro (alteração perceptível a olho nu) ou micro (alteração que não é perceptível a olho nu). Como observado nas comparações feitas anteriormente, o fogo, como agente transformador, modifica a microbiologia do solo de formas distintas, alterando a composição das comunidades microbiológicas presentes no solo, em diferentes profundidades: região da rizosfera (0-10 cm) e solo (10-20 cm).

O poder modificador das chamas é inegável. No entanto, a resistência e a adaptabilidade dos microrganismos presentes no solo, como um todo, a esse fenômeno é admirável. Como observado na comparação de beta diversidade, os microrganismos evoluem e adaptam-se para sobreviver em seus habitats, após cada evento transformador, o que confere à microbiota do solo uma nova composição de suas comunidades microbiológicas.

Essas comunidades sobreviventes evoluem, tornando-se cada vez mais resistentes aos efeitos das chamas. Essa evolução, como demonstrado nas análises, não ocorre da mesma forma em comunidades presentes, em um mesmo local atingido pelas queimadas. Como observado nas comparações realizadas, comunidades de um mesmo local, com profundidades de amostras distintas, apresentaram comportamentos diferentes em relação ao fogo.

Na comparação entre as amostras T5 e T6, observa-se uma maior alteração na comunidade presente, na região da rizosfera da planta após a queima, enquanto a comparação entre as amostras T7 e T8 revelou uma maior alteração na microbiota presente no solo.

De modo geral, as queimadas, como agentes transformadores, modificam a microbiologia do solo de formas distintas, seja essa mudança em relação a um mesmo local, dentro de uma determinada área ou em vários locais dentro de um mesmo espaço. Essa alteração na microbiota do solo, como observado, não ocorre de forma radical, as novas comunidades presentes no solo ainda preservam características das comunidades anteriores à queimada.

4.5.2 Comparação entre máquinas

Ao final de todas as baterias de testes, foi possível concluir que, mesmo máquinas modestas, em relação ao poder de processamento e memória RAM, podem conseguir executar alguns processos relacionados à bioinformática. No entanto, para executar esses processos que não exigem uma grande quantidade de recursos computacionais, uma quantidade razoável de tempo é demandada.

Como observado nos testes propostos no presente trabalho, as máquinas mais modestas foram capazes de realizar processos simples da bioinformática, como a importação e filtragem dos dados. Vale ressaltar que essas máquinas foram capazes de realizar tais processos com um conjunto de dados, que inclui 96 bibliotecas e 3.844.404 sequências pareadas após a filtragem. Conjuntos de dados, superiores ao utilizado nesse trabalho, podem não apresentar o mesmo resultado, se expostos ao mesmo experimento.

Mas como já era esperado, a limitação por falta de recursos computacionais tornou-se evidente muito rápido, excluindo as máquinas mais modestas dos testes realizados. Essa limitação, também excluiu um dos testes propostos à máquina 04, que foi apontada pelos estudos, como sendo capaz de realizar todos os testes propostos, o que demonstra a complexidade dos processos envolvidos nos estudos de bioinformática e como essa área de pesquisa demanda equipamentos e computadores potentes para poder concluir suas pesquisas.

Graças aos testes realizados, é possível afirmar que, para o estudo de microbiomas, utilizando a ferramenta QIIME 2, com um banco de dados semelhante ao utilizado neste trabalho, é necessário utilizar uma máquina com mais de 32 GB de memória RAM. Nos experimentos realizados, a máquina 05, que possui 12 vCPUs, 64 GB de RAM e 50 GB de armazenamento, demonstrou ser capaz de realizar todos os testes propostos com sucesso.

4.5.3 Relação custo-beneficio

Ao analisar o custo de execução de cada processo, em cada uma das máquinas, percebemos que apenas a máquina 05 foi capaz de realizar todos os testes, sendo, portanto, a escolha óbvia. No entanto, para fins de pesquisas universitárias, que enfrentam limitações orçamentárias ou até mesmo a ausência de financiadores, uma alternativa viável seria a combinação de diferentes máquinas para realizar os processos, a fim de reduzir os custos de execução.

Conforme observado nas baterias de testes realizadas, outras máquinas apresentaram desempenhos interessantes em comparação à máquina 05. Por exemplo, a máquina 04, que, embora tenha concluído 3/4 dos testes com êxito, possui

um custo de processamento menor que a máquina 05 e tempos de processamento semelhantes ou até menores, como observado na quarta bateria de testes.

Dessa forma, uma alternativa mais econômica para a realização de pesquisas no âmbito da bioinformática, utilizando máquinas virtuais contratadas, seria a combinação de diferentes máquinas para a execução dos processos.

4.6 Trabalhos futuros

Considerando os resultados apresentados neste trabalho e visando melhorias para trabalhos futuros, seria interessante ampliar o tempo de observação dos ciclos de queimadas do PNSC. Com um maior volume de dados, seria possível monitorar como a microbiota do solo está sendo modificada, ao longo do tempo.

Voltando o olhar para os resultados computacionais, seria interessante expor as máquinas criadas para este trabalho a bancos de dados maiores, com o intuito de identificar os limites que cada máquina não conseguiria ultrapassar, especialmente a máquina 05, que apresenta a maior disponibilidade de recursos de processamento e memória RAM.

REFERÊNCIAS

ABREU, K. C. de *et al.* **GRANDES FELINOS E O FOGO NO PARQUE NACIONAL DE ILHA GRANDE, BRASIL LaTeX**. 2004. Disponível em: https://revistas.ufpr.br/floresta/article/view/2389/1998. Acesso em: 28/04/2023.

ABSAM. **Absam 2024**. 2024. Disponível em: https://absam.io/.

AGUIAR, E. R. G. R. **GalaxyX**: – Software Gerador de Interfaces XML de Programas de Bioinformática para Integração no Framework Galaxy LaTeX. 2011. Disponível em: http://hdl.handle.net/1843/BUOS-94NLTW. Acesso em: 23/08/2023.

ALFONSO, E. T.; LEYVA, Á.; HERNÁNDEZ, A. Microorganismos benéficos como biofertilizantes eficientes para el cultivo del tomate (Lycopersicum esculentum, mill). **Revista Colombiana de Biotecnología**, v. 7, n. 2, p. 47–54, 2005.

ANACONDA, INC. **Miniconda**. 2024. Disponível em: https://docs.conda.io/projects/miniconda/en/latest/.

ANDREOTE, F. D. **Estrutura e função do microbioma de solos brasileiros**. 2014. Escola Superior de Agricultura Luiz de Queiroz. Disponível em: https://doi.org/10.1111/1574-6976.12035.

ARAÚJO, N. D. de *et al.* **A ERA DA BIOINFORMÁTICA: SEU POTENCIAL E SUAS IMPLICAÇÕES PARA AS CIÊNCIAS DA SAÚDE LaTeX**. 2008. Disponível em: https://biblat.unam.mx/hevila/Estudosdebiologia/2008/vol30/no70-72/16.pdf. Acesso em: 03/08/2023.

ATTWOOD, T. K. *et al.* **A global perspective on evolving bioinformatics and data science training needs LaTeX**. 2019. Disponível em: https://doi.org/10.1093/bib/bbx100. Acesso em: 06/06/2023.

BALD, D. R. *et al.* Microbiota do solo: a diversidade invisível e a sua importância. **Bio Diverso**, v. 1, n. 1, 2021. Disponível em: https://seer.ufrgs.br/index.php/biodiverso/article/view/120742.

BAYAT, A. **Science, medicine, and the future Bioinformatics LaTeX**. 2002.

BELL, T. *et al.* The contribution of species richness and composition to bacterial services. **Nature**, Nature Publishing Group, v. 436, p. 1157–1160, 2005.

BENÍCIO, R. M. A. *et al.* Um refúgio de Mata Úmida no interior do Nordeste brasileiro: estrutura e diversidades alfa e beta, 2023. Disponível em: https://doi.org/10.5902/1980509869097.

BISHOP, Ö. T. *et al.* **Bioinformatics Education**: -Perspectives and Challenges out of Africa LaTeX. 2015. Disponível em: https://doi.org/10.1093/bib/bbu022. Acesso em: 31/07/2023.

BOKULICH, N. A. *et al.* Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. **BMC Part of Springer Nature**, 2018. Disponível em: https://doi.org/10.1186/s40168-018-0470-z.

BOLYEN, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2 LaTeX. **Nature Biotechnology**, 2019. Disponível em: https://sci-hub.se/https://www.nature.com/articles/s41587-019-0209-9.

BOON, E. *et al.* Interactions in the microbiome: communities of organisms and communities of genes, 2014. Disponível em: https://doi.org/10.1111/1574-6976.12035.

BUYYA, R.; BROBERG, J.; GOSCINSKI, A. M. **Cloud Computing: Principles and Paradigms LaTeX**. 2011. Disponível em: http://books.google.com.br/books?id=S1NvRRd77rQC. Acesso em: 14/08/2023.

CAMPOS, H. d. S. Papel do genoma e do microbioma na patogenia e abordagem terapêutica da asma, 2021. Disponível em: 10.5935/2526-5393.20210039.

CARDOSO, E. J. B. N.; ANDREOTE, F. D. **Microbiologia do Solo**. 2016.

CARVALHO, F. A.; FELFILI, J. M. Diversidade alfa e beta como critérios para escolha de áreas prioritárias para conservação: um ensaio com as florestas estacionais deciduais sobre afloramentos calcários no brasil central. **SEB-Sociedade de Ecologia do Brasil**, 2007. Disponível em: https://www.seb-ecologia.org.br/revistas/indexar/anais/viiiceb/pdf/606.pdf.

CARVALHO, M.; SILVA, D. Sequenciamento de DNA de Nova Geração e suas Aplicações na Genômica de Plantas. **Revista de Ciência Rural**, Santa Maria, 2010. Disponível em: https://repositorio.unesp.br/bitstream/handle/11449/29442/S0103-84782010000300040.pdf?sequence=1%5C&isAllowed=y.

CATTLEY, S.; ARTHUR, J. W. **BioManager: the use of a bioinformatics web application as a teaching tool in undergraduate bioinformatics training LaTeX**. 2007. Disponível em: https://doi.org/10.1093/bib/bbm039. Acesso em: 03/08/2023.

CLC, G. W. **Manual for CLC Genomics Workbench 23.0.4 Windows, macOS and Linux July 20, 2023 LaTeX**. 2023. Disponível em: https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/User%5C_Manual.pdf.

COCK, P. J. A.; ANTAO, T. *et al.* **Biopython: freely available Python tools for computational molecular biology and bioinformatics LaTeX**. 2009. Disponível em: https://doi.org/10.1093/bioinformatics/btp163. Acesso em: 22/08/2023.

COCK, P. J. A.; FIELDS, C. J. *et al.* The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. **Nucleic Acids Research**, v. 38, n. 6, p. 1767–1771, 2009. ISSN 0305-1048. DOI: 10.1093/nar/gkp1137. eprint: https://academic.oup.com/nar/article-pdf/38/6/1767/16769834/gkp1137.pdf. Disponível em: https://doi.org/10.1093/nar/gkp1137.

COMMUNITY, T. G. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. **Nucleic Acids Res.**, Oxford University Press, v. 50, W1, p. W345–W351, 2022. DOI: 10.1093/nar/gkac247. Disponível em: https://doi.org/10.1093/nar/gkac247.

CORTES, V. S. M. P. **INTERFACES ACESSÍVEIS PARA USUÁRIOS QUE SÃO CEGOS: UM MÉTODO DE REPRESENTAÇÃO TÁTIL E SONORA DA INFORMAÇÃO EM BIOINFORMÁTICA ESTRUTURAL LaTeX**. 2019. Disponível em: https://repositorio.pucrs.br/dspace/bitstream/10923/18745/1/000500813-Texto%5C%2Bcompleto-0.pdf. Acesso em: 31/07/2023.

COSTA, C. N. **O uso de ferramenta de data analytics para o desenvolvimento de metodologia visando o mapeamento genético de doenças raras – Estudo de caso no rastreio do gene causador da cistinose**. 2020. Universidade Federal Fluminense, Rio de Janeiro. Disponível em: https://app.uff.br/riuff/handle/1/16381.

COTTA, S. R. O solo como ambiente para a vida microbiana. In: CARDOSO, E. J. B. N.; ANDREOTE., F. D. (Ed.). **MICROBIOLOGIA DO SOLO**. Brasília, DF: Embrapa, 2016. 10.11606/9788586481567. cap. 2, p. 23–35.

DAMASCENO, J. G. **Desenvolmento de um software para armazenamento e exploração de dados genômicos LaTeX**. 2023. Disponível em: https://www.teses.usp.br/teses/disponiveis/5/5144/tde-20012023-170800/en.php. Acesso em: 22/08/2023.

DIAS, N. S. T. **Rastreamento molecular por pirosequenciamento de mutações relacionadas à resistência a Lamivudina em subpopulações de HBV e HIV isoladas de pacientes coinfectados**. 2014. Trabalho de Conclusão de Curso – Universidade Federal Fluminense, Instituto Biomédico, Biomedicina. Disponível em: https://app.uff.br/riuff/handle/1/5200.

DOUGLAS, G. **Picrust2 Installation**. 2024. Disponível em: https://github.com/picrust/picrust2/wiki/Installation.

EDMUNDO, I. d. S. B. **Identificação de padrões de ocorrência de incêndios no Parque Nacional da Serra da Canastra a partir de mineração de dados**. 2014. Universidade Federal de Minas Gerais. Disponível em: http://hdl.handle.net/1843/IGCM-9UXHEN.

EISENSTEIN, M. Oxford Anúncio da Nanopore agita o setor de sequenciamento. **Nat Biotechnol**, v. 30, p. 295–296, 2012. DOI: 10.1038/nbt0412-295.

ESTAKI, M. *et al.* QIIME 2 Enables Comprehensive End-to-End Analysis of Diverse Microbiome Data and Comparative Studies with Publicly Available Data. **current protocols**, 2020. Disponível em: https://doi.org/10.1002/cpbi.100.

EVANGELISTA, C. R. *et al.* Atributos microbiológicos do solo na cultura de cana-de-açúcar sob manejo orgânico e convencional. **Semina: Ciências Agrárias**, Londrina, v. 34, n. 4, p. 1549–1562, 2013.

FERRARI, R. C.; GASTALDI, V. D. **Precisamos falar sobre a Bioinformática LaTeX**. Edição: Cláudia Maria Furlan. 2018. p. 246–261. Disponível em: https://scholar.google.com.br/scholar?oi=bibs%5C&cluster=5334918428068968362%5C&btnI=1%5C&hl=pt-BR. Acesso em: 08/08/2023.

FERRERO, A. P. d. S. Análise comparativa da estrutura do bacterioplâncton e biofilmes em lagoas costeiras rasas do RS. In. Disponível em: https://www.lume.ufrgs.br/bitstream/handle/10183/245018/Resumo%5C_75148.pdf?sequence=1.

FILGUEIRAS, A. W. N. *et al.* **LINUX EDUCACIONAL – UMA FERRAMENTA A SER DESCOBERTA NA REDE PÚBLICALaTeX**. 2016. Disponível em: http://www.periodicos.letras.ufmg.br/index.php/anais%5C_linguagem%5C_tecnologia/article/view/10509/9425. Acesso em: 21/08/2023.

FLÓRIA-SANTOS, M.; NASCIMENTO, L. C. **Perspectivas históricas do Projeto Genoma e a evolução da enfermagem LaTeX**. 2006. Disponível em: https://doi.org/10.1590/S0034-71672006000300020. Acesso em: 03/08/2023.

FRANCO, J. L. d. A. O conceito de biodiversidade e a história da biologia da conservação: da preservação da wilderness à conservação da biodiversidade, 2013. Disponível em: https://doi.org/10.1590/S0101-90742013000200003.

FRANCO, R. R. d. A. **Diversidade taxonômica e funcional da microbiota de fezes de macacos bugios (Alouatta spp.) de cativeiro e vida livre**. 2022. Universidade de São Paulo. Disponível em: https://doi.org/10.11606/T.95.2022.tde-14022023-130156.

FUNG, C. *et al.* Automation of QIIME2 Metagenomic Analysis Platform. **current protocols**, 2021. Disponível em: https://doi.org/10.1002/cpz1.254.

GALLOWAY-PEÑA, J.; HANSON, B. Tools for Analysis of the Microbiome. **Digestive Diseases and Sciences**, 2020. DOI: 10.1007/s10620-020-06091-y.

GAVRILESCU, M. Fate of pesticides in the environment and its bioremediation. **Engineering in Life Sciences**, v. 5, n. 6, p. 497–526, 2005.

GENTLEMAN, R. C. *et al.* **Bioconductor: open software development for computational biology and bioinformatics LaTeX**. 2004. Disponível em: https://doi.org/10.1186/gb-2004-5-10-r80. Acesso em: 22/08/2023.

GOI, S. R.; SOUZA, F. A. d. Diversidade de microrganismos do solo. **Floresta e Ambiente**, SciELO Brasil, v. 13, p. 46–65, 2006.

GOMES, D. I. d. S. *et al.* Estudo teórico das interações entre bases nitrogenadas do adn com o ÍON LI+ utilizando o modelo contínuo polarizável / Theoretical study of the interactions between dna nitrogen bases with the LI+ ION using the polarizable continuum model. **Brazilian Journal of Development**, v. 6, n. 6, p. 40970–40984, 2020. DOI: 10.34117/bjdv6n6-583. Disponível em: https://ojs.brazilianjournals.com.br/ojs/index.php/BRJD/article/view/12260.

HAWKINS, G. A. Analysis of Human Genetic Variations Using DNA Sequencing. In: BASIC Science Methods for Clinical Researchers. 2017. p. 77–98. DOI: 10.1016/b978-0-12-803077-6.00005-9.

HIRATA, H. V. **INVESTIGANDO O USO DE GPUS PARA APLICAÇÕES DE BIOINFORMÁTICA LaTeX**. 2011. Disponível em: http://rock.dcce.ibilce.unesp.br/spd/pubs/Mono%5C_Hector.pdf. Acesso em: 08/08/2023.

HOGEWEG, P. **The Roots of Bioinformatics in Theoretical Biology LaTeX**. 2011. Disponível em: https://journals.plos.org/ploscompbiol/article/file?id=10.1371/journal.pcbi.1002021%5C&type=printable. Acesso em: 31/07/2023.

ICMBIO. **Manejo Integrado do Fogo - Parque Nacional da Serra da Canastra**. 2016. São Roque de Minas: Parque Nacional da Serra da Canastra.

INPE, .-. I. N. d. P. E. **Portal do Monitoramento de Queimadas e Incêndios**. 2016. Disponível em: URL%20do%20portal.

JULI, L. B. D. **DIMINUIÇÃO DA DIVERSIDADE MICROBIANA DE SOLOS EM ÁREAS SOB ARENIZAÇÃO NO BIOMA PAMPA BRASILEIRO**. 2017. Disponível em: https://repositorio.unipampa.edu.br/jspui/handle/riu/4503.

KLEIN, E. N. M. **As inter-relações entre a nutrição, a bile e o microbioma intestinal humano: uma revisão narrativa**. 2021. Trabalho de Conclusão de Curso – Universidade Federal do Rio grande do Sul. Repositório Institucional Pantheon.

KNIGHT, R. *et al.* Best practices for analysing microbiomes. **Nature Reviews Microbiology**, Nature Publishing Group, v. 16, n. 7, p. 410–422, 2018. DOI: 10.1038/s41579-018-0029-9.

KOIDE, G. B. **Caracterização microbiológica do queijo Colonial artesanal de Seara-SC baseada em metodologias clássicas e metataxonômica**. 2022. UNIVERSIDADE FEDERAL DE SANTA CATARINA. Disponível em: https://repositorio.ufsc.br/handle/123456789/238055.

KUBIAK, B. B. **INFLUÊNCIA DE FATORES BIÓTICOS E ABIÓTICOS SOBRE O COMPORTAMENTO, ECOLOGIA E EVOLUÇÃO DA ESPÉCIE Ctenomys minutus (RODENTIA: CTENOMYIDAE)**. 2017. Tese (Doutorado) – Universidade Federal do Rio Grande do Sul, Porto Alegre. Disponível em: http://hdl.handle.net/10183/163673.

LANDER, E. S. **Impacto inicial do sequenciamento do genoma humano LaTeX**. 2011. Disponível em: https://doi.org/10.1038/nature09792. Acesso em: 31/07/2023.

LEAL, M. L. d. A. *et al.* Efeito dos sistemas de manejo e do uso do solo na população de microrganismos do solo. **Research, Society and Development**, Research, Society e Development, v. 10, n. 9, 2021.

LEHUGEUR, T. d. P.; MELO, H. C. S. **BIOINFORMÁTICA APLICADA NO DESENVOLVIMENTO DE NOVOS FÁRMACOS LaTeX**. 2018. Disponível em: https://www.lume.ufrgs.br/bitstream/handle/10183/77674/000896376.pdf?sequence=1%5C&isAllowed=y. Acesso em: 05/03/2023.

LESSA, R. N. T. **Ciclo do Nitrogênio**. 2007.

LIU, C.-H.; DI, Y. P. Analysis of RNA Sequencing Data Using CLC Genomics Workbench. In: **Molecular Toxicology Protocols**. Edição: Phouthone Keohavong, Kamaleshwar P. Singh e Weimin Gao. New York, NY: Springer US, 2020. p. 61–113. ISBN 978-1-0716-0223-2. DOI: 10.1007/978-1-0716-0223-2_4. Disponível em: https://sci-hub.se/https://link.springer.com/protocol/10.1007/978-1-0716-0223-2%5C_4.

LOBATO, F. M. F. **Abordagem probabilística para caracterização do sistema de marcação de sequenciamento multiplex na plataforma ABI SOLID**. 2011. p. 85. Mestrado em Engenharia Elétrica – Instituto de Tecnologia, Universidade Federal do Pará. Acesso em: 02/01/2024. Disponível em: http://repositorio.ufpa.br/jspui/handle/2011/2829.

MACHADO, D. M. *et al.* Atividades microbianas e as transformações no ciclo dos elementos no solo. **Enciclopédia Biosfera**, v. 8, n. 15, 2012.

MADIGAN, M.; MARTINKO, J.; BENDER, K. **Microbiologia de Brock**. 14th. Porto Alegre: Artmed, 2016.

MAGALHÃES, S. R. de; LIMA, G. S.; RIBEIRO, G. A. **Avaliação dos incêndios florestais ocorridos no Parque Nacional da Serra da Canastra-Minas Gerais**. 2012. Cerne, vol. 18, pp. 135–141.

MAGANA, A. J. *et al.* **A Survey of Scholarly Literature Describing the Field of Bioinformatics Education and Bioinformatics Educational Research LaTeX**. 2017. Disponível em: https://doi.org/10.1187/cbe.13-10-0193. Acesso em: 07/08/2023.

MAGURRAN, A. E. **Ecological diversity and its measurement**. Princeton university press, 1988.

MANHAES, C. M. C.; FRANCELINO, F. M. A. Biota do Solo e Suas Relações Ecológicas com o Sistema Radicular. **Nucleus (16786602)**, v. 10, n. 2, 2013.

MARCHESI, J. R.; RAVEL, J. The vocabulary of microbiome research: a proposal. **Microbiome**, v. 3, p. 31, 2015. DOI: 10.1186/s40168-015-0094-5.

MARCHETTI, M. M.; BARP, E. A. Efeito rizosfera: a importância de bactérias fixadoras de nitrogênio para o solo/planta–revisão. **IGNIS Periódico Científico de Arquitetura e Urbanismo Engenharias e Tecnologia de Informação**, p. 61–71, 2015.

MARCO, P. G. D. **Sequenciamento genético à luz do Direito: da digitalização da vida à identidade autêntica**. 2019. Tese (Doutorado) – UNIVERSIDADE FEDERAL DE MINAS GERAIS - FACULDADE DE DIREITO. Disponível em: http://hdl.handle.net/1843/30504.

MARTINS, A. M. **Sequenciamento de DNA, montagem de novo do genoma e desenvolvimento de marcadores microssatélites, indels e SNPs para uso em análise genética de Brachiaria ruziziensis**. 2013. p. viii, 190. Tese (Doutorado) – Universidade de Brasília, Brasília. Tese de Doutorado em Biologia Molecular. Disponível em: http://repositorio2.unb.br/jspui/handle/10482/15521.

MASSMAN, W. J.; FRANK, J. M.; MOONEY, S. J. **Advancing Investigation and Physical Modeling of First-Order Fire Effects on Soils LaTeX**. 2010. Disponível em: https://fireecology.springeropen.com/articles/10.4996/fireecology.0601036. Acesso em: 28/04/2023.

MATTOS, M. L. T. Microbiologia do Solo. In: NUNES, R. R.; REZENDE, M. O. O. (Ed.). **Recurso Solo: Propriedades e Usos**. São Carlos: Editora Cubo, 2015. p. 250–272.

MCKINNEY, W. **Python para Análise de Dados LaTeX**: Tratamento de dados com PANDAS, NUMPY E IPYTHON. Edição: Novatec Editora. 2018. p. 05–355. Disponível em: https://books.google.com.br/books?hl=pt-BR%5C&lr=%5C&id=Oj5FDwAAQBAJ%5C&oi=fnd%5C&pg=PA19%5C&dq=compara%5C%C3%5C%A7%5C%C3%5C%A3o+entre+as+linguagens+de+programa+Perl,+python+e+R%5C&ots=ZXP2qGQx%5C_t%5C&sig=FWrlMV2uTKgN5VXbh2zLEv-dOr0%5C#v=onepage%5C&q%5C&f=false. Acesso em: 21/08/2023.

MEDEIROS, M. B. de; FIEDLER, N. C. **INCÊNDIOS FLORESTAIS NO PARQUE NACIONAL DA SERRA DA CANASTRA: DESAFIOS PARA A CONSERVAÇÃO DA BIODIVERSIDADE LaTeX**. 2004. Disponível em: https://www.scielo.br/j/cflo/a/sC4Z8kqK9TGLJLCSwKmykFK/?lang=pt. Acesso em: 17/05/2023.

MELLO, F. C. A. *et al.* Detection of mixed populations of wild-type and YMDD hepatitis B variants by pyrosequencing in acutely and chronically infected patients. **BMC Microbiology**, v. 12, n. 1, p. 96, 2012. ISSN 1471-2180. DOI: 10.1186/1471-2180-12-96. Disponível em: https://doi.org/10.1186/1471-2180-12-96.

MELO, A. T. D. O. **MONTAGEM E CARACTERIZAÇÃO DO TRANSCRITOMA DE CANA-DE-AÇÚCAR (Saccharum spp.) UTILIZANDO DADOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO**. 2015. Tese (Doutorado) – Universidade Federal de Goiás. Disponível em: https://repositorio.bc.ufg.br/tede/handle/tede/4783.

MELO-MINARDI, R. C. de *et al.* **Caracterização dos Programas de Pós-graduação em Bioinform ática no Brasil LaTeX**. 2013. Disponível em: https://sol.sbc.org.br/index.php/brasnam/article/view/6833/6726. Acesso em: 31/07/2023.

MESSIAS, C. G.; FERREIRA, M. C. **Modelo Geoespacial para a Identificação de Áreas com Perigo de Propagação de Queimadas no Parque Nacional da Serra da Canastra LaTeX**. 2016. Disponível em: https://www.revistas.usp.br/rdg/article/view/153493/158200. Acesso em: 23/05/2023.

MOREIRA, F. M. d. S. *et al.* Bactérias Diazotróficas Associativas: Diversidade, Ecologia e Potencial de Aplicações. **Comunicata Scientiae**, v. 1, n. 2, p. 74–99, 2010.

MOREIRA, J. V. F. **DNA METABARCODING DA MICROBIOTA PRESENTE NO CHORUME DO ATERRO SANITÁRIO DA CIDADE DE FOZ DO IGUAÇU-PR VISANDO OS PROCESSOS DE BIORREMEDIAÇÃO**. 2019. Disponível em: http://dspace.unila.edu.br/handle/123456789/5625.

MOURA, W. B. D. ESTRUTURA DA DIVERSIDADE BETA E DETERMINANTES DE RIQUEZA DA METACOMUNIDADE FITOPLANCTÔNICA EM RESERVATÓRIOS, 2020. Disponível em: https://repositorio.bc.ufg.br/tede/handle/tede/10983.

NAIR, A.; NGOUAJIO, M. Soil microbial biomass, functional microbial diversity, and nematode community structure as affected by cover crops and compost in an organic vegetable production system. **Applied Soil Ecology**, v. 58, p. 45–55, 2012. ISSN 0929-1393. DOI: https://doi.org/10.1016/j.apsoil.2012.03.008. Disponível em: https://www.sciencedirect.com/science/article/pii/S0929139312000753.

NETO, A. Z.; CINTRA, L. C. **Análise de dados de RNA-Seq utilizando o Galaxy LaTeX**. 2016. Disponível em: http://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/1064217. Acesso em: 22/08/2023.

NHACUTOUO, S. A. **Virtualização de Servidores: Impacto LaTeX**. 2020. Disponível em: http://monografias.uem.mz/handle/123456789/2735. Acesso em: 14/08/2023.

NISHIYAMA, R. R. ESTUDO COMPARATIVO DE ABORDAGENS DE SEQUENCIAMENTO DE UM BANCO DE DADOS PÚBLICO PARA IDENTIFICAR O NÚCLEO DO MICROBIOMA DA SALIVA HUMANA, 2021. Disponível em: https://repositorio.unicamp.br/Busca/Download?codigoArquivo=552369.

OGUNYEMI, A.; JOHNSTON, K. **Is Server Virtualization Implementation in Business and Public Organizations a Worthwhile Investment? LaTeX**. 2017. Disponível em: https://doi.org/10.1142/S0219622017500146. Acesso em: 15/08/2023.

OLIVEIRA, É. L. D. **CONTRIBUIÇÕES DA BIOINFORMÁTICA PARA IDENTIFICAÇÃO DE ALVOS TERAPÊUTICOS NA PANDEMIA DO SARS-CoV-2: UMA ANÁLISE CIENCIOMÉTRICA LaTeX**. 2023. Disponível em: http://repositorio.utfpr.edu.br/jspui/handle/1/31486. Acesso em: 22/08/2023.

OLIVEIRA, L. N.; LUIZ SCHIAVONI, F. ávio. pdpk - Uma definição de pacote de software para o ambientePure Data LaTeX, 2018. Disponível em: https://sol.sbc.org.br/journals/index.php/reic/article/view/1069/939.

OLIVEIRA, T. **Aplicação de sequenciamento de nova geração no diagnóstico molecular de cardiomiopatia hipertrófica**. 2015. Tese (Doutorado) – Faculdade de Medicina da Universidade de São Paulo, São Paulo. DOI: 10.11606/D.5.2015.tde-29102015-115502.

OLIVEIRA, V. G. P. d. **Pipeline de Bioinformática para análise Metataxonômica de crianças autistas.** 2022. UNIVERSIDADE DE SÃO PAULO.

OTTO, T. D. *et al.* **A plataforma PDTIS de bioinformática: da seqüência à função LaTeX**. 2007. Disponível em: https://www.arca.fiocruz.br/bitstream/handle/icict/17601/ve%5C_Otto%5C_Thomas%5C_Dan%5C_etal%5C_2007%5C_pt.pdf?sequence=2%5C&isAllowed=y. Acesso em: 22/08/2023.

OUZOUNIS, C. A.; VALENCIA, A. **Early bioinformatics: the birth of a discipline– a personal view LaTeX**. 2003. Disponível em: https://doi.org/10.1093/bioinformatics/btg309. Acesso em: 06/06/2023.

PARANHOS, T. d. S. **Caracterização de bactérias presentes na carne bovina utilizando o sequenciamento da região V3/V4 do gene 16S rRNA**. 2019. UNIVERSIDADE FEDERAL DE SANTA CATARINA. Disponível em: https://repositorio.ufsc.br/handle/123456789/199736.

PARRON, L. M. *et al.* (Ed.). **Serviços Ambientais em Sistemas Agrícolas e Florestais do Bioma Mata Atlântica**. Brasília, DF: Embrapa, 2015. p. 372.

PEDROSA, M. V. *et al.* Importância ecológica dos microrganismos do solo. **Enciclopédia Biosfera**, v. 11, n. 22, 2015.

POOR, A. P. **Caracterização de patógenos bacterianos causadores de metrite em suínos através da cultura e análise metagenômica**. 2022. Faculdade de Medicina Veterinária e Zootecnia da Universidade de São Paulo. Disponível em: https://doi.org/10.11606/T.10.2022.tde-12092022-113605.

PORTO, A. L. M. *et al.* **COLETA, IDENTIFICAÇÃO, PRESERVAÇÃO DA MICROBIOTA DA ÁREA II DO CAMPUS DA USP**. 2013. Tese (Doutorado) – Instituto de Química de São Carlos.

PURVIS, A.; HECTOR, A. Getting the measure of biodiversity. **Nature**, 2000. Disponível em: https://sci-hub.se/https://www.nature.com/articles/35012221.

PYROSEQUENCING Technology and Platform Overview. https://www.qiagen.com/us/knowledge-and-support/knowledge-hub/technology-and-research/pyrosequencing-resource-center/pyrosequencing-technology-and-platform-overview. 28/12/2023.

QIAGEN. **System requirements**. 2023. Disponível em: https://digitalinsights.qiagen.com/technical-support/system-requirements/.

RAMOS, J. L. C. *et al.* Um estudo comparativo de classificadores na previsão da evasão de alunos em EAD. In. Disponível em: https://www.researchgate.net/profile/Alex-Gomes-11/publication/328615769%5C_Um%5C_estudo%5C_comparativo%5C_de%5C_classificadores%5C_na%5C_previsao%5C_da%5C_evasao%5C_de%5C_alunos%5C_em%5C_EAD/links/5bd8dc6892851c6b279b8472/Um-estudo-comparativo-de-classificadores-na-previsao-da-evasao-de-alunos-em-EAD.pdf.

REDIN, M. *et al.* **IMPACTOS DA QUEIMA SOBRE ATRIBUTOS QUÍMICOS, FÍSICOS E BIOLÓGICOS DO SOLO LaTeX**. 2011. Disponível em: https://www.scielo.br/j/cflo/a/jkprVJMw5mKbKjd9G4xyQ4p/?format=pdf%5C&lang=pt. Acesso em: 28/04/2023.

RIBEIRO, A. A. *et al.* Microbioma humano: uma interação predominantemente positiva. **Uningá Review**, v. 19, n. 1, 2014.

RIBEIRO, D. **diversidade**. Acesso em:11/11/2023. Disponível em: https://www.dicio.com.br/diversidade/.

RIBEIRO, I. **Identificação das variantes dos genes RHD e RHCE em pacientes com doença falciforme usando estratégia de sequenciamento de nova geração**. 2020. Tese (Doutorado) – Faculdade de Medicina da Universidade de São Paulo, São Paulo. DOI: 10.11606/D.5.2020.tde-23082021-103036.

RIBEIRO, J. C. M. **ESTUDO DA SUCESSÃO MICROBIANA EM AMBIENTES MARINHOS**. 2022. Disponível em: http://hdl.handle.net/11422/17416.

ROCHA, A. A. M.; VALE, V. S. do. DIVERSIDADE ALFA E BETA DE COMUNIDADES VEGETAIS DE CERRADO REMANESCENTES NAS BEIRAS DE ESTRADAS DAS MARGENS DE RODOVIAS. **GETEC Gestão Tecnologia e Ciência**, 2017. Disponível em: https://www.revistas.fucamp.edu.br/index.php/getec/article/view/1004.

RODRIGUES, J. A. M. **Otimização de alocação de máquinas virtuais em datacenter heterogêneo de sistema de computação em nuvem**. 2019. Dissertação (Mestrado em Ciência da Computação) – Universidade Estadual Paulista, São José do Rio Preto. Disponível em: http://hdl.handle.net/11449/182074. Acesso em: 14 ago. 2023.

RODRIGUES, P. C. M. **Obtenção de microrganismos solubilizadores com potencial valor ecológico para uma agricultura sustentável**. 2013. Diss. (Mestrado) – Universidade de Lisboa, Portugal.

RONCONI, L. **Microhcip filogenético: uma nova ferramenta molecular para o estudo da biodiversidade do solo**. 2008. Universidade Federal de Santa Catarina.

SALGADO, J. F. M. **O microbioma do intestino médio de *Aedes aegypti* e suas alterações em diferentes regimes alimentares**. 2021. Trabalho de Conclusão de Curso – Universidade Federal do Rio de Janeiro. Repositório Institucional Pantheon.

SANTI, D. B. **Análise metataxonômica de comunidades microbianas da rizosfera do milho após aplicação de nitrogênio e inoculação com bactérias**. 2019. Mestrado em Computação Aplicada – Universidade Estadual de Ponta Grossa, Ponta Grossa.

SANTOS, D. M. d. S. *et al.* Bactérias fixadoras de nitrogênio e molibdênio no cultivo do amendoim em solo do Cerrado. **Revista de Agricultura Neotropical**, Cassilândia-MS, v. 4, Suplemento 1, p. 84–92, 2017. ISSN 2358-6303.

SANTOS, W. F. *et al.* Sequenciamento de dna: métodos e aplicações. In: PROCEEDINGS of Safety, Health and Environment World Congress. 2013. v. 13, p. 139–141.

SCHMIDT, I. B. *et al.* Experiências internacionais de manejo integrado do fogo em áreas protegidas: recomendações para implementação de manejo integrado de fogo no cerrado. **Biodiversidade Brasileira**, v. 6, n. 2, p. 41–54, 2016.

SILVA, A. A. C. **Levantamento da comunidade bacteriana em diferentes tipos de solo impactados e não impactados por queimadas no parque nacional da serra da canastra – Minas Gerais.** 2019.

SILVA, F. C. d. *et al.* Quantificação da microbiota e diversidade ecológica da meso e macrofauna do solo sob diferentes usos no município de Urutaí (região Sudeste Goiano). **Multi-Science Journal**, v. 1, n. 4, p. 12–18, 2016.

SILVA, J. K. D. **SARA: DESENVOLVIMENTO E VALIDAÇÃO DE UM FLUXO DE ANÁLISES SEMIAUTOMÁTICO PARA ESTUDOS DE EXPRESSÃO GÊNICA**. 2022. Escola Bahiana de Medicina e Saúde Pública. Disponível em: https://repositorio.bahiana.edu.br:8443/jspui/handle/bahiana/6101.

SILVA, L. R. G. d. **Comunidades bacterianas presentes em reservatório de petróleo: análise de dados obtidos por sequenciamento de nova-geração (NGS) a partir de amostras de rocha-reservatório**. 2022. Disponível em: https://repositorio.usp.br/directbitstream/c44ab79e-2879-449e-83f7-d45ffd5d1c6b/Larissa%5C%20Reis%5C%20Gomes%5C%20da%5C%20Silva.pdf.

SILVA, P. I. T. **Descoberta e validação de marcadores SNPs por sequenciamento de alta performance do genoma estrutural e por genotipagem por sequenciamento (GBS) de arroz de sequeiro (Oryza sativa spp. japonica)**. 2012. Disponível em: http://www.realp.unb.br/jspui/handle/10482/14647.

SILVA, R. C.; LIMA, A.; SOUZA, L. C. d. S. Principais métodos de sequenciamento de DNA. **Scientific Electronic Archives**, 2022. DOI: 10.36560/15820221603. Disponível em: https://sea.ufr.edu.br/SEA/article/view/1603.

SOUZA, R. O. **Experimentos e Estudos de Manejo Integrado do Fogo em Unidade de Conservação do Cerrado – Parque Nacional da Serra da Canastra**. 2017. Diss. (Mestrado) – Programa de Pós Graduação em Sustentabilidade e Tecnologia Ambiental, Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG) Campus Bambuí.

SOUZA, R. O. *et al.* Estratégias de integração entre pesquisa e manejo do fogo no Parque Nacional da Serra da Canastra como parte do desenvolvimento de um Programa de Manejo Integrado do fogo. **Biodiversidade Brasileira**, v. 6, n. 2, p. 205–219, 2016.

STAJICH, J. E. *et al.* **The Bioperl Toolkit: Perl Modules for the Life Sciences LaTeX**. 2002. Disponível em: https://doi.org/10.1101/gr.361602. Acesso em: 22/08/2023.

STRÖHER, P. R. Sequenciamento de nova geração e entomologia: Novas perspectivas para antigos questionamentos. **Revista da Biologia**, v. 18, p. 27–36, 2018. DOI: 10.7594/revbio.18.01.04.

TAUK, S. M. Biodegradação de resíduos orgânicos no solo. **Brazilian Journal of Geology**, v. 20, n. 1, p. 299–301, 1990.

TEIXEIRA, N.; LEAL, J. P. Um Sistema de Processamento em Pipeline de XML. Disponível em: http://xata.fe.up.pt/2007/papers/14.pdf.

THOMPSON, L. R. *et al.* Tourmaline: A containerized workflow for rapid and iterable amplicon sequence analysis using QIIME 2 and Snakemake. **GigaScience**, v. 11, p. giac066, 2022. ISSN 2047-217X. DOI: 10.1093/gigascience/giac066. eprint: https://academic.oup.com/gigascience/article-pdf/doi/10.1093/gigascience/giac066/45181673/giac066.pdf. Disponível em: https://doi.org/10.1093/gigascience/giac066.

TORSVIK, V.; ØVREÅS, L.; THINGSTAD, T. F. Prokaryotic diversity - magnitude, dynamics, and controlling factors. **Science**, v. 296, n. 5570, p. 1064–1066, 2002. DOI: 10.1126/science.1071698.

VACONDIO, B. **Biodegradação do pesticida pentaclorofenol por uma linhagem de fungo marinho isolado da ascídia Didemnum ligulum**. 2014. Diss. (Mestrado) – Universidade Federal de São Carlos.

VAUGHAN-NICHOLS, S. **Linux totally dominates supercomputersLaTeX**: It finally happened. Today, all 500 of the world's top 500 supercomputers are running Linux. 2017. Disponível em: https://www.zdnet.com/article/linux-totally-dominates-supercomputers/. Acesso em: 21/08/2023.

VAZ, R. V. **Análise da diversidade taxonômica e funcional da microbiota presente no sedimento marinho da praia de Lucena-PB através da abordagem metagenômica e cultivo convencional**. 2021. Universidade Federal Rural de Pernambuco. Disponível em: http://www.tede2.ufrpe.br:8080/tede2/handle/tede2/8779.

VÁZQUEZ-BAEZA, Y. *et al.* EMPeror: a tool for visualizing high-throughput microbial community data, 2013. Disponível em: https://doi.org/10.1186/2047-217X-2-16.

VERLI, H. *et al.* **Bioinformática: da Biologia à Flexibilidade Molecular LaTeX**. 2014. Disponível em: https://www.lume.ufrgs.br/bitstream/handle/10183/166105/001012172.pdf?sequence=1. Acesso em: 31/07/2023.

VOGEL, T. M. *et al.* TerraGenome: a consortium for the sequencing of a soil metagenome. **Nature Reviews Microbiology**, v. 7, n. 4, p. 252, 2009.

WATSON, J. D.; CRICK, F. H. A Structure for Deoxyribose Nucleic Acid. **Nature**, v. 171, p. 737–738, 1953.

YAMAGUTI, M. **Isolamento de micoplasma de suínos com problemas de proteção e tipificação dos isolados pela PFGE e sequenciamento do gene 16S rRNA.** 2009. Tese (Doutorado) – Universidade de São Paulo. Disponível em: https://doi.org/10.11606/T.42.2009.tde-29102009-091226.

ZACARIA, R. **Infraestrutura de Tecnologia de Informação para Projetos de Sequenciamento de Genomas de Fungos**. 2018. Disponível em: https://repositorio.ucs.br/11338/3931.