

**UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS**

Larissa Reis Gomes da Silva

**Comunidades bacterianas presentes em reservatório de
petróleo: análise de dados obtidos por sequenciamento de
nova-geração (NGS) a partir de amostras de
rocha-reservatório**

São Carlos

2022

Larissa Reis Gomes da Silva

Comunidades bacterianas presentes em reservatório de petróleo: análise de dados obtidos por sequenciamento de nova-geração (NGS) a partir de amostras de rocha-reservatório

Trabalho de conclusão de curso apresentado ao Instituto de Física de São Carlos da Universidade de São Paulo para obtenção do título de Bacharel em Ciências Físicas e Biomoleculares.

Orientadora: Profa. Dra. Nelma Regina Segnini Bossolan

**São Carlos
2022**

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Silva, Larissa Reis Gomes da

Comunidades bacterianas presentes em reservatório de petróleo: análise de dados obtidos por sequenciamento de nova geração (NGS) a partir das amostras de rocha-reservatório / Larissa Reis Gomes da Silva; orientadora Nelma Regina Segnini Bossolan -- São Carlos, 2022.

34 p.

Trabalho de Conclusão de Curso (Bacharela em Ciências Físicas e Biomoleculares) -- Instituto de Física de São Carlos, Universidade de São Paulo, 2022.

1. Bacterias. 2. Rocha-reservatório. 3. Extremofilos.
I. Bossolan, Nelma Regina Segnini, orient. II. Título.

AGRADECIMENTOS

Ao meu namorado Lucas, pelo apoio incondicional, pela parceria, comprometimento e por aparecer quando menos esperava e se tornar tudo que eu preciso.

Ao meu filho Gabriel, por me mostrar todos os dias que posso ser mais paciente, mais responsável e a melhor mãe do mundo.

À minha mãe Cláudia, avó Esmeralda e irmã Gabrielle, por me permitirem sonhar e chegar aos meus objetivos, por me apoiarem e nunca me deixarem desistir.

À minha família, por todo cuidado e carinho. Ao meu pai Adriano, por compartilhar sua sabedoria e me ensinar a busca pelo conhecimento.

Aos meus amigos de graduação, carinhosamente chamados “Jubilados”, por me acompanharem nesta caminhada, criarem os momentos mais felizes e ajudarem nos momentos mais difíceis.

Às minhas amigas Gabriela e Raquel, que me esperavam chegar em casa cansada, ouviam meus problemas e me davam os melhores abraços, e Debora e Tamires, por me mostrarem que as amizades podem perdurar, apesar da distância.

À bateria universitária Gaperia e todos os amigos que fiz nela, pois proporcionaram os melhores anos da minha vida e levarei comigo para sempre.

Ao Clube de Astronomia do IFSC, que manteve minha paixão por astronomia acesa e me ensinou a ser melhor divulgadora científica.

Aos docentes do Instituto de Física de São Carlos, pela bagagem que me proporcionaram ao longo desses anos.

À professora Nelma, que me aceitou prontamente em seu laboratório, por sua atenção e paciência, orientação e conselhos, me auxiliando na descoberta de minha vocação.

Às agências de fomento, que se mantêm na missão de auxiliar a pesquisa brasileira e, principalmente, à PETROBRAS, que possibilitou a realização deste trabalho.

À todas e todos, muito obrigada!

RESUMO

SILVA, L. R. G. **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. 34p. Trabalho de Conclusão de Curso (Bacharel em Ciências Físicas e Biomoleculares) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2022.

Poucos são os estudos feitos a partir de amostras coletadas de rocha-reservatório, quer seja pelas limitações técnicas, ou pelos custos de pesquisa. O presente trabalho avaliou a diversidade microbiana de um ambiente extremófilico para fins de conhecimento e manejo. O ambiente amostrado foi um reservatório de petróleo hipertermófilico e hipersalino, localizado na Bacia de Campos, RJ, cujas amostras de rocha foram coletadas de profundidades superiores a 3.000 metros e cedidas ao IFSC no âmbito de um Termo de Cooperação com a PETROBRAS. A partir do DNA extraído das amostras de rocha, realizou-se o sequenciamento em larga escala das regiões V4-V5 e V6-V8 do gene 16S rRNA em plataforma Illumina (sequenciamento de alto rendimento), cujos resultados foram analisados por bioinformática, utilizando a plataforma QIIME2. Analisou-se os índices de diversidade da amostra, concluindo-se que o esforço amostral foi satisfatório e o número de sequências utilizadas nas análises foi suficiente para estudar a diversidade das comunidades. Os táxons identificados no sequenciamento foram correspondentes ao reino Bacteria, majoritariamente pertencentes aos filos Proteobacteria e Firmicutes a partir da região V4-V5, e Actinobacteria e Proteobacteria a partir da região V6-V8. A amplificação da região V6-V8 do gene 16S rRNA resultou em uma maior diversidade de Famílias identificadas.

Palavras-chave: Bactérias. Rocha-reservatório. Extremófilos.

ABSTRACT

SILVA, L. R. G. **Bacterial communities present in oil reservoir:** analysis of data obtained by next-generation sequencing (NGS) from reservoir rock samples. 2022. 34p. Conclusion course paper (Bachelor in Physics and Biomolecular Sciences) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2022.

Few studies are carried out from samples collected from reservoir rock, either due to technical limitations or research costs. This work evaluated the microbial diversity of an extremophilic environment for knowledge and management purposes. The sampled environment was a hyperthermophilic and hypersaline oil reservoir, located at the Bacia de Campos, RJ, whose rock samples were collected from depths greater than 3,000 meters and provided to the IFSC under a Cooperation Agreement with PETROBRAS. From the DNA extracted from the rock samples, large-scale sequencing of the V4-V5 and V6-V8 regions of the 16S rRNA gene was carried out on the Illumina platform (high-throughput sequencing), whose results were analyzed by bioinformatics, using the QIIME2 platform. The sample diversity indices were analyzed, concluding that the sampling effort was satisfactory and the number of sequences used in the analyzes was sufficient to study the diversity of communities. The taxa identified in the sequencing corresponded to the kingdom Bacteria, mostly belonging to the phyla Proteobacteria and Firmicutes for the V4-V5 region, and Actinobacteria and Proteobacteria for the V6-V8 region. Amplification of the V6-V8 region of the 16S rRNA gene resulted in a greater diversity of identified Families.

Keywords: Bacteria. Reservoir rock. Extremophiles.

LISTA DE FIGURAS

- Figura 1 – Para retirar as alíquotas de rocha, utilizou-se uma furadeira de bancada (imagem à esquerda), à qual foram acopladas brocas do tipo serra copo de 25 ou 50 mm (imagem central), com as quais foram feitas perfurações na face longitudinal de cada amostra (conforme esquema indicado na imagem à direita, que representa uma amostra de rocha de reservatório). 17
- Figura 2 – Curvas de rarefação indicando o efeito do esforço no sequenciamento para estimar o número de sOTUs em cada amostra. Análise das regiões V4-V5 (A) e V6-V8 (B) do gene 16S rRNA, sendo sOTUs observadas, em seguida sOTUs estimadas pelo índice de Chao1 e, por fim, OTUs estimadas pelo parâmetro filogenético PD. 22
- Figura 3 – Gráficos de duas coordenadas principais de PCoA, construídos a partir de matriz de distância UNIFRAC, sendo (I) qualitativo e (II) quantitativo, das sOTUs anotadas para (A) região V4-V5 e (B) região V6-V8 do gene 16S rRNA. 25
- Figura 4 – Frequência relativa dos Filos anotados a partir das sequências geradas com a amplificação das regiões V4-V5 (superior) e V6-V8 (inferior) do gene 16S rRNA, obtido das amostras de rocha das quatro profundidades estudadas (P7, P35, P46 e P53). 26
- Figura 5 – Frequência relativa das Famílias anotadas a partir das sequências geradas com a amplificação das regiões V4-V5 (superior) e V6-V8 (inferior) do gene 16S rRNA. 27

LISTA DE TABELAS

Tabela 1	– Leituras antes do tratamento das amostras (R1 e R2) e após os tratamentos computacionais (<i>read-joining</i> , <i>quality control</i> e Deblur), para as regiões V4-V5 e V6-V8 do gene 16S rRNA.	21
Tabela 2	– Quantidade de sOTUs observadas nas amostras, índices de riqueza (Chao1 e ACE) e cobertura (Good's Coverage) para as regiões V4-V5 e V6-V8 do gene 16S rRNA.	23
Tabela 3	– Índices de diversidade (Simpson e Shannon) e equitabilidade (Pielou) para as regiões V4-V5 e V6-V8 do gene 16S rRNA.	24

LISTA DE ABREVIATURAS E SIGLAS

ANP	Agência Nacional de Petróleo, Gás Natural e Biocombustíveis
API	American Petroleum Institute
ASV	Amplicon Sequence Variant
CENPES	Centro de Pesquisas, Desenvolvimento e Inovação Leopoldo Américo Miguez de Mello
DNA	Deoxyribonucleic acid
HTS	High-Throughput Sequencing
IFSC	Instituto de Física de São Carlos, Universidade de São Paulo
MAFFT	Multiple Alignment using Fast Fourier Transform
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
pb	Pares de bases
PCoA	Principal Coordinate Analysis
PCR	Polymerase Chain Reaction
PETROBRAS	PETRÓLEO BRASILEIRO S.A.
QIIME2	Quantitative Insights Into Microbial Ecology, versão 2022.2
RDP	Ribosomal Database Project
RJ	Rio de Janeiro, estado do Brasil
rRNA	ribosomal Ribonucleic acid
sOTU	sub-Operational Taxonomic Unit
UNIFRAC	Unique Fraction Metric

1 INTRODUÇÃO

A biosfera microbiana da subsuperfície profunda, continental e marinha, vem sendo objeto de interesse de diversos pesquisadores ao longo das últimas décadas, principalmente pelo interesse em identificar comunidades microbianas presentes nessas regiões e compreender seus mecanismos de sobrevivência e multiplicação em condições extremas, que são características destes ambientes. Esses estudos são possíveis por meio de coletas de amostras de água de fendas hidrotermais, por meio de expedições globais para perfurações exploratórias, com intuito de conhecer a estrutura da crosta terrestre (1), e por perfurações e coletas de fluidos de reservatórios de petróleo.(2)

O interesse em estudar reservatórios de petróleo reside na compreensão dos processos de biodegradação do óleo em reservatórios profundos (3), no controle de processos de corrosão microbiana de poços (4) e, principalmente, no uso biotecnológico da atividade microbiana e de seus subprodutos para o aumento da recuperação do petróleo.(5) Um fator limitante para a realização desses estudos é a dificuldade em se coletar as amostras sem contaminação externa e os custos para se realizar as perfurações, fazendo com que existam poucos trabalhos realizados diretamente a partir de amostras de rocha reservatório.(6–9)

Compreende-se por biosfera profunda um ambiente de temperaturas superiores a 50 °C e localizado abaixo de 1 metro da superfície (10), alcançando até 3.000 m abaixo da superfície da crosta terrestre, ou até 4.000 m, se considerarmos os registros de óleo biodegradado nesta profundidade.(3) Nos reservatórios, a presença do óleo caracteriza uma microbiota particular, uma vez que o mesmo está disponível para grupos microbianos que o acessam diretamente ou aos seus subprodutos.(11) Esta microbiota pode ser caracterizada por meio de estudos dependente e independente de cultivo, a partir de amostras de óleo ou de fluidos coletados nas saídas de poços de produção. As técnicas independentes de cultivo se baseiam na análise do material genético (Deoxyribonucleic acid - DNA) extraído diretamente das amostras ambientais. A maioria dos estudos com foco na diversidade microbiana e suas relações filogenéticas, ecológicas e evolutivas são os dependentes de primers, com amplificação por reação em cadeia da polimerase (Polymerase Chain Reaction - PCR) e sequenciamento do gene ribossomal 16S rRNA (ribosomal Ribonucleic acid) ou fragmentos dele.

O gene 16S rRNA é um dos marcadores moleculares mais utilizados em ecologia microbiana por possuir características como ubiquidade, função essencial no metabolismo celular, e propriedades evolutivas.(12) Componente da pequena subunidade 30S dos ribossomos procarióticos, corresponde a uma região com baixa taxa de mutação e o gene é encontrado em todos os microrganismos. Este gene possui particularidades que permitem o acesso à diversidade microbiana em amostras ambientais, tais como a presença de regiões

altamente conservadas e que não são significativamente diferentes entre espécies, permitindo o desenho de primers específicos, e regiões hipervariáveis (V1 a V9), que apresentam especificidade a cada espécie e que são utilizadas para inferir relações filogenéticas entre os microrganismos.(13,14) Assim, as regiões conservadas são idênticas para todas as espécies procarióticas e as variáveis são únicas para cada espécie, permitindo a utilização do gene 16S rRNA para distinção individual (fingerprint) de cada espécie dentro da comunidade microbiana. Para aumentar a fidedignidade da classificação das sequências, têm-se utilizado um sistema de amplificação combinado à amplificação e à análise de mais de uma região hipervariável.(15,16)

O sequenciamento do DNA microbiano pode ser realizado por meio de técnicas de sequenciamento de alto rendimento (High-Throughput Sequencing – HTS), denominadas sequenciamento de nova geração (Next Generation Sequencing – NGS) (17), que possuem como vantagem o sequenciamento direto dos amplicons, sem a necessidade de construção de biblioteca de clones e o grande volume de sequências produzidas em uma única corrida (de milhares a milhões), dependendo da plataforma utilizada. Alguns exemplos de plataformas de NGS são Roche 454, Illumina, Iron Torrent e Solid, que devem ser escolhidas de acordo com os objetivos. A plataforma utilizada neste trabalho foi a Illumina, que utiliza o sequenciamento paired-end, o qual produz fragmentos maiores (18), portanto essa estratégia torna-se interessante e tem sido muito usada atualmente em estudos de comunidade microbiana.

No âmbito de um Termo de Cooperação entre o Instituto de Física de São Carlos (IFSC) e a PETROBRAS (PETRÓLEO BRASILEIRO S.A.) (Termo de Cooperação 0050.0079046.12.9, regulado pela Agência Nacional de Petróleo, Gás Natural e Biocombustíveis - Resolução ANP 05/2015), a empresa forneceu amostras de rocha triturada coletadas durante uma atividade de perfuração exploratória em um reservatório de petróleo hipertermofílico e hipersalino (Campo de Tartaruga Verde, Bacia de Campos, RJ). A partir do DNA extraído destas amostras, foi realizado o sequenciamento do gene 16S rRNA em plataforma Illumina (sequenciamento de alto rendimento), cujos resultados serão analisados por meio de índices de diversidade e análises taxonômicas.

2 MATERIAIS E MÉTODOS

2.1 Obtenção das amostras de rocha-reservatório

O campo de onde as amostras foram coletadas é um reservatório de carbonato pós-sal de idade Albiano e que contém óleo bruto médio, de gravidade API de 26,9 °. A água de formação situava-se a 2.185 m abaixo do leito do mar, era hipersalina (127.000 mg/kg), com temperatura de aproximadamente 80 °C e pressão de 4.560 psi. Esta água ocorre naturalmente nos poros de uma rocha e é produzida junto com o petróleo.(19) Foram coletadas quatro amostras de 20 cm de comprimento, no intervalo de profundidade de 3.006,9 m a 3.053,2 m, incluindo a coluna d'água, de cerca de 942 m de altura. Conforme as colunas de perfuração eram abertas, as amostras foram removidas e limpas com papel embebido de álcool a 70%, embaladas em papel craft estéril e armazenadas sob refrigeração (aproximadamente 4 °C) até o processamento em laboratório.

Para obter as alíquotas de amostras para inoculação nos meios de cultivos, perfurou-se as amostras lateralmente com brocas do tipo serra copo, diamantadas na extremidade, acopladas a uma perfuradora de bancada. Cada perfuração foi feita desprezando-se os primeiros centímetros para evitar coleta de material atingido pelo fluido de perfuração, que é usado durante o procedimento de perfuração de um reservatório. Destaca-se que a própria pressão interna do reservatório impediu que o fluido de perfuração avançasse mais do que 1 cm para dentro da rocha, observação essa feita em campo. Após cada perfuração, as amostras foram retiradas da perfuradora com uma espátula estéril, dispostas em frascos de plástico estéreis e mantidas em jarra de anaerobiose (4 °C) até a extração do DNA. As ferramentas de perfuração foram limpas e esterilizadas anteriormente a cada extração de amostra de rocha.



Figura 1 – Para retirar as alíquotas de rocha, utilizou-se uma furadeira de bancada (imagem à esquerda), à qual foram acopladas brocas do tipo serra copo de 25 ou 50 mm (imagem central), com as quais foram feitas perfurações na face longitudinal de cada amostra (conforme esquema indicado na imagem à direita, que representa uma amostra de rocha de reservatório).

Fonte: Equipe técnica da Gerência de Biotecnologia do CENPES-PETROBRAS.

Foram enviados pela equipe técnica da Gerência de Biotecnologia do CENPES-PETROBRAS ao nosso laboratório 10 g de cada uma das quatro amostras obtidas, denominadas de P7 (3006,9 m a 3007,1 m), P35 (3035,1 a 3035,25 m), P46 (3046,5 m a 3046,7 m) e P53 (3053,0 m a 3053,2 m), de acordo com os intervalos de profundidade de onde foram extraídas.

2.2 Extração do DNA das amostras

Para a extração do DNA genômico das quatro amostras de rocha (P7, P35, P46 e P53) utilizou-se o kit NucleoSpin® Soil (Macherey-Nagel, Germany), seguindo as instruções do fabricante. Realizou-se 3 extrações a partir de 250 mg de cada amostra, estas foram ressuspensas no tampão de lise SL2 e suplementadas com Enhancer SX. O DNA extraído foi quantificado, com a produção de 5 a 10 ng/ μ l por extração por amostra. Por fim, as 3 repetições de cada amostra foram unidas.

2.3 Amplificação por PCR e sequenciamento do gene 16S rRNA das comunidades bacterianas das amostras

O gene 16S rRNA contém regiões altamente conservadas e aproximadamente 1500 pb. As regiões escolhidas para amplificação foram a região V4-V5, por esta abranger um maior número de espécies de Bactérias e Archaea, sem superestimativa (20), e a região V6-V8, por esta ser altamente específica para Eubacteria (21) e ser capaz de fornecer resultados quantitativos dos grupos predominantes. Para amplificar a região V4-V5, utilizou-se os pares de primers 515F (5'-CTGNCAGCMGCCGCGGTAA-3') e 926R (5'-CCGYCAATTYMTTTRAGTTT-3') (22), e para a região V6-V8, utilizou-se 967F (5'-GAACGCGAAGAACCTTACC-3') (23) e 1346R (5'-TAGCGCTTCCGACTTCA-3').(24) A análise de *self-dimers* e *cross primer dimers* entre os pares foi realizada pela ferramenta *Multiple Primer Analyzer* da Thermo Fisher Scientific e testou-se previamente os pares de iniciadores na base de dados RDP (Ribosomal Database Project) para verificação do número de hits alcançados por cada par, avaliando-se a abrangência destes.

A biblioteca do gene 16S rRNA das amostras foi preparada e sequenciada pela empresa Helixxa (Campinas, São Paulo, Brasil), utilizando a plataforma Illumina MiSeq a partir do Metagenomics workflow descrito pelo fabricante. O resultado do sequenciamento foi submetido na plataforma online BaseSpace® Sequence Hub Illumina, seguindo um workflow onde a corrida bruta foi demultiplexada, gerando os arquivos FASTQ para cada amostra, *paired-end* (R1 e R2). As sequências FASTQ (R1 e R2) das 4 amostras foram depositadas no NCBI Sequence Read Archive (SRA) no BioProject PRJNA407737, experimentos de código [SRA:SRX3217160] e [SRA:SRX3217213] para as regiões V4-V5 e V6-V8 do gene 16S rRNA, respectivamente.

2.4 Tratamento dos dados do sequenciamento do gene 16S rRNA e medidas da diversidade microbiana

Diversos são os softwares capazes de realizar o tratamento de sequências genômicas e metagenômicas, podendo estes combinarem entre si e gerarem diferentes algoritmos e ferramentas de preparação e processamento de sequências, muitas vezes em uma única plataforma de análises de sequenciamentos. O Quantitative Insights Into Microbial Ecology (QIIME), atualmente na versão 2022.2, é uma plataforma de bioinformática de microbioma de next-generation extensível, gratuita e de código aberto (<https://qiime2.org/>).⁽²⁵⁾ A plataforma possibilita que sejam feitos tratamentos desde demultiplexação (retirada de *barcodes*) e filtragem de qualidade, até análises de diversidade e taxonomia. Em sua documentação, disponível em <https://docs.qiime2.org/2022.2/>, é possível seguir um tutorial de como realizar cada uma dessas etapas.

Os sequenciamentos gerados no tópico 2.3 eram do tipo FASTQ, *paired-end*, demultiplexados e não seguiam os formatos tradicionais da plataforma, então estes foram importados no QIIME2 por meio do método “Fastq manifest”, onde importam-se os dados manualmente, criando-se um arquivo manifesto que mapeia identificadores de amostra e indica a direção das leituras *paired-end* em cada arquivo. Essa importação faz com que os dados de entrada sejam convertidos em artefatos do QIIME2 (formato .qza), que é o utilizado pela plataforma.

Sendo as sequências *paired-end*, ou seja, leituras diretas e reversas não unidas, utilizou-se a ferramenta VSEARCH (26) por meio do comando `qiime vsearch join-pairs` para juntá-las e fez-se o controle de qualidade das sequências unidas por meio de `qiime quality-filter q-score`, removendo possíveis adaptadores e bases com qualidade PHRED inferior a 20.⁽²⁷⁾ Um dos resultados fornecidos é o gráfico do índice de qualidade, cujo ponto onde se observa a queda desse índice é utilizado como parâmetro *-trim-length* para o comprimento da sequência no Deblur, que realiza a remoção de ruídos por meio do `qiime deblur denoise-16S`. Deblur é um método de sub unidade taxonômica operacional (sub-Operational Taxonomic Unit - sOTU) ou Amplicon Sequence Variant (ASV), que busca a identificação rápida e precisa de sequências exatas em estudos de amplicons.⁽²⁸⁾ Primeiro, ele remove artefatos conhecidos, como adaptadores, e em seguida, as sequências retidas passam por filtragem de ruído e remoção de sequências quiméricas.

Uma árvore pode ser gerada para analisar a diversidade filogenética das amostras, pois as métricas utilizadas para essas análises exigem uma árvore filogenética enraizada relacionando os recursos entre si. Para isso, utiliza-se o comando `qiime phylogeny align-to-tree-mafft-fasttree`, onde o *pipeline* usa o MAFFT (Multiple Alignment using Fast Fourier Transform) (29) para realizar um alinhamento de múltiplas sequências e criar um arquivo de sequências alinhadas artefato QIIME2. Em seguida, filtra o alinhamento para remover posições altamente variáveis e aplica o FastTree para gerar uma

árvore filogenética a partir do alinhamento mascarado. O programa FastTree cria uma árvore não enraizada, portanto, na etapa final desta seção, o enraizamento do ponto médio é aplicado para colocar a raiz da árvore no ponto médio da maior distância ponta a ponta na árvore não enraizada.(30)

Para saber se foi possível medir a diversidade nas amostras, não haver distorção nas análises estatísticas e possibilitar a comparação das comunidades microbianas, utilizou-se os comandos relacionados às diversidades alfa e beta. Primeiro, avaliou-se o esforço amostral por meio de curvas de rarefação das amostras, utilizando o comando `qiime diversity alpha-rarefaction`, variando-se a métrica de alfa diversidade para filogenética (Phylogenetic diversity), considerando a árvore enraizada obtida anteriormente, e não filogenéticas Chao1 e OTUs observadas.(31) Em seguida, os diferentes índices de diversidade alfa foram obtidos por meio de `qiime diversity alpha`, onde alterou-se a métrica para `observed_features`, `chao1`, `ace`, `goods_coverage`, `simpson`, `shannon` e `pielou_e` e os resultados foram tabelados para comparação.

A diversidade beta, por sua vez, foi obtida a partir do comando `qiime diversity beta-rarefaction`, que utiliza a árvore filogenética enraizada como uma das entradas para gerar gráficos de análise de coordenada principal (PCoA) com o índice UNIFRAC, calculado a partir de métrica qualitativa (*unweighted*), onde considera-se apenas a presença ou ausência das sOTUs observadas, e quantitativa (*weighted*), que considera a abundância das sOTUs, apresentando maior peso para proporção das espécies nas amostras.(32) Tanto a rarefação, quanto os índices alfa e beta foram calculados considerando o parâmetro *-max-depth* como o menor número de sequências da menor amostra após a filtragem com Deblur.

A análise taxonômica foi realizada por meio do comando `qiime feature-classifier classify-sklearn`, utilizando o classificador Naive Bayes (33) pré-treinado com sequências de referência do Greengenes (34,35), que pode ser utilizado para alguns alvos de genes marcadores comuns, como genes 16S rRNA. O classificador pré-treinado foi o chamado Greengenes 13_8 99%, com sequências completas de OTUs (36,37) para abranger as duas regiões analisadas e agrupadas em 99% de similaridade de sequência. O classificador foi aplicado às sequências obtidas após o tratamento com Deblur, e gerou-se gráficos de barra por meio do comando `qiime taxa barplot`.

3 RESULTADOS

3.1 Filtragem das sequências

Após as etapas de filtragem, pode-se observar que o número de leituras retidas diminuiu consideravelmente (Tabela 1). O comprimento de corte (parâmetro *-trim-length*) para o Deblur foi no valor 350 para as sequências da região V4-V5 e 360 para V6-V8, pois correspondia à região do gráfico onde observou-se a queda no índice de qualidade (discutido no tópico 2.4). Assim, o Deblur descartou qualquer sequência que contenha menos nucleotídeos que o comprimento de corte, uma vez que o algoritmo do Deblur exige que todas as leituras tenham o mesmo comprimento.(28)

Tabela 1 – Leituras antes do tratamento das amostras (R1 e R2) e após os tratamentos computacionais (*read-joining*, *quality control* e Deblur), para as regiões V4-V5 e V6-V8 do gene 16S rRNA.

Amostras	“reads” R1	“reads” R2	read-joining	quality control	Deblur
P7_v4v5	48354	48354	34815	30655	1641
P35_v4v5	59889	59889	41501	36566	1672
P46_v4v5	59801	59801	42119	37314	1658
P53_v4v5	57974	57974	43294	38566	1883
P7_v6v8	60847	60847	40206	37916	13802
P35_v6v8	74303	74303	47514	44668	14327
P46_v6v8	70344	70344	48278	44765	13126
P53_v6v8	67240	67240	48484	45307	13311

Fonte: Elaborada pela autora.

Ao final desta etapa de tratamento, obteve-se que os menores números de leituras de sequências para cada uma das regiões corresponde às amostras P7_v4v5 (1641 leituras), para a região V4-V5, e P46_v6v8 (13126 leituras), para a região V6-V8. Esses valores foram utilizados como parâmetro *-max-depth* para gerar as curvas de rarefação e os índices alfa e beta nas análises de diversidade microbiana.

3.2 Análises de diversidade microbiana

A rarefação é um método que ajusta as diferenças nos tamanhos das bibliotecas entre as amostras para ajudar nas comparações da diversidade alfa.(38) Seleciona-se um número especificado de sequências que é igual ou menor que o número de sequências na menor amostra e, em seguida, descarta-se aleatoriamente as leituras maiores até que o número de amostras restantes seja igual a este limiar.(39) Esse número corresponde ao parâmetro *-max-depth* utilizado para as curvas de rarefação. Baseando-se nessas subamostras de

tamanho igual, são calculadas métricas de diversidade que podem contrastar os ecossistemas, independentemente das diferenças nos tamanhos das amostras.(40)

Para verificar se o tamanho das amostras após a filtragem foi suficientemente grande para representar a comunidade em estudo, as curvas de rarefação encontram-se na Figura 2. Os valores utilizados para `-max-depth` foram 1640 e 13125 leituras para V4-V5 e V6-V8, respectivamente, conforme a Tabela 1.

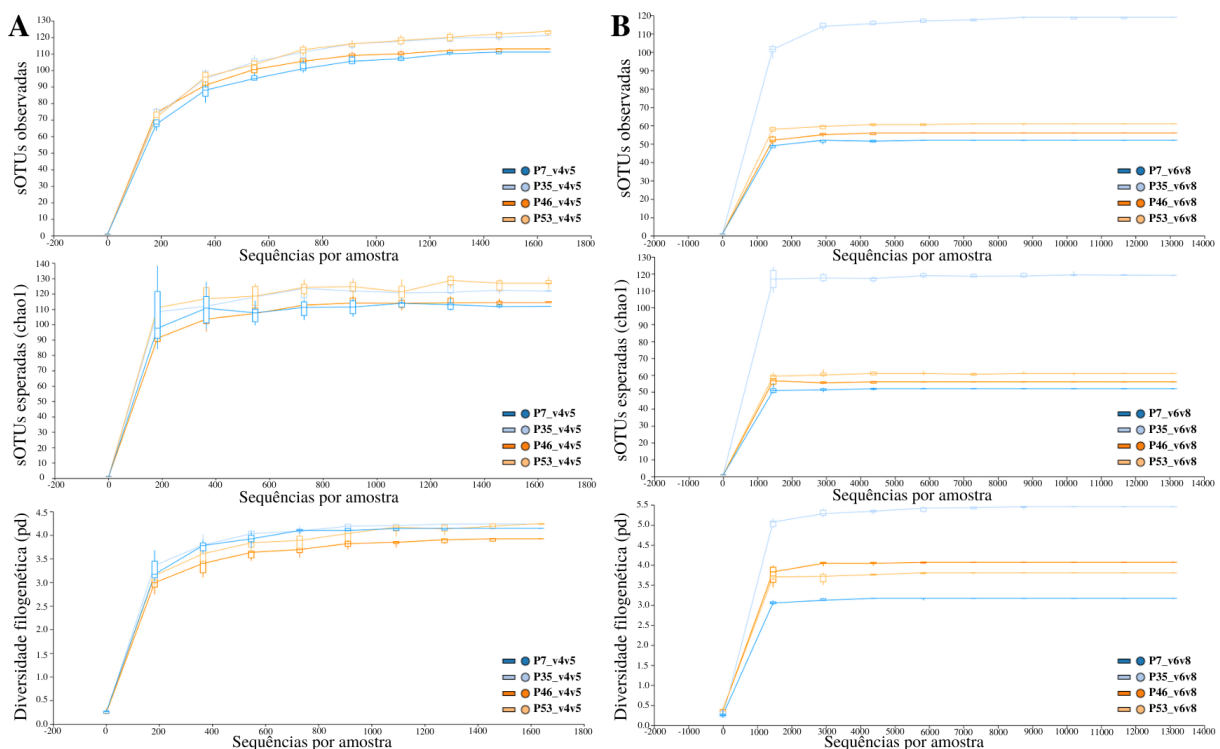


Figura 2 – Curvas de rarefação indicando o efeito do esforço no sequenciamento para estimar o número de sOTUs em cada amostra. Análise das regiões V4-V5 (A) e V6-V8 (B) do gene 16S rRNA, sendo sOTUs observadas, em seguida sOTUs estimadas pelo índice de Chao1 e, por fim, OTUs estimadas pelo parâmetro filogenético PD.

Fonte: Elaborada pela autora.

Observa-se que mesmo as curvas não sendo idênticas, elas possuem o mesmo comportamento, indicando uma estabilização próxima ao ponto de 400 sequências por amostra para a região V4-V5 e 1000 sequências por amostra para a região V6-V8. Podemos dizer, então, que a amostragem utilizada não necessita de um aumento no número de sequências, pois não haverá incremento de novas sOTUs, evidenciando que o esforço amostral foi satisfatório e o número de sequências utilizadas nas análises foi suficiente para estudar a diversidade das comunidades.

3.2.1 Diversidade Alfa

A diversidade alfa refere-se ao número e à abundância de espécies dentro de uma comunidade, a partir da qual podemos estudar a riqueza, diversidade e equitabilidade.(38) Nas Tabelas 2 e 3, é possível ver as medidas para cada um dos estimadores utilizados, conforme descrito no tópico 2.4. Os estimadores foram calculados considerando uma profundidade de amostras uniforme, assim como na rarefação, e serão discutidos juntamente às análises dos resultados.

Tabela 2 – Quantidade de sOTUs observadas nas amostras, índices de riqueza (Chao1 e ACE) e cobertura (Good's Coverage) para as regiões V4-V5 e V6-V8 do gene 16S rRNA.

Amostras	sOTUs observadas	Riqueza (Chao1)	Riqueza (ACE)	Cobertura (Good's coverage)
P7_v4v5	111	111.83	112.5734	0.9970
P35_v4v5	121	121.67	122.5653	0.9970
P46_v4v5	113	114.25	114.5399	0.9970
P53_v4v5	124	126.33	127.0393	0.9958
P7_v6v8	52	52.00	52.0000	1.0000
P35_v6v8	119	119.00	119.0000	1.0000
P46_v6v8	56	56.00	56.0000	1.0000
P53_v6v8	61	61.00	61.0000	1.0000

Fonte: Elaborada pela autora.

A riqueza da comunidade, também chamada de densidade de espécies, estima o número de espécies esperado caso todos os indivíduos da comunidade fossem amostrados. Os estimadores mais utilizados são o Chao1, que se baseia no número de espécies raras dentro de uma amostra, e o ACE (Abundance-based Coverage Estimator), que se baseia na abundância das espécies raras.(41) De acordo com os resultados para estes dois estimadores na Tabela 2, pode-se observar que não se obteve grandes diferenças no número de sOTUs esperadas estatisticamente com a quantidade observada nas amostras (coluna "sOTUs observadas"), indicando que o número de espécies nas comunidades foi amostrado praticamente em sua totalidade, considerando as regiões alvo estudadas.

O índice de cobertura (Good's coverage) foi calculado para estimar a porcentagem de diversidade capturada pelo esforço do sequenciamento.(42) Observa-se que os resultados para todas as amostras analisadas indicaram que mais de 99% da microbiota foi amostrada, concordando com o observado nas curvas de rarefação e os estimadores Chao1 e ACE, sugerindo que a diversidade microbiana foi contemplada na amostragem.

Os estimadores de diversidade se baseiam na abundância relativa das espécies.(38) O índice Simpson (**D**) trabalha com a dominância dos indivíduos em relação às espécies

Tabela 3 – Índices de diversidade (Simpson e Shannon) e equitabilidade (Pielou) para as regiões V4-V5 e V6-V8 do gene 16S rRNA.

Amostras	Diversidade Simpson (1-D)	Diversidade Shannon (H)	Equitabilidade (Pielou)
P7_v4v5	0.9724	5.9162	0.8707
P35_v4v5	0.9807	6.2339	0.9010
P46_v4v5	0.9796	6.1835	0.9066
P53_v4v5	0.9779	6.2012	0.8917
P7_v6v8	0.9169	4.4383	0.7786
P35_v6v8	0.8499	4.0140	0.5822
P46_v6v8	0.7117	3.0411	0.5237
P53_v6v8	0.7763	3.5042	0.5909

Fonte: Elaborada pela autora.

presentes, representando a probabilidade de, ao se selecionar um indivíduo aleatoriamente dentro de uma comunidade, ele já ter sido observado.(43) Quanto maior a dominância de algumas espécies em relação às outras, maior será o D. Este índice usualmente tem sido representado como $1 - D$ para ter uma representação direta com a diversidade. Já o índice de diversidade de Shannon (**H**) é uma medida quantitativa da riqueza da comunidade, avaliando a heterogeneidade das espécies da amostra, levando em consideração a proporção de indivíduos de cada espécie amostrada.(44) Ele expressa o grau de incerteza em se predizer a que espécie pertence um indivíduo tomado ao acaso em uma comunidade. Portanto, quanto maior o grau de incerteza, maior a diversidade.

Pelos resultados presentes na Tabela 3, como os valores para o índice Simpson são muito próximos de 1, temos que os valores obtidos para o estimador D foram muito pequenos para as amostras das regiões consideradas, ainda que um pouco maior para as amostras P46 e P53 da região V6-V8, se analisadas em relação às demais. Assim, pode-se concluir que não há grande dominância de espécies em relação às outras dentro das amostras. Ambos os índices Simpson e Shannon demonstraram valores aproximados entre as amostras, sugerindo que não há diferenças significativas nas diversidades das comunidades analisadas. Os menores índices de diversidade podem ser explicados por maior diferença nas abundâncias entre as sOTUs obtidas.

Finalmente, a equitabilidade mede a uniformidade da distribuição dos indivíduos nas espécies presentes na comunidade. Uma das formas de medir a equitabilidade é o índice de Pielou, derivado do índice de Shannon.(45) O índice de Pielou mede a uniformidade, e ele pode variar de 0 (mínimo) a 1 (máximo). A riqueza e a equitabilidade são altamente dependentes do esforço amostral. Ainda pela Tabela 3, pode-se concluir pelos resultados de equitabilidade que os indivíduos estão distribuídos uniformemente entre as espécies quando

considerada a região V4-V5, ou seja, as abundâncias entre as sOTUs representadas nas amostras estão distribuídas uniformemente, já que os valores se aproximam de 1, enquanto para a região V6-V8 temos uma distribuição desigual.

3.2.2 Diversidade Beta

A diversidade beta corresponde à diversidade biológica entre ambientes.(46) Consiste em uma análise multivariada, uma vez que temos diversas espécies de diferentes ambientes, e mede-se a variação entre as comunidades de uma região, considerando a diferença na composição de espécies entre as comunidades. Essa análise depende de uma matriz de distâncias, que é gerada a partir da árvore filogenética, e compara duas unidades amostrais com base em descritores, que podem ser as espécies presentes, abundância de espécies, biomassa de espécies, variáveis ambientais ou de localização espacial.

O índice UNIFRAC é uma métrica que incorpora a distância filogenética para realizar o cálculo de dissimilaridade entre os membros das comunidades (47), baseando-se no comprimento dos ramos de uma árvore filogenética compartilhada entre as comunidades analisadas. Com isso, tem-se que quanto menor a distância entre os ramos, mais similar a composição das comunidades, pois são compostas por linhagens que compartilham uma história evolutiva comum. Para visualização dos resultados foram construídos gráficos da análise de coordenada principal (PCoA) (Figura 3), utilizando a distância por UNIFRAC qualitativo (I) e quantitativo (II).

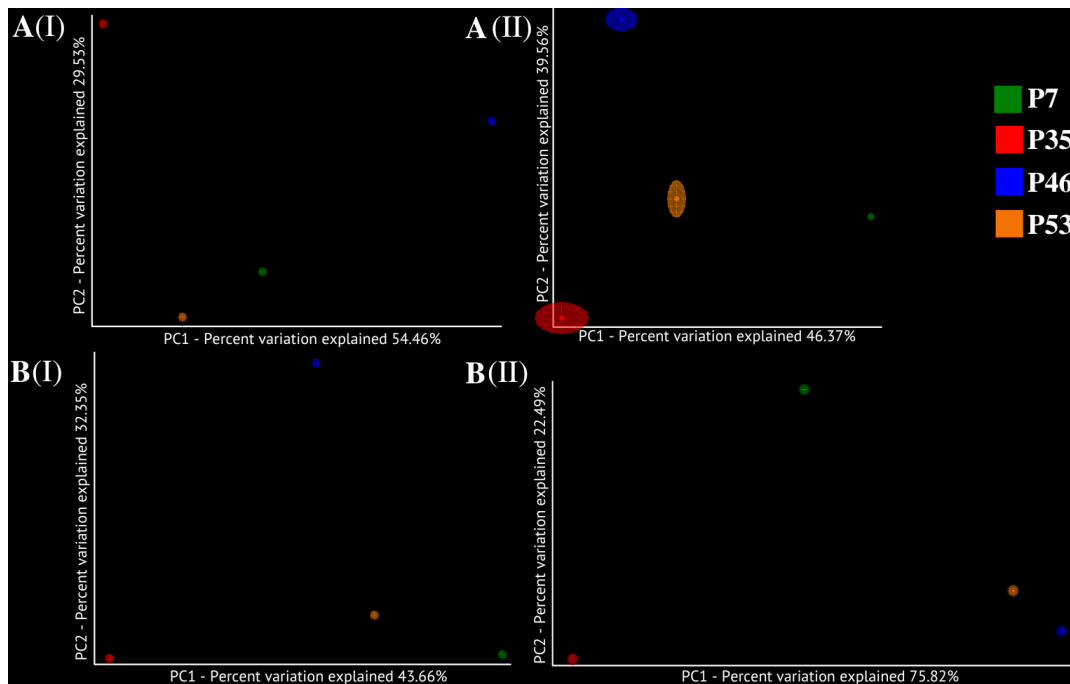


Figura 3 – Gráficos de duas coordenadas principais de PCoA, construídos a partir de matriz de distância UNIFRAC, sendo (I) qualitativo e (II) quantitativo, das sOTUs anotadas para (A) região V4-V5 e (B) região V6-V8 do gene 16S rRNA.

Fonte: Elaborada pela autora.

Na Figura 3, temos que as distâncias entre os pontos são próximas às dissimilaridades originais entre as comunidades microbiológicas, ou seja, as amostras que estão “mais próximas” são mais similares filogeneticamente. Pode-se observar que o eixo horizontal contém a maior parte da variação dos dados, já que representa os maiores valores percentuais, se comparado ao eixo vertical. Analisando os resultados para a região V4-V5, obteve-se que os fatores 1 e 2 do PCoA somados explicam 83.99% e 85.93% da variação da comunidade microbiana total nas amostras, não considerando (qualitativo) e considerando a abundância (quantitativo), respectivamente.

Apesar da proximidade entre os dois valores, a similaridade entre as profundidades foi diferente para essas medidas. Quando não considera-se a abundância, as amostras P7 e P53 apresentam maior proximidade e as demais se distanciam proporcionalmente, diferentemente de quando se considera a abundância, onde obteve-se maior dissimilaridade entre P7 e as demais amostras. Para a região V6-V8, os fatores explicam 76.01% e 98.31% com os parâmetros qualitativo e quantitativo, respectivamente. Pôde-se observar um padrão muito parecido entre as amostras, sendo que, ao desconsiderar-se a abundância, P7 e P53 estão mais próximas (assim como para V4-V5), enquanto a P53 encontra-se mais próxima da P46 quando considera-se a abundância das sOTUs.

3.3 Classificação taxonômica

Embora tenha-se optado pelo uso do primer para amplificar a região V4-V5 considerando sua abrangência para espécies dos Domínios Bacteria e Archaea, não foram identificadas sequências classificadas como Archaea, apenas Bacteria. Nas Figuras 4 e 5, podem ser observados os Filos e Famílias anotados para as sequências geradas com a amplificação das regiões V4-V5 e V6-V8 do gene 16S rRNA.

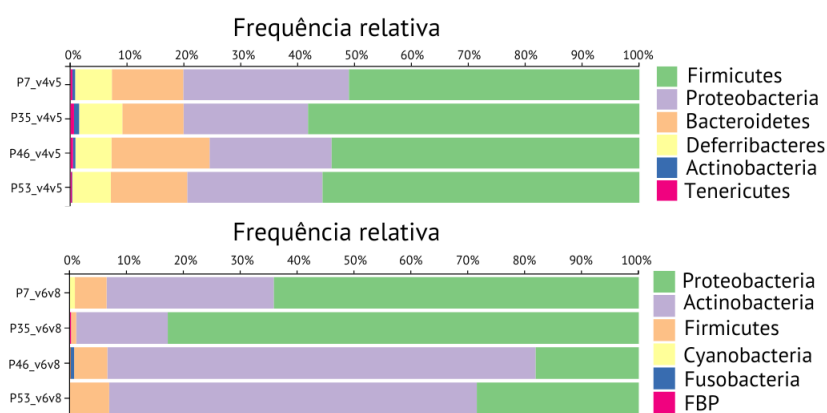


Figura 4 – Frequência relativa dos Filos anotados a partir das sequências geradas com a amplificação das regiões V4-V5 (superior) e V6-V8 (inferior) do gene 16S rRNA, obtido das amostras de rocha das quatro profundidades estudadas (P7, P35, P46 e P53).

Fonte: Elaborada pela autora.

Para a região V4-V5, observou-se uma predominância de representantes dos Filos Firmicutes, Proteobacteria, Bacteroidetes e Deferribacteres em todas as amostras (Figura 4 superior). Esses Filos representaram 99.20% (P7), 98.51% (P35), 99.15% (P46) e 99.69% (P53) do total em cada amostra. Considerando a região V6-V8, observou-se uma predominância de representantes dos Filos Proteobacteria, Actinobacteria e Firmicutes nas quatro amostras, representando 99.15% (P7), 99.83% (P35), 99.28% (P46) e 100% (P53) do total em cada amostra (Figura 4 inferior). Todas as amostras foram classificadas em algum dos Filos identificados.

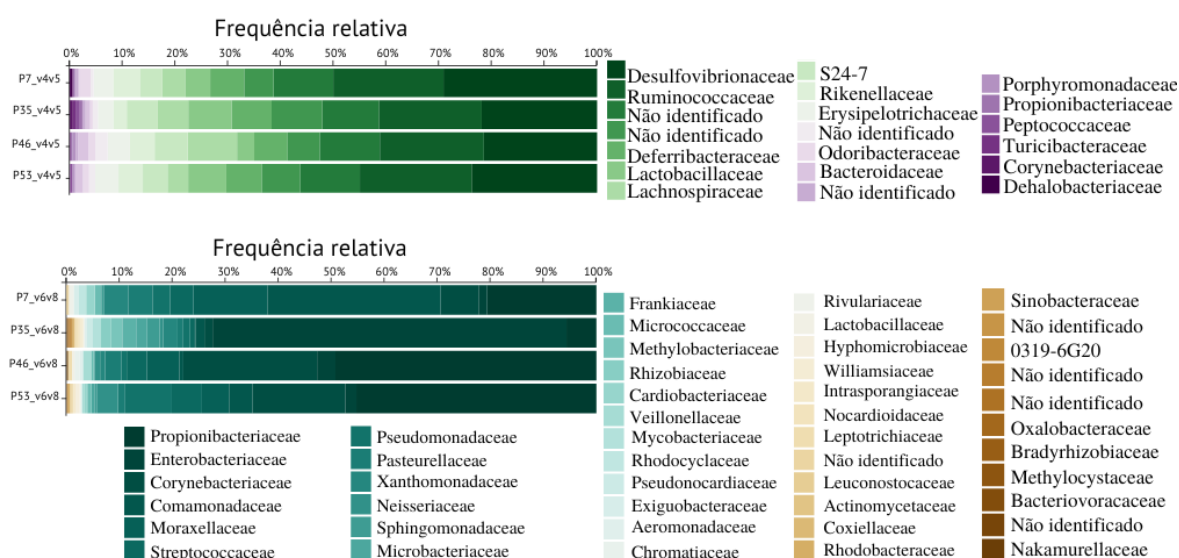


Figura 5 – Frequência relativa das Famílias anotadas a partir das sequências geradas com a amplificação das regiões V4-V5 (superior) e V6-V8 (inferior) do gene 16S rRNA.

Fonte: Elaborada pela autora.

Na Figura 5, pode-se observar que as duas famílias que predominaram nas amostras da região V4-V5 pertencem aos filos Proteobacteria e Firmicutes: Desulfovibrionaceae (29.13%) e Ruminococcaceae (21.19%), respectivamente. Os organismos da família Desulfovibrionaceae são bactérias redutoras de sulfato, aerotolerantes, comumente encontradas em ambientes aquáticos ricos em material orgânico e formam os principais membros da comunidade de aquíferos de rochas fraturadas graníticas profundas.(48, 49) Sua presença tem sido relatada em estudos de diversidade microbiana de reservatório a partir de rocha, como os feitos por Von der Weid *et al.* (2008) e Gales *et al.* (2016).(7,8)

A família Ruminococcaceae, por sua vez, possui representantes associados ao trato gastrointestinal de mamíferos (50) e pertence à classe Clostridia, assim como as duas famílias abundantes não identificadas no gráfico. Sabe-se que muitos membros dessa classe são capazes de formar esporos e podem mais raramente serem encontrados em solo.(51) Vale destacar a presença dos membros de Deferribacteraceae (7.60%), que são bastonetes anaeróbios, não formadores de esporos, possuindo algumas espécies termofílicas, halofílicas

ou halotolerantes, com metabolismo variado. Foram isolados de ambientes marinhos, fendas hidrotermais e reservatórios de óleo.(52)

As famílias que se destacaram para a região V6-V8 foram *Propionibacteriaceae* (49.25%) e *Corynebacteriaceae* (25.24%), ambas do filo *Actinobacteria* e ordem *Actinomycetales*, e da família *Comamonadaceae* (32.63%), do filo *Proteobacteria*. Os representantes de *Propionibacteriaceae* e *Corynebacteriaceae* são aeróbios ou anaeróbios facultativos, podem ocupar diversos habitats (lodo ativado, mar, solo contaminado) e alguns representantes são comensais ou causam doenças em humanos.(53, 54) Na família *Comamonadaceae*, os representantes são aeróbios, a maioria deles de habitats de água e solo e mesófilos, embora alguns tenham sido encontrados em habitats antárticos e fontes termais, e outros possam ser patógenos.(55)

A presença significativa de sequências anotadas em famílias com membros associados à microbiota de mamíferos, particularmente gastrointestinal, ou como agentes patogênicos, pode indicar possível contaminação das amostras (ocorrida no campo ou no processamento durante as análises laboratoriais). Se não descartarmos a presença dessas famílias, uma consideração a se fazer envolve a origem dessas rochas geradoras de petróleo, que são formadas pelo acúmulo de sedimentos depositados com matéria orgânica, algas e plâncton.(56) Pode-se sugerir que os representantes dessas famílias foram depositados nessas rochas por serem “levados” por correntes marítimas ou organismos hospedeiros. Seriam necessários mais estudos e coletas de amostras dessas regiões, além de um estudo sobre os representantes de hospedeiros marinhos que podem ser contaminados por essas bactérias, para chegar-se a essas conclusões.

São poucos os estudos de análise microbiana a partir de rocha de reservatório utilizando sequenciamento de larga escala, o que dificulta comparação direta entre os dados obtidos neste estudo e os obtidos em estudos similares. Para citar alguns, Tsesmetzis e colaboradores (2018) estudaram amostras de rochas de reservatório termofílico profundo usando sequenciamento de larga escala e verificaram a dominância das famílias *Oxalobacteraceae*, *Pseudomonadaceae* e *Comamonadaceae*.(57) Gales e colaboradores (2016) realizaram estudo de rocha de reservatório de óleo onshore hipersalino não explorado, por meio de biblioteca de clones e sequenciamento em larga escala, verificando que as famílias mais abundantes foram *Halomonadaceae*, *Desulfohalobiaceae* e *Pseudomonadaceae*; dentre as menos abundantes, detectaram membros da família *Comamonadaceae* e da ordem *Bacteroidales*.(8)

No presente estudo, muitas das famílias identificadas têm membros com metabolismo aeróbio, algumas com representantes termofílicos e outras com isolados de reservatório de petróleo. Uma relação entre a presença destes táxons e um possível quadro metabólico predominante no ambiente amostrado pode ser futuramente estudado a partir de uma caracterização geoquímica da rocha e química da água e do óleo circundantes.

4 CONCLUSÃO

Estudar as comunidades microbianas de reservatórios é um grande desafio, considerando as limitações de acesso a essas regiões e dos custos que envolvem perfuração, extração e coletas de amostras. Por isso, os poucos estudos desse tipo são de extrema importância para que se possa conhecer mais sobre a biosfera da subsuperfície profunda de nosso planeta. Graças a técnicas de sequenciamento em larga escala, atualmente é possível analisar a diversidade de organismos em um ambiente a partir de poucas amostras, utilizando modelos estatísticos para prever o comportamento de uma população microbiana.

Este trabalho conseguiu amostrar de maneira satisfatória a diversidade das comunidades coletadas, pois as curvas de rarefação indicaram que a quantidade de sequências foi suficiente para obter-se a quantidade de sOTUs necessárias para as análises taxonômicas. Além disso, os índices de riqueza e diversidade foram próximos aos esperados teoricamente, resultando em uma cobertura de aproximadamente 99% da microbiota, sugerindo que a diversidade microbiana foi contemplada na amostragem.

O gráfico de distância filogenética quantitativa indicou que as amostras P35, P46 e P53 da região V4-V5 estavam mais distantes da P7 e, para a região V6-V8, as amostras se distanciaram igualmente entre si, sendo que as amostras P46 e P56 se mantiveram mais próximas. Esses comportamentos foram observados e confirmados na quantidade de sOTUs amostradas em cada Família nos gráficos de taxonomia.

Na análise taxonômica, foram identificadas Famílias cujos representantes, em sua maioria, localizam-se no trato intestinal de mamíferos, como Ruminococcaceae e Lachnospiraceae. Ainda assim, observou-se famílias que representam organismos que se assemelham aos esperados para reservatórios, como Desulfovibrionaceae, Deferribacteraceae e Comamonadaceae.

Finalmente, pôde-se verificar que a plataforma utilizada foi satisfatória para realizar as análises propostas. Entretanto, existe a possibilidade de se repetir os procedimentos deste trabalho, utilizando outros softwares e parâmetros que sejam mais convenientes e considerando, inclusive, a exclusão de sOTUs que representam possíveis contaminações, assim como em alguns trabalhos mais sofisticados da literatura, citados anteriormente.

REFERÊNCIAS

- 1 EDWARDS, K. J.; BECKER, K.; COLWELL, F. The deep, dark energy biosphere: intraterrestrial life on earth. **Annual Review of Earth and Planetary Sciences**, v. 40, p. 551–568, 2012.
- 2 MAGOT, M. Indigenous microbial communities in oil fields. *In*: OLIVIER, B.; MAGOT, M. **Petroleum microbiology**, New York: Wiley, 2005, cap.2, p. 21-33.
- 3 HEAD, I. M.; SAUNDERS, J. R.; PICKUP, R. W. Microbial evolution, diversity, and ecology: a decade of ribosomal RNA analysis of uncultivated microorganisms. **Microbial Ecology**, v. 35, n. 1, p. 1-21, 1998.
- 4 CROLET, Jean-Louis. Microbial corrosion in the oil industry: a corrosionist's view. *In*: OLIVER B.; MAGOT, M. **Petroleum microbiology**, New York: Wiley, 2005, p. 143-169.
- 5 MCINERNEY, M.; NAGLE, D. P.; KNAPP, R. M. Microbially enhanced oil recovery: past, present, and future. *In*: OLIVER, B.; MAGOT, M. **Petroleum microbiology**, New York: Wiley, 2005, p. 215-237.
- 6 NUNOURA, Takuro *et al.* Culture-dependent and independent analyses of subsurface microbial communities in oil-bearing strata of the Sagara oil reservoir. **Island Arc**, v. 15, n. 3, p. 328-337, 2006.
- 7 VON DER WEID, I. *et al.* Molecular diversity of bacterial communities from seafloor rock samples in a deep-water production basin in Brazil. **Journal of Microbiology and Biotechnology**, v. 18, n. 1, p. 5-14, 2008.
- 8 GALES, Grégoire *et al.* Preservation of ancestral Cretaceous microflora recovered from a hypersaline oil reservoir. **Scientific Reports**, v. 6, n. 1, p. 1-10, 2016.
- 9 AN, Dongshan *et al.* Metagenomics of hydrocarbon resource environments indicates aerobic taxa and genes to be unexpectedly common. **Environmental Science Technology**, v. 47, n. 18, p. 10708-10717, 2013.
- 10 COLMAN, D. R. *et al.* The deep, hot biosphere: Twenty-five years of retrospection. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 114, n. 27, p. 6895–6903, 2017.
- 11 PEDERSEN, K. Exploration of deep intraterrestrial microbial life: current perspectives. **FEMS Microbiology Letters**, v. 185, n. 1, p. 9–16, 2000.
- 12 CASE, R. J. *et al.* Use of 16s rrna and rpob genes as molecular markers for microbial ecology studies. **Applied and Environmental Microbiology**, Am Soc Microbiol, v. 73, n. 1, p. 278–288, 2007.
- 13 DUARTE, R. T. D. **Micro-organismos em ambientes criogênicos: gelo glacial, solos expostos por recuo de geleiras, e permafrost polares**. 2010. Tese (Doutorado em Biotecnologia) — Instituto de Ciências Biomédicas da Universidade de São Paulo, São Paulo, 2010.

- 14 WARD, David M.; CASTENHOLZ, Richard W.; MILLER, Scott R. Cyanobacteria in geothermal habitats. *In*: WHITTON, B. A. ed. **Ecology of cyanobacteria II**. Dordrecht: Springer, 2012, p. 39-63.
- 15 YILDIRIM, S. *et al.* Characterization of the fecal microbiome from non-human wild primates reveals species specific microbial communities. **PloS One**, v. 5, n. 11, p. e13963, 2010.
- 16 LAZAREVIC, V. *et al.* Study of inter-and intra-individual variations in the salivary microbiota. **BMC Genomics**, v. 11, n. 1, p. 1–11, 2010.
- 17 SHENDURE, J.; JI, H. Next-generation dna sequencing. **Nature Biotechnology**, v. 26, n. 10, p. 1135–1145, 2008.
- 18 BARTRAM, A. K. *et al.* Generation of multimillion-sequence 16s rRNA gene libraries from complex microbial communities by assembling paired-end illumina reads. **Applied and Environmental Microbiology**, v. 77, n. 11, p. 3846–3852, 2011.
- 19 TELES, Danilo Ribeiro de Sá. *et al.* Caracterização isotópica de águas de formação hipersalinas de um campo de petróleo da Bacia do Recôncavo, Brasil. **Revista Brasileira de Geofísica**, v. 28, p. 291-301, 2010.
- 20 SUN, D.-L. *et al.* Intragenomic heterogeneity of 16s rRNA genes causes overestimation of prokaryotic diversity. **Applied and Environmental Microbiology**, v. 79, n. 19, p. 5962–5969, 2013.
- 21 SCHMALENBERGER, A.; SCHWIEGER, F.; TEBBE, C. C. Effect of primers hybridizing to different evolutionarily conserved regions of the small-subunit rRNA gene in PCR-based microbial community analyses and genetic profiling. **Applied and Environmental Microbiology**, v. 67, n. 8, p. 3557–3563, 2001.
- 22 QUINCE, C. *et al.* Removing noise from pyrosequenced amplicons. **BMC Bioinformatics**, v. 12, n. 1, p. 1–18, 2011.
- 23 HUSE, S. M. *et al.* Exploring microbial diversity and taxonomy using ssu rRNA hypervariable tag sequencing. **PLoS Genetics**, v. 4, n. 11, p. e1000255, 2008.
- 24 NÜBEL, U. *et al.* Sequence heterogeneities of genes encoding 16s rRNAs in *paenibacillus polymyxa* detected by temperature gradient gel electrophoresis. **Journal of Bacteriology**, v. 178, n. 19, p. 5636–5643, 1996.
- 25 BOLDYEN, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. **Nature Biotechnology**, v. 37, n. 8, p. 852–857, 2019.
- 26 ROGNES, T. *et al.* Vsearch: a versatile open source tool for metagenomics. **PeerJ**, v. 4, p. e2584, 2016.
- 27 BOKULICH, N. A. *et al.* Quality-filtering vastly improves diversity estimates from illumina amplicon sequencing. **Nature Methods**, v. 10, n. 1, p. 57–59, 2013.
- 28 AMIR, A. *et al.* Deblur rapidly resolves single-nucleotide community sequence patterns. **MSystems**, v. 2, n. 2, p. e00191–16, 2017.

-
- 29 KATOH, K.; STANDLEY, D. M. Mafft multiple sequence alignment software version 7: improvements in performance and usability. **Molecular Biology and Evolution**, v. 30, n. 4, p. 772–780, 2013.
- 30 PRICE, M. N.; DEHAL, P. S.; ARKIN, A. P. Fasttree 2—approximately maximum-likelihood trees for large alignments. **PloS One**, v. 5, n. 3, p. e9490, 2010.
- 31 MCDONALD, D. *et al.* The biological observation matrix (biom) format or: how i learned to stop worrying and love the ome-ome. **Gigascience**, v. 1, n. 1, p. 2047–217X, 2012.
- 32 LOZUPONE, C. A. *et al.* Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. **Applied and Environmental Microbiology**, v. 73, n. 5, p. 1576–1585, 2007.
- 33 RENNIE, J. D. *et al.* Tackling the poor assumptions of naive bayes text classifiers. *In*: INTERNATIONAL CONFERENCE ON MACHINE LEARNING (ICML-03). 20th, 2003. **Proceedings[...]** Washington, DC: 2003, p. 616-623.
- 34 MCDONALD, D. *et al.* An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. **ISME Journal**, v. 6, n. 3, p. 610–618, 2012.
- 35 DESANTIS, T. Z. *et al.* Greengenes, a chimera-checked 16s rRNA gene database and workbench compatible with arb. **Applied and Environmental Microbiology**, v. 72, n. 7, p. 5069–5072, 2006.
- 36 ROBESON, M. S. *et al.* Rescript: Reproducible sequence taxonomy reference database management. **PLoS Computational Biology**, v. 10, n. 2020.10, p. 05.326504, 2020.
- 37 BOKULICH, N. A. *et al.* Optimizing taxonomic classification of marker-gene amplicon sequences with qiime 2’s q2-feature-classifier plugin. **Microbiome**, v. 6, n. 1, p. 1–17, 2018.
- 38 WILLIS, A. D. Rarefaction, alpha diversity, and statistics. **Frontiers in Microbiology**, v. 10, p. 2407, 2019.
- 39 SANDERS, H. L. Marine benthic diversity: a comparative study. **American Naturalist**, v. 102, n. 925, p. 243–282, 1968.
- 40 WEISS, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. **Microbiome**, v. 5, n. 1, p. 1–18, 2017.
- 41 COLWELL, R. K.; CODDINGTON, J. A. Estimating terrestrial biodiversity through extrapolation. **Philosophical Transactions of the Royal Society of London, Series B: biological sciences**, v. 345, n. 1311, p. 101–118, 1994.
- 42 GOOD, I. J. The population frequencies of species and the estimation of population parameters. **Biometrika**, v. 40, n. 3-4, p. 237–264, 1953.
- 43 SIMPSON, E. H. Measurement of diversity. **Nature**, v. 163, n. 4148, p. 688–688, 1949.

- 44 SHANNON, C. E. A mathematical theory of communication. **Bell System Technical Journal**, v. 27, n. 3, p. 379–423, 1948.
- 45 PIELOU, E. C. The measurement of diversity in different types of biological collections. **Journal of Theoretical Biology**, v. 13, p. 131–144, 1966.
- 46 COUTINHO, L. B. **A influência do índice de confiança, pipeline de clusterização e classificadores em análises de metabarcoding**. Trabalho de conclusão de curso (Bacharelado em Biotecnologia), Universidade Federal do Rio Grande do Sul, Porto Alegre, 2019.
- 47 LOZUPONE, C.; KNIGHT, R. Unifrac: a new phylogenetic method for comparing microbial communities. **Applied and Environmental Microbiology**, v. 71, n. 12, p. 8228–8235, 2005.
- 48 KHELAIFIA, S. *et al.* Desulfovibrio piezophilus sp. nov., a piezophilic, sulfate-reducing bacterium isolated from wood falls in the mediterranean sea. **International Journal of Systematic and Evolutionary Microbiology**, v. 61, n. 11, p. 2706–2711, 2011.
- 49 ALAZARD, D. *et al.* Desulfovibrio hydrothermalis sp. nov., a novel sulfate-reducing bacterium isolated from hydrothermal vents. **International Journal of Systematic and Evolutionary Microbiology**, v. 53, n. 1, p. 173–178, 2003.
- 50 BIDDLE, A. *et al.* Untangling the genetic basis of fibrolytic specialization by lachnospiraceae and ruminococcaceae in diverse gut communities. **Diversity**, v. 5, n. 3, p. 627–640, 2013.
- 51 YUTIN, N.; GALPERIN, M. Y. A genomic update on clostridial phylogeny: G ram-negative spore formers and other misplaced clostridia. **Environmental Microbiology**, v. 15, n. 10, p. 2631–2641, 2013.
- 52 SPRING, S. *et al.* New insights into the energy metabolism and taxonomy of deferribacteres revealed by the characterization of a new isolate from a hypersaline microbial mat. **Environmental Microbiology**, v. 24, n. 5, p. 2543–2575, 2022.
- 53 STACKEBRANDT, Erko *et al.* The family Propionibacteriaceae: genera other than Propionibacterium. **Prokaryotes: actinobacteria**, p. 725–741, 2014.
- 54 FRASSETTO, Lynda A. Corynebacterium infections. eMedicine. November 18, 2004.
- 55 WILLEMS, A. The family comamonadaceae. *In*: ROSENBERG, E., DELONG, E.F, LORY, S; STACKEBRANDT, E.; THOMPSON, F. (eds) **The Prokaryotes**. Berlin: Springer, 2014.
- 56 WELTE, D. H.; TISSOT, P. B. **Petroleum formation and occurrence**. New York: Springer-Verlag, 1984.
- 57 TSESMETZIS, N. *et al.* Microbial community analysis of three hydrocarbon reservoir cores provides valuable insights for the assessment of reservoir souring potential. **International Biodeterioration & Biodegradation**, v. 126, p. 177–188, 2018.