UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS MATEMÁTICAS E DA NATUREZA
INSTITUTO DE QUÍMICA

ARACELI DE SOUSA PIRES

DIVERSIDADE MICROBIANA NA ÁGUA DE PRODUÇÃO DE UM CAMPO DE EXPLORAÇÃO PETROLÍFERA

Rio de Janeiro
2019

ARACELI DE SOUSA PIRES

DIVERSIDADE MICROBIANA NA ÁGUA DE PRODUÇÃO DE UM CAMPO DE EXPLORAÇÃO PETROLÍFERA

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Bioquímica, Instituto de Química, Universidade Federal do Rio de Janeiro, como requisito parcial à obtenção do título de Mestre em Ciências.

Orientadoras: Bianca Cruz Neves
Graciela Maria Dias

Rio de Janeiro
2019

ARACELI DE SOUSA PIRES

DIVERSIDADE MICROBIANA NA ÁGUA DE PRODUÇÃO DE UM CAMPO DE EXPLORAÇÃO PETROLÍFERA

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Bioquímica, Instituto de Química, Universidade Federal do Rio de Janeiro, como requisito parcial para a obtenção do título de Mestre em Ciências.

Rio de Janeiro, 14 de Março de 2019.

_____
Prof. Dra. Bianca Cruz Neves
Instituto de Química, Universidade Federal do Rio de Janeiro (UFRJ)

_____
Dra. Graciela Maria Dias
Instituto de Química, Universidade Federal do Rio de Janeiro (UFRJ)

_____
Prof. Dr. Rodrigo Volcan Almeida
Instituto de Química, Universidade Federal do Rio de Janeiro (UFRJ)

_____
Prof. Dra. Eliana Flávia Camporese Sérvulo
Escola de Química, Universidade Federal do Rio de Janeiro (UFRJ)

_____
Prof. Dra. Gizele Duarte Garcia
Universidade Federal do Rio de Janeiro (UFRJ) - Campus Macaé

Dedicatória

À minha família, vocês me moldaram na mulher que sou e por vocês dou meu melhor.

Agradecimentos

Eu tenho um coração agradecido e feliz, por mais essa conquista. Não foram dias fáceis e não poucos os dias em que este computador experimentou minhas lágrimas. Muitos desafios tiveram que ser superados, meu orgulho inclusive. Mas aqui estou, digitando essas páginas enquanto recordo com emoção esses dias tão complicados, mas que sem eles não teria aprendido tanto. Entendo que este momento é bastante particular e assim o farei.

Agradeço a Deus, autor da vida, a quem entrego toda a minha devoção e adoração. Agradeço à minha família pelo apoio e suporte – sem vocês não conseguiria; aos meus pais pelo exemplo dado de trabalho e dedicação, não conheço pessoas mais teimosas, mas a perseverança de vocês é o que os faz avançar em meio a tantas lutas. À minha mãe, pelas orações, pelas broncas, pelo amor incondicional, ao meu pai por ser incansável, por sempre me mimar sem deixar de dar aquele puxão de orelha! Ao meu irmão, pelos momentos de companheirismo, risos e choros compartilhados, séries "maratonadas" e segredos só nossos. À minha avó Alzelina, minha segunda mãe, exemplo de fé e disciplina, que tem sempre o braço acolhedor. Aos demais parentes, primos e tios que não terei como citar nomes pois somos muitos e mesmo assim somos amados.

Agradeço aos meus amigos, que em diversos momentos proporcionam o escape e alívio da rotina diária. Mesmo eu sendo uma pessoa dispersiva, vocês me aceitam como sou, sou grata por isso. Muito obrigada às minhas amigas da infância para a vida, Kamylla, Priscila e Joyce, sempre as carrego comigo. Obrigada, Lis, Hadassa, Mary, Gisela e Sara, não foram poucos os momentos que dividimos juntas, em confidências, risos, orações. Obrigada Thamiris, Taissa, Natália, Priscila, Bianca e Thayany, vocês são fofas e muito amadas, nossos encontros e fugas pelo corredor (fuga, que fuga?) sempre carregarei comigo, mesmo nunca saindo o encontro fora da faculdade.

Aos queridos do Laboratório LaMMP, mesmo os que não integram mais o grupo muito obrigada pelo crescimento. Aos professores Rodrigo Volcan e Márcia, que me aceitaram no grupo tão bem, à Michele, muito mais que colega de bancada e carona certa, obrigada por me ouvir, posso dizer que somos amigas, mesmo você gostando tanto de pagode e tendo opiniões tão contrárias às minhas. Camilla, fique pela França, o laboratório é tão quieto sem você por perto. Gabi, amor em pessoa, obrigada por estar sempre disponível e não ter tempo ruim. Giulia, obrigada por me introduzir ao admirável mundo novo, do qual não quero mais

sair. Juliana, obrigada pelos abraços matutinos, desenhos incríveis e pela risada bizarra do dia. Rosane, obrigada pelas conversas e conselhos, por provar que é possível ter fé e ser cientista, você é exemplo até sem falar nada. Isabelle obrigada por conselhos, conversas, por me aturar falar ao pé do ouvido e mesmo assim não ficar zangada comigo – ao menos não externamente. Dani, você é uma fofa, que me ilumina desde a graduação, muito obrigada por existir! Wilber, obrigada por ser essa enciclopédia ambulante, aprendi tanto contigo. Rafael, Vinícius, Francisco, meus companheiros no café, obrigada pelos bons momentos que dividimos. Cândida, por sua risada inconfundível e pelos bolos maravilhosos! Ao Tiago pelas conversas sem fim e aulas de Biologia Molecular.Aos agregados, que sempre aparecem para usar alguma coisa e acabam deixando um pouco de si também.

Às minhas orientadoras, Graciela você é mais que uma orientadora e amiga de bancada, professora, somos amigas e confidentes. Obrigada por, com paciência, me ensinar o caminho dos bioinformatas, a tela preta do computador e o vermelho pulsante do seu coração enorme. À professora Bianca, obrigada por apostar em mim, por me defender como uma leoa no momento que mais precisei, pelas conversas, conselhos, puxões de orelha, caronas, cultura, pela visão de mundo, por me desafiar a fazer mais e ir além. Por ensinar o significado da elegância e da classe, típicos de uma lady!

Agradeço muitíssimo aos membros da banca, que tão prontamente aceitaram meu convite, suas colações e opiniões são muito importantes para mim. Às professoras externas, Eliana Flávia e Gizele Garcia. Aos professores Anderson Pinheiro e Rodrigo Volcan.

Agradeço ao CENPES – Petrobras, por ter cedido a amostra e por financiar o projeto. À pessoa do Rubens Akamine, pelas trocas de informações e por tão solicitamente colaborar com esse trabalho.

Agradeço ao Instituto de Química, minha casa, lugar em que me graduei e concluo o mestrado. Passei por maus bocados, mas sobrevivi e estou mais forte por isso. Agradeço ao Departamento de Bioquímica.

Agradeço à Capes, pela bolsa de estudo e auxílio financeiro, assim como à FAPERJ.

"E ainda que tivesse o dom de profecia, e conhecesse todos os mistérios e toda a ciência, e ainda que tivesse toda a fé, de maneira tal que transportasse os montes, e não tivesse amor, nada seria."

1 Coríntios 13:2, versão Almeida Corrigida Fiel

"Se você conhece o inimigo e conhece a si mesmo, não precisa temer o resultado de cem batalhas. Se você se conhece mas não conhece o inimigo, para cada vitória ganha sofrerá também uma derrota. Se você não conhece nem o inimigo nem a si mesmo, perderá todas as batalhas."

A arte da Guerra, Sun Tzu

RESUMO

Um grave problema de saúde e segurança no trabalho, dentro de uma plataforma de petróleo, é decorrente da comunidade microbiana (microbioma) que cresce em tanques de armazenamento e tubulações. Este microbioma tende a formar biofilmes altamente corrosivos, que obstruem tubulações e possuem resistência a antimicrobianos, além de abrigar muitas espécies produtoras de sulfeto. Portanto, é de extrema importância para a indústria compreender a diversidade filogenética deste microbioma para que medidas de controle e intervenção específicas possam ser devidamente planejadas. Um dos principais genes usados para identificação taxonômica de bactérias e archaeas é o gene do rRNA 16S. O objetivo deste trabalho é caracterizar e comparar as regiões hipervariáveis do gene rRNA 16S da comunidade microbiana existente na água de produção de um tanque de armazenamento da plataforma P-37 da PETROBRAS, a fim de conhecer a sua diversidade. O DNA genômico foi extraído de uma amostra de água e submetido à amplificação pela PCR, com *primers* customizados para amplificar diferentes regiões do gene. Os produtos amplificados foram sequenciados por dois tipos de sequenciadores *Ion Torrent* (Thermo Fisher Scientific) e *HiSeq 2500* (Illumina). Como os protocolos para análise deste tipo de resultado estão em constante evolução, precisou-se fazer uma análise de qual fluxo de tratamento de dados seria mais apropriado para o caso em questão. Avaliou-se o programa a ser utilizado, a métrica para criação de unidades taxonômicas operacionais e o banco de dados para a atribuição da taxonomia. Foram gerados um total de 265.392 *reads* pela plataforma *Ion torrent* e 139.930 *reads* médios pelo *Illumina.* O número de OTU's criadas variou conforme o método escolhido. Dentre eles, o de referência aberta usando o *Greengenes* como banco de dados apresentou 1.781 e 1.082 OTU's pelo *Ion Torrent* e *Illumina*, respectivamente. Quando comparados os bancos de dados, pelo SILVA há uma diferença de quase 15% a mais de OTU's do que pelo *Greengenes.* Há uma predominância de Proteobactérias que chegam a cerca de 40% do total analisado, cujos gêneros mais abundantes são *Desulfovibrio*,18,36%, *Desulfoplanes*, 14,19%, e *Pelobacter,* 11,28%, todas bactérias redutoras de sulfato. A última análise realizada foi com relação ao perfil funcional do microbioma. Em torno de 1.5% dos genes previstos estão relacionados à degradação de xenobióticos, incluindo os derivados do petróleo. Com base nesses dados, é possível montar o perfil do microbioma encontrado, buscar entender sua dinâmica populacional e, por fim, desenvolver estratégias adequadas ao controle dessa população.

Palavras-Chave: Microbiologia do Petróleo, Sequenciamento, rRNA 16S, NGS, Microrganismos Redutores de Sulfato, Corrosão

ABSTRACT

A serious health and safety problem at work, within an oil platform, is due to the microbial community (microbiome) that grows in storage tanks and pipelines. This microbiome tends to form highly corrosive biofilms, which block pipes and have resistance to antimicrobials, in addition to harboring many species producing sulfide. It is therefore of utmost importance for the industry to understand the phylogenetic diversity of this microbiome so that specific control and intervention measures can be properly planned. One of the main genes used for taxonomic identification of bacteria and archaeas is the 16S rRNA gene. The objective of this work is to characterize and compare the hypervariable regions of the 16S rRNA gene of the microbial community existing in the production water of a storage tank of the P-37 platform of PETROBRAS, in order to know its diversity. Genomic DNA was extracted from a water sample and subjected to PCR amplification, with primers customized to amplify different regions of the gene. The amplified products were sequenced by two types of Ion Torrent (Thermo Fisher Scientific) and HiSeq 2500 (Illumina) sequencers. As the protocols for analysis of this type of result are constantly evolving, it was necessary to make an analysis of which flow of data treatment would be more appropriate for the case in question. It was evaluated the program to be used, the metric for the creation of operational taxonomic units and the database for the attribution of the taxonomy. A total of 265,392 reads were generated by the Ion torrent platform and 139,930 average reads by the Illumina. The number of OTUs created varied according to the method chosen. Among them, the open reference using Greengenes as database had 1,781 and 1,082 OTUs by Ion Torrent and Illumina, respectively. When comparing the databases, by SILVA there is a difference of almost 15% more of OTU's than by Greengenes. There is a predominance of Proteobacteria that reach about 40% of the total analyzed, whose most abundant genera are *Desulfovibrio*, 18,36%, *Desulfoplanes*, 14,19%, and *Pelobacter*, 11,28%, all reducing bacteria of sulfate. The last analysis was performed with respect to the functional profile of the microbiome. About 1.5% of predicted genes are related to the degradation of xenobiotics, including petroleum derivatives. Based on these data, it is possible to assemble the profile of the microbiome found, seek to understand its population dynamics and, finally, to develop adequate strategies to control this population.

Key words: Microbiology, Oil, Sequencing, 16S rRNA, NGS, Reducing Sulphate Microorganisms, Corrosion

Lista de figuras

Figura2.1. Fotografia laminar de um arenito.....19
Figura2.2. Fluxograma do processamento primário dos fluidos.....20
Figura2.3. Ciclo do Enxofre.....24
Figura2.4. Sitrofismo.....25
Figura2.5. Vias metabólicas do Nitrato.....25
Figura2.6. Amplificação da biblioteca de DNA.....32
Figura2.7. Sequenciamento *Single-Cell*.....32
Figura2.8. Estrutura Secundária do rRNA 16S.....34
Figura2.9. Representação de arquivos nos formatos FASTA e fastQ.....38
Figura3.1. Fluxograma de trabalho de diferentes abordagens para pré-tratamento dos dados sequenciados na plataforma *Ion Torrent*.....46
Figura3.2. Fluxograma de trabalho de diferentes abordagens para pré-tratamento dos dados sequenciados na plataforma *Illumina*.....47
Figura4.1. Controle de qualidade do DNA total.....54
Figura4.2. Produtos da PCR.....56
Figura4.3.Alfa diversidade da variação do método de criação de OTU's.....60
Figura4.4 Curvas de rarefação baseadas na diversidade alfa da variação do banco de dados.....63
Figura4.5 Gráficos de abundância relativa da variação do banco de dados.....64
Figura4.6 Alfa diversidade nas plataformas *Ion Torrent* e *Illumina*.....69
Figura4.7 Abundância relativa em nível de gênero nas plataformas *Ion torrent* e *Illumina*.....71
Figura4.8. Gráficos de rarefação e abundância relativa nas múltiplas regiões sequenciadas no *Ion Torrent*.....75
Figura4.9. Beta diversidade pelo índice Unifrac Ponderado nas múltiplas regiões sequenciadas no *Ion Torrent*.....78
Figura4.10. Gráfico de distância calculada de dois a dois, a partir do índice UniFrac não ponderado nas múltiplas regiões sequenciadas no *Ion Torrent*.....78
Figura4.11. Abundância relativa em nível de filo nas múltiplas regiões sequenciadas no *Ion Torrent*..80
Figura4.12. Gráfico de rarefação calculado pelo índice CHAO1 e o diagrama de *Whittaker* nas múltiplas regiões sequenciadas no *Illumina*.....86
Figura4.13. Gráfico de Coordenadas Principais obtidas a partir do cálculo de Beta diversidade pelo índice Unifrac Ponderado nas múltiplas regiões sequenciadas no *Illumina*.....90
Figura4.14. Abundância relativa em nível de filo nas múltiplas regiões sequenciadas no *Illumina*.....91
Figura4.15. Gráfico em pizza da classificação gênica do KEGG.....94
Figura4.16. Gráfico em Barras da contagem dos genes separados por via metabólica que participam 101

Figura4.17 Quantificação das subclasses presentes no metabolismo de xenobióticos.....106
Figura4.18 Quantificação de genes associados ao metabolismo xenobióticos.....107
FiguraS1. Contagem de Genes Relacionados a Doenças Humanas.....151
FiguraS2.Contagem de Genes Relacionados a produção de ATP.....151

Lista de tabelas

Tabela2.1. Vantagens e desvantagens dos Principais Sequenciadores de Nova Geração.....29
Tabela3.1. *Primers* utilizados para amplificação dos genes 16S rRNA e 18S rRNA.....43
Tabela3.2. Arquivo de Mapeamento utilizado em análises de beta diversidade.....51
Tabela4.1. Quantidade de leituras obtidas e filtradas durante o pré-tratamento dos dados, para análise dos dados gerados no *Ion Torrent*.....58
Tabela4.2. Quantidade de leituras obtidas e filtradas durante o pré-tratamento dos dados, para análise dos dados gerados no *Illumina*.....58
Tabela4.3. Percentuais de publicações separados por categorias globais no portal *'web of science'*….67
Tabela4.4. Valores totais das OTU's construídas com 97% de identidade com banco de dados *Greengenes*, na região V4 do gene 16S, no *Illumina* e no *Ion Torrent*.....69
Tabela4.5. Abundância dos gêneros identificados.....71
Tabela4.6. Probabilidade de Descobrimento Pontual para Cada Nova OTU.....75
Tabela4.7. Matriz de compartilhamento das OTU's idênticas.....76
Tabela4.8. Matriz de diversidade beta 77
Tabela4.9. Resumo taxonômico de cada região sequenciada do gene rRNA 16S pelo *Ion Torrent*.....81
Tabela4.10. Coeficiente de correlação e intervalo de confiança calculados par a par, a partir do resumo taxonômico em nível de gênero.....82
Tabela4.11. Probabilidade de Descobrimento Pontual para cada nova OTU nas múltiplas regiões sequenciadas no *Illumina*.....86
Tabela4.12. Matriz de compartilhamento das OTU's idênticas nas múltiplas regiões sequenciadas no *Illumina*.....88
Tabela4.13. Matriz de Beta diversidade nas múltiplas regiões sequenciadas no *Illumina*.....89
Tabela4.14. Resumo taxonômico de cada região sequenciada do gene rRNA 16S pelo *Illumina*.....92
Tabela4.15. Coeficiente de correlação e intervalo de confiança nas múltiplas regiões sequenciadas no *Illumina*.....93
Tabela4.16. Tabela com os dez caminhos mais presentes.....96
Tabela4.17. Tabela contagem dos principais grupos relacionados ao metabolismo energético e suas respectivas vias.....97
Tabela4.18. Dissimilaridade Média Calculada pelos índices Unifrac pondeado e *Jaccard*.....102
TabelaS1. Tabela de Identificação Taxonômica até o nível de gênero, da região 16SV4 sequenciada no Illumina e anotada tanto no banco de dados SILVA quanto no *Greengenes*.....126

TabelaS2. Relação metabolismo entre gene, e número de OTU's totais que possuem o gene.....143
TabelaS3. Tabela de OTU's compartilhadas entre as populações que possuem os genes relacionados às vias de redução de enxofre com as demais.....144
TabelaS4. Tabela de OTU's compartilhadas entre as populações que possuem alguns genes relacionados às vias de síntese de acetato.....145
TabelaS5. Tabela de OTU's compartilhadas entre as populações que possuem alguns genes relacionados às vias de produção de metano.....146
TabelaS6. Tabela de OTU's compartilhadas entre as populações que possuem alguns genes relacionados às vias do metabolismo do nitrogênio.....147
TabelaS7. Organização dos genes previstos pela estrutura de classificação do KEEG.....148
TabelaS8. Outros Subníveis previstos pelo KEGG.....150

Lista de Abreviaturas e Siglas

BMP - *Brazilian Microbiome Project*

CENPES - *Centro de Pesquisa e Desenvolvimento (CENPES), Petrobras*

DGGE - *Denaturing gradient gel eletrocphoresis*

DNA - *deoxyribonucleic acid*

dNTPs – Desoxirribonucleotídeos

EDTA - Ácido Etilenodiamino Tetra-Acético

EMP – Projeto Microbioma da Terra, do inglês *Eath Microbiome Project*

HMP – Projeto Microbioma do corpo Humano, do inglês *Human Microbiome Project*

IQ - Instituto de Química

LaMMP - Laboratório de Microbiologia Molecular e Proteínas

MEOR – Recuperação Melhorada do Petróleo por Microrganismos, do inglês *Microbial enhanced oil recovery*

MIC - Corrosão provocada por microrganismos, do inglês *Microbiologically Influenced Corrosion*

NGS - Nova Geração de Sequenciadores, *Next Generation Sequencing*

pb - pares de bases

PCoA – Análise de Coordanadas Principais, do inglês *Principal Coordinates Analysis*

PCR - Reação em Cadeia da Polimerase, *Polymerase Chain Reaction*

PMG – Máquina de Sequenciamento Pessoal, do inglês *Personal Machine Genome*

RNA - Ácido Ribonucleico

rRNA - Ácido Ribonucleico Ribossomal

rRNA 16S - RNA ribossomal 16S

UFRJ - Universidade Federal do Rio de Janeiro

SUMÁRIO

1. INTRODUÇÃO....15
Objetivo Geral....17
Objetivos Específicos....17
2. REVISÃO BIBLIOGRÁFICA....18
2.1. Microbiologia do Petróleo....18
2.1.1. Classificação microrganismos do petróleo....21
2.1.2. Identificação de microrganismos....26
2.2. Sequenciamento de DNA....27
2.2.1. Plataformas de sequenciamento....27
2.2.2. Metagenômica....31
2.2.3. Sequenciamento em massa por amplificação....32
2.3. Programas para tratamento dos dados....36
2.4. Tratamento dos dados....38
3. MATERIAL E MÉTODOS....41
3.1. Obtenção da amostra....41
3.2. Extração do DNA genômico....41
3.3. Amplificação e sequenciamento de regiões variáveis do gene rRNA 16S e 18S....42
3.3.1. Sequenciamento com o *Ion Torrent*....42
3.3.2. Sequenciamento pelo *Illumina*....44
3.4. Tratamentos dos dados....45
3.4.1. *Ion Torrent*....47
3.4.2. *Illumina*....48
3.4.3. Parâmetros Avaliados....49
3.4.3.1. Montagem das Unidades Taxonômicas Operacionais (OTU)....49
3.4.3.2. Banco de dados para Atribuição Taxonômica....50
3.4.3.3. Comparação dos dados obtidos nas diferentes plataformas....50
3.4.3.4. Região do rRNA 16S sequenciada....50
3.4.4. Análise de Diversidade....51
3.4.5. Predição Funcional....51
4. RESULTADOS E DISCUSSÃO....54
4.1. Caracterização da Amostra....54
4.2. Amplificação e sequenciamento de regiões variáveis do gene alvo rRNA 16S....55
4.3.1. Pré-Tratamento....56
4.3.2. Parâmetros Avaliados....58
4.3.2.1. Montagem das OTU....58
4.3.2.2. Emprego dos Banco de Dados....62

4.3.2.3. Plataformas de Sequenciamento....68
4.3.3. Análise da Diversidade das Regiões Hipervariáveis....74
4.3.3.1. *Ion Torrent*....74
4.3.3.2. *Illumina*....84
4.4. Predição Funcional....93
5. CONCLUSÕES E CONSIDERAÇÕES FINAIS....108
REFERENCIAS....114
ANEXO A – Tabelas Suplementares....126

1. INTRODUÇÃO

Microrganismos são capazes de crescer e proliferar nos mais diversos ambientes. Até mesmo em locais considerados mais inóspitos, já foram relatados alguns tipos de vida, principalmente procarionte. A capacidade que esses organismos têm de sobreviver aos mais adversos ambientes é notável e bem documentada, sejam eles indígenas de tais locais, conferindo-lhes alta especificidade, ou mesmo por possuírem a capacidade de adaptação ao serem introduzidos (BAKER; BANFIELD, 2003; SIERRA-GARCÍA et al., 2014). É o caso, por exemplo, de microrganismos que crescem nos tanques e dutos dentro de plataformas de petróleo, acarretando graves problemas operacionais. Esses microrganismos são oriundos do próprio reservatório explorado ou da água do mar utilizada em processos de recuperação secundária. Como a disponibilidade da matéria orgânica é abundante nesses ambientes, os microrganismos proliferam provocando danos como obstrução, corrosão, chegando até a liberação de gases tóxicos que representam um grande problema de saúde ocupacional. Por diversas vezes, o trabalho nesses locais precisa ser interrompido, seja para descontaminação ou mesmo manutenção/troca dos materiais danificados (BASTIN, 1926; FULLER; SURUDA, 2000; LI et al., 2016b).

Pode-se dizer que cada reservatório é único, em suas características geofisicoquímicas, tanto intrínsecas (e.g. temperatura, profundidade em que se encontra, salinidade e pH) como advindas da ação humana, tais como a forma de produção e tempo de exploração (RÖLING; HEAD; LARTER, 2003). No entanto, trabalhos prévios reportam uma similaridade em determinados grupos microbianos, possibilitando a determinação de um perfil comum entre os reservatórios (MAGOT; OLLIVIER; PATEL, 2000).

Existem diferentes técnicas capazes de identificar os microrganismos. Abordagens dependentes de cultivo, em que são abordadas as características fenotípicas, dependem do emprego de meios seletivos, controle de pressão e perfil metabólico, como a capacidade de utilização de diferentes compostos orgânicos. Tais abordagens permitem não apenas a identificação, mas também a identificação taxonômica destes organismos. No entanto, essa estratégia demanda um longo tempo de resposta, além de não contemplar a maior parte do microbioma local, que é composta por organismos não cultiváveis (KASTER et al., 2009; ORPHAN et al., 2000). Como uma alternativa, surgiram técnicas moleculares, independentes

de cultivo, como as que são baseadas na amplificação de genes alvos pela Reação em Cadeia da Polimerase (PCR), gel desnaturante (DGGE), e mais recentemente o sequenciamento de Nova Geração (NGS) (KUMARASWAMY et al., 2011). A combinação entre as duas estratégias, PCR e NGS, tem ajudado aos cientistas na elucidação de como esses microrganismos são capazes de sobreviver com sucesso a tais ambientes (HAMME; SINGH; WARD, 2003; ORPHAN et al., 2000).

A partir de diferentes métodos de identificação, já foram descobertos diversos grupos de microrganismos que contribuem de formas distintas na comunidade microbiana. Eles podem ser divididos de acordo com a função metabólica que possuem. Dentre esses grupos, o mais conhecido na literatura é o grupo das bactérias redutoras de sulfato (BRS), primeiramente descrito em 1926 (BASTIN, 1926; MAGOT; OLLIVIER; PATEL, 2000). O grupo é caracterizado por possuir um conjunto de genes conservados, responsável pela redução do sulfato (SO_4^{-2}) a sulfeto (S^{-2}) (BARTON; FAUQUE, 2009). Existem membros do domínio *Archaea* que também fazem parte desse grupo (ZVERLOV et al., 2005). Outro grupo bastante conhecido, é o dos microrganismos metanogênicos (LIU; WHITMAN, 2008; MAGOT; OLLIVIER; PATEL, 2000). Estima-se que metade do metano encontrado nos reservatórios têm origem microbiana, portanto esse grupo é de alta importância para a indústria {Formatting Citation}. Microrganismos fermentadores são fundamentais dentro da comunidade microbiana, assim como para aplicações biotecnológicas. Talvez o grupo mais importante seja o fermentativo, pois são capazes de metabolizar as complexas fontes de carbono e transformar em matérias mais simples e de fácil absorção pelos outros grupos. Atualmente esse grupo vem ganhando destaque, dentro da indústria petroquímica, pelo seu potencial biotecnológico (LI et al., 2017; YOUSSEF; ELSHAHED; MCINERNEY, 2009). Além desses, há redutores de nitrato, ferro III, iodeto, magnésio, perclorato, todos vivendo em equilíbrio sejam por interações simbiônticas ou competitivas com os demais (LIEBENSTEINER et al., 2014; MAGOT; OLLIVIER; PATEL, 2000; YOUSSEF; ELSHAHED; MCINERNEY, 2009).

Com o avanço da tecnologia e do poder de processamento dos dados torna-se possível realizar o sequenciamento em massa de diversos microbiomas, inclusive com detecção da biosfera rara. No entanto, o NGS, possui uma limitação, que é o comprimento das sequências a serem analisadas. Para análises *shotgun*, em que todo o DNA é sequenciado, é necessária

uma etapa de fragmentação. Já em análises do tipo *target*, na qual um gene alvo é utilizado, é necessária uma etapa prévia de amplificação. Geralmente é escolhida uma região com alto valor de sensibilidade filogenética para ser sequenciada, sendo que e a discussão na literatura é extensa a esse respeito (AMARAL-ZETTLER et al., 2009; CRUAUD et al., 2014; HUSE et al., 2008; MARTINEZ-PORCHAS et al., 2017; YANG; WANG; QIAN, 2016; YU et al., 2008). O gene 16S rRNA tem sido extensivamente usado da identificação taxonômica por tratar-se de um dos elementos mais conservados entre todos os domínios (GOOD, 1953; WARD; WELLER; BATESON, 1990). O gene ribossomal 16S codifica o RNA 16S, presente na subunidade ribossomal menor (30S).

A abordagem escolhida para esse estudo foi a identificação da comunidade microbiana através do sequenciamento de múltiplas regiões do gene 16S rRNA por duas tecnologias de NGS distintas: *Hi-Seq* (*Illumina*) e *Personal Magine Genome* (PMG) *Ion Torrent* (*Thermo*). A amostra foi retirada de uma plataforma de petróleo localizada no campo de Marlim, Bacia de Campos do Goytacazes, litoral norte do estado do Rio de Janeiro. A plataforma foi escolhida devido a problemas recorrentes de altos níveis de ácido sulfídrico (H_2S) e biocorrosão. Para isso, foi necessário uma análise comparativa das duas tecnologias de sequenciamento, o que inclui também o tratamento dos dados por bioinformática.

Objetivo Geral

Identificar a comunidade microbiana por sequenciamento do gene ribossomal 16S.

Objetivos Específicos

Estabelecer um protocolo para identificação da comunidade; comparando as metodologias existentes;

Comparar a especificidade das regiões do gene 16S rDNA sequenciadas;

Comparar os dados obtidos pelas plataformas NGS utilizadas;

Identificar grupos específicos que causem biocorrosão;

Identificar populações específicas dentro do microbioma;

Propor meios de combate à corrosão provocada por microrganismos.

2. REVISÃO BIBLIOGRÁFICA

2.1. Microbiologia do Petróleo

O petróleo é o resultado do acúmulo da matéria orgânica depositada no fundo de lagos e mares, que após anos de decomposição forma uma mistura complexa de hidrocarbonetos, em que o estado físico depende do tamanho das moléculas. Além do carbono e hidrogênio, podem existir outros constituintes como enxofre, nitrogênio, oxigênio e alguns metais. A composição dos reservatórios é basicamente invariável, o que muda é a abundância relativa de cada substância. Isto faz com que as características físico-químicas de cada reservatório seja única. O óleo fica aprisionado no interior de uma rocha porosa envolta por uma rocha impermeável, rocha selante. O conjunto de rochas é denominado jazida (PIRES, 2015; THOMAS, 2001).

Os hidrocarbonetos podem ser classificados de acordo com a sua estrutura química e em relação ao tipo de ligação: saturados e insaturados; ou ao tipo de cadeia: se aberta ou fechada; e ainda quanto a aromaticidade: se aromáticos ou não (JOHN MCMURRY, 2016). Na petroquímica, os compostos orgânicos são classificados em:

- Parafínicos, alcanos (hidrocarbonetos saturados) de cadeia fechada ou aberta, com ou sem ramificações;
- Olefínicos, hidrocarbonetos insaturados como alcenos e alcinos, os quais são altamente reativos;
- Aromáticos, que são hidrocarbonetos de cadeia fechada com insaturações intercaladas formando uma estrutura de ressonância conhecida como anel aromático, sendo o benzeno seu composto mais simples;
- Resinas e asfaltenos, os quais são as frações mais pesadas, com presença de heteroátomos, cuja diferença entre as duas a solubilidade no óleo, em que o primeiro aparece dissolvido, e o segundo grupo, na forma de coloide (SZKLO; ULLER; BONFÁ, 2012; THOMAS, 2001).

Além do óleo, a rocha reservatório pode conter gás natural, hidrocarbonetos com até seis átomos de carbono, livres ou dissolvidos no óleo. Com frequência também ocorre a presença de água, denominada *água conata*, a qual pode estar emulsionada ou livre de óleo como um aquífero no interior. Outros compostos, como alguns metais, os quais são considerados impurezas também ocorrem, além de material particulado em suspensão,

podendo ser proveniente da rocha, de processos corrosivos e incrustações. Além disso, existe a presença de microrganismos que podem ter origem endógena, ou exógena. São considerados exógenos aqueles que foram introduzidos por consequência da ação antropológica, seja no processo de perfuração, produção ou recuperação do poço. A Figura 2.1 trás o corte de uma rocha reservatório, uma fotografia em lâmina delgada de uma rocha reservatório, por Irene Gabardo em 2007, e um esquema generalizado de uma jazida com óleo, água e gás. Na fotografia é possível observar a rocha, o óleo em preto e marcando os poros um corante azul, espaço ocupado por água ou gás.

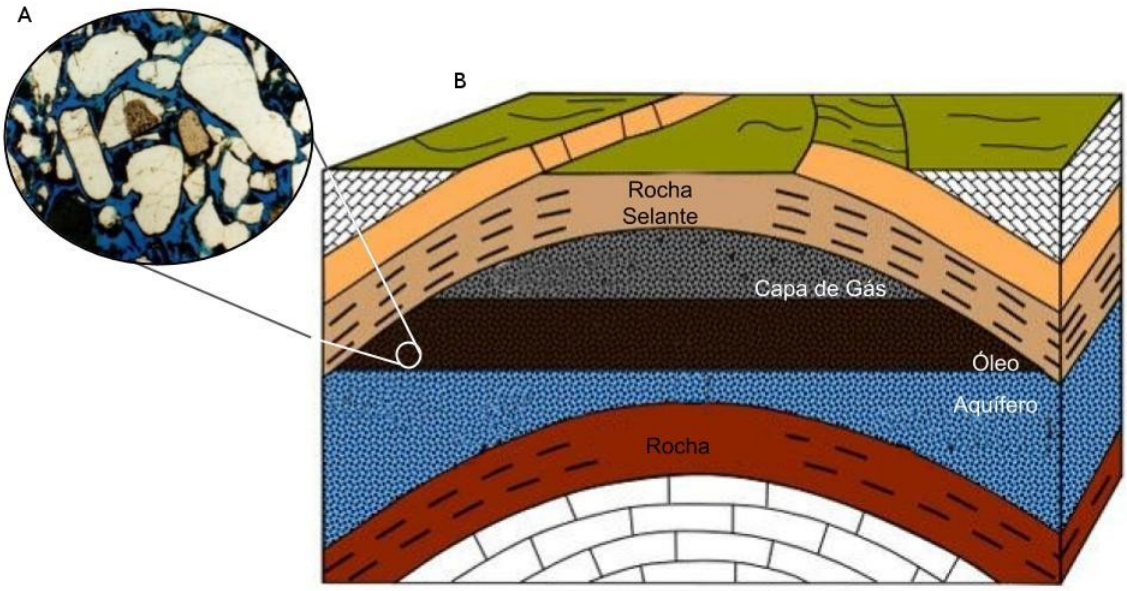

Figura 2.1. Fotografia laminar de um arenito. (A) No destaque, em branco a rocha, em azul corante marcando os poros e em preto o óleo; a direita **(GABARDO, 2007)**; (B) a esquerda uma reapresentação esquemática de uma jazida que possui capa de gás e aquífero adaptado da internet **(PETROPEDIA, 2018)**.

Antes de ser refinado, na própria plataforma em que é produzido o óleo é processado. O procedimento é chamado de processamento primário dos fluidos, cujo objetivo é separar os fluidos que são produzidos, que se encontram misturados. Além disso, deve-se condicionar o óleo e prepará-lo para o transporte para a refinaria, e tratar a água produzida para ser descartada ou reinjetada na poço. A presença de impurezas nos fluidos dificulta a separação, pois agem como agentes emulsificantes, requerendo etapas adicionais. A Figura 2.2 trás o

esquema simplificado do processamento primário de petróleo adaptado do livro “Fundamentos da Engenharia de Petróleo”, de José Eduardo Thomas (OLLIVIER; MAGOT, 2005; SZKLO; ULLER; BONFÁ, 2012; THOMAS, 2001). Etapas adicionais no processamento, quando associadas ao elevado tempo de residência nos tanques e dutos, juntamente à abundância de matéria orgânica em condições favoráveis de temperatura e pressão, proporcionam o ambiente perfeito para a formação de biofilmes bacterianos nos dutos e tanques de armazenamento da plataforma (OLLIVIER; MAGOT, 2005; XIAO et al., 2016b; YOUSSEF; ELSHAHED; MCINERNEY, 2009).

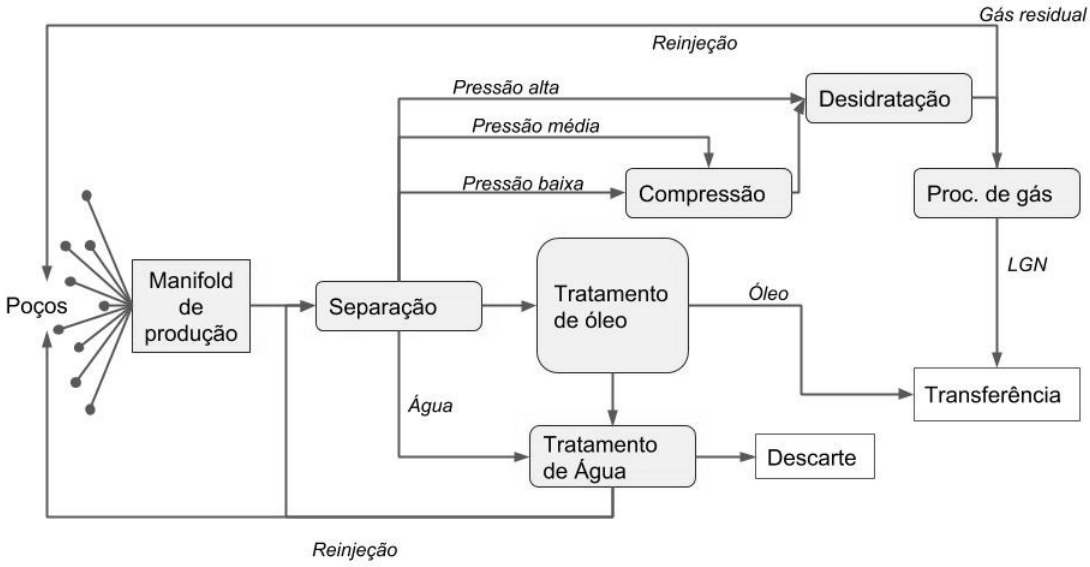

Figura 2.2. Fluxograma do processamento primário dos fluidos. O óleo é extraído do reservatório por poços de produção. O *manifold* de produção é reunião do óleo extraído em um único local, de lá segue para os tanques de separação, em que cada fase irá para um compartimento de tratamento próprio. Adaptado de Thomas, 2004.

Durante anos, acreditou-se que certos ambientes, como um reservatório, eram inóspitos, por não apresentarem as condições para o crescimento de organismos conhecidos à época. Credita-se a Bastin, em 1926, o primeiro trabalho sobre um microrganismo isolado a partir de um poço de petróleo, sendo este uma bactéria redutora de sulfato (BASTIN, 1926). Durante certo tempo, diziam que tais organismos deveriam ser exógenos, introduzidos no ambiente pela ação do homem. Desde então, diversos estudos foram realizados, nos quais

buscam identificar a origem desses organismos, seu perfil metabólico e seus mecanismos adaptativos em ambientes tão distintos; seja por sua capacidade de degradar o óleo, ou por provocar prejuízos como acidificação, corrosão, encrustação ou entupimento. Não apenas por seus prejuízos mas também pela potencialidade em conhecer e entender a dinâmica dessa comunidade, a fim de melhorar a produção ou usá-la de forma a remediar possíveis acidentes. A área ficou conhecida como Microbiologia do Petróleo (MAGOT; OLLIVIER; PATEL, 2000; OLLIVIER; MAGOT, 2005; ORPHAN et al., 2000).

2.1.1. Classificação microrganismos do petróleo

A partir do entendimento de que uma jazida não é um ambiente inóspito e estéril, foi possível isolar diferentes organismos. À medida que seu metabolismo foi conhecido, eles puderam ser categorizados de acordo com suas funcionalidades. O primeiro grande *review* na área é de Michel Magot em 2000, em que aborda estudos realizados com bactérias, endógenas ou exógenas, isoladas de campos de petróleo (MAGOT; OLLIVIER; PATEL, 2000). Em sequência, Wilfred F.M. Röling, em 2003, publica sobre características físico-químicas dos reservatórios dos quais foram encontradas atividades microbianas (RÖLING; HEAD; LARTER, 2003). No mesmo ano, Jonathan D. Van Hamme, divulga seu trabalho, desta vez focando nas descobertas metabólicas e suas aplicações (HAMME; SINGH; WARD, 2003) Em 2005, Michel Magot edita um livro intitulado “Microbiologia do Petróleo”, buscando unir não apenas o conhecimento microbiológico, como também aspectos de engenharia e geologia aplicados à area de petróleo (OLLIVIER; MAGOT, 2005). Mais recentemente, a fim de preencher as lacunas deixadas, Xiao-Xiao Li, em 2017, escreve um artigo não apenas retornando à linha de Michel Magot, destacando os trabalhos da área, como também introduzindo novos conceitos e técnicas mais avançadas de estudo (LI et al., 2017), através dos quais propõe-se compreender a dinâmica das comunidades microbianas e sua diversidade por toda a Terra. A partir de dados gerados desse tipo de estudo, surgem inúmeras aplicações de extrema importância, possibilitando novos estudos e sua correlação com dados já existentes. Torna-se possível categorizar os diferentes grupos de microrganismos de acordo com o tipo de metabolismo, além de correlacionar a presença frequente de determinados grupos, identificando-se comunidades em comuns a diversos reservatórios (MAGOT; OLLIVIER; PATEL, 2000; TARDY-JACQUENOD et al., 1996).. Essas comunidades

microbianas são compostas majoritariamente de anaeróbios e anaeróbios facultativos. Sendo que o anaeróbios restritos são considerados, com segurança, como microrganismos endógenos (MAGOT; OLLIVIER; PATEL, 2000). Nesses grupos, o metabolismo energético emprega como principais receptores de elétrons o sulfato, carbonato, ferro (III), nitrato e oxigênio, apesar desses três últimos serem mais limitados no interior dos reservatórios. Por outro lado, os principais substratos doadores de elétrons são hidrogênio, ácidos graxos voláteis (em destaque o acetato), hidrocarbonetos aromáticos e alifáticos, e sais inorgânicos, como sulfeto (YOUSSEF; ELSHAHED; MCINERNEY, 2009). No entanto os principais limitantes para o crescimento microbiano são os receptores de elétrons (HAMME; SINGH; WARD, 2003), e, com base nessas propriedades, os microrganismos podem ser classificados como: redutores de sulfato, metanogênicos, redutores de nitrato e fermentativos.

A Figura 2.3 ilustra o ciclo do enxofre, adaptado do livro “Diversidade Microbiana” de Oladele Ogunseitan. O enxofre é um dos elementos principais nos sistemas biológicos, presente nos aminoácidos cisteína e metionina e em coenzimas. Além disso, possui uma função importante na homeostase de metais, por sua eletroafinidade que possui o papel de auxiliar na estabilização do complexo formado. Encontrado em diferentes estados de oxidação, variando de 2- a 6+, o enxofre atua na regulação do pH. As reações envolvendo as espécies químicas do enxofre podem ser realizadas por diferentes microrganismos e estão relacionadas na figura 2.3 (OGUNSEITAN, 2014, cap. 8). No óleo cru, o enxofre é o terceiro elemento mais abundante, atrás apenas do carbono e hidrogênio, possuindo uma faixa de concentração de 200 a 80.000 ppm. Sua presença influencia diretamente no aumento da densidade do óleo, sendo o petróleo classificado como azedo quando possui teor de enxofre acima de 25.000 ppm, e doce quando abaixo de 5.000 ppm. Os compostos sulfurados mais abundantes são: mercaptanas, sulfetos, polissulfetos, benzotiofenos, moléculas policíclicas com nitrogênio e oxigênio, ácido sulfídrico (H_2S), dissulfeto de carbono, sulfeto de carbonila e muito raramente como enxofre elementar (SZKLO; ULLER; BONFÁ, 2012).

Microrganismos redutores de sulfato, em especial, ganham destaque pelos danos causados. Foram os primeiros a serem isolados de poços e são encontrados com frequência em uma variedade grande de ambientes. São capazes ainda de metabolizar diferentes fontes de carbono e vivem em sintrofia com fermentadores, sendo organismos obrigatoriamente anaeróbios (HAMME; SINGH; WARD, 2003; MAGOT; OLLIVIER; PATEL, 2000;

OLLIVIER; MAGOT, 2005, cap. 4). Eles são capazes de metabolizar a matéria orgânica mais complexa, transformando em mais simples e que facilmente é utilizada como substrato por outros organismos (LI et al., 2017).

A degradação completa da matéria orgânica em ambiente anóxico leva à geração de metano. Neste contexto, encontram-se os organismos metanogênicos. Existem três vias de redução do carbono ao metano: a partir do dióxido de carbono, usado como doador de elétrons (metanogênese hidrogenotrófica), a partir de compostos de um carbono, como metanol (metanogênese metilotrófica) e por fim a partir da redução do acetato (metanogênese acetilástica). As vias diferem apenas no substrato de entrada, sendo a etapa final comum a todas (MESLÉ; DROMART; OGER, 2013). A via acetilástica é a mais simples e gera menor quantidade de energia, ao passo que a hidrogenotrófica é mais complexa e possui o melhor rendimento (OGUNSEITAN, 2014, cap. 7). A via acetilástica ainda compete com a via de oxidação do acetato sintrófico, em que os produtos finais são H_2 e CO_2.

A degradação da matéria orgânica ocorre por uma cascata metabólica em que vários microrganismos participam (MESLÉ; DROMART; OGER, 2013). O início da degradação acontece através da ação de fermentadores, que produzem: ácidos graxos, ácidos orgânicos (acetato, succinato e lactato), álcoois, dióxido de carbono e hidrogênio. Finalmente estes últimos podem ser reduzidos a metano. Em reservatórios de petróleo, com frequência pode ocorrer a acetogênese, em conexão com oxidação sinérgica de acetato a dióxido de carbono e, por fim, a metanogênese hidrogenotrófica (MAGOT; OLLIVIER; PATEL, 2000; MESLÉ; DROMART; OGER, 2013). Altos teores de metano no reservatório podem ser associados à biodegradação (HAMME; SINGH; WARD, 2003). A Figura 2.4 mostra um esquema resumido adaptado do artigo "*Microbial methanogenesis in subsurface oil and coal*", de Margaux Meslé (MESLÉ; DROMART; OGER, 2013).

As condições físico-químicas das jazidas, como a composição elementar, a biodisponibilidade dos nutrientes, temperatura, pH e salinidade atuam como uma pressão seletiva sobre os microrganismos (OLLIVIER; MAGOT, 2005, cap. 5; RÖLING; HEAD; LARTER, 2003). O nitrogênio é o quarto elemento mais abundante no reservatório. A maior parte dos óleos crus possui menos de 2.000 ppm de nitrogênio. No entanto, o óleo produzido na bacia de Campos possui níveis acima dessa média. Os compostos orgânicos nitrogenados mais comuns são piridinas, quinolinas, pirróis, compostos policíclicos com oxigênio, enxofre

e metais. Eles atuam como emulsificantes dificultando a separação água/óleo e, assim como os compostos sulfurados, aumentam a densidade do óleo (ZUGAIB; ULLER; BONFÁ, 2016). A presença de organismos redutores de nitrato em comparação com os outros grupos de organismos é bastante tímida. Esse grupo é favorecido quando no processo de recuperação é usada injeção de nitrato. Já foram isolados de reservatórios organismos mesofílicos, microaerófilos, aeróbicos, anaerobióticos e facultativos, que oxidam nitrogênio, e comumente usam acetato(OLLIVIER; MAGOT, 2005, cap. 5). A Figura 2.5 mostra os caminhos metabólicos que podem ser favorecidos com a adição de nitrato, adaptado do artigo "*Microbial nitrate respiration - Genes, enzymes and environmental distribution*", de Beate Kraft (KRAFT; STROUS; TEGETMEYER, 2011).

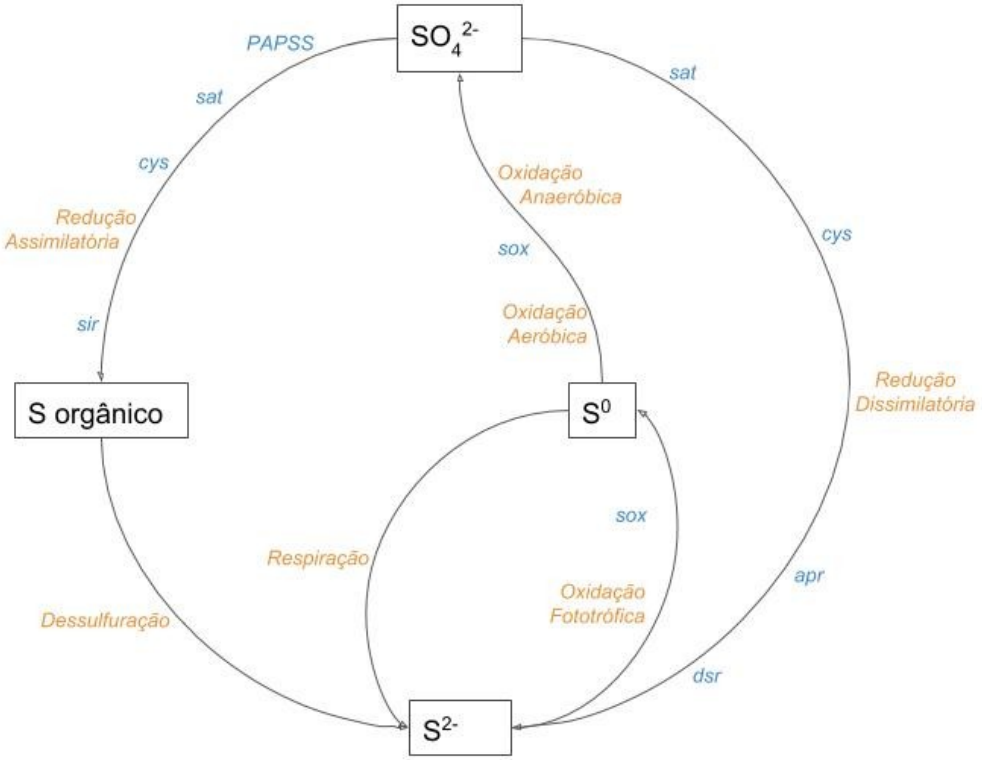

Figura 2.3. Ciclo do Enxofre. Em preto estão assinaladas as espécies de enxofre, em laranja o nome da via de transformação, e em azul os genes envolvidos na respectiva via. O KEGG aponta três vias envolvidas no ciclo do enxofre, Redução Assimilatória, Redução Dissimilatória e Oxidação.

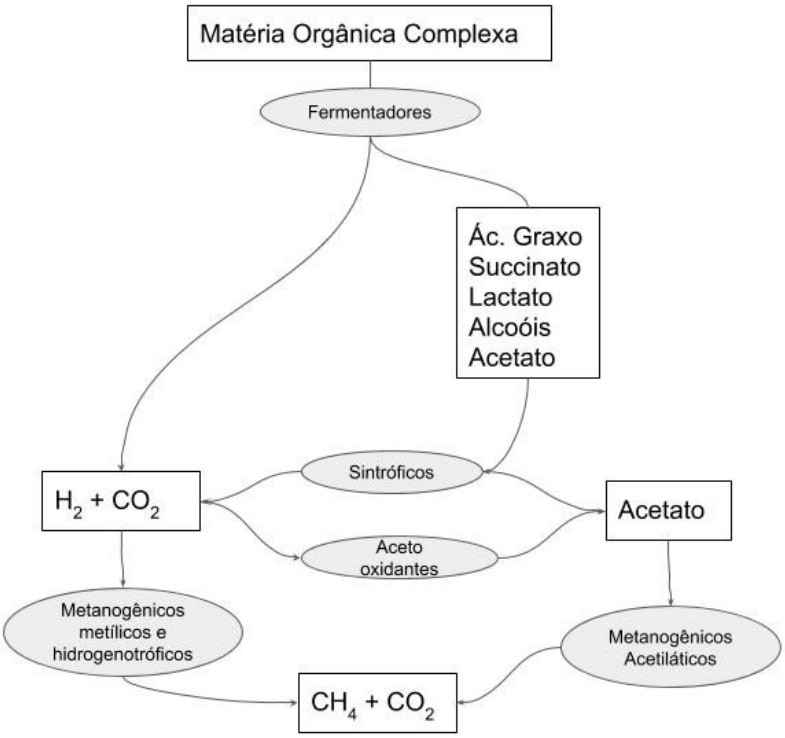

Figura 2.4. Sitrofismo. Esquema de degradação total da matéria orgânica complexa realizada por multiplos organismos em ambientes anoxidos Adaptado Margaux Meslé, 2013 (MESLÉ; DROMART; OGER, 2013)**.**

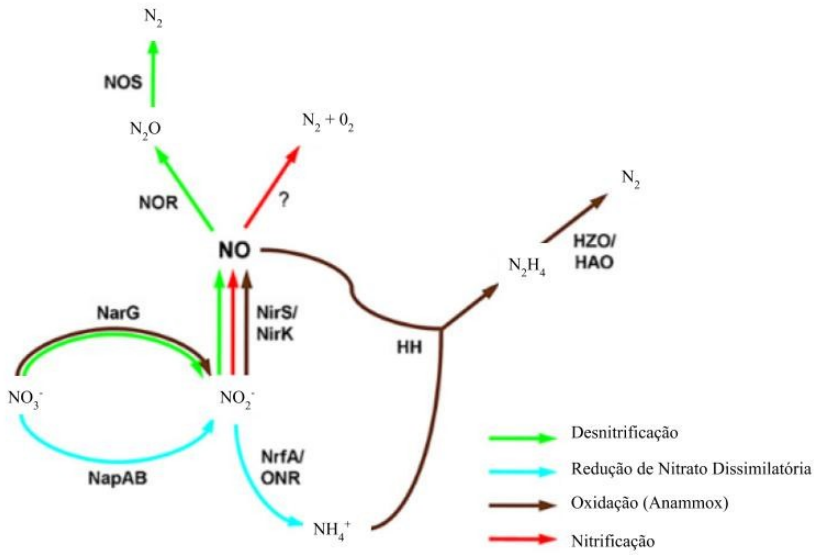

Figura 2.5. Vias metabólicas do Nitrato. As setas coloridas indicam as vias envolvidas no metabolismo do nitrogênio, em verde denitrificação, em ciano redução dissimilatória, em marrom oxidação e em vermelho nitrificação. As espécies de Nitrogênio são indicadas no ponto final e em negrito as enzimas envolvidas em cada transformação, o ponto de interrrogação as ainda não caracterizadas. Adaptado Beate Kraft, 2011 (KRAFT; STROUS; TEGETMEYER, 2011).

2.1.2. Identificação de microrganismos

Ao passo que o conhecimento sobre a microbiologia do petróleo avança, surgem novas técnicas de identificação de microrganismos. Estima-se que menos de um por cento dos microrganismos existentes na Terra sejam cultiváveis. A complexidade de determinados ambientes, como jazidas, solos e amostras de água, tornam desafiador o estudo dessas comunidades (STREIT; SCHMITZ, 2004).

O conhecimento sobre as comunidades microbianas e suas interações caminharam em um ritmo lento, devido principalmente aos estudos que dependiam do cultivo de microrganismos para a caracterização das espécies. Apesar dessa abordagem ser essencial na descoberta de novas espécies e sua caracterização, ela não responde sobre a diversidade do ambiente, sua abundância e significância (ORPHAN et al., 2000).

Os métodos dependentes de cultivo baseiam-se em características fenotípicas, como a capacidade em crescer em determinadas condições; caracterização de acordo com a morfologia, metabolismo e fisiologia. E por mais que as condições variem, a maior parte da comunidade permanece incultivável. Como fator positivo, essas abordagens são as melhores na caracterização de novas espécies. Já os métodos independentes de cultivo, podem passar desde técnicas baseadas em colorimetria para a identificação de grupos de microrganismos metabolicamente ativos a técnicas mais avançadas e dependentes de genética molecular, como eletroforese em gel gradiente de desnaturação e temperatura (DGGE), análise de espaçadores intergênicos ribossômicos, polimorfismo de conformação de fita única, polimorfismo de comprimento de fragmento de restrição transcrito interno, DNA polimórfico amplificado ao acaso e análise de restrição de DNA ribossomal amplificado, *microarray*, sequenciamento de fragmentos de DNA, por técnicas baseadas em clonagem e mais recentemente sequenciamento do DNA total obtido das comunidades. Se por um lado as técnicas independentes de cultivo proporcionaram uma visão global da comunidade microbiana, estas geram um grande volume de resultados, sendo necessárias ainda, em muitos casos, as comparações com bases de dados criadas a partir de métodos baseados em cultivo. Não se pode, portanto, preferir uma abordagem em detrimento da outra, sendo sempre necessário que ambas avancem (HAMME; SINGH; WARD, 2003; RÖLING; HEAD; LARTER, 2003; YOUSSEF; ELSHAHED; MCINERNEY, 2009).

2.2. Sequenciamento de DNA

2.2.1. Plataformas de sequenciamento

Em 1965, foi proposto que o efeito cumulativo em erros na replicação genética poderiam gerar distâncias evolutivas entre organismos. Tais distâncias gerariam linhagens diferentes e que poderiam ser rastreáveis pelo estudo de biomoléculas, especificamente ácidos nucleicos e proteínas. Abordagens moleculares têm sido utilizadas no estudo da evolução de espécies e, as inovações tecnológicas proporcionaram a análise de comunidades microbianas ((OGUNSEITAN, 2014, cap. 4).

Na década de 70 surgiram duas tecnologias para determinar a sequência de fragmentos de DNA. A primeira baseada em hidrólise química por Alan Maxam e Walter Gilbert. A segunda, por Frederick Sanger, por reações enzimáticas. Ambas as técnicas utilizam de elementos radioativos para marcar o DNA, ambas necessitavam realizar uma eletroforese em gel para visualizar o elemento radioativo. A diferença entre elas está na marcação, na primeira a molécula de DNA era toda marcada, clivada, e posteriormente, com os fragmentos, realizavam a eletroforese. Já na segunda, uma fita complementar à molde era sintetizada e os nucleotídeos terminais eram marcados, após a reação os fragmentos eram submetidos á eletroforese. Assim, enquanto a primeira técnica era permitido saber a sequência do DNA pelo perfil de clivagem; na segunda, era de acordo com o terminador do fragmento. Outro desenvolvimento tecnológico proporcionou que a técnica de Sanger avançasse mais rápido, a Reação em Cadeia da Polimerase (PCR), que possibilitou a replicação *in vitro* de um fragmento de DNA (FIETTO; MACIEL, 2015).

Lloyd M. Smith e colaboradores propuseram melhorias ao método de Sanger que até então era totalmente manual. A modificação permitiu que o sequenciamento fosse realizado de forma mais rápida e segura. Através da substituição de nucleotídeos radioativos por fluoróforos que emitissem luz em um comprimento de onda conhecido. A partir disso, surge a era dos sequenciadores automatizados, como ABI 377, que utiliza o método descrito acima, permitindo o sequenciamento de até 48 fragmentos de DNA, em cinco a seis horas. Novos aprimoramentos foram realizados, como a substituição do gel em placa por um em capilar, o que proporcionou que as leituras fossem realizadas muito mais rapidamente. O processo ficou cada vez mais automatizado, com a introdução do sistema de eletroinjeção, melhoramentos

nos detectores e nos *softwares* que analisam as sequências. Tais avanços permitiram que fossem sequenciados fragmentos de até 700 pares de bases de comprimento, além da injeção simultânea de 384 amostras em intervalos de tempo cada vez menores (FIETTO; MACIEL, 2015).

O projeto genoma humano foi o grande impulsionador para que novas técnicas e aprimoramentos surgissem (METZKER, 2010). Devido ao tamanho do DNA, a molécula deveria ser fragmentada, para isso se utilizaram enzimas de restrição ou clivagem mecânica aleatória. Na sequência, monta-se uma biblioteca genômica dos fragmentos ligados em vetores específicos. *Primers* projetados para alinhar nos vetores amplificavam e sequenciavam os fragmentos de DNA de interesse, enquanto programas de computadores foram construídos para montar o "quebra-cabeça". Por vezes era necessário desenhar *primers* que anelassem em parte do fragmento sequenciado para se obter toda a sequência do incerto no vetor, a metodologia ficou conhecida como *primer walkin*g (FIETTO; MACIEL, 2015). O projeto foi concluído e, apesar de todas as modificações e inovações, o sequenciamento do DNA ainda era realizado com base na técnica desenvolvida por Sanger. Finalmente, nos anos 2000 uma segunda onda de inovações culminaram nos chamados sequenciadores de nova geração (NGS, do inglês *Next Generation Sequencing*) (FIETTO; MACIEL, 2015; METZKER, 2010).

O primeiro sequenciador pós-Sanger, foi desenvolvido pela Roche em 2001, o 454-Roche, baseado na tecnologia de pirosequenciamento. Em seguida, pela *Applied Biosystems*, o sequenciador *SOLiD*, que utilizava o método *Polony*, e pela Illumina, foi desenvolvido o *Genome Analyser*, através do método de amplificação em ponte. Após a incorporação das empresas *Applied Biosystems* e *Roche*, a Thermo, empresa de biotecnologia, desenvolveu o *Ion Torrent*, que pode ser considerado um *upgrade* do antigo 454 (FIETTO; MACIEL, 2015; METZKER, 2010)). Essas novas tecnologias dispensam a clonagem, mas dependem de muitas etapas, como preparo da amostra, que consiste na fragmentação, seleção dos tamanhos próprios, ligação aos adaptadores, e da PCR, sequenciamento e análise dos dados gerados. A análise de dados é realizada por *softwares* específicos. Aliado aos novos sequenciadores, há novos computadores com capacidade de processamento de dados cada vez maior, o que permitiu não apenas que mais amostras fossem analisadas como também que se aumentasse a complexidade das mesmas (METZKER, 2010). A Tabela 2.1 mostra as diferenças entre as

tecnologias apontadas, com suas vantagens e desvantagens.

Tabela 2.1. Vantagens e desvantagens dos Principais Sequenciadores de Nova Geração

Plataforma	Preparo da Amostra	Sequenciamento	Vantagem	Desvantagem
Roche-454	PCR em emulsão	Emissão de luz pela luciferase na incorporação do pirofosfato liberado na adição de um nucleotídeo correspondente ao sequenciamento	Leituras mais longas melhoram o mapeamento em regiões repetitivas; tempos de execução rápida	Alto custo de reagente, e taxas de erro com homopolímero
Illumina	PCR em ponte na *flowcell*	Liberação de um fluoróforo na incorporação de um nucleotídeo	Possibilidade de sequenciar ambos os sentidos do fragmento DNA	Tempo de análise e tamanho do comprimento
Ion Torrent	PCR em emulsão	A cada incorporação de uma nova base é realizada a potenciometria do meio, medem o próton liberado na incorporação de um novo nucleotídeo	Menor custo com reagents	Tempo de corrida, erros por repetições, comprimentos curtos

O preparo da amostra para NGS é basicamente o mesmo para as diferentes plataformas. Se por um lado essas tecnologias proporcionaram o sequenciamento mais rápido, eficiente e com um custo reduzido, por outro o tamanho do fragmento precisa ser menor. Portanto, a primeira etapa consiste na fragmentação do DNA e na seleção dos tamanhos. Em seguida, é realizada a montagem da biblioteca, que é uma etapa de ligação aos adaptadores e posteriormente a amplificação da mesma por PCR. Na plataforma 454, da Roche, esta etapa inclui a PCR em emulsão. Há dois adaptadores, o primeiro (adaptador B) é usado na

ancoragem do fragmento a uma micropartícula magnética esférica, chamada de *bead*. Uma emulsão é formada, de tal forma que o interior da micela contenha uma polimerase, uma *bead* ligada a um único fragmento do DNA e os nucleotídeos, para que a reação de amplificação ocorra. No final da PCR em emulsão existirão várias *beads* cobertas por cópias de um mesmo fragmento de DNA. As microesferas são então depositadas no interior de uma placa com milhares de poços, o tamanho deles é tal que apenas uma ocupe seu lugar. Na *Illumina*, ocorre a PCR em ponte, em que ambos os adaptadores, P1 e P2, são ancorados por hibridização a uma placa de vidro, chamada de *flowcell*. Nucleotídeos não marcados são adicionados e a amplificação acontece, com a formação de uma fita complementar, na qual uma extremidade encontra-se livre e a outra ancorada. A parte livre forma uma nova ponte, ancorando-se na *flowcell*, através do adaptador, e o processo é repetido. Ao final de 35 ciclos, na superfície da placa tem-se a formação de zonas com aglomerados do mesmo fragmento, chamadas de *clusters*. De forma semelhante à 454 a PMG utiliza a metodologia da PCR em emulsão na amplificação da biblioteca. A Figura 2.6 esquematiza os dois processos acima citados (FIETTO; MACIEL, 2015; HEATHER; CHAIN, 2016; METZKER, 2010).

A evolução crescente de novas tecnologias levou rapidamente ao surgimento de sequenciadores de terceira geração. Estes não precisam de etapas de fragmentação do DNA, sendo capazes de sequenciar a molécula por completo de uma única vez, método que ficou conhecido como *single cell*. O primeiro sequenciador com essa abordagem no mercado foi o PacBio, desenvolvido pela *Pacific Biosciences* em 2010, cujo a técnica consiste na utilização de uma DNA polimerase imobilizada no interior de um chip. O sequenciamento ocorre a medida que o nucleotídeo é incorporado a fita complementar havendo liberação de um fluoróforo correspondente (HEATHER; CHAIN, 2016; METZKER, 2010; REUTER; SPACEK; SNYDER, 2015; RHOADS; AU, 2015). Em 2011, a Oxford lançou o sequenciador com tecnologia chamada de NanoPore. Um sequenciador portátil, como o da linha MinION, possui as dimensões de 105x23x33 mm, com entrada USB 3.0. Por esta metodologia, uma única molécula de DNA atravessa uma membrana elétrica através de um poro. A diferença no tamanho dos nucleotídeos faz com que haja uma diferença na amperagem detectada. A Figura 2.7 mostra a ilustração de como é realizado o sequenciamento por essas duas tecnologias. Independente da classificação, ambos os sequenciadores já estão bem estabelecidos no mercado. Os erros inerentes ao sequenciamento longo estão sendo corridos e seu custo

operacional aos poucos vem decaindo (HEATHER; CHAIN, 2016; LIU et al., 2012; LU; GIORDANO; NING, 2016; REUTER; SPACEK; SNYDER, 2015).

2.2.2. Metagenômica

Além da variedade de sequenciadores disponíveis, atualmente há diversas estratégias de análise desenvolvidas para o tipo de pergunta que é realizada. A alta capacidade de processamento dos computadores permitiu que fossem sequenciados ao mesmo tempo o DNA de comunidades inteiras, em uma única operação. O conjunto de análises genômicas de uma única comunidade microbiana é chamada de metagenômica, que compreende estratégias descritivas (metagenômica descritiva) e funcionais (metagenômica funcional). Na primeira, é onde ocorre o sequenciamento do DNA, montagem do metagenoma, identificação das espécies, e anotação gênica. Para alcançar esse objetivo, é escolhida a plataforma, podendo ser por *shotgun*, onde todo o conteúdo do metagenoma é fragmentado e sequenciado, ou *target*, conhecido como sequenciamento em que um gene alvo é amplificado por PCR, sendo posteriormente sequenciado. Esta segunda estratégia não é considera metagenômica. Isto é, por possuir a etapa de amplificação por PCR, e devido aos erros inerentes ao processo, alguns pesquisadores afirmam que o sequenciamento do tipo *target* não entra no conceito de metagenômica. No entanto, ela é específica para a identificação taxonômica (ELOE-FADROSH et al., 2016; OULAS et al., 2015; PRAKASH; TAYLOR, 2012; STREIT; SCHMITZ, 2004).

A metagenômica funcional, por sua vez, não está relacionada às sequências metagenômicas, mas às funções deduzidas a partir das mesmas. Com a sequência do DNA da amostra é construída uma biblioteca funcional de clones, a partir da qual é realizada uma prospecção de atividades enzimáticas de interesse. Nesses casos a identificação taxonômica da comunidade não é o objetivo principal, mas sim a busca por funções metabólicas (ESTHER M. GABOR; ALKEMA; JANSSEN, 2004; MARTIN et al., 2018; STREIT; SCHMITZ, 2004).

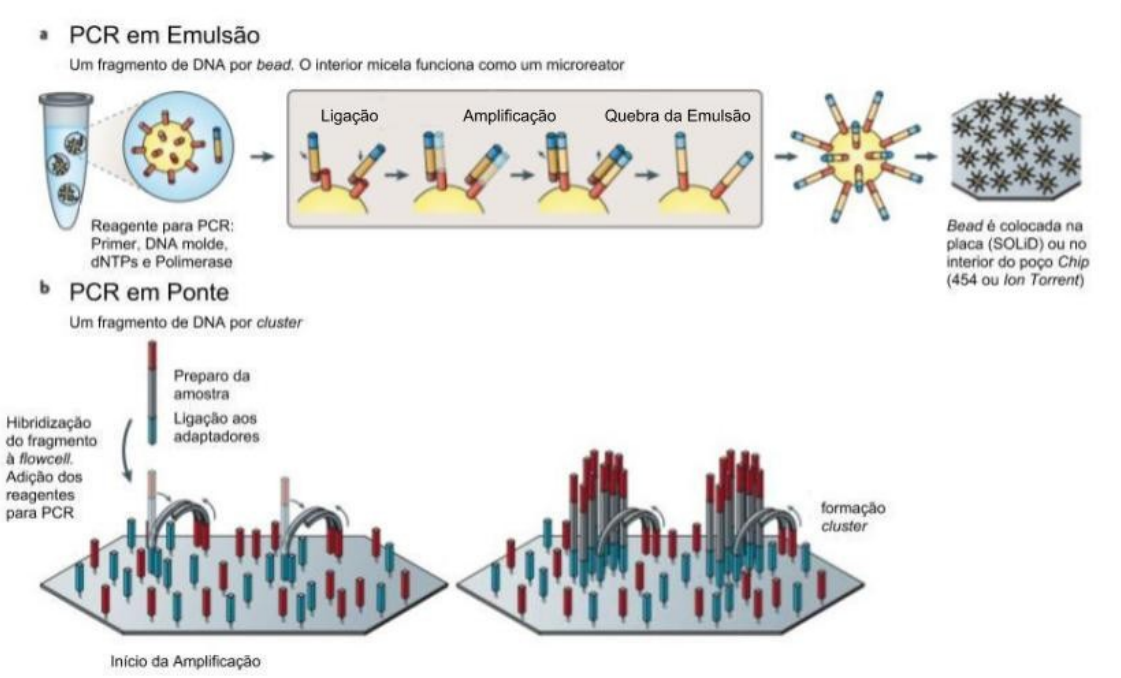

Figura 2.6. Amplificação da biblioteca de DNA.(A) PCR em Emulsão, (B) PCR em Ponte. Adaptado de Metzker (METZKER, 2010).

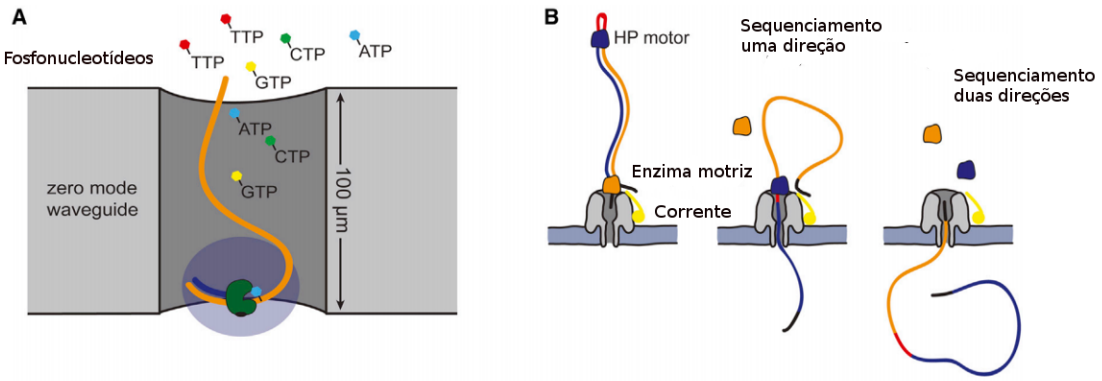

Figura 2.7. Sequenciamento *Single-Cell*. (A) Sequenciamento por incorporação de um fosfonucleotídeo marcado com um fluorófilo específico—PacBio. (B) Sequenciamento pela passagem do DNA por uma membrana – Nanopore (Oxford). (REUTER; SPACEK; SNYDER, 2015)

2.2.3. Sequenciamento em massa por amplificação

Em análises de diversidade taxonômica, o mais usual é o sequenciamento em massa de um gene alvo. O gene precisa possuir determinadas características para que a amostragem seja

representativa, devendo ser um gene conservado, para o alinhamento com *primers*, e ao mesmo tempo possuir variações que permitam identificações taxonômicas (KNIGHT et al., 2018). O gene mais utilizado para procariontes é o gene da subunidade ribossomal 16S, e, no caso de eucariontes o 18S. Trata-se de um gene muito informativo, amplamente empregado em estudos taxonômicos e evolutivos possuindo mais de um banco de dados específico (YARZA et al., 2014).

Conhecidos desde a década de 70 como “relógio moleculares”, os genes ribossomais são expressos constitutivamente e estão presentes em todas as formas de vida. Os genes rRNA 16S possuem uma alternância entre regiões conservadas (universais) e regiões hipervariáveis resultantes de diferentes taxas evolucionárias (KOZLOV et al., 2016; MARTINEZ-PORCHAS et al., 2017; YARZA et al., 2014). A conservação das regiões é a garantia na perpetuação da funcionalidade do gene (YARZA et al., 2014). Além da estrutura primária, a estrutura secundária do rRNA pode conter informações valiosas (SCHLOSS, 2010, 2013[a]). A Figura 2.8 ilustra a sequência primária do rRNA 16S de *Escherichia coli*, com as regiões variáveis.

O crescente avanço na tecnologia NGS tem promovido um aumento substancial de dados disponíveis em bancos públicos, este tipo de análise tem sido mais difundido para diferentes grupos de pesquisa. No entanto, uma das limitações da técnica é o comprimento do fragmento do DNA a ser sequenciado, de até 400 pares de bases tanto no *Ion Torrent* quanto no *Hi-Seq* do *Illumina*. O rRNA 16S possui 1.500 pares de bases (BAKER; SMITH; COWAN, 2003; SIMS et al., 2014) com uma característica evolutiva singular, o que é a sua evolução não linear, ou seja, ao longo de seu comprimento possui uma alternância entre regiões de alta conservação e outras altamente variáveis (BENÍTEZ-PÁEZ; PORTUNE; SANZ, 2016; MELO, 2008; SCHMIDT; MATIAS RODRIGUES; VON MERING, 2014; VĚTROVSKÝ; BALDRIAN, 2013). Ao todo, o gene possui nove dessas regiões, e a escolha de uma região inapropriada pode levar a conclusões biológicas questionáveis (KLINDWORTH et al., 2013; YANG; WANG; QIAN, 2016). A tecnologia do NGS, baseada na análise de fragmentos curtos permitiu uma maior sensibilidade e aprofundamento na composição das amostras, tornando possível a detecção e identificação de organismos em quantidades muito pequenas, que compõem uma biosfera rara. No entanto, a baixa resolução de pequenos segmentos analisados, principalmente para amostras ambientais e muito

complexas, impede o melhor conhecimento da amostra, devendo ser encarado como um resultado parcial (CAI et al., 2013; GUO et al., 2013). Ainda que a classificação taxonômica baseada no sequenciamento de 16S não seja conclusiva, a ponto de fornecer informações ao nível de espécie e subespécie, é considerada um avanço fundamental quando comparada à classificação baseada em fenótipos metabólicos (KIM; MORRISON; YU, 2011).

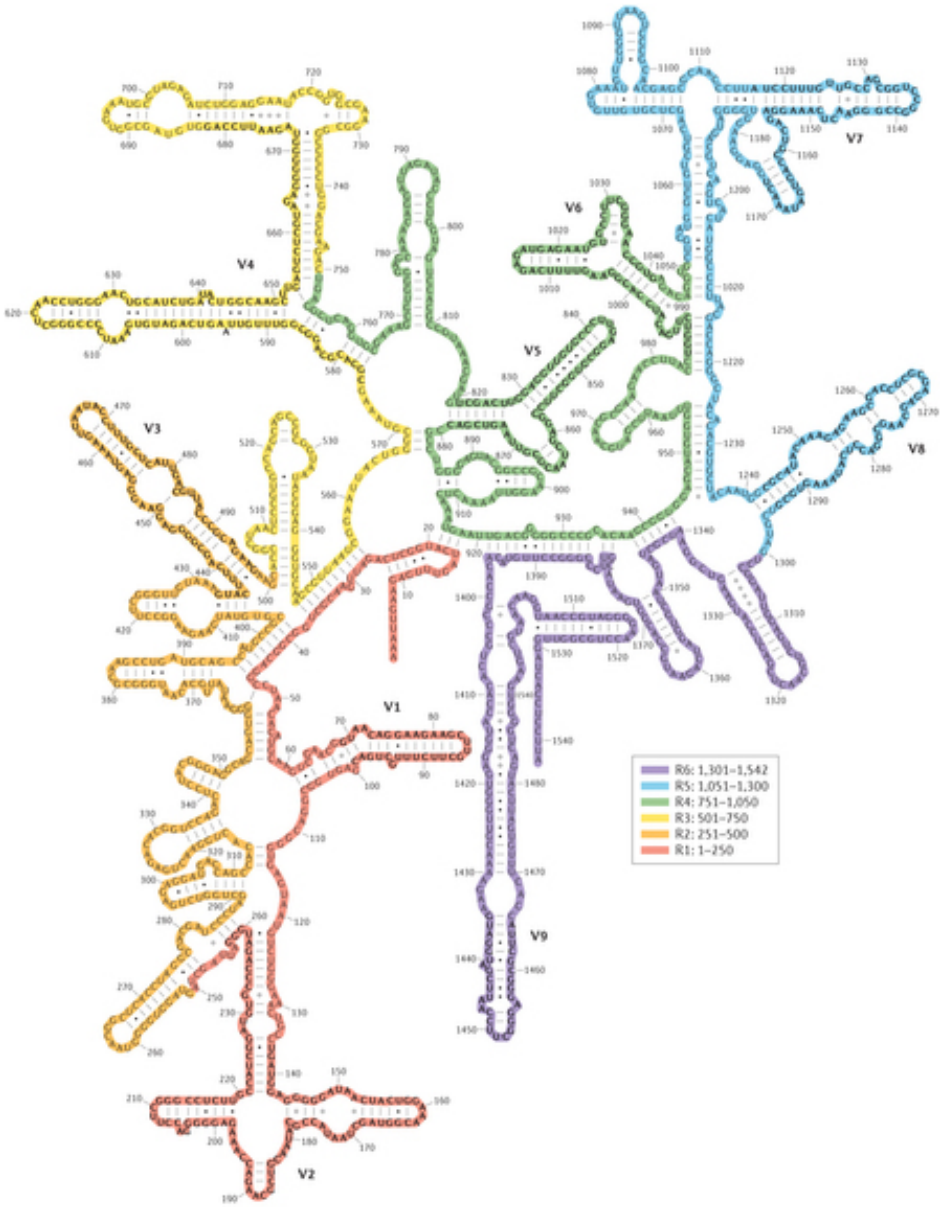

Figura 2.8. Estrutura Secundária do rRNA 16S. É possível observar na representação, as alças que se formam na estrutura secundária, e as representações das regiões hipervariáveis em seis cores diferentes; em vermelho, as regiões V1 e V2; em laranja, a região V3; em amarelo a região V4, em verde as regiões V5 e V6; em azul as regiões V7 e V8; e em roxo, a região V9. Imagem obtida de YARZA et al., 2014.

O sequenciamento de genes rRNA 16S é alvo constante de estudos tanto na pesquisa básica quanto aplicada. Ao passo em que há grupos de pesquisa dedicados à análise taxonômica, a partir de regiões específicas do gene rRNA 16S, para fins evolutivos (BENÍTEZ-PÁEZ; PORTUNE; SANZ, 2016; KLINDWORTH et al., 2013; SINGER et al., 2016; TIKHONOV; LEACH; WINGREEN, 2013), há outros grupos dedicados à análise de determinadas regiões variáveis, com o objetivo de determinar a composição de diferentes comunidades microbianas. Há um movimento de convergência para uma região em especial, a V4, sendo esta a escolhida para o projeto global de microbioma da Terra e do corpo humano (respectivamente *Earth Microbiome Project*, EMP, e *Human Microbiome Project*, HMP, (GILBERT; JANSSON; KNIGHT, 2014; LUKE K URSELL, JESSICA L METCALF, LAURA WEGENER PARFREY, 2013; LUNDBERG et al., 2012). No entanto, observa-se o emprego de diferentes regiões como alvos. Por exemplo, estudos na área de saúde oral frequentemente baseiam-se nas regiões V1-V3 (CHHIBBER-GOEL et al., 2017; SAMPSON et al., 2016; WOO et al., 2008; XIAO et al., 2016a), enquanto estudos baseados na região V4 são mais comuns na análise de microbiomas marinhos (GILBERT et al., 2012; LOGARES et al., 2014). Em eucariontes a região V9 do rRNA 18S é o alvo de escolha para a detecção e identificação de microrganismos (BARB et al., 2016). Na microbiologia do petróleo, encontram-se na literatura muitos estudos em que diferentes regiões foram selecionadas para a identificação taxonômica, sendo a V4 a mais frequentemente empregada (GAO et al., 2015; KORENBLUM et al., 2010; LI et al., 2016a). A escolha da região “alvo” afeta diretamente na diversidade e estimativas de abundância encontradas (BAKER; SMITH; COWAN, 2003; KIM; MORRISON; YU, 2011; TENG et al., 2018). No presente estudo, as diferentes regiões sequenciadas foram analisadas, quanto a diversidade por índices paramétricos e não paramétricos, e comparadas entre elas e com dados disponíveis na literatura específica.

Além da escolha das regiões alvos, dentro do gene rRNA 16S, fatores experimentais podem contribuir diretamente para os resultados obtidos, tais como subamostragem, eficiência diferencial de extração do DNA, presença de contaminante ou inibidor da PCR, condições de armazenamento da amostra, concentração do DNA, grau de conservação da molécula molde (rRNA 16S), conteúdo de GC, estrutura secundária do rRNA, ou mesmo erros intrínsecos à PCR, como a formação de quimeras. Há ainda a possibilidade de erros durante cada etapa do sequenciamento (ELOE-FADROSH et al., 2016; GOHL et al., 2016). Algumas pesquisas

sugerem que o sequenciamento da amostra de interesse ocorra paralelamente ao de uma comunidade artificial (e portanto conhecida) para que se possa estimar possíveis desvios durante o processo de sequenciamento (POLLOCK et al., 2018).

2.3. Programas para tratamento dos dados

À medida que aumenta a capacidade de geração de dados, através dos novos sequenciadores de DNA, tornam-se necessários programas de computador com maior capacidade de processamento para tender à demanda de toda a informação disponibilizada (NILAKANTA et al., 2014; OSMAN et al., 2018; POLLOCK et al., 2018).. Estima-se que a maior parte das descobertas de novas espécies e gêneros se darão por sequenciamento (YARZA et al., 2014). Portanto, o desenvolvimento de softwares robustos e confiáveis é tão necessário quanto a disponibilidade de novas tecnologias de sequenciamento do DNA.

Após o sequenciamento, um arquivo com os dados é gerado, com características próprias a cada tecnologia de NGS empregada, cada um com estratégias próprias da análise. A presente seção abordará as diferenças entre as plataformas *Illumina* e *Ion Torrent*, que foram empregadas no projeto que deu origem a essa dissertação Os sequenciadores automatizados, além de informara sequência e composição de bases, dão ainda informações importantes sobre a qualidade da leitura, isto é, sobre a qualidade da sequência (do inglês *read*), que corresponde a uma estimativa da confiabilidade da detecção, medida representada pelo "*Phred Quality Score*" (*Phred*), nome do programa que lhe deu origem, cuja fórmula é apresentada em Fórmula 2.1. Maiores valores do *Phred* indicam maiores qualidades de leitura, que variam entre 1 a 40, sendo que cada plataforma possui seu critério de pontuação (NILAKANTA et al., 2014).

$$Q = -10 \log P_e$$

Fórmula 2.1. *Phred*. P_e, é a probabilidade de erro na leitura da base.

O formato de sequência mais comumente empregado é o FASTA, muito utilizado no alinhamento de sequências de DNA e proteínas. É um formato de texto, cujo início é o cabeçalho da sequência iniciando pelo caractere ">" e na quebra de linha seguinte a sequência

de nucleotídeo ou aminoácidos, respectivamente. A sequência é iniciada, sem espaçamento, de acordo com a anotação padronizada pela *International Union of Pure and Applied Chemistry* (IUPAC). Linhas que iniciam pelo caractere “;” ou “#” são consideradas comentários. Acompanhando o arquivo FASTA, há um arquivo cuja extensão é “qual”, indicando a qualidade das leituras, cujo formato é similar: as sequências possuem o mesmo cabeçalho presente no formato FASTA mas, no lugar da representação dos nucleotídeos, há uma pontuação da qualidade das leituras individuais na escala *phred*. Apesar de simples, os arquivos FASTA e “.qual” em análises múltiplas geradas por NGS ocupam muita memória computacional. O arquivo de saída da plataforma *Illumina* é uma adaptação do FASTA, chamado de fastQ, que informa tanto a sequência quanto a qualidade da leitura, sendo composto da seguinte forma: cabeçalho, iniciado pelo caractere “@”,seguido da sequência de nucleotídeos, alinhada com o caractere “+” e por último a linha de pontuação codificada com os caracteres do tipo *American Standard Code for Information Interchange* (ASCII), que corresponde a um código binário, em que cada caractere terá um valor numérico diferente. A Figura 2.9 ilustra como são os arquivos, FASTA, qual e fastq. O *Ion Torrent* gera inicialmente um arquivo com uma extensão própria, que pode ser convertido nos formatos FASTA e “.qual” ((EDGAR, 2010; NILAKANTA et al., 2014).

Diversos programas são capazes de processar os dados de forma contínua, à medida que são gerados nas plataformas. De forma geral, o fluxo passa por filtragem que resulta na remoção de leituras com baixa qualidade, remoção de quimeras, retirada dos adaptadores e dos *primers*, seguindo-se a atribuição taxonômica e, em alguns casos, a análise de diversidade (POLLOCK et al., 2018). Muitos grupos de pesquisa em metagenômica se dedicaram a estudar os programas utilizados para esse fim. Dentre esses, a abordagem mais aceita é a de Nilakanta e colaboradores (NILAKANTA et al., 2014; POLLOCK et al., 2018). Em uma revisão mais recente, Pollock e colaboradores analisam os programas mais utilizados, QIIME, Mothur, UPARSE e MG-RAST, UPARSE, sugerindo que os dois primeiros como os melhores ((POLLOCK et al., 2018). O QIIME é programado na linguagem *Python*, e o Mothur em C++, mas na maioria dos casos não é necessária nenhuma programação adicional na utilização dessas ferramentas (SIEGWALD et al., 2017). Dentre os programas citados, o MG-RAST é o único que não utiliza linha de comando, no entanto é necessário um trabalho manual maior na análise dos dados (POLLOCK et al., 2018). Já o UPARSE é um programa licenciado, que

possui duas versões: uma aberta, mais limitada, e outra mais completa, que é comercializada. Outros pesquisadores criaram um programa similar, chamado VSEARSH, que funcionalidades semelhantes à versão comercializada do UPARSE, sendo, no entanto, aberto ao uso público (ROGNES et al., 2016).

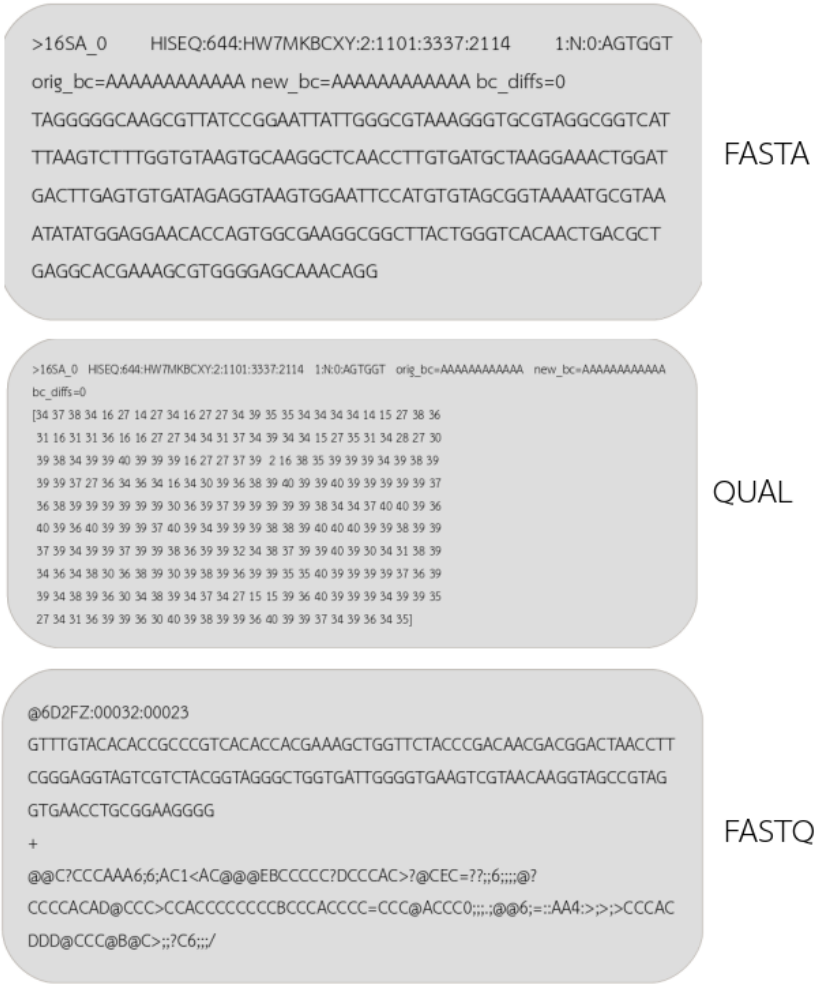

Figura 2.9. Representação de arquivos nos formatos FASTA e fastQ.

2.4. Tratamento dos dados

Independente do programa escolhido é necessário fazer uma boa filtragem na qualidade das leituras antes de fazer a atribuição taxonômica. Dela depende toda a

confiabilidade e reprodutibilidade da análise. Além disso, a escolha de parâmetros pode mudar significativamente os resultados obtidos. Há grupos de pesquisa que defendem a utilização de mais de um programa neste processo, aproveitando assim o melhor de cada ferramenta desenvolvida (PYLRO et al., 2014a).

Após tratamento inicial dos resultados, aqui denominado pré-tratamento, que entre outras funções remove as leituras com qualidade baixa, ocorre o agrupamento por meio de similaridade. *Reads* com homologia pré-determinada, presentes nos bancos de dados, são reunidos e os *clusters* de *reads* são denominados Unidades Taxonômicas Operacionais (OTU's). A princípio, cada OTU representa um elemento taxonômico que possui anotação em um banco de dados. Dessa forma, se o pré-tratamento não for realizado de forma adequada pode haver um número de OTU's superior à real diversidade do ambiente (EDGAR, 2018, 2013; HE et al., 2015; MYSARA et al., 2017; SCHLOSS; WESTCOTT, 2011). Diferentes bancos de dados podem ser utilizados na anotação taxonômica das OTU's, como o SILVA (QUAST et al., 2013; YILMAZ et al., 2014), o *Greengenes* (DESANTIS et al., 2006) ou o RDP (do inglês, *Ribossomal Database Project*, (WANG et al., 2007)), com metodologias de anotação característicos (KOO et al., 2017a; LIEBENSTEINER et al., 2014) A grande geração de dados através do sequenciamento requerem atualizações periódicas dos bancos, além das normativas de classificação taxonômica (YARZA et al., 2014).

Por fim, as sequências anotadas são submetidas a análises de diversidade. A primeira análise a ser feita refere-se à cobertura do sequenciamento, no qual é possível avaliar se a quantidade sequenciada é suficiente ou se são necessários novos ajustes. Esse parâmetro é designado curva de rarefação, que relaciona o número de OTU's encontradas com o número de *reads* obtidos (VARANI et al., 2015, cap. 12). Diferentes métricas podem ser utilizadas na construção dessa curva, dependendo da escolha do operador. Uma das mais antigas usadas em ecologia é a CHAO1, apresentada na Fórmula 2.2 (CHAO, 1984; CLAESSON et al., 2010; KUCZYNSKI et al., 2012). É possível, em análises multivariadas, comparar as diversidades entre as amostras, chamada de beta diversidade, por meio de métricas que dependem ou não de distâncias filogenéticas encontradas (LOZUPONE et al., 2007). Diferentes gráficos podem ser montados a partir das medidas de beta diversidade, inclusive com visualizações em 2D ou 3D. Grupos de amostras podem ser agrupados com base em diferentes fatores ou

características, a fim de avaliar suas influências na diversidade (VARANI et al., 2015, cap. 12)

$$G= 1 - (n/N)$$

Fórmula 2.2, em que n é o número de OTU's únicas, *singletons*, e N o número total de *reads (CLAESSON et al., 2010)*.

Uma nova ferramenta foi criada para enriquecer análise do sequenciamento em massa, visto que este é mais frequentemente concentrado na identificação taxonômica da comunidade microbiana, não abrangendo suas capacidades metabólicas ou funcionais. Por diferentes motivos nem sempre é possível realizar a análise por *shotgun*, que dá a melhor resposta quando se pretende investigar em mais profundidade a composição da amostra. Através de aproximações e correlações entre filogenia e funcionalidade, é possível fazer uma previsão das características metabólicas dos microrganismos presentes na amostra. Desenvolvida por Morgan Langille e colaboradores (LANGILLE et al., 2013), o PICRUSt é uma ferramenta analítica usa os dados taxonômicos obtidos, integrando-os a uma base de dados de vias metabólicas, como o KEGG. As previsões dependem da qualidade da anotação e da variação do número de cópias do gene alvo (LANGILLE et al., 2013). A robustez e qualidade dessa abordagem fazem com que o programa seja utilizado por diferentes grupos de pesquisa (DJEMIEL; GREC; HAWKINS, 2017; KOO et al., 2017b, 2017a; MUKHERJEE et al., 2017). Dessa forma, é possível obter uma estimativa da distribuição de vias metabólicas e funções existentes no microbioma, mesmo quando uma complexa população microbiana é caracterizada por *target*, como o gene rRNA 16S. Há outro programa com aplicação similar, Tax4Fun, que, no entanto, tem sido menos utilizado (ASSHAUER et al., 2015; KOO et al., 2017a).

3. MATERIAL E MÉTODOS

3.1. Obtenção da amostra

A amostra de água de produção foi coletada em abril de 2013, de um dos tanques de separação água/óleo de uma plataforma localizada no Campo de Marlim, com apoio logístico do CENPES/Petrobras. A amostra foi quarteada e acondicionada em galões e mantida refrigerada a 4 ºC até o momento da extração do DNA genômico. Descoberto em 1985, o Campo de Marlim ocupa uma área de 257,63 km^2 a 110 km do nordeste da Bacia de Campos, estado do Rio de Janeiro, Brasil. Sua lâmina d'água varia de 650 a 1.050 m e o óleo produzido é altamente biodegradado, de grau API entre 17 e 21° (ANP-Resolução nº 1001/2015; SCHOBBENHAUS; NEVES, 1990). O mecanismo primário de produção é o de gás em solução, com a utilização de injeção. O Campo possui uma das maiores produtividades do país, cerca de 2.021.580 bbl/dia, cerca de 12% da produção da Bacia em que está localizado, que representa 8% da produção nacional, dados estes calculados de 02/2009 a 02/2018 ("Boletim Mensal da Produção de petroleo e gas natural", 2018)

3.2. Extração do DNA genômico

Dois quintos da amostra, aproximadamente 400 mL, foram centrifugados a 16.000 *g*, por 10 minutos a 4 ºC. Após a centrifugação, a extração do DNA genômico foi feita através do *kit Power Water* (MO BIO-Qiagen, Hilden, Alemanha) de acordo com as instruções do fabricante. A amostra de DNA foi encaminhada do CENPES/Petrobras para o Laboratório de Microbiologia e Molecular e Proteômica (LaMMP), localizado no Instituto de Química (IQ) na Universidade Federal do Rio de Janeiro(UFRJ), onde deu-se continuidade à análise. A quantificação da amostra foi realizada por espectrometria a 260 nm no aparelho *NanoDrop* (Thermo, Massachusetts, EUA) e avaliado quanto à sua integridade por eletroforese em gel de agarose. 5 µL foram depositados em um gel com 1 % m/v de agarose em tampão Tris-Acetato-EDTA (TAE), que contém: 40 mM Tris (Sigma-Aldrich, Missouri, EUA), 20 mM ácido acético (Sigma-Aldrich, Missouri, EUA) e 1 mM EDTA (Sigma-Aldrich, Missouri, EUA), em pH 8,0. A eletroforese ocorreu por 40 minutos a 100 V. Em seguida, o gel foi corado em uma câmara contendo solução de brometo de etídio (Sigma-Aldrich, Missouri, EUA) a 1% v/v, por quinze minutos. A revelação foi realizada pelo fotodocumentador Min

Ibis Pro (DNR Bio Imaging Systems, Neve Yamin, Israel) com câmera de 1.3 Mpixel, que registra a imagem pelo software GelCapture Software, em luz ultravioleta.

3.3. Amplificação e sequenciamento de regiões variáveis do gene rRNA 16S e 18S

Foram utilizados neste trabalho duas plataformas de sequenciamento de Nova Geração (NGS). Uma delas, o *Ion Torrent* (Thermo, Massachusetts, EUA) está localizada noLADETEC, Instituto de Química, Universidade Federal do Rio de Janeiro. Para este, todas as etapas desde a ampliação gene por PCR até a análise final dos dados foi realizada no próprio Instituto de Química. Já para a segunda plataforma utilizada neste estudo, *Illumina HiSeq2500* (Ilumina, Califórnia, EUA), a amostra de DNA foi enviada à Novogene (UC Davis, Califórnia, Estados Unidos) que realizou a amplificação das diversas regiões do gene rRNA 16S ou 18S, o sequenciamento e o tratamento preliminar dos dados. A Tabela 3.1 mostra os *primers* utilizados, regiões amplificadas, seus tamanhos, ciclagem de PCR e as referências consultadas, as quais tiveram como prerrogativa o trabalho com microbiomas ambientais e estudos de correlação entre diferentes regiões do gene ribossomal.

3.3.1. Sequenciamento com o Ion Torrent

Para o sequenciamento na plataforma do Ion Torrent, a reação da PCR ocorreu nas seguintes condições: em um volume final de 20 µL, foram adicionados 4 µL DNTP 0,2 mM, 3 % DMSO, 100 ng DNA, 0,02 U da enzima *Phusion High-Fidelity* (Thermo, Massachusetts, EUA), equivalente a 0,2 µL, além de $MgSO_4$ a uma concentração final de 2,0 mM, para as regiões 16SV4, 16SV6 e 3,0 mM $MgSO_4$ para 16SV1-V2, 16SV9, 16SV6a e 18SV9. As reações foram realizadas em triplicatas e seus produtos foram purificados através do kit QIAquick PCR Purification (Qiagen, Hilden, Alemanha).

O preparo da biblioteca foi realizado com *Ion Xpress Plus Fragment Library kit e Xpress Barcode Adapters kit.* Para a sua quantificação foi usado o *Ion Library Quantification kit* (Thermo, Massachusetts, EUA). Para o preparo e enriquecimentos das *beads* foram empregados *Ion Sphere Quality Control, Ion PGM Enrichment Beads, Ion PGM Hi-Q View OT2* (Thermo, Massachusetts, EUA). O sequenciamento ocorreu na máquina *Ion Torrent PGM Hi-Q com o Ion PGM Hi-Q View Sequencing kit* (Thermo, Massachusetts, EUA). Todas as etapas foram seguidas de acordo com o protocolo do fabricante.

Tabela 3.1: Primers utilizados para amplificação dos genes 16S rRNA e 18S rRNA

Objetivo	*Primer*	Sequência (5'-3')	Gene	Pares Base(pb)	Ciclagem PCR	Referência
Ion Torrent	27F	GTAGMGTTYGATYMTGGCTCAG	16SV1-V2	341 pb	98°C/ 5' 25x (95°C/ 30", 64°C/ 30", 72°C/ 1') 72°C/ 10', 4°C	(FOUHY et al., 2016)
	R	CCTGCTGCCTCCCGTAGGAGT				
	515F	GTGTGYCAGCMGCCGCGGTAA	16SV4	292 pb	98°C/ 5' 25x (95°C/ 30", 50°C/ 30", 72°C/ 1') 72°C/ 10', 4°C	(GILBERT et al., 2010)
	806R	CCGGACTACNVGGGTWTCTAAT				
	926F	GTAAACTYAAAKGAATTGACGG	16SV6	207 pb	98°C/ 5' 25x (95°C/ 30", 50°C/ 30", 72°C/ 1') 72°C/ 10', 4°C	(TREMBLAY et al., 2015)
	1114R	CCGGGTTGCGCTCGTTGC				
	958arcF	GTAATTGGANTCAACGCCGG	16SV6a		98°C/ 5' 25x (95°C/ 30", 59°C/ 30", 72°C/ 1') 72°C/ 10', 4°C	(HUSE et al., 2007)
	1048arcR	CCCGRCRGCCATGYACCWC				
	1380ekF	GTCCCTGCCHTTTGTACACAC	18SV9	131 pb	98°C/ 5' 25x (95°C/ 30", 57°C/ 30", 72°C/ 1') 72°C/ 10', 4°C	(AMARAL-ZETTLER et al., 2009)
	1389F	GTTTGTACACACCGCCC	16SV9	141 pb	98°C/ 5' 25x (95°C/ 30", 57°C/ 30", 72°C/ 1') 72°C/ 10', 4°C	
	1510R	CCCCTTCYGCAGGTTCACCTAC				
Illumina	341F	CCTAYGGGRBGCASCAG	16SV3-V4	466 pb		(KOZICH et al., 2013)
	806R	GGACTACNNGGGTATCTAAT				
	515F	GTGCCAGCMGCCGCGGTAA	16SV4	292 pb		
	806R	GGACTACNNGGGTATCTAAT				

Objetivo	*Primer*	Sequência (5'-3')	Gene	Pares Base(pb)	Ciclagem PCR	Referência
	515F	GTGCCAGCMGCCGCGGTAA	16SV4-V5	393 pb		
	907R	CCGTCAATTCCTTTGAGTTT				
	U519F	CAGYMGCCRCGGKAAHACC	16SV4A	288 pb		
	806R	GGACTACNNGGGTATCTAAT				
	528F	GCGGTAATTCCAGCTCCAA	18SV4	179 pb		
	706R	AATCCRAGAATTTCACCTCT				
	1380F	CCCTGCCHTTTGTACACAC	18SV9	131 pb		
	1510R	CCTTCYGCAGGTTCACCTAC				
	ITS5-1737F	GGAAGTAAAAGTCGTAACAAGG	ITS1	307 pb		
	ITS2-2043R	GCTGCGTTCTTCATCGATGC				
	ITS3	GCATCGATGAAGAACGCAGC	ITS2	386 pb		
	ITS4	TCCTCCGCTTATTGATATGC				

3.3.2. Sequenciamento pelo Illumina

Para o sequenciamento na plataforma *Illumina*, a reação de PCR ocorreu com a enzima *Phusion High-Fidelity PCR Master Mix* (Thermo, Massachusetts, EUA). A purificação dos fragmentos foi relizada com o *Qiagen Gel Extraction Kit* (Qiagen, Hilden, Alemanha). O preparo da biblioteca foi realizado através do kit *TruSeq DNA PCR-Free Sample Preparation* (Illumina, Califórnia, USA), de acordo com os protocolos do fabricante. Através do Fluorímetro *Qubit* 2.0 (Thermo, Massachusetts, EUA) e *Agilent Bioanalyzer* 2100

(Life Technologies, Califórnia, EUA), a qualidade da biblioteca foi avaliada. Por último, o sequenciamento ocorreu na máquina *Illumina HiSeq 2500 pair-end* (Illumina, Califórnia, EUA) no modo *pair-end* lendo 250 bases.

3.4. Tratamentos dos dados

Dois computadores com capacidades de processamento diferentes foram utilizados durante as análises. O primeiro, um portátil da Dell, de memória 7,7 GiB, processador Pentium ® Dual Core CPCT4200@2.0Hz x2, gráfico Mobile Intel GM45Express Chip set, 64 bit, com sistema operacional Ubuntu 16.04 LTS, com os seguintes programas: Phyton 2.7.14, QIIME1 (v. 1.9.1.) (BOKULICH et al., 2012, 2013; SCHLOSS, 2010), conda (v.4.3.29), usearch (v. 10.240) (EDGAR, 2004, 2013), vsearch (v.2.7.1.) (ROGNES et al., 2016), PICRUSt (v. 1.1.3.) (LANGILLE et al., 2013). E o segundo, nos casos em que era necessário um maior poder de processamento, foi utilizado o de mesa da mesma empresa, de memória 15,6 GiB, processador Intel Core ™ i7-7700CPU@3.60GHz x8, 64-bit, com sistema operacional Ubuntu 17.04; com os mesmos programas instalados.

Seguindo uma tendência global de busca por uma metodologia que se enquadre no tipo de dado obtido a partir das diferentes plataformas (ALLALI et al., 2017; KOZICH et al., 2013; PYLRO et al., 2014b; SOH et al., 2013), foram testados diversos fluxos de trabalho, por mais de um programa ou pela combinação deles. As Figuras 3.2 e 3.3 esquematizam os variados caminhos, com indicação do programa e linha de comando, que foram executados durante o pré-tratamento dos dados gerados pelas plataformas *Ion Torrent* e *Illumina*, respectivamente.

Após a definição do fluxo de trabalho, buscou-se a avaliação de diferentes parâmetros durante a análise, tais como métrica para montagem de unidades taxonômicas operacionais (CHEN et al., 2013; RIDEOUT et al., 2014; WESTCOTT; SCHLOSS, 2015), banco de dados para anotação taxonômica (MARKOWITZ et al., 2012), plataforma de sequenciamento (D'AMORE et al., 2016; GLENN, 2011; KOZICH et al., 2013; PYLRO et al., 2014b; SALIPANTE et al., 2014), e, por último, diferenças entre as regiões sequenciadas (FOUHY et al., 2016; KOZICH et al., 2013; WANG; QIAN, 2009).

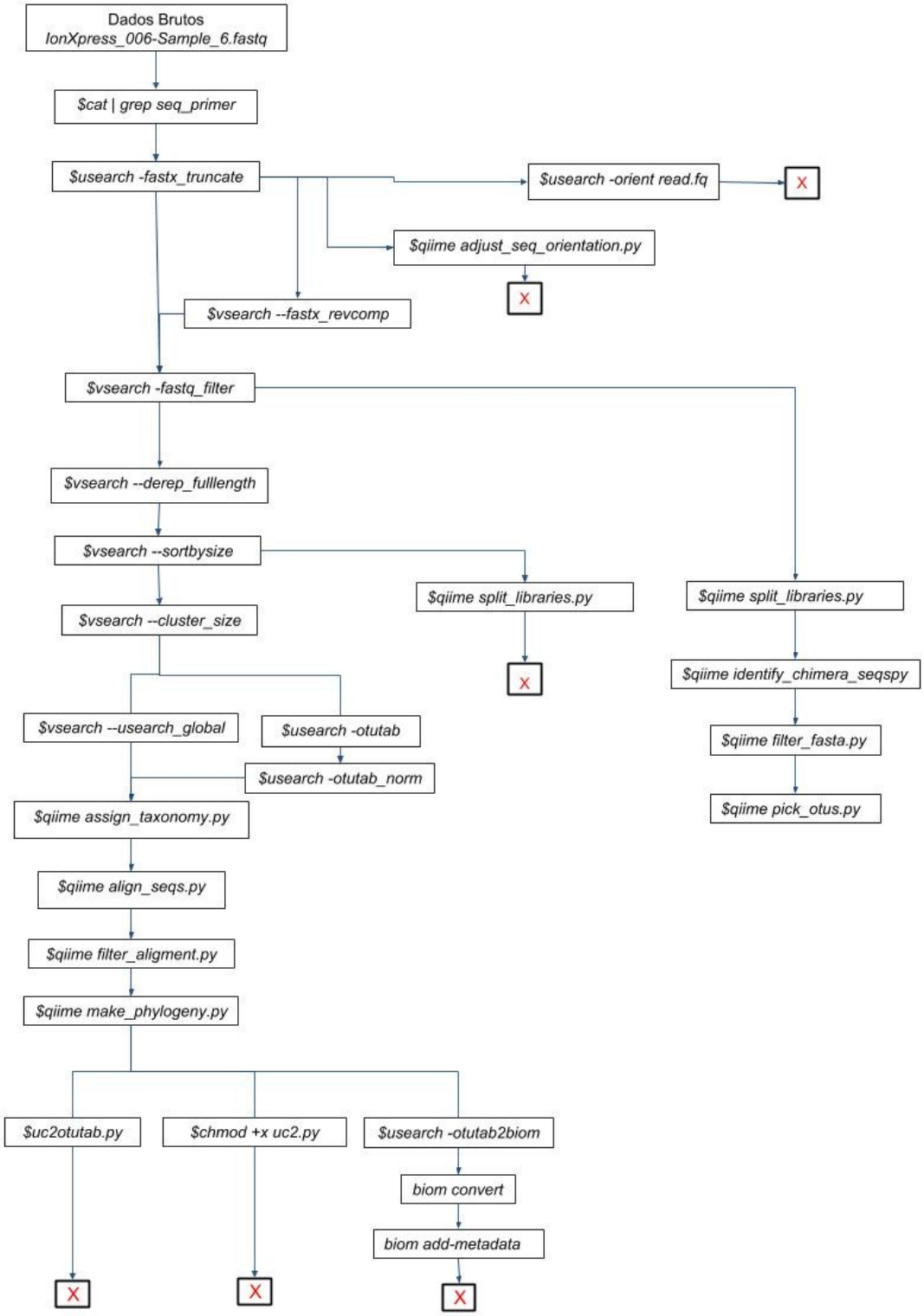

Figura 3.1. Fluxograma de trabalho de diferentes abordagens para pré-tratamento dos dados sequenciados na plataforma Ion Torrent. O X indica que o caminho adotado não gerou dados analisáveis.

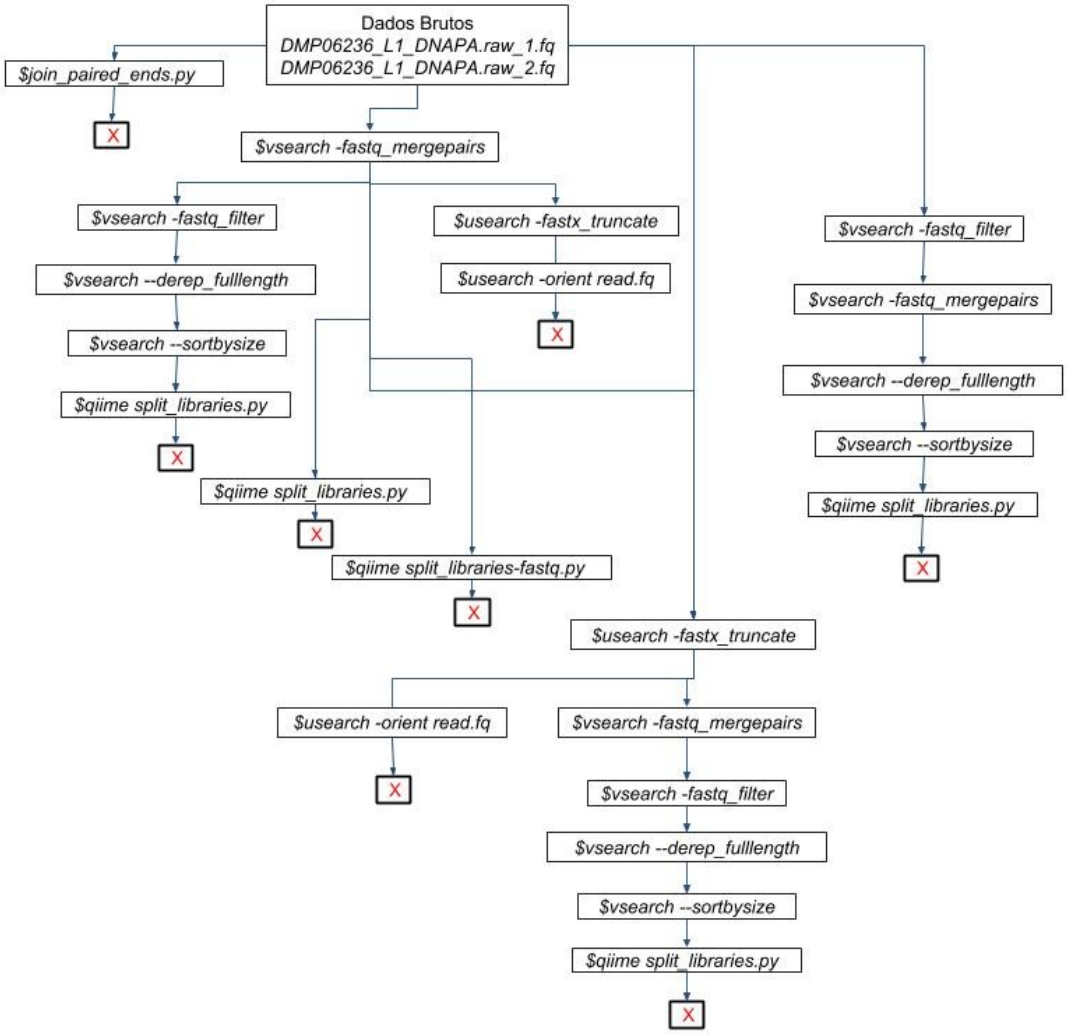

Figura 3.2. Fluxograma de trabalho de diferentes abordagens para pré-tratamento dos dados sequenciados na plataforma *Illumina*. O X indica que o caminho adotado não gerou dados analisáveis.

3.4.1. Ion Torrent

Para análise dos dados gerados pela plataforma *Ion Torrent*, o primeiro passo necessário foi separar as diferentes regiões sequenciadas. Para isso, foi realizada programação em *Shell*, para separar as sequências de acordo com os *primers*. Uma das etapas de pré sequenciamento é a ligação dos adaptadores à região amplificada, o qual acontece aleatoriamente (INC., 2014). Consequentemente, existe a probabilidade de ser sequenciado em ambos os sentidos da fita. No entanto, não é possível por bioinformática juntar as sequências complementares, como é feito com dados gerados pelo *Illumina*. Portanto, neste caso, cada região amplificada gerou duas sequências, que tiveram de ser tratadas de forma

independente. Para tal, o código estabelecido para a diferenciação das leituras, seguiu a seguinte ordem, o nome do gene (16S), seguido pela região sequenciada (V1,V4,V6 e V9), e, por último o sentido do *primer (*F e R, do inglês *forward* e *reverse*, respectivamente*).*

O primeiro pipeline seguido, foi proposto por Pylro e colaboradores através do Projeto Microbioma Brasileiro *(BMP, do inglês, Brazilian Microbiom Project)* (PYLRO et al., 2014b), o qual utiliza diferentes programas visando um resultado final de maior qualidade e confiabilidade. No entanto, esse projeto visa atender a um tipo de análise cujo sequenciamento engloba múltiplas amostras, neste caso, uma adaptação do *pipeline.* Para isso consistiu na remoção de alguns passos do *pipeline* original. Foram inseridos novos passos referentes à remoção *primers*, adaptadores e filtragem de acordo com o tamanho esperado do amplicon e com a qualidade de leitura. Posteriormente, foi realizado um processo de *dereplication*, em que leituras duplicadas são removidas. Dessa forma ocorre uma melhora no processamento dos dados o arquivo resultante é ordenado por abundânica. E, finalmente, é realizada a construção das OTU's. Essa primeira parte do processo é feita no programa *vsearch*, uma versão de algoritmo aberto, do *usearch*, que possui direitos autorais. A segunda parte da análise, é realizada no QIIME, também de código aberto. Nele, as leituras são alinhadas, agrupadas e preparadas para a atribuição taxonômica, a construção da árvore filogenética e por seguinte a análise de diversidade. No PiCRUST foi realizada a predição funcional.

3.4.2. Illumina

Devido ao tipo de saída da plataforma *Illumina* ser diferente da *Ion Torrent*, o fluxo de trabalho teve que novamente passar por uma otimização, desta feita considerando os mesmos critérios utilizados no *Ion Torrent*. A Figura 3.2 apresenta os diferentes caminhos tomados.

A junção da fita principal com a complementar foi realizada no software FLASH (V1.2.7, http://ccb.jhu.edu/software/FLASH/) (MAGOČ; SALZBERG, 2011), com a utilização a jusante do QIIME nos demais procedimentos. Para isso, foi utilizado o comando <split_libraries_fastq.py>, que é capaz de, com os parâmetros ajustados, remover sequências duplicadas, filtrar o tamanho esperado da leitura e realizar um filtro de qualidade por base. O arquivo de saída é então utilizado como entrada para montagem de unidades taxonômicas operacionais, assim como os dados gerados no PMG. De forma semelhante, o

código utilizado para diferenciar as reads no Illumina seguiu a seguinte ordem: gene (16S ou 18S), e região (V3V4, V4, V5V6, V9).

3.4.3. Parâmetros Avaliados

Para tornar viável as comparações em diferentes parâmetros, com tantas variáveis, foram realizados testes empregando as sequências obtidas na região V4 do gene rRNA 16S, sequenciados no *Illumina*. Desta forma, os métodos foram padronizados dentro dos programas. O pré-tratamento dos dados é descrito nas seções 3.4.1 e 3.4.2, para o *Ion Torrent* e *Illumina*, respectivamente.

Os dados gerados separadamente foram unidos através do comando <m*erge_otu_tables.py*>. Quando possível, houve a utilização de diversos comandos dentro do próprio QIIME para avaliar os parâmetros assinalados; como o *script* <*group_significance*> que calcula a abundância relativa das OTU's nas amostras, através de testes estatísticos. Foram realizadas também *scripts* de análise de diversidade como <*alpha_diversity*> e <*beta_diversity*>. Os índices utilizados foram CHAO-1 para diversidade Alfa, e a medida Unifrac ponderada e não ponderada para a diversidade Beta (parâmetros *default* no QIIME). (LOZUPONE; KNIGHT, 2005; VÁZQUEZ-BAEZA et al., 2011). Para gerar um gráfico de coordenadas principais, conhecido como PCoA, foi necessário usar um fluxo de trabalho com alguns comandos envolvidos, em que o arquivo de saída serve como entrada para o próximo, tais como: <*beta_diversity.py*>; <*principal_coordenates.py*> e <*make_2d_plots.py*>. As métricas usadas nesses casos são dependentes de fiologenia. Para verificar a correlação biológica entre os dados gerados pelos diferentes métodos, um outro *script* de comparação foi utilizado, <*compare_taxa_summaries.py*>, que calcula a correlação taxonômica entre dois arquivos. Isto é, o intervalo em que é possível percentualmente encontrar níveis taxonômicos iguais. Para isso, foram empregadas 999 permutações usando a transformação z de Fischer com confiança de 95%. A comparação é feita dois a dois, sem acréscimo, no qual a primeira anotação é fixada como referencial e a segunda a esperada.

3.4.3.1. Montagem das Unidades Taxonômicas Operacionais (OTU)

A partir das sequências em formato FASTA gerado pelo comando do QIIME <*split_libraries_fastq.py*>, procedeu-se à montagem das OTU's. Há três diferentes métodos

disponíveis, os quais foram utilizados, <*pick_close_reference_otus.py*>, <*pick_de_novo_otus.py*> e <*pick_open_reference_otus.py*>. No QIIME, o algoritmo utilizado para o agrupamento das OTU's foi o *uclust* (RIDEOUT et al., 2014), empregando a identidade mínima de 99% (CHEN et al., 2013). O alinhamento de sequências utilizado para a referência aberta e fechada foi feito pelo programa PyNAST (CAPORASO et al., 2010). O banco de dados adotado foi o *Greengenes Database* (v. 2013) (DESANTIS et al., 2006).

3.4.3.2. Banco de dados para Atribuição Taxonômica

Para a atribuição taxonômica da OTUs preditas foram utilizados os bancos *Greengenes* (v. 2013) (DESANTIS et al., 2006) e SILVA-132 (https://www.arb-silva.de/documentation/release-132/)(QUAST et al., 2013; YILMAZ et al., 2014).

A fim de avaliar as diferenças entre os dois maiores bancos de dados existentes, Greengenes e SILVA, ambos foram utilizados como referência no comando <*pick_open_reference_otus.py*> no QIIME, após junção dos *pair-ends* das leituras no FLASH. Cada anotação representativa foi comparada com os bancos de dados *Greengenes Database* e SILVA.

3.4.3.3. Comparação dos dados obtidos nas diferentes plataformas

Foi comparado o resultado de saída dos dados do *Illumina* e *Ion Torrent*, utilizando a montagem de OTU de referência aberta, usando como referência o banco de dados *Greengenes.* Como entrada foram utilizados os arquivos de saída do comando <*split_libraries.py*>. Para a comparação dos dados obtidos pelas plataformas *Illumina* e *Ion Torrent*, as OTU's geradas foram reunidas pelo comando <*merge_otu_tables.py*>.

3.4.3.4. Região do rRNA 16S sequenciada

Para comparar as diferentes regiões do gene rRNA 16S que foram sequenciadas, cada *amplicon* foi tradado individualmente, como descrito anteriormente. Na sequência, foram agrupados pelo comando <*merge_otu_tables.py*> e realizadas as análises de alfa e beta diversidade.

3.4.4. Análise de Diversidade

Para avaliar a riqueza e diversidade da comunidade, foi avaliada a rarefação e o cálculo de abundância relativa, através dos comandos <*alpha_rarefaction*> e <*plot_rank_abundance_graph*>, respectivamente. A rarefação compara a riqueza biológica observada entre as amostras de um mesmo habitat (EDGAR, 2010). O método utilizado nesse estudo foi o CHAO-1, que calcula a riqueza total da amostra com base na relação entre OTU's únicas, *singletons* e OTU's com duas sequências, *doubletons.* (CHAO, 1984).

3.4.5. Predição Funcional

Para predição funcional foi utilizado o *software* PICRUSt (V.1.1.3, http://picrust.github.io/picrust/) (KOO et al., 2017; MUKHERJEE et al., 2017; TRINGE; HUGENHOLTZ, 2008), o qual utiliza como arquivo de entrada a tabela OTU com anotação taxonômica gerada pelo banco de dados *Greengenes* e o arquivo de referência funcional, que neste estudo foio KEGG (KANEHISA et al., 2009). Para análise de anotação genômica foi utilizado o *script* <*metagenome_contributions.py*> que produz uma tabela que resume as contribuições de OTU para cada gene, previstas no KEEG. O *script* <*predict_metagenome.py*> foi utilizado para estimar o número de cópias dos genes, para assim conhecer o potencial do metagenoma analisado. Para a comparação entre as populações que fazem parte do microbioma, foi usado o arquivo de saída do <*metagenome_contributions.py*>. Foram gerados novos arquivos na extensão .biom e por eles foram realizadas análise de alfa e beta diversidade. A Tabela 3.2 mostra o resumo do arquivo de mapeamento utilizado.

Tabela 3.2. Arquivo de Mapeamento utilizado em análises de beta diversidade

Metabolismo	**Via**	**KEGG**	**Gene**
Acetato	ACDS	K00192	*cdhA*
Acetato	ACDS	K01895	*ACSS*
Acetato	ACDS	K14138	*acsB*
Acetato	ACDS	K00925	*ackA*

Metabolismo	**Via**	**KEGG**	**Gene**
Metano	Acetato_CO2-Metanol-Metano	K00399	*mcrA*
Metano	Acetato_CO2-Metanol-Metano	K00577	*mtrA*
Metano	Acetato-Metano	K00625	*[E.2.3.1.8.]*
Metano	Acetato-Metano	K13788	*pta*
Enxofre	Redução Assimilatoria/Dissimilarotia de Sulfato	K00955	*cysNC*
Nitrogênio	Redução Assimilatória Nitrato	K00367	*narB*
Nitrogênio	Redução Assimilatória Nitrato	K00372	*nasA*
Nitrogênio	Redução Assimilatória Nitrato	K00366	*nirA*
Nitrogênio	Redução Assimilatória Nitrato	K17877	*NIT-6*
Nitrogênio	Redução Assimilatória Nitrato	K10534	*NR*
Enxofre	Redução Assimilatória Sulfato	K00955	*cysC*
Metano	CO2-Metano	K00672	*ftr*
Metano	CO2-Metano	K00200	*fwdA*
Metano	CO2-Metano	K13942	*hmd*
Metano	CO2-Metano	K001499	*mch*
Metano	CO2-Metano	K00320	*mer*
Metano	CO2-Metano	K00319	*mtd*
Nitrogênio	Desnitrificação	K15877	*CYP55*
Nitrogênio	Desnitrificação	K00368	*nirK*
Nitrogênio	Desnitrificação	K15864	*nirS*
Nitrogênio	Desnitrificação	K04561	*norB*
Nitrogênio	Desnitrificação	K00376	*nosZ*
Nitrogênio	Redução Dissimilatoria/Desnitrificação	K02567	*napA*

Metabolismo	**Via**	**KEGG**	**Gene**
Nitrogênio	Redução Dissimilatoria/Desnitrificação	K00370	*narG*
Nitrogênio	Redução Dissimilatória Nitrato	K00362	*nirB*
Nitrogênio	Redução Dissimilatória Nitrato	K03385	*nrfA*
Enxofre	Redução Dissimilatória Sulfato	K11180	*dsrA*
Enxofre	Redução Dissimilatória Sulfato	K00394	*aprA*
Enxofre	Redução Dissimilatória Sulfato	K00958	*sat*
Enxofre	Redução Dissimilatória Sulfato	K00392	*sir*
House Keeping Genes	HKG	K09903	*pyrH*
Metano	Metano	K00440	*frhA*
Metano	Metano	K22516	*fdhA*
Metano	Metano	K22480	*hdrA1*
Metano	Metano	K14126	*mvhA*
Metano	Methanol-Metano	K14080	*mtaA*
Nitrogênio	Nitrificação	K10535	*hao*
Nitrogênio	Nitrificação	K10944	*PmoA-amoA*
Nitrogênio	Fixação de Nitrogênio	K00531	*anfG*
Nitrogênio	Fixação de Nitrogênio	K02586	*nifD*
Nitrogênio	Fixação de Nitrogênio	K22896	*nvfD*
Enxofre	SOX - Oxidação de Enxofre	K17222	*soxA*

4. RESULTADOS E DISCUSSÃO

4.1. Caracterização da Amostra

A amostra foi coletada de um separador água óleo da plataforma P37, condicionada sob refrigeração e encaminhada ao CENPES. Onde foi quarteada. O DNA extraído obteve concentração igual a 81,50 ηg/µL. Os valores das razões entre as absorbâncias 260/280 e 260/230 são 1,52 e 0,69 respectivamente. A primeira é referente a presença de proteínas, cujo máximo de absorção acontece no comprimento de onda de 280 nm, o valor da razão aceito é de aproximadamente 1,8. Outros tipos de contaminação são identificados pela razão 260/230, cujos valores aceitos como puros se encontram na faixa 2,0~2,2. Quando a razão é menor do que o esperado indica a presença de reagentes usados na extração. Tais fatores são importantes, pois são indicativos da qualidade do DNA. A Figura 4.1 mostra o gel de eletroforese realizado. É possível observar um arraste no poço equivalente à amostra, característico de degradação do DNA. A fragmentação do DNA provavelmente ocorreu durante o processo vigoroso de extração.

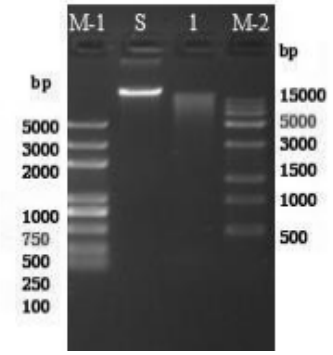

Figura 4.1. Controle de qualidade do DNA total. A figura mostra o gel de agarose (1% m/v) em tampão TAE, em que o poço identificado como S corresponde ao DNA_G de uma amostra padrão, enquanto o 1 à amostra com o DNA_G obtido da água produzida a ser analisado. Foram utilizados dois padrões de peso molecular, M-1 e M-2, cujos tamanhos em pares de bases (bp) estão assinaladas na imagem. A eletroforese ocorreu em 100 V, por 40 minutos.

Como discutido na Revisão Bibliográfica, a amostra de água de produção possui composição complexa, incluindo óleo, sais, sólidos em suspensão e gases dissolvidos

(MIDDLEDITCH, 1984; NEFF; LEE; DEBLOIS, 2011; SZKLO; ULLER; BONFÁ, 2012; THOMAS, 2001) (MIDDLEDITCH, 1984; NEFF; LEE; DEBLOIS, 2011; THOMAS, 2001). Tal mistura dificulta a extração do DNA, o que torna necessária uma ação mecânica mais vigorosa para obter um rendimento satisfatório no processo de extração, o que compromete em parte a sua integridade para uma análise em metagenômica por *shotgun* (CRUAUD et al., 2014; KNIGHT et al., 2018). No entanto, a moderada fragmentação observada, aliada à boa qualidade em termos de concentração e ausência significativa de contaminantes, não foi impeditivo para o sequenciamento dos produtos amplificados pela PCR (gene alvo rRNA 16S), sendo este o caminho utilizado.

4.2. Amplificação e sequenciamento de regiões variáveis do gene alvo rRNA 16S

A Figura 4.2 mostra o gel agarose produzido com os produtos de PCR da amplificação nas diferentes regiões do gene 16S rRNA. Através da Figura 4.2é possível observar que a reação foi bem-sucedida para as regiões V1-V2, V4, V6 e V9. Os *amplicons* que não mostraram desempenho satisfatório foram descartados, isto é, V6a (16SA) e V9ek (18SV9). Sendo assim, o sequenciamento específico para domínios *Archaea* e *Eukarya* foi efetuado exclusivamente no *Illumina.* Na própria figura 4.2, observando-se a intensidade dos produtos amplificados por PCR, é possível sugerir que há predominância do domínio *Bacteria* sobre os demais na amostragem questão. De acordo com o relatório fornecido pela empresa responsável pelo sequenciamento na plataforma *Illumina*, as regiões ITS1 e ITS2 foram as que tiveram os níveis mais baixos de amplificação pela PCR, sendo a ITS2 descartada pela falta de qualidade para prosseguir.

No sequenciamento pela plataforma *Ion Torrent* foi obtido um total de 759.897 leituras, com 157.633.490 bases sequenciadas, sendo 92% com *score* maior que Q20, em um comprimento médio de 207 bases. Já na *Illumina*, foram 630.589,73 leituras, com 157.790.475,00 bases sequenciadas, sendo 98,89% *score* maior que Q20, em um comprimento médio de 252 bases.

4.3. Tratamentos dos Dados

A tecnologia NGS revolucionou a pesquisa em Microbiologia, pois possibilitou um aumento exponencial dos dados gerados por análise (GLENN, 2011; LIU et al., 2012;

SANSCHAGRIN; YERGEAU, 2014) No decorrer dos anos vem se afirmando como uma opção viável para o conhecimento de microbiomas ainda desconhecidos (BARB et al., 2016; GEWIN, 2012). No entanto, a análise dos dados através da bioinformática ainda há muito o que caminhar. Não obstante, na literatura há diferentes trabalhos cujo objetivo é desenvolver abordagens viáveis, em que seja possível obter uma resposta confiável e reprodutível. Entretanto, muitos trabalhos sugerem estratégias distintas, cabendo ao pesquisador a decisão a respeito da linha analítica que seguirá (ALLALI et al., 2017; D'AMORE et al., 2016; HIERGEIST; REISCHL, 2016; KOZICH et al., 2013; MYSARA et al., 2017; NILAKANTA et al., 2014; POLLOCK et al., 2018; SALIPANTE et al., 2014; SIEGWALD et al., 2017; SOH et al., 2013)

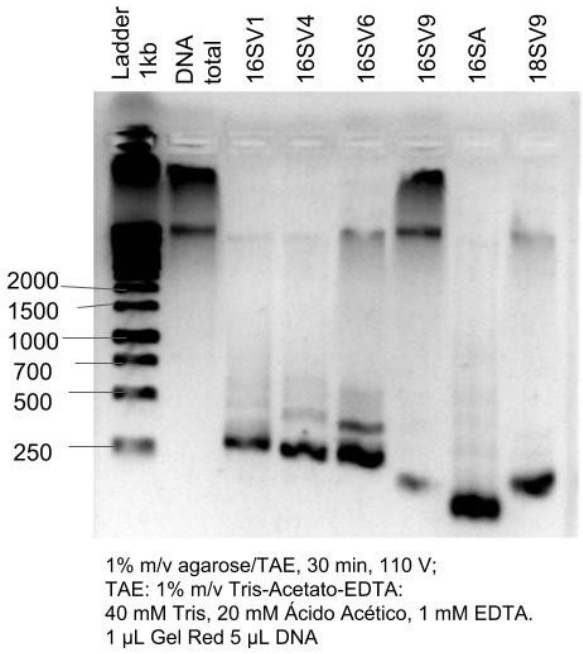

Figura 4.2. Produtos da PCR. Eletroforese em gel de agarose (1 %m/v), mostrando os produtos amplificados a partir de diferentes regiões do gene rRNA 16S.

4.3.1. Pré-Tratamento

O primeiro estágio da análise dos dados está relacionado ao pré-tratamento das sequências. Isto é, remoção de leituras com baixa qualidade, de dímeros, quimeras e filtragem

de acordo com o tamanho esperado do *amplicon*. Existem muitos programas disponíveis, (JU; ZHANG, 2015; KOZICH et al., 2013; PYLRO et al., 2016), sendo a escolha de qual programa utilizar dependente da linha analítica a ser tomada.

Em 2014, Pyrlo e colaboradores (PYLRO et al., 2014), publicaram um artigo do qual divulgavam a metodologia que viria ser a base da ferramenta computacional, por eles chamada de BMP (do inglês *Brazilian Microbiome Project*). Para isso, realizaram várias comparações com outras abordagens. Na primeira, utilizaram exclusivamente o programa QIIME sem nenhum tipo de adaptação. Na segunda, aumentaram a estringência da análise, ainda empregando o QIIME. Já na terceira, eles acoplaram ao fluxo de comandos do QIIME os *scripts* do UPARSE. Desenvolveram então o programa BMP, e criando uma metodologia que funcionasse independentemente da tecnologia de sequenciamento empregada (PYLRO et al., 2014). No entanto, como dito na seção 3.4 a máquina virtual do BMP, criada por Pylro e colaboradores ((PYLRO et al., 2014b) não foi capaz de processar todos os dados gerados por sequenciamento no presente projeto. Como alternativa, buscou-se seguir os mesmos princípios da execução do pré-tratamento.

A Figura 3.2 esquematiza os diferentes caminhos testados no pré-tratamento dos dados gerados no *Ion Torrent*. Observou-se que um maior número de etapas e filtros realizados no VSEARCH, programa de código aberto alternativo ao USEARCH, implicaria a remoção das leituras a ponto de inviabilizar etapas posteriores no QIIME. Foi então necessário diminuir a exigência, sem comprometer a análise, já que o percentual de corte nesta etapa foi alto, em torno de 95% das leituras brutas obtidas, como é possível notar na Tabela 4.1.

De forma semelhante os dados sequenciados na plataforma *Illumina* foram trabalhados. Como máquinas de sequenciamento possuem mecanismos diferentes, um outro tipo de ajuste foi realizado. A princípio buscou-se trabalhar com os dados brutos. Por fim, optou-se pela utilização do arquivo denominado "*merged_reads*", que contém as leituras que tiveram fita principal e complementar unidas no *software* FLASH. A região ITS1 apresentou um baixo número de leituras brutas e pouca qualidade nas mesmas, tendo sido portanto removidas da análise. Em média, 94% das leituras se mantiveram após a mescla com das fitas principal e complementar, como apresenta a Tabela 4.2.

Através desses dados é possível observar que a maior parte dos *reads* gerados pelo PMG é removido nos critérios iniciais, devido à metodologia distinta de sequenciamento.

Além disso, a necessidade de mesclar as fitas complementares, diminui processamento dos dados, acarretando um acréscimo na qualidade. Desta forma, os *reads* que não tiveram sua fita complementar sequencia da possuem menor confiabilidade, tendo sido portanto descartados.

Tabela 4.1 Quantidade de leituras obtidas e filtradas durante o pré-tratamento dos dados, para análise dos dados gerados no *Ion Torrent*.

Região	Leituras	Comprimento Médio	Percentual Filtrado Vsearch	Percentual Filtrado Qiime
16SV1V2F	61326	308	5,09%	94,91%
16SV1V2R	90094	289	7,03%	80,62%
16SV4F	70321	263	6,80%	92,13%
16SV4R	70774	274	6,20%	96,45%
16SV6F	75172	201	5,19%	89,88%
16SV6R	94337	187	5,30%	73,53%
16SV9F	121006	136	2,17%	0,30%
16SV9R	129149	138	2,14%	0,29%

Tabela 4.2: Quantidade de leituras obtidas e filtradas durante o pré-tratamento dos dados, para análise dos dados gerados no *Illumina*.

Região	Leituras	Percentual Combinado	Comprimento Médio	Percentual Filtrado Qiime
16SV3-V4	143305	89,07%	422	87,72%
16SV4	133045	99,01%	251	98,55%
16SV4-V5	136165	89,56%	372	88,35%
16SA	104089	99,24%	253	99,07%
18SV4	95134	93,07%	339	92,56%
18SV9	155657	96,48	141	96,78%

4.3.2. Parâmetros Avaliados

4.3.2.1. Montagem das OTU

A escolha do método de predição de OTU's não é considerado uma tarefa simples, ou trivial. Há inúmeras variáveis que interferem na qualidade do resultado (CHEN et al., 2013). Dentro do programa QIIME, existem três métodos de construção de uma OTU: referência fechada, *de novo* e referência aberta. Cada uma possui suas vantagens e desvantagens, com

uma ampla discussão na comunidade científica. Existem algumas formas de avaliar a qualidade dos métodos, um deles, é em relação ao desempenho computacional na criação dessas OTU's. Nesse sentido, a análise mais demorada, que exigiu mais poder computacional, foi a *de novo*, enquanto o método de referência fechada apresentou uma redução do tempo em 39,8% com relação ao *de novo*. O método de referência aberta chegou a diminuir o tempo em 81,4%.

As OTU's foram agrupadas de acordo com o algoritmo *uclust*, no qual ele primeiro seleciona uma sequência que funcionará como uma "semente" para a formação de um *cluster* inicial. Ocorre então uma busca nas sequências seguintes e, se houver similaridade dentro dos limites estabelecidos, há o agrupamento. Caso contrário, há formação de uma nova semente. Como vantagem, esse algoritmo é capaz de processar mais rápido os dados, possuindo a desvantagem de prejudicar a precisão quanto anotação taxonômica, ocasionando o aumento complexidade do microbioma (CHEN et al., 2013; SCHLOSS, 2010; SCHLOSS; WESTCOTT, 2011; WESTCOTT; SCHLOSS, 2015). Uma saída para contornar esse problema é uma modificação dos limites que definem uma OTU. Desta feita, em vez de alterar o algoritmo de agrupamento, comprometendo a capacidade de processamento, a dissimilaridade entre sequências dentro de um *cluster* passou de 3%, convencionalmente usada, para 1% (CHEN et al., 2013; KOEPPEL; WU, 2013). Em uma verificação recente, foi averiguado que esta mudança nos limites de dissimilaridade passou a integrar o manual da nova atualização do QIIME, v2.2018.8 (CAPORASO; DORRESTEIN, 2018).

Outros fatores influenciam na construção das OTU's como, por exemplo, a etapa de pré-tratamento. O sequenciamento por NGS, apesar de expandir o número de dados consideravelmente, gera numerosos erros, como também as etapas de extração do DNA e reação da PCR, sendo imperativo uma filtragem rigorosa, para que o resultado não seja inflado, dando uma falsa concepção da diversidade (BOKULICH et al., 2013; EDGAR, 2017; SCHLOSS, 2010; SINHA et al., 2015). Por este motivo, foi dedicada uma seção inteira na otimização dos dados.

A Figura 4.3A apresenta o gráfico de rarefação calculado a partir do índice CHAO1, que leva em consideração a relação entre leituras únicas, *singletons*, e duplicadas, *doubleton*, no agrupamento da OTU. A tipificação da curva é uma assíntota horizontal, na qual a medida que o número de *reads* é acrescentado, o número de OTU's formadas cresce linearmente, até

atingir um patamar máximo independente do aumento das leituras. Essa característica, além de dar uma noção de diversidade dentro da amostra, permite entender o quanto o universo foi contemplado na análise, ou seja, o nível de cobertura (EDGAR, 2010). A diferença apresentada entre curvas evidencia a capacidade de cobertura na amostra, através de cada métrica. Dos três métodos, o único que não apresenta tal esgotamento é o *de novo*, cuja curva encontra-se em uma crescente.

A Figura 4.3B mostra o *Diagrama de Whittaker* por número de espécies anotadas, em escala logarítmica. Esse gráfico pode ser interpretado como a riqueza da complexidade de uma amostra. Nesse sentido, a contagem total de espécies é maior no método de referência aberta, seguido pelo método *de novo*, e por último o de referência fechada. O método de referência fechada, além de possuir um número menor de espécies, também apresenta dificuldade em anotar as espécies minoritárias, possuindo um maior número de espécies com abundância maior que os outros métodos. É possível observar no gráfico (Figura 4.3B), que tanto o método de referência aberta quando o *de novo*, seguem uma trajetória similar. Entretanto, a curva da referência aberta demonstra que o método conseguiu anotar mais espécies minoritárias que os demais.

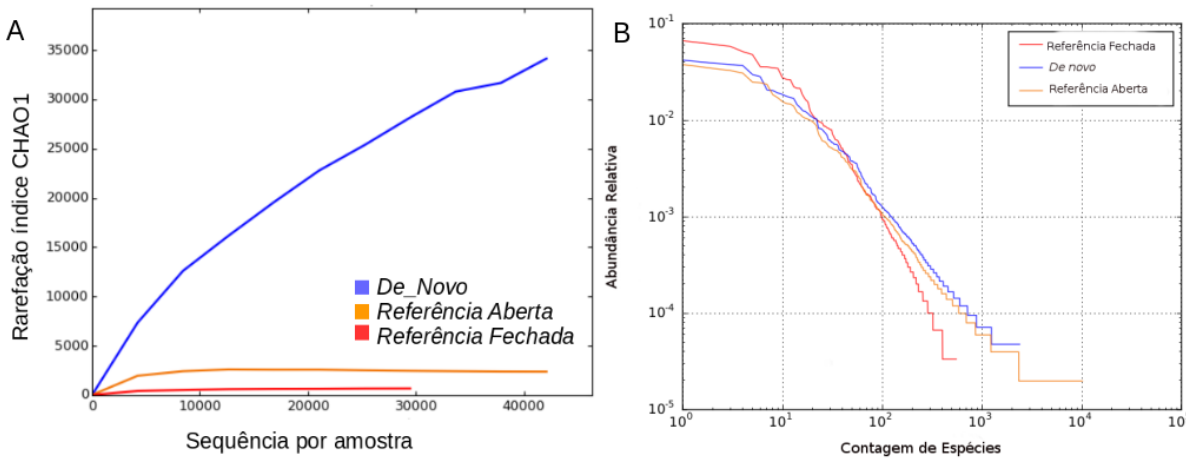

Figura 4.3.Alfa diversidade da variação do método de criação de OTU'S. (A) Curva de rarefação pelo índice CHAO1, a curva em azul representa o método de construção *de novo*, em laranja o de Referência Aberta e em vermelho o de Referência Fechada. (B) Pelo diagrama de Whittaker, a curva em azul representa o método de construção *de novo*, em laranja o de Referência Aberta e em vermelho o de Referência Fechada.

Vários estudos comparando métodos propõem que métodos robustos são capazes de gerar dados semelhantes entre si. Essas análises se baseavam em comparações de diversidade obtidos a partir de diferentes métricas utilizadas (EDGAR, 2004; HE et al., 2015; HUSE et al., 2007; RIDEOUT et al., 2014). Yan He, em 2015, conceitua a estabilidade de uma OTU. Isto é, um *cluster* que possui a mesma sequência independentemente da métrica utilizada (HE et al., 2015).

Neste espectro, foi testado o índice de diversidade beta a fim de encontrar as semelhanças entre os métodos. Em números totais, foram criadas 6.610 OTU's pelo método *de novo*, 1.541 em Referência Aberta e 620 em Referência Fechada. Dessas, há 289 OTU's em comum entre Referência Fechada e Aberta, mas nenhuma com *de novo*. Na diversidade beta *Unifrac* ponderada, métrica esta dependente da filogenia e da frequência das OTU's; foi registrado que há 14,61% de dissimilaridade entre os métodos de Referência Aberta e Fechada. E no índice *Bray-Curtis*, que é independente de filogenia, 0,32. Sendo as comparações supracitadas referentes apenas às sequências das OTU's, o método *de novo*, por ser único ao agrupar as OTU's, atribuindo-lhes taxonomia após essa etapa, acaba não sendo compatível em relação aos demais métodos. Yan He, em seu artigo chega a afirmar que o agrupamento das OTU's, sem que estas estejam atreladas a um banco de dados, as torna instáveis. No entanto, estabilidade não é sinônimo de qualidade. Os desenvolvedores do MOTHUR, programa similar ao QIIME, em diversos de seus artigos afirmam que a subjetividade na escolha do índice usado nessas análises é muito alta e dependendo da escolha feita pelo operador, esta pode direcionar a conclusão. A posição dos autores do MOTHUR é contrária a dos desenvolvedores do QIIME, ao defenderem que o método *de novo* é a melhor opção na construção de uma OTU; defendem ainda um algoritmo próprio para o agrupamento (PATRICK D. SCHLOSS, SARAH L. WESTCOTT, THOMAS RYABIN, JUSTINE R. HALL, MARTIN HARTMANN, EMILY B. HOLLISTER, RYAN A. LESNIEWSKI, BRIAN B. OAKLEY, DONOVAN H. PARKS, COURTNEY J. ROBINSON, JASON W. SAHL, BLAZ STRES, GERHARD G. THALLINGER, DAVID J. VAN HORN, 2009; SCHLOSS, 2010, 2013b; WESTCOTT; SCHLOSS, 2015). Já os desenvolvedores do QIIME recomendam fortemente o uso do método de Referência Aberta pois, na visão deles, além de criar OTU's mais estáveis, ainda é possível aproveitar as sequências que seriam descartadas pelo método de Referência Fechada. Além de possuir o melhor tempo de processamento (HE et al., 2015;

RIDEOUT et al., 2014). Os intervalos calculados dos coeficientes de correlação são: 0,7306-0,8047, 0,7546-0,8320 e 0,9274-0,9491, com valor médio 0,7702, 0,7966 e 0,9392, em Referência Fechada e *de novo*, Referência Fechada e Aberta, e Referência Aberta e *de novo*, respectivamente.

Apesar de divergências encontradas na literatura, o valor de correlação taxonômica entre as metodologias surpreende. Sem dúvida, o número de OTU's encontradas seguindo *de novo* é superior às outras. No entanto, ao atribuir uma classificação taxonômica, é utilizado o mesmo banco de dados em todas as estratégias, sendo que grande parte dessas OTU's anotadas como organismo desconhecido chegam a 3,3% do total das OTU's agrupadas, enquanto em Referência Aberta esse valor é igual a 2,9%. Para as próximas análises deu-se prosseguimento com a metodologia de Referência Aberta.

4.3.2.2. Emprego dos Banco de Dados

Foram geradas 1.225 OTU's quando no processo de anotação taxonômica foi utilizado o *Greengenes* como banco de dados. Já com o SILVA-132, foram 1.241, um aumento de 1,30% do número total de OTU's. A Figura 4.4 apresenta o gráfico de diversidade alfa de subamostragem, calculado de acordo com o índice CHAO1, com enraizamento a partir da junção dos arquivos <*rep_set.tre*>. É possível observar a proximidade entre as curvas, que não apresentam diferenças significativas.

De acordo com as Figuras 4.4 e 4.5 é possível ter uma noção da riqueza da amostra. Os arquivos gerados nos dois bancos de dados são similares quanto a complexidade, o número de espécies presentes e abundância relativa. Como mencionado brevemente na metodologia, cada banco de dados possui sua própria forma de anotação taxonômica. Que, por conseguinte, produz um obstáculo maior na comparação entre os arquivos, quando são utilizados os comandos tradicionais do QIIME. Por exemplo, a beta diversidade calculada de acordo com o índice *Unifrac* ponderado, a discrepância apresentada entre os arquivos é de 39,61%, com apenas 256 OTU's em comum entre os dois, com o mesmo enraizamento que a diversidade alfa. E, ao utilizar um índice independente de taxonomia, *Bray Curtis*, a divergência entre os arquivos chega a 0,98, e ao índice *Jaccard* igual a 0,8841.

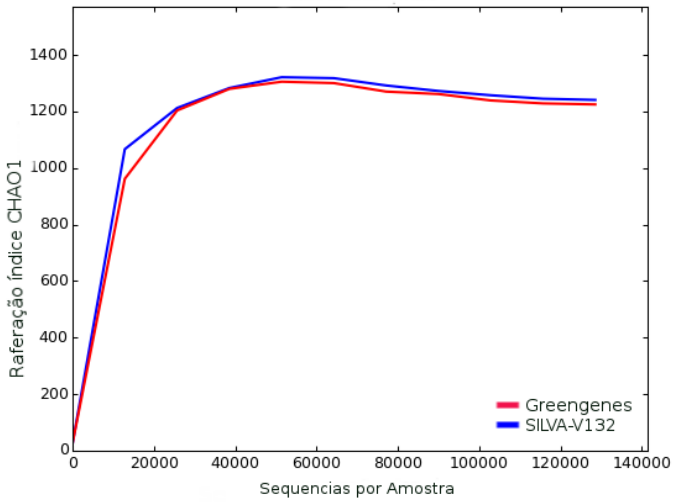

Figura 4.4 Curvas de rarefação baseadas na diversidade alfa da variação do banco de dados. Em vermelho *Greengenes* e em azul SILVA132.

Não obstante, a correlação taxonômica foi negativa e igual a -0,0121, com intervalo entre -0.0940 e 0.0698. Os valores encontrados não estão dentro do esperado, considerando a proximidade dos valores observada na diversidade alfa. Poderiam então os arquivos gerados possuírem uma quantidade similar de metadados e, ao mesmo tempo, um conteúdo tão divergente? A verificação manual das tabelas taxonômicas averiguou a existência de caracteres distintos que são usados como prefixos dos níveis taxonômicos. Por exemplo, no *Greengenes* o domínio *Bacteria* é descrito como *<bacteria>*, enquanto no SILVA-132 é *<_D_0__bacteria>*, o que torna os arquivos incompatíveis. A remoção de tais caracteres minimizou as incompatibilidades nos níveis taxonômicos mais elevados. Quando calculada em filo, a correlação taxonômica, passa a ser igual a 0,9992, com intervalo entre 0,9987 e 0,9995, sendo o maior valor encontrado entre todas as comparações realizadas nesta pesquisa. Mas, quando realizada em nível de gênero, o resultado foi de 0.3945, com intervalo entre 0.3204 e 0.4639.

Os bancos de referência taxonômica do rRNA 16S são montados a partir de uma combinação de bancos de genomas completos, os quais, por meio de metodologias distintas, resultam em tamanho e resolução característicos. Um banco de qualidade tem que possuir uma taxonomia normativa, que é oriunda de um alinhamento preciso e bem anotado. Os bancos mais utilizados pelos pesquisadores em seus trabalhos são o SILVA, em suas muitas

atualizações, e o *Greengenes* (SIEGWALD et al., 2017). Por mais criteriosos que sejam, os bancos possuem sequências taxonomicamente erradas, no inglês conhecidas como *mislabels*. São sequências que não foram devidamente anotadas no momento do depósito nos bancos públicos, propagando o erro para os bancos de referência e provocando anotações equivocadas (KOZLOV et al., 2016).

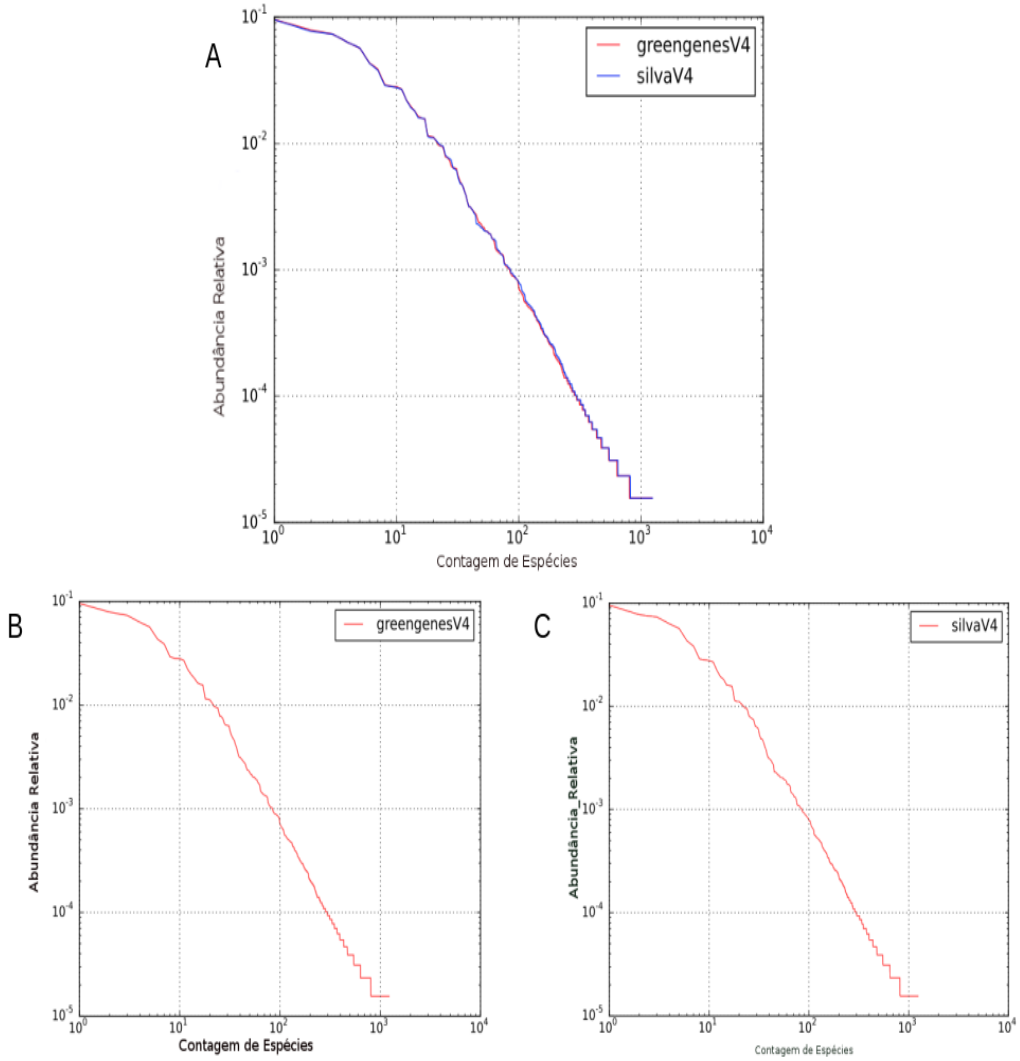

Figura 4.5 Gráficos de abundância relativa da variação do banco de dados. Na imagem superior, as curvas sobrepostas. No canto inferior esquerdo a curva em vermelho de abundância relativa do arquivo *Greengenes* isolado, e no canto inferior direito em azul do SILVA.

O *Greengenes* é anotado automaticamente. Para isso, bancos públicos de depósitos de genomas são utilizados como base para construção da árvore filogenética e anotação

taxonômica, sendo que o NCBI é a principal fonte. No processo de montagem, realizam uma filtragem de qualidade das sequências contidas no depósito, um alinhamento para remoção de dados redundantes, levando em consideração a estrutura secundária do rRNA, ainda contam com um algoritmo para identificar e remover as quimeras presentes. Em 2006, Desantis e colaboradores publicaram que com essas ferramentas removeram do NCBI cerca de 3% dos genes de microrganismos não cultiváveis e 0,2% dos cultiváveis (DESANTIS et al., 2006). O banco se dedica a cobrir procariotos, possuindo apenas os domínios *Archaea* e *Bacteria*, classificando-os nível de espécies. No entanto, sua última atualização foi realizada em 2013, ainda assim é o recomendado nas análises do QIIME, assim como do projeto global de mapeamento do Microbioma Humano (HMP, do inglês *Human Microbiome Project*) (BALVOČIUTE; HUSON, 2017; DESANTIS et al., 2006).

Por outro lado, há o banco de dados SILVA, o qual é curado manualmente, possui um registro próprio de genomas completos. Possui também um banco não-redundante para anotação e classificação taxonômica, incluindo as sequências de rRNA das duas subunidades ribossomais nos três domínios taxonômicos. As anotações seguem as normas de classificação vigentes no *Bergey's Manual of Systematic Bacteriology*, para *Archaea* e *Bacteria* enquanto paraeucariotos segue-se a normativa da *Sociedade Internacional de Protistologistas*. Além disso, o banco permite a descrição de taxonomia em mais de um táxon, denominado polifilético. Isto ocorre quando o gênero é descrito em um grupo, no entanto possui maior identidade com outro. E assim como no *Greengenes*, o alinhamento considera a estrutura secundária do produto gênico. Em 2016, Kozlov e colaboradores desenvolveram uma ferramenta analítica para determinar a quantidade presente de *mislabels* nos bancos, verificando que o percentual no SILVA-123 (v. 09/2015) era sete vezes maior que no *Greengenes* (v. 03/2013). Mesmo assim o percentual encontrado é inferior a 1,5 %, valor este, segundo os próprios pesquisadores, que não afeta a representatividade dos bancos, mas que indica que há espaço para melhoras. A ferramenta desenvolvida foi utilizada nas atualizações mais recentes do SILVA, realizada em dezembro de 2017 (v.132-12/2017), e assim diminuiu a presença das sequências enganosas (KOZLOV et al., 2016). O SILVA é o banco indicado tanto pelos desenvolvedores do MOTHUR quanto do USEARCH para usar no alinhamento das sequências de amostras desconhecidas (BALVOČIUTE; HUSON, 2017; QUAST et al., 2013; YILMAZ et al., 2014).

Uma característica comum aos bancos é a tendência a explorar mais os microrganismos com relevância clínica (POLLOCK et al., 2018; TRINGE et al., 2005). Em sua pesquisa, Siegwald verificou que tanto seguindo o *pipeline* do QIIME quanto do BMP, utilizando um conjunto de amostras conhecidas e padronizadas, ao mudar o banco do *Greengenes* para o SILVA, ocorreu um aumento na estimativa da riqueza no nível da família, que quase dobrou (SIEGWALD et al., 2017). O SILVA possui um método único de classificação, que inclui a introdução do táxon *Candidatus* para aqueles que ainda não foram publicados, o que acaba por possuir em suas referências nomes como “ambientais” ou “desconhecidas”, o que pode acarretar um aumento na aparente riqueza da amostra ((QUAST et al., 2013; YILMAZ et al., 2014). Em 2006, apenas no NCBI o número de sequências com a classificação “ambientais” ou “desconhecidas” de organismos não cultiváveis já superava a dos dados cultiváveis. Enquanto o SILVA categorizou tais organismos, criando novos nós na taxonomia, o *Greengenes* os evita (DESANTIS et al., 2006).

Originalmente, a taxonomia era uma atribuição fenotípica, gerada por observações laboratoriais. Com o aumento de dados gerados por sequenciamento genômico de microrganismos não cultiváveis, essa atribuição passou a ser pela identidade das sequências, resultando em dados conflitantes entre os bancos, sobretudo no caso de amostras ambientais, pois cada um possui sua metodologia de anotação. Como exemplo, Robert Edgar (2018), desenvolvedor do USEARCH, cita o caso comum dos gêneros *Escherichia* e *Shigella*, cuja identidade do gene 16S chegando a 99% ao longo de todo o comprimento (ESCOBAR-PÁRAMO et al., 2003) (DEVANGA RAGUPATHI et al., 2018). Nesse caso específico, o *Greengenes* classifica até a família *Enterobacteriaceae*, enquanto o SILVA classifica ambas as espécies como *Escherichia coli*, acrescentando um novo problema na classificação *Escherichia-Shigella.* Casos como este fazem com que a correlação taxonômica seja maior nos níveis taxonômicos mais elevados. Em seu artigo, para evidenciar as diferenças entre os bancos, Edgar (EDGAR, 2018) extraiu sequências idênticas e registradas em ambos os bancos. Neste trabalho foram encontrados diferentes tipos de divergências; tais como: 1) classificações em branco, onde, de acordo com Siegwald, apenas 7% das sequências anotadas no *Greengenes* possui classificação em nível de espécie, e no SILVA menos de 45% (SIEGWALD et al., 2017); 2) conflitos entre táxons inferiores com o mesmo endereçamento.

Como consequência, a discrepância estimada entre os bancos foi de 34%, resultando em que uma a cada cinco das anotações não confiram (EDGAR, 2018).

Uma rápida consulta por palavras-chaves realizada no portal *Web of Science*, base vinculada ao periódico CAPES, revela a tendência maior de publicação na área de metagenômica do gene 16S rRNA. Foram usados os seguintes termos nas buscas: na primeira, *Greengenes*, e na segunda, *SILVA database*, resultando em 140 e 131 publicações, respectivamente. O portal possui uma classificação própria das áreas em que os periódicos são publicados, as quais foram agrupadas e simplificadas conforme mostrado na Tabela 4.3. Uma nova busca foi realizada, desta vez utilizando as palavras: QIIME; MOTHUR; QIIME and *Greengenes*; QIIME and SILVA; MOTHUR and SILVA; MOTHUR and *Greengenes*. E foram encontrados respectivamente: 269, 114, 21, 7, 10 e 3. É possível observar que há um número maior de publicações relacionadas ao *Greengenes* que ao SILVA, assim como o QIIME é mais citado do que o MOHTUR. Na busca cruzada é visto que há maior relacionamento entre pessoas que citam ao mesmo tempo QIIME e *Greengenes* do que QIIME e SILVA, Relembrando que os pesquisadores tendem a seguir a orientação dos desenvolvedores ao utilizar um banco. Quanto à área de publicação os artigos tendem a ser classificados em áreas mais genéricas, como Bioquímica, Biologia Celular, ou mesmo Biologia Molecular, mas há um bom percentual classificado que publica em áreas relacionadas ao ambiente, como solo, amostras de água ou mesmo saúde, como no caso de algumas doenças, ou microbioma de algum órgão. Nestes casos, é possível observar que as publicações que citam o SILVA tentem a ter maior correlação com a área ambiental, assim como o MOTHUR. Ao passo que, trabalhos que abordam a área da saúde citam *Greengenes*, com banco de dados usados e o QIIME como programa.

Tabela 4.3: Percentuais de publicações separados por categorias globais no portal *web of science*.

Categorial Global	SILVA database %131	*Greengenes* %140	QIIME %269	MOTHUR %114
Genérica	79,4	86,4	67,7	87,7
Ambiental	13,7	25,7	13,8	18,4
Saúde	8,4	21,4	20,8	16,7
Computacional	11,5	4,3	7,4	19,3
Ecologia	6,9	3,6	8,2	7,0

A escolha do banco de dados é a chave para análises a jusante, a qual depende substancialmente do tipo de investigação pretendida. O SILVA é maior que o *Greengenes*, isto faz com que mais tempo seja necessário para processar os dados. Em compensação, um maior número de referências aumenta as possibilidades de uma melhor definição do microbioma. Além disso, o SILVA possui dados do rRNA 18S, os quais o *Greengenes* não possui. A atualização do SILVA é realizada periodicamente, já o *Greengenes* foi atualizado pela última vez em maio de 2013. Por outro lado, existem ferramentas integradas ao QIIME, como o PICRUSt, que executam uma predição funcional do microbioma, a partir de dados obtidos pelo sequenciamento de um gene alvo. Mas para isso, é preciso que a taxonomia seja anotada de acordo com o *Greengenes*. Há ainda a possibilidade de se fazer essa escolha de forma técnica, pela preferência por metodologia de montagem, anotação e agrupamento do banco. A facilidade de acoplamento ao *pipeline* do QIIME impulsionam o uso do *Greengenes* em alguns casos. No entanto, para análises de biodiversidade através do rRNA 18S, o SILVA foi utilizado no presente estudo. Vale ressaltar que a conclusão biológica, neste grupo de amostras, foi a mesma e independente do banco empregado.

4.3.2.3. Plataformas de Sequenciamento

O sequenciamento da região V4 no *Illumina* gerou 1.062 OTU's, já na PMG foram 1.781, conforme apresenta a Tabela 4.4. Cerca de 90% das OTU's construídas receberam anotação taxonômica em ambas as plataformas. É possível observar que o sequenciamento realizado pelo PMG, além de gerar mais OTU's, produz um maior percentual de dados desconhecidos. Os valores apresentados na tabela não levam em consideração a frequência das OTU's. O número menor de OTU's agrupadas neste ensaio é resultante do perfil menos estringente para o agrupamento das OTU's, que neste caso foi de 97%, no lugar do até então usado 99%.

Para se calcular a diversidade dentro de uma amostra, diversidade alfa, é utilizado um artifício chamado de rarefação, que consiste no cálculo do índice CHAO1. Assim, tem-se a noção da complexidade da amostra. A Figura 4.6 trás o resultado de alfa diversidade, em A mostra o gráfico de rarefação calculado pelo índice CHAO1, e em B o gráfico de abundância relativa. Neste caso, é observado que apesar de gerar um número de *reads* menor, o *Ion Torrent* apresenta maior riqueza. O que é possível observar também no gráfico da abundância

relativa por espécies observadas. A inclinação da curva do *Illumina*, é maior, indicando que o número de espécies decai de forma mais rápida do que com o Ion. Além disso, é maior o número de espécies com menor abundância no *Ion* do que no *Illumina*.

Tabela 4.4 Valores totais das OTU's construídas com 97% de identidade com banco de dados *Greengenes*, na região V4 do gene 16S, no *Illumina* e no *Ion Torrent*.

Plataforma	Total de OTU's	OTU's anotadas	OTU's comuns		OTU's exclusivas	
			Anotadas	Desconhecidas	Anotadas	Desconhecidas
Ion	1781	1530	237	29	1343	172
Illumina	1062	967			730	66

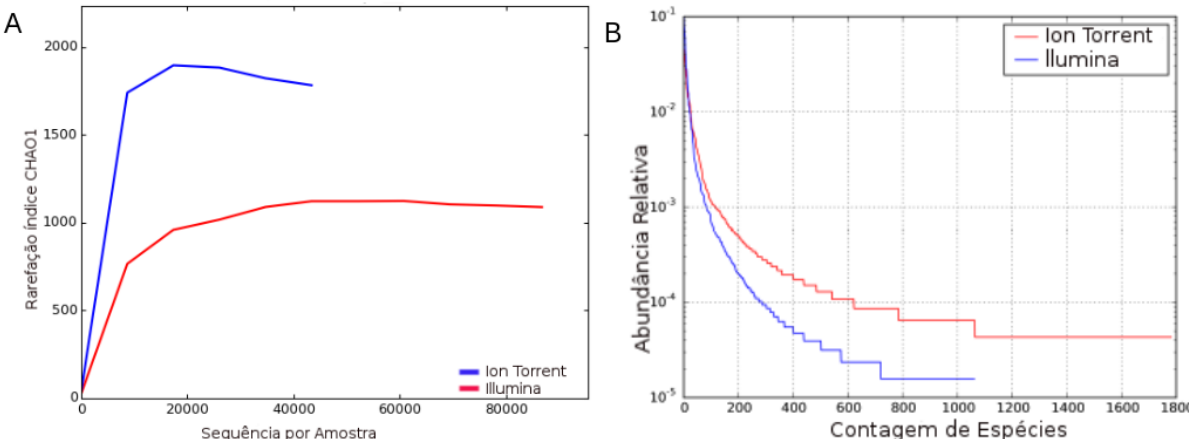

Figura 4.6 Alfa diversidade nas plataformas *Ion Torrent* e *Illumina*. (A) Curva de rarefação calculado pelo índice CHAO1, a curva em azul representa os dados sequenciados no *Ion Torrent*, e em vermelho no *Illumina*. (B) *Diagrama de Whittaker*, a curva em azul representa o método de construção *de novo*, em laranja o de Referência aberta e em vermelho o de Referência fechada.

A diversidade beta pelo índice *Unifrac* ponderada resultou em dissimilaridades em torno de 32%. Este cálculo considera, não apenas a distância filogenética entre as OTU's como também a frequência com que aparecem. Ao passo que a dissimilaridade calculada a partir do índice *Unifrac* não ponderado foi igual a 64%. Vale ressaltar que o diretório de saída do *<pick_open_reference_otus.py>*, gera uma árvore filogenética das OTU's criadas. Entretanto, foi utilizada a árvore padrão do *Greengenes* como base para o cálculo de diversidade, a fim de proporcionar um dado neutro, tendo em vista que as OTU's foram

montadas em arquivos distintos que, portanto, geraram árvores distintas. O cálculo da discrepância de acordo com índice *Bray-Curtis*, que é independente de taxonomia, foi igual a 0,88 e o índice *Jaccard*, 0,89.

A Figura 4.7 mostra um gráfico de abundância relativa, em que as OTU's estão identificadas, quando possível, até o nível de gênero. Cada uma é separada por cores, as mais abundântes estão presentes em ambas as plafaformas, como no caso o genero *Desulfovibrio*, que pelo I*llumina* foi identificado com 16,44% e pelo *Ion* 13,86%; em seguida a ordem Desulfovibrionales com 14,36% e 7,93%, o gênero *Pelobacter* 11,39% e 3,16%, a ordem Bacteroidales 7,72% e 4,79%, e a ordem Clostridiales 7,45% e 4,37%, respectivamente. No entanto, no *Ion Torrent* é observado um número maior de sequenciamento de gêneros minoritários. Através do comando <*group_significance.py*> foi calculado o quanto as OTU's estão diferenciadas significativamente, para isso foi utilizado o teste estatístico *Kruskal_wallis*, no qual possui a hipótese de que todas as OTU's estarão presentes, em uma mesma proporção nas amostras analisadas. Se a hipótese for verdadeira, retornará um *p value* alto, se as diferenças forem signicativas o *p value* será baixo (JOHN H. MCDONALD, 2014; OBREGON-TITO et al., 2015; PAULSON et al., 2013).

A Tabela 4.5 mostra os 30 gêneros mais abundantes. Vale ressaltar que as OTU's mais abundantes mas que não foram identificadas em nível de gênero não constam na tabela. Alguns relatos na literatura descrevem análises comparativas entre as duas plataformas de sequenciamento (*Ion Torrent* e *Illumina*), que cada vez mais se consolidam como as principais no mercado (ALLALI et al., 2017; GLENN, 2011; LIU et al., 2012; PYLRO et al., 2014b; SALIPANTE et al., 2014). No entanto, a dificuldade de tratamento dos dados da *Ion Torrent*, faz com que muitos optem por utilizar a *Illumina*. Muitos programas, inclusive a nova atualização do QIIME (v2.0. 2018) não trabalham com as leituras geradas no *Ion Torrent*, tornando uma dificuldade ainda maior para o tratamento dos dados. A empresa detentora dos direitos da PMG, atualmente a Thermo, desenvolveu um programa próprio para análises de amostras de origem clínica. O programa é limitado a usuários que possuam uma conta ativa com a Thermo, o que exige um pagamento anual, além de ser extremamente limitado e restritivo, forçando o analista utilizar também o pacote de *primers* preestabelecidos, cujas sequências não são disponibilizadas para o usuário, mostrando apenas o grupo de regiões que amplificam. Ademais, a qualidade dos dados *Illumina* tem se mostrado superior aos da *Ion*

Torrent (BARB et al., 2016; D'AMORE et al., 2016). Isso ocorre porque a plataforma *Illumina* permite que ambos os sentidos da fita do DNA sejam sequenciados (*pair ending*). No entanto, a *Illumina* possui como desvantagem o favorecimento de grupos majoritários (FOUHY et al., 2016), negligenciando aqueles que aparecem com pouca frequência; talvez por esse motivo, o número de OTU's que não foram identificadas pelo banco de dados seja, no caso da *Illumina*, menor do que na *Ion Torrent.* No entanto, este último apresenta uma vantagem: o sequenciamento na *Ion Torrent* tende a ser mais acessível, em termos de custo, comparado ao *Illumina*, tendo em vista o procedimento de detecção utilizado. O primeiro baseia-se em um potenciômetro e o segundo, em um fluorímetro (GLENN, 2011).

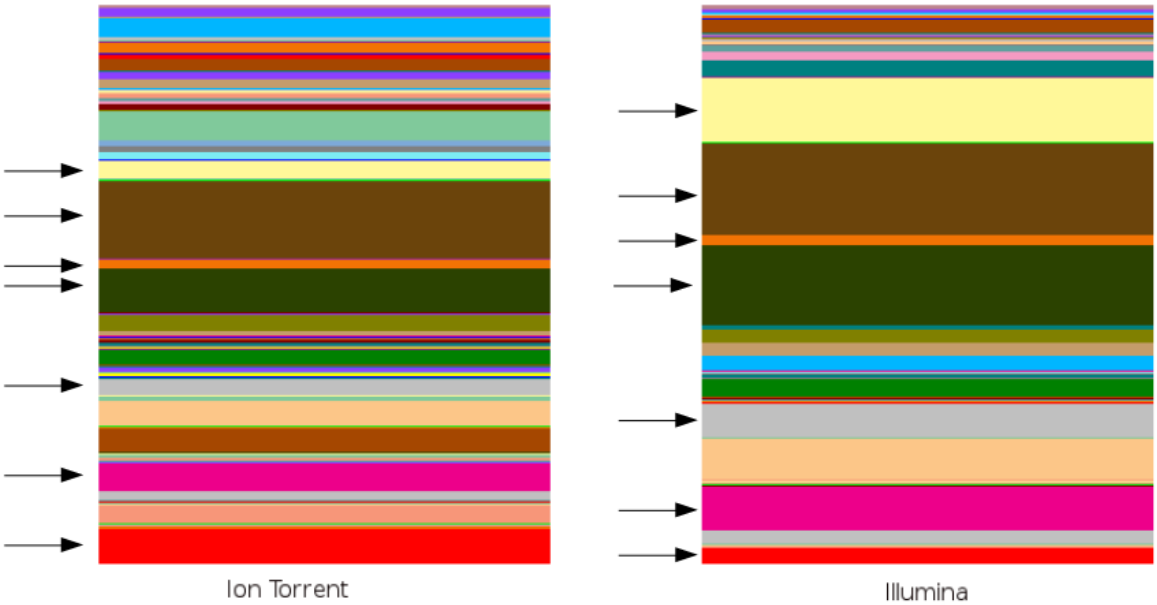

Figura 4.7 Abundância relativa em nível de gênero nas plataformas *Ion torrent* e *Illumina*. A barra à esquerda é representativo das leituras geradas no *Ion Torrent* e à direita, no *Illumina*. Indicados pelas setas estão alguns das OTU's anotadas, em vermelho, a abundância relativa do total das OTU's desconhecidas, em marrom *Desulfovibrio*, em verde musgo ordem *Desulfovibrionales*, em laranja *Desulfonauticos*, em amarelo pastel *Pelobacter*, em rosa choque ordem *Bacteroideae*, em cinza da família *Clostriciceae*.

Tabela 4.5: Abundância dos gêneros identificados

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***Ion***	***Illumina***
Desconhecido						*6,32%*	*2,90%*

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***Ion***	***Illumina***
Bacteria	Proteobacteria	Deltaproteobacteria	*Desulfovibrionales*	*Desulfovibrionaceae*	*Desulfovibrio*	*13,68%*	*16,44%*
Bacteria	Proteobacteria	Deltaproteobacteria	*Desulfuromonadales*	*Pelobacteraceae*	*Pelobacter*	*3,16%*	*11,39%*
Bacteria	Firmicutes	Clostridia	*Clostridiales*	*[Tissierellaceae]*	*ecb11*	*2,56%*	*2,89%*
Bacteria	Proteobacteria	Deltaproteobacteria	*Desulfobacterales*	*Desulfobacteraceae*	*Desulfobacter*	*2,97%*	*2,25%*
Bacteria	Thermotogae	Thermotogae	*Thermotogales*	*Thermotogaceae*	*Geotoga*	*2,25%*	*2,32%*
Bacteria	Firmicutes	Bacilli	*Bacillales*	*Bacillaceae*	*Natronobacillus*	*4,26%*	*0,01%*
Archaea	Euryarchaeota	Methanomicrobia	*Methanosarcinales*	*Methanosarcinaceae*	*Methanolobus*	*3,33%*	*0,23%*
Bacteria	Proteobacteria	Deltaproteobacteria	*Desulfovibrionales*	*Desulfohalobiaceae*	*Desulfonauticus*	*1,54%*	*1,89%*
Bacteria	Thermotogae	Thermotogae	*Thermotogales*	*Thermotogaceae*	*SC103*	*1,83%*	*0,65%*
Bacteria	WWE1	Cloacamonae	*Cloacamonales*	*Cloacamonaceae*	*W22*	*1,66%*	*0,45%*
Bacteria	Proteobacteria	Gammaproteobacteria	*Alteromonadales*	*Shewanellaceae*	*Shewanella*	*1,16%*	*0,01%*
Bacteria	Proteobacteria	Gammaproteobacteria	*Pseudomonadales*	*Pseudomonadaceae*	*Pseudomonas*	*1,06%*	*0,01%*
Bacteria	Chloroflexi	Anaerolineae	*Anaerolineales*	*Anaerolinaceae*	*T78*	*0,68%*	*0,32%*
Bacteria	Proteobacteria	Gammaproteobacteria	*Oceanospirillales*	*Halomonadaceae*	*Halomonas*	*0,96%*	*0,01%*
Bacteria	Thermotogae	Thermotogae	*Thermotogales*	*Thermotogaceae*	*Kosmotoga*	*0,80%*	*0,08%*
Bacteria	Firmicutes	Clostridia	*Clostridiales*	*[Acidaminobacteraceae]*	*Fusibacter*	*0,80%*	*0,01%*
Bacteria	Firmicutes	Clostridia	*Clostridiales*	*[Tissierellaceae]*	*Dethiosulfatibacter*	*0,28%*	*0,48%*

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***Ion***	***Illumina***
Bacteria	Firmicutes	Clostridia	*Halanaerobiales*	*Halanaerobiaceae*	*Halanaerobium*	*0,39%*	*0,00%*
Bacteria	Tenericutes	Mollicutes	*Acholeplasmatales*	*Acholeplasmataceae*	*Acholeplasma*	*0,05%*	*0,22%*
Bacteria	Firmicutes	Clostridia	*Clostridiales*	*Eubacteriaceae*	*Acetobacterium*	*0,01%*	*0,23%*
Bacteria	Proteobacteria	Deltaproteobacteria	*Desulfuromonadales*	*Desulfuromonadaceae*	*Desulfuromonas*	*0,06%*	*0,15%*
Bacteria	Firmicutes	Bacilli	*Lactobacillales*	*Streptococcaceae*	*Streptococcus*	*0,17%*	*0,04%*
Bacteria	Proteobacteria	Deltaproteobacteria	*Desulfarculales*	*Desulfarculaceae*	*Desulfarculus*	*0,06%*	*0,15%*
Bacteria	Firmicutes	Clostridia	*Clostridiales*	*Eubacteriaceae*	*Garciella*	*0,20%*	*0,00%*
Bacteria	Actinobacteria	Actinobacteria	*Actinomycetales*	*Corynebacteriaceae*	*Corynebacterium*	*0,19%*	*0,00%*
Bacteria	Thermotogae	Thermotogae	*Thermotogales*	*Thermotogaceae*	*Petrotoga*	*0,12%*	*0,06%*
Bacteria	Proteobacteria	Epsilonproteobacteria	*Campylobacterales*	*Campylobacteraceae*	*Arcobacter*	*0,16%*	*0,02%*
Bacteria	Firmicutes	Bacilli	*Bacillales*	*Staphylococcaceae*	*Staphylococcus*	*0,16%*	*0,00%*
Bacteria	Firmicutes	Clostridia	*Clostridiales*	*Clostridiaceae*	*Proteiniclasticum*	*0,15%*	*0,00%*
Archaea	Euryarchaeota	Methanococci	*Methanococcales*	*Methanococcaceae*	*Methanococcus*	*0,12%*	*0,02%*

Uma consequência direta da qualidade das leituras e do sequenciamento é o quanto foi removido no pré-tratamento dos resultados das duas plataformas: a remoção de leituras de baixa qualidade foi maior no *Ion* do que no *Illumina*, tabelas 4.1 e 4.2. Após a montagem e agrupamento das OTU's houve um maior aproveitamento dos dados gerados pelo *Illumina* em análises de predição funcional realizadas no PICRUSt. Apesar do imenso trabalho e da dificuldade de adaptação dos dados gerados na plataforma PMG, em se tratando da região V4 do rRNA (amplificada pelos *primers* 515F e 806R) a conclusão biológica é equivalente, tendo

em vista que ambas as plataformas identificaram os mesmos grupos taxonômicos. Em relação ao total dos custos, a análise do *Ion Torrent* foi cerca de 30% menos dispendiosa do que do *Illumina*.

4.3.3. Análise da Diversidade das Regiões Hipervariáveis

Em cada plataforma um grupo diferente de regiões foi analisada. Os resultados são apresentados nas seções seguintes.

4.3.3.1. Ion Torrent

Para a análise da diversidade nas diferentes regiões sequenciadas através da plataforma *Ion Torrent* o comando <*pick_open_reference.py*> foi utilizado com todos os parâmetros sem modificação. A identidade para o agrupamento das OTU's foi de 97%, foi utilizado o banco de dados *Greengenes*, e não houve ajuste no sentido da sequência em relação ao banco de dados para a atribuição taxonômica. Com isso, as regiões V1F, V1R, V4F, V4R, V6F, V6R e V9R, obtiveram respectivamente 125, 176, 239, 153, 383, 359 e 1 OTU's. A Figura 4.8 apresenta, na devida ordem, o gráfico de rarefação e o *Diagrama de Whittaker* para cada região sequenciada. A Tabela 4.6 mostra a estimativa pontual de probabilidade de cada região em descobrir uma nova OTU. O método foi criado por Lladser, Gouet e Reeder em 2011, que defendem que as medidas de diversidade são duvidosas e dependentes da escolha do parâmetro utilizado (LLADSER; GOUET; REEDER, 2011). A proposta do algoritmo é medir a probabilidade de se encontrar uma nova espécie na amostra sequenciada, o que pode ser realizado na forma de rarefação, em que é realizada uma subamostragem e a probabilidade é calculada para cada subconjunto, gerando um gráfico exponencial e similar ao *Diagrama de Whittaker*; ou ainda, como realizado aqui, partir-se do conjunto total para calcular apenas a probabilidade de existência de uma nova OTU, a partir das OTU's já construídas (LLADSER; GOUET; REEDER, 2011). Nesse sentido, os valores apresentados representam o potencial de novas descobertas, estando os baixos valores encontrados relacionados à biosfera rara; isto é, mesmo que uma nova OTU fosse encontrada, esta estaria com uma baixa abundância relativa. É possível notar a complementaridade dos dados, a qual, através de caminhos diferentes, como alfa rarefação, o *Diagrama de Whittaker* e a

probabilidade de novas descobertas, indicam que a região V6 é a aquela que mais conseguiu identificar OTU's diversas, com maior potencial para novas descobertas.

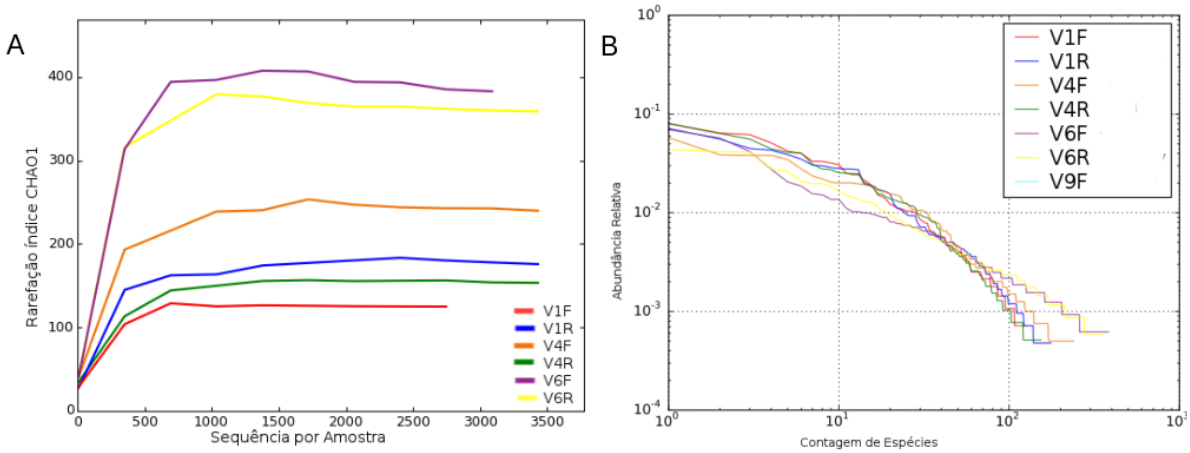

Figura 4.8. Gráficos de rarefação e abundância relativa nas multiplas regiões sequenciadas no *Ion Torrent*. (A) cálculo pelo índice CHAO1, (B) diagrama de *Whittaker.* Em A e B: a curva em vermelho representa a região V1F, em azul V1R, em laranja V4F, em verde V4R, em roxo V6F, e em amarelo V6R. V9F não aparece pelo baixo número de espécies presentes.

Tabela 4.6. Probabilidade de Descobrimento Pontual para Cada Nova OTU

Probabilidade de Descobrimento de uma nova OTU	
V1F	0,020
V1R	0,014
V4F	0,016
V4R	0,014
V6F	0,031
V6R	0,022
V9F	nan

Na busca por uma região do gene rRNA 16S que conseguisse trazer a mesma representatividade que o gene inteiro, Minseok Kim e colaboradores (KIM; MORRISON; YU, 2011) realizaram um estudo no qual utilizaram tanto amostras sequenciadas do gene integral quanto de suas diferentes regiões. Sua pesquisa, no entanto, envolveu sequências de comprimentos maiores, obtidas com outros tipos de sequenciadores que não os utilizados no

presente trabalho. Eles concluem que as regiões não respondem uniformemente à diversidade e riqueza amostral. Reconhecem o avanço do sequenciamento de nova geração, mas advertem que os dados devem ser interpretados com cautela (KIM; MORRISON; YU, 2011). É possível observar que de fato há uma grande variação nos valores calculados através do *índice CHAO1* para diversidade alfa, como também na estimativa da riqueza pelo *Diagrama de Whittaker*. Em ambos é possível observar que todas as amostras *forward* apresentam maiores valores de riqueza amostral, do que as *reverse*. E as regiões de maior a menor probabilidade de descobrimento de uma OTU, dentro das configurações da análise, são V6, V4, V1 e V9, respectivamente.

Para verificar o quanto as amostras (regiões sequenciadas) são similares entre si, comandos de comparação entre amostras diferentes foram realizados. O primeiro foi o <*shared_otus.py*> que revela quantas OTU's idênticas existem dentro do conjunto de amostras, o resultado é uma matriz quadrática apresentada na Tabela 4.7. É possível observar que, além de possuir um maior número de OTU's, a V6 ainda é a região que mais tem em compatilha, em termos percentuais, com as demais.

Tabela 4.7. Matriz de compartilhamento das OTU's idênticas.

	V1F	V1R	V4F	V4R	V6F	V6R	V9F
V1F	125	32	41	32	66	48	1
V1R	32	176	48	64	95	105	1
V4F	41	48	239	41	100	72	1
V4R	32	64	41	153	71	93	1
V6F	66	95	100	71	383	153	1
V6R	48	105	72	93	153	359	1
V9F	1	1	1	1	1	1	1

A Tabela 4.8 mostra os valores calculados em três índices de beta diversidade. O primeiro deles, *Unifrac* ponderado, é dependente tanto das distâncias filogenéticas quanto da frequência com que os táxons são encontrados, isto significa que a biosfera rara possui menor peso do que a majoritária. Os índices seguintes são independentes da taxonomia: o *Bray Curtis*, que quantifica a diferença na composição, mas que não é uma medida de distância como o *Jaccard*. Foi calculada também a significância, sendo que todos os casos

apresentaram um valor de *p* menor que um. A dissimilaridade média entre todos os *clusters* foi igual a 0,37 ± 0,15 no índice *Unifrac* ponderado, e nos índices independentes de taxonomia, *Bray Curtis* e *Jaccard*, 0,92 ± 0,06 e 0,88 ± 0,08, respectivamente.

Como nesta avaliação o número de amostras é maior, foi possível a partir dos dados de diversidade acima apresentados montar gráficos para visualizar as distâncias entre as regiões do rRNA 16S. A Figura 4.9 traz o gráfico de coordenadas principais, calculado a partir dos valores de beta diversidade com índice UniFrac ponderado, o qual é apresentado em três dimensões, e, para simplificar, a imagem possui três recortes em duas direções. É possível observar que na primeira dimensão, PC1 x PC2, as sequências no sentido *forward*, isto é 5'-3', estão mais agrupadas a exceção da V9, que possui a menor riqueza. Na segunda dimensão, PC2 x PC3, há uma nova divisão entre *forward* e *reverse*, com uma melhor aproximação do segundo grupo. E na terceira, e última dimensão, PC2 x PC3, ocorre um espalhamento, onde o V9 fica no centro, mais distantes as amostras *reverse*, e próximo a ele V1F e V4F, os quais mantiveram-se próximos em todo momento. A Figura 4.10, confirma as proximidades e distâncias entre as regiões, mostrando o gráfico em caixa de distâncias, construído a partir da diversidade beta com o índice UniFrac não ponderado, realizados por meio do teste t de *Student* de duas amostras com a hipótese de que o valor de diversidade do primeiro grupo é igual à do segundo.

Tabela 4.8. Matriz de diversidade beta

V	V1F			V1R			V4F			V4R			V6F			V6R		
	UP	BC	Jc	UP	BC	Jc	UP	BC	Jc	UP	BC	Jc	UP	BC	Jc	UP	BC	Jc
1F	0.0	0.0	0.0	0,29	0,93	0,88	0,20	0,78	0,87	0,29	0,94	0,85	0,23	0,85	0,85	0,26	0,93	0,89
1R	0,29	0,93	0,88	0,0	0,0	0,0	0,28	0,92	0,87	0,38	0,90	0,80	0,32	0,90	0,80	0,31	0,86	0,76
4F	0.20	0,78	0,87	0,28	0,92	0,87	0,0	0,0	0,0	0,26	0,90	0,81	0,18	0,85	0,81	0,26	0,93	0,86
4R	0,29	0,94	0,87	0,38	0,90	0,76	0,26	0,90	0,88	0,0	0,0	0,85	0,31	0,92	0,85	0,25	0,85	0,78
6F	0,23	0,85	0,85	0,32	0,90	0,80	0,18	0,85	0,81	0,31	0,92	0,0	0,0	0,0	0,0	0,26	0,84	0,74
6R	0,26	0,93	0,89	0,31	0,86	0,76	0,26	0,93	0,78	0,25	0,85	0,74	0,26	0,84	0,74	0,0	0,0	0,0

UP = *Unifrac* Ponderada, que é dependente de taxonomia e da frequência de cada táxon; BC = *Bray Curtis* e Jc = *Jaccard*, independentes de taxonomia, sendo o primeiro uma medida semimétrica e o segundo uma métrica que obedece a desigualdade triangular. Foi omitido na tabela a sequência V9F, pois só tinha apresentou uma única OTU. Em todos os casos o *p value* foi menor que 1.

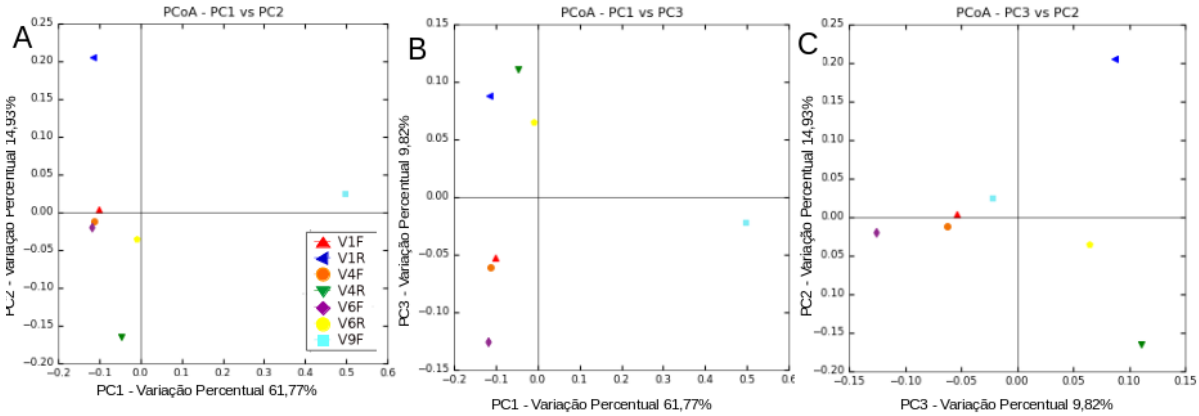

Figura 4.9. Beta diversidade pelo índice *Unifrac* Ponderado nas multiplas regiões sequenciadas no *Ion Torrent*. Gráfico de Coordenadas Principais obtidas a partir do cálculo de Beta diversidade pelo índice *Unifrac* Ponderado; em cortes de duas direções, os pontos são representados pelo sequenciamento da região V1F em vermelho, V1R em azul-escuro, V4F em laranja, V4R em verde, V6F em roxo, V6R em amarelo e V9F azul-piscina.

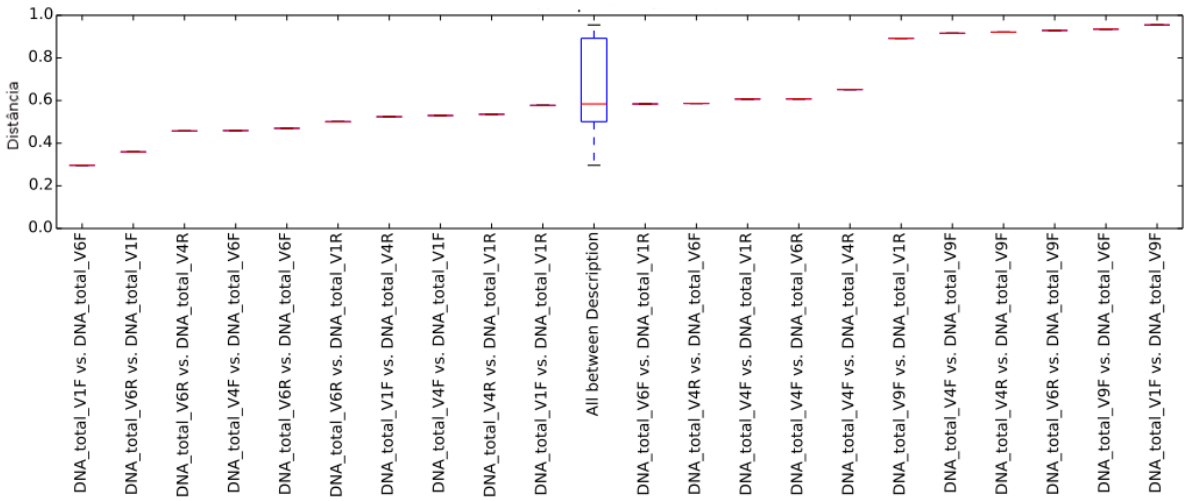

Figura 4.10. Gráfico de distância calculada de dois a dois, a partir do índice *UniFrac* não ponderado nas multiplas regiões sequenciadas no *Ion Torrent*. Emprego do teste t de *Student*, em que o primeiro grupo é igual ao segundo.

Guo et al, realizaram uma pesquisa similar com amostra microbiana complexa de lodo ativado, sem um modelo como prova real, assim como neste trabalho, sequenciando diferentes

regiões hipervariáveis do rRNA 16S. E utilizaram diferentes métodos para comparar seus resultados, com a premissa de que a consistência nos resultados traria maior confiabilidade. Mesmo assim, seus resultados indicam que não existe uma região perfeita, e concluem que para melhor avaliação de uma amostra ambiental é através da comparação dos resultados com outros trabalhos publicados. Por exemplo, por esta abordagem descobriram que a região V3V4 apresentou maior riqueza, mas omitiu gêneros que são frequentemente relatados na literatura, e que a região V1V2 foi a que melhor coincidiu com a literatura (CAI et al., 2013; GUO et al., 2013) (CAI et al., 2013; GUO et al., 2013). No presente trabalho, a partir do sequenciamento realizado no *Ion Torrent*, a região V6 apresentou maior riqueza.

A Figura 4.11 mostra o gráfico em barras de abundância relativa dos filos identificados nas amostras e a Tabela 4.9, a abundância relativa de cada um deles. É possível observar que os principais grupos são reconhecidos independente da região sequenciada. A Tabela 4.10 apresenta o coeficiente de correlação taxonômica e seu intervalo, o *p value* foi menor que 1, em todos os casos.

O diferencial entre as regiões, além da frequência das OTU's, é a identificação da biosfera rara. A região V6 é a que mais reconhece e identifica os táxons menos representativos, como visto na figura 4.8. A região V1 foi a que apresentou maior percentual de amostras não reconhecidas, apesar de estar próximo de V6 e V4, como mostram as análises de PCoA e *box-plot*. Enquanto V1R apresentou o maior percentual de organismos não reconhecidos pelo banco de dados, o V6F apresenta o menor valor, com uma diferença de quase dez vezes. E, de forma geral, os sentidos *reverse* apresentaram maior percentagem de organismos desconhecidos do que os *forwards*, dentro da mesma região. O próprio manual do QIIME adverte que o sentido em que o *read* é sequenciado determina a abrangência ou não da atribuição taxonômica, pois esta é dependente da orientação em que está anotado no banco. E, no caso desta análise, propositalmente não houve correção de orientação.

É possível observar na Tabela 4.9 que cerca de 50% dos filos identificados são representados em todas as regiões. Ao passo que V6 reconhece todos os filos apresentados, as regiões V1 e V4 excluem pelo menos um dos filos, como é o caso do *Chlorobi*, anaeróbio obrigatório e fotoautotrófico, e do *ZB3*, um organismo não cultivável e que ainda não recebeu uma classificação oficial. Apesar dos pares de *primers* usados no sequenciamento no *Ion Torrent* não serem específicos para *Archaea*, apenas a região V1 não identificou nenhum

organismo desse domínio. Como já visto nas análises de diversidade, a região V6 apresentou filos únicos, cuja ocorrência foi exclusiva e em baixa abundância. No total, cinco filos foram identificados em ambos os sentidos, sete em apenas um deles (três no *forward* e quatro no *reverse*).

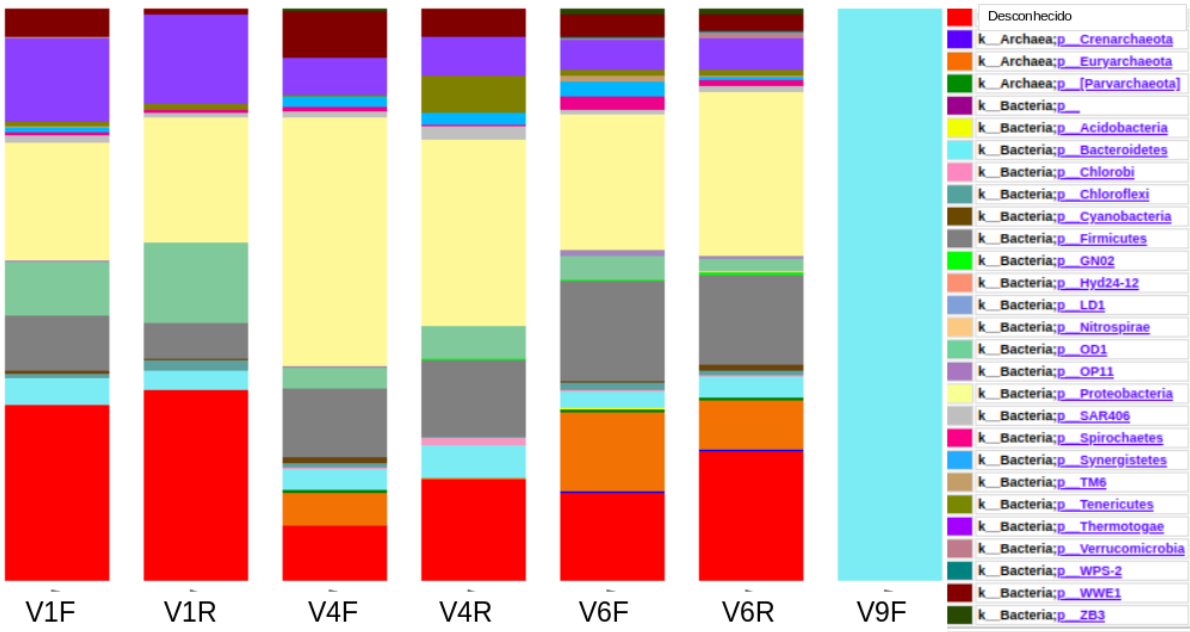

Figura 4.11. Abundância relativa em nível de filo nas multiplas regiões sequenciadas no *Ion Torrent*. . Da esquerda para direita, as colunas são representadas pelo sequenciamento da região V1F, V1R, V4F, V4R, V6F, V6R e V9F. Em vermelho, a abundância relativa do total das OTU's desconhecidas, a legenda com as cores dos filos identificados está à direita da imagem.

O índice de correlação taxonômica leva em consideração apenas os resumos dos táxons criados ao medir a contribuição de cada região, sendo interessante tendo em vista que os demais métodos de comparação, em suas métricas consideram também as OTU's que não foram identificadas, e este apenas o resumo classificação taxonômica. Para calcular o intervalo de confiança, o *script* considera a primeira amostra como padrão e a segunda o esperado, através de testes estatístico como no caso utilizado o *pearson*.

Tabela 4.9. Resumo taxonômico de cada região sequenciada do gene rRNA 16S pelo *Ion Torrent.*

Domínio	Filo	V1F	V1R	V4F	V4R	V6F	V6R
Desconhecido		30,82%	34,86%	8,33%	9,88%	3,07%	3,93%
Bacteria	*WWE1*	4,96%	0,74%	7,85%	6,47%	3,38%	3,34%
Bacteria	*Bacteroidetes*	4,50%	3,06%	7,68%	5,21%	1,52%	0,35%
Bacteria	*Thermotogae*	14,68%	16,00%	4,84%	4,13%	5,28%	3,49%
Bacteria	*Tenericutes*	0,79%	0,45%	1,61%	1,39%	1,46%	1,48%
Bacteria	*Synergistetes*	0,71%	0,60%	1,91%	2,67%	2,51%	2,65%
Bacteria	*Spirochaetes*	0,68%	0,31%	0,95%	0,33%	1,65%	2,04%
Bacteria	*SAR406*	1,36%	1,27%	1,08%	0,51%	1,46%	1,51%
Bacteria	*Proteobacteria*	20,50%	21,96%	39,91%	43,80%	26,05%	25,39%
Bacteria	*OD1*	9,46%	10,60%	2,16%	1,75%	3,63%	1,63%
Bacteria	*Firmicutes*	9,50%	9,19%	15,68%	14,94%	27,13%	29,83%
Bacteria	*Cyanobacteria*	0,50%	0,22%	0,05%	-	0,34%	0,41%
Bacteria	*Chloroflexi*	0,86%	0,36%	0,75%	0,80%	2,08%	0,90%
Bacteria	*Chlorobi*	0,07%	0,07%	-	-	0,25%	0,29%
Bacteria	*TM6*	0,21%	0,17%	0,13%	0,33%	0,09%	-
Bacteria	*Verrucomicrobia*	0,14%	0,10%	0,08%	-	0,90%	0,67%
Bacteria	*OP11*	0,25%	-	-	0,08%	0,96%	0,61%
Bacteria	*GN02*	-	0,05%	0,10%	-	0,28%	-
Bacteria	*ZB3*	-	-	0,63%	1,13%	1,18%	1,60%
Archaea	*Euryarchaeota*	-	-	5,75%	5,44%	15,68%	18,70%
Archaea	*[Parvarchaeota]*	-	-	0,53%	1,13%	0,25%	0,38%
Archaea	*Crenarchaeota*	-	-	-	-	0,31%	0,23%

Domínio	Filo	V1F	V1R	V4F	V4R	V6F	V6R
Bacteria	*Acidobacteria*	-	-	-	-	0,06%	0,06%
Bacteria	*Planctomycetes*	-	-	-	-	0,06%	0,06%
Bacteria	*OP3*	-	-	-	-	0,12%	0,09%
Bacteria	*LD1*	-	-	-	-	0,06%	0,06%
Bacteria	*Hyd24-12*	-	-	-	-	0,09%	-
Bacteria	*WPS-2*	-	-	-	-	0,06%	-
Bacteria	*BRC1*	-	-	-	-	0,06%	-
Bacteria	*[Caldithrix]*	-	-	-	-	-	0,06%
Bacteria	*Nitrospirae*	-	-	-	-	-	0,09%
Bacteria	*Lentisphaerae*	-	-	-	-	-	0,06%
Bacteria	*WS1*	-	-	-	-	-	0,06%

Tabela 4.10. Coeficiente de correlação e intervalo de confiança calculados par a par, a partir do resumo taxonômico em nível de gênero.

Regiões	Coeficiente de Correlação		
	Médio	Intervalo de Confiança	
V1F_V1R	0.9570	0.9251	0.9755
V4F_V4R	0.9055	0.8608	0.9363
V1F_V9F	0.9014	0.8187	0.9474
V1R_V9F	0.8962	0.8154	0.9428
V6F_V6R	0.8408	0.7905	0.8799
V4F_V1F	0.5712	0.4119	0.6966
V4R_V1F	0.5620	0.3918	0.6951
V4R_V1R	0.5427	0.3715	0.6782
V4F_V6F	0.5312	0.4126	0.6321
V4F_V1R	0.5004	0.3283	0.6404
V4F_V6R	0.4756	0.3493	0.5849

Regiões	Coeficiente de Correlação		
	Médio	Intervalo de Confiança	
V4R_V6F	0.4459	0.3115	0.5629
V4R_V9F	0.4399	0.2286	0.6117
V4R_V6R	0.3999	0.2609	0.5226
V4F_V9F	0.3826	0.1764	0.5566
V1F_V6F	0.3068	0.1490	0.4493
V1R_V6F	0.2895	0.1316	0.4330
V1F_V6R	0.2460	0.0859	0.3937
V1R_V6R	0.2252	0.0652	0.3740
V6F_V9F	0.1305	-0.0387	0.2924
V6R_V9F	0.1223	-0.0476	0.2854

A teoria da dispensabilidade dos genes diz que a evolução é inversamente proporcional à importância da função (YANG; WANG; QIAN, 2016). As regiões do gene rRNA 16S podem ser agrupadas em classes, de acordo com a estrutura terciária e função. A classe I é composta por V4, V5, V6 e V9, abrangendo o *hairpin* e o centro de decodificação. É relacionada à ligação do RNA transportador ao sítio P e interações com a alça do rRNA 23S. Por sua significância funcional, foi considerada a região com menor número de polimorfismos. A Classe II engloba as regiões periféricas V3 e V7, com funções não confirmadas. As regiões da Classe III, V2 e V8, ficam na parte inferior e superior da estrutura tridimensional e funcionam como estabilizadores funcionais. Bo Yang e colaboradores, em 2016, avaliaram a potencialidade, *in silico*, das regiões do gene rRNA 16S para caracterização em nível de filo. Para isso, eles extraíram do banco de dados SILVA-115 não redundante um conjunto de sequências do rRNA 16S do domínio *Bacteria* que obedecessem à critérios por eles estabelecidos e que fossem as mais amplas possíveis. Realizaram a fragmentação do gene e avaliaram quanto à capacidade de resolução filogenética. Por distância geodésica, concluíram que a região V4, de todas, apresentou melhor resolução para bactérias, em nível de filo. No entanto, a forma de segmentação adotada para o gene os fez abandonar as regiões V1 e V9. Eles concluem que as regiões centrais, de 515F a 1100R, foram mais adequadas para estudo de ambientes extremos, com novas linhagens bacterianas. Sugeriram que as regiões periféricas, classe II, poderiam ter um papel melhor no entendimento em nível de gênero, mas não ficou comprovado. Além disso, eles verificaram o potencial da resolução e não dos

primers. Pela escolha da análise, as regiões V1 e V9 não foram analisadas (YANG; WANG; QIAN, 2016)

A região V9 por ter identificado apenas um organismo apresenta, independente da análise, os resultados mais distintos do que das outras regiões. O usado *primer* foi projetado por Linda Amaral-Zettler, que em seu artigo afirma que a região por ser compartilhada entre os três domínios apresenta maior potencial de uso (AMARAL-ZETTLER et al., 2009). Em seu trabalho ela chegou a fazer um par específico para eucariotos, mas, no presente trabalho, esta amplificação não ocorreu como esperado, sendo o sequenciamento realizado apenas para procariotos. Apesar do potencial apontado para essa região, o *primer* não foi representativo para a amostra de água de produção. O comprimento do *amplicon* pode ser visto como um empecilho, tendo em vista que um maior comprimento favorece a identificação do táxon. Mesmo assim, a região V9 foi a que obteve o maior número de *reads*, como demonstrado na tabela 4.2. No entanto, a maior parte dessas leituras foi cortada por falta de qualidade.

Wang e Qian comentam que um *primer*, ainda que projetado em uma região conservada, pode possuir em seu comprimento um alto índice de polimorfismos, em determinados grupos. Nestes casos, a percentagem de espécies bacterianas em amostras ambientais não cultiváveis cai consideravelmente (WANG; QIAN, 2009) Outros pesquisadores sugerem que a projeção dos iniciadores seja feita de forma específica, para o nicho de onde a amostra foi retirada, tendo em vista que localidades similares possuem biomas comparáveis (GUO et al., 2013; ZHANG et al., 2018). Nesta perspectiva, não apenas a região sequenciada importa, mas também o desenho do iniciador que se anelará ao gene. Dessa forma, o resultado inesperado encontrado no sequenciamento da região V9 não é apenas consequência de uma provável falta de resolução da região. Mas parece depender, sobretudo, dos *primers* que aparentam não possuir, para este tipo de amostra, as características necessárias para uma maior representatividade. Na próxima seção apresentam-se os resultados do sequenciamento realizado pela plataforma *Illumina*, os quais, entre outros, possibilitou a análise da região V9.

4.3.3.2. Illumina

De forma semelhante à análise de diversidade realizada na plataforma *Ion Torrent*, os dados gerados no sequenciamento pela plataforma *Illumina* seguiram o mesmo fluxo de

tratamento. Com a utilização do comando <*pick_open_reference.py*> sem modificação, agrupamento das OTU's em 97% e utilização do banco de dados *Greengenes*. No entanto, como a plataforma permite sequenciar ambos os sentidos do fragmento de DNA, foi possível juntá-los criando uma sequência consenso. As regiões V3V4, V4, V4V5, obtiveram respectivamente 1.570, 1.225 e 3.513 OTU's. Além dessas regiões, foram analisadas regiões específicas para *Archaea* e eucariotos: 16SA, 18SV4 e 18SV9. As sequências geradas na região 18SV4, não puderam ser alinhadas pelo PyNAST, sendo esta região removida das análises. Para essas regiões, o número de OTU's geradas foi igual a 340 e 37 (16SA e 18SV9, respectivamente). Como o banco *Greengenes* é específico para procariotos, todos os dados gerados na plataforma *Illumina* foram analisadas com anotação taxonômica no banco de dados SILVA-v132. As diferenças entre os bancos foram discutidas na seção 4.3.2.2 e não serão retomadas. Na presente seção, a fim de facilitar a leitura, os resultados obtidos com o banco de dados SILVA-v132 ficarão disponíveis no anexo.

A Figura 4.12 mostra respectivamente o gráfico de rarefação (A) e o *Diagrama de Whittaker* (B) para cada região sequenciada e as Tabelas 4.11 e 4.12, os valores da estimativa pontual de probabilidade de cada região em descobrir uma nova OTU, além da Matriz de compartilhamento de OTU's idênticas.

Como já descrito nas seções anteriores, é possível correlacionar a riqueza amostral, em relação ao número de OTU's, através do conjunto de dados apresentados na Figuras 4.15 e na Tabela 4.11, tendo em vista que as análises são complementares. Desta feita, observa-se que a região rRNA 16S V4V5, cujo comprimento é de 393 pares de bases, apresenta maior riqueza em relação às demais, e a região específica para Archaea a menor riqueza. As três regiões específicas para bactérias, 16SV3V4, 16SV4, e 16V4V5 possuem a região central do gene 16S em comum, variando apenas no comprimento, fazendo com que as regiões adjacentes fossem sequenciadas, o que resulta em um acréscimo da riqueza e na maior probabilidade de se encontrar uma nova OTU, dado que vai ao encontro do trabalho do Bo Yang, publicado em 2016, que analisa o perfil de resolução ao nível de filo, em diferentes regiões do gene rRNA 16S. Nesse trabalho, concluem que entre as regiões V4 e V6 está o melhor perfil resolutivo (YANG; WANG; QIAN, 2016).

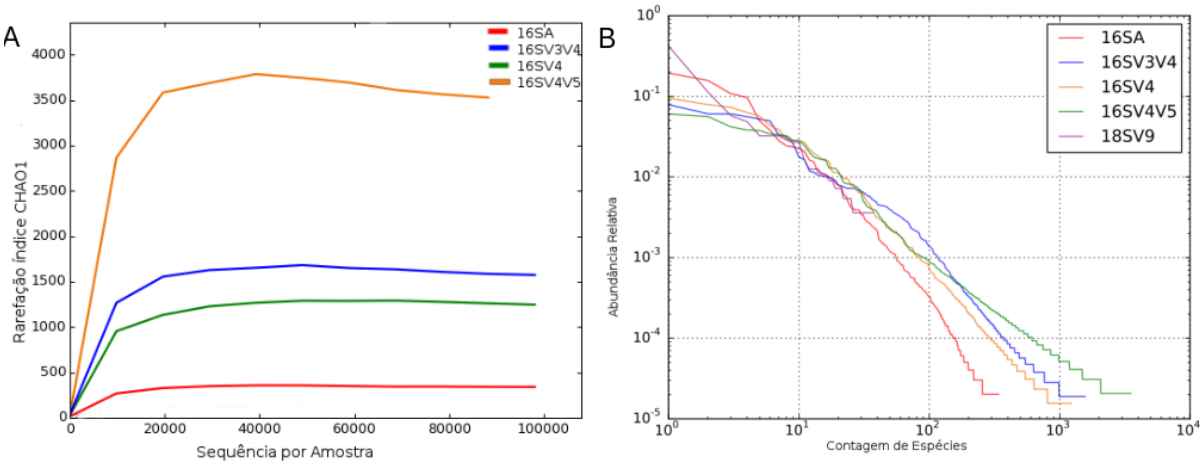

Figura 4.12. Gráfico de rarefação calculado pelo índice CHAO1 e o diagrama de *Whittaker* nas multiplas regiões sequenciadas no *Illumina.* Em ambas as imagens a curva em vermelho representa a região 16SA em azul, 16SV3V4, em laranja 16SV4V5, em verde e 16SV4, em laranja. O número de OTU's gerados em 18SV9 gerou uma curva que não aparece devido ao escalonamento das demais.

Tabela 4.11. Probabilidade de Descobrimento Pontual para cada nova OTU nas multiplas regiões sequenciadas no *Illumina.*

Região	Probabilidade ($x10^{-4}$)
16SA	5,8
16SV3V4	12,8
16SV4	10,4
16SV4V5	25,2
18SV9	572,2

Não apenas Yang, mas também Minseok Kim, por uma pesquisa *in silico* verificou que maiores comprimentos geram dados mais confiáveis tanto quanto à riqueza amostral como à anotação taxonômica. No entanto, o tamanho dos fragmentos dos quais obtiveram os melhores resultados são incompatíveis com as tecnologias NGS (KIM; MORRISON; YU, 2011). Martínez-Porchas e colaboradores avaliaram três regiões centrais que compararam com os dados do gene completo, por PCR virtual realizado a partir do banco de dados SILVA. As regiões avaliadas foram, V1V9, V3V5, V3V4 e V4V5, levando à conclusão de que a melhor região para ser analisada, tanto em número de riqueza como no baixo índice de falsos positivos, que possam ser sequenciadas por NGS, é a V3V4 (MARTÍNEZ-PORCHAS;

VILLALPANDO-CANCHOLA; VARGAS-ALBORES, 2016). Apesar de válidos, os dados obtidos *in silico* nem sempre conseguem ser reproduzidos. O tipo de amostra, as condições de PCR e até mesmo o desenho dos *primers* são determinantes no resultado. Nesse sentido, dois trabalhos podem ser destacados pois lavam a conclusões distintas, através do mesmo tipo de análise com amostras em origens diferentes. Claesson avaliou em fezes humanas a sensibilidade das regiões V3V4 e V4V5, concluindo que, apesar das análises *in silico*, que a região V4V5 foi melhor. Já Lin Cai usou amostras do sequenciamento de lodo ativo e seus resultados foram contrários aos do primeiro grupo (CAI et al., 2013; CLAESSON et al., 2010). Feng Guo, defende que não existe conjunto perfeito de *primers* e que é necessário avaliar os diferentes métodos e regiões, além se realizar uma verificação cruzada de múltiplas variáveis para a atribuição taxonômica (GUO et al., 2013).

Análises de beta diversidade são comparações entre duas amostras distintas, com o objetivo de verificar a correlação entre elas. Sejam no número de OTU's, ou mesmo na taxonomia entre os grupos. Ao se avaliar diferentes métodos para uma mesma amostra há de se tomar cuidado com a métrica escolhida para análise, para não chegar a conclusões equivocadas (SCHLOSS, 2010). Tendo isto em vista, diferentes abordagens foram utilizadas e comparadas entre si.

A Tabela 4.12 mostra a Matriz de compartilhamento das OTU's idênticas nas diferentes regiões e a Tabela 4.13, os valores da diversidade beta calculados par a par, nos índices Unifrac Ponderado, *Bray-Curtis* e Jaccard. As médias foram iguais a (0.81 ± 0.43), (0.95 ± 0.08) e (0.80 ± 0.17) respectivamente. Ao desconsiderar as regiões específicas para Archaea e eucariotos, tendo em vista que priorizam as espécies menos favorecidas nas outras regiões, os valores de diversidade beta média são iguais a (0,18 ± 0,04), (0,79 ± 0,05) e (0,59 ± 0,05). A Figura 4.17 apresenta o gráfico PCoA calculado a partir do índice Unifrac Ponderado.

O maior número de OTU's é observado na amostra 16SV4V5, que também é a região que mais compartilha OTU's com 16SV4. Assim como ocorreu com o sequenciamento da plataforma *Ion Torrent*, a região V9, com *primers* desenhados para alinhar mais com eucariotos, foi a região com menor número de OTU's, de forma que para região rRNA 18SV4 não foi possível realizar um alinhamento entre as sequências obtidos e o banco de dados, pelo PyNAST.

Tabela 4.12. Matriz de compartilhamento das OTU's idênticas nas multiplas regiões sequenciadas no *Illumina.*

	16SA	**16SV3V4**	**16SV4**	**16SV4V5**	**18SV9**
16SA	340	133	155	126	21
16SV3V4	133	1570	365	364	14
16SV4	155	365	1225	371	13
16SV4V5	126	364	371	3513	15
18SV9	21	14	13	15	37

A beta diversidade utiliza diferentes métricas das quais seja possível avaliar a dissimilaridade entre dois grupos. Portanto, é uma abordagem que avalia a distância entre amostras, como uma forma de mensurar a matriz de compartilhamento entre OTU's, que torna os dados complementares entre si. Percebe-se que os maiores valores registrados são das regiões que favorecem um domínio. Ao observar especificamente a região central do gene rRNA 16S, é possível observar que a aproximação é maior para a região do Clado I. Além disso, a região V4V5 foi a que mais se aproximou da região V9, em comparação às outras. Outra forma de observar, é através do gráfico PCoA, o qual é a representação em forma de gráfico das distâncias de beta diversidade, sendo neste caso usada a métrica UNIFRAC ponderada. É possível visualizar a proximidade entre a região central do gene 16S, a distância maior para a região V9 do domínio *Archaea*. É possível ainda verificar que a região V4V5 se aproxima mais da V9, ao passo que a região V3V4, do domínio *Archaea*. O *primer* desenhado para o domínio *Archaea* está localizado na região V4, sendo uma derivação do *primer* V4, conhecido como universal, qual também é englobado pela região V3V4, possibilitando que muitas OTU's fossem reconhecidas em ambas as regiões, mais do que V4V5 que, apesar de também incorporar a região V4, não compartilha do mesmo *primer*, principalmente no sentido *forward*.

A Figura 4.14 apresenta o gráfico em barras da abundância relativa em nível de filo de cada uma das regiões sequenciadas e a Tabela 4.14, os valores percentuais. A correlação taxonômica foi calculada par a par, a qual é mostrada na Tabela 4.15. Destaca-se o número de sequências que não obtiveram anotação, principalmente no domínio *Archaea*. São dois

prováveis motivos: deficiência dos bancos em espécies neste domínio, por desconhecimento das populações existentes nesse tipo de amostra, ou mesmo a presença de um número grande de quimeras e leituras de baixa qualidade que produzem medidas infladas de diversidade, impossibilitando a obtenção de uma correlação taxonômica. Ambas as situações evidenciam a necessidade de maiores estudos para conhecimento da população que prolifera em reservatórios de petróleo, seja por isolamento de novas espécies por cultivo e sua consequente caracterização, ou por sequenciamento do genoma completo. Dos quarenta e quatro filos que foram reconhecidos, apenas seis aparecem em todas as regiões, quinze não aparecem apenas na V9, nove apenas nos três principais para bactérias, e dez em pelo menos uma das regiões, sendo quer os três últimos listados na tabela 4.14 possuem abundância relativa menor 0,01%.

Tabela 4.13. Matriz de Beta diversidade nas multiplas regiões sequenciadas no *Illumina.*

	16SA			**16SV3V4**			**16SV4**			**16SV4V5**			**18SV9**		
	UP	**BC**	**Jc**	**UP**	**BC**	**Jc**	**UP**	**BC**	**Jc**	**UP**	**BC**	**Jc**	**UP**	**BC**	**Jc**
A	0,00	0,00	0,00	1,11	0,95	0,78	1,12	0,95	0,79	1,06	0,97	0,80	0,92	1,00	0,85
V34	1,11	0,95	0,78	0,00	0,00	0,00	0,12	0,72	0,53	0,21	0,83	0,65	0,49	1,00	0,91
V4	1,12	0,95	0,79	0,12	0,72	0,53	0,00	0,00	0,00	0,22	0,82	0,58	0,51	1,00	0,88
V45	1,06	0,97	0,80	0,21	0,83	0,65	0,22	0,82	0,58	0,00	0,00	0,00	0,46	1,00	0,98
V9	0,92	1,00	0,85	0,49	1,00	0,91	0,51	1,00	0,88	0,46	1,00	0,98	0,00	0,00	0,00

Índices: UP= *Unifrac* Ponderada, que é dependente de taxonomia e da frequência de cada táxon; BC = *Bray Curtis* e Jc = *Jaccard*, independentes de taxonomia, sendo o primeiro uma medida semimétrica e o segundo uma métrica que obedece a desigualdade triangular. Foi omitido na tabela a sequência V9F, pois só tinha apresentou uma única OTU. Em todos os casos o *p value* foi menor que 1.

Com os valores de abundância relativa de cada táxon em nível de gênero, e por meio do teste *pearson* com 999 permutações, a correlação taxonômica foi calculada com seus respectivos intervalos de confiança, cujos valores são apresentados na Tabela 4.15. Este resultado é mais uma alternativa ao cálculo de beta diversidade. É próximo o valor de correlação entre as três regiões 16SV3V4, 16SV4 e 16SV4V5, sendo que o maior é entre 16SV4 e V3V4.

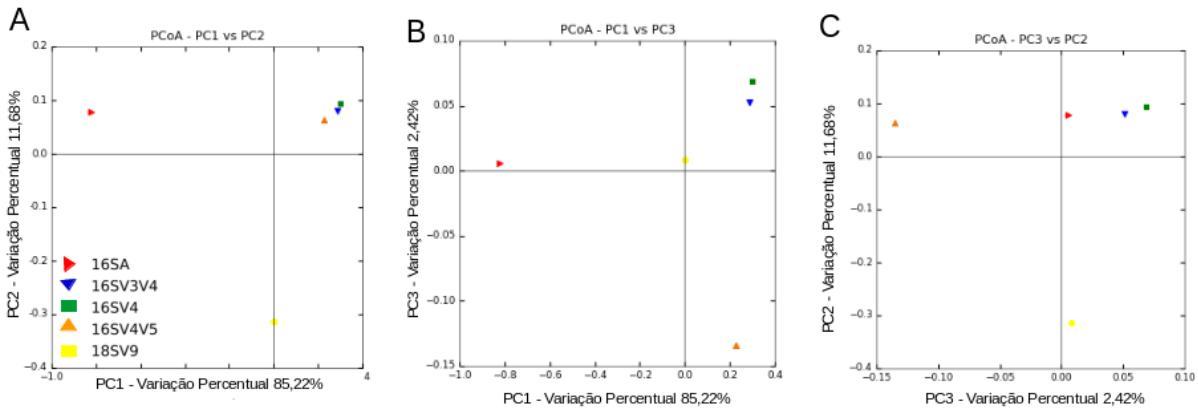

Figura 4.13. Gráfico de Coordenadas Principais obtidas a partir do cálculo de Beta diversidade pelo índice Unifrac Ponderado nas multiplas regiões sequenciadas no *Illumina*. Em cortes de duas direções, os pontos são representados pelo sequenciamento da região 16SA em vermelho, 16SV3V4 em azul, 16SV4V5 em laranja, 16SV4 em verde e 18SV9 em amarelo.

O aumento no comprimento na sequência permite uma maior resolução e definição dos diferentes níveis taxonômicos. Quando comparados em nível de gênero, a região V4V5 foi a que mais obteve anotações em relação às demais, e a V9, a que teve o menor número. Apesar disso, é necessário ressaltar que a parte do gene não pode substituir a análise do gene completo, como já foi demostrado em outros estudos (BARB et al., 2016; CASTELINO et al., 2017; KIM; MORRISON; YU, 2011; POLYMENAKOU et al., 2015; TENG et al., 2018; ZHANG et al., 2018). No entanto, a limitação da tecnologia NGS, em relação ao tamanho do *amplicon* a ser sequenciado, faz com que seja necessário a constante avaliação sobre a região analisada, em especial para amostras de ambientes diferentes (CAI et al., 2013; GUO et al., 2013; ZHANG et al., 2018). Os dados do sequenciamento no *Ion Torrent* para a amostra de água de produção direcionam que a região V4 é a melhor para se estudar, se ponderados a resolução da região, riqueza na amostra e qualidade da leitura. Uma análise mais profunda pelo sequenciamento no *Illumina* indica que não apenas a região central do gene é uma boa opção, mas também que o aumento no comprimento dessa região influenciaria em um incremento na riqueza e resolução. Neste trabalho, define-se resolução, como a capacidade de anotar as OTU's nos níveis taxonômicos mais profundos.

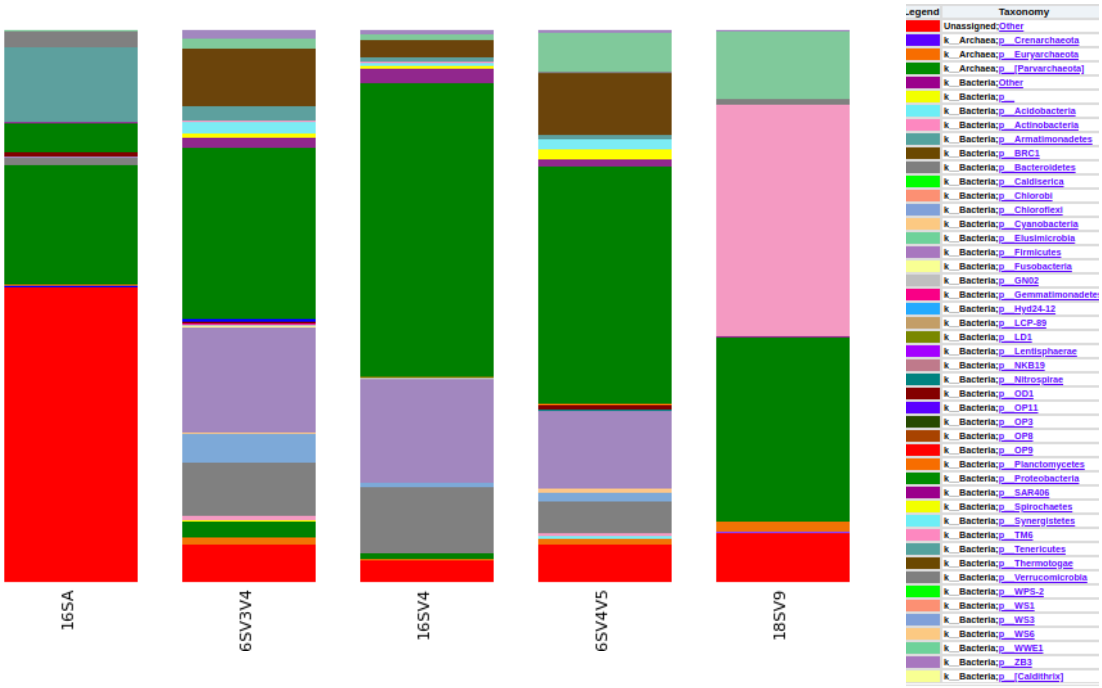

Figura 4.14. Abundância relativa em nível de filo nas multiplas regiões sequenciadas no *Illumina*. Da esquerda para direita, as colunas são representadas pelo sequenciamento da região 16SA, 16SV3V4, 16SV4, 16SV4V5, 18SV9. Em vermelho a abundância relativa do total das OTU's desconhecidas, a legenda com as cores dos filos identificados está mais à direita da imagem.

Tabela 4.14. Resumo taxonômico de cada região sequenciada do gene rRNA 16S pelo *Illumina*.

Domínio	Filo	16SA	16SV3V4	16SV4	16SV4V5	18SV9
Desconhecido		53,50%	6,73%	3,79%	6,74%	8,83%
Bacteria	*Proteobacteria*	5,07%	30,84%	52,92%	43,05%	33,33%
Bacteria	*SAR406*	0,17%	2,01%	2,62%	1,42%	0,36%
Bacteria	*WWE1*	0,27%	1,77%	0,88%	6,93%	12,07%
Bacteria	*ZB3*	0,04%	1,47%	0,76%	0,58%	0,36%
Bacteria	*Verrucomicrobia*	2,79%	0,11%	0,15%	0,22%	1,08%
Bacteria	*TM6*	0,01%	0,22%	0,14%	0,00%	41,98%
Bacteria	*Planctomycetes*	0,01%	0,03%	0,08%	0,06%	1,62%
Bacteria	*Lentisphaerae*	0,00%	0,01%	0,01%	0,04%	0,36%
Bacteria	*Firmicutes*	0,32%	19,06%	18,63%	13,89%	0,00%
Bacteria	*Bacteroidetes*	1,21%	9,56%	11,80%	5,62%	0,00%
Bacteria	*Thermotogae*	0,02%	10,20%	3,11%	11,13%	0,00%

Domínio	**Filo**	**16SA**	**16SV3V4**	**16SV4**	**16SV4V5**	**18SV9**
Archaea	*Parvarchaeota*	21,52%	2,91%	0,95%	0,12%	0,00%
Bacteria	*Chloroflexi*	0,02%	5,02%	0,83%	1,49%	0,00%
Bacteria	*Tenericutes*	13,55%	2,81%	0,81%	0,90%	0,00%
Bacteria	*Spirochaetes*	0,05%	0,68%	0,70%	1,65%	0,00%
Bacteria	*Synergistetes*	0,04%	2,05%	0,52%	1,83%	0,00%
Archaea	*Euryarchaeota*	0,21%	1,38%	0,36%	0,94%	0,00%
Bacteria	*Actinobacteria*	0,00%	0,79%	0,08%	0,42%	0,00%
Bacteria	*Caldithrix*	0,01%	0,05%	0,06%	0,03%	0,00%
Bacteria	*Acidobacteria*	0,00%	0,17%	0,04%	0,50%	0,00%
Bacteria	*OD1*	0,79%	0,36%	0,04%	0,98%	0,00%
Bacteria	*Other*	0,00%	0,04%	0,03%	0,11%	0,00%
Archaea	*Crenarchaeota*	0,27%	0,00%	0,01%	0,00%	0,00%
Bacteria	*GN02*	0,00%	0,42%	0,29%	0,06%	0,00%
Bacteria	*Chlorobi*	0,00%	0,12%	0,09%	0,08%	0,00%
Bacteria	*Cyanobacteria*	0,00%	0,38%	0,07%	0,93%	0,00%
Bacteria	*BRC1*	0,00%	0,01%	0,05%	0,01%	0,00%
Bacteria	*Gemmatimonadetes*	0,00%	0,11%	0,05%	0,05%	0,00%
Bacteria	*OP11*	0,00%	0,47%	0,04%	0,03%	0,00%
Bacteria	*Fusobacteria*	0,00%	0,03%	0,02%	0,01%	0,00%
Bacteria	*Nitrospirae*	0,00%	0,01%	0,02%	0,05%	0,00%
Bacteria	*WS3*	0,00%	0,02%	0,02%	0,02%	0,00%
Bacteria	*Armatimonadetes*	0,00%	0,00%	0,00%	0,01%	0,00%
Bacteria	*Elusimicrobia*	0,00%	0,03%	0,00%	0,01%	0,00%
Bacteria	*Hyd24-12*	0,01%	0,04%	0,00%	0,05%	0,00%
Bacteria	*LCP-89*	0,00%	0,00%	0,01%	0,00%	0,00%
Bacteria	*LD1*	0,00%	0,00%	0,00%	0,01%	0,00%
Bacteria	*NKB19*	0,00%	0,00%	0,00%	0,01%	0,00%
Bacteria	*OP3*	0,01%	0,00%	0,01%	0,01%	0,00%
Bacteria	*OP9*	0,09%	0,00%	0,00%	0,00%	0,00%
Bacteria	*WPS-2*	0,00%	0,02%	0,00%	0,02%	0,00%
Bacteria	*WS1*	0,00%	0,00%	0,00%	0,01%	0,00%
Bacteria	*WS6*	0,00%	0,06%	0,00%	0,00%	0,00%
Bacteria	*Caldiserica*	0,00%	0,00%	0,00%	0,00%	0,00%
Bacteria	*OP8*	0,00%	0,00%	0,00%	0,00%	0,00%

Tabela 4.15. Coeficiente de correlação e intervalo de confiança nas multiplas regiões sequenciadas no *Illumina*.

Região	Coeficiente de Correlação*		
	Média	Intervalo de Confiança	
16SA-16SV3V4	0,32	0,22	0,41
16SA-16SV4	0,13	0,00	0,24
16SA-16SV4V5	0,21	0,11	0,31
16SA-18SV9	0,03	-0,17	0,23
16SV3V4-16SV4	0,78	0,74	0,81
16SV3V4-16SV4V5	0,77	0,73	0,80
16SV3V4-18SV9	0,18	0,08	0,28
16SV4-16SV4V5	0,76	0,71	0,80
16SV4-18SV9	0,28	0,16	0,39
16SV4V5-18SV9	0,21	0,10	0,30

*Cálculos par a par, a partir do resumo taxonômico em nível de gênero. 999 permutações foram feitas com o teste estatístico de Pearson, e *p* value menor que 0,05, com exceção de 18SV9-16SA em que foi aproximadamente 0,8.

4.4. Predição Funcional

A análise de predição funcional foi realizada no programa PICRUSt, uma extensão do QIIME1, que utiliza os dados da anotação taxonômica cruzando-os com o banco de dados KEGG, o qual é uma base de dados que correlaciona vias metabólicas com genes e microrganismos. Para isso, é necessário que o arquivo inserido obedeça alguns critérios. A anotação taxonômica tem que ser realizada pelo método de referência fechada, com 97% de identidade e no banco de dados *Greengenes v08/2013*. Nas seções anteriores são discutidas as diferenças entre os métodos de montagem, banco de dados, região do rRNA 16S e plataforma de sequenciamento, que serviram de base para definir as escolhas tomadas na presente seção. Além de discutir os dados gerados no programa, os resultados serão correlacionados com trabalhos diretamente relacionados à corrosão e Microbiologia do Petróleo.

A plataforma *Illumina*, por possuir a capacidade de concatenar as sequências com sentidos complementares, produz leituras de maior qualidade das e por isso foi utilizada. Foi escolhida a região V4 do gene rRNA 16S pela relação entre resolução, riqueza e pela

distribuição entre abundância relativa das OTU's identificadas. O primeiro arquivo de saída mostra o perfil geral de como os genes são distribuídos no metagenoma de acordo com a classificação no KEGG. Para obtenção desses valores, o programa correlaciona a abundância da OTU com o número de cópias do gene no genoma e o organiza de acordo com a classificação da via metabólica no qual o gene participa. Esse resultado é útil para obter informações preliminares a respeito da comunidade estudada. A Figura 4.15 mostra o gráfico em pizza da classificação dos níveis mais altos, no qual cerca de 44% das contribuições estão relacionadas ao Metabolismo, outras 17% ao Processo de Informação Genética, e ainda há 14% no processo de Informação Ambiental. Sendo do Metabolismo, 19% dedicados à síntese de aminoácidos, 19% aos carboidratos e 15% à produção de energia. O KEGG possui 329 subcategorias, dentre as quais 35 não pontuaram, 50 ficaram abaixo de 1.000, 40 entre 1000 e 10.000, 73 entre 10.000 e 100.000, 109 de 100.000 a 1.000.000 e 21 de 1.000.000 a 5.000.000.

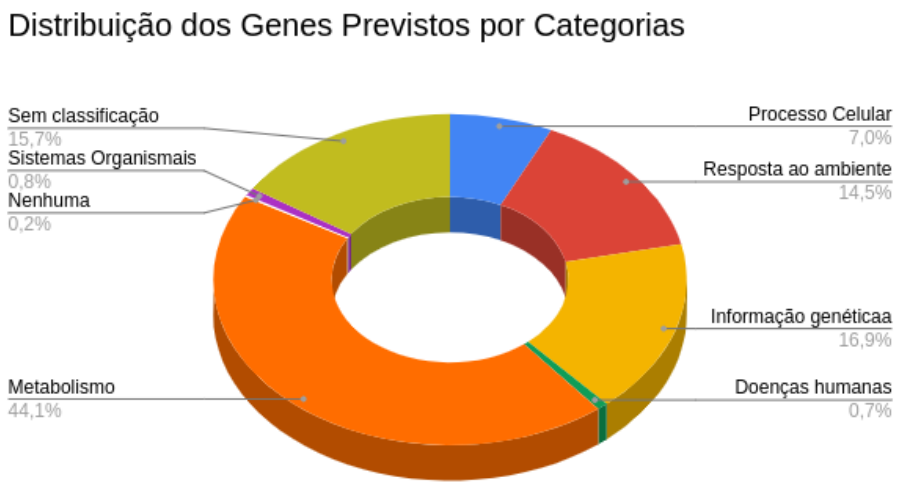

Figura 4.15. Gráfico em pizza da classificação gênica do KEGG.

A maior categoria é o Metabolismo, que envolve treze subclassificações, onde destacam-se a geração de energia e degradação de xenobióticos, algumas das quais estão presentes na Tabela 4.16. O Processamento de Informações Genéticas envolve processos de replicação, transcrição e tradução, inerentes a todos os seres vivos. O Processamento de

Informação Ambiental da via está relacionado a receptores de membrana e moléculas de sinalização.

Após a contagem dos genes por categorias, foi realizada a contagem individual dos mesmos, que foram filtrados de acordo com as funções acima assinaladas. Buscaram-se genes com o perfil marcador, ou seja, que fossem característicos de vias de interesse. O primeiro grupo de interesse são os procariotos redutores de sulfato. É possível observar na Tabela 4.16 a contagem de todos os genes relacionados ao metabolismo de sulfato, que está relacionado às vias assimilatória e dissimilatória de redução de sulfato, e no caminho contrário a oxidação do enxofre (SOX). A diferença entre as vias de redução é a liberação de sulfeto para o ambiente (OGUNSEITAN, 2014, cap. 8; OLLIVIER; MAGOT, 2005, cap. 3). Ambas as vias são descritas na Revisão Bibliográfica (Capítulo 2). A Figura 4.16 mostra a contagem das vias pertencentes ao metabolismo do enxofre, enquanto a identificação individual dos genesestá presente na Tabela 4.17. O segundo grupo, também relacionado à produção de energia, é o dos organismos produtores de metano, ou metanogênicos, encontrado dentro da subclasse Metabolismo de Metano, que envolve as seguintes vias: produção de metano a partir de óxido de carbono ($CO_2 \rightarrow CH_4$), de metanol ($CH_3OH \rightarrow CH_4$) ou de acetato ($C_2H_3O_2 \rightarrow CH_4$). Há também vias que consomem o metano, tendo o formaldeído como produto intermediário (OGUNSEITAN, 2014, cap. 7; OLLIVIER; MAGOT, 2005, cap. 4).

A Tabela 4.17 contém os genes relacionados à produção de metano, metabolismo do enxofre, nitrogênio e produção de acetato, e a Figura 4.16 o gráfico da contagem de suas vias. O terceiro grupo, relacionado ao metabolismo do nitrogênio, inclui os redutores de nitrato. E assim, como nos demais grupos, a subclasse correspondente a este grupo não possui uma única via, mas seis: fixação de nitrogênio, redução assimilatória de nitrato, redução dissimilatória, desnitrificação, nitrificação e oxidação anaeróbica de amônio (*anammox*), sendo que as duas últimas são vias de oxidação do nitrogênio (OGUNSEITAN, 2014, cap. 7; OLLIVIER; MAGOT, 2005, cap. 5). A Tabela 4.17 contém a contagem dos principais genes relacionados à redução de nitrato e a Figura 4.16 a contagem dos genes relacionados à vias pertencentes ao metabolismo do nitrato.

Tendo os dados das contagens dos genes e suas vias de atuação, foi possível prever quais os organismos deveriam conter esses genes, e assim correlacionar o perfil do microbioma da amostra em questão com os dados disponíveis na literatura sobre

Microbiologia no Petróleo. Aqui, vale destacar o local em que a amostra foi retirada e as características intrínsecas, tanto da plataforma quando do campo em que se encontra, pois todos esses dados influenciam grandemente a comunidade microbiana. A Tabela 4.18 mostra os valores de beta diversidade médio, que foram calculados entre as populações retiradas a partir de cada gene classificando dentro da subclasse pertencente ao metabolismo energético, isto é: metabolismo de enxofre, nitrogênio e metano. O cálculo foi realizado da seguinte forma: primeiro entre as populações que contém os genes pertencentes a mesma via (por exemplo, dentro da via dissimilatória de redução do sulfeto tem os genes *aprA*, *aprB*, *drsA* e *drsB*, para cada um foi extraído do arquivo todas as OTU's que os possuem, gerando quatro novos arquivos com uma listagem de OTU's, esses arquivos foram comparados e os valores médios de diversidade estão na tabela); depois foram das vias dentro do mesmo metabolismo (exemplo, dentro do metabolismo do enxofre há três vias: redução assimilatória, dissimilatória e oxidação, as populações foram classificadas dentro das vias a qual fazem parte o gene que elas possuem, e o conjunto desses arquivos serviu de comparação) e por último do metabolismo como um todo em relação a todo o microbioma (isto é, continuando no exemplo do metabolismo do enxofre, o conjunto, com os arquivos cujas populações as quais são classificados neste metabolismo, foi comparada a todo microbioma).

Tabela 4.16. Tabela com os dez caminhos mais presentes

KEGG_Nível- L1	KEGG_Nível- L2	KEGG_Nível- L3	16SV4
Metabolismo	Metabolismo Energético	Fixação de Carbono	1101417
Metabolismo	Metabolismo Energético	Metabolismo Metano	1183254
Metabolismo	Metabolismo Energético	Metabolismo Nitrogênio	796268
Metabolismo	Metabolismo Energético	Oxidação Fosforilativa	1035786
Metabolismo	Metabolismo Energético	Fotossíntese	373051
Metabolismo	Metabolismo Energético	Metabolismo Enxofre	234595
Metabolismo	Família de Enzimas	Citocromo P450	3
Metabolismo	Biodegradação de Xenobióticos	Degradação Amino benzoato	106733

KEGG_Nível- L1	KEGG_Nível- L2	KEGG_Nível- L3	16SV4
Metabolismo	Biodegradação de Xenobióticos	Degradação Benzoato	153165
Metabolismo	Biodegradação de Xenobióticos	Degradação Bisfenol	36023
Metabolismo	Biodegradação de Xenobióticos	Degradação de Cloro alcanos	192140
Metabolismo	Biodegradação de Xenobióticos	Degradação Etilbenzeno	30724
Metabolismo	Biodegradação de Xenobióticos	Degradação Fluobenzoato	1471
Metabolismo	Biodegradação de Xenobióticos	Metabolismo de Xenobióticos pelo citocromo P450	45776
Metabolismo	Biodegradação de Xenobióticos	Degradação Naftaleno	116369
Metabolismo	Biodegradação de Xenobióticos	Degradação Nitrotolueno	191291
Metabolismo	Biodegradação de Xenobióticos	Degradação Poliaromáticos	41590
Metabolismo	Biodegradação de Xenobióticos	Degradação Estireno	9542
Metabolismo	Biodegradação de Xenobióticos	Degradação Tolueno	96093
Metabolismo	Biodegradação de Xenobióticos	Degradação Xileno	13126

Tabela 4.17. Tabela contagem dos principais grupos relacionados ao metabolismo energético e suas respectivas vias

KEGG	Gene	Metabolismo	Via	Contagem
K00625	*[E.2.3.1.8.]*	Metano	Acetato – Metano	8392
K13788	*pta*	Metano	Acetato – Metano	15405
K00399	*mcrA*	Metano	Acetato/CO_2/Metanol - Metano	162
K00401	*mcrB*	Metano	Acetato/CO_2/Metanol - Metano	162
K00402	*mcrG*	Metano	Acetato/CO_2/Metanol - Metano	162
K00577	*mtrA*	Metano	Acetato/CO_2/Metanol - Metano	282
K00578	*mtrB*	Metano	Acetato/CO_2/Metanol - Metano	134
K00579	*mtrC*	Metano	Acetato/CO_2/Metanol - Metano	134
K00580	*mtrD*	Metano	Acetato/CO_2/Metanol - Metano	134

KEGG	Gene	Metabolismo	Via	Contagem
K00581	*mtrE*	Metano	Acetato/CO_2/Metanol - Metano	134
K00582	*mtrF*	Metano	Acetato/CO_2/Metanol - Metano	126
K00583	*mtrG*	Metano	Acetato/CO_2/Metanol - Metano	119
K00584	*mtrH*	Metano	Acetato/CO_2/Metanol - Metano	264
K03421	*mcrC*	Metano	Acetato/CO_2/Metanol - Metano	134
K03422	*mcrD*	Metano	Acetato/CO_2/Metanol - Metano	162
K00955	*cysNC*	Enxofre	Redução Dissimilatória/Assimilatória Enxofre	6956
K00956	*cysN*	Enxofre	Redução Dissimilatória/Assimilatória Enxofre	1622
K00957	*cysD*	Enxofre	Redução Dissimilatória/Assimilatória Enxofre	5437
K00200	*fwdA*	Metano	CO_2 - Metano	210
K00201	*fwdB*	Metano	CO_2 - Metano	335
K00202	*fwdC*	Metano	CO_2 - Metano	328
K00203	*fwdD*	Metano	CO_2 - Metano	273
K00204	*fwdH*	Metano	CO_2 - Metano	9
K00205	*fwdF*	Metano	CO_2 - Metano	257
K00319	*mtd*	Metano	CO_2 - Metano	143
K00320	*mer*	Metano	CO_2 - Metano	830
K00672	*ftr*	Metano	CO_2 - Metano	166
K11260	*fwdG*	Metano	CO_2 - Metano	150
K11261	*fwdE*	Metano	CO_2 - Metano	39116
K13942	*hmd*	Metano	CO_2 - Metano	9
K00368	*nirK*	Nitrogênio	Denitrificação	39
K00376	*nosZ*	Nitrogênio	Denitrificação	408
K02305	*norC*	Nitrogênio	Denitrificação	18
K04561	*norB*	Nitrogênio	Denitrificação	93
K00531	*anfG*	Nitrogênio	Fixação Nitrogênio	1
K02586	*nifD*	Nitrogênio	Fixação Nitrogênio	22122
K02588	*nifH*	Nitrogênio	Fixação Nitrogênio	24267
K02591	*nifK*	Nitrogênio	Fixação Nitrogênio	22527
K09903	*pyrH*	Nucleotide	Metabolismo Pirimidina	29402
K00440	*frhA*	Metano	Metano	24
K00441	*frhB*	Metano	Metano	8257
K00442	*frhD*	Metano	Metano	3278
K03388	*hdrA2*	Metano	Metano	95486
K03389	*hdrB2*	Metano	Metano	44809
K03390	*hdrC2*	Metano	Metano	31665

KEGG	Gene	Metabolismo	Via	Contagem
K08264	*hdrD*	Metano	Metano	2002
K08265	*hdrE*	Metano	Metano	110
K14126	*mvhA*	Metano	Metano	4917
K14127	*mvhD*	Metano	Metano	3792
K14128	*mvhG*	Metano	Metano	4504
K14080	*mtaA*	Metano	Metanol - Metano	221
K14081	*mtaC*	Metano	Metanol - Metano	263
K10535	*hao*	Nitrogênio	Nitrificação	27171
K10944	*PmoA amoA*	Nitrogênio	Nitrificação	0
K10945	*PmoB amoB*	Nitrogênio	Nitrificação	0
K10946	*PmoC amoC*	Nitrogênio	Nitrificação	0
K00360	*nasB*	Nitrogênio	Redução Assimilatória Nitrato	3
K00366	*nirA*	Nitrogênio	Redução Assimilatória Nitrato	2260
K00367	*narB*	Nitrogênio	Redução Assimilatória Nitrato	4
K00372	*nasA*	Nitrogênio	Redução Assimilatória Nitrato	543
K00380	*cysJ*	Enxofre	Redução Assimilatória Sulfato	1779
K00381	*cysI*	Enxofre	Redução Assimilatória Sulfato	103
K00390	*cysH*	Enxofre	Redução Assimilatória Sulfato	19934
K00955	*cysC*	Enxofre	Redução Assimilatória Sulfato	6956
K00370	*narG*	Nitrogênio	Redução Dissimilatória/Denitrificação	1528
K00371	*narH*	Nitrogênio	Redução Dissimilatória/Denitrificação	1837
K00374	*narI*	Nitrogênio	Redução Dissimilatória/Denitrificação	13075
K02567	*napA*	Nitrogênio	Redução Dissimilatória/Denitrificação	3109
K02568	*napB*	Nitrogênio	Redução Dissimilatória/Denitrificação	1511
K00362	*nirB*	Nitrogênio	Redução Dissimilatória Nitrato	5831
K00363	*nirD*	Nitrogênio	Redução Dissimilatória Nitrato	1583
K03385	*nrfA*	Nitrogênio	Redução Dissimilatória Nitrato	21004
K00392	*sir*	Enxofre	Redução Dissimilatória Sulfato	319
K00394	*aprA*	Enxofre	Redução Dissimilatória Sulfato	19682
K00395	*aprB*	Enxofre	Redução Dissimilatória Sulfato	17953
K00958	*sat*	Enxofre	Redução Dissimilatória Sulfato	26631
K11180	*dsrA*	Enxofre	Redução Dissimilatória Sulfato	17960
K11181	*dsrB*	Enxofre	Redução Dissimilatória Sulfato	17960
K00197	*cdhE*	Acetato	Via AcetilCoA/ Acetato - Metano	4664
K00925	*ackA*	Acetato	Via AcetilCoA/ Acetato-Metano	32959
K01895	*ACSS*	Acetato	Via AcetilCoA/ Acetato - Metano	56706

KEGG	Gene	Metabolismo	Via	Contagem
K14138	*acsB*	Acetato	Via AcetilCoA/ Acetato - Metano	11805

O Campo de Marlim, local em que a plataforma P37, de onde a amostra foi retirada, possui na composição de sua rocha reservatório arenitos com 30% de feldspatos, com o potássio como cátion mais presente (MILANI; ARAÚJO, 1990). Esse tipo de rocha possui muita friabilidade, ou seja, esfarela com muita facilidade. O fluxo de produção de um reservatório, isto é, a constante troca dos fluidos de injeção usados para incremento da produção petrolífera, faz com que o fluido produzido contenha muitos particulados de rocha. O óleo precisa seguir as recomendações mínimas antes de seguir para a refinaria. Para isso, o óleo é pré-processado (MILANI; ARAÚJO, 1990; SILVA et al., 2007; THOMAS, 2004). O aumento no tempo de residência nos tanques de separação faz com que proliferem organismos formadores de biofilme, havendo abundante matéria orgânica e receptores finais de elétrons, em condições propícias (DUNCAN et al., 2017; ENNING; GARRELFS, 2014; JOHNSON et al., 2017; TANJI et al., 2014; VIGNERON et al., 2016). Esses biofilmes, dentro dos tanques, tendem a se localizar na interface água/óleo onde são capazes de capturar pequenas moléculas polares (como propionato, butirato e ácidos orgânicos de até sete carbonos), sais, hidrogênio (JEANTHON et al., 2005; MAGOT; OLLIVIER; PATEL, 2000; VARJANI; GNANSOUNOU, 2017). A origem desses microrganismos é geralmente conhecida, sendo provenientes do material que é injetado, água do mar em sua maioria, ou do próprio reservatório (OLLIVIER; MAGOT, 2005, cap. 2).

Devido a constantes incidentes na plataforma durante a época de extração da amostra, o principal grupo investigado foi o de redutoras de sulfato. De fato, o gênero mais abundante em toda a amostra, independente da metodologia empregada, foi o *Dessulfovibrio*, que nas condições para predição funcional representa 18,36 % na região 16SV4 (97% ID, referência fechada anotado no *Greengenes*), que é um redutor de sulfato. Outros organismos redutores de sulfato podem ser encontrados, por exemplo, dentro dos filos, *Proteobacteria*, 55,8400%, *Firmicutes, 17,9889%, Nitrospira, 0,0039% e Euryarchaeota, 0,3073%* (YOUSSEF; ELSHAHED; MCINERNEY, 2009). É grande a diversidade desses organismos e eles podem ser encontrados em uma ampla variedade de ambientes (VIGNERON et al., 2018). Tanto entre organismos redutores de sulfato termófilos (provocam acidificação) quanto mesófilos

(causadores de corrosão), há a preferência por acetato, mas são capazes, dependendo do gênero, de metabolizar outros substratos. O *Desulfovibrio* (18,36%) metaboliza lactato, hidrogênio e piruvato, já o *Desulfobacter* (5,033%), apenas o acetato. (LI et al., 2017; MAGOT; OLLIVIER; PATEL, 2000; VAN HAMME; SINGH; WARD, 2003). O *Desulfacinum* (0,0023%) utiliza acetato, butirato, palmitato e alcoóis (OLLIVIER; MAGOT, 2005). Das 771 OTU's que possuem predição funcional disponível, 11% delas possuem os genes da via dissimilatória de redução de sulfeto. Os genes avaliados, *drsA*, *drsB*, *aprA e aprB*, representaram populações altamente similares, conforme apresenta a Tabela 4.18.

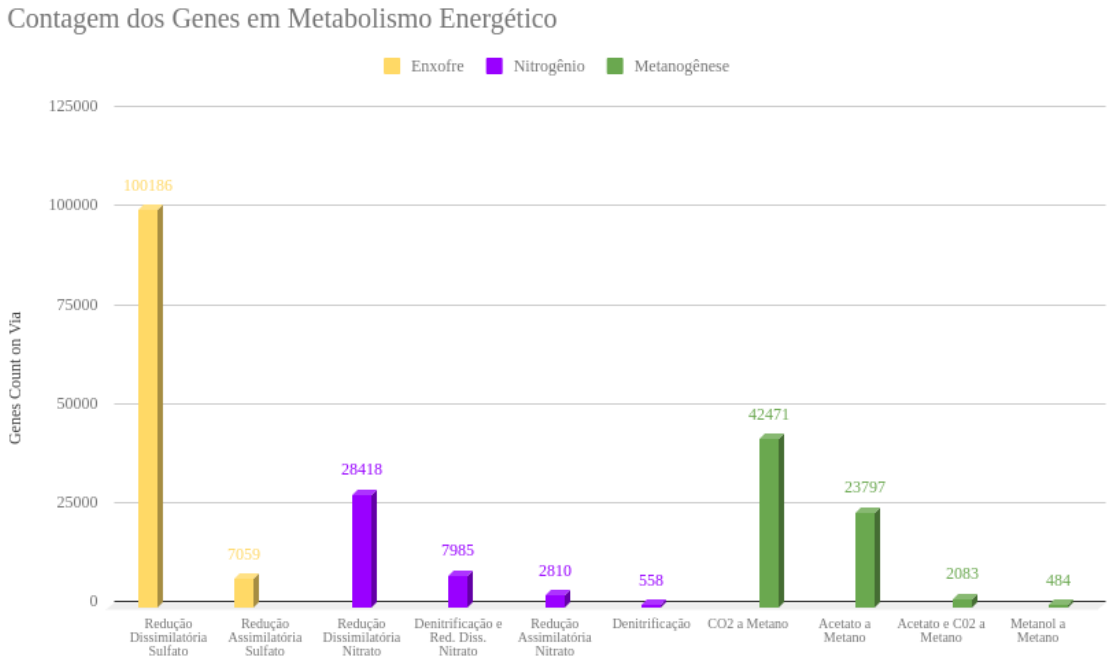

Figura 4.16. Gráfico em Barras da contagem dos genes separados por via metabólica que participam. A presença dos genes é correlacionada à abundância das OTU's que os contém. Os mesmos foram agrupados de acordo com a via a qual pertencem manualmente.

O metano pode ser produzido pela transformação, em anaerobiose, a partir de acetato, CO_2 ou metanol. A contagem dos genes relacionados a cada uma dessas vias foi de 123.870, 44.284, 1.240, respectivamente. Juntos eles representam 14% da contagem total dos genes envolvidos no metabolismo do metano, conforme mostrado na Tabela 4.17 e Figura 4.16.

Dentre essas vias, a maior é equivalente à produção de metano a partir do acetato. Esta via pode ser dividida em duas partes: a primeira em que o acetato é produzido e a segunda onde ele é consumido. Não são os mesmos organismos que participam do início ao fim do processo, por este motivo, na Tabela 4.17, houve uma separação na descrição dos genes. Os metanogênicos que utilizam acetato incluem indivíduos no domínio *Archaea* e não são caracterizados com frequência (YOUSSEF; ELSHAHED; MCINERNEY, 2009). São organismos mesofílicos, sensíveis a mudanças bruscas no ambiente, como pH e disponibilidade de oxigênio. Dos diferentes grupos relatados que vivem em reservatório de petróleo, os do domínio Archaea provavelmente são os mais sensíveis a mudanças externas (MAGOT; OLLIVIER; PATEL, 2000). Desenvolvem-se no estágio final da biodegradação do petróleo, principalmente quando há escassez dos demais receptores finais de elétrons (LI et al., 2017; OLLIVIER; MAGOT, 2005, cap. 4). Os metilotróficos são halotolerantes, como *Methanosarcina*, que além do metanol é capaz de transformar o acetato a metano. Dentro de *Euyarchaeota*, cinco ordens são reconhecidas como metanogênicas: *Methanomicrobiales, Methanobacteriales, Methanosarcinales, Methanococcales e Methanopirales* (LI et al., 2017; YOUSSEF; ELSHAHED; MCINERNEY, 2009).

Tabela 4.18. Dissimilaridade Média Calculada pelos índices Unifrac pondeado e Jaccard.

	Dissimilaridade Média					
	Em relação às demais vias		**Dentro da própria via**		**Em relação ao microbioma**	
	Unifrac	**Jaccard**	**Unifrac**	**Jaccard**	**Unifrac**	**Jaccard**
Enxofre	0,50 ± 0,24	0,96 ± 0,07	0,34 ± 0,31	0,66 ± 0,30	0,42 ± 0,29	0,82 ± 0,26
Nitrogênio	-	0,94 ± 0,05	-	0,85 ± 0,13	-	0,92 ± 0,08
Metano	0,56 ± 0,48	0,66 ± 0,30	0,30 ± 0,43	0,43 ± 0,32	0,48 ± 0,48	0,59 ± 0,33

Com origem em tratamento de águas residuais, a técnica de injeção de água com altos teores de nitrato ganhou um destaque especial em artigos publicados na última década. Ela permite a inibição do crescimento de redutoras de sulfato, o que por consequência diminui a corrosão nas plataformas e a liberação de sulfeto de hidrogênio no ambiente (BØDTKER et al., 2009; OLLIVIER; MAGOT, 2005, cap. 8; VARJANI; GNANSOUNOU, 2017; VOORDOUW, 2011). Há estudos que indicam a mudança no perfil global do microbioma

após esse tratamento. O interesse é grande devido aos altos custos com controle da corrosão e com equipamentos de proteção individual (BØDTKER et al., 2009; GITTEL et al., 2009; OKORO et al., 2014; SUNDE; TORSVIK; SUNDE, 2005; THRASHER; VANCE, 2016). Há quatro propostas para o mecanismo de inibição da produção do sulfeto pela adição de nitrato, todas passam pelo bioestímulo de redutoras de nitrato: a primeira é competição por nutrientes que ambos os organismos metabolizam; a segunda é a produção de intermediário tóxico durante a redução do nitrato, o nitrito, que ocorre tanto na via de redução dissimilatória quanto na desnitrificação; a terceira é a utilização do sulfeto como doador final de elétrons na redução do nitrato; a quarta e última é o desvio da rota metabólica nos microrganismos que empregam múltiplos receptores finais de elétrons, pois o nitrogênio é mais eletronegativo do que o enxofre (OLLIVIER; MAGOT, 2005, cap. 8). A dissimilaridade entre as populações que possuem os genes de redução de nitrato e sulfato nas vias dissimilatórias, é igual a 0.92 e 0.31 no índice Jaccard e UNIFRAC ponderado, respectivamente. Isso indica que poucas espécies reduzem os dois ânions. No entanto, a queda na dissimilaridade no índice UNIFRAC ponderado indica que, apesar de serem poucas as espécies, a abundância desses indivíduos é alta. As redutoras de nitrato são mesofílicas. O gênero *Marinobacter* (0,0047%) reduz nitrato a nitrito utilizando como substratos n-hexadecano, succinato, acetato e citrato (OLLIVIER; MAGOT, 2005, cap. 5). O nitrito é produto da reação de redução do nitrato catalisada pelas redutases Nar e Nap (LIEBENSTEINER et al., 2014). Destaca-se nesse grupo o gênero *Desulfobacter* (5,033%) por possuir ambos os genes.

Apesar de estudos anteriores apontarem para a redução de H_2S no ambiente (LYSNES et al., 2009; SUNDE; TORSVIK; SUNDE, 2005), ainda é preciso investigar mais a respeito da redução da corrosão, pois a presença de organismos que empregam múltiplos receptores de elétrons podem promover a corrosão induzida por microrganismos (MIC). Por exemplo, redutores do Ferro III ao Ferro II, através da oxidação completa do acetato, são *Pelobacter* (11,28%), *Desulfuromonas* (0,18%) e organismos da ordem *Geobacteraceae* (0,06%). O metal reduzido é depositado diretamente nas paredes dos tubos que são constituídos de aço-carbono. Vale destacar que os três grupos destacados também são redutores de sulfato (ZAPATA-PEÑASCO et al., 2013) De fato, a respiração por ferro foi retratada uma das primeiras formas de metabolismo microbiano, sendo que esse grupo metabólico inclui os

domínios procarióticos associados à degradação de diferentes fontes de carbono (WEBER; ACHENBACH; COATES, 2006)(LOVLEY, 1997).

Liebensteiner et al. (2014) publicaram um trabalho no qual discutem as possíveis reações redox que ocorrem em uma amostra de óleo, concluindo que todas essas reações devem ocupar um papel na MIC (LIEBENSTEINER et al., 2014). Com base nesse trabalho, é possível destacar alguns microrganismos presentes na presente amostra e que possuem os gêneros relacionados, como *Marinobacter* (0,0047% 16SV4) e *Arcobacter* (0,02% 16SV4) que podem ser redutores de (per) clorados em ambientes com alta salinidade. Outros organismos também relatados possuem uma enzima com capacidade redutora de (per) clorato, como algumas espécies de *Pseudomonas* (0,01% 16SV4), onde os doadores de elétrons são ácidos orgânicos, hidrocarbonetos aromáticos e álcoois.

O ferro oxidado por procariontes atua como um acelerador no processo de corrosão. Organismos autotróficos são capazes que oxidar o ferro, o que inclui produtores de metano, como os pertencentes ao filo *Euryarchaeota* (0,30% 16SV4) e ao gênero *Crenarchaeota* (0,004% 16SV4). A previsão funcional aponta que 8,10% dos genes previstos envolvidos no metabolismo energético são relacionados à fixação de carbono. Juntas, bactérias que degradam nafta (8,79% da previsão funcional da Biodegradação de Xenobióticos) promovem a formação de complexos orgânicos com o metal. Devido à sua alta eletronegatividade, o oxigênio sequestra o ferro do complexo e forma óxidos, o que foi observado por Pillay et al em 2014, em um teste de corrosão com uma amostra de água produzida após estimulação com nitrato. Eles verificaram que a corrosão foi acelerada e detectaram aumento de peso em óxidos de ferro (PILLAY; LIN, 2014)

Quanto menor a abundância de um organismo mais difícil é de determinar se este é um caso de falso positivo ou se está presente em quantidades ínfimas. A retratação desses casos deve ser feita com ressalva (EDGAR, 2017; LIEBENSTEINER et al., 2014; LU et al., 2015; WEISS et al., 2017). Isto posto, uma família comumente encontrada em ambientes marinhos, e que poderia ter sido introduzida naturalmente aos dutos da plataforma, a *Rhodobacteraceae* (0.0016%), ganha destaque pela capacidade de oxidar iodeto e prosperar na presença de oxigênio dissolvido, que não reduzem nitrato ou sulfato, mas contribuem no entupimento dos dutos (SUGAI et al., 2013). Este é apenas um dos casos de microrganismo relacionado à

corrosão e que não possui múltiplos receptores finais de elétrons, como os mencionados acima.

A contribuição metagenômica dá uma noção de como os genes são distribuídos dentro da categoria KEGG. A maioria dos genes está relacionada à geração de energia, que envolve a síntese de ATP e a degradação da matéria orgânica, havendo sempre os processos relacionados à duplicação celular, necessários para todos os organismos. O valor da contribuição está relacionado à presença do gene em cada OTU e à abundância da OTU que contém esse gene. (KOO et al., 2017b; MUKHERJEE et al., 2017). Vale ressaltar que os valores encontrados no PICRUSt são apenas uma aproximação e podem não representar a atividade real que o microbioma oferece. Às vezes, um organismo em menor porcentagem pode apresentar uma atividade que determinará um certo perfil funcional ao microbioma. Mas as previsões feitas por este programa foram robustas e pode servir como um guia para estudos mais direcionados.

Nesse sentido, é possível prever não apenas a geração de energia e consumo dos aceitantes finais, mas também a degradação da matéria orgânica, considerando que o processo ocorre em simbiose com vários organismos. A Figura 4.17 mostra a contagem de subclasses presentes no metabolismo de xenobióticos e a Figura 4.18 a contagem de alguns dos seus genes. É possível observar que à medida que a complexidade do substrato aumenta, o número de genes relacionados à sua degradação diminui e, consequentemente, o número de organismos capazes de metabolizá-lo. A refração de uma molécula pode inibir a ação de enzimas, bem como a passagem pelos canais de membrana, fazendo com que, durante a evolução, poucos organismos possam se especializar em sua metalização. Mesmo com a presença de genes putativos que podem conferir capacidade de degradação, a ação enzimática pode não ocorrer devido à produção de metabólitos tóxicos durante o processo, ou à presença ou ausência de qualquer sinalizador importante para ativar a expressão gênica do *operon* (VAN HAMME; SINGH; WARD, 2003).

Microrganismos que possuem genes *alk* têm o potencial de degradar alcanos aerobicamente, oxidando-os a aldeído e ácido, com um álcool como produto intermediário. A degradação anaeróbica pode ser por carboxilação ou por formação de succinato. Os produtos finais podem entrar na via da β-oxidação para ser degradado. Compostos com até dois anéis aromáticos podem sofrer clivagem e, dependendo do grupo de substituição, são transformados e incorporados em outras vias metabólicas (FUENTES et al., 2014). Tem sido relatado que

genes como *nah*, *ndo*, *pah* e *dox* estão relacionados à biodegradação de hidrocarbonetos aromáticos policíclicos (HAP). As enzimas geralmente têm baixa especificidade para esses compostos, o que lhes permite trabalhar com uma variedade de substratos pertencentes a essa classe. Além disso, também contribuem para diferentes compostos na presença de genes metabólicos de múltiplas vias que podem ser encontrados em um único indivíduo (HAMME; SINGH; WARD, 2003; WIDDEL; RABUS, 2001). Além disso, é muito importante conhecer a capacidade do microbioma de degradar o óleo, uma vez que essa característica tem uma relação profunda com a qualidade do óleo (OLLIVIER; MAGOT, 2005, cap. 6). Por outro lado, não só é importante saber do aspecto negativo, mas também do potencial de usar essas informações para futuras aplicações (YOUSSEF; ELSHAHED; MCINERNEY, 2009).

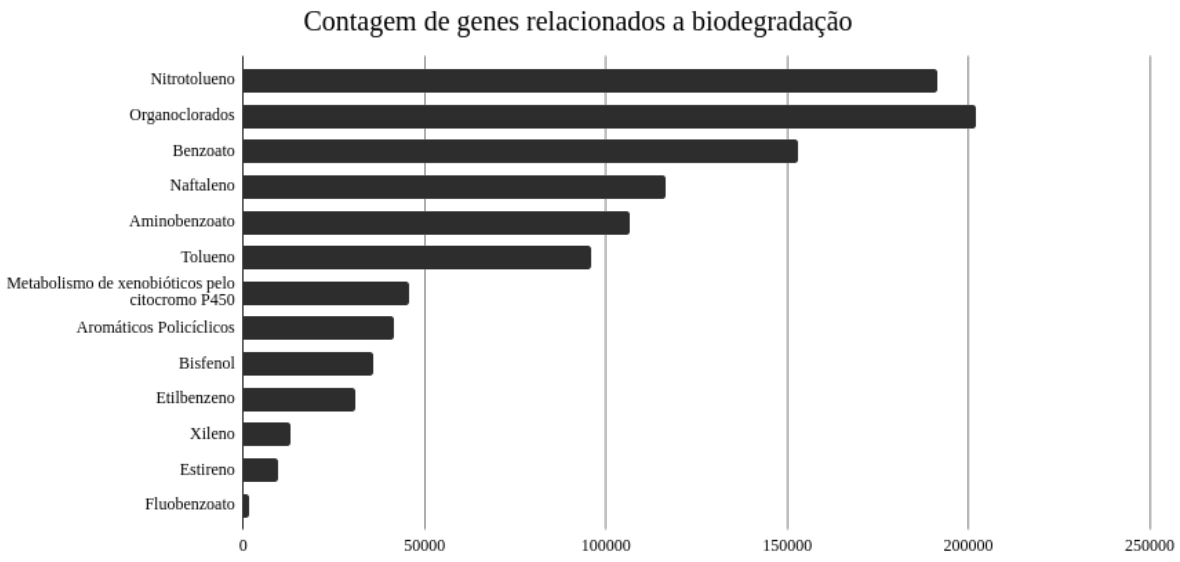

Figura 4.17 Quantificação das subclasses presentes no metabolismo de xenobióticos. Contagem de genes de acordo com a classificação no KEGG.

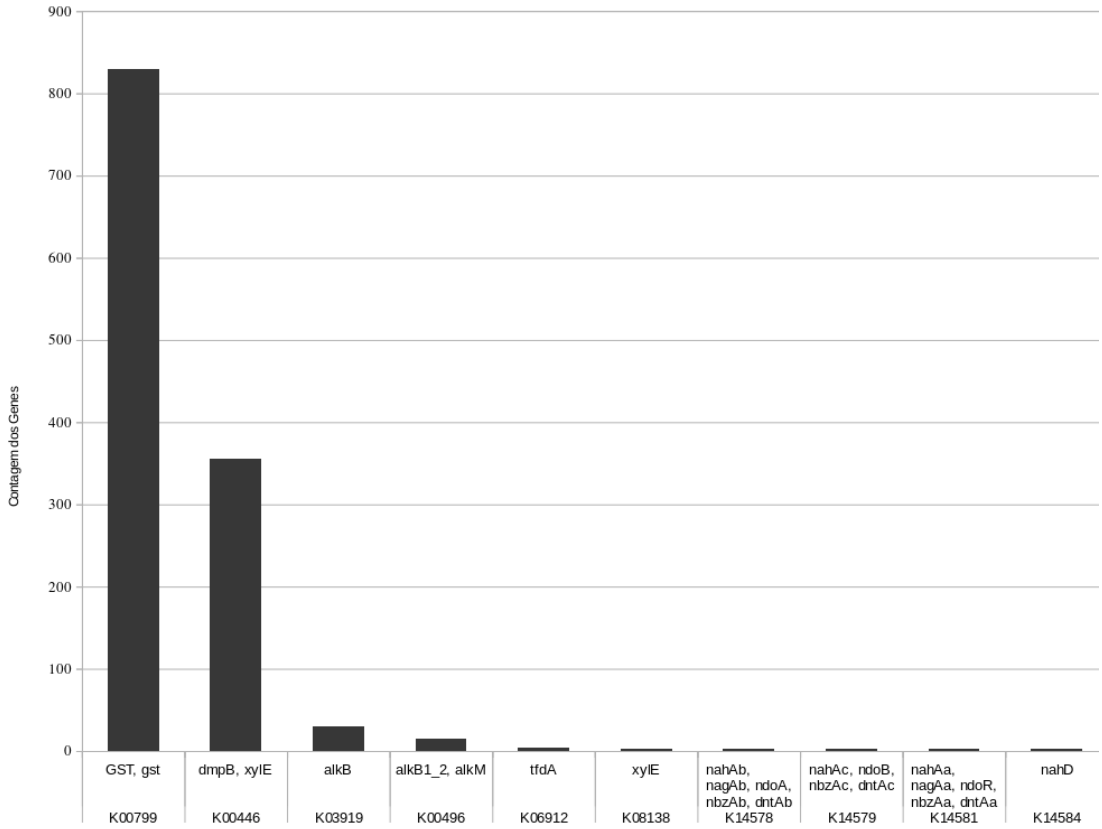

Figura 4.18 Quantificação de genes associados ao metabolismo xenobióticos Contagem individual da presença de alguns genes relacionados a biodegradação.

5. CONCLUSÕES E CONSIDERAÇÕES FINAIS

- O reservatório de petróleo caracterizado nesse projeto possui uma comunidade microbiana complexa, contendo diversas populações dependentes umas das outras, através de relações simbiontes ou competitivas;
- As populações caracterizadas apresentam potencial para a utilização do óleo como fonte de carbono, o que resulta em maior degradabilidade do óleo, o que, por sua vez, está relacionado à perda da qualidade para o refino. Geralmente, o óleo biodegradado possui maior acidez, densidade, evidenciado por maiores teores de asfaltenos e resinas;
- O processo de exploração e produção do óleo modifica o equilíbrio existente na dinâmica da comunidade, seja pela introdução de grupos exógenos, ou adição de alguns nutrientes que antes eram limitados, como oxigênio ou nitrato, ou mesmo a proporção entre o óleo e os demais fluidos presentes. Nesse contexto, o metabolismo anaeróbico, que era predominante, passa a ser substituído por metabolismos aeróbicos ou aerotolerantes, o que aumenta a velocidade de degradação do óleo;
- O fluido produzido precisa ser processado ainda na plataforma, separando a água, o óleo e o gás. Dependendo das condições, como estabilidade da emulsão, presença de sólidos em suspensão, entre outros, o processo pode ter mais etapas. Essas condições favorecem a proliferação de microrganismos, que no interior dos dutos e tanques favorecem a formação de biofilmes complexos, dificultando a ação de agentes antimicrobianos e antibiofilmes;
- A identificação dessas comunidades é fundamental não apenas para entender sua dinâmica e compreender todas modificações às quais são submetidas, mas, sobretudo, para empregar esse conhecimento de forma a contornar problemas operacionais associados. Além disso, abre caminho para a busca de novas vias metabólicas que possam ser utilizados tanto no incremento da produção, processo conhecido como MEOR (do inglês. *Microbial enhanced oil recovery*), como em biorremediação, em casos de acidentes de vazamento. Para isso, diferentes metodologias podem ser aplicadas, identificando assim o

microbioma. Uma das mais utilizadas é o sequenciamento do gene 16S rRNA por NGS. A técnica é relativamente barata, reprodutível e de rápida resposta. No entanto, as ferramentas analíticas utilizadas no tratamento dos dados estão em constante evolução. Os conceitos são novos e as frequentes atualizações dos programas requerem muita atenção do operador;

- Neste trabalho de dissertação, foi realizada a identificação do microbioma presente na plataforma, o que também envolveu tomadas de decisões operacionais que contribuíram para o resultado encontrado. São elas relacionadas abaixo:

  - ***Programas*** - Diversos programas com muitas funcionalidades são desenvolvidos todos os anos, cada um buscando cobrir as falhas dos anteriores. Por serem os pioneiros, tanto o QIIME como o Mothur e o USEARCH são referências na análise de sequenciamento massivo. O USEARCH é licenciado, exigindo pagamento de direitos ao desenvolvedor, enquanto o QIIME e o Mothur são gratuitos, independente da quantidade de dados a ser analisada. Além disso, existem ferramentas, como o BMP, que procura a melhor forma de tratar os dados na utilização de mais de um programa de bioinformática, aliando o que cada um tem de melhor. Ainda há outros programas que não foram citados, mas que são utilizados com frequência, como MS-RAST, DADA2, e ainda a versão mais recente do QIIME2, que mudou consideravelmente. A escolha do programa depende da adaptação do usuário e à sua interface e às operações. Nesse sentido, foi escolhido o programa QIIME1, versão que estava disponível na época em que foram realizadas as análises dos dados gerados tanto pelo *Illumina* como pelo *Ion Torrent*.
  - ***Pré-tratamento dos reads*** - Antes das leituras serem processadas, é necessário que passem por uma filtragem para remoção dos dados não confiáveis, que comprometem a reprodutibilidade e a qualidade

da análise. Nesse sentido, vários passos foram testados, utilizando um ou mais programas até encontrar um que realizasse uma boa limpeza.

- ***Agrupamento das leituras em OTU's*** - Três métodos de montagem foram testados, todos disponíveis no QIIME: *de novo*, Referência Fechada e Referência Aberta. O método *de novo*, por não utilizar nenhum parâmetro para construir a OTU, gera um número muito alto de OTUs e demanda um maior tempo de análise, o que para dados muito complexos representa um problema, pois nem todos os computadores conseguem processá-lo. O método de Referência Fechada é o mais limitado em termos de abrangência, pois agrupa as OTUs com base em um banco de referência. É mais rápido que o primeiro, mas descarta a maior parte das sequências. Como alternativa, os desenvolvedores do QIIME criaram um método chamado de Referência Aberta, cujo funcionamento engloba características dos dois anteriores, propondo-se a ser o intermediário entre os dois outros métodos. De fato, nos parâmetros aqui analisados, como diversidade alfa, número de OTUs e correlação taxonômica, o método faz o que promete. Apesar de ser bastante criticado por outros desenvolvedores contemporâneos ao QIIME, como Patrick Schloss do Mothur e Robert Edgar do USEARCH, o método de referência aberta mostra-se mais eficiente e prático.
- ***Banco de dados*** - Existem três bancos de dados utilizados na anotação taxonômica de sequências do gene rRNA 16S. Desses, dois foram testados: o *Greengenes*, opção padrão do QIIME1 e o SILVA. A diferença entre ambos vai além das atualizações. O primeiro, teve sua última atualização em 2013 e foi construído através de algoritmos que usam como base o banco de depósitos de genes do NCBI. Já o segundo, com atualizações anuais, é anotado manualmente. A diversidade da amostra é a mesma, independente

do banco utilizado. No entanto, cada um possui uma estrutura própria de organização dos táxons, como sua hierarquização, tratamento de espécies ambientais não cultiváveis, número de classificações em branco. Essas diferenças fazem com que o número de OTU's geradas a partir do SILVA seja ligeiramente maior do que no *Greengenes*. Como consequência disso, a correlação taxonômica decai à medida que a classificação é aprofundada. As atualizações periódicas do SILVA fazem com que este seja o banco mais utilizado. No entanto, ele é mais pesado, exigindo mais do computador e fazendo com que as análises demorem. Além disso, algumas ferramentas específicas para o QIIME, como a predição funcional, requerem que o banco de dados seja o *Greengenes* com o método de anotação seja o de referência fechada.

- ***Plataforma de Sequenciamento***—Dentro da segunda geração de sequenciadores, ainda predominam *Illumina* e *Ion Torrent*. plataformas NGS vem atendendo com qualidade às demandas. A *Illumina* produz leituras com menos erros, com maior confiabilidade. No entanto, a PMG atrai pelo menor custo nos reagentes e ainda pela possibilidade de capturar grupos minoritários, o que a *Illumina*, por seu processo de clusterização, acaba por negligenciar. O uso combinado de ambas as plataformas permite uma caracterização mais completa do microbioma.
- ***Regiões Hipervariáveis do gene rRNA 16S***—Cada plataforma cobriu um conjunto distinto de regiões do rRNA 16S. A princípio, pelo *Ion Torrent*, foram sequenciadas as regiões V1V2, V4, V6 e V9. Cada uma possui uma resolução filogenética diferente, sendo que a região V6 foi a que apresentou maior diversidade, enquanto a V9, a menor. A região V4 foi a que apresentou um menor número de leituras desconhecidas. A V4 localiza-se na região central do gene rRNA 16S, que ocupa um papel funcional importante, sendo a

região que entrega a melhor resposta, tanto em diversidade como em número de espécies reconhecidas e anotadas. Além disso, pelo processo de ligação do fragmento aos adaptadores, foram sequenciados os dois sentidos do DNA. Não há uma ferramenta em bioinformática que monte uma sequência consenso, em *reads* do PMG. Foi então necessário trabalhar cada sentido individualmente. Apesar da anotação no banco de dados ser no sentido principal, 5’3’, é possível verificar uma correlação entre eles. Ainda é possível verificar maior homologia nas regiões de *hairpin* como V6R e V4F. Outras regiões foram analizadas na plataforma *Illumina* - V3V4, V4 e V4V5 - variando assim o comprimento das sequências analisadas e sua posição no gene. De fato, o aumento no comprimento gerou maior diversidade. A região V4V5 foi a que apresentou maior riqueza e probabilidade de encontrar novas OTUs, mas, assim como a V6, gerou um percentual de amostras não anotadas mais alto do que a V4. É necessário ter um balanço entre riqueza de uma amostra e sua anotação. Por esses motivos a região V4, dentro das condições analisadas, foi a que melhor proporcionou essas respostas, sendo escolhida para análise da diversidade funcional.

- A análise funcional proporcionou um melhor entendimento da dinâmica da comunidade. Dentro dos grupos comumente estudados em microbiologia do petróleo, como redutores de sulfato, metanogênicos, redutores de nitrato e fermentativos, dominância é de redutoras de sulfato, não apenas em quantidade absoluta mas na abundância relativa que apresentam. Muitos desses organismos têm a capacidade de empregar múltiplos receptores finais de elétrons, representando um alto um risco para a geração de MIC. Sobre a capacidade de degradação da matéria orgânica, poucos organismos foram relatados com possuindo genes dentro das vias de degradação de orgânicos mais complexos. Se por um lado o resultado é esperando, tendo em vista que a maior parte dos fermentadores são anaeróbios restritos, por outro, deve-se considerar que a amostra foi retirada de um tanque de separação e que, portanto, já não

é representativa do interior do reservatório. No entanto, deve-se considerar a limitação da técnica, que apresenta as estimativas com base em organismos já anotados. Considerando o grande potencial dos organismos ainda não conhecidos, este estudo aponta apenas as potencialidades encontradas;

- A análise por sequenciamento é bastante recente e, ao mesmo tempo em que democratiza os estudos dos microbiomas, está em constante evolução. Nesse contexto, é necessário que investigações sejam realizadas com as tecnologias mais contemporâneas disponíveis, de forma a alcançar uma melhor elucidação da composição e comportamento metabólico e ecológico do complexo microbioma existente, em qualquer tipo de ambiente. Ainda que parcialmente, nesse projeto foi possível observar diferentes organismos com capacidade de degradação do óleo, potencialidades em bioremediação ou mesmo tratamentos *in situ*, como bioestímulo, usado para incrementar a produção ou mesmo diminuir a corrosão e produção de gases tóxicos como sulfeto de hidrogênio.

REFERENCIAS

ALLALI, I. et al. A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. **BMC Microbiology**, v. 17, n. 1, 2017.

AMARAL-ZETTLER, L. A. et al. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA Genes. **PLoS ONE**, v. 4, n. 7, p. e6372, 27 jul. 2009.

ASSHAUER, K. P. et al. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. **Bioinformatics (Oxford, England)**, v. 31, n. 17, p. 2882–4, 1 set. 2015.

BAKER, B. J.; BANFIELD, J. F. **Microbial communities in acid mine drainageFEMS Microbiology Ecology**, 2003.

BAKER, G. C.; SMITH, J. J.; COWAN, D. A. Review and re-analysis of domain-specific 16S primers. **Journal of Microbiological Methods**, v. 55, n. 3, p. 541–555, 2003.

BALVOČIUTE, M.; HUSON, D. H. SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare? **BMC Genomics**, v. 18, 2017.

BARB, J. J. et al. Development of an Analysis Pipeline Characterizing Multiple Hypervariable Regions of 16S rRNA Using Mock Samples. **PLoS ONE**, v. 11, n. 2, p. 18, 2016.

BARTON, L. L.; FAUQUE, G. D. Chapter 2 Biochemistry, Physiology and Biotechnology of Sulfate-Reducing Bacteria. In: **Advances in Applied Microbiology**. 1. ed. [s.l.] Elsevier Inc., 2009. v. 68p. 41–98.

BASTIN, E. S. The problem of the natural reduction of sulphates. **AAPG Bulletin**, v. 10, n. 12, p. 1270–1299, 1926.

BENÍTEZ-PÁEZ, A.; PORTUNE, K. J.; SANZ, Y. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. **GigaScience**, v. 5, n. 1, p. 4, 2016.

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.

Boletim Mensal da Produção de petroleo e gas natural. Disponível em: <http://www.anp.gov.br/publicacoes/boletins-anp/2395-boletim-mensal-da-producao-de-petroleo-e-gas-natural%0A>. Acesso em: 11 abr. 2018.

CAI, L. et al. Biased Diversity Metrics Revealed by Bacterial 16S Pyrotags Derived from Different Primer Sets. **PLoS ONE**, v. 8, n. 1, p. 1–11, 2013.

CAPORASO, J. G. et al. PyNAST: A flexible tool for aligning sequences to a template alignment. **Bioinformatics**, 2010.

CAPORASO, J. G.; DORRESTEIN, P. C. **Clustering sequences into OTUs using q2-vsearch**. Disponível em: <https://docs.qiime2.org/2018.8/tutorials/otu-clustering/>. Acesso em: 24 out. 2018.

CHAO, A. Non parametric estimation of the number of classes in a population. **Scandinavian Journal of Statistics**, v. 11, n. October 2014, p. 265–270, 1984.

CHEN, W. et al. A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs. **PLoS ONE**, v. 8, n. 8, p. e70837, 2013.

CHHIBBER-GOEL, J. et al. Linkages between oral commensal bacteria and atherosclerotic plaques in coronary artery disease patients. **NPJ Biofilms Microbiomes**, v. 7, p. 1–13, 2017.

CLAESSON, M. J. et al. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. **Nucleic Acids Research**, v. 38, n. 22, 2010.

CRUAUD, P. et al. Influence of DNA extraction method, 16S rRNA targeted hypervariable regions, and sample origin on microbial diversity detected by 454 pyrosequencing in marine chemosynthetic ecosystems. **Applied and Environmental Microbiology**, v. 80, n. 15, p. 4626–4639, 2014.

D'AMORE, R. et al. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. **BMC Genomics**, v. 17, n. 1, 2016.

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.

DEVANGA RAGUPATHI, N. K. et al. Accurate differentiation of Escherichia coli and Shigella serogroups: challenges and strategies. **New Microbes and New Infections**, v. 21, p. 58–62, 2018.

DJEMIEL, C.; GREC, S.; HAWKINS, S. Characterization of bacterial and fungal community dynamics by high-throughput sequencing (HTS) metabarcoding during flax dew-retting. **Frontiers in Microbiology**, v. 8, n. OCT, p. 1–16, 2017.

EDGAR, R. Taxonomy annotation and guide tree errors in 16S rRNA databases. **PeerJ**, v. 6, p. e5030, 2018.

EDGAR, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. **Nucleic Acids Research**, v. 32, n. 5, p. 1792–1797, 2004.

EDGAR, R. C. **USEARCH Manual**. Disponível em: <http://www.drive5.com/usearch/manual/>. Acesso em: 28 jan. 2018.

EDGAR, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. **Nature Methods**, v. 10, n. 10, 2013.

EDGAR, R. C. Accuracy of microbial community diversity estimated by closed- and open-reference OTUs. **PeerJ**, 2017.

ELOE-FADROSH, E. A. et al. Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. **Nature Microbiology**, v. 1, n. 4, p. 15032, 2016.

ESCOBAR-PÁRAMO, P. et al. The evolutionary history of Shigella and enteroinvasive Escherichia coli revised. **Journal of Molecular Evolution**, v. 57, n. 2, p. 140–148, 2003.

ESTHER M. GABOR; ALKEMA, W. B. L.; JANSSEN, D. B. Quantifying the accessibility of the metagenome by random expression cloning techniques. **Environ. Microbiol.**, v. 6, n. 9, p. 879–886, 2004.

FIETTO, J. L. R.; MACIEL, T. E. F. Sequenciando genomas. In: **Ciências genômicas: fundamentos e aplicações**. Ribeirão Preto: Sociedade Brasileira de Genética, 2015. p. 27–64.

FOUHY, F. et al. 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. **BMC Microbiology**, v. 16, n. 1, p. 123, 24 dez. 2016.

FULLER, D. C.; SURUDA, A J. Occupationally related hydrogen sulfide deaths in the United States from 1984 to 1994. **Journal of occupational and environmental medicine / American College of Occupational and Environmental Medicine**, v. 42, n. 9, p. 939–942, 2000.

GABARDO, I. T. **Caracterização Química e Toxicologia da água produzida descartada em plataformas de óleo e gás na costa brasileira e seu comportamento dispersivo no mar**. 2007. Tese (Doutorado em Química) Centro de Ciências Exatas e da Terra - Universidade Federal do Rio Grande do Norte.

GAO, P. et al. Microbial diversity and abundance in the Xinjiang Luliang long-term water-flooding petroleum reservoir. **MicrobiologyOpen**, v. 4, n. 2, p. 332–342, 2015.

GEWIN, V. Microbes en masse: The sequencing machine. **Nature**, v. 487, n. 7406, p. 156–158, 2012.

GILBERT, J. A. et al. The taxonomic and functional diversity of microbes at a temperate coastal site: A “multi-omic” study of seasonal and diel temporal variation. **PLoS ONE**, v. 5, n. 11, 2010.

GILBERT, J. A. et al. Defining seasonal marine microbial community dynamics. **The ISME Journal**, v. 6, n. 2, p. 298–308, 2012.

GILBERT, J. A.; JANSSON, J. K.; KNIGHT, R. The Earth Microbiome project: successes and aspirations. **BMC Biology**, v. 12, n. 1, p. 69, 2014.

GLENN, T. C. Field guide to next-generation DNA sequencers. **Molecular Ecology Resources**, 2011.

GOHL, D. M. et al. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. **Nature Biotechnology**, 2016.

GOOD, I. J. THE POPULATION FREQUENCIES OF SPECIES AND THE ESTIMATION OF POPULATION PARAMETERS. **Biometrika**, v. 40, n. 3–4, p. 237–264, 1953.

GUO, F. et al. Taxonomic Precision of Different Hypervariable Regions of 16S rRNA Gene and Annotation Methods for Functional Bacterial Groups in Biological Wastewater Treatment. **PLoS ONE**, v. 8, n. 10, 2013.

HAMME, J. D. VAN; SINGH, A.; WARD, O. P. Recent Advances in Petroleum Microbiology Recent Advances in Petroleum Microbiology. **Society**, v. 67, n. 4, 2003.

HE, Y. et al. Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. **Microbiome**, v. 3, n. 1, p. 34, 2015.

HEATHER, J. M.; CHAIN, B. The sequence of sequencers: The history of sequencing DNA. **Genomics**, v. 107, n. 1, p. 1–8, 2016.

HIERGEIST, A.; REISCHL, U. Multicenter quality assessment of 16S ribosomal DNA-sequencing for microbiome analyses reveals high inter-center variability. **International Journal of Medical Microbiology**, v. 306, n. 5, p. 334–342, 2016.

HUSE, S. M. et al. Accuracy and quality of massively parallel DNA pyrosequencing. v. 8, n. 7, 2007.

HUSE, S. M. et al. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. **PLoS Genetics**, 2008.

INC., L. T. **Prepare Amplicon Libraries without Fragmentation Using the Ion Plus Fragment Library Kit**, 2014. Disponível em: <http://ioncommunity.lifetechnologies.com/servlet/JiveServlet/downloadBody/3322-102-6-22243/MAN0006846_RevA_UB_3March2014.pdf>

JOHN H. MCDONALD. **Handbook of Biological Statistics**. (3 Ed) Sparky House Baltimore, Maryland: 2014.

JOHN MCMURRY. **Química Orgânica**. (9 ed) Cengage Learning, São Paulo: 2016.

JU, F.; ZHANG, T. **16S rRNA gene high-throughput sequencing data mining of microbial diversity and interactionsApplied Microbiology and Biotechnology**, 2015.

KANEHISA, M. et al. KEGG for representation and analysis of molecular networks involving diseases and drugs. **Nucleic Acids Research**, v. 38, n. SUPPL.1, p. 355–360, 2009.

KASTER, K. M. et al. Characterisation of culture-independent and -dependent microbial communities in a high-temperature offshore chalk petroleum reservoir. **Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology**, v. 96, n. 4, p. 423–429, 2009.

KIM, M.; MORRISON, M.; YU, Z. Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. **Journal of Microbiological Methods**, v. 84, n. 1, p. 81–87, 2011.

KLINDWORTH, A. et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. **Nucleic Acids Research**, v. 41, n. 1, p. 1–11, 2013.

KNIGHT, R. et al. Best practices for analysing microbiomes. **Nature Reviews Microbiology**, 2018.

KOEPPEL, A. F.; WU, M. Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. **Nucleic Acids Research**, v. 41, n. 10, p. 5175–5188, 2013.

KOO, H. et al. Comparison of two bioinformatics tools used to characterize the microbial diversity and predictive functional attributes of microbial mats from Lake Obersee, Antarctica. **Journal of Microbiological Methods**, 2017a.

KOO, H. et al. Microbial communities and their predicted metabolic functions in growth laminae of a unique large conical mat from Lake Untersee, East Antarctica. **Frontiers in Microbiology**, v. 8, n. AUG, p. 1–15, 2017b.

KORENBLUM, E. et al. Bacterial diversity in water injection systems of Brazilian offshore oil platforms. **Applied Microbiology and Biotechnology**, v. 85, n. 3, p. 791–800, 2010.

KOZICH, J. J. et al. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. **Applied and Environmental Microbiology**, v. 79, n. 17, p. 5112–5120, 2013.

KOZLOV, A. M. et al. Phylogeny-aware identification and correction of taxonomically mislabeled sequences. **Nucleic Acids Research**, v. 44, n. 11, p. 5022–5033, 2016.

KRAFT, B.; STROUS, M.; TEGETMEYER, H. E. Microbial nitrate respiration - Genes, enzymes and environmental distribution. **Journal of Biotechnology**, v. 155, n. 1, p. 104–117, 2011.

KUCZYNSKI, J. et al. Using QIIME to analyze 16S rRNA gene sequences from Microbial Communities. [s.d.].

KUMARASWAMY, R. et al. Molecular- and cultivation-based analyses of microbial communities in oil field water and in microcosms amended with nitrate to control H2S production. **Applied Microbiology and Biotechnology**, 2011.

LANGILLE, M. G. I. et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. **Nature Biotechnology**, v. 31, n. 9, p. 814–821, 25 ago. 2013.

LI, C. Y. et al. Diversity and abundance of ammonia-oxidizing bacteria (AOB) revealed by PCR amplification of amoA gene in a polyacrylamide transportation system of an oilfield. **International Biodeterioration and Biodegradation**, v. 115, p. 110–118, 1 nov. 2016a.

LI, X. X. et al. Dominance of Desulfotignum in sulfate-reducing community in high sulfate production-water of high temperature and corrosive petroleum reservoirs. **International Biodeterioration and Biodegradation**, v. 114, p. 45–56, 2016b.

LI, X. X. et al. Microbiota and their affiliation with physiochemical characteristics of different subsurface petroleum reservoirs. **International Biodeterioration and Biodegradation**, v. 120, p. 170–185, 2017.

LIEBENSTEINER, M. G. et al. Microbial redox processes in deep subsurface environments and the potential application of (per)chlorate in oil reservoirs. **Frontiers in Microbiology**, v. 5, n. SEP, p. 1–13, 2014.

LIU, L. et al. **Comparison of next-generation sequencing systemsJournal of Biomedicine and Biotechnology**, 2012. Disponível em: <http://www.hindawi.com/journals/bmri/2012/251364/>. Acesso em: 8 ago. 2017

LIU, Y.; WHITMAN, W. B. Metabolic, phylogenetic, and ecological diversity of the methanogenic archaea. **Annals of the New York Academy of Sciences**, v. 1125, p. 171–189, 2008.

LLADSER, M. E.; GOUET, R.; REEDER, J. Extrapolation of urn models via Poissonization: Accurate measurements of the microbial unknown. **PLoS ONE**, v. 6, n. 6, 2011.

LOGARES, R. et al. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. **Environmental Microbiology**, v. 16, n. 9, 2014.

LOZUPONE, C. A. et al. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. **Applied and environmental microbiology**, v. 73, n. 5, p. 1576–85, 1 mar. 2007.

LOZUPONE, C.; KNIGHT, R. UniFrac : a New Phylogenetic Method for Comparing Microbial Communities UniFrac : a New Phylogenetic Method for Comparing Microbial Communities. **Applied and environmental microbiology**, v. 71, n. 12, p. 8228–8235, 2005.

LU, H.; GIORDANO, F.; NING, Z. Oxford Nanopore MinION Sequencing and Genome Assemb. **Nuclear Physics, Section A**, n. 14, p. 265–279, 2016.

LUKE K URSELL, JESSICA L METCALF, LAURA WEGENER PARFREY, AND R. K. Definig the Human Microbiome. **NIH Manuscripts**, v. 70, n. Suppl 1, p. 1–12, 2013.

LUNDBERG, D. S. et al. Defining the core Arabidopsis thaliana root microbiome. **Nature**, v. 488, n. 7409, p. 86–90, 2012.

MA, J.; PRINCE, A.; AAGAARD, K. M. Use of whole genome shotgun metagenomics: A practical guide for the microbiome-minded physician scientist. **Seminars in Reproductive Medicine**, v. 32, n. 1, p. 5–13, 2014.

MAGOT, M.; OLLIVIER, B.; PATEL, B. K. C. Microbiology of petroleum reservoirs **Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology**, 2000.

MANDE, S. S.; MOHAMMED, M. H.; GHOSH, T. S. Classification of metagenomic sequences: Methods and challenges. **Briefings in Bioinformatics**, v. 13, n. 6, p. 669–681, 2012.

MARKOWITZ, V. M. et al. IMG: The integrated microbial genomes database and comparative analysis system. **Nucleic Acids Research**, v. 40, n. D1, p. 115–122, 2012.

MARTIN, T. C. et al. Conducting metagenomic studies in microbiology and clinical research. **Applied Microbiology and Biotechnology**, v. 102, n. 20, p. 8629–8646, 2018.

MARTINEZ-PORCHAS, M. et al. How conserved are the conserved 16S-rRNA regions? **PeerJ**, v. 5, p. e3036, 2017.

MELO, A. S. What do we win ‘confounding’ species richness and evenness in a diversity index? **Biota Neotropica**, v. 8, n. 3, p. 21–27, 2008.

MESLÉ, M.; DROMART, G.; OGER, P. Microbial methanogenesis in subsurface oil and coal. **Research in Microbiology**, v. 164, n. 9, p. 959–972, 2013.

METZKER, M. L. Sequencing technologies the next generation. **Nature Reviews Genetics**, v. 11, n. 1, p. 31–46, 2010.

MIDDLEDITCH, B. S. Ecological effects of produced water effluents from offshore oil and gas production platforms. **Ocean Management**, 1984.

MUKHERJEE, A. et al. Bioinformatic Approaches Including Predictive Metagenomic Profiling Reveal Characteristics of Bacterial Response to Petroleum Hydrocarbon Contamination in Diverse Environments. **Scientific Reports**, v. 7, n. 1, 2017.

MYSARA, M. et al. From reads to operational taxonomic units: An ensemble processing pipeline for MiSeq amplicon sequencing data **GigaScience**, 2017.

NEFF, J.; LEE, K.; DEBLOIS, E. M. Produced Water : Overview of Composition, Fates, and Effects. **Produced Water**. In: Lee K., Neff J. (eds) Springer**,** Nova York, NY. 2011.

NILAKANTA, H. et al. A review of software for analyzing molecular Sequences. **BMC Research Notes**, v. 7, n. 830, p. 9, 2014.

OBREGON-TITO, A. J. et al. Subsistence strategies in traditional societies distinguish gut microbiomes. **Nature Communications**, v. 6, p. 1–9, 2015.

OGUNSEITAN, O. **Microbial Diversiy**. (2 ed) John Whiley and Sons. Hoboken, New Jersey: 2008.

OLLIVIER, B.; MAGOT, M. **Petroleum Microbiology**. ASM Press. Washington – DC: 2005.

ORPHAN, V. J. et al. Culture-dependent and culture-independent characterization of microbial assemblages associated with high-temperature petroleum reservoirs. **Applied and Environmental Microbiology**, v. 66, n. 2, p. 700–711, 1 fev. 2000.

OSMAN, M. A. et al. 16S rRNA gene sequencing for deciphering the colorectal cancer gut microbiome: Current protocols and workflows. **Frontiers in Microbiology**, v. 9, n. APR, 2018.

OULAS, A. et al. Metagenomics: Tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. **Bioinformatics and Biology Insights**, v. 9, 2015.

PATRICK D. SCHLOSS, SARAH L. WESTCOTT, THOMAS RYABIN, JUSTINE R. HALL, MARTIN HARTMANN, EMILY B. HOLLISTER, RYAN A. LESNIEWSKI, BRIAN B. OAKLEY, DONOVAN H. PARKS, COURTNEY J. ROBINSON, JASON W. SAHL, BLAZ STRES, GERHARD G. THALLINGER, DAVID J. VAN HORN, AND C. F. W. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. **APPLIED ANDENVIRONMENTALMICROBIOLOGY**, v. 77, n. 23, p. 7537–7541, 2009.

PAULSON, J. N. et al. Differential abundance analysis for microbial marker-gene surveys. **Nature Methods**, v. 10, n. 12, p. 1200–1202, 2013.

PETROPEDIA. **Understanding Reservoir Drive Mechanisms**. Disponível em: <https://www.petropedia.com/understanding-reservoir-drive-mechanisms/2/9856>. Acesso em: 12 jan. 2019.

PIRES, A. **DESENVOLVIMENTO DE SISTEMAS ASSOCIATIVOS DE MICELAS ALONGADAS PARA RECUPERAÇÃO AVANÇADA DE PETRÓLEO**. [s.l.] Universidade Federal do Rio de Janeiro, 2015.

POLLOCK, J. et al. The madness of microbiome: Attempting to find consensus “best practice” for 16S microbiome studies. **Applied and Environmental Microbiology**, v. 84, n. 7, 2018.

PRAKASH, T.; TAYLOR, T. D. Functional assignment of metagenomic data: Challenges and applications. **Briefings in Bioinformatics**, v. 13, n. 6, p. 711–727, 2012.

PYLRO, V. S. et al. Brazilian Microbiome Project: Revealing the Unexplored Microbial Diversity-Challenges and Prospects. **Microbial Ecology**, v. 67, n. 2, p. 237–241, 2014a.

PYLRO, V. S. et al. Data analysis for 16S microbial profiling from different benchtop sequencing platforms. **Journal of Microbiological Methods**, v. 107, p. 30–37, dez. 2014b.

PYLRO, V. S. et al. BMPOS: a Flexible and User-Friendly Tool Sets for Microbiome Studies. **Microbial Ecology**, v. 72, n. 2, p. 443–447, 2016.

QUAST, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. **Psychiatria Danubina**, v. 41, n. 1, p. 590–596, 2013.

REUTER, J. A.; SPACEK, D. V.; SNYDER, M. P. High-Throughput Sequencing Technologies. **Molecular Cell**, v. 58, n. 4, p. 586–597, 2015.

RHOADS, A.; AU, K. F. PacBio Sequencing and Its Applications. **Genomics, Proteomics and Bioinformatics**, v. 13, n. 5, p. 278–289, 2015.

RIDEOUT, J. R. et al. Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. **PeerJ**, v. 2, p. e545, 21 ago. 2014.

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

RÖLING, W. F. M.; HEAD, I. M.; LARTER, S. R. The microbiology of hydrocarbon degradation in subsurface petroleum reservoirs: Perspectives and prospects. **Research in Microbiology**, v. 154, n. 5, p. 321–328, 2003.

SALIPANTE, S. J. et al. Performance comparison of Illumina and Ion Torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. **Applied and Environmental Microbiology**, v. 80, n. 24, p. 7583–7591, 2014.

SAMPSON, T. R. et al. Gut Microbiota Regulate Motor Deficits and Neuroinflammation in a Model of Parkinson's Disease. **Cell**, v. 167, n. 6, p. 1469–1480.e12, 2016.

SANSCHAGRIN, S.; YERGEAU, E. Next-generation Sequencing of 16S Ribosomal RNA Gene Amplicons. **Journal of Visualized Experiments**, n. 90, p. 3–8, 2014.

SCHLOSS, P. D. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. **PLoS Computational Biology**, v. 6, n. 7, p. 19, 8 jul. 2010.

SCHLOSS, P. D. Secondary structure improves OTU assignments of 16S rRNA gene sequences. **ISME Journal**, v. 7, n. 3, p. 457–460, 2013a.

SCHLOSS, P. D. Secondary structure improves OTU assignments of 16S rRNA gene sequences. **ISME Journal**, v. 7, n. 3, p. 457–460, mar. 2013b.

SCHLOSS, P. D.; WESTCOTT, S. L. Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. **Applied and Environmental Microbiology**, v. 77, n. 10, p. 3219–3226, maio 2011.

SCHMIDT, T. S. B.; MATIAS RODRIGUES, J. F.; VON MERING, C. Ecological Consistency of SSU rRNA-Based Operational Taxonomic Units at a Global Scale. **PLoS Computational Biology**, v. 10, n. 4, 2014.

SIEGWALD, L. et al. Assessment of common and emerging bioinformatics pipelines for targeted metagenomics. **PLoS ONE**, 2017.

SIERRA-GARCÍA, I. N. et al. New hydrocarbon degradation pathways in the microbial metagenome from brazilian petroleum reservoirs. **PLoS ONE**, v. 9, n. 2, 2014.

SIMS, D. et al. **Sequencing depth and coverage: Key considerations in genomic analysesNature Reviews Genetics**, 2014.

SINGER, E. et al. High-resolution phylogenetic microbial community profiling. **The ISME Journal**, v. 10, n. 8, p. 2020–2032, 9 ago. 2016.

SINHA, R. et al. The microbiome quality control project: Baseline study design and future directions. **Genome Biology**, 2015.

SOH, J. et al. Phoenix 2: A locally installable large-scale 16S rRNA gene sequence analysis pipeline with Web interface. **Journal of Biotechnology**, 2013.

STREIT, W. R.; SCHMITZ, R. A. **Metagenomics - The key to the uncultured microbesCurrent Opinion in Microbiology**, 2004.

SZKLO, A. S.; ULLER, V. C.; BONFÁ, M. H. P. **Fundamentos do Refino do Petróleo**. 3. ed. rio de janeiro: Interciência, 2012.

TARDY-JACQUENOD, C. et al. Characterization of sulfate-reducing bacteria isolated from oil-field waters. **Canadian Journal of Microbiology**, v. 42, n. 3, p. 259–266, mar. 1996.

THOMAS, J. E. **Fundamentos da Engenharia do Petroleo**. 1 ed ed. rio de janeiro: Interciência, 2001.

TIKHONOV, M.; LEACH, R. W.; WINGREEN, N. S. Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. 2013.

TREMBLAY, J. et al. Primer and platform effects on 16S rRNA tag sequencing. **Frontiers in Microbiology**, v. 6, 4 ago. 2015.

TRINGE, S. G. et al. Comparative metagenomics of microbial communities. **Science**, 2005.

VARANI, A. M. et al. **Ciências genômicas : fundamentos e aplicaçõe**. 1. ed. Ribeirão Preto: Sociedade Brasileira de Genética, 2015.

VÁZQUEZ-BAEZA, Y. et al. EMPeror : a tool for visualizing high - throughput microbial community data. p. 2–5, 2011.

VĚTROVSKÝ, T.; BALDRIAN, P. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. **PLoS ONE**, v. 8, n. 2, 2013.

WANG, Q. et al. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. **Applied and Environmental Microbiology**, v. 73, n. 16, p. 5261–5267, 2007.

WANG, Y.; QIAN, P. Y. Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. **PLoS ONE**, v. 4, n. 10, 2009.

WARD, D. M.; WELLER, R.; BATESON, M. M. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. **Nature**, v. 345, n. 6270, p. 63–65, 3 maio 1990.

WESTCOTT, S. L.; SCHLOSS, P. D. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. **PeerJ**, v. 3, p. e1487, 8 dez. 2015.

WOO, P. C. Y. et al. **Then and now: Use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratoriesClinical Microbiology and Infection**, 2008.

XIAO, C. et al. Bacterial diversity and community structure of supragingival plaques in adults with dental health or caries revealed by 16S pyrosequencing. **Frontiers in Microbiology**, v. 7, n. JUL, p. 1–15, 2016a.

XIAO, M. et al. Analysis of bacterial diversity in two oil blocks from two low-permeability reservoirs with high salinities. **Scientific Reports**, v. 6, n. December 2015, p. 1–10, 2016b.

YANG, B.; WANG, Y.; QIAN, P.-Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. **BMC Bioinformatics**, v. 17, n. 1, p. 135, 2016.

YARZA, P. et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. **Nature Reviews Microbiology**, v. 12, n. 9, 2014.

YILMAZ, P. et al. The SILVA and “all-species Living Tree Project (LTP)” taxonomic frameworks. **Nucleic Acids Research**, v. 42, n. D1, 2014.

YOUSSEF, N.; ELSHAHED, M. S.; MCINERNEY, M. J. Chapter 6 Microbial Processes in Oil Fields. Culprits, Problems, and Opportunities. In: **Advances in Applied Microbiology**. 1. ed. [s.l.] Elesvier Inc., 2009. v. 66p. 141–251.

YU, Z. et al. Evaluations of different hypervariable regions of archaeal 16S rRNA genes in profiling of methanogens by Archaea-specific PCR and denaturing gradient gel electrophoresis. **Applied and Environmental Microbiology**, v. 74, n. 3, p. 889–893, 2008.

ZHANG, J. et al. Evaluation of different 16S rRNA gene V regions for exploring bacterial diversity in a eutrophic freshwater lake. **Science of the Total Environment**, v. 618, p. 1254–1267, 2018.

ZUGAIB, M.; ULLER, V. C.; BONFÁ, M. H. P. **Fundamentos do refino de petróleo tecnologia e economia**. 3 ed ed. rio de janeiro: Editora Interciência, 2016.

ZVERLOV, V. et al. Lateral Gene Transfer of Dissimilatory (Bi) Sulfite Reductase Revisited. **Journal of Bacteriology**, v. 187, n. 6, p. 2203–2208, 2005.

ANEXO A – Tabelas Suplementares

Tabela S1. Tabela de Identificação Taxonômica até o nível de gênero, da região 16SV4 sequenciada no Illumina e anotada tanto no banco de dados *SILVA* quanto no *Greengenes*.

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***SILVA***	***Greengenes***
Desconhecido						*2,46%*	*2,93%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	*Desulfovibrionaceae*	*Desulfovibrio*	*18,89%*	*17,37%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	*Desulfomicrobiaceae*	*Desulfoplanes*	*14,25%*	*0,00%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales			*0,00%*	*14,24%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfuromonadales	*Desulfuromonadaceae*	*Pelobacter*	*11,29%*	*0,00%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfuromonadales	*Pelobacteraceae*	*Pelobacter*	*0,00%*	*11,23%*
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales			*0,00%*	*7,60%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Other*	*Other*	*0,05%*	*7,39%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfobacterales	*Desulfobacteraceae*	*Desulfobacter*	*5,05%*	*2,32%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*livecontrolB21*	*uncultured bacterium*	*7,36%*	*0,00%*
Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	*Lentimicrobiaceae*	*uncultured bacterium*	*6,65%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Clostridiaceae 4*	*Caminicella*	*5,84%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Clostridiaceae*		*0,00%*	*5,84%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	*Desulfohalobiaceae*	*Desulfonauticus*	*1,84%*	*1,84%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Clostridiales Incertae Sedis*	*Dethiosulfatibacter*	*3,36%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*[Tissierellaceae]*	*ecb11*	*0,00%*	*2,88%*

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***SILVA***	***Greengenes***
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfobacterales	*Desulfobacteraceae*	*Other*	*0,01%*	*2,75%*
Bacteria	Thermotogae	Thermotogae	Thermotogales	*Thermotogaceae*	*Geotoga*	*0,00%*	*2,30%*
Bacteria	Thermotogae	Thermotogae	Petrotogales	*Petrotogaceae*	*Geotoga*	*2,29%*	*0,00%*
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	*BA008*		*0,00%*	*2,22%*
Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	*E6aC02*	*uncultured bacterium*	*2,22%*	*0,00%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfobacterales	*Desulfobacteraceae*		*0,00%*	*2,21%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfobacterales	*Desulfobacteraceae*	*uncultured*	*2,21%*	*0,00%*
Bacteria	SAR406	AB16				*0,00%*	*1,62%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	*Other*	*Other*	*0,00%*	*1,61%*
Bacteria	Marinimicrobia (SAR406 clade)	uncultured organism	uncultured organism	*uncultured organism*	*uncultured organism*	*1,03%*	*0,00%*
Bacteria	SAR406	AB16	noFP_H7			*0,00%*	*1,03%*
Bacteria	Marinimicrobia (SAR406 clade)	Marinimicrobia bacterium 46_47	Marinimicrobia bacterium 46_47	*Marinimicrobia bacterium 46_47*	*Marinimicrobia bacterium 46_47*	*0,83%*	*0,00%*
Bacteria	Marinimicrobia (SAR406 clade)	Other	Other	*Other*	*Other*	*0,79%*	*0,00%*
Bacteria	ZB3					*0,00%*	*0,79%*
Bacteria	Tenericutes	Mollicutes	Acholeplasmatales	*Acholeplasmataceae*	*Acholeplasma*	*0,39%*	*0,39%*
Bacteria	Margulisbacteria	uncultured organism	uncultured organism	*uncultured organism*	*uncultured organism*	*0,75%*	*0,00%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfobacterales	*Desulfobulbaceae*		*0,00%*	*0,66%*
Bacteria	Proteobacteri	Deltaproteobacte	Desulfobact	*Desulfobulba*	*Desulfobulb*	*0,66%*	*0,00%*

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***SILVA***	***Greengenes***
	a	ria	erales	*ceae*	*us*		
Bacteria	Thermotogae	Thermotogae	Petrotogales	*Petrotogaceae*	*SC103*	*0,64%*	*0,00%*
Bacteria	Thermotogae	Thermotogae	Thermotogales	*Thermotogaceae*	*SC103*	*0,00%*	*0,64%*
Bacteria	Spirochaetes	Spirochaetes	Spirochaetales	*Spirochaetaceae*		*0,00%*	*0,51%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*[Tissierellaceae]*	*Dethiosulfatibacter*	*0,00%*	*0,48%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	*Desulfomicrobiaceae*	*Desulfomicrobium*	*0,45%*	*0,01%*
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	*Marinilabiliaceae*	*uncultured*	*0,46%*	*0,00%*
Bacteria	Cloacimonetes	Cloacimonadia	Cloacimonadales	*MSBL8*	*uncultured bacterium*	*0,46%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Eubacteriaceae*	*Acetobacterium*	*0,23%*	*0,23%*
Bacteria	WWE1	[Cloacamonae]	[Cloacamonales]	*MSBL8*		*0,00%*	*0,46%*
Archaea	Euryarchaeota	Methanomicrobia	Methanosarcinales	*Methanosarcinaceae*	*Methanolobus*	*0,22%*	*0,23%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	*Desulfomicrobiaceae*		*0,00%*	*0,43%*
Bacteria	Synergistetes	Synergistia	Synergistales	*Synergistaceae*	*Thermovirga*	*0,39%*	*0,00%*
Bacteria	Synergistetes	Synergistia	Synergistales	*Thermovirgaceae*		*0,00%*	*0,38%*
Bacteria	Cloacimonetes	Cloacimonadia	Cloacimonadales	*Cloacimonadaceae*	*LNR A2-18*	*0,36%*	*0,00%*
Bacteria	WWE1	[Cloacamonae]	[Cloacamonales]	*[Cloacamonaceae]*	*W22*	*0,00%*	*0,36%*
Archaea	Nanoarchaeaeota	Woesearchaeia	uncultured euryarchaeote	*uncultured euryarchaeote*	*uncultured euryarchaeote*	*0,34%*	*0,00%*
Archaea	Nanoarchaeaeota	Woesearchaeia	Other	*Other*	*Other*	*0,33%*	*0,00%*

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***SILVA***	***Greengenes***
Bacteria	Chloroflexi	Anaerolineae	Anaerolineales	*Anaerolinaceae*	*T78*	*0,00%*	*0,33%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfuromonadales	*Desulfuromonadaceae*	*Desulfuromonas*	*0,18%*	*0,15%*
Bacteria	Chloroflexi	Anaerolineae	Anaerolineales	*Anaerolineaceae*	*ADurb.Bin120*	*0,32%*	*0,00%*
Bacteria	Chloroflexi	Anaerolineae	OPB11			*0,00%*	*0,31%*
Bacteria	OD1					*0,00%*	*0,31%*
Bacteria	Patescibacteria	Parcubacteria	Candidatus Moranbacteria	*uncultured bacterium*	*uncultured bacterium*	*0,31%*	*0,00%*
Bacteria	Spirochaetes	Spirochaetia	Spirochaetales	*Spirochaetaceae*	*Sphaerochaeta*	*0,30%*	*0,00%*
Bacteria	Tenericutes	Mollicutes				*0,00%*	*0,30%*
Bacteria	GN02	BD1-5				*0,00%*	*0,29%*
Bacteria	Patescibacteria	Gracilibacteria	JGI 0000069-P22	*uncultured bacterium*	*uncultured bacterium*	*0,29%*	*0,00%*
Bacteria	Tenericutes	Mollicutes	Izimaplasmatales	*Izimaplasmataceae*	*Other*	*0,28%*	*0,00%*
Archaea	[Parvarchaeota]	[Parvarchaea]	YLA114			*0,00%*	*0,25%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfarculales	*Desulfarculaceae*	*Desulfarculus*	*0,07%*	*0,18%*
Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	*E6aC02*	*uncultured organism*	*0,24%*	*0,00%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfobacterales	*Desulfobacteraceae*	*Desulfococcus*	*0,09%*	*0,14%*
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	*SB-1*		*0,00%*	*0,22%*
Bacteria	Spirochaetes	MVP-15	PL-11B10			*0,00%*	*0,22%*
Bacteria	Spirochaetes	Spirochaetia	Spirochaetales	*Spirochaetaceae*	*Sediminispirochaeta*	*0,22%*	*0,00%*
Archaea	Nanoarchaeaeota	Woesearchaeia	uncultured archaeon	*uncultured archaeon*	*uncultured archaeon*	*0,21%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales			*0,00%*	*0,21%*

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***SILVA***	***Greengenes***
Bacteria	Firmicutes	Clostridia	Clostridiales	*Peptococcaceae*	*uncultured*	*0,21%*	*0,00%*
Bacteria	Spirochaetes	MVP-15	uncultured organism	*uncultured organism*	*uncultured organism*	*0,21%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Christensenellaceae*	*Christensenellaceae R-7 group*	*0,20%*	*0,00%*
Archaea	[Parvarchaeota]	[Parvarchaea]	WCHD3-30			*0,00%*	*0,19%*
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	*Prolixibacteraceae*	*uncultured*	*0,18%*	*0,00%*
Bacteria	Chloroflexi	Anaerolineae	SBR1031	*uncultured bacterium*	*uncultured bacterium*	*0,17%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*JTB215*		*0,00%*	*0,16%*
Bacteria	Synergistetes	Synergistia	Synergistales	*Synergistaceae*	*EBM-39*	*0,16%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Ruminococcaceae*		*0,00%*	*0,15%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Ruminococcaceae*	*uncultured*	*0,14%*	*0,00%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfobacterales	*Desulfobacteraceae*	*Desulfotignum*	*0,07%*	*0,07%*
Bacteria	Synergistetes	Synergistia	Synergistales	*Aminiphilaceae*		*0,00%*	*0,14%*
Bacteria	Chloroflexi	Anaerolineae	SBR1031	*uncultured Chloroflexi bacterium*	*uncultured Chloroflexi bacterium*	*0,13%*	*0,00%*
Bacteria	Chloroflexi	Anaerolineae	SBR1031	*uncultured organism*	*uncultured organism*	*0,13%*	*0,00%*
Bacteria	Dependentiae	Babeliae	Babeliales	*Babeliaceae*	*Other*	*0,13%*	*0,00%*
Bacteria	OD1	ABY1				*0,00%*	*0,13%*
Bacteria	TM6	SJA-4				*0,00%*	*0,13%*
Bacteria	Chloroflexi	Anaerolineae	SBR1031	*SHA-31*		*0,00%*	*0,12%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*EtOH8*		*0,00%*	*0,12%*
Bacteria	Proteobacteri	Deltaproteobacte	Desulfarcul	*Desulfarculac*	*Desulfocarb*	*0,11%*	*0,00%*

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***SILVA***	***Greengenes***
	a	ria	ales	*eae*	*o*		
Bacteria	Verrucomicrobia	Verruco-5	LD1-PB3			*0,00%*	*0,11%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*JTB215*	*uncultured bacterium*	*0,10%*	*0,00%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfarculales	*Desulfarculaceae*		*0,00%*	*0,10%*
Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	*Lentimicrobiaceae*	*Other*	*0,09%*	*0,00%*
Bacteria	Chloroflexi	Anaerolineae	Anaerolineales	*Anaerolineaceae*	*uncultured*	*0,09%*	*0,00%*
Bacteria	Patescibacteria	ABY1	Candidatus Falkowbacteria	*Other*	*Other*	*0,09%*	*0,00%*
Bacteria	Proteobacteria	Alphaproteobacteria				*0,00%*	*0,09%*
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	*uncultured*	*uncultured bacterium*	*0,09%*	*0,00%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfuromonadales	*Pelobacteraceae*	*Other*	*0,00%*	*0,09%*
Bacteria	Proteobacteria	Deltaproteobacteria	FW113	*uncultured organism*	*uncultured organism*	*0,09%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Christensenellaceae*		*0,00%*	*0,08%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Family XIII*	*Anaerovorax*	*0,08%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*[Mogibacteriaceae]*	*Anaerovorax*	*0,00%*	*0,08%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	*Desulfohalobiaceae*	*Desulfovermiculus*	*0,04%*	*0,04%*
Bacteria	Thermotogae	Thermotogae	Thermotogales	*Thermotogaceae*	*Kosmotoga*	*0,00%*	*0,08%*
Bacteria	Cloacimonetes	Cloacimonadia	MSBL2	*Other*	*Other*	*0,07%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Christensenellaceae*	*uncultured*	*0,07%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Dehalobacter*	*Dehalobacte*	*0,00%*	*0,07%*

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***SILVA***	***Greengenes***
				iaceae	*rium*		
Bacteria	Thermotogae	Thermotogae	Kosmotogales	*Kosmotogaceae*	*uncultured*	*0,07%*	*0,00%*
Bacteria	WWE1	MSBL2				*0,00%*	*0,07%*
Bacteria	Firmicutes	Bacilli	Lactobacillales	*Streptococcaceae*	*Streptococcus*	*0,03%*	*0,04%*
Bacteria	Cyanobacteria	4C0d-2	YS2			*0,00%*	*0,06%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Clostridiaceae*	*Other*	*0,00%*	*0,06%*
Bacteria	Firmicutes	Clostridia	Halanaerobiales	*Halanaerobiaceae*		*0,00%*	*0,06%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfuromonadales	*Other*	*Other*	*0,03%*	*0,03%*
Bacteria	Thermotogae	Thermotogae	Petrotogales	*Petrotogaceae*	*Petrotoga*	*0,06%*	*0,00%*
Bacteria	Thermotogae	Thermotogae	Thermotogales	*Thermotogaceae*	*Petrotoga*	*0,00%*	*0,06%*
Bacteria	Verrucomicrobia	Verrucomicrobiae	LD1-PB3	*Other*	*Other*	*0,06%*	*0,00%*
Archaea	Euryarchaeota	DSEG	HydGC-84-221A			*0,00%*	*0,05%*
Archaea	Nanoarchaeaeota	Nanohaloarchaeia	Deep Sea Euryarchaeotic Group(DSEG)	*uncultured archaeon*	*uncultured archaeon*	*0,05%*	*0,00%*
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	*Rikenellaceae*	*DMER64*	*0,05%*	*0,00%*
Bacteria	Bacteroidetes	Ignavibacteria	Ignavibacteriales	*Melioribacteraceae*	*Melioribacter*	*0,05%*	*0,00%*
Bacteria	Chlorobi	Ignavibacteria	Ignavibacteriales	*OTUb7*		*0,00%*	*0,05%*
Bacteria	Chloroflexi	Anaerolineae	Other	*Other*	*Other*	*0,00%*	*0,05%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*JTB215*	*Other*	*0,05%*	*0,00%*
Bacteria	Proteobacteri	Deltaproteobacte	Desulfobact	*Desulfobacte*	*Desulfatibac*	*0,05%*	*0,00%*

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***SILVA***	***Greengenes***
	a	ria	erales	*raceae*	*illum*		
Bacteria	Proteobacteri a	Deltaproteobacte ria	Desulfurom onadales	*Geobacterace ae*	*Geothermob acter*	*0,05%*	*0,00%*
Bacteria	Proteobacteri a	Gammaproteoba cteria	Betaproteob acteriales	*Neisseriaceae*	*Neisseria*	*0,05%*	*0,00%*
Bacteria	Verrucomicr obia	Verrucomicrobia e	LD1-PB3	*uncultured bacterium*	*uncultured bacterium*	*0,05%*	*0,00%*
Archaea	Euryarchaeot a	Methanococci	Methanococ cales	*Methanococc aceae*	*Methanococ cus*	*0,02%*	*0,02%*
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidale s	*Porphyromon adaceae*	*Porphyromo nas*	*0,02%*	*0,02%*
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidale s	*Prolixibacter aceae*	*Tangfeifania*	*0,04%*	*0,00%*
Bacteria	Calditrichaeo ta	Calditrichia	Calditrichal es	*Calditrichace ae*	*uncultured organism*	*0,04%*	*0,00%*
Bacteria	Chloroflexi	Anaerolineae	Anaerolinea les	*Anaerolinace ae*		*0,00%*	*0,04%*
Bacteria	Chloroflexi	Anaerolineae	Anaerolinea les	*Anaerolinace ae*	*C1_B004*	*0,00%*	*0,04%*
Bacteria	Chloroflexi	Anaerolineae	Anaerolinea les	*Anaerolinace ae*	*SHD-231*	*0,00%*	*0,04%*
Bacteria	Cloacimonet es	Cloacimonadia	Cloacimona dales	*NK-L14*	*uncultured organism*	*0,04%*	*0,00%*
Bacteria	Epsilonbacte raeota	Campylobacteria	Campylobac terales	*Sulfurospirill aceae*	*Sulfurospiril lum*	*0,04%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Clostridiacea e 2*	*Other*	*0,04%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Clostridiacea e*	*Clostridium*	*0,00%*	*0,04%*
Bacteria	Fusobacteria	Fusobacteriia	Fusobacteri ales	*Fusobacteria ceae*	*Fusobacteri um*	*0,02%*	*0,02%*
Bacteria	Gemmatimo nadetes	Gemmatimonad etes	Gemmatimo nadales	*Gemmatimon adaceae*	*uncultured*	*0,04%*	*0,00%*
Bacteria	Halanaerobia eota	Halanaerobiia	Halanaerobi ales	*Halanaerobia ceae*	*Halanaerobi um*	*0,04%*	*0,00%*
Bacteria	Margulisbact	Other	Other	*Other*	*Other*	*0,04%*	*0,00%*

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***SILVA***	***Greengenes***
	eria						
Bacteria	Proteobacteria	Alphaproteobacteria	BD7-3			*0,00%*	*0,04%*
Bacteria	Proteobacteria	Alphaproteobacteria	Micavibrionales	*Micavibrionaceae*	*uncultured*	*0,04%*	*0,00%*
Bacteria	Proteobacteria	Betaproteobacteria	Neisseriales	*Neisseriaceae*	*Neisseria*	*0,00%*	*0,04%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	*Desulfovibrionaceae*	*Desulfocurvus*	*0,04%*	*0,00%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfuromonadales	*Pelobacteraceae*		*0,00%*	*0,04%*
Bacteria	Proteobacteria	Epsilonproteobacteria	Campylobacterales	*Campylobacteraceae*		*0,00%*	*0,04%*
Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	*Pasteurellaceae*	*Haemophilus*	*0,02%*	*0,02%*
Bacteria	[Caldithrix]	KSB1	GW-22			*0,00%*	*0,04%*
Archaea	Euryarchaeota	Methanobacteria	Methanobacteriales	*WSA2*		*0,00%*	*0,03%*
Archaea	Euryarchaeota	Methanomicrobia	Methanosarcinales	*Methanosarcinaceae*	*Other*	*0,00%*	*0,03%*
Archaea	Euryarchaeota	Methanomicrobia	Methanosarcinales	*Methanosarcinaceae*	*Methanococcoides*	*0,03%*	*0,00%*
Archaea	Euryarchaeota	Thermococci	Methanofastidiosales	*uncultured*	*uncultured archaeon*	*0,03%*	*0,00%*
Bacteria						*0,00%*	*0,03%*
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	*Prolixibacteraceae*	*Roseimarinus*	*0,03%*	*0,00%*
Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	*ST-12K33*	*uncultured organism*	*0,03%*	*0,00%*
Bacteria	Bacteroidetes	Ignavibacteria	Ignavibacteriales	*Melioribacteraceae*	*IheB3-7*	*0,03%*	*0,00%*
Bacteria	Chlorobi	Ignavibacteria	Ignavibacteriales	*lheB3-7*		*0,00%*	*0,03%*
Bacteria	Chloroflexi	Anaerolineae	Anaerolineales	*Anaerolineaceae*	*Ornatilinea*	*0,03%*	*0,00%*
Bacteria	Chloroflexi	Anaerolineae	Anaerolinea	*Anaerolineac*	*Pelolinea*	*0,03%*	*0,00%*

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***SILVA***	***Greengenes***
			les	*eae*			
Bacteria	Cloacimonetes	Cloacimonadia	Cloacimonadales	*LK-44f*	*uncultured bacterium*	*0,03%*	*0,00%*
Bacteria	Cyanobacteria	Melainabacteria	Gastranaerophilales	*Other*	*Other*	*0,03%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Dehalobacteriaceae*		*0,00%*	*0,03%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Peptococcaceae*		*0,00%*	*0,03%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Peptostreptococcaceae*	*Tepidibacter*	*0,00%*	*0,03%*
Bacteria	Kiritimatiellaeota	Kiritimatiellae	WCHB1-41	*uncultured bacterium*	*uncultured bacterium*	*0,03%*	*0,00%*
Bacteria	OP11	WCHB1-64	d153			*0,00%*	*0,03%*
Bacteria	Patescibacteria	ABY1	Candidatus Falkowbacteria	*uncultured bacterium*	*uncultured bacterium*	*0,03%*	*0,00%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfuromonadales	*Desulfuromonadaceae*	*Other*	*0,03%*	*0,00%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfuromonadales	*Desulfuromonadaceae*		*0,00%*	*0,03%*
Bacteria	Verrucomicrobia	Verruco-5	LD1-PB3	*Other*	*Other*	*0,00%*	*0,03%*
Bacteria	WWE1	[Cloacamonae]	[Cloacamonales]			*0,00%*	*0,03%*
Archaea	Euryarchaeota	Methanomicrobia	Methanomicrobiales	*Methanomicrobiaceae*	*Methanogenium*	*0,01%*	*0,01%*
Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	*Flavobacteriaceae*	*Capnocytophaga*	*0,02%*	*0,00%*
Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	*Other*	*Other*	*0,02%*	*0,00%*
Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	*Flavobacteriaceae*	*Capnocytophaga*	*0,00%*	*0,02%*
Bacteria	Caldiserica	Caldisericia	Caldisericales	*WCHB1-02*	*uncultured bacterium*	*0,02%*	*0,00%*
Bacteria	Cyanobacteri	Melainabacteria	Gastranaero	*uncultured*	*uncultured*	*0,02%*	*0,00%*

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***SILVA***	***Greengenes***
	a		philales	*bacterium*	*bacterium*		
Bacteria	Epsilonbacteraeota	Campylobacteria	Campylobacterales	*Arcobacteraceae*	*Arcobacter*	*0,02%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Eubacteriaceae*		*0,00%*	*0,02%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Eubacteriaceae*	*Alkalibacter*	*0,01%*	*0,01%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Eubacteriaceae*	*Anaerofustis*	*0,02%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*GoM-GC232-4463-Bac1*	*Other*	*0,02%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Peptococcaceae*	*Candidatus Dichloromethanomonas*	*0,02%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Peptococcaceae*	*Dehalobacter_Syntrophobotulus*	*0,00%*	*0,02%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Peptococcaceae*	*Dehalobacterium*	*0,02%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Ruminococcaceae*	*Other*	*0,02%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Ruminococcaceae*	*Ruminococcaceae UCG-013*	*0,02%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Syntrophomonadaceae*	*Syntrophomonas*	*0,01%*	*0,01%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*uncultured*	*Other*	*0,02%*	*0,00%*
Bacteria	Gemmatimonadetes	Gemmatimonadetes				*0,00%*	*0,02%*
Bacteria	Halanaerobiaeota	Halanaerobiia	Halanaerobiales	*Halanaerobiaceae*	*Halocella*	*0,02%*	*0,00%*
Bacteria	Nitrospirae	Nitrospira	Nitrospirales	*0319-6A21*		*0,00%*	*0,02%*
Bacteria	Planctomycetes	vadinHA49	PeHg47			*0,00%*	*0,02%*
Bacteria	Planctomycet	vadinHA49	uncultured	*uncultured*	*uncultured*	*0,02%*	*0,00%*

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***SILVA***	***Greengenes***
	es		organism	*organism*	*organism*		
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	*Rhodospirillaceae*		*0,00%*	*0,02%*
Bacteria	Proteobacteria	Betaproteobacteria	MND1			*0,00%*	*0,02%*
Bacteria	Proteobacteria	Deltaproteobacteria	Other	*Other*	*Other*	*0,00%*	*0,02%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfobacterales	*Desulfobacteraceae*	*Desulfosalsimonas*	*0,02%*	*0,00%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	*Desulfonatronaceae*	*Desulfonatronum*	*0,02%*	*0,00%*
Bacteria	Proteobacteria	Deltaproteobacteria	Myxococcales	*0319-6G20*		*0,00%*	*0,02%*
Bacteria	Proteobacteria	Deltaproteobacteria	Oligoflexales	*0319-6G20*	*uncultured organism*	*0,02%*	*0,00%*
Bacteria	Proteobacteria	Epsilonproteobacteria	Campylobacterales	*Campylobacteraceae*	*Arcobacter*	*0,00%*	*0,02%*
Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	*Nitrosomonadaceae*	*MND1*	*0,02%*	*0,00%*
Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	*Pseudomonadaceae*	*Pseudomonas*	*0,01%*	*0,01%*
Bacteria	Rokubacteria	NC10	Rokubacteriales	*uncultured bacterium*	*uncultured bacterium*	*0,02%*	*0,00%*
Bacteria	Spirochaetes	Spirochaetes	Spirochaetales	*Spirochaetaceae*	*Treponema*	*0,00%*	*0,02%*
Bacteria	Spirochaetes	Spirochaetia	Spirochaetales	*Spirochaetaceae*	*Treponema 2*	*0,02%*	*0,00%*
Bacteria	Tenericutes	Mollicutes	Mycoplasmatales	*Mycoplasmataceae*	*Mycoplasma*	*0,01%*	*0,01%*
Bacteria	Thermotogae	Thermotogae	Thermotogales	*Fervidobacteriaceae*	*Thermosipho*	*0,02%*	*0,00%*
Bacteria	Thermotogae	Thermotogae	Thermotogales	*Thermotogaceae*	*Thermosipho*	*0,00%*	*0,02%*
Archaea	Euryarchaeota	Methanomicrobia	Methanosarcinales	*Methanosarcinaceae*	*Methanohalophilus*	*0,01%*	*0,00%*
Bacteria	Acidobacteri	Acidobacteria-6	iii1-15			*0,00%*	*0,01%*

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***SILVA***	***Greengenes***
	a						
Bacteria	Acidobacteria	BPC102	B110			*0,00%*	*0,01%*
Bacteria	Acidobacteria	Blastocatellia (Subgroup 4)	Pyrinomonadales	*Pyrinomonadaceae*	*RB41*	*0,01%*	*0,00%*
Bacteria	Acidobacteria	Subgroup 21	uncultured bacterium	*uncultured bacterium*	*uncultured bacterium*	*0,01%*	*0,00%*
Bacteria	Acidobacteria	Subgroup 6	uncultured bacterium	*uncultured bacterium*	*uncultured bacterium*	*0,01%*	*0,00%*
Bacteria	Acidobacteria	[Chloracidobacteria]	RB41			*0,00%*	*0,01%*
Bacteria	Actinobacteria	Thermoleophilia	Gaiellales	*Gaiellaceae*		*0,00%*	*0,01%*
Bacteria	Actinobacteria	Thermoleophilia	Gaiellales	*Gaiellaceae*	*Gaiella*	*0,01%*	*0,00%*
Bacteria	BRC1	NPL-UPA2				*0,00%*	*0,01%*
Bacteria	BRC1	uncultured organism	uncultured organism	*uncultured organism*	*uncultured organism*	*0,01%*	*0,00%*
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	*Prevotellaceae*	*Alloprevotella*	*0,01%*	*0,00%*
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	*[Paraprevotellaceae]*	*[Prevotella]*	*0,00%*	*0,01%*
Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	*Weeksellaceae*	*Bergeyella*	*0,01%*	*0,00%*
Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	*CMW-169*	*uncultured bacterium*	*0,01%*	*0,00%*
Bacteria	Bacteroidetes	Bacteroidia	Sphingobacteriales	*SBYC-2600*	*uncultured organism*	*0,01%*	*0,00%*
Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	*[Weeksellaceae]*		*0,00%*	*0,01%*
Bacteria	Bacteroidetes	Ignavibacteria	Kryptoniales	*MSB-3C8*	*metagenome*	*0,01%*	*0,00%*
Bacteria	Bacteroidetes	Ignavibacteria	OPB56	*uncultured organism*	*uncultured organism*	*0,01%*	*0,00%*
Bacteria	Calditrichaeota	Calditrichia	Calditrichales	*Calditrichaceae*	*SM23-31*	*0,01%*	*0,00%*

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***SILVA***	***Greengenes***
Bacteria	Chlorobi					*0,00%*	*0,01%*
Bacteria	Chlorobi	OPB56				*0,00%*	*0,01%*
Bacteria	Chloroflexi	Anaerolineae	Anaerolineales	*Anaerolineaceae*	*Anaerolineaceae UCG-001*	*0,01%*	*0,00%*
Bacteria	Epsilonbacteraeota	Campylobacteria	Campylobacterales	*Campylobacteraceae*	*Campylobacter*	*0,01%*	*0,00%*
Bacteria	Firmicutes	Bacilli	Bacillales	*Family XI*	*Gemella*	*0,01%*	*0,00%*
Bacteria	Firmicutes	Bacilli	Gemellales	*Gemellaceae*		*0,00%*	*0,01%*
Bacteria	Firmicutes	Bacilli	Lactobacillales	*Other*	*Other*	*0,00%*	*0,01%*
Bacteria	Firmicutes	Bacilli	Lactobacillales	*Carnobacteriaceae*	*Granulicatella*	*0,01%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Clostridiaceae 1*	*Oxobacter*	*0,01%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Clostridiaceae 4*	*Other*	*0,01%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Clostridiaceae 4*	*uncultured*	*0,01%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Clostridiaceae*	*Geosporobacter_Thermotalea*	*0,00%*	*0,01%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Clostridiaceae*	*Oxobacter*	*0,00%*	*0,01%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Clostridiales vadinBB60 group*	*Other*	*0,01%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Clostridiales vadinBB60 group*	*uncultured bacterium*	*0,01%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Lachnospiraceae*		*0,00%*	*0,01%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Peptococcaceae*	*Niigata-25*	*0,00%*	*0,01%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Peptococcaceae*	*Peptococcus*	*0,00%*	*0,01%*

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***SILVA***	***Greengenes***
Bacteria	Firmicutes	Clostridia	Clostridiales	*Peptostreptococcaceae*	*Peptoclostridium*	*0,01%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Peptostreptococcaceae*	*Sporacetigenium*	*0,01%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Ruminococcaceae*	*Ercella*	*0,01%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Ruminococcaceae*	*Ruminococcaceae UCG-010*	*0,01%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*Ruminococcaceae*	*Ruminococcaceae UCG-014*	*0,01%*	*0,00%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*[Acidaminobacteraceae]*	*Other*	*0,00%*	*0,01%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*[Acidaminobacteraceae]*		*0,00%*	*0,01%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*[Tissierellaceae]*	*Other*	*0,00%*	*0,01%*
Bacteria	Firmicutes	Clostridia	Clostridiales	*uncultured*	*uncultured bacterium*	*0,01%*	*0,00%*
Bacteria	Gemmatimonadetes	Gemm-1				*0,00%*	*0,01%*
Bacteria	Gemmatimonadetes	Gemmatimonadetes	N1423WL			*0,00%*	*0,01%*
Bacteria	Kiritimatiellaeota	Kiritimatiellae	WCHB1-41	*Other*	*Other*	*0,01%*	*0,00%*
Bacteria	Kiritimatiellaeota	Kiritimatiellae	WCHB1-41	*uncultured Verrucomicrobia bacterium*	*uncultured Verrucomicrobia bacterium*	*0,01%*	*0,00%*
Bacteria	Kiritimatiellaeota	Kiritimatiellae	WCHB1-41	*uncultured organism*	*uncultured organism*	*0,01%*	*0,00%*
Bacteria	Latescibacteria	Latescibacteria	Latescibacterales	*Latescibacteraceae*	*Other*	*0,01%*	*0,00%*
Bacteria	Marinimicrobia (SAR406	uncultured eubacterium	uncultured eubacterium	*uncultured eubacterium*	*uncultured eubacterium*	*0,01%*	*0,00%*

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***SILVA***	***Greengenes***
	clade)	AB16	AB16	*AB16*	*AB16*		
Bacteria	OP11	OP11-2				*0,00%*	*0,01%*
Bacteria	Patescibacteria	ABY1	Candidatus Kerfeldbacteria	*uncultured bacterium*	*uncultured bacterium*	*0,01%*	*0,00%*
Bacteria	Patescibacteria	Microgenomatia	Candidatus Shapirobacteria	*uncultured bacterium*	*uncultured bacterium*	*0,01%*	*0,00%*
Bacteria	Planctomycetes	Phycisphaerae	AKAU3564			*0,00%*	*0,01%*
Bacteria	Planctomycetes	Phycisphaerae	MSBL9			*0,00%*	*0,01%*
Bacteria	Planctomycetes	Phycisphaerae	MSBL9	*L21-RPul-D3*	*Other*	*0,01%*	*0,00%*
Bacteria	Planctomycetes	Phycisphaerae	Phycisphaerales	*AKAU3564 sediment group*	*uncultured organism*	*0,01%*	*0,00%*
Bacteria	Proteobacteria	Alphaproteobacteria	Other	*Other*	*Other*	*0,01%*	*0,00%*
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillales	*Azospirillaceae*	*Nitrospirillum*	*0,01%*	*0,00%*
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	*Sphingomonadaceae*	*Kaistobacter*	*0,00%*	*0,01%*
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	*Sphingomonadaceae*	*Sphingomonas*	*0,01%*	*0,00%*
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	*Burkholderiaceae*	*Lautropia*	*0,00%*	*0,01%*
Bacteria	Proteobacteria	Betaproteobacteria	Neisseriales	*Neisseriaceae*	*Other*	*0,00%*	*0,01%*
Bacteria	Proteobacteria	Deltaproteobacteria				*0,00%*	*0,01%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfobacterales	*Desulfobacteraceae*	*SEEP-SRB1*	*0,01%*	*0,00%*
Bacteria	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	*Desulfohalobiaceae*	*Other*	*0,01%*	*0,00%*
Bacteria	Proteobacteri	Deltaproteobacte	Desulfovibri	*Desulfovibrio*	*Other*	*0,01%*	*0,00%*

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***SILVA***	***Greengenes***
	a	ria	onales	*naceae*			
Bacteria	Proteobacteri a	Deltaproteobacte ria	Desulfurom onadales	*Geobacterace ae*	*Other*	*0,01%*	*0,00%*
Bacteria	Proteobacteri a	Deltaproteobacte ria	Desulfurom onadales	*Geobacterace ae*	*Geoalkaliba cter*	*0,01%*	*0,00%*
Bacteria	Proteobacteri a	Deltaproteobacte ria	Desulfurom onadales	*Geobacterace ae*	*Geobacter*	*0,00%*	*0,01%*
Bacteria	Proteobacteri a	Deltaproteobacte ria	Myxococcal es			*0,00%*	*0,01%*
Bacteria	Proteobacteri a	Deltaproteobacte ria	Myxococcal es	*VHS-B3-70*	*Other*	*0,01%*	*0,00%*
Bacteria	Proteobacteri a	Deltaproteobacte ria	Myxococcal es	*bacteriap25*	*uncultured bacterium*	*0,01%*	*0,00%*
Bacteria	Proteobacteri a	Deltaproteobacte ria	NB1-j	*NB1-i*		*0,00%*	*0,01%*
Bacteria	Proteobacteri a	Deltaproteobacte ria	NB1-j	*uncultured bacterium*	*uncultured bacterium*	*0,01%*	*0,00%*
Bacteria	Proteobacteri a	Deltaproteobacte ria	Syntrophob acterales	*Syntrophobac teraceae*		*0,00%*	*0,01%*
Bacteria	Proteobacteri a	Epsilonproteoba cteria	Campylobac terales	*Campylobact eraceae*	*Campylobac ter*	*0,00%*	*0,01%*
Bacteria	Proteobacteri a	Gammaproteoba cteria	Betaproteob acteriales	*Burkholderia ceae*	*Lautropia*	*0,01%*	*0,00%*
Bacteria	Proteobacteri a	Gammaproteoba cteria	Oceanospiri llales	*Nitrincolacea e*	*Marinobacte rium*	*0,01%*	*0,00%*
Bacteria	Proteobacteri a	Gammaproteoba cteria	Oceanospiri llales	*Oceanospirill aceae*		*0,00%*	*0,01%*
Bacteria	Proteobacteri a	Gammaproteoba cteria	Salinisphaer ales	*Solimonadac eae*	*Polycyclovor ans*	*0,01%*	*0,00%*
Bacteria	Proteobacteri a	Gammaproteoba cteria	Xanthomon adales	*Sinobacterac eae*		*0,00%*	*0,01%*
Bacteria	Proteobacteri a	Gammaproteoba cteria	Xanthomon adales	*Xanthomonad aceae*		*0,00%*	*0,01%*
Bacteria	Spirochaetes	MVP-15	uncultured Firmicutes bacterium	*uncultured Firmicutes bacterium*	*uncultured Firmicutes bacterium*	*0,01%*	*0,00%*

Reino	***Filo***	***Classe***	***Ordem***	***Família***	***Gênero***	***SILVA***	***Greengenes***
Bacteria	Spirochaetes	Spirochaetes	Sphaerochaetales	*Sphaerochaetaceae*		*0,00%*	*0,01%*
Bacteria	Synergistetes	Synergistia	Synergistales	*Other*	*Other*	*0,00%*	*0,01%*
Bacteria	Tenericutes	Mollicutes	Izimaplasmatales	*Izimaplasmataceae*	*uncultured bacterium*	*0,01%*	*0,00%*
Bacteria	Tenericutes	Mollicutes	Mollicutes RF39	*uncultured bacterium*	*uncultured bacterium*	*0,01%*	*0,00%*
Bacteria	Tenericutes	Mollicutes	RF39			*0,00%*	*0,01%*
Bacteria	Thermotogae	Thermotogae	Kosmotogales	*Kosmotogaceae*	*Kosmotoga*	*0,01%*	*0,00%*
Bacteria	Thermotogae	Thermotogae	Kosmotogales	*Kosmotogaceae*	*Mesotoga*	*0,01%*	*0,00%*
Bacteria	Thermotogae	Thermotogae	Petrotogales	*Petrotogaceae*	*Oceanotoga*	*0,01%*	*0,00%*
Bacteria	Verrucomicrobia	Verruco-5	LD1-PB3	*LD1-PA34*		*0,00%*	*0,01%*
Bacteria	Verrucomicrobia	Verruco-5	WCHB1-41	*WCHB1-25*		*0,00%*	*0,01%*
Bacteria	WS3	PRR-12	PBS-III-9			*0,00%*	*0,01%*
Bacteria	ZB3	BS119				*0,00%*	*0,01%*
Bacteria	[Caldithrix]	KSB1	WM105			*0,00%*	*0,01%*

Tabela S2. Relação metabolismo entre gene, e número de OTU's totais que possuem o gene.

Metabolismo	**VIA**	***gene***	***Número de OTU's***
Metano	ACDS	*ackA*	631
Metano	ACDS	*acsB*	162
Enxofre	DSR	*aprA*	93
Enxofre	DSR	*aprB*	52
Metano	ACDS	*cdhE*	95
Enxofre	ASR	*cysH*	261
Enxofre	ASR	*cysI*	65
Enxofre	ASR	*cysJ*	81
Enxofre	DSR	*drsA*	85
Enxofre	DSR	*drsB*	85

Metabolismo	**VIA**	***gene***	***Número de OTU's***
Metano	Todas	*frhB*	73
Metano	CO_2	*ftr*	10
Metano	CO_2	*fwdA*	12
Metano	CO_2	*mch*	30
Metano	Todas	*mcrA*	9
Metano	Todas	*mcrC*	12
Metano	CO_2	*mer*	50
Metano	MOH	*mtaC*	6
Metano	Todas	*mtrH*	10
Metano	Todas	*mvhA*	50
Metano	Todas	*mvhD*	67
Nitrogênio	DNR-DN	*napA*	28
Nitrogênio	DNR-DN	*narG*	81
Nitrogênio	ANR	*nasA*	37
Nitrogênio	ANR	*nirA*	52
Nitrogênio	DNR	*nirB*	115
Nitrogênio	DNR	*nirD*	68
Nitrogênio	DN	*nirK*	16
Nitrogênio	DN	*norB*	39
Nitrogênio	DN	*nosZ*	39
Nitrogênio	DNR	*nrfA*	157
Número total de OTU's			771

- Em que ACDS é Acetil-CoA Descaboxilase Sintase, DSR – Redução Dissimilatória de Sulfato, ASR – Redução Assimilatória de Sulfato, DNR – Redução Dissimilatória de Nitrato, ANR – Redução Assimilatória de Nitrato, DN - Denitrificação.

Tabela S3. Tabela de OTU's compartilhadas entre as populações que possuem os genes relacionados às vias de redução de enxofre com as demais

gene	***cysJ***	***sir***	***cysI***	***aprA***	***cysH***	***aprB***	***drsB***	***drsA***
ackA	78	1	54	84	206	43	66	66
acsB	0	0	0	29	28	21	43	43
aprA	0	0	0	93	69	52	54	54
aprB	0	0	0	52	32	52	52	52
cdhE	0	0	0	26	14	21	39	39
cysH	68	15	60	69	261	32	33	33
cysI	36	0	65	0	60	0	1	1

cysJ	81	1	36	0	68	0	1	1
drsA	1	0	1	54	33	52	85	85
drsB	1	0	1	54	33	52	85	85
frhB	0	0	0	18	31	18	26	26
ftr	1	1	0	0	9	0	0	0
fwdA	3	1	2	0	11	0	0	0
mch	1	15	0	1	28	0	0	0
mcrA	0	0	0	0	8	0	0	0
mcrC	0	0	0	0	8	0	0	0
mer	0	14	6	0	48	0	0	0
mtaC	0	0	0	0	5	0	1	1
mtrH	0	0	0	2	7	0	3	3
mvhA	0	0	1	22	10	22	42	42
mvhD	0	0	0	24	10	23	43	43
napA	5	0	1	7	7	2	2	2
narG	6	0	13	6	19	1	8	8
nasA	27	1	29	0	36	0	0	0
nirA	15	1	9	0	48	0	0	0
nirB	28	1	41	10	82	10	10	10
nirD	18	1	30	0	60	0	0	0
nirK	12	0	8	1	11	1	2	2
norB	14	0	21	0	23	0	1	1
nosZ	11	14	11	0	26	0	0	0
nrfA	11	14	9	35	74	34	43	43
Total	81	15	65	93	261	52	85	85

Tabela S4. Tabela de OTU's compartilhadas entre as populações que possuem alguns genes relacionados às vias de síntese de acetato

gene	***ackA***	***acsB***	***cdhE***
ackA	631	132	57
acsB	132	162	79
aprA	84	29	26
aprB	43	21	21
cdhE	57	79	95
cysH	206	28	14
cysI	54	0	0
cysJ	78	0	0

gene	*ackA*	*acsB*	*cdhE*
drsA	66	43	39
drsB	66	43	39
frhB	30	19	27
ftr	2	1	5
fwdA	4	1	5
mch	9	1	5
mcrA	1	1	5
mcrC	0	1	9
mer	10	3	5
mtaC	1	1	5
mtrH	3	4	5
mvhA	28	38	37
mvhD	44	57	54
napA	26	7	7
narG	75	12	12
nasA	34	0	0
nirA	47	0	1
nirB	90	22	10
nirD	59	16	0
nirK	14	1	1
norB	37	10	10
nosZ	18	4	4
nrfA	132	32	30
Total	631	162	95

Tabela S5. Tabela de OTU's compartilhadas entre as populações que possuem alguns genes relacionados às vias de produção de metano

gene	*frhB*	*fwdA*	*mer*	*mcrA*	*mtrH*	*ftr*	*mch*	*fwdE*	*mtaC*	*mvhA*	*mvhD*	*mcrC*
ackA	30	4	10	1	3	2	9	116	1	28	44	0
acsB	19	1	3	1	4	1	1	62	1	38	57	1
aprA	18	0	0	0	2	0	1	55	0	22	24	0
aprB	18	0	0	0	0	0	0	52	0	22	23	0
cdhE	27	5	5	5	5	5	5	40	5	37	54	9
cysH	31	11	48	8	7	9	28	95	5	10	10	8
cysI	0	2	6	0	0	0	0	0	0	1	0	0
cysJ	0	3	0	0	0	1	1	1	0	0	0	0

gene	***frhB***	***fwdA***	***mer***	***mcrA***	***mtrH***	***ftr***	***mch***	***fwdE***	***mtaC***	***mvhA***	***mvhD***	***mcrC***
drsA	26	0	0	0	3	0	0	55	1	42	43	0
drsB	26	0	0	0	3	0	0	55	1	42	43	0
frhB	73	9	19	8	7	9	8	26	5	21	22	12
ftr	9	10	8	8	7	10	9	8	5	1	1	8
fwdA	9	12	8	8	7	10	9	8	5	1	1	8
mch	8	9	21	7	7	9	30	22	5	0	0	7
mcrA	8	8	8	9	7	8	7	7	5	1	1	8
mcrC	12	8	8	8	7	8	7	7	5	2	2	12
mer	19	8	50	8	7	8	21	21	5	1	1	8
mtaC	5	5	5	5	6	5	5	6	6	1	1	5
mtrH	7	7	7	7	10	7	7	10	6	1	1	7
mvhA	21	1	1	1	1	1	0	24	1	50	48	2
mvhD	22	1	1	1	1	1	0	34	1	48	67	2
napA	0	0	0	1	0	0	0	2	0	2	2	0
narG	1	0	2	0	0	0	0	1	0	6	4	0
nasA	0	3	1	1	0	1	1	0	0	0	0	0
nirA	1	1	8	1	0	1	1	21	0	0	0	0
nirB	22	4	22	0	0	2	8	28	0	6	6	0
nirD	0	1	3	0	0	1	1	24	0	0	0	0
nirK	0	0	1	1	0	0	0	1	0	1	1	0
norB	0	0	0	1	0	0	0	12	0	1	8	0
nosZ	0	0	14	1	0	0	14	17	0	1	0	0
nrfA	9	0	15	0	0	0	14	76	0	19	36	0
Total	73	12	50	9	10	10	30	146	6	50	67	12

Tabela S6. Tabela de OTU's compartilhadas entre as populações que possuem alguns genes relacionados às vias do metabolismo do nitrogênio

gene	***nosZ***	***nirB***	***nirD***	***napA***	***nirA***	***nirK***	***norB***	***nasA***	***narG***	***nrfA***
ackA	18	90	59	26	47	14	37	34	75	132
acsB	4	22	16	7	0	1	10	0	12	32
aprA	0	10	0	7	0	1	0	0	6	35
aprB	0	10	0	2	0	1	0	0	1	34
cdhE	4	10	0	7	1	1	10	0	12	30
cysH	26	82	60	7	48	11	23	36	19	74
cysI	11	41	30	1	9	8	21	29	13	9
cysJ	11	28	18	5	15	12	14	27	6	11

gene	***nosZ***	***nirB***	***nirD***	***napA***	***nirA***	***nirK***	***norB***	***nasA***	***narG***	***nrfA***
drsA	0	10	0	2	0	2	1	0	8	43
drsB	0	10	0	2	0	2	1	0	8	43
frhB	0	22	0	0	1	0	0	0	1	9
ftr	0	2	1	0	1	0	0	1	0	0
fwdA	0	4	1	0	1	0	0	3	0	0
mch	14	8	1	0	1	0	0	1	0	14
mcrA	1	0	0	1	1	1	1	1	0	0
mcrC	0	0	0	0	0	0	0	0	0	0
mer	14	22	3	0	8	1	0	1	2	15
mtaC	0	0	0	0	0	0	0	0	0	0
mtrH	0	0	0	0	0	0	0	0	0	0
mvhA	1	6	0	2	0	1	1	0	6	19
mvhD	0	6	0	2	0	1	8	0	4	36
napA	4	2	8	28	3	2	4	3	13	11
narG	6	11	19	13	3	4	9	9	81	21
nasA	4	30	21	3	16	3	9	37	9	6
nirA	1	30	14	3	52	4	4	16	3	19
nirB	6	115	36	2	30	4	13	30	11	18
nirD	6	36	68	8	14	3	13	21	19	13
nirK	5	4	3	2	4	16	11	3	4	5
norB	10	13	13	4	4	11	39	9	9	16
nosZ	39	6	6	4	1	5	10	4	6	22
nrfA	22	18	13	11	19	5	16	6	21	157
Total	39	115	68	28	52	16	39	37	81	157

Tabela S7. Organização dos genes previstos pela estrutura de classificação do KEEG.

KEGG_Nível 1	**KEGG_Nível 2**	**Contagem de Genes**	**Percentual**
Processamento de Informação Ambiental	Transporte de Membrana	9234600	11,20%
Metabolismo	Metabolismo de Aminoácidos	7226314	8,77%
Metabolismo	Metabolismo dos Carboidratos	7218829	8,76%
Processamento de Informação Genética	Replicação e Reparo	5874717	7,13%
Metabolismo	Metabolismo Energético	5600239	6,79%
Processos Celulares	Motilidade Celular	5169853	6,27%
Não classificado	Mal caracterizada	4597943	5,58%

KEGG_Nível 1	KEGG_Nível 2	Contagem de Genes	Percentual
Processamento de Informação Genética	Tradução	4320081	5,24%
Metabolismo	Metabolismo de Cofactores e Vitaminas	3590903	4,36%
Não classificado	Processos e Sinalização Celular	3498466	4,24%
Metabolismo	Metabolismo nucleotico	2868785	3,48%
Processamento de Informação Ambiental	Transdução de Sinal	2664397	3,23%
Não classificado	Processamento de Informação Genética	2630686	3,19%
Não classificado	Metabolismo	2248568	2,73%
Metabolismo	Metabolismo Lipídico	2176698	2,64%
Processamento de Informação Genética	Dobra, Classificação e Degradação	2097565	2,54%
Metabolismo	Biossíntese e Metabolismo do Glicano	1967238	2,39%
Processamento de Informação Genética	Transcrição	1691683	2,05%
Metabolismo	Famílias Enzimáticas	1534020	1,86%
Metabolismo	Biodegradação e Metabolismo de Xenobióticos	1323622	1,61%
Metabolismo	Metabolismo de Terpenóides e Policetídeos	1275231	1,55%
Metabolismo	Metabolismo de outros aminoácidos	963571	1,17%
Metabolismo	Biossíntese de outros metabólitos secundários	696201	0,84%
Processos Celulares	Crescimento Celular e Morte	476582	0,58%
Doenças Humanas	Doenças infecciosas	299363	0,36%
Sistemas Organismais	Sistema endócrino	261491	0,32%
Sistemas Organismais	Adaptação Ambiental	176000	0,21%
Processos Celulares	Transporte e Catabolismo	171648	0,21%
Doenças Humanas	Doenças Neurodegenerativas	111849	0,14%
Doenças Humanas	Cancros	100100	0,12%
Processamento de Informação Ambiental	Moléculas e Interações de Sinalização	98977	0,12%
Doenças Humanas	Doenças Metabólicas	69930	0,08%

KEGG_Nível 1	KEGG_Nível 2	Contagem de Genes	Percentual
Sistemas Organismais	Sistema nervoso	55263	0,07%
Sistemas Organismais	Sistema imunológico	52266	0,06%
Sistemas Organismais	Sistema digestivo	43490	0,05%
Sistemas Organismais	Sistema Excretor	28587	0,03%
Doenças Humanas	Doenças do Sistema Imune	13721	0,02%

Tabela S8. Outros Subníveis previstos pelo KEGG.

KEGG_Nível 2	KEGG_Nível 3	Contagem
Adaptação Ambiental	Interação planta-patógeno	166647
Transporte de membrana	ABC transportadores	2688769
Transporte de membrana	Sistema de secreção bacteriana	584485
Moléculas de Sinalização e Interações	Toxinas Bacterianas	66440
Sistema de secreção	transporte de membrana	1554356
Transportadores	Transporte de Membrana	4265533
Transdução de sinal	Sistema de dois componentes	2562645
Comunicação Celular	Adere à Junção	3
Mobilidade Celular	Proteínas de motilidade bacteriana	2642453
Assembleia Flagelar	Mobilidade Celular	1027965
Processos Celulares e Sinalização	Mobilidade Celular e secreção	259672
Processos Celulares e Sinalização	Transportadores de transferência de elétrons	94835
Outros	Metabolismo Energético	842362

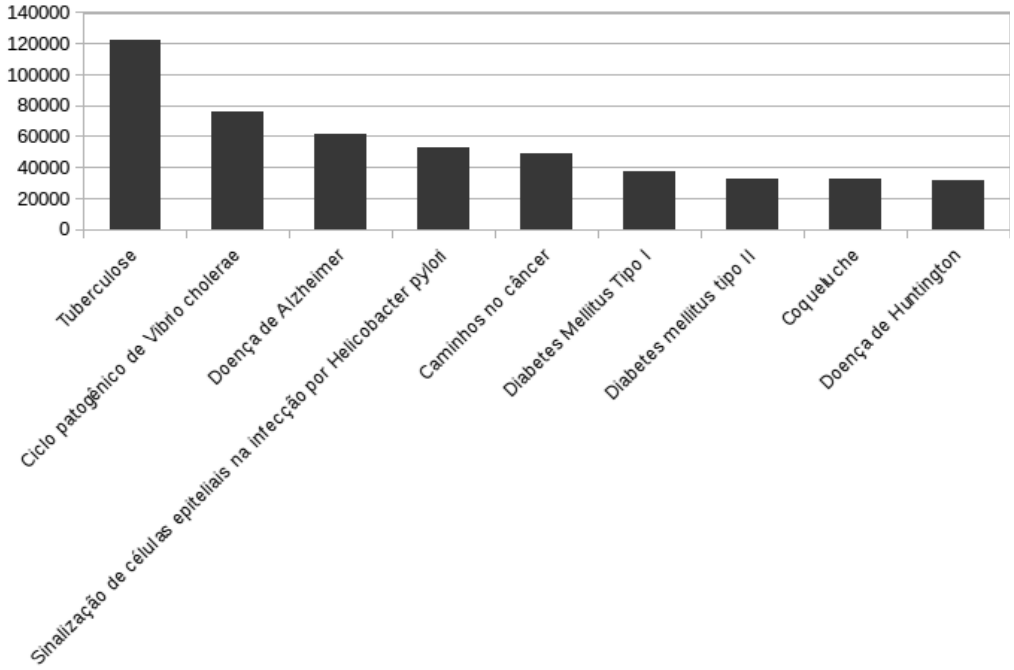

Figura S1. Contagem de Genes Relacionados à Doenças Humanas

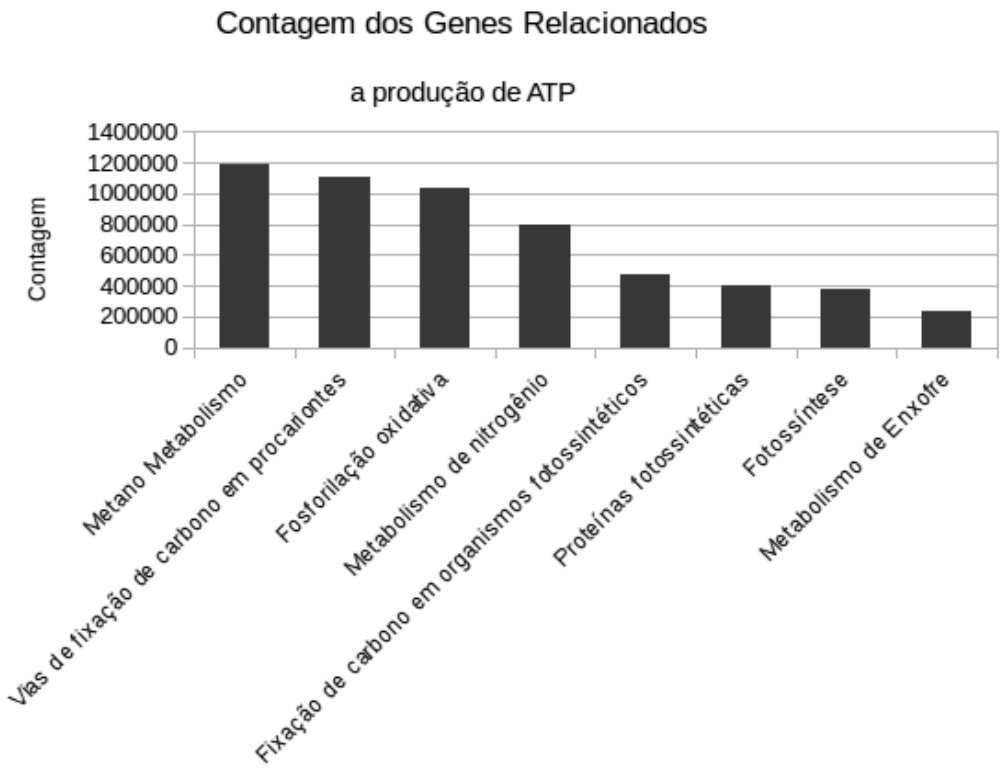

Figura S2.Contagem de Genes Relacionados a produção de ATP