Felipe Borim Corrêa

Estabelecimento de um *pipeline* para identificação de *Mycobacterium leprae* em microbiomas através do gene rRNA 16S

Belo Horizonte

Julho de 2017

Felipe Borim Corrêa

Estabelecimento de um *pipeline* para identificação de *Mycobacterium leprae* em microbiomas através do gene rRNA 16S

Dissertação apresentada ao Programa de Pós-Graduação em Bioinformática do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais como requisito para obtenção do título de Mestre em Bioinformática.

Universidade Federal de Minas Gerais - UFMG

Instituto de Ciências Biológicas

Programa de Pós-Graduação em Bioinformática

Orientador: Prof. Dr. Gabriel da Rocha Fernandes

Belo Horizonte

Julho de 2017

Felipe Borim Corrêa
Estabelecimento de um *pipeline* para identificação de *Mycobacterium leprae* em microbiomas através do gene rRNA 16S / Felipe Borim Corrêa. – Belo Horizonte, Julho de 2017-
78 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Gabriel da Rocha Fernandes

Dissertação (Mestrado) – Universidade Federal de Minas Gerais - UFMG
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Bioinformática, Julho de 2017.

1. Bioinformática. 2. Metagenômica. 3. 16S rRNA. I. Fernandes, Gabriel da Rocha. II. UFMG. III. ICB. IV. Título

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG

ATA DA DEFESA DE DISSERTAÇÃO

32/2017
entrada
2º/2015
CPF:
388.976.208-50

Felipe Borim Correa

Às quatorze horas do dia **03 de julho de 2017**, reuniu-se, no Instituto de Ciências Biológicas da UFMG, a Comissão Examinadora de Dissertação, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Estabelecimento de um pipeline para identificação de Mycobacterium leprae em microbiomas através do gene rRNA 16S"**, requisito para obtenção do grau de Mestre em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Gabriel da Rocha Fernandes**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	Instituição	CPF	Indicação
Dr. Gabriel da Rocha Fernandes	FIOCRUZ	05270621622	Aprovado
Dra. Andréa Maria Amaral Nascimento	UFMG	47734060625	Aprovado
Dr. Francisco Pereira Lobo	UFMG	02227373694	Aprovado

Pelas indicações, o candidato foi considerado: _____
O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.
Belo Horizonte, 03 de julho de 2017.

Dr. Gabriel da Rocha Fernandes - Orientador _____
Dra. Andréa Maria Amaral Nascimento Andréa Amaral
Dr. Francisco Pereira Lobo Francisco P. Lobo

Este trabalho é dedicado às crianças adultas que,
quando pequenas, sonharam em se tornar cientistas.

Agradecimentos

Primeiramente, agradeço aos meus pais Célia e José, e irmão Bruno que, através de tempestades e calmarias perseveraram e nunca deixaram o caçula ao relento. Juntamente, agradeço aos finados avós José Borim e Dalila, com os quais convivi a maior parte de minha vida e que permanecerão vivos em mim enquanto eu estiver caminhando nessa terra.

Não posso me esquecer do professor Dr. Leopoldo Sussumu Matsumoto que iniciou e acompanhou meu engatinhar na vida acadêmica, assim como a Dra. Mayra que teve o papel catalítico essencial ao me introduzir à Bioinformática.

Aqui em Belo Horizonte, onde resido enquanto escrevo essa dedicatória, começo agradecendo ao professor Dr. José Miguel Ortega e ao meu caro orientador Dr. Gabriel da Rocha Fernandes, pessoas que foram indispensáveis ao receberem dispostos um aspirante a cientista. À toda secretaria de pós-graduação em Bioinformática da UFMG, especialmente à Sheila Santana e Dr. Vasco Azevedo.

Também agradeço a Antonio Edson, Carlos Henrique e Dhiego, amigos bioinformatas que fiz durante esse trajeto. Agradecimentos especiais a Vanessa Monteiro, mulher inefável com a qual tive oportunidades de compartilhar momentos incríveis. Além desses, meus companheiros de republica Germain, Sérgio, Israel e Carlos Magno.

Aos meus queridos companheiros de laboratório Fausto, Daniel, Francislon, Juliana, Laura, Sara, Anderson, Julianne, Ângela, Douglas, Larissa, Victor, Amanda e Wagner que foram fundamentais sobre o desenvolvimento da minha pesquisa, sobre meu desenvolvimento pessoal e que se tornaram pessoas memoráveis. Por fim, agradeço à agência de fomento CAPES pelo concedimento da bolsa.

"Macacos me mordam antes que se extinguam."
(Cláudio Márcio de Souza Santos)

Resumo

Mycobacterium leprae é uma bactéria patogênica e agente etiológico da hanseníase, enfermidade que incide principalmente em países subdesenvolvidos. O estudo desse microorganismo não é trivial pois o mesmo não cresce em meios de cultura tradicionais, contudo, abordagens independentes de cultura baseadas no gene rRNA 16S podem ser usadas para o estudo comunidades microbianas. É sabido que existem vieses relacionados à amplificação e classificação taxonômica nessas abordagens, dessa forma, este trabalho tem como objetivo definir um *pipeline* que possibilite o estudo de *M. leprae* em microbiomas. A metodologia foi dividida em quatro etapas. Primeiro foi feita uma pré-seleção de iniciadores e seus respectivos amplicons simulados de *M. leprae* que permitiram a classificação taxonômica até nível de espécie desse organismo. A segunda análise avaliou a capacidade de amplificação dos iniciadores selecionados. A terceira etapa foi a avaliação da sensibilidade na classificação taxonômica de comunidades microbianas simuladas e, por fim, um modelo de regressão logística foi utilizado para identificar as regiões hipervariáveis mais informativas do rRNA 16S desse organismo. Apenas amplicons de iniciadores que cobriram as regiões V1-V2, V2-V3 e V6 de *M. leprae* puderam ser classificados até nível de espécie. No entanto, os iniciadores que flanqueiam V6 geraram amplicons que apresentaram maior sensibilidade na classificação de *M. leprae* das comunidades simuladas e possibilitaram a classificação de mais táxons que os demais. A regressão logística corroborou com os resultados, mostrando que as regiões V1, V2 e V6 são as mais informativas no rRNA 16S desse organismo. Com base nos resultados, foi possível chegar em algumas recomendações para a classificação de *M. leprae* em microbiomas, sendo sugerido o uso de iniciadores que flanqueiam a região V6, o uso bancos de dados de referência Silva e a classificação taxonômica a partir da definição de OTUs a 97% de similaridade.

Palavras-chave: bioinformática. metagenômica. hanseníase.

Abstract

Mycobacterium leprae is a pathogenic bacteria and the etiologic agent of leprosy, disease which affects mainly underdeveloped countries. The study of this microorganism is not trivial because it does not grow in traditional culture media; however, culture-independent approaches based on rRNA 16S gene can be used to study microbial communities. It is known that there are biases related to the amplification and taxonomic classification in these approaches; thus, the objective of this work was to define a pipeline which allows the study of *M. leprae* in microbiomes. The methods were divided into four parts. First was made a pre-selection of primers and its respective simulated amplicons of M. leprae that allowed the taxonomic classification at species level of this organism. The second analysis evaluated the amplification capacity of the selected primers. The third step was the evaluation of the sensitivity of the taxonomic classification of mock communities and, lastly, a logistic regression model was used to identify the most informative hypervariable regions of the 16S rRNA for *M. leprae*. Only amplicons of primers that covered the regions V1-V2, V2-V3 and V6 of *M. leprae* could be classified at species level. Though, primers which flank V6 regions generated amplicons that showed higher sensitivity in the classification of *M. leprae* from the mock communities and allowed the classification of more taxa than the others. Logistic regression corroborated with the results, showing that regions V1, V2 and V6 are the most informative in this organism 16S rRNA. Based on the results, were possible to reach in some recommendations for the classification of *M. leprae* in microbiomes, being suggested the use of primers for V6 region, Silva reference database and taxonomic classification using 97% similarity OTUs.

Keywords: bioinformatics. metagenomics. leprosy.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

1.1 *Mycobacterium leprae*

Mycobacterium leprae[1] é uma bactéria gram-positiva e agente causador da hanseníase, mal de Hansen ou lepra. Esse organismo foi descoberto pelo médico norueguês Gerhard Armauer Hansen, tendo sido o primeiro microrganismo patogênico a ser descrito (HANSEN, 1874). A descoberta quebrou o paradigma relacionado à hereditariedade da doença no século XIX, fato que era inquestionável até então (BECHLER, 2012).

Apesar de ter sido primeira bactéria patogênica a ser identificada, o histórico da doença é de conhecimento muito anterior. A hanseníase é uma doença citada desde os tempos bíblicos, possuindo registros de casos datados com mais de 3 mil anos. Sua origem geográfica ainda não é muito clara, já que não se sabe se a doença se originou no continente asiático ou africano. Acredita-se que a hanseníase tenha sido introduzida na Europa pela Índia através das tropas de Alexandre, o grande, em 300 a.C., e, presume-se que sua introdução nas Américas tenha ocorrido a partir da colonização francesa nos Estados Unidos, e da espanhola e portuguesa na América do Sul (LASTÓRIA; ABREU, 2014).

É uma doença infecciosa crônica que afeta principalmente o sistema nervoso periférico e a pele, além de outros tecidos como o sistema reticuloendotelial, ossos e articulações. Seus sinais clínicos são variados, ocorrendo na forma de poucas lesões até casos mais generalizados. Desde os anos 60, a classificação da doença se baseia em caracteres clínicos, histopatológicos, na carga bacterial e grau de resposta imune celular. Porém, a partir de 1982, para fins de tratamento através da terapia multidroga (MDT), a classificação da doença foi simplificada em dois grupos pela Organização Mundial de Saúde (OMS). De acordo com um índice baseado na carga bacteriana, a infecção foi classificada como Paucibacilar (PB) ou Multibacilar (MB) (TALHARI; TALHARI; PENNA, 2015).

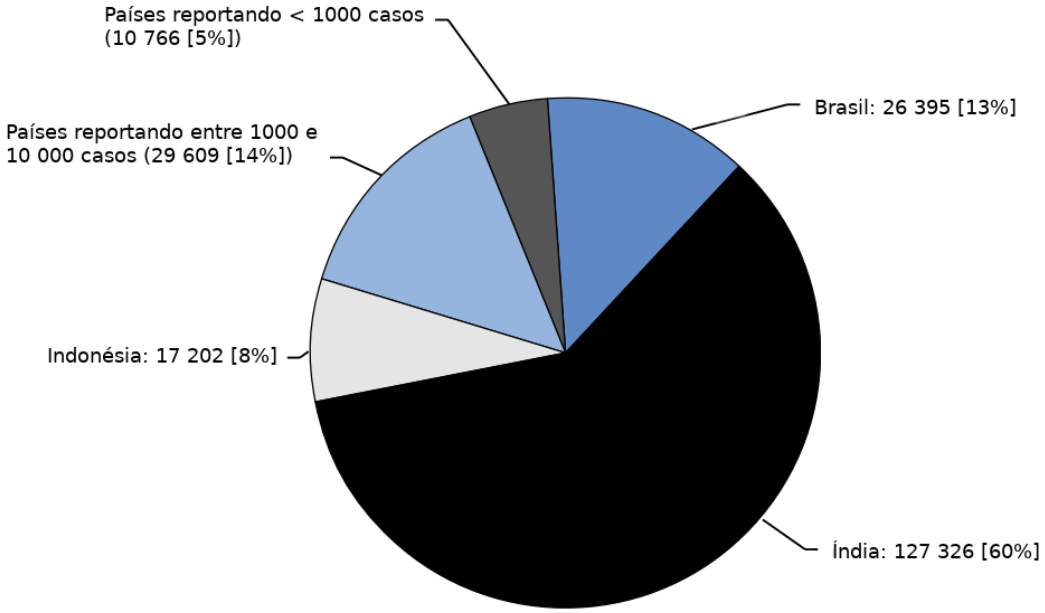
Além do tratamento convencional com a terapia multidroga, novos métodos vem sendo desenvolvidos com base em melhoramentos na MDT, assim como a criação de novas drogas. A busca pela vacina como arma para prevenção da hanseníase também está sendo estudada, porém ainda não foi estabelecido nenhum método definitivo até os dias de hoje. Dessa forma, hoje o padrão ouro para tratamento da hanseníase continua sendo a terapia multidroga desenvolvida pela Organização Mundial da Saúde (KAR; GUPTA, 2015).

Muito embora tenha havido uma grande diminuição de novos casos da doença a partir da introdução da terapia multidroga em meados dos anos 1980, atualmente, países subdesenvolvidos ainda tem registrado milhares de novos casos (WHO, 2016). Em 2015, o

[1] <http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1769>

número acumulado de novos casos na Índia, Brasil e Indonésia representaram mais de 80% dos novos casos no mundo (≈ 170 mil) (ver Figura 1).

Figura 1 – Gráfico da distribuição geográfica dos novos casos da hanseníase em 2015, segmentado por país.



Fonte: modificado de WHO (2016)

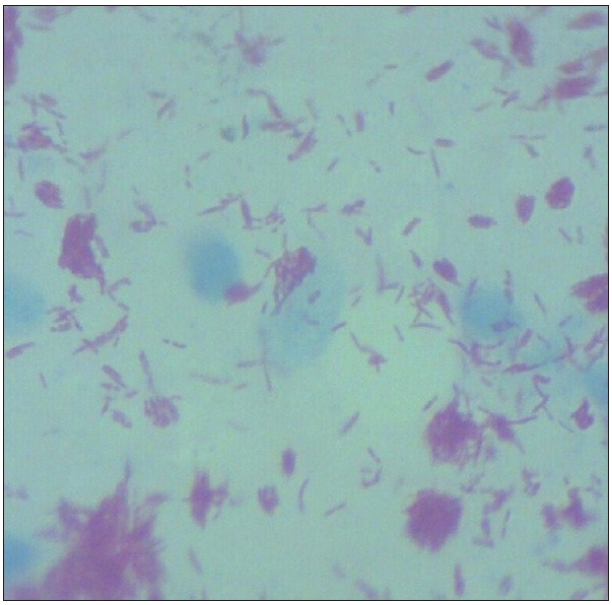
Sabe-se que o estudo de *M. leprae* é bastante atribulado pois o bacilo não cresce em meios de cultura tradicionais. É comum o cultivo em tecidos vivos como os de camundongos e tatus, mas ainda assim é um processo muito laborioso, extremamente demorado e restrito (TRUMAN; KRAHENBUHL, 2001). Um trabalho recente mostrou que *M. leprae* apresenta um tipo de crescimento muito atípico pois não acontece de forma exponencial (AMAKO et al., 2016).

O genoma de *M. leprae* sugere que seu crescimento extremamente lento tem relação com o grande número de pseudogenes e genoma reduzido em comparação com o genoma de *Mycobacterium tuberculosis*. Mais precisamente, apenas 49,5% do genoma de *M. leprae* contém genes codificadores de proteínas. Os demais 27% foram identificados como pseudogenes e os 23,5% restantes estão possivelmente ligados a funções regulatórias, entre outras. Esse tipo de evolução redutiva normalmente é vista em parasitos obrigatórios e endossimbiontes já que diversas funções vão se inativando durante a evolução pois ocorre um afunilamento e simplificação das necessidades metabólicas (COLE; EIGLMEIER; PARKHILL, 2001).

O diagnóstico da hanseníase é comumente baseado em sintomas e sinais clínicos. Em um país ou região endêmica um indivíduo deve ser considerado doente caso possua lesões na pele com perda definida da sensibilidade, apresentando, ou não, espessamento aparente dos nervos. Outra forma é a identificação dos bacilos após coloração do esfregaço da pele

através de métodos específicos (WHO, 2017). Como *M. leprae* não pode ser cultivada *in vitro*, o uso do microscópio ótico continua sendo a ferramenta de diagnóstico padrão. A partir do esfregaço de tecido ou biópsia de células em suspensão, uma vez espalhadas na lâmina, são coradas com o método *Ziehl-Neelsen*. Os bacilos adquirem uma coloração magenta sobre um fundo azul (REIBEL; CAMBAU; AUBRY, 2015). Um exemplo do método pode ser observado na Figura 2.

Figura 2 – Fotografia de esfregaço cutâneo mostrando os bacilos de *M. leprae*.



Método de coloração *Ziehl-Neelsen* modificado (x1000). Fonte: Kumaran et al. (2015)

A infecção não necessariamente leva a algum sintoma ou lesão. Na verdade, acredita-se que *M. leprae* não é uma bactéria muito patogênica e que a maioria das infecções não resultam em sintomas. Contudo, é considerada altamente infecciosa tendo o sistema respiratório como a principal via de entrada e saída do patógeno (VISSCHEDIJK et al., 2000).

Dados genômicos vem sendo utilizados para o desenvolvimento de métodos de detecção de *M. leprae* baseados na amplificação por PCR de uma região específica do gene marcador rRNA 16S (LAVANIA et al., 2008). Genes marcadores, tal como o rRNA 16S, podem ser utilizados como alvo para a identificação da presença de *M. leprae* em amostras ambientais de água e solo de regiões endêmicas (MOHANTY et al., 2016). Todavia, apesar desses métodos de detecção serem muito sensíveis, em comparação com a análise do esfregaço cutâneo de tecido humano (ARUNAGIRI et al., 2017; SIWAKOTI et al., 2016), utilizar a PCR para amplificar apenas o gene de uma única espécie como alvo não possibilita a caracterização de comunidades microbianas como um todo.

1.2 Metagenômica

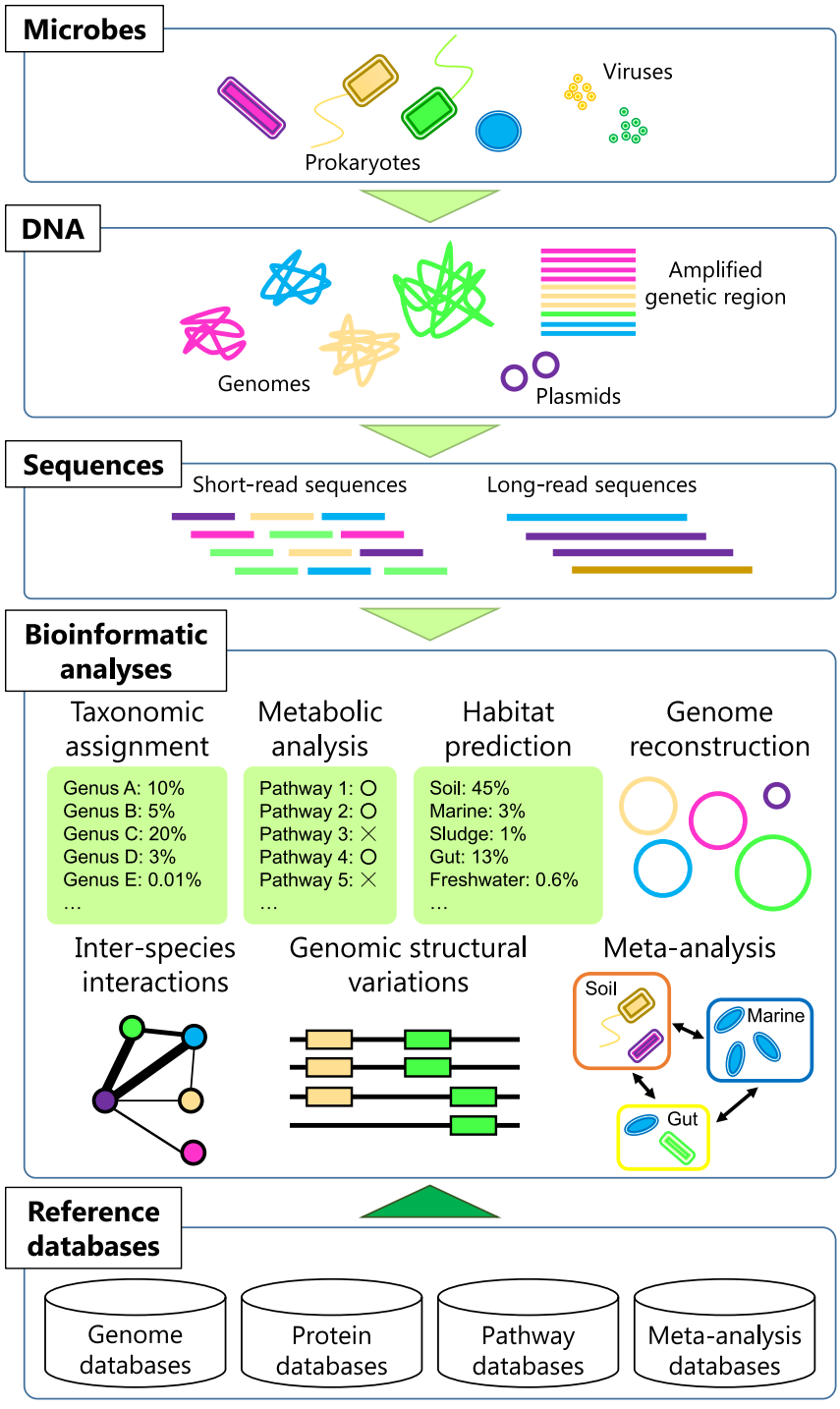
A cultura *in vitro*, na maioria das vezes, é o primeiro passo utilizado para o estudo de novos microrganismos. Infelizmente os métodos padrões de cultivo conseguem caracterizar a minoria dos organismos presentes no ambiente mantendo a vasta maioria fora do alcance dessas abordagens. Mesmo com os avanços na microbiologia clássica no cultivo *in vitro* de microrganismos, métodos moleculares vem se tornando indispensáveis para o estudo de centenas de milhares de espécies microbianas (RAVIN; MARDANOV; SKRYABIN, 2015).

As primeiras tecnologias de sequenciamento surgiram por volta dos anos 1970, transformando a biologia. De fato, o grande avanço no sequenciamento de DNA veio a partir do desenvolvimento das tecnologias chamadas de Sequenciadores de Nova Geração (NGS). Essas tecnologias, que, devido suas reações de sequenciamento em paralelo, chegaram ao mercado produzindo quantidades de dados muito superiores (*high throughput*) e numa velocidade extremamente maior que as tecnologias antecessoras. Além de todo esse progresso, o preço do sequenciamento caiu muito rapidamente chegando, em 2014, em torno de $1000 por genoma (DIJK et al., 2014).

Os sequenciadores NGS favoreceram os estudos genômicos em geral, incluindo também estudos de microbiomas, ou seja, o estudo de genomas presentes dentro de amostras ambientais. Um notável estudo nos primórdios da metagenômica se deu em 2004, quando Venter *et al.* sequenciaram amostras do Mar dos Sargaços através do método *Whole-genome shotgun sequencing* (WGS) e puderam identificar o conteúdo gênico e diversidade microbiana desse ambiente aquático. Nessa época foi possível gerar ≈1.5 Gb de sequências de DNA, e *reads* com média de 818 pares de base de comprimento (VENTER et al., 2004). Pouco adiante, outros trabalhos foram desenvolvidos caracterizando a diversidade taxonômica e funcional de alguns ambientes aquáticos e terrestres (TYSON et al., 2004; TRINGE et al., 2005), além do microbioma do sistema digestivo humano que começava a ser estudado através da metagenômica (GILL et al., 2006).

As abordagens de metagenômica são utilizadas na ecologia microbiana pois possibilitam um estudo mais profundo das comunidades de microrganismos. Essas metodologia se sustentam principalmente a partir do sequenciamento do DNA total do ambiente *Whole-genome shotgun sequencing* (WGS) e/ou a partir do sequenciamento de genes marcadores (HIRAOKA; YANG; IWASAKI, 2016). Essas abordagens muitas vezes são utilizadas em conjunto pois permitem a mais completa reconstrução dos genomas. O sequenciamento do genoma total WGS necessita da montagem das *reads* em *contigs*, fragmentos menores em maiores, para possibilitar a classificação taxonômica e a predição gênica dos genes descobertos nas amostras sequenciadas. Por outro lado, o sequenciamento de genes marcadores permite uma visão do perfil taxonômico e funcional a partir de regiões alvo específicas no genoma de cada organismo (RAVIN; MARDANOV; SKRYABIN, 2015) (ver Figura 3).

Figura 3 – Esquema mostrando as abordagens metagenômicas na ecologia microbiana.



Fonte Hiraoka, Yang e Iwasaki (2016)

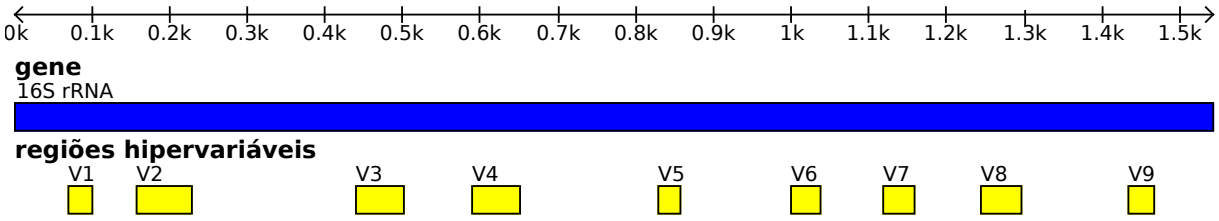
1.3 O rRNA 16S como gene marcador

O desenvolvimento da técnica da cadeia de reação da polimerase (PCR), os avanços na tecnologia de sequenciamento do DNA e uso do gene rRNA 16S como um marcador taxonômico tem possibilitado a identificação genética de Bacteria e Archaea presentes em amostras ambientais. O gene rRNA 16S, assim como outros genes marcadores possibilitaram

tanto identificar micróbios por similaridade entre sequências, tal como agrupá-los num contexto filogenético. Esses avanços permitiram identificar, inclusive, filos inteiros que não possuem representantes cultiváveis (GARZA; DUTILH, 2015). Além do mais, foi visto que árvores filogenéticas criadas a partir de sequências de rRNA 16S se assemelhavam muito àquelas criadas com base no conteúdo gênico total (KONSTANTINIDIS; TIEDJE, 2005).

Existem algumas características que fazem com que o rRNA 16S seja um bom marcador taxonômico. A primeira é o fato dele ser um gene *housekeeping*, o que confere alta conservação em sua sequência de nucleotídeos (GRAY; SANKOFF; CEDERGREN, 1984). A alta conservação permite o desenho de iniciadores (*primers*) "universais" que se ligam em uma ampla gama de organismos que possuem o gene rRNA 16S. Outro aspecto importante é que as regiões conservadas flanqueiam regiões que possuem alta variabilidade, conhecidas como regiões hipervariáveis. São essas regiões que conferem as peculiaridades dos grupos de organismos distintos, podendo ser utilizadas como identidades taxonômicas. Ao todo são 9 regiões hipervariáveis denotadas de V1 até V9, segundo Chakravorty et al. (2007), que podem ser vistas pela figura 4.

Figura 4 – Representação da estrutura do gene rRNA 16S e suas regiões hipervariáveis.



As nove regiões hipervariáveis do gene cobrem as posições 69-99, 137-242, 433-497, 576-682, 822-879, 986-1043, 1117-1173, 1243-1294 e 1435-1465, V1 até V9 respectivamente. As posições são baseadas na sequência do gene de rRNA 16S de *E. coli* (BROSIUS et al., 1978).

É muito comum que a classificação taxonômica das sequências de DNA em amostras de metagenômica seja feita a partir de agrupamento, ou *clustering*, das sequências em Unidades Taxonômicas Operacionais (OTUs). Esse conceito, a princípio, foi criado para se tratar do "taxa de menor nível em um estudo qualquer"(SNEATH; SOKAL et al., 1973), ou seja, essa unidade pode ser qualquer nível taxonômico. Hoje em dia, esse conceito é utilizado para se tratar de agrupamentos de sequências proximamente relacionadas a partir de um limiar de similaridade, não necessariamente se referindo ou representando diretamente um táxon específico.

Existem duas abordagens metodológicas para a caracterização de comunidades microbianas a partir do sequenciamento de DNA. A primeira compreende os métodos dependentes de taxonomias anotadas em bancos de dados, onde taxonomias de referência são atribuídas às sequências. O segundo método é baseado na definição de OTUs, que geralmente parte da utilização de uma matriz de distâncias com um limiar de similari-

dade entre as sequências. A grande vantagem do uso de OTUs é a não necessidade do conhecimento prévio das taxonomias em bancos de dados de referências, permitindo o agrupamento de sequências desconhecidas e portanto facilitando a descoberta de novos táxons (CHEN et al., 2013).

Diversos *pipelines* que facilitam a analise de dados brutos de sequenciamento até a obtenção de genomas e perfis taxonômicos estão disponíveis. Como exemplo temos as ferramentas QIIME (CAPORASO et al., 2010), MOTHUR (SCHLOSS et al., 2009) e USEARCH (EDGAR, 2010). Elas são normalmente customizáveis e, como exemplo, temos o pipeline QIIME que possibilita desde a definição das OTUs com diferentes métodos de clusterização. A atribuição taxonômica das OTUs encontradas pode ser inferida a partir de bancos de dados de OTUs de referência que são também totalmente arbitrários conforme quem está analisando os dados.

É sabido que existem alguns vieses relacionados a métodos metagenômicos baseados em genes marcadores. Esses estão principalmente relacionadas aos métodos de extração, amplificação do DNA e classificação taxonômica (BROOKS et al., 2015; FOUHY et al., 2016). Um problema comum na classificação taxonômica está relacionado com a amplificação de regiões hipervariáveis do rRNA 16S que não contém identidade o suficiente para separar grupos de sequências de táxons diferentes. Foi sugerido que a região V2 parece ser a melhor região para separar organismos do gênero Mycobacterium (CHAKRAVORTY et al., 2007), no entanto, ainda não existe nenhum estudo direcionado à classificação de *M. leprae* em abordagens metagenômicas.

Em um cenário ideal a sequência completa do rRNA 16S deveria ser sequenciada para cobrir todas as regiões hipervariáveis desse gene, que tem em média 1500 pares de base. Tecnologias de sequenciamento de *reads* longos ($\approx$ 5 mil pares de base), já estão sendo desenvolvidas e estão cada vez mais próximas de substituir as atuais líderes de mercado que produzem *reads* curtos. No entanto, elas ainda tem uma alta taxa de erro e o custo de sequenciamento não é competitivo. Por isso, a fabricante Illumina ainda se mantém líder com 65% da fatia de mercado, seguida pela Ion Torrent com 25% e o restante fica com as tecnologias de *reads* longos como a PacBio, mas também com outras tecnologias como a 454 Life Science e outros fabricantes mais antigos (STEINBOCK; RADENOVIC, 2015). Como atualmente os sequenciadores MiSeq da Illumina produzem fragmentos em par de até 2x 300 pares de base[2], esse trabalho será limitado ao alcance máximo dessa tecnologia em relação ao tamanho de fragmento sequenciado.

Com a evidente necessidade do estudo de *M. leprae*, as abordagens promissoras de metagenômica baseadas em genes marcadores e os vieses relacionados aos mesmos, esse trabalho visa a seleção de iniciadores e bancos de dados de referência para a definição de um pipeline sólido que possibilite futuros estudos de *M. leprae* em comunidades microbianas

[2] <https://www.illumina.com/systems/sequencing-platforms/miseq.html>

ambientais a partir da análise de sequências do gene marcador rRNA 16S.

1.4 Conceitos Fundamentais

Comunidade microbiana simulada

Uma comunidade microbiana simulada, ou *Mock Community*, é composta por um conjunto de organismos de abundância e taxonomia conhecidas. São muitas vezes construídas com a finalidade da avaliação de ferramentas de bioinformática voltadas para estudos de microbiomas. O consórcio *Human Microbiome Project* desenvolveu uma comunidade simulada devido a grande necessidade de padronização e *benchmarking* dos métodos de caracterização de microbiomas a partir de 16S rRNA. Uma comunidade microbiana simulada de 21 microrganismos do corpo humano foi desenvolvida antes do estabelecimento do protocolo oficial do consórcio para a tecnologia de sequenciamento 454 FLX Titanium (HMP 16S 454 *protocol*) (GROUP et al., 2012). Outros exemplos de comunidades microbianas simuladas são a *mockrobiota*[3] que disponibiliza diversos conjuntos de dados simulados de Bacteria, Archaea e Eukarya (BOKULICH et al., 2016), e também a MBARC-26 (*Mock Bacteria Archaea Community*) que é uma comunidade diversa de 26 organismos que também disponibiliza dados de sequenciamento Illumina e PacBio (sequenciadores de nova geração) (SINGER et al., 2016).

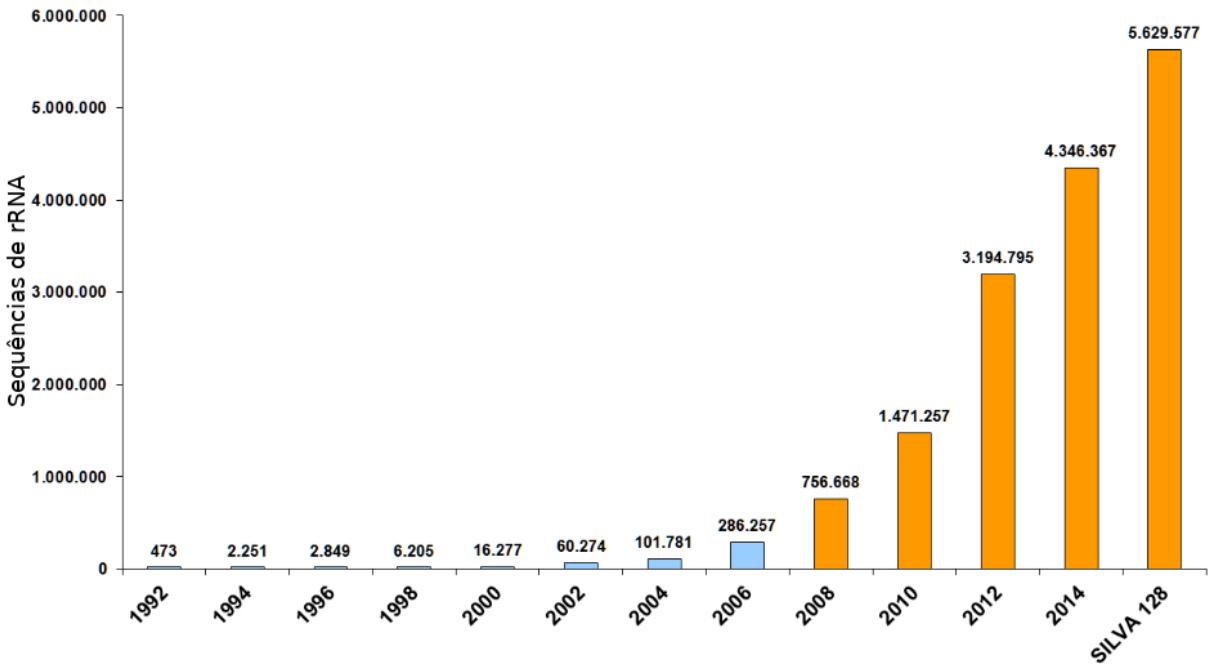
Bancos de dados do gene 16S rRNA

Os bancos de dados biológicos foram criados como elementos-chave para que os computadores manuseassem, analisassem e armazenassem dados de sequências e estruturas moleculares (ATTWOOD et al., 2011). Eles são convencionalmente divididos em dois grupos: primários e secundários. Os bancos de dados primários são aqueles que armazenam dados experimentais coletados por pesquisadores, como exemplo temos o GenBank[4] e o Protein Data Bank (PDB)[5]. Os bancos secundários representam aqueles que armazenam dados curados através do resultado de análises, processamento e interpretação de dados, geralmente, oriundos de bancos de dados primários (ZOU et al., 2015).

Entre os bancos de dados secundários, existem os bancos de dados da sequência do gene da subunidade ribossomal menor, que incluem as sequências do gene da subunidade 16S. Atualmente, os principais bancos de dados de sequências do gene 16S rRNA disponíveis são o Ribosomal Data Project (RDP) (COLE et al., 2009), GreenGenes (MCDONALD et al., 2012) e Silva (YILMAZ et al., 2013). É possível ver como a quantidade de sequências depositadas nos bancos de dados de rRNA tem crescido a cada ano (ver Figura 5).

[3] <http://mockrobiota.caporasolab.us/]>
[4] <https://www.ncbi.nlm.nih.gov/genbank/>
[5] <http://www.rcsb.org/pdb/home/home.do>

Figura 5 – Gráfico do crescimento dos bancos de dados de sequências do gene referente à subunidade ribossomal menor (RDP II & SILVA).



Fonte: Disponível em <https://www.arb-silva.de/documentation/release-128/>

2 Objetivos

2.1 Objetivo Geral

- Estabelecer um pipeline para análises de metagenômica baseadas no gene marcador rRNA 16S que possibilite a caracterização do perfil taxonômico de microbiomas, incluindo a identificação de *M. leprae*.

2.2 Objetivos Específicos

- Selecionar iniciadores e bancos de dados de referência que possibilitem a classificação taxonômica de amplicons de *M. leprae* até nível de espécie.
- Avaliar os candidatos selecionados com base na cobertura da amplificação do banco de dados Silva.
- Avaliar os candidatos selecionados com base na sensibilidade da classificação das *reads* de comunidades microbianas simuladas.
- Identificar as regiões hipervariáveis mais informativas do rRNA 16S de *M. leprae*
- Recomendar iniciadores, bancos de dados de referência de OTUs e outros requisitos para o pipeline.

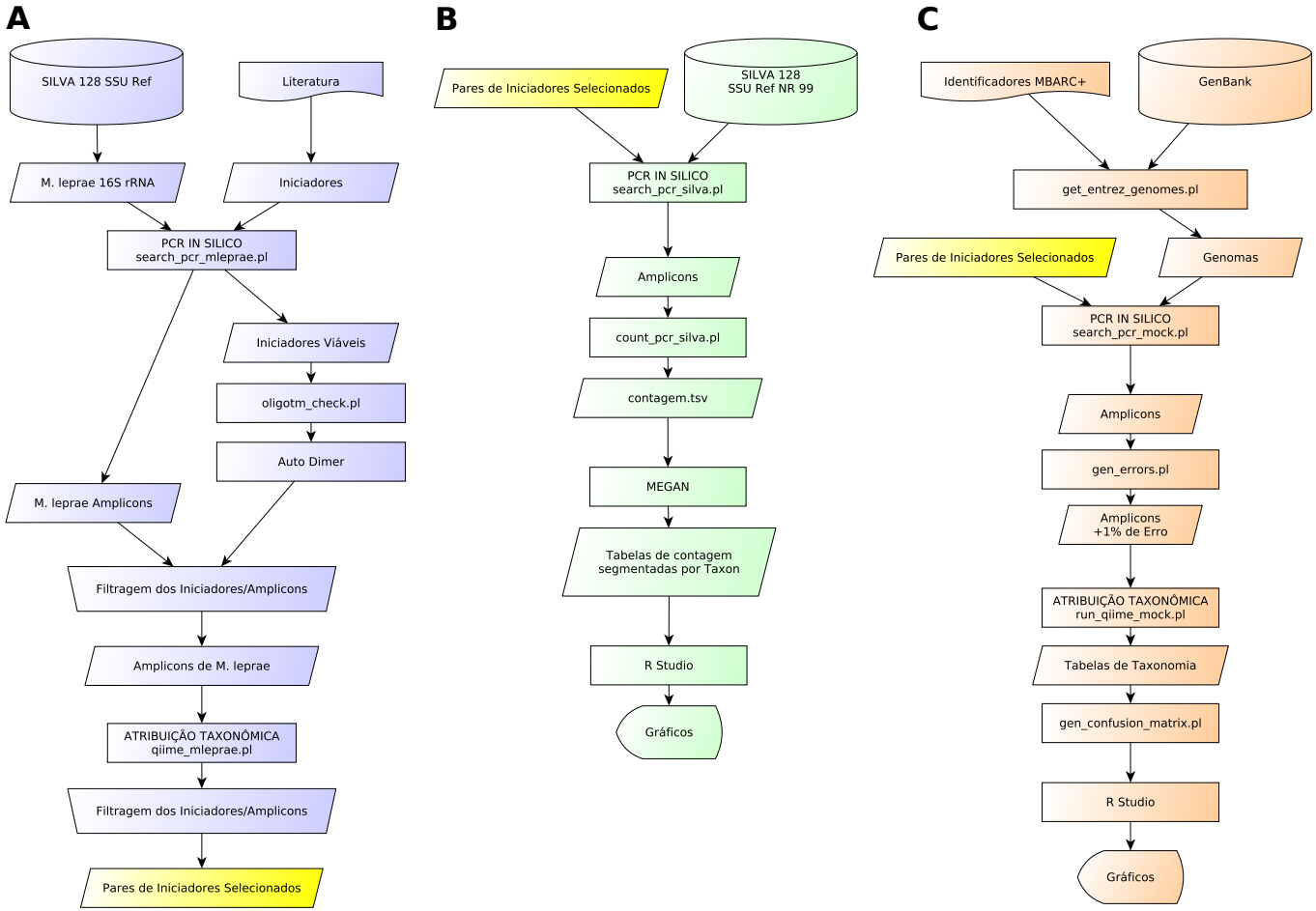
3 Métodos

O desenho experimental desse trabalho foi dividido em quatro partes. A primeira parte consiste na pré-seleção de iniciadores e seus respectivos amplicons simulados que possibilitam a classificação taxonômica em nível de espécie do organismo *M. leprae*, baseando-se na amplificação de diferentes regiões do gene rRNA 16S. A segunda parte compreende a avaliação dos iniciadores selecionados no passo anterior com base na eficiência da amplificação em um banco de dados referência. A terceira parte consiste na análise da sensibilidade na classificação taxonômica usando comunidades microbianas simuladas. Uma visão geral das três primeiras partes pode ser vista através da Figura 6. A quarta parte se trata da identificação das regiões hipervariáveis mais informativas do gene de rRNA 16S de *M. leprae* através de regressão logística.

Todos os arquivos e scripts desenvolvidos para a execução desse trabalho, e que serão citados a seguir, estão disponíveis com os mesmos nomes no repositório GITHUB[1].

[1] <https://github.com/felipeborim789/m_leprae>

Figura 6 – Fluxograma sistematizando as três primeiras partes da metodologia.



(A) Pré-seleção dos candidatos. (B) Avaliação da cobertura da amplificação dos iniciadores pré-selecionados no banco de dados Silva. (C) Avaliação da sensibilidade na classificação taxonômica utilizando comunidades microbianas simuladas.

3.1 Pré-seleção dos candidatos

Um conjunto de 44 iniciadores (22 *forward* e 22 *reverse*) foram coletados de uma coleção disponível na literatura. A construção dessa coleção de iniciadores foi feita a partir de uma busca na literatura de trabalhos que traziam experimentação dos mesmos. Posteriormente, todos os iniciadores escolhidos foram alinhados a conjuntos de dados construídos a partir do banco de dados GreenGenes (versão 11 de março 2011). Sendo esses conjuntos de dados representações de diversos ambientes, entre eles o microbioma intestinal humano, oceano, corais, solo, entre outros. Todos os iniciadores que não cobriram o mínimo de 40% de pelo menos um desses conjuntos de dados foram descartados. Grande parte dos descartes estavam relacionados a iniciadores específicos de Archeae ou que por algum motivo eram complementares a uma pequena fração das sequências (SOERGEL et al., 2012). Todos os iniciadores utilizados e suas respectivas sequências estão disponíveis no Apêndice A.

As sequências do gene rRNA 16S de *M. leprae* que foram escolhidas para essa primeira etapa são todas as oito sequências disponíveis no banco de dados Silva 128 SSU Ref[2]. O banco de dados Silva é um banco de dados curado de sequências de 16S rRNA onde todas as sequências disponíveis na versão SSU Ref possuem tamanho mínimo de 1200 nucleotídeos para Bacteria e Eukarya e mínimo de 900 nucleotídeos para Archaea, além de possuir filtragem por outros critérios de qualidade (QUAST et al., 2013).

PCR *in silico*

Para avaliar se os iniciadores selecionados poderiam amplificar as sequências de *M. leprae*, a reação da cadeia polimerase (PCR) foi simulada *in silico* através da ferramenta USEARCH (EDGAR, 2010). Essa ferramenta foi escolhida porque trabalha com degeneração dos iniciadores, em contraste com ferramentas baseadas em BLAST que durante o alinhamento de uma base degenerada com qualquer outra, reconhece esse alinhamento como *mismatch*. Isso foi constatado quando, a priori, a ferramenta Simulate_PCR havia sido testada para esse trabalho (GARDNER; SLEZAK, 2014).

A fim de testar todas as possíveis combinações de iniciadores, foi desenvolvido um script em Perl (search_pcr_mleprae.pl) que recebe como entrada um arquivo FASTA com todos os iniciadores e outro arquivo FASTA com as sequências de 16S de *M. leprae*. O script então utiliza um par de iniciadores por vez e executa a PCR in *silico* invocando o USEARCH com a *flag* -search_pcr. As outras opções definidas no USEARCH foram o número máximo de mismatches permitidos de 2, tamanho mínimo de nucleotídeos nos amplicons gerados de 30 e máximo de 550. Após a amplificação, foram selecionados todos

[2] Silva 128 SSU Ref <https://www.arb-silva.de/fileadmin/Silva_databases/release_128/Exports/Silva_128_SSURef_tax_Silva.fasta.gz>

os pares de iniciadores que geraram entre 7 a 8 amplicons, ou seja, que contemplaram pelo menos 87.5% das sequências selecionadas.

Filtragem dos iniciadores por atributos químicos

O primeiro atributo avaliado foi a temperatura de anelamento para cada par de iniciadores. Como critério de filtragem foi utilizada a diferença entre a temperatura de anelamento dos pares. O descarte foi feito quando encontrada uma diferença de temperatura entre os pares superior a 5º C. A predição da temperatura de anelamento dos oligonucleotídeos foi feita a partir da utilização da ferramenta oligotm, presente no pacote de ferramentas para desenho de iniciadores Primer3 (UNTERGASSER et al., 2012). Os parâmetros termodinâmicos para o calculo de temperatura de anelamento e a fórmula de correção de sais específica foram utilizados segundo SantaLucia (1998), conforme sugerido na ferramenta. A avaliação da diferença de temperatura e filtragem foi feita utilizando um script em Perl (oligotm_check.pl) que recebe como entrada um arquivo com os identificadores dos iniciadores, organizados em pares por linha, e outro com o arquivo FASTA contendo todas as sequências dos iniciadores.

Como segunda filtragem dos iniciadores, foi feita a predição de auto dimerização entre cada combinação de iniciadores através da ferramenta AutoDimer (VALLONE; BUTLER, 2004) para descarte dos dímeros. Foi utilizada a versão web da ferramenta[3]. Pelo fato dessa ferramenta não estar disponível em linha de comando, não foi possível automatizar essa análise através de script.

Atribuição taxonômica dos amplicons

A atribuição taxonômica foi feita para todos os amplicons originados pelos pares de iniciadores que não foram descartados nas etapas anteriores. Os amplicons gerados na etapa da PCR *in silico* foram submetidos ao pipeline do QIIME (CAPORASO et al., 2010).

Foi desenvolvido um script em Perl (qiime_mleprae.pl) que recebe como entrada um arquivo com uma lista contendo os identificadores das amostras (cada produto da PCR de cada par de iniciador) e uma pasta contendo cada arquivo FASTA correspondente aos identificadores. Esse script foi construído para que cada grupo de sequências, considerados como amostras independentes, fossem submetidos ao pipeline. Antes de executar o script, foi necessário fazer a conversão das sequências que estavam no formato de RNA (A/U/C/G) para DNA (A/T/C/G), tal conversão foi indispensável visto que o Silva, origem das sequências de rRNA 16S do primeiro passo desse trabalho, disponibiliza as sequências no formato de RNA.

[3] AutoDimer versão web <https://www-s.nist.gov/dnaAnalysis/primerToolsPage.do>

Em síntese, o pipeline do QIIME utilizado nesse trabalho consistiu na definição das OTUs através do algorítimo UCLUST (EDGAR, 2010) a partir de um banco de dados de OTUs de referência e em seguida faz a atribuição taxonômica dessas OTUs. Um limiar de 97% de similaridade foi utilizado na definição das OTUs e também na classificação taxonômica. Os seis diferentes bancos de dados de referência de OTUs utilizados foram as versões do GreenGenes: gg_29oct2010, gg_6nov2010, gg_4feb2011 e as versões do Silva: 111, 118 e 123. Apenas os bancos que continham OTUs com classificação taxonômica até nível de espécie para o organismo de interesse foram utilizados, justificando a ausência de versões mais recentes do banco de dados GreenGenes, que não possuíam nenhuma OTU de 97% com taxonomia *"Mycobacterium leprae"*.

Após a classificação taxonômica, foram selecionadas todas as amostras com taxonomia positiva para *M. leprae* e foram descartados todos os pares de iniciadores que não possibilitaram a classificação até nível de espécie para *M. leprae* em pelo menos um dos bancos de dados de referência utilizados. Para a execução dessa seleção, foi feita uma busca do termo *"Mycobacterium leprae"* em todos os arquivos "table.txt", um para cada amostra.

3.2 Avaliação da amplificação do banco de dados Silva

A PCR *in silico* foi utilizada através da ferramenta USEARCH, através da flag search_pcr, mas dessa vez com um número máximo de *mismatches* permitidos de 3. Foi construído um script em Perl (search_pcr_silva.pl) que usa todas as combinações de iniciadores selecionadas anteriormente e simula a amplificação do banco de dados Silva 128 SSU Ref NR 99 [4].

A versão NR (não redundante) do banco de dados Silva conserva apenas as sequências de maior tamanho quando são redundantes acima de 99% de similaridade. Logo, essa escolha teve como objetivo reduzir a chance de sequências com tamanhos incompletas não serem amplificadas.

Após a PCR *in silico*, foi utilizado um script desenvolvido em Perl (count_pcr_silva.pl) que recebe como entrada um arquivo com os cabeçalhos das sequências do Silva e um diretório contendo os amplicons. Esse script conta quais as sequências do banco de dados referência foram amplificadas e gera um arquivo tabular para ser carregado pela ferramenta MEGAN6. Essa ferramenta, que tem como função a analise interativa de dados metagenômicos (HUSON et al., 2007), foi utilizada aqui para gerar os arquivos de contagens de acordo com os táxons avaliados, no formato "csv" (valores separados por vírgulas). As contagens avaliadas foram as dos domínios Bacteria e Archaea. Em seguida, as tabelas

[4] <https://www.arb-silva.de/fileadmin/Silva_databases/release_128/Exports/Silva_128_SSURef_Nr99_tax_Silva.fasta.gz>

foram importadas no ambiente de desenvolvimento R Studio para o desenho dos gráficos.

3.3 Avaliação da sensibilidade da classificação taxonômica

A comunidade de referência utilizada foi a denominada MBARC-26 (*Mock Bacteria Archaea Community*). Os organismos escolhidos para sua composição foram derivados de diversos ambientes diferentes como humanos, água, solo, bovino e sapo. Esses possuem genomas com tamanho entre 1.8-6.5 Mb, conteúdo GC variando entre 28.4 e 72.7% e conteúdo repetitivo de 0-18.3%, compreendendo 10 filos e 14 classes, constituindo um perfil taxonômico diverso, totalizando 26 organismos (SINGER et al., 2016). Além dos 26 organismos, foi também incluído o genoma de referência de *M. leprae* (NC_002677.1). Como houve essa inclusão, portanto, nesse trabalho a comunidade simulada será mencionada como MBARC+.

Os 27 genomas de referência dos organismos foram baixados no formato FASTA através da ferramenta Entrez Direct[5] que facilita o acesso ao banco de dados GenBank. Um script desenvolvido em Perl (get_entrez_genomes.pl) recebe como entrada a lista de identificadores das sequências, baixa os genomas correspondentes e salva em um arquivo no formato FASTA. Uma tabela completa com os organismos pode ser visualizada no Apêndice B.

PCR *in silico* da comunidade microbiana simulada

Para a simulação dos amplicons, os genomas foram então submetidos a PCR *in silico* através da ferramenta USEARCH, com um máximo de *mismatches* permitidos de 3. Os pares de iniciadores utilizados nessa etapa foram apenas aqueles cujos amplicons da avaliação anterior tiveram taxonomia atribuída com todos os bancos de dados de OTUs de referência (6/6). O tamanho esperado dos amplicons para cada iniciador foi calculado com base nos tamanhos médios das sequências dos amplicons da primeira etapa, parte 1 dessa metodologia. Tendo esses valores em mãos, os valores de tamanho mínimo e máximo definidos na ferramenta foram o tamanho médio do amplicon menos 10 e mais 10 nucleotídeos respectivamente, de acordo com cada combinação de iniciadores.

Essa etapa foi executada através de um script desenvolvido em Perl para automatizar o processo (search_pcr_mock.pl) que recebe como entrada o arquivo FASTA dos iniciadores, o arquivo FASTA dos genomas e um arquivo de texto com as referências (identificador da amostra, identificador do iniciador *forward*, identificador do iniciador *reverse* e tamanho médio do amplicon).

[5] <https://www.ncbi.nlm.nih.gov/books/NBK179288/>

Depois da amplificação, cada arquivo FASTA contendo os amplicons foi replicado 2500 vezes com erros de sequenciamento simulados numa chance de 1 erro para cada 100 bases através de um script desenvolvido em Perl (gen_errors.pl) constituindo uma amostra para cada par de iniciadores. Esse script recebe como entrada o arquivo com os genomas e o um argumento com o número de vezes que se deseja replicar a saída.

Atribuição taxonômica aos amplicons

As amostras foram submetidas ao pipeline de atribuição taxonômica do QIIME, mas, dessa vez, sem o passo prévio da definição das OTUs, de modo que cada amplicon teve uma taxonomia atribuída individualmente para priorizar o cálculo de sensibilidade da classificação. No entanto, o mesmo método de clusterização por similaridade de 0.97% foi utilizado no passo da atribuição taxonômica.

A atribuição taxonômica foi feita com base em quatro bancos de dados de referência distintos, sendo esses duas versões do GreenGenes: gg_29nov2010, gg_4feb2011 e duas versões do Silva 118 e Silva 123, todos com OTUs construídas com 97% de similaridade. A versão 111 do Silva não foi utilizada nesse momento pois as taxonomias atreladas às sequências dessa versão não se encontravam em um formato definido de 7 níveis, muito menos indicavam os táxons de cada nível, o que inviabilizou o processamento dos resultados. A versão mais antiga do GreenGenes gg_16oct2010 também foi desconsiderada para diminuir a redundância de dados, já que é a mais antiga dos bancos recrutados e muito similar a versão gg_29nov2010. Para automatizar o pipeline do QIIME citado acima, um script foi desenvolvido em Perl (run_qiime_mock.pl).

Construção das matrizes de confusão

Cada banco de dados utilizado no passo anterior resultou em um arquivo tabular contendo os identificadores (amostras e amplicons) e as taxonomias atribuídas as mesmas. Esses arquivos foram processados e foram mantidas apenas das duas primeiras colunas. Em seguida, uma terceira coluna com os identificadores do NCBI foi introduzida ao arquivo. Como resultado, gerou-se um arquivo no formato tabular de três colunas no seguinte formato: identificador (amostra e amplicon), identificador do genoma de referência (NCBI accession) e taxonomia atribuída.

Para o cruzamento dos dados da atribuição taxonômica com os dados taxonômicos de referência presentes nos bancos de dados, uma busca manual foi feita dentro cada um dos bancos de OTUs utilizados e foram geradas tabelas de taxonomia de referência disponíveis no formato tabular de duas colunas, sendo a primeira coluna o identificador do genoma e a segunda a taxonomia por extenso (mbarc_tax.tar.gz).

Para reforçar a etapa anterior, cada taxonomia de referência foi checada com base nas respectivas taxonomias presentes no NCBI Taxonomy. Tendo em vista facili-

tar a consulta das taxonomias, um segundo script acessório foi desenvolvido em Perl (get_entrez_taxonomies.pl) que recebe como entrada a lista de identificadores da comunidade simulada e carrega as taxonomias como estão cadastradas no NCBI e faz montagem no formato de sete níveis (reino, filo, classe, ordem, família, gênero e espécie).

As tabelas de atribuição taxonômica foram processadas através de scripts desenvolvidos em Perl para gerar as matrizes de confusão (gen_confusion_matrix_gg.pl e gen_confusion_matrix_Silva.pl). Considerando-se que os formatos de taxonomia do Silva e do GreenGenes possuem algumas diferenças estruturais, foi indispensável o uso de dois scripts, um para cada. Ambos os scripts que geram as matrizes de confusão recebem como entrada um arquivo com a atribuição taxonômica no formato tabular de três colunas e um segundo arquivo com as taxonomias de referência no formato tabular de duas colunas.

Os quatro arquivos resultantes da etapa anterior foram concatenados em um único arquivo tabular. As taxonomias foram padronizadas nos seguintes formatos: "s___Mycobacterium leprae" para espécie e "g___Mycobacterium" para gênero. O arquivo então foi carregado através da linguagem de programação R e o cálculo de sensibilidade foi feito. Em seguida foi feito o desenho de dois gráficos, um para nível de gênero e o outro para espécie. A sensibilidade foi calculada segundo a equação a seguir, descrita por Baldi et al. (2000).

$$\text{Sensibilidade} = \frac{\text{Verdadeiro-positivos (VP)}}{\text{Verdadeiero-positivos (VP)} + \text{Falso-negativos (FN)}}$$

3.4 Identificação das regiões hipervariáveis informativas

Com o objetivo de identificar quais fragmentos da sequência do gene rRNA 16S de *M. leprae* são mais informativos para distinguir essa espécie de outras de mesmo gênero, foi construído um modelo classificador baseado em regressão logística, através da função *logit* ou *log-odds* (ZIEGLER, 2016).

Para a construção do modelo, foram utilizadas as sequências de *Mycobacterium spp.* presentes no banco de dados Silva versão 128[6]. Em seguida, as sequências de *Mycobacterium spp.* foram filtradas e aquelas que possuíam "N" em suas sequências de nucleotídeos ou sequências com taxonomias não informativas foram removidas (termos "*unclassified*", "*uncultured*" e "sp"), sem espécie atribuída. Das 8775 sequências de *Mycobacterium spp.* presentes no banco Silva, foram utilizadas 4946 após as filtragens, sendo 6 sequências de *M. leprae* e as demais 4940 sequências de *Mycobacterium spp.* (outas espécies desse gênero com exceção de *M. leprae*).

[6] <https://www.arb-silva.de/fileadmin/silva_databases/release_128/Exports/SILVA_128_SSURef_tax_silva.fasta.gz>

A construção do modelo foi feito através de um script na linguagem de programação R (regModel.R). O modelo de regressão logística foi construído a partir das frequências de janelas de nucleotídeos de tamanho 6 (matriz de 4096 colunas (*features*)), gerada pela biblioteca Biostrings (PAGES et al., 2017).

Para constatar a consistência dos resultados dessa análise, foi criado um modelo com cada uma das 6 sequências de *M. leprae*. Ou seja, o script gera um modelo com 1 das 6 sequências de *M. leprae* e as demais *Mycobacterium spp.*, repetindo esse passo 6 vezes. Para cada vez, logo após a construção do modelo, a sequência de *M. leprae* vigente é fragmentada em janelas de tamanho 100, aproximadamente 1400 janelas já que o tamanho médio do gene rRNA 16S é de 1500 nucleotídeos, e esses fragmentos são utilizados como *query* no modelo gerado. A seguir, um gráfico com as probabilidades de cada fragmento pertencer ao modelo foi gerado para cada uma das 6 sequências.

3.5 Seleção final

Tendo em mãos todos os resultados das etapas anteriores, foram avaliados quais os iniciadores, bancos de dados de OTUs de referência apresentaram os melhores resultados, ou seja, os maiores valores de sensibilidade e abrangência de organismos das comunidades simuladas. Partindo desse ponto, foi feita a recomendação das melhores características de um pipeline para caracterização de microbiomas, incluindo *M. leprae* em nível de espécie.

4 Resultados

4.1 Pré-seleção dos candidatos

A partir de todas as 484 combinações de iniciadores testadas para a amplificação simulada das sequências de rRNA 16S de *M. leprae*, aproximadamente metade (232) geraram amplicons válidos respeitando os critérios de tamanho mínimo de 30 e máximo de 550 nucleotídeos, e atingindo pelo menos 7 de 8 sequências do gene presentes em cada amostra (considerando que existiam 8 sequências de 16S de *M. leprae* no banco de dados).

Dos 232 iniciadores remanescentes, 96 foram considerados válidos após a análise das temperaturas de anelamento (ver Apêndice C) e autodimerização e os demais foram descartados. Entre os 96 remanescentes, apenas 37 pares de iniciadores originaram amplicons que foram atribuídos taxonomicamente até nível de espécie (*M. leprae)* em, pelo menos, um dos seis bancos de dados de OTUs de referência utilizados nesse passo.

4.2 Amplificação do banco de dados Silva 128 SSU Ref NR 99

A maioria dos iniciadores que foram selecionados apresentaram baixa ou nenhuma cobertura de Archaea no banco de dados Silva 128 SSU Ref NR 99. A figura 7 mostra a cobertura do banco de dados para cada um dos principais domínios que o gene de rRNA 16S é utilizado como marcador taxonômico: Bacteria e Archaea. Foi possível observar que os iniciadores tem maior afinidade com sequências do domínio Bacteria.

Dessa forma, a partir daqui analisamos a combinação de Bacteria e Archeae. Os pares de iniciadores que apresentaram a menor cobertura[1] (média aproximada de 30%) do banco de dados Silva (Bacteria e Archaea) foram aqueles que possuem os representantes 3'-5' (*forward*) que se ligam na posição 8 ou 9 da sequência alvo (E8Fa, E8Fb e E9F). Esses pares correspondem ao grupo que compreende as regiões hipervariáveis V1-V2 e V1-V3.

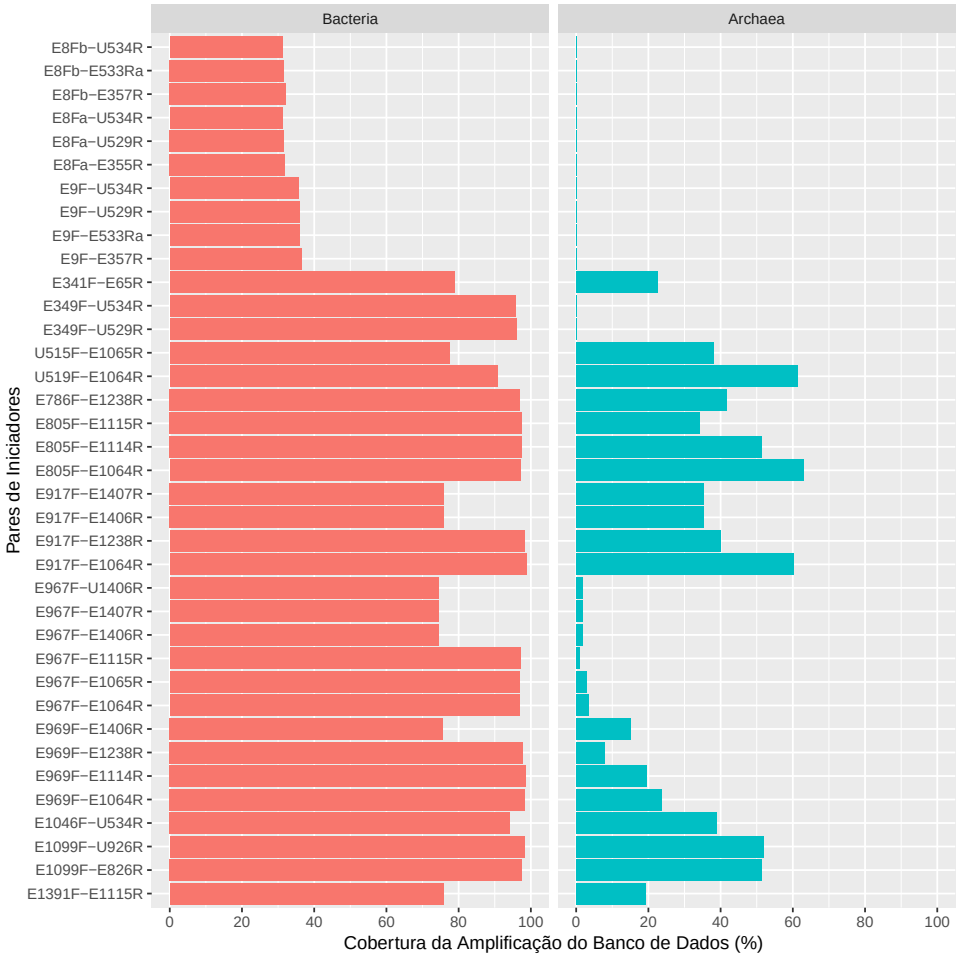
Como exceção em relação as regiões hipervariáveis compreendidas, o par de iniciadores E341F-E65R, que também se liga flanqueando a região V1-V2, teve uma cobertura do banco maior do que 75%. Porém, diferente dos anteriores, o iniciador *forward* se liga posição 341 e o reverse na posição 65.

Os demais iniciadores apresentaram cobertura maior do que 70%. Entre eles, pares de iniciadores das regiões V3, V4-V6, V5-V6, V6, V6-V7 e V6-V8. Entre os candidatos

[1] O termo cobertura utilizado nesse trabalho se refere à fração do banco de dados que os pares de iniciadores amplificaram *in silico*.

avaliados, apenas (10) os pares de iniciadores que cobrem as regiões hipervariáveis V1-V2, V2-V3 e V6 geraram amplicons que foram atribuídos taxonomicamente a OTUs de *M. leprae*. Uma síntese dos resultados descritos até essa parte pode ser visualizado na Figura 8.

Figura 7 – Gráfico da amplificação do banco de dados Silva 128 SSU NR 99 pelos pares de iniciadores selecionados, segmentado por domínios.

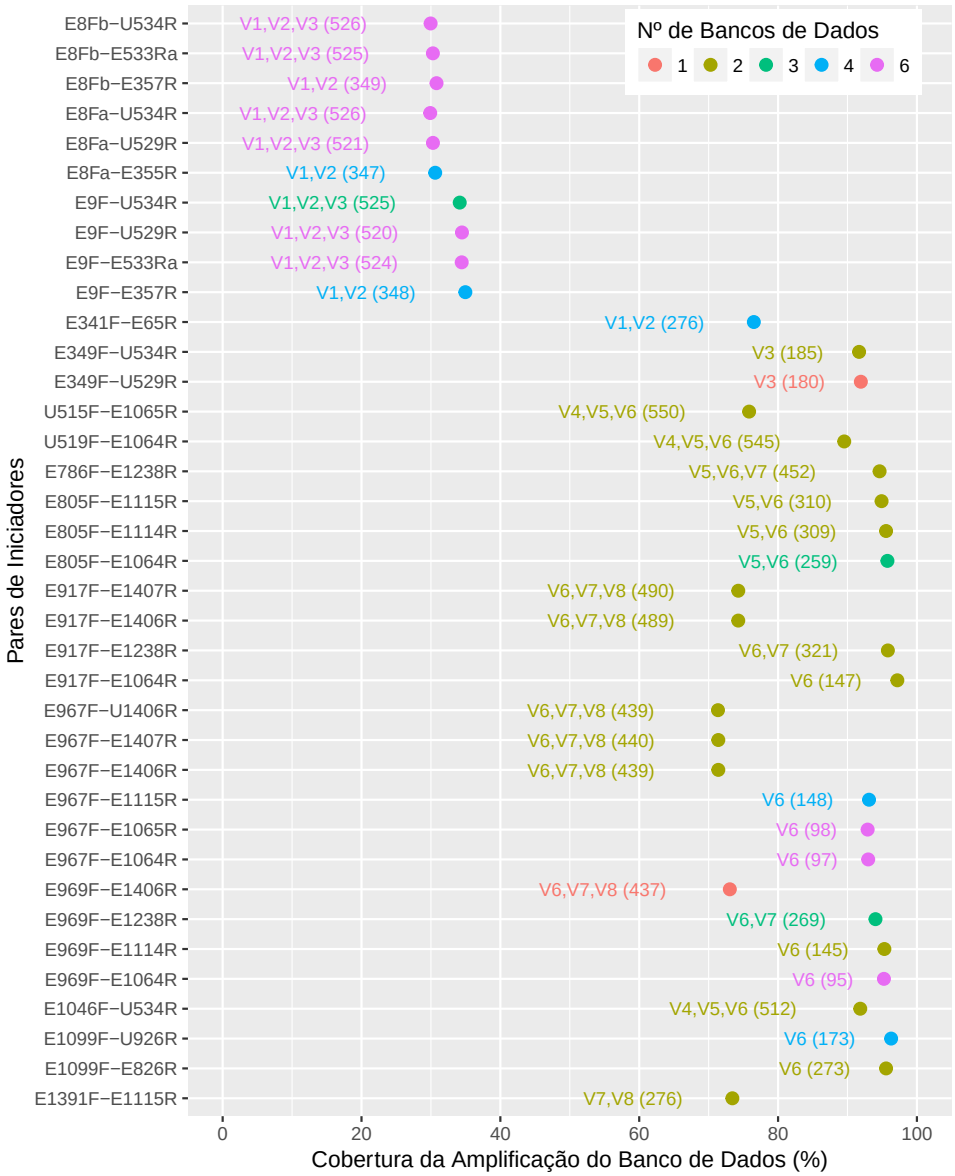


4.3 Sensibilidade da classificação taxonômica

Composição dos bancos de dados de referência

Inconsistências com relação a bancos de dados de OTUs distintos foram observadas. Não foi encontrada a presença da coleção completa correspondente aos organismos da MBARC+ em nenhum dos bancos de dados de OTUs. Temos que entre as versões do Greengenes analisadas, o banco gg_29nov2010 possuía 8 representantes e o banco gg_4feb2011 15 representantes. Por outro lado, Silva chegou mais próximo do total de epécies em MBARC+ (27). A versão do banco Silva 123 apresentou 21 e o Silva 118

Figura 8 – Gráfico da amplificação do banco de dados Silva 128 SSU NR 99 pelos pares de iniciadores selecionados de Bacteria e Archaea.



O número de bancos de dados representa em quantos bancos de dados de referência de OTUs (de 1 a 6) os amplicons foram classificados em nível de espécie (*M. leprae*). Ao lado dos pontos estão representadas quais as regiões hipervariáveis do gene rRNA 16S são cobertas pelos iniciadores e entre parênteses o tamanho esperado dos amplicons.

mostrou 25 das espécies, nesse último onde apenas a espécie *Nocardiopsis dassonvillei* e *Spirochaeta smaragdinae* não estavam inclusas. A única espécie que não tinha nenhuma OTU nos bancos de dados é a espécie *Spirochaeta smaragdinae*.

Em contrapartida, em nível taxonômico de gênero, os bancos apresentaram uma maior relação na presença/ausência nos bancos de dados de OTUs. O banco gg_29nov2010 mostrou 21 representantes e o banco gg_4feb2011 24 representantes. O banco Silva 118 apresentou 26 e no Silva 123 nenhum gênero ficou de fora. De fato, OTUs tanto de

Mycobacterium spp. quanto de *M. leprae* estavam presentes em todos os bancos de dados que foram utilizados. A tabela completa com a presença ou ausência pode ser observada através da Figura 1.

Tabela 1 – Comparação entre os bancos de dados de OTUs de 97% de similaridade com relação a presença (+) ou ausência (-) dos táxons de MBARC+.

	GG 29nov2010		GG 4feb2011		Silva 118		Silva 123	
Organismo	Gênero	Espécie	Gênero	Espécie	Gênero	Espécie	Gênero	Espécie
Terriglobus roseus	+	+	+	+	+	+	+	+
Corynebacterium glutamicum	+	-	+	-	+	+	+	+
Nocardiopsis dassonvillei	+	-	+	+	+	-	+	-
Olsenella uli	+	-	+	+	+	+	+	-
Segniliparus rotundus	-	-	+	-	+	+	+	-
Echinicola vietnamensis	+	-	+	-	+	+	+	+
Meiothermus Silvanus	+	-	+	+	+	+	+	+
Clostridium perfringens	+	+	+	+	+	+	+	+
Clostridium thermocellum	+	+	+	+	+	+	+	+
Desulfosporosinus acidiphilus	-	-	-	-	+	+	+	+
Desulfosporosinus meridiei	+	+	+	+	+	+	+	+
Desulfotomaculum gibsoniae	+	+	+	+	+	+	+	+
Streptococcus pyogenes	+	-	+	+	+	+	+	+
Thermobacillus composti	+	-	+	-	+	+	+	+
Escherichia coli	+	-	+	-	+	+	+	+
Frateuria aurantia	+	-	+	+	+	+	+	+
Hirschia baltica	+	+	+	+	+	+	+	+
Pseudomonas stutzeri	+	+	+	+	+	+	+	+
Salmonella bongori	-	-	+	+	+	+	+	+
Salmonella enterica	-	-	+	+	+	+	+	+
Spirochaeta smaragdinae	-	-	-	-	-	-	+	-
Fervidobacterium pennivorans	+	-	+	-	+	+	+	-
Coraliomargarita akajimensis	-	-	-	-	+	+	+	+
Halovivax ruber	+	-	+	-	+	+	+	-
Natronobacterium gregoryi	+	-	+	-	+	+	+	+
Natronococcus occultus	+	-	+	-	+	+	+	+
Mycobacterium leprae	+	+	+	+	+	+	+	+
Total de presentes	21	8	24	15	26	25	27	21

A presença/ausência se refere à representatividade dos táxons em OTUs e não necessariamente às sequências presentes nos bancos de dados.

Classificação taxonômica

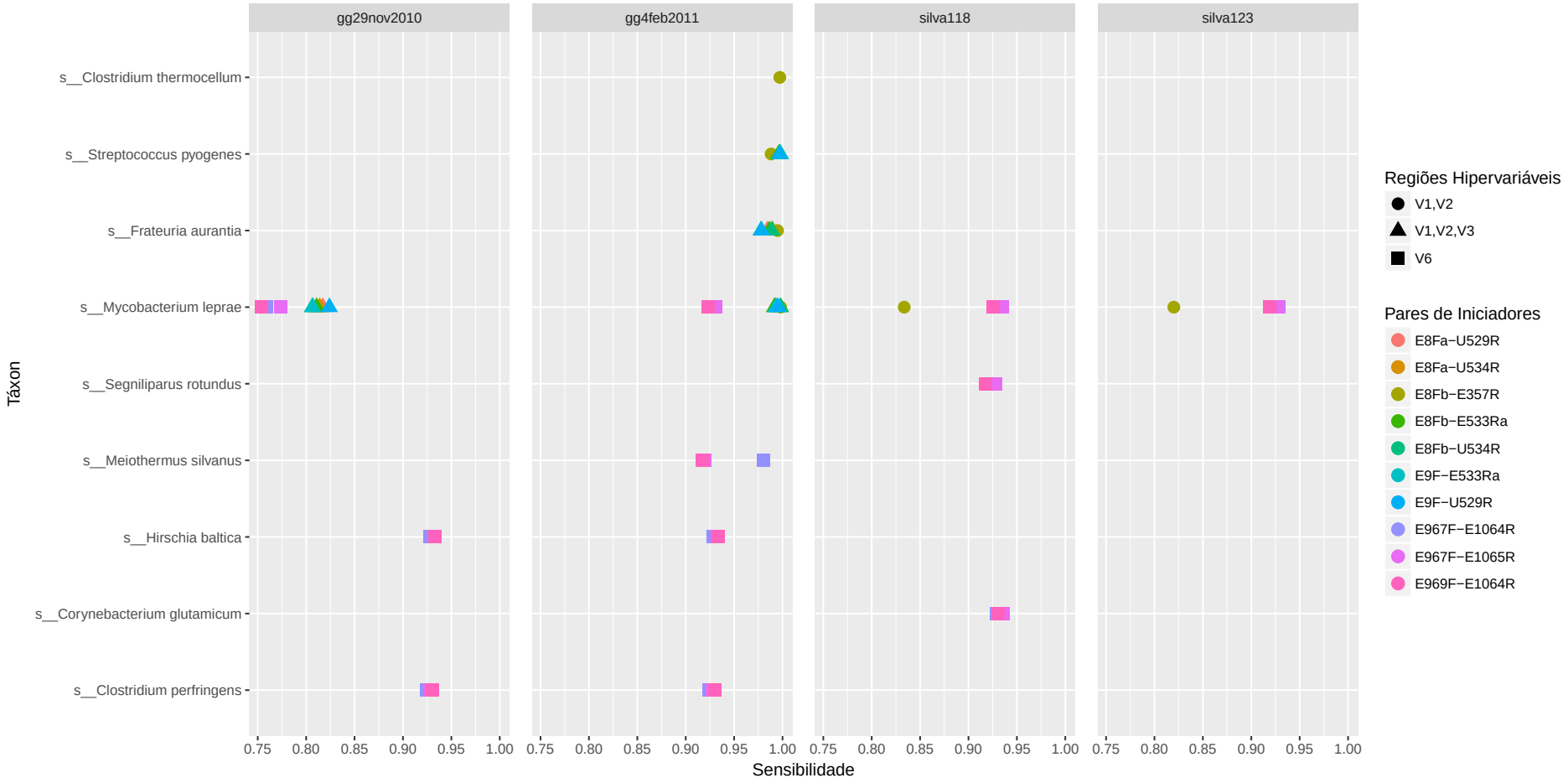
Com sensibilidade acima de 0.75 na classificação taxonômica, foi observado que menos da metade das espécies puderam ser classificadas até nível de espécie. Se considerarmos o resultado geral, com todos os iniciadores e bancos de dados de OTUs, 9 das 27 (33.3%) espécies foram classificadas. Num ponto de vista individual, ou seja, para cada banco de dados, foi possível perceber que *Hirschia baltica* e *Clostridium perfrigens* foram classificadas apenas nos bancos de dados GreenGenes.

Em uma visão mais restrita, vimos que as espécies *Clostridium thermocellum*, *Streptococcus pyogenes* e *Frateuria aurantia* só foram classificados através do banco GreenGenes versão 4feb2011 por iniciadores V1-V2 ou V1-V3. Também houveram outras peculiaridades, como exemplo a classificação de *Segniliparus rotundus* acima de 0.75 de sensibilidade que só apareceu no banco de dados Silva 118.

Assim como era esperado, a classificação de *M. leprae* em nível de espécie foi observada em todos os bancos de dados, no entanto, a sensibilidade variou bastante de acordo com os iniciadores utilizados e também conforme os bancos de dados de OTUs de referência. Com base em *M. leprae*, os bancos GreenGenes apresentaram maior sensibilidade em iniciadores V1-V3, já em Silva, os iniciadores V6 apresentaram maiores valores de sensibilidade (ver Figura 9).

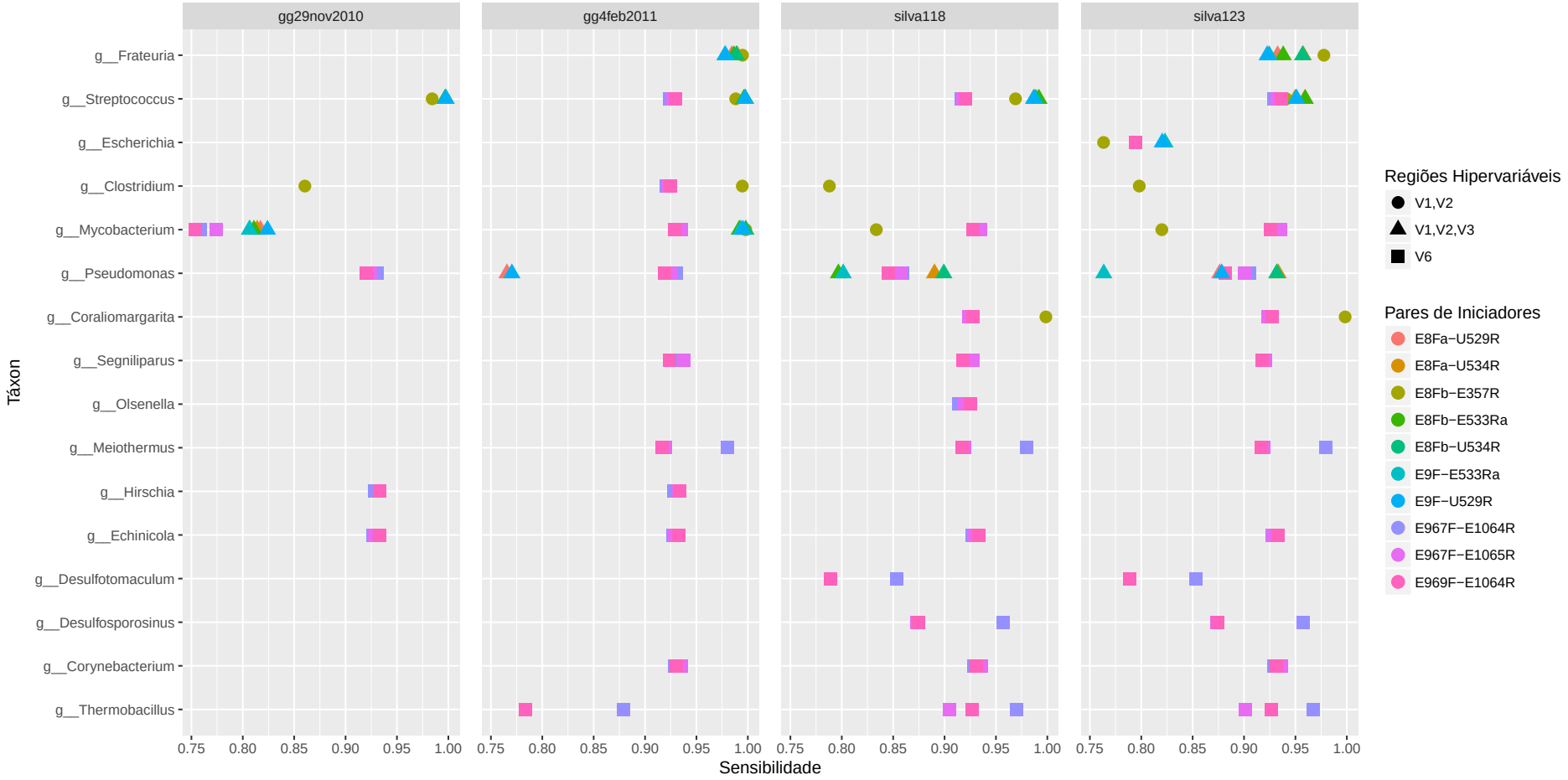
Com relação a classificação em nível de gênero observou-se um perfil bastante diferente. Numa visão geral, entre todos os iniciadores e bancos de dados de OTUs, 16 gêneros foram classificadas com sensibilidade acima de 0.75. Olhando individualmente cada banco de dados, foi observado que o banco de dados GreenGenes 29nov2010 classificou 6 gêneros e a versão 4feb2011 classificou 11. Em Silva observou-se que foram classificados 13 gêneros na versão 118 e 14 gêneros na versão 123. Através da Figura 10 é possível ver que diversos gêneros foram classificados através de iniciadores V6, não foram classificados através de iniciadores V1-V3 ou V1-V2.

Figura 9 – Gráfico da sensibilidade na classificação taxonômica dos amplicons gerados a partir da comunidade simulada MBARC+ por par de iniciadores em nível de espécie.



A classificação dos amplicons foi feita usando os bancos de dados de referência de OTUs gg29nov2010, gg4feb2011, Silva118 e Silva123, cada um representado em uma coluna do gráfico.

Figura 10 – Gráfico da sensibilidade na classificação taxonômica dos amplicons gerados a partir da comunidade simulada MBARC+ por par de iniciadores em nível de gênero.



A classificação dos amplicons foi feita usando os bancos de dados de referência de OTUs gg29nov2010, gg4feb2011, Silva118 e Silva123, cada um representado em uma coluna do gráfico.

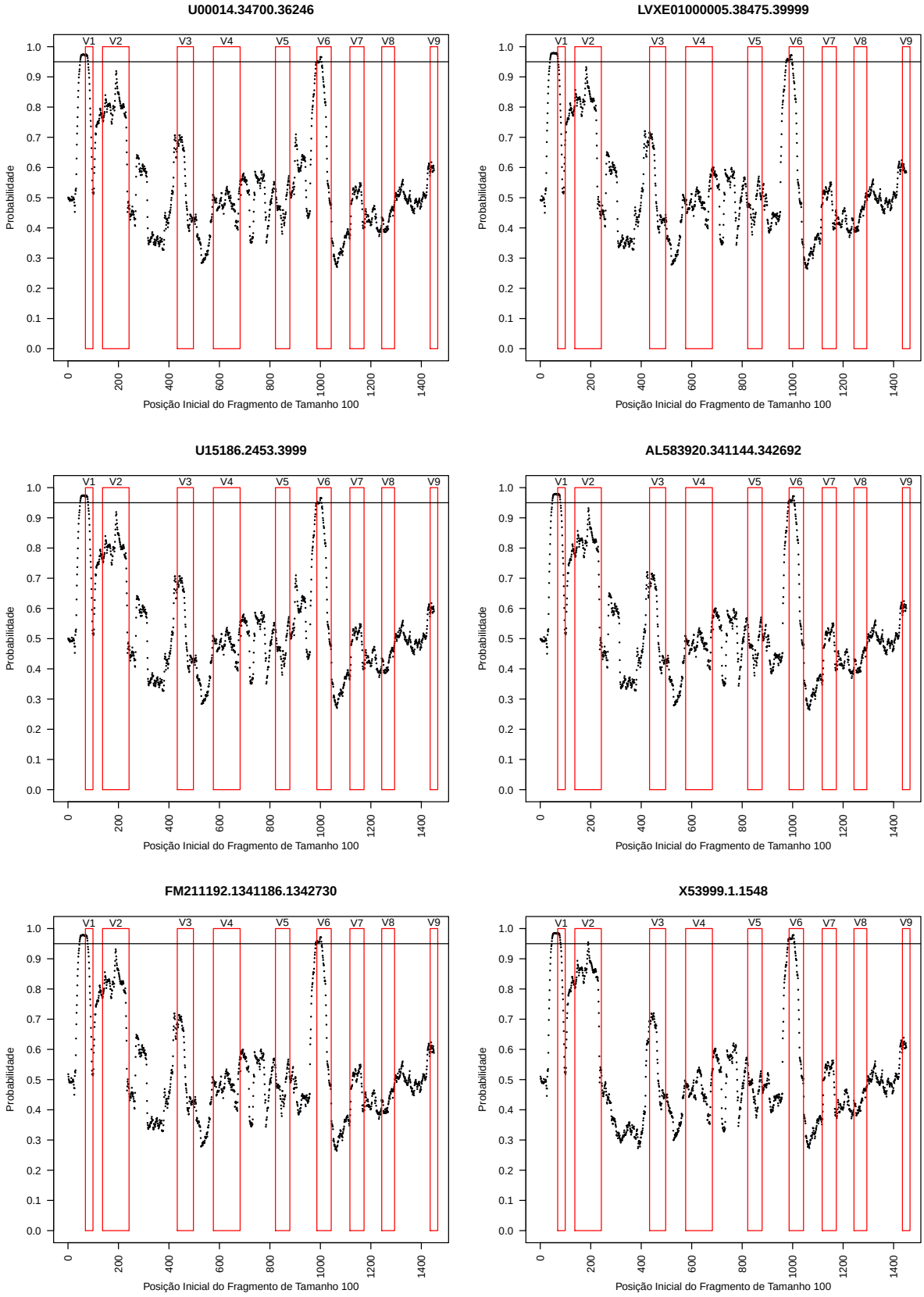
4.4 Identificação das regiões hipervariáveis informativas

Com base nos resultados observados nos modelos regressão logística construídos foi possível identificar que as regiões hipervariáveis do gene do rRNA 16S de *M. leprae* que apresentaram fragmentos mais informativas para diferenciar as sequências desses organismos de outros do mesmo gênero (*Mycobacterium spp.*) são as regiões hipervariáveis V1, V2 e V6 (ver Figura 11), para fragmentos de tamanho 100. Esse fato corrobora com o observado anteriormente nas análise da classificação taxonômica das comunidades simuladas, onde apenas iniciadores que flanqueavam essas regiões apresentaram maior sensibilidade na classificação de *M. leprae*.

Resultados finais

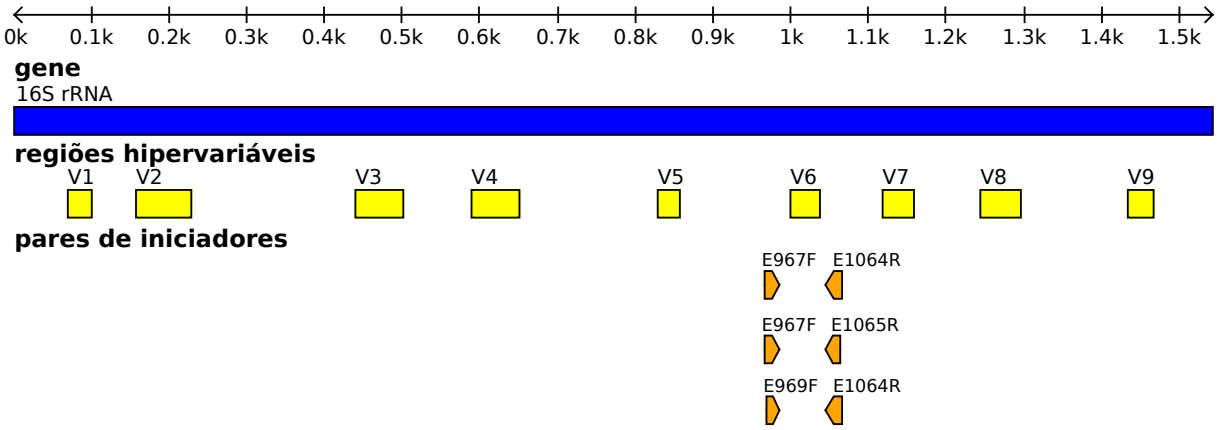
A partir dos resultados observados, foram escolhidas algumas características essenciais do pipeline para classificação taxonômica de *M. leprae*. Sendo assim fica definido que os bancos de dados de OTUs de referência Silva 118 e Silva 123 são indicados. Além disso fica sugerida a utilização dos pares de iniciadores que flanqueiam a região hipervariável V6 do rRNA 16S (ver Figura 12). O método de definição das OTUs e classificação taxonômica utilizado foi o UCLUST integrado ao QIIME com 97% de similaridade.

Figura 11 – Gráficos das probabilidades das *queries* nos modelos do gene rRNA 16S de *M. leprae*.



Cada gráfico representa um modelo construído com uma das seis sequências de *M. leprae*.

Figura 12 – Representação gráfica do gene rRNA 16S e dos pares de iniciadores escolhidos.



Em amarelo estão representadas as regiões hipervariáveis do gene mostrando que os pares de iniciadores selecionados, em laranja, flanqueiam a região hipervariável V6. As sequências desses iniciadores estão disponíveis no Apêndice A.

5 Discussão

O primeiro fato que merece destaque durante o desenvolvimento desse trabalho foi durante a reunião dos bancos de dados de OTUs de referência para a classificação taxonômica de *M. leprae*. Foi observado que as versões mais recentes do banco de dados GreenGenes[1] gg_12_8, gg_12_10 e gg_13_5, e também a versão gg_13_8 que está disponível como padrão na ferramenta QIIME versão 1.9.1, não possuem OTUs de *M. leprae*, nem sequências desse organismo nos bancos de dados FASTA. Dessa forma, o presente trabalho ficou limitado ao uso das versões mais antigas desse banco. Tal fato mostra como é importante conhecer o banco de dados de referência que está sendo utilizado antes da atribuição taxonômica das OTUs, já que algumas ferramentas como o QIIME já vem com um banco de dados de referência de OTUs padrão.

O banco de dados de OTUs de referência utilizado foi um dos motivos pelos quais seria impossível classificar até nível de espécie as possíveis *reads* de *M. leprae* presentes nas amostras do trabalho de Silva et al. (2015). Nesse trabalho, que foi o primeiro a apresentar a caracterização taxonômica de lesões cutâneas causadas por *M. leprae*, os autores utilizaram o banco de dados gg_13_5 (versão de Agosto de 2015), padrão do QIIME 1.9, que não possui OTUs representativas de *M. leprae*. Além disso, foi utilizado o método "*Pick Closed Reference OTUs*" na etapa de definição das OTUs. Esse método só consegue definir OTUs baseando-se na similaridade das sequências dos amplicons com OTUs já pré-definidas em um banco de dados, dessa forma, poderia ocorrer o descarte (atribuição como "Unassigned") das sequências que não compartilham similaridade dentro do *threshold* definido.

No mais, os autores utilizaram iniciadores que flanqueiam as regiões hipervariáveis V3-V4 do gene rRNA 16S. No presente trabalho foi visto que essa região hipervaríavel não poderia conferir a informação taxonômica suficiente para distinguir *M. leprae*. Esse fato se torna um problema quando o objetivo de uma análise hipotética seria identificar *M. leprae* em ambientes que não é possível esperar a presença de *M. leprae*, como água e solo, por exemplo. Como foi visto que no presente trabalho que essas regiões hipervariáveis utilizadas por Silva et al. (2015) não são tão informativas quanto as regiões V6 e V1-V2 para a classificação de *M. leprae*, o uso de iniciadores que flanqueiam as regiões V3-V4 talvez devessem ser evitados. Portanto, essas seriam algumas possíveis variáveis adversas no trabalho de Silva et al. (2015) já que os autores correlacionaram a sub representação ou ausência de *Mycobacterium spp.* com o método de fixação em formol das amostras, passo feito antes da preparação das bibliotecas de sequenciamento.

[1] <http://greengenes.secondgenome.com>

Outro fato que chamou bastante a atenção nos resultados do presente trabalho foi a baixa cobertura na amplificação do banco de dados Silva observada para os iniciadores que se ligam no início da sequência do gene rRNA 16S (antes da posição 10). A baixa amplificação corrobora com o que foi mostrado por Martinez-Porchas et al. (2017). Nesse artigo foi observado que as primeiras e últimas regiões conservadas (1 e 10) do gene rRNA 16S apresentam menos de 40% de conservação, enquanto as demais possuem aproximadamente mais de 75%. Esse estudo foi feito *in silico* através de uma análise com as sequências de 16S rRNA presentes no banco de dados Silva. Isto é, possivelmente grande parte das sequências presentes em uma amostra ambiental não serão complementares aos iniciadores que tem essas regiões como alvo e portanto poderiam subestimar os resultados antes mesmo do sequenciamento de DNA.

Dessa forma, a ciência de quais regiões os iniciadores escolhidos se ligam pode ser bastante importante no desenho experimental de estudos de metagenômica baseados em genes marcadores e PCR. A página do banco de dados Silva oferece uma ferramenta web denominada TestPrime[2] que avalia diversos aspectos dos iniciadores utilizando seu próprio banco como referência. Entre esses aspectos, a cobertura do banco de dados pode ser avaliada inserindo a sequência dos iniciadores (KLINDWORTH et al., 2013). A cobertura dos iniciadores no banco de dados Ribosomal Database Project (COLE et al., 2009) também pode ser avaliada pela ferramenta web Probe Match [3]. Ambas as ferramentas devem apresentar resultados semelhantes, mas não iguais, devido a diferente composição dos bancos de dados.

Apesar de já ter sido sugerido na literatura que a região V2 parece ser bastante informativa para separar organismos do gênero *Mycobacterium* (CHAKRAVORTY et al., 2007), no presente trabalho foi observado que, quando se trata de *M. leprae*, a região V1 junto com a V2 pode ser uma combinação mais eficiente para a identificação desse organismo em metagenômica. Porém, a região V6 também foi identificada como bastante informativa e em comparação com os demais iniciadores avaliados foi a que obteve melhor resultado na sensibilidade da classificação de *M. leprae* com o banco de dados Silva e também, os iniciadores cobriram muito mais sequências do banco de dados na avaliação da amplificação.

Iniciadores que flanqueiam a região V6 do rRNA 16S já foram sugeridos para o estudo de perfis de microbiomas pois permitem a geração *short reads* bastante informativos, sendo que também foi observada a alta capacidade de amplificação desses iniciadores (GLOOR et al., 2010). Todavia, foi mostrado *in silico* que a combinação de iniciadores que geram flanqueiam apenas a região V6 podem superestimar a diversidade microbiana observada (SUN et al., 2013). Anteriormente essa região também foi avaliada em um

[2] <https://www.arb-silva.de/search/testprime/>
[3] <https://rdp.cme.msu.edu/probematch/search.jsp>

experimento com pirosequenciamento onde também foi observada a diversidade super estimada. Essas discrepâncias na classificação taxonômica foram relacionadas com a quantidade de regiões variáveis, hipervariáveis e conservadas dentro do fragmento de 16S rRNA utilizado como alvo no sequenciamento (YOUSSEF et al., 2009). Ou seja, dependendo da região amplificada é possível agrupar mais ou menos sequências, de acordo com a quantidade de variabilidade nessa região, através dos métodos *clustering* em OTUs. Isso reforça o fato de que a padronização das metodologias é importante quando se trata de análises comparativas em metagenômica.

A avaliação da sensibilidade na classificação taxonômica das comunidades simuladas, dentro das condições experimentais desse trabalho, mostrou que o algorítimo de classificação utilizado (UCLUST) não conseguiu classificar, ou apresentou baixa sensibilidade para a maioria das espécies da comunidade MBARC+, muito embora tenha apresentado resultados satisfatórios para a classificação de sequências de *M. leprae*. De fato, o UCLUST foi utilizado nesse trabalho por ser o algoritmo padrão na ferramenta QIIME. Dessa forma, sugiro que cientistas utilizem diferentes algorítimos para classificação taxonômica. Também é importante notar que alguns classificadores, como o RDP Classifier, por exemplo, não classificam até espécie ou em um nível taxonômico menor (WANG et al., 2007), sendo esse detalhe bastante importante para o presente estudo.

Alguns classificadores vem sendo desenvolvidos com o objetivo de alcançar uma resolução mais profunda, e para isso não necessariamente requerem o agrupamento das sequências em OTUs como primeiro passo. A classificação taxonômica pode ser feita diretamente nos amplicons (CHAUDHARY et al., 2015), ou a interpretação dos dados metagenômicos pode ir além, baseando-se em pequenas diferenças dentro de grupos de sequências muito semelhantes (TIKHONOV; LEACH; WINGREEN, 2015; EREN et al., 2013; EREN et al., 2015).

De fato, o agrupamento em OTUs dentro de um limiar de 97% de similaridade é bastante arbitrário. Talvez essa abordagem, que é baseada em alinhamento, não seja a mais fiel para a classificação taxonômica pois as variações pontuais de nucleotídeos são todas transformadas em um percentual de similaridade, independente da posição e troca. Esse viés abre espaço para outros métodos de classificação independentes de alinhamento, onde cada carácter e sua posição conta, e não apenas o percentual de similaridade é levado em consideração.

A partir dessa ideia, junto com o grupo de pesquisas que participo, estamos desenvolvendo um classificador de dados de metagenômica baseados no gene 16S rRNA para Prokaryota, que tem como objetivo classificar até nível de espécie as sequências que possuem referência em banco de dados. Para isso, estamos utilizando o algoritmo de aprendizado de máquina *Random Forest* e geramos um conjunto de dados de treino para cada nível taxonômico, utilizando como entrada os amplicons simulados por iniciadores já

conhecidos.

6 Conclusão

Baseado nos achados desse trabalho, foi possível chegar em algumas recomendações para um *pipeline* que possibilite estudos de *M. leprae* através de metagenômica. As recomendações se iniciam a partir do uso de iniciadores que flanqueiam a região hipervariável V6 do gene rRNA 16S, o uso do banco de dados de referência de OTUs Silva e a classificação taxonômica dos amplicons através do QIIME com o algorítmo UCLUST utilizando-se um limiar de 97% de similaridade.

Este trabalho pôde fornecer algumas informações essenciais de como identificar regiões de rRNA 16S muito informativas para a classificação taxonômica em nível de espécie de *M. leprae*, além de evidenciar diversos vieses relacionados a amplificação de diferentes regiões hipervariávies do 16S. Considera-se ainda que cientistas usem diferentes parâmetros ou algorítmos para classificação taxonômica, mas a padronização do *pipeline* de modo geral é imprescindível para análises comparativas.

Por fim, o presente trabalho me faz acreditar que um experimento inteiramente *in silico* como esse não consegue alcançar ou avaliar todas as possíveis variáveis que um experimento que utiliza de um *workflow* completo poderia, e isso inclui desde a extração do DNA até a análise dos dados de sequenciamento. No entanto, é sabido que a cultura *in vitro* de *M. leprae* é inviável, e esse fato ocorre com a maioria dos microrganismos que conhecemos, portanto a execução de um *workflow* completo se torna ainda mais difícil. Dessa forma, um trabalho de natureza *in silico* é muito mais acessível e pode fornecer um conhecimento valioso, assim como o que foi aqui constatado.

Apêndices

APÊNDICE A – Iniciadores

Tabela 2 – Lista de iniciadores 3' "*forward*"

Identificador	Sequência 5'-3'	Tamanho	E. coli 5'	E. coli 3'
E8Fa	AGAGTTTGATCCTGGCTCAG	20	8	27
E8Fb	AGAGTTTGATCMTGGCTCAG	20	8	27
E9F	GAGTTTGATCCTGGCTCAG	19	9	27
E334F	CCAGACTCCTACGGGAGGCAGC	22	334	355
E338F	ACTCCTACGGGAGGCAGC	18	338	355
E341F	CCTACGGGNGGCNGCA	16	341	356
U341F	CCTACGGGRSGCAGCAG	17	341	357
E343F	TACGGRAGGCAGCAG	15	343	357
E349F	AGGCAGCAGTGGGGAAT	17	349	365
U515F	GTGCCAGCMGCCGCGGTAA	19	515	533
E517F	GCCAGCAGCCGCGGTAA	17	517	533
U519F	CAGCMGCCGCGGTAATWC	18	519	536
E786F	GATTAGATACCCTGGTAG	18	786	803
Eb787F	ATTAGATACCCTGGTA	16	787	802
E805F	GGATTAGATACCCTGGTAGTC	17	805	821
E917F	GAATTGACGGGGRCCC	16	917	932
E967F	CAACGCGAAGAACCTTACC	19	967	985
E969F	ACGCGARRAACCTTACC	17	969	985
E1046F	AGGTGCTGCATGGCTGT	16	1046	1061
U1053F	GCATGGCYGYCGTCAG	16	1053	1068
E1099F	GYAACGAGCGCAACCC	16	1099	1114
E1391F	TGTACACACCGCCCGTC	17	1391	1407

As colunas *E. coli* 5' e 3' se referem onde os iniciadores se alinham no gene de rRNA 16S de *E. coli*. Fonte: Soergel et al. (2012)

Tabela 3 – Lista de iniciadores 3' "*reverse*"

Identificador	Sequência 5'-3'	Tamanho	E. coli 5'	E. coli 3'
E65R	TCGACTTGCATGTRTTA	17	49	65
E355R	GCTGCCTCCCGTAGGAGT	15	341	355
E357R	CTGCTGCCTYCCGTA	15	343	357
U529R	ACCGCGGCKGCTGGC	15	515	529
E533Ra	TNACCGNNNCTNCTGGCAC	19	515	533
E533Rb	TTACCGCGGCTGCTGGCAC	19	515	533
E534R	ATTACCGCGGCTGCTGGC	18	517	534
U534R	GWATTACCGCGGCKGCTG	18	517	534
E826R	GACTACCAGGGTATCTAATCC	15	812	826
E926Ra	CCGNCNATTNNTTTNAGTTT	20	907	926
U926R	CCGTCAATTCCTTTRAGTTT	20	907	926
E926Rb	CCGTCAATTYYTTTRAGTTT	20	907	926
E939R	CTTGTGCGGGCCCCCGTCAATTC	23	917	939
E1064R	CGACARCCATGCASCACCT	19	1046	1064
E1065R	ACAGCCATGCAGCACCT	19	1047	1065
E1114R	GGGTTGCGCTCGTTRC	16	1099	1114
E1115R	AGGGTTGCGCTCGTTG	16	1100	1115
E1238R	GTAGCRCGTGTGTMGCCC	18	1221	1238
U1406R	GACGGGCGGTGTGTRCA	17	1390	1406
E1406R	GACGGGCGGTGWGTRCA	17	1390	1406
E1407R	GACGGGCGGTGTGTRC	16	1392	1407
E1492R	ACCTTGTTACGACTT	15	1478	1492

As colunas *E. coli* 5' e 3' se referem onde os iniciadores se alinham no gene de rRNA 16S de *E. coli*. Fonte: Soergel et al. (2012)

APÊNDICE B – Organismos da comunidade microbiana simulada

Tabela 4 – Lista de organismos presentes na MBARC+ e respectivos identificadores do GenBank

Organismo	**Identicador**
Terriglobus roseus	NC_018014.1
Corynebacterium glutamicum	NC_003450.3
Nocardiopsis dassonvillei	NC_014211.1
Olsenella uli	NC_014363.1
Segniliparus rotundus	NC_014168.1
Echinicola vietnamensis	NC_019904.1
Meiothermus Silvanus	NC_014212.1
Clostridium perfringens	NC_008261.1
Clostridium thermocellum	NC_009012.1
Desulfosporosinus acidiphilus	NC_018068.1
Desulfosporosinus meridiei	NC_018515.1
Desulfotomaculum gibsoniae	NC_021184.1
Streptococcus pyogenes	NC_002737.2
Thermobacillus composti	NC_019897.1
Escherichia coli	NC_000913.3
Frateuria aurantia	NC_017033.1
Hirschia baltica	NC_012982.1
Pseudomonas stutzeri	NC_019936.1
Salmonella bongori	NC_015761.1
Salmonella enterica	NC_010067.1
Spirochaeta smaragdinae	NC_014364.1
Fervidobacterium pennivorans	NC_017095.1
Coraliomargarita akajimensis	NC_014008.1
Halovivax ruber	CP003050.1
Natronobacterium gregoryi	NC_019792.1
Natronococcus occultus	NC_019974.1
Mycobacterium leprae	NC_002677.1

APÊNDICE C – Predição da temperatura de anelamento

Tabela 5 – Lista de iniciadores e temperaturas de anelamento

Iniciador 5'	**T (°C)**	**Iniciador 3'**	**T (°C)**	**Diferença (°C)**
U519F	46.73	E826R	48.20	1.48
U519F	46.73	E357R	44.62	2.10
U519F	46.73	E1064R	49.85	3.12
U515F	54.94	E355R	54.95	0.01
U515F	54.94	E1065R	53.36	1.58
U34IF	46.87	U534R	48.51	1.64
U34IF	46.87	U529R	51.00	4.13
U34IF	46.87	E826R	48.20	1.33
U34IF	46.87	E533Ra	44.88	1.99
U1053F	41.32	E926Rb	40.02	1.30
U1053F	41.32	E1238R	45.78	4.45
Eb787F	36.67	E926Rb	40.02	3.35
Eb787F	36.67	E926Ra	35.86	0.81
E9F	49.61	U534R	48.51	1.10
E9F	49.61	U529R	51.00	1.39
E9F	49.61	E533Ra	44.88	4.73
E9F	49.61	E357R	44.62	4.98
E969F	46.15	U926R	47.20	1.05
E969F	46.15	U534R	48.51	2.36
E969F	46.15	U529R	51.00	4.85
E969F	46.15	E826R	48.20	2.05
E969F	46.15	E533Ra	44.88	1.27
E969F	46.15	E1406R	49.04	2.90
E969F	46.15	E1238R	45.78	0.37
E969F	46.15	E1114R	49.02	2.88
E969F	46.15	E1064R	49.85	3.70
E967F	51.11	U926R	47.20	3.91
E967F	51.11	U534R	48.51	2.60
E967F	51.11	U529R	51.00	0.11
E967F	51.11	U1406R	52.71	1.60

Tabela 5 – Lista de iniciadores e temperaturas de anelamento

Iniciador 5'	**T (°C)**	**Iniciador 3'**	**T (°C)**	**Diferença (°C)**
E967F	51.11	E826R	48.20	2.91
E967F	51.11	E1407R	50.72	0.39
E967F	51.11	E1406R	49.04	2.06
E967F	51.11	E1115R	51.88	0.77
E967F	51.11	E1065R	53.36	2.25
E967F	51.11	E1064R	49.85	1.26
E917F	45.80	U534R	48.51	2.71
E917F	45.80	E826R	48.20	2.40
E917F	45.80	E533Ra	44.88	0.92
E917F	45.80	E1407R	50.72	4.91
E917F	45.80	E1406R	49.04	3.24
E917F	45.80	E1238R	45.78	0.03
E917F	45.80	E1114R	49.02	3.22
E917F	45.80	E1064R	49.85	4.04
E8Fb	44.10	U534R	48.51	4.41
E8Fb	44.10	E65R	39.57	4.54
E8Fb	44.10	E533Ra	44.88	0.77
E8Fb	44.10	E357R	44.62	0.52
E8Fa	51.08	U534R	48.51	2.57
E8Fa	51.08	U529R	51.00	0.08
E8Fa	51.08	E355R	54.95	3.86
E805F	48.20	U926R	47.20	1.01
E805F	48.20	U534R	48.51	0.31
E805F	48.20	U529R	51.00	2.80
E805F	48.20	E533Ra	44.88	3.32
E805F	48.20	E357R	44.62	3.58
E805F	48.20	E1238R	45.78	2.43
E805F	48.20	E1115R	51.88	3.68
E805F	48.20	E1114R	49.02	0.82
E805F	48.20	E1064R	49.85	1.65
E786F	41.54	E926Rb	40.02	1.52
E786F	41.54	E533Ra	44.88	3.34
E786F	41.54	E357R	44.62	3.08
E786F	41.54	E1238R	45.78	4.23
E517F	57.63	E355R	54.95	2.68
E517F	57.63	E1065R	53.36	4.27

Tabela 5 – Lista de iniciadores e temperaturas de anelamento

Iniciador 5'	**T (°C)**	**Iniciador 3'**	**T (°C)**	**Diferença (°C)**
E388F	54.95	U529R	51.00	3.94
E388F	54.95	E534R	57.83	2.88
E349F	52.11	U534R	48.51	3.60
E349F	52.11	U529R	51.00	1.11
E343F	46.83	U534R	48.51	1.68
E343F	46.83	U529R	51.00	4.17
E343F	46.83	E826R	48.20	1.37
E341F	42.50	E65R	39.57	2.93
E341F	42.50	E533Ra	44.88	2.38
E334F	60.37	E534R	57.83	2.54
E334F	60.37	E533Rb	60.33	0.04
E1391F	53.67	E1115R	51.88	1.79
E1391F	53.67	E1114R	49.02	4.65
E1391F	53.67	E1065R	53.36	0.32
E1391F	53.67	E1064R	49.85	3.83
E1099F	51.05	U926R	47.20	3.85
E1099F	51.05	U1406R	52.71	1.66
E1099F	51.05	E826R	48.20	2.85
E1099F	51.05	E1407R	50.72	0.33
E1099F	51.05	E1406R	49.04	2.00
E1099F	51.05	E1065R	53.36	2.31
E1099F	51.05	E1064R	49.85	1.20
E1046F	53.36	U534R	48.51	4.85
E1046F	53.36	U529R	51.00	2.36
E1046F	53.36	U1406R	52.71	0.65
E1046F	53.36	E534R	57.83	4.47
E1046F	53.36	E1407R	50.72	2.64
E1046F	53.36	E1406R	49.04	4.31
E1046F	53.36	E1115R	51.88	1.48
E1046F	53.36	E1114R	49.02	4.33

Referências

AMAKO, K. et al. Non-exponential growth of mycobacterium leprae thai-53 strain cultured in vitro. *Microbiology and Immunology*, v. 60, n. 12, p. 817–823, 2016. ISSN 03855600. Disponível em: <http://doi.wiley.com/10.1111/1348-0421.12454>. Citado na página 22.

ARUNAGIRI, K. et al. Nasal PCR assay for the detection of Mycobacterium leprae pra gene to study subclinical infection in a community. *Microbial Pathogenesis*, Elsevier Ltd, v. 104, p. 336–339, 2017. ISSN 08824010. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S088240101630540X>. Citado na página 23.

ATTWOOD, T. et al. Concepts, historical milestones and the central place of bioinformatics in modern biology: A european perspective. In: *Bioinformatics - Trends and Methodologies*. [S.l.]: InTech, 2011. Citado na página 28.

BALDI, P. et al. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, Oxford Univ Press, v. 16, n. 5, p. 412–424, 2000. Citado na página 40.

BECHLER, R. G. Hansen versus Neisser : controvérsias científicas na "descoberta"do bacilo da lepra. *História, Ciência, Saúde*, v. 19, n. 3, p. 815–841, 2012. ISSN 01045970. Citado na página 21.

BOKULICH, N. A. et al. mockrobiota: a public resource for microbiome bioinformatics benchmarking. *mSystems*, American Society for Microbiology Journals, v. 1, n. 5, 2016. Disponível em: <http://msystems.asm.org/content/1/5/e00062-16>. Citado na página 28.

BROOKS, J. P. et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiology*, v. 15, n. 1, p. 66, 2015. ISSN 1471-2180. Citado na página 27.

BROSIUS, J. et al. Complete nucleotide sequence of a 16s ribosomal rna gene from escherichia coli. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 75, n. 10, p. 4801–4805, 1978. Citado na página 26.

CAPORASO, J. G. et al. QIIME allows analysis of high-throughput community sequencing data. *Nature Publishing Group*, Nature Publishing Group, v. 7, n. 5, p. 335–336, 2010. ISSN 1548-7091. Disponível em: <http://dx.doi.org/10.1038/nmeth0510-335>. Citado 2 vezes nas páginas 27 e 36.

CHAKRAVORTY, S. et al. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J Microbiol Methods*, v. 69, n. 2, p. 330–339, 2007. ISSN 0167-7012. Citado 3 vezes nas páginas 26, 27 e 56.

CHAUDHARY, N. et al. 16s classifier: a tool for fast and accurate taxonomic classification of 16s rrna hypervariable regions in metagenomic datasets. *PloS one*, Public Library of Science, v. 10, n. 2, p. e0116106, 2015. Citado na página 57.

CHEN, W. et al. A comparison of methods for clustering 16s rrna sequences into otus. *PloS one*, Public Library of Science, v. 8, n. 8, p. e70837, 2013. Citado na página 27.

COLE, J. R. et al. The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, v. 37, n. SUPPL. 1, p. 141–145, 2009. ISSN 03051048. Citado 2 vezes nas páginas 28 e 56.

COLE, S.; EIGLMEIER, K.; PARKHILL, J. Massive gene decay in the leprosy bacillus. *Nature*, v. 409, p. 1007–1011, 2001. ISSN 0028-0836. Disponível em: <http://www.nature.com/nature/journal/v409/n6823/abs/4091007a0.html>. Citado na página 22.

DIJK, E. L. van et al. Ten years of next-generation sequencing technology. *Trends in genetics*, Elsevier, v. 30, n. 9, p. 418–426, 2014. Citado na página 24.

EDGAR, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, v. 26, n. 19, p. 2460–2461, 2010. ISSN 13674803. Citado 3 vezes nas páginas 27, 35 e 37.

EREN, A. M. et al. Oligotyping: differentiating between closely related microbial taxa using 16s rrna gene data. *Methods in Ecology and Evolution*, Wiley Online Library, v. 4, n. 12, p. 1111–1119, 2013. Citado na página 57.

EREN, A. M. et al. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *The ISME journal*, Nature Publishing Group, v. 9, n. 4, p. 968, 2015. Citado na página 57.

FOUHY, F. et al. 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. *BMC Microbiology*, BMC Microbiology, v. 16, n. 1, p. 123, 2016. ISSN 1471-2180. Citado na página 27.

GARDNER, S. N.; SLEZAK, T. Simulate_PCR for amplicon prediction and annotation from multiplex, degenerate primers and probes. *BMC bioinformatics*, v. 15, n. 1, p. 237, 2014. ISSN 1471-2105. Disponível em: <http://www.biomedcentral.com/1471-2105/15/237>. Citado na página 35.

GARZA, D. R.; DUTILH, B. E. *From cultured to uncultured genome sequences: Metagenomics and modeling microbial ecosystems*. [S.l.]: Springer Basel, 2015. 4287–4308 p. Citado na página 26.

GILL, S. R. et al. Metagenomic analysis of the human distal gut microbiome. *science*, American Association for the Advancement of Science, v. 312, n. 5778, p. 1355–1359, 2006. Citado na página 24.

GLOOR, G. B. et al. Microbiome profiling by illumina sequencing of combinatorial sequence-tagged pcr products. *PloS one*, Public Library of Science, v. 5, n. 10, p. e15406, 2010. Citado na página 56.

GRAY, M. W.; SANKOFF, D.; CEDERGREN, R. J. On the evolutionary descent of organisms and organelles: a global phylogeny based on a highly conserved structural core in small subunit ribosomal RNA. *Nucleic Acids Research*, v. 12, n. 14, p. 5837–5852, apr 1984. ISSN 0305-1048. Citado na página 26.

GROUP, J. C. H. M. P. D. G. W. et al. Evaluation of 16s rdna-based community profiling for human microbiome research. *PloS one*, Public Library of Science, v. 7, n. 6, p. e39315, 2012. Citado na página 28.

HANSEN, G. A. Undersøgelser angående spedalskhedens årsager (investigations concerning the etiology of leprosy). *Norsk Mag. Laegervidenskaben*, v. 4, p. 1–88, 1874. Citado na página 21.

HIRAOKA, S.; YANG, C.-c.; IWASAKI, W. Metagenomics and bioinformatics in microbial ecology: Current status and beyond. *Microbes and Environments*, Japanese Society of Microbial Ecology The Japanese Society of Soil Microbiology, v. 31, n. 3, p. 204–212, 2016. Citado 2 vezes nas páginas 24 e 25.

HUSON, D. H. et al. MEGAN analysis of metagenomic data. *Genome Research*, v. 17, n. 3, p. 377–386, 2007. ISSN 10889051. Citado na página 37.

KAR, H. K.; GUPTA, R. Treatment of leprosy. *Clinics in Dermatology*, Elsevier Inc., v. 33, n. 1, p. 55–65, 2015. ISSN 18791131. Disponível em: <http://dx.doi.org/10.1016/j.clindermatol.2014.07.007>. Citado na página 21.

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. ISSN 03051048. Citado na página 56.

KONSTANTINIDIS, K. T.; TIEDJE, J. M. Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, National Acad Sciences, v. 102, n. 7, p. 2567–2572, 2005. Citado na página 26.

KUMARAN, S. M. et al. Comparison of bacillary index on slit skin smear with bacillary index of granuloma in leprosy and its relevance to present therapeutic regimens. *Indian journal of dermatology*, Medknow Publications, v. 60, n. 1, p. 51, 2015. Citado na página 23.

LASTÓRIA, J. C.; ABREU, M. A. M. M. de. Leprosy: Review of the epidemiological, clinical, and etiopathogenic aspects - Part 1. *Anais Brasileiros de Dermatologia*, v. 89, n. 2, p. 205–218, 2014. ISSN 18064841. Citado na página 21.

LAVANIA, M. et al. Detection of viable Mycobacterium leprae in soil samples: Insights into possible sources of transmission of leprosy. *Infection, Genetics and Evolution*, v. 8, n. 5, p. 627–631, 2008. ISSN 15671348. Citado na página 23.

MARTINEZ-PORCHAS, M. et al. How conserved are the conserved 16S-rRNA regions? *PeerJ*, v. 5, p. e3036, 2017. ISSN 2167-8359. Disponível em: <https://peerj.com/articles/3036>. Citado na página 56.

MCDONALD, D. et al. An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME journal*, Nature Publishing Group, v. 6, n. 3, p. 610–618, 2012. Citado na página 28.

MOHANTY, P. et al. Viability of Mycobacterium leprae in the environment and its role in leprosy dissemination. *Indian Journal of Dermatology, Venereology, and Leprology*, v. 82, n. 1, p. 23, 2016. ISSN 0378-6323. Disponível em: <http://www.ijdvl.com/text.asp?2016/82/1/23/168935>. Citado na página 23.

PAGES, H. et al. *Biostrings: String objects representing biological sequences, and matching algorithms*. [S.l.], 2017. R package version 2.38.4. Citado na página 41.

QUAST, C. et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, v. 41, n. D1, p. 590–596, 2013. ISSN 03051048. Citado na página 35.

RAVIN, N. V.; MARDANOV, A. V.; SKRYABIN, K. G. Metagenomics as a tool for the investigation of uncultured microorganisms. *Russian Journal of Genetics*, v. 51, n. 5, p. 431–439, 2015. ISSN 1022-7954. Disponível em: <http://link.springer.com/10.1134/S1022795415050063>. Citado na página 24.

REIBEL, F.; CAMBAU, E.; AUBRY, A. Update on the epidemiology, diagnosis, and treatment of leprosy. *Medecine et Maladies Infectieuses*, Elsevier Masson SAS, v. 45, n. 9, p. 383–393, 2015. ISSN 17696690. Disponível em: <http://dx.doi.org/10.1016/j.medmal.2015.09.002>. Citado na página 23.

SANTALUCIA, J. A unified view of polymer, dumbbell, and oligonucleotide dna nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 95, n. 4, p. 1460–1465, 1998. Citado na página 36.

SCHLOSS, P. D. et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, Am Soc Microbiol, v. 75, n. 23, p. 7537–7541, 2009. Citado na página 27.

SILVA, P. E. et al. Leprous lesion presents enrichment of opportunistic pathogenic bacteria. *SpringerPlus*, v. 4, n. 1, p. 1–8, 2015. ISSN 2193-1801. Citado na página 55.

SINGER, E. et al. Next generation sequencing data of a defined microbial mock community. *Scientific Data*, v. 3, p. 160081, 2016. ISSN 2052-4463. Citado 2 vezes nas páginas 28 e 38.

SIWAKOTI, S. et al. Evaluation of Polymerase Chain Reaction (PCR) with Slit Skin Smear Examination (SSS) to Confirm Clinical Diagnosis of Leprosy in Eastern Nepal. *PLOS Neglected Tropical Diseases*, v. 10, n. 12, p. e0005220, 2016. ISSN 1935-2735. Disponível em: <http://dx.plos.org/10.1371/journal.pntd.0005220>. Citado na página 23.

SNEATH, P. H.; SOKAL, R. R. et al. *Numerical taxonomy. The principles and practice of numerical classification.* [S.l.: s.n.], 1973. Citado na página 26.

SOERGEL, D. A. W. et al. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *The ISME Journal*, Nature Publishing Group, v. 6, n. 7, p. 1440–1444, 2012. ISSN 1751-7362. Disponível em: <http://dx.doi.org/10.1038/ismej.2011.208>. Citado 3 vezes nas páginas 35, 64 e 65.

STEINBOCK, L.; RADENOVIC, A. The emergence of nanopores in next-generation sequencing. *Nanotechnology*, IOP Publishing, v. 26, n. 7, p. 074003, 2015. Citado na página 27.

SUN, D.-L. et al. Intragenomic heterogeneity of 16s rrna genes causes overestimation of prokaryotic diversity. *Applied and environmental microbiology*, Am Soc Microbiol, v. 79, n. 19, p. 5962–5969, 2013. Citado na página 56.

TALHARI, C.; TALHARI, S.; PENNA, G. O. *Clinical aspects of leprosy*. [S.l.]: Elsevier Inc., 2015. 26–37 p. Citado na página 21.

TIKHONOV, M.; LEACH, R. W.; WINGREEN, N. S. Interpreting 16s metagenomic data without clustering to achieve sub-otu resolution. *The ISME journal*, Nature Publishing Group, v. 9, n. 1, p. 68, 2015. Citado na página 57.

TRINGE, S. G. et al. Comparative metagenomics of microbial communities. *Science*, American Association for the Advancement of Science, v. 308, n. 5721, p. 554–557, 2005. Citado na página 24.

TRUMAN, R. W.; KRAHENBUHL, J. L. Viable M. leprae as a Research Reagent. *INTERNATIONAL JOURNAL OF LEPROSY and Other Mycobacterial Diseases*, v. 69, n. 1, 2001. Citado na página 22.

TYSON, G. W. et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, Nature Publishing Group, v. 428, n. 6978, p. 37–43, 2004. Citado na página 24.

UNTERGASSER, A. et al. Primer3-new capabilities and interfaces. *Nucleic Acids Research*, v. 40, n. 15, p. 1–12, 2012. ISSN 03051048. Citado na página 36.

VALLONE, P. M.; BUTLER, J. M. AutoDimer: A screening tool for primer-dimer and hairpin structures. *BioTechniques*, v. 37, n. 2, p. 226–231, 2004. ISSN 07366205. Citado na página 36.

VENTER, J. C. et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science (New York, N.Y.)*, v. 304, n. 5667, p. 66–74, apr 2004. ISSN 1095-9203. Citado na página 24.

VISSCHEDIJK, J. et al. Review: Mycobacterium leprae - Millennium resistant! Leprosy control on the threshold of a new era. *Tropical Medicine and International Health*, v. 5, n. 6, p. 388–399, 2000. ISSN 13602276. Citado na página 23.

WANG, Q. et al. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, Am Soc Microbiol, v. 73, n. 16, p. 5261–5267, 2007. Citado na página 57.

WHO. Leprosy: weekly epidemiological record, Septembre 2016. *World Health Organisation Weekly epidemiological record*, v. 91, n. 35, p. 405–420, 2016. ISSN 0049-8114. Citado 2 vezes nas páginas 21 e 22.

WHO. *Diagnosis of leprosy*. 2017. Accesso em 28 de abr. 2017. Disponível em: <http://www.who.int/lep/diagnosis/en/>. Citado na página 23.

YILMAZ, P. et al. The silva and "all-species living tree project (ltp)" taxonomic frameworks. *Nucleic acids research*, Oxford Univ Press, p. gkt1209, 2013. Citado na página 28.

YOUSSEF, N. et al. Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16s rrna gene-based environmental surveys. *Applied and environmental microbiology*, Am Soc Microbiol, v. 75, n. 16, p. 5227–5236, 2009. Citado na página 57.

ZIEGLER, A. An introduction to statistical learning with applications. rg james, d. witten, t. hastie, and r. tibshirani (2013). berlin: Springer. 440 pages, isbn: 978-1-4614-7138-7. *Biometrical Journal*, Wiley Online Library, v. 58, n. 3, p. 715–716, 2016. Citado na página 40.

ZOU, D. et al. Biological Databases for Human Research. *Genomics, Proteomics & Bioinformatics*, Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China, v. 13, n. 1, p. 55–63, feb 2015. ISSN 16720229. Disponível em: <http://dx.doi.org/10.1016/j.gpb.2015.01.006http://linkinghub.elsevier.com/retrieve/pii/S1672022915000078>. Citado na página 28.