UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE BIOCIÊNCIAS
BACHARELADO EM BIOTECNOLOGIA

LAURA BEZERRA COUTINHO

A INFLUÊNCIA DO ÍNDICE DE CONFIANÇA, *PIPELINE* DE CLUSTERIZAÇÃO E CLASSIFICADORES EM ANÁLISES DE *METABARCODING*

Porto Alegre
2019

LAURA BEZERRA COUTINHO

A INFLUÊNCIA DO ÍNDICE DE CONFIANÇA, *PIPELINE* DE CLUSTERIZAÇÃO E CLASSIFICADORES EM ANÁLISES DE *METABARCODING*

Trabalho de conclusão de curso apresentado como requisito parcial para a obtenção do título de Bacharel em Biotecnologia com ênfase em Bioinformática na Universidade Federal do Rio Grande do Sul.

Orientadora: Profª. Drª. Ursula Matte

Coorientador: Dr. Tiago Falcon Lopes

Porto Alegre

2019

AGRADECIMENTOS

Ao meu namorado, Felipe por me dar força e motivação, por compartilhar comigo meus sonhos e acalmar minhas tristezas, por ser meu porto seguro;

À minha segunda mãe, Mônica, por ter me acolhido na sua família e sempre feito de tudo para me ver bem;

Às minhas colegas, Natália e Isabela, por todos os momentos, incentivo, risadas e desabafos destes quatro anos;

Aos meus amigos, Betina e Felipe, por apesar de tudo sempre estarem ali comigo;

Ao meu time FEJERS, por ser o que me motiva a querer ser sempre a melhor versão de mim, por me permitir ter ousadia para sonhar e coragem para agir;

Aos meus irmãos, Larissa e Leandro, e meus pais, Maria e Antonio, que apesar da saudade sempre serão minha maior força;

À professora Ursula da Silveira Matte pela orientação e compreensão de sempre, e ao meu coorientador, Tiago Falcon Lopes, pela sua forma única de incentivar meu crescimento;

Às minhas colegas de laboratório Martiela, Cristal e Ana, que entre conversas e doces tornaram sempre meus dias mais leves;

A todos que contribuíram para que eu chegasse até aqui.

Ao CNPq, CAPES e FIPE/HCPA pelo apoio financeiro.

Muito obrigada!

RESUMO

As complexas comunidades microbiológicas presentes nos mais diversos ambientes - inclusive o corpo humano - são o que chamamos de microbiota e a compreensão de sua composição e funcionamento é alvo de grande interesse científico, tendo em vista a gama de interações e funções nas quais ela está envolvida. O *metabarcoding* a partir do 16S rDNA é, hoje, um método amplamente aplicado em estudos de composição da microbiota. Nesse contexto, as análises bioinformáticas dos dados gerados enfrentam o grande desafio da garantia de qualidade e reprodutibilidade dos resultados. *Pipelines* de análise padronizadas são uma alternativa para essa questão, no entanto, o uso de parâmetros apropriados podem gerar impacto direto nos resultados obtidos. O objetivo deste trabalho foi, portanto, elucidar a influência de parâmetros de análise nos resultados de análises de dados de 16S rDNA referentes à microbiota intestinal humana, utilizando como referência o *pipeline* do BMP - *Brazilian Microbiome Project.* Para isso foram comparados os resultados de diferentes combinações entre índices de confiabilidade, classificadores e *pipelines* para clusterização. Nossos resultados indicam diferenças claras entre a aplicação de diferentes parâmetros ao *pipeline*, gerando diferentes efeitos na quantidade de taxa identificados, de OTUs classificadas e na precisão da classificação. Resultados de combinações de classificadores e índices de confiança apresentam variações entre os dois *pipelines* de clusterização, no USEARCH havendo pouca diferenciação com a alteração dos classificadores e no VSEARCH apresentando maiores disparidades - com destaque para o Mothur, cujos resultados de número de taxa identificados e OTUs classificadas foram acima dos demais, não respondendo inclusive ao aumento do índice de confiança. Destacamos ainda o papel da remoção de sequências quiméricas na qualidade dos resultados. Com isso, salientamos a importância da inclusão em estudos de microbiota de detalhes dos parâmetros e métodos aplicados, garantindo a validação, qualidade e reprodutibilidade do resultado. Compreender os *pipelines* aplicados e os efeitos de seus parâmetros é essencial. Testar diferentes *pipelines* e variações dos parâmetros é recomendável.

Palavras-chave: Microbiota, *metabarcoding,* BMP, métodos de clusterização, classificadores, índice de confiança.

ABSTRACT

The complex microbiological communities found in many environments - including the human body - are the microbiota. The comprehension of its composition and operation is target to great scientific interest, in view of the scale of its interactions and functions in which it is involved. Metabarcoding using 16S rDNA data is, today, a method widely applied to studies of microbiota composition. In this context, bioinformatic analysis of the data generated face the great challenge of guaranteeing the quality and reproducibility of results. Standardized analysis pipelines are alternatives to this problem, however, using appropriate parameters may have direct impact in the results obtained. Therefore, the objective of this work is to clarify the influence of the parameters in the results of analysis of 16S rDNA from human gut, using as reference the BMP - Brazilian Microbiome Project - pipeline, and comparing the results of different combinations between minimum confidences, classifiers and clustering pipelines. Our results point to visible differences between the application of different parameters to the pipeline, having diverse effects in the quantity of identified taxa, classified OTUs and the precision of the classification. The results of combinations of classifiers and minimum confidences present variation between the two clustering pipelines, USEARCH showing little differences after altering the classifiers and VSEARCH great disparities - specially Mothur, whose results of number of identified taxa and classified OTUs were always above the others, even not responding to the increase in the minimum confidence. We also highlight the importance of the removal of chimeric sequences to the quality of results. Thereby, we point the importance of inclusion in microbiota studies of details of parameters and methods applied, guaranteeing the validation, quality and reproducibility of the result. It is essential to understand the applied pipelines and their parameters. Testing different pipelines and variations of parameters is recommendable.

Key-words: Microbiota, *metabarcoding,* BMP, clustering methods, classifiers, minimum confidence.

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Microbiota

Microrganismos estão em todos os lugares, espalhados no ambiente ou interagindo com um hospedeiro. Podem atuar na remediação de poluentes, além de serem capazes de tornar nutrientes, metais e vitaminas disponíveis para seus hospedeiros, sendo estes plantas ou animais. Podem ainda ser utilizados na produção de fármacos, recombinação gênica ou indústria alimentícia, apresentando uma imensa variedade de possíveis aplicações biotecnológicas (The National Academies Press, 2007). As complexas comunidades microbiológicas - compostas não só por bactérias mas também vírus, fungos e protozoários (Belkaid & Harrison, 2017) - responsáveis por tais atividades são o que chamamos de microbiota.

Em seres humanos, a microbiota, em número de células, supera a contagem de células do hospedeiro, expressando mais genes do que o genoma humano (Sender et al., 2016). Este conjunto de genomas microbianos proporciona ao ser humano características evolutivamente não desenvolvidas (Turnbaugh et al., 2007). A este conjunto de genomas chamamos microbioma (Barko et al., 2018).

Hoje, a compreensão da composição e funcionamento destas comunidades é objeto de diversos estudos (Pepper & Rosenfeld, 2012; Audebert et al., 2016; Vieira-Silva et al., 2016). É conhecido que existem múltiplos fatores capazes de influenciar a diversidade e distribuição dos microrganismos (Turnbaugh et al., 2007), incluindo o tipo de parto e de alimentação quando bebê, todo o processo de crescimento, composição da dieta, localização geográfica, medicações e estresse (Cresci & Bawden, 2015). Dependendo de todo este contexto, um mesmo microrganismo pode desenvolver uma relação mutualista, comensal ou parasítica com seu hospedeiro (Belkaid & Harrison, 2017), o que ressalta a importância do equilíbrio dessa comunidade para a saúde humana.

Em condições consideradas normais, a homeostase é mantida pela relação de mutualismo estabelecida entre o organismo do hospedeiro e a microbiota, na qual bactérias realizam uma série de outras ações além de fornecer ao ser humano parte dos nutrientes essenciais e metabolizar compostos indigeríveis, enquanto o organismo humano proporciona às bactérias nutrientes e um ambiente estável (Martín et al., 2013). Tal homeostase, quando quebrada, pode acarretar em disbioses - patologias associadas ao desbalanço da microbiota -

como casos de obesidade (Turnbaugh et al., 2006), alergias, doenças autoimunes, doenças inflamatórias (Belkaid & Harrison, 2017) e diabetes tipo 2 (Integrative HMP Research Network Consortium, 2014).

Entre outros papéis da microbiota no organismo humano pode-se citar a indução da função do sistema imune, definindo a patogênese e o resultado de infecções (Belkaid & Harrison, 2017), atuando na resposta a vacinas (Seekatz et al., 2013), suprimindo respostas inflamatórias a alimentos e outros antígenos ingeridos (Weiner et al., 2011), contribuindo para a hematopoiese e para o controle da função de células que saem da medula óssea (Maslowski et al., 2009; Chang et al., 2014) e, até mesmo, modulando o padrão de expressão gênica de macrófagos (Chang et al., 2014). Belkaid & Harrison (2017) associam a atividade da microbiota a potencial terapêutico aplicado à vacinação e à resistência contra bactérias resistentes a antibióticos. Existem ainda evidências da contribuição destes microrganismos para a resposta anti-tumoral (Belkaid & Harrison, 2017) e para a capacidade do hospedeiro de controlar tumores durante a imunoterapia (Iida et al., 2013). Além disso, Tanoue et al. (2019) apresentou a capacidade da microbiota de modular a resposta imune à órgãos transplantados em camundongos.

Ainda assim, diante da complexidade destas comunidades, tão interdependentes quanto adaptativas, ainda é pouco o que se sabe acerca da sua dinâmica. Como essas comunidades - incluindo bactérias, vírus, fungos e protozoários - cooperam e se influenciam permanece desconhecido (Belkaid & Harrison, 2017).

1.2 A microbiota intestinal

O trato gastrointestinal humano (TGI) se destaca por ser um dos ecossistemas mais complexos conhecidos, abrigando a maior coleção de microrganismos, entre bactérias, arqueas, fungos e vírus, em toda a sua extensão (Turnbaugh et al., 2007). O número total destes organismos pode chegar a 100 trilhões (Gill et al. 2006), com cerca de 1000 espécies bacterianas (Martín et al., 2013). A composição e função dessa comunidade estão atreladas às condições fisiológicas do TGI, que, por sua vez, são suscetíveis a uma série de fatores externos e internos, como dieta, exposição a antibióticos, exercício físico, gravidez, e até mesmo ao gênero do indivíduo (Maruvada et al., 2017; Laitinen & Mokkala, 2019). Esse conjunto de fatores torna a microbiota fecal de um indivíduo adulto única e, apesar da complexidade, estável (Martín et al., 2013).

Além disso, sabe-se hoje que o microbioma intestinal está envolvido em funções de todo o organismo humano, sendo reconhecidos os eixos de interação com o cérebro (Foster and Neufeld, 2013), fígado (Compare et al., 2012), pulmões (Budden et al., 2017; Hu et al., 2019), e o mais recentemente proposto, eixo intestino-músculos (Grosicki et al., 2018). Entre as funções às quais a microbiota intestinal está associada estão comportamentos relacionados ao estresse (Foster and Neufeld, 2013), o metabolismo da bile (Maruvada et al., 2017), alteração do metabolismo de nutrientes em resposta ao ritmo circadiano do hospedeiro (Leone et al., 2015), obesidade (Maruvada et al., 2017; Grosicki et al., 2018), desenvolvimento de doenças pulmonares, como asma e pneumonia (Hu et al., 2019), e a alteração de características musculares (Grosicki et al., 2018).

Tendo o alcance de suas interações no organismo em vista, a microbiota intestinal se tornou um importante objeto para diversos estudos clínicos acerca de disbioses (Pepper e Rosenfeld, 2012; Vieira-Silva et al., 2016), cujas consequências podem variar de complicações no próprio órgão - como doença inflamatória intestinal - a diabetes e até diferentes tipos de câncer (Pepper & Rosenfeld, 2012; Martín et al., 2013). Em contrapartida, o aprofundamento destes estudos já traz consigo novas oportunidades de prevenção e tratamento através terapias microbioma-direcionadas (ou *microbiome-targeted therapies*) para condições como ansiedade e depressão (Foster and Neufeld, 2013), tuberculose (Hu et al., 2019) e câncer (Frankel et al., 2019; Selber-Hnatiw et al., 2017), além de terapias como o transplante de microbiota fecal, cuja eficácia já foi comprovada para diversas patologias (Hsu et al., 2019).

Assim, tudo o que já se conhece acerca de possíveis aplicações e funcionalidades da microbiota intestinal reforça a importância e necessidade de métodos eficazes que possibilitem acessar a sua composição de forma precisa para detecção de possíveis alvos e validação de novos métodos.

1.3 *Metabarcoding*

As análises da comunidade bacteriana através do sequenciamento em larga escala (sequenciamento de nova geração) permitiu expandir o olhar de organismos únicos para comunidades, obtendo uma interpretação sem precedentes das relações de microrganismos em resolução espacial e temporal a nível de um indivíduo ou mesmo de ecossistemas ao redor do planeta (Caporaso et al., 2011).

Um dos métodos que permitiu o alcance de tal nível de interpretação é o *metabarcoding*, que Deiner et al. (2017) conceitua como a identificação taxonômica de múltiplas espécies extraídas de uma amostra mista que foi amplificada por PCR e sequenciada em uma plataforma de larga escala. Este tipo de análise é uma poderosa ferramenta capaz de acessar e caracterizar a riqueza de espécies em ecossistemas de forma não-invasiva e permitindo a superação da barreira antes imposta pela necessidade de cultivo dos organismos, prática que impedia o acesso a uma diversidade de espécies não-cultiváveis (Fiannaca et al., 2018; Deiner et al., 2017).

A sequência da região do gene codificador da subunidade 16S do rRNA (utilizaremos o termo 16S rDNA para indicar que o sequenciamento utiliza o DNA e não o RNA ribossomal) é a mais utilizada como marcador para caracterização de comunidades bacterianas (Fiannaca et al., 2018). O gene é encontrado em todos os microrganismos e consiste em nove regiões hipervariáveis (V1 a V9) - que apresentam especificidade a cada espécie - separadas por nove regiões altamente conservadas e não significativamente diferentes entre espécies (Song et al., 2019) (**Figura 1**). As regiões hipervariáveis são, em geral, o alvo em pesquisas de microbioma devido ao baixo custo e considerável acurácia (López-García et al., 2018), sendo a sub-região V4 a V6 a mais utilizada, adotada inclusive no *Human Microbiome Project* (Fiannaca et al., 2018).

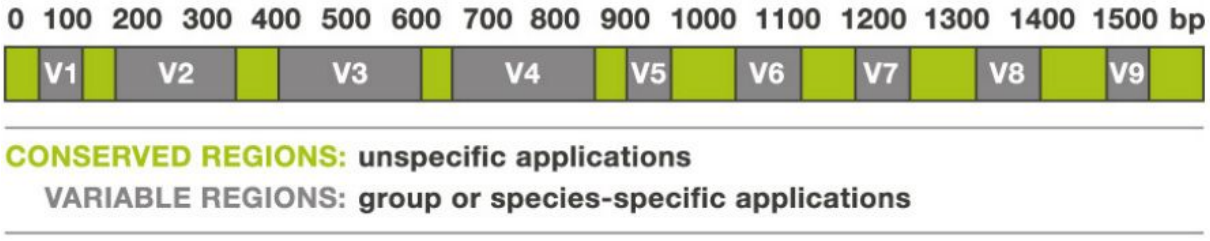

Figura 1 - Gene codificador do 16S rRNA. As regiões em verde são conservadas e são os alvos de primers para amplificação de todos os genes de rRNA de uma amostra por PCR, enquanto as regiões em cinza são espécie-específicas, e permitem analisar quais espécies estão presentes em uma comunidade. Disponível em <https://teachthemicrobiome.weebly.com/sequencing-the-microbiome.html> acesso em 5 de junho de 2019.

O *metabarcoding* a partir do 16S rDNA é, hoje, um método robusto aplicado em estudos de composição da microbiota de diversas espécies animais e vegetais, incluindo as utilizadas para consumo humano (Bruno et al., 2019), gerando informações sobre o impacto de uma série de condições na microbiota ao mesmo tempo que permite a identificação de

novas espécies (Allali et al., 2017). Todavia, a precisão da técnica ainda é afetada por fatores distribuídos por todo o seu fluxo de trabalho, desde a amostragem até a análise de bioinformática, última fase em que os arquivos resultantes do sequenciamento são computacionalmente processados.

Essa última fase geralmente é realizada em três etapas: clusterização de OTUs (*Operational Taxonomic Unit* - cada táxon propriamente dito), classificação taxonômica e análises de diversidade (Deiner et al., 2017). Na primeira, as *reads* já filtradas a partir de sua qualidade são agrupados de acordo com um limiar de similaridade determinado, gerando OTUs. Dessa forma são gerados clusters de *reads* que possivelmente representam um gênero ou espécie da amostra original, no entanto, destaca-se que mais de uma OTU pode se referir ao mesmo organismo, uma vez que a clusterização é realizada baseada somente na similaridade entre sequências da amostra, independentemente de referências taxonômicas (Tan et al., 2018). A segunda etapa diz respeito à comparação de sequências de clusters de OTUs anônimos com sequências de bancos de dados de referência para obtenção da sua classificação taxonômica. Por fim, o objetivo final da maior parte de estudos de *metabarcoding* é a caracterização da comunidade alvo, analisando sua riqueza e abundância diferencial de espécies por meio de índices calculados a partir de *softwares*. Tal resultado é o que permite a associação da análise com aspectos da ecologia do microbioma analisado.

O desafio atual no que se refere a análises bioinformáticas de dados de *metabarcoding* é a garantia de qualidade e reprodutibilidade dos resultados. Uma alternativa para essa questão é a padronização de *pipelines* de análise (Deiner et al., 2017).

1.4 O BMP

O *pipeline* referência para o desenvolvimento da análise deste trabalho foi a versão atualizada do BMP - *Brazilian Microbiome Project* (Pylro et al., 2014; atualização disponível em <https://www.brmicrobiome.org/clusteringmeth> acesso em 18 de maio de 2019). O projeto é inspirado pelo *Earth Microbiome Project* (EMP) e sua proposta é a criação de um banco de dados brasileiro de metagenômica, padronizando protocolos de análises e integrando pesquisas em andamento e futuras, garantindo sua reprodutibilidade.

O protocolo do BMP possui base no *pipeline* do UPARSE (Edgar, 2013), no qual são combinadas as etapas iniciais de filtro de qualidade (utilizando *reads* do mesmo tamanho) e descarte de *singletons* (*reads* que aparecem uma única vez na amostra), filtragem de quimeras

(sequências > 3% de erros da sequência biológica mais próxima) e clusterização *de novo* das OTUs. Como resultado, o *pipeline* apresenta melhor resultado (com menos quimeras ou outros contaminantes) que os softwares QIIME ou Mothur (Edgar, 2013).

No *pipeline* de análise dos dados de 16S rDNA do BMP, as etapas anteriormente citadas eram realizadas via USEARCH (Edgar, 2010; 2013) que, após recente atualização, foi substituído pelo VSEARCH (Rognes et al., 2016). Ambos são combinações de diferentes algoritmos e ferramentas de preparação e processamento de sequências genômicas e metagenômicas em uma só plataforma. No entanto, o VSEARCH, que surgiu com o objetivo de ser uma alternativa sem as limitações da versão aberta do USEARCH (como o tamanho máximo para arquivo inicial da análise de 4GB), possui uma extensão das funcionalidades deste, realizando outras operações e incluindo novos algoritmos (Rognes et al., 2016). Em contrapartida, durante a etapa de clusterização de OTUs, o VSEARCH não realiza a detecção de quimeras, diferindo do USEARCH e fazendo-se necessária a realização desta etapa adicional.

Após estas etapas é utilizado o QIIME para a determinação taxonômica das OTUs. Neste passo existem três possíveis parâmetros: o banco de dados de sequências de referência, o classificador utilizado, e o índice de confiança.

A comparação das *reads* é realizada em relação a sequências já classificadas presentes nos bancos de dados dedicados a genes de 16S rRNA, como o SILVA (Pruesse et al., 2007; Quast et al., 2013), Greengenes (DeSantis et al., 2006) e o RDP (*Ribosomal Database Project*) (Bacci et al., 2015). Quem realiza essa busca de sequências do banco de dados para treinar a classificação taxonômica das OTUs é o classificador, sendo os principais: o RDP *classifier* (*Ribosomal Database Project classifier*), o UCLUST e o Mothur.

O RDP *classifier* (Wang et al., 2007) utiliza o método Bayesiano e, para *reads* de 200 bases como as que foram utilizadas neste projeto, apresenta um percentual de mais de 99% de acerto para a classificação de taxa a nível de Filo e Classe, mais de 95% de acerto para Ordem e Família e 86,6% de acerto para o nível de Gênero, considerando-se a classificação do NCBI como referência. Enquanto o UCLUST (Edgar, 2010) possui como principal aplicação a clusterização de sequências, sendo capaz de clusterizar rapidamente um grande número de sequências e gerando resultado mais eficientemente que o software CD-HIT (Li e Godzik, 2006), previamente utilizado no pipeline do QIIME (Caporaso et al., 2011). Ele emprega o USEARCH como uma subrotina para classificação de sequências a clusters buscando

similaridade de sequências. Por fim, o Mothur (Schloss et al., 2009) é uma plataforma que reúne uma diversidade de algoritmos, permitindo a aplicação de diferentes abordagens e análises. Ele busca permitir análises de grandes conjuntos de dados com menos recursos computacionais, alcançando velocidade maior de processos por meio da redução da quantidade de memória necessária aliada à possibilidade de paralelização de comandos (Schloss et al., 2009).

O último parâmetro é o índice de confiança, que corresponde à confiança mínima necessária para que o algoritmo registre uma classificação taxonômica, ou seja, quanto maior o valor utilizado mais rigorosa se torna a classificação. Esse nível de exigência no tratamento dos dados durante todo o *pipeline* é essencial para a redução da porcentagem de reads mal-classificadas e para a geração de dados de alta qualidade (Pylro et al., 2014).

Destaca-se ainda que o *pipeline* do BMP emprega a versão 1 do QIIME em suas análises, que já não é mais suportada pelos autores do programa. A atual versão é o QIIME 2 (Bolyen et al., 2018), cuja aplicação não foi testada neste projeto.

A seleção de um *pipeline* e parâmetros apropriados podem levar a resultados próximo dos limites teóricos para um dado tipo de comunidade (Golob et al., 2017), ao mesmo tempo que a falta de atenção a estas variáveis pode levar a resultados de baixa qualidade e alta quantidade de *miscalls*. Em sua análise, Golob et al. (2017) apresenta que em muitos estudos é utilizado o método *default* ou padrão descrito em tutoriais para o respectivo *pipeline* aplicado, no entanto, os parâmetros sugeridos estiveram entre os de pior performance entre todos os resultados analisados.

Autores como Golob et al. (2017) já realizaram comparações de parâmetros para diferentes aspectos individuais da análise bioinformática de dados de *metabarcoding* a partir de 16S rDNA, como a comparação de classificadores (Almeida et al., 2018), de diferentes *pipelines* ou de bancos de dados de referência (López-García et al., 2018). Contudo, estudos que apliquem combinações destes diferentes parâmetros a um *pipeline* único ainda não foram realizados.

Este trabalho visa, portanto, elucidar o papel destes parâmetros no *pipeline* do Brazilian Microbiome Project e comparar resultados de diferentes combinações entre índices de confiabilidade, classificadores e *pipelines* para limpeza e clusterização dos dados de sequenciamento de nova geração de 16S rDNA provenientes da plataforma *Ion Torrent* referentes à microbiota intestinal humana, buscando ressaltar a diferenciação dos índices de

taxa identificados e/ou da precisão da sua classificação, aspectos essenciais para a associação da microbiota intestinal às diferentes condições fisiológicas.

2 OBJETIVOS

2.1 Objetivo geral

O principal objetivo deste trabalho é realizar comparações entre diferentes índices de confiança, classificadores e *pipelines* de clusterização de sequências, visando detectar seu impacto no índice de taxa identificados e/ou de OTUs classificadas e na precisão da classificação em análises de enriquecimento de 16S rDNA de amostras de fezes humanas.

2.2 Objetivos específicos

Para isso, os seguintes objetivos específicos deverão ser alcançados:

(i) Verificar efeito da aplicação de diferentes níveis de confiança no QIIME;

(ii) Comparar diferentes classificadores no QIIME (UCLUST, RDP e Mothur);

(iii) Comparar os resultados de clusterização OTUs pelos diferentes softwares já utilizados pelo *pipeline* do BMP para seus passos iniciais, o USEARCH e o VSEARCH, respectivamente.

3 MATERIAIS E MÉTODOS

3.1 Banco de dados

Para realizar os testes propostos, foram utilizados os dados de sequenciamento da plataforma *Ion Torrent PGM* (*Personal Genome Machine* – mesma plataforma disponível no Centro de Pesquisa Experimental do Hospital de Clínicas de Porto Alegre) disponíveis no *NCBI SRA Database* (<https://www.ncbi.nlm.nih.gov/sra/>) sob o número de acesso SRX1044553 (Oliveira et al., 2016), onde foi sequenciada a região V4 do gene codificador de 16S rRNA.

3.2 *Pipeline* inicial

As etapas do *pipeline* completo a serem aplicadas estão descritas na **Figura 2**. Os arquivos FASTQ passam por uma etapa de análise de qualidade das *reads* não prevista pelo *pipeline* do BMP antes do início da análise e após os filtros de qualidade, utilizando-se o FastQC (v.0.11.5) (Andrews, 2010).

Em seu *pipeline* padrão, o BMP baseia-se no banco de dados *Greengenes*, aplica o método *default* para classificação - o UCLUST -, índice de confiabilidade *default* igual a 0,5, e sua versão atualizada utiliza o software VSEARCH para as fases iniciais da análise enquanto sua versão anterior utilizava o USEARCH.

Para as etapas iniciais de processamento e clusterização dos dados foram testados os dois *pipelines* de análise que, em tese, deveriam gerar o mesmo resultado - VSEARCH e USEARCH. O uso do VSEARCH foi baseado nos passos apresentados pela atual *pipeline* do BMP, enquanto o USEARCH teve sua aplicação baseada na versão prévia do BMP que o utilizava como padrão. A etapa de clusterização de OTUs foi realizada aplicando limiar de 97% de similaridade.

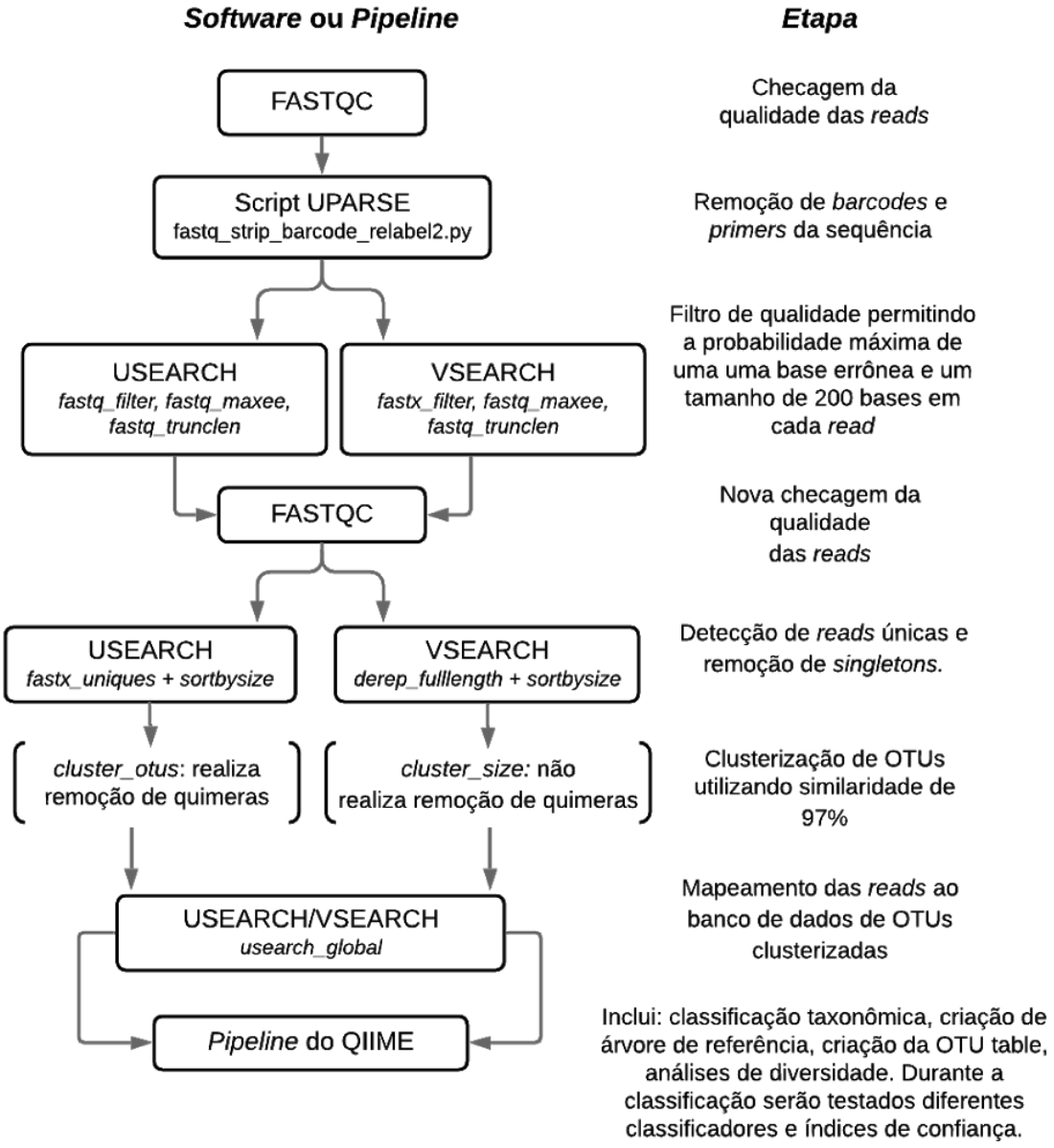

Figura 2 - Etapas do *pipeline* a serem aplicadas incluindo análises de qualidade da sequência não previstas pelo BMP. Nas caixas à esquerda estão os *softwares*, *pipelines* ou *scripts* utilizados, à direita as análises realizadas.

3.3 Classificação taxonômica

Na fase de identificação taxonômica das OTUs a partir do QIIME 1 (Caporaso et al., 2010), foi mantido o uso do *Greengenes* como banco de dados *default* para todas as análises realizadas. Foram testados ainda os três classificadores citados, RDP *classifier (Wang et al., 2007)*, UCLUST (Edgar, 2010) e Mothur (Schloss et al., 2009) e, por fim, o índice de confiança, para o qual foram utilizados os valores 0,5, 0,8, 0,9 e 0,95. Para o UCLUST pode ser aplicado somente o valor *default*, 0,5, como apresentado na descrição do script

assign_taxonomy.py do QIIME (Disponível em: <http://qiime.org/scripts/assign_taxonomy.html> acesso em 28 de maio de 2019).

Visando compreender o efeito da alteração dos parâmetros no índice de taxa identificados e na confiabilidade da classificação foram testadas todas as combinações entre os três conjuntos de parâmetros presentes na etapa de classificação do QIIME (**Figura 3**).

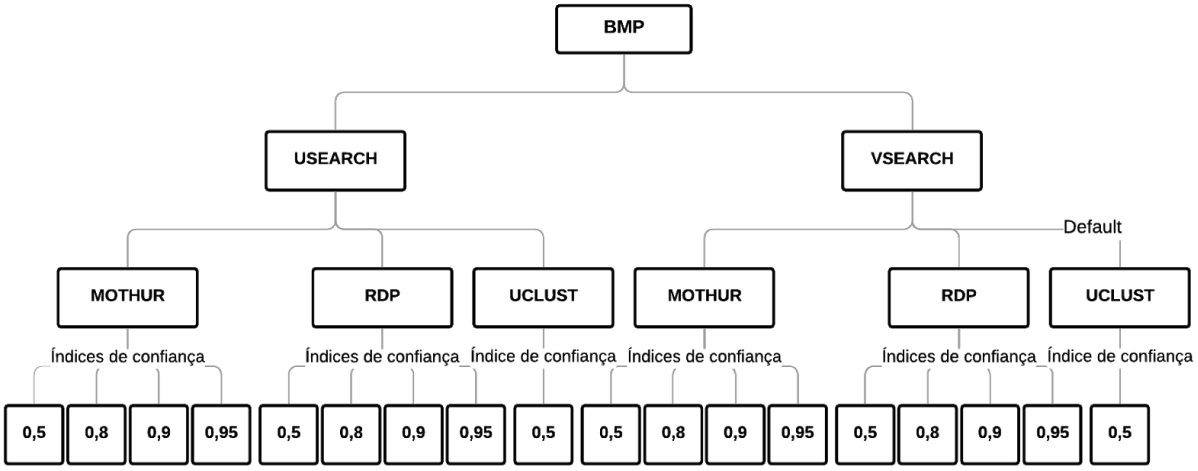

Figura 3 - Diagrama de todos as combinações de parâmetros do *pipeline* do BMP testados neste trabalho: dois *pipelines* de análises iniciais, três classificadores e quatro índices de confiança. Indicado como parâmetros *default* está o conjunto utilizado na atual versão do BMP.

Dois importantes parâmetros de comunidades são a α-diversidade (alfa-diversidade) - diversidade dentro de uma amostra, ou o número de espécies observadas em um ambiente -, e a β-diversidade (beta-diversidade), que corresponde à diversidade biológica entre ambientes, como o número de espécies compartilhadas entre duas amostras. Exemplos de medidas destes dois tipos de diversidade são as curvas de rarefação e análises de coordenada principal, medidas de alfa e beta-diversidade (Dada et al., 2014; Agnello et al., 2017), respectivamente, e na última etapa do *pipeline* do BMP o QIIME gera ambas. Como parâmetro para a análise foram testados os valores 500, 1000 e 10000 referentes à profundidade de sequenciamento a ser utilizada para amostragem (*sub-sampling*) equilibrada e profundidade máxima de rarefação.

Após término da aplicação do *pipeline* do BMP foram realizados dois passos adicionais. O primeiro para conversão da OTU *table* do formato BIOM para .spf a partir do script *biom_to_stamp.py* do *Microbiome Helper* (Comeau et al., 2016), repositório de *scripts*

para análises de microbiomas, de forma que o arquivo pode ser lido pelo STAMP (Parks et al., 2014), aberto no Excel separado por tabulação e salvo como arquivo texto. Por fim, foi utilizado o script *checkHierarchy.py* do STAMP (Parks et al., 2014) para conferência da hierarquia taxonômica, identificando entradas da árvore gerada que não são estritamente hierárquicas e que devem ser manualmente corrigidas.

4 RESULTADOS E DISCUSSÃO

4.1 *Pipeline* inicial

4.1.1 Quantificação de OTUs

O *pipeline* de análises iniciais, por incluir a etapa de clusterização de OTUs, possui influência direta no número de OTUs presentes na *OTU table* final de cada análise.

Ao todo, foram geradas 18 *OTU tables*, seguindo todas as combinações de parâmetros abordadas na **Figura 3**, e que estão disponíveis no seguinte link do google drive (Disponível em:<https://drive.google.com/drive/folders/1qaeudguZJbLv8Gr5H90B18Ev-3pNG7R4?usp=sharing>). Neste total, é possível observar dois grandes grupos, um gerado a partir do *pipeline* do USEARCH e o outro a partir do VSEARCH. O primeiro possui 653 OTUs em sua composição, enquanto o segundo atinge um total de 968 OTUs.

Esse resultado se deve ao método de clusterização aplicado pelos *pipelines*, enquanto o do USEARCH realiza a filtragem de quimeras (sequências > 3% de erros da sequência biológica mais próxima) e clusterização *de novo* das OTUs de forma única (Edgar, 2013), o VSEARCH não possui essa união das duas etapas, realiza somente a clusterização e a remoção de quimeras deve ser feita em uma etapa adicional prévia à clusterização (Rognes et al., 2016), de forma que os *reads* de quimeras deixam de ser considerados para a geração de OTUs .

Segundo definição do NCBI de quimeras de rRNA (disponível em <https://www.ncbi.nlm.nih.gov/genbank/rrnachimera/> acesso em 18 de maio de 2019), estima-se que até 30% das sequências de amostras ambientais podem ser quiméricas. Não realizar esta remoção permite a permanência destas sequências artefatos no conjunto de dados considerados para as análises seguintes, podendo influenciar significativamente durante a clusterização de OTUs, gerando não só mais OTUs pela maior disponibilidade de sequências mas também OTUs potencialmente errôneas, capazes de influenciar nos resultados de diversidade e no comparativo entre populações (Edgar et al., 2011).

Destaca-se que esta etapa adicional não está prevista pelo *pipeline* do BMP após a sua atualização para o uso do VSEARCH (atualização disponível em <https://www.brmicrobiome.org/clusteringmeth> acesso em 18 de maio de 2019). Portanto, os dados gerados neste trabalho aplicando o *pipeline* do VSEARCH e não realizando a remoção das quimeras refletem os resultados que seriam obtidos a partir do *pipeline* hoje

presente no BMP.

4.1.2 Análises de diversidade

Para a realização das análises de diversidade, o QIIME considera a quantidade de OTUs observadas e, por esse motivo estas análises foram aplicadas para validação e exemplificação do efeito do aumento no número de OTUs clusterizadas pela troca de USEARCH para VSEARCH no resultado final da análise.

Para eliminação de fatores de classificação taxonômica foram considerados para as análises de diversidade somente os resultados gerados a partir das análises utilizando os parâmetros de classificação *default* - classificador UCLUST e 0,5 de confiança mínima. O critério aplicado para a seleção da profundidade de sequenciamento utilizado foi a observação do alcance de um platô nos valores de OTUs observadas nas curvas de rarefação. Esse platô é atingido pois, conforme a profundidade de sequenciamento aumenta, o número de OTUs ou a quantidade de espécies detectadas também aumenta e a curva de rarefação tende a atingir um valor máximo, indicando que toda a diversidade de espécies amostradas foi capturada pelo sequenciamento (Song et al., 2019).

Nos testes realizados aplicando profundidades 500 e 1.000, foi observado que os valores de OTUs observadas continuaram crescentes, sem atingir um platô mesmo quando aplicado o valor 10.000, como ilustrado na **Figura 4**. Portanto, para a observação dos resultados de rarefação e de coordenada principal foram considerados os gráficos gerados a partir da aplicação do maior valor de profundidade testado - 10.000.

Para apresentação dos resultados e discussão das análises de diversidade destaca-se que as amostras obtidas a partir do banco de dados de sequências utilizado (Oliveira et al., 2016) estão subdivididas em um grupo controle e outro de indivíduos portadores de fenilcetonúria (PKU).

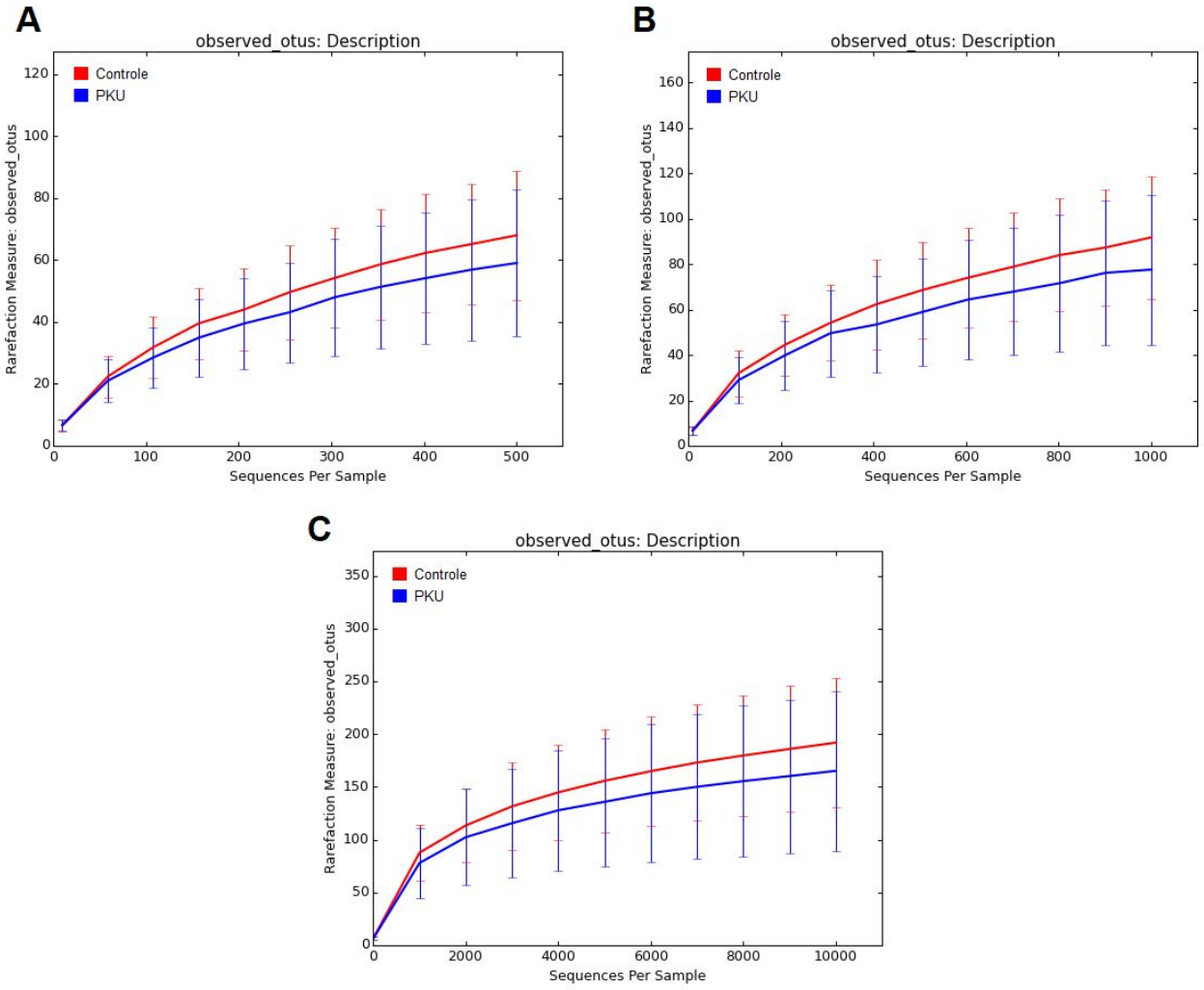

Figura 4 - Curvas de rarefação utilizando (A) 500, (B) 1000 e (C) 10000 como parâmetro de profundidade de sequenciamento.

Curvas de rarefação são utilizadas para estimar se toda a diversidade da comunidade original foi capturada pela análise realizada. Elas são baseadas tanto no número de OTUs/espécies quanto na proporção em que estas OTUs são representadas na comunidade, dessa forma, uma comunidade com alta alfa-diversidade possui alto número de espécies e as abundâncias destas são similares (Nipperess, 2016).

Neste trabalho, a análise da curva de rarefação foi realizada considerando o número de OTUs observadas de acordo com o número de sequências por descrição (Controle e PKU) e por amostra. Nas **Figuras 5** e **6** estão representados os dois conjuntos de gráficos gerados pelos pares USEARCH e VSEARCH para estes dois parâmetros.

Na **Figura 5** estão os gráficos de número de OTUs observadas por amostra e

observa-se que, além de um grande aumento do número de OTUs observadas, ocorre ainda uma inversão entre as curvas das amostras p7 e c6, de forma que a c6 - que pelo USEARCH possui menor medida de OTUs - passa a ter uma curva superior à p7 no *plot* a partir do VSEARCH, demonstrando uma alteração na abundância relativa de OTUs entre amostras. Nas curvas de rarefação por descrição (**Figura 6**), o número de OTUs observadas também aumenta significativamente, além de um aumento da distância entre as duas curvas (Controle/PKU), significando que haveria maior diferença entre o número de OTUs entre os grupos e, consequentemente, do número de espécies detectadas. Em ambos os casos a potencial presença de quimeras na detecção de OTUs pode estar relacionada a presença de falsos positivos, alterando as percepções da diversidade nas amostras, o que pode ter consequências no caso de estudos clínicos.

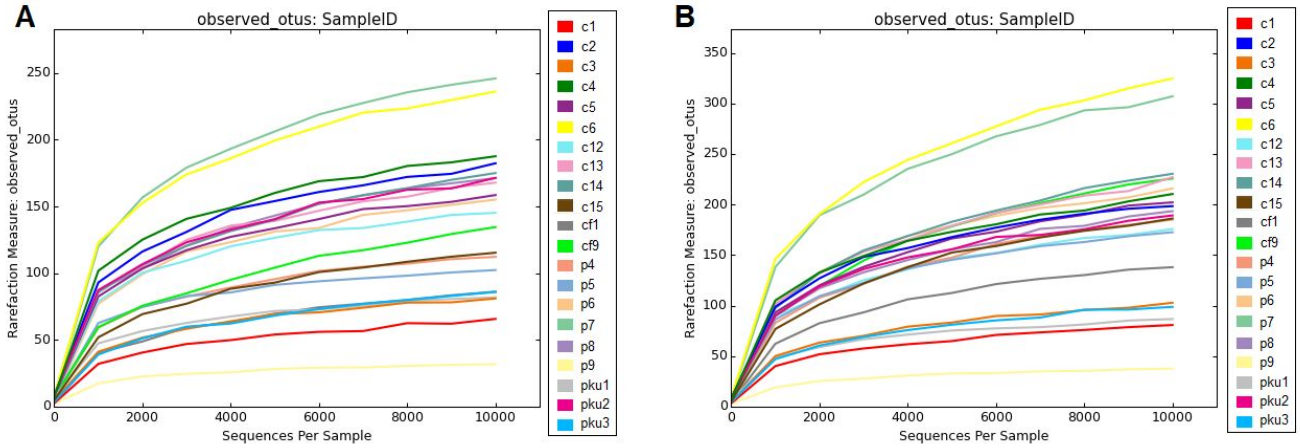

Figura 5 - Curvas de rarefação do número de OTUs observadas de acordo com o número de sequências por amostra. (A) A partir do *pipeline* do USEARCH. (B) A partir do *pipeline* do VSEARCH.

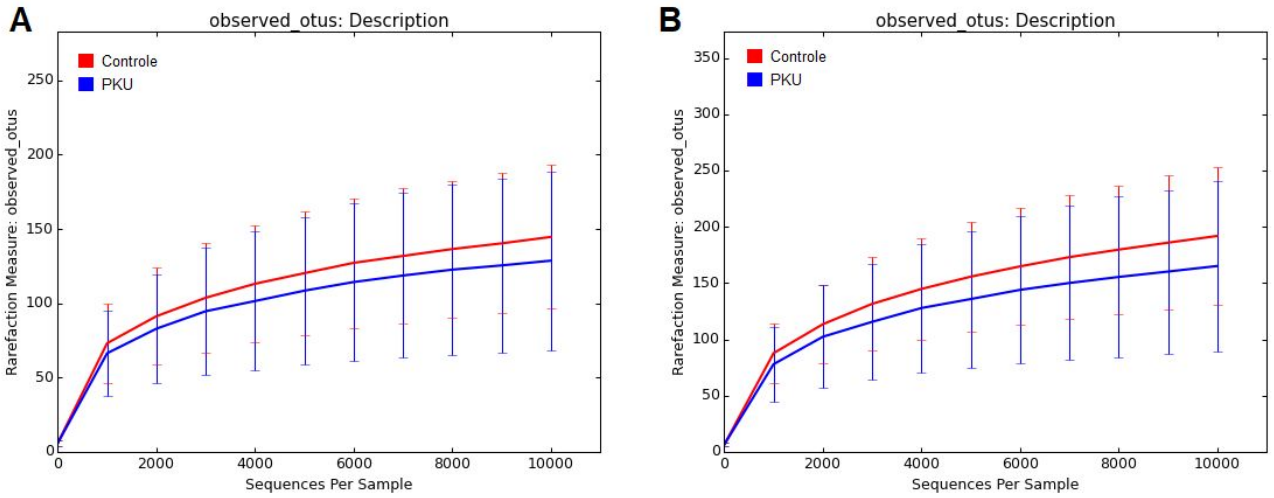

Figura 6 - Curvas de rarefação do número de OTUs observadas por descrição das amostras, sendo os grupos controle e PKU. (A) A partir do *pipeline* do USEARCH. (B) A partir do *pipeline* do VSEARCH.

Os gráficos da análise de coordenada principal (PCoA) possuem como objetivo mensurar a dissimilaridade entre comunidades de dois ambientes, amostras ou grupos, de maneira que, no gráfico, as distâncias entre os pontos são próximas às dissimilaridades originais entre as comunidades microbiológicas.

Foram gerados gráficos *weighted* e *unweighted*. O método *weighted* é uma medida quantitativa que leva em consideração informações de abundância de organismos observados, ou seja, o número de vezes que cada táxon é observado. Já o método *unweighted*, medida qualitativa, observa somente presença ou ausência dos taxa. O primeiro é o mais utilizado para observar diferenças entre comunidades causadas por mudanças na abundância relativa de taxa (Lozupone et al., 2007).

Tendo em vista que o efeito já observado da alteração do USEARCH pelo VSEARCH é o aumento do número de OTUs e, potencialmente, do número de taxa ou da abundância de OTUs representantes de um táxon, o método *weighted* foi utilizado para melhor observação do potencial impacto desta alteração na dissimilaridade entre os dois grupos observados. Apesar da sobreposição entre amostras PKU e controle se manter, o resultado apresentado na **Figura 7** corrobora a variação gerada pela alteração do *pipeline* inicial pela observação da modificação da posição relativa das amostras como um todo, como pode-se notar pelo maior grupamento de amostras (PKU e controle) que deixou de estar posicionado à esquerda para estar à direita, e pelo afastamento e aproximação de diferentes amostras no gráfico,

significando que ocorre alteração da similaridade entre diferentes amostras. Além disso, há considerável variação dos eixos, alterando o perfil de componentes taxonômicos principais e podendo influenciar diretamente em possíveis conclusões que levariam em conta a diferença de diversidade entre amostras.

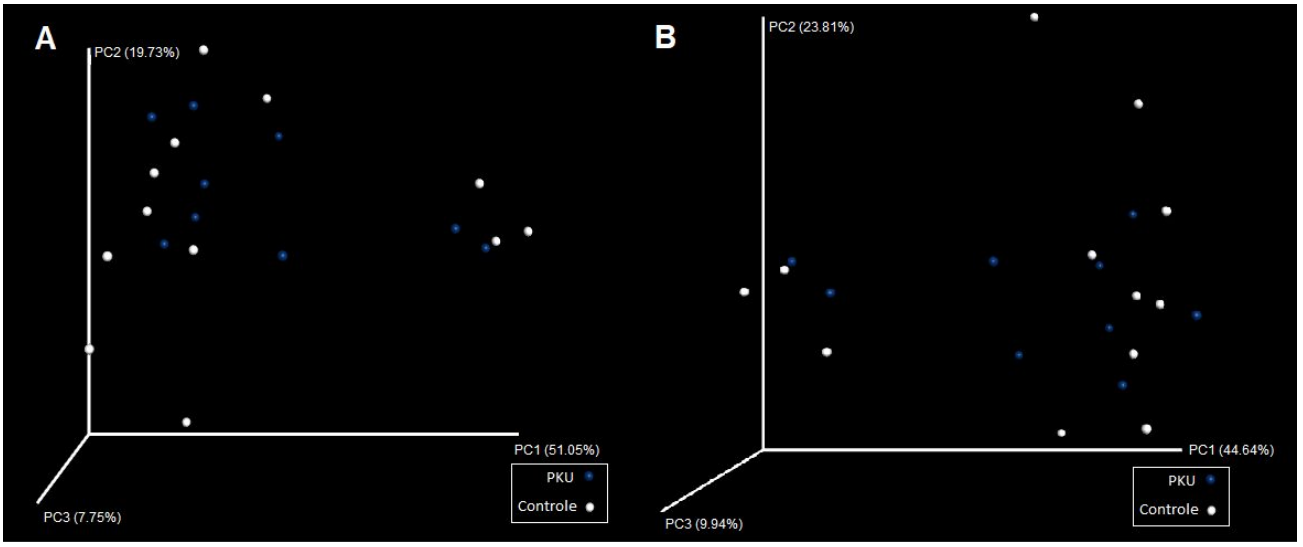

Figura 7 - Gráficos da análise de coordenada principal (PCoA) utilizando o método *weighted* aplicado aos dois grupos de amostras - Controle e PKU. (A) A partir do *pipeline* do USEARCH. (B) A partir do *pipeline* do VSEARCH.

4.2 Classificação taxonômica

4.2.1 Abundância relativa entre filos

Visando compreender o impacto dos parâmetros na precisão da classificação taxonômica, foi avaliada a correspondência dos resultados obtidos com os apresentados por Oliveira et al. (2016) para abundância de filos em amostras controle e de fenilcetonúria (PKU). Para esta validação da consistência, a análise foi realizada a partir do *pipeline* do BMP utilizando parâmetros de classificação *default*.

Os resultados gerados foram consistentes com os apresentados pela referência, sendo os filos *Bacteroidetes*, *Firmicutes* e *Proteobacteria* os três principais em abundância. Pouca diferença foi obtida entre resultados a partir do USEARCH e do VSEARCH, o que se justifica uma vez que a análise é baseada na abundância relativa entre os taxa presentes na *OTU table*.

Destaca-se ainda que no resultado de ambos houve a detecção do filo *Verrucomicrobia* em amostras controle, enquanto no trabalho referência este grupo só é detectado em amostras de PKU. Essa diferença pode ser evidência para presença de falsos positivos na classificação

devido ao baixo índice de confiança tido como *default* pelo BMP (0,5) - utilizado na análise dos dois *pipelines* testados -, já que é o aumento da confiança que reduz a possibilidade de tais falsos positivos e aumenta a confiabilidade do resultado encontrado (Pylro et al., 2014).

4.2.2 Heatmaps e gráficos

Tendo em vista o objetivo deste trabalho de observar o impacto do uso das diferentes combinações dos três parâmetros abordados na classificação e quantificação de taxa, a partir das *OTU tables* geradas para as dezoito possíveis combinações de parâmetros - nove entre diferentes classificadores e índices de confiança para cada uma dos dois *pipelines* iniciais - foram gerados dois *heatmaps*. O primeiro possui como parâmetro o número de diferentes taxa identificados na *OTU table* nos níveis taxonômicos filo, classe, ordem, família e gênero. O segundo considera para esses mesmos níveis taxonômicos o número de OTUs que foram efetivamente classificadas, ou seja, quantas não foram designadas como *unclassified*. O nível taxonômico espécie não foi considerado por em geral possuir baixa acurácia e grande variação.

Considerando que os *heatmaps* representam os dados de todas as dezoito combinações de parâmetros, duas fontes de variação para os resultados observados podem ser citadas: o potencial de classificação - alterado pela variação dos classificadores e dos índices de confiança - e o número de OTUs disponíveis para classificação - influenciado pelo *pipeline* aplicado na primeira etapa da análise, uma vez que, como já discutido previamente, o VSEARCH irá gerar maior número de OTUs a serem consideradas.

Para melhor observação das diferenças apresentadas de forma geral pelos *heatmaps*, foram construídos gráficos de barra para os mesmos parâmetros e considerando dois critérios: índices de confiança extremos superior (0,95) e inferior (0,5) dos dois classificadores passíveis dessa alteração (Mothur e RDP) - permitindo observar mais profundamente o impacto do índice de confiança no resultado obtido -, e a aplicação do índice de confiança *default* aos três classificadores, sendo possível observar a diferença decorrente do classificador aplicado. Os gráficos foram plotados considerando os resultados obtidos para o nível taxonômico de gênero por ser o nível em que maior variação pode ser observada.

O primeiro *heatmap*, considerando o total de taxa identificados, está representado na **Figura 8**. A alteração de classificadores e confiança utilizando o USEARCH não gerou diferença significativa no número de taxa. Além disso, tanto para o Mothur quanto para o

RDP, com o aumento do valor de confiança mínima o número de taxa reduz, conforme o esperado, considerando que o aumento da confiança reduz a possibilidade de falsos positivos e aumenta a confiabilidade do resultado encontrado (Pylro et al., 2014).

Já aplicando-se o VSEARCH, os valores em geral foram maiores, com exceção do nível filo, em que o resultado foram 13 filos identificados com todas as combinações de parâmetros - incluindo entre USEARCH e VSEARCH - exceto na aplicação do Mothur em VSEARCH, cujo resultado foi 14.

A partir desta observação destaca-se os demais resultados obtidos pelo Mothur, valores acima de todas as outras combinações de parâmetros - que se mantiveram na mesma faixa de valores/cores -, sendo essa diferença expressiva no nível taxonômico gênero, em que a classificação pelo Mothur com 0,9 de índice de confiança atingiu o valor máximo observado entre todas as combinações comparadas, sendo ainda maior que o atingido com 0,5. Ou seja, o Mothur foi o único a desviar da expectativa de redução do número de taxa identificados com aumento do limiar de confiança, uma vez que em seu resultado, a redução ocorre do nível 0,5 para 0,8, mas ao atingir 0,9 o valor volta a subir. Ainda não temos uma explicação plausível para isto.

Aplicado a ambos *pipelines* iniciais, o UCLUST apresenta o mesmo padrão de resultado para 0,5 de confiança do que os outros dois classificadores - no caso do VSEARCH somente com o RDP, já que os valores obtidos pelo Mothur fogem do padrão. Isso ocorre em todos os níveis taxonômicos exceto gênero, em que o valor apresentado pelo UCLUST foi mais próximo dos valores obtidos pelos outros utilizando índice 0,95.

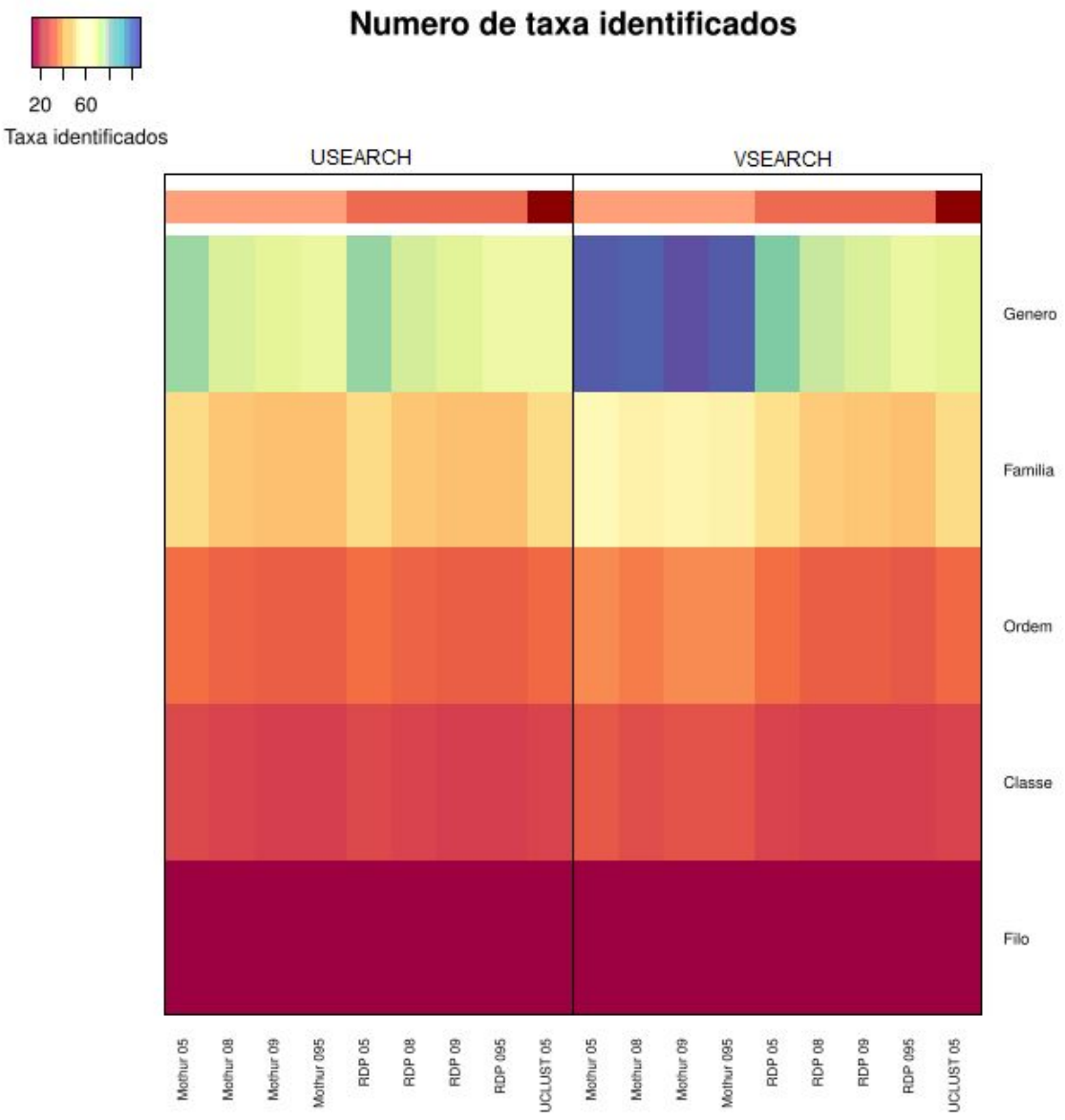

Figura 8 - *Heatmap* representando o resultado de número de taxa identificados em cada nível taxonômico (linhas) a partir da aplicação de cada combinação dos três parâmetros abordados (colunas). À esquerda no *heatmap*, resultados de combinações de classificador e índices de confiança para classificação de OTUs clusterizadas pelo USEARCH. À direita no *heatmap*, resultados para OTUs clusterizadas pelo VSEARCH.

Abaixo está representado o gráfico de barra para o parâmetro de taxa identificados e para os dois *pipelines* aplicados. Nos gráficos da **Figura 9**, sendo aplicado o critério de índices de confiança extremos, o USEARCH, assim como observado no *heatmap*, apresenta mínima diferença numérica entre as combinações de parâmetros, corroborando o resultado apresentado pelo *heatmap* para níveis taxonômicos específicos. Enquanto no VSEARCH o Mothur se destaca tanto por obter maior número de taxa do que o RDP, quanto por não sofrer

redução desse número com o aumento da confiança.

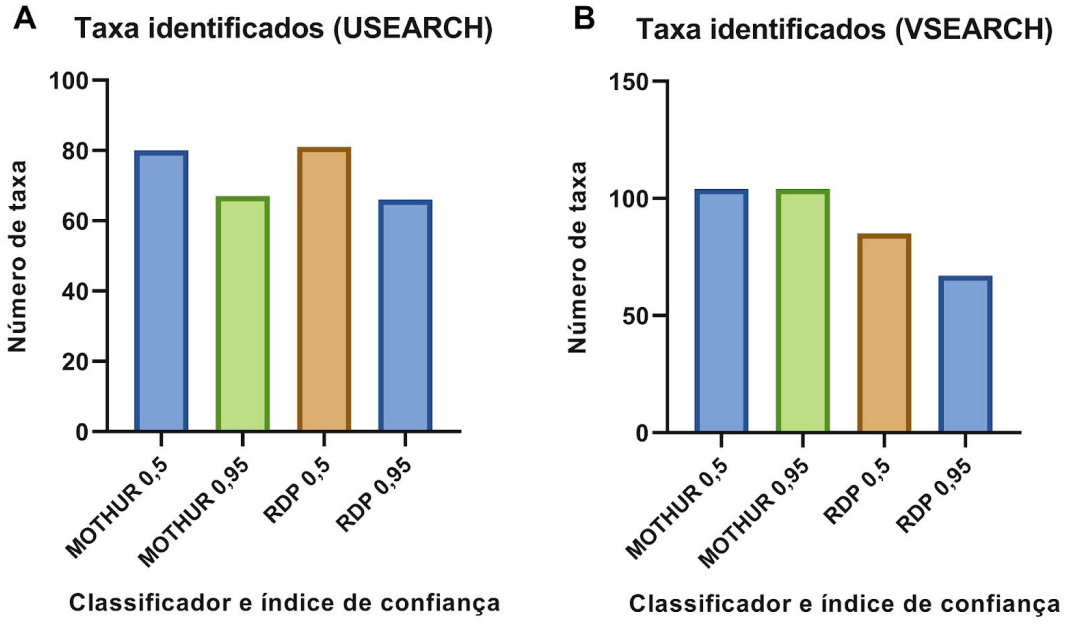

Figura 9 - Gráficos de barra do número de taxa identificados para comparação da aplicação de índices de confiança mínimo e máximo - 0,5 e 0,95 - nos classificadores passíveis desta alteração - Mothur e RDP. (A) Utilizando OTUs geradas a partir do USEARCH. (B) Utilizando OTUs geradas a partir do VSEARCH (note a diferença de escala entre as figuras).

Na aplicação do índice de confiança *default* aos três classificadores a partir do USEARCH, os resultados obtidos pelo Mothur e RDP são diferentes por somente um táxon identificado a mais pelo RDP - como demonstrado pela **Figura 9** -, enquanto o UCLUST se diferencia consideravelmente, detectando mais de dez taxa a menos que os demais, como é visível pelo *heatmap* da **Figura 8**. Aplicando o VSEARCH, o Mothur obteve o maior valor dentre os três, o RDP obteve o segundo maior valor e, assim como observado anteriormente, o UCLUST obteve o menor.

No segundo *heatmap* (**Figura 10**) o efeito do uso do USEARCH ou do VSEARCH se torna explícito uma vez que o impacto desta troca no número de OTUs disponíveis possui relação direta com o critério para geração do *heatmap* - OTUs classificadas -, dessa forma os valores obtidos para todos os níveis taxonômicos foram consideravelmente maiores para o VSEARCH.

Nos resultados pelo USEARCH o mesmo padrão do *heatmap* anterior é apresentado, com pouca diferença entre Mothur e RDP, redução dos resultados obtidos conforme aumento do nível de confiança aplicado, e resultados do UCLUST similares aos obtidos nos outros dois

classificadores com 0,5 de confiança exceto no nível gênero.

Enquanto no VSEARCH, novamente pode-se destacar os resultados apresentados pelo Mothur. Nos níveis taxonômicos filo, classe e ordem, há pouca variação entre os resultados com diferentes índices de confiança, sendo todos igualmente elevados. Em família e gênero um fenômeno diferente é observado, em que com o aumento do índice o número de OTUs classificadas não reduz, mas sim aumenta progressivamente, sendo o inverso do esperado. Enquanto o UCLUST a nível gênero, se aproxima dos resultados obtidos aplicando 0,8 de confiança com classificador RDP, não 0,95 como nos resultados anteriores.

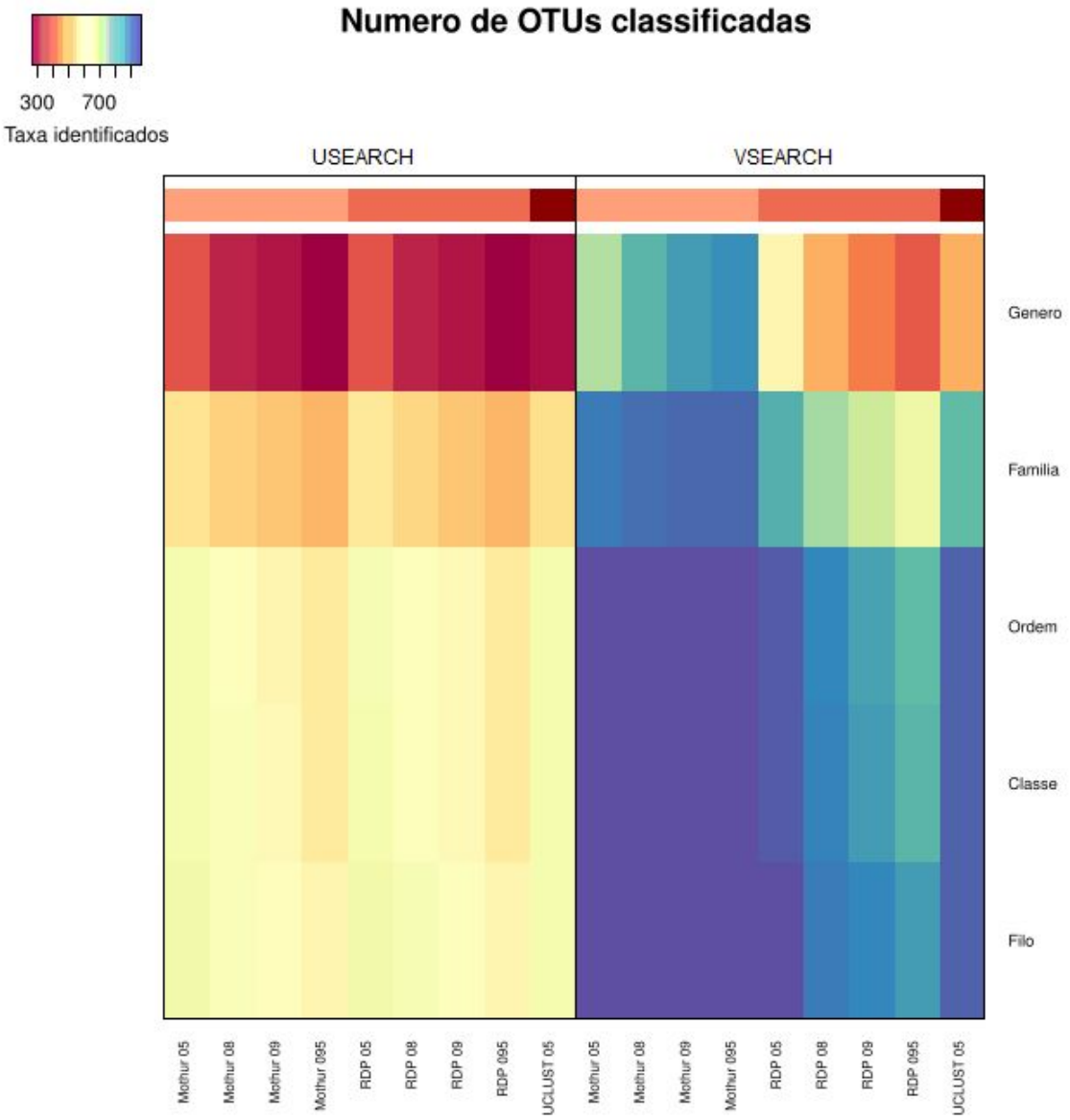

Figura 10 - *Heatmap* representando o resultado de número de OTUs totais classificadas em cada nível taxonômico a partir da aplicação de cada combinação dos três parâmetros abordados. À esquerda no *heatmap*, resultados de combinações de classificador e índices de confiança para classificação de OTUs clusterizadas pelo

USEARCH. À direita no *heatmap*, resultados para OTUs clusterizadas pelo VSEARCH.

Se analisados os gráficos de barra para o primeiro critério (índices de confiança máximo e mínimo)(**Figura 11**) observa-se o mesmo resultado obtido no último conjunto de gráficos para o USEARCH. Para o VSEARCH, confirma-se que o Mothur obtém valores maiores que o RDP para o índice *default*, no entanto, assim como observado no *heatmap*, o valor de OTUs classificadas com 0,95 de confiança supera o obtido para 0,5.

No caso da avaliação dos três classificadores com índice *default* o mesmo padrão observado na análise de taxa identificados se manteve.

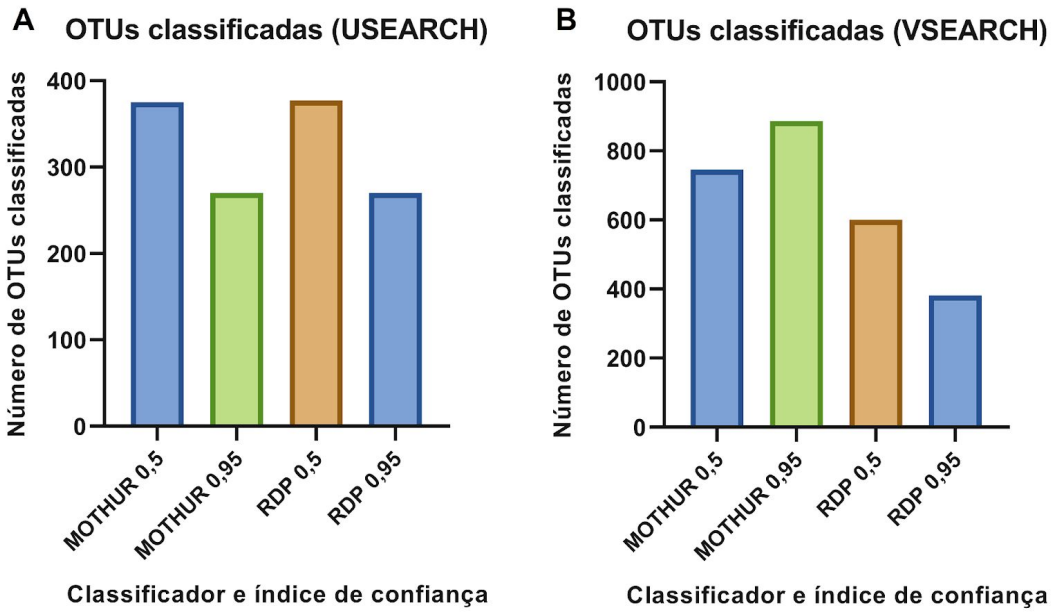

Figura 11 - Gráficos de barra do número de OTUs classificadas para comparação da aplicação de índices de confiança mínimo e máximo - 0,5 e 0,95 - nos classificadores passíveis desta alteração - Mothur e RDP. (A) Utilizando OTUs geradas a partir do USEARCH. (B) Utilizando OTUs geradas a partir do VSEARCH (note a diferença de escala entre as figuras).

A comparação de resultados de índice de taxa identificados e número de OTUs classificadas entre todas as combinações de parâmetros gerou diferenças claras. Para ambos os critérios, a partir do *pipeline* inicial do USEARCH foi atendida a expectativa da redução dos valores resultantes a partir do aumento do índice de confiança, uma vez que a análise passa a ter um critério mais rígido. Além disso, os resultados obtidos pelos classificadores Mothur e o RDP não apresentaram diferença visível, seguindo o mesmo padrão para todos os

índices de confiança aplicados. Essa similaridade é uma possível implicação da implementação do algoritmo utilizado pelo Mothur, sendo este uma reimplementação do classificador bayesiano do RDP (Almeida et al., 2018). Já os parâmetros do *pipeline default* - UCLUST e confiança 0,5 - obtém valores inferiores aos resultados dos outros dois.

Em todas as análises a partir do VSEARCH observa-se que os valores foram notadamente acima dos obtidos pelo USEARCH por conta da disponibilidade de maior número de OTUs para classificação. Os resultados obtidos pelo Mothur foram sempre maiores que os obtidos pelo RDP que, por sua vez, superaram os do UCLUST que, assim como no USEARCH, gera os menores valores dentre os três classificadores. Este resultado contrasta com os apresentados por Golob et al. (2017) em dados das plataformas MiSeq e 454, em que a proporção de OTUs que o Mothur não foi capaz de classificar superou a proporção apresentada pelo *default* do QIIME. No entanto, em Almeida et al. (2018), a partir de conjuntos de dados sintéticos, o Mothur demonstra maior taxa de *recall* se comparado ao UCLUST.

A aplicação do UCLUST no QIIME não permite a alteração do índice de confiança *default* utilizado, por isso, seus resultados tanto para USEARCH quanto para VSEARCH nos níveis taxonômicos filo, classe, ordem e família se assemelham aos obtidos pelos outros dois classificadores utilizando os mesmos 0,5 de confiança. Contudo, em nível de gênero, os resultados em geral se aproximaram dos obtidos nos níveis superiores (0,8 a 0,95).

O Mothur, por sua vez, apesar de manter resultados quase iguais ao RDP quando utilizado o USEARCH, com o VSEARCH obteve resultados muito destoantes dos demais. Seus valores resultantes foram acima de todas as outras combinações testadas. Além disso, seus resultados não atenderam ao previsto de redução dos valores devido ao aumento da confiança, sendo os valores obtidos com 0,95 de confiança iguais ou até maiores que os obtidos utilizando 0,5.

4.2.3 Avaliação de hierarquia taxonômica

O último passo do *pipeline* aplicado foi a utilização do *script checkHierarchy.py* do STAMP (Parks et al., 2014) para conferência da hierarquia taxonômica, ou seja, para identificação de classificações não estritamente hierárquicas e que devem ser manualmente corrigidas a partir da árvore filogenética criada pelo *pipeline*.

Os resultados obtidos foram, em geral, iguais para as árvores geradas a partir do

USEARCH e do VSEARCH. Os classificadores Mothur e RDP apresentam classificações não-hierárquicas, ou seja, em diferentes OTUs o mesmo gênero foi classificado em mais de uma família, enquanto o mesmo não foi apresentado pelo UCLUST. O aumento progressivo do índice de confiança na aplicação dos parâmetros diminui o número destas classificações, porém não elimina totalmente o erro em nenhum dos casos. Dessa forma, os resultados dos dois classificadores se diferenciam pelo número de OTUs classificadas desta forma e pelo impacto do aumento do índice de confiança na redução deste número, no entanto, em ambos os casos a variação entre os dois classificadores é pequena.

Golob et al. (2017) apresenta uma proposta de causa para este problema, sendo a possível presença de sequências duplicadas, mal-anotadas ou mal-sequenciadas em bancos de dados de referência, o que pode contribuir para erros em classificação. Além disso, um mesmo amplicon de 16S pode corresponder a várias entradas de bancos de referência com taxonomias diferentes e a maneira como o classificador lida com tal ambiguidade pode afetar a qualidade dos resultados, sendo ideal aqueles que refletem a ambiguidade mantendo a classificação a níveis taxonômicos mais elevados.

5 CONCLUSÃO E PERSPECTIVAS

A seleção de um *pipeline* e parâmetros apropriados para sua aplicação podem influenciar significativamente nos resultados obtidos para uma análise de determinada comunidade microbiológica. Nesse contexto, o BMP é uma iniciativa criada com o objetivo de unificar e padronizar estudos de microbioma brasileiros, garantindo a reprodutibilidade das pesquisas nele embasadas e, com este objetivo, traz *pipelines* com etapas e parâmetros padronizados e prontos para aplicação, testados em contextos específicos.

A prática de utilizar o método *default* de *pipelines* como o BMP ou o padrão descrito em tutoriais é extremamente comum. No entanto, em seu estudo, Golob et al., 2017 apresenta que os parâmetros sugeridos pelo *pipeline* analisado estiveram entre os de pior performance entre todas os testados. Tendo isso em vista, neste trabalho foram comparados os resultados de todas as combinações entre quatro índices de confiabilidade, três classificadores e dois *pipelines* para limpeza e clusterização de sequências, permitindo obter maior compreensão do papel destes parâmetros no *pipeline* do BMP e visando ressaltar a variação dos índices de taxa identificados e/ou do índice de OTUs classificadas pela alteração de parâmetros. É preciso ressaltar que o objetivo deste estudo não foi determinar qual ferramenta fornece melhor resultado da composição da microbiota, mas sim elucidar o impacto da variação destes parâmetros.

De acordo com os resultados apresentados, pode-se concluir que existem diferenças claras entre a aplicação de diferentes parâmetros ao *pipeline* do BMP, gerando diferentes efeitos na quantidade de taxa identificados, de OTUs classificadas e na precisão da classificação. Além disso, torna-se claro que o passo de remoção de quimeras é capaz de influenciar os resultados obtidos em análises de 16S rDNA de amostras de fezes humanas provenientes da plataforma Ion Torrent PGM, gerando efeitos como um número maior de OTUs detectadas - muitas possivelmente errôneas -, variação da performance de classificadores, estimativas errôneas de diversidade e de diferenças entre populações, como corroborado por Edgar et al. (2011).

Tendo isso em vista, sugere-se a importância da inclusão em estudos de microbioma de detalhes de parâmetros e métodos aplicados, garantindo a validação do resultado. Também é necessária uma maior compreensão por parte dos pesquisadores acerca dos *pipelines* computacionais aplicados e dos efeitos de seus parâmetros no resultado encontrado. Além disso, a realização de análises complementares quanto ao impacto dos demais parâmetros -

como o banco de dados de referência e o uso do QIIME2 -, e comparação com amostras de comunidades de referência (*Mock communities*) torna-se necessária e importante, visando conhecer a influência destes na acurácia de classificação e na quantificação de taxa identificados em análises de enriquecimento de 16S rDNA de amostras de fezes humanas.

6 REFERÊNCIAS

AGNELLO, M. et al. **Microbiome Associated with Severe Caries in Canadian First Nations Children**. Journal of Dental Research, v. 96, n. 12, p. 1378–1385, 2017.

ANDREWS, S. **FastQC: a quality control tool for high throughput sequence data**. Disponível em: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>. Acesso em 26 de maio de 2019.

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

ALMEIDA, A. et al. **Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments**. GigaScience, v. 7, n. 5, 2018.

AUDEBERT, C. et al. **Colonization with the enteric protozoa Blastocystis is associated with increased diversity of human gut bacterial microbiota**. Scientific Reports, v. 6, n. 1, 2016.

BACCI, G. et al. **Evaluation of the Performances of Ribosomal Database Project (RDP) Classifier for Taxonomic Assignment of 16S rRNA Metabarcoding Sequences Generated from Illumina-Solexa NGS**. Journal of Genomics, v. 3, p. 36–39, 2015.

BARKO, P. et al. **The Gastrointestinal Microbiome: A Review**. Journal of Veterinary Internal Medicine, v. 32, n. 1, p. 9–25, 2017.

BELKAID, Y.; HARRISON, O. J. **Homeostatic Immunity and the Microbiota**. Immunity, v. 46, n. 4, p. 562–576, 2017.

BOLYEN, E. et al. **QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science**. 2018.

BRUNO, A. et al. **Food Tracking Perspective: DNA Metabarcoding to Identify Plant Composition in Complex and Processed Food Products**. Genes, v. 10, n. 3, p. 248, 2019.

BUDDEN, K. F. et al. **Emerging pathogenic links between microbiota and the gut–lung axis**. Nature Reviews Microbiology, v. 15, n. 1, p. 55–63, 2016.

CAPORASO, J. G. et al. **Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample**. Proceedings of the National Academy of Sciences, v. 108, n. Supplement_1, p. 4516–4522, 2010.

CAPORASO, J. G. et al. **QIIME allows analysis of high-throughput community sequencing data**. Nature Methods, v. 7, n. 5, p. 335–336, 2010.

CHANG, P. V. et al. **The microbial metabolite butyrate regulates intestinal macrophage function via histone deacetylase inhibition**. Proceedings of the National Academy of Sciences, v. 111, n. 6, p. 2247–2252, 2014.

Chimera Detection in 16S rRNA Sequences. Disponível em: <https://www.ncbi.nlm.nih.gov/genbank/rrnachimera/>. Acesso em 26 de maio de 2019.

Clustering methods. Brazilian Microbiome Project. Disponível em: <https://www.brmicrobiome.org/clusteringmeth>. Acesso em 26 de maio de 2019.

COMEAU, A. M.; DOUGLAS, G. M.; LANGILLE, M. G. I. **Microbiome Helper: a Custom and Streamlined Workflow for Microbiome Research**. mSystems, v. 2, n. 1, 2017.

COMPARE, D. et al. **Gut–liver axis: The impact of gut microbiota on non alcoholic fatty liver disease**. Nutrition, Metabolism and Cardiovascular Diseases, v. 22, n. 6, p. 471–476, 2012.

CRESCI, G. A.; BAWDEN, E. **Gut Microbiome**. Nutrition in Clinical Practice, v. 30, n. 6, p. 734–746, 2015.

DADA, N. et al. **Comparative assessment of the bacterial communities associated with**

Aedes aegypti larvae and water from domestic water storage containers. Parasites & Vectors, v. 7, n. 1, p. 391, 2014.

DEINER, K. et al. **Environmental DNA metabarcoding: Transforming how we survey animal and plant communities**. Molecular Ecology, v. 26, n. 21, p. 5872–5895, 2017.

DESANTIS, T. Z. et al. **Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB**. Applied and Environmental Microbiology, v. 72, n. 7, p. 5069–5072, 2006.

EDGAR, R. C. **Search and clustering orders of magnitude faster than BLAST**. Bioinformatics, v. 26, n. 19, p. 2460–2461, 2010.

EDGAR, R. C. et al. **UCHIME improves sensitivity and speed of chimera detection**. Bioinformatics, v. 27, n. 16, p. 2194–2200, 2011.

EDGAR, R. C. **UPARSE: highly accurate OTU sequences from microbial amplicon reads**. Nature Methods, v. 10, n. 10, p. 996–998, 2013.

FIANNACA, A. et al. **Deep learning models for bacteria taxonomic classification of metagenomic data**. BMC Bioinformatics, v. 19, n. S7, 2018.

FOSTER, J. A.; NEUFELD, K.-A. M. **Gut–brain axis: how the microbiome influences anxiety and depression**. Trends in Neurosciences, v. 36, n. 5, p. 305–312, 2013.

FRANKEL, A. E. et al. **Cancer Immune Checkpoint Inhibitor Therapy and the Gut Microbiota**. Integrative Cancer Therapies, v. 18, p. 153473541984637, 2019.

GILL, S. R. et al. **Metagenomic Analysis of the Human Distal Gut Microbiome**. Science, v. 312, n. 5778, p. 1355–1359, 2006.

GOLOB, J. L. et al. **Evaluating the accuracy of amplicon-based microbiome computational pipelines on simulated human gut microbial communities**. BMC Bioinformatics, v. 18, n. 1, 2017.

GROSICKI, G. J.; FIELDING, R. A.; LUSTGARTEN, M. S. **Gut Microbiota Contribute to Age-Related Changes in Skeletal Muscle Size, Composition, and Function: Biological Basis for a Gut-Muscle Axis**. Calcified Tissue International, v. 102, n. 4, p. 433–442, 2017.

HSU, W. H.; WANG, J. Y.; KUO, C. H. **Current applications of fecal microbiota transplantation in intestinal disorders**. The Kaohsiung Journal of Medical Sciences, 2019.

HU, Y. et al. **The Gut Microbiome Signatures Discriminate Healthy From Pulmonary Tuberculosis Patients**. Frontiers in Cellular and Infection Microbiology, v. 9, 2019.

IIDA, N. et al. **Commensal Bacteria Control Cancer Response to Therapy by Modulating the Tumor Microenvironment**. Science, v. 342, n. 6161, p. 967–970, 2013.

Integrative HMP (iHMP) Research Network Consortium. **The Integrative Human Microbiome Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of Human Health and Disease**. Cell Host Microbe, v. 16, n. 3, p. 276–289, 2014.

LAITINEN, K.; MOKKALA, K. **Overall Dietary Quality Relates to Gut Microbiota Diversity and Abundance**. International Journal of Molecular Sciences, v. 20, n. 8, p. 1835, 2019.

LEONE, V. et al. **Effects of Diurnal Variation of Gut Microbes and High-Fat Feeding on Host Circadian Clock Function and Metabolism**. Cell Host Microbe, v. 17, n. 5, p. 681–689, 2015.

LEY, R. E. et al. **Human gut microbes associated with obesity**. Nature, v. 444, n. 7122, p. 1022–1023, 2006.

LI, W.; GODZIK, A. **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences**. Bioinformatics, v. 22, n. 13, p. 1658–1659, 2006.

LOZUPONE, C. A. et al. **Quantitative and Qualitative Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities**. Applied and

Environmental Microbiology, v. 73, n. 5, p. 1576–1585, 2007.

LÓPEZ-GARCÍA, A. et al. **Comparison of Mothur and QIIME for the Analysis of Rumen Microbiota Composition Based on 16S rRNA Amplicon Sequences**. Frontiers in Microbiology, v. 9, 2018.

MARTÍN, R. et al. **Role of commensal and probiotic bacteria in human health: a focus on inflammatory bowel disease**. Microbial Cell Factories, v. 12, n. 1, p. 71, 2013.

MARUVADA, P. et al. **The Human Microbiome and Obesity: Moving beyond Associations**. Cell Host Microbe, v. 22, n. 5, p. 589–599, 2017.

MASLOWSKI, K. M. et al. **Regulation of inflammatory responses by gut microbiota and chemoattractant receptor GPR43**. Nature, v. 461, n. 7268, p. 1282–1286, 2009.

The New Science of Metagenomics. 2007.

NIPPERESS D.A. **The Rarefaction of Phylogenetic Diversity: Formulation, Extension and Application.** Biodiversity Conservation and Phylogenetic Systematics. Topics in Biodiversity and Conservation, vol 14. Springer, Cham. 2016.

OLIVEIRA, F. P. D. et al. **Phenylketonuria and Gut Microbiota: A Controlled Study Based on Next-Generation Sequencing**. Plos One, v. 11, n. 6, 2016.

PARKS, D. H. et al. **STAMP: statistical analysis of taxonomic and functional profiles**. Bioinformatics, v. 30, n. 21, p. 3123–3124, 2014.

PEPPER, J. W.; ROSENFELD, S. **The emerging medical ecology of the human gut microbiome**. Trends in Ecology & Evolution, v. 27, n. 7, p. 381–384, 2012.

PRUESSE, E. et al. **SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB.** Nucleic Acids Research, v. 35, n. 21, p. 7188–7196, 2007.

PYLRO, V. S. et al. **Data analysis for 16S microbial profiling from different benchtop**

sequencing platforms. Journal of Microbiological Methods, v. 107, p. 30–37, 2014.

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, 2012.

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

SCHLOSS, P. D. et al. **Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities.** Applied and Environmental Microbiology, v. 75, n. 23, p. 7537–7541, 2009.

SEEKATZ, A. M. et al. **Differential Response of the Cynomolgus Macaque Gut Microbiota to Shigella Infection**. PLoS ONE, v. 8, n. 6, 2013.

SELBER-HNATIW, S. et al. **Human Gut Microbiota: Toward an Ecology of Disease**. Frontiers in Microbiology, v. 8, 2017.

SENDER, R.; FUCHS, S.; MILO, R. **Revised Estimates for the Number of Human and Bacteria Cells in the Body**. PLOS Biology, v. 14, n. 8, 2016

SONG, J. et al. **Analysis of microbial diversity in apple vinegar fermentation process through 16s rDNA sequencing**. Food Science & Nutrition, v. 7, n. 4, p. 1230–1238, 2019.

TAN, C. K. et al. **Comparative study of gut microbiota in wild and captive Malaysian Mahseer (*Tor tambroides*).** MicrobiologyOpen, v. 8, n. 5, 2018.

TANOUE, T. et al. **A defined commensal consortium elicits CD8 T cells and anti-cancer immunity**. Nature, v. 565, n. 7741, p. 600–605, 2019.

TURNBAUGH, P. J. et al. **The Human Microbiome Project**. Nature, v. 449, n. 7164, p. 804–810, 2007.

TURNBAUGH, P. J. et al. **An obesity-associated gut microbiome with increased capacity**

for energy harvest. Nature, v. 444, n. 7122, p. 1027–1031, 2006.

VIEIRA-SILVA, S. et al. **Species–function relationships shape ecological properties of the human gut microbiome.** Nature Microbiology, v. 1, n. 8, 2016.

WANG, Q. et al. **Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy.** Applied and Environmental Microbiology, v. 73, n. 16, p. 5261–5267, 2007.

WEINER, H. L. et al. **Oral tolerance**. Immunological Reviews, v. 241, n. 1, p. 241–259, 2011.