ESTIMAÇÃO IN-SILICO DO MICROBIOMA HUMANO EM CÂNCER GÁSTRICO

RODRIGO LUCENA BORGES

Dissertação apresentada à Fundação Antônio Prudente para a obtenção do título de Mestre em Ciências.

Área de concentração: Oncologia

Orientador: Dr. Israel Tojal da Silva, PhD

São Paulo

2021

RESUMO

Borges R. **Estimação in-silico do microbioma humano em câncer gástrico**. [Dissertação]. São Paulo; Fundação Antônio Prudente; 2021.

O microbioma pode influenciar a fisiologia humana, tanto na saúde quanto na doença. O uso das tecnologias de sequenciamento de última geração ajudou a ampliar o conhecimento a cerca de vários micro-organismos. No entanto, a detecção da diversidade em larga escala ainda apresenta resultados conflitantes devido a heterogeneidade dos protocolos e dos métodos de análise. Aqui, foi projetado e implementado um protocolo computacional para ser incorporado à *facility* de Bioinformática do A.C. Camargo Cancer Center para a detecção do microbioma humano em câncer gástrico. Avaliamos também os principais aspectos que possam influenciar o resultado final considerando o protocolo de sequenciamento, parametrização e algoritmos de classificação. O pipeline desenvolvido utiliza uma estrutura de gerenciamento de fluxo de trabalho (*Snakemake*), que permite obter uma análise automatizada e reprodutível, independente do ambiente de execução. Usando os resultados do sequenciamento de uma amostra controle com diversidade conhecida, mostramos que o pipeline proposto identifica com acurácia (TAR =0.89) os organismos a nível de gênero, bem como uma taxa de detecção sem falsos negativos (TDR=1). Como prova de conceito, usamos o pipeline em um conjunto de amostras provenientes de tumor primário e suco gástrico. De importância, ao avaliar como as várias etapas da análise podem influenciar os resultados, observamos que a combinação dos algoritmos *deblur* e *BLAST* durante a limpeza das leituras e classificação, respectivamente, apresentam os resultados mais próximos do esperado em relação à amostra controle. Diante da importância do microbioma em estudos de pesquisa básica e translacional, o pipeline proposto será útil na detecção da diversidade de bactérias em um ambiente computacional.

ABSTRACT

Borges R. **[In-silico estimation of the human cancer gastric]**. [Dissertation]. São Paulo; Fundação Antonio Prudente; 2021.

The human microbiome has a great potential to influence human physiology, both in health and in disease. The use of state-of-the-art sequencing technologies has helped to extend knowledge about microorganisms. However, diversity detection in large scale basis still presents conflicting results due to the heterogeneity of protocols and analysis methods. Here, a computational protocol was designed and implemented to be incorporated into the A.C. Camargo Cancer Center's Bioinformatics facility in order to detect the human gastric cancer microbiome. We also evaluate the main aspects that can influence final results according to the characteristics of the sequencing protocol, parameterization and classification algorithms. The developed pipeline uses a workflow management structure, called Snakemake, which allows automated and reproducible analysis, regardless of the execution environment. Using the sequencing oa control sample with known diversity, we show that the proposed pipeline accurately identifies organisms at the gender level (TAR = 0.89), as well as a detection rate without false negatvies (TDR = 1). As a proof of concept, we used the pipeline in a set of samples from primary tumor and gastric juice. Of importance, when evaluating how the various stages of the analysis can influence the results, we observed that the combination of the deblur and BLAST algorithms during the denoising and classification steps, respectively, presents the results closer to the expected regarding the control sample. Given the importance of the microbiome in basic and translational research studies, the proposed pipeline will be useful in detecting the microbiome diversity in a computational environment.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

ANCOM	*Analysis of composition of microbiomes*
ASV	*Amplicon sequence variant*
BLAST	*Basic Local Alignment Search Tool*
BWA	*Burrows-Wheeler Aligner*
OTU	*Operational taxonomic unit*
PCA	*Principal component analysis*
PCR	*Polymerase chain reaction*
ROE	*Quantidade Observada/Quantidade Esperada*
TAR	*Taxon Accuracy Rate*
TDR	*Taxon Detection Rate*

SUMÁRIO

1 INTRODUÇÃO

A evolução na área de sequenciamento de DNA em larga escala levou a uma redução significativa dos custos permitindo, assim, a prática dessa tecnologia em diferentes áreas de pesquisa básica e na prática clínica. Esse avanço, por sua vez, trouxe uma mudança de paradigma ao permitir uma revolução na medicina personalizada e de precisão, cada vez mais presente nos dias atuais.

Uma das áreas que sofreu grande impacto foi o estudo de micro-organismos presentes no corpo humano. Hoje, é possível pelo sequenciamento do material genético dos micro-organismos (microbioma) identificar com alta precisão a diversidade desse microbioma presente em indivíduos saudáveis e doentes.

1.1 MICROBIOMA

O material genético dos organismos encontrados no hospedeiro compõem o **microbioma**, sejam eles dos reinos *Bacteria*, *Archea*, *Fungi*, *Protozoa* ou os parasitas intra-celulares, como os vírus. A população de micro-organismos presentes em um hospedeiro compõem a **microbiota** (Shreiner et al. 2015).

Uma variedade de animais, incluindo os seres humanos, são habitados por comunidades de organismos essenciais para sua forma e função normal (Barko et al. 2018; Huttenhower et al. 2012). Em termos de composição celular, diversidade genética e capacidade metabólica, os animais poderiam ser considerados como um organismo híbrido multi-espécie composto pelo hospedeiro e células microbianas operando de forma dinâmica e simbiótica (Turnbaugh et al. 2007).

Estudos mostram que o número de genes associados ao microbioma é de 2 a 20 milhões, enquanto que os humanos têm aproximadamente 20 mil genes (Turnbaugh et al. 2007; Qin et al. 2010). Já foi estimado que as células microbianas superam o número de células humanas em uma razão de 10:1 de células microbianas em relação às humanas (Bianconi et al. 2013), mas um trabalho revisou e compilou alguns estu-

dos e atualizou a razão entre células bacterianas e humanas para aproximadamente 1:1 (Sender et al. 2016).

Por toda essa complexidade, a microbiota desempenha um importante papel no hospedeiro, interagindo indiretamente com a grande maioria das suas células. De particular interesse da comunidade científica, somente em 2017, aproximadamente 4000 artigos foram publicados onde o tema principal foi o estudo da microbiota humana e, ao olhar o intervalo de 2013 a 2018, mais de 12900 artigos foram publicados (Cani 2018). Esses números demonstram que, não somente é uma área de estudo relevante mas, principalmente, encontra-se em desenvolvimento.

1.2 A MICROBIOTA NA SAÚDE E NA DOENÇA

Adultos saudáveis hospedam mais de 1000 espécies de bactérias pertencentes, em sua maioria, aos filos *Bacteroidetes* e *Firmicutes* (Lozupone et al. 2012). Pesquisas mostram que 70% das espécies da população microbiota humana de indivíduos considerados saudáveis e, sem uso de antibióticos, podem permanecer estáveis por mais de um ano e poucas diferenças foram encontradas no período de cinco anos de acompanhamento. Observou-se, ainda, que as espécies são muito semelhantes entre indivíduos da mesma família mas não entre indivíduos não relacionados (Shreiner et al. 2015). Esses achados indicam que existe um papel complexo e ainda não muito conhecido do microbioma em indivíduos saudáveis.

O ambiente do estômago era dado como inóspito para os microrganismos devido a suas propriedades ácidas e outros fatores anti-microbiais, porém, com a descoberta da H.pylori e consequentes estudos percebeu-se que o ambiente gástrico contém DNA ribosomal de uma diversa variedade de bactérias como mostra o estudo de (Bik et al. 2006) que reportou 128 filotipos entre oito filos mostrando que a diversidade microbiana do sistema gástrico é diversa. O perfil microbial de fluidos gástricos tem uma diversidade maior comparada à amostras de mucosa gástrica e, além disso, mostram uma composição de filos diferentes (H. pylori, 66.5% vs. 3.3%, P = 0.033; Proteobac-

teria, 75.4% vs. 26.3%, P = 0.041 (Sung et al. 2016).

De particular importância, a microbiota gastro-intestinal é essencial para a nutrição e digestão dos hospedeiro auxiliando, por exemplo, na digestão de substratos indigestos aos humanos. Um desses substratos é o *x*iloglucano (presente na batata), uma fibra importante que ajuda no controle do colesterol e da glicose no sangue (Jenkins et al. 1978; Lairon 1996). Esse papel desempenhado pela microbiota demonstra a importância da simbiose. De maneira complementar, a microbiota também é capaz de influenciar a fisiologia humana de muitas outras formas. Por isso é relevante o estudo das mudanças da microbiota associada aos estados das doenças, essas mudanças são chamadas de disbiose. Atualmente, o relacionamento entre a disbiose e a patogenicidade da doença não é bem conhecida para a maioria das enfermidades. Nesse cenário, um dos grandes desafios é descobrir quais as mudanças da microbiota são associadas às doenças e a distinção entre causa e efeito (Shreiner et al. 2015).

Enquanto é intrigante especular que a disbiose esteja associada com doenças, também é importante notar que o estado da doença pode levar a mudanças da microbiota através de vários mecanismos, incluindo mudanças nos hábitos alimentares e na função do trato gástrico-intestinal bem como a administração de medicações, principalmente antibióticos (Shreiner et al. 2015).

Podemos ressaltar algumas das descobertas recentes sobre o papel da microbiota do trato gastro-intestinal e sua influencia em i) doenças cardiovasculares, ii) síndrome do intestino irritável iii) infecção por *Clostridium difficile*, iv) doenças inflamatórias intestinais e v) Psoríase (Benhadou et al. 2018)

1.3 MICROBIOTA E CÂNCER

De importância, por ano são diagnosticados, em média, 130 pacientes com câncer gástrico (CG) no ACCCC. As cirurgias ocorrem em cerca de 15% dos pacientes com diagnóstico precoce. Pacientes com tumores no estadio cT3/T4 em conjunto com N1,2,3 usualmente recebem tratamentos de quimioterapia e cirurgia após a laparosco-

pia. A resposta patológica completa (sem resíduo tumoral) é observada em 12% dos pacientes. A mediana de acompanhamento é de 41 meses e a mediana de sobrevida desde a quimioterapia adjuvante é de 73 meses. A análise de sobrevida global para 65% dos casos é de 5 anos (Bartelli et al. 2019). O câncer gástrico é um dos tipos de câncer mais comuns e está em quinto lugar em relação a mortalidade no mundo (Rawla e Barsouk 2019). No Brasil, é o tipo de câncer com maior mortalidade, em quarto lugar entre homens e o sexto lugar entre as mulheres (Câncer 2018). Esses números deixam claro que apesar da evolução da medicina personalizada e dos diagnósticos moleculares, ainda é necessário uma melhor compreensão de outros aspectos dessa doença.

Uma promissora abordagem no tratamento de pacientes com câncer é a imunoterapia, que modula a atividade de componentes do sistema imune e, atualmente, está transformando o tratamento do câncer. Recentes descobertas mostram que a microbiota influencia a resposta imune (Paulos et al. 2007). Dessa forma, o perfil molecular dos microrganismos presentes no trato gástrico pode representar uma oportunidade de identificação de preditores de resposta a tratamentos ou servir como coadjuvante aos desfechos clínicos. Nos tumores gástricos, por exemplo, a microbiota pode desempenhar um papel importante na resposta imunológica frente às diferentes terapias anti-tumorais. (Iida et al. 2013).

O sequenciamento do material genético, em larga escala, tem permitido a detecção de micro-organismos já conhecidos e também desconhecidos com uma alta sensibilidade (Paulos et al. 2007). Embora a detecção do microbioma representa um potencial no estudo de doenças, as abordagens analíticas ainda não estão consolidadas e é alvo de intensa investigação.

Embora o câncer seja usualmente associado à genética do indivíduo e a fatores ambientais, os micro-organismos são responsáveis por, aproximadamente, 20% das malignidades. Entre outras várias bacterias tendo associação com o câncer humano, apenas a *Helicobacter pylori* provou-se como carcinogênica pela agência de Pesquisa em Câncer (*Agency for Research on Cancer* - IARC) pois causa a gastrite, que por sua veza causa a úlcera, causando atrofia e, finalmente, o câncer gástrico. A infecção por

H. pylori não somente aflige as células do estomago, mas também altera o ambiente gástrico e sua barreira, aumentando a inflamação e alterando a microbiota, por outro lado, diminui o risco de adenocarcinoma esofágico em humanos, o que enfatiza os efeitos da microbiota na carcinogênese (Schwabe e Jobin 2013).

1.4 MICROBIOTA E IMUNOLOGIA

A interação imunológica entre a microbiota e os diferentes sistemas do metabolismo reconhecem e monitoram a presença de micro-organismos presentes no corpo (patógenos). Por exemplo, no trato gastro-intestinal (GI) as células epiteliais têm um papel fundamental como guardiãs que traduzem informações importantes para as células do sistema imune localizadas na *lamina propria* (Cani 2018).

As interações entre o microbioma gastro-intestinal e o sistema imune levaram a descoberta de funções previamente desconhecidas a respeito de como os componentes específicos da microbiota contribuem com os níveis de glucose e a homeostase lipídica (Duparc et al. 2017). Em 2007 foi identificado que constituintes de bactérias gram-negativas, como as *lipopolysaccharides* (LPS), eram o principal gatilho para a ativação de inflamações de baixo nível e resistência a insulina por mecanismos de interação entre a microbiota e o sistema imune inato (como o TLR-4 e CD14) (Smith et al. 2013). Os modelos genéticos tanto de obesidade induzida por dieta quanto de diabetes foram caracterizados por um aumento nos níveis de LPS circulante, uma condição chamada endotoxemia metabólica em humanos (Cani 2018).

Em adição às mudanças específicas na composição da microbiota intestinal, agora é bem aceito que vários fatores-chave contribuem para a translocação de componentes bacterianos do lúmen intestinal para o corpo. Notavelmente, a perda da tolerância imune está associada com a inflamação intestinal e dados recentes mostram que indivíduos obesos que consomem uma dieta rica em gorduras têm acúmulo de células T nos intestinos grosso e delgado, mostrando uma relação com a morbidade (Magalhaes et al. 2015).

Os meios pelos quais a microbiota contribui com a carcinogênese, seja por diminuir ou aumentar o risco do hospedeiro, podem ser resumidos em três categorias: (I) influenciando o metabolismo do hospedeiro, pelos insumos ingeridos e fármacos, (ii) alterando o equilíbrio entre proliferação celular e morte celular e (iii) guiando o função do sistema imune.

Muitas bactérias evoluíram seus mecanismos para causar danos ao DNA no intuito de matar os competidores e sobreviver. Esses mecanismos podem levar a eventos mutacionais que contribuem com a carcinogênese, como por exemplo, a Escherichia coli. Além de danificar o DNA, vários micróbios possuem proteínas que ativam vias metabólicas do humano envolvidas na carcinogênese, como a via canônica de Wnt. Outras bactérias, como a H. pylori causam proliferação celular, aumentando o risco de angiogênese. Outras bactérias secretam fatores que contribuem para o crescimento tumoral que regulam e ativam receptores como TLR e NLR causando inflamações crônicas mediadas por NF-kB e STAT3 e, potencialmente, causar a carcinogênese (Garrett 2015).

Para citar um exemplo de interação da microbiota com o desenvolvimento de câncer, temos o exemplo do câncer colorretal. Ao ocorrer a disbiose bactéria benéficas decaem e as prejudiciais aumentam, causando inflamação. Uma dessas bactérias prejudiciais é a *Fusobacterium nucleatum* (Fn). Foi demostrado que moléculas presentes em sua superfície quebram a barreira intestinal facilitando o contato dessa bactéria nas células epiteliais ativando cascatas pró-inflamatórias tornando o ambiente tumoral propício à ocorrência de CCR. Outra molécula também presente nesta bactéria ativa a via de sinalização E-caderina/β-catenina, que também contribui para a proliferação de células tumorais. Fn pode também modular a autofagia das células epiteliais intestinais ativando microRNAs reguladores, que estão correlacionados com mecanismos de quimiorresistência a quimioterápicos (Carvalho et al. 2019), tal interação de Fn pôde ser relacionada com a sobrevida de pacientes (Figura 1).

Em suma, a barreira intestinal é controlada por uma sensível comunicação que ocorre entre a microbiota intestinal e o sistema imune do hospedeiro. A complexidade

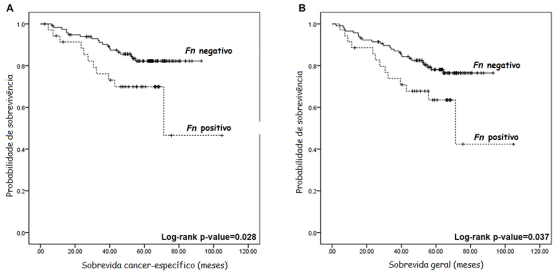


Figura 1 – Curvas de sobrevida Kaplan–Meier para câncer de colorretal (CCR) específico e sobrevida geral de acordo com a detecção de DNA de *Fusobacterium nucleatum em tecido colorretal*. (A) Curva de sobrevida de cinco anos de acompanhamento de cancer-específico foi de 69.9% para Fn positivo e 82.2% para Fn negativo. (B) Curva de sobrevida de cinco anos de acompanhamento geral foi de 63.5% para Fn positivo e 76.5% para Fn negativo (Carvalho et al. 2019).

dessas interações levantam a questão sobre o nível da compreensão que temos e eventualmente pode contribuir para explicar por que é relativamente desafiador desenvolver alvos terapêuticos específicos, como por exemplo espécies (ou variações) resistentes a antibióticos.

1.5 A IDENTIFICAÇÃO MICROBIANA

Regiões genômicas altamente conservadas durante a evolução são ferramentas muito utilizadas para detecção e comparação de organismos. Existem alguns genes que são exemplos dessas regiões, como os genes que definem os RNAs ribossomais (rRNA). A marioria dos procariotos contém três rRNAs, o 5S, 16S e 23S, e tais regiões podem ter uma ou mais cópias no genoma. Essas regiões têm tamanhos diferentes, tendo a região 5S 120bp, 16S 1500bp e a região 23S 2900bp (Stiegler et al. 1981).

Analisar tais genes permite a identificação e a comparação dos organismos encontrados nas amostras.

1.5.1 Taxonomia

Taxonomia é um termo grego derivado das palavras *taxis*, que significa ”arranjo”, e *nomus* que significa ”lei”e é utilizado para classificar biologicamente os organismos de uma forma hierárquica (níveis) (Cain 2020).

Dependendo da similaridade entre as espécies, nem sempre é possível definir com significância qual é o organismo presente quando sequenciamos apenas a região 16S, e por isso, a classificação em níveis taxonômicos menos específicos é utilizada (Figura 2).

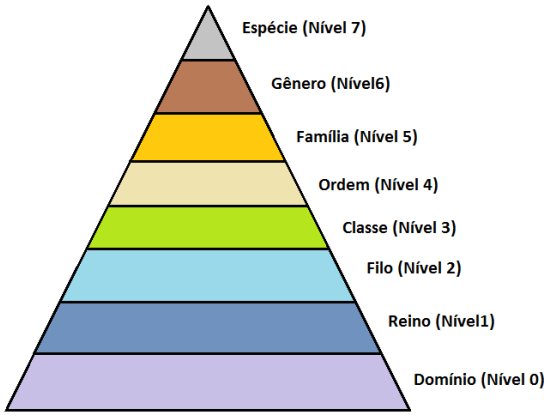


Figura 2 – O nível 7 de classificação taxonômica é o mais específico e o nível 0 (zero) é o menos específico.

1.5.1.1 A região 16S

Com o sequenciamento das bases do 16S RNA ribossomal (rRNA) é possível identificar micro-organismos e classifica-los taxonomicamente, analisar a filogenia e a diversidade do microbioma. O RNA ribossomal 16S (ou 16S rRNA) é o componente da subunidade 30S do ribossomo procariótico que liga-se com o sítio de ligação *Shine-Dalrgarno* presente nas *bactérias* e *arqueas*. Esse sítio de ligação recruta os ribossomos para o mensageiro RNA (mRNA), necessário para iniciar a síntese proteica.

Os genes que codificam o 16S são utilizados na reconstrução de filogenias devido à baixa taxa de evolução, no entanto, existem algumas regiões com alta variabilidade denominadas de V[1-9] e representados na Figura 3, onde o grau de conservação varia e seu sequenciamento permite a identificação da taxonomia em diferentes níveis, como espécie, família e gênero dos micro-organismos (Chakravorty et al. 2007); (Yang et al. 2016),(Woese e Fox 1977).

As regiões que flanqueiam as regiões variáveis V[1-9] são conservadas e permitem, assim, o desenho de *primers* para sua amplificação por meio da técnica de *P*CR,

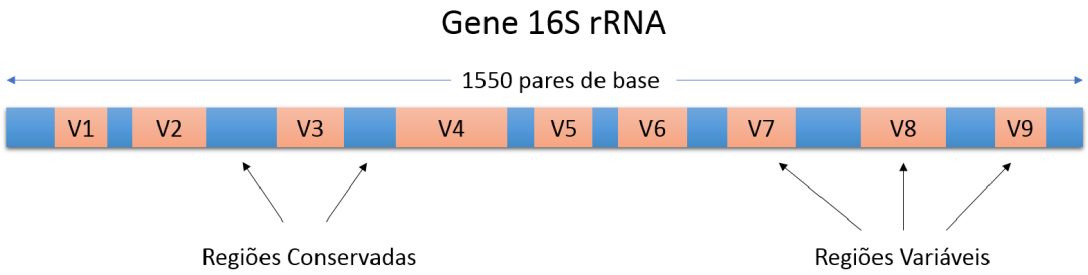


Figura 3 – As regiões conservadas encontram-se destacadas em azul enquanto as regiões variáveis estão destacadas em laranja V(1-9).

favorecendo o sequenciamento do DNA de alto desempenho NGS (Chakravorty et al. 2007). Como resultado, essas regiões sequenciadas podem ser analisadas pelas ferramentas analíticas para monitorar e estudar esses organismos (DeSantis et al. 2006).

As regiões pouco conservadas do 16S são geralmente utilizadas para detectar a presença de patógenos específicos, mas não são úteis para diferenciação de espécies. Como exemplo, a região V3 apresenta uma maior variabilidade a nível de gênero, enquanto a região V6 seria a mais assertiva para diferenciar espécies específicas (Chakravorty et al. 2007).

As regiões muito variadas do 16S rRNA são utilizadas em estudos taxonômicos, mas são inviáveis para diferenciar espécies muito relacionadas. A região V4 tem pequenas diferenças que podem chegar a apenas uma base de nucleotídeo, o que limita a busca em bancos de dados nos quais não existe um detalhamento tão profundo. Essa dificuldade pode agrupar as sequências de forma equivocada, e assim, subestimar a diversidade da amostra (Vëtrovsky e Baldrian 2013).

1.5.2 Desafios na detecção utilizando a região 16S

A detecção do microbioma de uma amostra representando uma microbiota ocorre por meio da amplificação do gene 16S rRNA seguido do sequenciamento do DNA. Após o pré-processamento das leituras sequenciadas, existe um importante processo de limpeza (*denoising*) que levam em consideração os possíveis erros que são inerentes às tecnologias de sequenciamento. Esses erros, por sua vez, podem levar a uma inferência incorreta da taxonomia do organismo sequenciado, nesse intuito alguns al-

goritmos foram desenvolvidos para corrigir esses erros como o *deblur* (Amir et al. 2017) e *Dada2* (Callahan et al. 2016). Em seguida, é possível fazer comparações entre grupos de amostras (casos versus controle, por exemplo) e classificar as sequências de acordo com sua taxonomia.

Outros desafios em relação a análise do microbioma utilizando a região 16S é a reprodutibilidade dos resultados, em particular os kits de extração de DNA, a construção de bibliotecas e até mesmo os equipamentos laboratoriais podem influenciar no resultado final devido a alta sensibilidade do método (Goodrich et al. 2014). Por isso, é imprescindível avaliar criteriosamente cada etapa da análise para obtenção de parâmetros críticos e modelar adequadamente todo o processo desde a obtenção das amostras até a análise dos resultados.

1.5.3 Métricas para avaliação da diversidade

Apesar dos desafios em relação à classificação taxonômica, saber exatamente a espécie presente no microbioma não é crucial dependendo da pergunta biológica, uma vez que as análises que medem a diversidade das amostras podem utilizar a filogenia entre as sequências, mesmo sem classifica-las taxonomicamente ou em níveis menos específicos de taxonomia e, desta forma, é possível diferenciar um organismo de outro.

O conhecimento acerca da diversidade da microbiota em uma amostra pode trazer informações importantes em estudos comparativos, por exemplo, em amostras de diferentes grupos (tratado versus não-tratado) ou avaliação da dinâmica do perfil microbiano ao longo de dias de tratamento. Particularmente, duas métricas são utilizadas, incluindo a diversidade *alfa* para mensurar a diversidade em uma amostra enquanto a diversidade beta é usada para mensurar a diversidade entre amostras.

1.5.3.1 Diversidade alfa

De particular importância, torna-se necessário avaliar a similaridade entre os organismos de uma dada amostra. Nesse contexto, dois índices são utilizados, enquanto o índice de Simpson (Hunter e Gaston 1988) representa a menor diversidade

quanto maior o seu valor, o índice de Shannon (Konopiński 2020) aumenta conforme a diversidade e, adicionalmente, leva em conta tanto a riqueza quanto a proporção das espécies.

Usualmente, para estimar a diversidade, utiliza-se a métrica de Shannon. Por outro lado, o índice de Simpson é utilizado para estimar a dominância de alguma espécie (Srivastava 2015).

A equalização de espécies (*Evenness*), leva em conta o i) índice de Shannon, ii) a quantidade de espécies e o iii) numero de indivíduos na amostra. Ao final, calcula a abundância relativa entre diferentes espécies e, quanto maior a equalização da abundância entre os organismos presentes em uma amostra, maior o *Evenness* (Pielou 1966).

De comum interesse, utiliza-se a diversidade filogenética para avaliar a biodiversidade genética considerando suas relações filogenéticas. Faith (Faith 1992) definiu como diversidade filogenética a soma dos tamanhos dos galhos de uma árvore filogenética. Essa abordagem permite medir o quanto as taxas encontradas em uma amostra são conservadas entre si. Na Figura 4 é ilustrado uma árvore filogenética e o respectivo cálculo de sua diversidade(Faith e Richards 2012).

1.5.3.2 Diversidade beta

Para lidar com a quantidade de informação e agrupar as amostras automaticamente, é comumente utilizada a redução de dimensão por meio da análise de componentes principais - PCA, *Principal Component of Analysis* (Hill et al. 1965). Esse método possibilita explorar e visualizar similaridade dos dados de uma forma mais intuitiva, como por exemplo um gráfico 2D ou 3D como ilustrado na Figura 5.

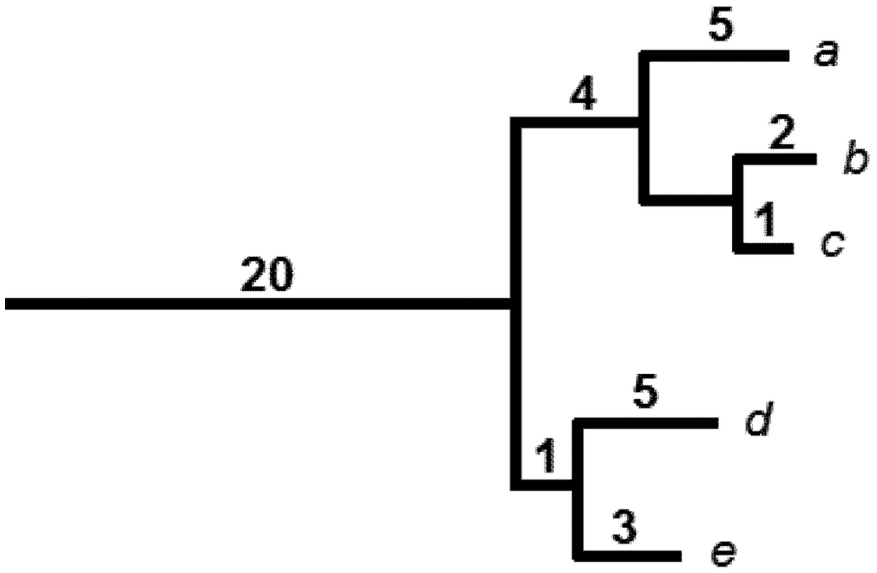


Figura 4 – Árvore filogenética hipotética para ilustrar espécies. Nos galhos estão os tamanhos (numéricos). O valor de diversidade é 41 para esse conjunto de espécies (20 + 5 + 4 + 2 + 1 + 5 + 1 + 3) (Faith e Richards 2012).

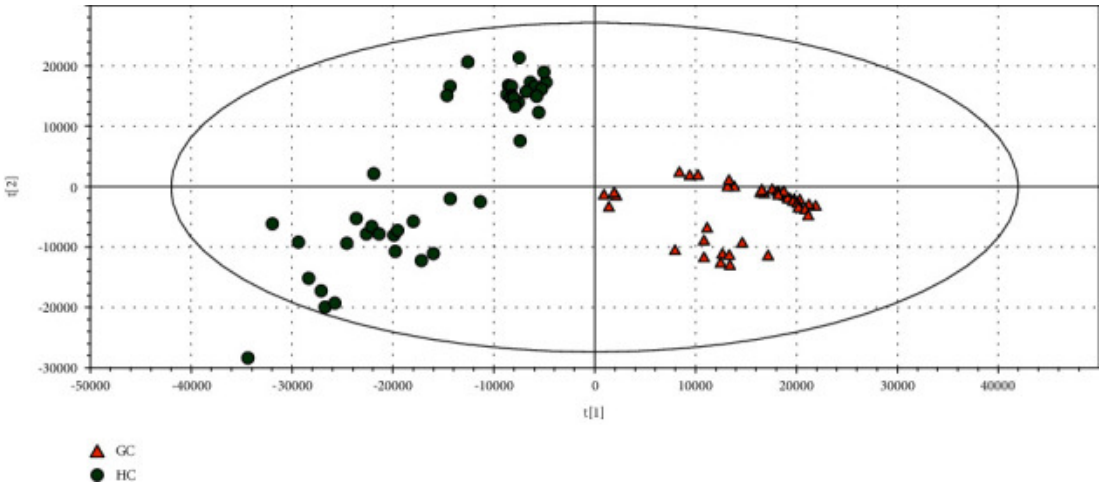


Figura 5 – *Principal component analysis* (PCA) de diferentes metabólitos séricos entre pacientes com câncer gástrico (em laranja, formato triangular) e controles saudáveis (em verde, formato circular) (Wu et al. 2020).

2 JUSTIFICATIVA

A microbiota exerce um papel crucial na doença e na saúde humana. Na última década, vários estudos visaram compreender sua interação com os aspectos moleculares e fisiológicos do hospedeiro. Em doenças como o câncer, por outro lado, esforços são ainda necessários para melhor avaliar o papel desses micro-organismos durante a carcinogênese, resposta imunológica e na terapia anti-tumoral (Meng et al. 2018). Nesse cenário, a caracterização da microbiota, em conjunto com as informações clínicas e patológicas, representa uma oportunidade para melhor compreender os aspectos moleculares envolvidos na doença e, assim, permitir um tratamento mais especializado ao paciente. Atualmente, não há uma abordagem definitiva para a caracterização da diversidade do microbioma (Tremblay et al. 2015), tornando-se necessário customizações de acordo com os protocolos utilizados de sequenciamento e dos métodos de análise (Bokulich et al. 2018).

Em um ambiente de pesquisa e/ou de rotina clínica como o AC Camargo Cancer Center, a acurácia e a capacidade de reprodutibilidade dos métodos são cruciais para prover resultados confiáveis, bem como permitir a validação de possíveis mudanças nos protocolos experimentais, tecnologias escolhidas, atualizações de softwares ou banco de dados (BioTechniques 2020; Marizzoni et al. 2020).

Coletivamente, avaliar os aspectos analíticos em conjunto com a intrínseca variabilidade de preparação dos protocolos que afetam os resultados permanece ser ainda melhor explorado (Bergsten et al. 2020; Bardenhorst et al. 2021).

De suma importância, o código aberto de uma ferramenta que leva esses pontos em consideração pode ser utilizada, por exemplo, em sistemas de saúde gratuitos como o SUS.

3 OBJETIVOS

3.1 OBJETIVO GERAL

Investigar os diversos aspectos que impactam a estimação da diversidade microbiana a partir da análise das regiões 16S. Como prova de conceito e guia para parametrização, utilizar uma amostra controle com resultados esperados. Avaliar um conjunto amplo de amostras geradas no projeto temático FAPESP (14/26897-0) obtidas a partir da saliva, suco gástrico e biópsia gástrica a fim de demonstrar os principais aspectos do pipeline.

3.2 OBJETIVOS ESPECÍFICOS

1. Desenvolvimento de um *pipeline* e sua integração em um ambiente computacional gerenciável;
2. Verificação dos aspectos analíticos que influenciam a acurácia considerando, para tanto, uma amostra controle;
3. Representar as diferenças significativas em relação às métricas de diversidade utilizando dados clínicos associados às amostras.

4 MATERIAIS E MÉTODOS

4.1 CASUÍSTICA

4.1.1 Aspectos Clínicos e epidemiológicos

A fim de avaliar vários aspectos genéticos e epidemiológicos de uma abrangente casuística no Brasil e tratados no ACCCC, pacientes de três capitais (São Paulo, Belém e Fortaleza) do Brasil foram recrutados para participar do projeto *Genomics and epidemiology for gastric adenocarcinomas* (GE4GAC) (Bartelli et al. 2019). No período de três anos (Fevereiro de 2016 a Maio de 2019) 1121 pacientes, com idade entre 35 e 74 anos foram entrevistados. Destes, 412 pacientes foram diagnosticados com câncer gástrico e 153 pacientes entraram para o grupo-controle recrutados durante as campanhas de prevenção ao câncer do ACCCC e 556 do Departamento de Endocopia do Hospital ACCCC. Desses pacientes, entre caso e controle, foram selecionadas 439 amostras para a detecção e análise da microbiota do suco gástrico, saliva e diferentes localizações e subtipos do câncer gástrico (C16/ICD-O3(Fritz et al. 2000)).

Adicionalmente, o banco de dados do Projeto Temático de Câncer Gástrico contém informações demográficas, hábitos diários e estilo de vida além dos dados clínicos relacionados ao câncer e ao seu tratamento (estadiamento, local da lesão, quimioterapia, resultado de exames, entre outros), tal conjunto de dados serão referenciados como *metadados* nos próximos capítulos.

4.1.2 Sequenciamento das amostras controle, saliva, tumor gástrico e suco gástrico

Para o estudo de 16S, 439 amostras foram coletadas e sequenciadas retrospectivamente no ambulatório de tumores gástricos do A.C. Camargo Cancer Center durante a execução do projeto temático, aprovado pela FAPESP (14/26897-0) e sob coordenação do Dr. Emmanuel Dias-Neto. Além disto, uma amostra controle comercial foi sequenciada seguindo os mesmos protocolos das amostras do projeto temático e servirá

de referência para a parametrização do *pipeline*. As amostras foram sequenciadas na plataforma MiSeq (Illumina) com o protocolo *paired-end*, ou seja, um fragmento de DNA é sequenciado iniciando-se em direção ao 3' até certo ponto e o final é sequenciado do final até o 5' em direção contrária.

4.2 CONTROLE DE QUALIDADE PÓS SEQUENCIAMENTO

O controle de qualidade é uma etapa crítica onde é possível observar problemas durante a execução do experimento, seja de bancada, amostra ou sequenciamento. Neste momento é crucial ter sequências de alta qualidade e confiáveis para trabalhar nos passos subsequentes do pipeline.

A avaliação e remoção de adaptadores é realizada pelo programa cutadapt (Martin 2011) bem como a identificação dos *primers* utilizados (região V escolhida) nas duas extremidades das leituras. Outro aspecto importante é certificar-se que, mesmo com os primers esperados identificados, pode haver semelhança ou amplificação inespecífica com o hospedeiro, sendo assim, o alinhamento com o genoma de referência do hospedeiro para eliminar sequências de alta semelhança é realizado.

Para sequenciamentos *paired-end* faremos a junção dos pares utilizando o algoritmo *VSEARCH* ((Rognes et al. 2016) e a remoção de bases de baixa qualidade utilizando o *plugin* do *qiime2 quality-filter* (Bokulich et al. 2018).

Resumidamente, o *pipeline* proposto irá envolver as seguintes etapas e o fluxo de dados é ilustrado na Figura (Figura 6):

- Remoção de adaptadores de sequenciamento;
- Identificação de primers utilizados nas pontas 3' e 5';
- Remoção de sequências do hospedeiro;
- Junção de leituras *paired-end*;
- Remoção de bases de baixa qualidade.

Cada um dos passos acima citados pode ser automatizado para seguir a análise ou alertar o usuário de que alguma característica inesperada foi encontrada, desde que os parâmetros de entrada sejam configurados, como por exemplo, as regiões amplificadas, o genoma hospedeiro e tamanho esperado das leituras.

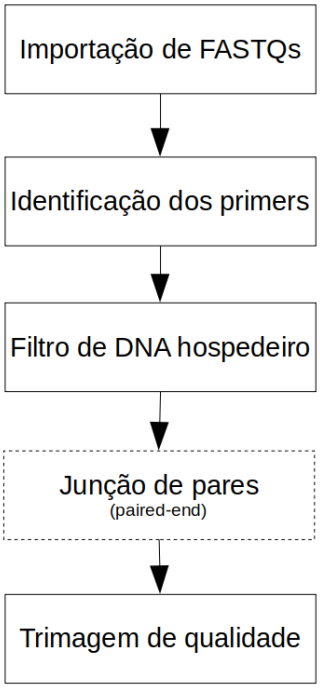


Figura 6 – Fluxograma do pré-processamento: os dados brutos no formato FASTQ serão processados para remoção de sequências do protocolo de sequenciamento, remoção do DNA humano, junção das reads e, ao final, a qualidade final será avaliada. O Passo de junção de leituras *paired-end* será realizado neste momento quando o algoritmo de *denoise* dada2 não for utilizado, uma vez que o dada2 aplica a junção de leituras.

4.2.1 Indentificação dos *primers*

O desenho experimental deste projeto tem um par de *primers* para a região V3 e V4, sendo assim, o pipeline exige que sejam encontrados os *primers* senso (*forward* - F(V3)) e reverso (*reverse* - R(V4)) nas extremidades das leituras sequenciadas, caso ambos os *primers* não sejam detectados, tal leitura será excluída da análise.O desenho esperado dos *primers* em relação ao inserto está ilustrado na Figura 7.

A mesma representação ilustrada na Figura 7 em forma de texto é:

5' F - sequência de interesse - R 3'

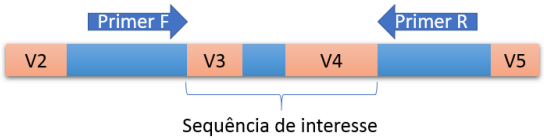


Figura 7 – A região de interesse é o alvo de amplificação interrogada pelos *primers*, esperado e média o tamanho de 550 pares de base.

Levando em consideração que no processo de amplificação (PCR) a fita reverso complementar (RC) também é levada ao sequenciamento, espera-se, também, o seguinte constructo:

5' RC(R) - RC(sequencia de interesse) - RC(F) 3'

Tal configuração se mantém para sequenciamentos não pareados, porém, para sequenciamentos *paired-end*, a configuração esperada é:

5' F – sequencia de interesse ... RC(sequencia de interesse) - RC(reverse primer) 3'

Para a fita anti-senso do produto de PCR:

5' R - RC(sequência de interesse) ... sequencia de interesse - RC(F) 3'

4.3 CARACTERIZAÇÃO DO MICROBIOMA

O sequenciamento do DNA das amostras são analisados através de algoritmos para identificação de micro-organismos implementados pelo pacote *qiime2* (*Quantitative Insights Into Microbial Ecology*) (Bokulich et al. 2018). Os dados de sequenciamento são pré-processados para selecionar apenas as leituras (*reads*) de alta qualidade. Em seguida, as leituras são analisadas por meio de algoritmos que corrigem erros inerentes ao sequenciamento, como por exemplo, regiões de dinucleotídeos repetitivos (Callahan et al. 2016). Quimeras são identificadas e removidas (Amir et al. 2017). Ao final, é realizada a classificação taxonômica utilizando anotações dos bancos de dados Greengenes (DeSantis et al. 2006), SILVA (Quast et al. 2013) e diferentes classificadores foram comparados dinamicamente para decisão do banco e classificador mais reprodutível.

O fluxograma com as etapas propostas de análise é ilustrado na Figura 8.

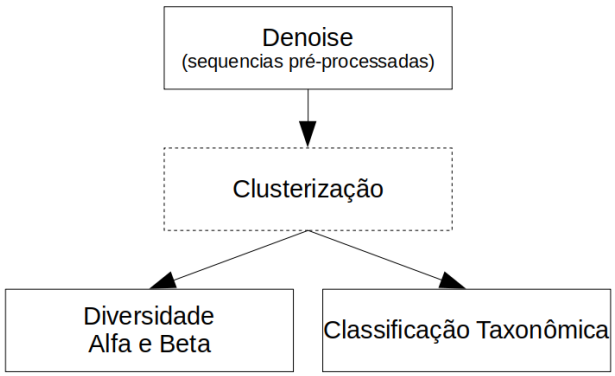


Figura 8 – Fluxograma do pipeline: Após o pré-processamento e eliminação de ruídos resultantes do sequenciamento, as sequências são agrupadas para posterior análise de complexidade entre e dentro dos grupos e, finalmente, a anotação taxonômica. A clusterização em OTUs é um passo opcional cuja a necessidade é avaliada no projeto.

4.3.1 OTUs, ASVs e limpeza das leituras

Algumas espécies são muito semelhantes em algumas regiões variáveis do 16S, e essa característica agregada a erros de sequenciamento mostraram a necessidade em *clusterizar* as sequências em unidades similares chamadas de *Operational Taxonomic Units* (OTUs). As OTUs são úteis para facilitar o processamento nos passos seguintes e mitigam o problema de que erros de sequenciamento que poderiam interferir na diferenciação da espécie, sendo assim, as OTUs são *clusterizadas* (ou agrupadas) com uma identidade (semelhança), por exemplo, de 97%. Porém, houve uma evolução tanto nos equipamentos de sequenciamento de DNA quanto no conhecimento sobre erros inerentes à tecnologia e os algoritmos de *denoising* (eliminação de ruídos), como os já citados *Dada2* e *Deblur*, que são comumente utilizados e referenciados na literaruta, permitem a utilização das sequências sem *clusterização*, tais sequências são chamadas de *Amplicon Sequence Variant* (ASV) e podem trazer uma resolução significante uma vez que uma variante no gene pode ter significado biológico importante (Callahan et al. 2017), porém, apesar de aumentar a capacidade analítica mostrando maior riqueza das amostras, não é o suficiente para diferenciar espécies melhor do que o método de OTUs (Barnes et al. 2020).

Neste projeto também foram comparadas as diferenças dos resultados de classificação taxonômica utilizando OTUs (com uma identidade de 97%) com as ASVs.

É importante citar que ambos os métodos testados (*deblur* e *dada2*) são capazes de excluir leituras quiméricas (artefatos de amplificação da PCR) e tal recurso é utilizado nesse projeto.

4.3.2 Classificação taxonômica

Os métodos de classificação disponíveis no pacote do *qiime2* estão listados abaixo e foram comparados nesse projeto:

- *SKLEARN* (Pedregosa et al. 2011)
- *Naive Bayes* (Webb 2010)
- BLAST+ (Camacho et al. 2009)
- *VSEARCH* (Rognes et al. 2016)
- Híbrido: *VSEARCH-sklearn* (Bokulich et al. 2018)

O algoritmo *VSEARCH* apresentou um melhor resultado na comparação de comunidades *mock* a nível de espécie em relação a outros métodos comumente utilizados para esse fim, no entanto o estudo deixa claro que o ajuste fino de parâmetros é essencial para o sucesso do classificador e foi avaliado neste projeto (Bokulich et al. 2018).

4.3.3 Treinamento do classificador

Os classificadores taxonômicos baseados em aprendizado supervisionado performam melhor quando são treinados levando em consideração aspectos da preparação das amostras e parâmetros de sequenciamento, como por exemplo, tamanho da sequencia, região amplificada, *primers* utilizados e método de sequenciamento (Bokulich et al. 2018).

4.3.4 Análise da diversidade da microbiota

As comparações intra (alfa) e inter (beta) diversidade para as amostras clínicas do projeto foram realizadas utilizando algoritmos de agrupamento não supervisionado, análise de componentes principais (PCA) e medidas de desordem (entropia).

4.4 AMOSTRA CONTROLE

Utilizar um controle nos experimentos para ajudar a detectar problemas que podem passar desapercebidos no preparo do experimento é essencial em caso de testes e melhoria nos protocolos desde a extração da amostra, a amplificação das regiões de interesse, mudança de *primers* e também durante a análise em relação a atualização dos pacotes e softwares ou mudança de parâmetros. Um controle pode ser comercial ou uma amostra caracterizada, ou seja, com um perfil conhecido e comprovado.

Neste projeto, o controle guiou os ajustes iniciais dos parâmetros de modo que o maior número de taxonomias concordantes em nível de gênero foi alcançado, e em segundo plano o nível esperado de abundância em níveis superiores.

O controle *ZymoBIOMICS® Microbial Community DNA Standard D6300* contém 8 espécies de bactérias para detecção por 16S em mesma equivalência (abundância) de 12,5, como ilustrado na Figura 9.

4.5 MEDINDO A CLASSIFICAÇÃO TAXONÔMICA

Para a comparação dos resultados com a amostra controle foram utilizadas as seguintes métricas no nível 6 (Gênero) para o controle Zymo. A razão de acurácia de Taxonomia (TAR - de *Taxon Accuracy Rate*) é o número de taxonomias verdadeiros positivos divido pelo número de taxonomias encontradas. A razão de detecção de taxonomia (TDR - de *Taxon Detection Rate*) é o número de taxonomias verdadeiros positivos dividido pelo total de taxonomias esperadas. A razão entre a quantidade observada sobre a esperada é ROE. Quanto mais próximo de 1 para as métricas, mais próximo do resultado esperado.

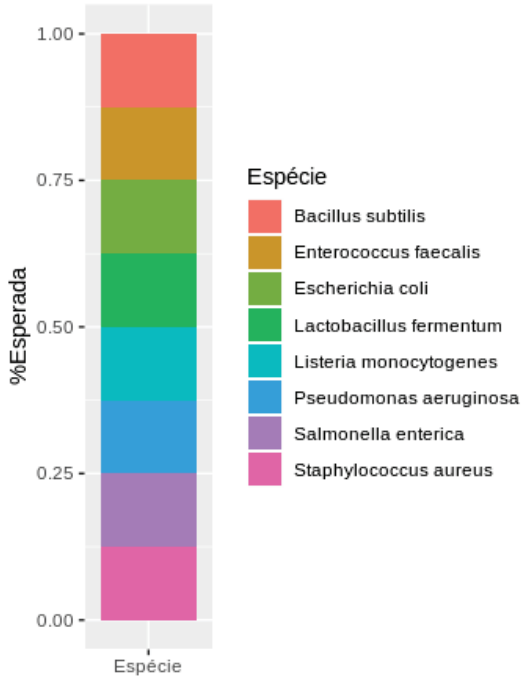


Figura 9 – Controle Zymo - Abundância esperada para cada organismo é de 12,5%.

$$TAR = verdadeiros\ positivos/(verdadeiros\ positivos + falsos\ positivos)$$

$$TDR = verdadeiros\ positivos/(verdadeiros\ positivos + falsos\ negativos)$$

$$ROE = Quantidade\ Observada/Quantidade\ Esperada$$

4.6 ESTRUTURA DE EXECUÇÃO

Para a execução da grande escala de amostras do projeto e aproveitamento da melhor forma dos recursos disponíveis no LBCB , bem como a automação do processo de análise foi utilizada a plataforma de execução Snakemake (Köster e Rahmann 2018) e o código do projeto com controle de versionamento, permitindo assim, que o projeto seja utilizado ou reproduzido em máquinas pessoais ou *clusters* configurando apenas o Snakemake e suas dependências e que os recursos computacionais sejam adequados.

4.6.1 Controle de versão

O *pipeline* foi desenvolvido com controle de versão *git* e o repositório será disponibilizado para a comunidade científica.

5 RESULTADOS

A sessão de resultados está dividida em 4 blocos. Da sessão 5.1 ao 5.5 os resultados de pré-processamento; do 5.6 ao 5.10 os resultados de comparação de algoritmos para *denoizing*, classificadores taxonômicos e as comparações com a amostra controle. Na sessão 5.10 alguns pontos sobre o processamento das amostras clínicas e no 5.12 o *pipeline* desenvolvido.

5.1 REMOÇÃO E IDENTIFICAÇÃO DE PRIMERS

O número de sequencias *paired-end* geradas foi de 68422 e a leitura R1 foi de 305bp, e a leitura R2 foi de 205bp. A detecção de primers com essas exigências resultou em 98,29% (67249) das leituras adequadas, ou seja, com os primers detectados tanto na região 3' quanto na região 5' seguindo as especificações apresentadas no item 4.2.1.

5.2 QUALIDADE DAS LEITURAS APÓS TRIMAGEM

A média de qualidade das bases foi calculada e são apresentadas através de um boxplot por base nas Figuras 10 e 11.

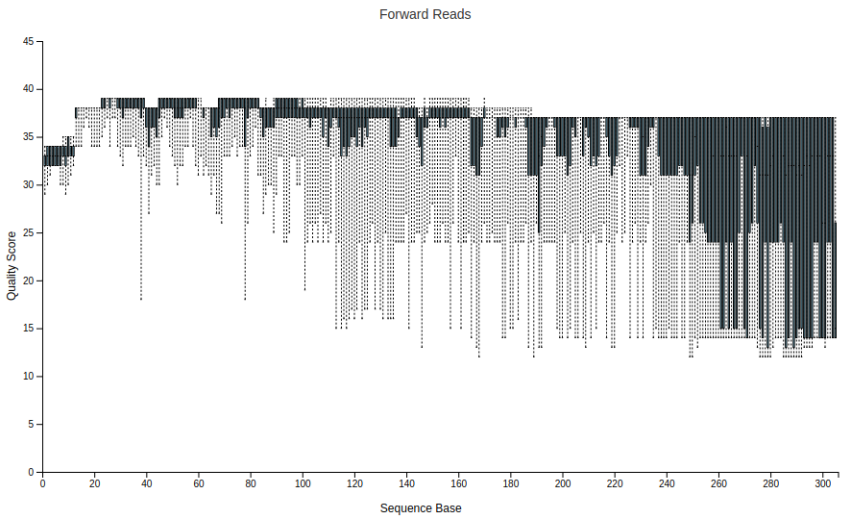


Figura 10 – Boxplot por base das leituras do Controle Zymo - Forward (R1). No eixo Y a qualidade da base na escala Phred e no eixo X as bases em relação ao início da leitura.

Podemos notar a queda da qualidade (eixo Y - em Phred Score) da leitura no final dos reads, tal perfil é típico de leituras da tecnologia de sequenciamento utilizada no projeto.

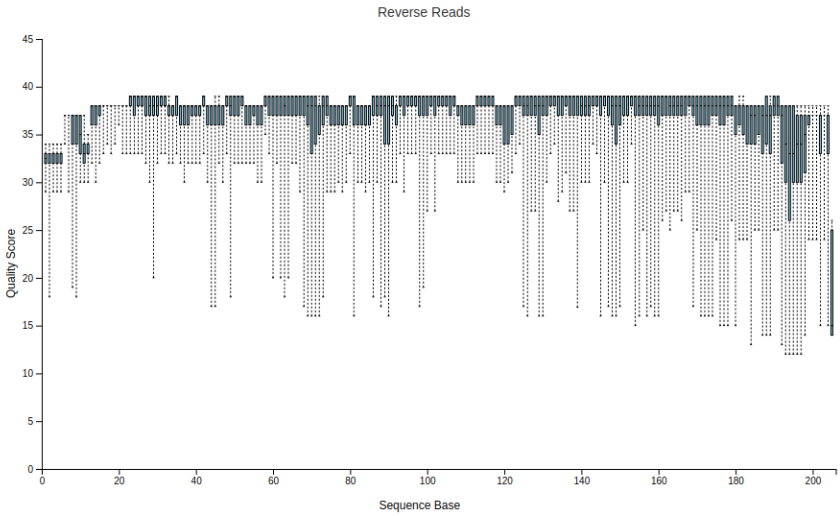


Figura 11 – Boxplot por base das leituras do Controle Zymo - Forward (R2). No eixo Y a qualidade da base na escala Phred e no eixo X as bases em relação ao início da leitura.

A queda de qualidade no final das leituras também pode ser observada na leitura *reverse* (ou R2). Considerando a última base da leitura R2 na Figura 11, a mediana de qualidade é de Q14 (aproximadamente 96% de acurácia), o que é considerado de baixa qualidade, mas vale lembrar que o processo de junção dos pares R1 e R2 irá trazer mais confiança sobre a chamada da base já que haverá sobreposição das bases finais de R1 com as bases finais de R2 e o resultado da junção será discutido no item 5.4.

5.3 REMOÇÃO DE LEITURAS DO HOSPEDEIRO

O programa bwa (Li e Durbin 2009) foi utilizado para o alinhamento das sequências com o genoma do hospedeiro (hg19) para detectar possíveis regiões genômicas sequenciadas e remover possíveis amplificações inespecíficas (ou quiméricas). Um filtro de qualidade de mapeamento de 60 foi aplicado resultando na remoção de 1,36% (916) das leituras consideradas do genoma do hospedeiro, mantendo 66333 sequências para análise e próximas etapas.

5.4 JUNÇÃO DE PARES

Para sequenciamentos *paired-end* a etapa de junção das leituras R1 e R2 é necessária para os próximos passos exceto para o plugin *dada2* que aplica essa etapa automaticamente.

O plugin *vsearch* (Rognes et al. 2016) foi utilizado para os testes com exceção do método de denoise dada2. Quando o *vsearch* foi utilizado, 91,35% (60596) das leituras foram unificadas, Já, utilizando o plugin *dada2*, 70,91% (47043) das leituras foram unificadas.

Um mínimo de 10bp de sobreposição entre R1 e R2 foi exigido e a mediana do resultado final foi de 430bp.O perfil da qualidade das bases é ilustrado na Figura 12 e podemos observar que a perda de qualidade no final do R2 na Figura 11 é mitigada pela sobreposição com a região final do R1.

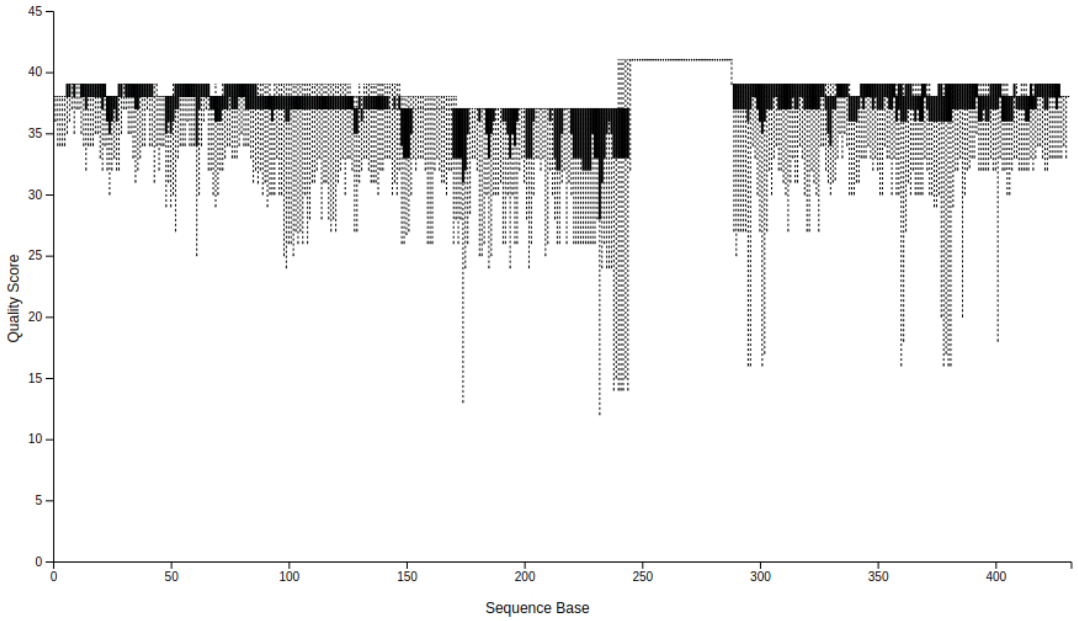


Figura 12 – Boxplot de qualidade da base entre todas as leituras unificadas (método *vsearch join-pair*). No eixo Y o boxplot da qualidade em escala Phred e no eixo X a posição da base.

5.5 TRIMAGEM DE QUALIDADE DAS BASES

A alta qualidade das bases das reads unificadas resultou em 100% das sequências inalteradas quando um filtro de Q10 foi aplicado.

5.6 LIMPEZA DE RUÍDOS DAS LEITURAS - ASVS

No intuito de avaliar o impacto da etapa de *denoising* foram comparados os métodos *dada2* e *deblur*.

Tabela 1 – Resultados das principais métricas de *denoising* para a amostra controle. O número de leituras finais para o *deblur* foi de 20271, para o *dada2* foi de 29461.

Métrica	deblur	dada2
Quimeras	15,82%	26,51%
Artefatos e filtros	50,85%	28,98%
Leituras finais(denoised)	33,33%	44,51%

5.7 CLUSTERIZAÇÃO - OTUS

Para a clusterização das leituras foi utilizado o plugin *vsearch-cluster* com a opção de clusterização com o método de-novo (Rognes et al. 2016) e uma similaridade de 97% entre as leituras para formar uma OTU.

5.8 TREINAMENTO DO CLASSIFICADOR TAXONÔMICO

Para o treinamento do método de classificação *naive-bayes* é necessário realizar o treinamento do algoritmo e uma etapa crucial é a de extração das sequências referência do banco de dados respeitando as sequências dos *primers* referentes às regiões 16S utilizadas no experimento e limitando, assim, o tamanho da sequencia de referências com a região de interesse, neste projeto as regiões V3 e V4. É possível fixar o tamanho das sequências de referência, porém, a junção de pares e a trimagem de qualidade pode resultar em sequências de tamanhos variados. Sendo assim, foram testados e comparados o treinamento do classificador utilizando-se o método naive-bayes com duas abordagens:

1. Limitando o tamanho com a média das leituras unificadas;
2. Sem limite de tamanho permitindo a variabilidade de tamanho das leituras unificadas.

O tamanho de 430bp foi escolhido para o item (1) pela mediana de tamanho das leituras encontradas no processo de junção de pares.

5.9 COMPARAÇÕES DE TAXONOMIAS

Os algoritmos e parâmetros testados mostraram diferenças importantes ao analisar o controle.

Para melhor visualização das comparações as combinações seguem as seguintes siglas:

- *Denoise*
    - Deblur : *deblur*
    - Dada2 : *dada2*
- Clusterização por OTU : cl
- Classificação Taxonômica (SILVA)
    - Naive-Bayes : *nvb*
    - BLAST: *blast*
    - VSEARCH: *vsrch*

O identificador se dará pela combinação das siglas separadas por *sobrelinha* ("_") e na ordem de execução do pipeline: <denoise>_<clusterização>_<classificação>.

Como exemplos: uma análise com *dada2*, sem clusterização e que usa o BLAST como classificador taxonômico: **dada2_blast**. Da mesma forma uma análise com *deblur*, clusterização em OTU e o classificador naive-bayes: **deblur_cl_nvb**.

Os valores de TAR para cada combinação nos níveis zero a seis estão ilustrados na Figura 13.

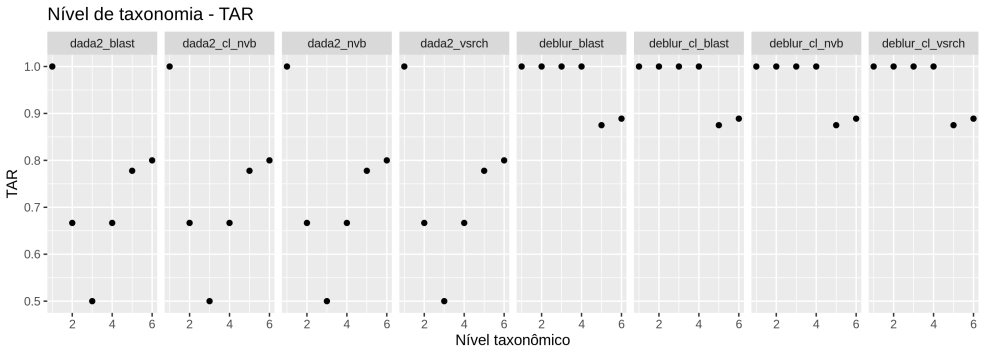


Figura 13 – TAR - Valores para cada combinação dos parâmetros e algoritmos (blocos).No eixo Y o valor de TAR e no eixo X o nível taxônomico avaliado. Quanto mais próximo de 1, melhor o resultado.

Na Figura 13 podemos observar que o *deblur* teve o melhor valor de TAR 0.89, ou seja, com o maior número no nível gênero identificadas corretamente. Como a amostra controle tem espécies definidas foi possível analisar o nível de espécie que demonstrou que o classificador BLAST teve melhor TAR com 0.3 em relação aos outros classificadores naive-bayes e vsearch (0 e 0.1, respectivamente), e não houve diferença na clusterização por OTU.

O algoritmo *dada2* apresentou uma queda importante já nos primeiros níveis taxonômicos independente do classificador, resultado de falsos-positivos encontrados e o melhor TAR de 0.80.

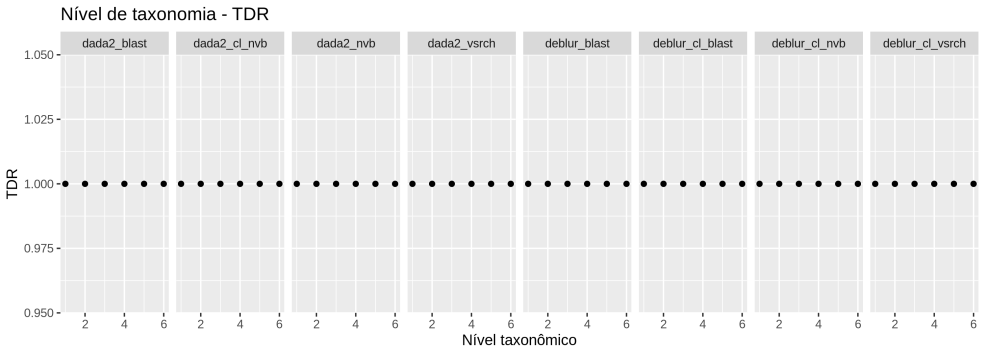


Figura 14 – TDR - Valores para cada combinação dos parâmetros e algoritmos (blocos). No eixo Y o valor de TDR e no eixo X o nível taxônomico avaliado. Quanto mais próximo de 1 o valor de TDR, melhor o resultado.

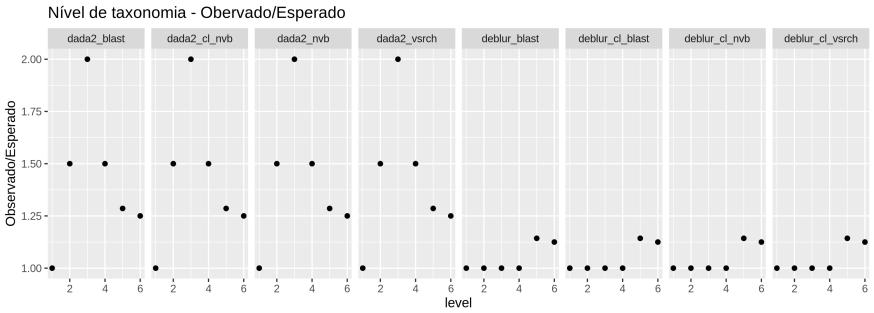


Figura 15 – Quantidade observada/Quantidade esperada para cada combinação de algoritmos por nível taxonômico. Quanto mais próximo de 1, melhor o resultado.

Observa-se na Figura 14 que não há diferenças entre os níveis 1 a 6, isto é, até o nível de gênero não houve diferenças, todas as combinações não tiveram falsos negativos. A nível de espécie o BLAST teve melhor resultado com TDR de 0.375 (independente de denoise ou clusterização).

A combinação que apresentou o melhor resultado na comparação das taxas esperadas e encontradas é observada na Figura. 16

5.10 PARÂMETROS E ALGORITMOS SELECIONADOS

A combinação de parâmetros, métodos e algoritmos que mais se aproximaram ao controle, apresentando o maior número de verdadeiros positivos e apresentou menos falsos positivos foi a configuração utilizando o *denoising deblur*. Diferentes classificadores não mostraram diferença até o nível de gênero, sendo assim, selecionamos como padrão a classificação por BLAST uma vez que seu uso é mais prático para o usuário final (não exige treino).

Todas as comparações foram feitas utilizando-se o banco de dados SILVA versão 132.

A possibilidade de parametrização de todo o *pipeline* tem uma vasta oportunidade de melhoria e a implementação na plataforma *snakemake* a fim de aumentar a escalabilidade, favorecendo a automatização das comparações que será desenvolvida nos próximos passos.

Quando comparamos as taxas de abundância na Figura 16 é possível observar

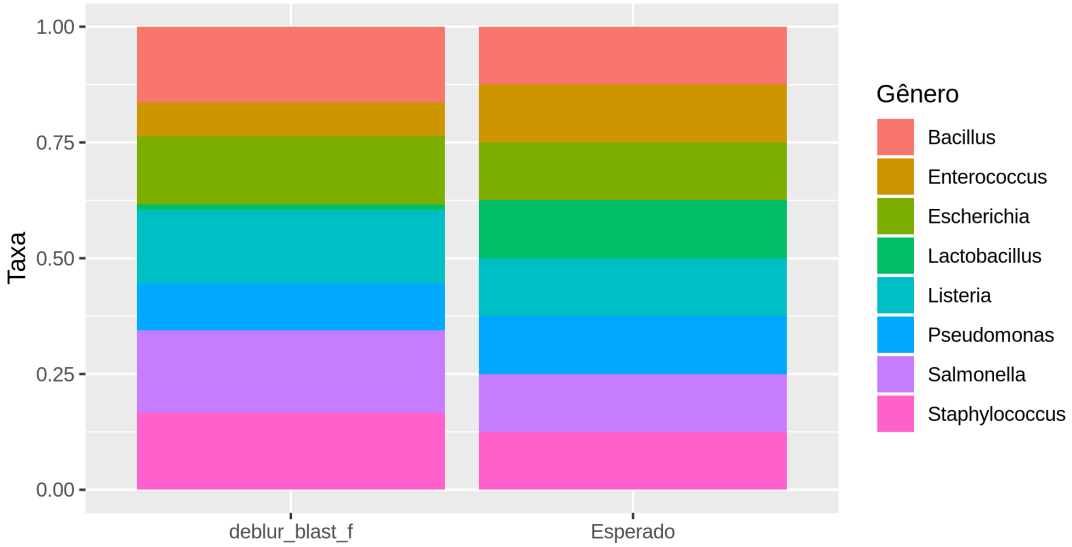


Figura 16 – Abundância de taxonomia esperada e encontrada - Controle Zymo - Nível taxonômico 6 (Gênero).

que o conjunto deblur_blast aproxima-se do esperado demonstrando que apesar de ser possível melhorar os parâmetros, os métodos já se aproximam do resultado esperado com exceção dos Lactobacillos (esperado de 12,5% e encontrado de 1,16%) e os falsos positivos que são muito pouco abundantes e foram filtrados (abundância inferior a 1% [1]. Correlação de Pearson=0.71 e p-value=0.02 com os menos abundantes. Após o fitlro de 1% não foi possível calcular a correlação pois o desvio padrão é igual a zero.

[1] recomendado pelo fabricante do controle

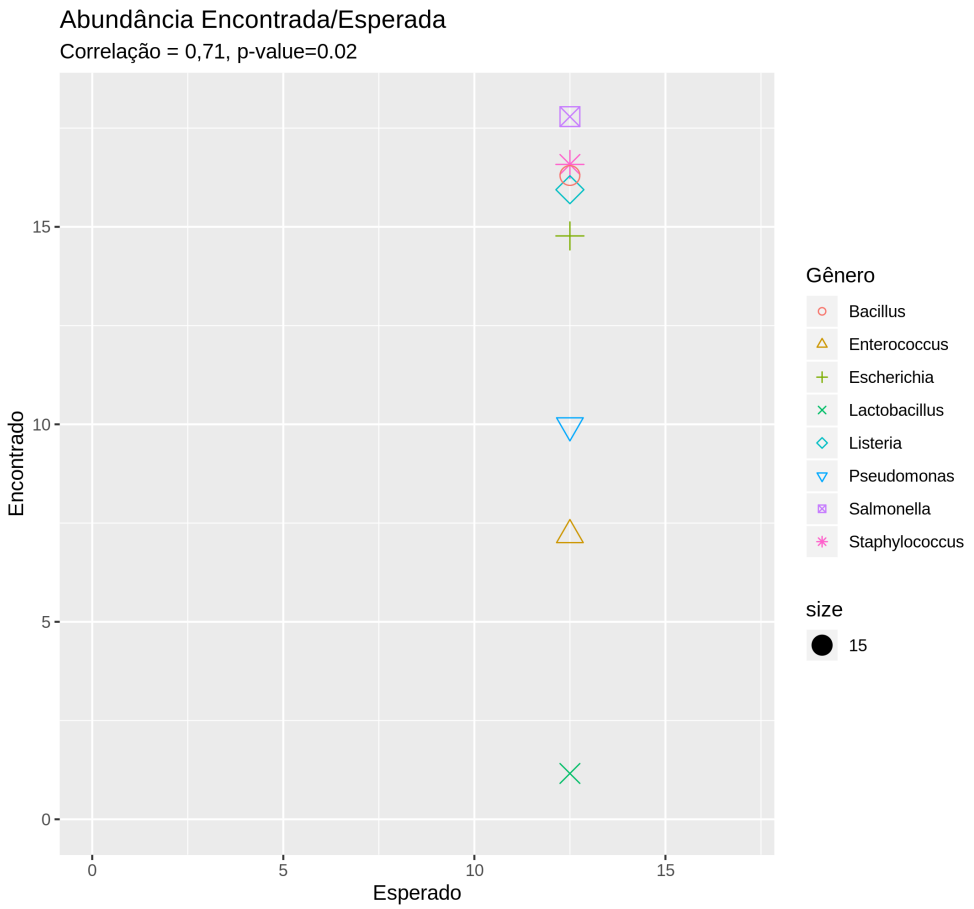


Figura 17 – Correlação da Abundância de taxonômia esperada e encontrada - Controle Zymo - Nível taxonômico 6 (Gênero).

5.11 ANÁLISE DAS AMOSTRAS CLÍNICAS

Foram utlizadas 439 amostras do projeto temático, o pipeline foi executado com os parâmetros que geraram melhor resultado para a amostra controle (seção 5.10). O *pipeline* fornece arquivos no formato qza (qiime2) e também exportações em PNG e html para rápida visualização das métricas de controle de qualidade e testes estatísticos de diversidades alfa e beta. Os resultados das tabelas também são exportados e estão disponíveis para o usuário final. Todos os gráficos e resultados desta sessão são gerados pelo pipeline, com pequenas alterações para melhor formatação a este relatório.

5.11.1 Controle de qualidade dos sequenciamentos

No diretório de resultados de controle de qualidade, é disponibilizado um gráfico com os detalhes das etapas de processamento (Figura 18) além de uma tabela (arquivo texto separado por tabulação) com detalhes .

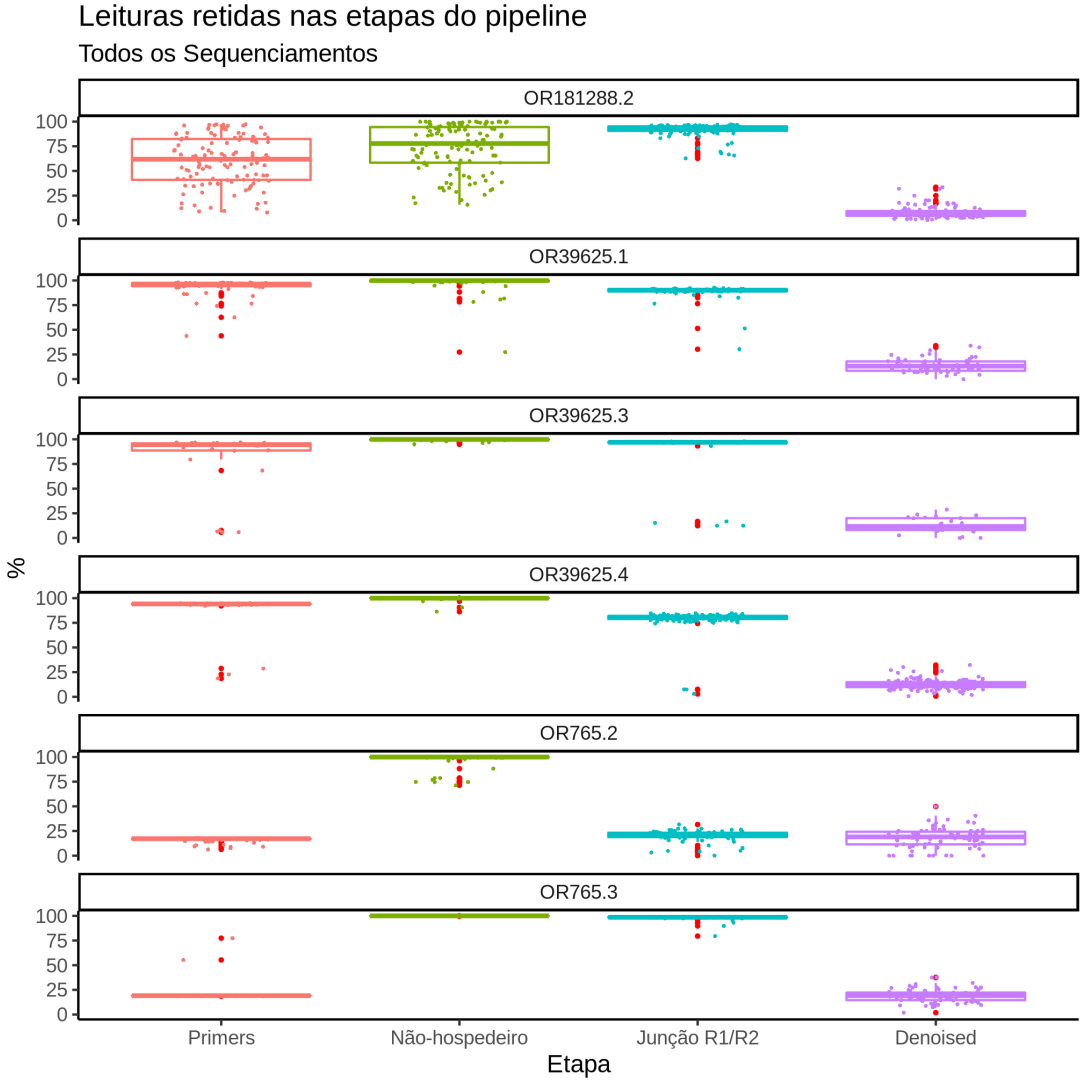


Figura 18 – Porcentagem de leituras que se mantiveram para análise após cada etapa do processamento. (i) "Primers" % de sequências com primers esperados; (ii) "Não hospedeiro" % de leituras não alinhadas com o hospedeiro (humano); (iii) "Junção R1/R2" % de leituras com sucesso na junção paired-end/ (iv) "Denoised" % de leituras com denoise aplicado e que passaram nos filtros do deblur.

Como ilustrado na Figura 18 é possível observar o perfil de cada sequenciamento em cada etapa, auxiliando o usuário a analisar o resultado e confrontar com o que seria esperado de acordo com o experimento laboratorial. É com essas informações

que o pipeline reporta ao usuário possíveis problemas e pontos de ação.

Como um exemplo, o controle de qualidade do sequenciamento OR765.2 na Figura 19 tem uma quantidade baixa de primers esperados, que compromete todas as etapas seguintes (cada etapa refere-se a porcentagem em relação à etapa anterior), ou seja, uma perda de mais de 75% das sequências logo na primeira etapa.

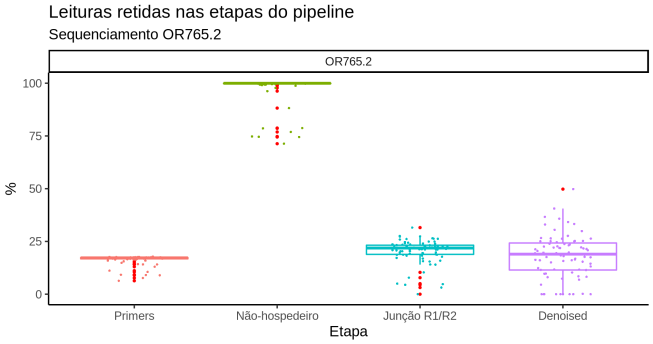


Figura 19 – Controle de qualidade do sequenciamento OR765.2: Poucos primers detectados no experimento, sugestão de falha do protocolo ou do sequenciamento.

Tal observação é importante tanto para que o pesquisador possa avaliar o sequenciamento inicial e tomar alguma ação (como repetição) ou para entender o impacto da qualidade do experimento na análise final ou até mesmo decidir se será menos restrito no controle de qualidade. Por exemplo, poderia configurar a etapa de identificação dos primers para não remover tais sequencias, apenas reportar tal achado. Esse exemplo também mostra a importância de rodar uma amostra controle para tomar a decisão e ter confiança no resultado final, assim, se o usuário escolhe ser menos estrito, tem a possibilidade de confrontar o resultado esperado da amostra controle e seguir com a análise das amostras e suas conclusões.

Ao verificar a qualidade do sequenciamento podemos observar na Figura 20 que a qualidade das leituras R2 ficaram baixas, causando a falha de detecção dos primers. Ainda que o primer seja detectado há também perda considerável na junção R1/R2, provavelmente causada pela baixa qualidade do sequenciamento.

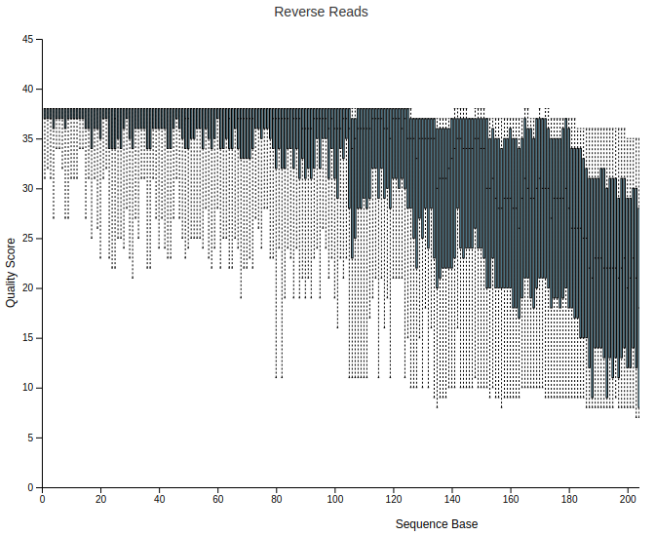


Figura 20 – Valores de qualidade por base do par R2 do Sequenciamento OR765.2.

Com os mesmos parâmetros de análise, temos um exemplo do processamento o sequenciamento OR39625.4 (Figura 21) em que quase a totalidade das leituras contém o primer esperado e mais de 75% das leituras tem a junção dos pares inferindo alta qualidade tanto nas leituras R1 quanto na leituras R2. Apesar da etapa denoise aparentemente ter poucas leituras, tal comportamento é esperado para leituras longas devido à natureza do deblur (quanto maior a leitura, maior a chance de erro de sequenciamento) e os resultados com a amostra-controle visto na Tabela 1.

Tais achados podem ser relacionados à região V3-V4 resultar em leituras quiméricas com grande frequência, tanto pelas suas características biológicas quanto pela natureza das amostras somáticas que têm maior chance de degradação (Walker et al. 2020; Shiu et al. 2018).

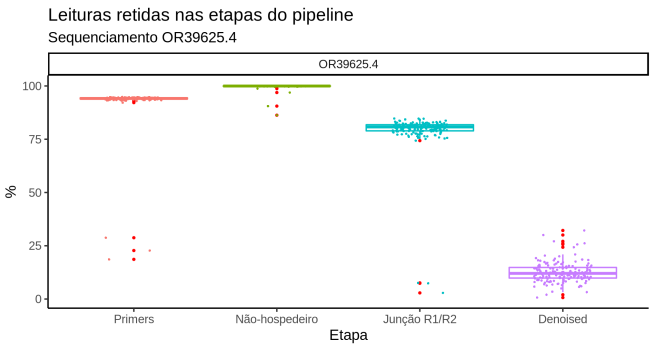


Figura 21 – Controle de qualidade do sequenciamento OR39625.4.

5.11.2 Equalização por amostragem de cobertura

Os testes estatísticos das próximas etapas necessitam de uma equalização no número de ASVs entre as amostras para evitar viés estatístico, por isso, uma das etapas do qiime2 é a amostragem de cobertura (nos tutoriais do qiime2 refere-se ao termo em inglês *Sampling Depth*) em que um número de ASVs é escolhido e aplicado em um sorteio caso a amostra tenha mais do que o mínimo escolhido, por outro lado, se uma amostra não tem o mínimo escolhido, será excluída da análise.

Para determinar o mínimo de ASVs necessário para os testes estatísticos, podemos analisar a curva de rarefação. O algoritmo sorteia ASVs em várias profundidades e aplica a métrica escolhida. Sendo assim, é possível observar se há saturação de ASVs. Em outras palavras, indica que mesmo com mais sequenciamento (ou profundidade/-cobertura) não seriam observados mais ASVs/organismos. (Figura 22)

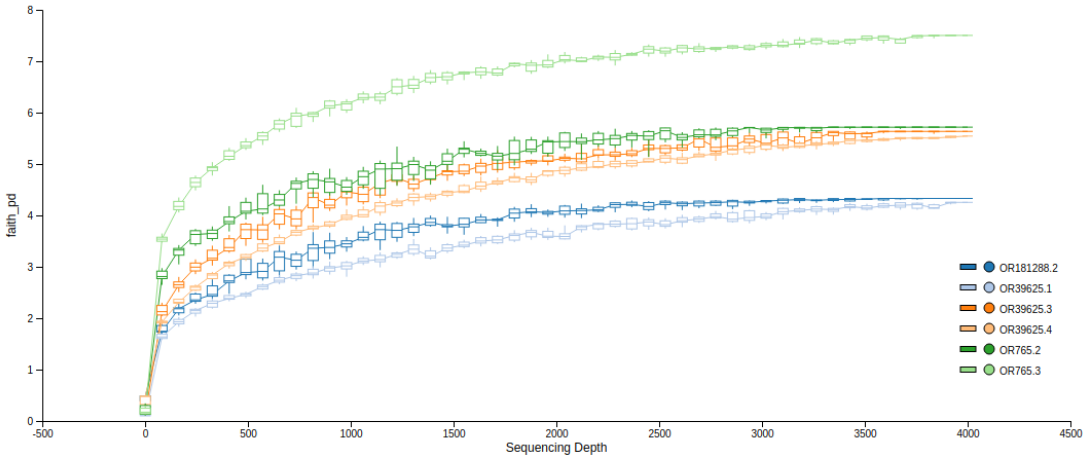


Figura 22 – Curva com métrica de diversidade filogenética de Faith (Chao et al. 2010) no eixo Y - Tal métrica pode ser interpretada como uma generalização filogenética de riqueza de espécies nas amostras. Nota-se no eixo X (profundidade de sequenciamento) que a partir de 3500 é perceptível uma tendência de saturação. Cada ponto da curva é um *boxplot* de todas as amostras do sequenciamento (legenda). Nota: É possível escolher qualquer metadado para desenhar a curva.

Dada a heterogeneidade do número de sequencias utilizáveis em cada amostra, a equalização impacta na exclusão de amostras que não tiveram o mínimo escolhido, resultando em um número menor de amostras utilizáveis para os testes de diversidade. Ao selecionar um mínimo de 4000, algumas amostras são excluídas (resumo na Tabela

Tabela 2 – Contagem de amostras em cada sequenciamento, desde a importação (Total de amostras), passando pelo controle de qualidade (Total importadas) e finalmente após o filtro de mínimo de 4000 (Amostragem de cobertura 4000).

Sequenciamento	Total de amostras	Total Importadas (denoised)	Amostragem de cobertura (4000)	Mantidas para análise
OR181288.2	119	73	20	27.40%
OR39625.1	64	57	50	87.72%
OR39625.3	27	23	21	91.30%
OR39625.4	160	157	147	93.63%
OR765.2	80	52	8	15.38%
OR765.3	81	77	63	81.82%
Total		439	309	70.39%

2), limitando o poder de análise de algumas características clínicas/epidemiológicas, por outro lado, garantindo que as amostras que estão sendo analisadas representam a diversidade do tecido estudado com sequencias de alta qualidade.

5.11.3 Análise de diversidade

Os testes implementados no pipeline estão listados abaixo e são executados automaticamente utilizando-se todos os metadados incluídos na configuração do pipeline.

- Diversidade alfa
    - Significância Shannon
    - Significância Simpson
    - Significância da Uniformidade
    - Diversidade filogenética de Faith
    - Curva de rarefação (Número de ASVs, Pielou, Simpson, Shannon, Faith PD)

Para ilustrar um dos resultados do pipeline, pode-se observar os campos *Metric* (shannon selecionado)e *Sample Metadata Column* (Tipo da Amostra selecionado) na Figura 23, o resultado de curva de rarefação e a opção de download da tabela em formato CSV. Esta é a mesma ferramenta utilizada para gerar a Figura 22 de forma dinâmica, apenas selecionando a métrica e metadado diferentes.

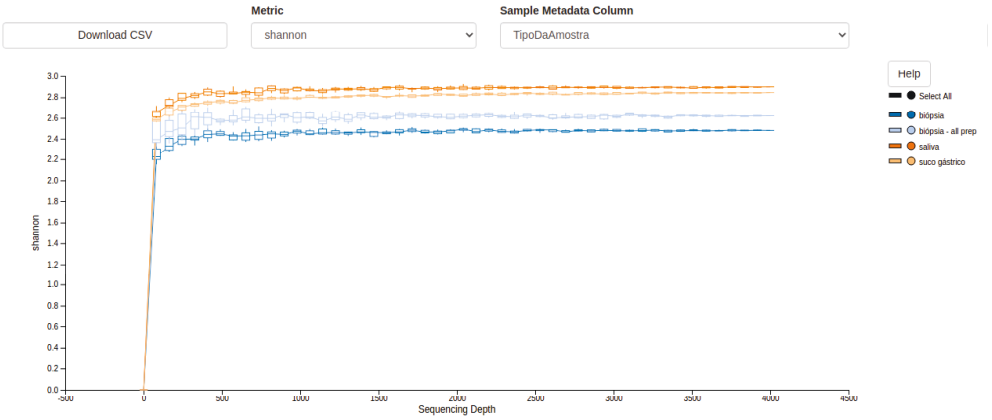


Figura 23 – Curva de rarefação com métrica escolhida “Shannon” para o metadado escolhido “Tipo Da Amostra”. É possível observar que a saliva e fluido gástrico têm um índice de diversidade Shannon maior do que a “biópsia” e “biópsia - all prep” (este último um teste de protocolo, essas amostras serão removidas da análise)

Outro exemplo para ilustração do resultado são os testes estatísticos realizados com os metadados, como por exemplo o teste Kruskal-Wallis entre todos os grupos na Figura 24 e também pareado na Figura 25. Desta forma, o usuário tem uma gama de resultados em que pode-se explorar navegando nas métricas e nos metadados de uma forma dinâmica utilizando as facilidades fornecidas pelo qiime2 com métodos já implementados no pipeline.

Kruskal-Wallis (all groups)

	Result
H	10.271644314031164
p-value	0.01639273866712334

Figura 24 – Teste Kruskal - alfa shannon (todos os grupos)

Kruskal-Wallis (pairwise)

Download CSV

		H	p-value	q-value
Group 1	**Group 2**			
biópsia (n=64)	**biópsia - all prep (n=2)**	2.468284	0.116165	0.185708
	saliva (n=59)	7.752700	0.005363	0.017197
	suco gástrico (n=314)	7.632600	0.005732	0.017197
biópsia - all prep (n=2)	**saliva (n=59)**	2.368507	0.123805	0.185708
	suco gástrico (n=314)	1.302558	0.253747	0.304496
saliva (n=59)	**suco gástrico (n=314)**	0.007317	0.931831	0.931831

Figura 25 – Teste Kruskal - alfa shannon (pareado)

Além da diversidade alfa, também estão implementados testes para análise de diversidade beta.

- Diversidade beta
    - Unweighted-unifrac (permanova pareado)
    - Unweighted-unifrac-emperor (PCA)

O usuário pode configurar o pipeline para realizar o teste pareado Permanova (Anderson 2001) que testa se as distâncias entre as amostras de um grupo são mais similares entre si do que com amostras de outros grupos. Tal teste pode ser aplicado com quantos itens do metadado for necessário (esse processamento é custoso computacionalmente, por isso, não é aplicado automaticamente em todos os metadados como nas análises alfa), mas está implementado no pipeline e é configurável adicionando explicitamente quais metadados serão testados.

Na Figura 26 estão ilustrados os resultados do teste permanova para o metadado ”Tipo de amostra”, que mostrou significância na diferença entre o fluido gástrico, biópsia e saliva com q-value de 0.002.

Pairwise permanova results

Download CSV

		Sample size	Permutations	pseudo-F	p-value	q-value
Group 1	Group 2					
biópsia	biópsia - all prep	12	999	1.266599	0.334	0.334
	saliva	65	999	8.216039	0.001	0.002
	suco gástrico	254	999	5.883369	0.001	0.002
biópsia - all prep	saliva	55	999	2.320478	0.016	0.024
	suco gástrico	244	999	1.455716	0.035	0.042
saliva	suco gástrico	297	999	7.715684	0.001	0.002

Figura 26 – Teste permanova - Diversidade entre beta entre os grupos, cada linha mostras os resultados da comparação entre o Grupo 1 e Grupo 2

Um dos métodos bastante utilizados na análise da composição da microbiota utilizando metadados é a exploração de Coordenadas de Componentes Principais (ou do inglês *principal component analysis* - PCA) e é um dos objetos de resultado do pipeline e um exemplo está ilustrado na Figura 27, também utilizando-se o metadado "Tipo de Amostra".

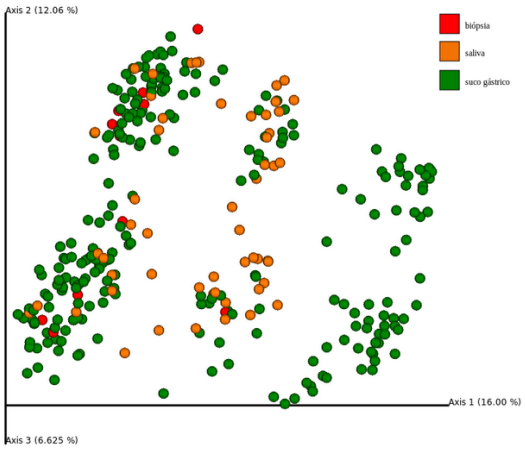


Figura 27 – PCA Tipo de Amostra com métrica *Unweighted Unifrac* para o metadado "Tipo de Amostra". Na exportação html e no qzv é possível rotacionar os eixos observando os pontos em 3 dimensões, bem como escolher qualquer outro metadado para observar a distância das amostras em relação ao metadado selecionado (cor dos pontos).

Por fim, um exemplo da classificação taxonomia reportada pelo qiime2 que também está disponível como resultado do pipeline. Tal ferramenta permite a exploração das taxonomias classificadas e sua proporção. É possível ordenar as amostras pelos metadados e mudar o nível taxonômico (entre zero e sete), além de fazer o download da tabela com contagem por amostra, em CSV.

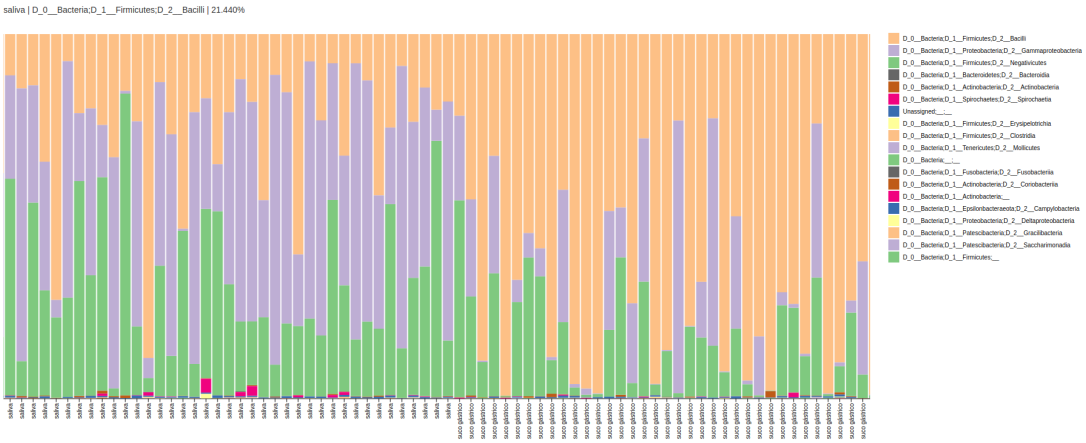


Figura 28 – Taxonomia ordenada pelo "Tipo da Amostra" no nível 3 (classe). Podemos perceber uma diferença na composição entre saliva e fluido gástrico

5.11.3.1 Comparação de diversidade alfa

A fim de avaliar as diferenças significativas ($p<0.05$) entre grupos de interesse de acordo com os metadados, o *pipeline* integra esses resultados em uma mapa de calor (*heatmap*) contendo a significância das comparações utilizam os valores da diversidade alfa.

Na Figura 29 pode-se observar que, independente da métrica, os grupos *ph_gastric_wash* e *ph_range* foram agrupadas. Tais metadados estão altamente relacionados já que um tem a medida de pH e o outro o a faixa de pH medido, demonstrando que a representação proposta pelo mapa de calor destaca a diversidade no contexto dos metadados apontando possíveis associações a serem melhor exploradas.

Concluímos que a representação proposta para avaliar os metadados endossam os resultados esperados com conhecimento prévio.

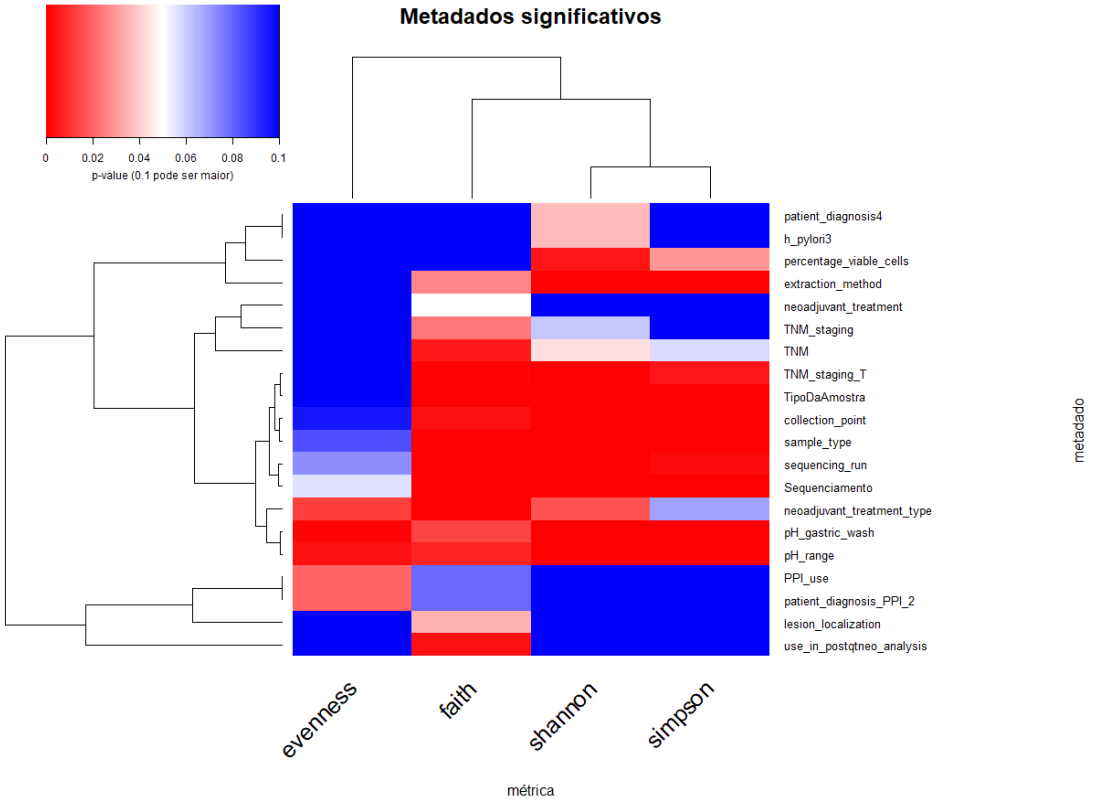


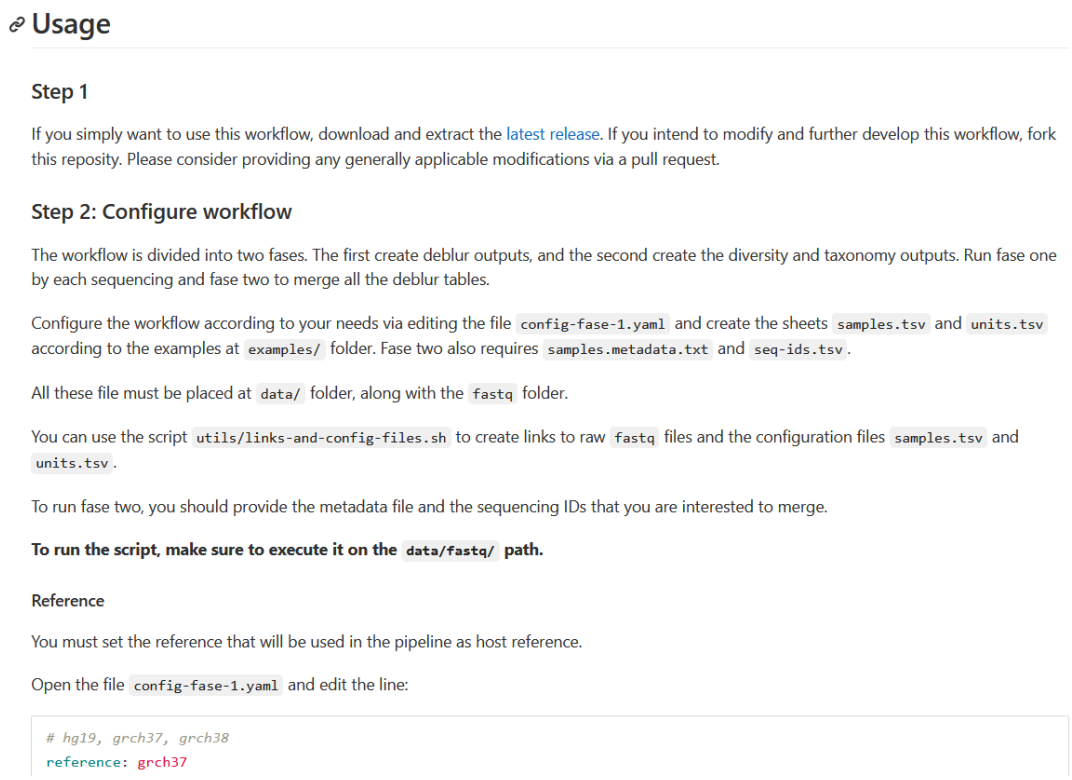
Figura 29 – Heatmap de valor de p, dos metadados nos quais houve diferença estatística entre grupos de p<0.05 em pelo menos uma das métricas alfa (shannon, evenness, simpson ou faith-PD

5.12 PIPELINE E DOCUMENTAÇÃO

O código fonte do pipeline e sua documentação com instruções de uso estão disponíveis no endereço https://gitlab.com/lbcb/metagenomics-analysis, porém, ainda com acesso privado até a finalização deste projeto.

A documentação tem as instruções de uso e configuração do pipeline para correta parametrização e utilização, um excerto é mostrado na Figura 30.

O *pipeline* foi codificado nas linguagens python(63,9%), shell(28,3%), R(4,6%) e PERL(3,2%).



Usage

Step 1

If you simply want to use this workflow, download and extract the latest release. If you intend to modify and further develop this workflow, fork this reposity. Please consider providing any generally applicable modifications via a pull request.

Step 2: Configure workflow

The workflow is divided into two fases. The first create deblur outputs, and the second create the diversity and taxonomy outputs. Run fase one by each sequencing and fase two to merge all the deblur tables.

Configure the workflow according to your needs via editing the file `config-fase-1.yaml` and create the sheets `samples.tsv` and `units.tsv` according to the examples at `examples/` folder. Fase two also requires `samples.metadata.txt` and `seq-ids.tsv`.

All these file must be placed at `data/` folder, along with the `fastq` folder.

You can use the script `utils/links-and-config-files.sh` to create links to raw `fastq` files and the configuration files `samples.tsv` and `units.tsv`.

To run fase two, you should provide the metadata file and the sequencing IDs that you are interested to merge.

To run the script, make sure to execute it on the `data/fastq/` path.

Reference

You must set the reference that will be used in the pipeline as host reference.

Open the file `config-fase-1.yaml` and edit the line:

```
# hg19, grch37, grch38
reference: grch37
```

Figura 30 – Excerto de arquivo README.md com as instruções de configuração do pipeline e de sua utilização.

O Fluxograma do *pipeline* é apresentado nas Figuras 31 e 32 ilustrando o fluxo de dados implementado com seus métodos de entrada e saída.

Figura 31 – A primeira etapa do pipeline é realizada para cada sequenciamento (ou corrida) contendo uma ou mais amostras, onde os adaptadores são removidos, leituras humanas descartadas, junção das leituras *paired-end* e a etapa de *denoising* são aplicados.

Figura 32 – A etapa final compreende na junção de N sequenciamentos e em seguida na reconstrução filogenética, a estimação da diversidade e a classificação do microbioma.

6 DISCUSSÃO

Uma vez que o entendimento da microbiota humana e sua relação com o hospedeiro é cada vez mais evidente em setores da saúde, inclusive na oncologia, a estimação de sua diversidade é fundamental em atividades de pesquisa com grande potencial para auxiliar no suporte à decisão e/ou na compreensão do desfecho clínico. Assim, à medida que os dados e estudos do microbioma se tornam mais comuns, análises confiáveis e extensíveis tornam-se cada vez mais necessárias. Por isso, a implementação de uma abordagem integrada, escalável, reprodutível e segura que suporte os estudos científicos com grande potencial para ser expandindo à rotina clínica, em conjunto com os aspectos que afetam os resultados foi explorada neste projeto.

Desenvolvemos um *pipeline* para estimação do microbioma que utiliza um sistema de gerenciamento de fluxo de trabalho possibilitando, assim, uma análise de dados reprodutível e escalonável em ambientes que suportam processamento em paralelo de alto desempenho.

De importância, esse projeto traz à tona os desafios relacionados ao estudo do microbioma utilizando-se a região 16S e destaca, ainda, aspectos não resolvidos diante da alta heterogeneidade e natureza dos protocolos experimentais e de sequenciamento. Nossos achados se revestem de uma importância maior pelo fato da implementação do *pipeline* ocorrer em um cenário onde há disponibilidade de uma amostra controle com resultados já esperados e confiáveis. Desta forma, foi possível uma avaliação crítica de cada passo do experimento e os respectivos controles de qualidade, escolha de algoritmos e suas parametrizações que são dependentes da escolha dos protocolos (kits de extração, quantidade de ciclos de amplificação, tecnologia de sequenciamento).

A disponibilidade de uma amostra controle por sequenciamento e o controle negativo são fundamentais para que seja possível detectar contaminações e outros problemas no contexto de preparação das amostras que possam influenciar na interpretação final dos resultados. Embora tais práticas são conhecidas em ambientes de rotina clí-

nica, nossas observações endossam que a escolha de kits de sequenciamento, regiões do 16S (primers) influenciam na escolha de parâmetros e algoritmos a serem utilizados.

Ao comparar o desempenho dos resultados obtidos da amostra controle, observamos que não há diferença na classificação taxonômica ao utilizar a forma clássica de clusterização (OTUs) em relacão aos ASVs, embora ASVs sejam mais informativas em análises específicas que necessitam de uma maior resolução.

Os algoritmos de *denoizing* são de suma importância para obtenção de ASVs de alta qualidade, e foram comparados os algoritmos (*dada2* e *deblur*). A maior concordância em relação à amostra controle foram observados nos testes com o *deblur* em conjunto com o alinhador BLAST. É interessante notar que esta escolha é baseada nos sequenciamentos realizados para esse projeto, especificamente, utilizando-se as regiões V3-V4 que, dependendo da tecnologia de sequenciamento e *primers* utilizados, tal escolha deve ser revisitada, reforçando a a relevância da padronização e utilização de amostra controle.

Os dados obtidos demonstram como a parametrização e a avaliação do desenho experimental tem influência direta nos resultados. Apesar de sutil, as diferenças podem resultar em conclusões imprecisas se não forem bem avaliadas e as limitações da análise por 16S não forem criteriosamente discutidas.

Como não houve diferenças nas classificações em OTUs ou ASVs sugere-se que a vantagem de maior resolução dos ASVs deve ser mantida e a clusterização abandonada, uma vez que SNVs podem trazer informações importantes como por exemplo ena correlação com resistência a antibioóticos.

Por fim, buscamos ampliar a interpretação dos principais achados ao incorporar um mapa de calor com as diferenças significativas observadas entre os grupos de acordo com os metadados. Ao compilar esses resultados, destacamos as alterações mais relevantes a serem exploradas que permitem apoiar as decisões e previsões de forma assertiva, bem como gerar hipóteses baseadas na diversidade microbiana.

7 CONCLUSÃO

O desenvolvimento de um *pipeline* e a avaliação dos principais aspectos experimentais e analíticos que podem afetar a estimação do microbioma foram apresentados nesse trabalho. Os resultados obtidos a partir de um conjunto de dados, com amostra controle, possibilita auxiliar na parametrização e avaliação de estudos e aplicações adicionais. Nosso estudo demonstrou que a qualidade do sequenciamento nas amostras estudadas e a região do gene 16S influenciam na quantidade de leituras utilizáveis refletindo na etapa de *denoising* que, por sua vez, terá a capacidade de filtrar quimeras e artefatos restringindo ainda mais as leituras utilizáveis permitindo que, com uma amostra controle, os algoritmos e seus respectivos parâmetros sejam definidos de forma adequada. A associação dos testes estatísticos com os metadados no mapa de calor servem de base para demonstrar que os aspectos analíticos avaliados e selecionados resultaram de forma positiva nas amostras clínicas. A identificação da composição da microbiota em situações clínicas requer um banco de dados extensivo e análise criteriosa para sua correlação com a resposta clínica. Em um ambiente multi-disciplinar e complexo, esperamos que esta ferramenta forneça visibilidade e transparência no fluxo de dados, desde o controle de qualidade até a apresentação dos resultados.

8 REFERÊNCIAS

Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, et al. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. mSystems. 2017; 2:.

Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecology. 2001; 26:32–46.

Kleine Bardenhorst S, Berger T, Klawonn F, Vital M, Karch A, Rübsamen N. Data Analysis Strategies for Microbiome Studies in Human Populations-a Systematic Review of Current Practice. mSystems. 2021 Feb; 6:.

Barko PC, McMichael MA, Swanson KS, Williams DA. The Gastrointestinal Microbiome: A Review. J Vet Intern Med. 2018 Jan; 32:9–25.

Barnes CJ, Rasmussen L, Asplund M, Knudsen SW, Clausen ML, Agner T, et al. Comparing DADA2 and OTU clustering approaches in studying the bacterial communities of atopic dermatitis. J Med Microbiol. 2020 Nov; 69:1293–1302.

Bartelli TF, Senda de Abrantes LL, Freitas HC, Thomas AM, Silva JM, Albuquerque GE, et al. Genomics and epidemiology for gastric adenocarcinomas (ge4gac): a brazilian initiative to study gastric cancer. Applied Cancer Research. 2019 Oct; 39:12.

Benhadou F, Mintoff D, Schnebert B, Thio HB. Psoriasis and Microbiota: A Systematic Review. Diseases. 2018 Jun; 6:.

Bergsten E, Mestivier D, Sobhani I. The Limits and Avoidance of Biases in Metagenomic Analyses of Human Fecal Microbiota. Microorganisms. 2020 Dec; 8:.

Bianconi E, Piovesan A, Facchin F, Beraudi A, Casadei R, Frabetti F, et al. An estimation of the number of cells in the human body. Ann Hum Biol. 2013; 40:463–471.

Bik EM, Eckburg PB, Gill SR, Nelson KE, Purdom EA, Francois F, et al. Molecular analysis of the bacterial microbiota in the human stomach. Proceedings of the National Academy of Sciences. 2006; 103:732–737.

BioTechniques. *Michael Weinstein on the reproducibility of microbiome data*. 2020. ’www.biotechniques.com/interview/microbiome_sptl-michael-weinstein-on-the-reproducibility-of-microbiome-data/’ [Online; acessado em 26/07/2020].

Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2’s q2-feature-classifier plugin. Microbiome. 2018 May; 6:90.

Cain A. *Taxonomy*. 2020. Https://www.britannica.com/science/taxonomy [Online; acessado em 25/07/2020].

Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017 12; 11:2639–2643.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016 07; 13:581–583.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec; 10:421.

Cani PD. Human gut microbiome: hopes, threats and promises. Gut. 2018 09; 67: 1716–1725.

de Carvalho AC, de Mattos Pereira L, Datorre JG, Dos Santos W, Berardinelli GN, Matsushita MM, et al. in Colorectal Cancer Patients of Barretos Cancer Hospital. Front Oncol. 2019; 9:813.

Chakravorty S, Helb D, Burday M, Connell N, Alland D. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J Microbiol Methods. 2007 May; 69:330–339.

Chao A, Chiu CH, Jost L. Phylogenetic diversity measures based on Hill numbers. Philos Trans R Soc Lond B Biol Sci. 2010 Nov; 365:3599–3609.

Instituto Nacional do Câncer I. *Atlas On-line de Mortalidade*. 2018. Https://mortalidade.inca.gov.br/MortalidadeWeb/pages/Modelo01/consultar.xhtml [Online; acessado em 26/10/2018].

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006 Jul; 72:5069–5072.

Duparc T, Plovier H, Marrachelli VG, Van Hul M, Essaghir A, Stahlman M, et al. Hepatocyte MyD88 affects bile acids, gut microbiota and metabolome contributing to regulate glucose and lipid metabolism. Gut. 2017 04; 66:620–632.

Faith DP. Conservation evaluation and phylogenetic diversity. Biological Conservation. 1992; 61:1–10.

Faith DP, Richards ZT. Climate change impacts on the tree of life: Changes in phylogenetic diversity illustrated for acropora corals. Biology. 2012; 1:906–932.

Fritz A, Percy C, Jack A, Shanmugaratnam K, Sobin LH, Parkin DM, et al. *International classification of diseases for oncology*. 2000. Online None.

Garrett WS. Cancer and the microbiota. Science. 2015; 348:80–86.

Goodrich JK, Di Rienzi SC, Poole AC, Koren O, Walters WA, Caporaso JG, et al. Conducting a microbiome study. Cell. 2014 Jul; 158:250–262.

Hill LL, Silvestri LG, Ihm P, Farchi G, Lanciani P. Automatic classification of staphylococci by principal-component analysis and a gradient method. J Bacteriol. 1965 May; 89:1393–1401.

Hunter PR, Gaston MA. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J Clin Microbiol. 1988 Nov; 26: 2465–2466.

Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, et al. Structure, function and diversity of the healthy human microbiome. Nature. 2012 Jun; 486:207–214.

Iida N, Dzutsev A, Stewart CA, Smith L, Bouladoux N, Weingarten RA, et al. Commensal bacteria control cancer response to therapy by modulating the tumor microenvironment. Science. 2013 Nov; 342:967–970.

Jenkins DJ, Wolever TM, Leeds AR, Gassull MA, Haisman P, Dilawari J, et al. Dietary fibres, fibre analogues, and glucose tolerance: importance of viscosity. Br Med J. 1978 May; 1:1392–1394.

Konopiński MK. Shannon diversity index: a call to replace the original Shannon's formula with unbiased estimator in the population genetics studies. PeerJ. 2020; 8: e9391.

Köster J, Rahmann S. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics. 2018 10; 34:3600.

Lairon D. Dietary fibres: effects on lipid metabolism and mechanisms of action. Eur J Clin Nutr. 1996 Mar; 50:125–133.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul; 25:1754–1760.

Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. Nature. 2012 Sep; 489:220–230.

Magalhaes I, Pingris K, Poitou C, Bessoles S, Venteclef N, Kiaf B, et al. Mucosal-associated invariant T cell alterations in obese and type 2 diabetic patients. J Clin Invest. 2015 Apr; 125:1752–1762.

Marizzoni M, Gurry T, Provasi S, Greub G, Lopizzo N, Ribaldi F, et al. Comparison of Bioinformatics Pipelines and Operating Systems for the Analyses of 16S rRNA Gene Amplicon Sequences in Human Fecal Samples. Front Microbiol. 2020; 11:1262.

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011; 17:10–12.

Meng C, Bai C, Brown TD, Hood LE, Tian Q. Human Gut Microbiota and Gastrointestinal Cancer. Genomics Proteomics Bioinformatics. 2018 02; 16:33–49.

Paulos CM, Wrzesinski C, Kaiser A, Hinrichs CS, Chieppa M, Cassard L, et al. Microbial translocation augments the function of adoptively transferred self/tumor-specific CD8+ T cells via TLR4 signaling. J Clin Invest. 2007 Aug; 117:2197–2204.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. Journal of machine learning research. 2011; 12: 2825–2830.

Pielou E. The measurement of diversity in different types of biological collections. Journal of Theoretical Biology. 1966; 13:131–144.

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010 Mar; 464:59–65.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013 Jan; 41:D590–596.

Rawla P, Barsouk A. Epidemiology of gastric cancer: global trends, risk factors and prevention. Przeglad Gastroenterologiczny. 2019; 14:26–38.

Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016; 4:e2584.

Schwabe RF, Jobin C. The microbiome and cancer. Nat Rev Cancer. 2013 Nov; 13: 800–812.

Sender R, Fuchs S, Milo R. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. Cell. 2016 Jan; 164:337–340.

Shiu JH, Ding JY, Tseng CH, Lou SP, Mezaki T, Wu YT, et al. A Newly Designed Primer Revealed High Phylogenetic Diversity of Endozoicomonas in Coral Reefs. Microbes Environ. 2018 Jul; 33:172–185.

Shreiner AB, Kao JY, Young VB. The gut microbiome in health and in disease. Curr Opin Gastroenterol. 2015 Jan; 31:69–75.

Smith PM, Howitt MR, Panikov N, Michaud M, Gallini CA, Bohlooly-Y M, et al. The microbial metabolites, short-chain fatty acids, regulate colonic Treg cell homeostasis. Science. 2013 Aug; 341:569–573.

Srivastava V. *Which biodiversity index is more reliable orbetter and how can it be applied to a mixed vegetation How might I differentiate diversity indices and IVI*. 2015. ”testetest2”.

Stiegler P, Carbon P, Zuker M, Ebel JP, Ehresmann C. Structural organization of the 16S ribosomal RNA from E. coli. Topography and secondary structure. Nucleic Acids Res. 1981 May; 9:2153–2172.

Sung J, Kim N, Kim J, Jo H, Park JH, Nam R, et al. Comparison of gastric microbiota between gastric juice and mucosa by next generation sequencing method. Journal of Cancer Prevention. 2016 03; 21:60–65.

Tremblay J, Singh K, Fern A, Kirton E, He S, Woyke T, et al. Primer and platform effects on 16s rrna tag sequencing. Frontiers in Microbiology. 2015; 6:771.

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature. 2007 Oct; 449:804–810.

Větrovsky T, Baldrian P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PLoS ONE. 2013; 8:e57923.

Walker SP, Barrett M, Hogan G, Flores Bueso Y, Claesson MJ, Tangney M. Non-specific amplification of human DNA is a major challenge for 16S rRNA gene sequence analysis. Sci Rep. 2020 10; 10:16356.

Webb GI. Naïve bayes. In: ____. *Encyclopedia of Machine Learning*. Boston, MA: Springer US, 2010. 713–714. ISBN 978-0-387-30164-8. Disponível em: https://doi.org/10.1007/978-0-387-30164-8_576.

Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA. 1977 Nov; 74:5088–5090.

Wu J, Zhang C, Xu S, Xiang C, Wang R, Yang D, et al. Fecal Microbiome Alteration May Be a Potential Marker for Gastric Cancer. Dis Markers. 2020; 2020:3461315.

Yang B, Wang Y, Qian PY. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC Bioinformatics. 2016 Mar; 17:135.