

UNIVERSIDAD  
**NACIONAL**  
DE COLOMBIA

# **Estrategia bioinformática para el análisis taxonómico y funcional de la microbiota subgingival de pacientes colombianos con periodontitis**

**Yineth Neuta Poveda**

Universidad Nacional de Colombia

Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial

Bogotá D.C, Colombia

2025

# **Estrategia bioinformática para el análisis taxonómico y funcional de la microbiota subgingival de pacientes colombianos con periodontitis**

**Yineth Neuta Poveda**

Tesis o trabajo de investigación presentada(o) como requisito parcial para optar al título de:

**Magister en Bioinformática**

Director:

PhD. Emiliano Barreto Hernández

Universidad Nacional de Colombia

Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial

Bogotá D.C., Colombia

2025

## Resumen

**Título:** Estrategia bioinformática para el análisis taxonómico y funcional de la microbiota subgingival de pacientes colombianos con periodontitis

El análisis de datos a partir de genes marcadores se ha centrado en estrategias bioinformáticas que agrupan diversas herramientas desarrolladas en los últimos años, incluyendo múltiples pasos. Esas estrategias se han utilizado para analizar la microbiota de la cavidad oral a partir de datos genómicos basados en marcadores como el gen 16S rRNA, sin embargo, existen pocos estudios que evalúen la microbiota oral en pacientes colombianos con periodontitis, de acuerdo con la última clasificación de la enfermedad, por lo que el objetivo del presente trabajo fue implementar la estrategia bioinformática más adecuada para el análisis taxonómico y funcional de datos metagenómicos del gen 16s rRNA, obtenidos a partir de muestras de placa subgingival de pacientes colombianos con periodontitis. Se determinó realizar el análisis con QIIME2, que se caracteriza por ser de código abierto y en el cual se pueden realizar análisis a partir de datos de secuencias de amplicones incluyendo en su análisis demultiplexación y filtrado de calidad, asignación taxonómica por ASV, y reconstrucción filogenética, utilizando en complemento DADA 2. Adicionalmente se realizó predicción de la función con PICRUST2, a partir de los ASV identificados. Se concluyó que QIIME 2 es un framework que permite realizar análisis múltiples, lo que confirma su uso para análisis de comunidades microbianas, así mismo en la clasificación, el uso de base de datos especializadas como HOMD para evaluar la microbiota oral, permite focalizar los resultados al ambiente determinado, logrando una clasificación hasta el nivel de especie, evidenciaron microorganismos específicos más frecuentes asociados con periodontitis tales como *Filifactor alocis*, *Fretibacterium*, *Eubacterium nodatum group*, *Eubacterium saphenum*, *Eubacterium brachy*, *Dialister invisus*, *Porphyromonas gingivalis*, *Desulfubulbus*, *Selenomonas*, *Sneathia*, *Treponema*, *Tannerella forythia* y *Parvimonas*. El análisis con PICRUST2 permitió observar vías metabólicas que se asociaron principalmente con los grupos de periodontitis, lo que podría sugerir una alta actividad metabólica de los microorganismos asociados con esta condición clínica.

**Palabras clave:** Bioinformática, secuenciación, análisis del microbioma, periodontitis

## Abstract

**Title:** Bioinformatics strategy for the taxonomic and functional analysis of the subgingival microbiota of Colombian patients with periodontitis

Data analysis from marker genes has focused on bioinformatics strategies that group various tools developed in recent years, including multiple steps. Different strategies have been used to analyze the oral cavity microbiota from genomic data based on markers such as the 16S rRNA gene; however, few studies evaluate the oral microbiota in Colombian patients with periodontitis, according to the latest classification of the disease. Therefore, the objective of the present work was to implement the most appropriate bioinformatics strategy for the taxonomic and functional analysis of metagenomic data of the 16s rRNA gene, obtained from subgingival plaque samples of Colombian patients with periodontitis. It was decided to perform the analysis with QIIME2, which is characterized by being open source and in which analyses can be performed from amplicon sequence data, including demultiplexing and quality filtering, taxonomic assignment by ASV, and phylogenetic reconstruction, using the DADA 2 plugin. Additionally, function prediction was performed with PICRUST2, based on the identified ASVs. It was concluded that QIIME 2 is a framework that allows multiple analyses, which confirms its use for microbial community analysis, also in the classification, the use of specialized databases such as HOMD to evaluate the oral microbiota, allows focusing the results to the specific environment, achieving a classification up to the species level, they showed more frequent specific microorganisms associated with periodontitis such as *Filifactor alocis*, *Fretibacterium*, *Eubacterium nodatum* group, *Eubacterium saphenum*, *Eubacterium brachy*, *Dialister invisus*, *Porphyromonas gingivalis*, *Desulfubulbus*, *Selenomonas*, *Sneathia*, *Treponema*, *Tannerella forythia* and *Parvimonas*. The analysis with PICRUST2 allowed us to observe metabolic pathways that were associated with periodontitis groups, which could suggest a high metabolic activity of the microorganisms associated with this clinical condition.

**Keywords:** Bioinformatics, sequencing, microbiome analysis, periodontitis

Este Trabajo Final de maestría fue calificado en junio de 2025 por el(la) siguiente evaluador(a):

ANDRÉS MAURICIO PINZÓN SOLANO  
Profesor Facultad de Ciencias  
Universidad Nacional de Colombia, Sede Bogotá

# Contenido

<b>Resumen</b> .....	<b>III</b>
<b>Abstract</b> .....	<b>IV</b>
<b>Lista de figuras</b> .....	<b>VIII</b>
<b>Lista de Tablas</b> .....	<b>IX</b>
<b>Introducción</b> .....	<b>10</b>
<b>1. Capítulo 1</b> .....	<b>12</b>
1.1 Marco teórico .....	12
1.1.1 Herramientas bioinformáticas para análisis de datos metagenómicos .....	12
1.2 Análisis de genes marcadores .....	13
1.2.2 Preprocesamiento .....	13
1.2.3 Análisis taxonómico .....	14
1.2.4 Análisis funcional.....	16
<b>2. Capítulo 2</b> .....	<b>19</b>
2.1 Estado del arte .....	19
2.1.1 Enfermedad periodontal .....	19
2.1.2 Estudios de microbioma en cavidad oral.....	20
<b>3. Capítulo 3</b> .....	<b>23</b>
3.1 Objetivos .....	23
3.1.1 Objetivo general .....	23
3.1.2 Objetivos específicos.....	23
<b>4. Capítulo 4</b> .....	<b>24</b>
4.1 Diseño Metodológico .....	24
4.2 Identificación de pasos y herramientas bioinformáticas para el análisis de datos.....	24
4.3 Instalación y configuración de herramientas bioinformáticas.....	24
4.3.1 Preprocesamiento, agrupación y clasificación taxonómica.....	24
4.3.2 Análisis de diversidad .....	25
4.3.3 Análisis de funcionalidad .....	25
4.3.4 Análisis estadísticos .....	25
4.4 Implementación del flujo de trabajo .....	25
<b>5. Capítulo 5</b> .....	<b>28</b>
5.1 Resultados y discusión .....	28
5.1.2 Selección de la estrategia a implementar .....	32

5.1.3. Implementación de la estrategia .....	34
<b>6. Capítulo 6.....</b>	<b>57</b>
6.1 Conclusiones .....	57
6.2 Recomendaciones.....	58
<b>Anexo 1: Script .....</b>	<b>59</b>
<b>Bibliografía .....</b>	<b>62</b>

## Lista de figuras

Figura 1. Análisis de alfa y beta diversidad realizados en Qiime2 .....	31
Figura 2. Flujo de trabajo de PICRUST2 tomado de Douglas et al., 2020 .....	32
Figura 3 Representación de la calidad de las secuencias forward y reverse. ....	36
Figura 4. Diagrama de barras de frecuencia relativa de los fillos identificados utilizando la base de datos SILVA.....	40
Figura 5. Diagrama de barras de frecuencia relativa de los fillos identificados utilizando la base de datos HOMD .....	41
Figura 6. Diagrama de barras de frecuencia relativa de los géneros identificados utilizando la base de datos SILVA.....	43
Figura 7. Diagrama de barras de frecuencia relativa de los géneros y especies identificados utilizando la base de datos HOMD .....	44
Figura 8. Análisis STAMP de las diferencias en la abundancia bacteriana entre grupo salud/gingivitis y los grupos de periodontitis, utilizando la prueba t de Welch. $p < 0.05$ .....	46
Figura 9. Análisis STAMP de las diferencias en la abundancia bacteriana entre grupo salud/gingivitis y los grupos de periodontitis, utilizando la prueba t de Welch. $p < 0.05$ .....	46
Figura 10. Análisis de las diferencias en la abundancia bacteriana entre grupo salud/gingivitis y los grupos de periodontitis, utilizando HOMD (prueba t de Welch. $p < 0.05$ ). ....	48
Figura 11. Análisis de las diferencias en la abundancia bacteriana entre grupo periodontitis I-II y periodontitis III-IV, utilizando HOMD (prueba t de Welch. $p < 0.05$ ). ....	49
Figura 12. Curva de rarefacción.....	50
Figura 13. Análisis de alfa diversidad utilizando los índices de uniformidad, diversidad de Faith e índice de Shannon. Comparaciones entre grupos con Kruskal Wallis $**p < 0.001$ , $*p < 0,05$ .....	51
Figura 14. Análisis de beta diversidad evaluando los índices Bray -Curtis (arriba-izquierda), Jaccard (arriba-derecha), Unifrac no ponderado (abajo-izquierda), Unifrac ponderado (abajo-derecha) .....	53
Figura 15. Comparaciones de distancias entre grupos utilizando PERMANOVA para el índice de Jaccard. $** p < 0.001$ .....	53
Figura 16. Predicción de vías metabólicas que presentaron diferencias significativas entre el grupo salud/gingivitis y los grupos de periodontitis.....	56

## Lista de Tablas

Tabla 1. Características generales asociadas a las secuencias seleccionadas.....	27
Tabla 2. Resumen de pasos y herramientas utilizadas en la estrategia a implementar.....	33
Tabla 3. Resumen de recuentos de secuencias demultiplexadas .....	35
Tabla 4. Resumen de recuentos de secuencias posterior al control de calidad .....	36
Tabla 5. Resumen de rangos de secuencias que se conservaron después de denoising .....	37
Tabla 6. Secuencias representativas .....	38
Tabla 7. Comparación de resultados obtenidos en la clasificación taxonómica de acuerdo con la aproximación utilizada y la base de datos .....	38
Tabla 8. Comparación de características observadas (riqueza) e índice de Shannon entre grupos con Kruskall Wallis ( $p < 0.005$ ).....	51
Tabla 9. Análisis de beta diversidad comparando entre grupos utilizando análisis multivariado (PERMANOVA).....	53
Tabla 10. Principales vías metabólicas basadas en MetaCyc identificadas en la predicción realizada con Picrust2.....	54

## Introducción

Los estudios genómicos realizados a partir de muestras de la cavidad oral, y en general a nivel de microbioma, se han concentrado principalmente en identificar los microorganismos asociados que componen una comunidad, pero para entender mejor el rol que cumplen es necesario realizar estudios más profundos que incluyan un análisis funcional del papel que cumplen en dicho entorno, lo cual es muy útil para entender cómo ese microbioma se asocia con diversas patologías (Niu et al., 2018).

Las tecnologías de secuenciación han permitido dilucidar cada vez más el papel de los microorganismos en relación con diferentes enfermedades, principalmente a nivel gastrointestinal, aunque en la actualidad se evalúa el microbioma de casi cualquier ambiente. A nivel oral, la enfermedad periodontal es una de las enfermedades más comunes que afecta a la población global. En Colombia según el reporte del ENSAB IV (Estudio Nacional de Salud Bucal IV) (Amaya et al., 2014), aproximadamente el 61,8% de la población presenta periodontitis. La etiología de la enfermedad se relaciona con la presencia del biofilm dental, el cual está formado por diversos microorganismos inmersos en una matriz de polímeros que colonizan la superficie de los dientes (Listgarten, 1988). La periodontitis se considera el resultado del desequilibrio entre los microorganismos y los tejidos del hospedador, que con el paso del tiempo afectan los tejidos de soporte de los dientes, lo cual podría llegar a generar pérdida de las piezas dentarias (López, Socransky, Da Silva, Japlit, & Haffajee, 2004).

Entre las metodologías utilizadas para la identificación de microorganismos, la secuenciación de última generación (NGS) se ha convertido en una poderosa metodología que permite identificar la diversidad de la microbiota y su interacción en la cavidad oral y otros ambientes. Se ha observado que el uso de secuenciación de amplicones de las regiones variables del gen 16S rRNA, es un método que permite la identificación de microorganismos cultivables y no cultivables, aunque también se ha encontrado que puede tener varias desventajas, tales como introducir sesgos y omitir organismos y

elementos funcionales del análisis (McIntyre et al., 2017). La secuenciación del genoma completo permite obtener más información y comprensión sobre las comunidades microbianas debido a que se secuencia todo el genoma de los microorganismos que hacen parte de un microbioma particular, permitiendo identificar un gran número características genómicas importantes, sin embargo este tipo de análisis puede considerarse más robusto debido a la gran cantidad de datos que se generan, lo que requiere un gran poder computacional para el análisis, y una mayor profundidad de secuenciación (Mitchell et al., 2018).

El procesamiento de los datos requiere de distintas fases que van desde el pre-procesamiento, en donde se consideran análisis que permiten identificar y filtrar artefactos de secuenciación, evaluando y seleccionando las lecturas obtenidas por la tecnología NGS utilizada de acuerdo con su calidad, hasta análisis para clasificación taxonómica y funcional. En la actualidad se han desarrollado múltiples programas para el análisis de los datos que resultan de los procesos de secuenciación, herramientas que ofrecen diversas ventajas y desventajas, de acuerdo con la finalidad de cada investigador y cada estudio propuesto, por lo que se deben considerar aspectos como: los recursos computacionales disponibles, las bases de datos de referencia, entre otros, para la selección de las herramientas a utilizar, los cuales pueden influir notablemente en los resultados obtenidos (Abellan-Schneyder et al., 2021).

Lo anterior lleva a la necesidad de definir una estrategia bioinformática que permita el análisis confiable de datos a partir de secuenciación genómica para identificar y comprender las diferencias entre el componente microbiano a nivel taxonómico y funcional de muestras de placa subgingival de pacientes colombianos con periodontitis, lo que puede contribuir con información relevante para el diagnóstico y tratamiento, a partir del perfil funcional de los microorganismos que se relacionen con la enfermedad, basada en su actual clasificación.

# 1. Capítulo 1.

## 1.1 Marco teórico

### 1.1.1 Herramientas bioinformáticas para análisis de datos metagenómicos

La genómica incluye la caracterización a nivel del genoma de los miembros de una comunidad (Applications, 2007), este análisis ha permitido ampliar el conocimiento que se había obtenido y que era limitado, debido a las técnicas utilizadas como el cultivo. Los análisis metagenómicos se han enfocado a partir del aislamiento de ADN de organismos que componen una comunidad y que proporcionan información valiosa favoreciendo la comprensión en la interacción de los microorganismos (Handelsman, 2004), a partir de la diversidad microbiana. La utilización de genes marcadores como el 16S rRNA para bacterias, es uno de los enfoques más utilizados debido a características propias del gen, ser un marcador común para las bacterias y arqueas, y ya que contiene información que permite discriminar entre linajes, siendo útil para la clasificación taxonómica, la cual es basada en las regiones conservadas e hipervariables (Niu et al., 2018).

El análisis de datos a partir de genes marcadores se ha centrado en estrategias bioinformáticas que agrupan diversas herramientas que permiten pasar de secuencias de genes sin procesar a perfiles taxonómicos o medidas de diversidad que implican una serie de transformaciones de los datos (Hall & Beiko, 2018). Entre las herramientas más utilizadas se encuentran QIIME, UPARSE, MOTHRU, y DADA (Prodan et al., 2020). QIIME se caracteriza por realizar análisis a partir de datos de secuencia de Sanger, Illumina u otras plataformas, y su análisis incluye demultiplexación y filtrado de calidad, asignación taxonómica por OTU o ASV, y reconstrucción filogenética, siendo de código abierto, el cual puede ser instalado con el paquete ANACONDA (Hall & Beiko, 2018). Dentro de QIIME se puede utilizar DADA2, el cual es un paquete R de código abierto que mejora el algoritmo DADA. El paquete DADA2 implementa el flujo de trabajo completo de amplicón, como el filtrado, la desrepleción, la inferencia de muestras, la identificación de quimeras, la fusión de lecturas de extremos emparejados, etc, esta herramienta incluye análisis de eliminación de ruido por separado en las lecturas directa y reversa, y basa la clasificación taxonómica en variantes de secuencia del amplicón (ASV) que resuelven diferencias en un solo nucleótido (Callahan et al., 2016).

UPARSE es un enfoque desarrollado para producir grupos (OTU) a partir de lecturas de secuenciación de próxima generación de genes marcadores como 16S rRNA, su énfasis está en construir OTU de novo, con la finalidad de lograr alta precisión en la recuperación de secuencias biológicas, lo cual puede mejorar las estimaciones de riqueza (Edgar, 2013). MOTHUR es un paquete de programas que incorporan los algoritmos de herramientas anteriores e integra funciones adicionales, como parámetros ecológicos, visualización y selección de colecciones de secuencias basadas en la calidad (Schloss, 2020). Otras herramientas como MicrobiomeAnalyst, en su versión 2.0, disponible de forma gratuita en [microbiomeanalyst.ca](http://microbiomeanalyst.ca), incluye una amplia gama de métodos de análisis, facilita el análisis estadístico, así como la visualización, analizando secuencias de genes marcadores con una estrategia informática automatizada basada en DADA2, y posterior procesamiento para la anotación taxonómica utilizando bases de datos como SILVA, RDP y Greengenes. La plataforma también explora las relaciones entre los perfiles del microbioma y sus productos metabólicos implementando estrategias asociadas con reducción de la dimensionalidad, análisis de la red metabólica y análisis de correlación, elaborando perfiles funcionales basados en la predicción de Tax4Fun2 (Lu et al., 2023).

La selección de las herramientas a utilizar también depende en gran medida de los recursos computacionales que se dispongan, para lo cual se debe tener en cuenta que sean programas de fácil instalación y/o uso.

## **1.2 Análisis de genes marcadores**

El análisis de los datos de secuenciación incluye diferentes fases en donde se inicia por la evaluación de calidad de las lecturas obtenidas, realizando recortes de las secuencias, unión de las lecturas cuando se realizar secuenciación pareada, agrupación de las secuencias, construcción de tablas de abundancia, anotación taxonómica y análisis del perfil funcional.

### **1.2.2 Preprocesamiento**

El preprocesamiento y el control de calidad de los datos son de los primeros pasos que se realizan al analizar datos metagenómicos, ya que los datos de secuenciación pueden generar ruido de fondo,

debido a la contaminación de los adaptadores de secuenciación, distribución de base desequilibrada, calidad de secuenciación y errores introducidos durante los experimentos. En el preprocesamiento se debe realizar la filtración de lecturas de baja calidad en función del contenido de GC, la aparición de N, la longitud de lectura, las puntuaciones de calidad, y la eliminación de adaptadores de secuenciación que pueden provocar errores. Se han desarrollado varias herramientas con diferentes funciones para manejar archivos FASTQ producidos por la secuenciación en plataformas como Illumina, y se ha observado que algunas herramientas sólo ofrecen funciones para manejar archivos FASTQ o FASTA, entre las que se encuentran Seqtk, FastQ, u otras más generales como PIQA, PRINSEQ, FASTX-Toolkit y NGS QC Toolkit (Abellan-Schneyder et al., 2021; Zhou et al., 2023).

Otras herramientas incluyen Trimmomatic, el cual incluye una variedad de pasos de procesamiento para el filtrado y recorte de lectura, pero las principales innovaciones algorítmicas están relacionadas con la identificación de secuencias adaptadoras y el filtrado de calidad (Bolger et al., 2014).

En QIIME 2 se inicia con la importación de los datos en formato FASTQ, los cuales pueden contener lecturas de extremo único o pareado, por lo que, en el caso de las lecturas pareadas se debe realizar una demultiplexación, lo cual se puede realizar con q2-demux. Posteriormente se puede ejecutar el complemento de DADA 2 para filtrar las lecturas, comprobando quimeras y minimizando las tasas de error en las lecturas directas e inversas, lo cual facilita su fusión, utilizando qiime dada2 denoise-paired. En el proceso de filtrado de calidad se tiene en cuenta las puntuaciones de calidad, en donde se selecciona preferiblemente una puntuación mayor a 30, o de 20 si la calidad de la secuencia es demasiado baja.

### **1.2.3 Análisis taxonómico**

La clasificación taxonómica está basada en la búsqueda de la similitud genómica, la cual inicia con el agrupamiento taxonómico, que permitirá simplificar la cantidad de secuencias, considerando la variabilidad de estas, para posteriormente identificar los organismos que componen las comunidades microbianas. La agrupación de las lecturas puede realizarse en unidades taxonómicas operacionales (OTU), utilizando un umbral de secuencia superior al 97% de similitud. Como OTU se consideran a las lecturas que se derivan de un mismo clado, y esta agrupación va a condicionar directamente a composición microbiana (Gwak & Rho, 2020). Recientemente se han utilizado otras estrategias para

clasificar las lecturas, basadas en variantes de secuencia exacta o variantes de secuencia del amplicón (ASV), en este método las secuencias se diferencian desde un solo nucleótido, y tienen el objetivo de permitir una mayor precisión que las OTU, sin embargo, no son equivalente totalmente a las OTU. Se ha observado que la elección del nivel de agrupamiento que se seleccione también es crucial para los resultados (Chiarello et al., 2022).

Para identificar el perfil taxonómico, lo cual permite evidenciar la composición de cada muestra analizada en función de los taxones presentes y la abundancia relativa de los organismos, se han desarrollado varios programas que han permitido una detección de especies más rápida y el descubrimiento de especies nuevas. Debido a la gran cantidad de datos que se generan se han presentado desafíos computacionales para una adecuada clasificación taxonómica, por lo que es importante comprender cómo funcionan estas diferentes herramientas, denominadas en general clasificadores, y cómo determinar el mejor enfoque para un tipo de muestra, grupo microbiano o aplicación determinados. Esto hace que sea necesaria la evaluación comparativa continua del conjunto de clasificadores para obtener las mejores características de rendimiento en múltiples dimensiones: precisión de clasificación, velocidad y requisitos computacionales (Ye et al., 2019).

Los clasificadores se basan en bases de datos de referencia que están compuestas por secuencias genómicas microbianas previamente secuenciadas. Entre las bases de datos más utilizadas se encuentran RefSeq (RefSeq CG) (Goldfarb et al., 2025), para especies microbianas, SILVA para 16S rRNA (Quast et al., 2013), y asociada a la cavidad oral se encuentra la base HOMD (Human oral microbiome database) (Chen et al., 2010). La selección de la base de datos también puede condicionar los resultados, teniendo en cuenta que la búsqueda en estas bases puede generar falsos positivos debido al gran número de posibles taxones como los que se comparan las secuencias, así mismo la falta de información respecto a especies aún no descubiertas puede dar lugar a clasificaciones falsas negativas. Estas bases de datos pueden usar fuentes completamente diferentes para datos de secuencias o, incluso cuando comparten una fuente común para secuencias (por ejemplo, RefSeq), las actualizaciones continuas y la adición de nuevas secuencias significarán que las bases de datos creadas en diferentes momentos tendrán contenidos diferentes (Qian et al., 2020; Tanca et al., 2016).

Los flujos de trabajo como QIIME (Bolyen et al., 2018; Caporaso et al., 2010) y MOTHUR (Schloss et al., 2009), basados en una serie de scripts escritos en lenguajes de programación como Python, integran análisis para la clasificación taxonómica. Adicionalmente, estos flujos de trabajo generalmente permiten la integración de módulos de R, que facilitan la realización de los análisis estadísticos de los datos para evaluar la composición microbiana de las comunidades, calculando índices de biodiversidad, datos de abundancia, y permiten realizar análisis de inferencia de acuerdo con la composición del microbioma para determinar asociación entre las variables de interés. Entre los análisis realizados se encuentran análisis de componentes principales, análisis multivariados, comparaciones múltiples, entre otros (Calle, 2019).

### **1.2.4 Análisis funcional**

La relación entre taxonomía y función se considera de mayor importancia en los últimos tiempos, por lo que muchos estudios se han basado en la comprensión sobre la composición funcional que permitan inferir la relación entre los perfiles de los microbiomas y sus funciones (Dias et al., 2020). Existen varias herramientas disponibles: PICRUSt, Tax4Fun, Piphillin, Faprotax y paprica, utilizadas para la predicción de perfiles funcionales inferidos a partir de datos de secuencias del gen 16S rRNA. Aunque estas herramientas no pueden reemplazar la evaluación funcional obtenida por secuenciación metagenómica, han brindado conocimientos únicos sobre las capacidades funcionales de las comunidades en diversos hábitats (Wemheuer et al., 2020). Herramientas como PICRUS o Tax4fun utilizan la abundancia relativa de taxones dentro de la comunidad para predecir la funcionalidad, teniendo en cuenta el contenido genético, basados en el genoma de referencia de cada taxón (Galloway-Peña & Hanson, 2020).

PICRUSt es un paquete de software diseñado para predecir el contenido funcional del metagenoma a partir de genes marcadores, es gratuito bajo licencia de GLP, y consta principalmente de dos flujos de trabajo: inferencia de contenido genético e inferencia de metagenoma. Actualmente se encuentra disponible la versión PICRUST2 (Langille et al., 2013), la cual contiene una base de datos actualizada que ha incrementado sustancialmente el número de genomas bacterianos (de 19.493 a 26.868) así como el número de anotaciones funcionales presentes y más grande de familias de genes y genomas de referencia, e integra herramientas de código abierto existentes para predecir genomas de secuencias de genes 16S rRNA muestreadas ambientalmente, la actualización está preinstalada con PICRUST2 a partir de la versión 2.6.0 (Wright & Langille, 2025). Se incluyen más de 20000 genes completos de ARNr 16S, que se utilizan como base de las predicciones funcionales, teniendo en cuenta ASV. En PICRUST2 se incluyen HMMER ([www.hmmmer.org](http://www.hmmmer.org)) para ubicar los ASV, EPA-

ng8 para determinar la posición óptima de estos ASV ubicados en una filogenia de referencia y GAPPA9 para generar un nuevo árbol que incorpore el ASV. A partir del árbol filogenético que contiene genomas de referencia y microorganismos muestreados ambientalmente se predice el número de copias de familias de genes individuales para cada ASV. Los ASV se corrigen por su número de copias del gen 16S rRNA y luego se multiplican por sus predicciones funcionales para producir un metagenoma predicho. PICRUSt2 también proporciona la contribución de ASV de cada función predicha, lo que permite realizar análisis estadísticos basados en taxonomía. Por último, las abundancias de vías se infieren en función de mapeos de vías estructuradas (Douglas et al., 2020).

Tax4Fun2 es una versión nueva y mejorada de Tax4fun, funciona basada en R, se considera fácil de usar y muy eficiente en memoria. Se centra en procariotas, aunque incorpora algunos datos de eucariotas. En análisis comparativos realizados se ha observado que esta herramienta permite calcular la redundancia funcional, lo cual favorece la predicción funcional, siendo más precisa de PICRUSt. Dentro de las fortalezas que tiene la herramienta se considera la incorporación de datos definidos por el usuario y específicos del hábitat, lo que mejora la precisión de las predicciones (Wemheuer et al., 2020).

Existen otros paquetes de software como Piphillin, disponible públicamente, el cual predice el contenido metagenómico funcional basado en la frecuencia de las secuencias del gen 16S rRNA detectadas, correspondientes a genomas en bases de datos de genomas funcionalmente anotadas y actualizadas periódicamente. El algoritmo utiliza la coincidencia directa del vecino más cercano entre los amplicones y los genomas de ARNr 16S para predecir los genomas representados (Iwai et al., 2016). Las secuencias de ácido nucleico representativas de las unidades taxonómicas operativas (OTU) candidatas se comparan directamente con las secuencias del gen 16S rRNA de los genomas en la base de datos para inferir el contenido del genoma y, por lo tanto, el potencial funcional. Piphillin también puede predecir el contenido metagenómico de los ASV corregidos por DADA2, aunque aún no está claro cómo funcionan estas técnicas en procesos metagenómicos de predicción de contenido funcional basadas en ASV (Narayan et al., 2020).

Faprotax es una base de datos asociada con la anotación funcional de taxones procariotas, al convertir los perfiles taxonómicos de comunidades microbianas en posibles perfiles funcionales, basados en los taxones identificados en una muestra a partir de tablas OTU, por medio de scripts de Python. Se centra en funciones que se relacionan con el ciclo del azufre, el nitrógeno, el hidrógeno y el carbono, aunque también se incluyen otras funciones, pero se describe que probablemente falten funciones, o estén parcialmente incluidas en la base de datos (Louca et al., 2016). Otros métodos como Paprica

determinan la estructura de la comunidad microbiana, y predicen las rutas metabólicas con enfoques filogenéticos. Requiere de la instalación de algunos programas, los cuales están disponibles en entornos Linux y para usuarios Mac OSX (Erazo et al., 2021).

## 2. Capítulo 2

### 2.1 Estado del arte

#### 2.1.1 Enfermedad periodontal

La cavidad oral alberga gran cantidad de microorganismos, que cumplen funciones importantes a nivel de salud y enfermedad, lo cual se ha asociado con la diversidad y abundancia de las especies. El desequilibrio entre los microorganismos habitantes normales de la cavidad oral, y los tejidos del hospedador favorecen el desarrollo de enfermedades como la periodontitis. De acuerdo con la última clasificación, la periodontitis se caracteriza por una inflamación mediada por el hospedador y asociada a microorganismos, lo que da como resultado la pérdida de la inserción periodontal, lo que conduce a la pérdida de las piezas dentarias (Tonetti et al., 2018). En la clasificación actual se tienen en cuenta cuatro categorías (etapas 1 a 4) que consideran varias variables incluyendo pérdida de inserción clínica, cantidad y porcentaje de pérdida ósea, profundidad al sondaje, presencia y extensión de defectos óseos angulares y afectación de furca, movilidad y pérdida dental. Adicionalmente se incluyen tres niveles (grado A: bajo riesgo, grado B: riesgo moderado, grado C: alto riesgo de progresión) y abarca, además de los aspectos relacionados con la progresión de la periodontitis, el estado general de salud y otras exposiciones, como fumar o el nivel de control metabólico en la diabetes (Caton et al., 2018). La naturaleza infecciosa de la enfermedad periodontal fue reconocida desde los años 60 y con el desarrollo de las técnicas de cultivo, numerosos estudios alrededor del mundo reconocieron algunos microorganismos anaerobios y microaerofílicos que crecían significativamente en el ambiente subgingival de pacientes con periodontitis comparado con pacientes sin enfermedad periodontal. *P. gingivalis*, *Tannerella forsythia*, *Aggregatibacter actinomycetemcomitans*, *Eikenella corrodens*, *Campylobacter rectus*, *Treponema denticola* y *Prevotella intermedia* fueron reconocidos como aquellos asociados a la periodontitis (Hong et al., 2015). Más tarde, Socransky et al., en 1998, analizaron los distintos microorganismos, para examinar las diferencias entre las especies bacterianas en pacientes diagnosticados con periodontitis en

comparación con pacientes sanos periodontalmente, mediante la tecnología de hibridación con sondas de DNA para el gen 16S rRNA. Se lograron establecer 5 complejos bacterianos por homología genética y de estos uno se asoció con el aumento de la profundidad de la bolsa y la pérdida de inserción clínica periodontal (Haffajee, Socransky, Patel, & Song, 2008).

Muchos de los estudios realizados han relacionado una baja diversidad de algunos microorganismos presentes en el ecosistema oral. Algunos estudios sólo han evaluado el componente microbiano de bacterias cultivables y no cultivables en el biofilm subgingival en pacientes con periodontitis basados en técnicas que no son muy sensibles o específicas, aunque actualmente con el uso de nuevas tecnologías como la Secuenciación de última generación (NSG de sus siglas en inglés Next Generation Sequencing), basadas en la secuenciación simultánea del gen 16S rRNA de las bacterias presentes en el biofilm en la cavidad oral a nivel subgingival, se ha determinado el hallazgo de filos de bacterias periodontales no cultivables, 11 en el dominio de bacterias del microbioma oral y periodontal, y dos actualmente sin nombre, SR1 y TM74, así como la verificación de la presencia de *Filifactor alocis*, *Eubacterium saphenum*, *Porphyromonas endodontalis*, *Prevotella denticola*, *Parvimonas micra*, especies de *Bacteroidetes*, *Peptostreptococcus*, *Desulfobulbus*, *Dialister*, y de *Synergistetes* (Pérez-Chaparro et al., 2014). En los últimos años se han realizado estudios que evalúan la prevalencia de microorganismos identificados en pacientes con periodontitis, y se han logrado identificar otras bacterias que anteriormente no se conocían asociadas a las alteraciones periodontales como algunos representantes de los géneros *Filifactor*, *Megasphaera*, *Desulfobulbus*, *Fretibacterium fastidiosum*, *Anaeroglobus geminatus*, *Eubacterium saphenum*, *Porphyromonas endodontalis* y *Prevotella denticola*, y otros como *Anaeroglobus geminatus*, *Dialister invisus*, *Dialister pneumosintes*, *Eubacterium brachy*, *Mogibacterium timidum*, *Slackia exigua*, las cuales han cobrado importancia y han permitido ampliar la información de los microorganismos que se han relacionado con la enfermedad periodontal, (Antezack et al., 2023).

### **2.1.2 Estudios de microbioma en cavidad oral**

A nivel de la cavidad oral se han realizado diversos estudios evaluando la composición de las comunidades, entendiendo que este hábitat lo componen cientos de especies bacterianas y otros microorganismos, los cuales cumplen un papel importante en la preservación de la salud y/o el desarrollo de la enfermedad. La mayoría de estudios realizados han evaluado principalmente la diversidad microbiana, lo cual es importante y un punto inicial para el entendimiento de las comunidades orales, lo cual estaría más relacionado con estudios de microbioma que se asocian con

responder ¿Quién está ahí?, las técnicas actuales de secuenciación han permitido con el paso del tiempo comprender más características y en mayor profundidad a las comunidades, lo cual depende también de la técnica o plataforma de secuenciación utilizada y la profundidad de los análisis que se realicen. En 2012 Belda et al., realizaron un estudio utilizando piro secuenciación a partir de muestras de placa supragingival de individuos con diferentes diagnósticos orales. Iniesta et al., en 2023, caracterizaron el microbioma subgingival de pacientes con diferentes estados periodontales, incluyendo pacientes españoles con periodontitis, diagnosticados con base en la última clasificación. Entre los resultados observaron mayor riqueza en los grupos de periodontitis y gingivitis comparados con los pacientes sanos, aunque no encontraron diferencias significativas entre los estadios de la periodontitis a nivel de la estructura de la comunidad. Más del 95% de las secuencias representativas se asignaron a siete filos principales, *Firmicutes*, *Bacteroidetes*, *Fusobacteria*, *Proteobacteria*, *Actinobacteria*, *Spirochaetes* y *Synergistetes*, y se encontró que a nivel de género se presentó una mayor abundancia relativa de *Porphyromonas*, *Treponema*, *Tannerella*, *Fretibacterium* y *Filifactor* en el grupo de periodontitis III-IV, en comparación con el grupo de salud. También Lee et al., en 2024 evaluaron pacientes con distintos diagnósticos orales, incluyendo 28 con periodontitis, observando que se encontró mayor abundancia de variantes de secuencias en el grupo con periodontitis, asociadas a bacterias como *Fretibacterium fastidiosum*, *Sinanaerobacter chloroacetimidivorans*, *Filifactor alocis* y *Treponema denticola*

En Colombia existen pocos estudios que evalúen la microbiota oral asociada a la periodontitis, estos estudios han sido principalmente basados en técnicas de cultivo y en técnicas de biología molecular como la reacción en cadena de la polimerasa (PCR). Estudios como el reportado por Mayorga *et al.*, evaluó por cultivo la presencia y concentración subgingival de diversos microorganismos en pacientes con periodontitis crónica y agresiva, en comparación con sujetos sanos (Mayorga-Fayad et al., 2007)), también Lafaurie *et al.*, en 2007, evaluó los aspectos sociodemográficos, clínicos y microbiológicos de pacientes colombianos con periodontitis en diferentes regiones geográficas (Lafaurie *et al.*, 2007), y más recientemente evaluaron la microbiota periodontal comparando entre pacientes colombianos y españoles (Lafaurie et al., 2022).

### **2.1.3 Análisis funcional del microbioma oral asociado con periodontitis**

El potencial funcional de los microorganismos permite identificar características clave que dilucida cómo los microorganismos asociados con la enfermedad contribuyen de manera significativa. Los principales análisis asociados con la función de los microorganismos utilizan la secuencia del genoma completo, o del transcriptoma obtenido por medio de RNA-Seq. Algunas investigaciones

evaluaron un conjunto de datos metatranscriptómicos para identificar especies bacterianas asociadas con periodontitis, indagando las características funcionales de la comunidad asociada a placa subgingival. Se identificaron 52 familias de genes, de las cuales 50 fueron más abundantes en periodontitis, en comparación con el grupo de pacientes sanos. Se observó que el perfil funcional se asociaba principalmente con transporte y secreción transmembrana, el metabolismo de aminoácidos, la síntesis de proteínas de superficie y flagelos, el metabolismo energético y el superenrollamiento del ADN, los cuales se asocian con bacterias periodontopatógenas como *Porphyrromonas*, *Tannerella*, *Treponema* y *Desulfubulbus* (Huang et al., 2021).

Otros estudios incluyen información asociada con la funcionalidad basada en predicciones, utilizando herramientas bioinformáticas que analizan la información obtenida de la secuenciación del gen 16S rRNA, es así como Zhao et al., en 2022 evaluaron la microbiota oral de pacientes femeninas con periodontitis en estadios I y III, determinando las bacterias predominantes en cada grupo, y las vías de señalización que se relacionarán con la composición microbiana, utilizando la base de datos de genes y genomas de Kioto (KEGG) (Kotera et al., 2015). Entre los resultados se observó que el microorganismo más representativo asociado al estadio III fue *Prevotella*, encontrando que el cambio en la composición microbiana de estas pacientes puede estar asociado con el procesamiento de proteínas.

## **3. Capítulo 3**

### **3.1 Objetivos**

#### **3.1.1 Objetivo general**

Implementar la estrategia bioinformática más adecuada para el análisis taxonómico y funcional de datos metagenómicos obtenidos a partir de muestras de placa subgingival de pacientes colombianos con periodontitis

#### **3.1.2 Objetivos específicos**

- Identificar las principales herramientas bioinformáticas disponibles para el análisis taxonómico y funcional de datos metagenómicos
- Seleccionar la herramienta bioinformática más adecuada para el análisis taxonómico y funcional de datos metagenómicos obtenidos a partir de muestras de placa subgingival
- Implementar la estrategia bioinformática seleccionada

## **4. Capítulo 4**

### **4.1 Diseño Metodológico**

Estudio descriptivo en el cual se seleccionan herramientas bioinformáticas y se utilizan para el análisis de datos metagenómicos a partir del gen 16 S rRNA.

### **4.2 Identificación de pasos y herramientas bioinformáticas para el análisis de datos**

Se realizó una búsqueda bibliográfica de los pasos y las herramientas más utilizadas en los 5 últimos años para realizar el análisis taxonómico y funcional, a partir de secuencias del gen 16S rRNA obtenidos utilizando la tecnología NGS Illumina. Como restricción se buscaron herramientas de acceso libre para su uso, y actualizadas, así como estar disponibles para ejecutarlas en entorno LINUX.

### **4.3 Instalación y configuración de herramientas bioinformáticas**

Se instaló y configuró QIIME 2 y PICRUST2 en un entorno LINUX, en un servidor OpenSuse leap 42.2. El equipo debía cumplir con las características mínimas para su ejecución, tales como espacio en disco de al menos 25 GB, y memoria RAM de 16 GB.

#### **4.3.1 Preprocesamiento, agrupación y clasificación taxonómica**

Para realizar el análisis, se organizaron los datos de las secuencias de ARNr 16S, los cuales fueron generados posterior a la secuenciación y estaban en formato FASTQ (. fq). Adicionalmente se organizó el archivo de metadatos en el cual se incluye información con datos básicos de las secuencias a analizar.

Para el paso de procesamiento se realizó filtrado de la calidad, eliminando lecturas con calidad baja, de acuerdo con el umbral establecido y eliminación de secuencias correspondientes a primers o adaptadores. Se realizó identificación y exclusión de quimeras. Posteriormente se desmultiplexaron las secuencias, para eliminar los cebadores y marcadores añadidos a las secuencias.

Las secuencias fueron agrupadas en ASV utilizando el complemento DADA 2, generando una tabla ASV para los análisis posteriores. A partir de la tabla ASV generada, las secuencias representativas fueron clasificadas, utilizando clasificadores de taxonomía con las bases de datos HOMD y SILVA.

### **4.3.2 Análisis de diversidad**

Se realizó normalización de los datos y análisis de la curva de rarefacción alfa para explorar la integridad y diversidad de cada muestra para obtener resultados imparciales en los análisis posteriores de diversidad.

Se evaluó la diversidad alfa y la diversidad beta utilizando diferentes índices. Para alfa diversidad se utilizaron los estimadores de uniformidad, diversidad de Shannon, y distancia de Faith, mientras que la beta diversidad fue basada en los índices de Unifrac ponderado y no ponderado, distancia de Jaccard y disimilitud de Bray- Curtis.

### **4.3.3 Análisis de funcionalidad**

Se utilizaron los archivos ASV (seq.fna) generados por QIIME 2 como entrada en PICRUSt 2 para predecir las funciones biológicas de las comunidades, basadas en ortólogos (KO), de la Enciclopedia de Kioto de Genes y Genomas (KEGG), números de clasificación de enzimas, Grupos de genes ortólogos (COG), familias de proteínas (Pfam), y la base de datos de familias de proteínas del Instituto de Investigación Genómica. PICRUSt fue instalado en un entorno Bioconda.

### **4.3.4 Análisis estadísticos**

Se realizó comparación de alfa diversidad utilizando Kruskal-Wallis para establecer diferencias, y se evaluó mediante un análisis de varianza multivariado permutacional (PERMANOVA) la diferencia entre grupos. Se tuvo en cuenta un  $p < 0,05$ , para considerar las diferencias significativas, y cuando  $p < 0,01$  y  $p < 0,001$ , las diferencias se consideraron altamente significativas. Para comparar los perfiles taxonómicos en cada grupo evaluado se utilizó la prueba t de Welch con el software STAMP 2.1.3.

## **4.4 Implementación del flujo de trabajo**

Se realizó la implementación de las herramientas bioinformáticas seleccionadas en un servidor Dell Power Edge M640 con 96Gb de RAM, doble procesador Xeon con sistema operativo Linux OpenSuse 15.6. Utilizando Miniconda se realizó la instalación de la distribución de amplicones QIIME2 y de PICRUSt2.

## **4.5 Análisis de microbiomas orales de pacientes colombianos con y sin gingivitis**

Se seleccionaron secuencias asociadas con muestras de cavidad oral de pacientes colombianos con periodontitis y el grupo control sano/gingivitis. Los datos de secuenciación fueron descargados de la base de datos NCBI Sequence Read Archive (SRA) identificados con número de acceso PRJNA828047.

Las muestras secuenciadas fueron extraídas utilizando un kit comercial de extracción (QIAamp Mini Extraction Kit), y se secuenciaron la región V4 del gen 16S rRNA, en la plataforma Illumina Miseq (secuenciación pareada, lecturas 250 nucleótidos). Corresponden a placa subgingival de 20 pacientes diagnosticados con periodontitis, los cuales fueron clasificados en 4 estadios de periodontitis (periodontitis estadio I, periodontitis estadio II, periodontitis estadio III y periodontitis estadio IV), agrupados en periodontitis I-II y periodontitis III-IV y 10 pacientes diagnosticados periodontalmente como sanos/gingivitis, considerado el grupo control. Las características generales de las secuencias que fueron descargadas de la base de datos SRA se observan en la tabla 1.

Se prepararon los archivos requeridos para el procesamiento en QIIME 2 incluyendo las secuencias (forward y reverse), el archivo de metadatos y el de manifiesto en los formatos especificados por la documentación de QIIME 2.

Tabla 1. Características generales asociadas a las secuencias seleccionadas

<b>Nombre</b>	<b>Código</b>	<b>Diagnóstico</b>	<b>Muestra</b>	<b>Origen</b>
COL1	SRR18809707	Health&Gingivitis	placa subgingival	cavidad oral humana
COL2	SRR18809706	Health&Gingivitis	placa subgingival	cavidad oral humana
COL3	SRR18809671	Health&Gingivitis	placa subgingival	cavidad oral humana
COL4	SRR18809660	Health&Gingivitis	placa subgingival	cavidad oral humana
COL5	SRR18809697	Health&Gingivitis	placa subgingival	cavidad oral humana
COL6	SRR18809686	Health&Gingivitis	placa subgingival	cavidad oral humana
COL7	SRR18809651	Health&Gingivitis	placa subgingival	cavidad oral humana
COL8	SRR18809650	Health&Gingivitis	placa subgingival	cavidad oral humana
COL9	SRR18809649	Health&Gingivitis	placa subgingival	cavidad oral humana
COL10	SRR18809648	Health&Gingivitis	placa subgingival	cavidad oral humana
COL11	SRR18809705	I-II_Periodontitis	placa subgingival	cavidad oral humana
COL12	SRR18809680	I-II_Periodontitis	placa subgingival	cavidad oral humana
COL13	SRR18809679	I-II_Periodontitis	placa subgingival	cavidad oral humana
COL14	SRR18809678	I-II_Periodontitis	placa subgingival	cavidad oral humana
COL15	SRR18809677	I-II_Periodontitis	placa subgingival	cavidad oral humana
COL16	SRR18809676	I-II_Periodontitis	placa subgingival	cavidad oral humana
COL17	SRR18809675	I-II_Periodontitis	placa subgingival	cavidad oral humana
COL18	SRR18809674	I-II_Periodontitis	placa subgingival	cavidad oral humana
COL19	SRR18809673	I-II_Periodontitis	placa subgingival	cavidad oral humana
COL20	SRR18809672	I-II_Periodontitis	placa subgingival	cavidad oral humana
COL21	SRR18809670	I-II_Periodontitis	placa subgingival	cavidad oral humana
COL22	SRR18809669	I-II_Periodontitis	placa subgingival	cavidad oral humana
COL23	SRR18809668	III-IV_Periodontitis	placa subgingival	cavidad oral humana
COL24	SRR18809667	III-IV_Periodontitis	placa subgingival	cavidad oral humana
COL25	SRR18809666	III-IV_Periodontitis	placa subgingival	cavidad oral humana
COL26	SRR18809665	III-IV_Periodontitis	placa subgingival	cavidad oral humana
COL27	SRR18809664	III-IV_Periodontitis	placa subgingival	cavidad oral humana
COL28	SRR18809663	III-IV_Periodontitis	placa subgingival	cavidad oral humana
COL29	SRR18809662	III-IV_Periodontitis	placa subgingival	cavidad oral humana
COL30	SRR18809661	III-IV_Periodontitis	placa subgingival	cavidad oral humana

## **5. Capítulo 5**

### **5.1 Resultados y discusión**

#### **5.1.1 Principales pasos y herramientas para el análisis taxonómico y funcional de datos metagenómicos**

En la literatura se describe que el análisis de secuencias del gen 16S rRNA consta de varios pasos, entre los que se encuentran el preprocesamiento de secuencias, la clasificación taxonómica, la comparación de la diversidad taxonómica, el análisis de abundancia diferencial y el análisis funcional (Bokulich et al., 2018).

Para el preprocesamiento, en el cual se eliminan datos no informativos como adaptadores, cebadores de PCR y bases de baja calidad, se han desarrollado diversas herramientas integradas. De acuerdo con lo reportado en la literatura, QIIME 2 es uno de los frame work más utilizados para realizar el preprocesamiento de secuencias del gen 16S rRNA (Srivastava et al.,2024), proceso que se realiza posterior a la importación de las secuencias dentro del ambiente qiime2 (figura 1), que incluye la comprobación de calidad de la secuencia, la corrección de errores, filtrado y la asignación de las secuencias (Hall & Beiko, 2018), proceso que es completado en menor tiempo en comparación con otras herramientas como MOTHUR (Marizzoni et al., 2020).. En la mayor parte de los reportes de la literatura se utiliza una estrategia basada en DADA2 (Callahan et al., 2016) para el preprocesamiento también conocida como denoising de secuencias, que identifica Exact ASVs (Amplicon Sequence Variants) con resolución de un solo nucleótido, por lo general utilizando el complemento de denoising qiime DADA2 (q2-dada2), incorporado en QIIME2. El denoising comienza con la demultiplexación, sigue con el filtrado y recorte de extremos de baja calidad de las secuencias, el modelado del error a partir del perfil de los errores de las secuencias, con lo que genera un modelo estadístico que le permite inferir las secuencias originales verdaderas. El proceso DADA2 continua con el solapamiento de las secuencias pareadas (merging) y concluye con la eliminación de

quimeras. DADA2 ha demostrado tener buena sensibilidad al detectar ASV verdaderos, al lograr una alta precisión en la cuantificación de la abundancia, y se consideró como la mejor opción para estudios que requieren una mayor resolución biológica (Prodan et al., 2020), mejorando la clasificación taxonómica de los organismos identificados en los microbiomas estudiados (Callahan et al., 2017). Se ha demostrado que los flujos de trabajo basados en ASV presentan mayor resolución en comparación con OTU, y tiene una mejor especificidad, permitiendo una mejor integración de las características biológicas (Prodan et al., 2020). Se ha encontrado que los OTU conducen proporcionalmente a una subestimación en los indicadores de diversidad, y un comportamiento distorsionado de índices asociados con dominancia y uniformidad, en comparación con los ASV, por lo que el agrupamiento basado en ASV presenta ventajas significativas en comparación con OTU (Fasolo et al., 2024). El análisis resultante de ASV genera una tabla de abundancias que contiene los recuentos absolutos o relativos de los microorganismos observados en las muestras analizadas, evidenciando el número de veces que se observó un ASV en cada muestra y una tabla que proporciona una secuencia representativa para cada característica. Posterior al análisis se visualiza el resumen de las lecturas para observar el número de lecturas que se mantuvieron en cada paso utilizado en el análisis, revisando de manera general la integridad del análisis. A partir de los resultados se pueden utilizar filtros adicionales que seleccionen secuencias con recuentos bajos, para evitar interferencias disminuyendo el riesgo de que se genere influencia sobre los resultados generales de composición (Reitmeier et al., 2021).

En cuanto a la asignación taxonómica uno de los más utilizados es el clasificador Naive Bayes de aprendizaje automático para asignar taxonomías probables a las lecturas, el cual asigna la taxonomía con un método preentrenado por QIIME2 con la base de datos seleccionada, que permite determinar la afiliación taxonómica más cercana con un grado de confianza o consenso (Bokulich et al., 2018). Se ha descrito que los métodos basados en aprendizaje automático pueden superar a otros métodos de clasificación como los clasificadores basados en el alineamiento de las secuencias los cuales pueden no ser tan adecuados, ya que puede haber secuencias que tengan coincidencias parecidas, pero en con anotaciones taxonómicas diferentes (Kaehler et al., 2019). Para la asignación de la clasificación taxonómica es necesario seleccionar una la base de datos especializada asociada al entorno biológico específico del estudio, entorno que para el presente trabajo es la cavidad oral, y para el que se ha implementado la base de datos HOMD, que proporciona una representación de los microorganismos que suelen encontrarse en dicha cavidad. Aunque esta base de datos suele ser recomendada, al complementarla con bases de datos más amplias se puede mejorar la asignación especialmente de especies raras (Regueira-Iglesias et al., 2023), por lo que una buena estrategia es la comparación de los resultados obtenidos utilizando una base de datos especializada como HOMD

y SILVA, las cuales proporcionan un conjunto de datos completos, de calidad y son actualizadas periódicamente, considerando que SILVA es utilizada dentro de QIIME2, y que en estudios con comunidades simuladas demostró que predecía el número correcto de géneros en comparación con otras bases de datos .

Para evaluar la cantidad de especies presentes en las muestras y comparar la diversidad y composición del microbioma entre los grupos se realizan análisis de alfa y beta diversidad, los cuales tienen en cuenta el número de microorganismos, su abundancia y su relación filogenética. En la literatura se reporta a menudo el uso constante de diversos índices para evaluar diversidad, la cual se puede calcular a partir de las ASV inferidas utilizando QIIME 2 con el complemento de SampleData [AlphaDiversity], que contiene estimaciones de la diversidad alfa para cada muestra de la tabla de características, y se realizan comparaciones estadísticas basadas en pruebas de Kruskal-Wallis para comparar entre grupos. Los análisis de diversidad requieren previamente una normalización de los recuentos de las lecturas utilizando un método filogenético de métricas centrales, basada en una profundidad específica por medio de qiime diversity alpha-rarefaction, con la finalidad de corregir las diferencias en la profundidad de lectura. Así mismo algunas métricas de diversidad requieren la construcción de un árbol filogenético para calcular la diversidad teniendo en cuenta la similitud filogenética (Regueira-Iglesias et al., 2023). Teniendo en cuenta los reportes de la literatura, en los que se menciona que es importante considerar métricas que reflejen resultados de riqueza, dominancia, información y relaciones filogenéticas (Cassol et al., 2025), se incluyen los índices más usuales que se utilizan en los estudios, tales como índice de Shannon, uniformidad y diversidad de Faith para la diversidad alfa. Los análisis de beta diversidad son basados en matrices de distancia / disimilitud, y para evaluar diferencias se incluyen análisis estadísticos como PERMANOVA (Anderson 2017), utilizando el comando beta-group-significance. Los índices más utilizados para calcular la diversidad beta son distancia de Jaccard, distancia de Bray-Curtis y distancia de UNIFRAC ponderada y no ponderada, siendo algunos índices basados en la proporción de especies y otros que solo contemplan ausencia y presencia de taxones, o que combinan las dos características, e incluso en índices como Unifrac se tiene en cuenta adicionalmente información sobre las distancias filogenéticas entre las especies observadas (Kers & Saccenti, 2022). Para estimar la diversidad, basada en alfa y beta diversidad, se utilizó el complemento q2-diversity de QIIME 2, en el cual se incluyen diferentes pasos para calcular y visualizar resultados de diversidad (Figura 1).

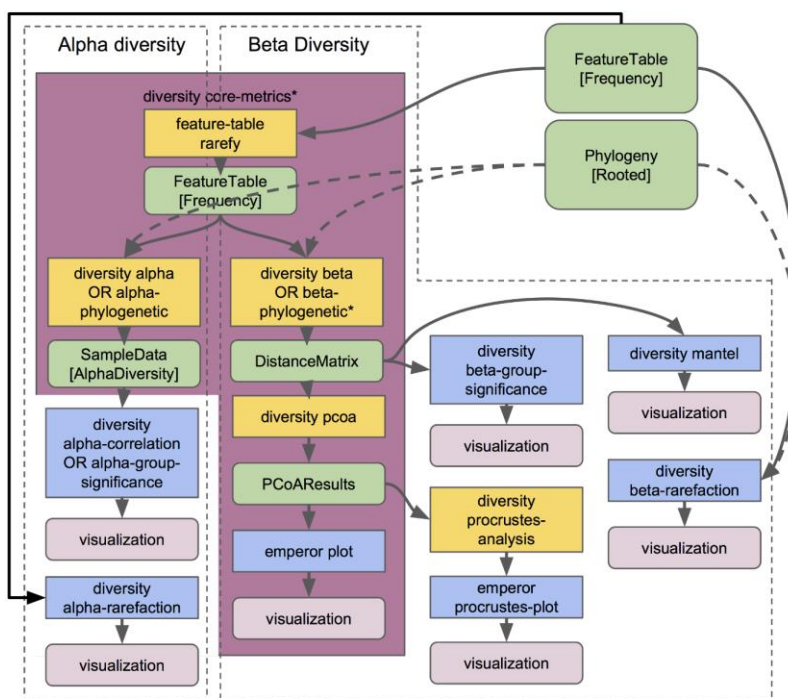


Figura 1. Análisis de alfa y beta diversidad realizados en Qiime2

Tomada de <https://docs.qiime2.org/2024.10/tutorials/overview/#demultiplexing>

Para la estimación de la función del microbioma basado en marcadores moleculares como el gen 16S rRNA, PICRUST2 (Douglas et al., 2020) es ampliamente utilizado, y ha demostrado mayor precisión al evaluar perfiles metagenómicos, en comparación con otras herramientas como Tax4Fun2 (Wemheuer et al., 2020). Así mismo se ha observado que tiene un buen desempeño general en la predicción de la presencia de términos funcionales KO (ortología KEGG), y es una herramienta que utiliza referencias recientes, tales como la base de datos Integrated Microbial Genomes and Microbiomes (IMG/M), para predecir el contenido funcional de las comunidades microbianas (Matchado et al., 2024). En el flujo de trabajo de PICRUST2 se realiza un análisis filogenético, predicción del estado oculto y tabulación de la abundancia de genes y vías por muestra. Las secuencias y abundancias de ASV se toman como entrada, y las abundancias de familias de genes y vías son la salida (Figura 2). La salida de PICRUST2 puede analizarse utilizando STAMP, programa a partir del cual se pueden realizar análisis estadísticos y representaciones gráficas del análisis del perfil funcional predicho (Parks et al., 2014)

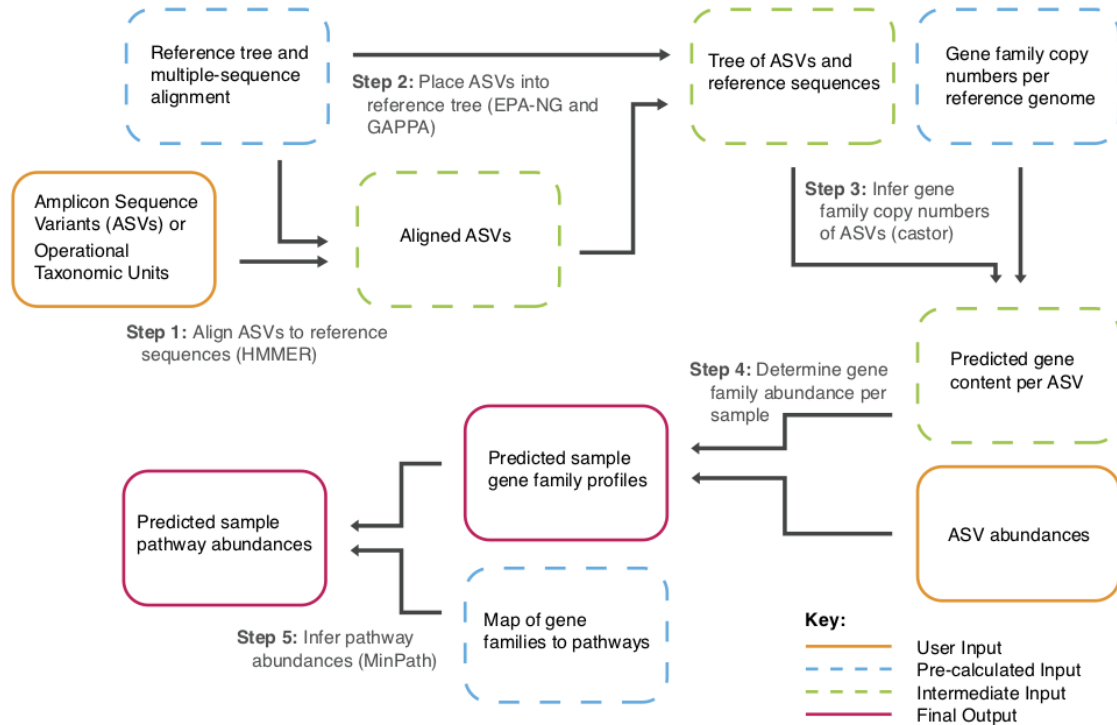


Figura 2. Flujo de trabajo de PICRUSt2 tomado de Douglas et al., 2020

### 5.1.2 Selección de la estrategia a implementar

Teniendo en cuenta los resultados obtenidos en el objetivo 1, la estrategia seleccionada fue basada principalmente en QIIME 2 2024.10.1, utilizando diferentes plugins (Tabla 2). Se realizó el preprocesamiento de las secuencias sin procesar en formato FASTQ, obtenidas de la secuenciación del gen 16S rRNA, las cuales se demultiplexaron y se analizó su calidad utilizando el plugin q2-demux, posteriormente para la agrupación y clasificación de las secuencias se utilizó DADA2 mediante el plugin q2-dada2, lo que permitió identificar las secuencias de variante de amplicón (ASV), obtenidos utilizando DADA2 para asignar etiquetas taxonómicas utilizando qiime feature-classifier classify-sklearn y las base de datos HOMD y SILVA.

Para el análisis de alfa diversidad se construyó un árbol filogenético que tiene en cuenta la relación evolutiva entre las secuencias de ADN, y se establecieron las métricas de diversidad alfa y beta comparando los resultados obtenidos por grupos, utilizando qiime diversity core-metrics-phylogenetic y qiime diversity alpha-group-significance. Para diversidad alfa se evaluaron los índices de diversidad de Shannon, características observadas y diversidad filogenética de Faith, y

para diversidad beta se analizaron distancia de Jaccard, Distancia Bray Curtis, Distancia UniFrac ponderada y no ponderada

Para el análisis de funcionalidad se utilizaron los resultados obtenidos en QIIME2 que corresponden a las secuencias representativas de los ASV obtenidos, en formato fasta, y la tabla de abundancia de ASV, que fueron obtenidos en los análisis de denoising con DADA2, y se obtuvo el perfil funcional predictivo utilizando PICRUST2, el cual fue analizado utilizando STAMP 2.1.3 para evaluar el perfil funcional comparando entre grupos.

Tabla 2. Resumen de pasos y herramientas utilizadas en la estrategia a implementar

<b>Paso/Herramienta</b>	<b>Características</b>	<b>Resultado</b>
Preprocesamiento QIIME 2 plugin q2-demux	Se demultiplexan las secuencias	Secuencias sentido y antisentido
Comprobación de primers y adaptadores utilizados para la amplificación de la región V4 del gen 16S rRNA	Eliminar secuencias de primers o adaptadores	Secuencias filtradas
Control de calidad DADA 2 plugin q2-dada2 qiime dada2 denoise-paired qiime metadata tabulate qimme feature-table summarize qimme feature-table tabulate-seqs	Establecer ASV, eliminación de quimeras, filtrado de abundancia, unión de lecturas pareadas.	Resumen de estadísticas con las secuencias filtradas Tabla de características y secuencias representativas Secuencias definidas como ASV
Filtrado de secuencias con recuentos bajos QIIME2 Qiime feature-table filter-samples	Se filtran secuencias que contienen pocas secuencias de ASV, que representan inconvenientes en la secuenciación	ASV filtradas

Clasificación taxonómica QIIME 2 qiime feature-classifier classify-sklearn Base de datos HOMD Base de datos SILVA	Determinar la clasificación de los taxones a nivel de filos, géneros y especies comparando con una base de datos de referencia de secuencias con composición taxonómica conocida	Anotación taxonómica de cada secuencia determinada como ASV
Comparar diferencias entre grupos STAMP	Análisis de perfiles taxonómicos, comparando estadísticamente pares o grupos de muestras	Gráficos exploratorios de resultados estadísticos (gráficos de barra extendida)
Análisis de diversidad QIIME 2 q2-diversity qiime diversity core-metrics-phylogenetic	Evaluar similitud o diferencias entre muestras grupos.  Construcción de árbol filogenético	Estimaciones de diversidad para cada muestra Matriz de distancia/disimilitud Resultados de análisis de coordenadas principales PCoA
Análisis de funcionalidad PICRUST2	Predecir las funciones biológicas de las comunidades	Predicciones metagenómicas de vías funcionales
Análisis de diferencias funcionales del microbioma entre grupos y análisis de vías STAMP	Análisis de perfiles funcionales, comparando entre grupos	Principales vías funcionales Comparación entre grupos de las vías metabólicas predichas

### 5.1.3. Implementación de la estrategia

Considerando que qiime2 incluye herramientas para realizar todos o casi todos los pasos de la estrategia, se instaló Qiime 2 v.2024.10.1 en un servidor Dell Power Edge M640 con 96Gb de RAM,

doble procesador Xeon con sistema operativo Linus OpenSuse 15.6, utilizando un entorno miniconda, por la línea de comando.

### 5.1.4 Análisis de microbiomas orales de pacientes colombianos con y sin gingivitis

- **Análisis de calidad de las secuencias**

Se obtuvo un total de 1.410.021 lecturas brutas a partir de las secuencias pareadas con una longitud promedio de 251 nucleótidos, mediante secuenciación de alto rendimiento de la región V4 del gen ARNr 16S de 30 muestras, 20 correspondientes a pacientes con periodontitis y 10 a pacientes con diagnóstico periodontal de salud/gingivitis (Tabla 3). La calidad de las secuencias fue el resultado del plugin qiime demux summarize, el cual permite obtener un resumen de las secuencias demultiplexadas con sus puntuaciones de calidad.

Tabla 3. Recuentos de secuencias demultiplexadas

	Lecturas directas	Lecturas reversas
<b>Mínimo</b>	2.879	2.879
<b>Mediana</b>	27.396	27.396
<b>Promedio</b>	47.000,7	47.000,7
<b>Máximo</b>	174.361	174.361
<b>Total</b>	1.410.021	1.410.021

- **Control de calidad de las secuencias y construcción de tablas de características**

Teniendo en cuenta la calidad de las secuencias (figura 3), la cual fue en promedio de 38 para todas las secuencias sentido y antisentido, se consideró que no era necesario realizar la eliminación de nucleótidos de baja calidad de los extremos de las lecturas ya que se presentaron en una muy baja proporción nucleótidos con valores de calidad por debajo de 20.

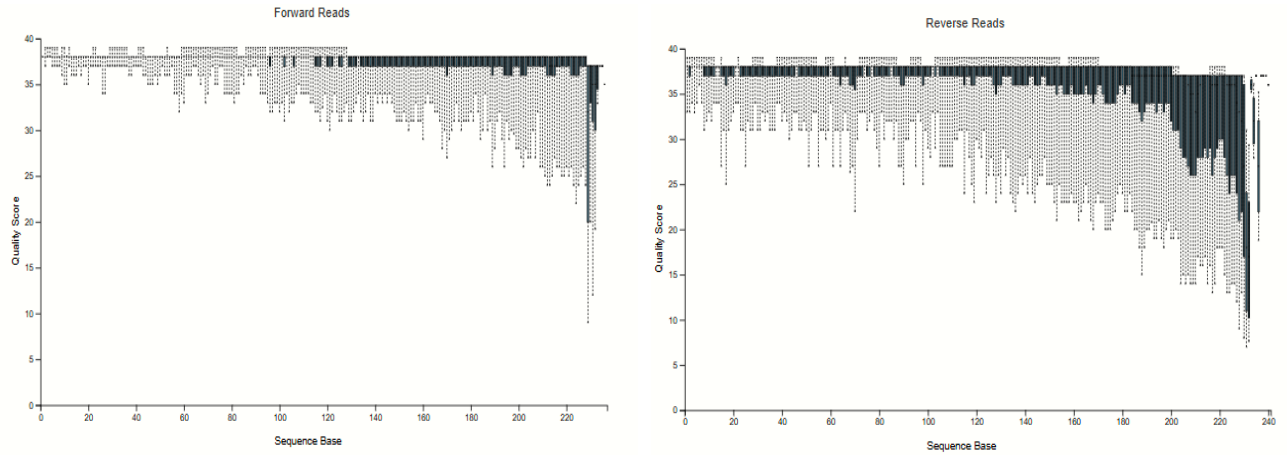


Figura 3. Representación de la calidad de las secuencias forward y reverse.

Se realizó comprobación de la presencia de primers y adaptadores teniendo en cuenta los primers utilizados para la amplificación de la región V4 del gen 16S rRNA. Posterior al ajuste de calidad de las secuencias se obtuvieron 1352171, obteniendo secuencias libres de primers que puedan generar interferencia con los análisis posteriores (Tabla 4).

Tabla 4. Resumen de recuentos de secuencias posterior al control de calidad

	<b>Lecturas directas</b>	<b>Lecturas reversas</b>
<b>Mínimo</b>	2677	2677
<b>Mediana</b>	26590.0	26590.0
<b>Promedio</b>	45072.366667	45072.366667
<b>Máximo</b>	166906	166906
<b>Total</b>	1352171	1352171

#### ▪ Denoising

Se realizó denoising con DADA2. Al finalizar el filtrado con DADA2, se observó un total de 1524 ASV. En la tabla se observan los resultados del proceso de denoising, denotando los resultados según

cada muestra analizada utilizando el comando qiime metadata tabulate, en el cual se observa información que de las lecturas que pasaron los diversos pasos del filtrado de calidad (Tabla 5).

Tabla 5. Resumen de rangos de secuencias que se conservaron después de denoising

sample-id	input	filtered	percentage of input passed filter	denoised	merged	percentage of input merged	non-chimeric	percentage of input non-chimeric
SRR18809648	17213	16463	95.64	16294	16087	93.46	15831	91.97
SRR18809649	16081	14958	93.02	14651	14103	87.7	13845	86.1
SRR18809650	25338	23381	92.28	23099	22503	88.81	22215	87.67
SRR18809651	9796	9117	93.07	8875	8495	86.72	8417	85.92
SRR18809660	26131	23929	91.57	23723	23139	88.55	22735	87
SRR18809661	12174	10872	89.31	10740	10606	87.12	10438	85.74
SRR18809662	76565	64313	84	63701	62527	81.67	58493	76.4
SRR18809663	77640	64412	82.96	63515	61681	79.44	58317	75.11
SRR18809664	108965	91229	83.72	90510	89291	81.94	88032	80.79
SRR18809665	2677	2323	86.78	2187	2108	78.74	2108	78.74
SRR18809666	34093	27354	80.23	26731	26016	76.31	25804	75.69
SRR18809667	64387	53673	83.36	52698	51226	79.56	49744	77.26
SRR18809668	91808	75733	82.49	75204	74137	80.75	73768	80.35
SRR18809669	23254	21892	94.14	21541	20828	89.57	20477	88.06
SRR18809670	166906	137390	82.32	136121	133799	80.16	131794	78.96
SRR18809671	18773	17167	91.45	16922	16397	87.34	16360	87.15
SRR18809672	43178	35846	83.02	35475	34928	80.89	34116	79.01
SRR18809673	9738	9066	93.1	8890	8468	86.96	8371	85.96
SRR18809674	47908	39932	83.35	39488	38832	81.06	37976	79.27
SRR18809675	25741	24225	94.11	23939	23388	90.86	23320	90.59
SRR18809676	70030	58023	82.85	57332	56210	80.27	55248	78.89
SRR18809677	93758	77799	82.98	76854	75605	80.64	69993	74.65
SRR18809678	14041	12466	88.78	12229	11883	84.63	11443	81.5
SRR18809679	67895	55594	81.88	54812	53515	78.82	52328	77.07
SRR18809680	85758	71595	83.48	71173	70339	82.02	69563	81.12
SRR18809686	27049	25060	92.65	24714	23973	88.63	23789	87.95
SRR18809697	24901	23292	93.54	22989	22436	90.1	22196	89.14
SRR18809705	33808	31651	93.62	31150	29720	87.91	29036	85.88
SRR18809706	20271	18357	90.56	18004	17219	84.94	17000	83.86
SRR18809707	16294	15134	92.88	14978	14617	89.71	14425	88.53

Se realizó filtrado de la tabla de características para filtrar las secuencias que aparecen solo en una muestra, y así reducir el número de secuencias final y dejar las más representativas, adicionalmente

se realizó filtrado excluyendo taxones específicos como mitocondrias y cloroplastos, obteniendo al final 780 ASV y una frecuencia total de 975.675 (Tabla 6). Eliminar las secuencias con bajos recuentos disminuye los tiempos de procesamiento y evita secuencias que suelen representar muestras con amplificación o secuenciación incorrecta, así mismo, el eliminar secuencias correspondientes a mitocondrias o cloroplastos, que no son parte del microbioma bacteriano, disminuye la distorsión de la abundancia relativa de los taxones.

Tabla 6. Secuencias representativas

<b>Summary Statistic</b>	<b>Value</b>
Number of samples	30
Number of features	780
Total frequency	975.675

- **Composición taxonómica**

Al utilizar el clasificador taxonómico preentrenado Naive Bayes con la base de datos SILVA, y entrenar el calificador con la base de datos HOMD se realizaron las anotaciones taxonómicas a partir de los ASV establecidos. Se evaluaron las diferencias generales de cada base de datos teniendo en cuenta el número de filios identificados con cada base de datos, así como el número de géneros y número de especies específicas, de acuerdo con los resultados de cada base de datos en las muestra totales (tabla 7).

Tabla 7. Comparación de resultados obtenidos en la clasificación taxonómica de acuerdo con la aproximación utilizada y la base de datos

	<b>SILVA</b>	<b>HOMD</b>
Número de filios identificados	15	11
Número de géneros identificados	151	289
Número de especies identificadas	0	247

Se realizaron análisis adicionales con un enfoque en la composición de bacterias y la abundancia relativa de cada taxón utilizando el clasificador classify-sklearn con la base de datos SILVA y se obtuvieron 15 filos (figura 4), los cuales correspondieron a *Firmicutes* (40,8%), *Fusobacteriota* (28,8%), *Bacteroidota* (10,3%), *Proteobacteria* (7,4%), *Actinobacteriota* (4,2%), *Spirochaetota* (3,1%) *Desulfobacterota* (1,9%), *Campylobacterota* (1,7%) *Synergistota* (1,3%) *Patescibacteria* (0,4%) *Chloroflexi* (0,1%), *Euryarchaeota* (0,01%), *Verrucomicrobiota*, (0,01%), *Myxococcota* (0,002%), y *Elusimicrobiota* (0,002%). Con la base de datos HOMD, que es específica para microorganismos de la cavidad oral se encontraron 11 filos (Figura 5), que a nivel general correspondieron a *Firmicutes* (40.8%), *Fusobacteria* (28.8%), *Proteobacteria* (11.1%), *Bacteroidetes* (10.3%), *Actinobacteria* (4.2%), *Spirochaetes* (3.1%), *Synergistetes* (1.3%), *Absconditabacteria\_(SR1)* (0.2%), *Saccharibacteria\_(TM7)* (0.2%), *Chloroflexi* (0.1), y *Euryarchaeota* (0,01%).

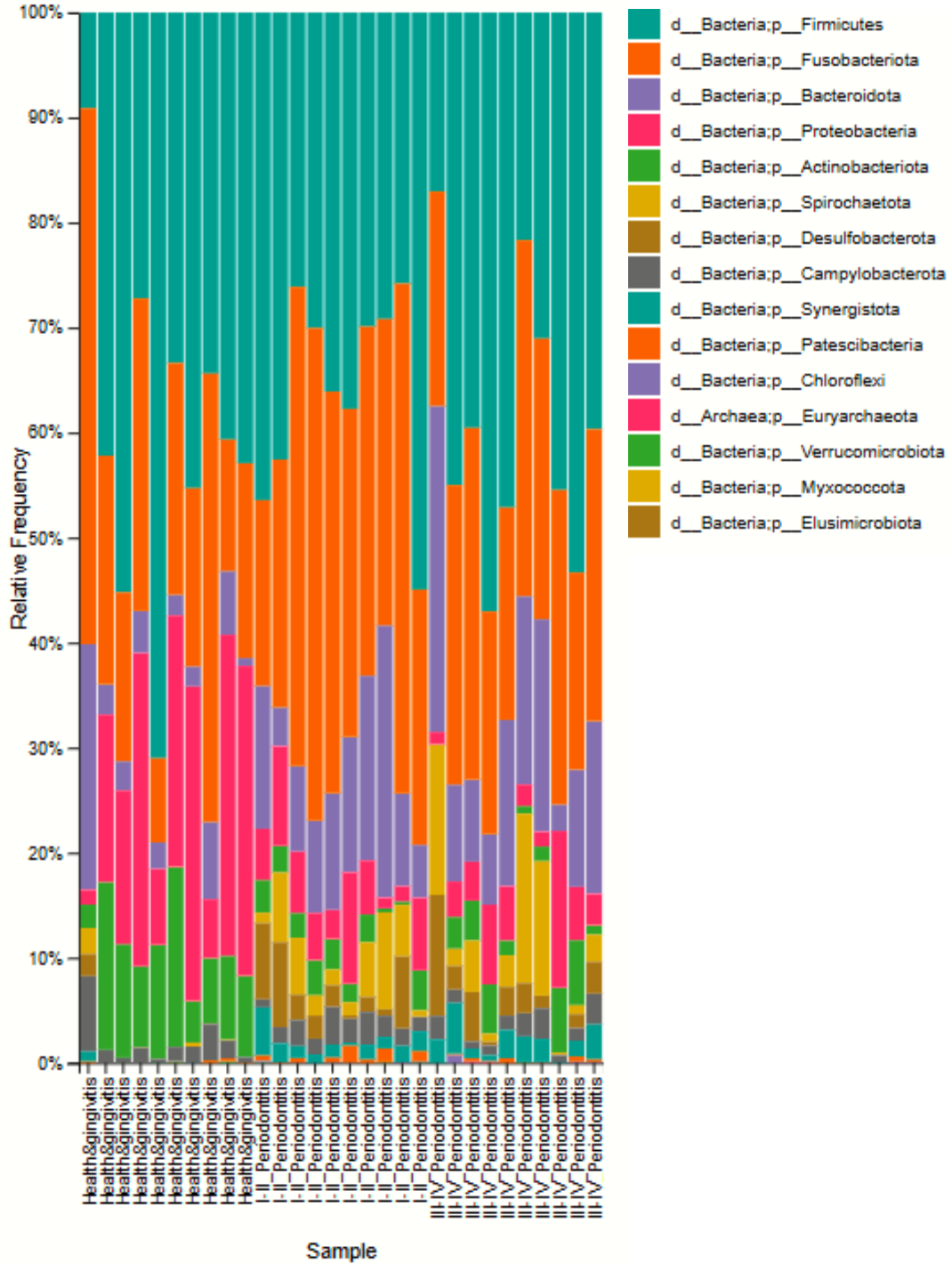


Figura 4. Diagrama de barras de frecuencia relativa de los filios identificados utilizando la base de datos SILVA.

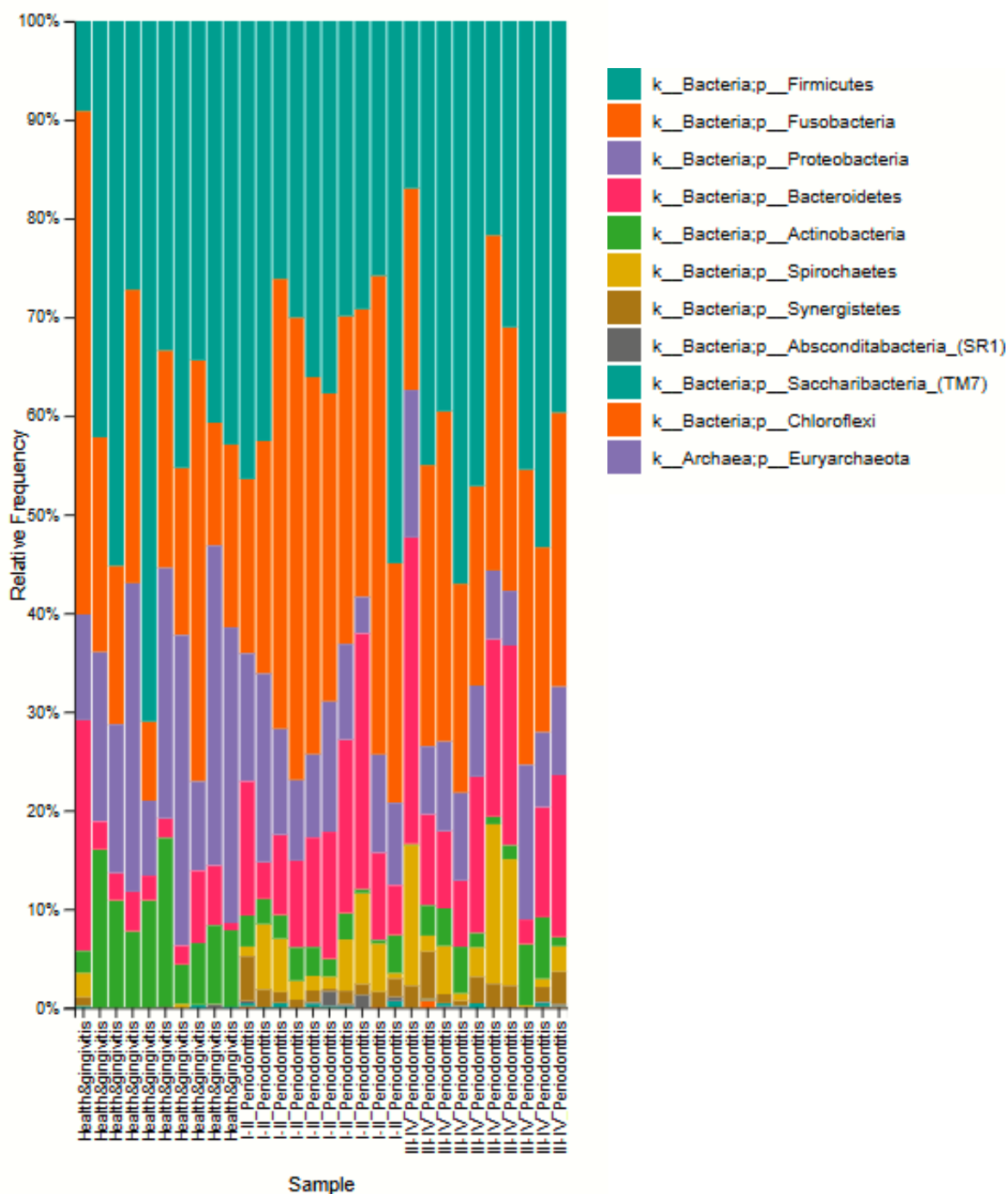


Figura 5. Diagrama de barras de frecuencia relativa de los filios identificados utilizando la base de datos HMD

Con la base de datos SILVA la clasificación solo fue posible hasta el nivel de género, encontrando principalmente *Fusobacterium* (22.1%), *Streptococcus* (11.2%), *Leptotrichia* (4.5%), *Porphyromonas* (3,8%), *Filifactor* (3,3%), *Treponema* (3,1%), *Tannerella* (2.9%), *Parvimonas* (2.6%), *Veillonella* (2.4%), *Gemella* (2%), *Campylobacter* (1,7%), *W5053* (1,6%), *Desulfobulbus* (1.5%), *[Eubacterium]\_saphenum\_group* (1.3%), *Fretibacterium* (1.3%), *Centipeda* (1.2%), *Selenomonas* (1.2%), *Rothia* (1.2%), *Lachnoanaerobaculum* (1,1%), *F0058* (1,1%), *Corynebacterium* (1.1%) y *Actinomyces* (1,2%) ( Figura 6), mientras que con la base de datos HOMD se logró la identificación hasta especie encontrando la presencia de *Fusobacterium* (22.1%), *Streptococcus* (9.9%), *Filifactor alocis* (3.2%), *Porphyromonas gingivalis* (2.6%), *Sneathia* (2.3%), *Tannerella forsythia*, (2.2%), *Haemophilus parainfluenzae* (2%), *Parvimonas* (2%), *Gemella* (2%), *Selenomonas* (2%), *Treponema* (1.9%), *Campylobacter* (1.6%), *Veillonella dispar* (1.6%), *Peptoniphilaceae [G-1]bacterium\_HMTbacterium\_HMT\_113* (1.6%), *Desulfobulbus HMT\_041* (1.5 %) *Peptostreptococcaceae [XI][G-5]Eubacterium\_saphenum* (1.3%), *Leptotrichia HMT\_417* (1.2%) y *Bacteroidales [G-2]bacterium\_HMT\_274* (1.1%) y *Corynebacterium* (1.1%) (Figura 7).

Los resultados obtenidos con las diferentes bases de datos demuestran que HOMD fue más exhaustiva al lograr identificar a nivel de especie, encontrando algunos géneros en común entre las bases de datos con una frecuencia similar. Se identificaron microorganismos que se han reportado actualmente asociados con periodontitis, considerados como nuevos periodonto patógenos como el caso de *Filifactor alocis* y *Desulfobulbus sp.\_HMTsp.\_HMT\_041*. Estos microorganismos pueden tener un papel importante en la enfermedad (Antezack et al., 2023).

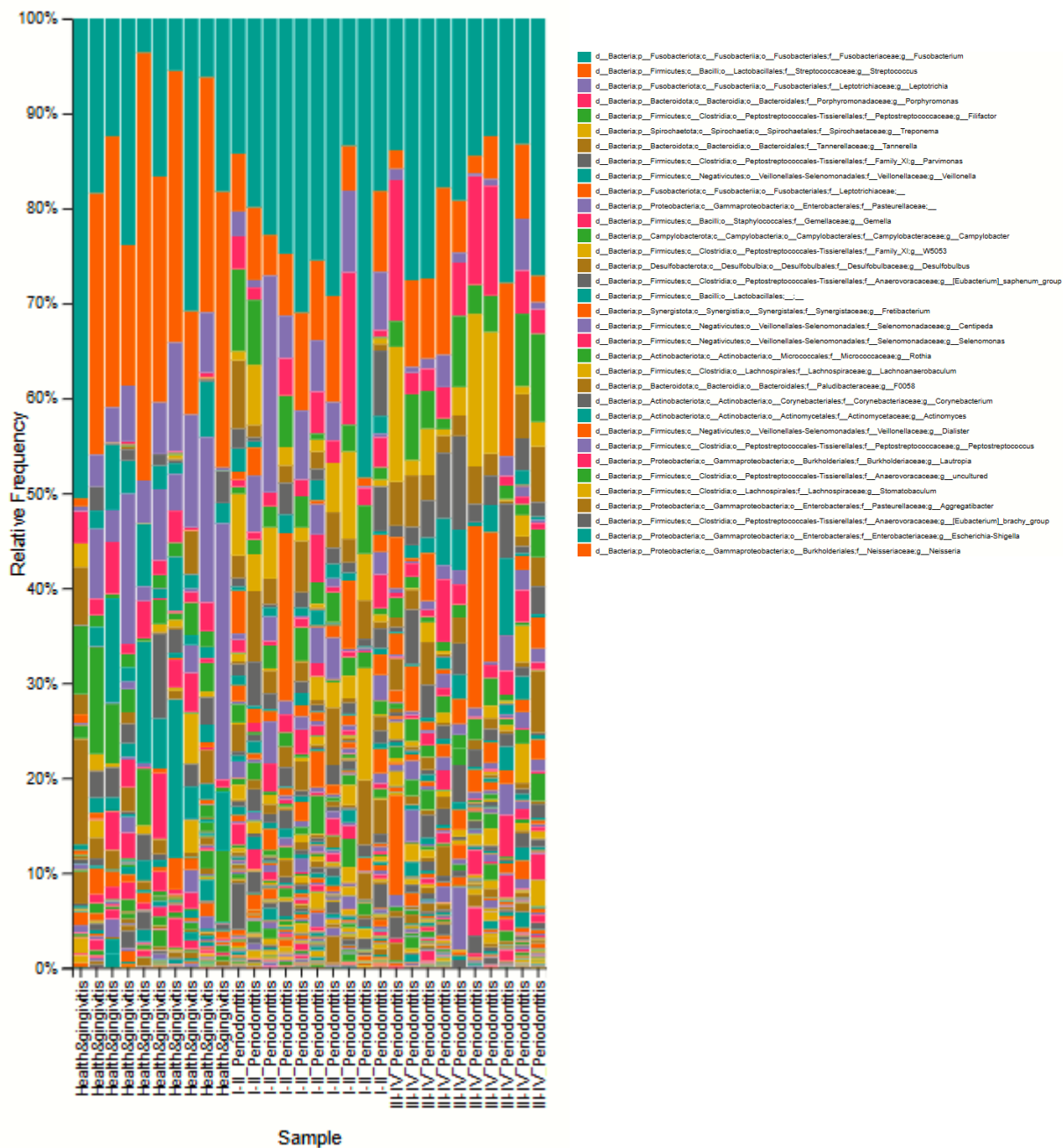


Figura 6. Diagrama de barras de frecuencia relativa de los géneros identificados utilizando la base de datos SILVA



Utilizando STAM se realizaron comparaciones analizando las bacterias identificadas con la base de datos SILVA que presentaron diferencias entre el grupo de salud/gingivitis y los grupos en general de periodontitis (figura 8). Se observó que en salud se presentaba mayor abundancia de microorganismos como *Streptococcus*, *Rothia*, *Corynebacterium* y *Actinomyces*, géneros que se consideran microbiota normal de la cavidad oral (Caselli et al., 2020). Los principales microorganismos que se asociaron con periodontitis, los cuales presentaron diferencias estadísticamente significativas, fueron *Filifactor*, *Fretibacterium*, *Eubacterium nodatum group*, *Eubacterium saphenum group*, *Dialister*, *Porphyromonas*, *Desulfubulbus*, *Tannerella* y *Parvimonas*. Estos géneros bacterianos se han asociado con la enfermedad periodontal, algunos de ellos considerados en la actualidad como nuevos periodonto patógenos (Huang et al., 2021). Analizando los resultados obtenidos con SILVA, pero comparando únicamente entre los grupos de periodontitis se observaron diferencias estadísticamente significativas en *Clostridia\_UCG-014*, y *Leptotrichia*, asociados principalmente con el grupo de periodontitis I y II, estas bacterias que se han asociado con periodontitis (Figura 9) (Singh, 2017).



Figura 8. Análisis STAMP de las diferencias en la abundancia bacteriana entre grupo salud/gingivitis y los grupos de periodontitis, utilizando la prueba t de Welch.  $p < 0.05$

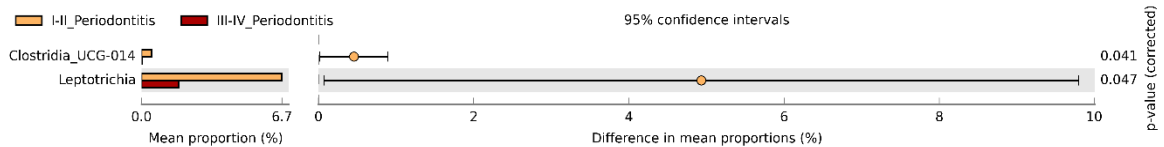


Figura 9. Análisis STAMP de las diferencias en la abundancia bacteriana entre grupo salud/gingivitis y los grupos de periodontitis, utilizando la prueba t de Welch.  $p < 0.05$

Al evaluar con HOMD, en la comparación de las bacterias que presentaron diferencias estadísticamente significativas ( $p < 0.05$ ) entre el grupo de salud/gingivitis y los grupos en general de periodontitis (Figura 10), se encontró que las bacterias que se asociaban con salud/gingivitis fueron *Streptococcus*, *Corynebacterium*, *Haemophylus parainfluenzae* y *Rothia dentocariosa*. Estos microorganismos se asocian en general con salud periodontal, y son resultados coincidentes con lo encontrado con la base de datos SILVA. En cuanto a los microorganismos que se asociaron con

periodontitis principalmente se observaron especies como *Filifactor alocis*, *Eubacterium saphenum*, *Dialister invisus*, *Fretibacterium*, *Parvimonas*, *Treponema*, *Desulfubulbus* sp. HMT 168, *Porphyromonas gingivalis*, *Eubacterium brachy*, *Tannerella forythia*, *Selenomonas*, *Sneathia*, *Peptoniphilaceae* [G1] bacterium HMT 094, y *Porphyromonas* sp. HMT 257.

Al analizar los resultados utilizando HOMD entre los grupos de periodontitis, se puede observar en la figura 11, que los microorganismos a nivel de especie que presentaron diferencias estadísticamente significativas asociándose con periodontitis I-II fueron *Leptotrichia* sp. HMT 417, *Ruminococcaceae* [G-1] bacterium HMT 075, *Capnocytophaga*, *Selenomonas* sp. HMT 137 y *Aggregatibacter actinomycetemcomitans*. Los géneros *Leptotrichia* y *Capnocytophaga* se han asociado tanto con la enfermedad como con la salud, lo que sugiere que estas bacterias podría ser potenciales patógenos y variar su comportamiento de acuerdo con el nicho en el que se encuentren (Chen et al., 2018). En cuanto a *Aggregatibacter actinomycetemcomitans*, esta bacteria se considera periodonto patógeno de baja abundancia, el cual se ha encontrado en pacientes con periodontitis agresiva, de acuerdo con la clasificación periodontal de 1999 (Fine et al., 2019)



Figura 10. Análisis de las diferencias en la abundancia bacteriana entre grupo salud/gingivitis y los grupos de periodontitis, utilizando HOMD (prueba t de Welch.  $p < 0.05$ ).

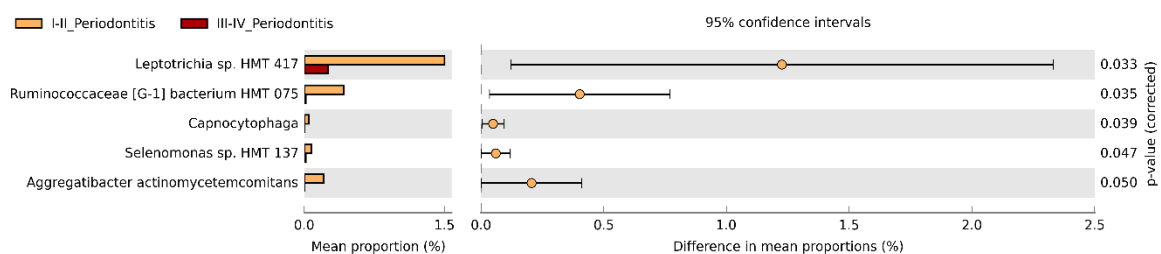


Figura 11. Análisis de las diferencias en la abundancia bacteriana entre grupo periodontitis I-II y periodontitis III-IV, utilizando HOMD (prueba t de Welch.  $p < 0.05$ ).

Teniendo en cuenta los resultados generales a partir de las bases de datos utilizadas para la clasificación se observan diferencias importantes en la clasificación principalmente a nivel de géneros y especies, encontrando un número superior de estas categorías taxonómicas en la clasificación realizada con HOMD. Se ha discutido que las bases de datos de referencias son cruciales en la anotación taxonómica y pueden influenciar los resultados generando diferentes composiciones taxonómicas. Sierra et al, en 2020 evaluaron la influencia de las bases de datos sobre en análisis de la comunidad microbiana encontrando que el uso de bases personalizadas asociadas al microbioma específico podría evitar anotaciones erróneas y mejorar la asignación taxonómica en niveles inferiores como la especie, aunque también podrían omitirse algunos taxones no contenidos en las bases de datos especializadas por lo que utilizar estrategias utilizando bases de datos complementarias podría mejorar la asignación taxonómica.

#### ▪ Análisis de Diversidad

En los análisis de diversidad realizados con QIIME2 utilizando *qiime diversity alpha-rarefaction*, se realizó previamente el ajuste de las diferencias en la profundidad de secuenciación utilizando una curva de rarefacción, la cual se estabilizó en los 3 grupos, lo cual indicó una profundidad de secuenciación suficiente (Figura 12). El análisis de rarefacción permite observar la riqueza de la comunidad teniendo en cuenta el número de secuencias por muestra.

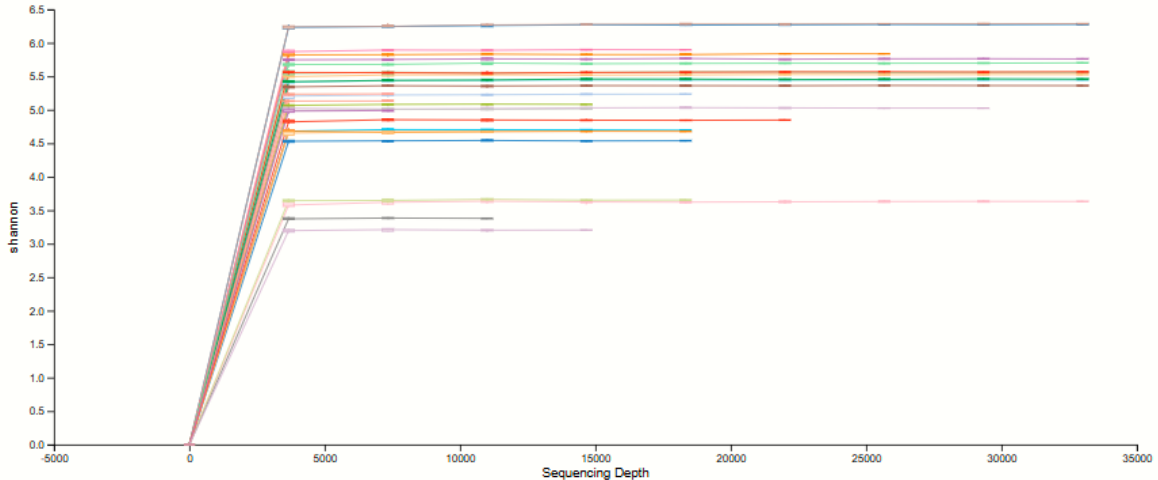


Figura 12. Curva de rarefacción

Utilizando la tabla de característica anteriormente obtenida, el árbol filogenético construido y los metadatos se calculó la diversidad  $\alpha$  con QIIME2, comparando los índices de uniformidad, índice de Shannon y diversidad de Faith, entre los grupos, con Kruskal Wallis ( $p < 0.05$ ), encontrando diferencias estadísticamente significativas en el índice de uniformidad al comparar el grupo de salud/gingivitis con periodontitis I-II, y diferencias en el índice Faith e índice de Shannon al comparar el grupo de salud/gingivitis con el grupo de periodontitis I-II y el grupo de periodontitis III-IV, lo que se puede observar en la figura 13 y la tabla 8. Cada índice utilizado denota algunas características específicas de la comunidad siendo el índice de riqueza correspondiente al número de microorganismos diferentes entre las muestras, el índice de Faith la medida de biodiversidad basada en la distancia filogenética y el índice de Shannon un índice de información el cual refleja la cantidad de ASV diferentes que hay en cada muestra o grupo teniendo en cuenta que tan uniformemente se distribuyen, por lo que se aborda desde varias perspectivas la diversidad alfa, lo que permite obtener un análisis más preciso y significativo en relación con la composición de la microbiota (Cassol et al., 2025). Los resultados obtenidos son comparables con otros estudios que han encontrado mayor o menor diversidad alfa en comparación con pacientes sanos, aunque en el índice de riqueza se ha reportado que los pacientes sanos presentan menor riqueza microbiana en comparación con pacientes con periodontitis, independiente del estadio (Lafaurie et al., 2022). Otros autores como Iniesta et al., en 2023 observaron mayor diversidad bacteriana con los índices riqueza y Shannon en pacientes españoles con gingivitis y periodontitis, en comparación con el grupo de salud, siendo similares los resultados observados con los pacientes de periodontitis en el presente estudio, aunque se encontraron diferencias con el índice de diversidad de Faith, en el cual no se

observaron diferencias estadísticamente significativas entre los grupos evaluados en los pacientes españoles.

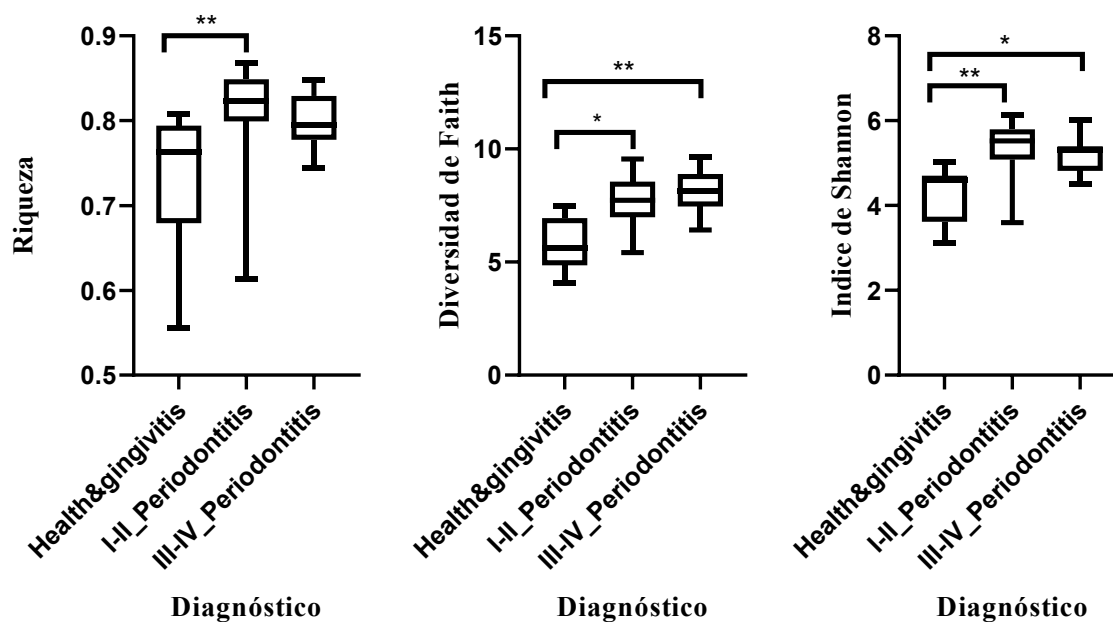


Figura 13. Análisis de alfa diversidad utilizando los índices de uniformidad, diversidad de Faith e índice de Shannon. Comparaciones entre grupos con Kruskal Wallis \*\* $p < 0.001$ , \* $p < 0,05$

Tabla 8. Comparación de características observadas (riqueza) e índice de Shannon entre grupos con Kruskal Wallis ( $p < 0.005$ )

Grupo	Riqueza p-value	Faith_PD p-value	Shannon p-value
Salud/gingivitis- Periodontitis I-II	0.007	0.011	0.001
Salud/gingivitis- Periodontitis III-IV	0.160	0.001	0.033
Periodontitis I-II- Periodontitis III-IV	0.824	>0.999	>0.999

#### ▪ Análisis de beta diversidad

Las muestras fueron agrupadas mediante análisis de componentes principales utilizando el complemento de QIIME2 qimme emperor plot, el cual permite visualizar los datos en gráficos de

ordenación. Al evaluar los índices de beta diversidad se puede observar en los índices de Bray curtis, Jaccard y Unifrac ponderado, dos agrupaciones principales, una correspondiente a las muestras del grupo sano/gingivitis (blanco), y otra de las muestras de periodontitis I-II y periodontitis III-IV (naranja-rojo) (Figura 14), encontrando diferencias entre los grupos de salud/gingivitis y periodontitis, y mayor similitud en los microorganismos asociados a las muestras de los grupos de periodontitis, considerando la abundancia relativa de los microorganismos.

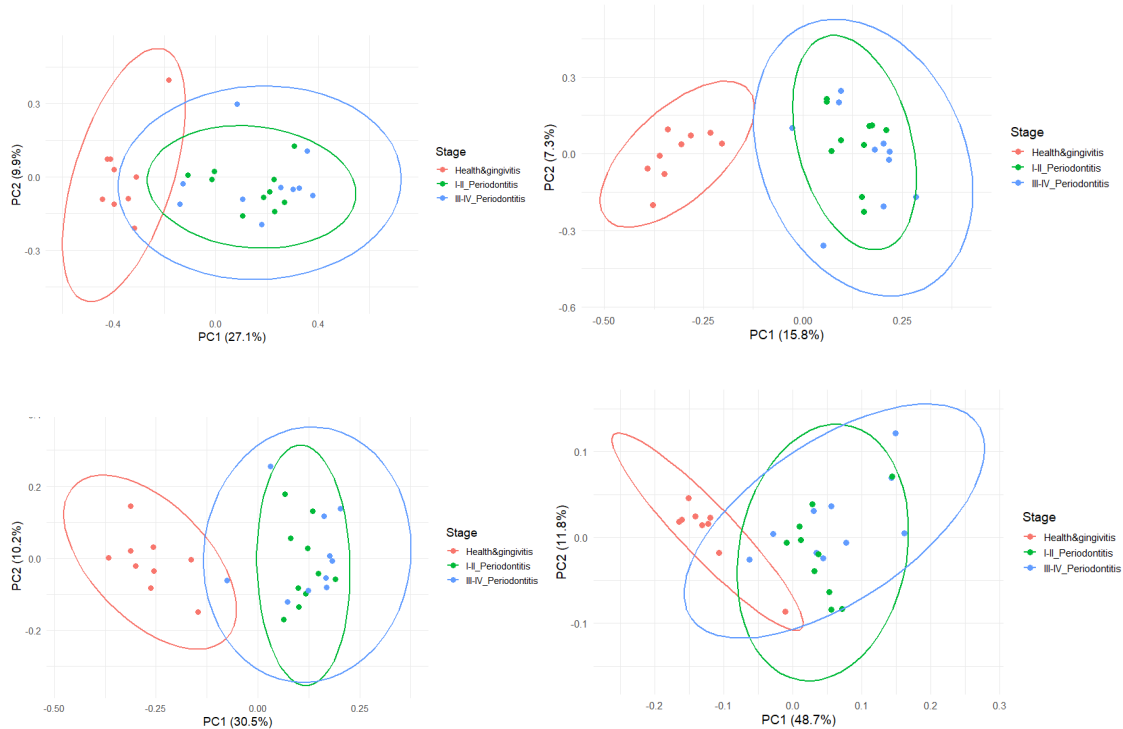


Figura 14. Análisis de beta diversidad evaluando los índices Bray -Curtis (arriba-izquierda), Jaccard (arriba-derecha), Unifrac no ponderado (abajo-izquierda), Unifrac ponderado (abajo-derecha)

En el índice Bray-Curtis, el cual evalúa la diferencia en la abundancia de especies, el eje 1 explica el 27.05% y el eje 2 el 9,9% de la variación total, observándose una agrupación más cercana entre los grupos de periodontitis, aunque se observa una superposición parcial entre el grupo de salud/gingivitis y los grupos de periodontitis, lo que podría asociarse con la disimilitud encontrada en los pacientes con periodontitis. Un patrón similar se observó al analizar los índices de Unifrac ponderado que se basa en las distancias filogenéticas considerando la abundancia relativa de los taxones y Unifrac no ponderado que solo considera presencia o ausencia. Respecto al índice de Jaccard, que tiene en cuenta la presencia o ausencia de especies, se observó una separación más clara

ente grupos, observando que el grupo de salud/gingivitis se separa de los grupos de periodontitis. De manera general también se observa que en el grupo de salud/gingivitis hay más homogeneidad dentro del grupo, mientras que el grupo de periodontitis III-IV presenta mayor alta variabilidad.

Considerando la agrupación diferencial en el índice Jaccard se realizó un análisis comparando entre los grupos, utilizando un análisis multivariado (PERMANOVA), y se evidenciaron diferencias significativas de las comunidades microbianas entre los grupos de periodontitis comparados con el de salud/gingivitis (figura 15, tabla 9). Estos resultados son similares a los reportados por Lafaurie et al., en 2022, aunque otros estudios han encontrado que el microbioma de los pacientes con periodontitis presentó una alta abundancia de microorganismos en comparación con el grupo de salud, al realizar comparaciones utilizando un análisis PERMANOVA, aunque sin encontrar diferencias estadísticamente significativas entre los dos grupos (Narayanan et al., 2023).

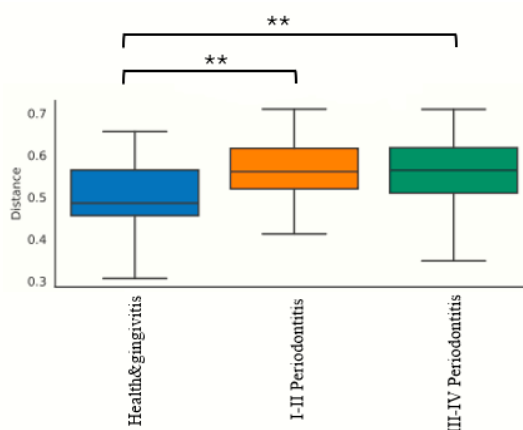


Figura 15 Comparaciones de distancias entre grupos utilizando PERMANOVA para el índice de Jaccard. \*\*  $p < 0.001$

Tabla 9. Análisis de beta diversidad comparando entre grupos utilizando análisis multivariado (PERMANOVA)

Grupo	p-value
Salud/gingivitis- Periodontitis I-II	0.001
Salud/gingivitis- Periodontitis III-IV	0.001
Periodontitis I-II- Periodontitis III-IV	0.197

### ▪ Predicción de la funcionalidad

Utilizando el método PICRUSt2, basado en la base de datos KEGG, se realizó la predicción de las posibles funciones bioquímicas asociadas a variaciones de la microbiota oral en las diferentes condiciones de los pacientes del estudio.

Se identificaron 27 vías que presentaron diferencias estadísticamente significativa entre los grupos de pacientes, las cuales se asociaron con biosíntesis y metabolismo de glucanos, metabolismo de cofactores y vitaminas, procesamiento de información genética; plegamiento, clasificación y degradación, metabolismo de terpenoides y policétidos, biosíntesis de otros metabolitos secundarios, biodegradación y metabolismo de xenobióticos, adaptación, motilidad celular y procesamiento de información genética (transcripción). Algunas otras vías se asociaron con enfermedades humanas incluyendo cáncer y enfermedades infecciosas causadas por bacterias (Tabla 10).

Tabla 10. Principales vías metabólicas basadas en MetaCyc identificadas en la predicción realizada con Picrust2.

Característica	p_values	p_ajustado	Nombre de la vía
ko00563	0.0014436446	0.020211025	Biosíntesis del ancla de glicosilfosfatidilinositol (GPI)
ko00740	0.0035384748	0.031083073	Metabolismo de la riboflavina
ko00760	0.0051331996	0.040345148	Metabolismo del nicotinato y la nicotinamida
ko04141	0.0016407863	0.022274918	Procesamiento de proteínas en el retículo endoplásmico
ko04974	0.0054673911	0.042230883	Digestión y absorción de proteínas
ko01051	0.0024200750	0.025213805	Biosíntesis de ansamicinas
ko00524	0.0018133961	0.023521618	Biosíntesis de neomicina, kanamicina y gentamicina
ko00950	0.0009374990	0.015649846	Biosíntesis de alcaloides isoquinolínicos
ko05120	0.0002704654	0.008077900	Señalización de células epiteliales en la infección por <i>Helicobacter pylori</i>
ko00625	0.0057568353	0.043712919	Degradación de cloroalcanos y cloroalquenos
ko00624	0.0003570358	0.008771306	Degradación de hidrocarburos aromáticos policíclicos
ko01053	0.0011916756	0.017546092	Biosíntesis de péptidos no ribosómicos del grupo sideróforo
ko05215	0.0021856976	0.024928191	cáncer de próstata
ko04914	0.0021794982	0.024928191	Maduración de ovocitos mediada por progesterona
ko03050	0.0023370179	0.024928191	Proteasoma
ko04626	0.0009987664	0.015980262	Interacción planta-patógeno
ko00983	0.0061086387	0.044863445	Metabolismo de fármacos: otras enzimas
ko02030	0.0004209581	0.009429462	Quimiotaxis bacteriana
ko04612	0.0022383242	0.024928191	Procesamiento y presentación de antígenos
ko04621	0.0020879773	0.024928191	Vía de señalización del receptor tipo NOD
ko05014	0.0058693281	0.043824316	esclerosis lateral amiotrófica

ko03020	0.0023227173	0.024928191	ARN polimerasa
ko00860	0.0009431827	0.015649846	Metabolismo de las porfirinas
ko00908	0.0048344069	0.039378441	Biosíntesis de zeatina
ko04930	0.0071014510	0.048945385	Diabetes mellitus tipo II
ko00900	0.0064405765	0.046538359	Biosíntesis de la estructura principal de los terpenoides
ko02040	0.0031913355	0.029785798	Ensamblaje flagelar

Utilizando STAMP se identificaron 23 vías funcionales que presentaron diferencias significativas entre los grupos comparados por la prueba t de Welch ( $p < 0.05$ ) (Figura 16). La mayoría de estas vías se asociaron con los grupos de periodontitis, lo que podría sugerir una alta actividad metabólica de los microorganismos asociados con esta condición clínica. Las vías que se asociaron con periodontitis se relacionan con glicogénesis, biosíntesis de cobalamina, gluconeogénesis, biosíntesis de difosfato de geranylgeranilo II (Vía MEP), fermentación de piruvato a acetona, biosíntesis de coenzima A I, biosíntesis de novo de desoxirribonucleótidos de pirimidina III y supervía de L-aspartato y L-asparagina. Se ha mencionado que las vías metabólicas del microbioma tienen un papel importante en la conformación de la microbiota oral disbiótica que se asocia con periodontitis, algunas de estas vías se relacionan con elementos esenciales que se requieren para funciones vitales, tales como la biosíntesis de pirimidinas, proceso necesario para la síntesis de nucleótidos y proliferación microbiana, la síntesis de cofactores que están involucrados en diversos procesos metabólicos (Moradi et al., 2025) y la supervía de L-aspartato y L-asparagina que participan en el proceso metabólico que es esencial para la viabilidad bacteriana y están involucradas en bacterias principalmente periodonto patógenas (Li et al., 2023).

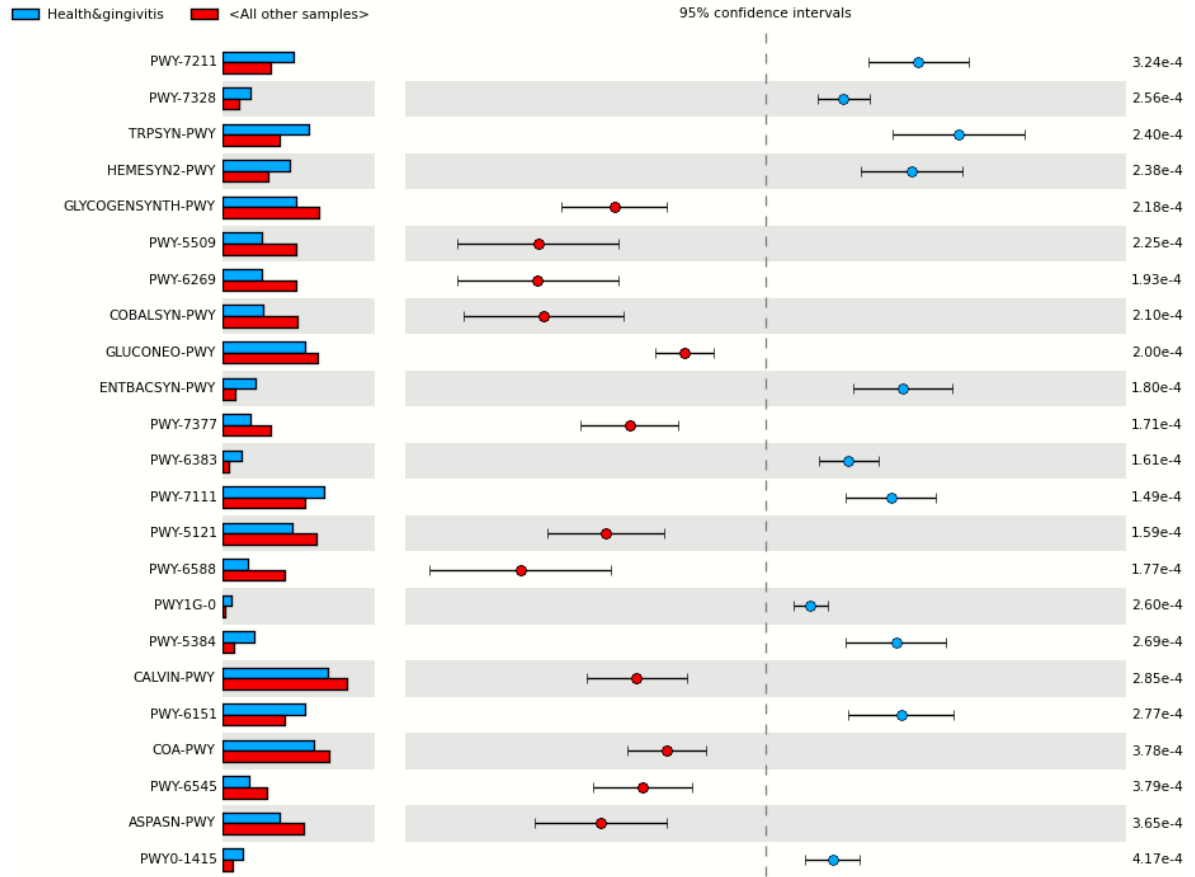


Figura 16. Predicción de vías metabólicas que presentaron diferencias significativas entre el grupo salud/gingivitis y los grupos de periodontitis.

## 6. Capítulo 6

### 6.1 Conclusiones

Qiime 2 es un framework que permite realizar análisis múltiples, incluyendo genes marcadores como 16S rRNA, lo que confirma su uso para análisis de comunidades microbianas en conjunto con complementos y herramientas adicionales.

La estrategia utilizada se basó en QIIME 2, DADA2 y PICRUST2, considerando los pasos necesarios para obtener secuencias de calidad adecuada que permitieran análisis de clasificación taxonómica, diversidad y predicción funcional, logrando obtener visualizaciones interpretables validados y análisis estadísticos.

La base de datos HOMD se consideró como la mejor opción para estudios de la microbiota de cavidad oral que requieren una mayor resolución biológica, en comparación con SILVA.

Se evidenciaron microorganismos específicos más frecuentes asociados con periodontitis tales como *Filifactor alocis*, *Fretibacterium*, *Eubacterium nodatum group*, *Eubacterium saphenum*, *Eubacterium brachy*, *Dialister invisus*, *Porphyromonas gingivalis*, *Desulfubulbus*, *Selenomonas*, *Sneathia*, *Treponema*, *Tannerella forythia* y *Parvimonas*.

*Leptotrichia sp. HMT 417*, *Ruminococcaceae [G-1] bacterium HMT 075*, *Capnocytophaga*, *Selenomonas sp. HMT 137* y *Agregatibacter actinomycetemcomitans* fueron especies asociadas con periodontitis I- II, constituyendo posibles marcadores de estos estadios de la enfermedad.

La mayoría de las vías metabólicas predichas con PICRUST2 se asociaron con los grupos de periodontitis, lo que podría sugerir una alta actividad metabólica de los microorganismos asociados con esta condición clínica.

## 6.2 Recomendaciones

Se debe tener en cuenta el clasificador a utilizar, el cual debe corresponder a la región secuenciada, y utilizar bases de datos actualizadas o complementarias para mejorar los resultados en la clasificación taxonómica.

Aumentar el número de secuencias podría mejorar los resultados, evidenciando mayores asociaciones entre los microorganismos identificados y la condición periodontal en la población local.

El uso de plataformas de secuenciación de lecturas largas del gen 16S rRNA pueden mejorar la resolución taxonómica, obteniendo mejores resultados en los análisis.

Se podrían implementar scripts automatizados para mejorar el procesamiento de los datos utilizando QIIME2, DADA2 y PICRUST2

## Anexo 1: Script

```
# Importar secuencias
qiime tools import \
--type 'SampleData[PairedEndSequencesWithQuality]' \
--input-path q2_manifest.csv \
--output-path importar/seq_import.qza \
--input-format PairedEndFastqManifestPhred33
# Resumen de secuencias importadas
qiime demux summarize \
--i-data importar/import.qza \
--o-visualization importar/seq_import.qzv
# Denoising con DADA2
mkdir 02_denoising
qiime dada2 denoise-paired \
--i-demultiplexed-seqs importar/seq_import.qza \
--p-trunc-len-f 0 \
--p-trunc-len-r 0 \
--p-n-threads 2 \
--o-representative-sequences 03_denoising/secuencias_asv.qza \
--o-table 03_denoising/asv-table.qza \
--o-denoising-stats 03_denoising/dada2-stats.qza
# Secuencias representativas
qiime feature-table summarize \
--i-table 03_denoising/asv-table.qza \
--m-sample-metadata-file metadata.tsv \
--o-visualization 03_denoising/feature-table-summ.qzv
qiime feature-table tabulate-seqs \
--i-data 03_denoising/asv-sequences.qza \
--o-visualization 03_denoising/asv-sequences-summ.qzv
# Clasificar secuencias
qiime feature-classifier classify-sklearn \
```

```

--i-classifier gg-13-8-99-V4-nb-classifier.qza \
--i-reads 03_denoising/asv-sequences.qza \
--o-classification 05_taxonomia/taxonomy.qza
qiime metadata tabulate \
--m-input-file 05_taxonomia/taxonomy.qza \
--o-visualization 05_taxonomia/taxonomy.qzv
#Arbol filogenético
qiime alignment mafft \
--i-sequences asv-sequences.qza \
--o-alignment aligned-asv-rep-seqs.qza
qiime phylogeny fasttree \
--i-alignment aligned-asv--rep-seqs.qza \
--o-tree rooted-tree.qza
#Análisis de diversidad
qiime diversity core-metrics-phylogenetic \
--i-phylogeny phylogeny-align-to-tree-mafft-fasttree/rooted_tree.qza \
--i-table filtered-table.qza \
--p-sampling-depth \
--p-n-jobs-or-threads 8 \
--m-metadata-file metadata.tsv \
--output-dir 07_diversidad diversity-core-metrics-phylogenetic
qiime diversity alpha-group-significance \
--i-alpha-diversity 07_diversidad/diversity-core-metrics-
phylogenetic/observed_features_vector.qza \
--m-metadata-file metadata.tsv \
--o-visualization 07_diversidad/alpha-group-sig-obs-feats.qzv
#Análisis de beta diversidad
qiime diversity beta-group-significance \
--i-distance-matrix 07_diversidad/diversity-core-metrics-
phylogenetic/unweighted_unifrac_distance_matrix.qza \
--m-metadata-file metadata.tsv \
--m-metadata-column 'Stage' \

```

```
--p-method 'permanova' \  
--o-visualization 07_diversidad/permanova_results.qzv  
#Análisis de funcionalidad  
#Exportar secuencias  
picrust2_pipeline.py -s exported-file/dna-sequences.fasta -i phyloseq/feature-table.biom -o  
picrust2_out_pipeline -p 1  
biom head -i phyloseq/feature-table.biom  
biom summarize-table -i phyloseq/feature-table.biom  
biom convert -i phyloseq/feature-table.biom -o feature-table.tsv --to-tsv  
#Análisis con Picrust2  
place_seqs.py -s exported-file/dna-sequences.fasta -o out.tre -p 4 --intermediate place_seqs  
metagenome_pipeline.py -i exported-file/dna-sequences.fasta -m 16S_predicted_and_nsti.tsv.gz -f  
COG_predicted.tsv.gz -o COG_metagenome_out --strat_out  
picrust2_pipeline.py -s exported-file/dna-sequences.fasta -i phyloseq/feature-table.biom -o  
picrust2_out_pipeline -p 16 --in_traits COG,EC,KO --remove_intermediate --stratified
```

## Bibliografía

Abellan-Schneyder, I., Machado, M. S., Reitmeier, S., Sommer, A., Sewald, Z., Baumbach, J., List, M., & Neuhaus, K. (2021). Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing. *mSphere*, 6(1), e01202-20. <https://doi.org/10.1128/mSphere.01202-20>

Amaya, J., Peñaloza Quintero, E., Palacio Basto, Y., Carlos Gómez Serrano, L., María Becerra Garnica, A., Suárez Zúñiga, E., Alejandro Garnica, J., Sánchez García, H., & Uzaheta, A. (2014). IV ESTUDIO NACIONAL DE SALUD BUCAL. ENSAB IV. Metodología y Determinación Social de la Salud Bucal.

Anderson, M. J. (2017). Permutational Multivariate Analysis of Variance (PERMANOVA). En *Wiley StatsRef: Statistics Reference Online* (pp. 1-15). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781118445112.stat07841>

Antezack, A., Etchecopar-Etchart, D., La Scola, B., & Monnet-Corti, V. (2023). New putative periodontopathogens and periodontal health-associated species: A systematic review and meta-analysis. *Journal of Periodontal Research*, 58(5), 893-906. <https://doi.org/10.1111/jre.13173>

Belda-Ferre, P., Alcaraz, L. D., Cabrera-Rubio, R., Romero, H., Simón-Soro, A., Pignatelli, M., & Mira, A. (2012). The oral metagenome in health and disease. *The ISME Journal*, 6(1), 46-56. <https://doi.org/10.1038/ismej.2011.85>

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* (Oxford, England), 30(15), 2114-2120. <https://doi.org/10.1093/bioinformatics/btu170>

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., ... Caporaso, J. G. (2018). QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science (e27295v2). PeerJ Inc. <https://doi.org/10.7287/peerj.preprints.27295v2>

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A., & Gregory Caporaso, J. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, 6(1), 90. <https://doi.org/10.1186/s40168-018-0470-z>

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13(7), 581-583. <https://doi.org/10.1038/nmeth.3869>

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, 11(12), 2639-2643. <https://doi.org/10.1038/ismej.2017.119>

Calle, M. L. (2019). Statistical Analysis of Metagenomics Data. *Genomics & Informatics*, 17(1). <https://doi.org/10.5808/GI.2019.17.1.e6>

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., Fierer, N., & Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences*, 108(supplement\_1), 4516-4522. <https://doi.org/10.1073/pnas.1000080107>

Cassol, I., Ibañez, M., & Bustamante, J. P. (2025). Key features and guidelines for the application of microbial alpha diversity metrics. *Scientific Reports*, 15(1), 622. <https://doi.org/10.1038/s41598-024-77864-y>

Caton, J. G., Armitage, G., Berglundh, T., Chapple, I. L. C., Jepsen, S., Kornman, K. S., Mealey, B. L., Papananou, P. N., Sanz, M., & Tonetti, M. S. (2018). A new classification scheme for periodontal and peri-implant diseases and conditions – Introduction and key changes from the 1999 classification. *Journal of Clinical Periodontology*, 45(S20), S1-S8. <https://doi.org/10.1111/jcpe.12935>

Chen, T., Yu, W.-H., Izard, J., Baranova, O. V., Lakshmanan, A., & Dewhirst, F. E. (2010). The Human Oral Microbiome Database: A web accessible resource for investigating oral microbe taxonomic and genomic information. *Database: The Journal of Biological Databases and Curation*, 2010. <https://doi.org/10.1093/database/baq013>

Chen, C., Hemme, C., Beleno, J., Shi, Z. J., Ning, D., Qin, Y., Tu, Q., Jorgensen, M., He, Z., Wu, L., & Zhou, J. (2018). Oral microbiota of periodontal health and disease and their changes after nonsurgical periodontal therapy. *The ISME Journal*, 12(5), 1210-1224. <https://doi.org/10.1038/s41396-017-0037-1>

Chiarello, M., McCauley, M., Villéger, S., & Jackson, C. R. (2022). Ranking the biases: The choice of OTUs vs. ASVs in 16S rRNA amplicon data analysis has stronger effects on diversity measures

than rarefaction and OTU identity threshold. *PLoS ONE*, 17(2), e0264443. <https://doi.org/10.1371/journal.pone.0264443>

Dias, C. K., Starke, R., Pylro, V. S., & Morais, D. K. (2020). Database limitations for studying the human gut microbiome. *PeerJ Computer Science*, 6, e289. <https://doi.org/10.7717/peerj-cs.289>

Douglas, G. M., Maffei, V. J., Zaneveld, J. R., Yurgel, S. N., Brown, J. R., Taylor, C. M., Huttenhower, C., & Langille, M. G. I. (2020). PICRUSt2 for prediction of metagenome functions. *Nature biotechnology*, 38(6), 685-688. <https://doi.org/10.1038/s41587-020-0548-6>

Edgar, R. C. (2013). UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, 10(10), 996-998. <https://doi.org/10.1038/nmeth.2604>

Erazo, N. G., Dutta, A., & Bowman, J. S. (2021). From microbial community structure to metabolic inference using paprica. *STAR Protocols*, 2(4), 101005. <https://doi.org/10.1016/j.xpro.2021.101005>

Fasolo, A., Deb, S., Stevanato, P., Concheri, G., & Squartini, A. (2024). ASV vs OTUs clustering: Effects on alpha, beta, and gamma diversities in microbiome metabarcoding studies. *PLOS ONE*, 19(10), e0309065. <https://doi.org/10.1371/journal.pone.0309065>

Fine, D. H., Patil, A. G., & Velusamy, S. K. (2019). *Aggregatibacter actinomycetemcomitans* (Aa) Under the Radar: Myths and Misunderstandings of Aa and Its Role in Aggressive Periodontitis. *Frontiers in Immunology*, 10. <https://doi.org/10.3389/fimmu.2019.00728>

Galloway-Peña, J., & Hanson, B. (2020). Tools for Analysis of the Microbiome. *Digestive diseases and sciences*, 65(3), 674-685. <https://doi.org/10.1007/s10620-020-06091-y>

Goldfarb, T., Kodali, V. K., Pujar, S., Brover, V., Robbertse, B., Farrell, C. M., Oh, D.-H., Astashyn, A., Ermolaeva, O., Haddad, D., Hlavina, W., Hoffman, J., Jackson, J. D., Joardar, V. S., Kristensen, D., Masterson, P., McGarvey, K. M., McVeigh, R., Mozes, E., ... Murphy, T. D. (2025). NCBI RefSeq: Reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Research*, 53(D1), D243-D257. <https://doi.org/10.1093/nar/gkae1038>

Gwak, H.-J., & Rho, M. (2020). Data-Driven Modeling for Species-Level Taxonomic Assignment From 16S rRNA: Application to Human Microbiomes. *Frontiers in Microbiology*, 11. <https://www.frontiersin.org/articles/10.3389/fmicb.2020.570825>

Haffajee, A. D., Socransky, S. S., Patel, M. R., & Song, X. (2008). Microbial complexes in supragingival plaque. *Oral microbiology and immunology*, 23(3), 196-205.

Hall, M., & Beiko, R. G. (2018). 16S rRNA Gene Analysis with QIIME2. *Methods in Molecular Biology* (Clifton, N.J.), 1849, 113-129. [https://doi.org/10.1007/978-1-4939-8728-3\\_8](https://doi.org/10.1007/978-1-4939-8728-3_8)

Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*, 68(4), 669-685. <https://doi.org/10.1128/MMBR.68.4.669-685.2004>

Hong, B.-Y., Furtado Araujo, M. V., Strausbaugh, L. D., Terzi, E., Ioannidou, E., & Diaz, P. I. (2015). Microbiome profiles in periodontitis in relation to host and disease characteristics. *PloS One*, 10(5), e0127077. <https://doi.org/10.1371/journal.pone.0127077>

Iwai, S., Weinmaier, T., Schmidt, B. L., Albertson, D. G., Poloso, N. J., Dabbagh, K., & DeSantis, T. Z. (2016). Piphillin: Improved Prediction of Metagenomic Content by Direct Inference from Human Microbiomes. *PLOS ONE*, 11(11), e0166104. <https://doi.org/10.1371/journal.pone.0166104>

Kaehler, B. D., Bokulich, N. A., McDonald, D., Knight, R., Caporaso, J. G., & Huttley, G. A. (2019). Species abundance information improves sequence taxonomy classification accuracy. *Nature Communications*, 10(1), 4643. <https://doi.org/10.1038/s41467-019-12669-6>

Kers, J. G., & Saccenti, E. (2022). The Power of Microbiome Studies: Some Considerations on Which Alpha and Beta Metrics to Use and How to Report Results. *Frontiers in Microbiology*, 12. <https://doi.org/10.3389/fmicb.2021.796025>

Kotera, M., Moriya, Y., Tokimatsu, T., Kanehisa, M., & Goto, S. (2015). KEGG and GenomeNet, New Developments, Metagenomic Analysis. En K. E. Nelson (Ed.), *Encyclopedia of Metagenomics: Genes, Genomes and Metagenomes: Basics, Methods, Databases and Tools* (pp. 329-339). Springer US. [https://doi.org/10.1007/978-1-4899-7478-5\\_694](https://doi.org/10.1007/978-1-4899-7478-5_694)

Lafaurie, G. I., Contreras, A., Barón, A., Botero, J., Mayorga-Fayad, I., Jaramillo, A., Giraldo, A., González, F., Mantilla, S., Botero, A., Archila, L. H., Díaz, A., Chacón, T., Castillo, D. M., Betancourt, M., Del Rosario Aya, M., & Arce, R. (2007). Demographic, clinical, and microbial aspects of chronic and aggressive periodontitis in Colombia: A multicenter study. *Journal of Periodontology*, 78(4), 629-639. <https://doi.org/10.1902/jop.2007.060187>

Lafaurie, G. I., Neuta, Y., Ríos, R., Pacheco-Montealegre, M., Pianeta, R., Castillo, D. M., Herrera, D., Reyes, J., Diaz, L., Castillo, Y., Sanz, M., & Iniesta, M. (2022). Differences in the subgingival microbiome according to stage of periodontitis: A comparison of two geographic regions. *PLOS ONE*, 17(8), e0273523. <https://doi.org/10.1371/journal.pone.0273523>

- Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Clemente, J. C., Burkepille, D. E., Vega Thurber, R. L., Knight, R., Beiko, R. G., & Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, 31(9), 814-821. <https://doi.org/10.1038/nbt.2676>
- Lee, Y.-H., Park, H. J., Jeong, S.-J., Auh, Q.-S., Jung, J., Lee, G.-J., Shin, S., & Hong, J.-Y. (2024). Oral microbiome profiles of gingivitis and periodontitis by next-generation sequencing among a group of hospital patients in Korea: A cross-sectional study. *Journal of Oral Biosciences*, 100591. <https://doi.org/10.1016/j.job.2024.100591>
- Li, X., Zeng, Z., Cheng, Z., Wang, Y., Yuan, L.-J., Zhai, Z., & Gong, W. (2023). Common pathogenic bacteria-induced reprogramming of the host proteinogenic amino acids metabolism. *Amino Acids*, 55(11), 1487-1499. <https://doi.org/10.1007/s00726-023-03334-w>
- Louca, S., Parfrey, L. W., & Doebeli, M. (2016). Decoupling function and taxonomy in the global ocean microbiome. *Science (New York, N.Y.)*, 353(6305), 1272-1277. <https://doi.org/10.1126/science.aaf4507>
- Lu, Y., Zhou, G., Ewald, J., Pang, Z., Shiri, T., & Xia, J. (2023). MicrobiomeAnalyst 2.0: Comprehensive statistical, functional and integrative analysis of microbiome data. *Nucleic Acids Research*, 51(W1), W310-W318. <https://doi.org/10.1093/nar/gkad407>
- Marizzoni, M., Gurry, T., Provasi, S., Greub, G., Lopizzo, N., Ribaldi, F., Festari, C., Mazzelli, M., Mombelli, E., Salvatore, M., Mirabelli, P., Franzese, M., Soricelli, A., Frisoni, G. B., & Cattaneo, A. (2020). Comparison of Bioinformatics Pipelines and Operating Systems for the Analyses of 16S rRNA Gene Amplicon Sequences in Human Fecal Samples. *Frontiers in Microbiology*, 11. <https://www.frontiersin.org/articles/10.3389/fmicb.2020.01262>
- Mayorga-Fayad, I., Lafaurie, G. I., Contreras, A., Castillo, D. M., Barón, A., & Aya, M. del R. (2007). Microflora subgingival en periodontitis crónica y agresiva en Bogotá, Colombia: Un acercamiento epidemiológico. *Biomédica*, 27(1), 21-33.
- Matchado, M. S., Rühlemann, M., Reitmeier, S., Kacprowski, T., Frost, F., Haller, D., Baumbach, J., & List, M. (2024). On the limits of 16S rRNA gene-based metagenome prediction and functional profiling. *Microbial Genomics*, 10(2), 001203. <https://doi.org/10.1099/mgen.0.001203>
- Moradi, J., Berggreen, E., Bunæs, D. F., Bolstad, A. I., & Bertelsen, R. J. (2025). Microbiome composition and metabolic pathways in shallow and deep periodontal pockets. *Scientific Reports*, 15(1), 12926. <https://doi.org/10.1038/s41598-025-97531-0>

Narayan, N. R., Weinmaier, T., Laserna-Mendieta, E. J., Claesson, M. J., Shanahan, F., Dabbagh, K., Iwai, S., & DeSantis, T. Z. (2020). Piphillin predicts metagenomic composition and dynamics from DADA2-corrected 16S rDNA sequences. *BMC Genomics*, 21(1), 56. <https://doi.org/10.1186/s12864-019-6427-1>

Narayanan, A., Söder, B., Meurman, J., Lundmark, A., Hu, Y. O. O., Neogi, U., & Yucel-Lindberg, T. (2023). Composition of subgingival microbiota associated with periodontitis and diagnosis of malignancy—A cross-sectional study. *Frontiers in Microbiology*, 14. <https://doi.org/10.3389/fmicb.2023.1172340>

The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet. (2007). National Academies Press. <https://doi.org/10.17226/11902>

Niu, S.-Y., Yang, J., McDermaid, A., Zhao, J., Kang, Y., & Ma, Q. (2017). Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes. *Briefings in Bioinformatics*. <https://doi.org/10.1093/bib/bbx051>

Parks, D. H., Tyson, G. W., Hugenholtz, P., & Beiko, R. G. (2014). STAMP: Statistical analysis of taxonomic and functional profiles. *Bioinformatics (Oxford, England)*, 30(21), 3123-3124. <https://doi.org/10.1093/bioinformatics/btu494>

Pérez-Chaparro, P. J., Gonçalves, C., Figueiredo, L. C., Faveri, M., Lobão, E., Tamashiro, N., Duarte, P., & Feres, M. (2014). Newly identified pathogens associated with periodontitis: A systematic review. *Journal of Dental Research*, 93(9), 846-858. <https://doi.org/10.1177/0022034514542468>

Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A. H., Nieuwdorp, M., & Levin, E. (2020). Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLOS ONE*, 15(1), e0227434. <https://doi.org/10.1371/journal.pone.0227434>

Qian, X.-B., Chen, T., Xu, Y.-P., Chen, L., Sun, F.-X., Lu, M.-P., & Liu, Y.-X. (2020). A guide to human microbiome research: Study design, sample collection, and bioinformatics analysis. *Chinese Medical Journal*, 133(15), 1844-1855. <https://doi.org/10.1097/CM9.0000000000000871>

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, 41(Database issue), D590-D596. <https://doi.org/10.1093/nar/gks1219>

Regueira-Iglesias, A., Balsa-Castro, C., Blanco-Pintos, T., & Tomás, I. (2023). Critical review of 16S rRNA gene sequencing workflow in microbiome studies: From primer selection to advanced data analysis. *Molecular Oral Microbiology*, *38*(5), 347-399. <https://doi.org/10.1111/omi.12434>

Reitmeier, S., Hitch, T. C. A., Treichel, N., Fikas, N., Hausmann, B., Ramer-Tait, A. E., Neuhaus, K., Berry, D., Haller, D., Lagkouvardos, I., & Clavel, T. (2021). Handling of spurious sequences affects the outcome of high-throughput 16S rRNA gene amplicon profiling. *ISME Communications*, *1*(1), 1-12. <https://doi.org/10.1038/s43705-021-00033-z>

Sierra, M. A., Li, Q., Pushalkar, S., Paul, B., Sandoval, T. A., Kamer, A. R., Corby, P., Guo, Y., Ruff, R. R., Alekseyenko, A. V., Li, X., & Saxena, D. (2020). The Influences of Bioinformatics Tools and Reference Databases in Analyzing the Human Oral Microbial Community. *Genes*, *11*(8), Article 8. <https://doi.org/10.3390/genes11080878>

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Horn, D. J. V., & Weber, C. F. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*, *75*(23), 7537-7541. <https://doi.org/10.1128/AEM.01541-09>

Schloss, P. D. (2020). Reintroducing mothur: 10 Years Later. *Applied and Environmental Microbiology*, *86*(2), e02343-19. <https://doi.org/10.1128/AEM.02343-19>

Tanca, A., Palomba, A., Fraumene, C., Pagnozzi, D., Manghina, V., Deligios, M., Muth, T., Rapp, E., Martens, L., Addis, M. F., & Uzzau, S. (2016). The impact of sequence database choice on metaproteomic results in gut microbiota studies. *Microbiome*, *4*(1), 51. <https://doi.org/10.1186/s40168-016-0196-8>

Tonetti, M. S., Greenwell, H., & Kornman, K. S. (2018). Staging and grading of periodontitis: Framework and proposal of a new classification and case definition. *Journal of Clinical Periodontology*, *45*(S20), S149-S161. <https://doi.org/10.1111/jcpe.12945>

Wang, J., Qi, J., Zhao, H., He, S., Zhang, Y., Wei, S., & Zhao, F. (2013). Metagenomic sequencing reveals microbiota and its functional potential associated with periodontal disease. *Scientific Reports*, *3*, 1843. <https://doi.org/10.1038/srep01843>

Wemheuer, F., Taylor, J. A., Daniel, R., Johnston, E., Meinicke, P., Thomas, T., & Wemheuer, B. (2020). Tax4Fun2: Prediction of habitat-specific functional profiles and functional redundancy

based on 16S rRNA gene sequences. *Environmental Microbiome*, 15(1), 11. <https://doi.org/10.1186/s40793-020-00358-7>

Wright, R. J., & Langille, M. G. I. (2025). PICRUSt2-MPGA: An update to the reference database used for functional prediction within PICRUSt2 (p. 2025.01.27.635123). *bioRxiv*. <https://doi.org/10.1101/2025.01.27.635123>

Ye, S. H., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell*, 178(4), 779-794. <https://doi.org/10.1016/j.cell.2019.07.010>

Zhou, Q., Su, X., & Ning, K. (2014). Assessment of quality control approaches for metagenomic data analysis. *Scientific Reports*, 4. <https://doi.org/10.1038/srep06957>