UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería

Implementación de análisis metagenómico utilizando hipertensión como caso de estudio

Trabajo de graduación presentado por Jennifer Daniela Sandoval Rivas para optar al grado académico de Licenciada en Ingeniería en Bioinformática

Guatemala,

2022

UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería

Implementación de análisis metagenómico utilizando hipertensión como caso de estudio

Trabajo de graduación presentado por Jennifer Daniela Sandoval Rivas para optar al grado académico de Licenciada en Ingeniería en Bioinformática

Guatemala,

2022

Vo.Bo.:

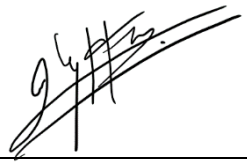
(f) 

MSc. Luis Augusto Franco

Tribunal Examinador:

(f) 

Msc. Luis Augusto Franco

(f) 

Msc. Jorge Chang

(f) 

MSc. Douglas Barrios

Fecha de aprobación: Guatemala, 08 de diciembre de 2022.

Prefacio

El presente trabajo surge a partir de mi interés por el aprendizaje, compartir mis conocimientos, mi deseo de aplicar mis conocimientos y retarme a implementar metodologías y herramientas nuevas que son útiles en el ámbito de la bioinformática. Mi principal objetivo era presentar un tema de interés para Guatemala e incentivar al uso de la bioinformática como método de estudio en temas relevantes en la salud.

Este proyecto representa la culminación de una etapa y un sueño que comenzó hace 5 años. Agradezco primeramente a Dios por permitirme cumplir mis objetivos, por brindarme sabiduría y todo lo que me fue necesario a lo largo de este proceso. Reconozco que todo mi esfuerzo no habría valido la pena si Él no me hubiera acompañado.

También agradezco a la Universidad del Valle de Guatemala por darme las herramientas necesarias para desarrollar mi potencial y por todo el apoyo brindado a lo largo de mi carrera. Así mismo, agradezco a todos los catedráticos que formaron parte de mi proceso de formación universitaria.

Especialmente agradezco a mi asesor Msc. Augusto Franco, quien me apoyó no únicamente a lo largo de este proyecto sino a lo largo de mi carrera. Agradezco su apoyo, guía, disposición y motivación. Es una fuente de inspiración como persona y como profesional.

En lo personal, agradezco a mi familia por todo su apoyo y ánimo a lo largo de toda mi trayectoria académica. Principalmente agradezco a mi mamá Rosa Nely Rivas. Sin ella nada de esto hubiera sido posible. Le agradezco por confiar en mi capacidad y apoyarme siempre en todas las áreas de mi vida. También agradezco a mis hermanas María José Sandoval y Sandy Sandoval por su apoyo incondicional.

Finalmente, agradezco a todos mis amigos y compañeros de estudio que han sido parte de este proceso y que me han apoyado y animado a seguir adelante.

Índice

Lista de figuras

Lista de cuadros

Resumen

En el siguiente documento se presenta la información necesaria para llevar a cabo la implementación de un análisis metagenómico utilizando datos obtenidos de muestras del microbioma del intestino humano con el propósito de conocer si es posible determinar asociación entre el microbioma y la hipertensión. El objetivo principal del análisis metagenómico fue obtener microorganismos o familias de genes que se encuentran correlacionados positiva o negativamente a la enfermedad, así como obtener las familias de proteínas que se encuentran en mayor abundancia para ambos tipos de muestras.

La importancia de este estudio radica principalmente en que la hipertensión es una enfermedad que afecta a una gran proporción de la población no solo en Guatemala si no a nivel mundial, por lo que conocer a mayor profundidad los factores de riesgo de padecer esta enfermedad es de relevancia para su continuo estudio, supervisión y desarrollo de tratamientos con mayor efectividad.

Las herramientas bioinformáticas permiten realizar un análisis de conjuntos extensos de datos como lo son los datos obtenidos de secuenciación de metagenómica. Además, el uso de modelos de *machine learning* mejora el proceso de predicción en la asociación de microbiomas y fenotipos presentes en el húesped. Por lo que en este documento se describe la metodología que se utilizó para el análisis de los datos de metagenoma del microbioma intestinal, las herramientas utilizadas en el flujo de trabajo del análisis de datos y cómo posteriormente se utilizó un modelo de machine learning para describir su asociación a la hipertensión. Así mismo, se describen hallazgos relevantes en cuanto a los resultados de las familias de proteínas y organismos más abundantes.

Abstract

The following document presents the necessary information to carry out the implementation of a metagenomic analysis using data obtained from samples of the human intestinal microbiome in order to know if it is possible to determine an association between the microbiome and hypertension. The main objective of the metagenomic analysis was to obtain microorganisms or families of genes that are positively or negatively correlated to the disease, as well as to obtain the families of proteins that are found in greater abundance for both types of samples.

The importance of this study lies mainly in the fact that hypertension is a disease that affects a large proportion of the population not only in Guatemala but worldwide. Therefore, it is important to study in depth the risk factors that are related to hypertension in order to monitor the disease, its development and the study of more effective treatments.

Bioinformatic tools allow analysis of large data sets such as data obtained from metagenomic sequencing. In addition, the use of machine learning models improves the prediction process in the association of microbiomes and phenotypes present in the host. Therefore, this document describes the methodology used for the analysis of the gut microbiome metagenome data, the tools used in the data analysis workflow and how a machine learning model was subsequently used to describe the association with hypertension. Likewise, relevant findings are described regarding the results of the most abundant families of proteins and organisms.

CAPÍTULO 1

Introducción

En el presente informe se detalla el desarrollo de la implementación de una *pipeline* para el análisis metagenómico. En esta *Pipeline* se utilizan datos de hipertensión como caso de estudio. El próposito de este estudio es implementar la *pipeline* para caracterizar microorganismos o proteínas que esten asociados a hipertensión.

Para analizar la relación existente entre los microorganismos en la Microbiota intestinal y la hipertensión se hace uso de herramientas bioinformáticas que especificamente hacen uso de datos metagenómicos. Los datos metagenómicos contienen un conjunto de secuencias que son obtenidos a partir de muestras analizadas de la microbiota intestinal y estas muestras son comparadas con genomas de referencia conocidas de microorganismos con el objetivo de identificar qué organismos o grupos de proteínas, presentan una mayor abundancia en muestras de personas que tienen hipertensión.

Es importante reconocer la relación existente entre la microbiota intestinal y la hipertensión, ya que esta relación puede ser determinante en cuanto a la detección y supervisión de la enfermedad. A la misma vez se tiene un mayor conocimiento en cuanto a la hipertensión y los factores que influyen en su desarrollo.

Otro de los própositos de este estudio es el poder validar las herramientas utilizadas en la *pipeline* para el análisis metagenómico y a su vez modificar la metodología tradicional utilizando una herramienta de *Machine learning* para poder determinar las asociaciones entre los datos metagenómicos y la hipertensión.

CAPÍTULO 2

Antecedentes

Es importante reconocer la relevancia del análisis de genómica computacional, específicamente en metagenómica para identificar la asociación de genes del microbioma humano a enfermedades que afectan a un gran porcentaje de la población mundial. Por otro lado es importante analizar los estudios de metagenómica que ya han sido realizados debido a que esto permite innovar en la realización del proyecto y a la vez conocer las metodologías y herramientas que se utilizan actualmente en este ámbito.

2.1. Un estudio de asociación de todo el metagenoma de la microbiota intestinal en la diabetes tipo 2

Este fue el primer estudio publicado en el cual fue posible definir la metodología de análisis de datos de metagenoma para poder asociar las variaciones en el microbioma a la enfermedad de diabtes tipo 2. La diabetes de tipo 2 se considera como un trastorno metabólico que se caracteriza por hiperglucemia y resistencia a la insulina. A pesar de que se tenían identificados algunos genes que aumentaban el riesgo de padecer de esta enfermedad, estos solo representaban una pequeña porción de todos los factores que influían al riesgo de este padecimiento.

Por medio de este estudio se ha demostrado que el microbioma intestinal afecta la fisiología del huésped, se propusó que este fuera tomado en cuenta como un factor ambiental que contribuye al riesgo de desarrollar diabetes tipo 2 [1].

Un análisis MWAS, por sus siglas en inglés *Metagenome Wide Association*, hace referencia a un estudio en el cual se lleva a cabo la asociación de genomas completos. Siendo estudios de metagenomas, se analizan múltiples genomas a la vez. Este análisis permite la identificación de genes asociados a una enfermedad o rasgo en particular [2]. En este estudio se utilizó un análisis MWA el cual demostró que los pacientes con diabetes tipo 2 se caracterizaban por un grado moderado de disbiosis microbiana intestinal. La disbiosis hace

referencia a un desbalance en el equilibrio de la microbiota normal.

En este estudio se observó una disminución en la abundancia de algunas bacterias universales productoras de butirato y un aumento de varios patógenos oportunistas, así como un enriquecimiento de otras funciones microbianas que confieren reducción de sulfatos y resistencia al estrés oxidativo [1].

2.2. El estudio de asociación de todo el metagenoma del microbioma intestinal reveló una nueva etiología de la artritis reumatoide en la población japonesa

La causalidad y el mecanismo patogénico de la composición del microbioma siguen siendo poco estudiados y conocidos en muchas enfermedades, incluidas las enfermedades autoinmunes como la artritis reumatoide (ar). Este estudio tenía como objetivo dilucidar el papel del microbioma intestinal en la patología de la artritis reumatoide mediante un estudio integral de asociación de todo el metagenoma [3].

Las pruebas filogenéticas de asociación de casos y controles mostró una alta abundancia de múltiples especies pertenecientes al género *Prevotella* (por ejemplo, *Prevotella denticola*) en el metagenoma del caso ar. En este estudio utilizaron un modelo de machine learning no-lineal el cual les permitió controlar la discrepancia filogenética [3]. Como se puede observar en la Figura 1 se presentan los principales resultados del análisis de la asociación metagenómica de la artritis reumatoide.

2.3. Un estudio de asociación del metagenoma del microbioma intestinal en pacientes con esclerosis múltiple reveló una nueva patología de la enfermedad

En este experimento se llevó a cabo un estudio completo del metagenoma del microbioma intestinal para asociar su efecto en una población japonesa con casos de personas con esclerosis múltiple y casos control. El análisis se basó en tres pipelines de análisis bioinformático: análisis filogenético, análisis de genes funcionales y análisis de vías metabólicas.

Los resultados obtenidos contribuyeron a conocer más acerca de la enfermedad, dado que se observaron en las pruebas filogenéticas de asociación de casos y controles discrepancias de ocho clados, la mayoría de los cuales estaban relacionados con el sistema inmunitario por ejemplo, *Erysipelatoclostridium* sp. y *Gemella morbillorum*. Por otro lado, las pruebas de asociación de genes encontraron una mayor abundancia de un gen putativo de deshidrogenasa (Clo1100_2356) y un gen relacionado con el transportador ABC (Mahau_1952) en el metagenoma de los casos de esclerosis múltiple en comparación con los controles [4].

En este caso se puede resaltar la importancia de los estudios en metagenómica y cómo estos pueden llegar a proporcionar información relevante para una enfermedad y en algunos casos los microorganismos o genes de ciertos clados pueden ser considerados como biomar-

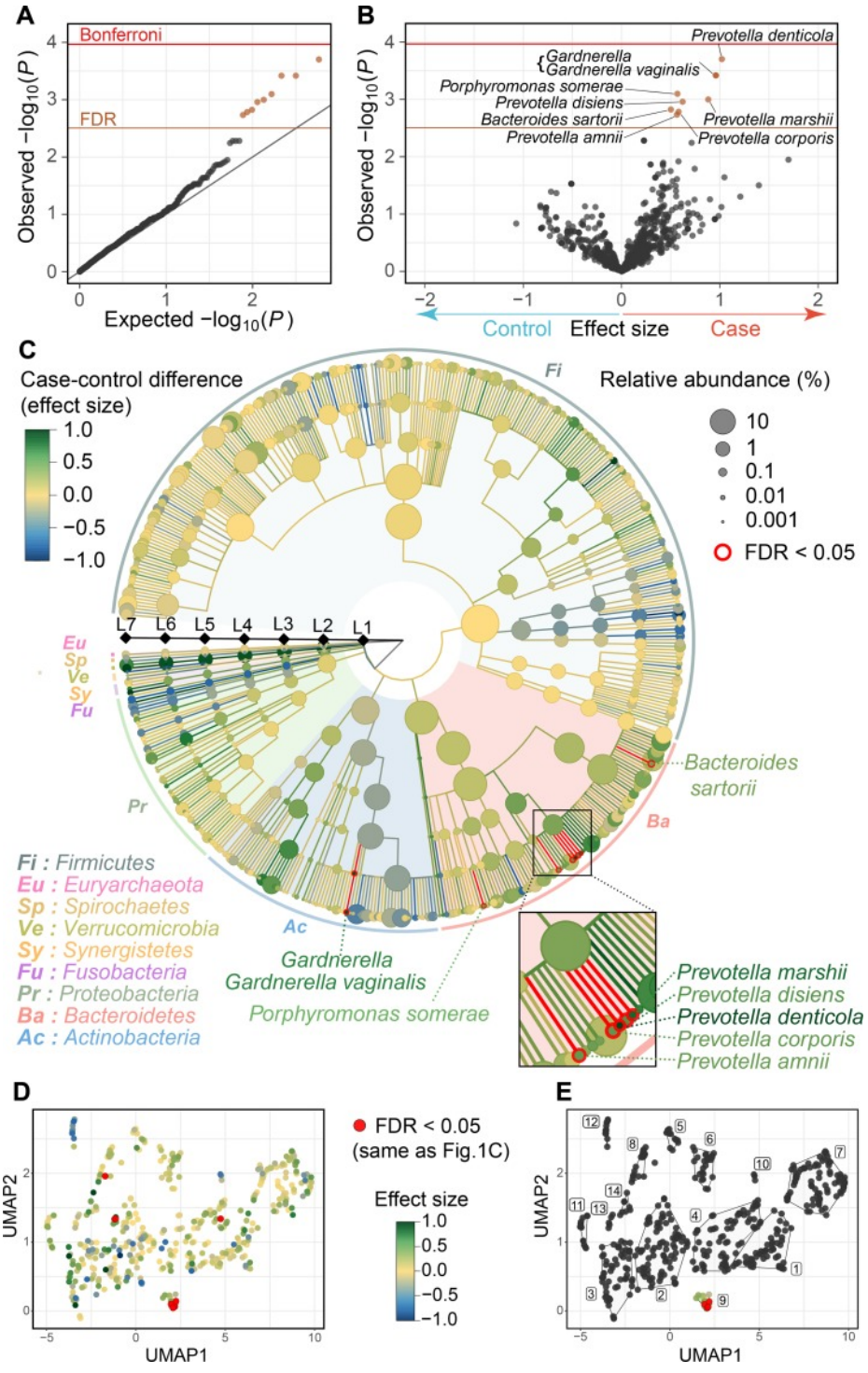


Figura 1: Resultados de MWAS de las pruebas de asociación filogenética de casos y controles de AR [3]

cadores para diagnosticar la enfermedad, ver su progreso o bien para evaluar la respuesta a tratamientos.

2.4. Los microbiomas orales e intestinales se alteran en la artritis reumatoide y se normalizan parcialmente después del tratamiento

En este estudio se llevó a cabo una secuenciación de escopeta metagenómica y un estudio de asociación de todo el metagenoma (MWAS) de muestras fecales, dentales y salivales de una muestra de personas con artritis reumatoide (AR) y controles sanos. Se detectó disbiosis en los microbiomas intestinales y orales de pacientes con AR, pero se resolvió parcialmente después del tratamiento para la AR. En particular, *Haemophilus* spp. se agotaron en individuos con AR en los tres sitios y se correlacionaron negativamente con los niveles de autoanticuerpos séricos, mientras que *Lactobacillus salivarius* estuvo sobrerrepresentado en individuos con AR en los tres sitios y estuvo presente en cantidades aumentadas en casos de AR muy activa [5].

Las alteraciones en el microbioma intestinal, dental o de la saliva distinguieron a las personas con AR de los controles sanos, se correlacionaron con medidas clínicas, por lo que se sugiere que podrían usarse para estratificar a las personas en función de su respuesta a la terapia [5].

2.5. El estudio de asociación del metagenoma reveló un panorama específico de la enfermedad lupus eritematoso sistémico

El objetivo de este estudio era investigar acerca de la relación entre los cambios en el microbioma intestinal y la enfermedad de lupus eritematoso sistémico, que anteriormente se había descrito que se encontraban relacionados, pero no se conocía a profundidad la relación. Por lo que se llevó a cabo un estudio de asociación de todo el genoma [6].

Los resultados indican un aumento de *Streptococcus intermedius* y *Streptococcus anginosus* en los pacientes con lupus eritematoso sistémico. El análisis de genes microbianos reveló aumentos de genes derivados de *Streptococcus*, incluido uno involucrado en la reacción redox [6].

2.6. Fuentes de referencia de datos biológicos

Para este tipo de análisis es importante utilizar bases de datos de referencia con las cuales se pueden comparar los datos estudiados, lo que permite identificar los diferentes microorganismos sobre los cuales ya se tienen los datos de su genoma descritos de forma oficial en estas bases de datos.

La base de datos utilizada generalmente para los estudios de metagenómica es:

- NCBI: Es una base de datos que cuenta con distintos tipos de datos, desde secuen-

ciaciones de genomas completos, secuencias de genes, proteínas, bio proyectos, entre otros. Es una fuente confiable de datos y cuenta con información para diversos organismos tanto eucariotas como procariotas. Además, cuenta con diferentes herramientas que pueden ser utilizadas en la metodología de procedimientos in silico. Enlace: https://www.ncbi.nlm.nih.gov/

CAPÍTULO 3

Justificación

Los ensayos clínicos requieren de grandes costos y en la industria de la salud estos ensayos son fundamentales en el proceso de confirmación de un hallazgo científico. La microbiología tradicional generalmente implica obtener un cultivo de microorganismos como un paso importante en cualquier estudio. Sin embargo, se estima que las técnicas estándar de cultivo de laboratorio brindan información sobre el 1 % o menos de la diversidad bacteriana en una muestra ambiental determinada [7]. Además, las condiciones nunca serán exactamente iguales a las condiciones en las cuales se encuentran en su estado natural.

Actualmente, una buena práctica que se aplica en este tipo de estudios es realizar análisis apoyandose en herramientas informáticas para que posteriormente se puedan desarrollar ensayos clínicos comprobando los resultados obtenidos en el análisis informático. Las herramientas tecnológicas actualmente nos permiten realizar análisis y estudios a un menor costo y proveen mayor información de lo que se podría obtener en ensayos clínicos. La cantidad de información de secuenciación genómica que se produce ha aumentado considerablemente. Anteriormente, el principal problema para las investigaciones radicaba la poca disponibilidad o acceso a los datos necesarios, sin embargo, esto ya no es un problema dado al avance de la tecnología y herramientas para poner a disposición los datos al público.

El creciente problema al cual nos enfrentamos es que, a pesar de la gran cantidad de datos que se generan, estos no están siendo utilizados como recursos y muy pocas veces son analizados e interpretados. Uno de los principales objetivos de la bioinformática es poder aplicar tecnologías para el análisis de datos biológicos, por lo que por medio de la bioinformática se puede atacar el problema del análisis de datos biológicos disponibles. La bioinformática tiene diferentes ramas dentro de las cuales se puede mencionar la rama de genómica computacional, la cual tiene gran relevancia actualmente por el hecho de que es una herramienta que permite analizar grandes tramos de información genética en cualquier organismo de estudio.

Al utilizar genómica computacional es posible analizar una gran cantidad de datos genéticos y a pesar de que no se provean resultados exactos estos resultados permiten tener un acercamiento más específico. Esto puede resultar útil en el desarrollo de ensayos clínicos

que se encuentren enfocados en el estudio de un biomarcador para el desarrollo de fármacos más efectivos o bien como indicadores de la enfermedad y su progreso. La metegenómica es una rama de estudio que permite estudiar microorganismos mediante el análisis de su ADN adquirido directamente de una muestra ambiental, sin necesidad de obtener un cultivo puro. Con esta tecnología se analiza en su conjunto el ADN de los microorganismos de una población.

La secuenciación y el análisis del ADN metagenómico total pueden proporcionar información sobre varios aspectos de la muestra, lo que permite caracterizar mejor la vida microbiana en un entorno determinado. No solo puede revelar la identidad de las especies presentes, sino que también puede proporcionar información sobre las actividades metabólicas y los roles funcionales de los microbios presentes en una población determinada [8].

Existen diversas aplicaciones para las cuales el análisis metagenómico puede llegar a ser útil. Una de las aplicaciones más relevantes de metagenómica es en la industria farmacéutica ya que estos análisis permiten conocer más acerca de los microorganismos con el objetivo de poder desarrollar fármacos más efectivos o en algunos casos para estudiar la relación entre microorganismos y el padecimiento de una enfermedad. Actualmente este tipo de asociaciones se ha vuelto más común y se han estudiado una diversidad de enfermedades como cáncer, artritis, lupus, entre otras.

De esta tendencia reciente en los estudios de metagenómica surgió el interés de poder estudiar la posible asociación entre los microorganismos de la microbiota intestinal y sus posibles efectos o influencia en el padeciemiento de hipertensión. Ya que el estudio de esta enfermedad en esta área ha sido poco explorada y esta es una enfermedad de relevancia tanto a nivel mundial como en la región de latinoamérica.

La hipertensión es una enfermedad crónica que afecta de un 30 a 40 % de la población de la región de América y es responsable en un 30 % de las muertes de esta región. Según la OPS la mayoría de personas que fallece por esta enfermedad se encuentra por debajo de los 70 años y por esta razón se considera como una muerte prematura y que puede ser prevenible. Además, se sabe que las personas con menos recursos socioeconómicos tienden a tener más riesgo de padecer enfermedades cardíacas y accidentes cerebrovasculares, y tienen menos acceso a la prevención o al tratamiento [9].

En Guatemala, según los datos tomados del 2012 al 2020 por el Ministerio de Salud Pública y Asistencia Social, se conoce que la enfermedad no infecciosa con mayor cantidad de muertes es la hipertensión como se puede observar en la tabla de resultados en la Figura 2 [10].

A pesar de que existen tratamientos actuales que actúan para regular la presión sanguínea, específicamente la presión alta, se ha vuelto más claro que las respuestas a los medicamentos antihipertensivos estén respaldadas por la diversidad genética. Por esta razón, algunos fármacos pueden llegar a ser efectivos para algunos pacientes pero, poco efectivos para otros.

Por otro lado, es importante recalcar que cuando un paciente se ve afectado por una enfermedad, esta enfermedad no se desarrolla en un ambiente aislado sino en un ambiente en el cual intervienen diversos factores como lo es el cuerpo humano y a pesar de que se conozcan posibles condiciones genéticas y físicas que contribuyan a la predisposición de adquirir una

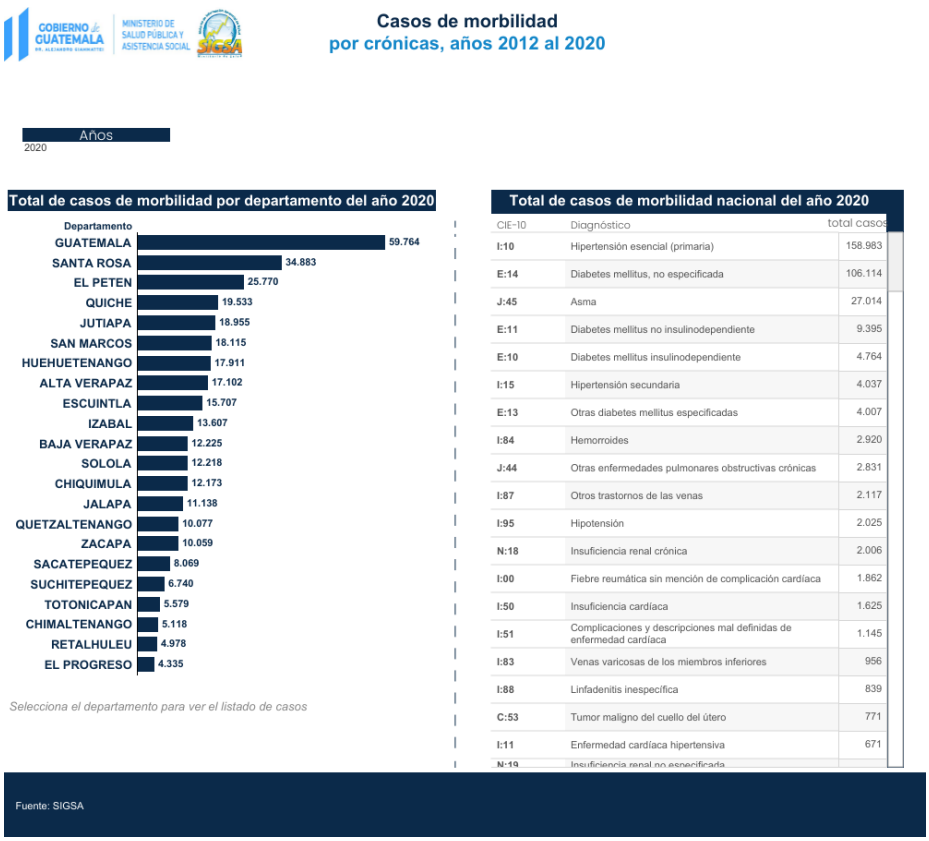


Figura 2: Casos de morbilidad por enfermedades crónicas del año 2012 al 2020

enfermedad, puede que estas condiciones no sean las únicas variables que intervengan.

El análisis de las variantes genéticas que puedan dar una referencia acerca de cómo hacer más eficientes los fármacos en condición de diversidad genética, es un gran aporte para el avance de la medicina personalizada y así mismo, de la mejoría en la efectividad de los medicamentos y una mejor calidad de vida para las personas que padecen esta enfermedad crónica.

CAPÍTULO 4

Objetivos

4.1. Objetivo general

Implementar un flujo de trabajo para analizar datos metagenómicos asociados a hipertensión.

4.2. Objetivos específicos

- Interpretar los resultados obtenidos del análisis metagenómico para definir si es posible determinar una asociación entre la microbiota intestinal y la hipertensión.
- Identificar los microorganismos y grupos de proteínas que presentan una mayor abundancia relativa con respecto a la hipertensión.
- Utilizar un modelo de *machine learning* para la inferencia estadística de asociaciones entre comunidades microbianas y fenotipos del huésped.

CAPÍTULO 5

Alcance

El objetivo principal del proyecto es poder implementar una metodología para el análisis metagenómico utilizando como caso de estudio la hipertensión. La metodología a seguir consta de una *pipeline* en la cual se utilizan diferentes herramientas para poder llevar a cabo el análisis completo de las muestras de metagenoma. Este análisis incluye el control de calidad de las muestras, ensamblaje de los metagenomas, anotación funcional, alineamiento y finalmente la inferencia de las asociaciones obtenidas a partir del uso de una herramienta de *machine learning*.

A partir de las asociaciones obtenidas el objetivo es analizar las proteínas con mayor abundancia encontrados en el análisis y comparar los grupos de estudio que en este caso constan de muestras de metagenoma correspondientes a pacientes saludables y pacientes con hipertensión.

El objetivo de implementar un análisis metagenómico es poder validar las diferentes herramientas utilizadas para el análisis con datos de estudio, en este caso la hipertensión, para que esta metodología pueda ser replicada para cualquier conjunto de datos de metagenómica. Específicamente, se tomó como caso de estudio la hipertensión dado que es una enfermedad relevante en el contexto mundial y mayormente en la región de latinoamérica. Por esta razón, se espera que los resultados obtenidos generen interés de investigación y análisis en beneficio de quienes padecen esta enfermedad, utilizando datos provenientes de la región de latinoamérica.

CAPÍTULO 6

Marco teórico

6.1. Microbiota y microbioma

La Microbiota humana puede definirse como el conjunto de apróximadamente 10-100 billones de células microbianas simbióticas que se encuentran albergadas por una persona y específicamente se considera Microbioma a los genes que albergan estas células. La microbiota humana se encuentra distribuida en una persona en diferentes áreas dentro de las cuales se puede mencionar la biota intestinal, cutánea, vaginal, oral, ocular y biliar [11]. Un aspecto importante a tomar en cuenta es que se sabe que el microbioma presente en los humanos sobrepasa la cantidad de genes que pueden ser encontrados en el genoma humano. La mayor parte de concentración de microorganismos en la microbiota humana se encuentra localizada en la microbiota intestinal y por esta razón este conjunto de organismos es el mayormente estudiado en cuanto a su relación con diversos padecimientos.

Como se mencionó anteriormente, los microorganismos que conforman la microbiota tienen una relación simbiótica con los humanos, esto quiere decir que ambos organismos se benefician mutuamente. Los microorganismos obtienen los nutrientes necesarios de los humanos para sobrevivir y reproducirse, por otro lado, los humanos se benefician de estos organismos para el correcto funcionamiento de algunos órganos o procesos y podría asegurarse que ciertas necesidades fisiológicas de los humanos han sido influidas por las microbiotas que han prevalecido durante su adaptación y evolución [12].

En las últimas décadas, la microbiota intestinal ha sido estudiada con mayor intensidad, dado que se ha demostrado que además de afectar de forma positiva en su relación simbiótica con los humanos, también se ha visto que es un factor influyente en varias enfermedades [12]. Actualmente, se conoce que la microbiota intestinal se compone en un 90 % de dos divisiones principales, Firmicutes (grampositivos) y Bacteroidetes (gramnegativos) como se puede observar en el Cuadro 1.

La clasificación e identificación de las distintas especies que conforman la microbiota intestinal se realiza por medio del uso del gen 16S ribosomal, este es un componente de la

Phylum	Género más representativo
Firmicutes	*Ruminococcus*
Enterococcus	
Clostridium	
Peptostreptococcus	
Lactobacillus	
Bacteroidetes	*Bacteroides*
Proteobacterias	*Desulfovibrio*
Escherichia	
Helicobacter	
Actinobacteria	*Bifidobacterium*
Actinomyces	
Verrucomicrobia	*Verrucomicrobium*

Cuadro 1: Principales divisiones o phyla de la microbiota del tracto digestivo humano [12].

subunidad 30s de los ribosomas en los procariotas. El gen 16S rRNA se caracteriza por sus propiedades evolutivas, que le permiten convertirse en un importante marcador molecular en la ecología microbiana. Dado que el gen 16S rRNA se conserva en las bacterias y contiene regiones hipervariables que pueden proporcionar secuencias características específicas de la especie, la secuenciación 16S rRNA se usa ampliamente en la identificación de bacterias y estudios filogenéticos [13].

En su mayor parte, las asociaciones entre la microbiota y las enfermedades humanas se han estudiado utilizando la secuenciación del amplicón del gen 16S rRNA. Estos estudios han sugerido que la disbiosis, que hace referencia a un desequilibrio en la microbiota normal, ya sea en cambios cuantitativos, cualitativos, composición o cambios en su funcionamiento, puede ser un factor de riesgo ambiental clave para muchas enfermedades humanas, aunque la gravedad de la disbiosis varía según la enfermedad [14].

La secuenciación de escopeta metagenómica, en la que se secuencia el genoma completo de los microorganismos que componen el microbioma en lugar de un solo gen marcador taxonómico, puede superar las limitaciones de un único marcador genético, como el 16s, al proporcionar información sobre la abundancia de genes en vías funcionales y en todos los niveles taxonómicos [14].

6.2. Secuenciación de escopeta

El método de escopeta de genoma completo implica secuenciar muchos fragmentos de ADN superpuestos en paralelo y luego usar una computadora para ensamblar los fragmentos pequeños en contigs más grandes y, finalmente, en cromosomas. La secuenciación de escopeta permite evaluar todos los genes de todos los organismos presentes en una muestra compleja. Esta técnica es de mucha utilidad en la microbiología, dado que permiten evaluar diversidad y abundancia microbiana en diversos ambientes. Además, permite estudiar microorganismos incultivables, que de otra forma sería casi imposible de estudiar [15].

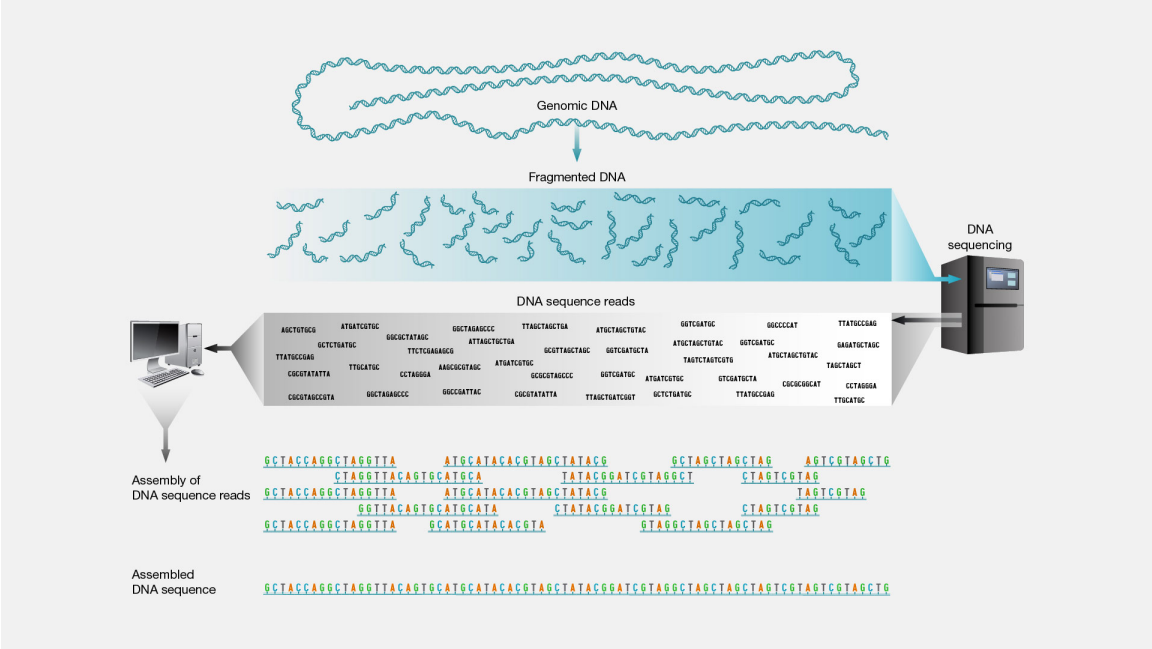


Figura 3: Proceso de secuenciación de escopeta [16]

.

6.3. Perfilación taxonómica

La taxonomía se considera como un sistema de categorías y relaciones que presenta un orden jerárquico. En el ámbito biológico la taxonomía se considera como la organización de organismos en un sistema de clasificación compuesto por una jerarquía de taxones. Esta clasificación se encuentra bastante relacionada al orígen evolutivo de los organismos. Existen ocho niveles de clasificación que van de lo más general a lo más específico siendo el más específico el nivel que abarca un único tipo de organismo. Los niveles son: dominio, reino, filo, clase, orden, familia, género y especie. En la Figura 4 se puede observar la jerarquía de estos niveles taxonómicos [17].

El perfil taxonómico permite tener una idea de la composición taxonómica de cada muestra analizada, es decir, permite conocer qué organismos se encuentran en una muestra. El reconocimiento de estos organismos puede ser en cualquiera de los niveles taxonómicos. En la construcción de perfiles taxonómicos aparte de la identificación de taxones presentes en una muestra también se estiman las abundancias relativas de estos organismos. La abundancia relativa se refiere al cálculo de la proporción de lecturas secuenciadas que pertenecen a un mismo organismo, es decir, la proporción representativa de este organismo en el metagenoma. Como resultado, el perfil taxonómico contiene una lista de taxones detectados y sus abundancias relativas estimadas [18].

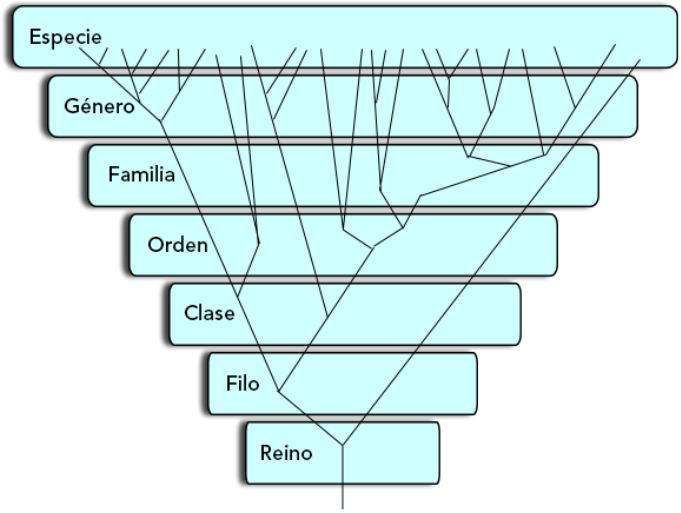


Figura 4: Representación de jerarquía taxonómica

6.4. Genómica computacional

La genómica computacional es una rama de la bioinformática que se encarga del análisis computacional y estadístico de las secuencias del genoma de un organismo. Como su nombre lo dice esta rama se enfoca principalmente en el genoma el cual se considera como el conjunto de instrucciones genéticas o genes que se encuentran en una célula [19]. En esta rama se utiliza la secuenciación del ADN y herramientas de informática para secuenciar, ensamblar y analizar la estructura y funcionamiento del genoma. El objetivo principal de la genómica es relacionar la secuencia de ADN de los genes con su función, así como relaciones e interacciones que puedan verse relacionadas a enfermedades o diferencias entre genomas [20].

Existen algunos pasos generales que seguir en cuanto a el análisis metagenómico de los cuales se pueden mencionar:

- **Control de calidad:** El proceso de control de calidad en análisis bioinformático se conoce como la etapa en la cual se evaúan los sets de datos que serán utilizados para poder descartar cualquier tipo de contaminación o de error en cuanto al proceso previo de secuenciación. Generalmente este es el primer paso para llevar a cabo cualquier tipo de análisis bioinformático en el cual se utilicen secuencias como datos de entrada.
- **Ensamblaje de genomas:** El ensamblaje del genoma se refiere al proceso de tomar una gran cantidad de secuencias cortas de ADN y volver a unirlas para crear una representación de los cromosomas originales a partir de los cuales se originó el ADN.

  En un proyecto de secuenciación del genoma, el ADN del organismo objetivo se divide en millones de pequeños fragmentos y se lee en una máquina de secuenciación. Estas

"lecturas"varían de 20 a 1000 pares de bases de nucleótidos (pb) de longitud según el método de secuenciación utilizado. Por lo general, para la secuenciación de lectura corta de tipo Illumina, se producen lecturas de una longitud de 36 a 150 pb. Estas lecturas pueden ser de “extremo único” como se describe anteriormente o “extremo emparejado” [21].

Las lecturas de extremos emparejados se producen cuando el tamaño del fragmento utilizado en el proceso de secuenciación es mucho mayor (normalmente de 250 a 500 pb de longitud) y los extremos del fragmento se leen hacia el medio. Esto produce dos lecturas "pareadas". Uno del extremo izquierdo de un fragmento y otro del derecho con una distancia de separación conocida entre ellos. Esta información adicional contenida en las lecturas finales emparejadas puede ser útil para ayudar a unir piezas de secuencia durante el proceso del ensamblaje [21].

6.5. Metagenómica

Metagenómica se encuentra compuesta de dos palabras, "genómica"la cual es una rama de bioinformática que tiene como principal objetivo obtener las secuencias de ADN de los organismos estudiados, "meta"por otro lado, hace referencia a que se están obteniendo las secuencias de ADN de varios organismos al mismo tiempo [22].

La metagenómica se define como una rama de las ciencias que se encarga de estudiar la estructura y función de todas las secuencias de nucleótidos aisladas y analizadas de todos los organismos en una muestra grande [22].

Esta rama es importante en el estudio de poblaciones que no pueden ser estudiadas de forma individual, como lo son las poblaciones de microorganismos. Estas poblaciones conviven en un mismo ecosistema, ya pueda ser terrestre, acuático o dentro de otro organismo que los alberga, y estos se encuentran fuertemente relacionados entre sí y con su ecosistema, por lo que su estudio en conjunto tiene una mayor utilidad porque se toman en cuenta las interacciones entre todos los microorganismos y factores que conforman el ecosistema.

Estudios de asociación amplia del metagenoma

La asociación amplia del metagenoma o por sus siglas en inglés MWAS (*Metagenome Wide Association Studies*) tiene como objetivo la identificación de variantes genéticas en el metagenoma de una población humana que se encuentran asociadas con un fenotipo que generalmente se describe como una enfermedad. En la metodología actual de estos estudios la abundancia relativa de un gen en el metagenoma se utiliza para establecer su asociación con una enfermedad de interés [14].

6.6. *Machine learning:* aplicaciones en metagenómica

Machine learning se refiere a un tipo de inteligencia artificial que por medio de modelos le permite aprender de los datos que se le proporcionan, es decir, tiene un aprendizaje au-

tomático. En este tipo de modelos la cantidad de datos es importante, ya que los modelos mejoran su proceso de aprendizaje al alimentarlos con conjuntos grandes de datos. El resultado principal de este tipo de algoritmos se describe como el modelo que ha sido entrenado con una cantidad significativa de datos y que por medio de este aprendizaje es capaz de proporcionar una predicción o prónostico sobre nuevos datos que sean ingresados [23].

Las técnicas de *machine learning* tienen diversas aplicaciones debido a la flexibilidad de su composición, sin embargo, existen algunas técnicas que han demostrado proporcionar mejores resultados para el procesamiento de cierto tipo de datos. Además, existen diferentes agrupaciones de algoritmos en *machine learning* que pueden clasificarse en modelos de aprendizaje supervisados, semisupervisados, no supervisados e incluso aprendizaje reforzado[24].

Actualmente, la inteligencia artificial ha demostrado un gran avance en cuanto a la diversidad de aplicaciones que se pueden desarrollar utilizando diferentes técnicas. La metegenómica es un área de las ciencias en el cual se está explorando el uso de estos algoritmos. Como se mencionó anteriormente, en los estudios de metagenómica se genera una gran cantidad de datos, en este caso de secuenciación, los cuales dan paso a la posibilidad de utilizar modelos de predicción de *machine learning* por la facilidad de utilizar grandes conjuntos de datos que permitan mejorar el proceso de aprendizaje de los modelos generados.

La ecología microbiana se ha basado durante mucho tiempo en análisis estadísticos tradicionales para resumir datos, probar hipótesis e interpretar interacciones entre características y respuestas en conjuntos de datos microbianos. Sin embargo, las aplicaciones de *machine learning* tienen cierta ventaja sobre los modelos estadísticos comunes. La principal diferencia es que los modelos estadísticos se basan en describir e inferir las relaciones entre variables, mientras que los modelos de *machine learning* están diseñados para optimizar la capacidad de predecir un resultado en un conjunto de datos externo [24].

6.7. Hipertensión

6.7.1. ¿Qué es la hipertensión?

La hipertensión también conocida como presión alta es una afección en la que la fuerza que ejerce la sangre contra las paredes de las arterias es lo suficientemente alta como para poder causar problemas cardíacos. La presión arterial está determinada por dos factores: cantidad de sangre bombeada por el corazón y el grado de resistencia de flujo de sangre en las arterias. Lo que provoca una presión alta es la consecuencia del bombeo excesivo de sangre y arterias estrechas que impiden el flujo normal de sangre [25]. Existen dos tipos de presión alta:

- **Hipertensión primaria (esencial)**: Este tipo de presión alta no tiene una causa principal, pero se sabe que se da principalmente en adultos y que esta se va desarrollando con el paso de los años debido a que los vasos sanguíneos se vuelven más rígidos. Así mismo, hay factores que contribuyen a desarrollar esta enfermedad como lo es la cantidad de agua o sal presente en el cuerpo, niveles hormonales y el estado de los riñones, sistema cardiovascular o nervioso.

- **Hipertensión secundaria**: Este tipo de presión alta se da en consecuencia de la presencia de otra enfermedad y en muchas ocasiones tiende a ser más riesgosa que la hipertensión primaria. Algunas causas de esta hipertensión pueden ser problemas de tiroides, enfermedad renal, consumo de medicamentos, entre otras. [25]

Algo que es importante mencionar de esta enfermedad es que es considerada como una enfermedad crónica, es decir, que no tiene una cura absoluta, sino que únicamente puede ser regulada y tratada para mejorar la calidad de vida de la persona afectada.

CAPÍTULO 7

Materiales y métodos

7.1. Materiales

Este proyecto fue principalmente desarrollado haciendo uso de herramientas bioinformáticas, por lo que los principales materiales que se utilizaron se pueden considerar como las muestras de estudio que se listan en el Cuadro 2, las cuales fueron tomadas de un estudio previamente investigado y verificado. También se hace mención de las principales herramientas o softwares utilizados en el Cuadro 3, de estos se hace mención en el diagrama de flujo de trabajo. Finalmente se enlista en el Cuadro 4 un listado de las especificaciones del equipo en el cual se llevó a cabo el análisis. Esto con el fin de que la implementación del análisis pueda ser replicado y a su vez se tome en consideración la capacidad del equipo respecto a las diferentes herramientas utilizadas.

7.2. Métodos

7.2.1. Selección de muestras

Se consultaron distintas bases de datos en donde se encontró una colección de muestras de metagenomas del intestino humano de personas que padecen de hipertensión así como de personas saludables para poder realizar el análisis comparativo de ambos casos. En este paso fue importante investigar colecciones que contaran con una cantidad de muestras estadísticamente significativa y que fueran de orígen confiable, además el estudio debía tener la aprobación de un comité de ética.

Finalmente se seleccionaron las muestras correspondientes al estudio *Gut microbiota dysbiosis contributes to the development of hypertension* [26]. Este estudio tenía como objetivo

ID	Clasificación
ERR1398068	Hipertensión
ERR1398168	
ERR1398221	
ERR1398076	
ERR1398077	
ERR1398085	
ERR1398129	Control saludable
ERR1398206	
ERR1398263	
ERR1398078	
ERR1398257	
ERR1398089	

Cuadro 2: Listado de muestras de metagenoma utilizadas para la implementación de la pipeline

Herramienta	Uso
Trim Galore	Control de calidad y filtro
Megahit	Ensamblaje de metagenoma
Superfocus	Anotación funcional
Kraken	Clasificador de secuencias taxonómicas
Krona	Visualización metaegenómica interactiva
STAMP	Visualización de abundancias relativas e inferencia estadística
Seaborn	Librería para visualización de datos
SIAMCAT	Uso de modelos de *machine learning*

Cuadro 3: Listado de softwares utilizados para el procesamiento y análisis de datos metagenómicos

Especificaciones equipo de cómputo	
Procesador	Intel(R) Core(TM) i7-8565U CPU @ 1.80GHz 1.99 GHz
Memoria RAM	24GB
Sistema operativo	Ubuntu, subsistema de Linux para Windows 11
Memoria de almacenamiento	500 GB libres

Cuadro 4: Especificaciones del equipo de cómputo en el cual se implementó el análisis metagenómico

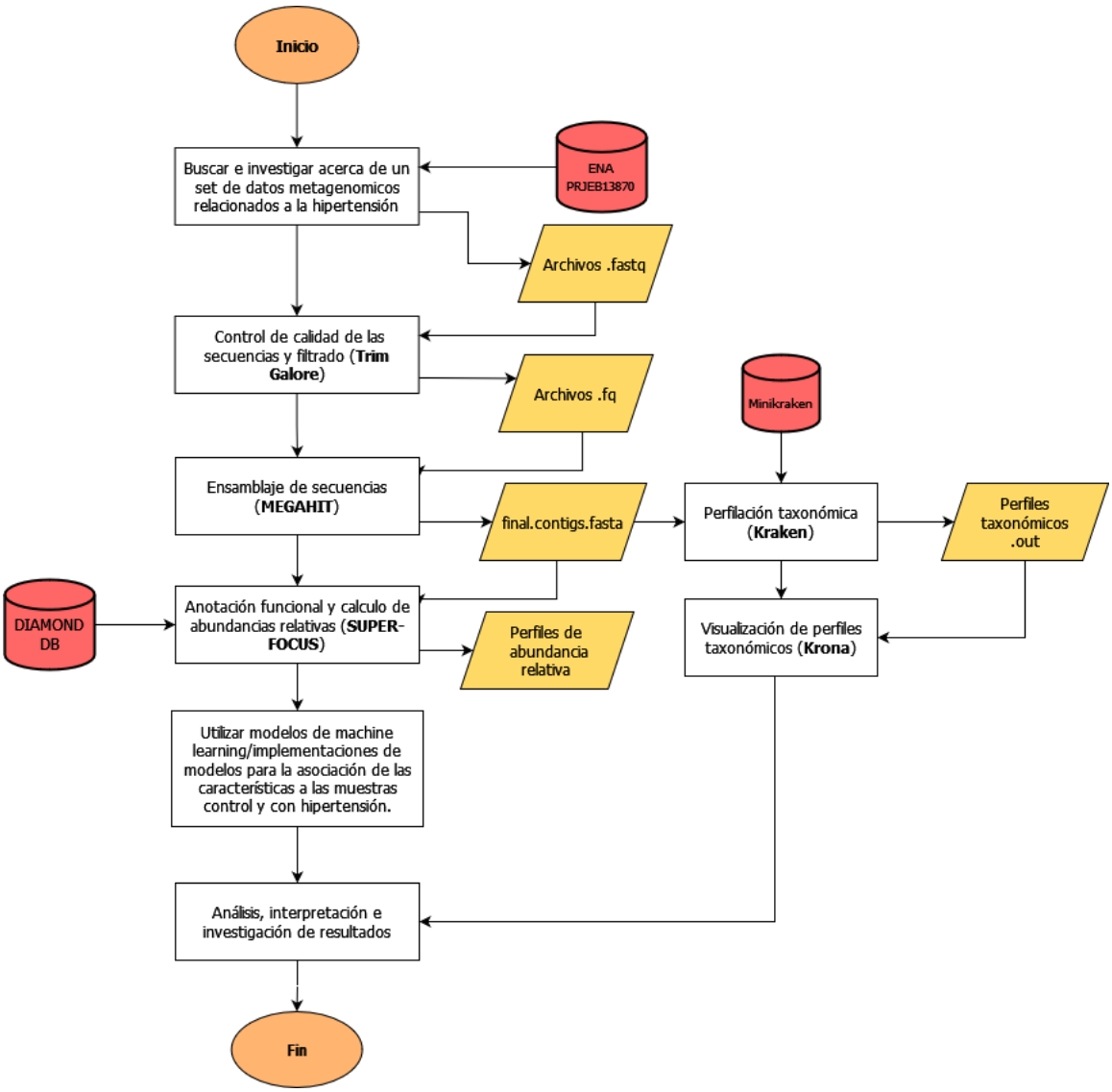


Figura 5: Diagrama de flujo del análisis metagenómico

utilizar datos de metagenómica para investigar a profundidad la hipertensión por lo que se utilizó una cohorte conformada por 41 pacientes control, es decir saludables y 99 pacientes con hipertensión. De estas secuencias metagenómicas se utilizaron 6 secuencias de cada grupo para realizar el análisis comparativo.

7.2.2. Creación y preparación de entorno virtual

Un entorno virtual permite tener un ambiente aislado en el cual se tiene un conjunto de paquetes con las versiones espcíficas que se utilizarán de estos paquetes, sin que estos afecten el entorno general y no existan conflictos entre las versiones utilizadas. Para esto se utilizó el manejador de paquetes Anaconda [27], el cual permite la creación de ambientes y además, proporciona una mayor facilidad en la instalación de paquetes y librerias.

Para la implementación del análisis metagenómico fue necesario crear un ambiente por medio de Anaconda en el cual se instalaron los paquetes que serían utilizados como se muestran en el Cuadro 3.

7.2.3. Control de calidad y filtración de secuencias

Cuando la colección de muestras de metagenoma fue seleccionada, se realizó un un control de calidad de cada muestra y lectura, para esto se definen los estándares óptimos de calidad de secuencias según los parámetros utilizados en el proceso de secuenciación y únicamente se toman en cuenta para el análisis las muestras que superan los límites establecidos. Para este proceso se utilizó la herramienta de Trim galore [28]. Esta herramienta permite utilizar en conjunto el software fastQC [29] para observar el control de calidad de las secuencias y Cutadapt [30].

FastQC es una herramienta desarrollada por Babraham Bioinformatics Institute que tiene como objetivo realizar comprobaciones para el control de cálidad para datos de secuencia sin procesar. Esta herramienta provee información acerca del estado de la lectura y esto permite tomar una decisión sobre si la secuencia puede ser utilizada para futuros análisis.

Cutadapt es una herramienta que encuentra y elimina secuencias adaptadoras, cebadores, colas poli-A y otros tipos de secuencias no deseadas de sus lecturas de secuenciación de alto rendimiento. El próposito de esto es poder eliminar cualquier contaminante que puedan tener las secuencias y que pueda interferir en el uso de estas.

7.2.4. Ensamblaje de genomas

El ensamblaje de genomas se refiere al proceso de colocar las secuencias de nucleótidos en el orden correcto. Esto es necesario dado que en el proceso de secuenciación las longitudes de las lecturas generadas son mucho menores a la longitud de un genoma completo. Por esto, es necesario utilizar bases de datos de referencia que permitan inferir en qué posición se deben ordenar las lecturas para formar la secuencia de los genomas completos y en este caso la secuencia de los diferentes genomas que conforman el metagenoma [31].

Actualmente existen diferentes softwares que permiten realizar el proceso de ensamblaje por medio de un algoritmo y consultas a las bases de datos existentes para tomar como referencia los genes y genomas ya registrados. Para este proceso se utilizó la herramienta Megahit [32].

7.2.5. Anotación funcional y cálculo de abundancia

La anotación funcional es el proceso de identificar elementos funcionales en la secuencia de un genoma para darles significado. A partir de los datos sin procesar se identifican las regiones del genoma que pertenecen a un gen.

La anotación del genoma es esencial debido a que la secuenciación del genoma o del ADN genera información acerca de la secuencia pero, no se tiene información acerca de

la funcionalidad de esta secuencia. Después de secuenciar el genoma, se debe realizar la anotación para brindar información lógica sobre sus características estructurales y roles funcionales [33].

En la anotación funcional es posible predecir los marcos de lectura abiertos por sus siglas en inglés ORF *Open Reading Frames*. La predicción de los marcos de lectura se realiza por medio de la anotación funcional de los genes o proteínas identificados en las secuencias se puede realizar buscando su similitud con secuencias bien verificadas experimentalmente disponibles en las bases de datos.

Para el proceso de anotación funcional se utilizó el flujo de trabajo de SUPER-FOCUS [34] y como bases de datos de referencia se utilizaron las bases de datos preconstruidas para el software de alineamiento utilizado por superfocus, DIAMOND [35]. Estas bases de datos preconstruidas son provistas en el repositorio del proyecto de superfocus.

7.2.6. Uso de herramienta de machine learning

Generalmente en los análisis de metagenómica, las inferencias acerca de la relación del fenotipo, en este caso la enfermedad de estudio, y los genes presentes en el metagenoma, se utiliza estadística.

En este caso se utilizó SIAMCAT [36], una herramienta basada en *machine learning* la cual permite realizar asociaciones entre comunidades microbianas y fenotipos del huésped.

7.2.7. Análisis e interpretación de resultados

Finalmente una vez procesados los datos y habiendo obtenido los resultados correspondientes del proceso, se llevó a cabo un análisis e interpretación de estos resultados, consultando la literatura para poder proporcionar información que sea relevante para conocer con mayor profundidad la enfermedad de hipertensión y cómo la microbiota intestinal puede ser un factor de riesgo.

CAPÍTULO 8

Resultados

8.1. Implementación de análisis metagenómico

El principal objetivo del presente proyecto era poder llevar a cabo la implementación de un flujo de trabajo que permitiera realizar el análisis metagenómico utilizando como caso de estudio la hipertensión.

El principal resultado de esta implementación se encuentra en el repositorio de GitHub del cual se adjunta el enlace en anexos. En este repositorio se pude encontrar el flujo de trabajo utilizado, especificando las herramientas utilizadas y scripts adicionales, esto con el fin de que el procedimiento de este análisis sea reproducible para distintos datos de tipo metagenómicos.

8.2. Anotación funcional

En la Figura 6 se puede observar un Análisis de Componentes Principales (PCA), el cual nos permite reducir la cantidad de variables orginales por medio de la transformación de estas a componentes princiaples. Los componentes principales son una combinación lineal de todas las variables originales.

Según el caso de estudio, el gráfico debería mostrar una diferenciación en cuanto a la agrupación de los fenotipos representados, sin embargo, es posible observar que no existe alguna agrupación formada por las muestras, que indique algun tipo de diferenciación en cuanto a las variables y la clasificación de las muestras.

Es posible observar en la Figura 7 una representación de la abundancia relativa de cada

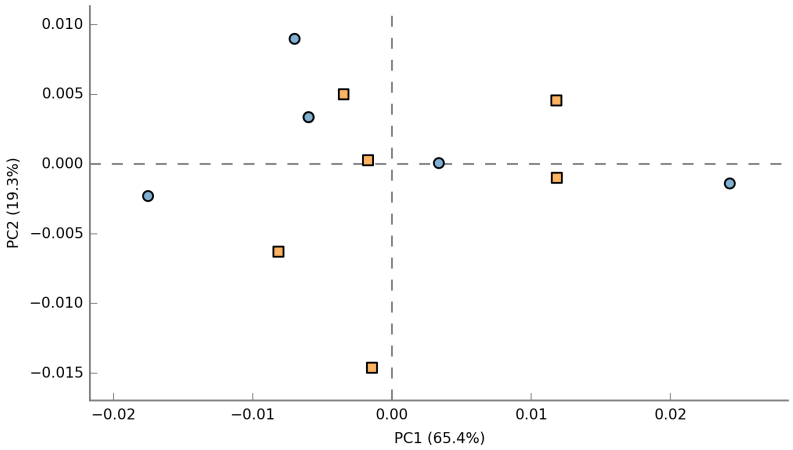


Figura 6: Gráfico PCA para proteínas del subsistema nivel 1. Muestras control en azul y muestras con hipertensión en naranja

uno de los grupos de proteínas encontradas en común en las muestras. Los grupos provienen de la anotación funcional realizada y a su vez la clasificación se encuentra descrita por la especificación de subsistemas definidos por la herramienta de SUPER-FOCUS. Un Subsistema puede definirse como conjuntos de proteínas que implementan un proceso biológico específico o complejo estructural. SUPER-FOCUS implementa estas agrupaciones en 3 niveles, siendo el primer nivel un subsistema más general y el tercer subsistema más específico.

Como se puede observar en el mapa de calor, no se presenta una diferencia significativa en cuanto a las muestras de hipertensión y las muestras control. Esto puede deberse a que la clasificación a nivel 1 de este subsistema es bastante general, por lo que estos grupos abarcan una gran cantidad de proteínas sin ser muy específicos.

Por otro lado, se observa que en este nivel el subsistema más abundante para todas las muestras es el que se encuentra relacionado a los carbohidratos. Este subsistema representa una mayor abundancia dado que las muestras provienen de la microbiota intestinal y la digestión y absorción de carbohidratos se lleva a cabo principalmente en los intestinos [37].

Como se mencionó anteriormente SUPER-FOCUS presenta 3 niveles de subsistemas que representan la especificidad de la clasificación de las proteínas encontradas. Se decidió evaluar el subsistema de nivel 2 y 3 para poder comparar los resultados con el subsistema de nivel 1.

En la Figura 8 se puede observar similar al nivel 1 que no se presenta una diferenciación en cuanto a la agrupación de los componentes por lo que no se puede determinar una diferencia estadistíca entre ambos tipos de muestra.

En la Figura 9 se puede observar un nivel mayor de diferenciación en cuanto a las muestras de hipertensión y muestras control. Sin embargo, esta diferenciación aun no llega a ser significativa, lo que puede deberse al tamaño de la muestra.

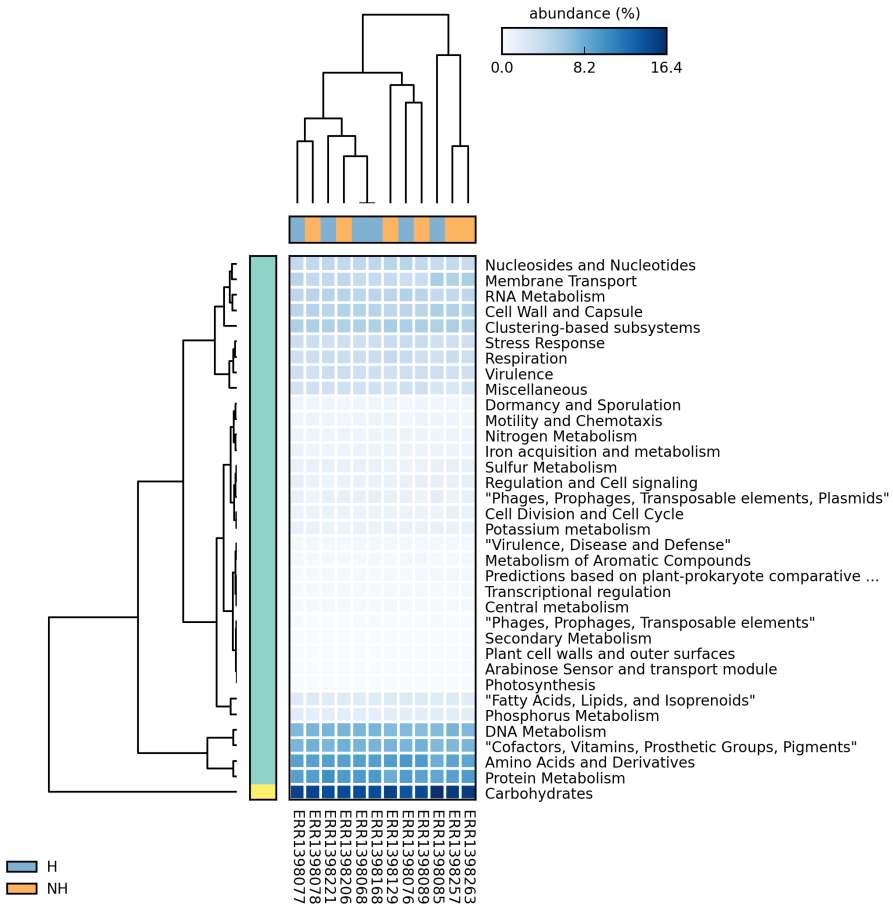


Figura 7: Mapa de calor de la representación de abundancias relativas para la clasificación de proteínas en el subsistema nivel 1

Debido a que el subsistema de nivel 3 es un nivel más específico se decidió realizar una comparación en cuanto a las 20 familias de proteínas más abundantes presentes en cada uno de los grupos de muestras como se puede observar en las figuras 10 y 11. Los subsistemas encontrados para ambos casos no difieren significativamente entre sí, sin embargo, es posible notar un caso interesante en las muestras de hipertensión se presenta dentro de los subsistemas más abundantes los transposones conjugativos bacteroidales y estos no se presentan para las muestras de control saludable.

Este caso es particularmente interesante dado que en las especies de Bacteroides, los plásmidos conjugativos y los transposones conjugativos se han estudiado intensamente durante décadas debido a la importancia de estos elementos móviles en la transferencia de genes de resistencia a los antibióticos. Estudios sugieren que los genomas de nuestras bacterias intestinales son modificados sustancialmente por otros miembros corresidentes del ecosistema, lo que da como resultado cepas de Bacteroidales altamente personalizadas, probablemente únicas para ese individuo [38].

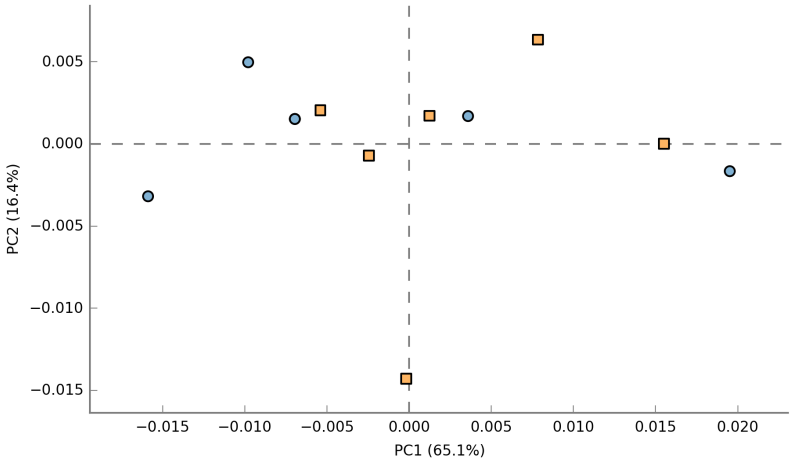


Figura 8: Gráfico PCA para proteínas del subsistema nivel 2. Muestras control en azul y muestras con hipertensión en naranja

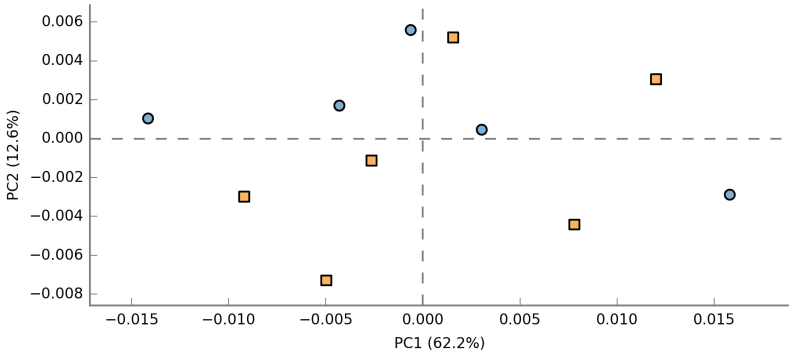


Figura 9: Gráfico PCA para proteínas del subsistema nivel 3. Muestras control en azul y muestras con hipertensión en naranja

La abundancia de este tipo de proteínas debería ser mayormente estudiada, dado que como se comentó anteriormente la transferencia de ADN entre especies podría dar una idea acerca de nuevos rasgos en las bacterias presentes en la microbiota que podrían tener una asociación con la hipertensión.

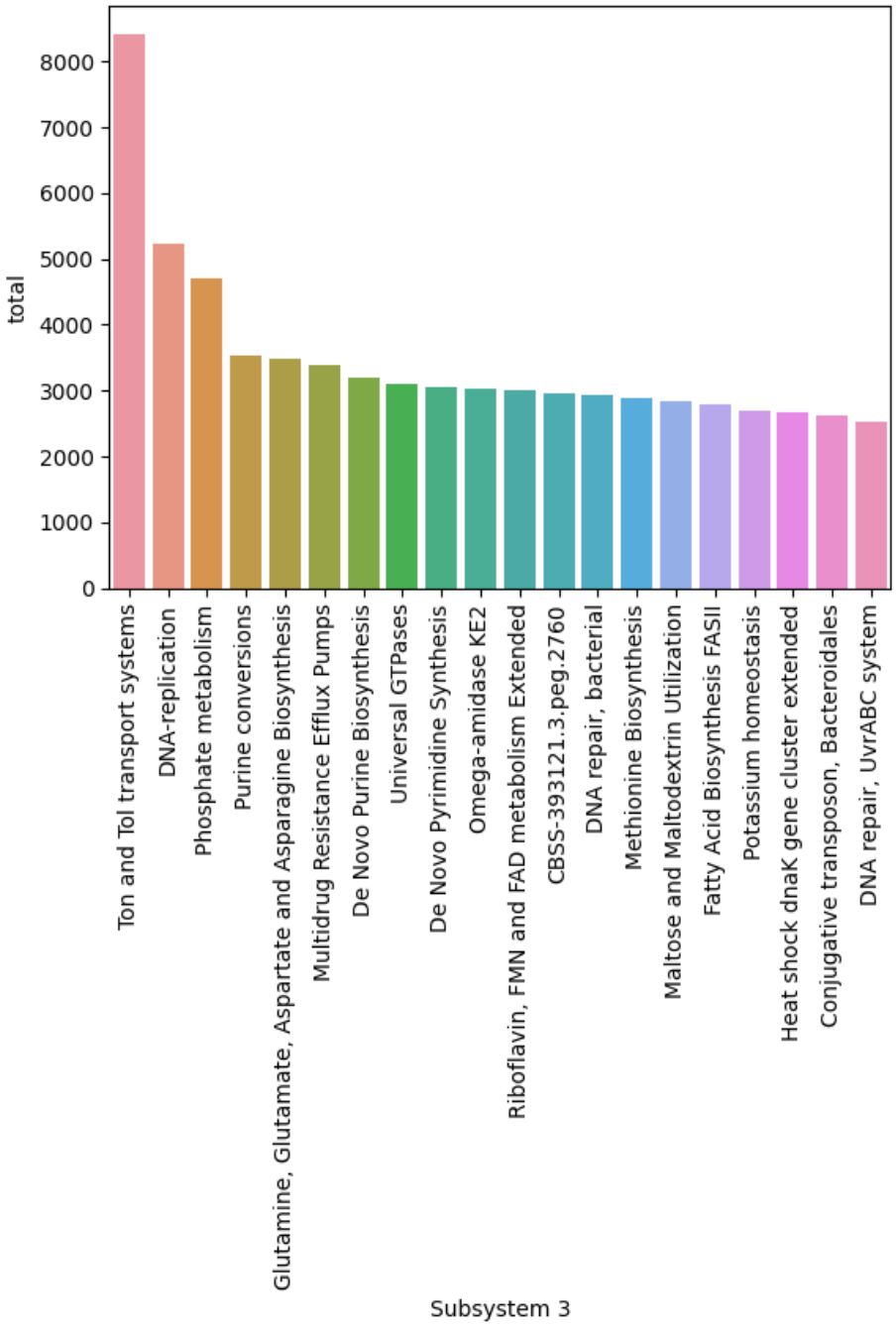


Figura 10: Grupos de proteínas más abundantes para las muestras de hipertensión del subsistema de nivel 3

8.3. Perfilación taxonómica

Para la clasificación taxonómica se utilizó la herramienta de Kraken [39] el cual es un sistema para la asignación de etiquetas taxonómicas a secuencias cortas de ADN utilizando alineaciones exactas de k-mers (ver anexo 23).

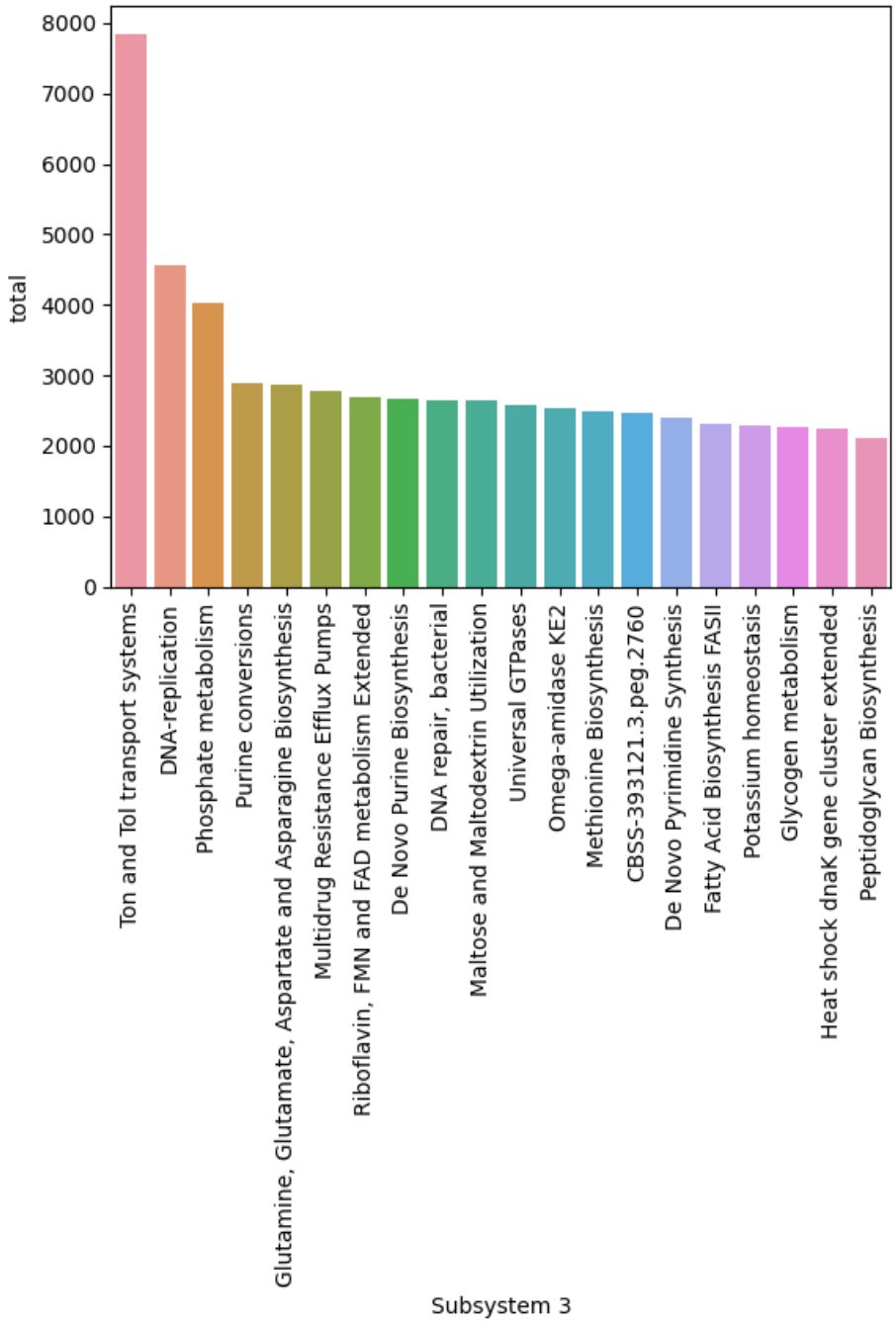


Figura 11: Grupos de proteínas más abundantes para las muestras de control saludable del subsistema de nivel 3

A partir de esta clasificación se obtuvieron los perfiles taxonómicos de cada muestra. El objetivo de esta perfilación era poder obtener los organismos que se presentaban con mayor o menor abundancia en las muestras y determinar si estos tienen alguna relación en cuanto al fenotipo de la hipertensión.

Es posible notar en la Tabla 12 que para todas las muestras existe un gran porcentaje de

Name	Number of raw reads	Classified reads	Unclassified reads	Microbial reads	Bacterial reads	Viral reads
ERR1398221	227,402	17.8%	82.2%	17.8%	17.8%	0.0022%
ERR1398206	156,182	23.3%	76.7%	23.3%	23.2%	0%
ERR1398068	155,818	26%	74%	25.9%	25.9%	0.00193%
ERR1398168	155,818	26%	74%	25.9%	25.9%	0.00193%
ERR1398263	152,454	25.7%	74.3%	25.7%	25.7%	0.00131%
ERR1398076	132,462	24.2%	75.8%	24.2%	24.1%	0.00226%
ERR1398077	125,691	33.9%	66.1%	33.9%	33.9%	0.00159%
ERR1398078	124,431	28.7%	71.3%	28.7%	28.6%	0.00402%
ERR1398089	120,295	33.2%	66.8%	33.1%	33.1%	0.000831%
ERR1398085	91,061	39.1%	60.9%	39.1%	39.1%	0.00439%
ERR1398257	89,907	29%	71%	29%	29%	0.00556%
ERR1398129	74,554	31.9%	68.1%	31.9%	31.9%	0%

Figura 12: Resumen de la composición de las muestras estudiadas [40].

lecturas que no fueron clasificadas, esto puede ser explicado principalmente a que se eligió la herramienta Kraken por su sensibilidad y precisión en la clasificación taxonómica. Sin embargo, esto implica que muchas de las lecturas no serán clasificadas porque no se alinean con precisión a las secuencias utilizadas provenientes de las bases de datos.

Por otro lado, es importante mencionar que en este análisis la capacidad de computo es limitada, por lo que se utilizó una base de datos de referencia simplificada (mini kraken) y esto también puede afectar en cuanto a la clasificación taxonómica. Además, otro factor que influye en cuanto a la cantidad de lecturas no clasificadas es que se estima que es posible encontrar aproximadamente 1000 especies bacterianas en el intestino con 2000 genes por especie lo cual resulta en una estimación de 2,000,000 de genes bacterianos [41]. Esta magnitud de genes sobrepasa incluso la cantidad de genes presentes en el genoma humano y muchas de las especies bacterianas aún no han sido caracterizadas o la información aún no ha sido digitalizada en bases de datos.

Para todas las muestras la totalidad de lecturas pertenecientes a microorganismos son de Bacterias y también es posible observar que se presentan pequeños porcentajes de lecturas pertenecientes a virus. Debido a que los microorganismos más abundantes son las bacterias, se decidió analizar a profundidad los organismos más abundantes pertenecientes a este dominio y realizar la comparación entre los grupos saludables y con hipertensión.

En las figuras 13 y 14 para las muestras de hipertensión y grupo control, respectivamente, se puede observar que no se presenta una mayor diferencia en cuanto a los 5 filos de bacterias más abundantes. Una de las principales diferencias que se puede observar en cuanto a los filos más abundantes es que en las muestras de hipertensión el filo Fusobacteria figura dentro de los filos más abundantes.

Las bacterias pertenecientes al filo Fusobacteria son miembros bastante comunes en la microbiota oral y generalmente tienen una relación simbiótica con sus huéspedes. Sin embargo, algunas de las especies pertenecientes a este filo son conocidos por encontrarse relacionadas a infecciones o enfermedades. Por ejemplo, F. *nucleatum* es un organismo que se encuentra asociado a diversas enfermedades. Este organismo se ha aislado de muestras clínicas en una variedad de enfermedades, que incluyen apendicitis, abscesos cerebrales, osteomielitis, pericarditis y resultados adversos del embarazo como la corioamnionitis [42].

La presencia abundante de este filo en las muestras de hipertensión debe ser estudiada con mayor profundidad dado que se ha demostrado que ciertas especies se encuentran relacionadas a infecciones y enfermedades, es posible que exista una relación entre la presencia de este filo y la hipertensión. Los demás filos representados en estos gráficos (Bacteroidetes, Firmicutes, Proteobacteria, Verrucomicrobia y Actinobacteria) pertenecen a los filos comunes que se encuentran en la microbiota intestinal, como se mostró anteriormente en la Tabla 1.

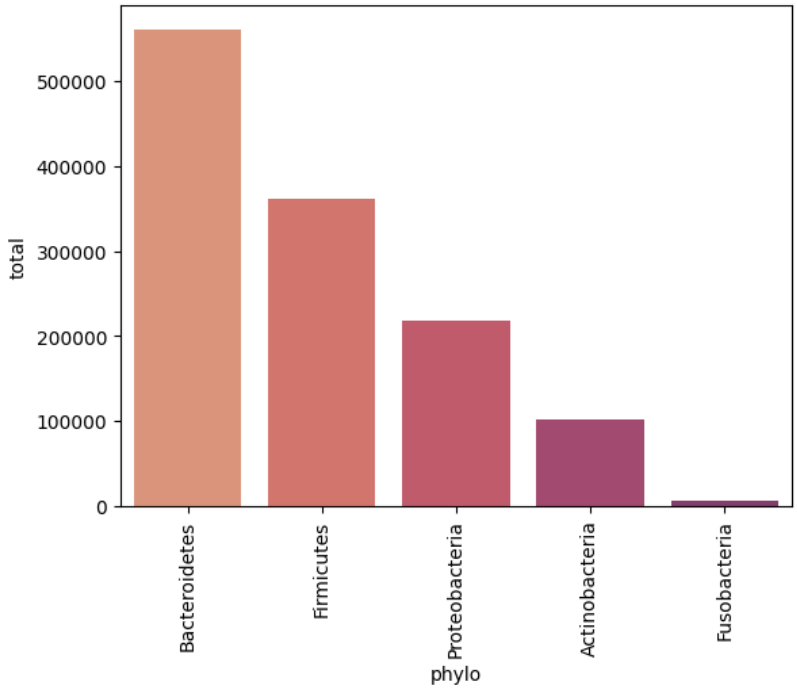


Figura 13: Filos del dominio bacteria con mayor abundancia en muestras de hipertensión. *Eje y representa la abundancia total, eje x la clasificación de filos.*

En cuanto a las clases observadas en las figuras 15 y 16 se puede evidenciar que no existe una diferencia significativa entre los diferentes tipos de muestras. Además, las clases más abundantes son pertenecientes a organismos que se encuentran comúnmente en la microbiota intestinal.

8.4. Asociaciones utilizando *machine learning*

Finalmente como parte del objetivo del proyecto se utilizó la herramienta de SIAMCAT, la cual es una pipeline que permite inferir las asociaciones entre comunidades microbianas y fenotipos del huésped.

El objetivo de utilizar una herramienta de machine learning era poder aprovechar la cantidad de datos que se generan en este ámbito de estudio. Cada una de las muestras cuenta con una cantidad extensa de datos, en este caso por cada muestra se tiene una cantidad extensa de microorganismos identificados.

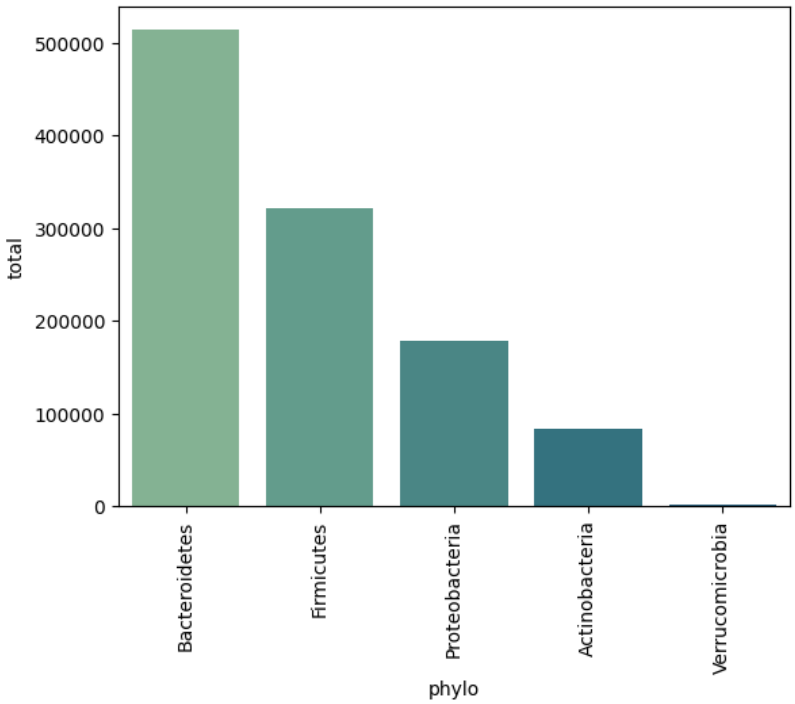


Figura 14: Filos del dominio bacteria con mayor abundancia en muestras de control saludable. *Eje y representa la abundancia total, eje x la clasificación de filos*

Para este paso se utilizó la función de check.associations, esta función calcula tres medidas de asociación entre las características, en este caso los microorganismos, y el fenotipo que para este estudio podían ser H(hipertensión) NH(No hipertensión). Las medidas de asociación que se calculan son:

- Prueba de Wilcoxon para la significancia
- El cambio de pliegue generalizado (gFC) es un pseudo cambio de pliegue que se calcula como la media geométrica de las diferencias entre los cuantiles para las diferentes clases que se encuentran en la etiqueta.
- El cambio de prevalencia entre las dos clases diferentes que se encuentran en la etiqueta.

En este caso se utilizó el nivel significancia de 0.05. A partir de esto no fue posible obtener ningun tipo de asociación del software. Esto puede deberse a que como se mencionaba anteriormente se necesita una muestra de datos más grande para poder utilizar el modelo de machine learning utilizado por la herramienta.

SIAMCAT también cuenta con una función para la creación de modelos de machine learning usando como base los datos de especies microbianas que se le proporcionen. Se utilizó esta función para crear un modelo que se adecuará a los datos proporcionados.

Con esta función se probaron los modelos de lasso y ridge. Los modelos no presentan una precisión de predicción. La curva ROC nos dice qué tan bien un modelo puede distinguir

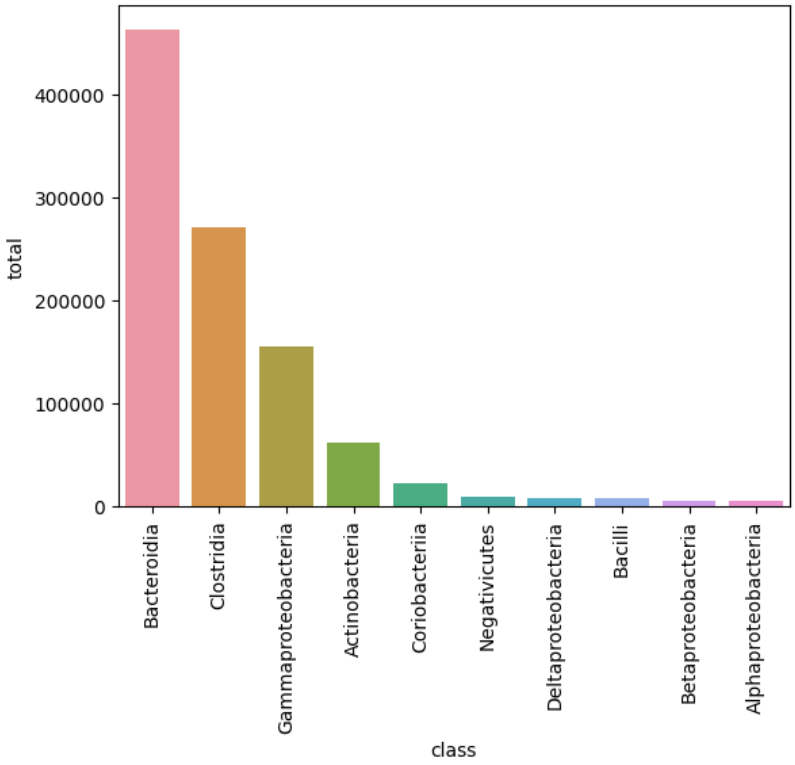


Figura 15: Clases del dominio bacteria con mayor abundancia en muestras de hipertensión *Eje y representa la abundancia total, eje x la clasificación de clases.*

entre las clasificaciones definidas. En las figuras 17 y 19 se puede observar que los gráficos para la curva ROC presentan un nivel muy bajo.

Así mismo en las figuras 18 y 20 se presenta la precisión del modelo de predicción la clasificación de nuevas muestras a partir de las variables definidas en el modelo. En este caso también se observa un porcentaje de precisión bastante bajo lo que nos indica que el modelo generado no puede ser utilizado para clasificación o asociación de variables.

Estos resultados pueden atribuirse a que el tamaño de muestras para el modelo es pequeño o bien a que realmente las variables, en este caso los microorganismos, no representan ninguna asociación hacia la clasificación del fenotipo.

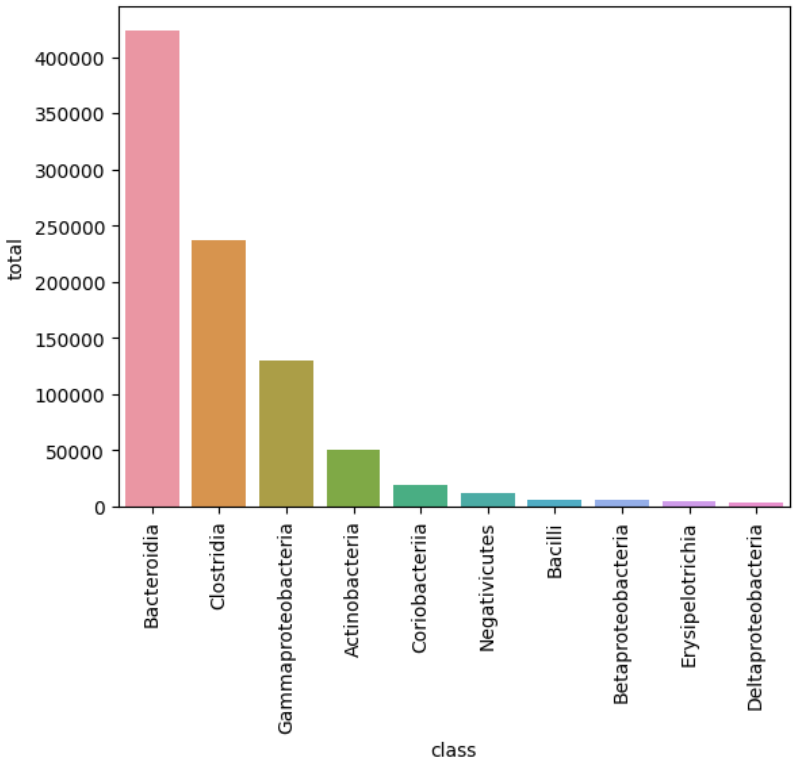


Figura 16: Clases del dominio bacteria con mayor abundancia en muestras de control saludable. *Eje y representa la abundancia total, eje x la clasificación de clases.*

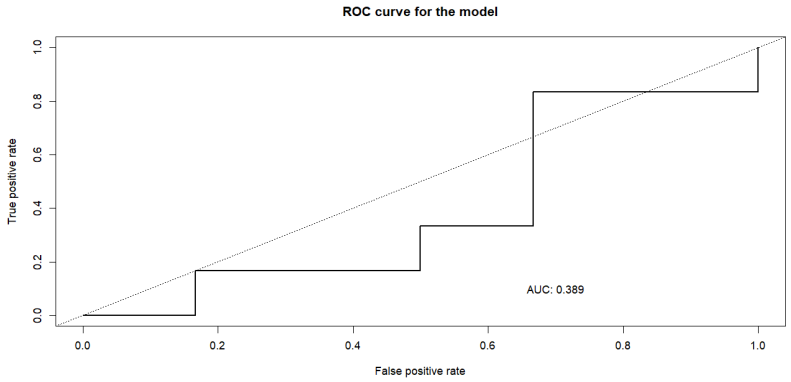


Figura 17: Gráfico ROC modelo lasso

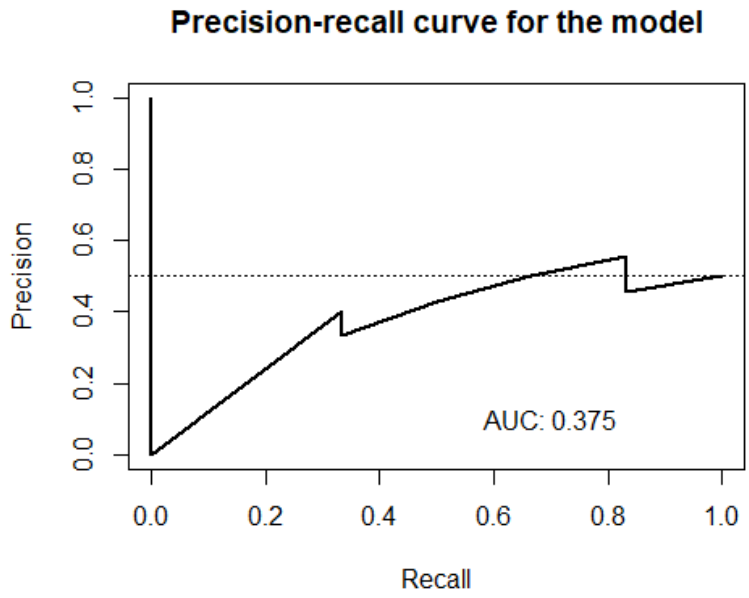


Figura 18: Gráfico precisión modelo lasso

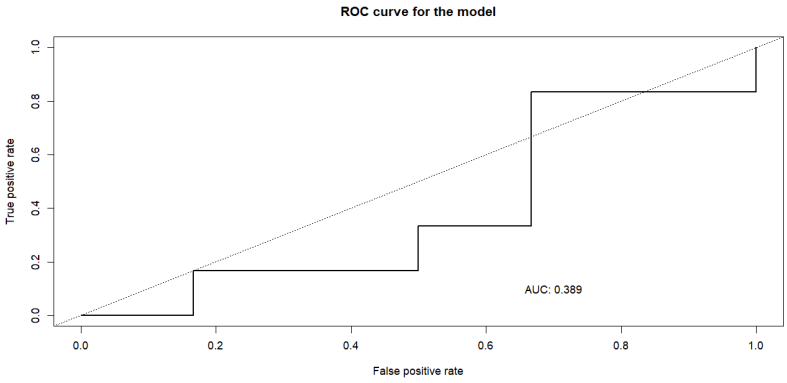


Figura 19: Gráfico ROC modelo Ridge

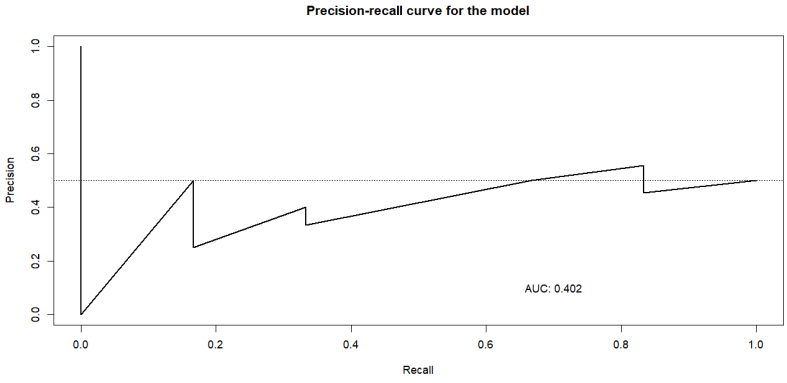


Figura 20: Gráfico precisión modelo Ridge

CAPÍTULO 9

Conclusiones

- Fue posible implementar un flujo de trabajo para el análisis metagenómico utilizando como caso de estudio la hipertensión.
- Se evidenció que no se presenta alguna diferencia estadísticamente significativa en la abundancia de las proteínas encontradas en las muestras de hipertensión.
- Para el subsistema de nivel 1 de proteínas no se encontró una diferencia significativa en la abundancia de las proteínas pero, se evidenció que el subsistema más abundante en ambos casos fue la familia de proteínas que se encuentra relacionada a los carbohidratos.
- Para el subsistema de nivel 3 se encontró una abundancia mayor en el subsistema de transposones conjugativos bacteroidales en las muestras de hipertensión. Este hallazgo es de especial interés dado que estass proteínas pueden encontrarse asociadas a la variabilidad genética de bacterias y tener una posible asociación a la hipertensión.
- Se observó que a partir de los resultados del análisis metagenómico los microorganismos que se presentaron mayormente en ambos tipos de muestra fueron las bacterias.
- En la comparación de filos de bacterias más abundantes se encontró que el filo Fusobacteriota se encuentra con mayor abundancia en las muestras de hipertensión. Este hallazgo es de particular interés debido a la asociación entre especies de este filo y enfermendades o infecciones.
- Se demostró que para el uso de herramientas o modelos basados en machine learning la inferencia estadística de asociaciones entre comunidades microbianas y fenotipos del huésped es necesario un tamaño de muestra mayor.

CAPÍTULO 10

Recomendaciones

- Utilizar un conjunto de muestras más grande, tomando en cuenta que esto requiere de mayor potencia computacional en los equipos en los cuales se lleve a cabo el análisis.
- Analizar los genes o proteínas que presentan una mayor abundancia en relación a la hipertensión como posibles biomarcadores de la enfermedad.
- Realizar un análisis comparativo de los resultados utilizando diferentes metodologías y herramientas para el análisis de los datos metagenómicos.
- Utilizar un conjunto de datos que cuente con metadata relevante que permita enriquecer el estudio y conocer cómo afectan los factores externos a la microbiota.
- Utilizar el flujo de análisis metagenómico para el estudio de muestras pertenecientes a la región de latinoamérica.

CAPÍTULO 11

Bibliografía

[1] c. y. e. a. Qin li, "A metagenome-wide association study of gut microbiota in type 2 diabetes," *Nature*, vol. 490, págs. 55-60, 2012. dirección: https://doi.org/10.1038/nature11450.

[2] NIH. "Genome-Wide Association Studies Fact Sheet." (), dirección: https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet.

[3] N. y. e. a. Kishikawa Maeda, "Metagenome-wide association study of gut microbiome revealed novel aetiology of rheumatoid arthritis in the Japanese population," *Annals of the Rheumatic Diseases*, vol. 79, págs. 103-111, 2020. dirección: doi:10.1136/annrheumdis-2019-215743.

[4] M. y. e. a. Kishikawa Ogawa, "A Metagenome-Wide Association Study of Gut Microbiome in Patients With Multiple Sclerosis Revealed Novel Disease Pathology," *Frontiers in cellular and infection microbiologys*, vol. 10, pág. 585 973, 2020. dirección: https://doi.org/10.3389/fcimb.2020.585973.

[5] F. y. e. a. Zhang Jia, "The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment.," *Nature medicine*, vol. 21, págs. 895-905, 2015. dirección: doi:10.1038/nm.3914.

[6] Y. Tomofuji, Y. Maeda, E. Oguro-Igashira y col., "Metagenome-wide association study revealed disease-specific landscape of the gut microbiome of systemic lupus erythematosus in Japanese," *Annals of the Rheumatic Diseases*, vol. 80, n.º 12, págs. 1575-1583, 2021, ISSN: 0003-4967. DOI: 10.1136/annrheumdis-2021-220687. eprint: https://ard.bmj.com/content/80/12/1575.full.pdf. dirección: https://ard.bmj.com/content/80/12/1575.

[7] G. J. Torsvik V. y D. F., "High diversity in DNA of soil bacteria," vol. 56, págs. 782-787, 1990.

[8] L. Coughlan, P. Cotter, C. Hill y A. Alvarez-Ordóñez, "Biotechnological applications of functional metagenomics in the food and pharmaceutical industries," *Frontiers in Microbiology*, vol. 6, 2015, ISSN: 1664-302X. DOI: 10.3389/fmicb.2015.00672. dirección: https://www.frontiersin.org/articles/10.3389/fmicb.2015.00672.

[9] O. P. de la Salud. "Hipertensión." (2022), dirección: https://www.paho.org/es/temas/hipertension (visitado 2022).

[10] M. de Salúd Pública y Asistencia Social de Guatemala. "Casos de Morbilidad y Mortalidad por Crónicas." (2020), dirección: https://sigsa.mspas.gob.gt/datos-de-salud/morbilidad/enfermedades-cronicas (visitado 2022).

[11] P. y. K. Ursell Metcalf, "Defining the Human Microbiome," *Nutrition reviews*, vol. 1, págs. 38-44, 2012. dirección: https://doi.org/10.1111/j.1753-4887.2012.00493.x.

[12] A. y García, "El microbioma humano. Su papel en la salud y en algunas enfermedades," *Cirugía y cirujanos*, vol. 84, págs. 31-35, 2016. dirección: https://www.elsevier.es/es-revista-cirugia-cirujanos-139-articulo-el-microbioma-humano-su-papel-X0009741116539900.

[13] C. Genomics. "Why Do We Perform 16S rRNA Sequencing." (2018), dirección: https://www.cd-genomics.com/blog/why-do-we-perform-16s-rrna-sequencing/ (visitado 2022).

[14] W. y Jia, "Metagenome-wide association studies: fine-mining the microbiome," *Nature reviews microbiology*, vol. 14, págs. 508-522, 2016. dirección: doi:10.1038/nrmicro.2016.83.

[15] Illumina. "Shotgun Metagenomic Sequencing." (2018), dirección: https://www.illumina.com/areas-of-interest/microbiology/microbial-sequencing-methods/shotgun-metagenomic-sequencing.html.

[16] E. Green. "Shotgun Sequencing." (2022), dirección: https://www.genome.gov/genetics-glossary/Shotgun-Sequencing.

[17] K. A. y. e. a. Judd W. Campbell C., ed., *Plant systematics: a phylogenetic approach.* Sinauer Axxoc, 2022, ISBN: 0-87893-403-0.

[18] S. K. y. V. P. Aleksandar Danicic Nemanja Vucic. "Taxanomic Profiling of Metagenomics Samples." (2018), dirección: https://www.sevenbridges.com/taxonomic-profiling-of-metagenomics-samples/.

[19] F. Collins. "Genoma." (2022), dirección: https://www.genome.gov/es/genetics-glossary/Genoma (visitado 2022).

[20] Oncogenomics. "Genómica en la salud." (2022), dirección: https://www.oncogenomics.es/genomica/ (visitado 2022).

[21] S. Gladman. "De novo Genome Assembly for Illumina Data." (), dirección: https://www.melbournebioinformatics.org.au/tutorials/tutorials/assembly/assembly-protocol/ (visitado 2022).

[22] genomegov. "Metagenómica." (2022), dirección: https://www.genome.gov/es/genetics-glossary/Metagenomica (visitado 2022).

[23] IBM. "¿Qué es Machine Learning?" (2022), dirección: https://www.ibm.com/mx-es/analytics/machine-learning.

[24] G. y Techtmann, "Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring," *Computational and Structural Biotechnology Journal*, vol. 19, págs. 1092-1107, 2021. dirección: https://doi.org/10.1016/j.csbj.2021.01.028.

[25] Mayoclinic. "Presión arterial alta (hipertensión)." (2021), dirección: https://www.mayoclinic.org/es-es/diseases-conditions/high-blood-pressure/symptoms-causes/syc-20373410 (visitado 2022).

[26] Z. y. e. a. Li, "Gut microbiota dysbiosis contributes to the development of hypertension," *Microbiome*, vol. 5, 2017. dirección: https://doi.org/10.1186/s40168-016-0222-x.

[27] *Anaconda Software Distribution*, ver. Vers. 2-2.4.0, 2020. dirección: https://docs.anaconda.com/.

[28] F. Krueger, F. James, P. Ewels, E. Afyounian y B. Schuster-Boeckler, *FelixKrueger/TrimGalore: v0.6.7 - DOI via Zenodo*, ver. 0.6.7, jul. de 2021. DOI: 10.5281/zenodo.5127899. dirección: https://doi.org/10.5281/zenodo.5127899.

[29] Andrews. "FastQC: A Quality Control Tool for High Throughput Sequence Data." (2015), dirección: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[30] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal*, vol. 17, n.º 1, págs. 10-12, 2011, ISSN: 2226-6089. DOI: 10.14806/ej.17.1.200. dirección: https://journal.embnet.org/index.php/embnetjournal/article/view/200.

[31] B. Foxman, "Chapter 5 - A Primer of Molecular Biology," en *Molecular Tools and Infectious Disease Epidemiology*, B. Foxman, ed., San Diego: Academic Press, 2012, págs. 53-78, ISBN: 978-0-12-374133-2. DOI: https://doi.org/10.1016/B978-0-12-374133-2.00005-8. dirección: https://www.sciencedirect.com/science/article/pii/B9780123741332000058.

[32] L. y. e. a. Luo Liu, "MEGAHIT v1.0: A Fast and Scalable Metagenome Assembler driven by Advanced Methodologies and Community Practices," 2016.

[33] M. M. y. R. K. Aditya Harbola Deepti Negi, *Chapter 27 - Bioinformatics and biological data mining*, D. B. Singh y R. K. Pathak, eds. Academic Press, 2022, págs. 457-471, ISBN: 978-0-323-89775-4. DOI: https://doi.org/10.1016/B978-0-323-89775-4.00019-5. dirección: https://www.sciencedirect.com/science/article/pii/B9780323897754000195.

[34] D. B. E. R. Silva G. Green K., "SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data.," *Bioinformatics (Oxford, England)*, vol. 32, págs. 354-361, 2016. DOI: https://doi.org/10.1093/bioinformatics/btv584.

[35] R. K. y. D. H. Buchfink B., "Sensitive protein alignments at tree-of-life scale using DIAMOND," *Nature methods*, vol. 18, págs. 366-368, 2021. DOI: https://doi.org/10.1038/s41592-021-01101-x.

[36] E. y. e. a. Wirbel Zych, "Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox," *Genome Biol*, vol. 22, pág. 93, 2021. dirección: https://doi.org/10.1186/s13059-021-02306-1.

[37] P. P. GarcÃa Luna y G. LÃ Gallardo, "EvaluaciÃde la absorciÃy metabolismo intestinal," es, *NutriciÃHospitalaria*, vol. 22, págs. 05-13, mayo de 2007, ISSN: 0212-1611. dirección: http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0212-16112007000500002&nrm=iso.

[38] M. J. Coyne, N. L. Zitomersky, A. M. McGuire, A. M. Earl y L. E. Comstock, "Evidence of Extensive DNA Transfer between Bacteroidales Species within the Human Gut," *mBio*, vol. 5, n.º 3, e01305-14, 2014. DOI: 10.1128/mBio.01305-14. eprint: https://journals.asm.org/doi/pdf/10.1128/mBio.01305-14. dirección: https://journals.asm.org/doi/abs/10.1128/mBio.01305-14.

[39] W. D. S. S., "Kraken: ultrafast metagenomic sequence classification using exact alignments.," vol. 15, págs. 354-361, 2014. DOI: https://doi.org/10.1186/gb-2014-15-3-r46.

[40] F. P. Breitwieser y S. L. Salzberg, "Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification," *Bioinformatics*, vol. 36, n.º 4, págs. 1303-1304, sep. de 2019, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz715. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/4/1303/38712594/btz715.pdf. dirección: https://doi.org/10.1093/bioinformatics/btz715.

[41] B. M. y. e. a. Gilbert J., "Current understanding of the human microbiome," *Nature medicine*, vol. 24, págs. 392-400, 2018. DOI: https://doi.org/10.1038/nm.4517.

[42] B. C. y Garrett W., "Fusobacterium nucleatum - symbiont, opportunist and oncobacterium," *Nature reviews*, vol. 17, págs. 155-166, 2019. DOI: https://doi.org/10.1038/s41579-018-0129-6.

[43] B. N. y. P. Ondov B., "Interactive metagenomic visualization in a Web browser," *Bioinformatics*, vol. 12, pág. 385, 2011. DOI: https://doi.org/10.1186/1471-2105-12-385.

CAPÍTULO 12

Anexos

12.1. Repositorio de GitHub

A continuación se presenta el enlace al repositorio de GitHub en el cual se puede encontrar la documentación acerca de la implementación del análisis metagenómico y el conjunto de archivos de los resultados principales.

https://github.com/JennsiS/MWASHypertension

12.2. Herramientas utilizadas

SUPER-FOCUS

La herramienta de SUPER-FOCUS fue utilizada para el procesamiento de las muestras de metagenómica, específicamente para llevar a cabo la anotación funcional de estas. Esta herramienta permite identificar la abundancia de los genes funcionales de los organismos presentes en la muestra original. SUPER-FOCUS utiliza un enfoque ágil basado en la homología que utiliza una base de datos de referencia reducida para informar los subsistemas presentes en los conjuntos de datos metagenómicos y perfilar sus abundancias.

Una de las razones principales de utilizar esta herramienta es por su capacidad de análisis de muestras y su rapidez en comparación con herramientas similares. En la Figura 21 se puede observar el flujo de trabajo general de super-focus que se emplea para todas las secuencias en un set de muestras de metagenoma.

Por otro lado, en la Figura 22 se puede observar una representación de cómo se definen los subsistemas de super-focus en sus distintos niveles. El primer nivel es la clase más general y el tercer nivel es la clase más específica, y un rol funcional.

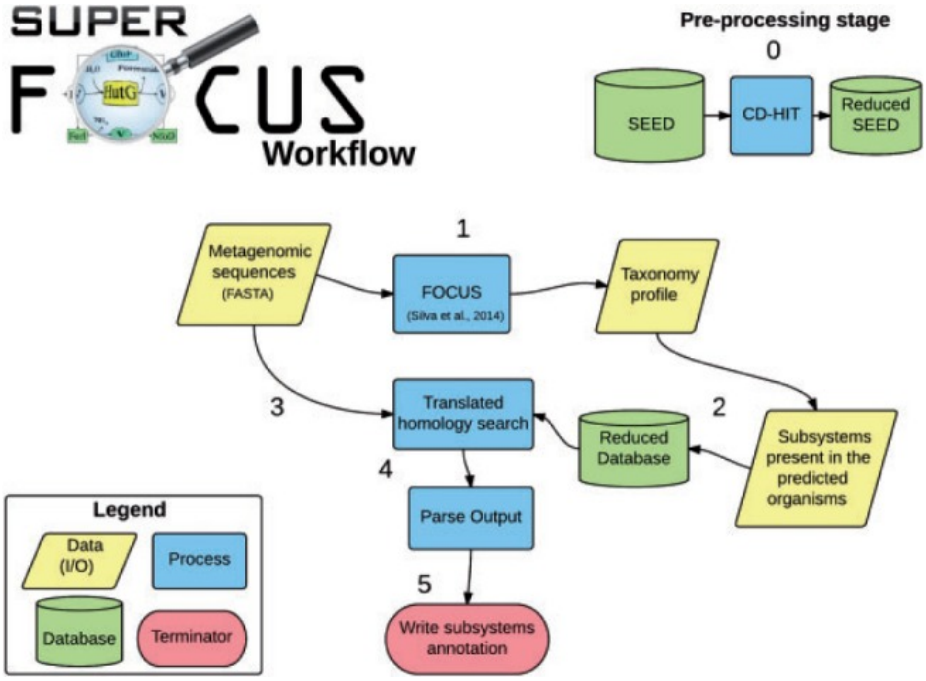


Figura 21: Flujo de trabajo de la herramienta superfocus[34]

Kraken

Kraken es un programa rápido y de alta precisión para asignar etiquetas taxonómicas a secuencias de ADN metagenómico. El funcionamiento principal de esta herramienta se basa en que en el núcleo de Kraken hay una base de datos que contiene registros que consisten en un k-mer de todos los organismos cuyos genomas contienen ese k-mer. Esta base de datos, construida utilizando una biblioteca de genomas especificada por el usuario, permite una búsqueda rápida del nodo más específico en el árbol taxonómico que está asociado con un k-mer determinado. Las secuencias se clasifican consultando la base de datos para cada k-mer en una secuencia y luego utilizando el conjunto resultante de taxones para determinar una etiqueta adecuada para la secuencia se puede observar el flujo general de esta herramienta en la Figura 23.

12.3. Perfilación taxonómica

En los diagramas de la figura 24 a la 66 se presentan perfiles taxónomicos más específicos en los cuales se detalla la composición de cada una de las muestras desglosando los organismos según su orden taxonómico. Los diagramas representados fueron creados con la herramienta de Krona [43] y su versión interactiva puede ser encontrada en el repositorio de GitHub.

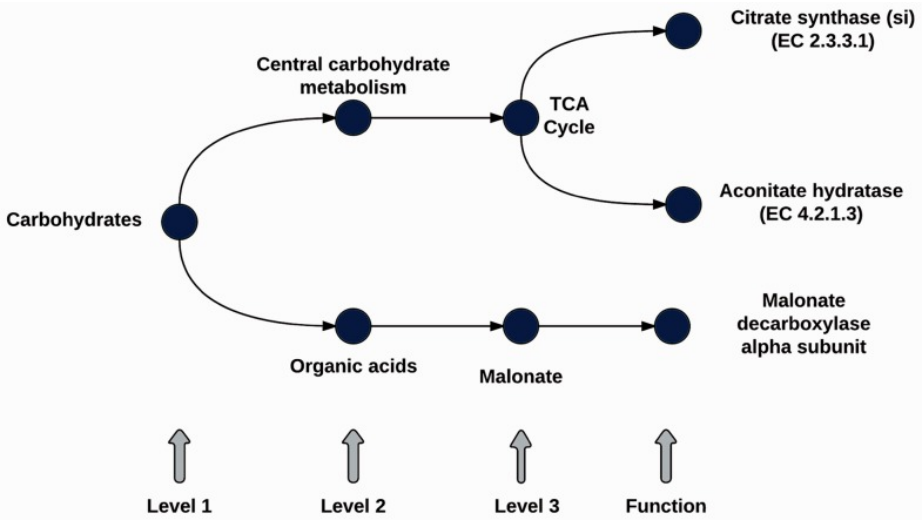


Figura 22: Representación de una estructura de subsistema (Niveles 1-3 clasificaciones y Función)[34]

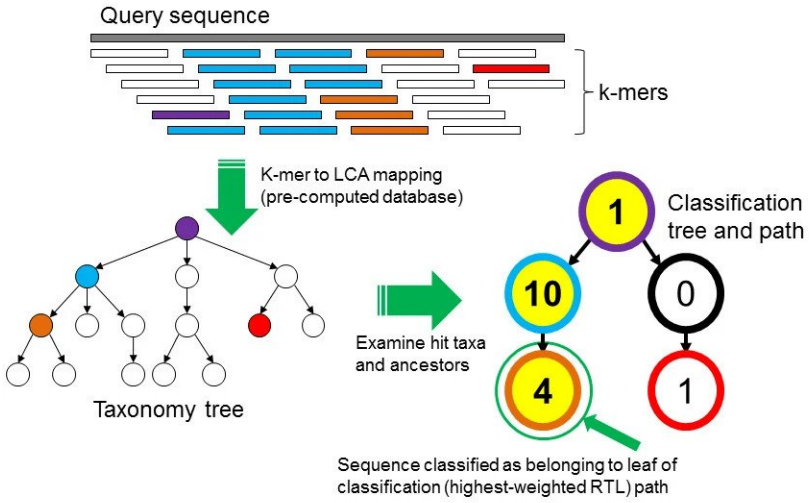


Figura 23: Algoritmo de clasificación de secuencias utilizado por Kraken[39]

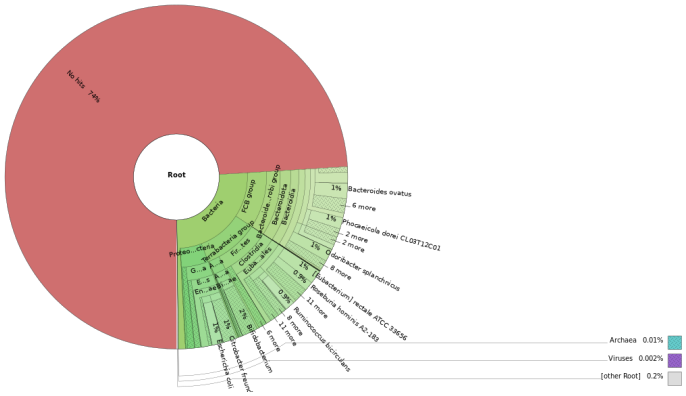


Figura 24: Perfil taxonómico muestra hipertensión ERR1398068

Figura 25: Perfil taxonómico de virus muestra hipertensión ERR1398068

Figura 26: Perfil taxonómico de archaea muestra hipertensión ERR1398068

Figura 27: Perfil taxonómico de bacterias muestra hipertensión ERR1398068

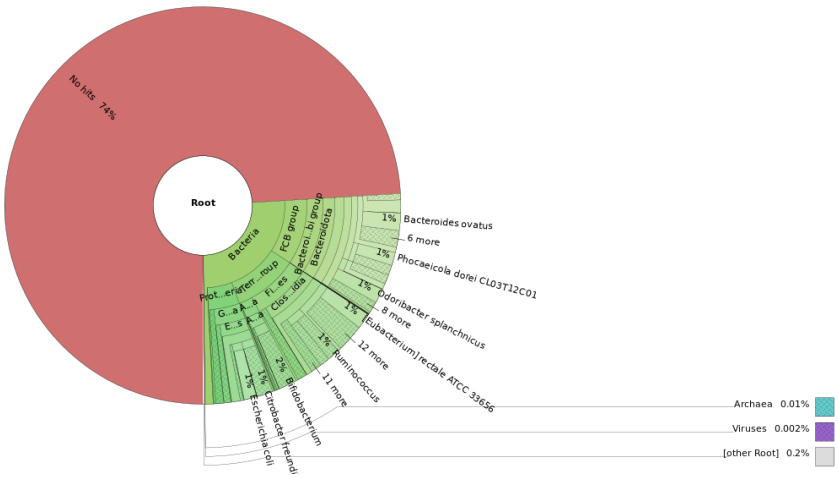


Figura 28: Perfil taxonómico muestra hipertensión ERR1398168

Figura 29: Perfil taxonómico de virus muestra hipertensión ERR1398168

Figura 30: Perfil taxonómico de archaea muestra hipertensión ERR1398168

Figura 31: Perfil taxonómico de bacterias muestra hipertensión ERR1398168

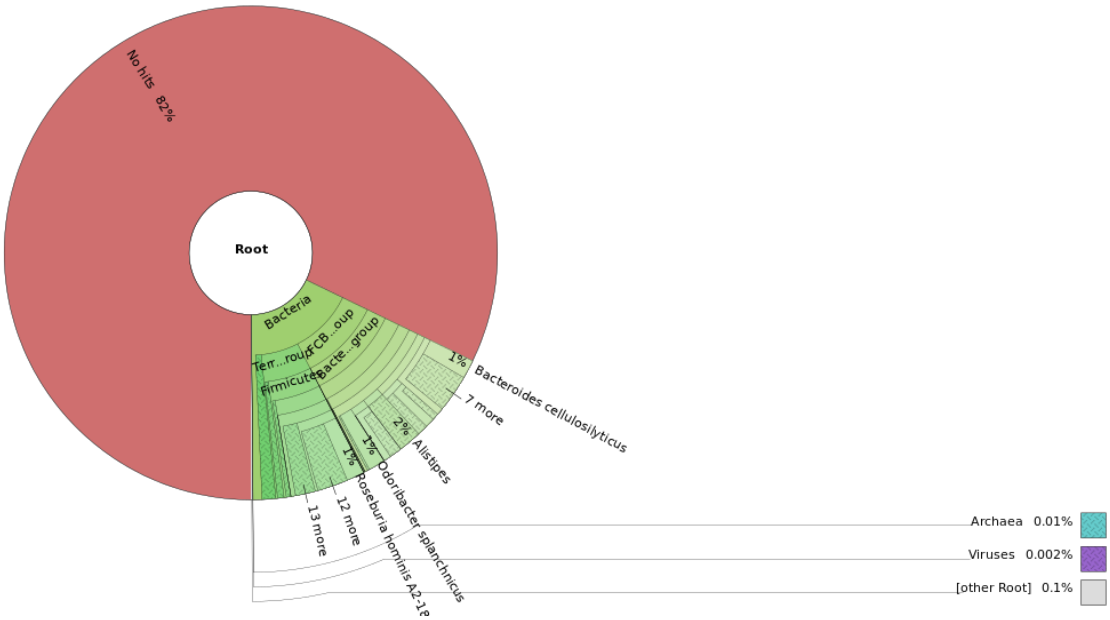


Figura 32: Perfil taxonómico muestra hipertensión ERR1398221

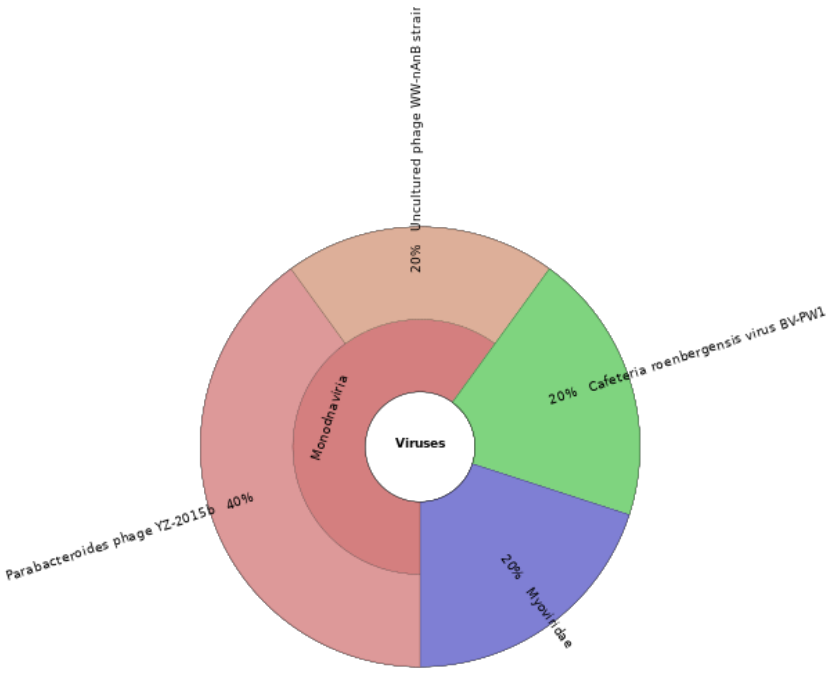


Figura 33: Perfil taxonómico de virus muestra hipertensión ERR1398221

Figura 34: Perfil taxonómico de archaea muestra hipertensión ERR1398221

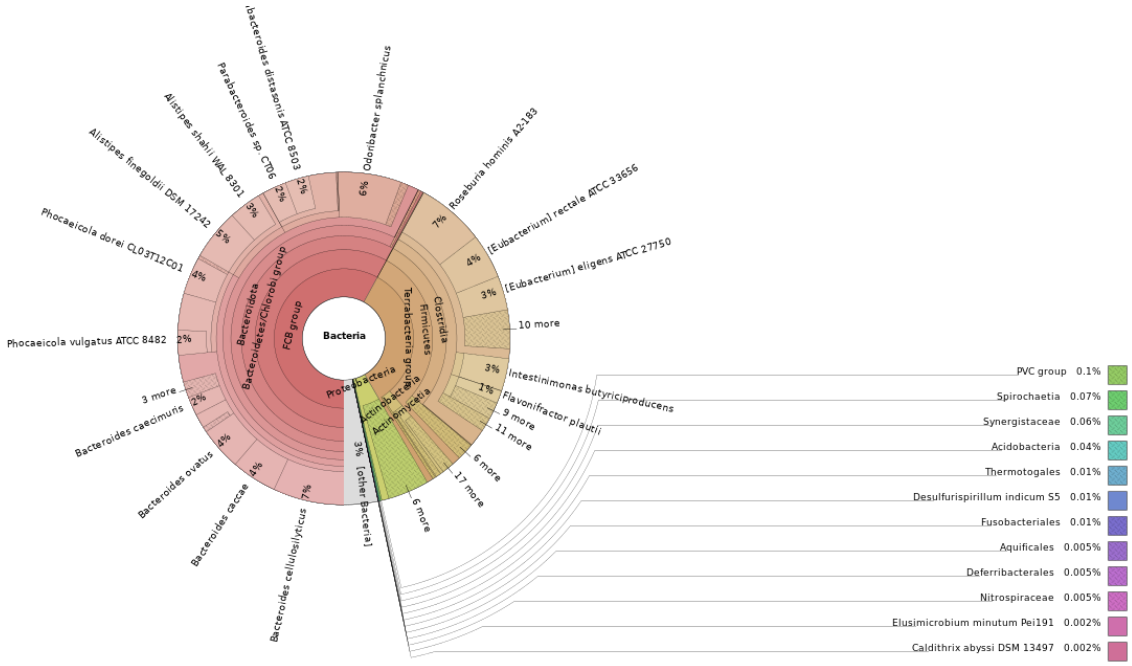


Figura 35: Perfil taxonómico de bacterias muestra hipertensión ERR1398221

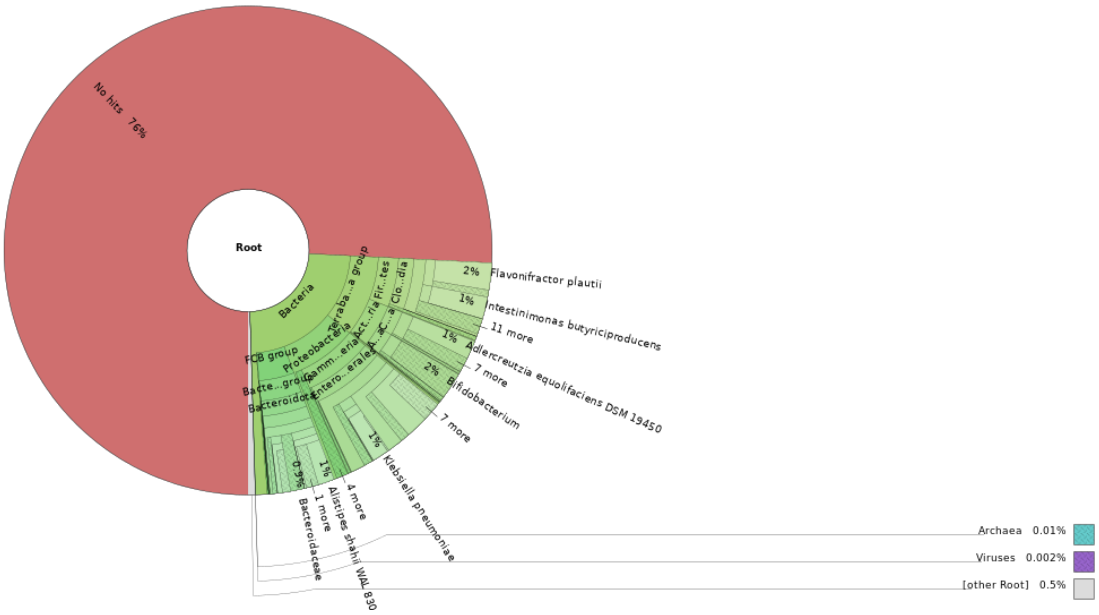


Figura 36: Perfil taxonómico muestra hipertensión ERR1398076

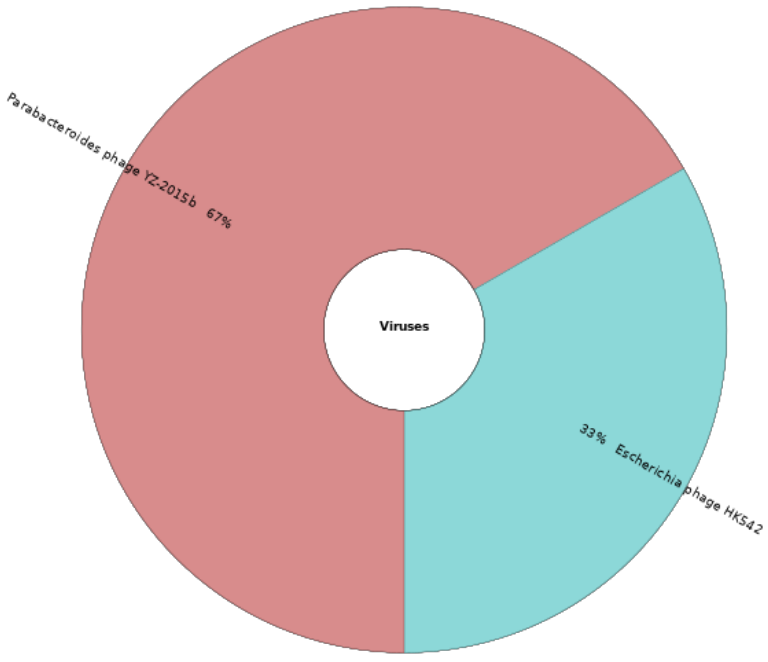


Figura 37: Perfil taxonómico de virus muestra hipertensión ERR1398076

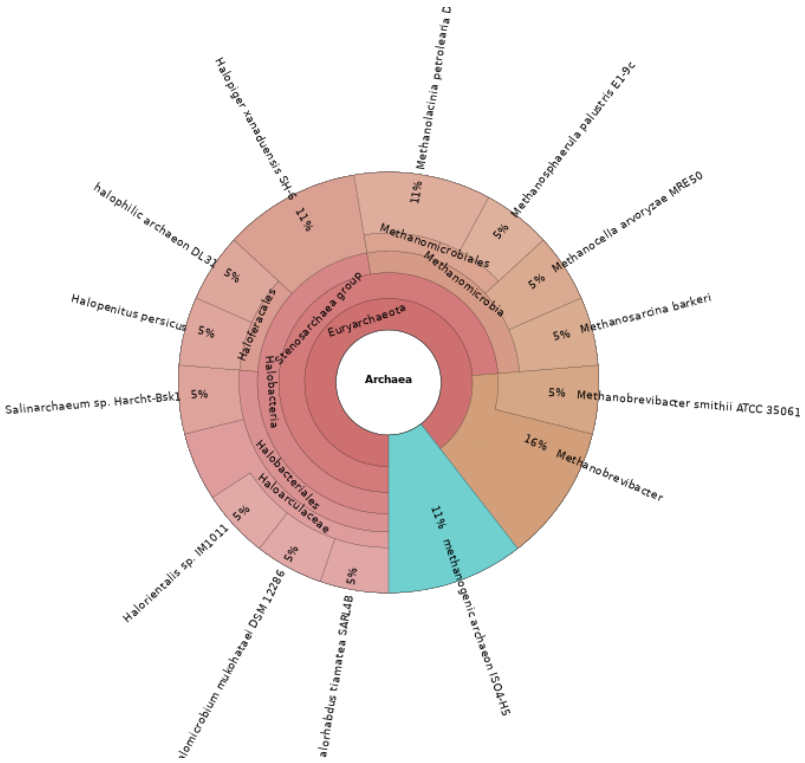


Figura 38: Perfil taxonómico de archaea muestra hipertensión ERR1398076

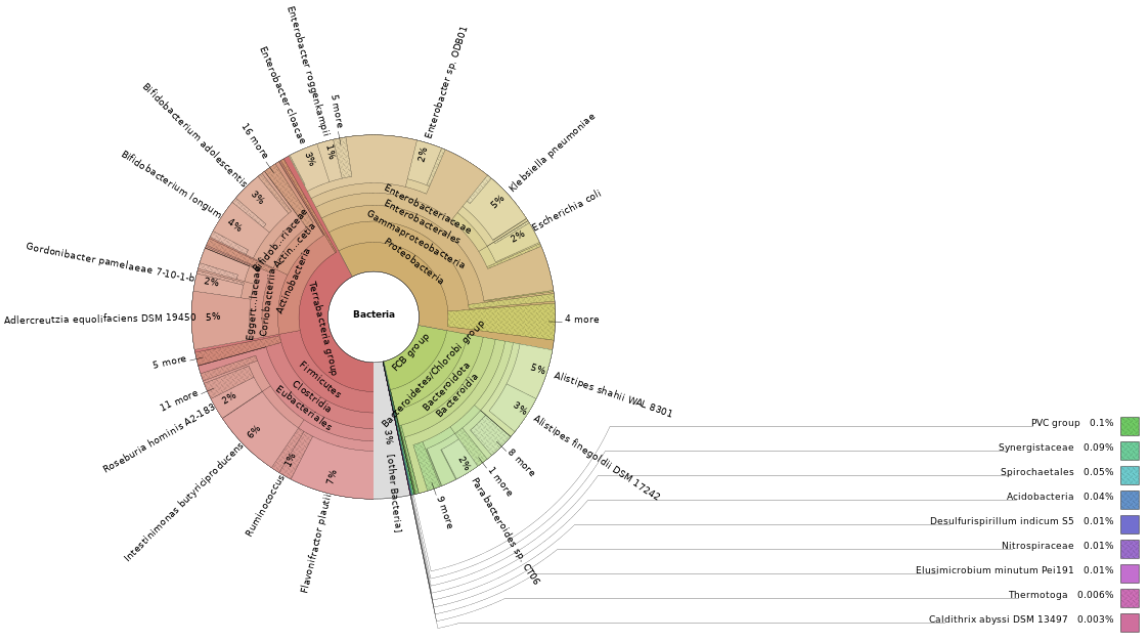


Figura 39: Perfil taxonómico de bacterias muestra hipertensión ERR1398076

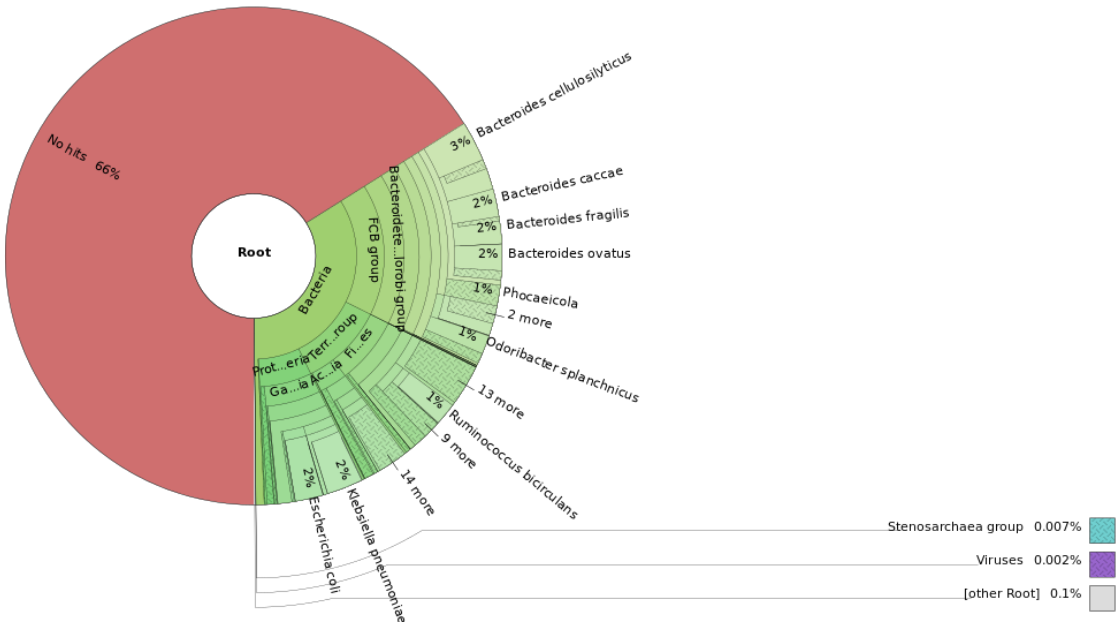


Figura 40: Perfil taxonómico muestra hipertensión ERR1398077

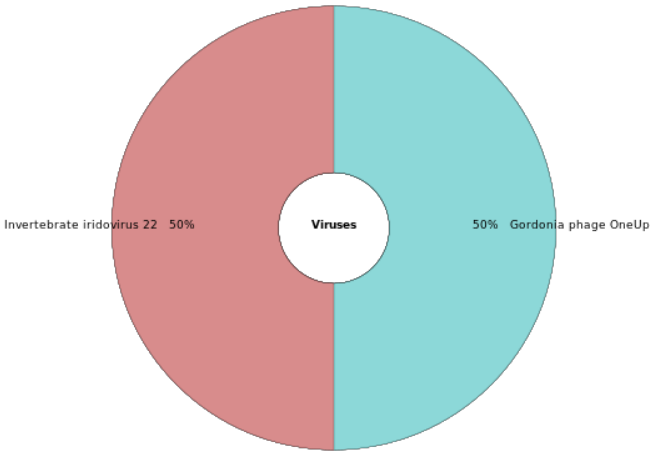


Figura 41: Perfil taxonómico de virus muestra hipertensión ERR1398077

Figura 42: Perfil taxonómico de archaea muestra hipertensión ERR1398077

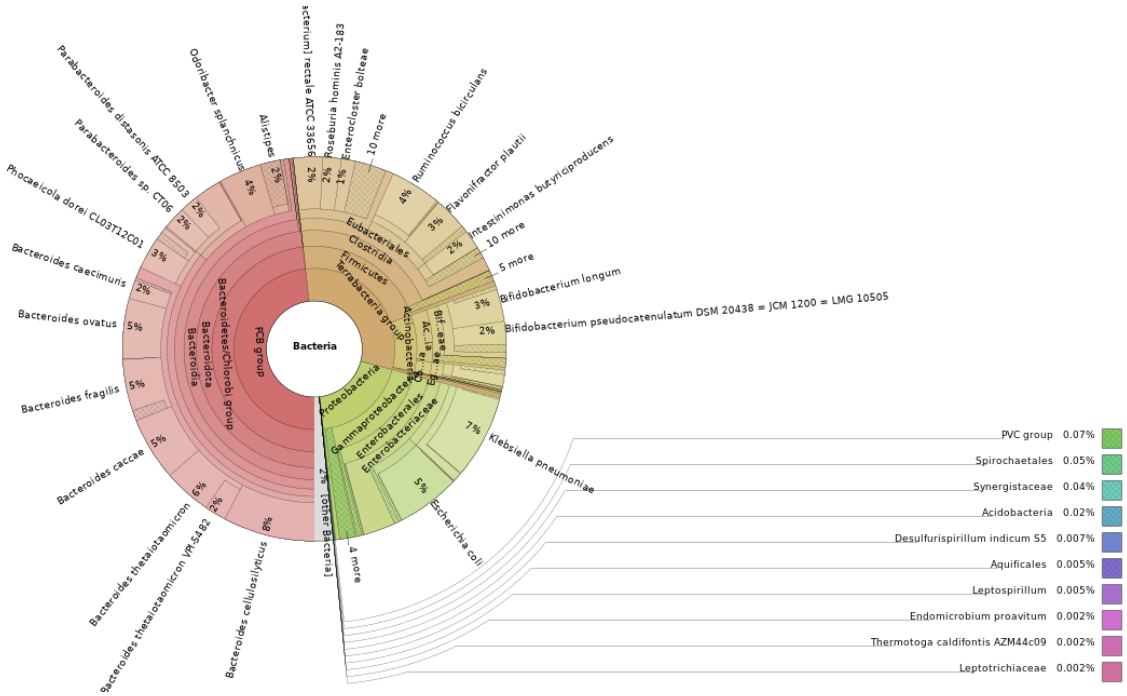


Figura 43: Perfil taxonómico de bacterias muestra hipertensión ERR1398077

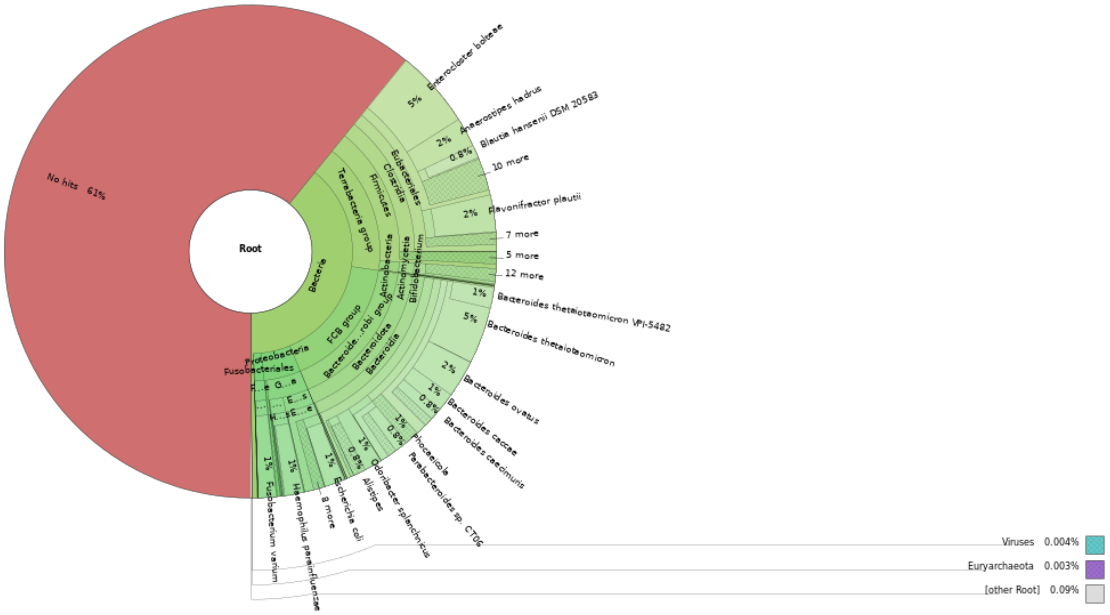


Figura 44: Perfil taxonómico muestra hipertensión ERR1398085

Figura 45: Perfil taxonómico de virus muestra hipertensión ERR1398085

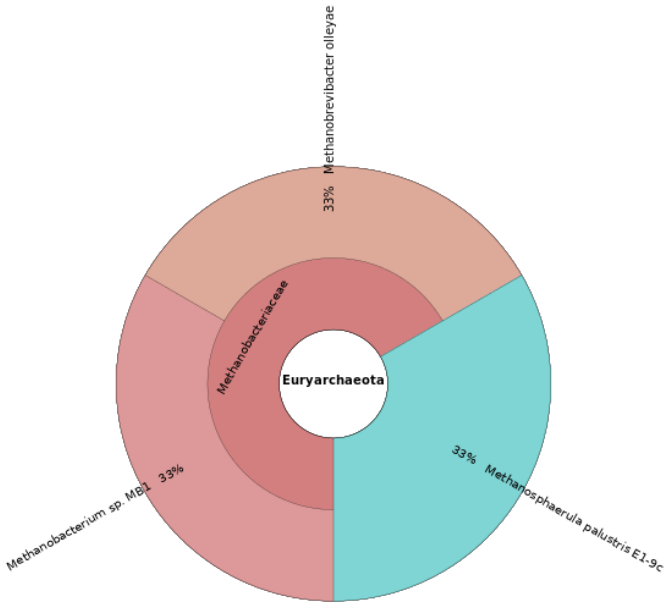


Figura 46: Perfil taxonómico de archaea muestra hipertensión ERR1398085

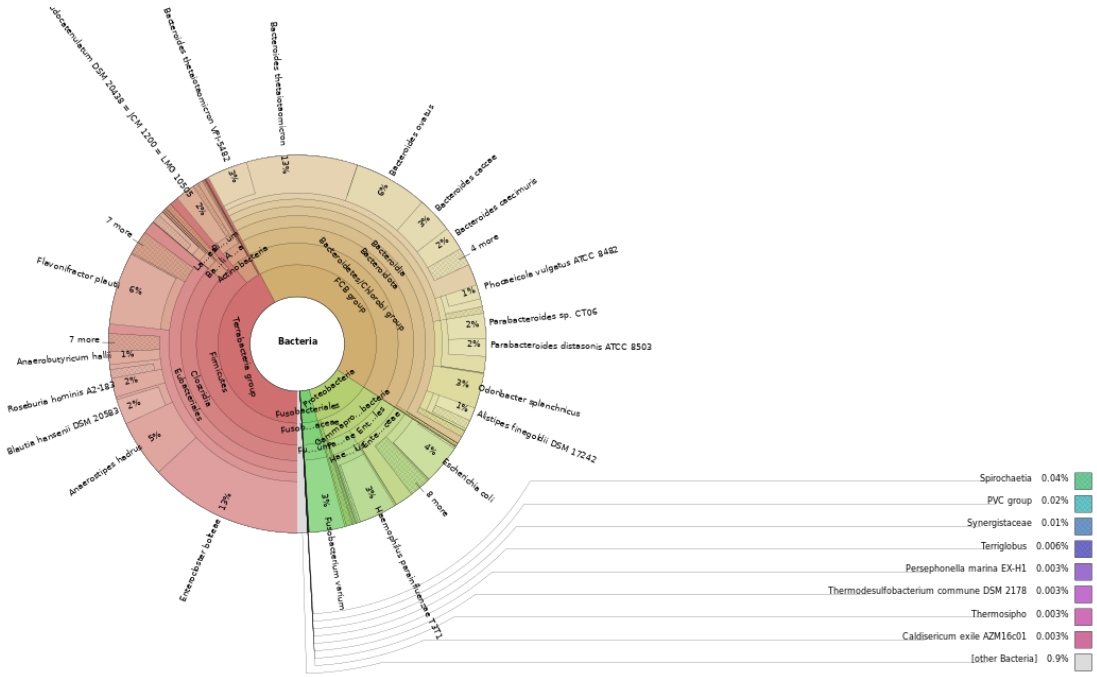


Figura 47: Perfil taxonómico de bacterias muestra hipertensión ERR1398085

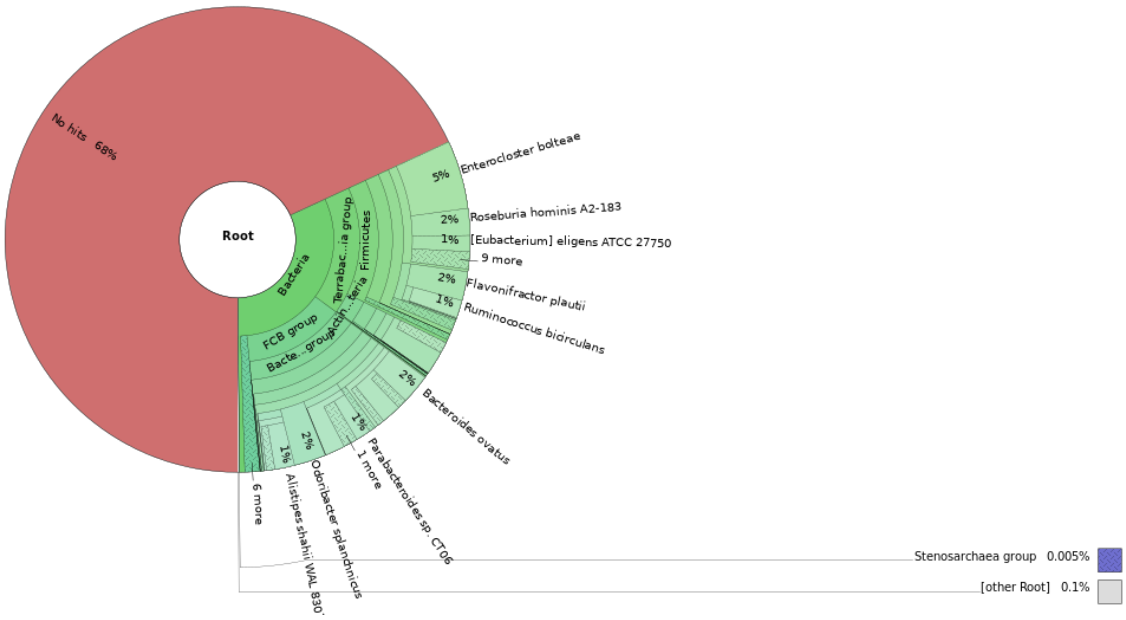


Figura 48: Perfil taxonómico muestra control saludable ERR1398129

Figura 49: Perfil taxonómico de archaea muestra control saludable ERR1398129

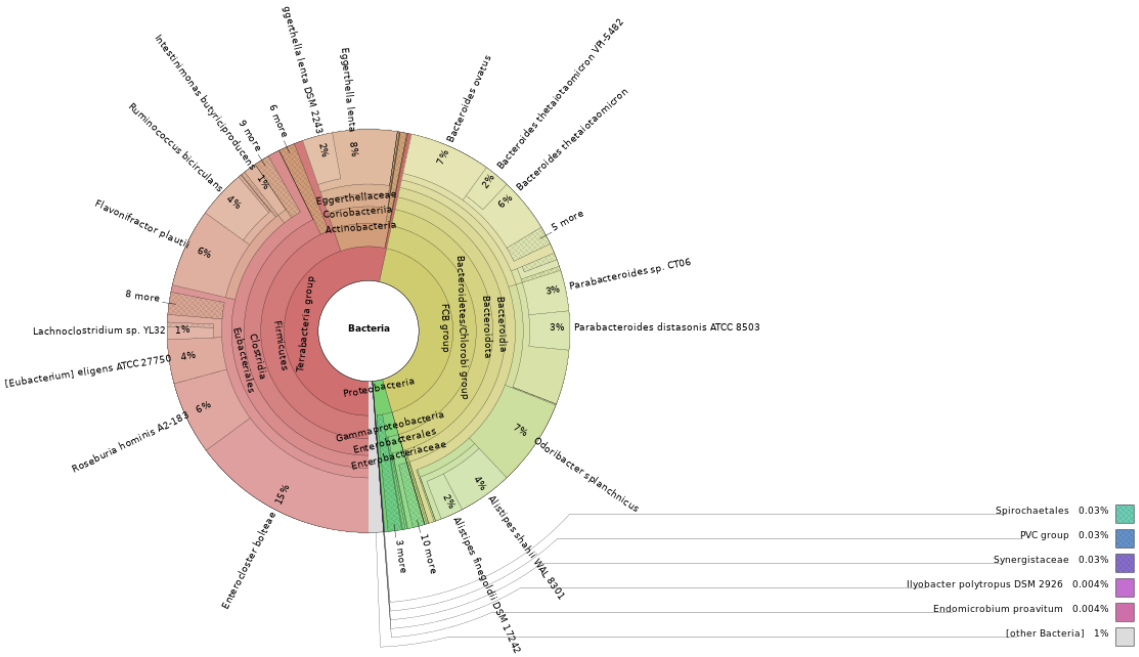


Figura 50: Perfil taxonómico de bacterias muestra control saludable ERR1398129

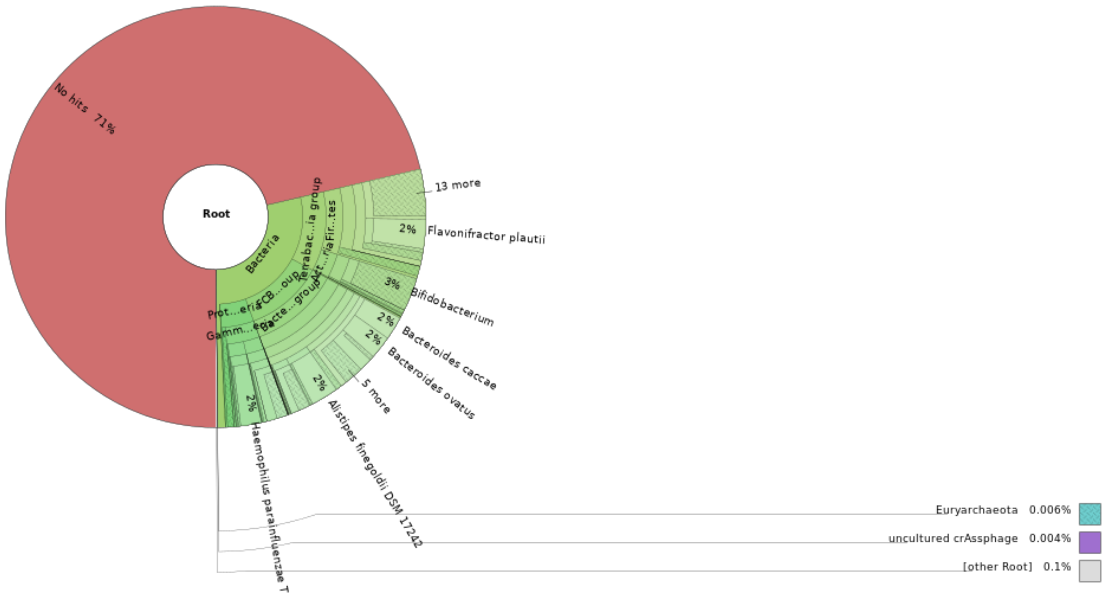


Figura 51: Perfil taxonómico muestra control saludable ERR1398078

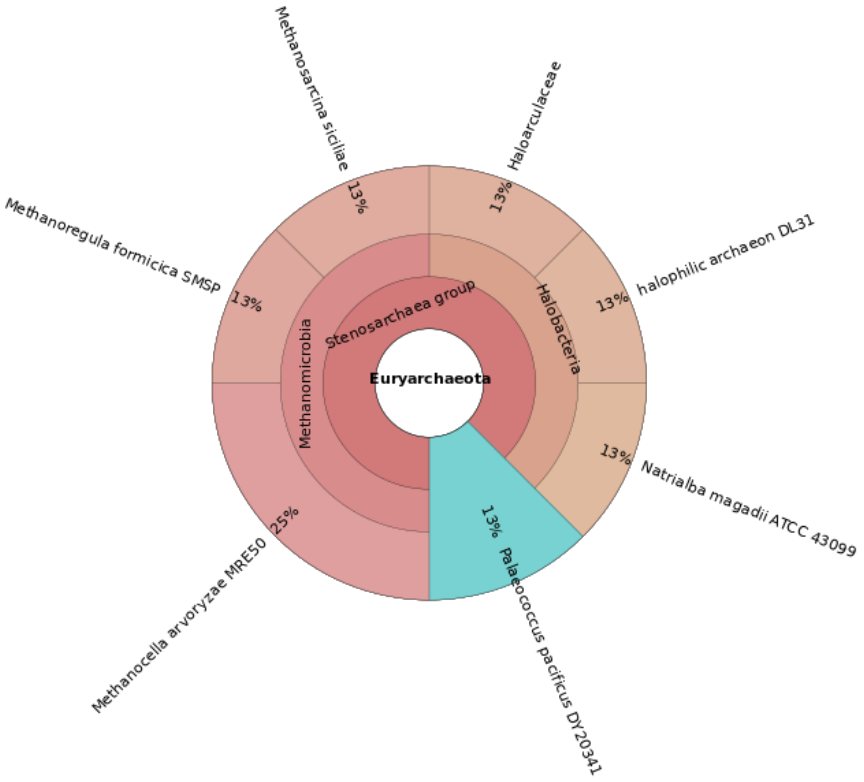


Figura 52: Perfil taxonómico de archaea muestra control saludable ERR1398078

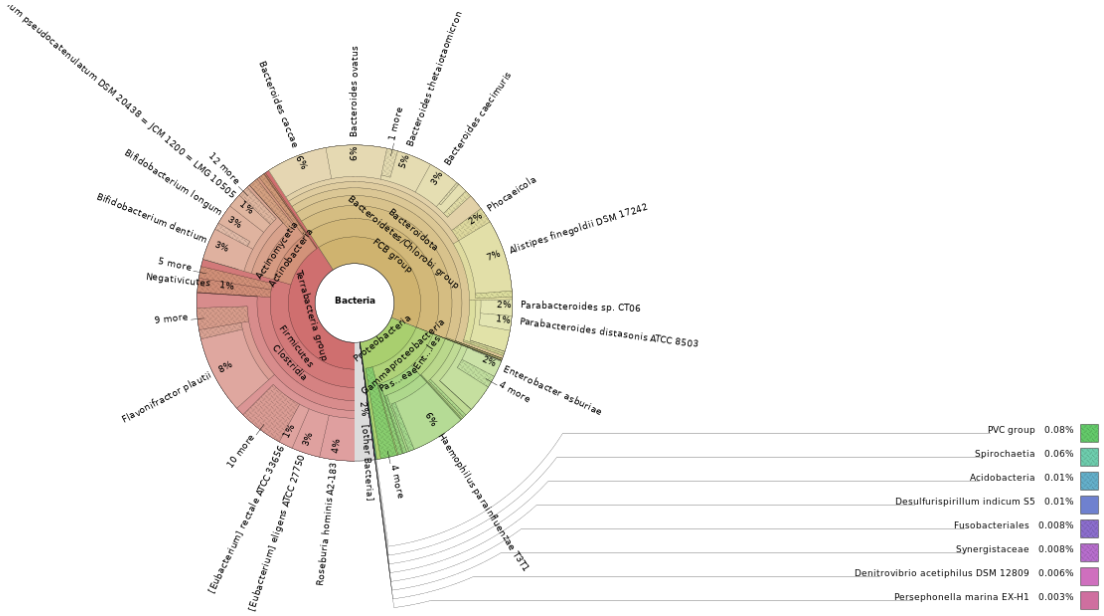


Figura 53: Perfil taxonómico de bacterias muestra control saludable ERR1398078

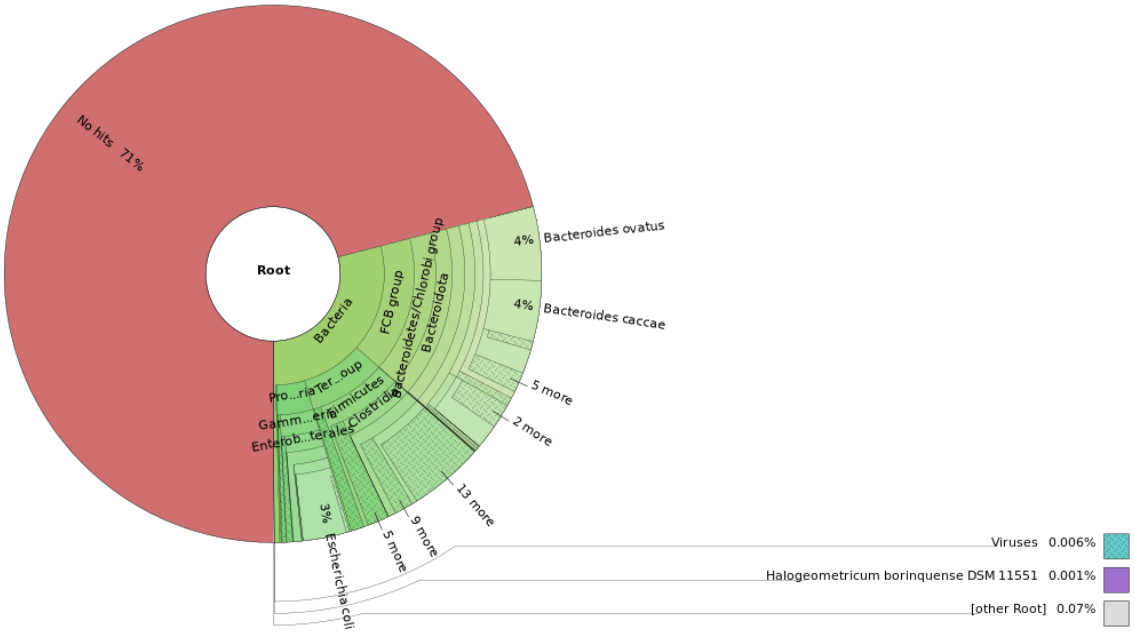


Figura 54: Perfil taxonómico muestra control saludable ERR1398257

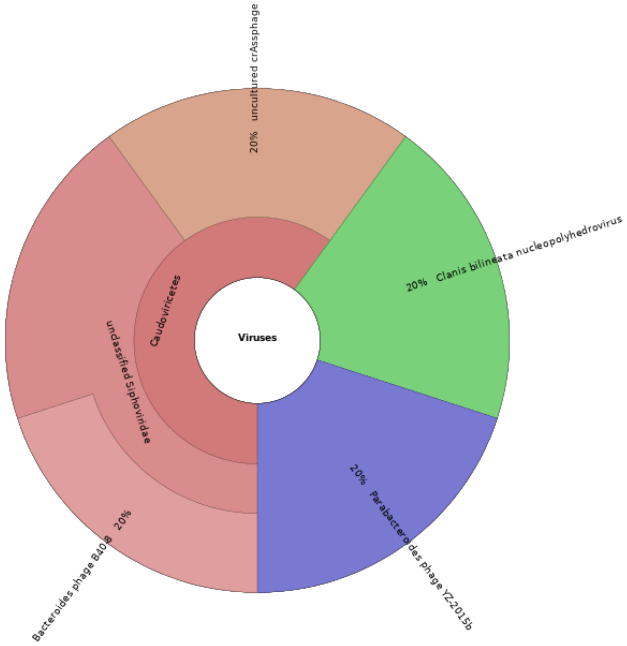


Figura 55: Perfil taxonómico de virus muestra control saludable ERR1398257

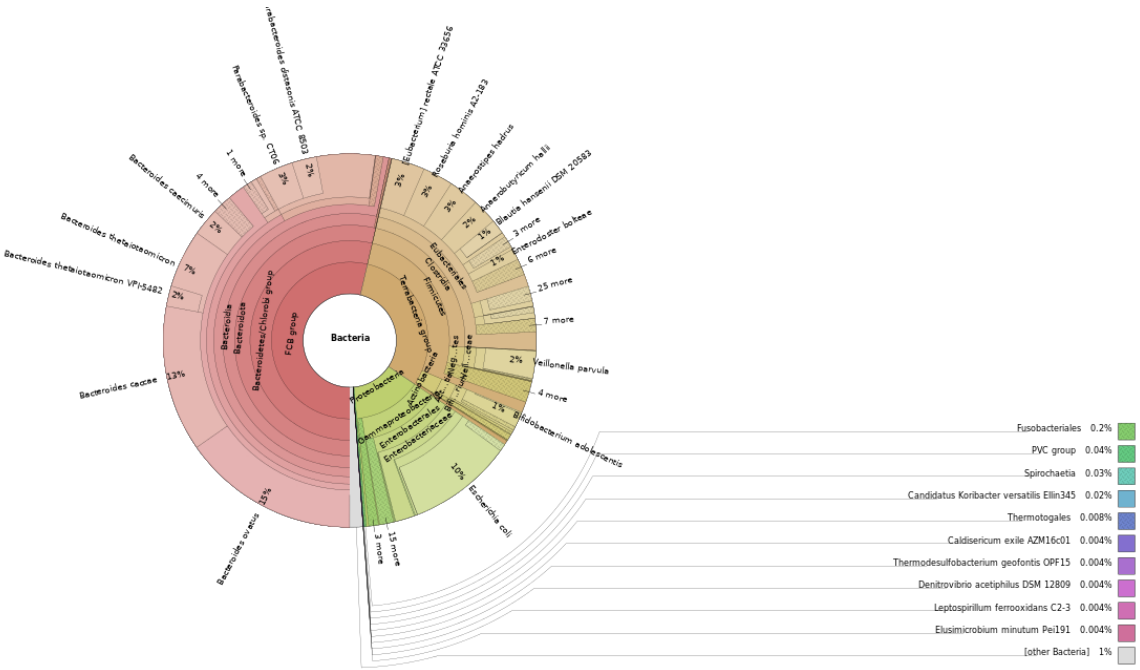


Figura 56: Perfil taxonómico de bacterias muestra control saludable ERR1398257

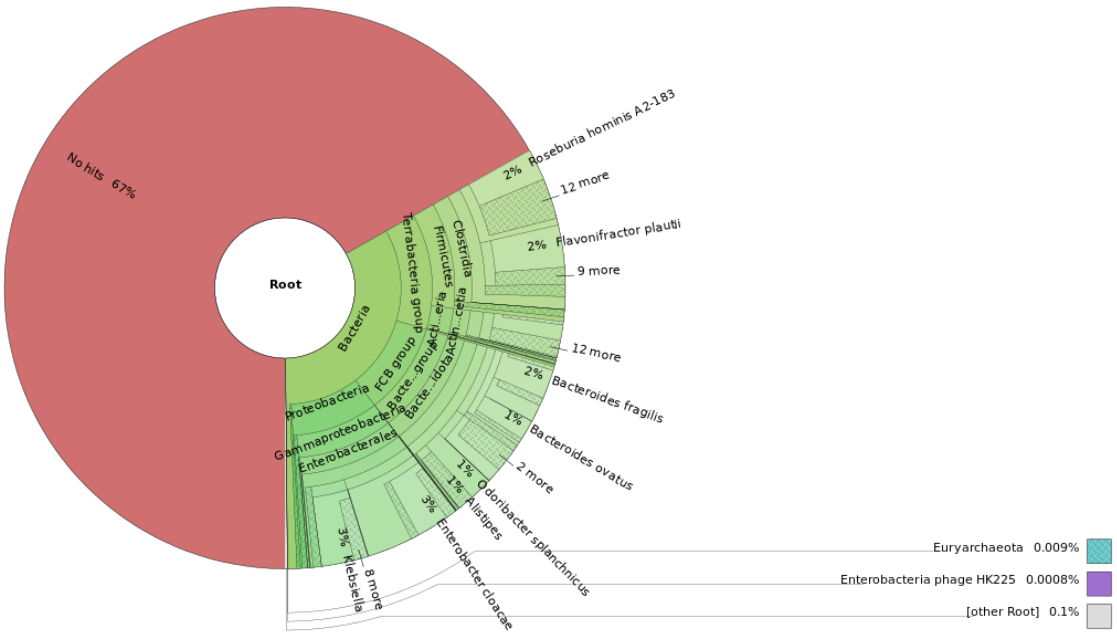


Figura 57: Perfil taxonómico muestra control saludable ERR1398089

Figura 58: Perfil taxonómico de archaea muestra control saludable ERR1398089

Figura 59: Perfil taxonómico de bacterias muestra control saludable ERR1398089

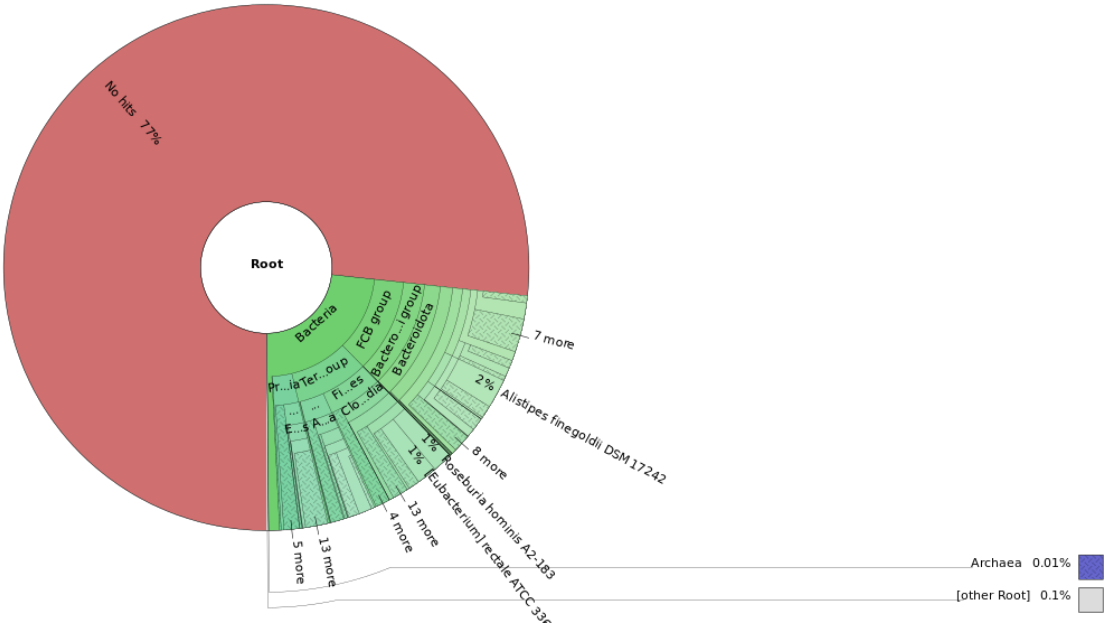


Figura 60: Perfil taxonómico muestra control saludable ERR1398206

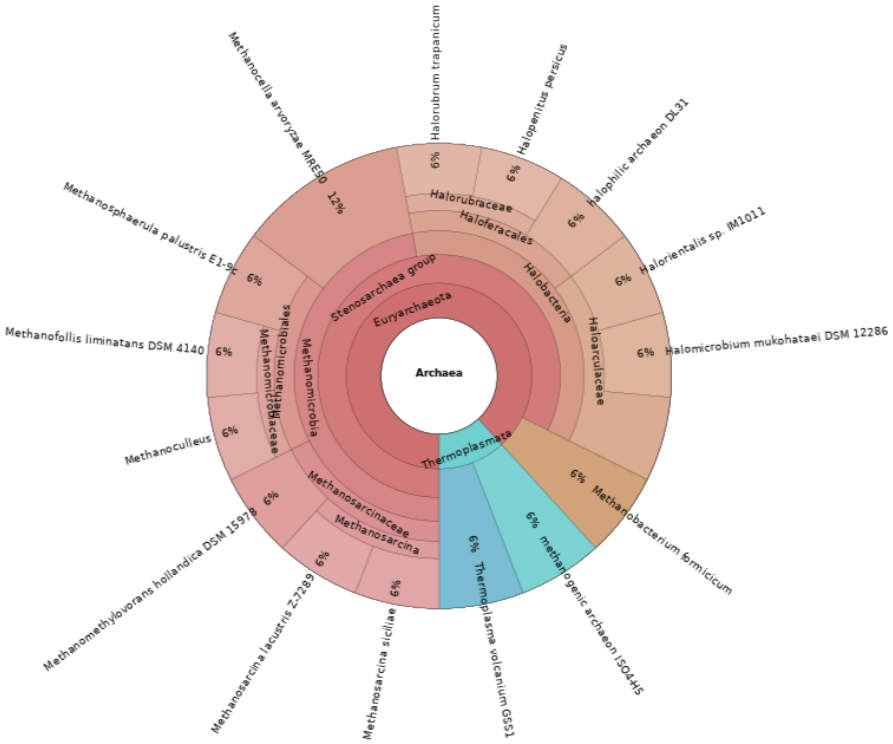


Figura 61: Perfil taxonómico de archaea muestra control saludable ERR1398206

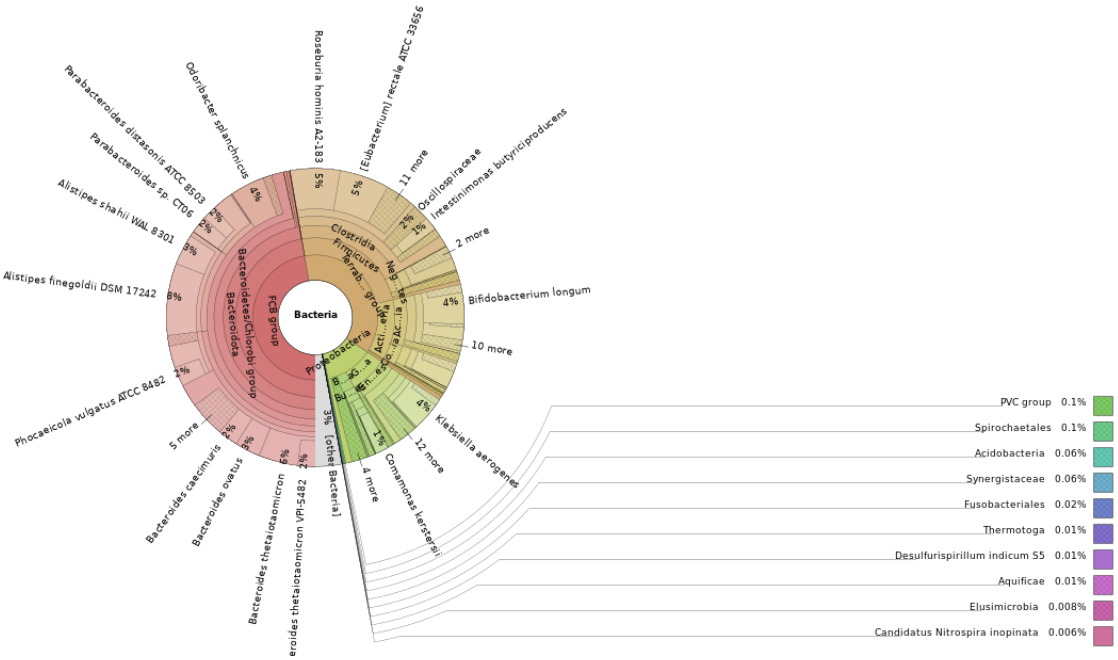


Figura 62: Perfil taxonómico de bacterias muestra control saludable ERR1398206

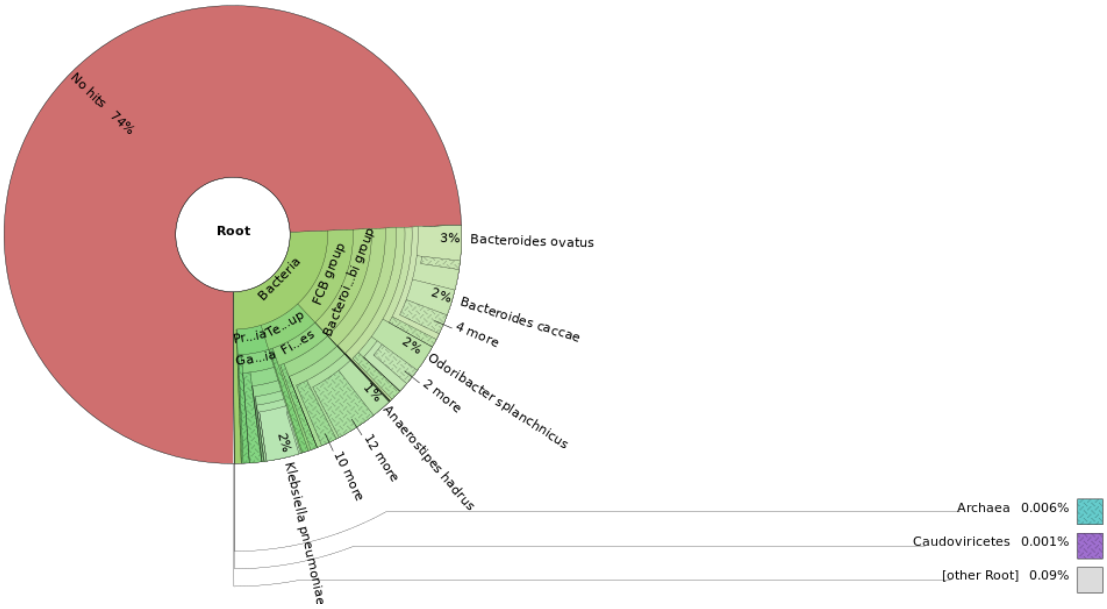


Figura 63: Perfil taxonómico muestra control saludable ERR1398263

Figura 64: Perfil taxonómico de archaea muestra control saludable ERR1398263

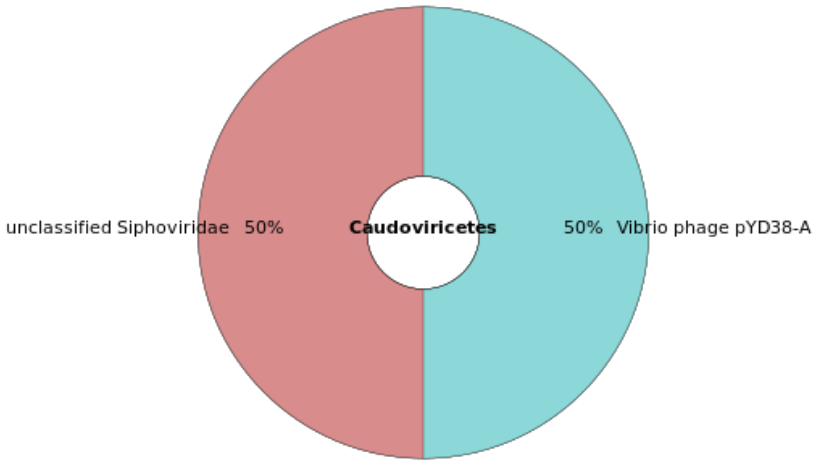


Figura 65: Perfil taxonómico de virus muestra control saludable ERR1398263

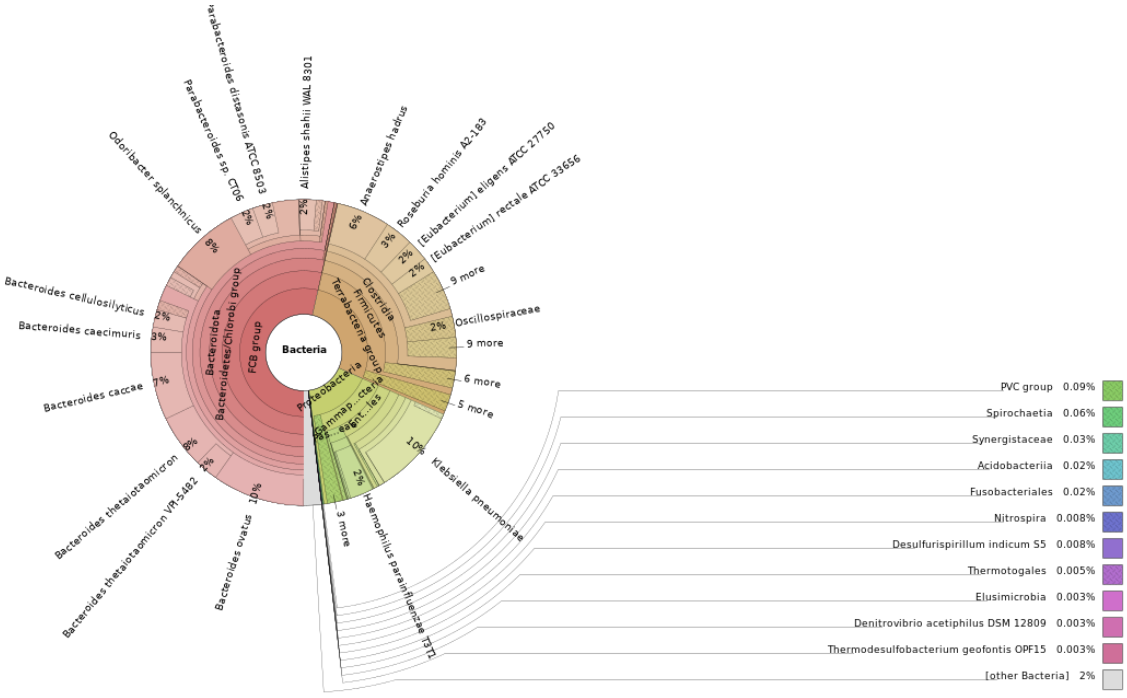


Figura 66: Perfil taxonómico de bacterias muestra control saludable ERR1398263

Glosario

Machine learning: Aprendizaje automático o aprendizaje de máquinas es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan. 1

Microbioma: Conjunto de microbios (bacterias, arqueas, virus, hongos y protistas) incluyendo sus genes y metabolitos, así como las condiciones ambientales que les rodean. 17

Microbiota: Conjunto de microorganismos que residen en nuestro cuerpo. 1, 17

ORF: Los Open Reading Frames o marcos de lectura abiertos son tramos de secuencia de ADN entre los codones de inicio y finalización. Se consideran como fragmentos de la secuencia de ADN que con frecuencia son parte de un gen, es decir, una secuencia que directamente codifica una proteína. 29

Pipeline: Una pipeline de bioinformática es una serie de algoritmos de software que procesan datos de secuenciación sin procesar y generan interpretaciones a partir de estos datos. 1

Subsistema: En el contexto biológico y específicamente en el estudio de proteínas, se hace referencia a los subsistemas como conjuntos de familias de proteínas que tienen una función similar. 32