Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Maestría en Explotación de Datos y Descubrimiento de Conocimiento

Interacciones del microbioma intestinal humano en diferentes comunidades europeas: un enfoque complementario desde el aprendizaje automático

Tesis presentada para optar al título de Magister de la Universidad de Buenos Aires en Explotación de Datos y el Descubrimiento de Conocimiento

Dra. Valeria Laura Burgos

Directora de Tesis: Dra. María Laura Fernández

Co-Director: Dr. Marcelo Risk

Tesis desarrollada en el Instituto de Medicina Traslacional e Ingeniería Biomédica (IMTIB)

CONICET - Instituto Universitario Hospital Italiano - Hospital Italiano de Buenos Aires

Buenos Aires, 19 de Diciembre de 2023

Dedicatoria

Esta Tesis está dedicada a las memorias de Pablo F. Argibay y Robert J. Barros.

Agradecimientos

A mi directora, María Laura Fernández, por acompañarme y aconsejarme en las diversas fases por las que atravesó la realización de esta Tesis.

A mi co-director, Marcelo Risk, por su constante ayuda y estímulo para crecer profesionalmente.

A Pilar Ávila, Amalia Guaymás y a mis compañeros de cursada que hicieron que la maestría haya sido una gran experiencia.

A mi hija, Sophia Barros.

A Eduardo San Román.

A mis amigos Guido Guzmán y Nicolás Quiroz.

Resumen

El tracto gastrointestinal humano está colonizado por una abundancia de comunidades de bacterias, virus, hongos y arqueas, conviviendo en equilibrio entre sí y con el hospedador humano. Son millones de microorganismos que forman la denominada microbiota intestinal, con roles muy importantes en el bienestar, mantenimiento de la salud y aparición de varias enfermedades que van desde el síndrome de colon irritable, cáncer hasta depresión.

En la presente Tesis se propone un enfoque de análisis desde la minería de datos (*data mining*) sobre un conjunto de datos de microbioma intestinal provenientes de individuos europeos, utilizando algoritmos de aprendizaje automático (*machine learning*) que permitan identificar relaciones con relevancia biomédica entre las variables demográficas de los sujetos de estudio y algún patrón de abundancia de bacterias.

El estudio se realizó sobre varias especies de bacterias pertenecientes a los dos grupos (*phyla*) de mayor abundancia en la microbiota intestinal humana, tal como *Bacteroidetes* y *Firmicutes*, teniendo en cuenta además información sobre la edad, índice de masa corporal, nacionalidad y sexo de los individuos. Ambos subconjuntos de abundancias de bacterias junto a las variables de metadatos fueron sujetos a un análisis exploratorio, usando posteriormente análisis de reducción de la dimensionalidad y algoritmos de agrupamiento (*clustering*) con el fin de caracterizar la presencia de agrupamientos naturales en los datos, además de evaluar la variabilidad de los mismos.

Se usó también el algoritmo de mapas auto-organizados cuya visualización representa una valiosa herramienta de análisis ya que permite integrar información multivariada a través de diferentes planos componentes.

Por último, para estudiar la diversidad de las poblaciones de bacterias en estudio y su

asociación con datos demográficos de los individuos, se usó un métodos de ensamble como *random forest* para modelar la importancia de las variables sobre la diversidad biológica de las comunidades.

Los resultados muestran que las variables más importantes para explicar la diversidad bacteriana serían la edad y el índice de masa corporal. Curiosamente, varias especies de bacterias conocidas por estar asociados con la dieta y la obesidad, fueron identificadas como características relevantes. El análisis topológico de los mapas auto-organizados identificó ciertos grupos de nodos de características similares en los metadatos de los sujetos y grupos de bacterias.

Se concluye que los resultados representan un enfoque que podría incorporarse en futuros estudios, como un potencial complemento en reportes de salud para ayudar a los profesionales de la salud a personalizar el tratamiento del paciente o como apoyo para la toma de decisiones.

Palabras clave: microbiota intestinal, mapas auto-organizados, agrupamiento, *random forest*, medicina personalizada, bioinformática.

Interactions of the human gut microbiome in various European communities: a complementary approach using machine learning

The human gastrointestinal tract is colonized by abundant communities of bacteria, viruses, fungi, and archaea, living in balance with each other and the human host. Millions of microorganisms make up the so-called gut microbiota, with significant roles in the well-being, and health maintenance as well as the appearance of various diseases ranging from irritable bowel syndrome, and cancer to depression.

This thesis proposes a data mining approach to analyze a set of gut microbiota data of individuals from several European regions, using machine learning algorithms to identify biomedically relevant relationships between demographic and biomedical variables of the subjects and patterns of abundance of bacteria.

The study focused on the two most abundant human gut microbiota groups (phyla), *Bacteroidetes* and *Firmicutes*. Both subsets of bacterial abundances together with the metadata variables were subjected to an exploratory analysis, subsequently using dimensionality reduction techniques and clustering algorithms to characterize the presence of natural clusters in the data, in addition to evaluating their variability.

The self-organizing map algorithm was also used, whose visualization of outcome represents a valuable analysis tool since it integrates multivariate information through different component

planes.

Finally, to evaluate the relevance of the variables on the biological diversity of the microbial communities, an ensemble-based method such as random forest was used.

Results showed that age and body mass index were important in explaining bacteria diversity. Interestingly, several species of bacteria associated with diet and obesity were also identified as relevant features. The topological analysis of self-organizing maps identified specific groups of nodes with similarities in subject metadata and gut bacteria.

This study represents an approach that could be considered in future studies as a potential complement in health reports to help healthcare professionals personalize patient treatment or support decision-making.

Keywords: gut microbiota, self-organizing maps, clustering, random forest, personalized medicine, bioinformatics.

Lista de abreviaturas

ADN Ácido Desoxiribonucleico

ARN Ácido Ribonubleico

BMI *Body Mass Index* – Índice de Masa Corporal

GI Gastrointestinal

HGP *Human Genome Project* – Proyecto Genoma Humano

MAE *Mean Absolute Error* – Error Absoluto Medio

ML *Machine Learning* – Aprendizaje Automático

NA *Not Available* – Valor faltante

NGS *Next Generation Sequencing* – Secuenciación de Nueva Generación

PC *Principal Component* – Componente Principal

PCA *Principal Component Analysis* – Análisis de Componentes Principales

RF *Random Forest*

SOM *Self-Organizing Maps* – Mapas Auto-Organizados

UMAP *Uniform Manifold Approximation and Projection*

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Los seres vivos están clasificados dentro de tres grandes dominios: Bacterias, Arqueas y Eucariotas, tal cual indica el modelo conocido como árbol de la vida, que explora la evolución y las relaciones de todos los organismos (Figura 1.1). Una rama del árbol representa a los eucariotas, que incluye animales, plantas, hongos, algas y protozoos. La segunda rama describe a las bacterias; microorganismos unicelulares (algunas de los cuales serán detallados más adelante). Y la tercer rama abarca a arqueas, microorganismos presentes en ambientes considerados extremos para la vida (en términos de temperatura, pH, salinidad y anaerobiosis) y no extremos, como océanos, sedimentos de agua dulce y suelos [Chaban et al., 2006].

Desde organismos simples compuestos por una única célula (como arqueas, bacterias, protozoos, algunas algas y hongos) hasta los complejos organismos multicelulares (tal como los humanos), las instrucciones necesarias para la creación de biomoléculas y mantenimiento de funciones celulares durante la vida de un organismo están codificadas en la molécula de ácido desoxiribonucleico (ADN). Esta molécula está formada por un par de cadenas largas pareadas, conformada por cuatro tipos de monómeros (Adenina [A], Timina [T], Citosina [C] y Guanina [G]) unidos en una secuencia lineal. El ADN contiene las instrucciones biológicas para producir, por ejemplo, proteínas (un tipo de biomolécula con diversos roles en la célula, como transporte, comunicación, catálisis, formación de estructuras, almacenamiento, etc). También se incluyen instrucciones para mecanismos de reproducción, funciones especializadas, interacción con otras células y respuestas a señales del entorno, etc. En definitiva, las instrucciones para mantener la vida y hacer única a cada especie.

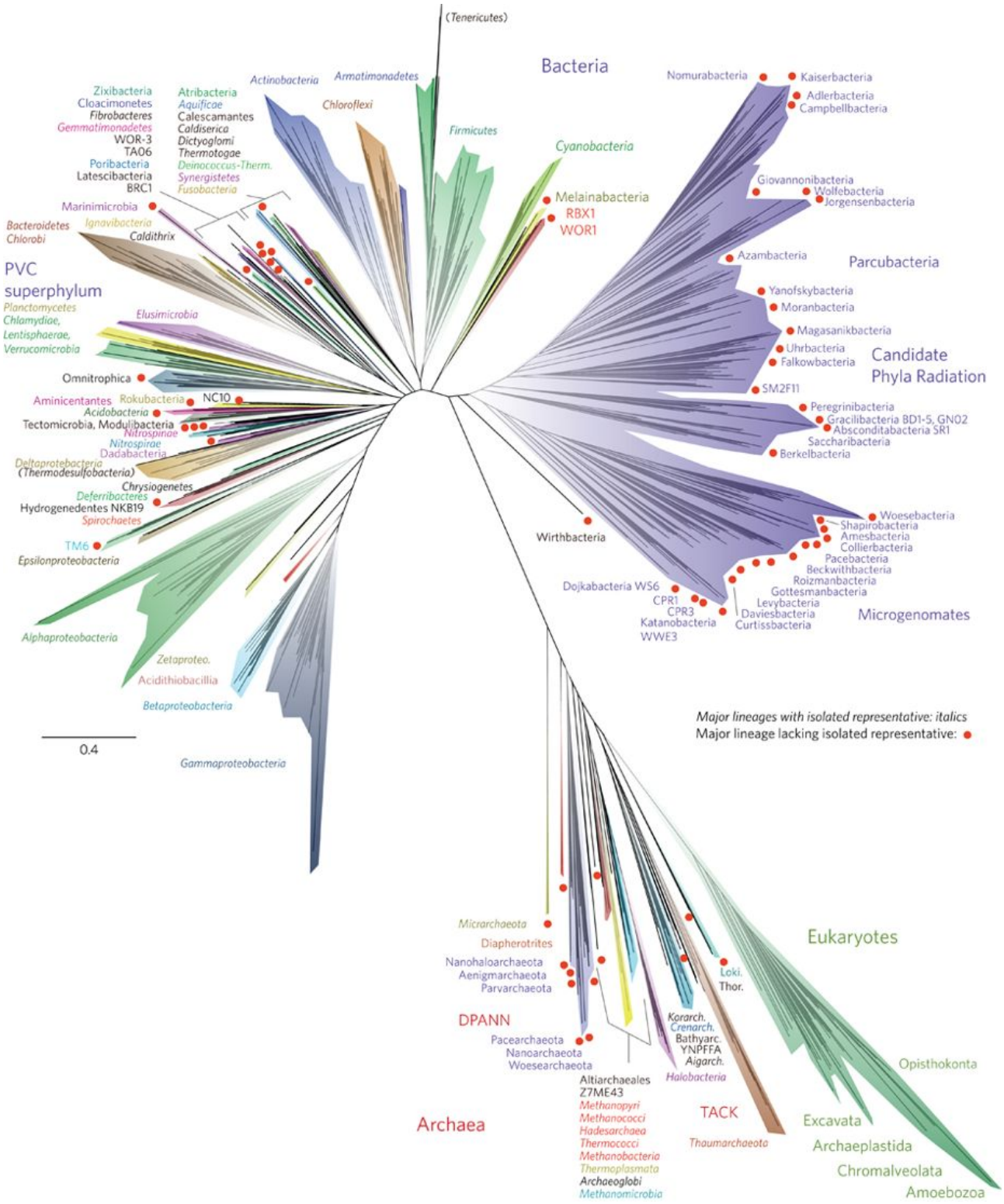

Figura 1.1: Esquema actual del árbol de la vida, comprendiendo la diversidad total representada por genomas secuenciados. El árbol incluye 92 nombres de divisiones (*phyla*) de Bacterias, 26 *phyla* de Arquea y los 5 supergrupos de Eucariotas. Adaptado de [Hug et al., 2016].

Estas instrucciones en forma de combinaciones de los cuatro monómeros A, T, C y G (conocidos como nucleótidos) están almacenadas en forma de **gen**, definido como una unidad

funcional de ADN que codifica moléculas cuyas funciones dirigen todas las actividades de la vida [Alberts et al., 2002]. Los genes funcionan como elementos contenedores de información que determinan las características de una especie en su totalidad y de los individuos dentro de la misma.

Teniendo en cuenta que esta Tesis estudia la abundancia de bacterias intestinales y su interacción con el hospedador humano, es necesario describir muy brevemente algunas diferencias entre ambos tipos de organismos desde el punto de vista de la organización celular de la molécula de ADN.

En el interior celular de las bacterias, el ADN se encuentra libre sin un compartimento nuclear definido que lo separe del resto de los componentes intracelulares y existe como molécula desnuda, lo que quiere decir que no está acompañado en su estructura por ningún complejo proteico u otras macromoléculas. Estas características son típicas de organismos clasificados como procariotas, a los cuales pertenecen bacterias y arqueas (Figura 1.2).

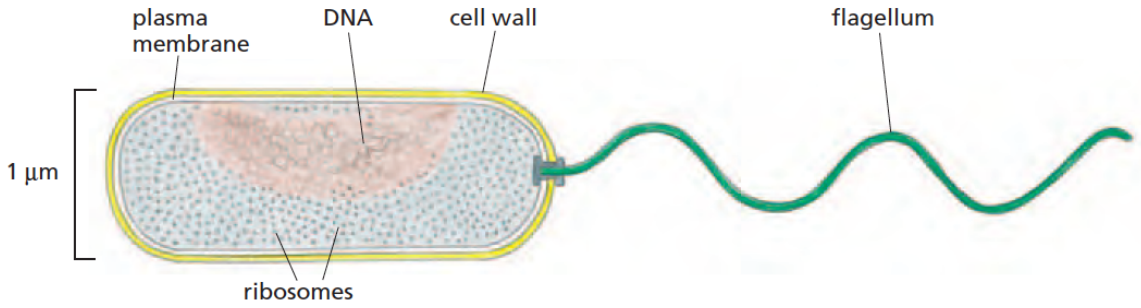

Figura 1.2: Esquema de una célula procariota. En este ejemplo, la estructura de la bacteria *Vibrio cholerae* mostrando su organización interna simple. Como muchas otras especies, *V. cholerae* tiene un flagelo que rota como propulsor de movimiento de la célula. Adaptado de [Alberts et al., 2002].

En comparación, en el interior de una célula eucariota (por ejemplo, la humana) la molécula de ADN sí está en un compartimento definido llamado núcleo, compuesto por una envoltura nuclear que permite la comunicación y tránsito de moléculas entre el espacio intracelular y el interior del núcleo (Figura 1.3). Además, el ADN está acompañado en su estructura por un complejo proteico formando cromosomas (Figura 1.4). Estas son características de organismos clasificados como Eucariotas (a los cuales pertenece el ser humano).

En las diferencias mencionadas a nivel de organización del ADN subyacen a su vez diferentes y complejos mecanismos de regulación de expresión de genes en un producto funcional. Sin

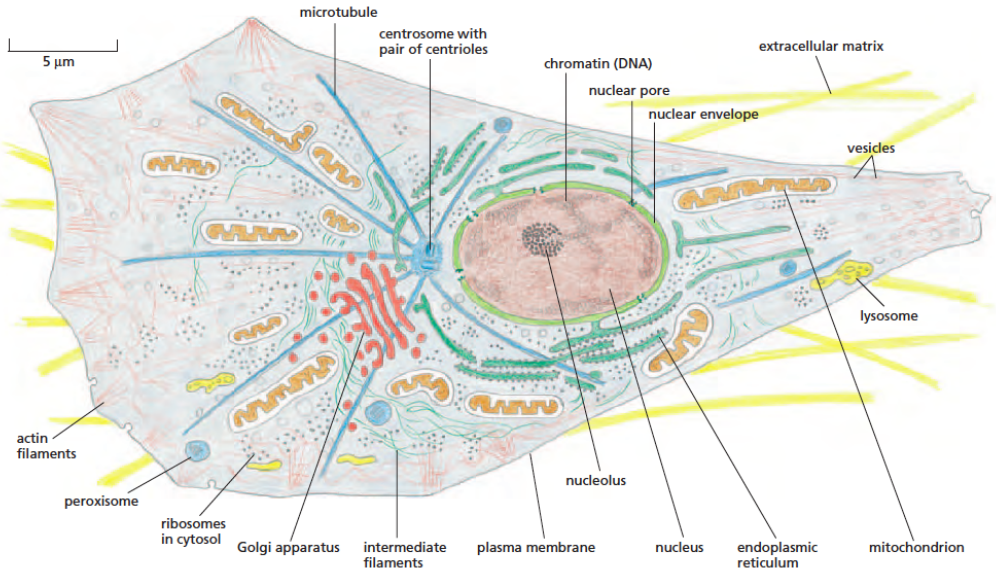

Figura 1.3: Esquema de una célula eucariota. El esquema corresponde a una célula animal, pero casi todos los componentes se encuentran también en plantas y hongos, además de eucariotas unicelulares como levaduras y protozoos. Adaptado de [Alberts et al., 2002].

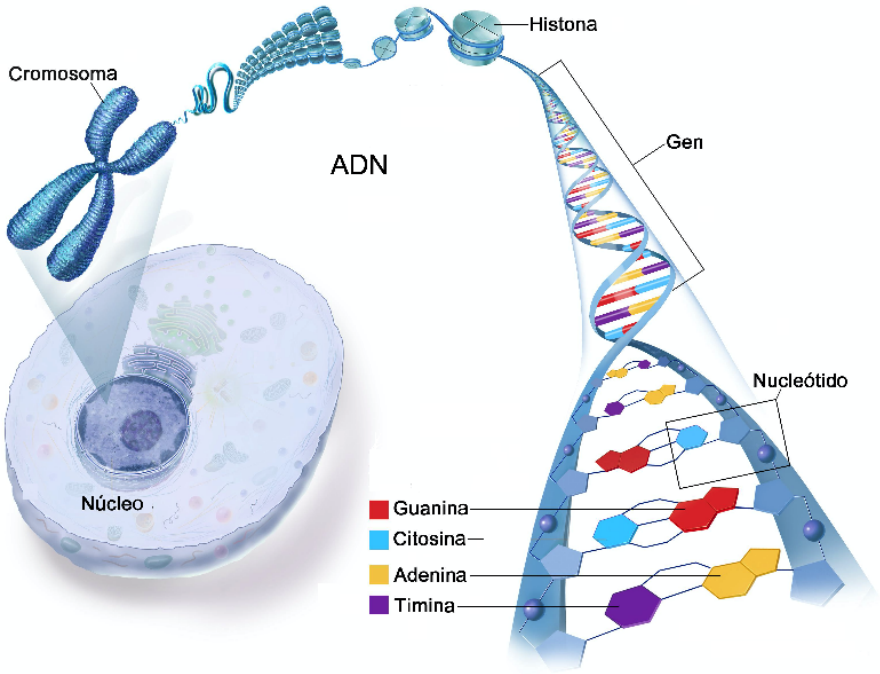

Figura 1.4: Estructura del ADN en una célula eucariota. El ADN se encuentra en el interior del núcleo de una célula, donde forma los cromosomas. Los cromosomas contienen proteínas llamadas histonas sobre las cuales se enrolla el ADN. Adaptado de www.cancer.gov.

embargo, en todas las células de organismos vivos (tanto eucariotas como procariotas) rige un principio (en gran parte fundamental, aunque existen excepciones) denominado *dogma central de la biologia molecular*. El mismo se refiere al flujo principal de la información genética que va

desde el ADN (que contiene la información necesaria para la producción), al ácido ribonucleico (ARN, que actúa como un intermediario) y a las proteínas. Cuando la célula necesita una proteína determinada, la secuencia de nucleótidos del gen (ubicado en el ADN) es copiada primero en una molécula de ARN (proceso denominado transcripción). Estas copias de ARN son usadas directamente como moldes para dirigir la síntesis de la proteína en cuestión (proceso conocido como traducción). Por lo tanto, el flujo de información genética en las células es usualmente desde el ADN a ARN a proteína (Figura 1.5). En todas las células, la expresión de genes individuales es un proceso altamente regulado: una célula no produce el repertorio total de posibles proteínas a toda velocidad todo el tiempo sino que ajusta la tasa de transcripción y traducción de diferentes genes de manera independiente, de acuerdo a la necesidad [Alberts et al., 2002].

Esta descripción muy general sobre conceptos fundamentales de biología molecular es necesaria para brindar un contexto a términos como 'gen' y 'expresión génica'. Los datos de microbiota intestinal analizados en esta Tesis fueron obtenidos mediante *microarrays* filogenéticos, una técnica de biología molecular descripta en secciones posteriores.

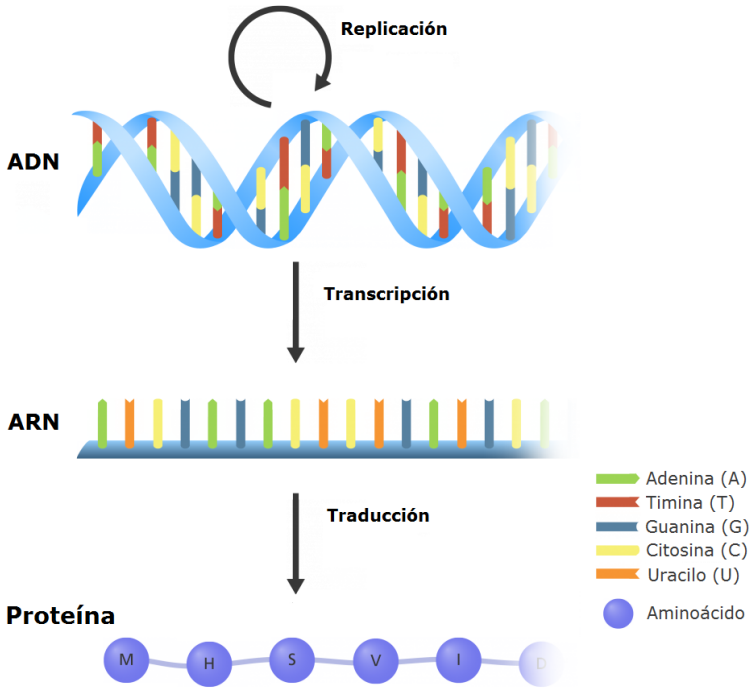

Figura 1.5: Flujo de la información genética contenida en el ADN y su relación con las proteínas.

1.1. Microbioma

Una gran diversidad de microorganismos abundan en las superficies e interior del cuerpo humano, desde la piel, cavidades orales y genitales, tracto respiratorio y el sistema gastrointestinal. Los microorganismos incluyen a bacterias, arqueas, virus y hongos, los que en conjunto constituyen la **microbiota** [Lloyd-Price et al., 2016]. Las comunidades de microorganismos se distribuyen en diferentes composiciones dependiendo del sitio en el hospedador (Figura 1.6).

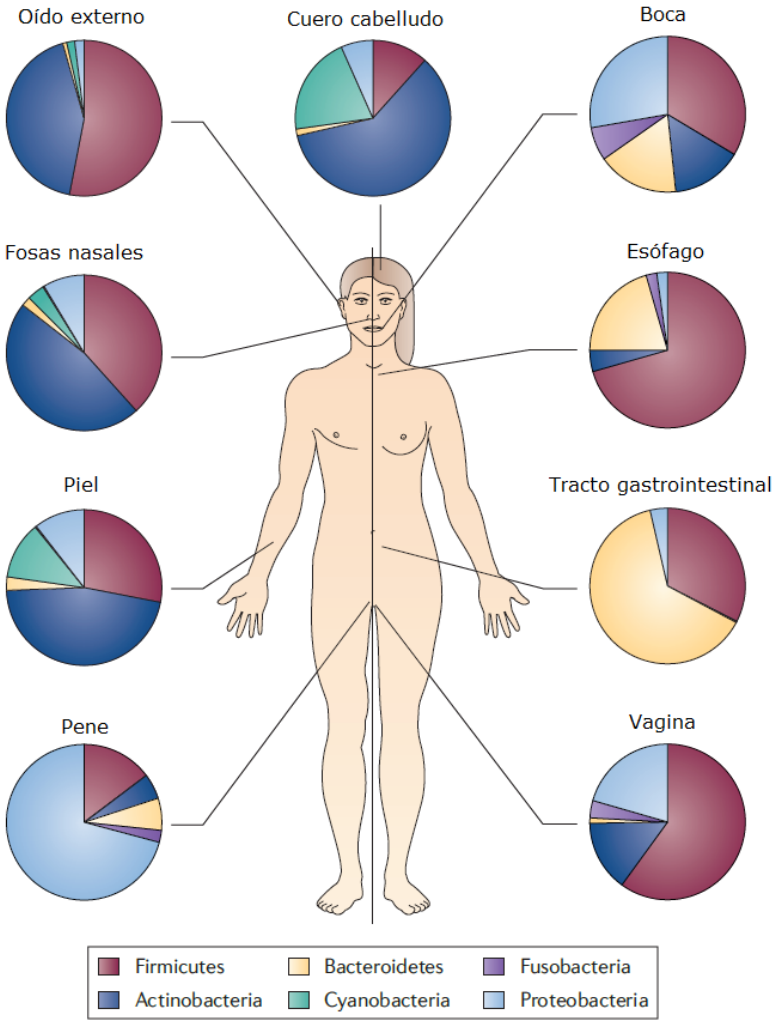

Figura 1.6: Composición de comunidades de microorganismos en diferentes partes del cuerpo en un individuo sano. Se indican las abundancias relativas de los seis grupos dominantes de bacterias. Adaptado de [Spor et al., 2011].

El creciente entendimiento de la microbiota humana ha significado un cambio de paradigma en las ciencias biomédicas en los últimos años, dado que los microorganismos eran considerados principalmente como patógenos [Sansonetti, 2011]. De esta manera, se tiene cada vez más conocimiento de las complejas y dinámicas interacciones de convivencia entre la mi-

crobiota y el hospedador, es decir, relaciones biológicas simbióticas. La simbiosis se refiere a la relación cercana entre dos organismos diferentes y abarca al comensalismo, mutualismo y parasitismo. Las bacterias tienen una larga historia de simbiosis: en el comensalismo, la relación entre ambos organismos es neutra ya que ninguna de las partes saca provecho de la otra ni se provocan perjuicio mutuo alguno. En el mutualismo, ambos organismos sacan provecho de la relación. Y en el parasitismo, un organismo se aprovecha del otro mientras lo daña [European Society Neurogastroenterology and Motility, 2020].

Existe una variedad de interacciones simbióticas entre el humano y su microbiota en el contexto del mantenimiento de la homeostasis. Por ejemplo, en el ecosistema vaginal de mujeres en edad reproductiva, las bacterias pertenecientes principalmente al género *Lactobacillus* parecen tener un papel fundamental en el sistema de defensa del nicho: se hipotetiza que mediante la producción de ácido láctico, producción de compuestos bacteriostáticos y bactericidas o por exclusión competitiva se logra disminuir el pH, lo cual actúa como una protección contra la colonización excesiva por microorganismos patógenos, tal como la levadura *Candida albicans* [Tortora et al., 2007],[Ravel et al., 2011]. La composición de la microbiota en cualquier nicho del cuerpo puede ser perturbada ante la presencia de factores externos, provocando una disrupción en las relaciones simbióticas entre el hospedador y los microorganismos asociados. Este desequilibrio persistente, conocido como **disbiosis**, está asociado a enfermedades [Floch et al., 2016]. Por ejemplo, la composición de la microbiota intestinal en la cirrosis (una enfermedad hepática) está caracterizada por una disminución de bacterias potencialmente beneficiosas, tal como *Ruminococcaceae*, y un aumento de bacterias patogénicas, como *Staphylococcaeae*. Tales cambios están asociados al progreso de la enfermedad [Bajaj et al., 2014].

La presencia de diversas comunidades de bacterias en el cuerpo humano representa una importante expansión funcional del genoma hospedador, es decir, un aporte de características que se suman a las codificadas en el ADN humano (revisado en [Gill et al., 2006]). Por ejemplo, el genoma bacteriano aporta varios tipos de enzimas que participan en la digestión de ciertos compuestos que no pueden ser digeridos en el estómago, como componentes vegetales o ciertos azúcares [Larsbrink et al., 2014]. De esta manera, existe una contribución relevante al metabolismo y fisiología del hospedador por parte de algunas bacterias. El genoma del total de

los microorganismos que habitan los diferentes nichos de un organismo vertebrado es lo que se conoce como **microbioma**.

La microbiota es una comunidad muy dinámica y como tal, las tasas de crecimiento y supervivencia de sus poblaciones componentes pueden fluctuar. Se considera que la microbiota presenta dos características muy importantes: plasticidad y resiliencia [Ruggles et al., 2018, Liu et al., 2019]. La resiliencia se pone en evidencia cuando nos exponemos a agentes estresores temporales, tales como cambios en la dieta o el consumo de antibióticos: la microbiota responde adaptándose y recuperando el estado basal una vez que cesa el estresor [Greenhalgh et al., 2016]. Sin embargo, también presenta la característica de flexibilidad, porque en su constitución influyen varios factores, tales como la genética del hospedador, el entorno, el estado nutricional, edad y estilo de vida, entre otros (Figura 1.7).

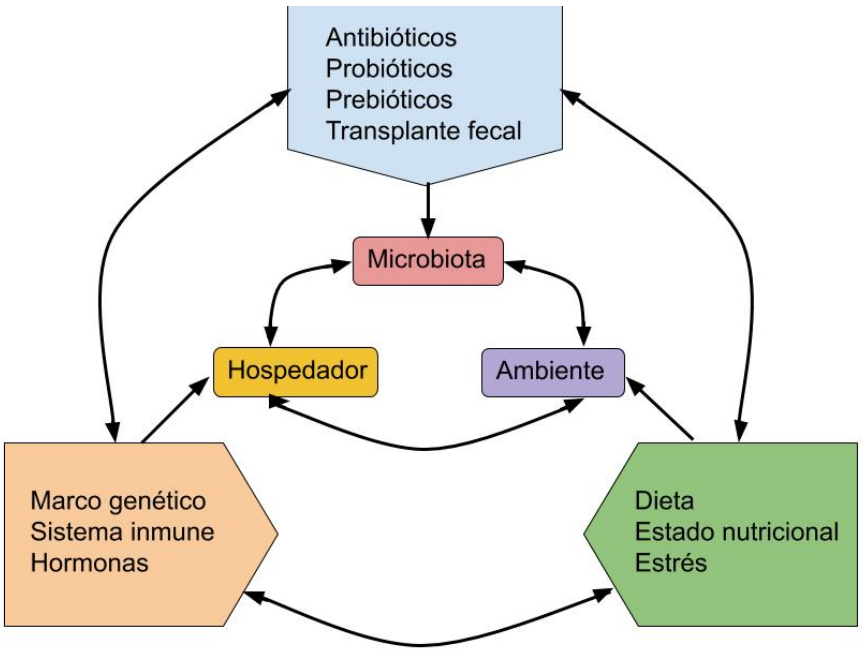

Figura 1.7: Esquema de la complejidad de interacciones entre el hospedador, la microbiota y el ambiente. Adaptado de [Eloe-Fadrosh and Rasko, 2013].

1.1.1. Microbioma intestinal

En humanos, el tracto gastrointestinal (GI) comienza en la cavidad oral y continúa a través del estómago e intestinos, finalizando en el ano. El complejo consorcio de millones de microorga-

nismos que habitan el tracto GI se ha convertido en un tópico de gran interés debido a su importante influencia en estados de fisiológicos o patológicos [Manos, 2022, Vijay and Valdes, 2022].

Las diferentes especies de bacterias que componen la microbiota intestinal humana están clasificadas taxonómicamente en géneros, familias, órdenes y *phyla* (Figura 1.8).

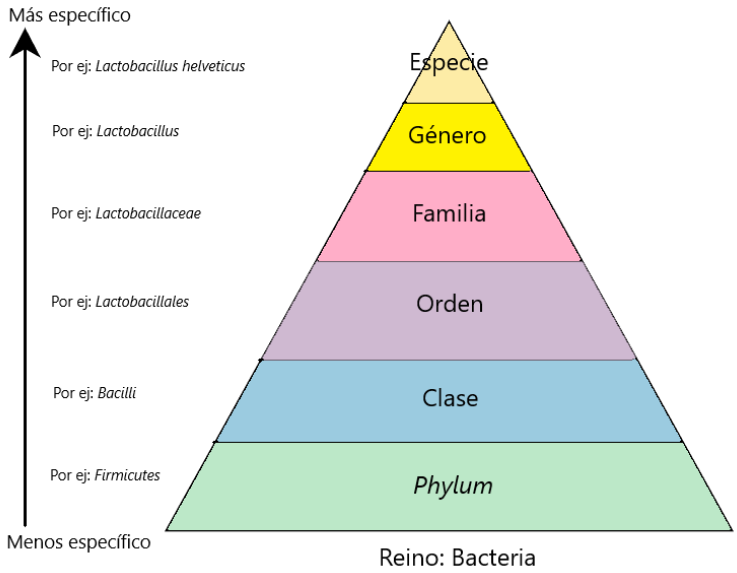

Figura 1.8: Jerarquía taxonómica de bacterias. A modo de ejemplo se utiliza a la especie *Lactobacillus helveticus*. Adaptado de [Van Ameringen et al., 2019].

Los *phyla* presentes en la microbiota intestinal son:

- *Actinobacteria*
- *Proteobacteria*
- *Fusobacteria*
- *Verrucomicrobia*
- *Bacteroidetes*
- *Firmicutes*

A lo largo del tracto GI se diferencian nichos fisiológicos asociados a una densidad y composición de microorganismos característicos. Esta estratificación espacial heterogénea se distribuye desde la boca hasta el recto y también a nivel transversal (desde la mucosa hacia el lumen, Figura 1.9). Algunos de los factores que influyen en la heterogeneidad de distribución espacial de la microbiota del tracto GI son el pH, nivel de oxígeno, gradientes químicos y de nutrientes, péptidos antimicrobiales y características físicas del intestino, entre otros [Donaldson et al., 2016, Tropini et al., 2017].

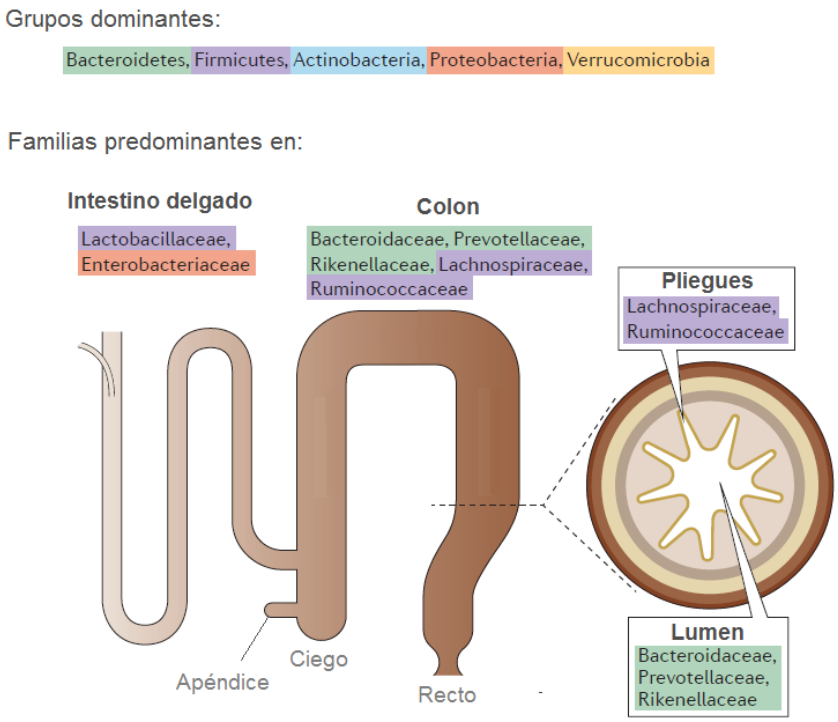

Figura 1.9: Hábitats de la microbiota en la porción baja del tracto GI. Adaptado de [Donaldson et al., 2016]

Diversos trabajos en modelos animales y humanos han destacado la asociación de la disbiosis intestinal con una lista de enfermedades crónicas que incluyen obesidad, síndrome de colon irritable, diabetes, cáncer, y enfermedades neurodegenerativas como Parkinson, Alzheimer y esclerosis múltiple, entre otras [Bäckhed et al., 2004, Matsuoka and Kanai, 2015, Dutta et al., 2019, Zhang et al., 2022].

Mientras que muchas bacterias están asociadas a enfermedades, la microbiota intestinal tiene una participación esencial en diversas funciones beneficiosas para la salud, tal como la digestión, síntesis de vitaminas y aminoácidos esenciales, absorción de calcio, magnesio y hierro, fermentación de componentes no digeribles y protección del intestino ante patógenos, entre otros [Bäckhed et al., 2005, Egert et al., 2006].

Se cree que durante la gestación, el feto humano se desarrolla en un ambiente libre de bacterias y al momento del parto, el recién nacido es expuesto a un amplio espectro de microorganismos que empiezan a colonizar de una manera altamente sincronizada el intestino neonatal, contribuyendo así a las funciones de protección, inmunes y metabólicas. El pasaje a través del canal de parto, el contacto con la microbiota vaginal, fecal y de piel de la madre son factores

que determinan la adquisición y establecimiento de la microbiota inicial en los recién nacidos. En este sentido, existe evidencia de que los nacidos mediante parto vaginal adquieren comunidades bacterianas similares a la microbiota vaginal de su madre, con predominio de especies de *Lactobacillus*, *Prevotella* o *Sneathia*. Mientras que en los nacidos por cesárea predominan comunidades bacterianas similares a las encontradas en la superficie de la piel, con dominancia de *Staphylococcus*, *Corybacterium* y *Propionibacterium* spp [Dominguez-Bello et al., 2010].

La diversidad de las poblaciones que componen la microbiota intestinal es baja en los primeros períodos postnatales y va incrementando con el transcurso del tiempo. Es probable que el aumento en diversidad sea el reflejo del aumento de tamaño del intestino, lo cual va asociado también con una mayor tasa de interacciones con nuevas bacterias y la proliferación de nichos ecológicos [Koenig et al., 2011].

El período de la infancia humana representa para la microbiota intestinal una etapa de rápida colonización de microorganismos que está influenciada por diversos factores, como tipo de parto, alimentación con leche materna o fórmula, duración del período de lactancia, introducción de alimentos sólidos, enfermedad, uso de antibióticos, etc. (Figura 1.10)

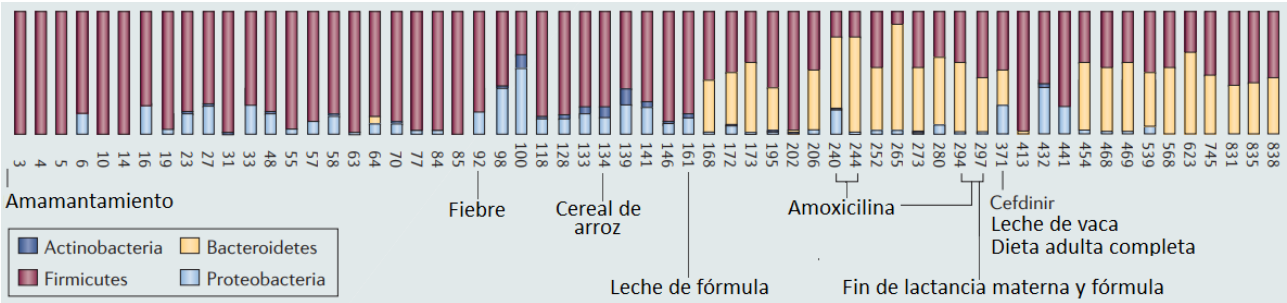

Figura 1.10: Variación de la abundancia de los principales grupos de bacterias (*Actinobacteria*, *Bacteroidetes*, *Firmicutes* y *Proteobacteria*) en el microbioma intestinal en la fase temprana de la niñez ante diversos eventos como consumo de leche materna, fórmula, antibióticos (amoxicilina y cefdinir) y alimentos procesados. Los números indican días postnacimiento. Adaptado de [Spor et al., 2011].

De esta manera, el entendimiento de los factores y mecanismos que afectan al ensamblado y mantenimiento de la microbiota intestinal permite aportar un perfil de comprensión adicional al tratamiento de enfermedades complejas.

1.1.2. ¿Cómo se estudia el microbioma?

La biodiversidad de la microbiota humana ha estado por muchos años subestimada debido fundamentalmente a la limitación de los métodos de aislamiento y cultivo de poblaciones de bacterias. Sin embargo, la innovación tecnológica de los últimos años ha permitido profundizar el conocimiento de las características, diversidad y funciones de la gran cantidad de microorganismos intestinales en el humano [Arnold et al., 2016].

La investigación de la microbiota intestinal puede ser abordada mediante estudios basados en el ADN y pueden ser clasificados en dos categorías principalmente:

- Los que acceden directamente al contenido genético de comunidades enteras de microorganismos en muestras de ambiente natural (suelo, agua, intestino, etc), área conocida como metagenómica.

- Los enfocados en uno o unos pocos genes marcadores. Esto representa una herramienta clave de estudio, dado que a través del perfil de expresión de un gen marcador, es posible caracterizar la composición y abundancia de los microorganismos presentes. Un gen marcador es un segmento de ADN del cual se conoce su ubicación en el genoma y puede ser rastreado a través de la herencia en sucesivas generaciones.

En esta última categoría, el análisis del gen marcador se centra principalmente en el gen del ARN ribosomal (ARNr) de la subunidad pequeña (16S), referido a partir de ahora como **16S rRNA**. La importancia de utilizarlo como una suerte de 'código de barras' se debe a la expresión ubicua del gen dado que todas las bacterias y arqueas lo poseen.

La estructura del gen 16S RNAr presenta varias regiones conservadas (región conservada refiere a una secuencia idéntica o similar entre especies mantenida por selección natural) y nueve regiones hipervariables (Figura 1.11). De esta manera, las regiones conservadas reflejan las relaciones filogenéticas entre las especies mientras que las regiones altamente variables representan suficiente diversidad de secuencia para realizar la identificación y clasificación de bacterias en diferentes grupos.

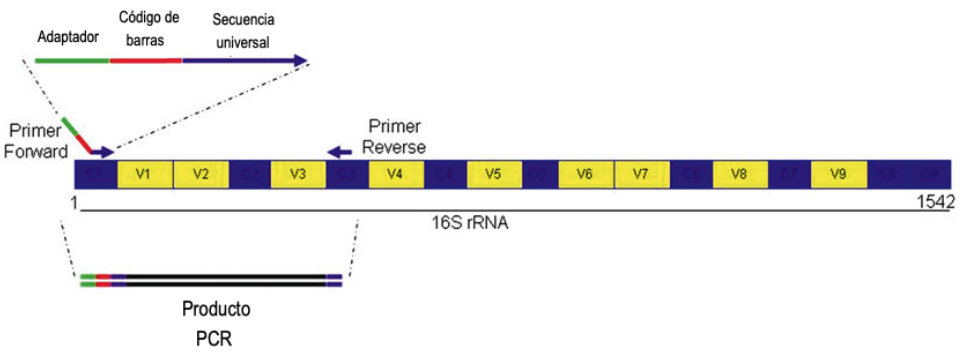

Figura 1.11: Esquema del gen 16S RNAr, indicando las regiones conservadas (en violeta) e hipervariables V1 a V9 (en amarillo). En este ejemplo, la región comprendida entre las flechas *primer forward* y *reverse* es la región *target* a amplificar y analizar (producto PCR). Las diferencias encontradas en ese fragmento amplificado permitirán determinar las diferencias de abundancias de bacterias en una muestra. Adaptado de [Del Chierico et al., 2015]

Las características mencionadas del gen permiten cuantificar la abundancia de cada grupo de bacterias en base a la variabilidad de secuencia en una determinada región hipervariable en estudio. Así, es posible identificar la composición taxonómica de la microbiota (Figura 1.12).

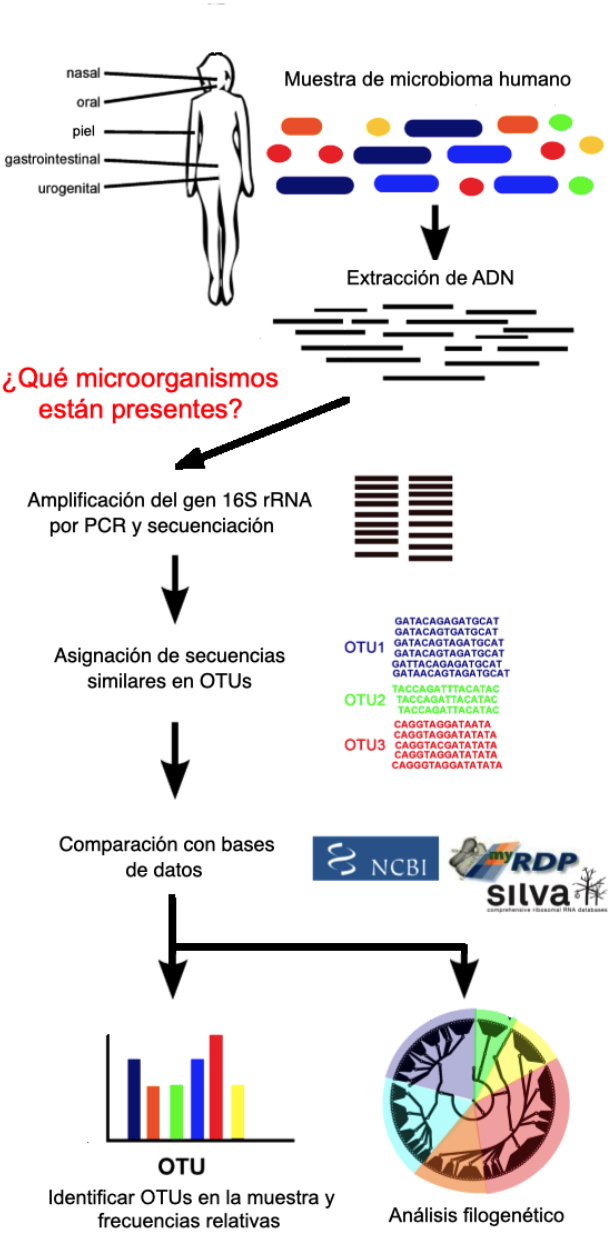

Figura 1.12: La composición de la comunidad de microorganismos puede ser determinada mediante el análisis del gen 16S RNAr. El esquema indica los pasos involucrados en la secuenciación de dicho gen. Las secuencias muy similares se agrupan en unidades taxonómicas operativas (OTU), las cuales son comparadas con bases de datos de organismos reconocidos y analizadas en términos de abundancia o diversidad filogenética. Adaptado de [Morgan et al., 2013].

Las técnicas de biología molecular para caracterizar la microbiota intestinal independientes del cultivo celular son:

- Tecnologías de secuenciación a gran escala, como la secuenciación de próxima generación (NGS).
- Microarreglos o *microarrays* filogenéticos.

A continuación se describirán las características de los *microarrays* filogenéticos dado que los datos usados en esta Tesis fueron obtenidos mediante dicha técnica.

1.1.3. *Microarrays* filogenéticos

La tecnología de *microarray* emplea fragmentos de secuencias conocidas de ácidos nucleicos (sondas) inmovilizados y organizados de manera predeterminada en una placa de vidrio o sílice (denominado *chip*). Las sondas son diseñadas para ser complementarias a una secuencia específica de ADN o ARN proveniente de la muestra en estudio (células o tejido) (Figura 1.13).

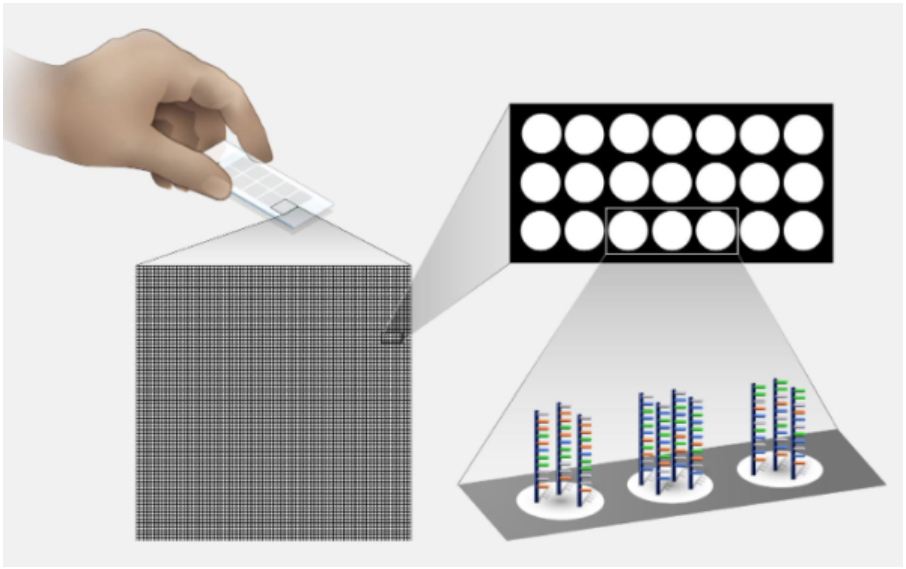

Figura 1.13: Esquema de un *microarray* básico. Cada *spot* contiene múltiples copias idénticas de una secuencia específica de ácido nucleico. La fluorescencia emitida en cada *spot* luego de la hibridación es recolectada y procesada para crear una imagen de color del *microarray*. Adaptado de Genome.gov

Por ejemplo, si se quiere analizar la expresión de un gen de interés, se extrae todo el conjunto de transcriptos que son expresados en ese momento por las células. Los mismos son marcados con fluorescencia y, luego de incubar junto al *chip* en condiciones favorables, los transcriptos presentes que hayan hibridado con las sondas inmovilizadas emitirán una señal de fluorescencia. La intensidad de fluorescencia emitida en cada posición predeterminada en el *chip* refleja la abundancia del transcripto correspondiente a esa secuencia.

Aunque el estudio de la expresión de genes fue la motivación original en el desarrollo de *microarrays*, su gran versatilidad ha permitido adaptar esta tecnología para otros usos, incluyendo el análisis filogenético. Así, los *microarrays* filogenéticos son útiles para determinar la

composición y diversidad de la microbiota humana, dado que es posible cuantificar con gran precisión bacterias conocidas y previamente detectadas. Contienen sondas complementarias a regiones de un gen o un grupo de genes que están presentes de forma ubicua en las especies de interés, una característica importante para el análisis filogenético. En el caso de bacterias se utiliza el gen 16S RNAr. La hibridación del gen 16S RNAr se refleja directamente en el nivel de intensidad de la fluorescencia para cada grupo de bacteria lo que permite hacer una comparación directa de la abundancia relativa de cada grupo en las muestras.

Varios tipos de *microarrays* filogenéticos fueron desarrollados, dependiendo de la elección de la microbiota a estudiar (intestinal, vaginal, oral), nivel de grupo taxonómico alcanzado (especie, género, familia), plataforma, cantidad de sondas, entre otros parámetros [Paliy and Agans, 2012, Paliy et al., 2014].

1.2. Bioinformática

La finalización del Proyecto Genoma Humano en 2003 y los posteriores resultados de secuenciación del genoma de otros organismos representó un cambio fundamental en las Biociencias debido a la abundante disponibilidad de datos biológicos. Esta situación generó, a su vez, una demanda de herramientas computacionales altamente eficientes para el almacenamiento en bases de datos específicas, análisis e interpretación. La combinación de varias disciplinas como Ciencias Matemáticas, Estadística, Ciencias de la Computación, Biología y Medicina ha posibilitado el desarrollo de la Bioinformática (Figura 1.14).

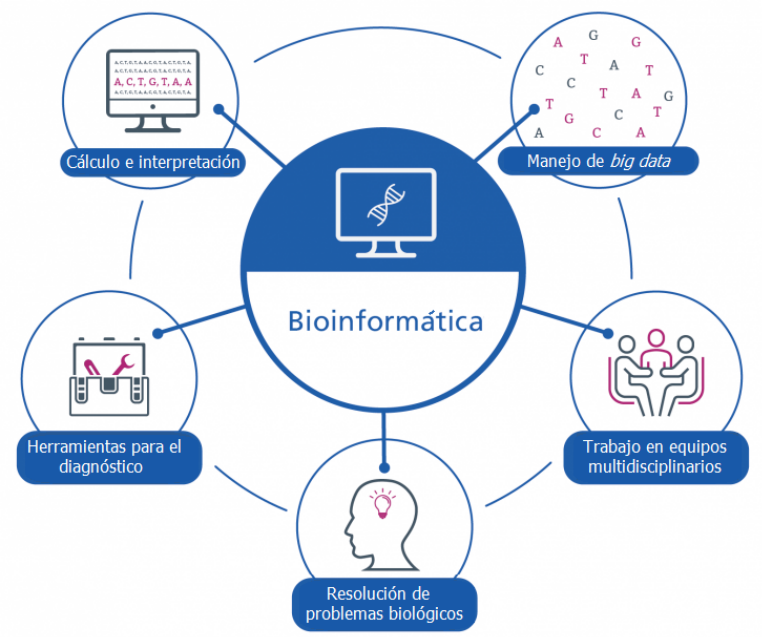

Figura 1.14: Algunas de las áreas de trabajo incluidas en la bioinformática.

En este sentido, la bioinformática podría definirse como una subdisciplina de la biología y las ciencias de la computación que se encarga de adquirir, almacenar, analizar y diseminar la información biológica, en gran parte correspondiente a las secuencias de ácidos nucleicos y aminoácidos. En bioinformática se usan y desarrollan programas informáticos que tienen importantes aplicaciones, tal como: determinar las funciones de genes y proteínas o establecer relaciones evolutivas. (revisado en [Genome.gov]).

El enfoque bioinformático mediante el uso de herramientas de aprendizaje automático permite una integración exhaustiva de datos clínicos multiómicos, interviniendo en etapas como el manejo y análisis de información genética, interpretación de la funcionalidad de genes, mecanismos de regulación celular, selección y diseño de drogas *target*, identificación de enfermedades, entre otras.

1.3. Aprendizaje Automático

La generación de conocimiento biológico a partir del gran flujo de datos generados por las nuevas tecnologías en las ciencias biomédicas requiere el uso de herramientas y técnicas provenientes de las ciencias de la computación para complementar el análisis bioinformático.

Aprendizaje Automático o *Machine Learning* (ML) es una disciplina de la Inteligencia

Artificial e incluye un conjunto de métodos que apuntan a que un sistema aprenda un modelo a partir de una cantidad determinada de datos, tal que pueda identificar uno o más patrones contenidos en los mismos. Los patrones identificados (aprendidos) pueden ser usados para hacer estimaciones sobre datos nuevos, similares a los datos usados para armar el modelo. Es decir, ML confiere a una computadora la capacidad de adaptarse a nuevas circunstancias, detectar y extrapolar patrones [Russell, 2010].

Los algoritmos de ML se dividen principalmente en cuatro categorías: supervisado, no supervisado, semi-supervisado y por refuerzo.

- El **aprendizaje supervisado** implica un conjunto de p predictores X_1, X_2, ..., X_p, medidos en n observaciones y una respuesta Y medida también en las mismas n observaciones (datos etiquetados). El propósito es predecir Y usando los p predictores [Gareth et al., 2013].

- El **aprendizaje no supervisado** utiliza un conjunto de predictores X_1, X_2, ..., X_p registrados en n observaciones pero no existe una variable de respuesta Y asociada (datos no etiquetados). El objetivo es descubrir patrones interesantes en las mediciones de X_1, X_2, ..., X_p [Gareth et al., 2013].

- El **aprendizaje semi-supervisado** combina las dos categorías mencionadas, cuando en un conjunto de datos con predictores X_1, X_2, ..., X_p, no todas las n observaciones tienen una respuesta Y asociada. Entonces, el uso de algoritmos semi-supervisados es útil en la construcción de un modelo que combina datos etiquetados y no etiquetados [Sarker, 2021].

- En el **aprendizaje por refuerzo** el sistema aprende un comportamiento óptimo mediante prueba y error usando *feedback* de sus propias acciones y experiencias. Está basado en la recompensa y la penalización en un ambiente interactivo. Este tipo de aprendizaje puede ayudar a incrementar la automatización u optimización de eficiencia en sistemas sofisticados [Sarker, 2021].

Debido a su particular capacidad para manejar grandes y complejos conjuntos de datos y hacer predicciones sobre los mismos mediante la generación de modelos, el uso de ML se

expandió rápidamente en la comunidad de biociencias y en particular en el análisis bioinfomático [Lai et al., 2018, Shastry and Sanjay, 2020]. En el campo del microbioma intestinal, un número cada vez más creciente de investigaciones se ha visto beneficiado por el uso de métodos de ML [Zhou and Gallins, 2019, Curry et al., 2021].

Sin embargo, la selección de un método de ML que mejor se ajuste a un conjunto de datos representa un desafío, teniendo en cuenta que existe un extenso repertorio de algoritmos y, en consecuencia, múltiples modelos que pueden ser generados. Por ello, a continuación se describirán los principios de los algoritmos de ML usados en esta Tesis:

- Análisis de componentes principales
- *Uniform Manifold Approximation and Projection* (UMAP)
- Análisis de agrupamiento (*Clustering*)
- Mapas auto-organizados
- *Random forest*

1.3.1. Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA) es una técnica de reducción de la dimensionalidad lineal de una matrix $X_{n\times p}$ de datos, donde n es el número de observaciones y p el número de variables. La reducción de la dimensión genera un nuevo espacio de componentes independientes (llamados componentes principales, PC) el cual es simplemente una rotación del espacio original, manteniendo cualquier relación presente en los datos.

El sistema de PC generado son combinaciones lineales de las variables p de tal manera que gran parte de la información contenida en las variables originales es almacenada en los primeros componentes. PCA es una herramienta útil en el análisis exploratorio de datos (Figura 1.15).

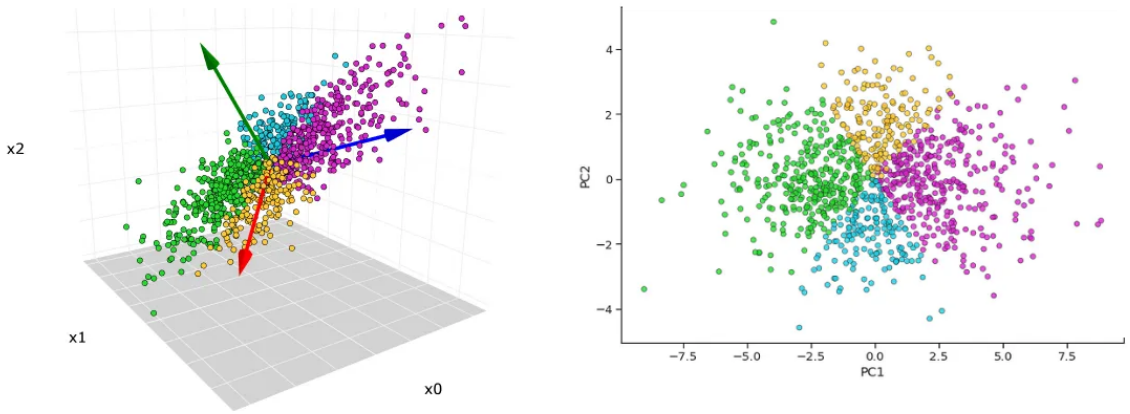

Figura 1.15: Análisis de Componentes Principales. Izquierda: Un conjunto de datos hipotético tiene una distribución 3-D. Las flechas indican la orientación de los tres primeros PC. Derecha: Luego del PCA, se redujo de 3 a 2 dimensiones.

Los principales pasos en PCA son:

1. Escalar las variables iniciales tal que cada una contribuya de igual manera al análisis (a cada valor se le resta la media y se divide por el desvío estándar).

2. Calcular la matriz de covarianza la cual define la dispersión (varianza) y la orientación (covarianza) de los datos, es decir, resumir las correlaciones entre todos los posibles pares de variables.

3. Determinar los autovectores y autovalores de la matriz de covarianza: los autovectores indican las direcciones de los ejes que contienen la dispersión mayor de los datos (los PC) mientras que los autovalores son los coeficientes asociados a los autovectores e indican la magnitud de la varianza en cada componente principal.

4. Calcular los componentes principales: Luego de realizar PCA, se obtendrá un número N de componentes principales, donde N es igual a la dimensionalidad de los datos originales.

De la lista de PC obtenidos, el interés no es enfocarse en todos sino en unos pocos que expliquen la mayor parte de la variabilidad del sistema y permitan una buena comprensión de los datos. Sin embargo, no existe una forma objetiva para decidir cuántos PC son suficientes. Esta cuestión depende del área específica de aplicación y del conjunto de datos [James et al., 2013]. De todos modos, puede ayudar:

- Considerar de manera arbitraria un porcentaje de varianza explicada (por ejemplo, tomar solamente los PC que en conjunto expliquen más del 70 %).

- Analizar la herramienta visual que ofrece un gráfico de sedimentación, que muestra la curva de la varianza explicada en función de los componentes principales. Se busca el punto de corte en el cual la curva muestra un descenso empinado, es decir, la proporción de la varianza explicada cae bruscamente y aparece un codo en el gráfico.

En esta Tesis, los resultados de PCA fueron usados como una herramienta en el análisis exploratorio de los datos.

1.3.2. UMAP

Cuando los métodos de reducción de dimensionalidad lineales no son efectivos sobre un conjunto de datos, se asume que la estructura subyacente de los datos está sobre subespacios conocidos como *manifolds*. En matemática, un *manifold* es un objeto geométrico estándar que generaliza la noción intuitiva de 'curva' (1-*manifold*) y de 'superficie' (2-*manifold*) a cualquier dimensión y sobre cuerpos diversos. De esta manera, la reducción de la dimensionalidad en estos casos puede ser estimada mediante enfoques de reducción no lineales. Estos métodos comparten el concepto central de reducir la dimensionalidad mediante la incrustación (*embedding*) de un grafo de vecinos cercanos (construido en el espacio de alta dimensionalidad donde residen los datos) en un *manifold* latente de menor dimensión. Los pasos básicos que emplean son:

1. Construcción de un grafo de k-vecinos en el espacio de alta dimensión.

2. Proyección del grafo sobre el espacio objetivo de menor dimensión

UMAP (*Uniform Manifold Approximation and Projection*) es un algoritmo perteneciente a este grupo de métodos, desarrollado recientemente [McInnes et al., 2018]. Si bien no se pretende hacer un desarrollo matemático del algoritmo, básicamente usa análisis de datos topológicos y mapeo de un grafo de alta dimensionalidad en un grafo de dimensiones menores manteniendo la estructura de los datos originales lo más similar posible (Figura 1.16). UMAP crea un vecindario alrededor de cada punto mediante un enfoque de k-vecinos cercanos estocástico. El mapeo

óptimo de la incrustación de baja dimensión es resuelto por UMAP mediante un descenso de gradiente estocástico usando como función de costo la entropía cruzada, es decir, mediante la minimización de las representaciones topológicas de los datos en alta y baja dimensión [Coenen, 2019].

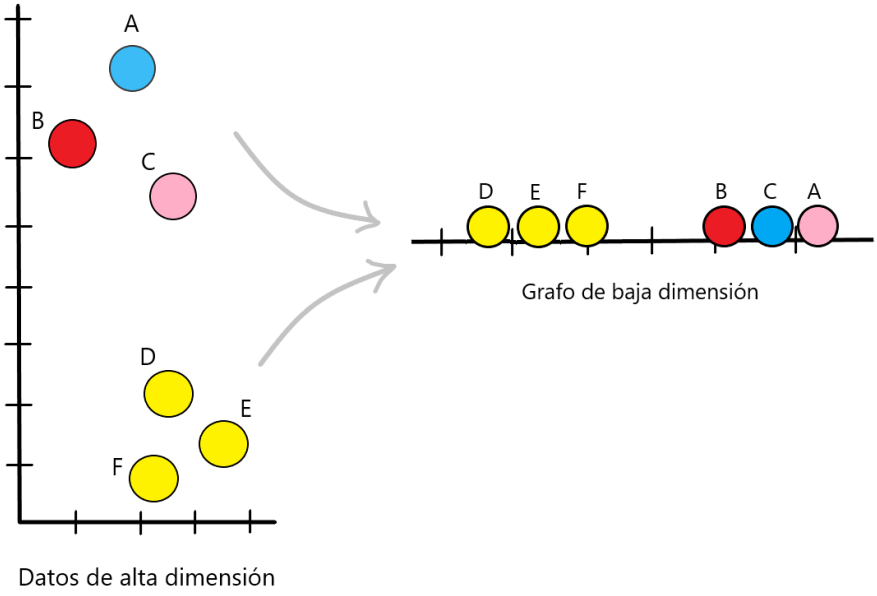

Figura 1.16: Esquema general de UMAP.

Para equilibrar entre la estructura local y global de la proyección final, UMAP requiere el ajuste de varios hiperparámetros que impactan significativamente en la estructura final siendo dos de ellos los más importantes:

- Número de vecinos: usados para contruir el grafo inicial de alta dimensión. Este parámetro controla cuán compacto el algoritmo agrupa los datos. Valores bajos hacen que el algoritmo esté más centrado en la estructura local mientras que valores altos ponen más énfasis en la representación global, no centrado en detalles finos.

- Distancia mínima: la distancia mínima efectiva entre los puntos incrustados. Valores bajos de este parámetro resultan en incrustaciones más compactas, mientras que valores altos resultan en puntos más dispersos enfocados en la preservación de la estructura topológica general.

Existe un parámetro adicional que es el número de componentes, el cual define la dimensión del resultado de UMAP (1-dimensión, 2-dimensiones, etc). En la mayoría de los casos,

$n\_componentes = 2$ es la mejor opción dado que es más fácil interpretar un mapa 2D que 1D, 3D o más.

Cuando los métodos lineales no son efectivos para aportar resultados concluyentes, el uso de una técnica multidimensional como UMAP permite abordar la no linealidad de un conjunto de datos y encontrar una representación de los mismos en dos dimensiones.

1.3.3. Análisis de agrupamiento

El agrupamiento o *clustering* es una importante herramienta en el aprendizaje no supervisado e incluye un amplio conjunto de técnicas que apuntan a identificar grupos o k *clusters* de objetos similares en un conjunto de n datos, donde $k \ll n$.

La finalidad de un análisis de agrupamiento es que los objetos dentro de un *cluster* estén lo más estrechamente relacionados entre sí (es decir, una alta similitud intra-clase) en comparación con los objetos de otros *clusters* (baja similitud entre-clase). Un punto central a todos los algoritmos de *clustering* es la noción del grado de similitud (o disimilitud) entre los objetos individuales a ser agrupados: se intenta agrupar los objetos en base a la definición de similitud proporcionada [James et al., 2013]. *Clustering* se basa en aprendizaje no supervisado dado que se intenta descubrir estructuras en los datos, es decir grupos distintos.

Existen diferentes tipos de algoritmos de *clustering*, siendo los más frecuentes:

- Partición: minimizan un determinado criterio de agrupamiento mediante la reubicación iterativa de los datos entre los *clusters* hasta obtener una partición óptima. Requieren que se especifique el valor de k. Por ejemplo: *k-means*, *k-medoids*, etc.

- Jerárquicos: no requieren que se especifique el valor de k. Producen una serie de particiones donde, en cada paso, dos *clusters* son unidos (aglomerativo) o un *cluster* es dividido en dos (disociativo).

A continuación se describirán las principales características de los métodos *k-means* y jerárquico.

1.3.3.1. *K-means*

K-means es un método simple de partición de datos en un conjunto de k grupos no superpuestos donde cada uno de estos está representado por su centroide (el valor medio de las observaciones). Primero debe especificarse el número de k grupos. Luego, el algoritmo *k-means* va alternando entre los siguientes pasos:

- Asignación de los datos: Cada dato es asignado a su centroide más cercano, en base a una medida de distancia o similaridad.
- Actualización del centroide: En este paso, se recalculan los centroides mediante el promedio de todos los datos asignados a cada *cluster*.

El algoritmo se detiene cuando los centroides se han estabilizado (es decir, no hay más cambio que se obtenga de agrupamientos previos) y se ha alcanzado el número definido de iteraciones. El valor de k debe ser especificado con anticipación y su determinación depende en general del conocimiento del dominio y de los datos, entre otros factores.

Sea $C_1 ... C_k$ los conjuntos que contienen observaciones en cada *cluster*, estos satisfacen dos propiedades:

1. $C_1 \cup C_2 \cup \cdots \cup C_k = \{1, \ldots, n\}$. Es decir, cada observación pertenece al menos a uno de los k $clusters$.

2. $C_k \cap C_{k'} = \emptyset$ para todo $k \neq k'$. Es decir, ninguna observación pertenece a más de un $cluster$.

El algoritmo *k-means* mueve los centroides de manera iterativa con el fin de minimizar la varianza total intra-*cluster*. La varianza intra-*cluster* para el *cluster* C_k es una medida $W(C_k)$ de cuánto difieren las observaciones entre sí dentro de un *cluster*. Generalmente, se usa la distancia euclídea para medir la variación intra-*cluster*.

1.3.3.2. *Clustering* jerárquico

El *clustering* jerárquico es un método que busca armar una jerarquía de *clusters*. A diferencia de los métodos de partición como *k-means*, es necesario especificar una medida de disimilitud entre los grupos de observaciones. Existen dos categorías principales:

- Aglomerativo: este tipo de agrupamiento toma a cada observación como un *cluster* individual y de manera iterativa va uniendo a cada uno de ellos hasta que el *cluster* final contiene todos las observaciones. Este enfoque es considerado de abajo hacia arriba (*bottom-up*).
- Divisivo: esta técnica sigue un enfoque de arriba-abajo (*top-down*) el cual comienza a partir de un único *cluster* con todas las observaciones y de manera iterativa divide el *cluster* en grupos más pequeños hasta que cada uno contiene una única observación.

El *clustering* jerárquico resulta en una estructura similar a un árbol, el cual representa de manera gráfica las relaciones entre todas las observaciones. En el tipo de agrupamiento aglomerativo, el dendrograma se construye iniciando en las hojas (cada observación individual) y combinando y fusionando *clusters* hasta formar el tronco (que contiene todas las observaciones), completando la estructura final del dendograma.

Sin embargo, a medida que el algoritmo va fusionando *clusters* de abajo hacia arriba, el concepto de disimilitud entre pares de observaciones necesita ser extendido a pares de grupos de observaciones. Esto implica el concepto de *linkage*, el cual especifica cómo calcular la distancia entre dos grupos de observaciones. Los tipos más comunes de *linkage* son: completo, *average*, *single* y centroide.

1.3.3.3. Evaluación del número óptimo de *clusters*

El componente de aprendizaje no supervisado del análisis de *clustering* implica que no hay una variable de respuesta para entrenar y encontrar relaciones entre las observaciones. Por lo tanto, al no existir información *a priori* sobre los grupos obtenidos, es importante la estimación y validación de los mismos mediante métodos estadísticos. Para ello existen métodos directos y de evaluación estadística que ayudan a determinar el valor de k.

Los métodos directos apuntan a la optimización de un criterio e incluyen:

- Método del codo (*elbow method*): Este método se basa en calcular la suma de errores cuadrados intra-grupo (WCSS, *Within Cluster Sums of Squares*) en función de k. WCSS calcula la distancia promedio al cuadrado de todos los puntos en un grupo al centroide del mismo, es decir, mide cuán compacto es el agrupamiento. El valor de k a partir del cual WCSS empieza a disminuír (codo en la curva) sugiere el número óptimo de k a usar.

- Coeficiente de *Silhouette*: Evalúa la calidad del *cluster* y estima las distancias de cada punto intra e inter *cluster*, mediante la similitud y disimilitud.

La idea principal en estos métodos directos es lograr una máxima cohesión de los *clusters*, es decir, que la variación total dentro de los mismos sea mínima: a medida que k aumenta, cada *cluster* tiene menos objetos que lo conforman y los mismas estarán más cerca de sus respectivos centroides. Es decir, la variación promedio va disminuyendo pero en algún punto la mejora se vuelve insignificante.

Por otra parte, el análisis de *Silhouette* (S_i) permite determinar qué tan bien un dato fue agrupado en el *cluster*. S_i es calculado de la siguiente manera:

$$S_i = (b_i - a_i)/max(a_i, b_i) \tag{1.1}$$

- Para cada observación i, se calcula la disimilitud a_i entre i y todos las otras observaciones del *cluster* al que pertenece i.

- Para los restantes *clusters* C, a los cuales i no pertenece, se calcula la disimilitud promedio *d(i,C)* de i a todas las observaciones de C. El menor de estos *d(i, C)* se define como $b_i = min_C d(i, C)$. El valor de b_i puede verse como la disimilitud entre i y sus *clusters* "vecinos", es decir, el más cercano al cual no pertenece.

El coeficiente de *Silhouette* oscila entre -1 y 1. Los valores cercanos a 1 indican que la observación está lejos de los *clusters* vecinos. Valores cercanos a 0 indican *clusters* solapados. Valores negativos generalmente indican que la observación ha sido asignada al *cluster* incorrecto.

En cuanto a los métodos de evaluación estadística para determinar el valor de k:

- Estadístico de *Gap*: Se basa en comparar la variación total intra-*cluster* para diferentes valores de k con respecto a los valores esperados bajo una distribución uniforme de referencia sin agrupamientos obvios. La estimación del k óptimo es el valor que maximiza el estadístico *gap*, lo que significa que la estructura de agrupamiento está alejada de la distribución uniforme al azar ([Tibshirani et al., 2001]).

1.3.4. Tendencia de *clustering*

La determinación de la tendencia a formar *clusters* en los datos puede ser evaluada previo a la implementación de algún algoritmo para tal fin. Para ello existen pruebas estadísticas, como el estadístico de *Hopkins* o métodos visuales, como el algoritmo VAT (*Visual Assessment of cluster Tendency*) (revisado en [Han et al., 2011]).

En cuanto al estadístico de Hopkins (H), éste analiza la distribución espacial aleatoria de los datos mediante el cálculo de la probabilidad de que los datos provengan de una distribución uniforme. Por ejemplo, sea D un conjunto de datos, el estadístico H se calcula de la siguiente manera:

- Se selecciona una muestra uniforme de n observaciones $(p_1, ..., p_n)$ a partir de D.

- Para cada observación p_i en D, se calcula la distancia x_i a su vecino más cercano p_j, siendo $x_i = dist(p_i, p_j)$.

- Se genera un conjunto de datos simulados de tamaño n a partir de una distribución al azar (q_1, ..., q_n) con la misma variación que el conjunto D.

- Se mide la distancia y_i para cada dato simulado q_i al dato q_j más cercano, siendo $y_i = dist(q_i, q_j)$

- Se calcula el estadístico H como la media de las distancias de vecinos más cercanos en el conjunto simulado, dividido por la suma de las medias de las distancias vecimas más cercanas de D y el conjunto simulado.

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

El estadístico H puede interpretarse así: si D siguiera una distribución uniforme, entonces $\sum_{i=1}^{n} x_i$ y $\sum_{i=1}^{n} y_i$ estarían muy cercanos, resultando en $H \approx 0.5$. Por el contrario, si existieran *clusters* en D, entonces las distancias entre los datos simulados ($\sum_{i=1}^{n} y_i$) serían mayores que las distancias entre los datos reales ($\sum_{i=1}^{n} x_i$), resultando en un incremento en el valor de H. Entonces considerando a $H = 0.5$ como el valor umbral:

- Si $H < 0.5 \Rightarrow$ es poco probable que los datos tengan *clusters* estadísticamente significativos.

- Si H es cercano a 1 $\Rightarrow$ el conjunto de datos es significativamente agrupable en *clusters*.

1.3.5. Mapas auto-organizados (SOM)

Los mapas auto-organizados (SOM, por sus siglas en inglés) son un tipo de red neuronal artificial que usa un proceso de aprendizaje competitivo no supervisado para representar observaciones multidimensionales (vectores de entrada) en un mapa bidimensional de neuronas o nodos. Los SOM, también conocidos como mapas de Kohonen, fueron descriptos por primera vez por Teuvo Kohonen en Finlandia en 1982. Son usados frecuentemente para tareas de clasificación y *clustering* en diversas áreas como reconocimiento del lenguaje, reconocimiento de imágenes, control de robots y diagnóstico médico, entre otros [Ruan et al., 2011, Gandía-Aguiló et al., 2017].

Como toda red neuronal, los SOM consisten en una capa de neuronas de entrada y otra de salida. Mediante aprendizaje competitivo, las neuronas de salida compiten entre ellas para ser activadas, con el resultado de que sólo una es activada a la vez. Tal competencia puede ser inducida o implementada mediante conexiones inhibitorias laterales entre las neuronas, forzando a las mismas a organizarse (Figura 1.17).

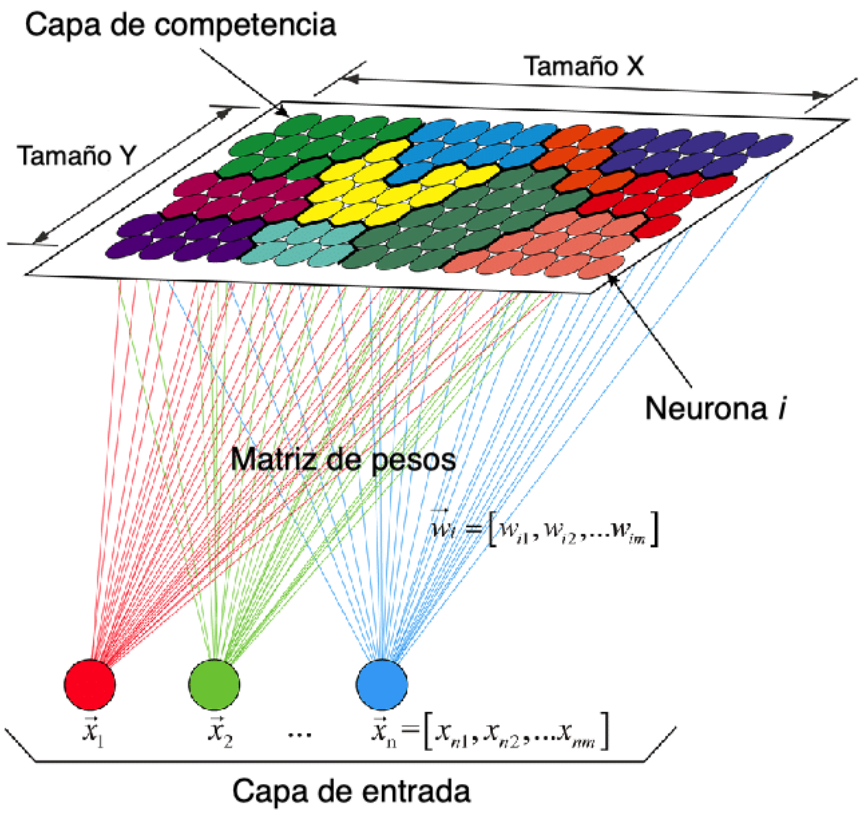

Figura 1.17: Estructura de una red SOM. Los n vectores de la capa de entrada son mapeados en una capa de competencia 2D, representeda por vectores que contienen los pesos. Cada neurona i tiene un vector de peso w de la misma dimensionalidad m que los vectores de entrada(x_n). Adaptado de [Han et al., 2019]

La representación en dos dimensiones es un **mapa topográfico**. Este concepto proviene del área de la neurobiología, el cual establece que las diferentes entradas o *inputs* sensoriales tienen una representación ordenada en su correspondiente área en la corteza cerebral.

Los SOM tienen una arquitectura de red de dos capas:

- Capa de vectores de entrada, donde ingresa la información de los vectores originales a la red durante el proceso de entrenamiento.

- Capa de nodos de aprendizaje (capa de competencia), en la cual interesan las relaciones topológicas entre los nodos. Esta capa contendrá finalmente la información acerca de la representación resultante.

Cada nodo tiene una posición en el mapa 2D así también como una posición asociada en el espacio de entrada, toma la forma de un vector de pesos n-dimensional $m = [m_1, ..., m_n]$ donde n es el número de dimensiones de los vectores de entrada.

1.3.5.1. Algoritmo

El espacio de datos de entrada está conformado por vectores n-dimensionales $x \in \mathbb{R}^n$ mientras que el conjunto ordenado de nodos tiene la forma $m_i \in \mathbb{R}^n$.

- Inicialización de los pesos de cada nodo con valores aleatorios entre $[0, 1]^n$. Luego, se selecciona aleatoriamente un vector de entrada y se calcula la distancia del mismo a través de todos los vectores de peso.

- Existen diversos métodos de determinación de la distancia, aunque el más usado es la **distancia euclídea**. El peso con la distancia más corta es el ganador.

$$||x(t) - m_c|| \leq ||x(t) - m_i|| \forall i \tag{1.2}$$

- El nodo ganador, m_c, es denominado *Best Matching Unit* (BMU) para $x(t)$. Luego, se calcula el vecindario del BMU: a mayor cercanía de un nodo al BMU, su peso será más influenciado por este mientras que, a mayor distancia del BMU, el nodo estará menos influenciado y aprende menos.

- Mediante N iteraciones a partir de la selección aleatoria de un vector de entrada el algoritmo continua hasta que cada vector de entrada ha sido asignado a un nodo. De esta manera, el mapa va creciendo y toma diferentes formas: cuadrada, rectangular o hexagonal en el espacio 2D.

- En cada iteración, todos los nodos m_i y m_c son ajustados de acuerdo a una tasa de aprendizaje indicada en 1.3.

$$m_i(t + 1) = m_i(t) + h_{ci}(x(t) - m_i(t)) \tag{1.3}$$

  donde t es el índice de la iteración, $x(t)$ es un vector de entrada elegido al azar en esa iteración, c es el índice del BMU para $x(t)$ y h_{ci} representa la tasa de aprendizaje (que disminuye a medida que la distancia converge a cero) [Asan and Ercan, 2012].

Luego del entrenamiento, el resultado es configurado de manera tal que mantiene la estructura topológica de los datos de entrada, produciendo un mapa de dos dimensiones en el cual los valores similares son mapeados en el mismo nodo o cercanos.

1.3.6. *Random forest*

Random forest (RF) es un algoritmo de aprendizaje supervisado muy utilizado en ML que genera, incluso sin ajuste de hiperparámetros, resultados robustos y de impacto en problemas de toma de decisiones, tanto para clasificación como para regresión. RF está basado en árboles de decisión los cuales tienen una estructura similar a la de un árbol: cada nodo interno representa un atributo, cada rama representa una decisión y cada nodo hoja terminal representa un resultado categórico o contínuo (es decir, la decisión tomada luego de calcular todos los atributos). El camino desde el nodo raíz hasta el nodo hoja representan reglas de clasificación.

Los algoritmos inductores de árboles de decisión son considerados los métodos de aprendizaje supervisado más usados. Todos comparten el objetivo de encontrar el árbol óptimo, de tal manera que cada partición binaria va dividiendo el árbol de manera recursiva en nodos terminales cada vez más homogéneos en relación a una variable *target*.

Un esquema típico para un algoritmo de inducción *top-down* es el que se indica en la Figura 1.18. En cada iteración, el algoritmo considera la partición del conjunto de entrenamiento usando la salida de una función discreta de los atributos de inicio. Luego de la selección de la partición apropiada, cada nodo continúa subdividiendo el conjunto de entrenamiento en subconjuntos más pequeños, hasta que se satisface algún criterio previamente establecido.

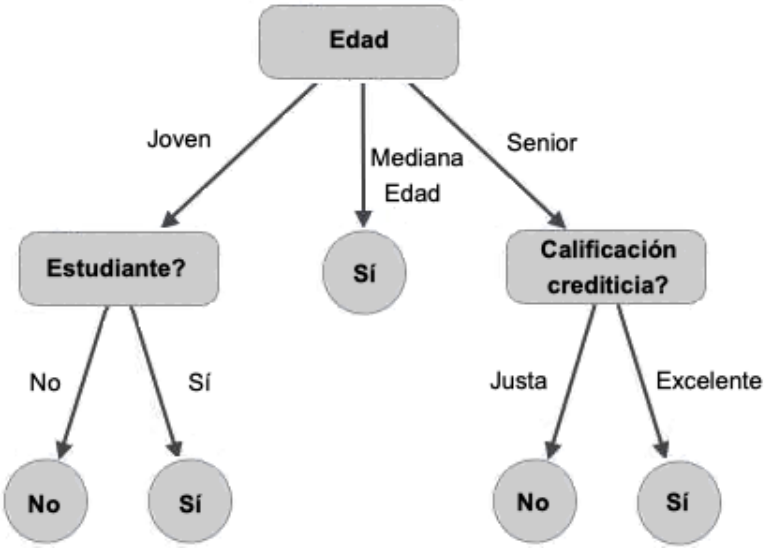

Figura 1.18: Ejemplo de árbol de decisión.

Sin embargo, un árbol de decisión simple con poca profundidad resulta en un modelo débil al momento de analizar conjuntos de datos complejos y grandes. Por ello, para obtener mejoras en la eficiencia de predicción se recurre a los ensambles, que son una combinación de varios modelos en un modelo fuerte. Los métodos de ensamble son efectivos ya que la combinación de resultados de múltiples modelos reduce el error de varianza.

Un procedimiento de ensamble es *bagging* (o *bootstrap aggregation*) que toma al azar varios subconjuntos de datos a partir de los datos de entrenamiento. Así, cada colección de subconjuntos es usada para entrenar los árboles de decisión, generando como resultado un ensamble de diferentes modelos. Sin embargo, *bagging* implica que los árboles son similares entre ellos dado que, ante la presencia de un predictor muy fuerte junto a otros predictores más moderados en un conjunto de datos, cada árbol que sea construído tomará el predictor fuerte en su nodo raíz.

Teniendo en cuenta esto, en el 2001 Leo Breiman de la Universidad de California, Berkeley, desarrolló una extensión de *bagging* que mejoró significativamente la precisión de clasificación [Breiman, 2001], conocido como *random forest* (Figura 1.19).

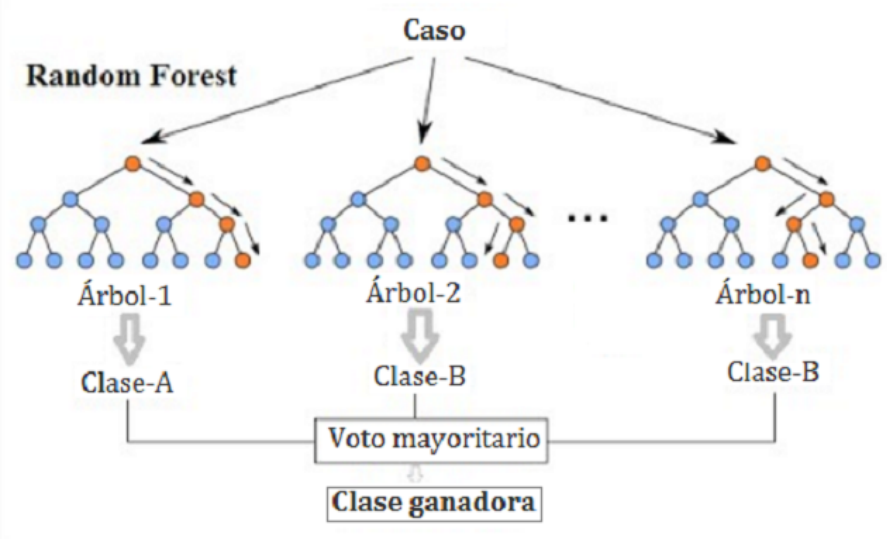

Figura 1.19: Esquema de *random forest*. El algoritmo construye una cantidad n de árboles de decisión en paralelo durante la etapa de entrenamiento. La salida o voto mayoritario del sistema es la moda de las clases (clasificación) o la predicción media (regresión).

La metodología RF involucra dos pasos de aleatoridad en la formación de los árboles componentes:

- El primer paso toma al azar subconjuntos de datos de entrenamiento (selección con reemplazo). Así, para el entrenamiento de cada árbol se usa un subconjunto distinto de datos mientras que para la evaluación del modelo se usan los datos remanentes (conocidos como *out-of-bag*).

- Luego, en cada partición el algoritmo considera solamente una fracción de las variables predictoras en vez de todas juntas. Ésto resulta en árboles con diferentes predictores en la partición inicial y, por lo tanto, en árboles no correlacionados y un resultado promedio más confiable. Este incremento de diversidad hace que RF sea más robusto ante variables predictoras correlacionadas [Horning et al., 2010].

Tanto en problemas de regresión como de clasificación, el desempeño de RF en una observación i generalmente es cuantificada mediante una función de pérdida, la cual mide la diferencia entre la respuesta verdadera y_i y la respuesta de predicción $\hat{y}_i$.

La métrica clásica y más directa es definida para una observación i como:

$$e_i = (y_i + \hat{y}_i)^2 = L(y_i, \hat{y}_i) \tag{1.4}$$

donde *L(.,.)* es la función de pérdida $L(x,y) = (x - y)^2$. En el caso de regresión, esto corresponde al error cuadrado.

El error cuadrado medio (MSE) es una métrica que mide el promedio de la ecuación 1.4, pero es sensible a *outliers*.

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i + \hat{y}_i)^2 \tag{1.5}$$

La raíz cuadrada del error cuadrado medio (RMSE) es una métrica ampliamente usada en problemas de regresión. El rango va desde 0 hasta infinito: un valor de 0 significa que los valores predichos coinciden con los valores reales, valores bajos de RMSE indican que el modelo se ajusta bien a los datos y tiene predicciones más precisas. Por el contrario, valores más altos sugieren mayor error y predicciones menos precisas. Al igual que MSE, RMSE también es sensible a *outliers*.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i + \hat{y}_i)^2} \tag{1.6}$$

El error absoluto medio (MAE) es el promedio de los errores absolutos y representa una métrica robusta a los valores *outliers*. Conceptualmente, MAE es la métrica de evaluación más sencilla en problemas de regresión que indica qué tan alejadas estuvieron las predicciones.

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|x_i - x| \tag{1.7}$$

donde N es el número total de errores y $|x_i - x|$ son los errores absolutos.

1.4. Objetivos

El principal objetivo de la tesis es generar una herramienta que permita encontrar asociaciones entre los datos de abundancia de la microbiota intestinal y las características obtenidas a partir de cada individuo mediante el uso de algunos de los principales métodos de minería de datos y aprendizaje automático. El objetivo es principalmente metodológico, buscando ofrecer un recurso de procesamiento, análisis de datos y variables así como de soporte en la toma de decisiones dentro de un proyecto médico.

Los objetivos específicos son:

- Clasificar la importancia biomédica de las relaciones entre las variables demográficas de los pacientes y la información obtenida a partir del microbioma intestinal.
- Seleccionar las herramientas de aprendizaje automático más útiles para analizar y visualizar los resultados de inteligencia artificial aplicados en salud.

1.5. Sinopsis

La Tesis está organizada de la siguiente manera:

- En el Capítulo 1 (página 16) se presentan los conceptos básicos de biología molecular y microbioma intestinal necesarios para establecer el marco teórico de esta Tesis, además de aprendizaje automático.
- En el Capítulo 2 (página 53) se describen los materiales y métodos usados.
- En el Capítulo 3 (página 58) se realiza una descripción y análisis exploratorio del conjunto de datos.
- En el Capítulo 4 (página 69) se realiza un análisis combinado de reducción de la dimensionalidad y *clustering*.
- En el Capítulo 5 (página 89) se implementa el método no supervisado de Mapas Auto-Organizados (SOM).
- En el Capítulo 6 (página 101) se implementa el algoritmo supervisado de *Random Forest*.
- En el Capítulo 7 (página 108) se discuten los resultados obtenidos y su significancia biomédica en el estudio del microbioma humano.
- Finalmente, en el Capítulo 8 (página 112) se resumen los resultados y se proponen recomendaciones para análisis futuros.

Para llevar a cabo los objetivos planteados, se muestra a continuación un esquema de los pasos y métodos usados (Figura 1.20):

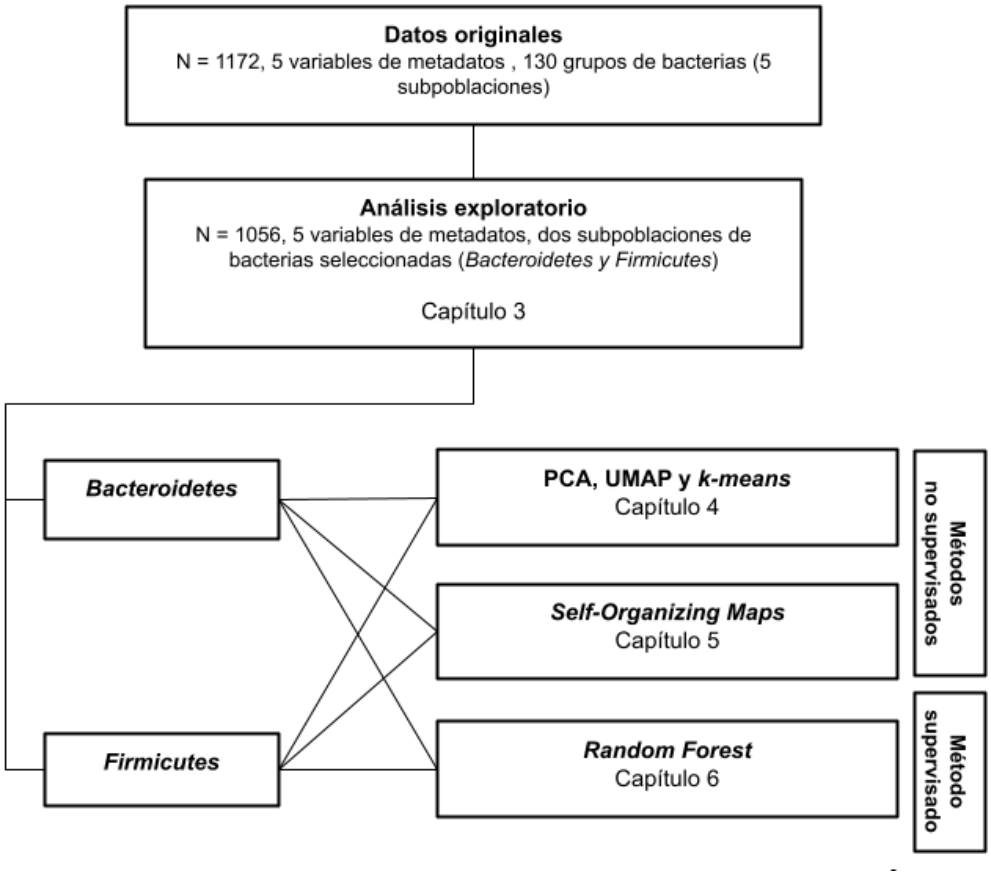

Figura 1.20: Diagrama del flujo de trabajo general realizado en esta Tesis.

Capítulo 2

Materiales y Métodos

2.1. Conjunto de datos

Los datos utilizados en esta Tesis fueron obtenidos del repositorio digital DRYAD que ofrece diversos tipos de datos curados para uso libre en investigación. El conjunto de datos original usado incluye información de 1172 individuos adultos provenientes de 15 países occidentales agrupados en 6 regiones:

- Europa Central: Bélgica, Dinamarca, Alemania, Países Bajos.
- Europa del Este: comprendido solamente por Polonia.
- Países nórdicos: Finlandia, Noruega, Suecia (a lo largo de la Tesis será referido como Escandinavia).
- Europa del Sur: Francia, Italia, Serbia, España.
- Reino Unido e Irlanda (será referido como UK/Irlanda).
- Estados Unidos (US).

.

A partir de muestras de materia fecal de los individuos, se determinó la abundancia relativa para 130 grupos de bacterias intestinales a través de las intensidades de señal del gen del ARN ribosomal (ARNr) de la subunidad pequeña (16S), referido a partir de ahora como 16S RNAr.

Para ello se usó la tecnología de *microarrays* filogenéticos *HITChip*. Estos tipos de *microarrays* ofrecen una plataforma de análisis sensible mediante la detección de las regiones hipervariables V1 y V6 del gen 16S RNAr, pudiendo detectar 1033 tipos de bacterias que representan la mayoría de la diversidad bacteriana en el intestino humano.

Los metadatos correspondientes a las muestras de la matriz de *HITChip* también están disponibles con las siguientes variables:

- Edad (años).
- Sexo (masculino/femenino).
- Nacionalidad (región geográfica, referido en el texto como Nacionalidad).
- Diversidad (índice de diversidad de *Shannon*, en base a las señales de las sondas en el *microarray*).
- BMI: clasificación del índice de masa corporal. Bajo peso: <18.5; peso normal (delgado): 18.5 - 25; sobrepeso: 25 - 30; obesidad tipo I (obeso): 30 - 35; obesidad tipo II (obeso severo): 35 - 40; obesidad grado III (mórbido obeso): 40 - 45.

2.2. Procesamiento y limpieza

El conjunto de datos usado contenía valores faltantes (denominados NA en adelante). Si bien existen varios métodos de imputación para tratar los valores NA, se decidió omitir los mismos y no incluirlos en los posteriores análisis. La justificación para esto es que no es posible determinar la naturaleza del tipo de NA: los mismos fueron tomados de un repositorio público sin información adicional sobre el modo en el cual fue recopilada la información o indicaciones que justifiquen la presencia de algunas observaciones vacías. De esta manera, los registros que contenían al menos un valor faltante fueron removidos.

Se controló que la eliminación de los valores NA no alterara la distribución de las variables categóricas nacionalidad, BMI y sexo mediante el análisis de la tabla de contingencia con el *test* de χ^2, antes y después del procedimiento. La hipótesis nula considera que no existen diferencias entre las distribuciones mientras que la hipótesis alternativa las considera diferentes.

El conjunto de datos con el cual se inició el análisis consitía en 1056 registros completos con información de abundanciaa de bacterias y metadatos de los individuos.

2.3. PCA

PCA se usa comúnmente para la reducción de dimensionalidad mediante el uso de dato en solo los primeros componentes principales (en la mayoría de los casos, primera y segunda componente) para obtener datos de dimensiones inferiores y al mismo tiempo mantener la mayor variación posible de los datos.

El algoritmo fue implementado usando la función prcomp() del paqueteFactoMineR en R.

2.4. UMAP

UMAP es una técnica de reducción de dimensionalidad no lineal y se utiliza a menudo para visualizar conjuntos de datos de alta dimensión. El algoritmo fue ejecutado usando el paquete de R *umap*.

2.5. Agrupamiento *k-means*

K-means es uno de los algoritmos de agrupamiento particional más frecuentemente usado. En líneas generales, existen cuatro pasos principales: primero, especificar el valor de k. Luego, seleccionar al azar k objetos del conjunto de datos como los centros de los *cluster* iniciales. Asignar cada observación al centroide más cercano. Para cada *cluster*, actualizar el centroide mediante el cálculo de la media de los datos en el *cluster*. Por último, iterar hasta que las asignaciones a los *clusters* no sigan cambiando.

Para implementar el algoritmo, se usó la función estándar *kmeans* de R (del paquete *statspackage*). Se usaron las variables numéricas correspondientes a las abundancias de bacterias de los subconjuntos *Bacteroidetes* y *Firmicutes*.

2.6. SOM

Los mapas auto-organizados (SOM) son un tipo de red neuronal no supervisada que consiste en m neuronas o nodos totalmente conectados y organizados en un mapa bidimensional. Al momento de determinar el tamaño del mismo, se espera una distribución de observaciones relativamente uniforme a través de la grilla. Las dimensiones de la grilla de nodos y la tasa α de entrenamiento son los principales hiperparámetros que definen diferentes modelos en el SOM y en consecuencia, la cantidad y el tamaño de los agrupamientos en los datos. La configuración de la tasa de aprendizaje α define la velocidad de convergencia del algoritmo a una solución. Se exploraron distintas dimensiones de la grilla, además del número de iteraciones, para definir el SOM más apropiado. Finalmente se decidió por una topología hexagonal regular y se entrenó el SOM usando 49 neuronas, en grillas de 7 x 7 y 1000 iteraciones.

Se usaron las variables numéricas correspondientes a edad, diversidad y abundancias de bacterias (estas últimas fueron escaladas usando escala logarítmica previo al entrenamiento en el SOM). Las variables categóricas fueron convertidas en variables indicadoras para poder representar de forma numérica las distintas categorías en cada atributo.

La escala de colores en cada mapa va desde el azul (baja diversidad, edad joven o nivel bajo de abundancia de bacterias), pasando por el verde (valores medios) hasta llegar al rojo (máximo valor de diversidad, edad mayor o abundancia alta de bacterias). En el caso de la variable nacionalidad, el color azul corresponde a la categoría Europa Central, seguido en la escala por Escandinavia, Europa del Sur, Reino Unido/Irlanda y finalizando en color rojo correspondiente a Estados Unidos. Para la variable BMI, el color azul corresponde a bajo peso seguido en la escala por delgado, sobrepeso, obeso, obeso severo y finalizando en color rojo correspondiente a mórbido obeso.

2.7. *Random forest*

Random forest (RF) es una técnica de aprendizaje por ensamble usado en problemas de clasificación y regresión. En esta Tesis, el algoritmo de RF fue utilizado para generar un modelo de regresión para predecir la variable diversidad. Las variables numéricas de abundancia de

bacterias y los metadatos fueron usados como variables explicativas.

RF fue implementado mediante la herramienta DRF (*Distributed Random Forest*) provista por la interfase R de *h2o.ai* [h2o, 2020].

Se mantuvieron los valores por defecto de los hiperparámetros a excepción de N_{tree}, del cual se fueron buscando varios valores hasta converger en métricas de error estables.

No se consideró separar en conjuntos de entrenamiento y evaluación independientes sino que la evaluación del modelo de regresión surge de *out-of-bag*.

Capítulo 3

Análisis exploratorio de la población en estudio

La exploración de datos es un paso importante al iniciar un análisis de datos ya que permite comprender las características de los mismos, evaluar posibles relaciones, ver la integridad, identificar valores vacíos y valores atípicos (*outliers*), transformar variables, etc.

El objetivo de este capítulo es explorar la información asociada a los individuos en estudio e identificar los tipos de bacterias presentes para enfocar luego los datos a ser usados en los análisis posteriores.

3.1. Descripción de las variables de metadatos

La edad de los sujetos en estudio oscila entre los 18 a los 77 años, con una media de 45 años. Se observa que la población joven está distribuida de manera homogénea mientras que existe una predominancia de edades entre los 50-60 años. Por otra parte, se observa una baja representación de adultos mayores (mayores a 70 años). En cuanto a la representación de mujeres y hombres en el conjunto de datos, se observa que existe una distribución pareja a lo largo de todo el rango etario (Figura 3.1).

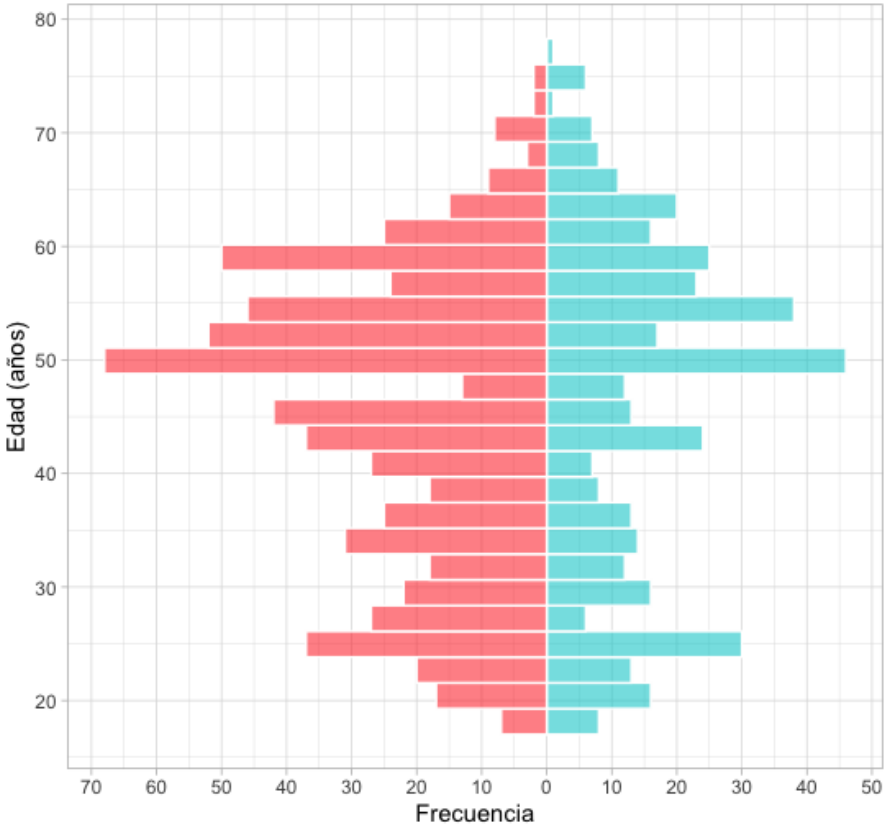

Figura 3.1: Distribución de hombres (celeste) y mujeres (coral) de acuerdo a la edad.

Cinco regiones europeas están representadas en el conjunto de datos: Europa del Sur, del Este y Central abarcan de manera extensa todo el espectro de edades observado en el conjunto de datos, Escandinavia tiende a concentrar individuos de edad media (alrededor de 50 años), mientras que Reino Unido/Irlanda concentra individuos jóvenes en su mayoría (entre 20 y 30 años). Por otra parte, en este conjunto de datos existe representación de una población no europea: Estados Unidos (US) abarca también edades jóvenes y adultas (Figura 3.2).

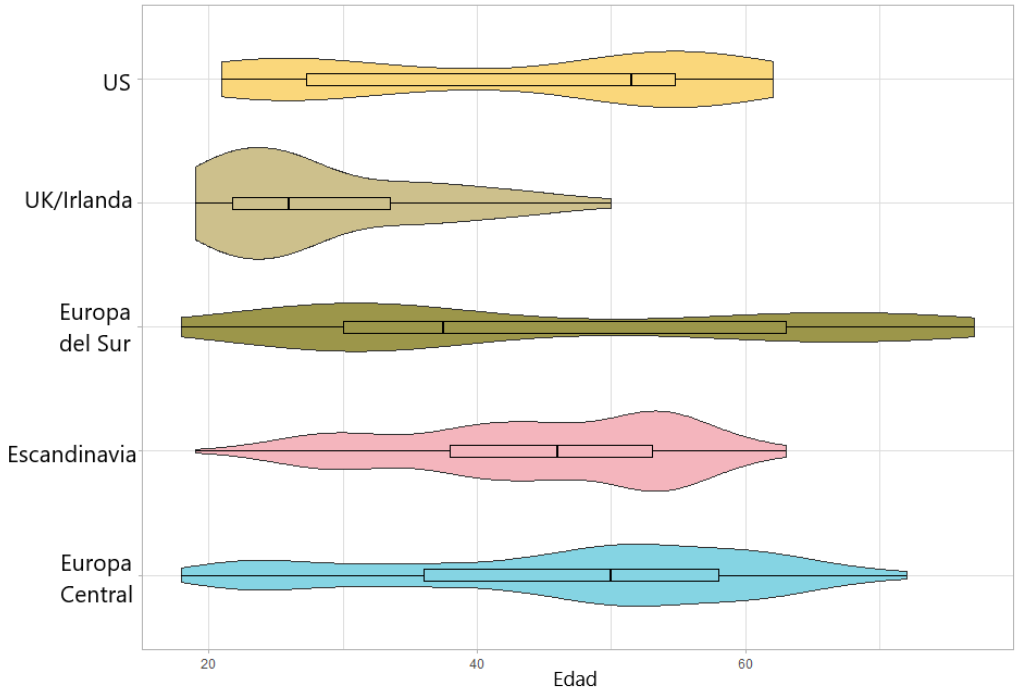

Figura 3.2: Distribución de las franjas etarias en las distintas regiones geográficas del conjunto de datos.

Una variable fisiológica que representa el estado nutricional de un individuo es el índice de masa corporal (BMI), definido por el peso (en kg) dividido por el cuadrado de la altura (en metros). La escala de valores de BMI están basados en el efecto que la grasa corporal en exceso tiene sobre el riesgo de enfermedad: a medida que aumenta el BMI, también aumenta el riesgo de algunas enfermedades [World Health Organization, 2021].

Los datos mostraron que los individuos delgados se distribuyeron de manera homogénea en todas las franjas etarias, mientras que los individuos con sobrepeso, obesos y obeso severos se concentraron entre los 45-60 años. Una gran proporción de población de bajo peso estuvo representada entre los 20-30 años de edad (Figura 3.3).

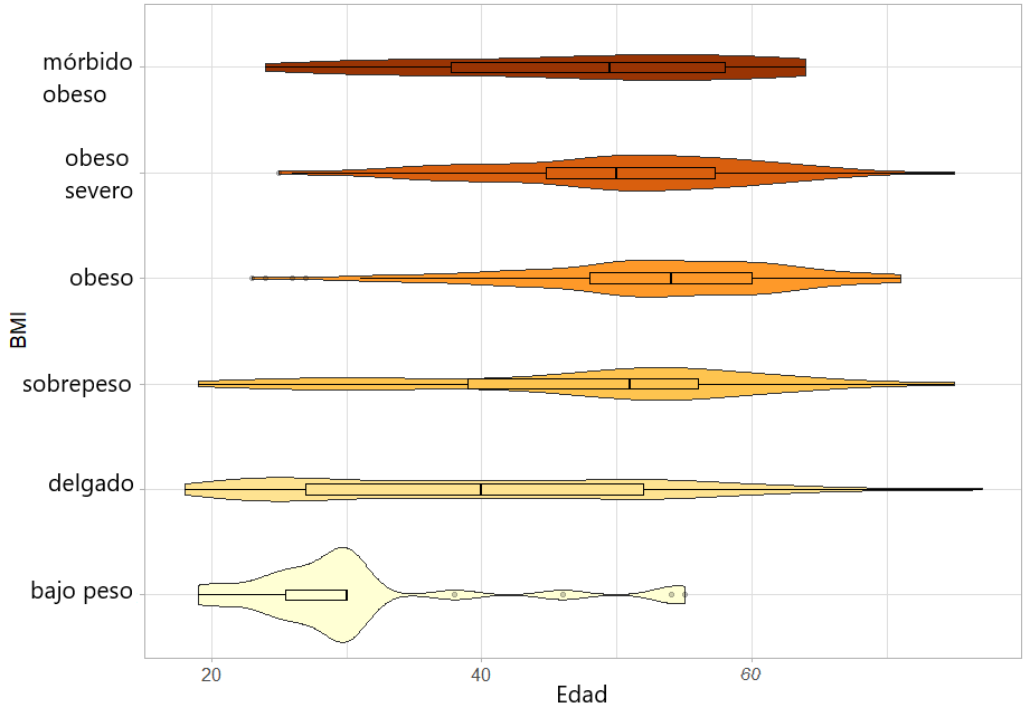

Figura 3.3: Distribución de las categorías de BMI en función de la edad.

La distribución de las diferentes categorías de BMI en cada región geográfica mostró una representación predominante de individuos delgados y proporciones variables de los restantes niveles de BMI. Tanto en Europa del Sur como en UK/Irlanda, los individuos delgados representaron un poco más de la mitad de esas subpoblaciones, seguidos en cantidad por individuos con sobrepeso. En cambio, en Europa Central, Escandinavia y US, la proporción de individuos delgados representó un poco menos de la mitad de esas subpoblaciones, seguidas por individuos obesos. En Europa Central se observó representación variable de todas las categorías de BMI (Figura 3.4).

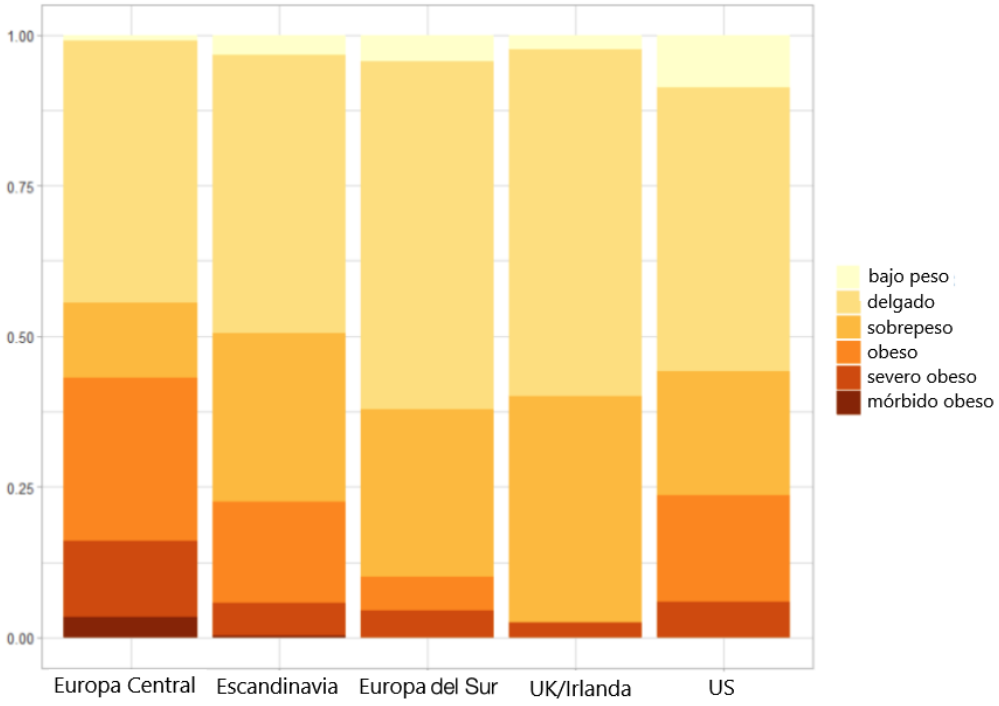

Figura 3.4: Proporción de índices de masa corporal (BMI) de acuerdo a la región geográfica.

El conjunto de datos incluye además una variable que representa la diversidad en términos del índice de *Shannon*. En el análisis de comunidades ecológicas, la diversidad es una propiedad ampliamente estudiada porque tiene en cuenta la riqueza (la cantidad de clases) y uniformidad (la distribución de individuos entre las clases) de los miembros. El entendimiento de la diversidad en la comunidad microbiana intestinal permite además comprender el impacto de diversos factores tales como el uso de antibióticos, tipo de dieta, grado de obesidad, intervenciones médicas y factores ambientales, entre otros (revisado en [Reese and Dunn, 2018]). De las seis categorías de BMI presentes en el conjunto de datos, no hay diferencias significativas entre los individuos delgados y los de bajo peso. Sin embargo, hay diferencias entre los individuos delgados y las restantes categorías de individuos con exceso de peso (Figura 3.5).

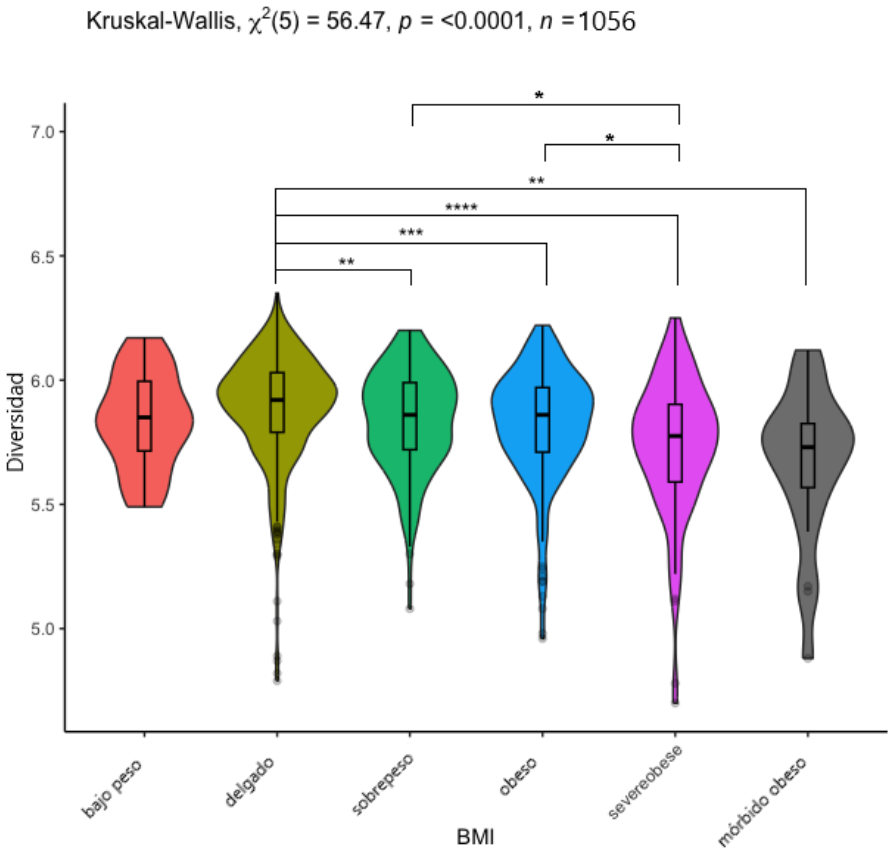

Figura 3.5: Distribución de la diversidad bacteriana en relación a las categorías de BMI. Se usó el *test* de *Kruskal-Wallis* para determinar diferencias entre las medias de diversidad para los grupos. Se corrigieron los valores de p mediante el método de *Holm*. * p <0.05, ** p <0.01, *** p <0.001, **** p <0.0001

3.2. Detección de valores faltantes

El problema de la existencia de datos faltantes en conjuntos de datos reales es relativamente común y puede tener un efecto significativo en las conclusiones a extraer. El análisis inicial del conjunto de datos total mostró que la edad, nacionalidad, sexo y BMI fueron las únicas variables que presentaron valores faltantes (NAs). El conjunto de datos contiene una cantidad de 231 NAs de los cuales el 9 % está representado por la variable BMI, alrededor del 5 % por Edad y aproximadamente el 3 % por Sexo y Nacionalidad (Cuadro 3.1).

Variable	Cantidad	%
BMI	106	9
Edad	56	4.7
Sexo	37	3.1
Nacionalidad	32	2.7

Cuadro 3.1: Variables del conjunto de datos original que presentan valores faltantes.

Dichos valores faltantes se distribuyeron de manera combinada en las variables mencionadas: de un total de 1172 invididuos, 49 de ellos tenían valores vacíos solamente en la variable BMI; 22 individuos presentaron valores faltantes en combinación con BMI, Edad, Sexo y Nacionalidad; 19 individuos presentaron valores faltantes combinados en BMI y Edad; 15 casos faltantes de manera conjunta entre BMI, Edad y Sexo; 9 casos presentaron valores faltantes en la variable Nacionalidad solamente y 1 caso presentó valor faltante de manera conjunta en las variables BMI y Nacionalidad (Figura 3.6). En total, 115 individuos presentaron algún valor faltante.

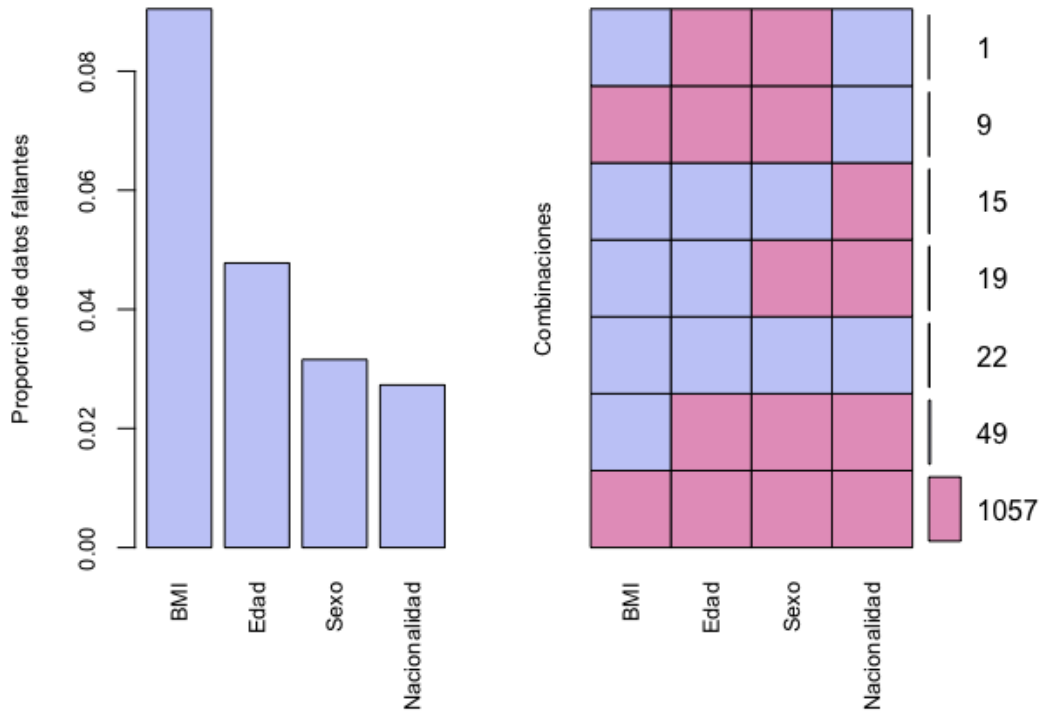

Figura 3.6: Proporción de valores faltantes en las variables BMI, edad, sexo y nacionalidad (izquierda). El gráfico de la derecha indica las combinaciones de las cuatro variables con valores vacíos (en violeta) y valores completos (rosa).

Se decidió no considerar valores NA en los posteriores análisis y trabajar entonces con un conjunto de datos completo. Una manera de controlar que el impacto de la eliminación de 115 NAs no afecte al conjunto total, es decir que no produzca ningún sesgo, es mediante la comparación de las variables antes y después del procedimiento. Se observó que la remoción de los 115 datos vacíos no alteró las categorías de distribución de las variables Nacionalidad, BMI ni Sexo, a excepción de la categoría "Europa del Este" que quedó representado solamente por un único registro completo (Cuadro 3.2). Dado que dicha categoría ya no es representativa con un único registro en el conjunto total, se decidió eliminarla de los análisis posteriores.

Categoría	N original	N sin NA
Nacionalidad		
Europa Central	650	617
Europa del Este	15	1
Escandinavia	291	275
Europa del Sur	90	90
Reino Unido/Irlanda	50	40
Estados Unidos	44	34
NA	32	0
BMI		
Delgado	493	487
Mórbido obeso	22	22
Obeso	224	224
Sobrepeso	204	201
Obeso severo	100	100
Bajo peso	23	23
Sexo		
Femenino	680	646
Masculino	455	411

Cuadro 3.2: Distribución de individuos antes (N original) y después (N sin NA) de la eliminación de los registros que presentaban campos vacíos. No hubo diferencias significativas entre ambos grupos para Nacionalidad (χ^2 gl = 4, 1.52, p = .82), BMI (χ^2 gl = 5, .02, p = 0.91) ni Sexo (χ^2 gl = 1, 0.33, p = .56). Tampoco se observaron diferencias entre ambos grupos para Edad (p >.05, *t test*).

Resumiendo, el conjunto de datos a utilizar en los siguientes capítulos considerará $N = 1056$ registros completos.

3.3. Selección de grupos de bacterias

La inspección inicial del conjunto de datos indicó que los seis *phyla* de bacterias (mencionados en la Sección 1.1.1 de Introducción) están representados en los 97 grupos de bacterias del *microarray* filogenético: 13 tipos de bacterias pertenecen a *Bacteroidetes*, con predominancia de los géneros *Bacteroides* y *Prevotella* (ver Anexo B.1). En cuanto a *Firmicutes*, se identificaron 42 miembros (ver Anexo B.2). *Actinobacteria* está representado por 8 miembros, mientras

que *Fusobacteria* y *Verrucomicrobia* tienen un único miembro cada uno. Los restantes 32 tipos pertenecen al *phylum Proteobacteria* (Figura 3.7).

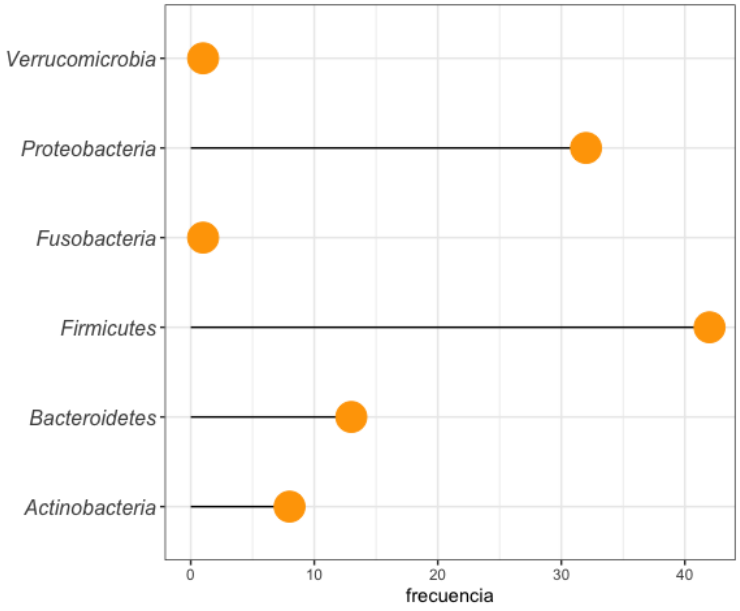

Figura 3.7: Frecuencias de los *phyla* presentes en el conjunto de datos.

De los grupos mencionados, se decidió seleccionar a las bacterias pertenecientes a *Bacteroidetes* y *Fimicutes* para que formen parte, junto a los metadatos, del conjunto de datos a ser analizado en los capítulos siguientes.

Además, se decidió no incluír a bacterias cuyas abundancias relativas son bajas o raras (indicadas en un trabajo previo con el mismo conjunto de datos [Lahti et al., 2014]). De esta manera, los datos a analizar en los capítulos siguientes está formado (además de los metadatos) por las abundancias de 13 tipos de bacterias de *Bacteroidetes* y 42 tipos de bacterias pertenecientes a *Firmicutes*, ambos para una población de individuos de $N = 1056$.

3.4. Conclusiones

La decisión de enfocar el análisis en dos grupos principales de bacterias, *Bacteroidetes* y *Firmicutes*, está basada en la importancia de ambos en el ámbito biomédico:

- Numerosos estudios indican que el microbioma intestinal está compuesto en más de un 90 % por especies pertenecientes a *Firmicutes* y *Bacteroidetes*, mientras que los restantes grupos tienen muy baja predominancia ([Qin et al., 2010, Arumugam et al., 2011].

- La relación *Firmicutes*/*Bacteroidetes* (F/B) es un índice ampliamente aceptado como indicador de disbiosis, cuyas variaciones están asociados a condiciones de salud tal como obesidad, inflamación y cáncer, entre otras ([Stojanov et al., 2020, An et al., 2023]).

- Por otra parte, los miembros de *Proteobacteria* tienen baja abundancia en el instestino humano saludable [Shin et al., 2015], por lo cual se decidió no considerarlos.

En cuanto a la decisión de remover los casos con NAs, se tuvo en cuenta que el conjunto de datos fue tomado de un repositorio público y que no existe información adicional sobre el modo en el cual fue recopilada la información o indicaciones que justifiquen la presencia de algunos campos vacíos. Por lo tanto no es posible determinar la naturaleza del tipo de dato faltante en los datos, lo cual limita la decisión de usar algún método de imputación apropiado.

Capítulo 4

Reducción de la dimensionalidad y agrupamiento

La cantidad de variables en un conjunto de datos representa la dimensionalidad del mismo. Uno de los primeros pasos en el análisis es caracterizar la composición y estructura de las variables. En principio, un conjunto de datos multivariado puede incluir la posiblidad de que la mayoría estén correlacionadas, lo que implica redundancia de información. Además, dado que es muy difícil visualizar o hacer predicciones con un alto número de variables, se puede recurrir a la reducción de la dimensionalidad.

La reducción de la dimensionalidad es una técnica que reduce el espacio de variables mediante la selección de algunas de ellas o de nuevas variables como combinaciones de las originales, preservando la mayor cantidad de información posible. Dependiendo del tipo de datos, existen diversos métodos que pueden ser divididos principalmente en lineales y no lineales. En este capítulo se usará PCA y UMAP como representantes de ambos grupos.

Por otra parte, el análisis de agrupamiento se enfoca en la reducción del espacio de filas mediante la detección de grupos de observaciones con características similares. El método de agrupamiento a usar es *k-means*.

El objetivo de este capítulo es realizar un análisis combinado de reducción de la dimensionalidad (mediante un enfoque lineal y otro no lineal) seguido de un análisis de agrupamiento sobre los datos reducidos. Para ello, se hace un análisis por separado a los subconjuntos de *Bacteroi-*

detes y *Firmicutes*, representados en dos matrices de datos (las columnas son las abundancias de bacterias y las filas son los individuos).

4.1. Análisis de Componentes Principales

El análisis de componentes principales (PCA) calcula un nuevo conjunto de variables (denominadas PC, componentes principales) las cuales representan la misma cantidad de información que las variables originales. La varianza total del sistema se redistribuye entre los PC de forma tal que el primer PC explica la mayor parte de la varianza. Dado que no existe una regla específica para seleccionar una determinada cantidad de PC, la reducción de la dimensionalidad de un conjunto de datos implica seleccionar de manera arbitraria en base a dos criterios:

- Un primer criterio se enfoca en considerar una determinada proporción de varianza acumulada. Este criterio es puramente empírico, dependiendo de cuánta información se busca conservar y el grado de interpretabilidad de los datos por parte de los PC seleccionados. El gráfico de sedimentación (descripto en la sección 1.3.1 de la Introducción) brinda una herramienta visual, porque el punto de quiebre (codo) en la curva descendiente da una aproximación del número de PC a seleccionar.

- Un segundo criterio es mantener aquellos componentes cuyos autovalores están por encima del promedio, es decir, mayores a 1 ([Abdi and Williams, 2010]).

PCA también posee una importante capacidad de análisis, enfocado en la contribución de cada variable sobre los PC, magnitud y dirección de los coeficientes de las variables originales (también llamados *loadings*). Cuanto mayor sea el valor absoluto del coeficiente, más importante será la variable correspondiente en el cálculo del componente. Una herramienta gráfica que contribuye a la evaluación exploratoria es el *biplot*, el cual permite visualizar simultáneamente las direcciones de las variables y las observaciones dentro de un marco de componentes principales.

El interés en usar PCA sobre los datos de *Bacteroidetes* y *Firmicutes* es identificar cuántas bacterias aportan a la varianza. En otras palabras, identificar qué bacterias aportan en distinto grado a la dinámica de las relaciones en cada grupo.

4.1.1. *Bacteroidetes*

Luego del análisis de PCA sobre la matriz de datos de *Bacteroidetes* se observa que, si bien la varianza del sistema se concentra en los dos primeros PC, estos representan de manera conjunta apenas el 33 % de la varianza total (Cuadro 4.1).

PC	Autovalor	Porcentaje de varianza	% Varianza acumulada
1	2,74	21,07	21,07
2	1,54	11,90	32,97
3	0,98	7,62	40,60
4	0,94	7,24	47,84
5	0,91	7,02	54,86
6	0,87	6,74	61,61
7	0,82	6,35	67,97
8	0,81	6,21	74,19
9	0,80	6,16	80,36
10	0,73	5,64	86,00
11	0,70	5,44	91,44
12	0,65	5,04	96,48
13	0,45	3,51	100

Cuadro 4.1: Aporte de cada componente principal a la varianza total en el subconjunto *Bacteroidetes*.

La selección del número de PC para la reducción dimensional puede basarse en el criterio de los autovalores mayores a 1: se observa que PC1 y PC2 satisfacen este criterio (2.74 y 1.54, respectivamente). Por otra parte, según el criterio en base al gráfico de sedimentación, se observa un cambio en la pendiente entre PC2 y PC3, sugiriendo considerar 2 o 3 componentes (Figura 4.1).

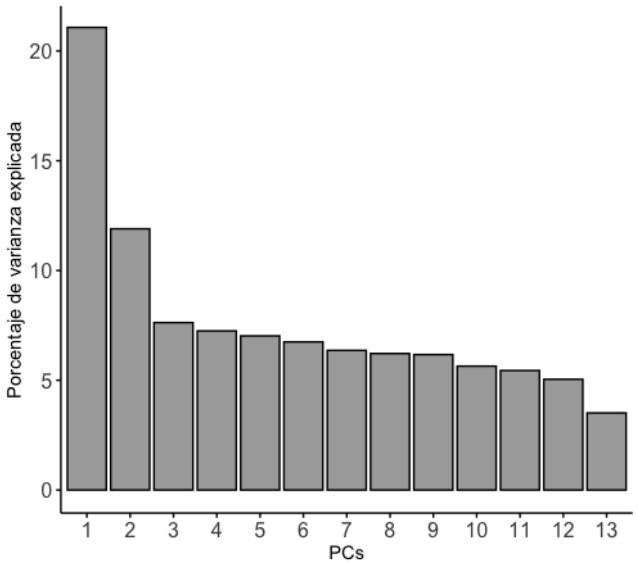

Figura 4.1: Gráfico de sedimentación de los autovalores de cada componente principal (PC) para el subconjunto de *Bacteroidetes*.

La idea de hacer un análisis de PCA es poder concentrar la mayor proporción de varianza de los datos en la menor cantidad de componentes principales. Sin embargo, los resultados muestran que para explicar un porcentaje representativo (por ejemplo, el 80 %, un porcentaje frecuentemente usado) se necesitan al menos nueve PC (Cuadro 4.1), una cantidad elevada para poder ser interpretada. Entonces, de acuerdo a los criterios basados en el gráfico de sedimentación y autovalores mayores a 1, se decide **seleccionar los dos primeros PC de *Bacteroidetes* que explican el 33 % de la varianza total. Los mismos serán usados en un análisis posterior de *clustering***.

En cuanto al aspecto exploratorio de PCA, la inspección de la magnitud, correlación y el signo de las coordenadas de las variables originales permiten determinar el grado de contribución de las mismas sobre los componentes PC1 y PC2 seleccionados. Las columnas del Cuadro 4.2) muestran una combinación lineal de las variables originales: el primer PC contiene la combinación $PC1 = 0,31 \times \textit{Allistipes} + 0,33 \times \textit{B.fragilis} + ... + 0,29 \times \textit{Tannerella}$.

La información sobre la magnitud, dirección y correlación de las variables originales puede ser mejor representada en un *biplot*, que brinda una mejor inspección. De esta manera, el *biplot* para PC1 y PC2 indica que:

- PC1: Todas las bacterias muestran una dirección positiva aunque con diferentes magnitudes. Las bacterias del género *Bacteroides* tienen una máxima influencia en este compo-

Variable	PC1	PC2
Allistipes	0,31	-0,05
Bacteroides fragilis	0,33	0,06
Bacteroides ovatus	0,26	-0,05
Bacteroides plebeius	0,25	0,06
Bacteroides splachnicus	0,24	0,02
Bacteroides stercoris	0,20	-0,02
Bacteroides uniformis	0,33	**-0,16**
Bacteroides vulgatus	**0,38**	-0,01
Parabacteroides distasonis	0,28	0,03
Prevotella melaninogenica	**0,06**	**0,69**
Prevotella oralis	0,09	0,68
Prevotella tannerae	0,32	-0,03
Tannerella	0,29	-0,04

Cuadro 4.2: Coordenadas de las variables de *Bacteroidetes* para los dos primeros componentes principales. En negrita se indican los valores máximos y mínimos obtenidos para cada PC.

nente (*B. vulgatus* es la de mayor efecto). En cambio, las bacterias del género *Prevotella* muestran una contribución menor (Figura 4.2, eje horizontal). Podría interpretarse entonces que **PC1 podría servir como una dimensión representativa del género *Bacteroides*** (con la excepción de *P. tannerae*).

- PC2: Las bacterias del género *Prevotella* tienen una influencia predominante en PC2 (*P. melaninogenica* y *P. oralis* fueron las de mayor aporte). En cambio, varias bacterias del género *Bacteroides* tuvieron una contribución casi nula o negativa (Figura 4.2, eje vertical). Esto sugiere que el componente **PC2 podría servir como una dimensión representativa del género *Prevotella***.

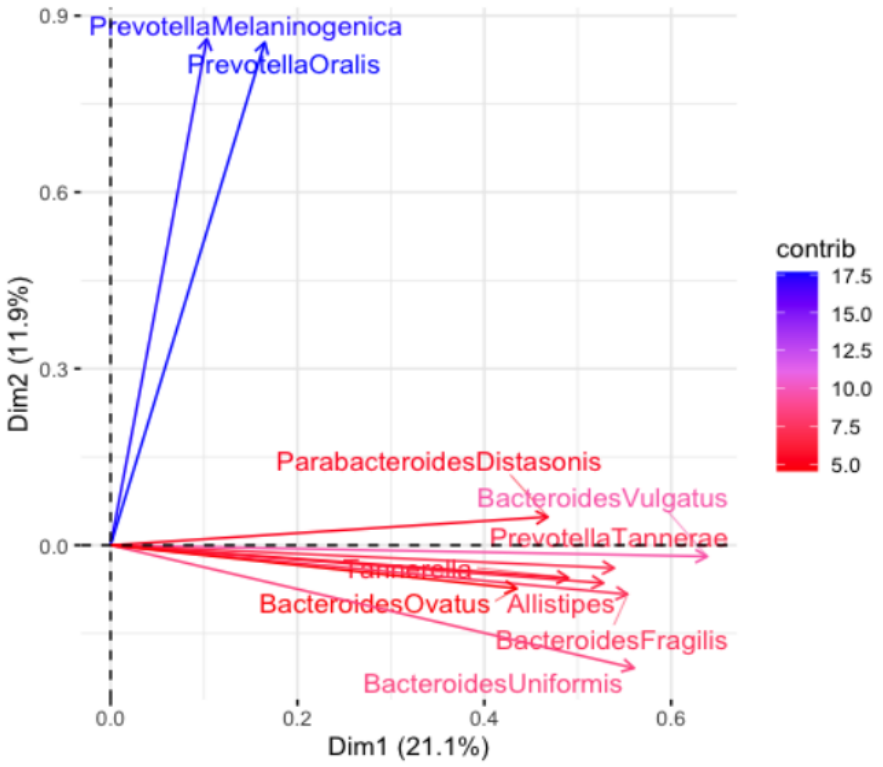

Figura 4.2: *Biplot* de correlación y dirección de variables de *Bacteroidetes* sobre PC1 y PC2. El gradiente de color indica el grado de contribución.

El grado de contribución de las variables descriptas en los *biplots* permite identificar las características que definen a cada componente principal. Esta información también puede ser presentada en otro formato de visualización: un gráfico de barras que indica el porcentaje de contribución de las variables en orden descendiente sobre cada PC (Figura 4.3).

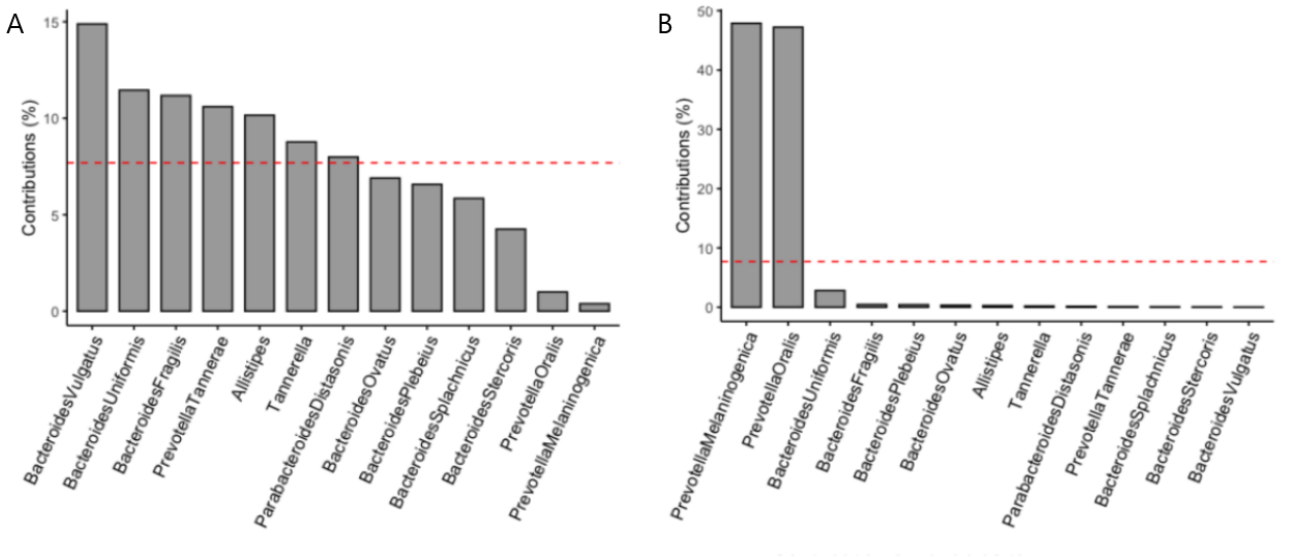

Figura 4.3: Contribuciones de las variables de *Bacteroidetes* en los dos primeros PCs. La línea punteada indica la contribución promedio esperada. Si el aporte de las variables fuera uniforme, el valor esperado sería 7.7 % (calculado como 1/cantidad de variables). Para un PC determinado, se considera importante a una variable con una contribución mayor a este punto de corte.

4.1.2. *Firmicutes*

El análisis de PCA sobre la matriz de datos de *Firmicutes* muestra que, aunque la varianza del sistema se concentra en los dos primeros PC, éstos representan solamente un 11 % de la varianza total (ApéndiceC).

La selección del número de PC para la reducción dimensional en base al gráfico de sedimentación no sugiere ningún punto de quiebre destacado en la curva descendiente gradual, lo cual no permite considerar ningún componente principal (Figura 4.4). Sin embargo, si se considera el criterio de mantener aquellos PC cuyos autovalores sean mayores a 1, quedan seleccionados los primeros quince componentes principales (ver Apéndice C). Estos quince PC representan en conjunto el 47.6 % de la varianza total.

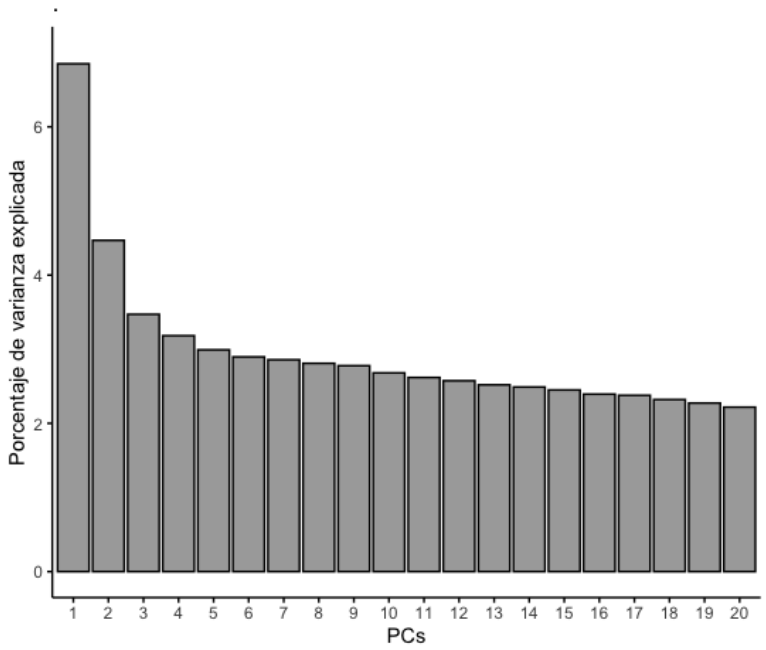

Figura 4.4: Gráfico de sedimentación de los autovalores de cada componente principal (PC) para el subconjunto de *Firmicutes*. Se grafican solamente los primeros 20 PC.

Tal como se mencionó anteriormente, uno de los principales objetivos de un análisis de PCA es concentrar la mayor proporción de varianza de los datos en la menor cantidad de PC. El Apéndice C muestra que para explicar un porcentaje significativo de la información, tal como 70 % o 80 %, se necesitan al menos 25 a 30 componentes principales respectivamente, una cantidad difícil de manejar y explicar simultáneamente. De esta manera, de acuerdo al criterio basado en los autovalores mayores a 1, se decide **seleccionar los primeros quince PC de *Firmicutes*. Los mismos serán usados posteriormente en un análisis de *clustering***.

En cuanto al enfoque exploratorio de PCA, el análisis que examina la magnitud y dirección de los coeficientes de las variables originales sobre los quince componentes seleccionados se torna forzado de describir, debido a que la varianza explicada por los primeros componentes es muy baja. Por lo cual, cualquier interpretación del efecto de las variables originales de *Firmicutes* sobre el espacio de los quince PC seleccionados carece de significado.

4.2. UMAP

Teniendo en cuenta que la implementación de PCA no resultó eficiente al momento de conservar la mayor cantidad de información posible luego de la reducción dimensional, se elige

usar un método de reducción no lineal denominado UMAP (*Uniform Manifold Approximation and Projection*). A diferencia de PCA (que realiza una descomposición de autovalores centrado en las relaciones lineales de los datos) UMAP está enfocado en captar estructuras no lineales en los datos multidimensionales preservando la estructura subyecente de los mismos.

Para la aplicación de UMAP se evalúan distintos valores de los dos hiperparámetros más frecuentemente usados: el número de vecinos (*n_ neighbours*) y la distancia mínima (*min_ dist*), cuyos ajustes permiten manejar un equilibrio entre las estructuras locales y preservar la estructura global. El resultado visual de UMAP se presenta en dos dimensiones (el valor por defecto).

El enfoque que se considera es crear una grilla con algunas combinaciones posibles de valores para los hiperparámetros mencionados usando ambos subconjuntos de *Bacteroidetes* y *Firmicutes*.

4.2.1. *Bacteroidetes*

La reducción de la dimensionalidad no lineal es realizada usando diferentes combinaciones de *n_ neighbours* y *min_ dist*: las columnas indican valores de *min_ dist* desde 0.05 a 1 mientras que las filas indican valores para *n_ neighbours* comenzando en 5 hasta 50 (Figura 4.5).

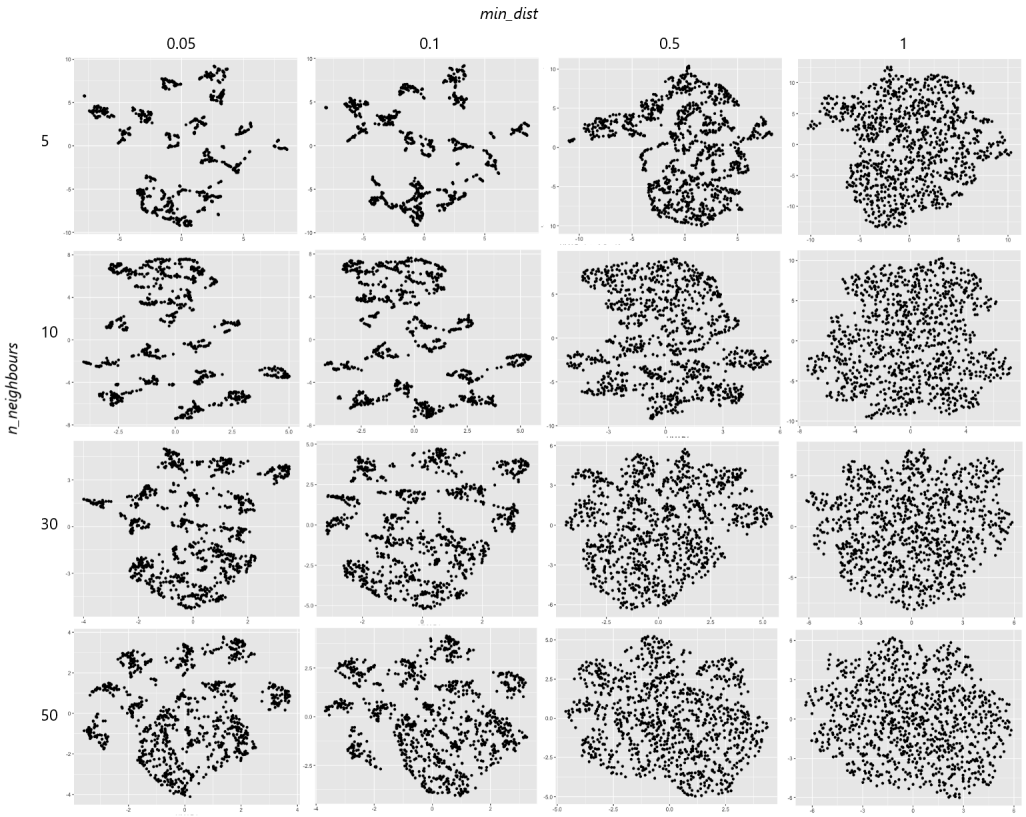

Figura 4.5: Combinaciones de los hiperparámetros de UMAP, *n_ neighbours* (filas) y *min_ dist* (columnas), para el subconjunto de *Bacteroidetes*.

Se observa que a medida que aumenta *n_ neighbours*, más puntos vecinos se van conectando entre sí, lo cual mejora la estructura de los datos proyectados en grupos definidos. Por otra parte, *min_ dist* controla cuán apartados están los puntos unos de otros: a mayor valor de este parámetro, los puntos están más dispersos enfocados en la preservación de la estructura topológica general.

Luego de la inspección visual de las combinaciones, **se elige la proyección UMAP correspondiente a los valores *min_ dist* = 0.05 y *n_ neighbour* = 5** porque se considera que son una buena representación de los datos de *Bacteroidetes*. **Esta proyección UMAP seleccionada es usada en un análisis posterior de *clustering***.

4.2.2. *Firmicutes*

Del mismo modo, la reducción dimensional no lineal de *Firmicutes* fue realizada con diferentes combinaciones de *n_ neighbours* y *min_ dist* (Figura 4.6). Se observa que a medida que aumenta *n_ neighbours*, la proyección de los datos va adquiriendo gradualmente una estructura

más visible. Sin embargo, a distintos valores de *min_ dist* no parecen observarse cambios en la estructura.

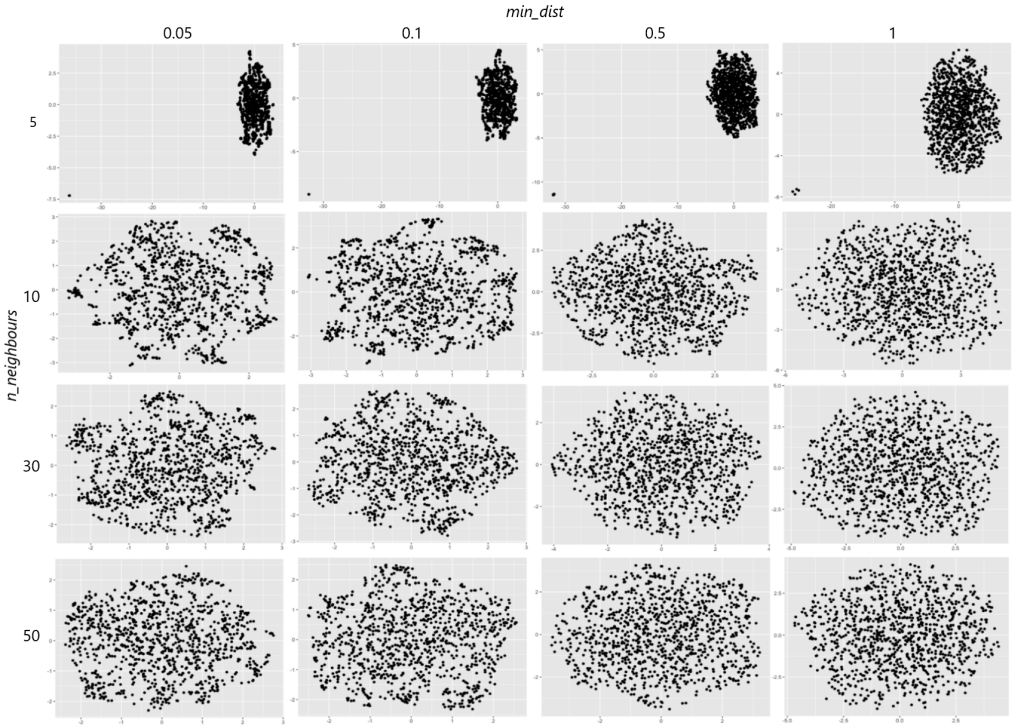

Figura 4.6: Combinaciones de los hiperparámetros de UMAP, *n_ neighbours* (filas) y *min_ dist* (columnas), para el subconjunto de *Firmicutes*.

A partir de la inspección visual de la grilla de combinaciones, **se selecciona la proyección UMAP correspondiente a *min_ dist* = 0.05 y *n_ neighbour* = 10. Esta proyección UMAP de *Firmicutes* es usada en un análisis posterior de *clustering***.

4.3. *Agrupamiento* con *k-means*

Uno de los métodos de agrupamiento más usado es *k-means*, un algoritmo de partición que requiere predefinir el número de grupos (k). Para determinar el valor óptimo de k se consideran tres métricas (descriptos en la Introducción (Sección 1.3.3.3). Brevemente, la primera está basada en minimizar la variación total intra-grupos (WCSS por *Within-Cluster Sum of Squares*), para medir cuán lejanos están los datos de un *cluster* de su respectivo centroide (valores altos de WSS indican mayor dispersión, valores bajos indican *clusters* más compactos). La segunda es el coeficiente de *Silhouette* (S), que mide la calidad del *cluster*. El valor óptimo de k es aquel que maximiza el coeficiente S. Mientras la tercera es el coeficiente de *Gap*, cuyo

valor máximo significa que la estructura del *clustering* es muy diferente a una distribución uniforme al azar.

A continuación, se describen los resultados de aplicar *k-means* a las variables originales de abundancias de bacterias *Bacteroidetes* y *Firmicutes*. Además, luego de la reducción dimensional con PCA y UMAP, se realiza un análisis de *clustering* sobre los resultados seleccionados de ambos con la idea de complementar la exploración y análisis de los datos.

Sin embargo, antes de la implementación de *k-means* sobre cada uno de los datos mencionados, es importante determinar la tendencia de los mismos a formar *clusters*. Para tal fin, se evaluará el valor del estadístico de *Hopkins* (H) descripto en la sección 1.3.4. El Cuadro 4.3 indica que los datos de *Bacteroidetes*, tanto los originales como los obtenidos de la reducción dimensional de PCA y UMAP, son agrupables (los valores H = 0.8 - 0.9 están por encima del valor umbral de 0.5). Mientras que los datos de *Firmicutes* originales y los obtenidos por PCA (y en menor medida UMAP) no son agrupables (H = 0.6 - 0.7, cercanos al valor umbral de 0.5).

	Original		PCA		UMAP	
	Bacteroidetes	*Firmicutes*	*Bacteroidetes*	*Firmicutes*	*Bacteroidetes*	*Firmicutes*
H	0.8	0.6	0.8	0.6	0.9	0.7

Cuadro 4.3: El estadístico H indica la tendencia de los datos a formar *clusters*.

Aunque los datos de *Firmicutes* no muestran tendencia al agrupamiento, de todos modos se aplicará *k-means* para corroborar el estadístico H.

4.3.1. *Bacteroidetes*

4.3.1.1. Datos originales

El conjunto de datos consiste en las 13 variables originales correspondientes a las abundancias de bacterias. La determinación del valor óptimo de k mediante la inspección de la curva WCSS sugiere un punto destacable de inflexión a partir de $k = 3$ (Figura 4.7 A). Por otra parte, la curva S indica que se obtiene un valor máximo del coeficiente S para $k = 2$ (Figura 4.7 B). Finalmente, la curva del estadístico de *Gap* no indica de manera clara un valor de k máximo para seleccionar (Figura 4.7 C). De acuerdo a estas observaciones, se toma $k = 2$ como el número óptimo de *clusters* para implementar *k-means*. Los resultados indican que los dos *clusters* obtenidos están conformados por $N = 657$ y $N = 399$ observaciones cada uno, siendo el segundo grupo más compacto en su distribución. Se observan también varios valores *outliers* (Figura 4.7 D).

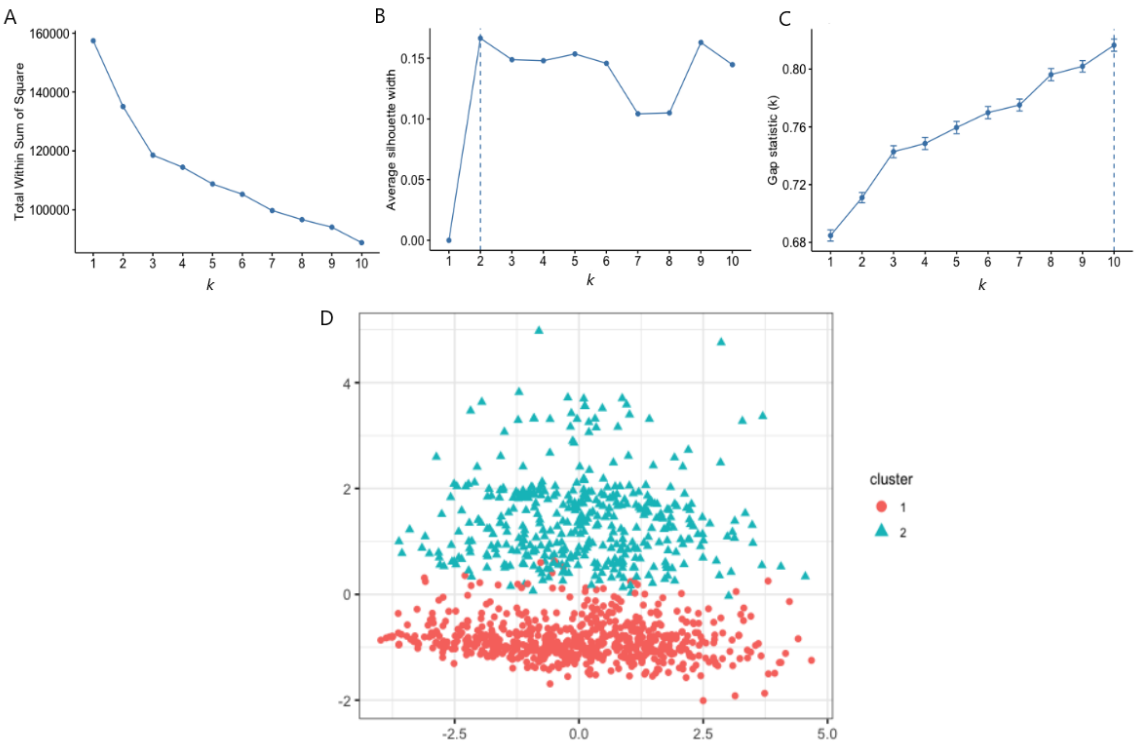

Figura 4.7: Métricas para determinar el valor k para *k-means* sobre los variables originales de *Bacteroidetes*. (A) Método de *elbow*, (B) Coeficiente de *Silhouette* y (C) Método de *Gap*. (D) *Clustering* mediante *k-means* con $k = 2$.

4.3.1.2. Componentes seleccionados de PCA

Los datos usados consisten en dos componentes principales obtenidos por análisis de PCA (PC1 y PC2 representan el 33 % de la varianza total). Para la determinación del valor óptimo de k, la curva WCSS sugiere que a partir de $k = 3$ o 4 se observa una diminución no tan pronunciada del valor de WCSS (Figura 4.8 A). Por otra parte, la curva S indica que se obtiene un valor máximo del coeficiente S para $k = 3$ (Figura 4.8 B). La curva del estadístico de *Gap* muestra que para $k = 3$ se obtiene un valor máximo del mismo (Figura 4.8 C). De acuerdo a estas observaciones, se toma $k = 3$ como el número óptimo de *clusters* para aplicar *k-means* en PC1 y PC2 conjuntamente. Los resultados indican que los tres *clusters* obtenidos no están separados entre sí, conformados por $N = 297$, 382 y 377 observaciones cada uno (Figura 4.8 D).

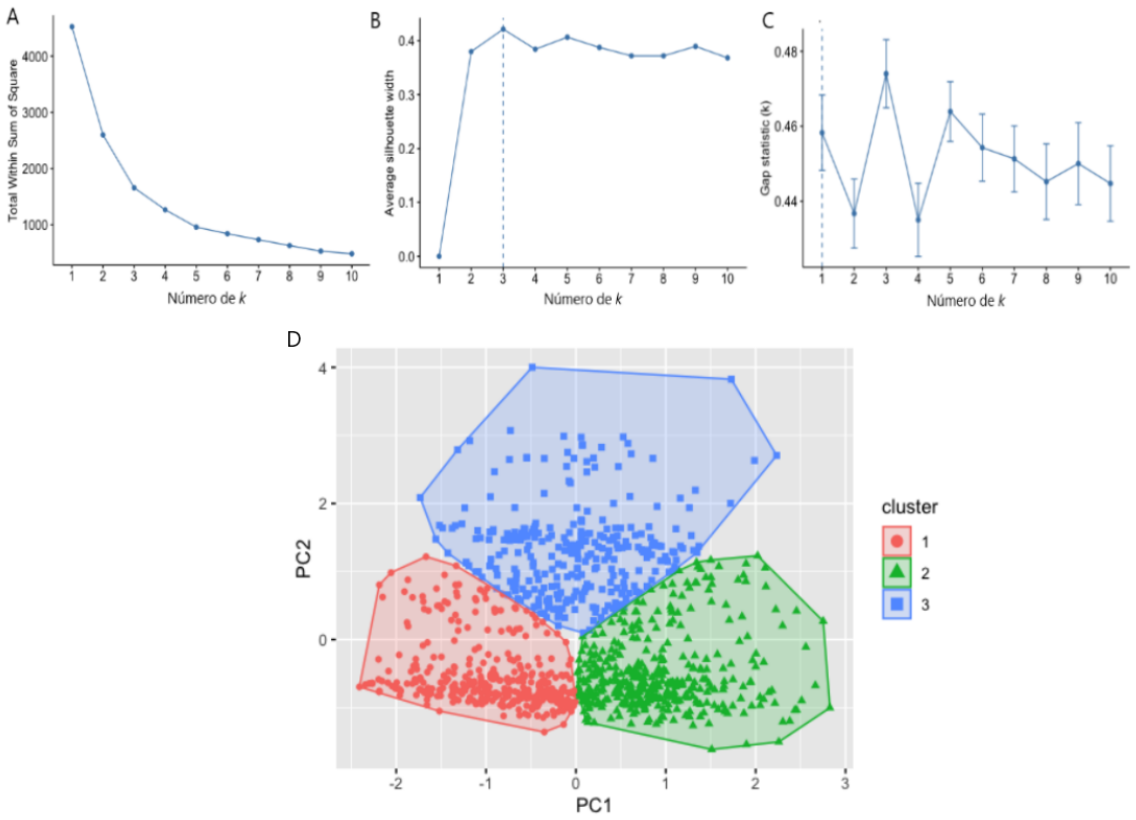

Figura 4.8: Métricas para determinar el valor k para *k-means* sobre los componentes PC1 y PC2 de *Bacteroidetes*. (A) Método de *elbow*, (B) Coeficiente de *Silhouette* y (C) Método de *Gap*. (D) *Clustering* mediante *k-means* con $k = 3$.

4.3.1.3. Proyección seleccionada de UMAP

Los datos usados consisten en la reducción UMAP de 2 dimensiones usando los parámetros $n\_neighbours = 5$ y $min\_dist = 0.05$. La determinación del valor óptimo de k mediante la inspección de la curva WCSS sugiere un punto de inflexión a partir de $k = 4$ (Figura 4.9 A). En tanto que la curva S muestra un valor máximo del coeficiente S para $k = 4$ (Figura 4.9 B). Finalmente, la curva del estadístico de *Gap* no indica de manera clara un valor de k máximo para seleccionar (Figura 4.9 C). De acuerdo a estas observaciones, se toma un valor de $k = 4$ como el número óptimo de *clusters* para implementar *k-means*. Los resultados indican que los cuatro *clusters* están separados entre sí, conformados por $N = 210$, 230, 278 y 338 observaciones cada uno (Figura 4.9 D).

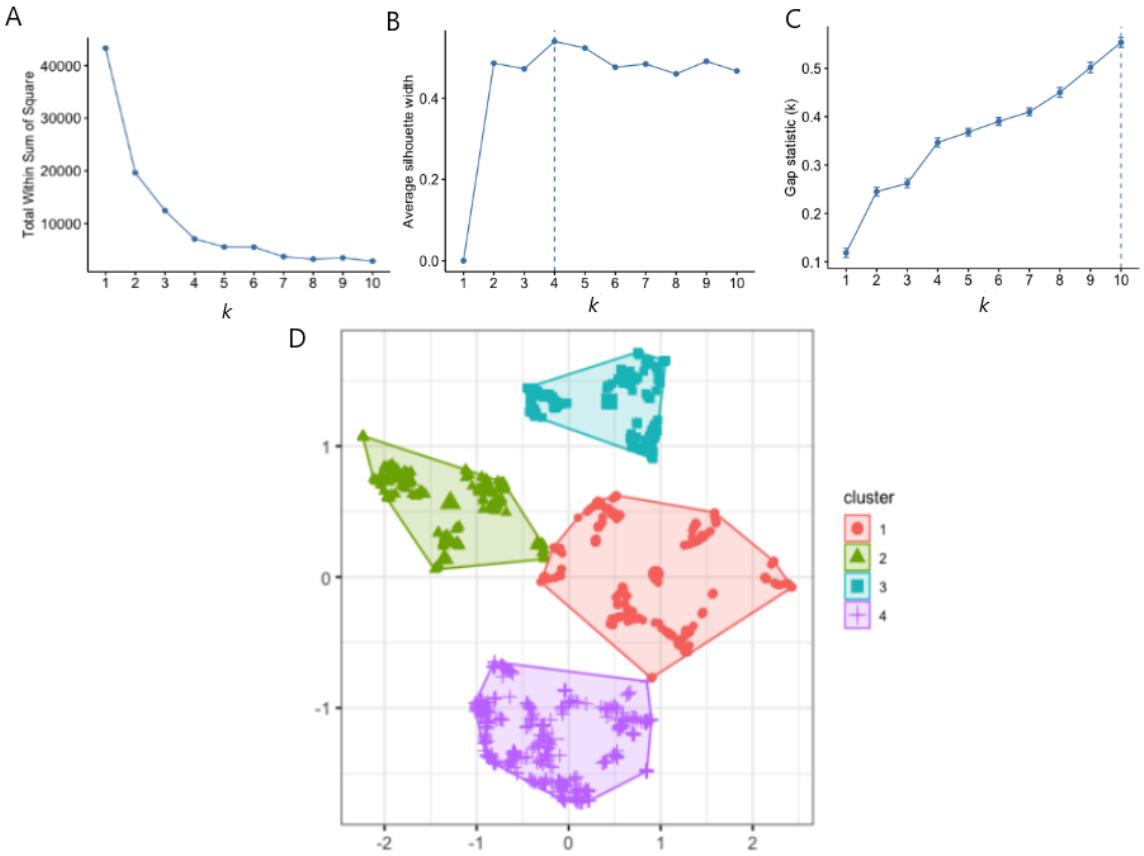

Figura 4.9: Métricas para determinar el valor k para aplicar *k-means* sobre la proyección UMAP de *Bacteroidetes*.(A) Método de *elbow*, (B) Coeficiente de *Silhouette* y (C) Método de *Gap*, (D) *Clustering* mediante *k-means* con $k = 3$

.

4.3.2. *Firmicutes*

4.3.2.1. Datos originales

El conjunto de datos consiste en las 42 variables originales correspondientes a las abundancias de bacterias. La determinación del valor óptimo de k mediante la inspección de la curva WCSS no sugiere un punto destacable de inflexión (Figura 4.10 A). Por otra parte, la curva S indica que se obtiene un valor máximo del coeficiente S para $k = 2$ (Figura 4.10 B). Finalmente, la curva del estadístico de *Gap* no indica de manera clara un valor de k máximo para seleccionar (Figura 4.10 C). De acuerdo a estas observaciones, aunque no hay consenso entre los estadísticos evaluados, se toma $k = 2$ como el número óptimo de *clusters* para implementar *k-means*. Los resultados indican que los dos *clusters* obtenidos no están separados entre sí y están conformados por $N = 524$ y 534 observaciones cada uno (Figura 4.10 D).

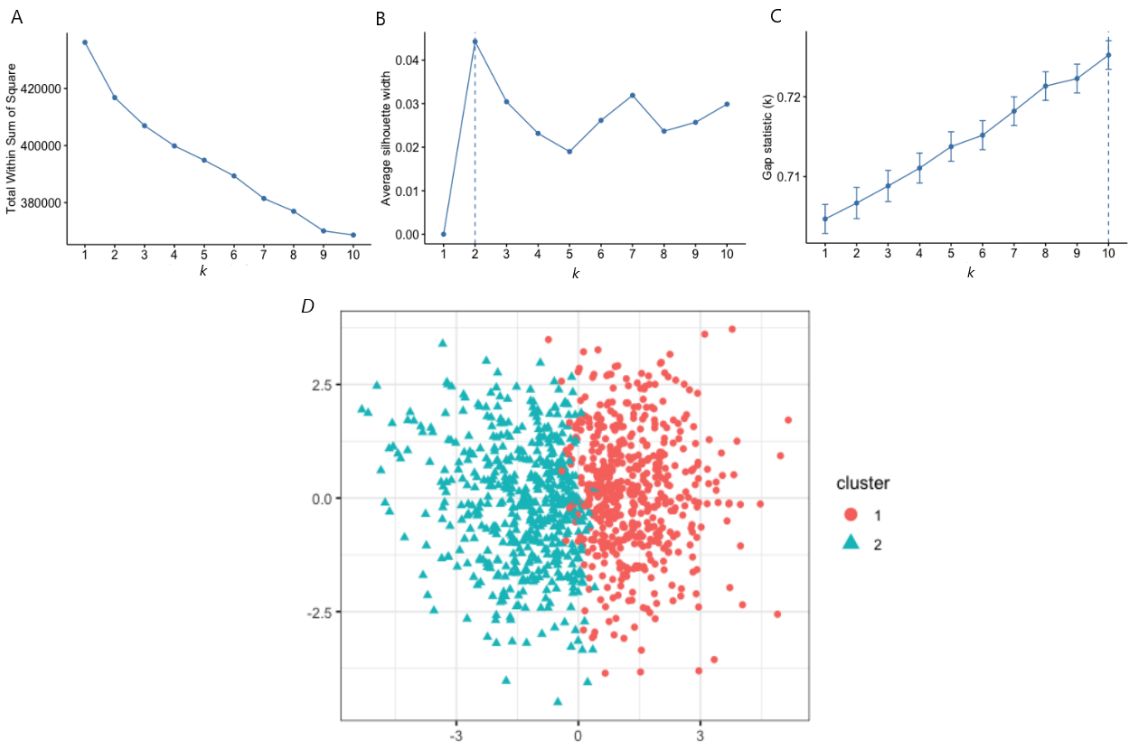

Figura 4.10: Métricas para determinar el valor k a ser usado en *k-means* sobre los variables originales de *Firmicutes*. (A) Método de *elbow*, (B) Coeficiente de *Silhouette* y (C) Método de *Gap*. (D) *Clustering* mediante *k-means* con $k = 2$.

4.3.2.2. Componentes seleccionados de PCA

El conjunto de datos consiste en quince componentes principales obtenidos luego de PCA (PC1-PC15 en conjunto representan el 47.6 % de la varianza total). La determinación del valor óptimo de k mediante la inspección de la curva WCSS no sugiere un punto destacable de inflexión en la curva (Figura 4.11 A). Por otra parte, el coeficiente de *Silhouette* destacó $k = 2$ (Figura 4.11 B) y finalmente el estadístico de *Gap* no indicó de manera clara un valor de k máximo a ser seleccionado (Figura 4.11 C). Aunque dos de las tres métricas analizadas no muestran un valor definido de k a elegir, se decide implementar *k-means* con $k = 2$. Los resultados indican una importante superposición de los *clusters*, sin una partición definida (Figura 4.11 D). De esta manera, los PC1-PC15 de *Firmicutes* que engloban alrededor del 50 % de la varianza no representan un conjunto que pueda ser particionado adecuadamente mediante *k-means*.

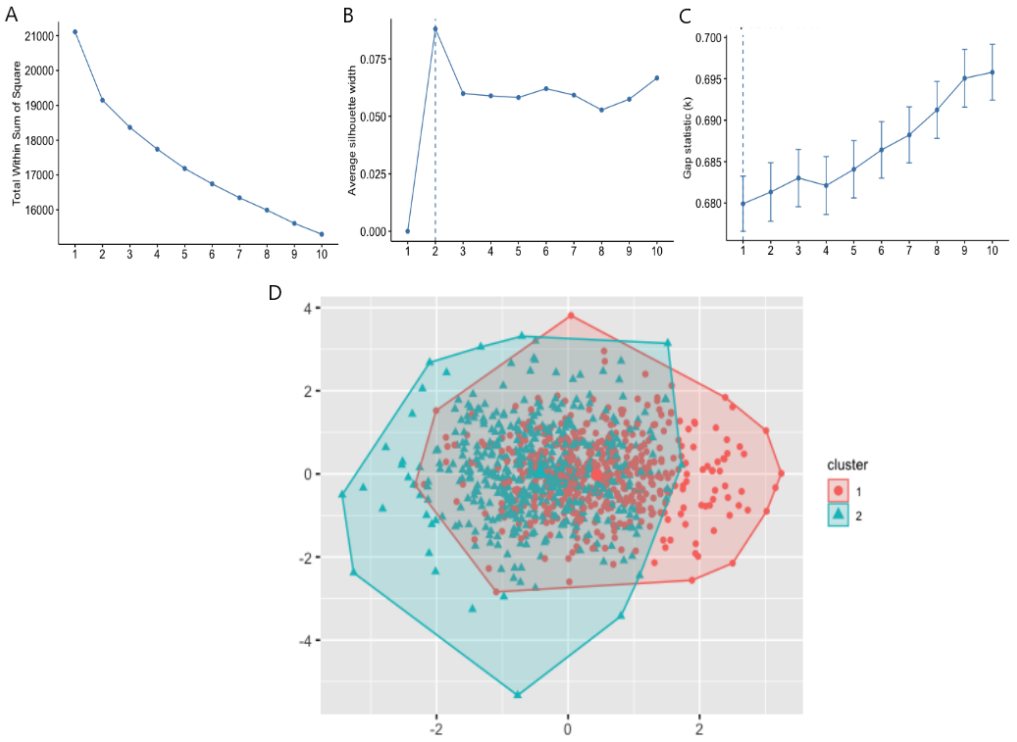

Figura 4.11: Determinación del valor óptimo de k para *k-means* sobre PC1-PC15 de *Firmicutes* mediante (A) el método de *elbow*, (B) coeficiente de *Silhouette* y (C) *gap method* para el subconjunto de *Firmicutes*. (D) *Clustering* mediante *k-means* con $k = 2$.

4.3.2.3. Proyección seleccionada de UMAP

Los datos usados consisten en la proyección UMAP de 2 dimensiones usando los parámetros $n\_neighbours = 10$ y $min\_dist = 0.05$. La determinación del valor óptimo de k mediante la inspección de la curva WCSS parece sugerir un punto de inflexión a partir de $k = 6$ (Figura 4.12 A). En tanto que la curva de *Silhouette* no muestra un valor máximo del coeficiente S (Figura 4.12 B). Finalmente, la curva del estadístico de *Gap* no indica de manera clara un valor de k máximo para seleccionar (Figura 4.12 C). Aunque las métricas no son concretas en definir un valor de k, se toma un valor de $k = 6$ como el número óptimo de *clusters* para implementar *k-means*. Los resultados muestran seis *clusters* uno al lado de otro sin una partición separada, conformados por $N = 133$, 157, 175, 182, 200 y 209 observaciones cada uno (Figura 4.12 D).

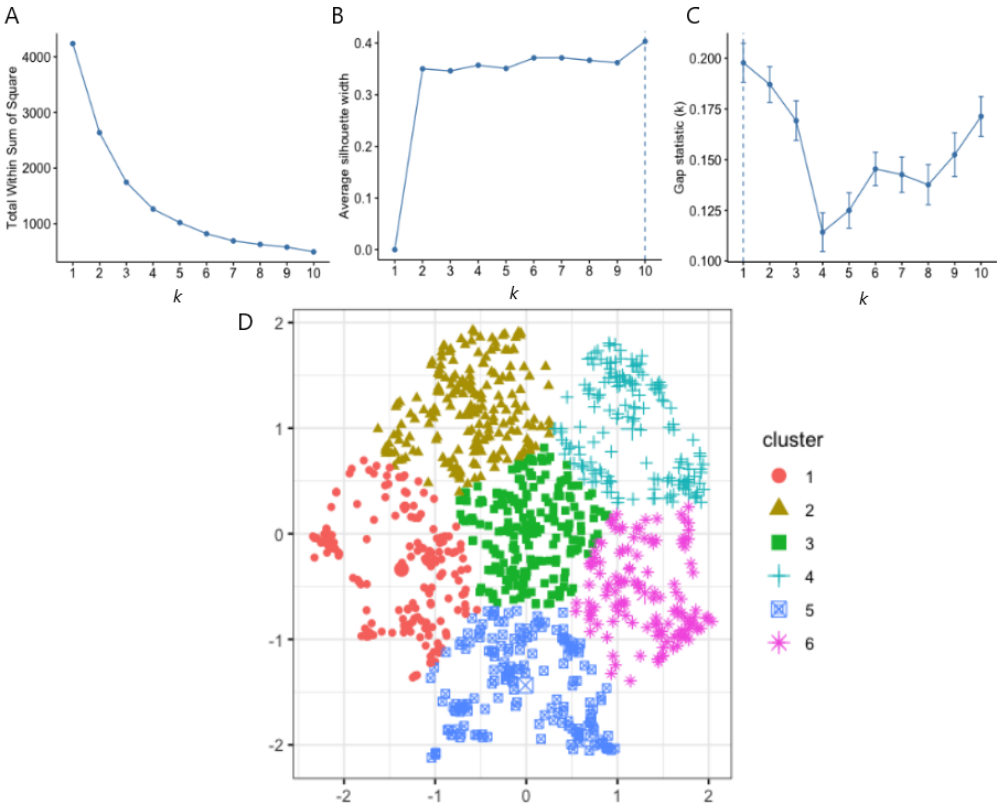

Figura 4.12: Métricas para determinar el valor k para *k-means* sobre la proyección UMAP de *Firmicutes*.(A) Método de *elbow*, (B) Coeficiente de *Silhouette* y (C) Método de *Gap*, (D) *k-means* con $k = 10$

.

4.4. Conclusiones

Si bien los métodos de reducción de la dimensionalidad y agrupamiento consideran diferentes objetivos, el uso combinado de ambos en el análisis de datos multivariados apunta a identificar y resumir propiedades y relaciones relevantes en los datos.

A partir del análisis realizado, se desprende que el subconjunto *Bacteroidetes* tiene una distribución compleja cuyo análisis por métodos lineales no resulta en una clara identificación de grupos. Esto pudo observarse con la combinación de *k-means* sobre los datos originales y los componentes obtenidos por PCA, que resultaron en *clusters* no tan definidos y con cierta superposición.

Sin embargo, la combinación de UMAP y *k-means* resultó en una separación definida y clara de *clusters*. Estos podrían estar reflejando los subtipos de bacterias de *Bacteroidetes*, aunque este análisis no fue explorado en esta parte de la Tesis.

Por otra parte, ningún agrupamiento fue evidente para los datos del subconjunto de *Firmicutes*. Es probable que la distribución compacta, homogénea y sin ningún *cluster* natural refleje el hecho de que la gran mayoría de estas bacterias oscilan entre perfiles de abundancia significativa o por el contrario casi nulas en la mayoría de los individuos, dificultando la detección de agrupamientos claros (al menos mediante el algoritmo utilizado acá). Se sabe que las abundancias relativas de estas bacterias siguen una distribución bimodal variable [Lahti et al., 2014]. Esta característica podría explicar además la necesidad de una cantidad elevada de componentes principales que agrupe un porcentaje significativo de varianza en cada subconjunto.

Capítulo 5

Mapas Auto-Organizados

El uso de mapas auto-organizados (SOM) en el análisis de conjuntos de datos multidimensionales ofrece una herramienta de visualización muy útil porque permite representar la información en un sistema de menor dimensión, preservando las relaciones topológicas de los datos de entrada.

Estas relaciones topológicas pueden ser mejor representadas en un mapa de nodos, que muestran la distribución de los datos de cada variable. Todos los mapas están unidos por posición (es decir, un nodo en una determinada ubicación corresponde a la misma unidad en otro mapa). De esta manera, es posible observar las superposiciones de los agrupamientos generados luego del entrenamiento de SOM. Los mapas de calor (*heatmaps*) son la visualización típica del SOM, los cuales muestran la distribución de una variable a través del mapa de nodos.

El objetivo de este capítulo es aprovechar la poderosa herramienta de visualización que ofrece SOM e identificar de manera precisa las relaciones entre las diferentes variables de abundancias de bacterias y metadatos de los sujetos en estudio. Por ello, se usarán los datos originales de *Bacteroidetes* y *Firmicutes*: las columnas de cada matriz de datos corresponden a los niveles de abundancias de bacterias individuales y las filas corresponden a cada sujeto de estudio. Se incluyen además las variables de metadatos descriptas en Materiales y Métodos.

5.1. *Bacteroidetes*

El progreso de aprendizaje de la red SOM puede ser visualizado mediante una curva que relaciona las distancias entre los nodos y el BMU en función del número de iteraciones. Se espera que a medida que progresa el entrenamiento de la red, esta distancia vaya disminuyendo continuamente hasta llegar a una meseta. La configuración del mapa con 49 neuronas (7 filas x 7 columnas) y 1000 iteraciones indicó que se alcanzó un mínimo estable a partir de 800 iteraciones (Figura 5.1).

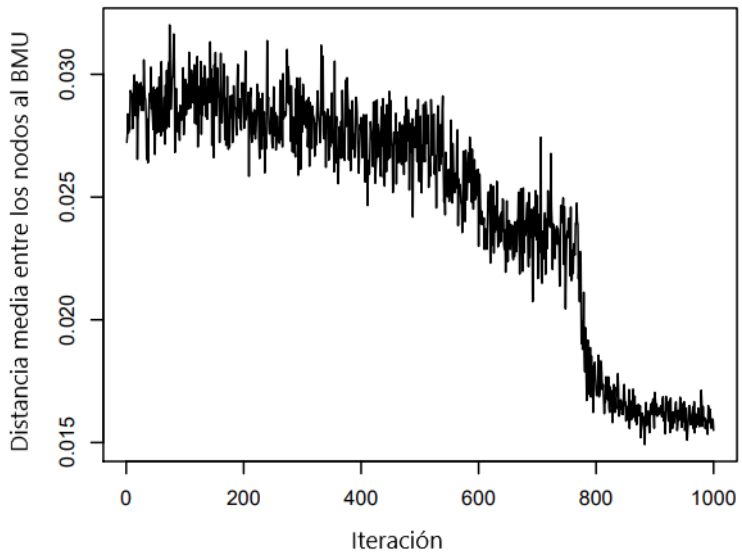

Figura 5.1: Progreso del entrenamiento de la red neuronal SOM en el subconjunto de datos de *Bacteroidetes*.

5.1.1. Metadatos

La distribución de la variable edad a través del SOM indicó que los nodos correspondientes a la franja más joven (alrededor de 20 años) se agruparon en un sector acotado en el mapa (nodos 21, 31, 35, 41 y 42). Mientras que la franja etaria de mayor edad ($>$ 65 años) se posicionó en un único nodo en el mapa (nodo 12). Los individuos del rango de edad media (40-50 años) fueron mapeados en nodos de manera dispersa por todo el mapa (Figura 5.2 A).

La variable nacionalidad está mayormente representada por individuos de Europa Central, los cuales mapearon en nodos homogeneamente distribuidos a través de todo el mapa. Grupos de nodos aislados mapearon a individuos de Escandinavia (nodos 1, 4, 7, 11, 17 y 47). Mientras

que los individuos de UK/Irlanda mapearon en un nodo aislado (nodo 42) al igual que los individuos de Estados Unidos (nodo 49) (Figura 5.2 B).

Un índice de diversidad ecológica muy utilizado para calcular la diversidad dentro de una población es el índice de diversidad de *Shannon*, el cual calcula el grado de distribución uniforme de las bacterias en una muestra. Se observó que los valores bajos en la escala de diversidad quedaron representados en dos nodos separados (nodos 15 y 43). Por el contrario, los nodos con valores altos de diversidad mapearon en grupos aislados más grandes y con mayor distribución a través del mapa de la red (nodos 17, 23, 24, 27, 29, 30, 35 y 44) (Figura 5.2 C).

Los individuos pertenecientes a categorías de BMI bajos (bajo peso - delgados) mapearon en pequeños grupos de nodos (nodos 5, 13, 15, 21, 23, 27, 28, 31, 35, 41 y 44). Mientras que los individuos con la categoría más alta de BMI (obesos severos) se distribuyeron principalmente nodos aislados, bien separado de las demás categorías (nodos 16, 24 y 43) (Figura 5.2 D).

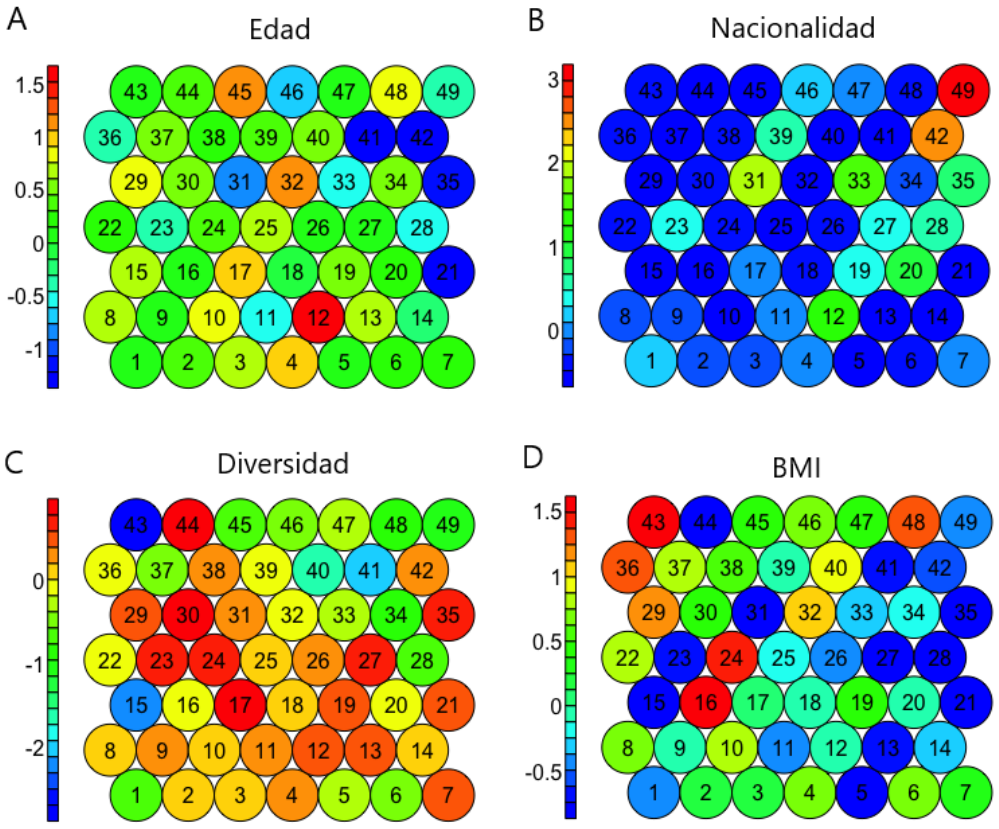

Figura 5.2: Distribución de los metadatos asociados a *Bacteroidetes* luego de SOM. El índice de color está establecido en base a los valores para cada variable. A) Edad, azul = más jóvenes, rojo = adultos mayores. B) Nacionalidad, azul = Europa Central, rojo = US. C) Diversidad, azul = baja, roja = alta. D) BMI, azul = bajo peso, rojo = obesos severos.

Al comparar de manera conjunta los mapas de Diversidad y BMI, se observa que los dos nodos que representan valores de baja diversidad (ver nodos 15 y 43 en la Fig.5.2 C) corresponden a individuos de BMIs extremos: obesos severos (nodo 43 en la Fig. 5.2 D) y bajo peso (nodo 15 en la Fig.5.2 D). El nodo 15 no será tenido en cuenta porque está representado por un único individuo (ver Figura 5.5). Además, al agregar a esta comparación el mapa de Edad, se observó que son individuos de mediana edad (Figura 5.2 A).

En cuanto a los nodos de alta diversidad (por ejemplo, los nodos 23, 27 y 35) se observa correspondencia con individuos de bajo BMI. Sin embargo, el nodo 24 que corresponde a individuos de alto BMI y alta diversidad, comprende una cantidad baja de individuos en comparación con el nodo 43 (ver Figura 5.5).

5.1.2. Abundancias de bacterias

La distribución de los valores de abundancia para cada especie de bacteria presente en el subconjunto *Bacteroidetes* mostró una gran proporción de nodos que mapearon valores de abundancia baja o nula en todo el mapa (en los mapas son los nodos de color azul). Mientras que los valores de máxima abundancia se concentraron en pequeños nodos aislados (indicados en color rojo). Los *heatmaps* para todas las bacterias de *Bacteroidetes* están en el Apéndice D.

Al comparar los mapas de bacterias con el mapa de BMI, se podría relacionar la abundancia máxima de bacterias con alguna categoría de BMI, individualizando cada especie. A modo de ejemplo, en la Figura 5.3 se representan algunos ejemplos de especies de *Bacteroidetes* y su relación con el BMI. Cabe destacar que, dado que algunos nodos presentan una cantidad baja de individuos, se necesitaría un n más grande para poder hacer afirmaciones robustas.

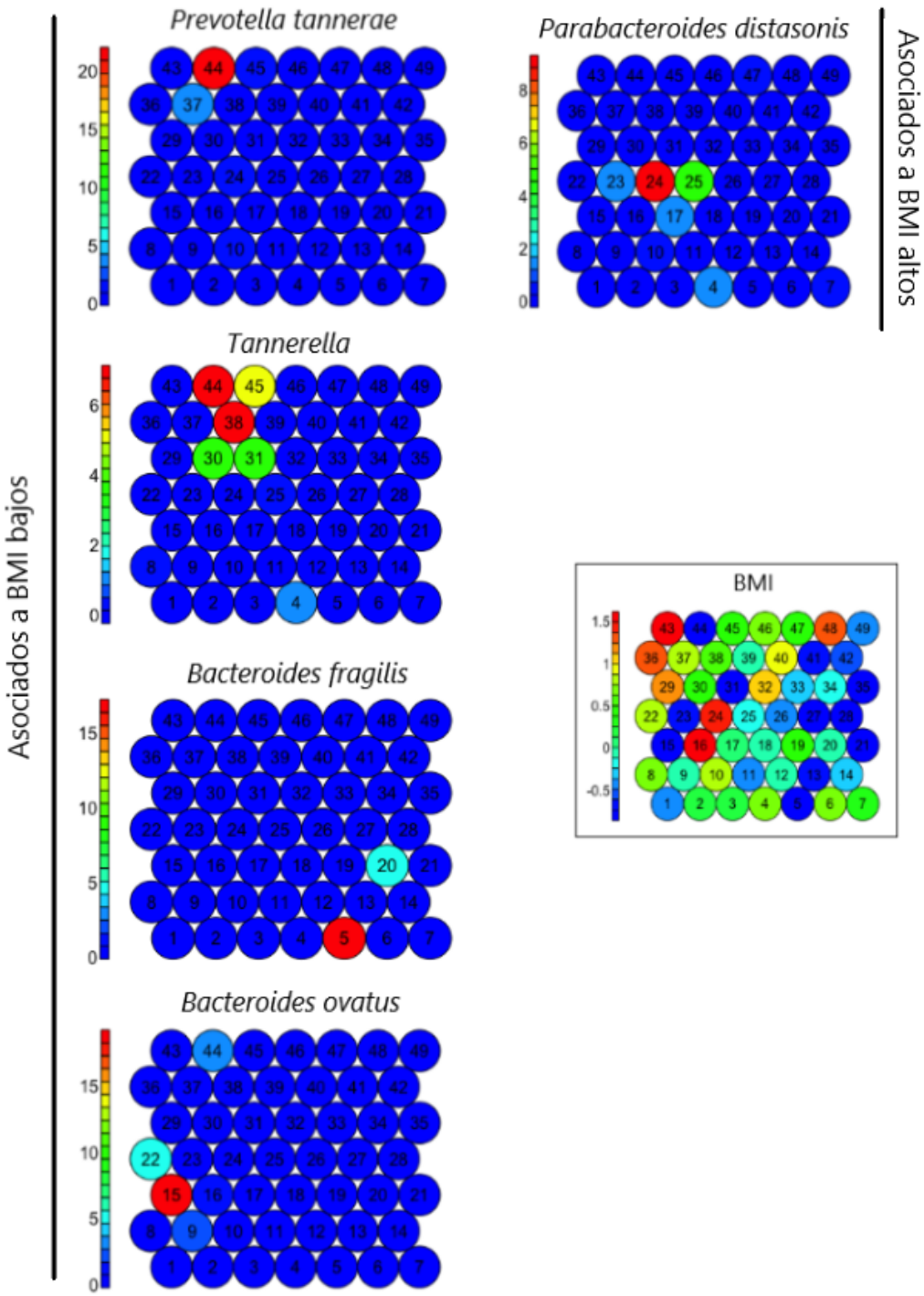

Figura 5.3: Mapeo coincidente de bacterias de *Bacteroidetes* con BMI bajos y altos.

Si bien la red neuronal SOM genera una proyección 2D de los datos que ayuda a determinar visualmente características y probables agrupamientos, es posible complementar el enfoque con el uso de un algoritmo de *clustering* aplicado a los vectores prototipo obtenidos luego del entrenamiento de la SOM y así dividir los datos del mapa 2D en *clusters*. De esta manera, el análisis de los mapas de calor 2D para cada variable en estudio se completa mediante la

identicación manual de los *clusters*, y en conjunto, se les da sentido a los resultados acerca de las diferentes áreas en el mapa.

Los 49 vectores de nodos del SOM son usados como valores *input* en un método agregativo de agrupamiento, tal como *clustering* jerárquico. El resultado final es un árbol jerárquico (Figura 5.4). Inicialmente se probaron varias condiciones para seleccionar el número de *clusters* en el árbol, llegando a la condición n=6 grupos (marcado por la línea de color).

El análisis de la longitud de las ramas muestra que tempranamente se separan dos *clusters*: el primero de ellos contiene el nodo 44 mientras que el segundo contiene los nodos 5 y 15 (denominados *cluster* 4, 5 y 6). A medida que se va bajando en la estructura del árbol, se separa un *cluster* conformado por los nodos 45 y 46 (*cluster* 1). Y de esta bifurcación, surgen luego dos grandes grupos claramente separados (*cluster* 2 y 3).

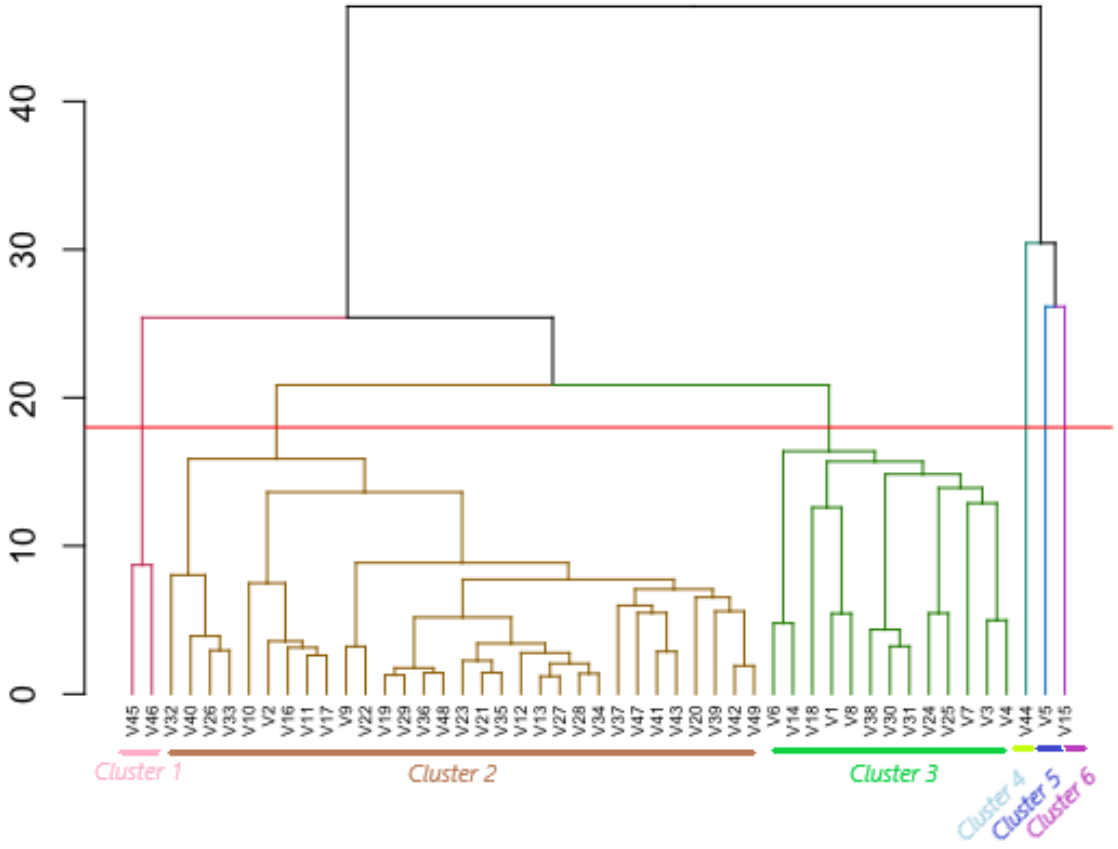

Figura 5.4: *Clustering* jerárquico aglomerativo de los nodos de SOM para *Bacteroidetes*, método de *Ward's linkage*.

La implementación de los *clusters* sobre el mapa de distribución de muestras de SOM sugiere que los *clusters* 1, 4, 5 y 6 corresponden a valores *outliers* ya que los nodos mapearon 1 o 2

muestras solamente (Figura 5.5).

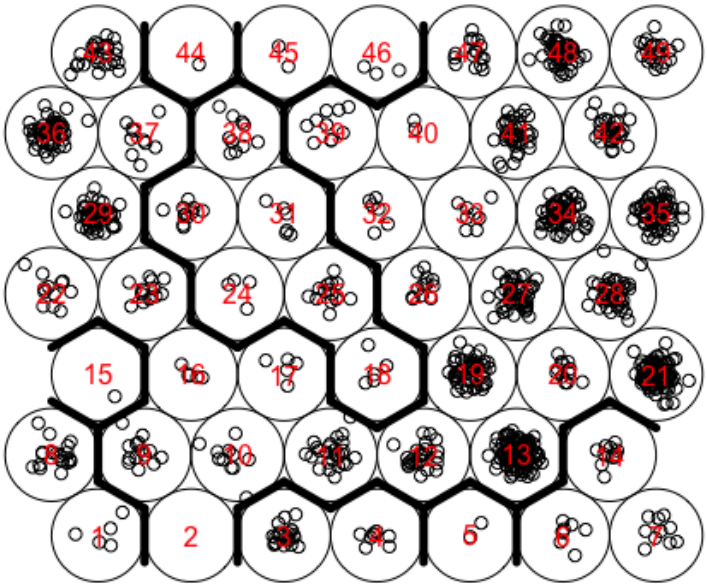

Figura 5.5: Datos de *Bacteroidetes* mapeados en 49 nodos.

Una vez realizado el agrupamiento jerárquico, es posible observar que algunos *clusters* tienen características más definidas que otros y, por lo tanto, es más directa su descripción. La información de todos los mapas generados por SOM con los límites marcados de los seis *clusters* está disponible en el Anexo E.

La visualización conjunta de los mapas indica, por ejemplo, que en el *cluster* 2 existen nodos con predominancia de *Bacteroides plebeius* y bajos valores de BMI. Mientras que en el *cluster* 3 predominan las abundancias de *Prevotella melaninogenica*, *Bacteroides vulgatus*, *Tannerella*, *Parabacteroides distasonis* y *Allistipes*. Además, hay una tendencia a individuos jóvenes con valores altos de BMI.

5.2. *Firmicutes*

El progreso del entrenamiento de la red SOM puede ser visualizado en la curva descendiente que relaciona las distancias entre los nodos y el BMU en función del número de iteraciones. La configuración del mapa con 49 neuronas (7 filas x 7 columnas) y 1000 iteraciones indicó que se alcanzó un mínimo estable a partir de 800 iteraciones (Figura 5.6)

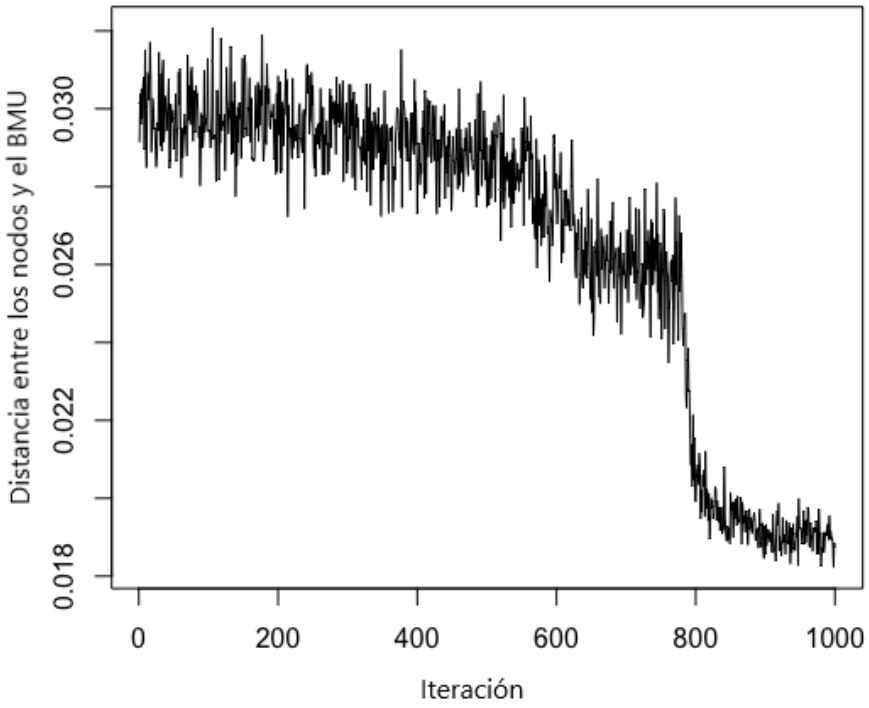

Figura 5.6: Progreso del entrenamiento de la red neuronal SOM en los datos de *Firmicutes*.

5.2.1. Abundancias de bacterias

El *phylum Firmicutes* está conformado por alrededor de 250 géneros diferentes de bacterias, tales como *Lactobacillus*, *Bacillus*, *Clostridium*, *Enterococcus* y *Ruminicoccus*, entre otros. En el conjunto de datos analizado se identificaron 42 miembros pertenecientes a este *phylum*, cada uno de los cuales generó un mapa de calor con los resultados del entrenamiento en la SOM (los *heatmaps* para todos los miembros están en el Apéndice F). La interpretación visual de todos los mapas en búsqueda de una superposición biológica interesante resulta difícil de realizar a simple vista. Es por eso que se complementará el análisis con el uso de un algoritmo de *clustering* jerárquico utilizando como valores *input* a los 49 vectores de nodos del SOM.

En base a las características de la estructura del árbol jerárquico obtenido, se eligió separar en 9 *clusters* (Figura 5.7, línea de color). El análisis de la longitud de las ramas indica que tempranamente se separan varios *clusters* separados y distantes de la estructura principal del árbol. Estos fueron denominados "varios" porque cada uno está conformado por un único nodo (nodos 41, 47, 31, 43, 4 y 29). A medida que se va bajando en el árbol, se observan 3 grupos definidos (Figura 5.7, *clusters* 1, 2 y 3).

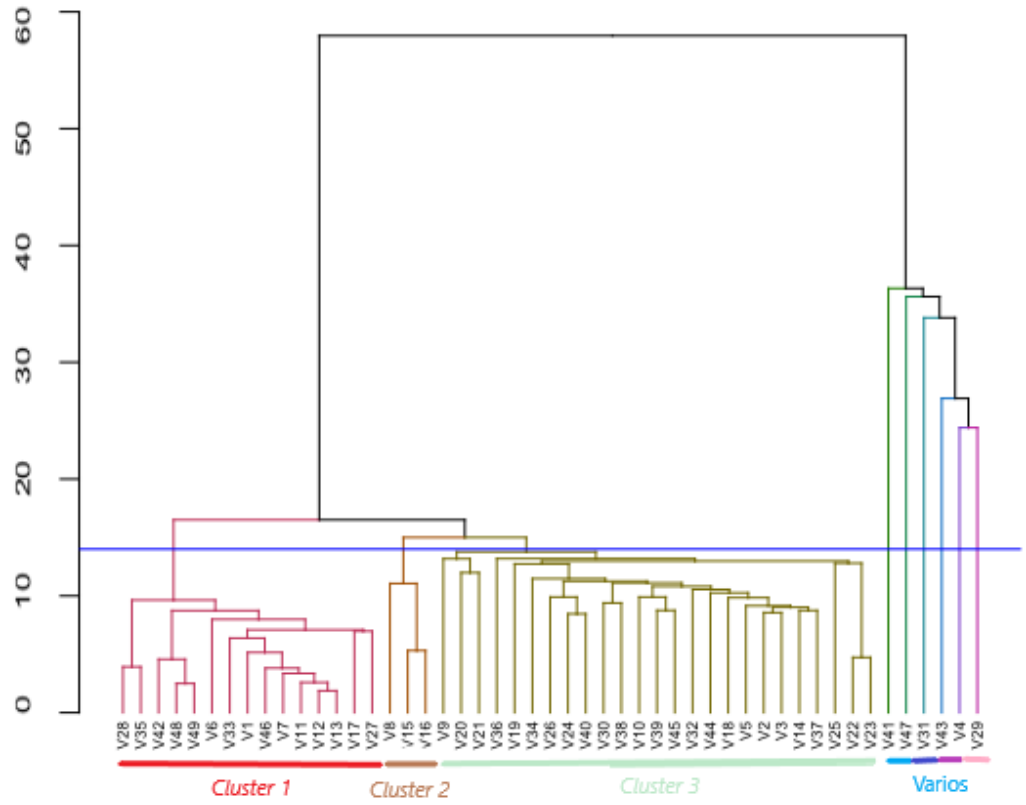

Figura 5.7: *Clustering* jerárquico aglomerativo de los nodos de SOM de *Firmicutes*, método de *Ward's linkage*.

La implementación de los *clusters* sobre el mapa SOM de distribución de casos sugiere que las ramas agrupadas bajo el término "varios" corresponden a valores *outliers*, dado que cada uno de sus nodos mapearon 1 a 2 casos solamente (Figura 5.8).

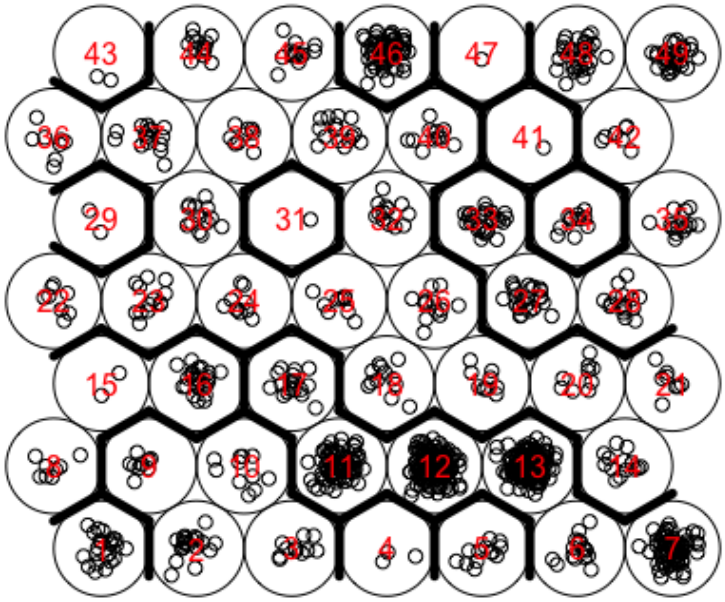

Figura 5.8: Datos de *Firmicutes* mapeados en 49 nodos.

El análisis de los mapas de bacterias pertenecientes a los *clusters* 1, 2 y 3 no indicó un perfil de bacterias definido. Sin embargo, es interesante destacar que la abundancia máxima de ciertas bacterias coincidió con valores altos de diversidad. Por ejemplo, esto se observó para *Lachnobacillus bovis* (nodo 10) y para *Staphylococcus*, cuya máxima expresión coincidió con individuos de bajo BMI y alta diversidad (nodo 42) (ver Figura 5.9 parte superior). Por otra parte, se observó la superposición de valores máximos de abundancias para las bacterias *Coprococcus eutactus*, *Eubacterium biforme* y *Ruminococcus lactaris* en el nodo 44 (Figura 5.9 parte inferior).

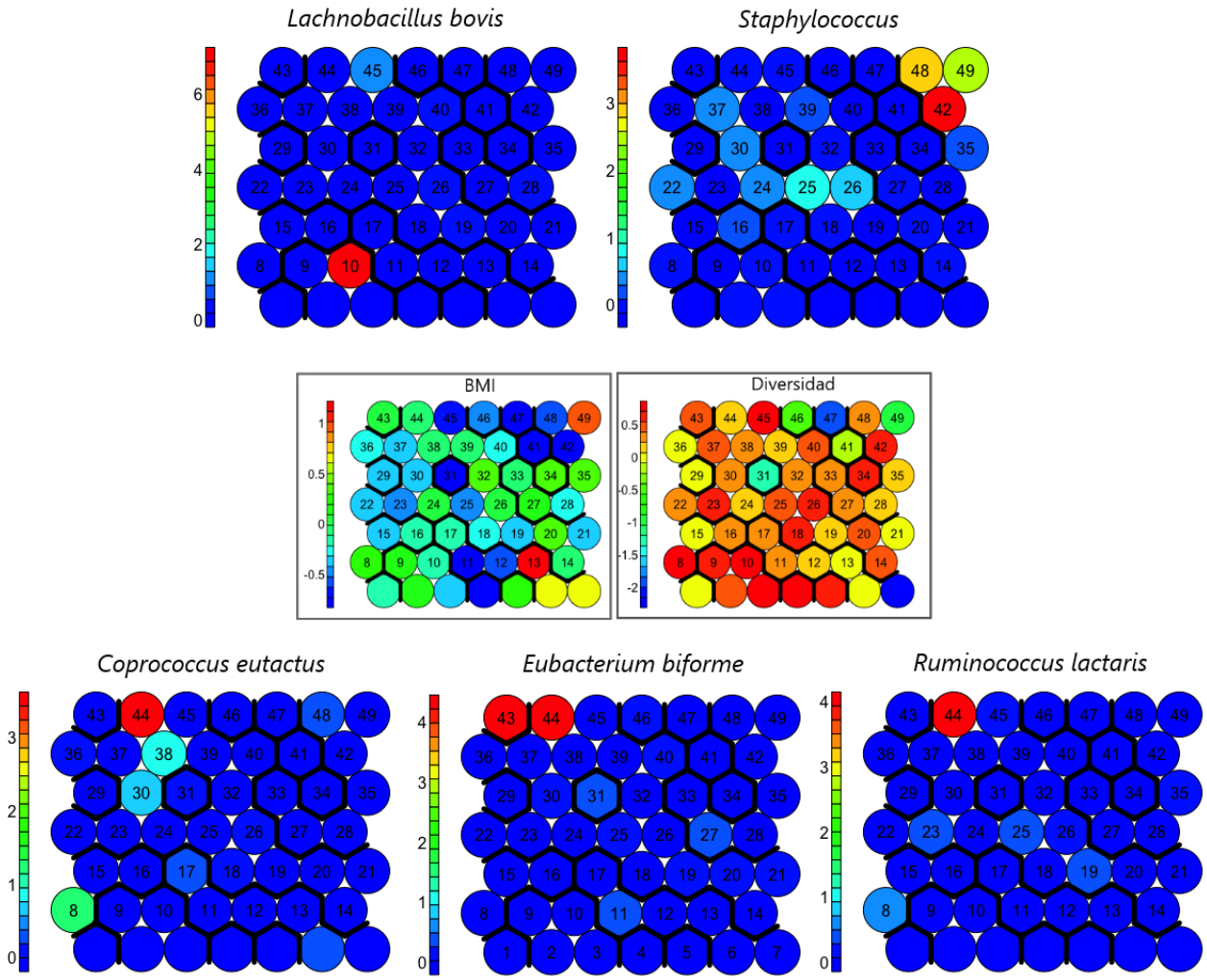

Figura 5.9: Mapeo coincidente de abundancias de bacterias de *Firmicutes* con BMI y diversidad luego de SOM. El índice de color está establecido en base a los valores para cada mapa. BMI: azul-bajo peso, rojo-obesos severos. Diversidad: azul-baja, roja-alta.

5.3. Conclusiones

El análisis de SOM es relevante dado que combina dos aspectos, el de la proyección y el agrupamiento. Mientras que el agrupamiento brinda información sobre patrones de expresión similares, la proyección de datos a un plano sirve como un método de visualización y permite estudiar las relaciones entre los datos.

El tamaño de la grilla de nodos requiere ser determinada por prueba y error. Tanto para *Bacteroidetes* como para *Firmicutes*, se fueron probando varias dimensiones de filas y columnas de la grilla de tal manera de que los datos se distribuyeran sin dejar nodos vacíos.

Las asociaciones detectadas mediante el SOM son interesantes teniendo en cuenta que la baja diversidad es considerada un indicador de una microbiota no saludable: los resultados mostraron un alto número de individuos con obesidad severa asociados a baja diversidad. Esto está en concordancia con trabajos previos que indican que la condición de obesidad severa está asociada a una menor diversidad en comparación con individuos sanos [Wan et al., 2020].

La combinación de SOM y *clustering* jerárquico aglomerativo permitiría resaltar asociaciones interesantes de bacterias y metadatos de los individuos.

Aunque los *clusters* no fueron validados, es posible observar que los dendogramas para *Bacteroidetes* y para *Firmicutes* muestran *clusters* definidos.

Capítulo 6

Random Forest

Random forest es un ensamble de árboles de decisión (o de regresión, en el caso de una variable *target* continua), donde cada árbol está formado por un subconjunto de datos tomados a partir de datos de entrenamiento con reemplazo (*bootstrap*). El algoritmo considera además una fracción de las variables predictoras en cada partición, favoreciendo la diversidad y reduciendo al mismo tiempo la correlación entre los árboles componentes del ensamble.

Un resultado muy útil de RF para interpretar el modelo de predicción obtenido es una escala con el grado de importancia de las variables. Ésta representa la influencia relativa de cada variable sobre la variable *target*: tiene en cuenta si una determinada variable fue seleccionada en cada paso de partición durante el proceso de construcción del árbol y cuánto mejoró (disminuyó) el error cuadrado (sobre todos los árboles).

En estudios de la microbiota intestinal, la diversidad representa una variable que caracteriza el grado de distribución de las bacterias en una muestra. A continuación, se implementa el algoritmo RF para generar un modelo de regresión de la diversidad bacteriana usando de manera independiente las variables de abundancias de miembros de *Bacteroidetes* y *Firmicutes* en conjunto con los metadatos. En base al modelo RF armado para cada caso, se busca identificar las variables relevantes al mismo.

6.1. Bacteroidetes

Los resultados indicaron que la edad de los individuos fue el factor más importante en el modelo para predecir la diversidad de la microbiota intestinal. En cuanto a las abundancias de bacterias, se observó que dos especies tuvieron la mayor importancia relativa en la identificación de la diversidad: *Prevotella tannerae* y *Bacteroides ovatus*. Por otra parte, de las restantes variables de metadatos en el subconjunto de datos, el BMI de los individuos y la nacionalidad fueron las siguientes en el listado. Se observó que la variable sexo no fue importante en la predicción de la diversidad (Cuadro 6.1).

Variable	Importancia relativa	Importancia escalada
Edad	222,61	1,00
Prevotella tannerae	187,01	0,84
BMI	153,52	0,68
Bacteroides ovatus	108,42	0,48
Nacionalidad	104,61	0,47
Bacteroides plebeius	82,35	0,37
Parabacteroides distasonis	77,08	0,34
Bacteroides stercoris	75,19	0,33
Bacteroides splachnicus	73,41	0,33
Allistipes	68,61	0,30
Tannerella	66,79	0,30
Prevotella oralis	64,48	0,28
Bacteroides vulgatus	62,33	0,28
Bacteroides uniformis	55,36	0,24
Bacteroides fragilis	53,81	0,24
Sexo	49,97	0,22
Prevotella melaninogenica	39,91	0,17

Cuadro 6.1: Listado de la importancia relativa de las variables de *Bacteroidetes* y metadatos sobre la diversidad luego de RF.

Se obtiene una interpretación más directa de las variables en una escala con respecto a la variable de mayor importancia (Figura 6.1).

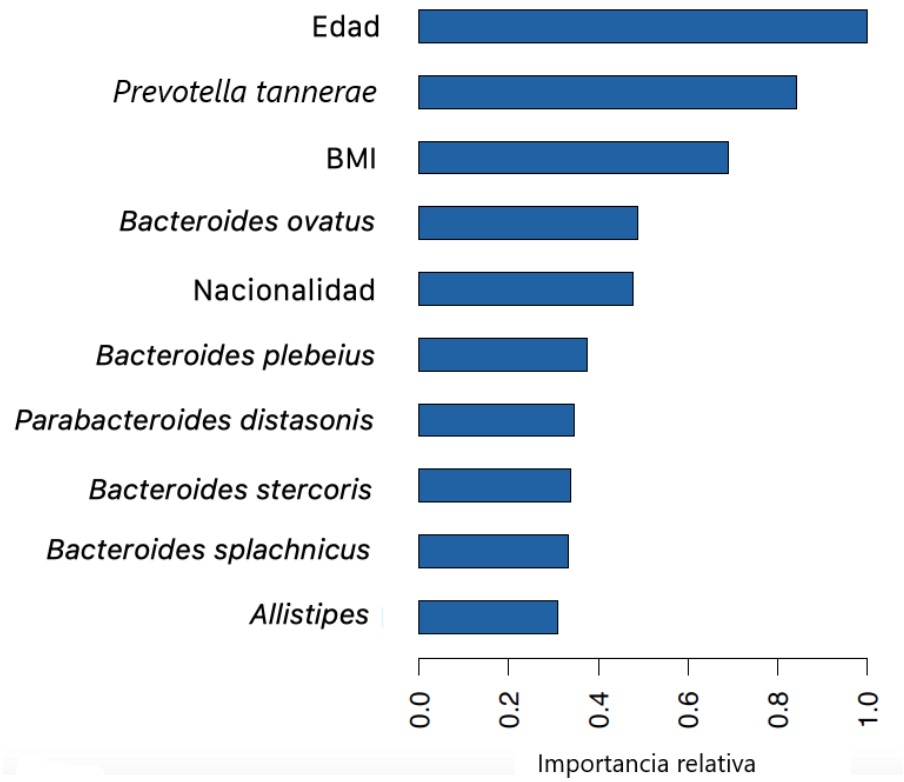

Figura 6.1: Diez primeras variables importantes de *Bacteroidetes* y metadatos sobre la diversidad luego de implementar RF.

En cuanto a las métricas de evaluación del modelo, se observó que comenzando con un valor inicial de $N_{tree} = 50$ y ascendiendo, MSE empezó a disminuír hasta mantenerse constante con $N_{tree} = 150$. A medida que aumentaba N_{tree}, los valores de RMSE disminuyeron hasta $N_{tree} = 250$. Los valores de MAE disminuyeron desde $N_{tree} = 50$ hasta $N_{tree} = 200$ (Cuadro 6.2).

Número de árboles	MSE	RMSE	MAE	RMLE
50	0,040	0,201	0,153	0,030
100	0,039	0,199	0,152	0,029
150	0,038	0,196	0,151	0,029
200	0,038	0,196	0,150	0,029
250	0,038	0,195	0,150	0,029
300	0,038	0,196	0,151	0,029

Cuadro 6.2: Comparación de diversas métricas en la evaluación de RF de regresión dependiente del número de árboles en el subconjunto *Bacteroidetes*. MSE = *Mean Squared Error*, RMSE = *Root Mean Squared Error*.

Dado que MSE y RMSE son métricas sensibles a valores *outliers*, se decidió tener en cuenta los valores de MAE en el ajuste de hiperparámetros, porque esta métrica es robusta. De esta manera, se determinó que el modelo de predicción de la diversidad biológica en *Bacteroidetes* mediante *random forest* es óptimo con $N_{tree} = 200$.

6.2. *Firmicutes*

En el modelo de predicción de la diversidad, las variables de metadatos fueron las más importantes, comenzando con el BMI de los individuos, seguida por la nacionalidad y la edad. En cuanto a las abundancias de los 42 tipos de bacterias de *Firmicutes*, se observa que dos miembros tuvieron la mayor importancia relativa en la predicción de la diversidad: *Lachnobacillus bovis* y *Anaerovorax odorimutans* (Cuadro 6.3). Algunas de las restantes especies de *Firmicutes*, listadas en orden decreciente de importancia para el modelo, han sido asociadas a obesidad [Salonen et al., 2014, Rabot et al., 2016]. En la sección de Conclusiones de este capítulo se desarrollará más sobre estos resultados.

Variable	Importancia relativa	Importancia escalada
BMI	99,44	1,00
Nacionalidad	81,66	0,82
Edad	69,81	0,70
Lachnobacillus bovis	60,99	0,61
Anaerovorax odorimutans	54,89	0,55
Ruminococcus obeum	53,50	0,53
Sporobacter termitidis	47,89	0,48
Eubacterium biforme	45,28	0,45
Clostridium leptum	33,35	0,33
Bryantella formatexigens	32,80	0,33

Cuadro 6.3: Listado de la importancia relativa de las diez primeras variables de *Firmicutes* y metadatos sobre la diversidad luego de RF. Las restantes variables están en el apéndice G.

La interpretación de la influencia de las variables en el modelo de regresión es más directa al representar en una escala con respecto a la variable de mayor importancia (Figura 6.2).

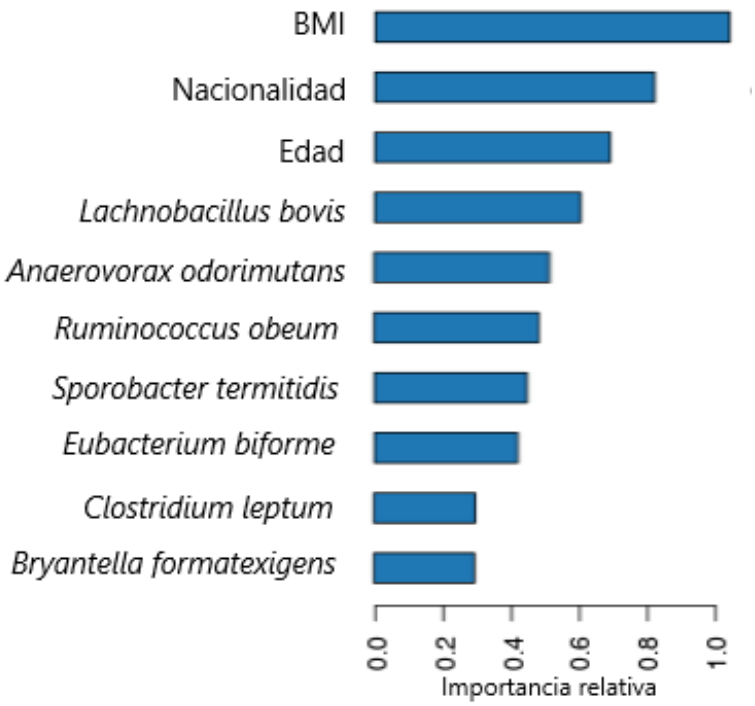

Figura 6.2: Diez primeras variables importantes de *Firmicutes* y metadatos sobre la diversidad luego de implementar RF.

En cuanto a las métricas de evaluación del modelo, se observó que con un valor de inicial de $N_{tree} = 50$ y ascendiendo, los valores de MSE, RMSE y MAE fueron variando levemente (Cuadro 6.4). Sin embargo, dado que MSE y RMSE no presentan grandes variaciones y que además son sensibles a *outliers*, se eligió a MAE como medida de evaluación. Así, se determinó que el modelo de predicción de la diversidad biológica en *Firmicutes* mediante *random forest* es óptimo con $N_{tree} = 50$.

Número de árboles	MSE	RMSE	MAE	RMLE
50	0,030	0,175	0,136	0,026
100	0,031	0,176	0,136	0,026
150	0,031	0,176	0,137	0,026
200	0,030	0,175	0,137	0,026
250	0,030	0,175	0,137	0,026
300	0,030	0,176	0,137	0,026

Cuadro 6.4: Comparación de diversas métricas en la evaluación de RF de regresión dependiente del número de árboles en el subconjunto *Firmicutes*. MSE = *Mean Squared Error*, RMSE = *Root Mean Squared Error*.

6.3. Conclusiones

El objetivo de este capítulo fue armar un modelo de regresión de la diversidad bacteriana mediante *random forest*, usando de manera independiente los datos de bacterias de *Bacteroidetes* y *Firmicutes* (junto a las variables de metadatos).

Si bien la diversidad bacteriana indica la diversidad de especies en una comunidad, se decidió mantener el enfoque de análisis de *Bacteroidetes* y *Firmicutes* por separado con el fin de caracterizar la contribución de ambos subconjuntos al modelado de diversidad,

Independientemente de las métricas de evaluación en cada modelo de regresión, los resultados obtenidos sobre la importancia de determinadas variables son relevantes al momento de predecir la diversidad de la población.

Tanto la Edad y el BMI de los individuos aparecen como las dos variables más relevantes en los modelos generados para *Bacteroidetes* y *Firmicutes*. Estos resultados están avalados por trabajos previos que indican una importante influencia de estos factores en la dinámica del microbioma intestinal [Yatsunenko et al., 2012, Egert et al., 2006].

Con respecto a las bacterias del subconjunto *Bacteroidetes*, dos especies tuvieron mayor importancia en el modelo de predicción:

- *Prevotella tannerae* pertenece a un género de bacterias cuya presencia determina las respuestas de individuos a suplementos dietarios [Chung et al., 2020].

- *Bacteroides ovatus* es una especie dominante en el intestino y ha sido identificada en trabajos previos como un probiótico de próxima generación debido a sus efectos preventivos en inflamación intestinal (revisado en [Tan et al., 2018]).

Del subconjunto de bacterias de *Firmicutes*, dos miembros tuvieron mayor importancia en el modelo de predicción: *Lachnobacillus bovis* y *Anaerovorax odorimutans*. Si bien no existen trabajos basados exclusivamente en estas dos especies, las abundancias relativas de las mismas están influenciadas por el tipo de dieta y uso de antibióticos en individuos obesos, además de asociaciones con el BMI de gemelos monocigóticos [Tims et al., 2013, Salonen et al., 2014, Reijnders et al., 2016].

Aunque el algoritmo RF tiene un buen desempeño con muy pocos ajustes de hiperparámetros, en este capítulo se evaluó solamente el ajuste de un único hiperparámetro: N_{tree}. Por otra parte, en el análisis de las importancias relativas de las variables sobre modelo no se tuvo en cuenta el potencial efecto de la colinealidad de las mismas. La multicolinealidad ocurre cuando dos o más variables están altamente correlacionadas, lo cual implica información redundante.

A futuro, se espera poder evaluar el modelo de regresión mediante el ajuste de otros hiperparámetros y determinar alguna estrategia a considerar para la colinealidad de las variables.

Capítulo 7

Discusión

A lo largo del tracto gastrointestinal se aloja un vasto sistema de millones de bacterias, hongos y virus, cuyas comunidades residentes conviven en un balance complejo entre sí y con el hospedador. La composición de especies de bacterias en el microbioma intestinal humano es muy variable. Tal diversidad depende de factores como la dieta, el ambiente, la genética del hospedador, la edad, entre otros y su estudio representa entonces un desafío importante para las ciencias biomédicas debido a que desbalances en la microbiota intestinal están asociados a enfermedades.

Los avances tecnológicos hicieron posible que la información biológica pueda ser cuantificada en paralelo y a escala masiva, impulsando a que las áreas de las biociencias se enfrenten a cantidades cada vez mayores de datos ómicos (referidos al conjunto total de biomoléculas como ADN, ARN, proteínas, metabolitos, etc). Dentro de la genómica, una de las primeras tecnologías que ha permitido el análisis masivo y en paralelo de la expresión de genes es el *microarray*. En esta Tesis se usaron datos generados en *microarrays* filogenéticos que permiten evaluar la presencia y abundancia relativa de miles de bacterias de la microbiota intestinal proveniente de individuos mediante la cuantificación de un gen marcador, como es el 16S RNAr.

Los microorganismos que componen la microbiota intestinal humana están clasificados taxonómicamente en seis grandes grupos o *phyla*, que contienen a su vez cientos de miembros. En esta Tesis se abordó el análisis de dos de estos grupos, *Bacteroidetes* y *Firmicutes*, mediante el uso de algoritmos de aprendizaje automático. Si bien el conjunto de datos usado ya fue descripto previamente [Lahti et al., 2014], el enfoque del presente trabajo fue caracterizar la abundancia

relativa de bacterias de los dos grupos mencionados con nuevas herramientas de análisis para evaluar posibles asociaciones con información adicional de los individuos.

El conjunto de datos cuenta con información sobre el índice de masa corporal (BMI) de los individuos, además del sexo, nacionalidad y edad. La decisión inicial de enfocar el estudio en *Bacteroidetes* y *Firmicutes* se basó en el hecho de que varios géneros de ambos grupos son predominantes en el microbioma intestinal. Además existe una relación con la obesidad, tanto en humanos como en modelos animales, la microbiota de individuos obesos está mayormente enriquecida con bacterias del grupo *Firmicutes* y en menor medida las del grupo *Bacteroidetes* (Ley et al, 2006; Castaner et al, 2018).

Se buscó en principio determinar si es posible identificar bacterias con predominancia de tal manera que representen una gran proporción de la variabilidad. Es decir, reducir el número de variables a unas pocas para empezar a tener una idea sobre la estructura de los datos. Para ello, PCA a menudo es usado como una herramienta de análisis exploratorio para reducir la dimensión de variables antes de armar modelos de predicción. Puede ser usado para reducir una gran cantidad de variables predictoras en un número menor de componentes principales, en particular si se aplica en conjuntos de datos no muy limpios o con variables fuertemente correlacionadas. En ambos subconjuntos de datos no fue posible explicar un porcentaje de variabilidad importante sin considerar un gran número de componentes. Aunque menor a la cantidad original en cada caso, el no poder reducir a dos a tres componentes (como suele ocurrir en conjuntos) implica que la variabilidad no está dominada por unas pocas bacterias sino que justamente revela el comportamiento complejo y dinámico de la comunidad ecológica.

Un aspecto de gran relevancia en la minería de datos ómicos al momento de presentar resultados es el análisis visual. Por ello, se eligió implementar el algoritmo de SOM (mapas auto-organizados) que usa como modelo el espacio de vectores para representar datos en mapas de dos dimensiones.

Los resultados mostraron que cada capa o mapa representa un tipo de dato: un mapa bidimensional para cada variable de metadatos (BMI, nacionalidad, edad y diversidad) del mismo modo que un mapa para cada bacteria (cuya abundancia relativa está representada en niveles de expresión del gen 16S RNAr). La interpretación de los resultados de SOM es mediante

la exploración de las relaciones geométricas entre los nodos (dado que cada mapa preserva la forma y densidad) lo que favorece la identificación más directa de similitudes y diferencias entre las capas.

Sin dejar de tener en cuenta que la relación entre la obesidad y el microbioma es compleja y variable, el tipo de visualización de resultados que ofrecen los mapas autoorganizados representa un elemento que podría sumar en el análisis de una población en estudio dentro del marco de un proyecto biomédico.

El análisis de la diversidad de microorganismos que habitan en el cuerpo humano es fundamental para comprender la estructura, biología y ecología de las comunidades de los mismos; dicho análisis representa un primer componente crítico en el estudio de la microbiota.

Aún cuando en condiciones de salud se mantiene un núcleo central de bacterias relativamente estable en el tracto gastrointestinal a lo largo de la etapa adulta, la diversidad bacteriana depende de un número de factores tanto intrínsecos como extrínsecos del hospedador.

Por lo tanto, para determinar si era posible predecir la diversidad biológica del sistema se usó un algoritmo de aprendizaje supervisado: *Random Forest*. Es un método de ensamble, basado en entrenar múltiples modelos débiles (*weak learners*) para resolver el mismo problema y así generar un modelo fuerte con mejor desempeño que sus componentes individuales. Luego de entrenar los datos con RF es posible tener una idea de las variables con mayor influencia en el modelo.

En el modelo generado por RF para el subconjunto de *Bacteroidetes* se observó que entre las primeras variables más importantes estaban la edad, el BMI y la nacionalidad. Y en cuanto a la abundancia de bacterias, *Prevotella tannerae* y *Bacteroides ovatus*.

La progresión de la edad del hospedador es un factor importante sobre la diversidad de la microbiota intestinal. Con la edad, las funciones beneficiosas que aporta una microbiota intestinal saludable empiezan a disminuir y empieza a aumentar, en cambio, la frecuencia de procesos inflamatorios y enfermedad especialmente en las personas de edad avanzada [Nagpal et al., 2018, Xu et al., 2019].

En cuanto a la influencia del BMI, diversos estudios que comparan la microbiota intestinal entre individuos obesos y delgados indican que la variación en el grado de diversidad está

asociada al peso corporal: los individuos obesos (BMI alto) presentan una diversidad baja [Turnbaugh et al., 2009, Bäckhed et al., 2012].

El origen geográfico de un individuo (representado en el conjunto de datos a través de la variable nacionalidad) involucra una cantidad de factores ambientales que ejercen un efecto específico en la adquisición, composición y estabilidad de la microbiota intestinal. Varios trabajos muestran las diferencias en composición y diversidad de las microbiotas intestinales humanas en individuos saludables a nivel poblacional [Nishijima et al., 2016, Gupta et al., 2017].

Capítulo 8

Conclusiones

La evaluación de la composición de la microbiota intestinal humana facilita en gran medida la comprensión de su papel en la fisiopatología humana. El conocimiento de los diversos factores que influyen sobre la composición, dinámica y funciones de la microbiota podría servir como una herramienta complementaria en la práctica médica.

El creciente interés en investigar el microbioma intestinal y su influencia en la salud y enfermedad ha acelerado la necesidad de conocimiento y manejo de técnicas que permitan acceder a información valiosa de conjuntos de datos biológicos. En el ámbito de la salud es necesario cada vez más el trabajo de grupos multidisciplinarios, incluyendo a especialistas en el manejo, análisis e interpretación de datos y métodos (limpieza, filtrado, elección de algoritmos, interpretación, etc.). Así, la ciencia de datos puede contribuir a la traducción de resultados innovadores en conocimientos valiosos que brinden apoyo a la toma de decisiones, algo muy requerido en la **medicina de precisión**.

Una contribución de esta Tesis es aportar conocimiento sobre el manejo, interpretación y posibles usos de herramientas de minería de datos y aprendizaje automático aplicados al estudio de datos de microbioma intestinal humano.

Bibliografía

[h2o, 2020] (2020). *h2o: R Interface for H2O*. R package version 3.30.0.6.

[Abdi and Williams, 2010] Abdi, H. and Williams, L. J. (2010). Overview-principal component analysis. *WIREs Computational Statistics*, 2.

[Alberts et al., 2002] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., et al. (2002). Molecular biology of the cell, 6th edn new york. *NY: Garland Science*.

[An et al., 2023] An, J., Kwon, H., and Kim, Y. J. (2023). The firmicutes/bacteroidetes ratio as a risk factor of breast cancer. *Journal of Clinical Medicine*, 12(6):2216.

[Arnold et al., 2016] Arnold, J. W., Roach, J., and Azcarate-Peril, M. A. (2016). Emerging technologies for gut microbiome research. *Trends in microbiology*, 24(11):887–901.

[Arumugam et al., 2011] Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., et al. (2011). Enterotypes of the human gut microbiome. *nature*, 473(7346):174–180.

[Asan and Ercan, 2012] Asan, U. and Ercan, S. (2012). An introduction to self-organizing maps. In *Computational Intelligence Systems in Industrial Engineering*, pages 295–315. Springer.

[Bäckhed et al., 2004] Bäckhed, F., Ding, H., Wang, T., Hooper, L. V., Koh, G. Y., Nagy, A., Semenkovich, C. F., and Gordon, J. I. (2004). The gut microbiota as an environmental factor that regulates fat storage. *Proceedings of the national academy of sciences*, 101(44):15718–15723.

[Bäckhed et al., 2012] Bäckhed, F., Fraser, C. M., Ringel, Y., Sanders, M. E., Sartor, R. B., Sherman, P. M., Versalovic, J., Young, V., and Finlay, B. B. (2012). Defining a healthy human gut microbiome: current concepts, future directions, and clinical applications. *Cell host & microbe*, 12(5):611–622.

[Bäckhed et al., 2005] Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A., and Gordon, J. I. (2005). Host-bacterial mutualism in the human intestine. *science*, 307(5717):1915–1920.

[Bajaj et al., 2014] Bajaj, J. S., Heuman, D. M., Hylemon, P. B., Sanyal, A. J., White, M. B., Monteith, P., Noble, N. A., Unser, A. B., Daita, K., Fisher, A. R., et al. (2014). Altered profile of human gut microbiome is associated with cirrhosis and its complications. *Journal of hepatology*, 60(5):940–947.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[Burgos et al., 2021] Burgos, V., Piñero, T., Fernández, M. L., and Risk, M. (2021). A complementary approach in the analysis of the human gut microbiome applying self-organizing maps and random forest. In *International Conference on Applied Informatics*, pages 97–110. Springer.

[Chaban et al., 2006] Chaban, B., Ng, S. Y., and Jarrell, K. F. (2006). Archaeal habitats—from the extreme to the ordinary. *Canadian journal of microbiology*, 52(2):73–116.

[Chung et al., 2020] Chung, W. S. F., Walker, A. W., Bosscher, D., Garcia-Campayo, V., Wagner, J., Parkhill, J., Duncan, S. H., and Flint, H. J. (2020). Relative abundance of the prevotella genus within the human gut microbiota of elderly volunteers determines the inter-individual responses to dietary supplementation with wheat bran arabinoxylan-oligosaccharides. *BMC microbiology*, 20(1):1–14.

[Coenen, 2019] Coenen (2019). Understanding umap. `https://github.com/PAIR-code/understanding-umap`.

[Curry et al., 2021] Curry, K. D., Nute, M. G., and Treangen, T. J. (2021). It takes guts to learn: machine learning techniques for disease detection from the gut microbiome. *Emerging Topics in Life Sciences*, 5(6):815–827.

[Del Chierico et al., 2015] Del Chierico, F., Ancora, M., Marcacci, M., Camma, C., Putignani, L., and Conti, S. (2015). Choice of next-generation sequencing pipelines. In *Bacterial Pangenomics*, pages 31–47. Springer.

[Dominguez-Bello et al., 2010] Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., and Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences*, 107(26):11971–11975.

[Donaldson et al., 2016] Donaldson, G. P., Lee, S. M., and Mazmanian, S. K. (2016). Gut biogeography of the bacterial microbiota. *Nature Reviews Microbiology*, 14(1):20–32.

[Dutta et al., 2019] Dutta, S. K., Verma, S., Jain, V., Surapaneni, B. K., Vinayek, R., Phillips, L., and Nair, P. P. (2019). Parkinson's disease: the emerging role of gut dysbiosis, antibiotics, probiotics, and fecal microbiota transplantation. *Journal of neurogastroenterology and motility*, 25(3):363.

[Egert et al., 2006] Egert, M., de Graaf, A. A., Smidt, H., de Vos, W. M., and Venema, K. (2006). Beyond Diversity: Functional Microbiomics of the Human Colon. *Trends Microbiol*, 14(2):86–91.

[Eloe-Fadrosh and Rasko, 2013] Eloe-Fadrosh, E. A. and Rasko, D. A. (2013). The human microbiome: from symbiosis to pathogenesis. *Annual review of medicine*, 64:145–163.

[European Society Neurogastroenterology and Motility, 2020] European Society Neurogastroenterology and Motility (2020). Comensal (bacteria) - editado por european society of neurogastroenterology and motility. `https://www.gutmicrobiotaforhealth.com/es/glossary/comensal-bacteria/`, Last accessed on 2021-11-08.

[Floch et al., 2016] Floch, M. H., Ringel, Y., and Walker, W. A. (2016). *The microbiota in gastrointestinal pathophysiology: implications for human health, prebiotics, probiotics, and dysbiosis*. Academic Press.

[Gandía-Aguiló et al., 2017] Gandía-Aguiló, V., Cibrián, R., Soria, E., Serrano, A.-J., Aguiló, L., Paredes, V., and Gandía, J.-L. (2017). Use of self-organizing maps for analyzing the

behavior of canines displaced towards midline under interceptive treatment. *Medicina oral, patologia oral y cirugia bucal*, 22(2):e233.

[Gareth et al., 2013] Gareth, J., Daniela, W., Trevor, H., and Robert, T. (2013). *An introduction to statistical learning: with applications in R*. Spinger.

[Gill et al., 2006] Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., and Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *science*, 312(5778):1355–1359.

[Greenhalgh et al., 2016] Greenhalgh, K., Meyer, K. M., Aagaard, K. M., and Wilmes, P. (2016). The human gut microbiome in health: establishment and resilience of microbiota over a lifetime. *Environmental microbiology*, 18(7):2103–2116.

[Gupta et al., 2017] Gupta, V. K., Paul, S., and Dutta, C. (2017). Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. *Frontiers in microbiology*, 8:1162.

[Han et al., 2011] Han, J., Kamber, M., and Pei, J. (2011). Data mining concepts and techniques third edition. *Morgan Kaufmann*.

[Han et al., 2019] Han, L., Yang, G., Dai, H., Yang, H., Xu, B., Li, H., Long, H., Li, Z., Yang, X., and Zhao, C. (2019). Combining self-organizing maps and biplot analysis to preselect maize phenotypic components based on uav high-throughput phenotyping platform. *Plant methods*, 15(1):1–16.

[Horning et al., 2010] Horning, N. et al. (2010). Random forests: An algorithm for image classification and generation of continuous fields data sets. In *Proceedings of the International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences, Osaka, Japan*, volume 911.

[Hug et al., 2016] Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., et al. (2016). A new view of the tree of life. *Nature microbiology*, 1(5):16048.

[James et al., 2013] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

[Koenig et al., 2011] Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., Angenent, L. T., and Ley, R. E. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4578–4585.

[Lahti et al., 2014] Lahti, L., Salojärvi, J., Salonen, A., Scheffer, M., and De Vos, W. M. (2014). Tipping elements in the human intestinal ecosystem. *Nature communications*, 5(1):1–10.

[Lai et al., 2018] Lai, K., Twine, N., O'brien, A., Guo, Y., and Bauer, D. (2018). Artificial intelligence and machine learning in bioinformatics. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 55:272.

[Larsbrink et al., 2014] Larsbrink, J., Rogers, T. E., Hemsworth, G. R., McKee, L. S., Tauzin, A. S., Spadiut, O., Klinter, S., Pudlo, N. A., Urs, K., Koropatkin, N. M., et al. (2014). A discrete genetic locus confers xyloglucan metabolism in select human gut bacteroidetes. *Nature*, 506(7489):498–502.

[Liu et al., 2019] Liu, H., Han, M., Li, S. C., Tan, G., Sun, S., Hu, Z., Yang, P., Wang, R., Liu, Y., Chen, F., et al. (2019). Resilience of human gut microbial communities for the long stay with multiple dietary shifts. *Gut*, 68(12):2254–2255.

[Lloyd-Price et al., 2016] Lloyd-Price, J., Abu-Ali, G., and Huttenhower, C. (2016). The healthy human microbiome. *Genome medicine*, 8(1):1–11.

[Manos, 2022] Manos, J. (2022). The human microbiome in disease and pathology. *Apmis*, 130(12):690–705.

[Matsuoka and Kanai, 2015] Matsuoka, K. and Kanai, T. (2015). The gut microbiota and inflammatory bowel disease. In *Seminars in immunopathology*, volume 37, pages 47–55. Springer.

[McInnes et al., 2018] McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

[Morgan et al., 2013] Morgan, X. C., Segata, N., and Huttenhower, C. (2013). Biodiversity and functional genomics in the human microbiome. *Trends in genetics*, 29(1):51–58.

[Nagpal et al., 2018] Nagpal, R., Mainali, R., Ahmadi, S., Wang, S., Singh, R., Kavanagh, K., Kitzman, D. W., Kushugulova, A., Marotta, F., and Yadav, H. (2018). Gut microbiome and aging: Physiological and mechanistic insights. *Nutrition and healthy aging*, 4(4):267–285.

[Nishijima et al., 2016] Nishijima, S., Suda, W., Oshima, K., Kim, S.-W., Hirose, Y., Morita, H., and Hattori, M. (2016). The gut microbiome of healthy japanese and its microbial and functional uniqueness. *DNA Research*, 23(2):125–133.

[Paliy and Agans, 2012] Paliy, O. and Agans, R. (2012). Application of phylogenetic microarrays to interrogation of human microbiota. *FEMS microbiology ecology*, 79(1):2–11.

[Paliy et al., 2014] Paliy, O., Shankar, V., and Sagova-Mareckova, M. (2014). Phylogenetic microarrays. *Bio-informatics and data analysis in microbiology. ÖT Bishop (ed.). Caister Academic Press, Norfolk*, pages 207–230.

[Qin et al., 2010] Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *nature*, 464(7285):59–65.

[Rabot et al., 2016] Rabot, S., Membrez, M., Blancher, F., Berger, B., Moine, D., Krause, L., Bibiloni, R., Bruneau, A., Gérard, P., Siddharth, J., et al. (2016). High fat diet drives obesity regardless the composition of gut microbiota in mice. *Scientific reports*, 6(1):1–11.

[Ravel et al., 2011] Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S., McCulle, S. L., Karlebach, S., Gorle, R., Russell, J., Tacket, C. O., et al. (2011). Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4680–4687.

[Reese and Dunn, 2018] Reese, A. and Dunn, R. (2018). Drivers of Microbiome Biodiversity: A Review of General Rules, Feces, and Ignorance. *mBio*, 9(4):e01294–18.

[Reijnders et al., 2016] Reijnders, D., Goossens, G. H., Hermes, G. D., Neis, E. P., van der Beek, C. M., Most, J., Holst, J. J., Lenaerts, K., Kootte, R. S., Nieuwdorp, M., et al. (2016). Effects of gut microbiota manipulation by antibiotics on host metabolism in obese humans: a randomized double-blind placebo-controlled trial. *Cell metabolism*, 24(1):63–74.

[Ruan et al., 2011] Ruan, X., Gao, Y., Song, H., Chen, J., et al. (2011). A new dynamic self-organizing method for mobile robot environment mapping. *Journal of Intelligent Learning Systems and Applications*, 3(04):249.

[Ruggles et al., 2018] Ruggles, K. V., Wang, J., Volkova, A., Contreras, M., Noya-Alarcon, O., Lander, O., Caballero, H., and Dominguez-Bello, M. G. (2018). Changes in the gut microbiota of urban subjects during an immersion in the traditional diet and lifestyle of a rainforest village. *mSphere*, 3(4):e00193–18.

[Russell, 2010] Russell, S. J. (2010). *Artificial intelligence a modern approach*. Pearson Education, Inc.

[Salonen et al., 2014] Salonen, A., Lahti, L., Salojärvi, J., Holtrop, G., Korpela, K., Duncan, S. H., Date, P., Farquharson, F., Johnstone, A. M., Lobley, G. E., et al. (2014). Impact of diet and individual variation on intestinal microbiota composition and fermentation products in obese men. *The ISME journal*, 8(11):2218–2230.

[Sansonetti, 2011] Sansonetti, P. (2011). To be or not to be a pathogen: that is the mucosally relevant question. *Mucosal immunology*, 4(1):8–14.

[Sarker, 2021] Sarker, I. (2021). Machine learning: Algorithms, real-world applications and research directions. sn comput sci 2, 160.

[Shastry and Sanjay, 2020] Shastry, K. A. and Sanjay, H. (2020). Machine learning for bioinformatics. In *Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications*, pages 25–39. Springer.

[Shin et al., 2015] Shin, N.-R., Whon, T. W., and Bae, J.-W. (2015). Proteobacteria: microbial signature of dysbiosis in gut microbiota. *Trends in biotechnology*, 33(9):496–503.

[Spor et al., 2011] Spor, A., Koren, O., and Ley, R. (2011). Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Reviews Microbiology*, 9(4):279–290.

[Stojanov et al., 2020] Stojanov, S., Berlec, A., and Štrukelj, B. (2020). The influence of probiotics on the firmicutes/bacteroidetes ratio in the treatment of obesity and inflammatory bowel disease. *Microorganisms*, 8(11):1715.

[Tan et al., 2018] Tan, H., Yu, Z., Wang, C., Zhang, Q., Zhao, J., Zhang, H., Zhai, Q., and Chen, W. (2018). Pilot safety evaluation of a novel strain of bacteroides ovatus. *Frontiers in genetics*, 9:539.

[Tibshirani et al., 2001] Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

[Tims et al., 2013] Tims, S., Derom, C., Jonkers, D. M., Vlietinck, R., Saris, W. H., Kleerebezem, M., De Vos, W. M., and Zoetendal, E. G. (2013). Microbiota conservation and bmi signatures in adult monozygotic twins. *The ISME journal*, 7(4):707–717.

[Tortora et al., 2007] Tortora, G. J., Funke, B. R., and Case, C. L. (2007). *Introducción a la microbiología*. Ed. Médica Panamericana.

[Tropini et al., 2017] Tropini, C., Earle, K. A., Huang, K. C., and Sonnenburg, J. L. (2017). The gut microbiome: connecting spatial organization to function. *Cell host & microbe*, 21(4):433–442.

[Turnbaugh et al., 2009] Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P., et al. (2009). A core gut microbiome in obese and lean twins. *nature*, 457(7228):480–484.

[Van Ameringen et al., 2019] Van Ameringen, M., Turna, J., Patterson, B., Pipe, A., Mao, R. Q., Anglin, R., and Surette, M. G. (2019). The gut microbiome in psychiatry: A primer for clinicians. *Depression and Anxiety*, 36(11):1004–1025.

[Vijay and Valdes, 2022] Vijay, A. and Valdes, A. M. (2022). Role of the gut microbiome in chronic diseases: A narrative review. *European Journal of Clinical Nutrition*, 76(4):489–501.

[Wan et al., 2020] Wan, Y., Yuan, J., Li, J., Li, H., Yin, K., Wang, F., and Li, D. (2020). Overweight and underweight status are linked to specific gut microbiota and intestinal tricarboxylic acid cycle intermediates. *Clinical nutrition*, 39(10):3189–3198.

[World Health Organization, 2021] World Health Organization (2021). Body mass index - bmi. `https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi`, Last accessed on 2021-11-08.

[Xu et al., 2019] Xu, C., Zhu, H., and Qiu, P. (2019). Aging progression of human gut microbiota. *BMC microbiology*, 19(1):1–10.

[Yatsunenko et al., 2012] Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., et al. (2012). Human gut microbiome viewed across age and geography. *nature*, 486(7402):222–227.

[Zhang et al., 2022] Zhang, H., Chen, Y., Wang, Z., Xie, G., Liu, M., Yuan, B., Chai, H., Wang, W., and Cheng, P. (2022). Implications of gut microbiota in neurodegenerative diseases. *Frontiers in Immunology*, 13:325.

[Zhou and Gallins, 2019] Zhou, Y.-H. and Gallins, P. (2019). A review and tutorial of machine learning methods for microbiome host trait prediction. *Frontiers in genetics*, 10:579.

Apéndice A

Artículo en ICAI 2021

Parte de los resultados de esta Tesis fueron presentados en la *4th International Conference on Applied Informatics* (ICAI2021) llevada a cabo del 28 al 30 de Octubre del 2021 en Buenos Aires, Argentina. El trabajo fue seleccionado para su publicación [Burgos et al., 2021].

A Complementary Approach in the Analysis of the Human Gut Microbiome Applying Self-organizing Maps and Random Forest

Valeria Burgos[1( )], Tamara Piñero[1], María Laura Fernández[2], and Marcelo Risk[1]

[1] Instituto de Medicina Traslacional e Ingeniería Biomédica (IMTIB) - CONICET, Hospital Italiano de Buenos Aires, Instituto Universitario del Hospital Italiano, Buenos Aires, Argentina
valeria.burgos@hospitalitaliano.org.ar

[2] CONICET - Universidad de Buenos Aires, Instituto de Física del Plasma (INFIP), Buenos Aires, Argentina

Abstract. The human gastrointestinal tract is colonized by millions of microorganisms that make up the so-called gut microbiota, with a vital role in the well-being, health maintenance as well as the appearance of several diseases in the human host. A data mining analysis approach was applied on a set of gut microbiota data from healthy individuals. We used two machine learning methods to identify biomedically relevant relationships between demographic and biomedical variables of the subjects and patterns of abundance of bacteria. The study was carried out focusing on the two most abundant human gut microbiota groups, *Bacteroidetes* and *Firmicutes*. Both subsets of bacterial abundances together with the metadata variables were subjected to an exploratory analysis, using self-organizing maps that integrate multivariate information through different component planes. Finally, to evaluate the relevance of the variables on the biological diversity of the microbial communities, an ensemble-based method such as random forest was used. Results showed that age and body mass index were among the most important features at explaining bacteria diversity. Interestingly, several bacteria species known to be associated to diet and obesity were identified as relevant features as well. In the topological analysis of self-organizing maps, we identified certain groups of nodes with similarities in subject metadata and gut bacteria. We conclude that our results represent a preliminary approach that could be considered, in future studies, as a potential complement in health reports so as to help health professionals personalize patient treatment or support decision making.

Keywords: Gut microbiome · Self-organizing maps · Random forest · Precision medicine · Bioinformatics

H. Florez and M. F. Pollo-Cattaneo (Eds.): ICAI 2021, CCIS 1455, pp. 97–110, 2021.
https://doi.org/10.1007/978-3-030-89654-6_8

1 Introduction

Data science comprises different scientific fields of knowledge to target the analysis of complex and massive data. In particular, the increased interest on the application of machine learning algorithms to extract hidden associations or patterns in electronic health records, processing of medical images, prediction of a health situation or classification of patients has demonstrated the need for machine learning tools for reliable decision-making in healthcare and handling of biological data. The human gastrointestinal tract harbors millions of microorganisms which includes bacteria, archaea, fungi and viruses, interacting in symbiotic relationships between the host and each microbial community. This is known as the gut microbiota, while the collective genome of all symbiotic and pathogenic microorgnisms represents the gut microbiome. The establishment of a large part of the component communities that will remain in the adult life occurs at birth and during the first years of life [1,2] and its composition is shaped not only by the host genetics but also by environmental factors, nutritional status, age and lifestyle. Importantly, the gut microbiome plays an essential role in a number of health-beneficial functions (digestion, synthesis of essential vitamins and amino acids, absorption of calcium, magnesium and iron, fermentation of indigestible components, protection against pathogens, etc.) [3].

The rates of growth and survival of its component populations may fluctuate in response to temporary stressors, such as changes in diet or the consumption of antibiotics [4]. This potential for dynamic restructuring involves two important characteristics of the gut microbiome: plasticity and resilience [5,6]. Ongoing research in human and animal models highlights the importance of a healthy gut microbiome since persistent disbalances in composition and stability, known as dysbiosis, are associated to the onset and progression of chronic diseases that include obesity, irritable bowel syndrome, diabetes, cancer, and neurological diseases such as Parkinson's, among others [7].

The generation of biological knowledge from the large flow of data generated by new technologies in biomedical sciences has accelerated their transformation into data-centered fields. Thus, the study of the human gut microbiome represents a major challenge since it requires an interdisciplinary work between computer science and medicine. The interaction between these two fields will help obtain knowledge about gut bacteria interactions in human health and disease.

In the present work, we analyzed microbiome abundance data and the associated metadata using a machine learning approach: we used the visualization capabilities offered by self-organizing maps to identify patterns of multivariate data stored in multiple layers and additionally, we applied random forest to model the prediction of microbial diversity. For each analysis, we focused on the abundance levels of two major groups of the human gut bacteria, such as *Bacteroidetes* and *Firmicutes*.

2 Methods

2.1 Dataset

Microarray profiling data of human gut microbiota and anonymized metadata were obtained from the Dryad Digital Repository, as described by [8]. Briefly, the data matrix contained 1172 intestinal samples of western adults. In each sample, bacterial abundances were quantified using the HITChip phylogenetic microarray. This technology allows the assessment of relative abundances of gut bacteria through signal intensities of the targeted 16S rRNA gene, frequently used for the identification of poorly described or non cultured bacteria. Data contained hybridization signals for 130 genus-like phylogenetic groups. Subject metadata included age, sex, nationality, probe-level Shannon diversity, BMI group and subjectID. Geographical origin of the study subjects were: Central Europe (Belgium, Denmark, Germany, the Netherlands), Eastern Europe (Poland), Scandinavia (Finland, Norway, Sweden), Southern Europe (France, Italy, Serbia, Spain), United Kingdom/Ireland (UK, Ireland) and the United States (US). We used VIM and tidyverse R packages [9,10] to check for the presence of missing values (NAs). Records containing NAs were carefully removed without causing bias in the dataset. During the cleaning process, the category 'Eastern Europe' was turned out since it was represented by only one complete case. The final dataset to be used was represented by 1056 complete patient records containing 130 bacterial abundance data and subject metadata.

2.2 Self-organizing Maps (SOMs)

Self-organizing maps (also known as Kohonen maps) represent an optimal option to organize multidimensional data in a two-dimensional space by using a neural network. SOM uses the vector space as a model to represent data in a two-dimensional lattice: each value through N samples could be referred to as a data point in an N-dimensional space. Thousands of data points would therefore form data clouds in space, with a intrinsic topology due to geometric relationships. From this it follows that the greater the similarity in the data value level, the closer is the geometric space they occupy. To visualize the trained SOM, we used heatmaps for each variable to plot the degree of connectivity between adjacent output neurons through the use of a color intensity panel. In the case of multivariate datasets, the visualization of different heatmaps allows an overall analysis of the relations between the variables since maps are linked to each other by position: in each map, a node in a given location corresponds to the same unit in another map. The SOM map can be implemented in different topologies. Data were divided into two subsets by major groups of gut bacteria (*Bacteroidetes* and *Firmicutes*). We used a regular hexagonal 2D grid consisting of 750 neurons, in 30×25 grids. Data was logarithmically-scaled before training. We used the kohonen R package, which provides a standardized framework for SOMs.

2.3 Random Forest

Random forest (RF) is an ensemble learning method that can solve regression and classification problems. The algorithm uses a random subset of the training samples for each tree and a random subset of predictors in each step during the training process. These two sources of randomization make the algorithm robust to correlated predictors and more reliable at obtaining average outputs into a model. Data were divided into two subsets by major groups of gut bacteria (*Bacteroidetes* and *Firmicutes*). Bacterial abundance data and metadata were used as RF regressors to generate a diversity prediction model, which was performed using the RF regression algorithm provided by the R interface for h2o [11]. We used 10-fold cross validation for training the regression models and their performance were evaluated using Mean Absolute Error (MAE) as the error metric. After parameter tuning (mainly focused on the number of trees) through cross validation, the best RF regression model was selected.

3 Results

3.1 Characteristics of the Study Population

The degree of obesity is a relevant aspect in gut microbiome studies in terms of its influence on the microbiota composition [12,13]. This parameter, that can be obtained through the body mass index (BMI), indicates the nutritional status of an individual. Descriptive analysis of the study population showed that lean individuals were homogeneously distributed in all age groups, while overweight, obese and severe obese categories were more abundant in 45–60 year-old individuals. A large proportion of the underweight population was represented between 20–30 years old (Fig. 1).

The distribution of the different BMI categories in each geographic region showed that lean individuals represented approximately half of the proportions for all locations. For Scandinavia, Southern Europe, UK/Ireland (UKIE) and the US, the following proportion was represented by overweight subjects. In contrast, in Central Europe, the second proportion after lean individuals was represented by obese individuals. Morbid obese subjects were present only in Central Europe (Fig. 2).

3.2 SOM Analysis

Each component plane or map in a SOM represents one type of data: a two-dimensional lattice for each metadata variable (BMI, nationality, age, sex and diversity) as well as for each bacteria (whose relative abundance is represented in expression levels of the 16S rRNA gene). Since each map preserves shape and density, exploration of the geometric relationships between nodes allows a direct identification of similarities and differences between the layers.

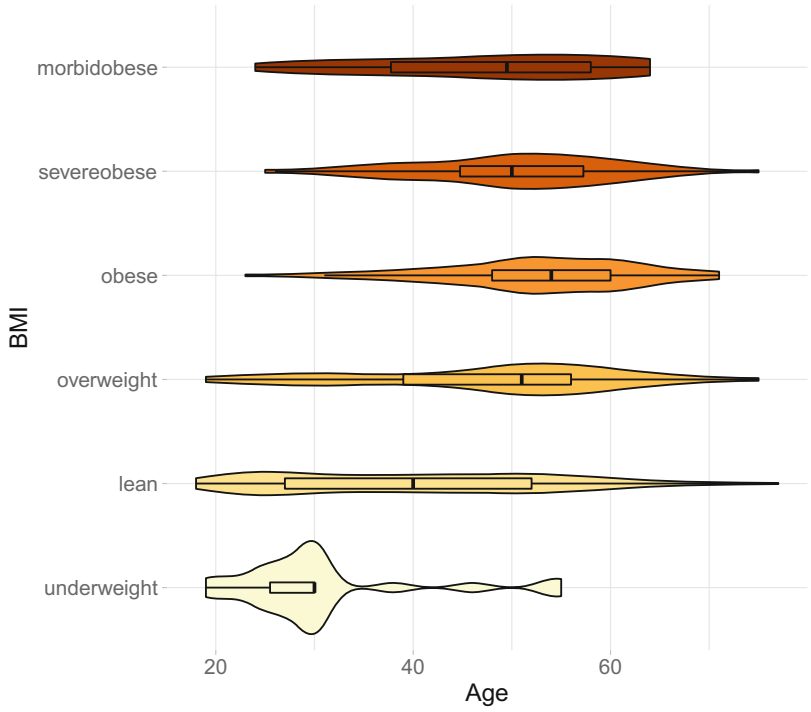

Fig. 1. Distribution of the BMI categories across age intervals in the study population.

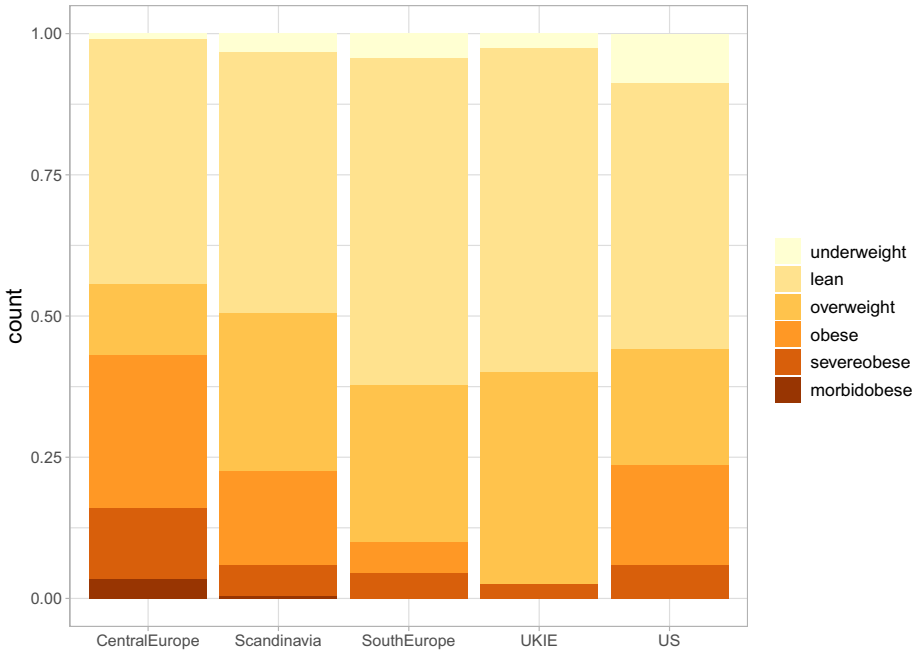

Fig. 2. Proportion of each BMI category of the study subjects across different geographic locations.

Bacteroidetes. After running the SOM algorithm using the *Bacteroidetes* subset, the different regions in each map indicated that the age distributed into two well-defined subregions of nodes for younger ages (around 20 years old), quite separated from a small group of nodes representing older individuals (older than 65 years old). Nodes representing 40–50 year old subjects were scattered throughout the map (Fig. 3A). Men and women were clearly distributed in two large, separated areas in the map (Fig. 3B). Scandinavian individuals mapped in a few groups of isolated nodes while Central European subjects were homogeneously distributed. Interestingly, the map identified an isolated group of nodes corresponding to individuals from the United States (Fig. 3C). Additionally, underweight and lean subjects mapped in two wide groups, while severe and morbid obese individuals mapped mainly in a small and defined group of nodes (Fig. 3E). The distribution of microbial diversity showed that two small and defined groups of nodes mapped low diversity values while higher values distributed into larger and clearly defined subsets of nodes across the map (Fig. 3D).

After SOM training, the abundance levels of each bacteria species of the *Bacteroidetes* phylum was also represented in a map. In the present dataset, several members belonging to this phylum were identified but no overlay between any bacterial map was observed (data not shown). However, since many members of the *Bacteroidetes* community have a relevant role in the host health, we chose to analyze two prominent bacteria whose relative abundances are known to be influenced by the host lifestyle and diet: a high fat and protein intake is associated with elevated microbial presence of *Bacteroides* species, while a high fiber intake is associated with high microbial levels of *Prevotella* species [14,15]. It appears that the abundances of these two *Bacteroidetes* members showed no overlay between any host metadata map because there are no coincidences in location (Fig. 3F and G).

The superimposition of the multiple maps described above allows to obtain some clear aspects of the data from the perspective of the *Bacteroidetes* subset:

- Low diversity values overlaps with a lower BMI (lean subjects) corresponding to young men from Central Europe and Scandinavia (indicated by a red circle in Fig. 3A–E).
- Interestingly, another subset of low diversity values overlaps with high BMI values (that is, severe to morbid obese) corresponding to middle-aged female individuals from Central Europe (indicated by a black circle in Fig. 3A–E).
- The small proportion of young subjects is equally represented in both men and women, while older individuals correspond exclusively to men.
- There are no women older than 65 years.
- US nationality corresponds to lean women in the range of medium values of microbial diversity.

Firmicutes. When subject metadata was trained using the *Firmicutes* subset, the different layers showed a more disperse variable distribution in the maps: middle-age individuals were homogeneously represented throughout the lattice while younger (around 20 years old) and older ages (older than 65 years) mapped

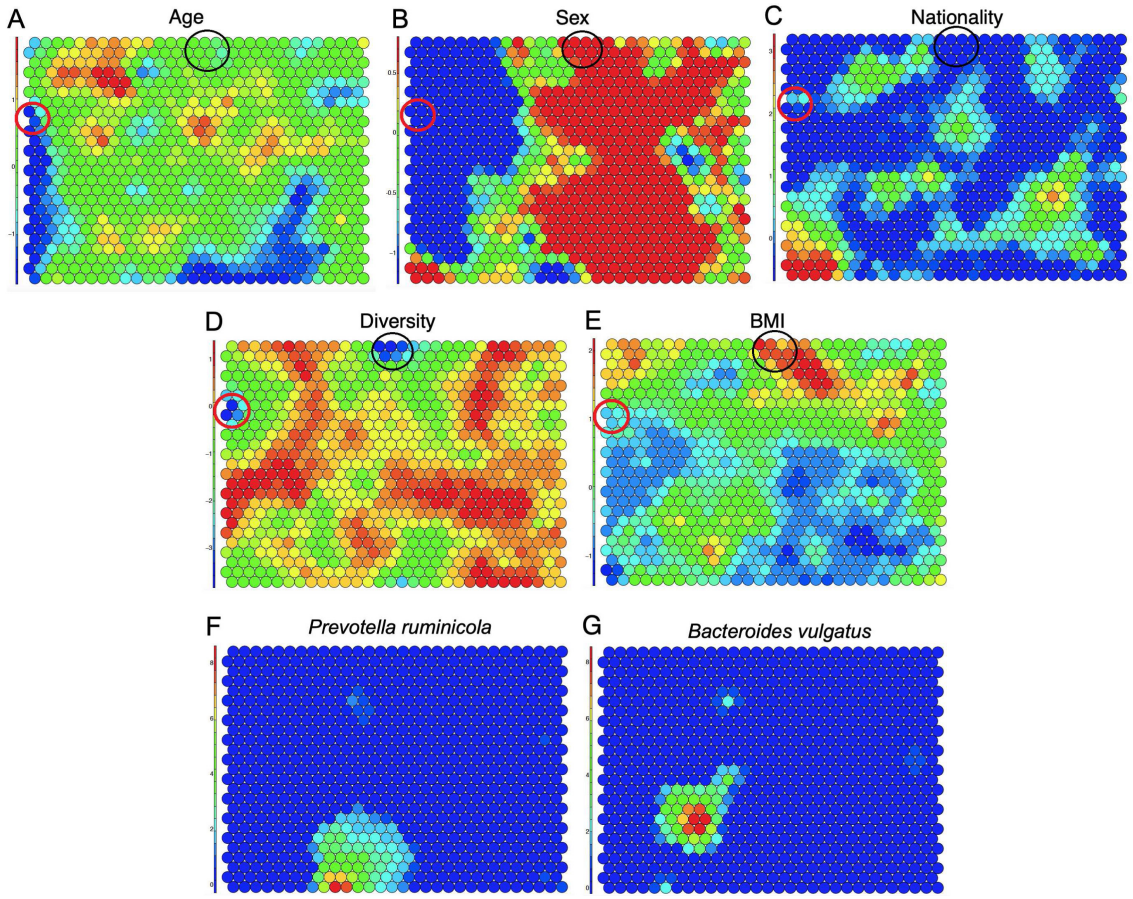

Fig. 3. SOM results for metadata variables associated with *Bacteroidetes*. The color index of each map is established based on the values for each variable. A) Age, scale adjusted for a range from 18 to 77 years old, blue = younger, red = older adults. B) Sex, blue = male, red = female. C) Nationality, color gradient beginning at Central Europe (level 0, color blue), Scandinavia (level 1, light blue), Southern Europe (level 2, green), UK/Ireland (level 3, orange) and the US (level 4, red). D) Diversity, scale adjusted for a range of 4.7 to 6.3 diversity index values beginning at blue (lowest diversity) to red (highest diversity). E) BMI, color gradient beginning at underweight (color blue), lean (light blue), overweight (green), obese (yellow), severe obese (orange) and morbid obese (red). In F) *P. ruminicola* and G) *B. vulgatus* the color gradient represents the level of abundance, blue = low, red = high. (Color figure online)

in small, discrete groups of nodes (Fig. 4A). Men and women were not as clearly separated in their node distribution as in the *Bacteroidetes* subset (Fig. 4B). Regarding nationality, a node pattern similar to *Bacteroidetes* was observed (Fig. 4C). High microbial diversity values predominated throughout the map while only three nodes mapped for low diversity values (Fig. 4D). Additionally, lean and underweight subjects predominated in most of the nodes and only a very small group of nodes grouped the highest BMI values, corresponding to the severe obese category (Fig. 4E).

The phylum *Firmicutes* is made up of around 250 different genera of bacteria, such as *Lactobacillus*, *Bacillus*, *Clostridium*, *Enterococcus*, and *Ruminicoccus*,

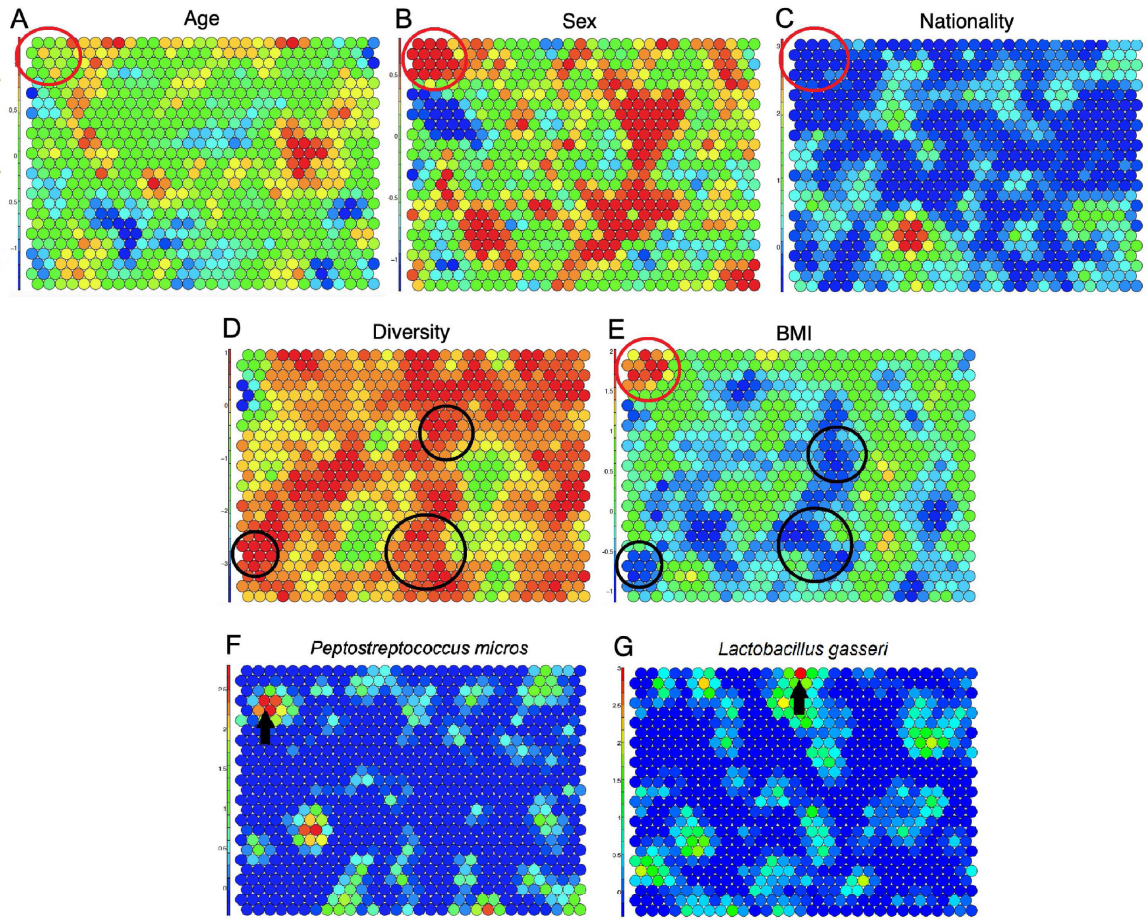

Fig. 4. SOM results for metadata variables associated with *Firmicutes*. The color index of each map is established based on the values for each variable. A) Age, scale adjusted for a range from 18 to 77 years old, blue = younger, red = older adults. B) Sex, blue = male, red = female. C) Nationality, color gradient beginning at Central Europe (level 0, color blue), Scandinavia (level 1, light blue), Southern Europe (level 2, green), UK/Ireland (level 3, orange) and the US (level 4, red). D) Diversity, scale adjusted for a range of 4.7 to 6.3 diversity index values beginning at blue (lowest diversity) to red (highest diversity). E) BMI, color gradient beginning at underweight (color blue), lean (light blue), overweight (green), obese (yellow), severe obese (orange) and morbid obese (red). In F) *P. micros* and G) *L. gasseri* the color gradient represents the level of abundance, blue = low, red = high. (Color figure online)

among other important members. In the present dataset, 74 members belonging to this phylym were identified and consequently, each generated a heatmap after SOM training (data not shown). However, superimposing multiple maps revealed that only one node corresponding to overweight individuals slightly coincided with higher values of abundance of a single species, *Peptostreptococcus micros* (indicated by an arrow in Fig. 4F). This represents an interesting result since several other members of *Firmicutes* have previously been associated to obesity [16,17]. Additionally, an overlay between high microbial diversity and higher values of abundance for *Lactobacillus gasseri* was observed (indicated by an arrow in Fig. 4G).

The map overlay between diversity and BMI categories allowed to observe that:

- Severe to morbid obese individuals were middle-age women from Central Europe (indicated by a red circle in Fig. 4A, B, C and E).
- Many of the maximum diversity values superimpose with the lowest BMI categories (indicated by a black circle in Fig. 4D and E).

3.3 Random Forest

Diversity is an important variable in microbiome research because it describes the richness (the number of classes) and distribution of the component microorganisms among classes. Understanding diversity in the intestinal microbial community also allows us to understand the impact of factors on bacteria distribution, such as the use of antibiotics, type of diet, degree of obesity, medical interventions and environmental factors, among others (reviewed in [18]). We developed a RF regression model of microbiome diversity. A very useful visual way to interpret RF results of the prediction model is through a ranking list of feature importance, which refers to the relative influence of each feature on the target variable. It considers whether a variable was selected to split and how much the squared error (over all trees) improved as a result.

Bacteroidetes. We observed that the age of the individuals was the most important factor in the regression model to predict the diversity of the gut microbiota. Regarding the bacterial abundances, it was observed that two species had the greatest relative importance in the identification of diversity: *Prevotella ruminicola* and *Bacteroides ovatus*. On the other hand, of the remaining metadata variables, the BMI of individuals had a remarkable position in the ranking of importance, followed by the geographic location. Gender was not important in the prediction of diversity. While the list of importance for each variable is informative, a better interpretation is obtained after scaling between 0 and 1 in a descending order of importance (Fig. 5A).

Firmicutes. When data was RF-trained considering this subset, the first three positions of the list were occupied by the BMI of the individuals, followed by Nationality and Age. Regarding the abundances of the 74 types of bacteria, it was observed that two members had the greatest relative importance in predicting diversity: *Lachnobacillus bovis* and *Granulicatella* (Fig. 5B).

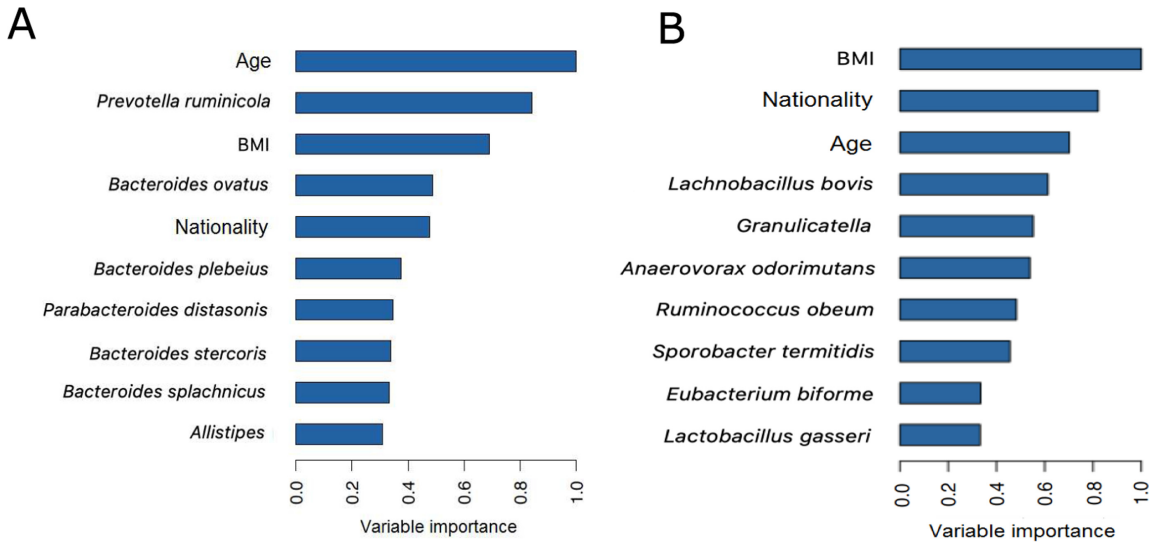

Fig. 5. Scale of the relative importance of the first ten variables on the diversity of *Bacteroidetes* (A) and *Firmicutes* (B) after implementing random forest.

4 Discussion

We sought to characterize the relations between subject metadata and specific bacterial members of the gut microbiome in a thousand western adults through the use of two machine learning methods that provide robust analytical visualizations, such as self-organizing maps and random forest. Bacteria of the human intestinal microbiome are taxonomically classified into six large groups or phyla, which in turn are subclassified into classes, orders, families, genera and species. The present work addressed the analysis on two of the most abundant phyla, *Bacteroidetes* and *Firmicutes*, which represent 90% of the gut microbiota [19].

The configuration of the SOM outcome maintains the topological structure of the input multidimensional data, in which similar values are mapped in the same or near node in a two dimensional map. Such topological preservation is of particular significance in the exploratory phase of omics data mining since there is generally no *a priori* knowledge of data structure. We presented the visualization of different superimposed heatmaps that allowed the exploration of relationships between input variables. This way of presenting SOM outcomes is similar to previous studies [20].

Our results showed that only one species of *Firmicutes*, *Peptostreptococcus micros*, was slightly associated to nodes that grouped overweight individuals, mostly middle-aged women. Notably, several previous studies have shown that *Peptostreptococcus micros* (later classified as *Parvimonas micra*) is one of various colorectal cancer (CRC) microbial markers [13,21]. This bacteria has also been reported to have a pathogenic role in periodontal diseases [22]. Considering the behavior of the oral microbiome during periodontal infection and its influence in health complications, such as diabetes, cardiovascular disease, and obesity [23], the significance of the results obtained here provides a basis for future studies on the possible role of gut bacteria as biological markers of a

developing overweight condition. Among bacteria with known beneficial roles, we observed that high abundance of *Lactobacillus gasseri* was related to nodes that grouped high diversity values. This result supports the previously reported role of *Lactobacillus gasseri* in the management of obesity and probiotic properties [24,25]. The topological structures of the metadata variables were slightly different depending whether *Bacteroidetes* or *Firmicutes* subsets were used for SOM training. For both age and sex, mapping distribution of the study subjects was more effective using the *Bacteroidetes* data. In the case of the different BMI categories, subject distribution was more effective using *Firmicutes* data.

Although the SOM results presented here allowed us to gain insight into the different regions of matching information underlying host metadata and the relative abundance of bacteria, we consider that a deeper approach of the use of SOM is needed, in terms of parameter configuration, such as size, dimensionality, shape, learning rate, among others. Considering that the relationships between gut microbiome and host BMI are dynamic and complex, self-organizing maps provide an excellent tool of visualization and dimensionality reduction that could serve as a complementary tool in a biomedical report.

Analysis of the microbiome diversity in the human body is essential to understand the structure, biology and ecology of its component communities. This analysis represents a critical first step in microbiome studies. When supervised learning through a regression random forest algorithm was used to determine which variables were important in the prediction of microbial diversity, we observed that both age and BMI category of the individuals appeared as the most relevant in the regression models generated for *Bacteroidetes* and *Firmicutes* subsets. The contribution of these physiological factors in shaping the gut microbiome has been reported previously: with age, the beneficial functions provided by a healthy gut microbiome begin to decrease in association to an increasing frequency of inflammatory processes and disease, especially in the elderly. Regarding the influence of BMI, various studies that compare the intestinal microbiota between obese and lean individuals indicate that the variation in the degree of diversity is associated with body weight (obese individuals present a low diversity, which means a higher BMI) [26–29].

Random forest regression also indicated that two *Bacteroidetes* species were the most relevant on diversity: *Prevotella ruminicola* is involved in the response of individuals to dietary supplements [30] and *Bacteroides ovatus* is a dominant species in the human intestine, previously identified as a next generation probiotic due to its preventive effects on intestinal inflammation (reviewed in [31]). On the other hand, two members of the *Firmicutes* group were identified as important at predicting diversity in this subset: *Lachnobacillus bovis* and *Granulicatella* whose relative abundances are reported to be influenced by the type of diet and the use of antibiotics in obese individuals [32,33]. Hence, some of the results obtained in both RF regression models are consistent with published microbiome research, indicating the robustness of the RF regression algorithm. A further analysis is needed that involves a complete parameter tuning so as to characterize the most accurate RF setting for a microbiome project.

In the last decade, the impulse provided by innovative developments in technology and the generation of large volumes of microbiome data has caused an increase in the use of machine learning methods in this field, such as microbial ecology, identification of certain bacteria to cancer and forensics, among others [34–36]. We consider that our study represents the start of a contribution to the vast field of microbiome research, although we need further refinements of the methodology used in order to validate the obtained models and improve performances.

The accumulating research of the gut microbiome and its influence on health and disease has accelerated the need for integration of multidisciplinary fields in its analysis. In general, health professionals (medical doctors, nurses, biochemists) are not prepared to work in all the steps along the data analysis process (cleaning, filtering, choice of algorithms, interpretation, etc.). Therefore, data science and machine learning can contribute to the translation of innovative results into valuable knowledge that provide decision support in microbiome-based precision medicine.

5 Conclusions

We used two robust computer science-based methods, such as self-organizing maps and random forest, to study the relationships between gut microbiome data and host information. Our results represent a preliminary approach that could be considered, in future studies, as a potential complement in health reports so as to help health professionals to individualize patient treatment or support decision making. Additionally, this work contributes to the increasingly growing area of gut microbiome interactions on human health and disease. However, further studies using other machine learning algorithms to validate the results obtained here are required.

References

1. Dominguez-Bello, M.G., et al.: Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proc. Natl. Acad. Sci. **107**(26), 11971–11975 (2010)
2. Koenig, J.E., et al.: Succession of microbial consortia in the developing infant gut microbiome. Proc. Natl. Acad. Sci. **108**(Supplement 1), 4578–4585 (2011)
3. Rowland, I., et al.: Gut microbiota functions: metabolism of nutrients and other food components. Eur. J. Nutr. **57**(1), 1–24 (2017). https://doi.org/10.1007/s00394-017-1445-8
4. Kiewiet, M., et al.: Flexibility of gut microbiota in ageing individuals during dietary fiber long-chain inulin intake. Molecular Nutrition & Food Research, p. 2000390 (2020)
5. Ruggles, K.V., et al.: Changes in the gut microbiota of urban subjects during an immersion in the traditional diet and lifestyle of a rainforest village. Msphere, vol. 3(4) (2018)
6. Liu, H., et al.: Resilience of human gut microbial communities for the long stay with multiple dietary shifts. Gut **68**(12), 2254–2255 (2019)

7. Floch, M.H., Ringel, Y., Walker, W.A.: The microbiota in gastrointestinal pathophysiology: implications for human health, prebiotics, probiotics, and dysbiosis. Academic Press (2016)
8. Lahti, L., Salojärvi, J., Salonen, A., Scheffer, M., De Vos, W.M.: Tipping elements in the human intestinal ecosystem. Nat. Commun. **5**(1), 1–10 (2014)
9. Kowarik, A., Templ, M.: Imputation with the R package VIM. J. Stat. Softw. **74**(7), 1–16 (2016). https://doi.org/10.18637/jss.v074.i07
10. Wickham, H., et al.: Welcome to the tidyverse. J. Open Source Softw. **4**(43), 1686 (2019). https://doi.org/10.21105/joss.01686, http://dx.doi.org/10.21105/joss.01686
11. LeDell, E., et al.: h2o: R interface for 'h2o'. R package version 3(0.2) (2018)
12. Haro, C., et al.: Intestinal microbiota is influenced by gender and body mass index. PLoS ONE **11**(5), e0154090 (2016)
13. Yu, J., et al.: Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut **66**(1), 70–78 (2017)
14. De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J.B., Massart, S., Collini, S., Pieraccini, G., Lionetti, P.: Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc. Natl. Acad. Sci. **107**(33), 14691–14696 (2010)
15. Wu, G.D., et al.: Linking long-term dietary patterns with gut microbial enterotypes. Science **334**(6052), 105–108 (2011)
16. Koliada, A., et al.: Association between body mass index and firmicutes/bacteroidetes ratio in an adult ukrainian population. BMC Microbiol. **17**(1), 120 (2017)
17. Crovesy, L., Masterson, D., Rosado, E.L.: Profile of the gut microbiota of adults with obesity: a systematic review. Eur. J. Clin. Nutr. **74**(9), 1251–1262 (2020)
18. Reese, A.T., Dunn, R.R.: Drivers of microbiome biodiversity: a review of general rules, feces, and ignorance. MBio **9**(4), e01294-18 (2018)
19. Arumugam, M., et al.: Enterotypes of the human gut microbiome. Nature **473**(7346), 174–180 (2011)
20. Qian, J., et al.: Introducing self-organized maps (som) as a visualization tool for materials research and education. Results Mater. **4**, 100020 (2019)
21. Xu, J., et al.: Alteration of the abundance of parvimonas micra in the gut along the adenoma-carcinoma sequence. Oncol. Lett. **20**(4), 1 (2020)
22. Nagarajan, M., Prabhu, V.R., Kamalakkannan, R.: Metagenomics: implications in oral health and disease. In: Metagenomics, pp. 179–195. Elsevier (2018)
23. Goodson, J., Groppo, D., Halem, S., Carpino, E.: Is obesity an oral bacterial disease? J. Dent. Res. **88**(6), 519–523 (2009)
24. Selle, K., Klaenhammer, T.R.: Genomic and phenotypic evidence for probiotic influences of lactobacillus gasseri on human health. FEMS Microbiol. Rev. **37**(6), 915–935 (2013)
25. Mahboubi, M.: Lactobacillus gasseri as a functional food and its role in obesity. Int. J. Med. Rev. **6**(2), 59–64 (2019)
26. Turnbaugh, P.J., et al.: A core gut microbiome in obese and lean twins. Nature **457**(7228), 480–484 (2009)
27. Yatsunenko, T., et al.: Human gut microbiome viewed across age and geography. Nature **486**(7402), 222–227 (2012)
28. Dominianni, C., et al.: Sex, body mass index, and dietary fiber intake influence the human gut microbiome. PLoS ONE **10**(4), e0124599 (2015)
29. Bosco, N., Noti, M.: The aging gut microbiome and its impact on host immunity. Genes & Immunity, pp. 1–15 (2021)

30. Chung, W.S.F., et al.: Relative abundance of the prevotella genus within the human gut microbiota of elderly volunteers determines the inter-individual responses to dietary supplementation with wheat bran arabinoxylan-oligosaccharides. BMC Microbiol. **20**(1), 1–14 (2020)
31. Tan, H., et al.: Pilot safety evaluation of a novel strain of bacteroides ovatus. Front. Genet. **9**, 539 (2018)
32. Salonen, A., et al.: Impact of diet and individual variation on intestinal microbiota composition and fermentation products in obese men. ISME J. **8**(11), 2218–2230 (2014)
33. Reijnders, D., et al.: Effects of gut microbiota manipulation by antibiotics on host metabolism in obese humans: a randomized double-blind placebo-controlled trial. Cell Metab. **24**(1), 63–74 (2016)
34. Ai, D., Pan, H., Han, R., Li, X., Liu, G., Xia, L.C.: Using decision tree aggregation with random forest model to identify gut microbes associated with colorectal cancer. Genes **10**(2), 112 (2019)
35. Thompson, J., Johansen, R., Dunbar, J., Munsky, B.: Machine learning to predict microbial community functions: an analysis of dissolved organic carbon from litter decomposition. PLoS ONE **14**(7), e0215502 (2019)
36. Topçuoğlu, B.D., Lesniak, N.A., Ruffin, M.T., IV., Wiens, J., Schloss, P.D.: A framework for effective application of machine learning to microbiome-based classification problems. MBio **11**(3), e00434-20 (2020)

Apéndice B

Bacterias

Se indican las bacterias repesentantes de cada *phylum* presentes en el conjunto de datos: *Bacteroidetes* (Cuadro B.1 y *Firmicutes* (Cuadro B.2).

Especie / Género		
Allistipes sp.	*Bacteroides fragilis*	*Bacteroides ovatus*
Bacteroides plebeius	*Bacteroides stercoris*	*Bacteroides splachnicus*
Bacteroides vulgatus	*Bacteroides uniformis*	*Parabacteroides distasonis*
Prevotella melaninogenica	*Prevotella oralis*	*Prevotella tannerae*
Tannerella sp.		

Cuadro B.1: Miembros del subconjunto *Bacteroidetes* presentes en los datos.

Especie / Género		
Anaerostipes caccae	*Anaerotruncus colihominis*	*Anaerovorax odorimutans*
Bryantella formatexigens	*Butyrivibrio crossotus*	*Clostridium leptum*
Clostridium sensu stricto	*Clostridium stercorarium*	*Clostridium cellulosi*
Clostridium nexile	*Clostridium orbiscindens*	*Clostridium symbiosum*
Clostridium difficile	*Clostridium colinum*	*Clostridium sphenoides*
Coprococcus eutactus	*Dialister*	*Dorea formicigenerans*
Eubacterium biforme	*Eubacterium rectale*	*Eubacterium hallii*
Eubacterium ventriosum	*Faecalibacterium prausnitzii*	*Lachnobacillus bovis*
Lachnospira pectinoschiza	*Lactobacillus plantarum*	*Oscillospira guillermondii*
O. Clostridium XIVa	*Papillibacter cinnamivorans*	*Roseburia intestinalis*
Ruminococcus lactaris	*Ruminococcus gnavus*	*Ruminococcus bromii*
Ruminococcus obeum	*Ruminococcus callidus*	*Sporobacter termitidis*
Staphylococcus	*Streptococcus mitis*	*Streptococcus bovis*
Subdoligranulum variable	*UCI I*	*UCI II*

Cuadro B.2: Miembros del subconjunto *Firmicutes* presentes en los datos.

Apéndice C

PCA - *Firmicutes*

PC	Autovalor	Porcentaje de varianza	% Varianza acumulada
1	2,87	6,85	6,85
2	1,87	4,46	11,31
3	1,45	3,47	14,79
4	1,33	3,18	17,97
5	1,25	2,99	20,96
...	...	...	...
10	1,12	2,68	34,98
11	1,09	2,61	37,60
12	1,08	2,57	40,18
13	1,05	2,51	42,70
14	1,04	2,48	45,19
15	1,02	2,45	47,64
16	1,00	2,39	50,03
17	0,99	2,37	52,41
18	0,97	2,32	54,73
19	0,95	2,27	57,00
20	0,93	2,21	59,22
...	...	...	...
25	0,86	2,05	69,86
26	0,85	2,04	71,91
27	0,84	2,02	73,93
28	0,83	1,99	75,92
29	0,81	1,93	77,86
30	0,80	1,92	79,78
31	0,80	1,90	81,69
...	...	...	...
38	0,69	1,64	93,93
39	0,66	1,57	95,51
40	0,65	1,56	97,08
41	0,63	1,52	98,60
42	0,58	1,39	100,00

Cuadro C.1: Varianza total explicada por cada componente principal del subconjunto *Firmicutes*.

Apéndice D

SOM - *Bacteroidetes*

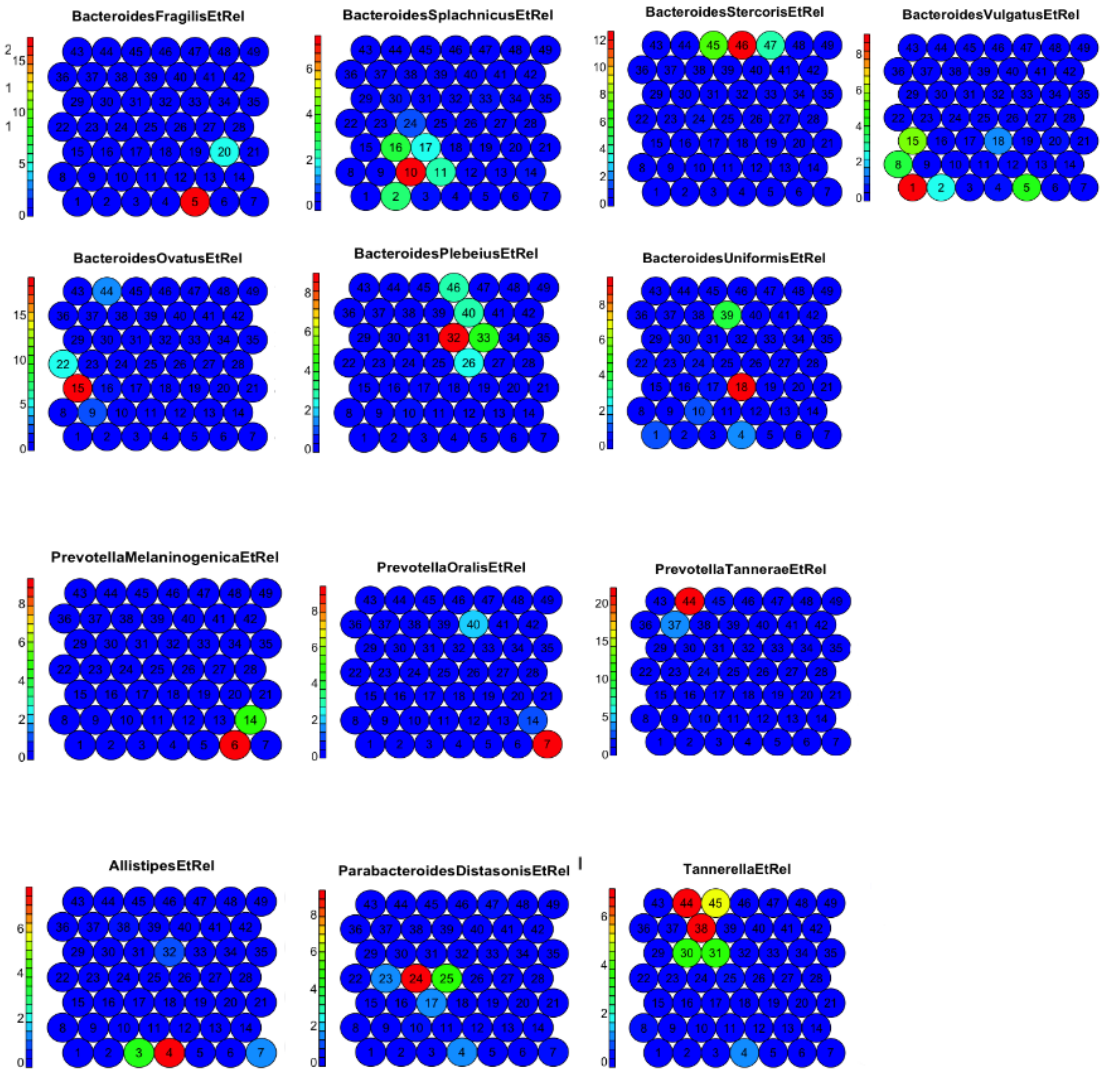

Figura D.1: Mapas de calor luego del entrenamiento SOM de las abundancias de bacterias pertenecientes a *Bacteroidetes*.

Apéndice E

Combinación de *clustering* jerárquico y SOM - *Bacteroidetes*

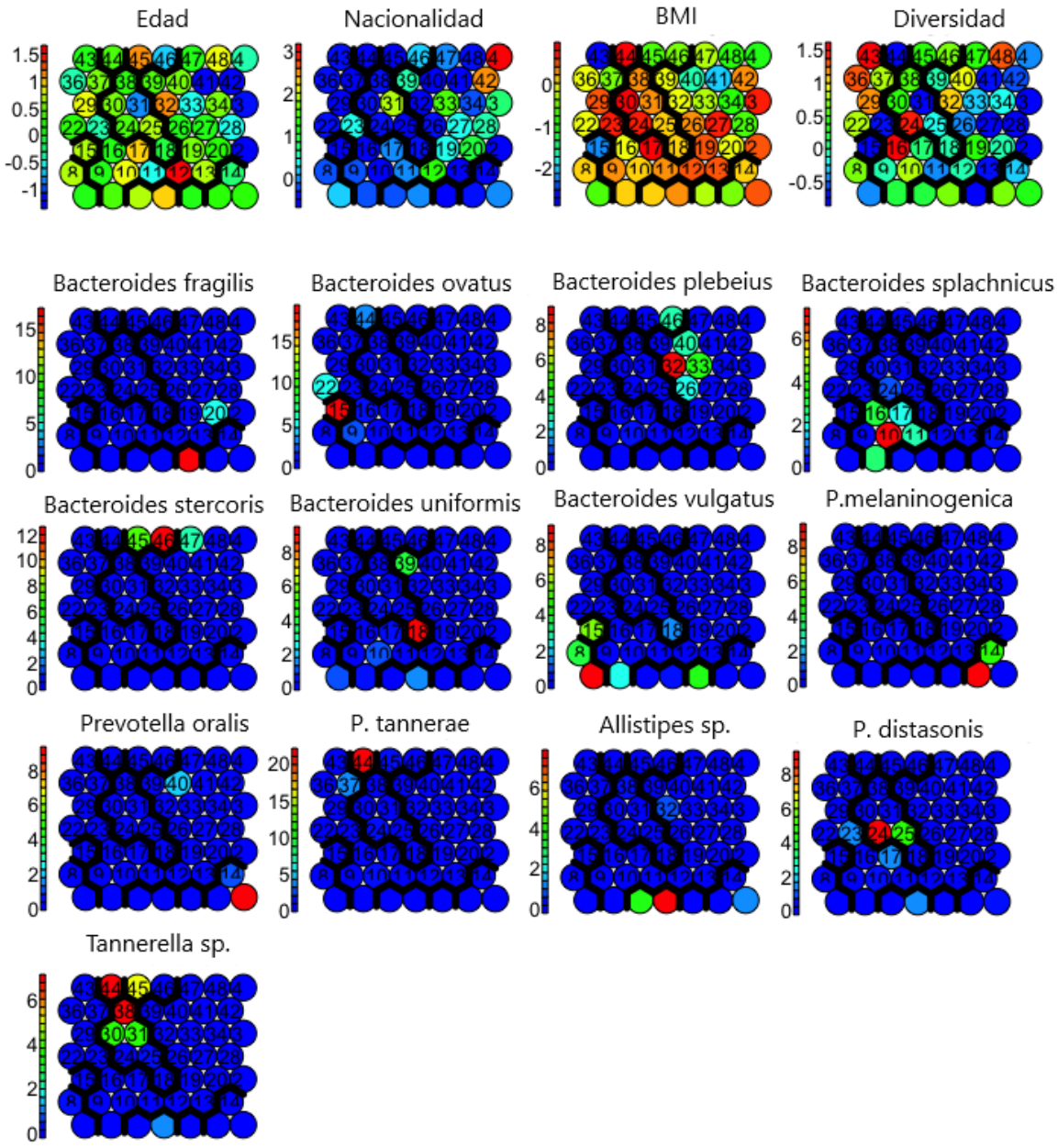

Figura E.1: Mapas de SOM luego de la elección del número de *clusters* en *Bacteroidetes*.

Apéndice F

SOM - *Firmicutes*

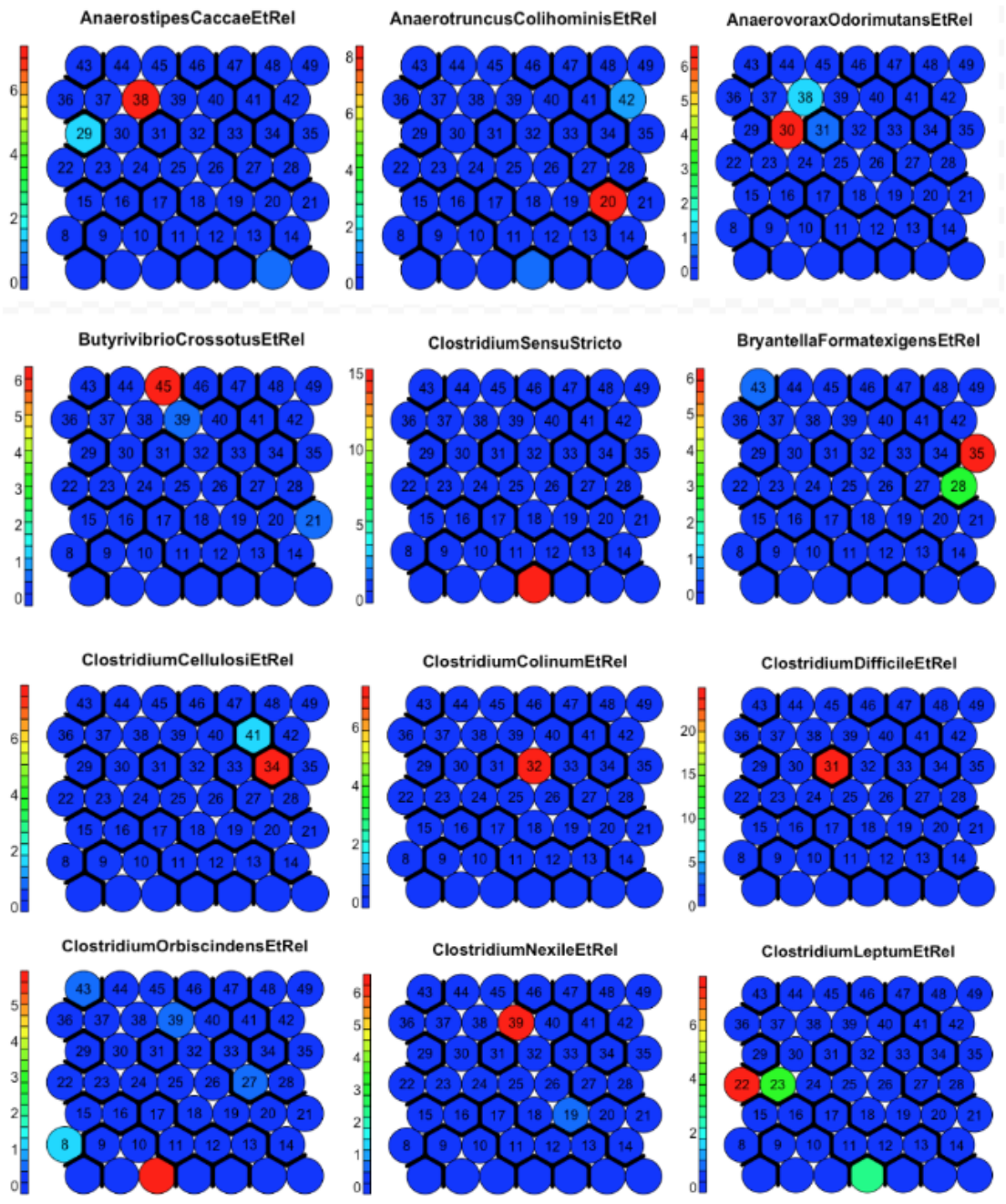

Figura F.1: Mapas de calor luego del entrenamiento SOM de las abundancias de bacterias pertenecientes a *Firmicutes*.

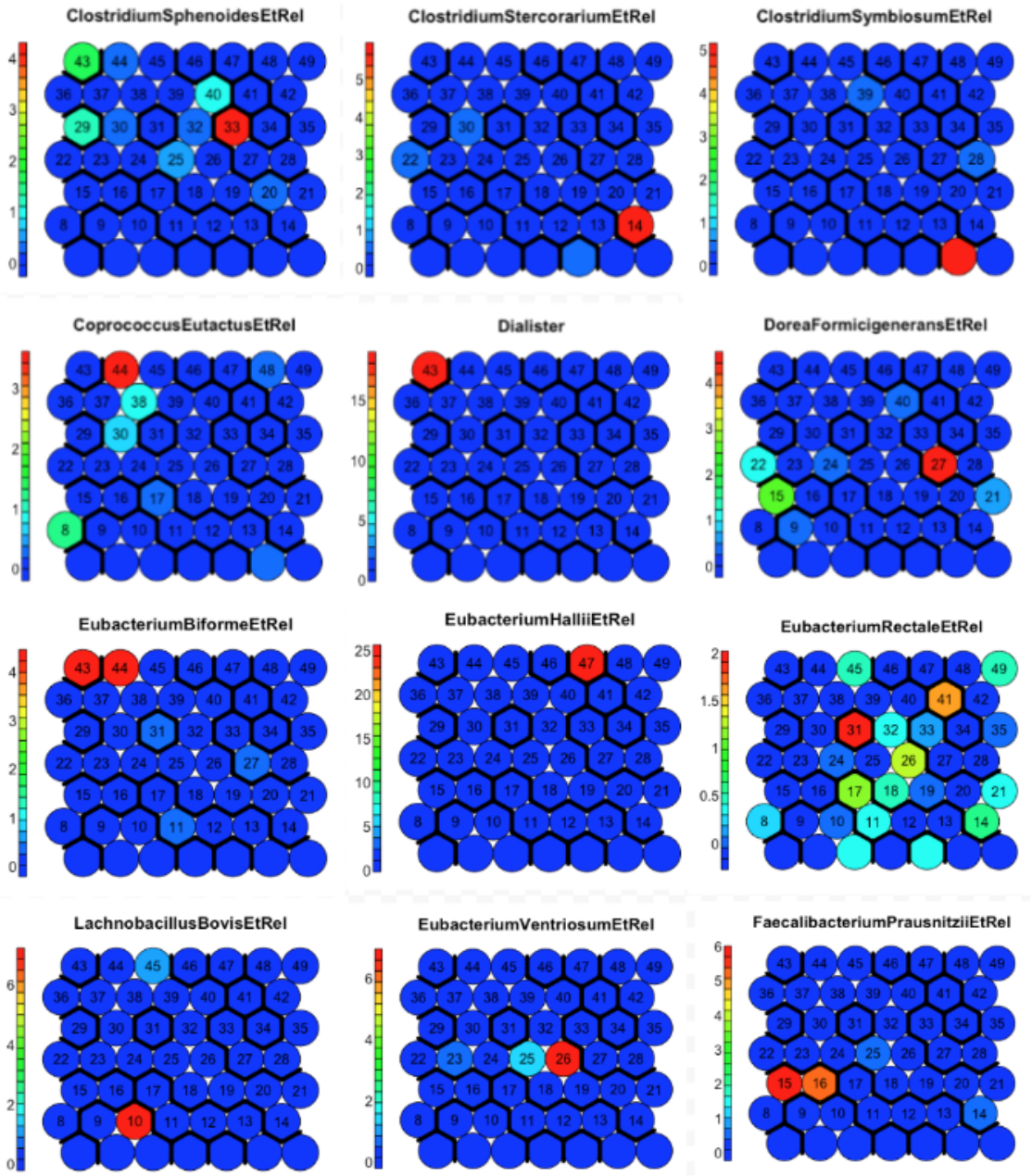

Figura F.2: Mapas de calor luego del entrenamiento SOM de las abundancias de bacterias pertenecientes a *Firmicutes*.

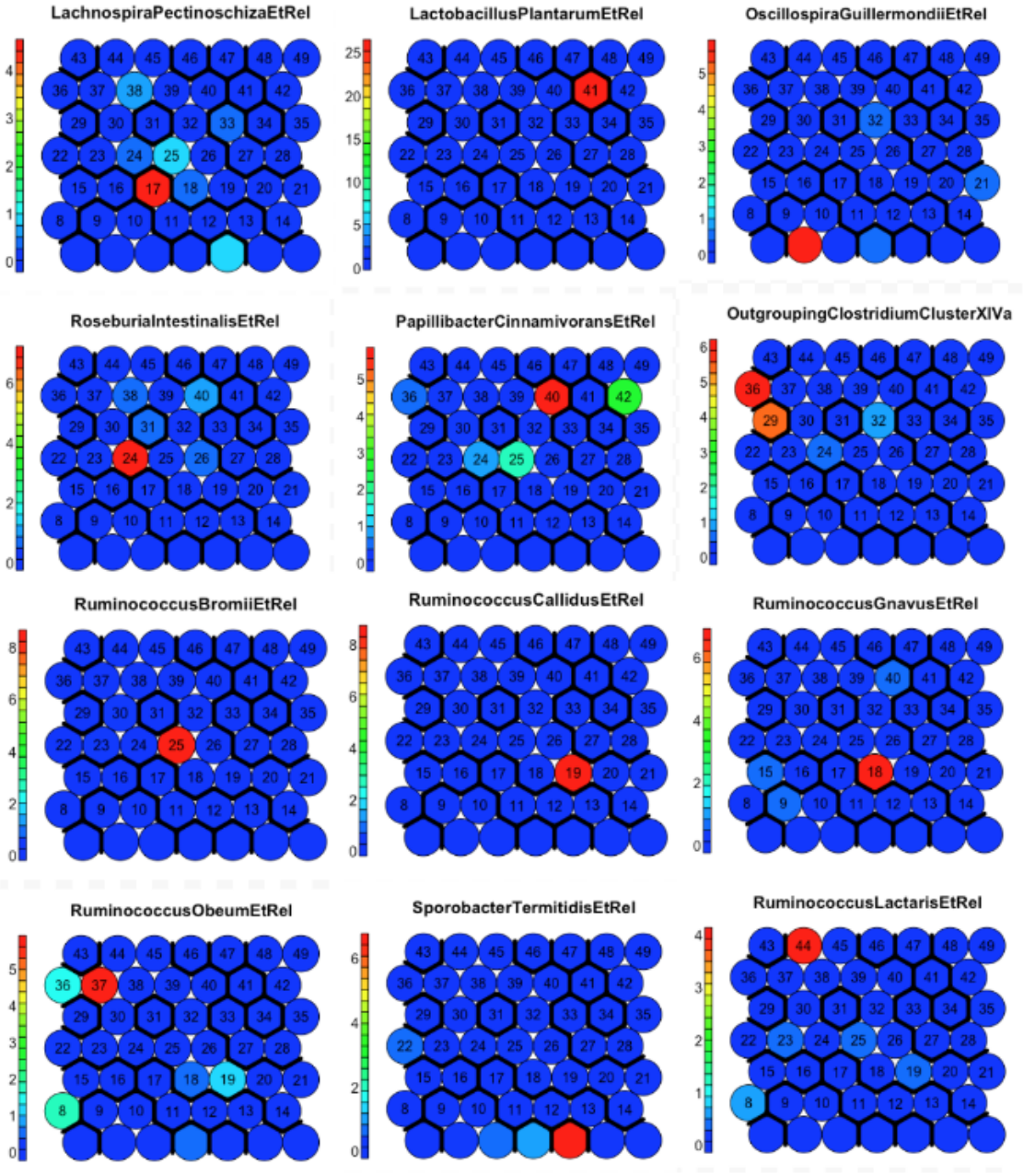

Figura F.3: Mapas de calor luego del entrenamiento SOM de las abundancias de bacterias pertenecientes a *Firmicutes*.

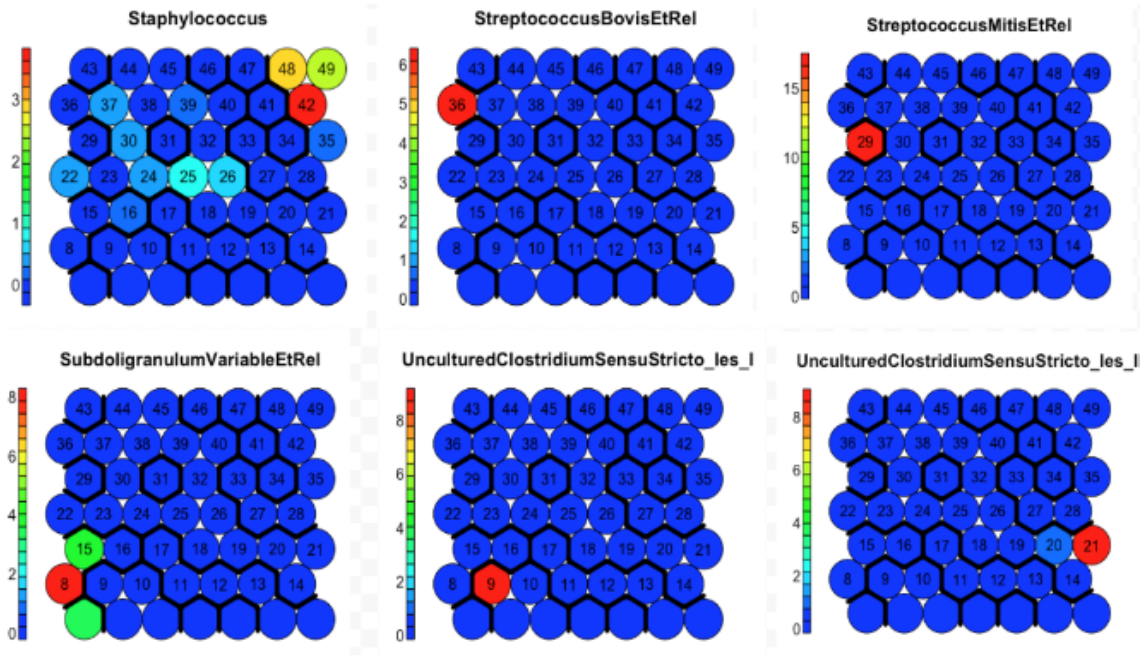

Figura F.4: Mapas de calor luego del entrenamiento SOM de las abundancias de bacterias pertenecientes a *Firmicutes*.

Apéndice G

Random forest - Firmicutes.

Cuadro G.1: Listado de importancia de variables sobre la diversidad luego de la implementación de RF.

Variable	Importancia relativa	Importancia escalada
BMI	99,44	1,00
Nacionalidad	81,66	0,82
Edad	69,81	0,70
Lachnobacillus bovis	60,99	0,61
Anaerovorax odorimutans	54,89	0,55
Ruminococcus obeum	53,56	0,53
Sporobacter termitidis	47,89	0,48
Eubacterium biforme	45,28	0,45
Clostridium leptum	33,35	0,33
Bryantella formatexigens	32,80	0,33
Anaerotruncus colihominis	26,16	0,26
Butyrivibrio crossotus	26,07	0,26
Oscillospira guillermondii	26,06	0,26
Clostridium stercorarium	23,72	0,23
Clostridium cellulosi	20,98	0,21
RuminococcusLactaris	20,36	0,20

(Sigue)

Variable	Importancia relativa	Importancia escalada
UCI I	19,80	0,19
Dorea formicigenerans	19,05	0,19
Eubacterium hallii	18,86	0,18
Clostridium nexile	17,55	0,17
Anaerostipes caccae	17,41	0,17
Clostridium orbiscindens	16,15	0,16
Roseburia intestinalis	16,15	0,16
Papillibacter cinnamivorans	16,10	0,16
Lachnospira pectinoschiza	14,47	0,14
Eubacterium rectale	13,96	0,14
Ruminococcus gnavus	13,88	0,13
Clostridium symbiosum	12,49	0,12
Subdoligranulum variable	12,05	0,12
Clostridium difficile	12,02	0,12
UCI II	11,08	0,11
Faecalibacterium prausnitzii	10,74	0,10
Outgrouping Clostridium Cluster XIVa	10,66	0,10
Ruminococcus bromii	10,60	0,10
Coprococcus eutactus	10,33	0,10
Clostridium sensu stricto	10,27	0,10
Sexo	10,10	0,10
Ruminococcus callidus	9,93	0,09
Streptococcus mitis	9,45	0,09
Streptococcus bovis	9,11	0,09
Staphylococcus	7,77	0,07
Dialister	7,53	0,07
Eubacterium ventriosum	7,18	0,07

(Sigue)

Variable	Importancia relativa	Importancia escalada
Clostridium colinum	7,07	0,07
Clostridium sphenoides	7,05	0,07
Lactobacillus plantarum	6,95	0,06