Escola Tècnica Superior d'Enginyeria ETSE-UV

MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA

VNIVERSITAT DE VALÈNCIA

TRABAJO DE FIN DE MÁSTER

EVALUACIÓN COMPARATIVA DE ESTRATEGIAS DE ANÁLISIS DE DATOS EN LOS ESTUDIOS DE MICROBIOMA

AUTORA:

MARÍA MULET FERNÁNDEZ

TUTORES:

FRANCISCO GARCÍA GARCÍA

MARIAM DE LA IGLESIA VAYÁ

MARÍA PASCUAL MORA

MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA

TRABAJO DE FIN DE MÁSTER

EVALUACIÓN COMPARATIVA DE ESTRATEGIAS DE ANÁLISIS DE DATOS EN LOS ESTUDIOS DE MICROBIOMA

AUTORA:

MARÍA MULET FERNÁNDEZ

TUTORES:

FRANCISCO GARCÍA GARCÍA

MARIAM DE LA IGLESIA VAYÁ

MARÍA PASCUAL MORA

TRIBUNAL:

PRESIDENTE/A:	VOCAL 1:
VOCAL 2:	**FECHA DE DEFENSA:** **CALIFICACIÓN:**

Resumen

El microbioma está adquiriendo una creciente relevancia en la salud y en el desarrollo humano. Por ello cada día aparecen más estudios metagenómicos en los que se busca la relación del microbioma con la salud. El método más común de análisis es el de tipo *gene-marker* en el que se analiza el gen 16S ARNr de los microorganismos. Para analizar este gen hay una gran cantidad de herramientas bioinformáticas, sin embargo, es necesaria una evaluación de los recursos y estrategias disponibles, que permita conocer los procedimientos adecuados para cada grupo de datos.

Por ello, el objetivo de este trabajo consiste en la revisión y comparación de una selección de tres herramientas de análisis que aplicaremos sobre dos conjuntos de datos diferentes. A continuación se evaluarán los resultados obtenidos en cada una de ellas, para finalmente decidir cuál es la mejor estrategia que cubre los objetivos de los estudios.

Palabras clave: Microbioma, Metataxonomía, Bioinformática, Pipeline, 16S ARNr.

Abstract

The microbiome is having more relevance in human health and development. Therefore metagenomic studies where the relation between microbiome and health is studied are increasingly appearing. The most common method of analysis is the *gene-marker* method. In this kind of method is analysed the gene 16S rRNA of microorganisms. There are many bioinformatic tools to analyse this gene. However, a detailed assessment of available strategies and resources is needed to have a better knowledge about appropriate procedures for each group of data.

Therefore, the main objective of this work is to review and compare three of these tools that we are going to use on two different conjunto de datos. Then, the obtained results will be evaluated for at last decide which is the one that fits better to our study goals.

Keywords: Microbiome, Metataxonomy, Bioinformatics, Pipeline, 16S rRNA.

Agradecimientos

La realización de este proyecto ha sido posible gracias a toda la ayuda recibida por la gente que me rodea.

Gracias a mis compañeros del Centro de Investigación Príncipe Felipe, en especial a Raúl Pérez por contestar a cada una de mis miles de dudas y sacar siempre un hueco para ayudarme. A Paco García, por brindarme esta oportunidad, ayudarme y, sobretodo, por darme ánimos cada semana para seguir adelante con el proyecto a pesar de las dificultades con las que nos hemos visto envueltos este año.

A mis compañeros del máster, en especial a Miguel por estar en las buenas y en las malas, pero sobretodo por saber sacar siempre una sonrisa gracias a sus ingeniosas canciones y a sus bromas.

A mis amigos, por aguantar todas mis quejas, apoyarme cuando no podía más y conseguir hacerme ver que los sacrificios, al final, se ven recompensados.

Por último a mi familia, por darme las oportunidades que me han dado y me han llevado a donde estoy ahora.

Índice

Índice de Figuras

Índice de Tablas

Glosario de acrónimos

ADN	Ácido desoxirribonucleico
ARN	Ácido Ribonucleico
ASV	Variantes de Secuencia de Amplicones
CNG	China National Genebank: Microbiome Database
eHOMD	Human Oral Microbiome Database
ENA	European Nucleotide Archive
FS	Figura Suplementaria
GABA	Ácido Gamma-Aminobutírico
GEO	Gene Expression Omnibus
HMPDP	Microbiome Project Data Portal
MAMPs	Patrones Moleculares Asociados a Microbios
MHC	Complejo Mayor de Histocompatibilidad
MS	Esclerosis Múltiple
NGS	Next Generation Sequencing
NMDS	Escalamiento Multidimensional No-Métrico
NMR	Resonancia Magnética Nuclear
NRC	National Research Council
OTU	Unidades Taxonómicas Operacionales
PCR	Reacción en Cadena de la Polimerasa
RDP	Ribosomal Database Project
SNC	Sistema Nervioso Central
TS	Tabla Suplementaria.
WGS	Whole Genome Shotgun

1. Introducción

El estudio de las comunidades microbianas ha cambiado desde el primer descubrimiento de microbios por Leeuwenhoek en 1676, hasta su caracterización utilizando las técnicas moleculares actuales (Escobar-Zepeda, 2015).

A finales de los años 70, Carl Woese propuso el uso de los genes ribosómicos del ácido ribonucleico (ARN) como marcadores moleculares para la clasificación taxonómica (Woese, 1977). Esta idea junto con el método ya conocido de secuenciación automatizada de Sanger (Sanger, 1977) revolucionó el estudio y la clasificación de los microorganismos. Algunas décadas más tarde, los avances en las técnicas moleculares fueron utilizados para realizar la descripción de la diversidad microbiana (Escobar-Zepeda, 2015).

Pese a ser objeto de estudio desde hace décadas, no ha sido hasta hace unos años que se ha descubierto la importancia de la microbiota que coloniza el cuerpo humano en la salud. Esto ha sido gracias al uso de diferentes técnicas de shotgun que han proporcionado evidencias del estado de la microbiota y su relación con algunas enfermedades (Korem, 2015).

1.1. ¿Qué es el microbioma?

Como ya se ha mencionado, hay una gran cantidad de microorganismos que habitan el cuerpo humano y para ello forman comunidades que incluyen eucariotas, bacterias, arqueas, protozoos y virus, a este conjunto se le llama microbiota. Estos microorganismos pueden diferenciarse en comensales, mutualistas y patógenos (del Campo-Moreno, 2018), y se encuentran en diversos sitios del cuerpo humano (piel, boca, vías respiratorias, vagina, tracto gastrointestinal) con diferentes condiciones fisiológicas (Otero, 2016).

De hecho, teniendo en cuenta la composición celular, la diversidad genética y la capacidad metabólica, el huésped que habitan puede considerarse como un organismo híbrido compuesto por el propio huésped y las células microbianas que lo habitan, trabajando ambos en un equilibrio simbiótico (Barko, 2018; Shreiner, 2015; Turnbaugh, 2007).

Se ha estimado que el número medio de células bacterianas que se encuentran en el cuerpo humano es diez veces mayor al número de células humanas. Aunque, debido a su pequeño tamaño, la microbiota humana supone solamente un 1-3% de la masa corporal (Otero, 2016).

Debido a esta gran cantidad de microorganismos que viven en nuestro cuerpo, y gracias a su estudio por diversas técnicas moleculares, se empezó a

tener en cuenta su relevancia en la salud humana y se creó el concepto de microbioma. Este hace referencia a todo el hábitat, es decir, al conjunto de genomas, tanto microbianos como humano, y a las condiciones ambientales en las que se encuentran (del Campo-Moreno, 2018; Dietert, 2015).

El microbioma es dinámico y se ve sometido a importantes cambios durante la vida del huésped debido a una gran variedad de factores que incluyen la dieta, el entorno y las enfermedades (Barko, 2018).

1.2. ¿Qué impacto tiene su estudio en biomedicina?

El concepto de híbrido o superorganismo por el que se considera al conjunto humano-microbioma nos lleva a redefinir qué constituye un organismo humano y a ampliar nuestra forma de estudiar las enfermedades y la salud en los humanos (Dietert, 2015).

Todo esto se ve apoyado por las pruebas que hay del elevado riesgo de enfermedades específicas relacionadas con la disbiosis microbiana y por la, cada vez mayor, evidencia de que los componentes del microbioma ejercen una gran influencia sobre el comportamiento, el metabolismo o la inmunidad (Chung, 2012; Borre, 2014).

Por ello, el microbioma debe empezar a considerarse como un foco de protección de la salud humana (Dietert, 2015).

Estas pruebas sobre su importancia desafían el modelo de entorno propuesto por el *National Research Council* (NRC, 1987), en el que se distingue entre eventos externos e internos en el organismo. Pero su importancia no sólo reside en su papel en el funcionamiento del organismo huésped, además se ha visto que presenta particularidades y características propias de cada individuo, pudiendo variar en función de la dieta, la base genética y la interacción con el medio ambiente (del Campo-Moreno, 2018).

Una de las principales evidencias del papel de la microbiota en la salud es su influencia en el desarrollo del sistema inmune. Cuando nacemos, las células del sistema inmune carecen de estímulos y por tanto reconocen a todos los antígenos de su entorno como parte del organismo, por lo que no generan ningún tipo de respuesta inflamatoria frente a ellos. De esta forma, el contacto que tienen las células inmunológicas sin diferenciar con los microorganismos de la placenta y de la pared vaginal de la madre es muy importante ya que va a ayudar a distinguir lo propio de lo externo (Alarcón Cavero, 2016).

El órgano con mayor proporción de microorganismos es el intestino ya que en él ejercen funciones muy relevantes tales como ayudar a digerir los

alimentos, producir algunas vitaminas que el cuerpo humano no puede sintetizar, previenen la colonización por patógenos y, como ya se ha mencionado, estimulan el sistema inmune (Figura 1). Pero esto no es todo, hay estudios que han descrito la existencia del eje cerebro-intestino, que conecta el sistema nervioso central con la microbiota intestinal a través del nervio vago, el sistema parasimpático, los metabolitos bacterianos y el sistema endocrino asociado al tracto digestivo (Foster, 2013; Fung, 2017).

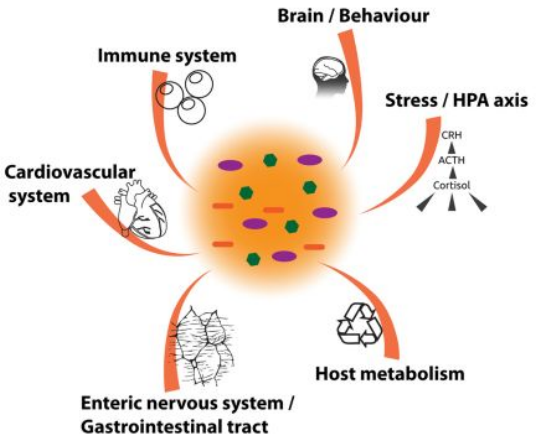

Figura 1. Implicaciones del microbioma del intestino en la salud (Clarke, 2014).

La vía de comunicación entre el sistema nervioso y el intestino permite al cerebro controlar las funciones gastrointestinales y algunas funciones inmunes, entre otras actividades, pero también la influencia de la microbiota y sus metabolitos en las funciones cerebrales (Fig. 2). De hecho, algunos ejemplos de producción de estos metabolitos serían el de los lactobacilos y las bifidobacterias que sintetizan ácido gamma-aminobutírico (GABA), uno de los principales neurotransmisores cerebrales, o el de E. coli, Saccharomyces y Bacillus que producen norepinefrina (Alarcón Cavero, 2016).

Por todo esto, la microbiota ya no sólo jugaría un papel relevante en enfermedades como la obesidad o la diabetes, sino que puede tener relación con enfermedades del Sistema Nervioso Central (SNC) como el autismo, la depresión o la ansiedad.

1.3. La microbiota y el sistema nervioso

Además del papel que desempeña la microbiota con la generación de metabolitos sobre el sistema nervioso, existe un punto de conexión en su

influencia tanto en el sistema nervioso como en el inmune y este punto lo constituyen las células inmunes más abundantes del cerebro, la microglía. Aunque los mecanismos por los que la microbiota influye en la maduración y la función de la microglía son desconocidos, sí que parece haber una modulación ya que hay estudios en los que se ha comprobado que ratones Germ-Free adultos presentan déficits funcionales (mala respuesta a virus y activación inmune atenuada) (Matcovitch-Natan, 2016).

Otras células neuronales que se ven afectadas por la microbiota son los astrocitos. Los astrocitos son las células gliales más abundantes del cerebro y ejercen diversas funciones (regulación de la integridad de la barrera hematoencefálica, el balance de gradiente de iones, transporte de nutrientes, etc.) (Matcovitch-Natan, 2016).

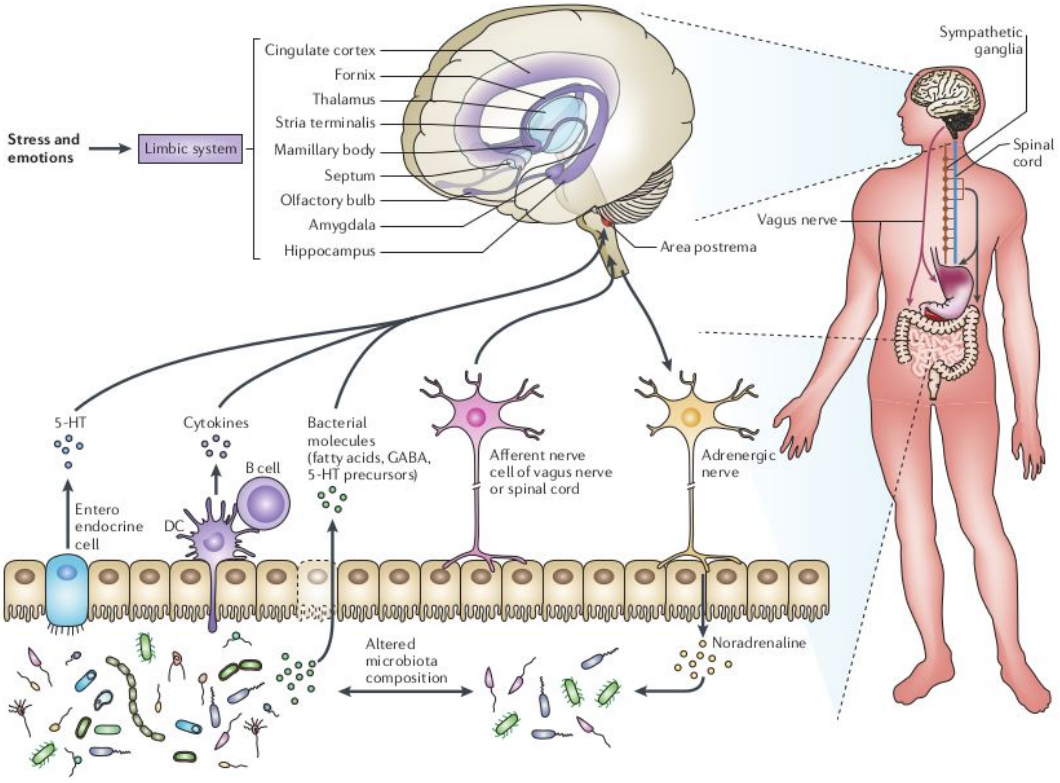

Figura 2. Eje bidireccional microbiota-intestino-cerebro (Collins, 2012).

Además, los astrocitos integran información de las células adyacentes (glía, neuronas, células vasculares e inmunes) para regular la excitabilidad neuronal y la formación sináptica (Rossi, 2015; Barres, 2008).

A pesar de que los astrocitos no forman parte de las células inmunes residentes del SNC, sí que ejercen funciones relacionadas con la inmunidad (expresando receptores que detectan los *Patrones Moleculares Asociados a Microbios* (MAMPs) y modulando la respuesta neuroinflamatoria a través de la producción de citocinas y de la presentación de antígenos mediante el Complejo mayor de Histocompatibilidad (MHC II) (Dong, 2001). La actividad de

los astrocitos se ve influida por la microbiota intestinal a través de los metabolitos que generan y que activan los receptores de hidrocarburos de arilos (Rothhammer, 2016).

Esta información tiene especial interés en este proyecto ya que forma parte del proyecto “Share Your Brain”, en el que, entre otros estudios, va a analizarse la microbiota de una gran parte de la población para establecer las diferencias entre mujeres y hombres y así poder usar estos datos para el diagnóstico de posibles enfermedades mentales.

1.4. Tecnologías de alto rendimiento y Next Generation Sequencing (NGS).

En las últimas décadas ha habido grandes avances en todas las áreas de conocimiento de la biología gracias al desarrollo de tecnologías de alto rendimiento para el estudio de las diferentes ómicas.

Una ómica, es un campo interdisciplinario dentro de la biología molecular en el que el estudio se centra en el conjunto de algo (genes, metabolitos, proteínas, etc.). Dentro de las diferentes ómicas, la que más se ha estudiado es la genómica en la que se estudia la estructura, la función, el desarrollo y el mapeo de los genomas. El genoma es el conjunto de información relacionada con el ácido desoxirribonucleico (ADN), un polímero de nucleótidos que contiene las instrucciones genéticas para que un organismo funcione (Watson, 1953). Conforme se ha ido avanzando en el estudio de la genómica, se ha comenzado a estudiar otras ómicas hasta llegar, entre otras, al estudio de los genomas combinados de los conjuntos de poblaciones de organismos o metagenómica. El conjunto de organismos que se estudian pueden ser tanto bacterianos, víricos, como eucariotas (Ghurye, 2016). Y es gracias a la aparición de esta ómica que se ha empezado a estudiar el microbioma (Gevers, 2012; Li, 2015).

Para estudiar estas áreas se han ido desarrollando las tecnologías de alto rendimiento que precisamente se encargan de analizar simultáneamente la actividad de miles de unidades biológicas. Concretamente, se ha avanzado en el estudio de nuestro microbioma gracias al desarrollo de las técnicas moleculares de secuenciación masiva de segunda generación o NGS. Estas técnicas destacan principalmente porque con ellas se puede secuenciar directamente el ADN sin necesidad de clonarlo, por su menor coste, por la reducción de tiempo de secuenciación y por la generación de una gran cantidad de secuencias de alta calidad.

Hay dos métodos principales para la secuenciación de alto rendimiento de microbiomas: los estudios *gene-marker* (Fig. 3) y los *Whole Genome Shotgun* (WGS). En los primeros se generan primers para amplificar mediante PCR un gen concreto (16S ARNr en este caso) de todos los genomas que aparecen en la muestra y esto después es lo que se secuencia. Las secuencias generadas se agrupan entonces en *Unidades Taxonómicas Operacionales* (OTUs) o en *Variantes de Secuencia de Amplicones* (ASVs) y estas son comparadas entre las diferentes muestras (Roumpeka, 2017). El segundo método, el WGS, aplicado a la metagenómica, fue acuñado por primera vez por Handelsman et al. (1998) como el análisis funcional de una colección de ADN microbiano extraído de muestras de suelo. Este método se basa en la secuenciación de los genomas de todos los organismos en lugar de un único gen de interés, proporcionando información sobre la estructura, la función y la organización de los genomas de la comunidad microbiana (Roumpeka, 2017).

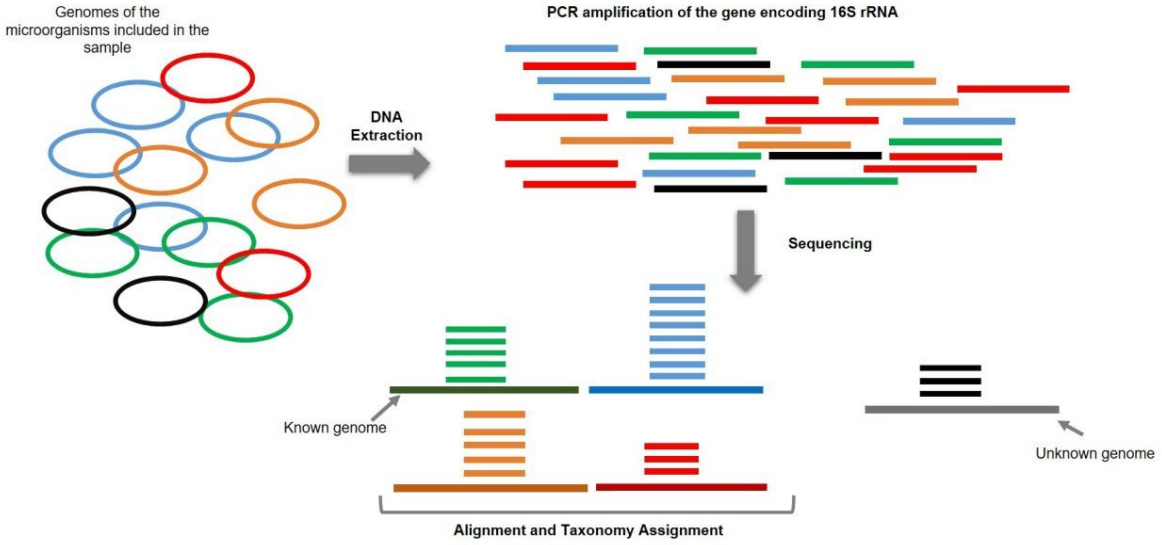

Figura 3. Esquema de la amplificación del gen 16S.
(https://www.france-genomique.org/technological-expertises/metagenomics/shotgun-metagenomics/?lang=en)

Las tecnologías NGS, incluyen tecnologías como Illumina, Ion torrent o Roche 454 (Mardis, 2008). La tecnología más utilizada es la de Illumina y se basa en la detección de la fluorescencia. En ella, fragmentos cortos de ADN se unen a una pieza de cristal y son amplificados para formar clusters. Después se etiquetan mediante fluorescencia los nucleótidos y se incorporan de forma complementaria a la secuencia de ADN, se detecta la fluorescencia y se revela la secuencia.

La tecnología de Ion Torrent se basa en la liberación de protones. Los fragmentos de ADN son amplificados mediante PCR en emulsión e introducidos en micropocillos. Los nucleótidos van moviéndose por los pocillos y se incorporan a la cadena complementaria liberando un protón (H+) que se mide

por el cambio de voltaje. Esta tecnología genera grandes lecturas en poco tiempo pero no se usa tanto debido a su alta tasa de error de homopolímeros.

Otra de las tecnologías que se usan es la de Roche 454 basada en la pirosecuenciación. En este caso se realiza la detección quimioluminiscente del pirofosfato liberado durante la elongación de la cadena complementaria de ADN. Esta tecnología ha sido la más usada durante mucho tiempo ya que las lecturas que generaba eran muy largas, sin embargo también presenta errores en la lectura de homopolímeros.

1.5. Principales abordajes de análisis bioinformáticos.

Como ya se ha comentado en el apartado anterior, existen diversas ómicas que se han ido desarrollando conforme se ha generado interés en ese campo de estudio. La principal y una de las más estudiadas hasta ahora es la genómica, pero existen otras que cada vez están adquiriendo mayor relevancia. Entre estas, las más utilizadas son la proteómica, la transcriptómica, la metabolómica, la interactómica y la metagenómica.

La proteómica es el estudio de un gran número de proteínas en diferentes sistemas biológicos a la vez. Gracias a esta ómica, se pueden responder una gran cantidad de preguntas básicas sobre las proteínas (abundancia, localización, interacciones,, etc.). Sin embargo, la mayoría de estudios de proteómica se centran en responder preguntas relacionadas con la abundancia proteica y las modificaciones translacionales (McArdle, 2020).

La transcriptómica es el estudio del conjunto de ARN que existe en una célula, tejido u órgano. Este campo de estudio ha ido adquiriendo interés debido a que se ha intentado averiguar cómo los diferentes grupos de genes son activados por diferentes células dando lugar a distintas funciones. La transcriptómica ayuda precisamente a resolver la correlación entre el destino celular y su función a través de los patrones de expresión de genes (Morozova, 2009).

La metabolómica es el estudio de los procesos químicos que implican a los metabolitos. Los metabolitos son moléculas utilizadas o producidas durante el metabolismo. El metabolismo juega un papel central en todas las áreas de biología y cada vez más se están estudiando desde este punto de vista. Estos estudios están siendo posibles gracias a los avances en tecnologías de medida de metabolitos como son la espectrometría de masas o la Resonancia Magnética Nuclear (NMR). Sin embargo, la medida de las concentraciones de

metabolitos sólo nos da una parte de la información. Es por tanto igual de importante entender la actividad de las rutas (Jang, 2018).

La interactómica es el estudio del conjunto de interacciones entre distintas biomoléculas en un entorno determinado (célula, organismo, proceso fisiológico…), pero principalmente se centra en el estudio de las interacciones entre proteínas. Esta ómica trata de ordenar y conectar la información existente sobre genómica y proteómica, para finalmente integrarla en un contexto biológico asignando funciones a las proteínas a través de sus interacciones.

Y finalmente, la metagenómica que, como ya se ha comentado, es el estudio de la diversidad y la disbiosis microbiana de muestras ambientales. Muchos de estos estudios se realizan sobre el microbioma humano y sirven para ver la relación del mismo con la salud. La metagenómica funcional puede identificar genes funcionales nuevos, rutas microbianas, genes de resistencia a antibióticos, etc (Wang, 2015).

2. Objetivos

Debido a la relevancia y el impacto que está adquiriendo el estudio del microbioma durante los últimos años, es necesario conocer las diferentes estrategias, herramientas y recursos bioinformáticos para el análisis de este tipo de datos. Por ello, el objetivo de este trabajo es la realización de una evaluación y comparativa de métodos de análisis de datos de microbioma, que nos permita conocer y decidir cuáles son los procedimientos óptimos en cada estudio biomédico.

Para alcanzar este objetivo principal, se cumplirán los siguientes objetivos específicos:

1. Revisión de los repositorios de datos existentes sobre microbiota.
2. Selección de conjunto de datos de interés que nos sirvan para reproducir el análisis de los mismos con diferentes herramientas.
3. Revisión de cada una de las estrategias de análisis de datos de microbioma.
4. Comparación de estas estrategias para comprobar su nivel de adecuación y optimización en los estudios biomédicos.

Los resultados obtenidos en este trabajo, serán utilizados en el proyecto “Share Your Brain”, donde se caracterizarán las diferencias de sexo en el microbioma de población control sana. Esta información será de gran utilidad para el estudio de las enfermedades mentales.

3. Material y Métodos

A continuación se detallan todos los procedimientos llevados a cabo para la comparación de los diferentes tipos de abordaje en el análisis de datos de microbioma, así como los conjuntos de datos empleados para su evaluación y comparación.

3.1. Revisión de repositorios y bases de datos sobre estudios de microbioma

Se ha llevado a cabo una revisión sistemática de todos los repositorios de datos sobre microbioma. Para realizar correctamente este procedimiento, es necesario tener unos objetivos claros que ayuden a encontrar estudios que cumplan con unos criterios de selección concretos que se ajusten a nuestros objetivos (Liberati, 2009).

En este caso, lo primero que se hizo fue buscar en internet todas las bases de datos que tuvieran información sobre microbioma, microbiota o microorganismos en general y que además presentaran los fastq descargables. El formato fastq es un formato basado en texto plano para el almacenamiento de una secuencia biológica con la puntuación a nivel de la calidad de la secuenciación. Las bases de datos encontradas fueron MicrobiomeDB, Human Microbiome Project Data Portal (HMPDP), Human Oral Microbiome Database (eHOMD), MGnify, China National Genebank: Microbiome Database (CNG) y Gene Expression Omnibus (GEO). A continuación se resume la información recogida en cada una:

- **MicrobiomeDB.** Es una herramienta de descubrimiento que permite a los investigadores aprovechar sus metadatos experimentales para construir consultas sobre los conjuntos de datos de microbiomas. Esta base de datos, además de proporcionar su propio pipeline con el que analizar tus datos, contiene la información de 12 estudios relacionados con el microbioma tanto de humanos como de otros seres vivos como ratones o perros.

  Los datos están separados por estudios y se puede realizar la búsqueda de los mismos por tipo de hospedador o directamente por la muestra. En esta base de datos se pueden descargar los datos referentes a las tablas de conteos pero no las secuencias de los estudios, que están depositados en el SRA (Francislon, 2018).

- **HMPDP.** Este es el portal de datos del Proyecto del Microbioma Humano. Este proyecto se centra en la colección de todos los microorganismos que

viven en asociación con el cuerpo humano (Human Microbiome Project Consortium, 2012).

En este portal se recoge la información de 18 estudios. Los datos de estos estudios pueden gestionarse por archivos o por muestras. Las muestras se pueden filtrar por zona corporal, tipo de estudio o género, y los archivos por formato, tipo de dato, anotación y tipo de matriz.

Los datos son descargables en diferentes formatos (Fastq, Standard flowgram file, tsv y Biological observation matrix). Los estudios recogidos en este portal se diferencian en estudios de microbiota en adultos sanos y estudios de casos vs. control en algunas enfermedades.

- **eHOMD.** En esta base de datos se recoge toda la información sobre las diferentes especies que habitan tanto el tracto respiratorio como el tracto digestivo humano. En total contiene datos de 771 especies, y de ellas se pueden descargar los genomas, proteomas o secuencias de 16S ARNr (Escapa, 2018).

- **MGnify.** Forma parte del Emb-Ebi. Ofrece un pipeline automatizado para el análisis y archivamiento de datos de microbiomas para ayudar a determinar la diversidad taxonómica y el potencial funcional y metabólico de muestras. Los usuarios pueden subir sus propios datos para el análisis o utilizar los conjunto de datos públicos. Contiene los datos de diferentes tipos de biomas y en diferentes especies. Además seleccionando cualquier estudio de interés puedes descargar los datos en formato fastq a partir del ENA (European Nucleotide Archive) (Mitchell, 2020).

- **CNG.** Es una base de datos que contiene información sobre diferentes tipos de bioma y diferentes especies. En ella puedes acceder a 534 proyectos, los cuales presentan enlace de acceso al ENA para descargar los fastq (Microbiome DataBase).

- **GEO.** Es un repositorio público de datos genómicos en el cual se puede encontrar y descargar perfiles de genes de expresión curados (Edgar, 2002). Aquí se recogen los datos de estudios de todo tipo, nosotros nos centramos en los estudios de microbiomas en humanos y en total había 14 estudios.

Una vez revisados los repositorios, se procedió a la búsqueda de conjunto de datos en los que preferentemente se estudiara población control entre sexos, o en segundo lugar presentaran un diseño caso vs. control (diseños que se utilizarán en el proyecto “Share your Brain”). Principalmente la búsqueda se realizó en GEO, debido a su facilidad en la selección de estudios y el acceso a

los datos. Se emplearon las palabras clave “microbiome”, “human microbiome”, “metagenome” y “human metagenome”. Puesto que para el proyecto de “Share your Brain” necesitamos desarrollar un pipeline de estudio metagenómico en humanos, se buscaron sólo estudios en humanos. Tras la aplicación de estos criterios de inclusión, sólo se encontraron un total de 8 estudios, de los cuales algunos no proporcionaban los fastq (tipo de archivo necesario para el análisis) y por tanto las opciones con las que trabajar se redujeron aún más, seleccionando finalmente dos estudios de interés.

3.2. Descripción de conjunto de datos.

Se han seleccionado dos conjunto de datos sobre estudios de microbioma en humano, que se utilizarán para aplicar las diferentes estrategias de análisis y realizar su comparación.

El primer estudio seleccionado fue *Alterations of the human gut microbiome in multiple sclerosis* (Jangi, 2016). En este experimento se estudia el papel que ejerce el microbioma intestinal en la esclerosis múltiple (MS). Para ello analizan la región V4 del gen 16S ARNr tanto en casos (MS, n=61) como en controles (n=43). Además, el grupo de casos se divide en tres grupos: dos tratados (unos tratados con interferón y otros tratados con copaxona) y otro sin tratamiento. Los no tratados debían estar con tratamiento placebo o con un tratamiento no esteroide en el mes previo al estudio. Los individuos tratados son los que han recibido un tratamiento de beta-interferón o de copaxona durante al menos 6 meses. Además, todos los individuos fueron sometidos a una dieta específica. Se tomó una muestra fecal única de cada paciente y se realizó la extracción del ADN para secuenciar las muestras mediante secuenciación por fluorescencia de Illumina MiSeq. Las secuencias generadas fueron analizadas siguiendo el protocolo MiSeq SOP de Mothur (Schloss, 2011).

En este estudio utilizan el programa Mothur (v.1.34.2) para realizar los análisis. Nosotros vamos a reproducirlo utilizando tanto este programa como otros que se explican a continuación.

El segundo estudio seleccionado ha sido *Characterization of antibiotic resistance and host-microbiome interactions in the human upper respiratory tract during influenza infection* (Zhang, 2020). En este trabajo se estudia la expresión de los genes con resistencia a antibióticos de tracto respiratorio superior en un grupo de personas con gripe (n=33) para entender mejor el papel que esta resistencia a antibióticos ejerce en la patogénesis de las infecciones bacterianas secundarias asociadas a la gripe. Las muestras de garganta se recogieron al pasar 1-2 días desde que empezara la enfermedad. En este estudio se

secuenció la región V4 del gen 16S ARNr con Illumina MiSeq. Para analizar el conjunto de datos utilizaron QIIME2 (Caporaso, 2010), además de otros programas como FastQC o Trimmomatic para el control de calidad y filtrados previos.

3.3. Revisión de métodos de análisis de datos de microbioma.

La forma de analizar los datos procedentes de microbioma difiere en función del tipo de secuenciación que se haya hecho. Los dos procedimientos principales son la metataxonomía y la metagenómica (del Campo-Moreno, 2018).

La metataxonomía es la estrategia que más suele utilizarse para caracterizar la composición de las comunidades microbianas presentes en un microbioma. Este tipo de abordaje se realiza cuando se ha hecho una secuenciación de tipo *gene-marker*, en la que, como ya se ha mencionado, se amplifica el gen ARNr 16S y después se secuencian masivamente los amplicones. Generalmente se secuencia una o dos de las nueve regiones hipervariables de este gen (Fig. 4).

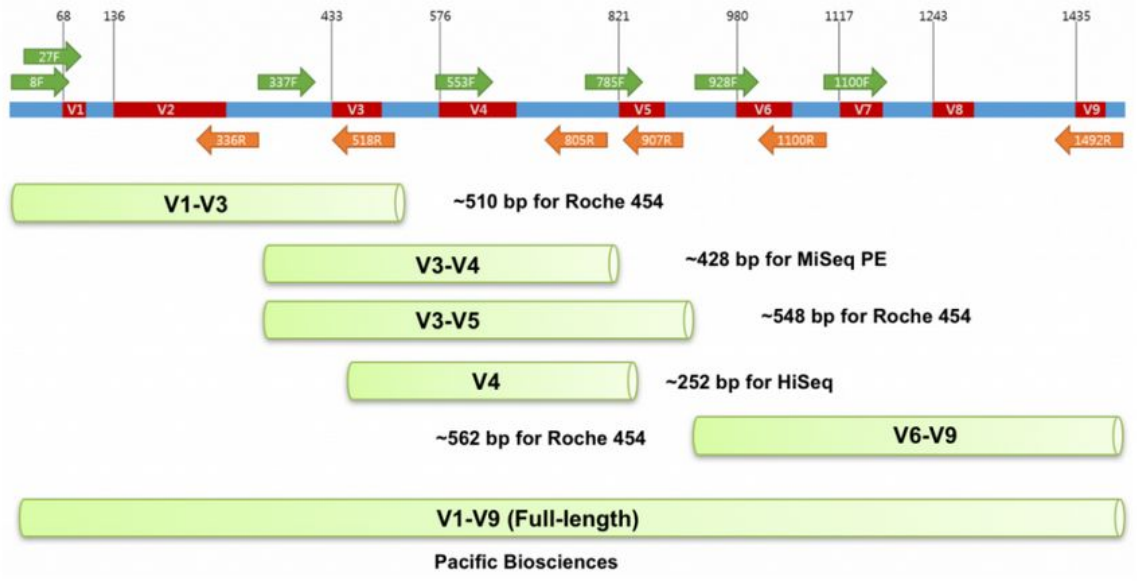

Figura 4. Regiones hipervariables del gen 16S ARNr y las tecnologías que las secuencian. (https://help.ezbiocloud.net/16s-rrna-and-16s-rrna-gene/).

En este caso, durante el análisis bioinformático se busca la asignación de cada secuencia a un grupo taxonómico mediante su comparación con una base

de datos. Los pasos habituales suelen ser 1) control de calidad de las secuencias, 2) filtrado y eliminación de quimeras (secuencia híbrida generada por la combinación de dos templates no relacionados durante una PCR), 3) clustering o agrupamiento de las secuencias por similitud, 4) asignación taxonómica y 5) análisis estadístico para determinar las diferencias significativas.

El clustering se ha realizado tradicionalmente en Unidades Taxonómicas Operacionales (OTUs), pero recientemente está surgiendo una nueva forma de hacerlo mediante el uso de Variantes de Secuencia de Amplicones (ASVs).

Los OTUs concretamente son lecturas que se agrupan porque difieren entre ellas por menos de un umbral establecido (normalmente el 97%), esto es un problema debido a que se pueden perder especies microbianas que tengan pocas lecturas ya que se acabarían agrupando dentro de un cluster mayor. Hay dos métodos para definir OTUs, el método *de novo* y el de referencia cercana. El más común es el de referencia cercana en el que las lecturas más semejantes a una secuencia en una base de datos de referencia son reclutadas dentro del OTU correspondiente.

Con los nuevos métodos que se han desarrollado, el resultado final son los ASVs de los amplicones en los que no se imponen los umbrales de diferencias que definen a los OTUs moleculares (Callahan, 2017).

Los métodos de ASV infieren las secuencias biológicas en las muestras antes de la introducción de la amplificación y los errores de secuenciación, y distingue las variantes de secuencia que difieren tan sólo por un nucleótido. Concretamente lo que hacen es que en vez de ir agrupando secuencias que estén por encima de un determinado valor umbral de similitud, sólo agruparán lecturas en un ASV dado cuando sean idénticas a nivel de secuencia, esto es posible porque los algoritmos de ASV tienen un modelo estadístico de *denoising,* capaz de detectar los errores de secuenciación producidos por Illumina.

Los métodos de ASV han demostrado tener una sensibilidad y especificidad tan buena o incluso mejor que los métodos basados en agrupación en OTUs (Eren *et al.*, 2013; Eren *et al.,* 2015).

La validez de estos tipos de agrupaciones puede observarse en la figura 5, en la que el eje X representa todas las variaciones biológicas que existen en el locus genético secuenciado y el eje Y representa todos los amplicones generados para ese locus y todos los datos futuros que pueden ser generados. La región sombreada sería la validez. Con esta gráfica podemos ver las limitaciones de los OTUs. Los OTUs *de novo* no son válidos fuera del grupo de datos en el que son definidos. Los OTUs de referencia cercana no pueden recoger variaciones biológicas que están fuera de la base de datos de referencia

utilizada en su creación. Por el contrario, los ASVs sí que recogen todas las variaciones biológicas presentes en los datos y pueden ser reproducidos en futuros conjunto de datos y por tanto comparados de forma válida.

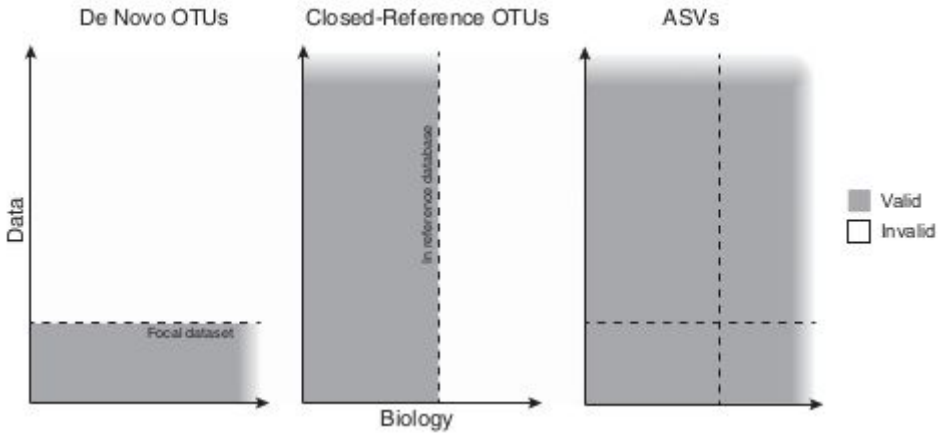

Figura 5. Validez de los OTUs *de novo*, los OTUs de referencia cercana y los ASVs de un conjunto de datos. La zona sombreada representa la validez del método (Callahan, 2017).

Por otra parte, la metagenómica se utiliza para estudiar cómo las alteraciones en la composición de la microbiota influyen en el contenido de los genes y su expresión. Este abordaje se emplea cuando se realiza una secuenciación WGS en el que se secuencia todo el ARN. El conjunto de secuencias obtenidas se considera representativo de los genomas bacterianos presentes en la microbiota. Esta estrategia proporciona una visión más fiel de la distribución taxonómica existente de los microorganismos. Los pasos bioinformáticos básicos para el análisis de este tipo de muestras suelen ser 1) control de calidad, 2) ensamblaje de las lecturas obtenidas, 3) anotación de las lecturas que corresponden a las regiones contiguas del genoma y 4) caracterización taxonómica de las lecturas o contigs (Thomas, 2012).

Generalmente suele utilizarse más el abordaje metataxonómico debido a su bajo coste y a su alta precisión. Además, el gen 16S ARNr está altamente conservado y tiene muchas regiones variables. Las regiones altamente conservadas nos permiten identificar genes más fácilmente entre organismos, mientras que las regiones con alta variabilidad nos permiten distinguir entre especies.

Debido a esto, y a que en los estudios seleccionados los datos de secuenciación son del gen 16S ARNr, hemos decidido centrarnos en el desarrollo de un pipeline para analizar este tipo de datos.

Existen diversas herramientas con las que abordarlos. De hecho, hay herramientas específicas para cada uno de los pasos a realizar, generando numerosas combinaciones de estrategias de análisis con diferentes recursos.

Por ello, decidimos seleccionar aquellas herramientas que permitían la ejecución completa de todos los pasos en el análisis de microbiomas.

Para seleccionar estos recursos o herramientas, revisamos en la literatura diversas evaluaciones de análisis metataxonómico.

En general, las dos herramientas más utilizadas son Mothur y Qiime. Sin embargo, los estudios que comparan estas herramientas en conjunto de datos reales son escasos. Además, otro método ampliamente usado es DADA2, un pipeline nativo de R. Estos 3 recursos son los que nosotros utilizaremos para realizar el benchmarking y analizar los estudios escogidos. A continuación, se explica paso a paso el procedimiento llevado a cabo con cada herramienta.

Mothur. Hemos utilizado la versión 1.44.1. El primer paso del pipeline (Figura 6) es generar los contigs a partir de las lecturas forward y reverse, además en este paso se genera también un archivo fasta con los contigs resultantes combinados. Después se procede a realizar un control de calidad, para ello se comprueba si ha habido algún mal ensamblaje o si hay ambigüedad y se procede a la eliminación de los mismos. Puesto que Mothur aporta información sobre las bases ambiguas o los homopolímeros, pero no sobre las calidad de la secuencia en sí, antes de continuar con el procedimiento, realizamos un análisis con FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Este programa nos proporcionará, entre otros resultados, unas gráficas que mostrarán la calidad base a base de las secuencias.

Ya que las muestras de microbiomas contienen grandes cantidades del mismo organismo, se espera encontrar muchas secuencias idénticas en nuestros datos. Por tanto, vamos a determinar las secuencias únicas y después guardar cuántas veces cada una de estas secuencias es observada en el conjunto de datos original. Este paso genera dos archivos: un archivo fasta que contiene las secuencias únicas y el otro es un *names file* que contiene el nombre de secuencia para cada secuencia única y el resto de nombres de secuencia que son idénticos a la de la primera.

Una vez realizado el control de calidad, se procede al alineamiento de las secuencias. Este alineamiento se realiza frente a la base de datos SILVA (Quast, 2013), pero sólo contra una región concreta de la misma, en este caso con la región V4 del gen 16S ARNr. Este paso genera como output un fasta alineado. A continuación, se lleva a cabo un segundo control de calidad en el que se eliminan los homopolímeros que superen en tamaño a los de nuestras muestras y los gap terminales. Después de realizar este filtrado, cabe la posibilidad de que hayan vuelto a aparecer duplicados y por tanto volvemos a hacer selección de las secuencias únicas.

El siguiente paso es el *pre-clustering*, vamos a separar las secuencias por grupos y ordenarlas por abundancia (de más a menos), entonces se identifican las secuencias que estén a dos o menos nucleótidos de las otras y se agrupan. Este paso generará un nuevo fasta. A continuación, se procede a la eliminación de las posibles quimeras. Durante la amplificación por PCR es posible que dos templates no relacionados se combinen para formar una secuencia híbrida, esto es una quimera. Éstas se identifican con la herramienta *vsearch* que incluye el programa y se eliminan.

Una vez hecho esto, se realiza la clasificación taxonómica. Para ello se utiliza un clasificador bayesiano que usa por defecto Mothur, y un training set proporcionado por Schloss basado en la referencia taxonómica *Ribosomal Database Project* (RDP) (Cole, 2014).

A pesar de todo lo que hemos hecho para mejorar la calidad de nuestros datos, aún puede haber fragmentos 18S ARNr, o fragmentos 16S de arqueas, cloroplastos, etc., que se hayan escapado en los pasos de limpieza. Para eliminarlos, primero se clasifican las secuencias y a continuación podemos indicar que se elimine todos los que entren dentro de las clasificaciones que le indiquemos (Chloroplast, Mitochondria, unknown, Archaea, Eukaryota, etc). Estos nombres se basan en los nombres con los que están clasificados los taxones en la referencia RDP, si se usa otra base de datos, habría que adaptar los nombres.

Otro paso opcional es el cálculo de las tasas de error usando una Mock Community. Esto son comunidades microbianas de las cuales se conoce su diversidad y su abundancia con exactitud, de forma que se secuencian y se pasan por la herramienta de análisis para ver si las predicciones de la herramienta son correctas. De esta forma podemos obtener la tasa de error relativo de nuestro protocolo, si nuestro workflow se ejecuta correctamente en la muestra mock, tendremos una mayor confianza en la precisión de los resultados en el resto de las muestras.

El siguiente paso es el *clustering* o agrupación en OTUs de nuestras secuencias de interés. La herramienta agrupará las diferentes secuencias en OTUs por su similaridad y además proporcionará la clasificación taxonómica de cada uno. Puesto que algunas de nuestras muestras pueden contener más secuencias que otras, se normaliza el conjunto de datos mediante el submuestreo. Para ello le indicamos el número de secuencias de la muestra más pequeña y lo que hace esto es que ahora todos los grupos tendrán ese número de secuencias.

Una vez realizados todos estos pasos, la misma herramienta de Mothur te permite realizar algunos análisis estadísticos como son el análisis de diversidad alfa (complejidad ecológica de una única muestra) y beta (complejidad ecológica entre muestras).

DADA2. Paquete de R (v. 3.6.1) (Callahan, 2016). Para su uso se requiere disponer de los archivos fastq separados por muestras, los nucleótidos no biológicos deben de haber sido eliminados y si los datos son apareados, los fastq forward y reverse deben contener lecturas emparejadas en orden. Este paquete es muy útil y fácil de utilizar ya que tiene implementada una función para cada uno de los pasos principales del análisis metataxonómico.

Como ya se ha indicado antes, el primer paso es el control de calidad de nuestros archivos. Primero cargamos los fastq y visualizamos gráficamente los perfiles de calidad. Con estas representaciones podemos valorar la calidad de nuestras muestras para saber cuáles son los parámetros que nos interesará incluir en el filtrado.

A continuación, procedemos a realizar el filtrado y el recorte de las muestras. Usamos parámetros de filtrado estándar: maxN=0 (número máximo de Ns admitidas), truncQ=2 (eliminar lecturas con un valor de calidad menor al valor indicado), rm.phix=TRUE (eliminar el *phix,* cadenas sintéticas de ADN que se utilizan para calibrar la muestra) y maxEE=2 (elimina las lecturas con un valor de *Error Esperado* mayor al indicado). Normalmente se establece un maxEE más relajado para las lecturas reversas ya que suelen tener peor calidad, sobretodo al final de la lectura.

El siguiente paso es la asignación de las tasas de error. El algoritmo de DADA2 hace uso de un modelo de error paramétrico (*err*) y cada conjunto de datos de amplicones tiene un set diferente de tasas de error. El método *learnErrors* aprende este error de los datos, lo hace alternando la estimación de las tasas de error y la inferencia de la composición de las muestras hasta que convergen en una solución consistente. Como en muchos problemas de machine learning, el algoritmo debe empezar con una suposición inicial, para la que la tasa de error máxima posible de los datos sea usada.

Después representamos las tasas de error en gráficas. Concretamente se muestran las tasas de error para cada posible transición ($A \rightarrow C$, $A \rightarrow G$, …).

A continuación, se ejecuta una función para deducir el número de secuencias variantes dentro del total de secuencias únicas. Y después se unen las secuencias apareadas para obtener las secuencias sin ruido. La mezcla se realiza alineando las lecturas reversas sin ruido con las reversas-complementarias de las lecturas reversas correspondientes, y entonces construyendo el contig. Por defecto, las secuencias unidas son el único output si las lecturas forward y reverse solapan por al menos 12 bases, y son idénticas la una a la otra en la región solapada. El objeto de la mezcla es una lista de data frames para cada muestra. Cada data frame contiene la secuencia unida, su abundancia y los índices de las secuencias variantes

forward y reverse que son unidas. Las lecturas apareadas que no solapen exactamente son eliminadas.

Después construimos una tabla de secuencias, que será una matriz cuyas filas corresponden a las muestras y las columnas corresponden a las secuencias variantes. Una vez hecho esto se eliminan las quimeras. El método *core* de DADA corrige los errores de sustitución y de indels, pero las quimeras permanecen. Afortunadamente, la precisión de las secuencias variantes después de eliminar el ruido hace la identificación de ASVs quiméricas más simple que cuando se trata de OTUs. Las secuencias quiméricas son identificadas si pueden ser reconstruidas exactamente combinando el segmento izquierdo y el derecho de las dos secuencias parentales más abundantes. La frecuencia de las secuencias quiméricas varía sustancialmente de un conjunto de datos a otro, y depende de factores que incluyen los procedimientos experimentales y la complejidad de la muestra.

Ahora es cuando se hace un último chequeo para ver si hay algún error, en caso de haberlo, habría que revisar los parámetros utilizados en funciones anteriores y corregir hasta obtener los resultados deseados.

Por último, una vez comprobado que todo está bien, realizamos la asignación taxonómica. El paquete DADA2 proporciona una implementación nativa del método de clasificación Bayesiana. La función *assignTaxonomy* toma como input un set de secuencias para ser clasificadas y un set de secuencias de entrenamiento con una taxonomía conocida (Mock) y nos devuelve dos objetos. El primero será una tabla de ASVs por conteos por lectura y el segundo será la clasificación taxonómica para los ASVs.

Aquí, igual que con Mothur, se puede realizar el paso opcional de evaluar la precisión de nuestro pipeline a partir de una “Mock community”. Si el análisis de las secuencias Mock nos da el número de cepas que sabemos que contiene esa muestra podemos afirmar que nuestro protocolo funciona correctamente.

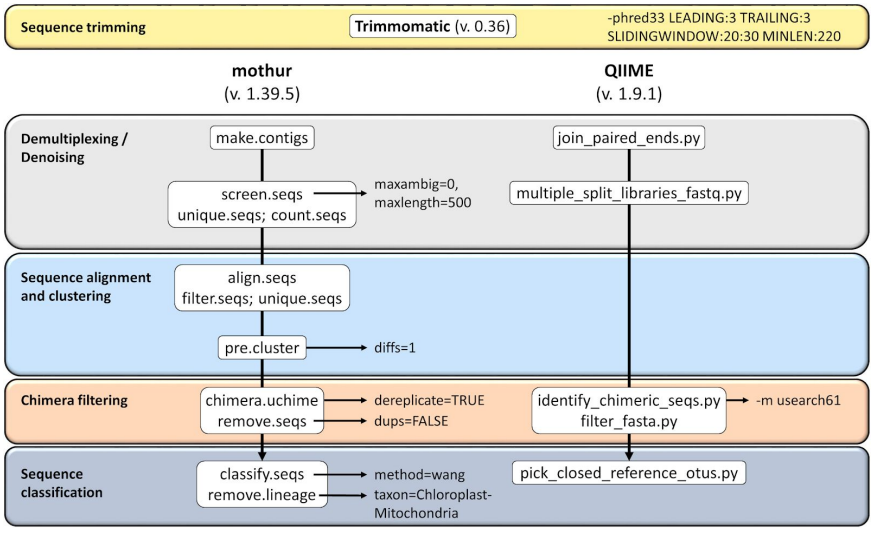

Figura 6. Comparación pipelines Mothur y QIIME2 (López-García, 2018).

QIIME2. Es la tercera herramienta con la que hemos analizado los datos para hacer el estudio de benchmarking. Hemos utilizado la versión 2019.10.0 (Boylen, 2019).

Lo primero que hacemos es crear los *artefactos* con los que vamos a trabajar. Estos *artefactos* no son lo que se suele conocer como tal, sino que son un tipo de archivos que contiene nuestros datos y sus metadatos correspondientes. Por tanto, para generarlos le pasamos los fastq y los metadatos en formato tsv.

A continuación, representamos la calidad de las secuencias para saber dónde habría que truncar. Una vez conocida la calidad, QIIME2 hace uso de DADA2 para quitar el ruido a nuestras secuencias. Este proceso genera como resultado 3 archivos. El primero es un artefacto *qza* que describe los cambios que han ido sufriendo las secuencias en cada etapa de tratamiento. El segundo archivo es un artefacto *FeatureData* que incluye las secuencias representativas, es decir, las secuencias únicas. Además, proporciona información sobre el tamaño, el mínimo, el máximo, el promedio, etc. de las secuencias. A estas secuencias se les asigna un *FeatureId* por lo que cada una de ellas ya está mapeada. El último sería un archivo tipo *FeatureTable*, el cual contiene la relación entre los *FeatureId* y su abundancia.

Después se procede con la asignación taxonómica de las secuencias. En este caso elegimos otra vez la base de datos SILVA y utilizamos un archivo *classifier* que podemos descargar de la misma página web de la herramienta. Una vez clasificados, podemos filtrar nuestros datos para eliminar los taxones que no nos interesen, como por ejemplo cloroplastos y mitocondrias.

El último paso es crear una *FeatureTable* nueva, pero esta vez anotada según la taxonomía. En este paso es importante indicar el nivel taxonómico al cual vamos a crear la tabla, en este caso lo haremos a nivel 7 que es el nivel de especie.

Puesto que todos los archivos generados por QIIME2 son de tipo *artefacto*, debemos transformarlos en objetos visualizables, los cuales se abrirán mediante nuestro buscador.

Ya que cada método de análisis genera unos archivos de salida con formatos diferentes, se han generado diversos scripts con R con los que poder obtener, en caso de que el propio programa no diera la opción directamente, los gráficos de las medidas que nos interesan. Estas medidas serían la diversidad alfa, las curvas de rarefacción, el Escalamiento Multidimensional No-Métrico o NMDS y la abundancia diferencial.

La obtención de estos resultados varía en función del método de análisis. En el caso de Mothur y QIIME2 se han obtenido las curvas de rarefacción y las gráficas NMDS, además de generar unas pie charts con Krona (Ondov, 2011)

en las que se visualiza la proporción de la abundancia taxonómica. Con DADA2 en cambio, se han obtenido gráficas de diversidad alfa, rarefacción, NMDS y una tabla con la abundancia diferencial, pero no se ha podido generar la visualización con Krona.

La diversidad alfa es una medida que determina la diversidad de las especies microbianas encontradas en un ecosistema dado. Para realizar este cálculo, las métricas más utilizada son Chao1 (Chao, 1984) y Shannon (Pla, 2006). La primera determina la riqueza de un ecosistema, es decir, la cantidad de especies encontradas, de hecho, da más peso a las especies menos representadas. Por tanto, una Chao1 elevada indicaría un ecosistema con una gran cantidad de especies raras. La segunda, determina la abundancia de las especies y por tanto, le da más importancia a la riqueza (que haya muchas especies) y a la igualdad (que el número de individuos entre especies sea similar). El índice de Shannon suele oscilar entre los valores 2 y 4. Se suele interpretar que valores menores a 2 pertenecen a ecosistemas con una diversidad de especies relativamente baja, mientras que los mayores a 3 tendrán una diversidad alta.

La rarefacción es una técnica utilizada para evaluar la riqueza de especies en un muestreo. Esta técnica realiza el cálculo basándose en la construcción de las llamadas curvas de rarefacción. Estas curvas son una representación gráfica del número de especies como una función de número de muestras. Las curvas de rarefacción generalmente crecen rápido al principio, ya que es cuando se van detectando las especies más comunes, pero acaban dibujando una meseta conforme se limita su capacidad para detectar las especies restantes. Por tanto, estas curvas nos permiten estudiar cómo varía la diversidad alfa en base a la profundidad de la secuenciación. Además, generamos unas curvas de rarefacción. Estas curvas muestran el número de especies como una función del número de individuos muestreados. De esta forma podemos ver si todas las especies han sido muestreadas.

El NMDS es una técnica multivariante de independencia que representa en el espacio la proximidad existente entre un conjunto de objetos o, en este caso, de grupos de estudio. Es decir, nos ayuda a resumir las diferencias entre muestras y encontrar grupos homogéneos de muestras, por lo que en caso de haber disimilitudes significativas entre grupos de estudio los veríamos claramente separados en el espacio. Para calcular esta disimilitud hemos utilizado en todos los casos el método de Bray-Curtis. Este método varía entre 0 y 1, siendo 1 el resultado de dos comunidades que no comparten ninguna especie.

Para obtener los análisis de diversidad alfa, rarefacción y NMDS se ha utilizado el paquete de R *phyloseq* (McMurdie, 2013).

El cálculo de abundancia diferencial nos sirve para saber, comparando dos grupos de estudio, en cuál de ellos es más abundante cada ASV. Para ello se utiliza el paquete de R *DESeq2* a partir de los resultados obtenidos con DADA2. De hecho, utiliza el objeto phyloseq creado anteriormente y se le aplican las funciones de *DESeq2*.

Todos los pipelines han sido ejecutados utilizando el cluster HPC (High Performance Computing) del Centro de Investigación Príncipe Felipe (https://bioinfo.cipf.es/ubb/?page_id=338).

4. Resultados

Los resultados presentados a continuación siguen el orden de análisis realizado con cada uno de los pipelines explicados anteriormente en el apartado de Material y Métodos. Para cada paso de análisis se mostrará el resultado obtenido por cada herramienta haciendo comparación entre ellos.

4.1 Revisión de repositorios

Se buscaron conjunto de datos en la base de datos GEO que se pudieran utilizar para comparar los 3 pipelines y que pertenecieran a estudios realizados en humanos. Combinando diferentes palabras clave se encontraron un total de 8 estudios. Tras su evaluación, la mayoría fueron descartados por no tener los archivos fastq disponibles en la base de datos. Finalmente se seleccionaron los dos estudios descritos anteriormente, los cuales, además de disponer los fastq para su descarga, presentaban un diseño experimental de interés (caso vs. control y mujer vs. hombre).

4.2 Análisis de calidad.

Se llevó a cabo una evaluación de la calidad de las secuencias, en cada uno de los métodos de análisis seleccionados.

4.2.1 Mothur y FastQC

En el caso de Mothur, el paquete no incorpora la opción para generar un análisis de calidad *per se,* sino que proporciona únicamente una tabla resumen con las características de nuestros datos (TS. 1). Concretamente nos muestra el número total de secuencias, que sería de 11.502.947 en el estudio de MS y de 135.287 en el estudio de gripe, el tamaño de las cuales varía en su mayoría entre 151-254 pares de bases (pb) y 250-258 pb respectivamente. Llegando a presentar algunas lecturas mucho más grandes, de unos 502 pb. Estas tablas también muestran el número de bases ambiguas, que en promedio es menor en el estudio de MS, aunque su número máximo por lectura es mayor. Y por último muestra el número de homopolímeros por lectura: en ambos casos la mayoría de las lecturas tienen homopolímeros de unas 6 pb de largo.

Para completar el análisis de la calidad de las secuencias, utilizamos el software FastQC. Los resultados obtenidos están descritos en la figura 7. Estas representaciones aportan información de la calidad de cada uno de los nucleótidos que forman la secuencia, según su posición. Como puede observarse en ambos estudios la calidad media de las lecturas está por encima de 30 (Phred Quality Score) lo que indica una buena calidad de las mismas, siendo un poco mayor la calidad de las lecturas del estudio de MS. Además, se puede apreciar cómo disminuye la calidad conforme nos acercamos al final de la lectura, esto es común y se debe a la propia naturaleza de los cebadores, los cuales se van despegando y la secuenciación pierde calidad. También se puede apreciar una peor calidad de las secuencias reverse, esto suele ser normal en la secuenciación realizada con Illumina, debido al protocolo experimental usado en el laboratorio ya que la R2 se secuencia más tarde que la R1 y entonces los reactivos que hay dentro del secuenciador se empiezan a degradar (Tan, 2019).

El análisis realizado con fastQC nos proporciona más información acerca de nuestras secuencias, como por ejemplo la proporción de contenido de bases (C, G, T, A) por posición de la secuencia, la proporción de contenido en GC o las secuencias sobrerrepresentadas.

4.2.2 QIIME2

Al contrario que Mothur, QIIME2 sí que incluye un análisis detallado de la calidad de las secuencias como el que nos proporciona FastQC (FS. 1). En este caso se observa una buena calidad en las secuencias forward pero una

calidad bastante baja en las reverse, coincidiendo con los resultados anteriores en que la calidad de las lecturas del estudio de gripe es, en líneas generales, más baja.

Este primer análisis de calidad con QIIME2 nos proporciona también unas gráficas con el número de secuencias y la frecuencia con la que éstas aparecen (Fig. 8). Como se puede observar en las gráficas, en el estudio de MS la frecuencia media del número de secuencias es mayor, de unas 11.000, mientras que en el estudio de gripe la frecuencia media es bastante más baja, de unas 3.000.

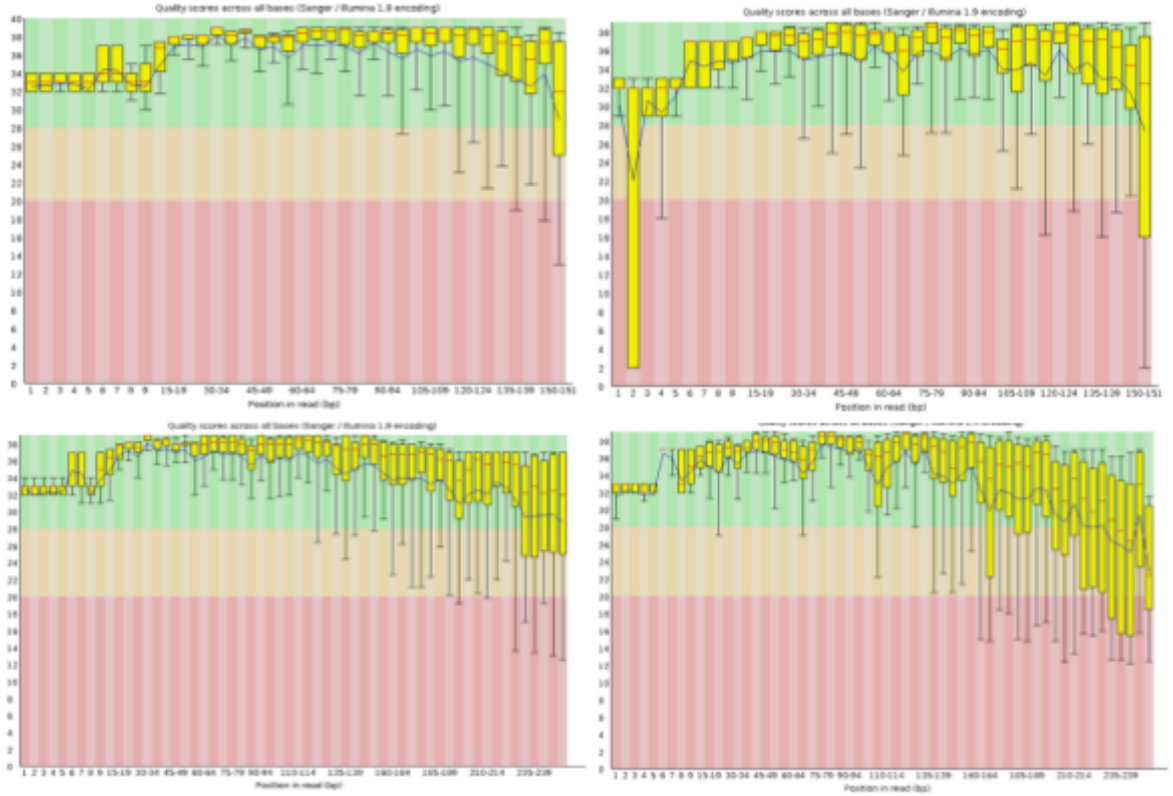

Figura 7. Análisis de calidad de FastQC. Análisis de las lecturas forward y reverse de dos muestras de los estudios de MS (a y b) y de gripe (c y d). Como puede observarse la calidad de las lecturas disminuye hacia el final de la secuencia y es bastante más baja en las reverse.

4.2.3 DADA2

Con el paquete DADA2 de R se pueden obtener también las gráficas que muestran la calidad base a base de las secuencias de nuestras muestras (FS. 2). Además, muestra la frecuencia de cada score de calidad para cada base. Los resultados son los mismos que con los anteriores métodos, una

calidad media mayor en las secuencias del estudio de MS y una calidad que disminuye bastante en las secuencias reverse.

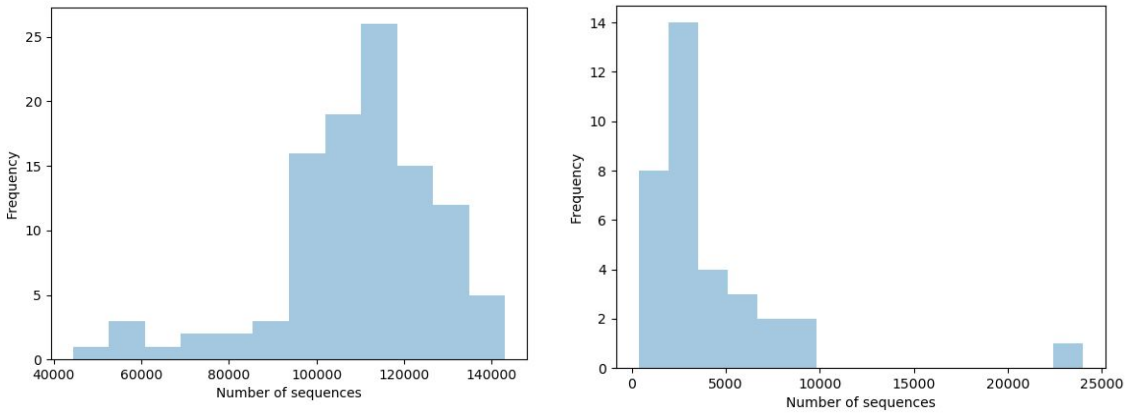

Figura 8. Frecuencia del número de secuencias por muestra obtenida con QIIME2. Las gráficas muestran el número de secuencias del estudio de MS (a) y del estudio de gripe (b).

4.3 Diversidad alfa y rarefacción

Como ya se ha mencionado, la diversidad alfa sirve para medir la diversidad de especies de una muestra y las curvas de rarefacción se utilizan para representar cómo varía esta diversidad conforme aumenta la profundidad de las lecturas. A continuación, se muestran los resultados obtenidos por los distintos pipelines en el cálculo de la diversidad alfa y de las curvas de rarefacción.

4.3.1 Mothur

En este caso, el programa produce únicamente las curvas de rarefacción en las que, como ya se ha mencionado, podemos ver cómo varía la diversidad alfa. Mothur nos da la opción de calcularlas con distintas métricas, sin embargo no presenta opción de calcular el índice de Shannon por lo que sólo hemos calculado el de Chao1. En la figura suplementaria 3 aparecen representadas las curvas de rarefacción. Como puede observarse, hay una diversidad mucho mayor en el estudio de MS que en el de gripe. Aunque en el caso de la MS han sido detectados una gran cantidad de OTUs, mucho mayor que con el resto de análisis.

4.3.2 QIIME2

Con QIIME2 sucede lo mismo que con Mothur, únicamente nos da la opción de generar las curvas de rarefacción y con ellas poder ver cómo varía la diversidad.

Como puede observarse en las gráficas en las que se representa el índice Chao1, la cantidad de especies totales encontradas también es mayor en el estudio de MS (FS. 4) que en el estudio de gripe (FS. 5). En el estudio de MS hay grandes diferencias en cuanto a la cantidad de especies encontradas entre unas muestras y otras, pero en todas ellas se siguen detectando especies nuevas a una profundidad más elevada de secuenciación. Sin embargo, en el caso del estudio de gripe la cantidad de especies encontradas en las muestras se satura a una profundidad de secuenciación relativamente baja en la mayoría de ellas salvo en dos, las cuales muestran una mayor cantidad de especies.

En las gráficas correspondientes al índice de Shannon, en el caso del estudio de la MS (FS. 4) hay una mayor diversidad que en el estudio de gripe (FS. 5), habiendo una diferencia de unos 2 puntos entre ambas. Además, en el estudio de gripe podemos observar cómo hay más diferencias entre muestras, siendo que gran parte de ellas se saturan y alcanzan el máximo de diversidad a una profundidad de secuenciación más baja.

4.3.3 DADA2

Con este paquete de R pueden obtenerse de forma sencilla diferentes métricas, como ya se ha mencionado anteriormente.

Lo primero que nos permite representar es, sin utilizar las curvas de rarefacción, una representación de la diversidad alfa a través de los índices de Shannon y Chao1 (F S. 6). En estas gráficas puede observarse que los resultados coinciden con los obtenidos anteriormente, siendo que hay una mayor diversidad y abundancia en los grupos del estudio de MS que en el de gripe.

En el estudio de MS no se aprecian diferencias significativas entre grupos en cuanto a la diversidad de especies y la abundancia de las mismas. En el estudio de gripe tampoco se observan diferencias significativas en cuanto a la diversidad y abundancia medias, pero sí que se puede ver cómo en el grupo de hombres hay una muestra que presenta una mayor cantidad

de especies según marca el índice de Chao1, coincidiendo con los resultados hallados anteriormente.

Además de poder representar de esta forma la diversidad alfa, DADA2 nos permite conocer su variación a través de las curvas de rarefacción. De hecho, nos permite visualizar esta variación por muestra y por grupo de estudio.

En estas gráficas (Fig. 9) puede observarse claramente qué grupos del estudio o muestras tienen una mayor diversidad. En el estudio de la MS no se aprecian diferencias significativas entre grupos del estudio, como ya habíamos visto. Tampoco se aprecian grandes diferencias entre los dos grupos del estudio de la gripe, pero una vez más se puede ver cómo en el grupo de los hombres hay dos muestras que tienen una diversidad de especies mucho más alta que el resto.

4.4 NMDS

Como ya se ha explicado, el NMDS es una técnica que nos ayuda a resumir las diferencias entre muestras y a encontrar grupos homogéneos. A continuación se muestran los resultados obtenidos con cada tipo de análisis.

4.4.1 Mothur

En este caso, Mothur nos generaba un archivo de salida una vez aplicado el análisis NMDS que consiste en una tabla de 3 columnas en la que la primera columna corresponde al nombre de las muestras y la segunda y tercera columna corresponden con las coordenadas de posición de cada muestra en los ejes X e Y. Este archivo se procesó con un script de R y se generaron las gráficas de la figura suplementaria 7.

Como puede observarse en las gráficas, las muestras están uniformemente distribuidas y por tanto no hay una disimilitud evidente entre los diferentes grupos de estudio en ninguno de los dos estudios.

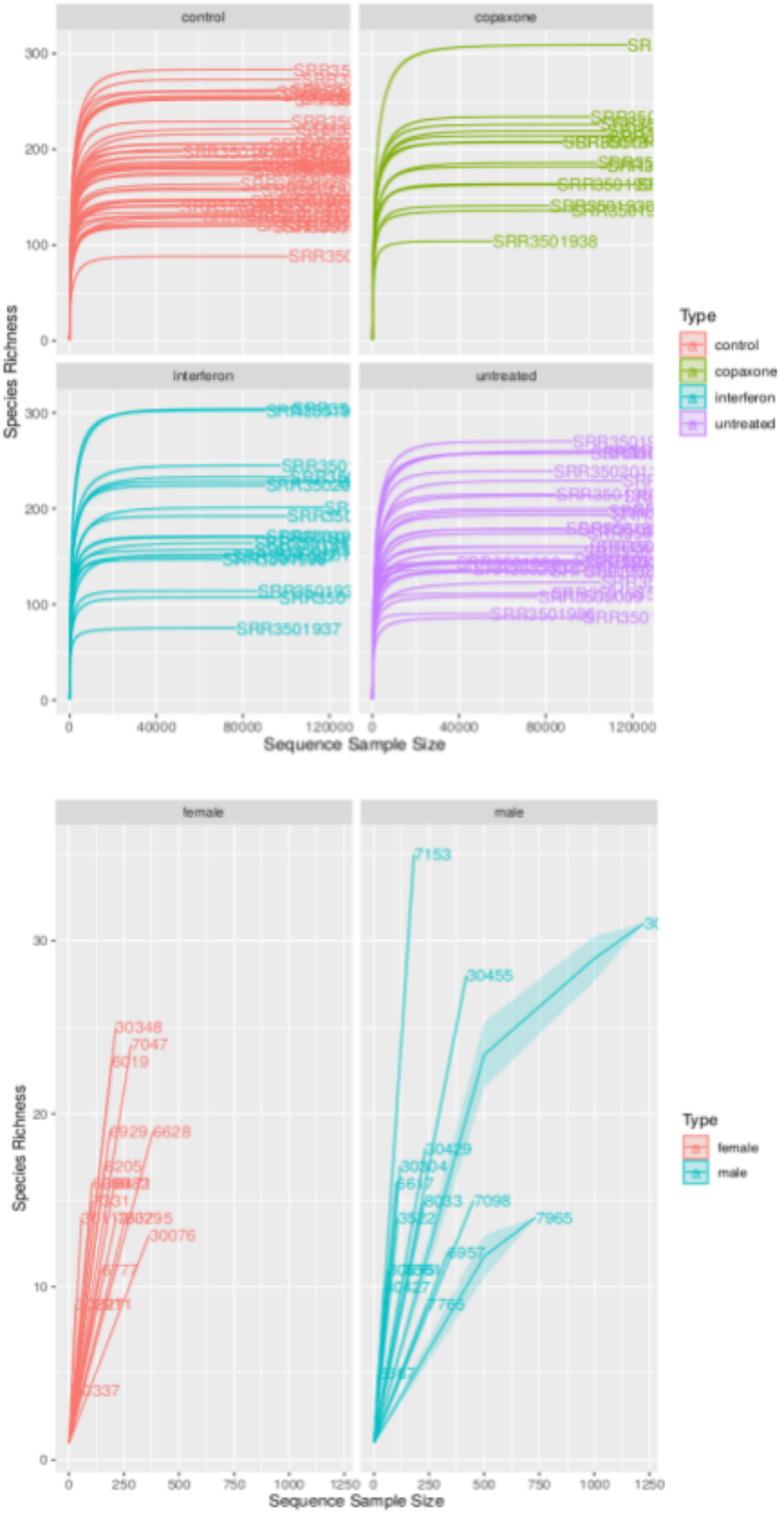

Figura 9. Curvas de rarefacción. Representan la variación de la diversidad entre muestras y grupos de estudio conforme aumenta la profundidad de secuenciación en los estudios de MS (a) y gripe (b). Obtenidas con DADA2.

4.4.2 QIIME2

En este caso, con QIIME2, tras aplicar el cálculo de distancia basado en el método Bray-Curtis, generaba un *artefacto* con una matriz de distancias entre muestras. A partir de este artefacto se obtiene un archivo *tsv* que contiene esta matriz de distancias y al cual le pasamos un script de R para generar las gráficas NMDS (FS. 8).

Como puede apreciarse en las gráficas los resultados son los mismos a los obtenidos con Mothur. Hay una dispersión homogénea de las muestras indicando que no hay disimilitudes significativas entre los grupos de estudio.

4.4.3 DADA2

En este caso, al ser un paquete de R, no hubo que transformar los datos, sino que simplemente se utilizó el paquete *phyloseq* que crea un objeto tipo phyloseq y lo grafica directamente con una función que aplica el método de Bray-Curtis. Los resultados de este análisis se muestran en la figura 10.

De la misma forma que con los análisis anteriores, en este caso tampoco se aprecia una disimilitud significativa entre los grupos de estudio analizados.

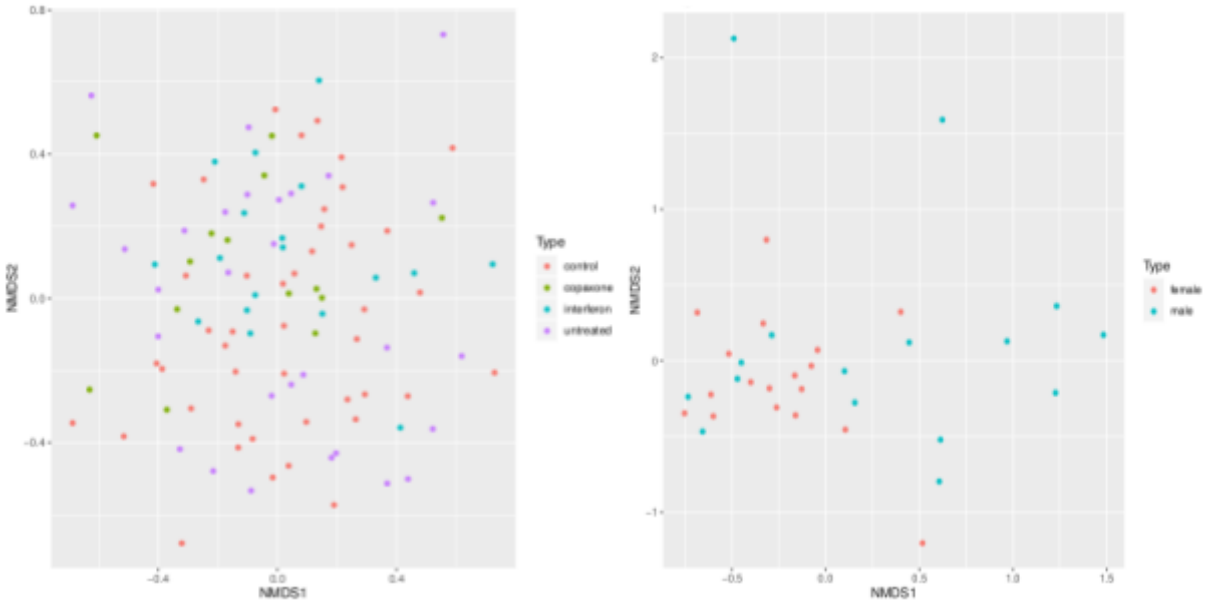

Figura 10. Representación de disimilitudes entre muestras a través del método NMDS de DADA2. En los estudios de MS (a) y gripe (b). Cada color corresponde a un grupo de estudio.

4.5 Abundancia diferencial

Hasta ahora hemos realizado análisis sobre la diversidad de especies en nuestros estudios. Sin embargo, también queremos determinar exactamente qué taxa está más representada en una condición versus otra y en qué medida. Los resultados están disponibles en las figuras suplementarias 9, 10 y 11.

En estas figuras aparece representado el log2FoldChange en cada grupo comparado. Esta medida es interesante ya que indica en qué grupo es más abundante cada ASV. En el estudio sobre MS se han hecho dos comparaciones: control vs MS sin tratamiento y control vs MS tratados. Si el log2FoldChange es positivo indicaría que el ASV es más abundante en el primer grupo comparado, mientras que si el valor es negativo indica que es más abundante en el segundo grupo.

En el caso de la gripe hemos realizado el contraste de mujeres vs hombres, por tanto, un log2FoldChange positivo indica que ese ASV es más abundante en mujeres, mientras que uno negativo indica que es más abundante en hombres.

Como complemento a estos cálculos de la abundancia diferencial, se han obtenido unas gráficas (Figura 11) que representan la abundancia de los primeros 15 ASV en cada una de las muestras de cada grupo de estudio. Estas representaciones pueden servir para ver algún patrón de similitud de abundancia de especies entre los grupos de estudio comparados.

En el estudio de la gripe parece que no hay patrones similares en la abundancia de estos ASV entre mujeres y hombres.

4.6 Visualización con Krona

Además de los cálculos mostrados, para los análisis con Mothur y QIIME2 se han generado unas pie charts con Krona en las que aparece una por cada muestra y se ve la proporción de microorganismos. Estos resultados pueden verse en las figuras suplementarias 12 y 13.

Como puede observarse en estas figuras, hay grandes diferencias en la proporción de especies mostradas por un pipeline y por otro. Siendo que con el de Mothur aparece una mayor variedad de especies.

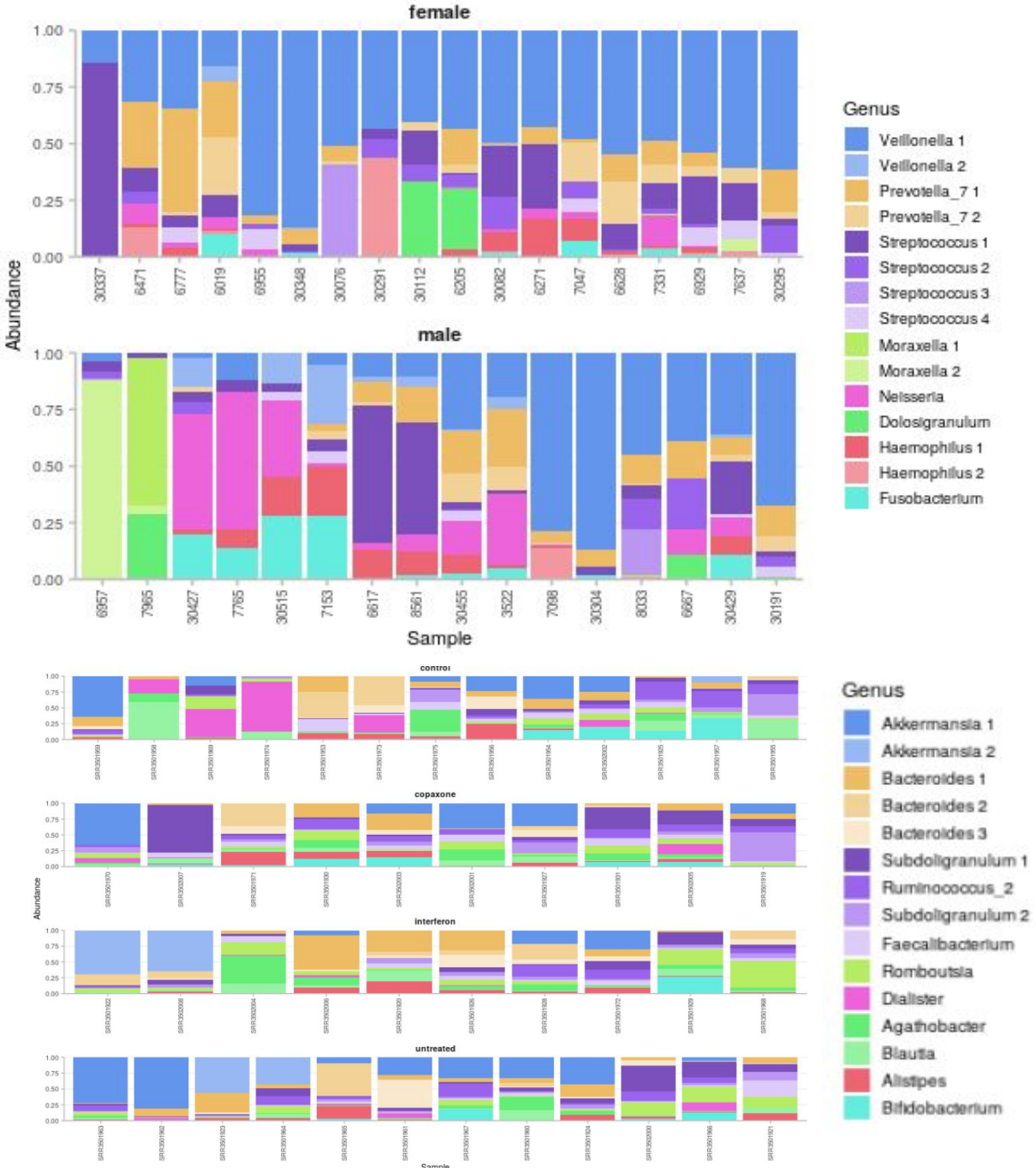

Figura 11. Representación gráfica de la abundancia de los primeros 15 ASV en cada una de las muestras en cada grupo de estudio.

5. Discusión

El microbioma es el conjunto de genomas de los microbios que viven en el cuerpo humano, junto con el del propio humano y también comprende las condiciones ambientales en las que estos se encuentran. Ya se ha demostrado su relevancia en la salud humana y por ello hay un continuado y creciente análisis de la microbiota en el estudio de diversas enfermedades.

Este análisis se lleva a cabo mediante la metagenómica y más comúnmente mediante la metataxonomía. Nosotros hemos desarrollado un estudio en el que comparamos tres métodos de análisis computacionales de secuenciación tipo *gene-marker* de microbiomas. Estos métodos de análisis son los más frecuentemente utilizados en los trabajos realizados durante los últimos años.

Como en el estudio de Olson, et al. (2020), en el que comparan precisamente las tres mismas herramientas que nosotros. En concreto, desarrollan un marco de trabajo que utiliza una mezcla de muestras medioambientales y que evalúa las características cualitativas y cuantitativas de las tablas de recuento. La evaluación cualitativa evalúa la presencia/ausencia de ciertas características, mientras que la evaluación cuantitativa evalúa la abundancia diferencial y relativa de las características comparando lo observado con lo esperado. Sus resultados muestran que en la evaluación cualitativa Mothur y QIIME presentan más falsos positivos, mientras que la de DADA2 presenta más falsos negativos. En el caso de la evaluación cuantitativa las tres herramientas presentan buenos resultados. Nosotros hemos realizado una evaluación parecida pero con un enfoque más biomédico, ya que nos centramos en la evaluación de las diferencias en variedad y cantidad de especies encontradas, en cada grupo de estudio y considerando cada una de las herramientas seleccionadas. Además, hemos valorado otras características de las herramientas como son su manejabilidad, si son intuitivas o la compatibilidad de los resultados para aplicar análisis estadísticos.

Nosotros hemos analizado dos conjunto de datos descargados de la base de datos GEO en los que se analiza caso vs control en MS y mujer vs hombre en personas con gripe. Estos datos han sido analizados por las tres herramientas ya descritas y sus resultados han sido comparados.

Lo primero que evaluamos en cualquier análisis metataxonómico, es la calidad de nuestras lecturas, ya que ésta va a tener un peso importante en los resultados finales. Observamos que ambos estudios presentan una calidad relativamente alta de las lecturas, proporcionando robustez en los resultados que se generen.

Después de realizar el filtrado y el clustering, se ha obtenido una serie de medidas con las que podemos ver la diversidad de especies de nuestras muestras y la abundancia de las mismas. En primer lugar teníamos los análisis de

diversidad alfa y rarefacción. Con estos veíamos la riqueza de especies, basándonos en la riqueza de especies raras (Chao1) y en la abundancia e igualdad de especies (Shannon), cuya variación se veía representada con las curvas de rarefacción conforme aumentaba la profundidad de secuenciación.

Los resultados generales de estos análisis llevaban a la misma conclusión: que en el estudio de MS había una diversidad mayor. De hecho podemos ver cómo más o menos los resultados coinciden entre los análisis hechos con las diferentes herramientas. Como puede verse que en el estudio de la gripe hay una diversidad media menor pero en el grupo de los hombres hay un par de individuos que tienen una diversidad mucho mayor que el resto. Por tanto aunque sean análisis diferentes y los resultados no sean exactamente iguales, lo que obtenemos al final se asemeja bastante.

Luego se realizaron los análisis de NMDS para saber el nivel de disimilitud entre los grupos comparados, y comprobamos que en ninguno de los dos estudios había diferencias estadísticamente significativas, ya que las muestras estaban dispersas de forma homogénea en ambos grupos.

Por último se llevó a cabo el análisis de abundancia diferencial en el caso del análisis con DADA2, comparando los grupos de estudio dos a dos para conocer dónde eran más abundantes los ASVs definidos. Este análisis no pudo realizarse con los outputs generados por las herramientas de Mothur y de QIIME2, debido a la incompatibilidad de formatos. Además de la abundancia diferencial, en el caso de DADA2 se obtuvieron unas gráficas que representaban la abundancia de los 15 ASVs más comunes en cada grupo de estudio. Por otra parte, se hicieron visualizaciones con Krona de los resultados de Mothur y de QIIME2 en los que podía verse representado en forma de pie chart las especies más representadas en cada muestra.

Si comparamos los resultados obtenidos con Mothur y QIIME2 a través de Krona y los obtenidos con DADA2 a través del paquete *phyloseq*, vemos que hay algunas especies que coinciden (menos del 50%). En el caso del estudio de gripe coinciden sólo las especies *Veillonella, Prevotella, Streptococcus,* y *Neisseria.* En el estudio de MS coinciden las especies *Bacteroides, Ruminococcus* y *Diallister.* Esto muestra claramente las grandes diferencias que pueden producirse en función del método de análisis empleado, además de cómo afectan también los parámetros usados en cada metodología.

A continuación compararemos los resultados obtenidos en este estudio con los obtenidos en los estudios originales.

En el estudio sobre la MS determinaron también la diversidad de especies (Shannon) y la PCoA. La PCoA es un método de análisis diferente al NMDS pero en el cual se pretende calcular lo mismo. Las figuras 12 y 13 muestran las

gráficas de estos resultados, respectivamente. Como podemos observar, en el análisis de la diversidad alfa, no se aprecian diferencias significativas entre los tres grupos de comparaciones. En cada caso hay una diversidad de especies similar entre cada grupo comparado. En el análisis de PCoA no se produce separación clara entre los grupos comparados (caso vs control) igual que sucedía con nuestro análisis por NMDS. Además de estos análisis, en el estudio original encontraron, como resultado final, que los pacientes tratados tenían un aumento en los microorganismos *Prevotella y Sutterella*. Ellos sugieren que la reducción de estos microorganismos en los pacientes no tratados, puede normalizar algunos cambios relacionados con la MS en la microbiota. Sería necesario realizar más comparaciones, principalmente con respecto a la abundancia relativa de las especies en cada grupo de estudio, para ver si finalmente llegáramos a obtener los mismos resultados que en el estudio original.

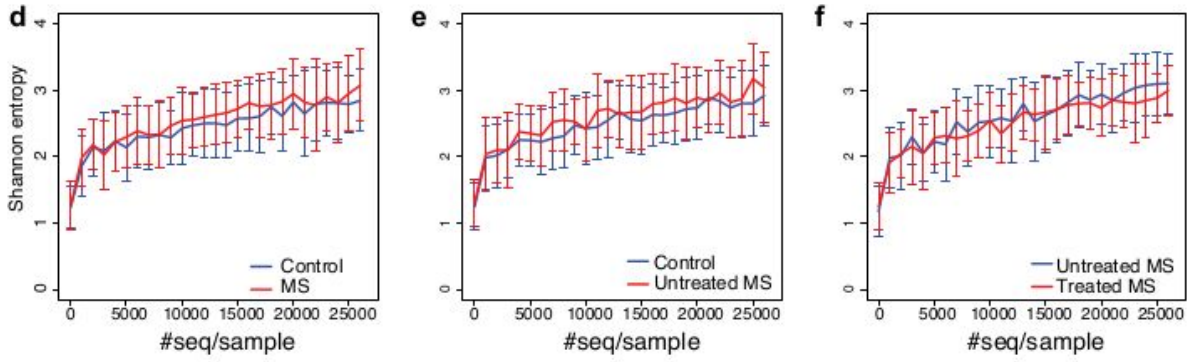

Figura 12. Análisis de diversidad alfa mediante el método de Shannon del estudio original del MS. En las gráficas aparece la comparación entre Controles vs Casos (d), Controles vs casos sin tratar (e) y Casos sin tratar vs Casos tratados (f) (Jangi, 2016).

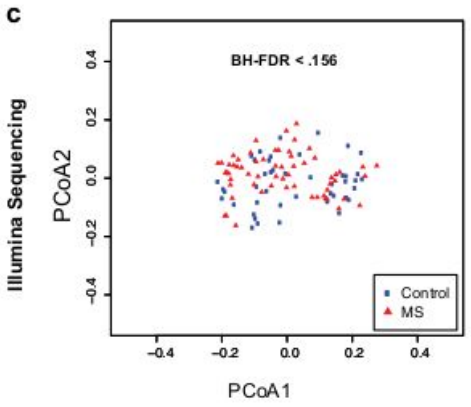

Figura 13. Análisis de PCoA del estudio original de MS. Se comparan Casos vs Controles (Jangi, 2016).

En el estudio sobre la gripe sólo muestran los resultados de la abundancia relativa presentados en la figura 14. En este caso comparan dos grupos generados por la diferencia de expresión de algunos genes, al contrario que nosotros, que hemos realizado la comparación por sexos. En el estudio, el grupo 1 contiene los individuos que poseen genes con resistencia a un medicamento. Esta figura muestra especies diferencialmente expresadas entre grupos como *Corynebacterium, Neisseria* y *Haemophilus*.

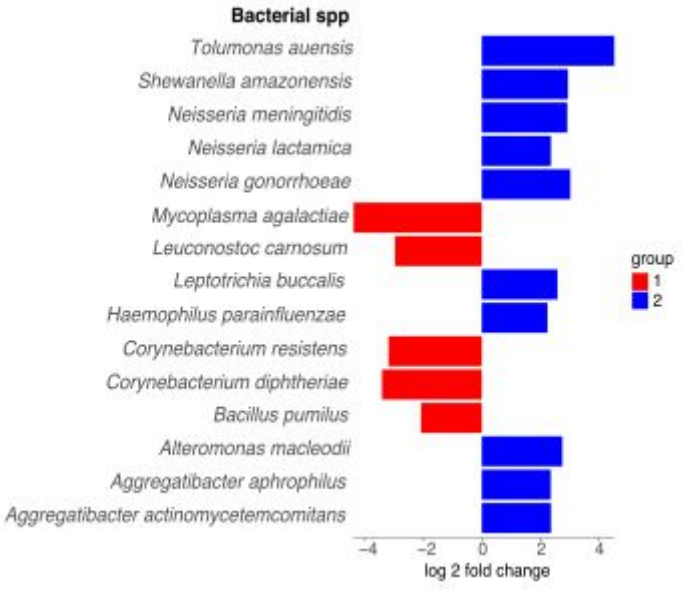

Figura 14. Representación de la abundancia relativa del estudio original de gripe. Aparecen las especies diferencialmente expresadas de los grupos comparados (Zhang, 2020).

En los resultados obtenidos por nuestro análisis con respecto a la abundancia relativa, aparece una mayor cantidad de especies expresadas diferencialmente entre ambos grupos y ninguna de ellas coincide con las expresadas diferencialmente entre los grupos del estudio original. Esto se debe a que los grupos comparados son diferentes y por tanto es esperable que no hay una coincidencia de resultados. En el estudio original concluyen que en la microbiota del tracto respiratorio, los genes resistentes a antibióticos se ven afectados indirectamente por la respuesta del huésped a la gripe. Puesto que nosotros hemos realizado un análisis con una aproximación diferente, no podemos comparar los resultados pero es interesante ya que hemos generado una nueva dimensión de información sobre estos mismos grupos de datos.

Como ya se ha mencionado al principio del trabajo, pese a que las tres herramientas comparadas son las más comúnmente utilizadas, existen otras herramientas de análisis metataxonómico que también han sido comparadas, junto con las nuestras, en otros estudios de benchmarking.

Como por ejemplo, el estudio de López-García, et al. (2018), en el que realizaron una comparación entre dos de estas herramientas en amplicones del gen 16S ARNr del rumen de vacas. Concretamente compararon Mothur y QIIME2. Además, utilizaron dos bases de datos de referencia con ambas herramientas: GreenGenes (GG, versión de mayo 2013) (DeSantis, 2006) y SILVA (versión 132). Lo que hallaron en este estudio es que sí que hay diferencias significativas en el análisis de los datos utilizando diferentes herramientas, pero estas diferencias dependían principalmente de la base de datos seleccionada.

En el estudio de Almeida, et al. (2018) comparan las herramientas MAPseq, Mothur, QIIME y QIIME2 utilizando conjunto de datos sintéticos simulados compuestos por los géneros más abundantes encontrados en el intestino humano, el océano y el suelo. En este estudio evalúan su precisión alineando los datos con dos bases de datos diferentes y secuenciando diferentes regiones hipervariables del gen 16S ARNr. Como conclusiones finales recomiendan el uso de QIIME2 o de MAPseq para una evaluación genética del ARNr.

En otro estudio, realizado por Prodan, et al. (2020), comparan seis herramientas de análisis metagenómico, tres de ellas agrupan en OTUs (Mothur, QIIME y USEARCH-UPARSE) y otras tres que agrupan en ASVs (QIIME2, DADA2 y USEARCH-UNOISE3). El estudio lo realizan sobre una Mock community y sobre un conjunto de datos de muestras fecales. Se centran en comparar la sensibilidad, la especificidad y el grado de consenso de los diferentes outputs. En sus resultados hallaron que las herramientas que agrupan por ASV tienen una mayor especificidad y que de éstas, DADA2 es la que presenta una mayor sensibilidad.

Los resultados de estos estudios, junto con los nuestros, muestran la gran variedad de opciones que hay a la hora de analizar el gen 16S ARNr de un grupo de datos de microbioma. Además, son un reflejo de que cada paso realizado en el análisis y cada filtro aplicado, va influyendo en gran medida en el resultado final que obtengamos. Por ello,a la hora de realizar cualquier análisis, es preciso conocer bien nuestros datos y sus características, además de conocer la mecánica de cada paso. Ya que cada herramienta presenta algunas características específicas, siendo importante que tengamos claro lo estrictos que queremos ser en el filtrado y *trimming*, para poder llegar a obtener los mejores resultados. Como observamos en los estudios ya nombrados, también es importante la base de datos seleccionada para realizar el alineamiento y que esa base de datos esté lo más actualizada posible.

Con todas estas consideraciones, debemos ser capaces de seleccionar la herramienta que mejor se ajuste al análisis que queremos hacer. Para ello, hay una serie de elementos que debemos tener en cuenta a la hora de elegir (Lindgreen, 2016). Entre otros, el *tiempo de análisis* es importante, ya que el tiempo de ejecución puede ser un cuello de botella significativo. La *facilidad de uso,* ya que cuanto más fácil sea de comprender y utilizar, más probable es que el usuario optimice su uso. La *información proporcionada*, ya que cuanta más información proporcione la herramienta sin tener que recurrir a otras, menor serán los posibles errores añadidos por el manejo y transformación de outputs. La *disponibilidad*, una herramienta de acceso libre estará al alcance de una mayor cantidad de personas y probablemente haya más información. Y por último también es importante la disposición de manuales detallados de uso

Teniendo en cuenta todos estos elementos que hacen que la herramienta que los posea sea considerada mejor, y analizando las comparaciones, el manejo y los resultados obtenidos por las tres herramientas utilizadas en este estudio, llegamos a la conclusión de que la mejor herramienta para hacer un análisis metataxonómico es DADA2. Esta herramienta cumple con todos los elementos ya que es rápida, consume menos recursos computacionales, es fácil de manejar y comprender. Tiene un manual muy intuitivo y claro. Al ser un paquete de R es la herramienta que tiene más facilidad para proporcionar una mayor cantidad de información gracias a la compatibilidad directa de formatos. Además es una herramienta gratuita y sobre la cual hay una gran cantidad de información de otros usuarios en internet.

En la tabla 1 podemos ver un resumen de los diferentes análisis que pueden obtenerse a partir de los outputs generados por cada una de las herramientas explicadas en este trabajo.

	mothur	**QIIME2**	**DADA2**
Diversidad alfa	Sí	Sí	Sí
Rarefacción	Sí	Sí	Sí
NMDS	Sí	Sí	Sí
Abundancia Diferencial	No	No	Sí
Krona	Sí	Sí	No

Tabla 1. Resumen de los resultados que pueden obtenerse con cada pipeline.

6. Conclusiones

1. Existen diversos repositorios de datos de microorganismos y sólo unos pocos con datos de microbioma. Es necesario el incremento de datos en estos recursos, posibilitando su uso accesible y compartido, y favoreciendo la relación de estudios in silico, así como la comparativa con otros estudios experimentales propios.
2. Hay una gran cantidad de herramientas y combinaciones orientadas al análisis de los datos procedentes del microbioma. Conocer el funcionamiento de las mismas, las diferencias entre cada una de ellas y su adecuación a los grupos de datos, es esencial para responder a las preguntas de investigación que se planteen.
3. La herramienta de análisis metataxonómico DADA2 es una de las herramientas más rápidas, cómodas y fáciles de utilizar. Debido a que es un paquete de R, se puede obtener de forma sencilla cualquier tipo de análisis que queramos realizar mediante el uso de otros paquetes gracias a la compatibilidad de formatos.
4. El benchmarking realizado en este trabajo ha sido de gran ayuda para conocer el manejo y análisis detallado de los datos de microbioma. Además, ha servido para poder desarrollar varios pipelines diferentes y poder elegir de forma segura el que queremos utilizar en el futuro para analizar los datos del proyecto “Share your Brain”.

Bibliografía

1. Alarcón Cavero T, D'Auria G, Delgado Palacio S, Del Campo Moreno, R, Ferrer Martínez, M. Microbiota. 59. Del Campo Moreno R (coordinadora). Procedimientos en Microbiología Clínica. Cercenado Mansilla E, Cantón Moreno R (editores). Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica (SEIMC). 2016.
2. Almeida, A., Mitchell, A. L., Tarkowska, A., & Finn, R. D. (2018). Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. *GigaScience*, *7*(5), giy054.
3. Barko, P. C., McMichael, M. A., Swanson, K. S., & Williams, D. A. (2018). The gastrointestinal microbiome: a review. *Journal of veterinary internal medicine*, *32*(1), 9-25.
4. Barres, BA. The mystery and magic of glia: a perspective on their roles in health and disease. Neuron 60, 430–440 (2008).
5. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, and Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37: 852–857.
6. Borre, Y. E., O'Keeffe, G. W., Clarke, G., Stanton, C., Dinan, T. G., and Cryan, J.F. (2014). Microbiota and neurodevelopmental windows: Implications for brain disorders. Trends Mol. Med. 20, 509–518.
7. Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nature methods*, *13*(7), 581.
8. Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME journal*, *11*(12), 2639.
9. Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7, 335–336. doi: 10.1038/nmeth. f.303.
10. Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of statistics*, 265-270.

11. Chung, H., Pamp, S. J., Hill, J. A., Surana, N. K., Edelman, S. M., Troy, E. B., Reading, N. C., Villablanca, E. J., Wang, S., Mora, J. R., et al. (2012). Gut immune maturation depends on colonization with a host-specific microbiota. Cell 149, 1578–1593.
12. Clarke, G., Stilling, R. M., Kennedy, P. J., Stanton, C., Cryan, J. F., & Dinan, T. G. (2014). Minireview: gut microbiota: the neglected endocrine organ. *Molecular endocrinology*, *28*(8), 1221-1238.
13. Cole, J. R., Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske, and J. M. Tiedje. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis *Nucl. Acids Res.* 42(Database issue):D633-D642; doi: 10.1093/nar/gkt1244.
14. Collins, S. M., Surette, M., & Bercik, P. (2012). The interplay between the intestinal microbiota and the brain. *Nature Reviews Microbiology*, *10*(11), 735-742.
15. del Campo-Moreno, R., Alarcón-Cavero, T., D'Auria, G., Delgado-Palacio, S., & Ferrer-Martínez, M. (2018). Microbiota en la salud humana: técnicas de caracterización y transferencia. *Enfermedades Infecciosas y Microbiología Clínica*, *36*(4), 241-245.
16. DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., ... & Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, *72*(7), 5069-5072.
17. Dietert, R. R., & Silbergeld, E. K. (2015). Biomarkers for the 21st century: listening to the microbiome. *Toxicological Sciences*, *144*(2), 208-216.
18. Dong, Y. & Benveniste, EN. Immune function of astrocytes. Glia 36, 180–190 (2001).
19. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002 Jan 1;30(1):207-10
20. Edwards, R., & Edwards, J. A. (2019). fastq-pair: efficient synchronization of paired-end fastq files. *BioRxiv*, 552885.
21. Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., & Sogin, M. L. (2013). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution*, *4*(12), 1111-1119.
22. Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H., & Sogin, M. L. (2015). Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *The ISME journal*, *9*(4), 968-979.
23. Escapa, I. F., Chen, T., Huang, Y., Gajare, P., Dewhirst, F. E., & Lemon, K. P. (2018). New insights into human nostril microbiome from the expanded Human

Oral Microbiome Database (eHOMD): a resource for the microbiome of the human aerodigestive tract. *Msystems*, *3*(6).
24. Escobar-Zepeda, A., Vera-Ponce de Leon, A., & Sanchez-Flores, A. (2015). The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. *Frontiers in genetics*, *6*, 348.
25. Foster JA, McVey Neufeld KA. Gut-brain axis: How the microbiome influences anxiety and depression. Trends Neurosci. 2013;36:305–12.
26. Francislon S Oliveira, John Brestelli, Shon Cade, Jie Zheng, John Iodice, Steve Fischer, Cristina Aurrecoechea, Jessica C Kissinger, Brian P Brunk, Christian J Stoeckert, Jr, Gabriel R Fernandes, David S Roos, Daniel P Beiting, MicrobiomeDB: a systems biology platform for integrating, mining and analyzing microbiome experiments, *Nucleic Acids Research*, Volume 46, Issue D1, 4 January 2018, Pages D684–D691.
27. Fung TC, Olson CA, Hsiao EY. Interactions between the microbiota, immune and nervous systems in health and disease. Nat Neurosci. 2017;20:145–55.
28. Gevers D, Knight R, Petrosino JF. et al. The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome. PLoS Biol. 2012;10(8):6–10.
29. Ghurye, J. S., CEPEDA-ESPINOZA, V., & POP, M. Metagenomic assembly: Overview, challenges and applications. 2016. *Yale Journal of Biology and Medicine. ISBN*, 1551-4056.
30. Handelsman J., Rondon M. R., Brady S. F., Clardy J., Goodman R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* 5 R245–R249. 10.1016/s1074-5521(98)90108-9.
31. Human Microbiome Project Consortium (2012). A framework for human microbiome research. *Nature*, *486*(7402), 215–221. https://doi.org/10.1038/nature11209
32. Janda, J. M., and Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J. Clin. Microbiol. 45, 2761–2764. doi: 10.1128/JCM.01228-07.
33. Jang, C., Chen, L., & Rabinowitz, J. D. (2018). Metabolomics and isotope tracing. *Cell*, *173*(4), 822-837.
34. Jangi, S., Gandhi, R., Cox, L. M., Li, N., Von Glehn, F., Yan, R., ... & Cook, S. (2016). Alterations of the human gut microbiome in multiple sclerosis. *Nature communications*, *7*(1), 1-11.
35. Korem, T., Zeevi, D., Suez, J., Weinberger, A., Avnit-Sagi, T., Pompan-Lotan, M., ... & Sirota-Madi, A. (2015). Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science*, *349*(6252), 1101-1106.
36. Li J, Jia H, Cai X. et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotech. 2014;32(8):834–841.

37. Liberati, A. et al. The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. PLOS Medicine. 2009; 6: 1-28.
38. Lindgreen, S., Adair, K. L., & Gardner, P. P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific reports*, *6*, 19233.
39. López-García, A., Pineda-Quiroga, C., Atxaerandio, R., Pérez, A., Hernández, I., García-Rodríguez, A., & González-Recio, O. (2018). Comparison of Mothur and QIIME for the analysis of rumen microbiota composition based on 16S rRNA amplicon sequences. *Frontiers in microbiology*, *9*, 3010.
40. Matcovitch-Natan O et al. Microglia development follows a stepwise program to regulate brain homeostasis. Science 353, aad8670 (2016).
41. Mardis, E. R. (2008). Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. 9, 387–402. doi: 10.1146/annurev.genom.9.081307. 164359.
42. McArdle AJ, Menikou, S. What is proteomics? Archives of Disease in Childhood - Education and Practice Published Online First: 02 April 2020. doi: 10.1136/archdischild-2019-317434.
43. McMurdie, P. J., & Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS one*, *8*(4), e61217.
44. Microbiome Database（MDB). db.cngb.org/microbiome
45. Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., ... & Sakharova, E. (2020). MGnify: the microbiome analysis resource in 2020. *Nucleic acids research*, *48*(D1), D570-D578.
46. Morozova, O., Hirst, M., & Marra, M. A. (2009). Applications of new sequencing technologies for transcriptome analysis. *Annual review of genomics and human genetics*, *10*, 135-151.
47. NRC (National Research Council) (1987). Biological markers in environmental health research. Environ. Health Perspect. 74, 3–9.
48. Ondov BD, Bergman NH, and Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011 Sep 30; 12(1):385.
49. Olson, N. D., Kumar, M. S., Li, S., Braccia, D. J., Hao, S., Timp, W., ... & Bravo, H. C. (2020). A framework for assessing 16S rRNA marker-gene survey data analysis methods using mixtures. *Microbiome*, *8*(1), 1-18.
50. Otero, L. L. (2016). *Características del microbioma gástrico e intestinal en relación al estado de Helicobacter pylori en una población pediátrica* (p. 1). Universidad Complutense de Madrid.
51. Pla, L. (2006). Biodiversidad: Inferencia basada en el índice de Shannon y la riqueza. *Interciencia*, *31*(8), 583-590.
52. Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A. H., Nieuwdorp, M., & Levin, E. (2020). Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *Plos one*, *15*(1), e0227434.

53. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucl. Acids Res. 41 (D1): D590-D596.*
54. Rossi, D. Astrocyte physiopathology: at the crossroads of intercellular networking, inflammation and cell death. Prog. Neurobiol 130, 86–120 (2015).
55. Rothhammer, V. et al. Type I interferons and microbial metabolites of tryptophan modulate astrocyte activity and central nervous system inflammation via the aryl hydrocarbon receptor. Nat. Med 22, 586–597 (2016).
56. Roumpeka, D. D., Wallace, R. J., Escalettes, F., Fotheringham, I., & Watson, M. (2017). A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. *Frontiers in genetics*, *8*, 23.
57. Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. 74, 5463–5467. doi: 10.1073/pnas.74.12.5463.
58. Schloss PD, Westcott SL (2011). Assessing and improving methods used in OTU-based approaches for 16S rRNA gene sequence analysis. Appl Environ Microbiol 77:3219.
59. Shreiner AB, Kao JY, Young VB. The gut microbiome in health and in disease. Curr Opin Gastroenterol 2015;31:69–75. 2.
60. Tan, G., Opitz, L., Schlapbach, R., & Rehrauer, H. (2019). Long fragments achieve lower base quality in Illumina paired-end sequencing. *Scientific reports*, *9*(1), 1-7.
61. Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. Microb Inform Exp. 2012;1:3.
62. Turnbaugh PJ, Ley RE, Hamady M, et al. The human microbiome project. Nature 2007;449:804–810.
63. Wang, W. L., Xu, S. Y., Ren, Z. G., Tao, L., Jiang, J. W., & Zheng, S. S. (2015). Application of metagenomics in the human gut microbiome. *World journal of gastroenterology: WJG*, *21*(3), 803.
64. Watson, JD. and Crick, FHC. A Structure for Deoxyribose Nucleic Acid. Nature. 1953; 171: 737-738.
65. Woese, C. R., and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U.S.A. 74, 5088–5090. doi: 10.1073/pnas.74.11.5088
66. Zhang, L., Forst, C. V., Gordon, A., Gussin, G., Geber, A. B., Fernandez, P. J., ... & Bonneau, R. (2020). Characterization of antibiotic resistance and host-microbiome interactions in the human upper respiratory tract during influenza infection. *Microbiome*, *8*(1), 1-12.

Anexo A – Tablas

	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	1	151	151	0	2	1
2.5%-tile:	1	252	252	0	4	287574
25%-tile:	1	253	253	0	4	2875737
Median:	1	253	253	0	4	5751474
75%-tile:	1	253	253	1	5	8627211
97.5%-tile:	1	254	254	4	6	11215374
Maximum:	1	502	502	249	243	11502947
Mean: 1	253	253	0	4		
# of Seqs:	11502947					

	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	1	250	250	0	3	1
2.5%-tile:	1	252	252	0	3	3383
25%-tile:	1	253	253	23	4	33822
Median:	1	254	254	34	4	67644
75%-tile:	1	256	256	44	4	101466
97.5%-tile:	1	258	258	61	6	131905
Maximum:	1	502	502	81	185	135287
Mean: 1	256	256	32	4		
# of Seqs:	135287					

Tabla Suplementaria 1. Resumen de las características de nuestras lecturas obtenidas con Mothur. En estas tablas se muestra un resumen de las características de las lecturas de los estudios de MS (arriba) y gripe (abajo) antes de aplicarles los filtros de calidad. Concretamente muestra el tamaño mínimo, máximo y medio de las lecturas, el número de bases ambiguas, el número de homopolímeros y el número total de secuencias.

Anexo B - Figuras

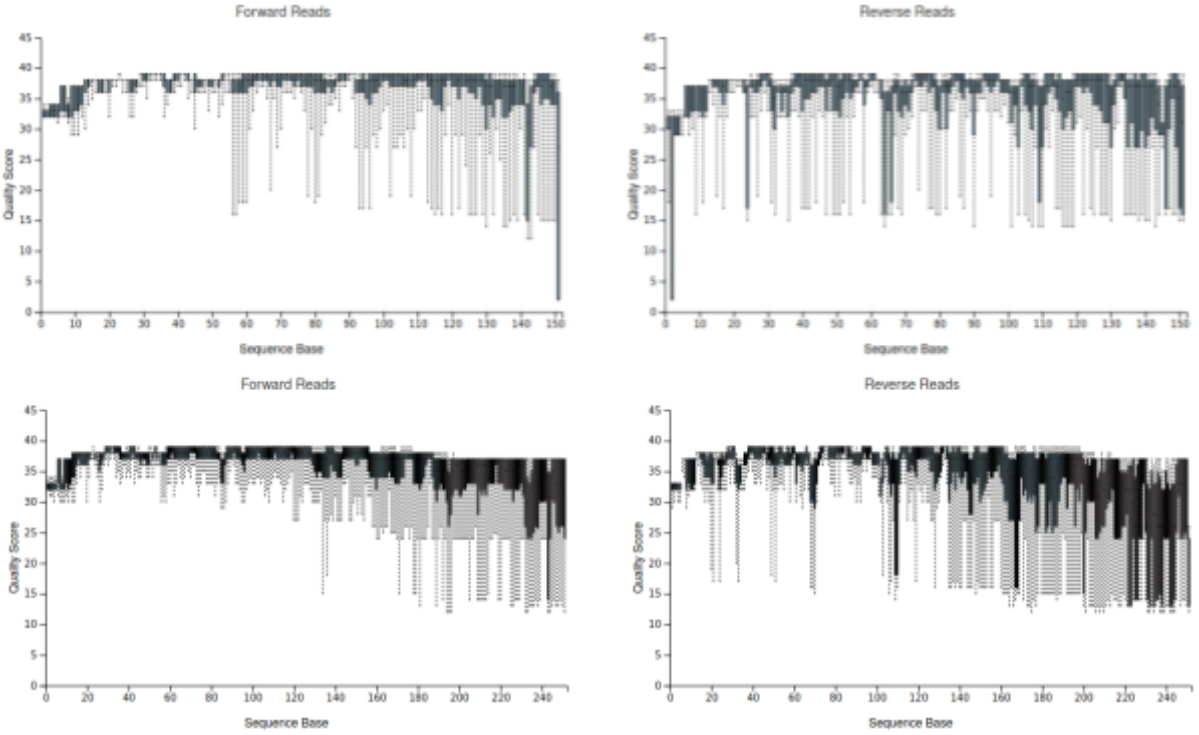

Figura Suplementaria 1. Análisis de calidad de cada base en las secuencias forward y reverse obtenido con QIIME2. Las gráficas de arriba representan las lecturas del estudio de MS y las de abajo del estudio de gripe.

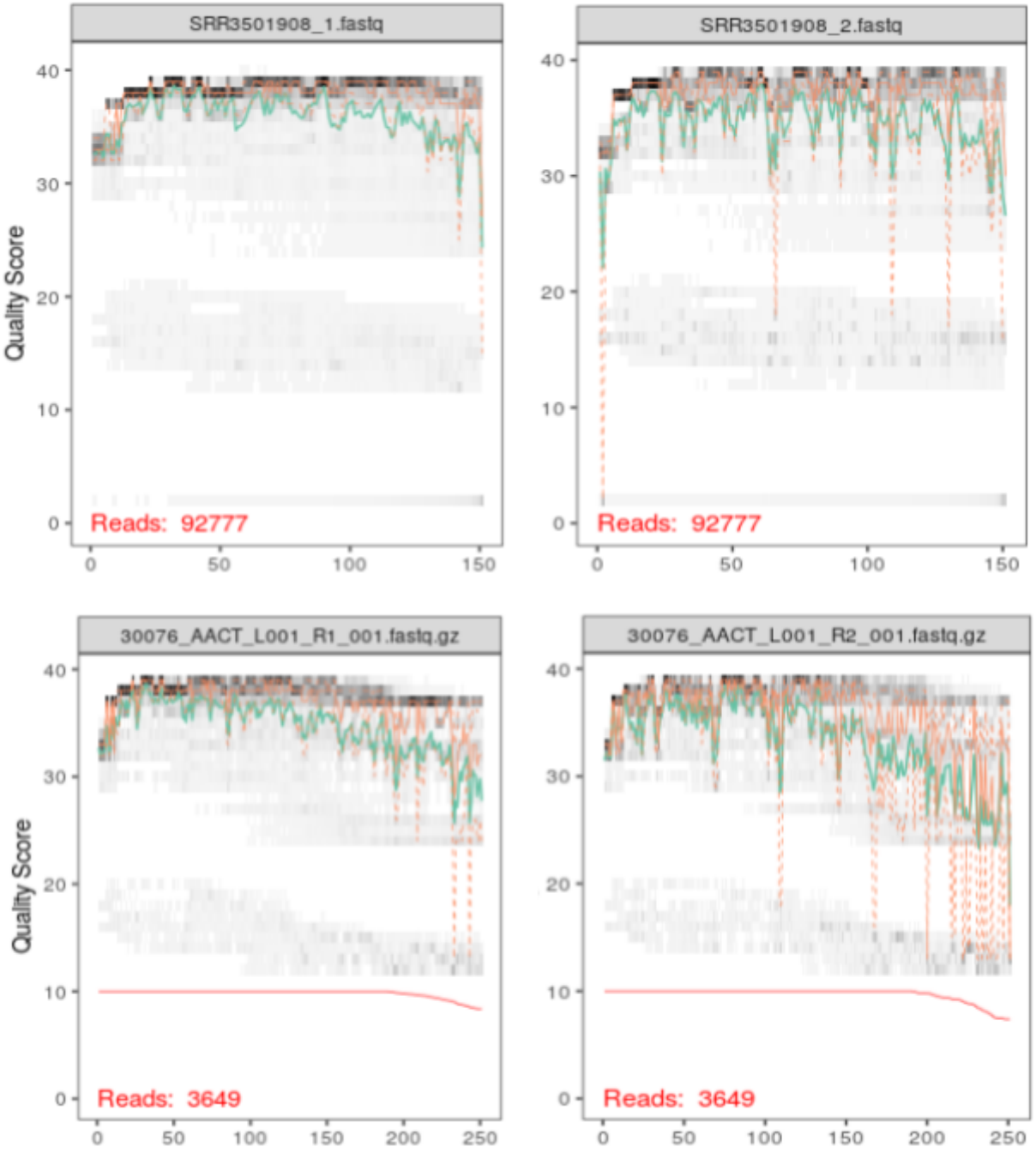

Figura Suplementaria 2. Análisis de calidad con DADA2 de las secuencias forward y reverse de MS (arriba) y de gripe (abajo). En escala de grises aparece un heatmap de la frecuencia de cada score de calidad para cada base. El score de calidad medio para cada posición se muestra con la línea verde y los cuartiles de la distribución del score de calidad con la línea naranja. La línea roja muestra la proporción escalada de las lecturas que se extienden al menos hasta esa posición.

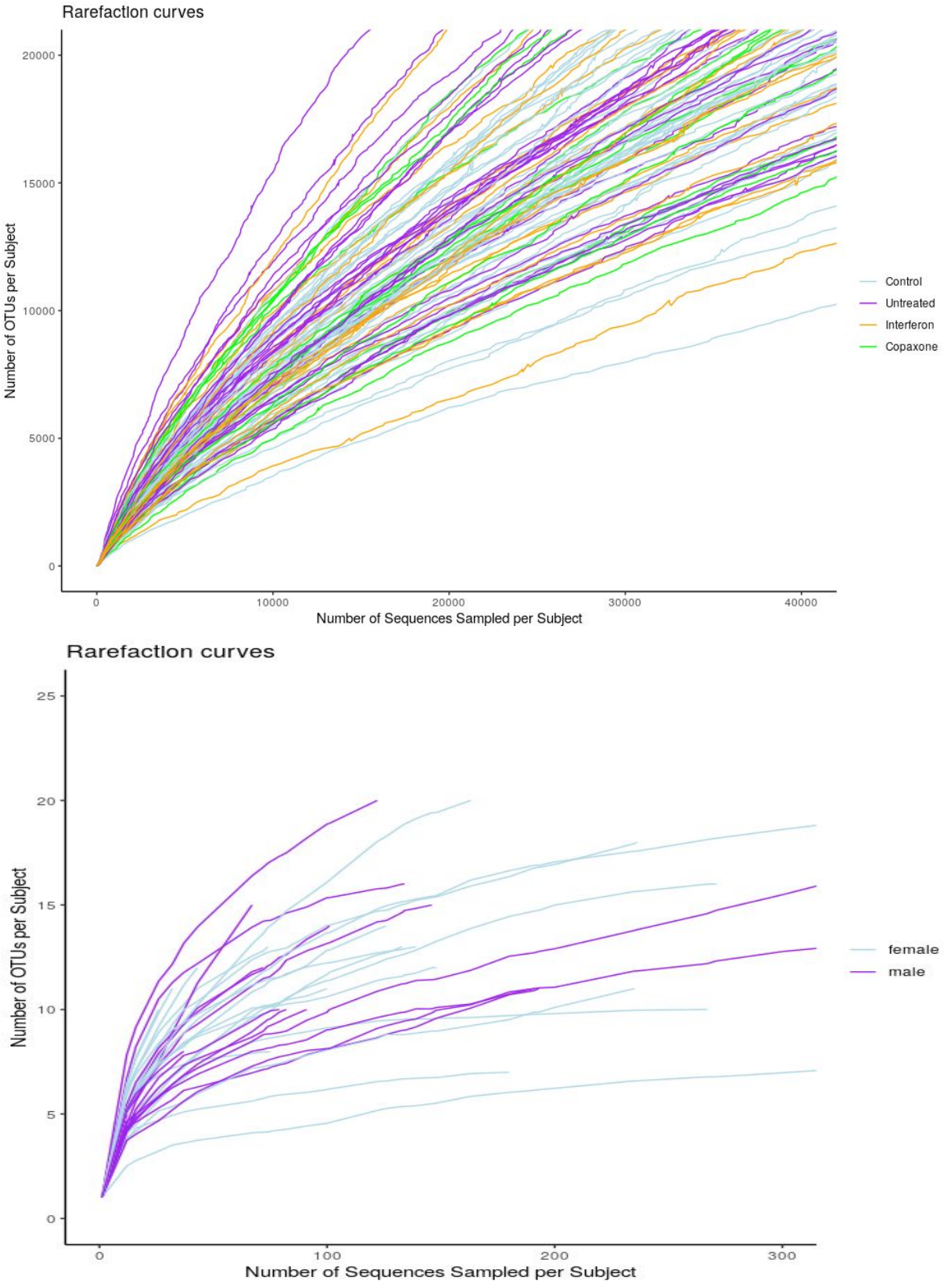

Figura suplementaria 3. Curvas de rarefacción del estudio de MS (arriba) y gripe (abajo) obtenidas a partir de Mothur. Cada línea representa una muestra del estudio.

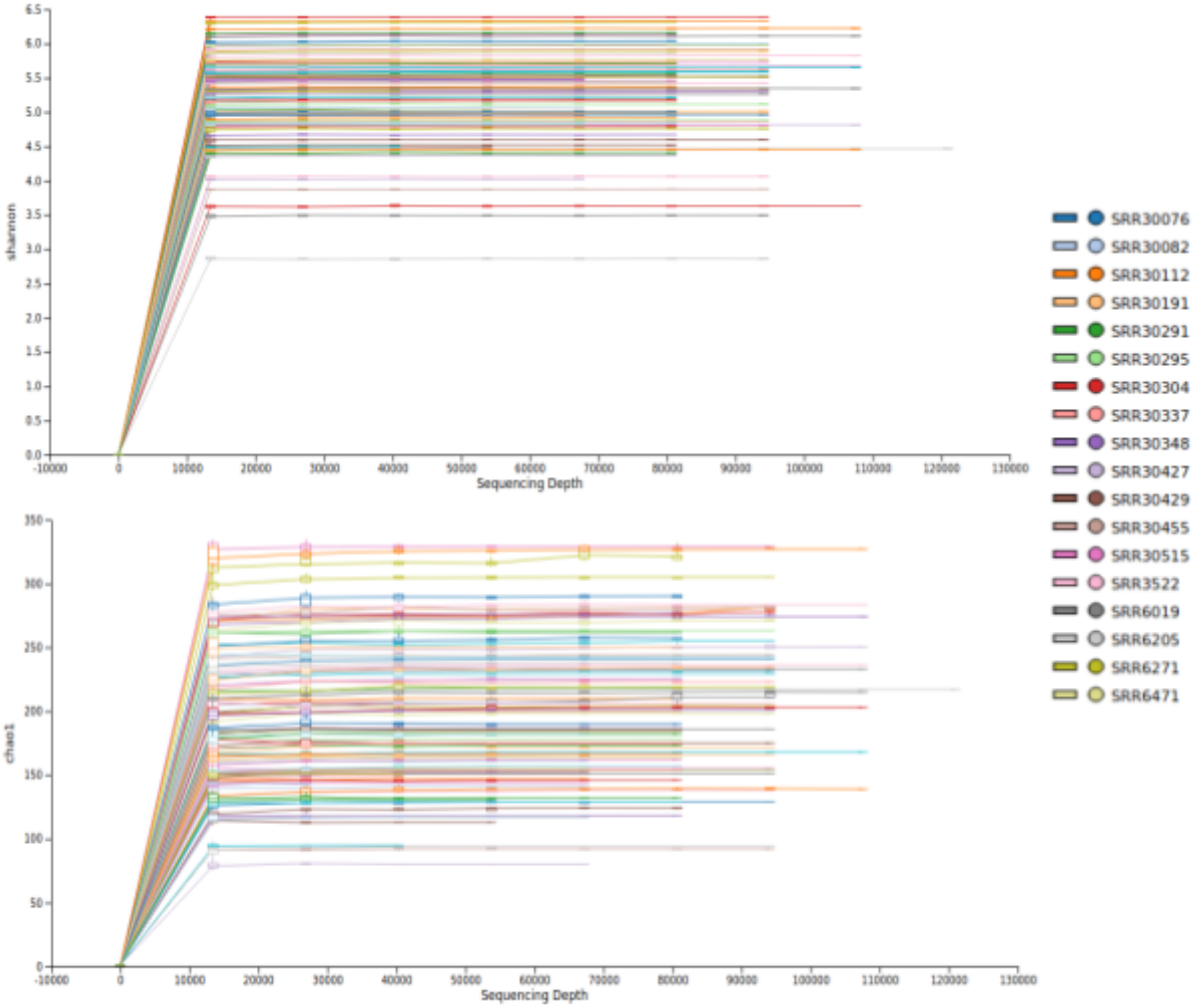

Figura Suplementaria 4. Curvas de rarefacción del estudio de MS aplicando las métricas de Shannon (arriba) y Chao1 (abajo). Cada línea representa una muestra del estudio. Se muestran los nombres de una pequeña parte de las muestras ya que el gran número de las mismas impide mostrarlas todas.

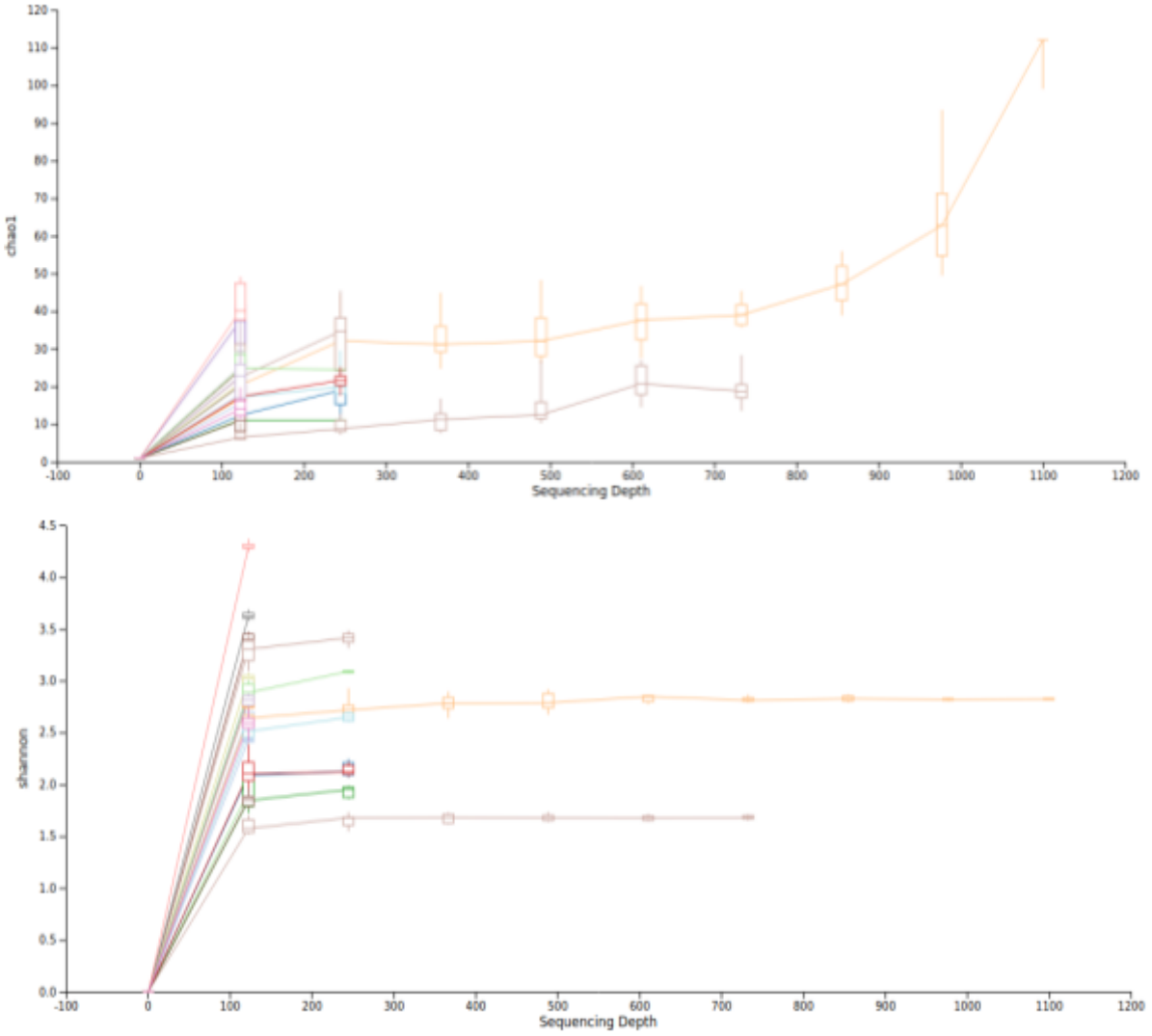

Figura Suplementaria 5. Curvas de rarefacción del estudio de gripe aplicando las métricas de Shannon (arriba) y chao1 (abajo).

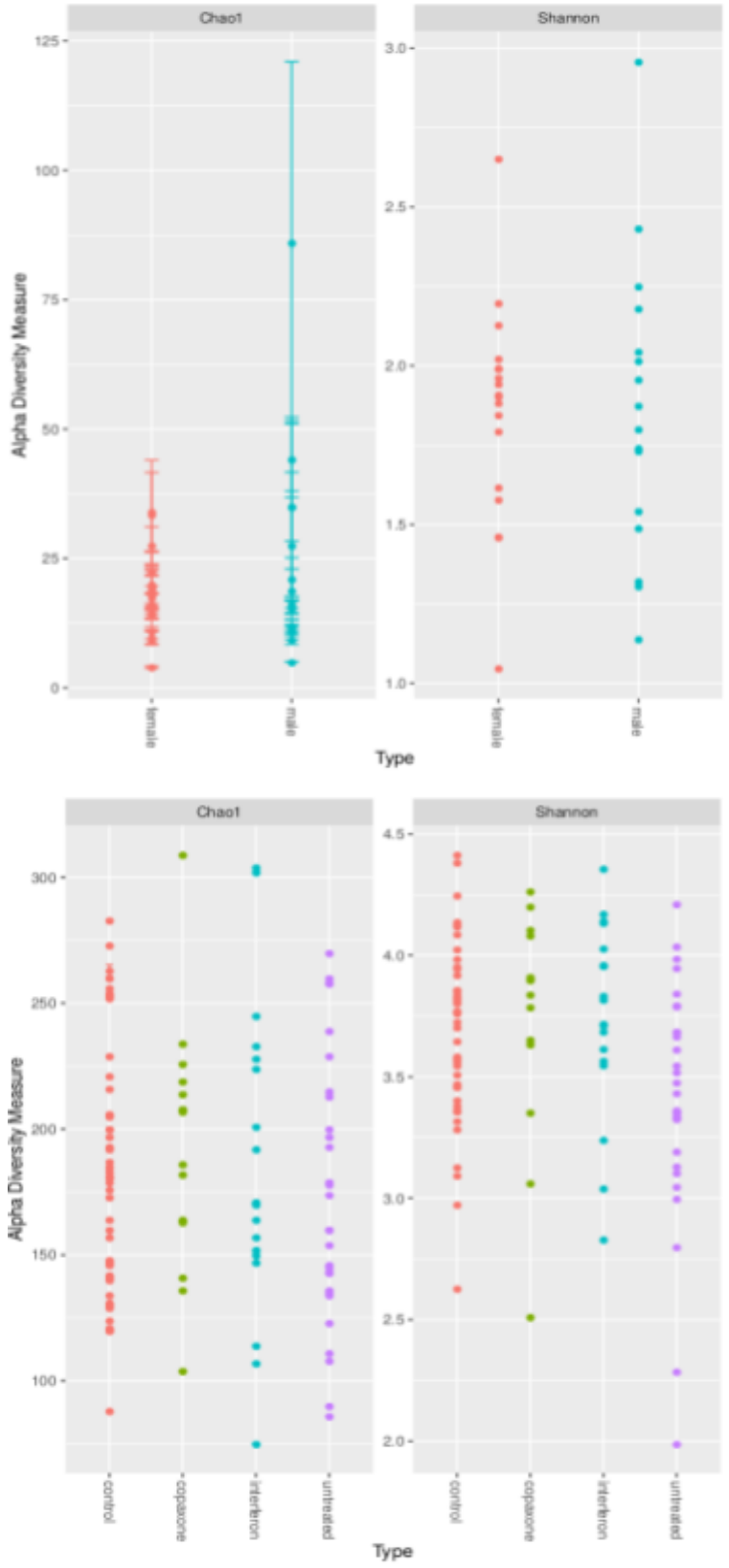

Figura Suplementaria 6. Representación de la alfa diversidad a través de los índices de Chao1 y Shannon en los estudios de MS (arriba) y gripe (abajo) con DADA2.. Cada color representa un grupo de estudio

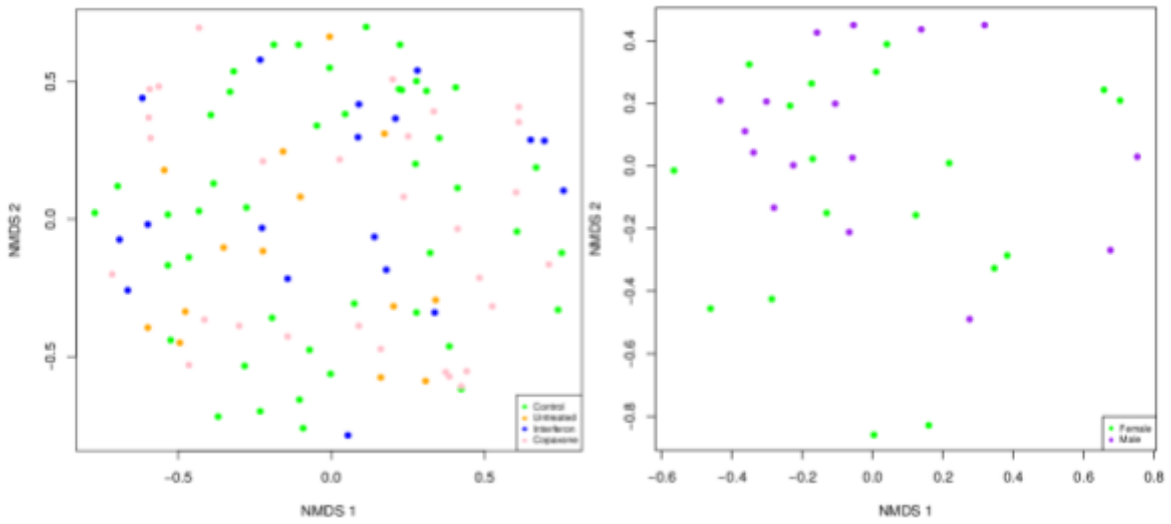

Figura Suplementaria 7. Representación de disimilitudes entre muestras a través del método NMDS de Mothur de los estudios de MS (izquierda) y gripe (derecha), obtenido con Mothur. Cada color corresponde a un grupo de estudio.

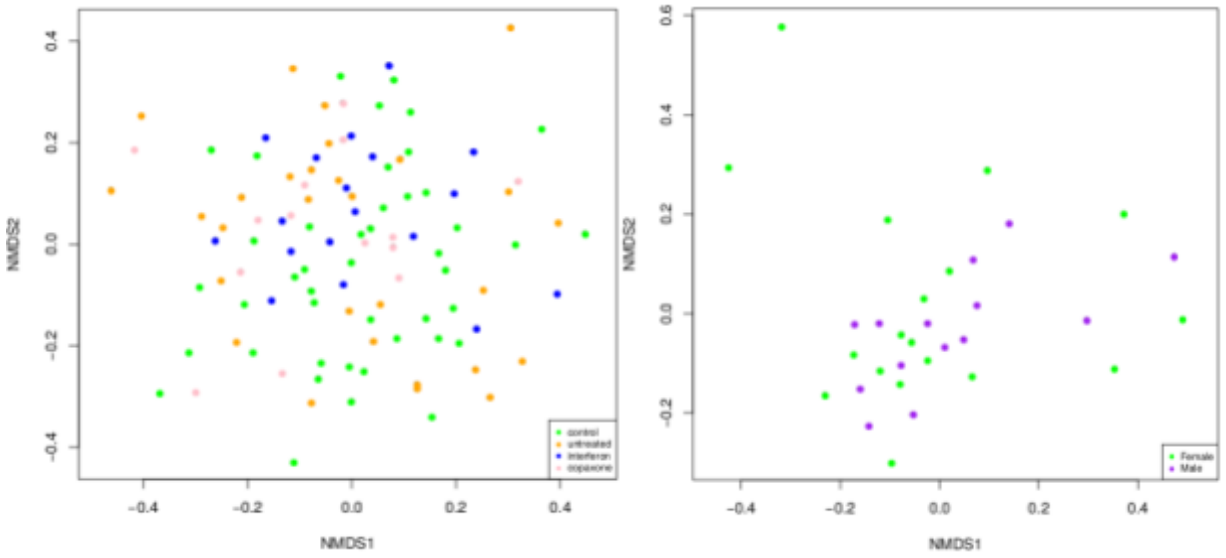

Figura Suplementaria 8. Representación de disimilitudes entre muestras a través del método NMDS de QIIME2 de los estudios de MS (izquierda) y gripe (derecha), obtenido con QIIME2. Cada color corresponde a un grupo de estudio.

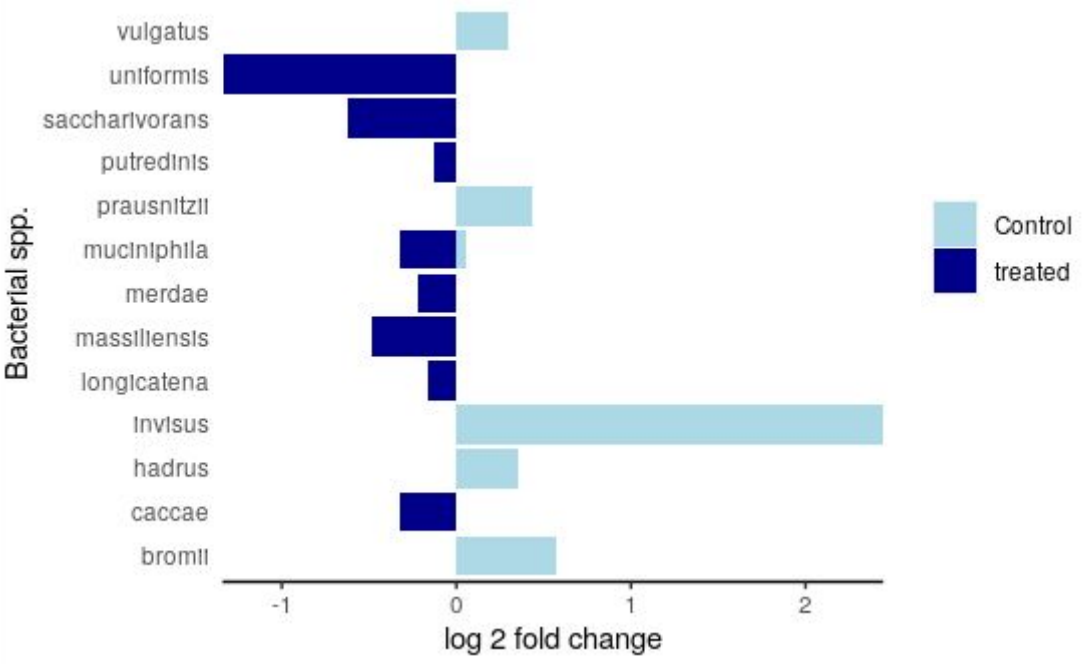

Figura Suplementaria 9. Abundancia relativa del estudio de MS en el que se compara el grupo control con el grupo de pacientes MS tratados.

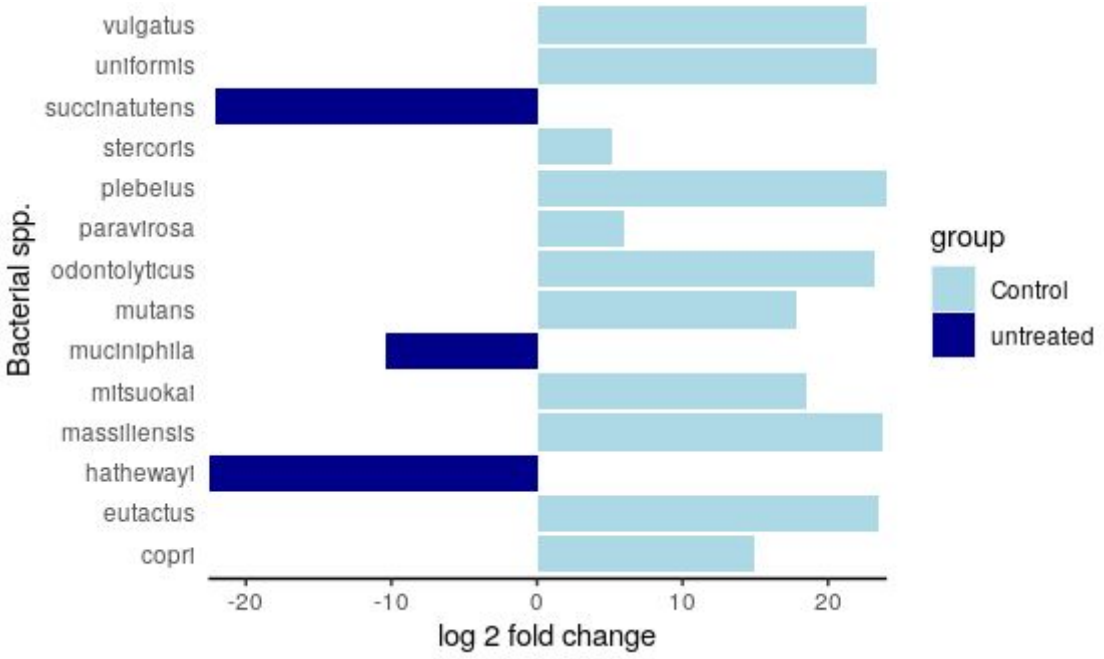

Figura Suplementaria 10. Abundancia relativa del estudio de MS en el que se compara el grupo control con el grupo de pacientes MS no tratados.

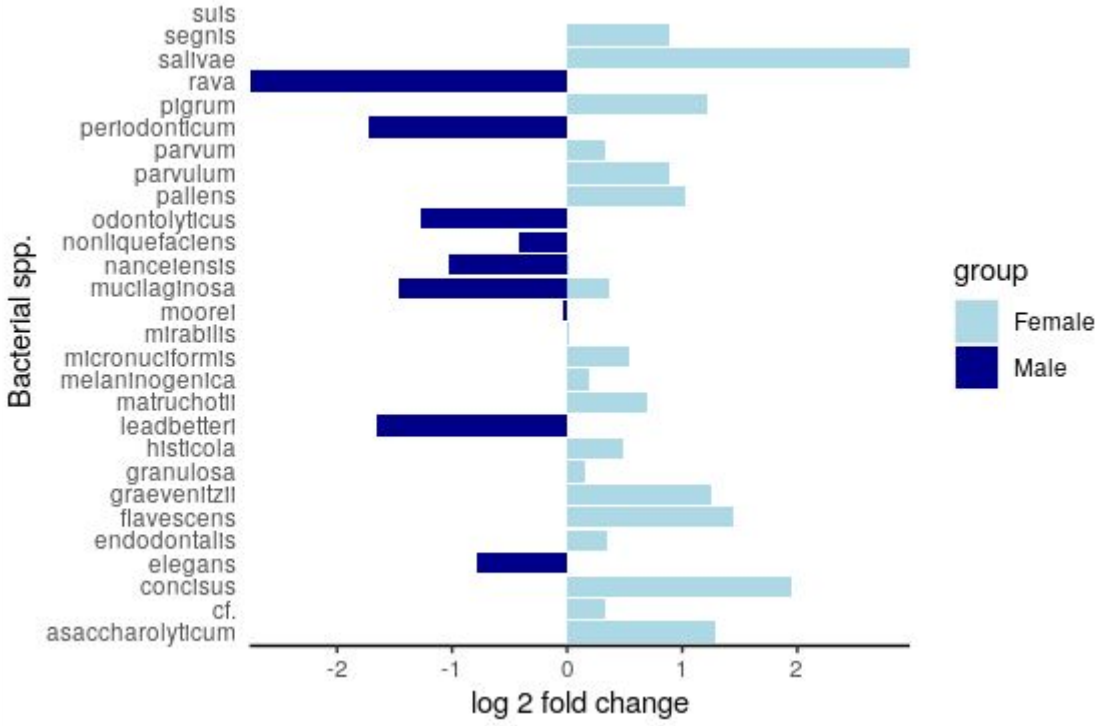

Figura Suplementaria 11. Abundancia relativa del estudio de gripe en el que se compara el grupo de mujeres con el grupo de hombres.

Figura suplementaria 12. Visualización con Krona de las especies halladas en una muestra del estudio de gripe. Arriba las halladas por QIIME2 y abajo las halladas por Mothur.

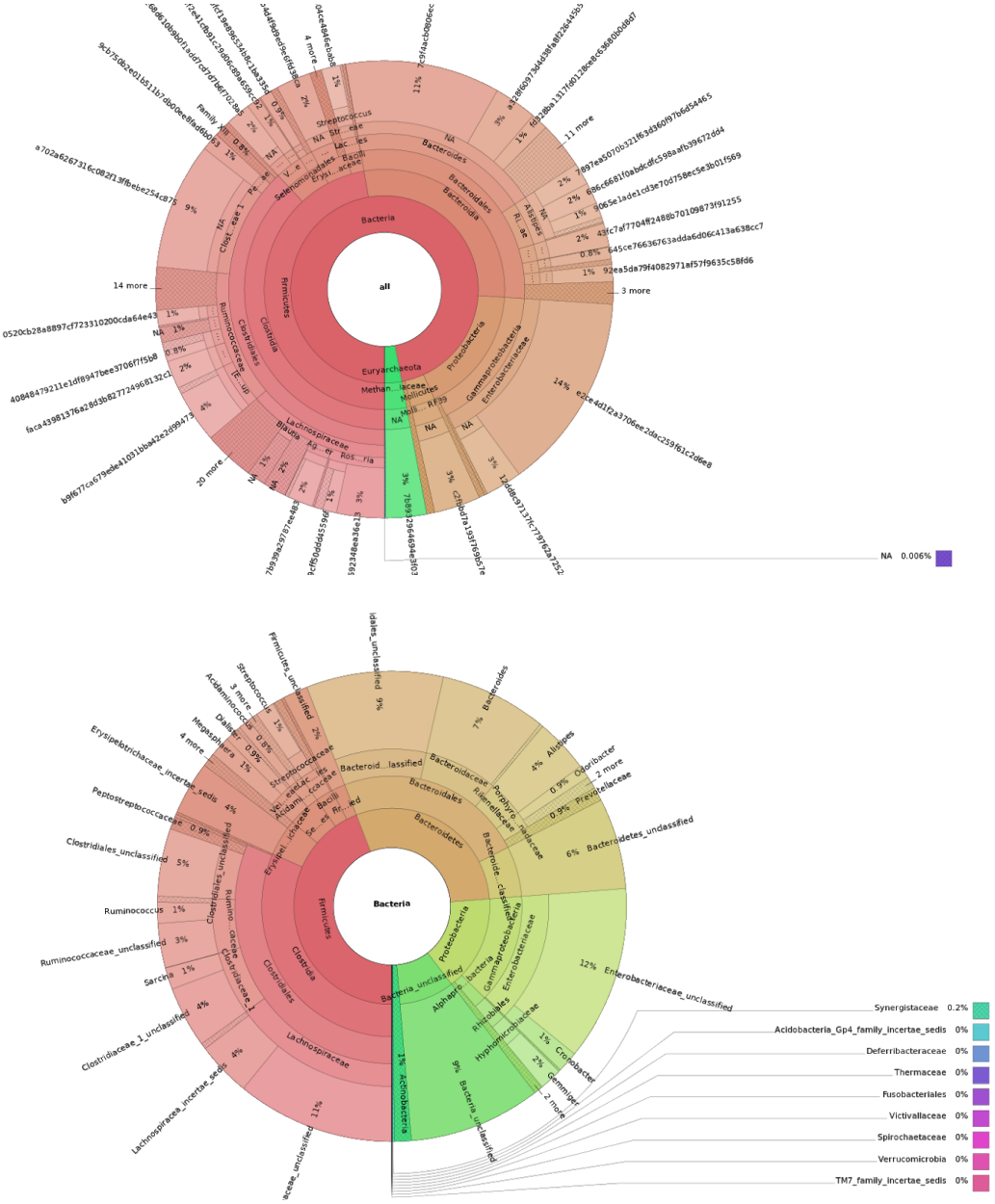

Figura suplementaria 13. Visualización con Krona de las especies halladas en una muestra del estudio de MS. Arribas las halladas por QIIME2 y abajo las halladas por Mothur.

Información Suplementaria

Todos los pipelines seguidos en este trabajo y los scripts empleados para transformar los archivos de salida y realizar los análisis correspondientes se encuentran en el repositorio de GitHub (https://github.com/mamufer2/metagenomics)

Los datos utilizados para hacer la comparación de estas herramientas se encuentran en GEO, bajo el número de acceso GSE126900 (BioProject PRJNA523620) para el estudio de la gripe y el número de acceso GSE81279 (BioProject PRJNA321051) para el estudio de la MS.