

# Estudio de la comunidad microbiana asociada al intestino de la corvina

**Edgar Bertomeu Primo**

Máster universitario en Bioinformática y Bioestadística

Análisis de datos ómicos

**Paloma María Pizarro Tobías**

**Antoni Pérez Navarro**

Diciembre 2021



Esta obra está sujeta a una licencia de  
Reconocimiento-NoComercial-SinObraDerivada  
[3.0 España de Creative Commons](https://creativecommons.org/licenses/by-nc-nd/3.0/es/)

## FICHA DEL TRABAJO FINAL

<b>Título del trabajo:</b>	<i>Estudio de la comunidad microbiana asociada al intestino de la corvina</i>
<b>Nombre del autor:</b>	<i>Edgar Bertomeu Primo</i>
<b>Nombre del consultor/a:</b>	<i>Paloma María Pizarro Tobías</i>
<b>Nombre del PRA:</b>	<i>Antoni Pérez Navarro</i>
<b>Fecha de entrega (mm/aaaa):</b>	12/2021
<b>Titulación:</b>	<i>Máster universitario en Bioinformática y Bioestadística UOC-UB</i>
<b>Área del Trabajo Final:</b>	<i>Análisis de datos ómicos</i>
<b>Idioma del trabajo:</b>	Español
<b>Número de créditos:</b>	15
<b>Palabras clave</b>	<i>microbioma, corvina, acuicultura</i>
<b>Resumen del Trabajo (máximo 250 palabras):</b> <i>Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.</i>	

Este estudio tiene como objetivo explorar la microbiota intestinal de la corvina (*Argyrosomus regius*) y evaluar el efecto de la dieta en la composición, abundancia y diversidad de esta.

En los últimos años, el estudio del microbioma ha cobrado relevancia debido a su influencia e implicación en diversas funciones del hospedador relacionadas con el metabolismo y la salud. Conocer la composición, abundancia y diversidad de la microbiota es necesario para la comprensión de su implicación en estos procesos. Entre los factores extrínsecos que modulan la microbiota, la dieta juega un papel muy importante. En este estudio se alimentaron 750 corvinas con cinco dietas basadas en fuentes de proteínas distintas [dieta control (DC), dieta de microalgas (DM), dieta de harina de insectos (DI), dieta de residuos del enlatado del atún (DP.L) y dieta mezcla de las tres anteriores (DMix)] y se analizaron seis muestras de cada grupo. Se secuenció un fragmento del gen *16S ARNr* a través de la tecnología MiSeq® de Illumina®, y se realizó un análisis bioinformático con la herramienta QIIME2, y un análisis bioestadístico con el paquete *microeco* de R.

El estudio reveló que *Mycoplasma*, *Enterobacter*, *Bacillus* y *Flavobacterium* fueron los géneros más abundantes. El grupo DM mostró un mayor índice de Shannon ( $H = 2,54 \pm 0,91$ ). Según las distancias de Bray-Curtis, las comparaciones entre grupos no fueron estadísticamente significativas.

Aunque los resultados entre las diferentes dietas no fueron significativos, este estudio describe por primera vez la composición de la comunidad microbiana del intestino de la corvina.

**Abstract (in English, 250 words or less):**

This study aims to explore the meagre (*Argyrosomus regius*) microbiota, in order to evaluate the effect of the diet on the composition, abundance and diversity of its microbiota.

In the least years, the study of the microbiome has gained relevance due to its influence and involvement in various host functions related to metabolism and health. Comprehend the composition, abundance, and diversity of the microbiota is necessary for understanding its involvement in these processes. Moreover, diet plays a very important role between the extrinsic factors that modulate the microbiota. In this study, 750 meagre were fed with five diets based on different protein sources [control diet (CD), microalgae diet (MD), insect meal diet (ID), protein and oil from water of tuna canning process (P&LD) and mixture diet (MixD)] and six samples from each group were analyzed. A fragment of the 16S rRNA gene was sequenced using the MiSeq® technology from Illumina®. The bioinformatic and biostatistical analysis were performed with QIIME2 and *microeco* package from R, respectively.

The study revealed that *Mycoplasma*, *Enterobacter*, *Bacillus* and *Flavobacterium* were the most abundant genus. The higher Shannon index was revealed by the MD group ( $H = 2,54 \pm 0,91$ ). According to Bray-Curtis distances, the comparisons between groups were not statistically significant.

Although the results between the different diets were not significant, this study describes the composition of the microbial community of the meagre gut for the first time ever.

## Índice

1. Resumen.....	2
2. Introducción.....	3
2.1. Contexto y justificación del trabajo.....	3
2.2. Objetivos del trabajo.....	6
2.3. Enfoque y método seguido.....	6
2.4. Planificación del trabajo.....	7
2.5. Breve resumen de productos obtenidos.....	8
2.6. Breve descripción de los otros capítulos de la memoria.....	8
3. Estado del arte.....	8
4. Metodología.....	9
4.1. Ética.....	9
4.2. Procedimiento experimental.....	9
4.3. Recogida de muestras, extracción de ADN y secuenciación.....	11
4.4. Bioinformática.....	12
4.5. Análisis bioestadístico.....	26
5. Resultados.....	32
5.1. Análisis taxonómico y abundancia diferencial.....	32
5.2. Índice de diversidad alfa.....	37
5.3. Índice de diversidad beta.....	39
6. Discusión.....	42
7. Conclusiones.....	44
7.1. Conclusiones.....	44
7.2. Líneas de futuro.....	44
7.3. Seguimiento de planificación.....	45
8. Glosario.....	46
9. Bibliografía.....	48
10. Anexos.....	52
11. Agradecimientos.....	68

## Lista de figuras

- Figura 1. Cambios de la industria en la formulación de piensos para acuicultura
- Figura 2. Corvina (*Argyrosomus regius*)
- Figura 3. Regiones del gen 16S ARNr
- Figura 4. Resumen del contaje de lecturas
- Figura 5. Calidad de secuenciación de las secuencias sentido y antisentido
- Figura 6. Visualización de QIIME2 de la tabla de características
- Figura 7. Visualización de QIIME2 de los detalles por muestra de la tabla de características
- Figura 8. Visualización de QIIME2 de las secuencias representativas
- Figura 9. Resumen estadístico con DADA2
- Figura 10. Visualización de QIIME2 de las curvas de rarefacción de la alfa diversidad
- Figura 11. Profundidad de muestreo en función del grupo experimental y número de muestras
- Figura 12. Flujo general de trabajo del paquete *microeco*
- Figura 13. Visualización sobre la composición taxonómica a nivel de familia
- Figura 14. Visualización sobre la composición taxonómica a nivel de género
- Figura 15. Asignación taxonómica con la abundancia relativa media a nivel de género
- Figura 16. Distribución de los valores de la alfa diversidad
- Figura 17. Diagrama de caja y bigotes de la alfa diversidad según la métrica de Shannon
- Figura 18. Gráfico PCoA de la beta diversidad con la métrica de Bray-Curtis
- Figura 19. Diagrama de caja y bigotes de la beta diversidad según las distancias medias de Bray-Curtis

## **Lista de tablas**

Tabla 1. Planificación del trabajo

Tabla 2. Formulación (%) y composición de proteínas y lípidos crudos de las dietas experimentales usadas para el estudio

Tabla 3. Estructura del archivo de metadatos

Tabla 4. Contaje de lecturas para cada una de las muestras en función de la taxonomía asignada

Tabla 5. Contrastes de abundancia diferencial a nivel de familia

Tabla 6. Contrastes de abundancia diferencial a nivel de género

Tabla 7. Kurskal-Wallis de la alfa diversidad con la métrica de Shannon

Tabla 8. PERMANOVA de la variable grupo

Tabla 9. PERMANOVA entre grupos

Tabla 10. PERMANOVA de la variable grupo y tanque

# 1. Resumen

Este estudio tiene como objetivo explorar la microbiota intestinal de la corvina (*Argyrosomus regius*) y evaluar el efecto de la dieta en la composición, abundancia y diversidad de esta.

En los últimos años, el estudio del microbioma ha cobrado relevancia en el ámbito de la producción animal, y en particular en la acuicultura, debido a su influencia e implicación en diversas funciones del hospedador relacionadas con su metabolismo y la salud. Conocer la composición, abundancia y diversidad de la microbiota intestinal es de crucial importancia para la comprensión de su implicación en estos procesos. Entre los factores extrínsecos que modulan la microbiota, la dieta juega un papel muy importante. En este estudio se alimentaron 750 corvinas con cinco dietas basadas en fuentes de proteína distintas [dieta control (DC), dieta de microalgas (DM), dieta de harina de insectos (DI), dieta de residuos del enlatado del atún (DP.L) y dieta mezcla de las tres anteriores (DMix)] durante un periodo de 60 días, y al final de su engorde se analizaron 6 muestras de cada grupo con el fin de evaluar el efecto de las distintas fuentes de proteína sobre la microbiota intestinal. Para este fin, se secuenció un fragmento del gen *16S ARNr* a través de la tecnología MiSeq® de Illumina®, y se realizó un análisis bioinformático con la herramienta QIIME2, y un análisis bioestadístico con el paquete *microeco* implementado en lenguaje R.

El estudio reveló que *Mycoplasma*, *Enterobacter*, *Bacillus* y *Flavobacterium* fueron los géneros más abundantes en la microbiota de la corvina, independientemente de la dieta considerada. El análisis del índice de alfa diversidad señaló que el grupo de peces alimentados con microalgas (DM) presenta un mayor índice de diversidad alfa ( $H = 2.54 \pm 0.91$ ). Según las distancias de Bray-Curtis ninguna de las comparaciones entre grupos fue estadísticamente significativa. Estos resultados indican que pese a un cambio en la composición de la dieta con relación al origen de la fuente de proteína no se produjo ningún cambio relevante en la microbiota ni situación de disbiosis que pudiera afectar a la condición y salud intestinal de los animales. A pesar de que los resultados entre las diferentes dietas no fueron significativos, este estudio describe por primera vez la composición de la comunidad microbiana del intestino de la corvina.

## 2. Introducción

### 2.1. Contexto y justificación del trabajo

La creciente demanda global de productos acuáticos sanos y nutritivos es un reto al que sólo ha sido posible hacer frente sumando la producción de la acuicultura a la pesca, dos actividades que seguirán de la mano, al menos en las próximas décadas. En este contexto, en los últimos años, la acuicultura ha adquirido una gran relevancia. La demanda creciente de consumo de pescado y marisco, la protección de las existencias de animales silvestres y el fin de ofrecer una fuente de proteínas saludables y sostenibles para el consumo humano, han sido los motivos principales del incremento de su demanda. Los alimentos de origen acuático son una de las más importantes fuentes de proteína animal del mundo. Según la FAO (Organización de las Naciones Unidas para la Alimentación y la Agricultura), los productos acuáticos suponen el 17,0% de la ingesta de proteína animal mundial y el 6,7% de toda la proteína consumida<sup>1</sup>. Se prevé que la acuicultura suministre la mayoría de las proteínas alimentarias en 2050.

El progreso de la acuicultura en las cuatro últimas décadas revela no sólo la vitalidad de esta actividad como técnica productiva, sino también la capacidad de innovación, emprendimiento y aprovechamiento sostenible de los recursos disponibles. La FAO considera que la acuicultura contribuye a la utilización eficaz de los recursos naturales, a la seguridad alimentaria y al desarrollo económico, con un limitado y controlable impacto sobre el medio ambiente. Es por ello por lo que el desarrollo de esta industria requiere del desarrollo de sistemas de producción sostenibles y entre ellos la alimentación de los peces de crianza es de gran relevancia, y en particular el tipo de materias primas que se utilizan para la formulación de los piensos. Para proporcionar estos grandes volúmenes de proteína animal de forma sostenible, se debe tener en cuenta el impacto de la acuicultura en la integridad ambiental, en la salud y bienestar de los animales de cultivo y, además, en la salud humana y seguridad alimentaria<sup>2</sup>.

Cada vez se están buscando más alternativas a los ingredientes habituales usados para la fabricación de los piensos de acuicultura. La harina de pescado y los aceites de pescado han sido las principales fuentes de proteínas y ácidos grasos de los piensos debido a su alto contenido en estas macromoléculas<sup>3,4</sup>. No obstante, la obtención de estos productos no es sostenible, porque depende de las explotaciones de cultivo de peces pelágicos y, por tanto, es necesaria la búsqueda de alternativas a estos ingredientes<sup>5</sup>. En la Figura 1 se muestran los cambios de cómo la industria ha abordado la formulación de piensos para la acuicultura y el uso de distintas materias primas<sup>6</sup>. En los años 90 la principal materia para la fabricación de los piensos era la harina y el aceite de pescado; como se observa esta tendencia ha disminuido a más de la mitad, y actualmente, es la segunda fuente principal de materia prima en los piensos. Se espera que en un futuro su utilización sea mínima, y la mayor parte de materia prima para la fabricación de los piensos sean proteínas de origen vegetal.

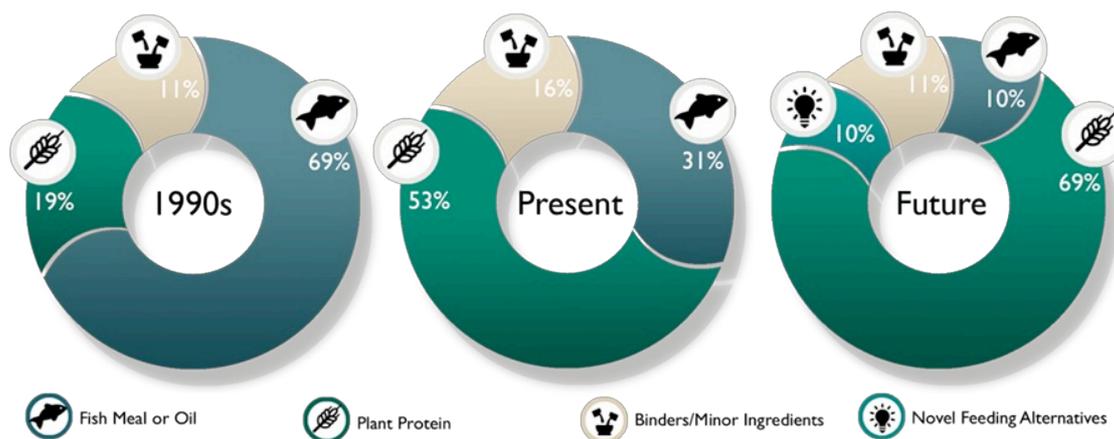


Figura 1. Cambios de la industria en la formulación de piensos para acuicultura

Las investigaciones de los últimos años se han centrado en productos derivados de la soja, como alternativa a la harina y al aceite de pescado<sup>7</sup>. A pesar de ello, se están buscando otros ingredientes para la elaboración de los piensos, entre ellos podemos destacar las harinas de insectos y de algas<sup>8-11</sup>. Otro ingrediente que ha despertado interés en los últimos años es el agua de cola, un subproducto del enlatado del atún (TWC). Este subproducto contiene proteína soluble que se puede aprovechar como fuente de proteína en la fabricación del pienso, promoviendo así modelos de economía circular y el aprovechamiento de subproductos de la pesca<sup>12</sup>.

El uso de nuevos ingredientes para piensos acuícolas está en crecimiento y el estudio de su eficacia como ingredientes alternativos es esencial para la industria<sup>13</sup>. Evaluar los efectos de estos ingredientes en las distintas especies de cultivo puede ser el objetivo de muchas investigaciones.

En los últimos años, se ha tomado una mayor conciencia sobre la importancia del microbioma de los peces y cómo este afecta a las distintas funciones del organismo y de su desarrollo<sup>14-18</sup>. Se ha demostrado que la colonización temprana del intestino de los peces está causada por la absorción de las bacterias presentes en el agua<sup>19</sup> y también por la composición de los piensos<sup>20</sup>. La dieta y las variaciones de los factores ambientales, como la temperatura y la salinidad, pueden influir en los tipos de microorganismos que colonizan el intestino<sup>21,22</sup>.

La corvina (*Argyrosomus regius*), ver Figura 2, es una de las especies seleccionadas para la diversificación de la acuicultura en el Mediterráneo por sus características adaptativas y su alto valor nutricional<sup>23</sup>.



*Figura 2. Corvina (Argyrosomus regius)*

En una investigación previa de este equipo se evaluaron los efectos de distintos ingredientes alternativos estudiando el rendimiento, crecimiento, digestibilidad y calidad muscular en la corvina<sup>24</sup>. No obstante, su microbioma no está explorado, siendo de gran relevancia su caracterización por el papel importante que juegan las comunidades microbianas del hospedador en funciones relacionadas con el metabolismo y la salud.

El objetivo de este estudio es explorar la comunidad bacteriana asociada al intestino de la corvina y evaluar el efecto de diferentes ingredientes frente al uso clásico de la harina de pescado como fuente de proteína en la dieta, en la composición, abundancia y diversidad microbiana de esta. Para ello se ha usado la secuenciación del amplicón del gen *16S ARNr* utilizando la tecnología MiSeq® de Illumina®.

## 2.2. Objetivos del trabajo

### 2.2.1. Objetivo general

- Evaluar el efecto de distintos ingredientes (fuentes de proteína alternativas como las harinas de insecto, de microalgas y subproductos de la pesca) en la dieta de la corvina, en condiciones de cultivo, con relación a su microbiota intestinal.

### 2.2.2. Objetivos específicos

- Caracterizar la microbiota intestinal de la corvina.
- Estudiar el efecto de la dieta acerca de la composición, riqueza y diversidad de la microbiota intestinal de la corvina.

## 2.3. Enfoque y método seguido

Actualmente, existen multitud de herramientas y estrategias para el análisis de datos de microbioma.

En este estudio se utilizó QIIME2 para el análisis bioinformático, ya que se trata de una plataforma de código abierto con herramientas interactivas<sup>25</sup> ampliamente utilizada en este tipo de estudios. En comparación con otras herramientas, QIIME2 es fácil de instalar, se puede ejecutar en máquinas virtuales como VirtualBox, es reproducible, sus resultados pueden ser visualizados de forma virtual, admite características personalizadas y se une al proceso de análisis<sup>26</sup>.

El análisis bioestadístico se realizó con el *software* libre R, empleando el paquete *microeco*. R fue diseñado para hacer análisis estadísticos y gráficos. Se puede descargar y usar sin problemas en un ordenador personal, y el aprendizaje es sencillo comparado con otros lenguajes de programación.

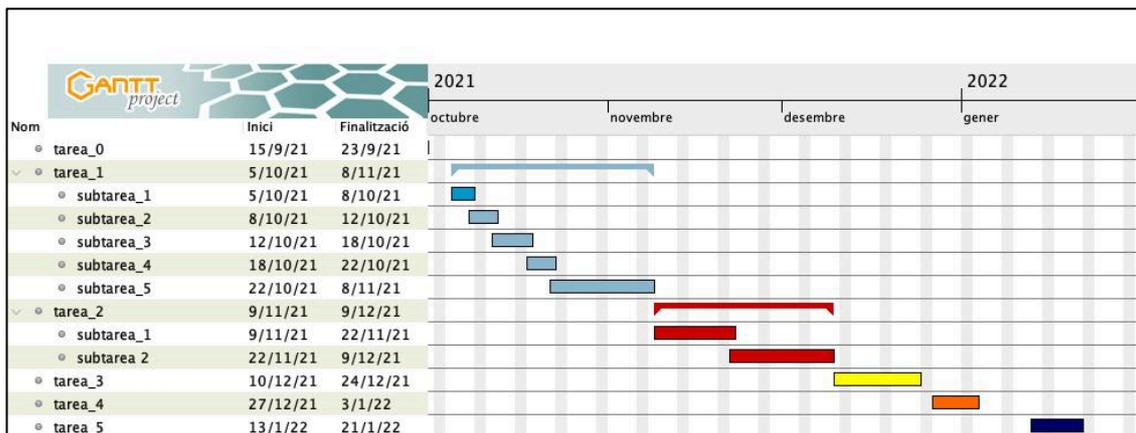
Las técnicas de secuenciación de alto rendimiento producen una gran cantidad de datos complejos, por ello los análisis y la gestión de esos datos son un desafío. Existen muchos paquetes de R para el análisis de los perfiles de los microbiomas; sin embargo, sigue siendo difícil hacer una minería de datos de manera rápida y eficiente. Por ese motivo, se creó el paquete *microeco*<sup>27</sup>, que es un referente en el campo de la ecología microbiana por su rapidez, flexibilidad y facilidad de uso<sup>28</sup>.

## 2.4. Planificación del trabajo

A continuación, se detallan las tareas y las subtareas seguidas para la elaboración del trabajo.

- Tarea 0: Definición del contenido del trabajo.
- Tarea 1. Dividida en cinco subtareas:
  - Subtarea 1: Preparación del entorno de trabajo, instalación de la máquina virtual y de QIIME2. Descarga de los datos.
  - Subtarea 2: Importación de los datos y demultiplexación de las muestras.
  - Subtarea 3: Filtrado de las muestras.
  - Subtarea 4: Control de calidad de las secuencias con DADA2, construcción de la tabla de características y generación de los gráficos de rarefacción.
  - Subtarea 5: Clasificación taxonómica y generación de un árbol filogenético.
- Tarea 2. Dividida en dos subtareas:
  - Subtarea 1: Análisis de la alfa y beta diversidad.
  - Subtarea 2: Análisis de la composición taxonómica y abundancia diferencial. Interpretación y revisión de los resultados.
- Tarea 3. Finalización de la memoria.
- Tarea 4. Elaboración de la presentación.
- Tarea 5. Defensa pública del trabajo.

Tabla 1. Planificación del trabajo



## 2.5. Breve resumen de contribuciones y productos obtenidos

Los productos obtenidos del trabajo constituyen los requeridos para un trabajo final de máster como son la memoria de trabajo, los planes de trabajo y su seguimiento. Finalmente, se hará una defensa pública del trabajo con una presentación audiovisual.

Aunque los resultados obtenidos no han mostrado diferencias significativas, no se descarta una publicación científica, ya que esta sería, a fecha de hoy, la primera publicación científica que describe la microbiota intestinal de la corvina, una nueva especie para la acuicultura con la que se han llevado a cabo trabajos interesantes de nutrición a efectos de rendimiento, crecimiento, digestibilidad y calidad muscular<sup>24</sup>.

## 2.6. Breve descripción de los otros capítulos de la memoria

En el capítulo tres se informa sobre el estado del arte, y a continuación, en el capítulo cuatro se incluye la metodología seguida para realizar el presente estudio, desde el procedimiento experimental, hasta el análisis bioinformático y bioestadístico. La idea principal es poder reproducir este tipo de análisis a partir de otros datos de ensayos parecidos, proporcionando los aspectos más importantes durante el proceso. A continuación, el capítulo cinco muestra los resultados obtenidos con la metodología descrita y también su interpretación. En el capítulo seis, se valora el estudio procedido en función de los resultados obtenidos; mientras que en el capítulo siete se concluyen los aspectos más fundamentales del trabajo efectuado, así como futuras mejoras y líneas de trabajo. Los cuatro últimos capítulos constituyen el glosario, la bibliografía, los anexos con el código usado para poder ejecutar el análisis y los agradecimientos.

# 3. Estado del arte

La microbiota intestinal tiene grandes beneficios para el hospedador y participa en diferentes funciones del organismo y del desarrollo<sup>14-18</sup>. El estudio de la microbiota de los organismos de cultivo acuícola ha cobrado relevancia en los últimos años, pero no hay mucha información acerca de la microbiota de la corvina. El uso de las técnicas convencionales microbiológicas para el estudio y caracterización de la microbiota no permite su correcto estudio, pues una fracción muy baja de las bacterias que hay en el organismo son cultivables, por lo que el uso de herramientas de tipo ómico, como las relacionadas con la secuenciación masiva son en la actualidad una de las únicas herramientas para llevar a cabo este tipo de estudios. Por ese motivo, el presente estudio es importante, ya que nos va a permitir conocer la composición microbiana y valorar cómo la dieta afecta a la composición, abundancia y diversidad de la microbiota intestinal de la corvina.

## 4. Metodología

### 4.1. Ética

Las manipulaciones de los peces se realizaron de acuerdo con las Directrices del Consejo de la Unión Europea (2010/63 / UE) y la Legislación Española para uso en animales de laboratorio.

### 4.2. Procedimiento experimental

Cinco dietas experimentales fueron formuladas por DIBAQ Acuicultura (Segovia, España) y fabricadas por el Centro Tecnológico CARTIF (Valladolid, España) utilizando las mismas instalaciones y parámetros de extrusión para todas ellas. Las dietas fueron las siguientes: (1) dieta control similar a un pienso comercial (DC), que contenía harina de pescado como principal fuente proteica en el alimento; (2) dieta de microalgas (DM) que contenía 10% de una mezcla de cuatro microalgas marinas (*Nannochloropsis gaditana*, *Tisochrysis lutea* (CCAP 927/14), *Rhodomonas lens* (ECC030) e *Isochrysis galbana* (CCAP927/1)) producidas por ANFACO-CECOPESCA (Vigo, España); (3) dieta de insectos (DI) en la que se incluyó un 15% de harina de *Acheta domesticus* producida por Nustrinsect (Navarra, España); (4) dieta con un 7% de proteína y 11% de lípidos (DP & L) procedente del agua de cola del enlatado del atún (TWC), recuperada por ANFACO-CECOPESCA (Vigo, España); y (5) dieta mixta (DMix) basada en la inclusión de los tres ingredientes (10% harina de microalgas, 15% harina de insectos, y la fracción proteica y lipídica del proceso de fabricación del atún enlatado). Todas las dietas fueron isoprotéicas, isolipídicas e isoenergéticas en comparación con la dieta control. La Tabla 2 muestra la formulación y composición de las dietas experimentales usadas para el estudio<sup>24</sup>.

Tabla 2. Formulación (%) y composición de proteínas y lípidos crudos (% de peso seco, PS) de las dietas experimentales usadas para el estudio

<b>Ingredientes</b>	<b>DC</b>	<b>DM</b>	<b>DP&amp;L</b>	<b>DI</b>	<b>DMix</b>
Harina de calamar	-	-	-	-	4,41
Harina de pescado	23,76	20,00	21,63	21,46	20,00
Harina de insecto	-	-	-	15,00	15,00
Microalgas	-	10,00	-	-	10,00
Proteínas (TWC)	-	-	7,00	-	2,00
Lípidos (TWC)	-	-	10,89	-	9,43
Almidón de guisante	6,68	6,32	7,14	7,97	1,14
Gluten de trigo	3,57	15,54	4,23	-	15,00
Trigo	6,00	6,00	6,00	6,00	17,12
Soja	6,16	4,63	6,18	5,18	0,15
Harina de gluten de maíz	15,00	15,00	10,57	8,32	1,61
Harina de guar	6,00	-	6,00	6,00	-
Aceite de salmón	7,20	10,92	-	8,55	-
Aceite de krill	4,00	-	-	-	-
Mix de AA (Aminopro)	5,00	5,00	5,00	5,00	2,29
Lisina	0,18	0,50	0,50	0,27	0,50
Treonina	0,00	0,32	0,29	0,07	0,50
Metionina	0,07	0,05	0,18	0,14	0,09
Colina	0,16	0,16	0,16	0,16	0,16
Taurina	0,16	0,17	0,18	0,18	0,15
Ácido butírico	0,10	0,10	0,10	0,10	0,10
Antibacterianos	0,15	0,15	0,15	0,15	0,15
Antifúngicos	0,015	0,015	0,015	0,015	0,015
Antioxidantes	0,07	0,07	0,07	0,07	0,07
Atrayente	0,10	0,10	0,10	0,10	0,10
Betaína anhidra	0,07	0,07	0,07	0,07	0,07
Conc. Minerales orgánicos	0,10	0,10	0,10	0,10	0,10
Conc. Vitaminas	0,10	0,10	0,10	0,10	0,10
Humedad (%)	7,07	9,38	6,22	11,99	5,98
Proteínas brutas (%PS)	47,08	47,11	47,63	44,18	46,83
Lípidos brutos (%PS)	14,88	16,14	14,87	15,55	16,81

DC: dieta control, DI: dieta con harina de insecto, DM: dieta con harina de microalgas, DP&L: dieta con proteínas y lípidos del enlatado del atún (TWC), DMix: dieta con la mezcla de todos los ingredientes.

### 4.3. Recogida de muestras, extracción de ADN y secuenciación

A lo largo del ensayo no se produjo ninguna mortalidad ni síntoma de enfermedad que tuviera que ser tratada. Todos los peces estuvieron sanos durante el ensayo. Al finalizar el experimento (60 días) y tras un ayuno de 24 horas todos los peces de los tanques se muestrearon a fin de obtener su peso final, así como el peso de las vísceras, hígado y filete que fueron usados para estimar los índices somáticos, incluidos en el estudio previo<sup>24</sup>. Además, se recolectaron un total de cinco intestinos de corvina por tanque (3 tanques por dieta experimental). Estos cinco intestinos fueron recolectados en un único tubo y se mantuvieron en RNAlater™ (Invitrogen, Massachusetts, Estados Unidos) durante 24 horas a 4°C y posteriormente fueron congelados a -80°C hasta su procesamiento. Las muestras de cada tanque se separaron de forma homogénea en dos submuestras y se trituraron con un homogeneizador de tejidos (IKA ULTRA-TURRAX®, Staufen, Alemania).

Se obtuvieron un total de 6 muestras por grupo experimental y finalmente se realizó una extracción de ADN con el kit QIAamp® Fast DNA Stool siguiendo las instrucciones del fabricante (QIAGEN, Hilden, Alemania). El ADN (ácido desoxirribonucleico) extraído se envió a la empresa MR DNA (Texas, Estados Unidos) para la preparación y secuenciación de las muestras.

Los cebadores, V3\_F357\_N: 5'-CCTACGGGNGGCWGCAG-3' y V4\_R805: 5'-GACTACHVGGGTATCTAATCC-3' se usaron para amplificar un fragmento del gen *16S ARNr* bacteriano, concretamente la región V3-V4, tal y como se muestra en la Figura 3. Los amplicones se obtuvieron a través de la secuenciación de extremo emparejado (*paired-end*, 2 x 300 bp) con la tecnología MiSeq® de Illumina® (Illumina, California, Estados Unidos).

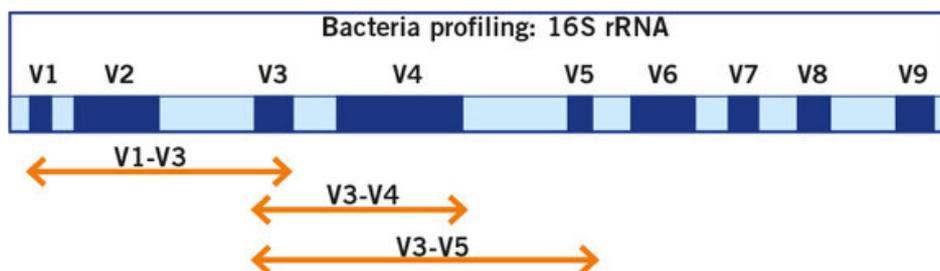


Figura 3. Regiones del gen *16S ARNr*

## 4.4. Bioinformática

### 4.4.1. Obtención de los datos

Los datos analizados corresponden a un conjunto de muestras de intestino de corvina obtenidas a los 60 días del experimento. Los archivos de las secuencias originales de este estudio pertenecen a la región V3 y V4 del gen *16S ARNr* y fueron secuenciadas a una profundidad mínima de 20K lecturas con el método de extremo emparejado. Los amplicones se secuenciaron con una máquina Illumina® a través de la tecnología MiSeq® a una longitud de lectura de 300 pares de bases (bp).

Las secuencias fueron obtenidas a través de la plataforma BaseSpace de Illumina® y se procesaron con el FASTQ Processor (MR DNA, Texas, Estados Unidos). Las secuencias de los códigos de barras, de los cebadores y de los adaptadores se eliminaron usando esta herramienta. Posteriormente, las secuencias fueron analizadas con QIIME2.

Se integró información sobre el experimento al archivo de metadatos de la secuenciación. El archivo resultante es de suma importancia, ya que contiene la identidad de las secuencias de los códigos de barras (*BarcodeSequence*) asociados a las muestras (ver Tabla 3).

Tabla 3. Estructura del archivo de metadatos (solo se muestran las primeras filas como ejemplo de la matriz de datos generada).

sample-metadata								
SampleID	BarcodeSequence	LinkerPrimerSequence	BarcodeName	ReversePrimer	ProjectName	Description	Group	Tank
#q2:types	categorical	categorical	categorical	categorical	categorical	categorical	categorical	categorical
CORVM.1.1	CAGTTCAT	CCTACGGGNGGCWGCAG	illwhoN341bar1	GACTACHVGGGTATCTAATCC	052421ARwhoI341F	CORVM.1.1	DC	T1
CORVM.1.2	CAGTTGCA	CCTACGGGNGGCWGCAG	illwhoN341bar2	GACTACHVGGGTATCTAATCC	052421ARwhoI341F	CORVM.1.2	DC	T1
CORVM.2.1	CATAACAA	CCTACGGGNGGCWGCAG	illwhoN341bar3	GACTACHVGGGTATCTAATCC	052421ARwhoI341F	CORVM.2.1	DPL	T2
CORVM.2.2	CATAATAG	CCTACGGGNGGCWGCAG	illwhoN341bar4	GACTACHVGGGTATCTAATCC	052421ARwhoI341F	CORVM.2.2	DPL	T2

El archivo de metadatos está formado por la columna *SampleID* que contiene el nombre de la muestra y la columna *BarcodeSequence*, que es la secuencia nucleotídica del código de barras que permite identificar las muestras. También se incluyen las columnas *LinkerPrimerSequence* y *ReversePrimer* con las secuencias del adaptador y del cebador antisentido, respectivamente. La columna *BarcodeName* contiene el nombre del código de barras usado. En la columna *ProjectName* figura el nombre del proyecto, y la columna *Description* contiene la misma información que la columna *SampleID*. Las dos últimas columnas son las más relevantes, ya que contienen información descriptiva sobre los organismos, *Group* y *Tank*. En la primera encontramos el nombre del grupo al que pertenece la muestra y en la segunda el tanque en el que habitaban los animales. Estas dos columnas van a ser de suma importancia, pues nos permitirán conectar los datos de la secuenciación con nuestras características experimentales.

#### 4.4.2. Importación de los datos a QIIME2

A continuación, se importaron los datos de las secuencias de ADN a QIIME2. Todos los datos utilizados y generados por QIIME2, a excepción de los metadatos, existen como artefactos de QIIME2 y utilizan la extensión de archivo *.qza*. Los artefactos son archivos comprimidos que contienen datos en formatos habituales, como FASTQ. Los artefactos de QIIME2 describen las diversas características de los datos, como el tipo semántico, el formato y las citas relevantes para los pasos del análisis<sup>29</sup>.

Al importar los datos a QIIME2, proporcionamos detalles sobre los datos, incluyendo el formato del archivo y el tipo semántico. Actualmente, el tipo más común de datos brutos de la secuenciación de alto rendimiento es el formato FASTQ. Este formato almacena la secuencia de nucleótidos con sus puntuaciones de calidad codificadas con un solo carácter con código ASCII.

Indicamos que los datos eran de extremo emparejado para su importación a QIIME2. Proporcionamos a QIIME2 la ubicación de los archivos, las secuencias R1, las secuencias R2 y el archivo de metadatos, para poder proceder a la importación.

#### 4.4.3. Demultiplexación de los datos

La mayoría de las máquinas de secuenciación de próxima generación tienen la capacidad de analizar cientos o incluso miles de muestras en un solo carril. Esto se consigue gracias al multiplexado de las muestras, que es una palabra elegante para referirnos a mezclar un montón de cosas. Pero ¿cómo sabemos de qué muestra proviene cada lectura? Esto se consigue añadiendo una secuencia única de código de barras a uno o ambos extremos de cada secuencia. Detectar estas secuencias de los códigos de barras y mapearlas de nuevo a las muestras a las que pertenecen nos permitió demultiplexar las secuencias<sup>30</sup>.

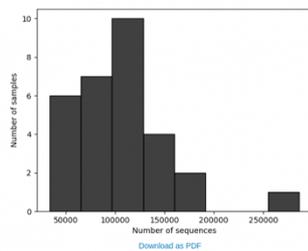
El artefacto demultiplexado permitió crear un resumen interactivo de los datos. Este resumen nos proporcionó información útil para evaluar la calidad de la ejecución de la secuenciación, incluido el número de lecturas obtenidas por muestra y la distribución de las puntuaciones de calidad de la secuencia en cada posición nucleotídica. Transformamos el artefacto a formato *.qzv* para poder visualizarlo con QIIME2 view<sup>31</sup>.

Generamos un subconjunto de las muestras, ya que en nuestro caso los archivos contenían datos de otros experimentos. Después volvimos a transformar el nuevo artefacto para poder visualizarlo.

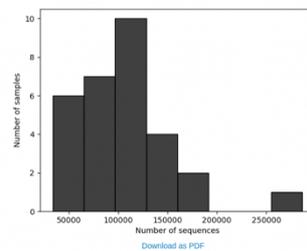
#### Demultiplexed sequence counts summary

	forward reads	reverse reads
Minimum	33672	33672
Median	106073.0	106073.0
Mean	105110.9	105110.9
Maximum	286626	286626
Total	3153327	3153327

Forward Reads Frequency Histogram



Reverse Reads Frequency Histogram



#### Per-sample sequence counts

Total Samples: 30 (forward) 30 (reverse)

sample ID	forward sequence count	reverse sequence count
CORV.M.7.2	286626	286626
CORV.M.13.2	165920	165920
CORV.M.18.2	161964	161964

Figura 4. Resumen del conteo de lecturas

En la parte superior de la Figura 4 podemos ver un resumen estadístico del recuento de las secuencias de las muestras en cada uno de los sentidos de secuenciación y sus distribuciones. Y en la parte inferior de la misma Figura 4 un desglose del recuento de secuencias (sentido y antisentido) por muestra. En la pestaña *Interactive Quality Plot*, podemos ver la gráfica de calidad de las secuencias, tal y como se muestra en la Figura 5. En ella se puede observar una gráfica para cada uno de los sentidos de secuenciación, donde en el eje X está representada la posición de cada base nucleotídica y en el eje Y el valor de calidad medio para cada una de las posiciones. Calcular las calidades para todos los datos puede llevar mucho tiempo, por eso se hace a partir de un subconjunto de lecturas seleccionadas al azar. El valor predeterminado es de 10000 secuencias. Estos gráficos sirven para determinar si hay una disminución en la calidad de alguna posición de las secuencias, lo que puede ser útil para elegir los parámetros de truncamiento y recorte.

#### 4.4.4. Control de calidad de las secuencias

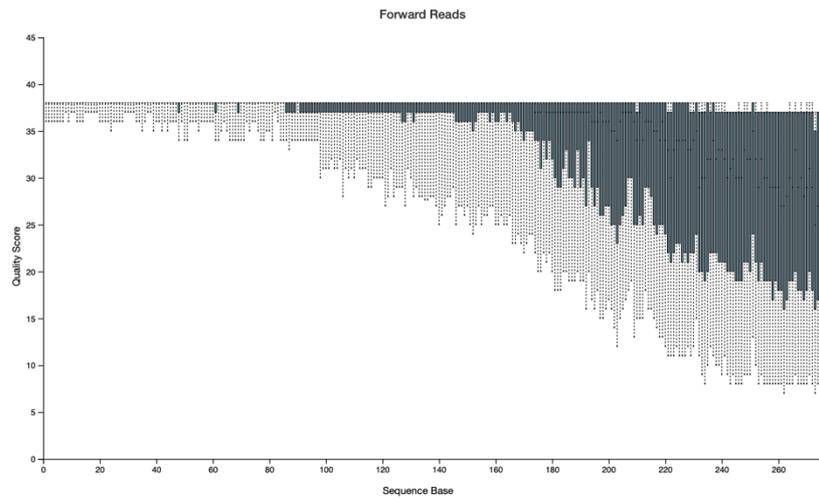
Tradicionalmente, el control de calidad de las secuencias se realizaba recortando y filtrando las secuencias en función de sus puntuaciones de calidad<sup>28-30</sup>, y a continuación se agrupaban en unidades taxonómicas operativas (OTU) basadas en un umbral de diferencia fijo, normalmente del 97%<sup>33</sup>.

Actualmente, existen mejores métodos para el control de calidad que corrigen los posibles errores de la secuencia nucleotídica del amplicón, creando lo que se conoce como variantes de secuencia de amplicón (*Amplicon sequence variant*, ASV), que en comparación con las OTU resuelven diferencias de tan solo un nucleótido. Las OTU y las ASV también se conocen como características.

Para filtrar las secuencias utilizamos el complemento DADA2. Este método consiste en una serie de pasos que permiten (1) el filtraje de las secuencias que no tienen una longitud mínima requerida (*triming*), (2) la combinación de todas las lecturas idénticas en secuencias únicas, asignando la abundancia a cada secuencia (*denoising*) y (3) la eliminación de secuencias quiméricas por el método de consenso, que es el que viene establecido por defecto. Los parámetros de corte se obtienen a partir de los gráficos de calidad de las secuencias (ver Figura 5).

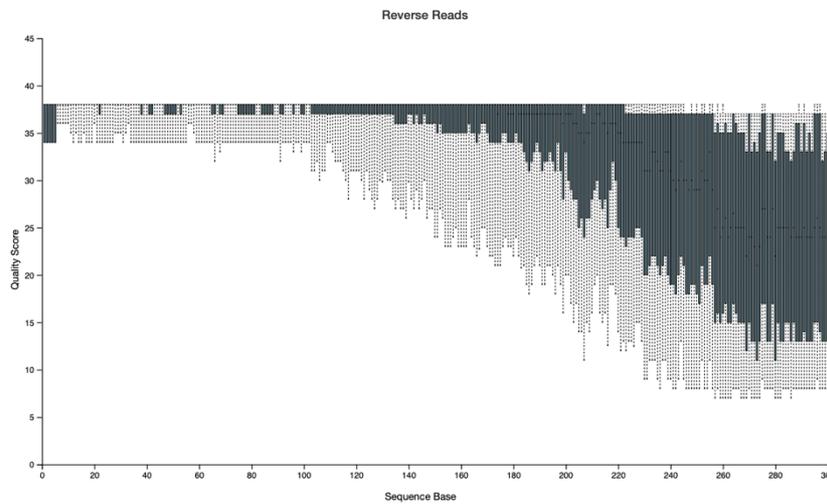
En este paso es importante elegir unos valores adecuados de recorte y truncamiento. Se recomienda establecer el valor de truncamiento a una longitud en la que la calidad media comience a disminuir por debajo de 30, o de 20 si la calidad de secuenciación general es demasiado baja.

En el gráfico interactivo, equivalente a la Figura 5, pudimos comprobar que en la posición del nucleótido 250 de las lecturas sentido, la calidad empezaba a disminuir por debajo de Q30. En el caso de las lecturas antisentido esta disminución se observó a la posición 240.



The plot at position 250 was generated using a random sampling of 9983 out of [object Object] sequences without replacement. The minimum sequence length identified during subsampling was 53 bases. Outlier quality scores are not shown in box plots for clarity.

Parametric seven-number summary for position 250		
Box plot feature	Percentile	Quality score
(Not shown in box plot)	2nd	7
Lower Whisker	9th	9
Bottom of Box	25th	20
Middle of Box	50th (Median)	32
Top of Box	75th	37
Upper Whisker	91st	37
(Not shown in box plot)	98th	38



The plot at position 240 was generated using a random sampling of 9993 out of [object Object] sequences without replacement. The minimum sequence length identified during subsampling was 116 bases. Outlier quality scores are not shown in box plots for clarity.

Parametric seven-number summary for position 240		
Box plot feature	Percentile	Quality score
(Not shown in box plot)	2nd	7
Lower Whisker	9th	9
Bottom of Box	25th	19
Middle of Box	50th (Median)	31
Top of Box	75th	37
Upper Whisker	91st	38
(Not shown in box plot)	98th	38

Figura 5. Calidad de secuenciación de las secuencias sentido y antisentido

La disminución de calidad es debida a los errores de secuenciación. Estos errores son más frecuentes en los métodos de secuenciación NGS (*Next-Generation Sequencing*) que en los métodos convencionales. Estos errores pueden conducir a falsos positivos de asignación de bases nucleotídicas, afectando a la asignación taxonómica de las secuencias.

En general estos errores se producen al azar, pero a veces, en función de la plataforma utilizada, se pueden correlacionar con algunas características típicas. Por ejemplo, en la plataforma Illumina®, los errores se correlacionan con la posición de la base en la secuencia, acumulándose estos errores con mayor frecuencia hacia el final de las lecturas.

Como valor de recorte usamos 20, aunque los cebadores y los adaptadores ya se eliminaron con el FASTQ processor previamente. Nos aseguramos de que las secuencias no biológicas no influyeran en el análisis.

DADA2 genera tres tipos de salidas (artefactos), (1) artefacto del tipo FeatureTable[Frequency], (2) artefacto del tipo FeatureData[Sequence] y (3) artefacto con el resumen estadístico de DADA2.

El artefacto del tipo FeatureTable[Frequency] está formado por tres pestañas. En la pestaña *Overview* (Figura 6) vemos un resumen de la tabla de características, la frecuencia por muestra y la frecuencia por característica. En la parte superior de la figura vemos que el número de muestras es 30, el número de características (ASV) es de 1.774 y la frecuencia total de 1.945.657. La frecuencia se refiere al número total de lecturas que han pasado el filtraje de DADA2. En la parte donde se describe la frecuencia en función de las muestras podemos ver la frecuencia mínima (16.354), este parámetro es el número mínimo de lecturas presentes en una muestra. También observamos el promedio de la frecuencia en función de las muestras y el valor máximo, entre otros. Es importante entender este tipo de información y su histograma. En él observamos que la mitad de las lecturas (64.855,23) están presentes aproximadamente en la mitad de las muestras (14 muestras). De la misma forma, observamos la frecuencia en función de las características en la parte inferior de la figura. En esta parte es relevante destacar que la ASV más rara está presente en los datos con una frecuencia de una lectura. Al igual que también podemos apreciar, con el valor máximo de frecuencia, que la ASV más común está presente en los datos con una frecuencia de 693.810 lecturas. La mayoría de las ASV están representadas en las muestras con una frecuencia de 1.096,76 lecturas.

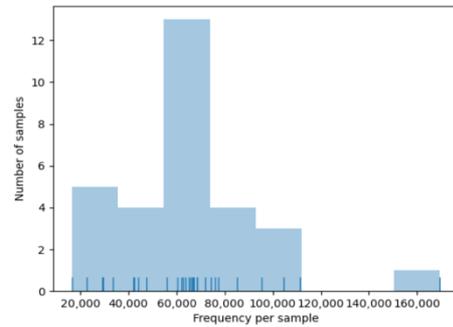
### Table summary

Metric	Sample
Number of samples	30
Number of features	1,774
Total frequency	1,945,657

### Frequency per sample

	Frequency
Minimum frequency	16,354.0
1st quartile	45,014.75
Median frequency	65,520.0
3rd quartile	73,749.25
Maximum frequency	169,570.0
Mean frequency	64,855.2333333333

Frequency per sample detail ([csv](#) | [html](#))

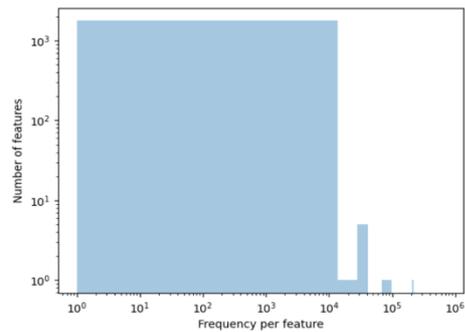


[Download as PDF](#)

### Frequency per feature

	Frequency
Minimum frequency	1.0
1st quartile	16.0
Median frequency	59.0
3rd quartile	248.0
Maximum frequency	693,810.0
Mean frequency	1,096.7626832018038

Frequency per feature detail ([csv](#) | [html](#))



[Download as PDF](#)

Figura 6. Visualización de QIIME2 de la tabla de características

En la pestaña *Interactive Sample Detail* visualizamos el conteo de las lecturas de cada muestra (Figura 7). En la parte superior de la figura se observa un histograma donde en el eje de las X se representan los distintos grupos experimentales y en el eje de las Y el número de muestras. Si nos fijamos, aparecen 6 muestras para cada grupo experimental del estudio. El número total de lecturas, para cada muestra, aparece en la parte inferior del histograma (se muestran solo los conteos para las tres primeras muestras). La pestaña *Feature Detail*, que no visualizamos en este documento, contiene una lista de todos los ASV (1.774) con la frecuencia en la que aparecen en las muestras. A modo de ejemplo, la primera ASV (la más abundante) aparece en una frecuencia de 693.810 a todas las muestras.

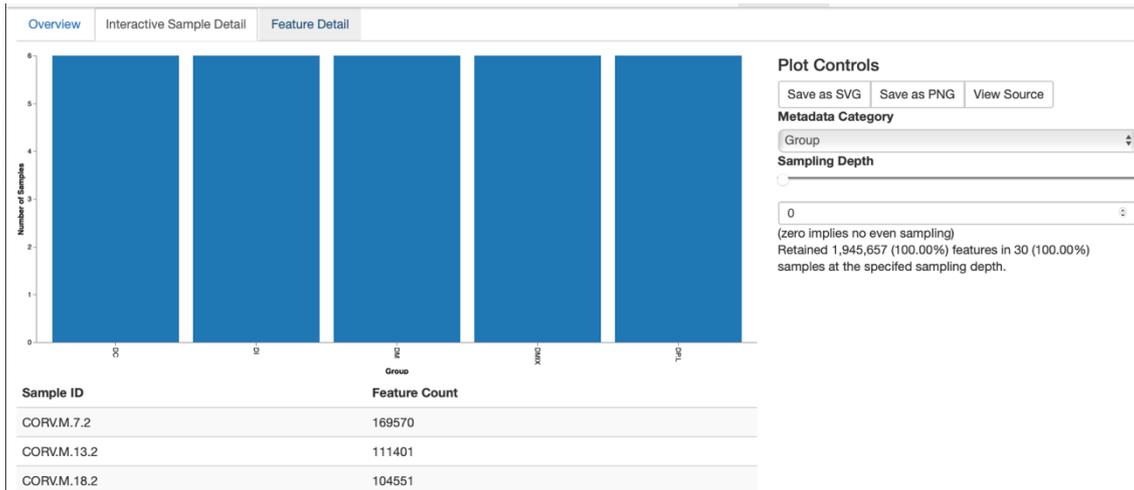


Figura 7. Visualización de QIIME2 de los detalles por muestra de la tabla de características

La Figura 8 nos muestra el artefacto FeatureData[Sequence]. En la parte superior vemos un resumen de siete números sobre la longitud de las secuencias y en la parte inferior una tabla en la que se representa un ID asignado para cada una de las secuencias, y la longitud de estas. La secuencia nos permite acceder directamente a la herramienta BLAST. Este artefacto es útil para la clasificación taxonómica. QIIME2 utiliza el *hash* de MD5 para la asignación del ID a la secuencia de forma predeterminada.

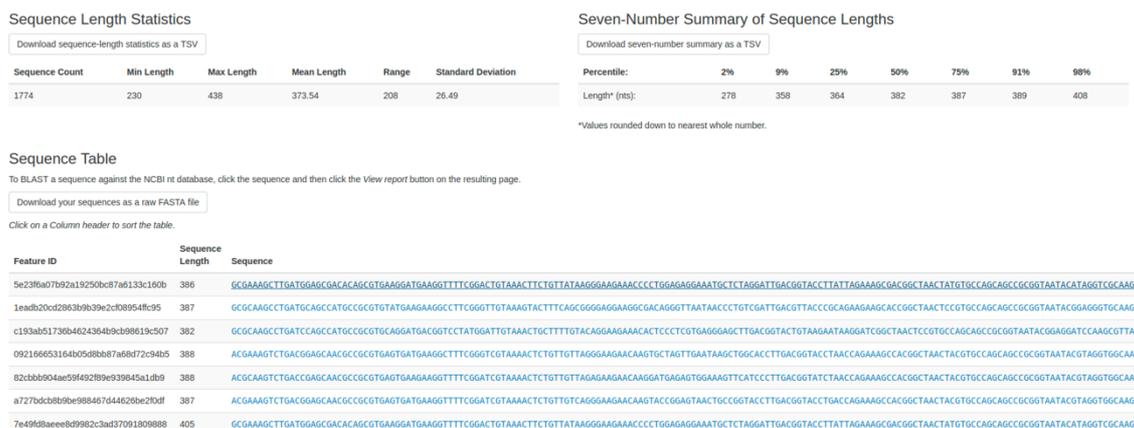


Figura 8. Visualización de QIIME2 de las secuencias representativas

En la Figura 9 visualizamos el resumen estadístico de DADA2. Se trata de una tabla formada por nueve columnas y tantas filas como muestras analizadas. Las columnas de la tabla contienen información sobre el filtraje realizado con DADA2, (1) identidad de las muestras, (2) lecturas iniciales, (3) lecturas filtradas según el valor de calidad Q, (4) porcentaje de lecturas que han pasado el filtraje, (5) lecturas a las que se han eliminado los errores debidos a la secuenciación (*denoised*), (6) lecturas que han superado la superposición (*merged*); aquí una caída importante del número de lecturas indicaría que las secuencias son demasiado cortas y no se pueden superponer y/o los extremos no se alienan, (7) porcentaje de lecturas que han superado el filtraje *denoised* y *merged*, (8) lecturas a las que se han eliminado las quimeras (subproductos no reales de la secuenciación) y (9) porcentaje de lecturas que han pasado todos los filtrajes y que vamos a analizar. El número medio de lecturas a analizar es del 62% sobre el *input* inicial. Este valor se calculó exportando los datos del artefacto de DADA2 a Excel, y calculando la media.

sample-id	input	filtered	percentage of input passed filter	denoised	merged	percentage of input merged	non-chimeric	percentage of input non-chimeric
seq.types	numeric	numeric	numeric	numeric	numeric	numeric	numeric	numeric
CORVM.1.1	106670	69248	64.92	68555	66735	62.56	65880	61.76
CORVM.1.2	48084	30472	63.37	30146	29758	61.89	29504	61.36
CORVM.10.1	72770	45456	62.47	44807	43063	59.18	41957	57.66
CORVM.10.2	115839	71620	61.83	70597	66764	57.64	63893	55.16
CORVM.11.1	70413	48965	69.54	48492	47935	68.08	47459	67.4
CORVM.11.2	105476	73805	69.97	73308	72440	68.68	71956	68.22
CORVM.13.1	151816	100565	66.24	99446	97070	63.94	95218	62.72
CORVM.13.2	165920	115557	69.65	114753	112990	68.1	111401	67.14
CORVM.14.1	86987	61033	70.16	60700	60317	69.34	60192	69.2
CORVM.14.2	115784	73169	63.19	72244	70094	60.54	68443	59.11
CORVM.17.1	50203	35156	70.03	34584	33551	66.83	33443	66.62
CORVM.17.2	103497	69131	66.8	68231	66164	63.93	65160	62.96
CORVM.18.1	120101	71309	59.37	70315	68284	56.86	66778	55.6
CORVM.18.2	161964	108701	67.11	108011	106622	65.83	104551	64.55
CORVM.2.1	131458	88775	67.53	87971	86120	65.51	85176	64.79
CORVM.2.2	91118	63164	69.32	62693	62231	68.3	62149	68.21

Figura 9. Resumen estadístico con DADA2

#### 4.4.5. Filtraje de los datos

QIIME2 proporciona numerosos métodos para filtrar los datos. En este caso se realizaron los siguientes filtrajes de la tabla de características; uno basado en la frecuencia total y el otro basado en contingencias. El filtrado basado en la frecuencia total se utiliza para filtrar las muestras en función de la frecuencia con la que se representan en la tabla de características, de esta manera conseguimos eliminar las características de baja abundancia de las muestras. En nuestro caso eliminamos aquellas características con una abundancia del 0,001%. El parámetro más complicado de elegir es la frecuencia mínima de las muestras. Para calcular este valor se siguió el tutorial de soporte de QIIME2<sup>34</sup>, donde se explica que para filtrar ASV con una abundancia relativa inferior al 0,001% se tiene que multiplicar la frecuencia total obtenida de la tabla de características, (ver Figura 6 (1.945.657)), por 0,00001. Se obtuvo el valor aproximado de 20. De esta forma nos aseguramos de que como mínimo una característica apareciera con una frecuencia mínima de 20 veces por muestra.

En la filtración por contingencia filtramos las muestras en función del número de características presentes. Este filtrado se usa para filtrar características que aparecen en una o en pocas muestras, basándose en la sospecha de que pueden no representar una diversidad biológica real, sino más bien errores. Para el filtraje de contingencias excluimos aquellas características presentes en menos de tres muestras.

#### 4.4.6. Curvas de rarefacción

Un paso importante antes de realizar el análisis del microbioma es evaluar la profundidad de muestreo, ya que variará entre las muestras, porque el número de secuencias generadas no se distribuyen de forma uniforme entre las muestras ni se correlacionan con la biomasa de la muestra y, por lo tanto, para evitar sesgos, se tiene que normalizar antes del análisis. Para explorar cómo la profundidad de muestreo afecta a las estimaciones de la diversidad alfa, concepto que veremos más adelante, se utilizaron las curvas de rarefacción con el complemento  $q_2$ -diversity. En las curvas de rarefacción, podemos visualizar la alfa diversidad como una función de la profundidad de muestreo utilizando el visualizador que nos proporciona QIIME2.

Para generar las curvas de rarefacción elegimos el valor de la profundidad máxima de secuenciación para representar el gráfico. Este valor se determinó revisando la información en la pestaña *Frequency per sample* de la tabla de características. En general, la elección de este valor suele ser la frecuencia media, ya que parece funcionar bien.

La Figura 10 muestra las curvas de rarefacción. La imagen superior es una gráfica de rarefacción alfa según el número de características observadas, y se utiliza principalmente para determinar si la diversidad de las muestras ha sido capturada por completo. Si las líneas en el gráfico parecen nivelarse, es decir, si su pendiente es cero, a cierta profundidad de muestreo a lo largo del eje X, sugiere que es poco probable que la recopilación de secuencias adicionales resulte en cambios significativos en la diversidad estimada de las muestras. Si las líneas del gráfico no se nivelan, es posible que no toda la diversidad de las muestras haya sido capturada. La imagen inferior ilustra el número de muestras en función de la profundidad de muestreo. Si una profundidad de muestreo dada  $d$  es mayor que la frecuencia total de una muestra  $s$  (número de secuencias que se obtuvieron para las muestras  $s$ ), no es posible calcular la métrica de diversidad para las muestras  $s$  a la profundidad de muestreo  $d$ . Si muchas muestras en un grupo tienen frecuencias totales más bajas que  $d$ , la diversidad promedio presentada para ese grupo a la profundidad de muestreo  $d$  en la gráfica superior, no será fiable porque se habrá calculado en pocas muestras. En el gráfico se observa que a una profundidad de muestreo de 20.000 lecturas solo perdemos una muestra perteneciente al grupo de peces alimentados con microalgas (DM). Excluimos la muestra CORV.M.21.2 de los siguientes análisis y capturamos la mayor parte de características, ya que las pendientes a esta profundidad de muestreo se nivelaron. Aumentando este valor habríamos podido capturar más características, pero habríamos perdido demasiadas muestras.

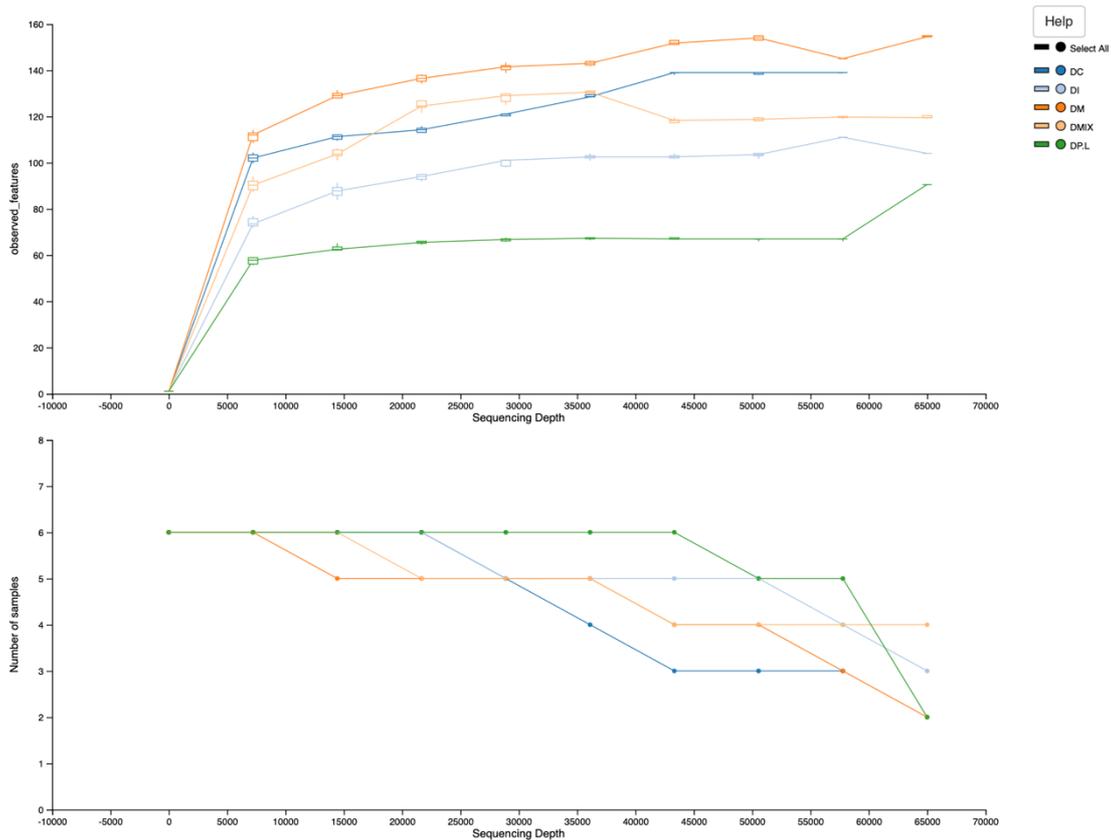


Figura 10. Visualización de QIIME2 de las curvas de rarefacción de la alfa diversidad

Aunque la profundidad de secuenciación en una muestra no se relaciona directamente con la biomasa original en una comunidad, se sabe que tiene un gran impacto en las comunidades observadas<sup>35</sup>. Para que la mayoría de las métricas de diversidad sean comparables entre sí se necesita aplicar un enfoque de normalización, esta normalización es la que se conoce como rarefacción y pretende estimar los índices de diversidad a una misma profundidad de secuenciación para todas las muestras.

Acotamos el valor de profundidad de muestreo con la visualización de los resultados de la tabla de características. En la pestaña *Interactive Sample Detail* encontramos una barra con el nombre de *Sampling Depth* que en función del valor elegido nos dice el número de muestras que conservamos y las características que retenemos de las muestras. Nos interesa elegir un valor lo más alto posible, para que conserve el máximo número de características por muestra, y al mismo tiempo que se excluya la menor cantidad de muestras posible. Elegimos el valor de 21.640, de esta manera se retuvieron 627.560 de las características (34%) en 29 muestras (97,67%) (ver Figura 11). Es un valor bajo, ya que solo conservamos el 34% de las características, pero con este valor solo perdimos una muestra. El valor de profundidad de muestreo de 21.640 fue necesario para el análisis bioestadístico. De forma complementaria, se intentó aumentar el valor de profundidad de muestreo, pero no se consiguieron retener muchas características más y se perdían muchas muestras.

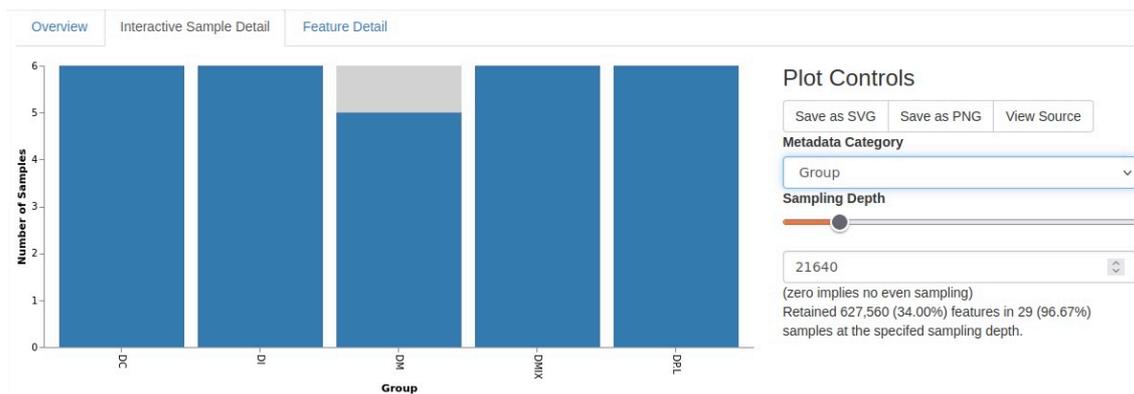


Figura 11. Profundidad de muestreo en función del grupo experimental y número de muestras

#### 4.4.7. Clasificación taxonómica

Las secuencias representativas filtradas con el complemento DADA2 proporcionaron la mayor resolución posible de las características en las muestras. El próximo paso fue conocer la afiliación taxonómica de los microorganismos de los que se obtuvieron estas secuencias. La afiliación taxonómica se consiguió gracias a la comparación de las secuencias obtenidas con las secuencias de referencia de una base de datos con una composición taxonómica determinada. La identificación de la alineación más cercana no es suficiente, ya que secuencias coincidentes pueden tener diferentes anotaciones taxonómicas. En estos casos se usan los clasificadores taxonómicos para determinar la afiliación taxonómica más cercana con algún grado de confianza o consenso, basado en la alineación.

QIIME2 proporciona varios métodos para la clasificación taxonómica. Los métodos de clasificación basados en el aprendizaje están disponibles a través del método Classify-Sklearn<sup>36</sup>. Estos clasificadores deben ser entrenados agregando un paso adicional al proceso de clasificación. El entrenamiento del clasificador es específico de la base de datos de referencia y del gen marcador en sí. También existen una serie de clasificadores pre-entrenados de uso público.

Para llevar a cabo la clasificación taxonómica utilizamos el método Classify-Sklearn con un clasificador de Naive-Bayes<sup>37</sup>. Este método es más difícil, requiere algunos pasos de capacitación adicional y unos requerimientos de RAM mayores que los habituales. Por ese motivo se usó una supercomputadora externa. La estación de trabajo en la que se obtuvo el clasificador entrenado fue en el mismo IRTA, concretamente en el centro de Torre Marimon (Barcelona, España). Se empleó una computadora 24 CPU Intel Xeon® E5-2640 2.50 GHz de 32 GB de RAM.

Para la clasificación taxonómica, en primer lugar, obtuvimos e importamos el conjunto de los datos de referencia. Para entrenar el clasificador de Naive-Bayes, se necesitaron dos archivos: (1) las secuencias de referencia y (2) las taxonomías de referencia.

Existen multitud de bases de datos donde podemos adquirir este tipo de archivos, aunque las más utilizadas son SILVA y Greengenes.

Greengenes es una base de datos que incluye información del gen *16S ARNr* de bacterias y arqueas, mientras que SILVA es una base de datos que además de contener información de bacterias y arqueas tiene información del gen 18S que representa a los microorganismos eucariotas (como por ejemplo los protozoos). SILVA también tiene base de datos 23S/28S. Ambas bases de datos contienen taxonomías de referencia que incluyen anotaciones a todos los niveles taxonómicos, incluso a nivel de especie. Sin embargo, una de las limitaciones de la secuenciación del fragmento de un gen, en lugar de un gen completo, es que no se pueden obtener clasificaciones taxonómicas precisas a nivel de especie.

No existe una respuesta clara sobre qué base de datos de referencia elegir, ya que cada una tiene sus ventajas y sus limitaciones, y dependerá del investigador la elección. En este estudio nos decidimos por la base de Greengenes, puesto que los datos estaban formados por secuencias del fragmento 16S. Descargar la base de datos de SILVA requiere más tiempo y más espacio en nuestra computadora, pues contiene más información. La limitación principal de la base de datos que usamos fue que la última versión de Greengenes fue lanzada en agosto del 2013, mientras que SILVA proporciona actualizaciones más regulares<sup>38</sup>. Greengenes se alimenta con datos de bases públicas como NCBI.

Obtenemos la última versión de Greengenes 13\_8 y descargamos los archivos para entrenar al clasificador con las OTU de Greengenes con un 99% de similitud<sup>39</sup>. Una vez importados los archivos de referencia el siguiente paso fue extraer las lecturas de referencia a partir de la base de datos. Se ha demostrado que la precisión de la clasificación taxonómica de las secuencias del gen *16S ARNr* mejora cuando un clasificador de Naive-Bayes se entrena solo con la región de las secuencias objetivo que se secuenciaron<sup>40</sup>. Esto es importante y crucial para obtener una clasificación taxonómica de calidad. En este paso indicamos las secuencias de los cebadores que se usaron para la secuenciación.

V3\_F357\_N: 5'-CCTACGGGNGGCWGCAG

V4\_R805: 5'-GACTACHVGGGTATCTAATCC

Para extraer las secuencias de referencia se indicó que el tamaño a recortar fuese de 500 bp, ya que nuestro amplicón tenía un tamaño inferior. El resultado fue la obtención de aquellas secuencias de la base de datos de referencia basadas en las coincidencias con este par de cebadores. Es fundamental que la secuencia de los cebadores usada para extraer las secuencias de referencia sea únicamente la secuencia real de unión al ADN biológico. No deben contener ninguna secuencia no biológica, como, por ejemplo, secuencias de los adaptadores o del código de barras.

Obtenidas las secuencias de referencia, procedimos a entrenar al clasificador de Naive-Bayes utilizando las lecturas y la taxonomía de referencia. Finalmente, aplicamos el clasificador entrenado a las secuencias representativas. El resultado final fue un artefacto que nos relacionó las secuencias obtenidas con la identidad de los microorganismos presentes en las muestras.

Con la taxonomía obtenida, se realizó un filtraje de la tabla de características, eliminando todas las características anotadas como mitocondrias y cloroplastos.

#### 4.4.8. Generación de un árbol para el análisis de la diversidad filogenética

Aunque el análisis de datos de microbioma se puede analizar sin un árbol filogenético, muchos métodos de análisis de diversidad, como la diversidad filogenética de Faith, requieren uno. Para usar estos métodos debemos construir un árbol filogenético que nos permita considerar la relación evolutiva entre las secuencias de ADN.

QIIME2 ofrece varios métodos para la generación de árboles filogenéticos basados en las características encontradas en los datos. En nuestro caso empleamos el *pipeline* align-to-tree-mafft-fasttree para la generación del árbol con raíz. El *pipeline* realiza una serie de pasos; (1) alinea las secuencias utilizando MAFFT (*Multiple Alignment using Fast Fourier Transform*), (2) elimina las áreas hipervariables que aportan ruido al árbol, (3) construye un árbol filogenético sin raíz para ver la relación entre especies sin un antepasado común y (4) convierte el árbol sin raíz a un árbol con raíz, el cual tiene un ancestro común que conecta todos los organismos. El artefacto del árbol filogenético del tipo Phylogeny[Rooted] se puede visualizar fácilmente usando la herramienta interactiva de QIIME2 q2-empress o con la herramienta online iTOL's<sup>41</sup> (*Interactive Tree Of Life*).

Llegados a este punto teníamos todos los elementos para proceder al análisis estadístico. Es importante destacar que el análisis bioestadístico también se hubiera podido realizar con QIIME2.

Los archivos necesarios para el análisis estadístico fueron los siguientes:

1. La tabla de características (ASV)
2. La tabla de taxonomía obtenida con el clasificador de Naive-Bayes.
3. El árbol filogenético
4. Las secuencias representativas
5. El archivo de metadatos.

#### 4.5. Análisis bioestadístico

El paquete *microeco* tiene muchas ventajas en comparación con otros paquetes de R. Además, con la extensión *file2meco* podemos importar los datos obtenidos con QIIME2 de una forma muy fácil<sup>42</sup>. La Figura 12 nos muestra una visión global sobre el funcionamiento del paquete *microeco*. En ella figura el *input* necesario, el preprocesado de los datos para la creación del objeto de la clase *microtable* y una serie de análisis extendidos. En nuestro caso usamos cuatro de estos análisis extendidos; *trans\_abund* para el análisis de diversidad taxonómica, *trans\_alpha* y *trans\_beta* para el análisis de la alfa y beta diversidad, respectivamente y *trans-diff* para los estudios de abundancia diferencial a diferentes niveles taxonómicos.

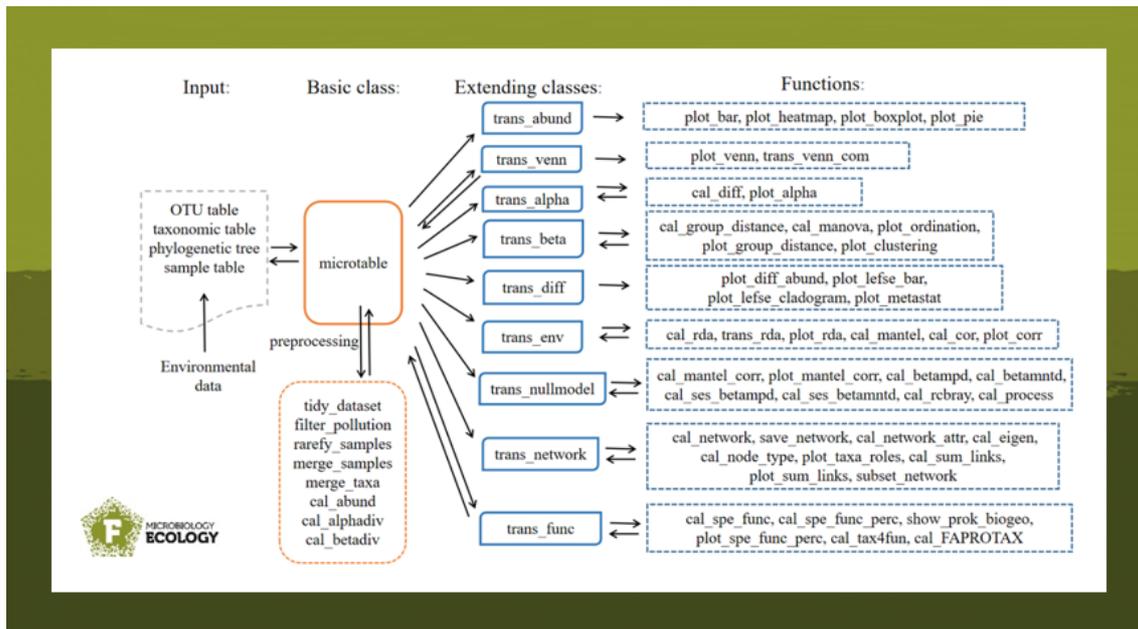


Figura 12. Flujo general de trabajo del paquete *microeco*

##### 4.5.1. Preparación del entorno de trabajo

Siguiendo las instrucciones del tutorial del paquete *microeco*, instalamos todos los paquetes necesarios para poder trabajar, e importamos los archivos necesarios a un directorio de trabajo específico de R Studio<sup>43</sup>.

Una característica relevante que merece ser mencionada es la clase de objeto *microtable*. Este tipo de clase está diseñada para almacenar todos los datos básicos para su posterior análisis.

Creamos el objeto de la clase *microtable*, incluyendo los archivos obtenidos a partir del análisis bioinformático realizado con QIIME2, con la función *qiime2meco* del paquete *file2meco*.

Para que la especie y la información de las muestras fueran consistentes en los diferentes archivos del objeto, usamos la función *tidy\_dataset* para recortar el conjunto de datos. A continuación, eliminamos las ASV que no estaban asignadas al Reino Archaea o Bacteria, y eliminamos las ASV con las asignaciones taxonómicas de mitocondrias y/o de cloroplastos. Este último paso se realizó también con QIIME2. Después, igualamos las ASV en *otu\_table*, *tax\_table* y *phylo\_tree*, con la función *tidy\_dataset*. Finalmente, se comprobó el rango del número de lecturas de las muestras, y obtuvimos el valor mínimo de 13.091 y el máximo de 154.018.

Normalizamos las muestras para que el número de lecturas fuera el mismo en todas ellas, de esta forma se consiguió reducir los efectos de la profundidad de secuenciación en las mediciones de los índices de diversidad. Para normalizar las muestras se usó la función *rarefy\_samples*, indicando que el valor al que se quería normalizar las muestras fuera de 21.640. Este valor se obtuvo durante el análisis con QIIME2 (véase 4.4.6. Curvas de rarefacción).

Calculamos la abundancia de taxones en cada rango taxonómico de interés utilizando la función *cal\_abund*. Esta función nos proporcionó una lista llamada *taxa\_abund* con varios marcos de datos que contenían la información de abundancia en cada rango taxonómico. La lista se almacenó automáticamente en el objeto *microtable*.

A continuación, se calculó la diversidad alfa. El resultado también se almacenó automáticamente en el objeto *microtable*. Por último, calculamos la matriz de distancia de la diversidad beta empleando la función *cal\_betadiv*. Se proporcionaron cuatro índices: Bray-curtis, Jaccard, Unifrac ponderado y Unifrac no ponderado.

En este momento disponíamos de todos los elementos para empezar con los análisis de interés. En este estudio realizamos un análisis taxonómico y un análisis de la diversidad alfa y beta.

#### 4.5.2. Análisis taxonómico

Creamos el objeto *trans\_abund* para poder transformar los datos de abundancia taxonómica en gráficos de abundancia relativa con el paquete *ggplot2*. Nuestro interés era conocer los taxones a rango de familia (véase Figura 13 en el apartado de Resultados) y género (véase Figura 14 en el apartado de Resultados). Finalmente, se representó la composición taxonómica de las muestras de forma individual en un gráfico de barras.

Este tipo de gráficos son interesantes, ya que muestran la abundancia relativa de las características observadas en cada muestra. Además de la representación por muestras, nos interesó graficar la media de las abundancias relativas de cada grupo a nivel de género (véase Figura 15 en el apartado de Resultados). Este gráfico de barras podría ser interesante si al final alguna de las comparaciones resultase significativa.

Se usó una combinación de artefactos de QIIME2 y algunas funciones de R con el fin de obtener los porcentajes de abundancia relativa. A partir de la tabla de características de QIIME2, agrupamos las muestras en función de la columna categórica de interés grupo e indicamos que nos calculara los valores medios de abundancia relativa de las muestras para cada grupo. A continuación, colapsamos la tabla a nivel de género con la función *taxa collapse* de QIIME2. Después exportamos esta tabla a formato *.biom*. Seguidamente, importamos la tabla a R con la función *read\_biom* del paquete *biomformat*, y finalmente, exportamos esa misma tabla a formato *.csv*. De esta forma obtuvimos una tabla con tantas filas como microorganismos identificados a nivel de género, y el conteo de lecturas medio de ese mismo taxón para cada grupo de estudio. A partir de esta tabla se calculó el porcentaje relativo de abundancia de cada uno de los géneros para poder hacer una descriptiva de los datos.

#### 4.5.3. Análisis del índice de diversidad alfa

Se entiende como alfa diversidad el número efectivo de especies presentes en una muestra. La diversidad dentro de una muestra, se puede medir según el número efectivo de especies presentes o según la distancia filogenética que hay entre ellas.

Las métricas de diversidad alfa utilizadas con mayor frecuencia son: (1) la  $S_{obs}$ , métrica la cual cuenta el número de características (OTU o ASV) únicas que se encuentran en las muestras; se trata de una medida cualitativa de la riqueza. (2) El índice de Chao1, una medida cuantitativa del número de especies basada en el número de especies raras en una muestra. (3) El índice Shannon. Una medida cuantitativa de la riqueza, que tiene en cuenta la cantidad de especies que existen en una muestra y la cantidad relativa de individuos que hay para cada una de las especies; contempla la riqueza y la abundancia de las especies y por último (4) la diversidad filogenética de Faith, una medida basada en la longitud total de la rama de un árbol filogenético que comprende una muestra.

Las métricas de diversidad son sensibles a los errores introducidos durante la PCR y la secuenciación<sup>44</sup>.

Los valores de la alfa diversidad se transformaron y graficaron usando un objeto de la clase *trans\_alpha*. La creación del objeto de la clase *trans\_alpha* invocó los datos *alpha\_diversity* del objeto de la clase *microtable* creado anteriormente. El objeto *trans\_alpha* estaba formado por dos marcos de datos, (1) *alpha\_data* y (2) *alpha\_stat*. Este último contiene todos los valores de las métricas de alfa diversidad.

En este estudio usamos el índice de Shannon para el análisis de la alfa diversidad. Este índice se representa como H y se expresa con un número positivo, que en la mayoría de los ecosistemas naturales varía entre 0,5 y 5, aunque su valor normal está entre 2 y 3. Es uno de los índices más empleados para cuantificar la biodiversidad específica. El índice tiene en cuenta la cantidad de especies que existen en una muestra (riqueza) y la cantidad relativa de individuos que hay para cada una de las especies (abundancia).

Para poder ver si nuestros resultados eran significativos, a nivel de alfa diversidad, realizamos una prueba de comparaciones por parejas. Utilizamos la prueba de Kruskal-Wallis como enfoque no paramétrico, ya que nuestros datos no cumplían el supuesto de normalidad (véase Figura 16 en el apartado de Resultados) para poder aplicar un análisis de las varianzas (ANOVA).

Finalmente, representamos los valores medios de abundancia según el índice de Shannon (H) con la ayuda de un diagrama de caja y bigotes (véase Figura 17 en el apartado de Resultados).

#### 4.5.4. Análisis del índice de diversidad beta

La diversidad beta mide las diferencias en la composición o estructura de las comunidades microbianas entre muestras. Estas diferencias se pueden medir de forma cuantitativa o de forma cualitativa.

Las métricas de diversidad beta más utilizadas son la disimilitud de Bray-Curtis y el análisis UniFrac no ponderado. La métrica UniFrac es una métrica filogenética empleada con frecuencia para comparar la historia evolutiva de las diferentes comunidades microbianas. El hecho de ser no ponderada significa que la métrica usa solo la riqueza de las características y no la abundancia. La distancia de Bray-Curtis a partir de una matriz de distancias calcula cómo de diferentes son las muestras entre sí. Esta métrica ópera de manera independiente a los datos de los árboles filogenéticos, mientras que UniFrac es una métrica filogenética.

Transformamos y graficamos los valores de la matriz de distancia de Bray-Curtis usando el objeto de la clase *trans\_beta*. A continuación, visualizamos los resultados con un gráfico de análisis de coordenadas principales (PCoA)<sup>44</sup>.

El análisis de coordenadas principales se utiliza para observar la similitud o disimilitud entre muestras. Cuando más similares sean las muestras respecto a sus características, más cerca entre sí estarán en la gráfica de ordenación de PCoA.

Creamos el objeto *trans\_beta* para invocar la matriz de distancia calculada a partir del índice de disimilitud de Bray-Curtis. Ordenamos la matriz de distancia de Bray-Curtis por el método PCoA y finalmente graficamos los resultados (véase Figura 18 en el apartado de Resultados). También representamos en un diagrama de caja y bigotes los valores medios de la matriz de distancia según la disimilitud de Bray-Curtis en función del grupo experimental (véase Figura 19 en el apartado de Resultados).

Para ver la significancia de la beta diversidad se realizó un PERMANOVA. El PERMANOVA nos permitió medir si los grupos de muestras eran significativamente diferentes entre sí mediante una prueba estadística centrada en las permutaciones<sup>45</sup>, basada en el método Anderson<sup>46</sup> y la función *adonis* del paquete *vegan* de R<sup>47</sup>. La significancia probabilística se estimó usando 999 permutaciones. Este enfoque no requirió ningún supuesto de distribución. Se efectuó un análisis basado en permutaciones del grupo, para ver la significancia estadística del factor grupo, otro análisis entre grupos para ver si alguna de las comparaciones era significativa y, por último, un análisis donde se añadió al modelo el factor tanque para ver la importancia de esta variable en la variabilidad de las muestras.

#### 4.5.5. Análisis de abundancia diferencial

Las pruebas de abundancia diferencial son una parte muy importante en este tipo de análisis. Se utilizan para encontrar taxones significativos en la determinación de las diferencias de la comunidad microbiana entre los distintos grupos.

En el contexto del análisis diferencial de abundancia se corrigió, por el tamaño desigual de las muestras, el número de lecturas y la correlación de la variabilidad de los datos con la media de estos<sup>48</sup>. El objeto de clase *trans\_diff* del paquete *microeco*, tiene cuatro enfoques para hacer este tipo de corrección.

En el presente estudio se usó el enfoque Metastat. Este enfoque se basa en dos supuestos: (1) que los datos corresponden a dos poblaciones diferentes, cada una compuesta por múltiples individuos o muestras y (2) que para cada muestra se proporcionan datos de recuento que representan la abundancia relativa de las características específicas dentro de cada muestra. El *input* corresponde a la matriz de abundancia de características, cuyas filas corresponden a las características y cuyas columnas corresponden a las muestras. El objetivo de este enfoque es distinguir entre las dos poblaciones las características individuales cuya abundancia sea diferente.

Antes de proceder al análisis de la abundancia diferencial Metastat realiza una normalización de los datos. Para tener en cuenta los diferentes niveles de muestreo entre los distintos individuos se realiza una transformación de la medida de la abundancia bruta en una fracción que representa la contribución relativa de cada característica a cada uno de los individuos (o muestras en nuestro caso). Esto da como resultado una versión normalizada de la matriz de abundancia de entidades descrita anteriormente.

Para cada característica  $i$ , compara su abundancia entre las dos poblaciones (o grupos) calculando una prueba  $t$  para dos muestras independientes (para probar si las medias de dos grupos son iguales o no). Específicamente se calcula la proporción de la media y de la varianza para cada uno de los tratamientos (en nuestro caso, el grupo). Después, se calcula el estadístico  $t$  de dos muestras. Se puede concluir que las características cuyo estadístico  $t$  superen un umbral especificado son diferencialmente abundantes en los dos tratamientos (prueba  $t$  de dos colas). Se elige el umbral del estadístico  $t$  para que se minimice el número de falsos positivos (es decir, la determinación incorrecta de características como diferencialmente abundantes cuando realmente no lo son). Se trata de controlar el  $p$ -value (la probabilidad de observar una estadística  $t$  dada por casualidad).

Normalmente, se calcula el  $p$ -value utilizando una distribución  $t$  de Student con unos grados de libertad determinados. Sin embargo, una suposición implícita de este procedimiento es que la distribución subyacente es normal. En este enfoque se estima una distribución nula del estadístico  $t$  de forma no paramétrica empleando un método de permutación, como describen Storey y Tibdhirani <sup>50</sup>

Este procedimiento, también conocido como prueba  $t$  no paramétrica, ha demostrado proporcionar estimaciones precisas cuando las distribuciones no son normales. Específicamente, se permutan de manera aleatoria las etiquetas de los tratamientos (grupos) de las columnas de la matriz de abundancia y se recalculan los estadísticos  $t$ . De esta forma, para cada fila (característica) el  $p$ -value asociado con el estadístico  $t$  observado se contrasta contra la fracción de las pruebas permutadas con una estadística  $t$  más grande o igual al  $t_i$  observado para determinar si el observado es diferente del que esperamos por azar (permutado). El número de permutaciones, a menos que se indique lo contrario, es de 1000. Para reducir el número de falsos positivos, se usa el enfoque FDR (*False Discovery Rate*). Por ese motivo, la importancia de la prueba se mide por el valor  $q$  ( $p$ -value corregido), una medida individual del FDR para cada prueba.

Se creó el objeto `trans_diff` con el método `Metastat` señalando que las comparaciones las realizara a nivel de familia, después se repitió el mismo análisis a nivel de género. De esta manera la función nos calculó la abundancia diferencial en los diez contrastes. La función internamente calculó la media, la varianza y el error estándar de cada uno de los grupos comparados para finalmente proporcionarnos el  $p$ -value y el  $q$ -value. Por último, guardamos los resultados de las tablas de abundancia diferencial en formato `.csv`.

## 5. Resultados

### 5.1. Análisis taxonómico y abundancia diferencial

El análisis del microbioma de las muestras intestinales de corvina se evaluó a los niveles taxonómicos de familia y género. La asignación taxonómica a nivel de familia reveló un total de 123 familias para los cinco grupos experimentales, aunque la Figura 13 solo muestra las 10 familias más abundantes. Si nos fijamos se observa *Others*, que hace referencia a la suma del resto de familias no representadas en la Figura 13. Aunque el número de familias observado fue muy alto, se puede reparar en que la microbiota intestinal de la corvina está dominada principalmente por las familias Mycoplasmataceae y Enterobacteriaceae, excepto en algunas muestras. Se observa además que la abundancia relativa de las familias Bacillaceae y Flavobacteriaceae fue muy variable entre grupos, a pesar de ser dos familias dominantes.

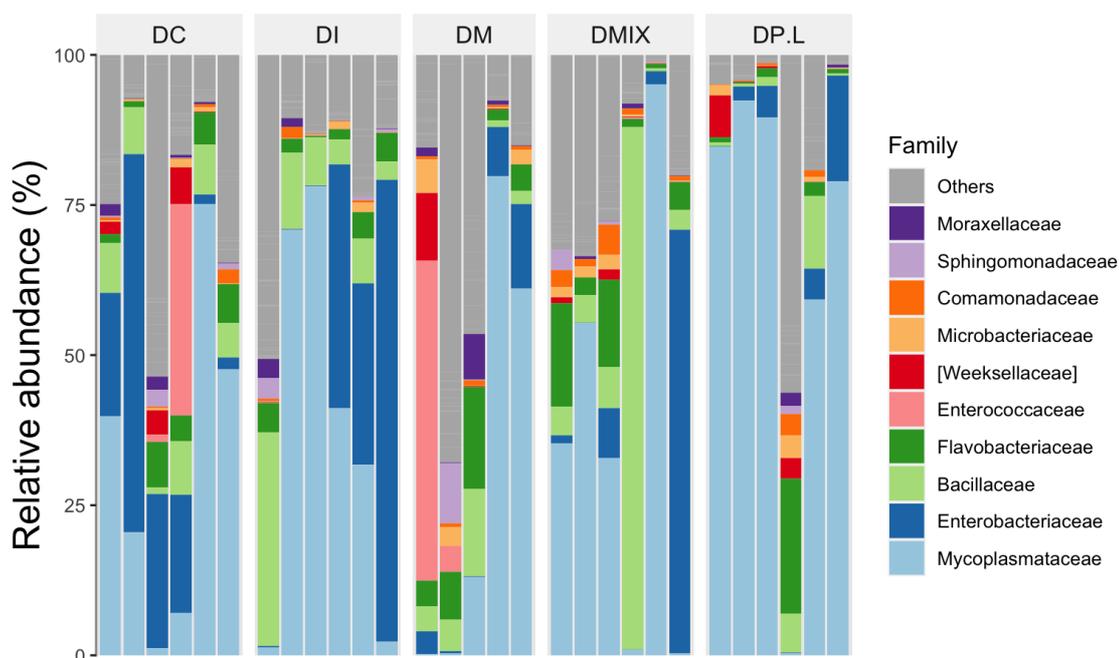


Figura 13. Visualización sobre la composición taxonómica a nivel de familia

En la Figura 14 se representa la asignación taxonómica en función del género. Un total de 179 géneros fueron identificados. Observamos que los principales géneros que dominaron la microbiota intestinal de la corvina fueron *Mycoplasma*, *Enterobacter*, *Bacillus* y *Flavobacterium*.

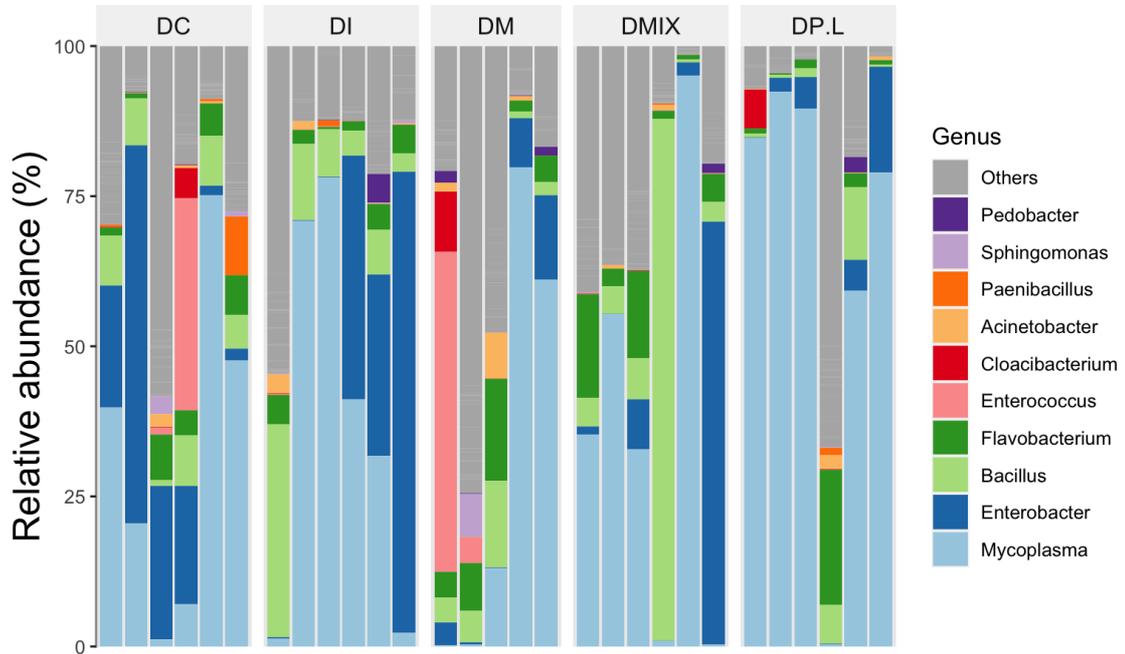


Figura 14. Visualización sobre la composición taxonómica a nivel de género

Con los resultados obtenidos de la tabla de colapso de los géneros (véase Tabla 4), calculamos los porcentajes relativos de abundancia media y contrastamos la información con la Figura 15.

Tabla 4. Contaje de lecturas para cada una de las muestras en función de la taxonomía asignada

taxa_group	DC	DP.L	DI	DMIX	DM
k_Bacteria;p_Tenericutes;c_Mollicutes;o_Mycoplasmatales;f_Mycoplasmataceae;g_Mycoplasma	14387	40411	24722	27701	16494
k_Bacteria;p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae;g_Enterobacter	8239	3249	13475	9072	2845
k_Bacteria;p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Enterococcaceae;g_Enterococcus	3497	26	46	33	5944
k_Bacteria;p_Firmicutes;c_Bacilli;o_Bacillales;f_Bacillaceae;g_Bacillus	3049	2635	7733	5422	5437

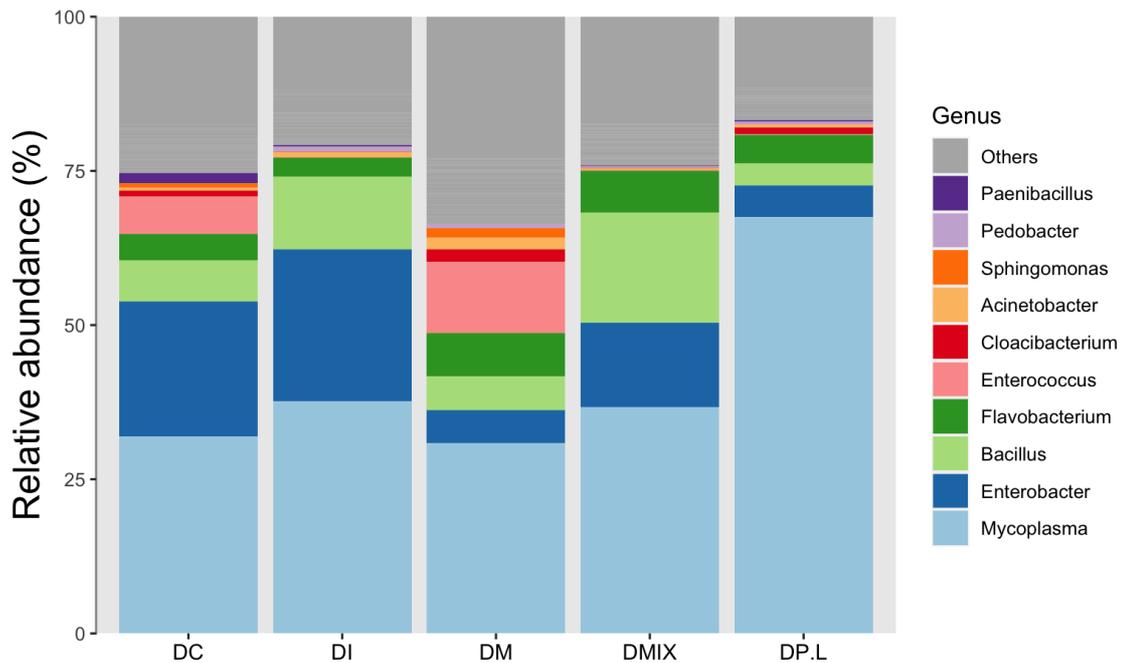


Figura 15. Asignación taxonómica con la abundancia relativa media a nivel de género

En todos los grupos, el principal género que predomina en la microbiota intestinal de la corvina es el género *Mycoplasma*. El género *Enterobacter* está presente en todos los grupos, siendo más abundante en el grupo de peces alimentados con harina de insectos (DI), (21,5%), y menos abundante en el grupo DM (4,3%). Una observación interesante es que el género *Enterococcus* es más abundante al grupo de peces alimentados con microalgas (DM) (8,9%) y en el grupo de peces control (DC) (7,6%), mientras que los otros tres grupos casi no presentan bacterias de este género. Este resultado podría ser de especial relevancia, pues algunas especies del género *Enterococcus* como *Enterococcus faecalis* han sido asociadas con enfermedades de peces<sup>51</sup>. Los géneros *Bacillus* y *Flavobacterium* también están representados en todos los grupos, siendo más abundante el género *Bacillus* en el grupo DI (12,4%), y menos abundante en el grupo DP.L (4,3%). La presencia de bacterias del género *Bacillus* es también de interés, pues un amplio número de especies pertenecientes a este género han demostrado tener propiedades probióticas y moduladoras de la respuesta inmunitaria en peces<sup>52</sup>. Mientras que el género *Flavobacterium* es más abundante en el grupo DM (9,5%), y menos abundante en el grupo DI (3,0%).

Otros géneros con menor abundancia, pero con diferencias entre grupos, fueron las bacterias del género *Staphylococcus* (2,6%) o del género *Paenibacillus* (2,2%) en el grupo DC, a diferencia de los otros grupos que no presentaron bacterias de este género o lo hicieron con un porcentaje muy bajo (<0,2%). De forma similar el grupo DM presentó un porcentaje alto del género *Acinetobacter* (3,2%) y de *Arcobacter* (1,0%), en comparación con el resto de los grupos.

El orden Rickettsiales, es muy abundante en el grupo de peces alimentados con una mezcla de los tres ingredientes principales (microalgas, harina de insectos y restos del enlatado del atún) (DMIX) (2,7%), mientras que el resto de los grupos presentan un porcentaje muy bajo de este tipo de bacterias. Por último, he de comentar que algunos de los taxones solo se han encontrado en algunos grupos en concreto, por ejemplo, miembros del género *Segetibacter* solo están presentes en el grupo DM (1,4%), la familia Rhodobacteraceae en el grupo DI (1,5%), el género *Clostridium* en el grupo DM (1,4%), la familia Parachlamydiaceae en el grupo DMix (1,4%) y el género *Rubrobacter* en el grupo DM (1,2%).

A nivel de familia ninguna de las comparaciones de abundancia diferencial mostró diferencias significativas. El valor mínimo del parámetro *q-value* fue de 0,110. Si establecemos como *q-value* significativo un valor más relajado, *q-value* = 0,20, dos de las comparaciones fueron significativas (Ver Tabla 5).

En el primer contraste se compararon los valores de abundancia para la familia Jonesiaceae entre el grupo DMix y el grupo DM y en el contraste se obtuvo un *q-value* de 0,110. La familia Jonesiaceae del Filo Actinobacteria es común en microbiota humana<sup>53</sup>, no habiéndose detectado antes en especies de peces de cultivo. En términos de abundancia, esta familia no parece estar presente en el grupo DP.L, sin embargo; en el grupo DM presenta una abundancia de  $2,2 \cdot 10^{-4}$ . Lo mismo ocurre en el segundo contraste, DP.L vs. DM, con un *q-value* de 0,13.

Tabla 5. Contrastes de abundancia diferencial a nivel de familia

trans_diff_family										
	compare	taxa	mean(group1)	variance(group1)	std.err(group1)	mean(group2)	variance(group2)	std.err(group2)	pvalue	qvalue
1	DMIX vs DM	f__Jonesiaceae	0,0E+00	0,0E+00	0,0E+00	2,2E-04	2,7E-08	7,3E-05	1,4E-03	1,1E-01
2	DP.L vs DM	f__Jonesiaceae	0,0E+00	0,0E+00	0,0E+00	2,2E-04	2,7E-08	7,3E-05	1,9E-03	1,3E-01

A nivel de género, observamos que ninguna de las comparaciones de abundancia diferencial fue significativa. El valor mínimo del parámetro *qvalue* fue de 0,091. En la Tabla 6 podemos ver los 5 contrastes más significativos con un *qvalue* < 0.3.

Tabla 6. Contrastes de abundancia diferencial a nivel de género

trans_diff_genus										
	compare	taxa	mean(group1)	variance(group1)	std.err(group1)	mean(group2)	variance(group2)	std.err(group2)	pvalue	qvalue
1	DI vs DM	g__Limnohabitans	3,3E-05	6,7E-09	3,3E-05	4,9E-03	1,6E-05	1,8E-03	1,3E-03	8,2E-02
2	DI vs DM	g__Salinisphaera	6,2E-05	9,3E-09	3,9E-05	3,2E-04	5,8E-08	1,1E-04	6,1E-03	1,9E-01
3	DP.L vs DI	g__Ruegeria	0,0E+00	0,0E+00	0,0E+00	3,4E-04	1,2E-07	1,4E-04	1,8E-03	2,0E-01
4	DC vs DI	g__Ruegeria	0,0E+00	0,0E+00	0,0E+00	3,4E-04	1,2E-07	1,4E-04	2,2E-03	2,3E-01
5	DP.L vs DM	g__Limnohabitans	1,1E-04	6,9E-08	1,1E-04	4,9E-03	1,6E-05	1,8E-03	2,5E-03	2,7E-01

En el contraste entre el grupo DI vs. DM pudimos ver que el género *Limnohabitans* fue más abundante en el grupo DM ( $q\text{-value} = 0,082$ ). En ese mismo grupo, en el contraste DI y DM se observó una mayor abundancia del género *Salinisphaera* ( $q\text{-value} = 0,19$ ). En el tercer y cuarto contrastes, se observó que el grupo DI presentaba bacterias del género *Ruegeria*, mientras que el grupo DP.L ( $q\text{value} = 0,2$ ) y DC ( $q\text{-value} = 0,23$ ) no presentaban bacterias de este género. En el último contraste se observó, que el género *Limnohabitans* fue más abundante en el grupo DM que en el grupo DP.L ( $q\text{-value} = 0,27$ ).

Los datos obtenidos de abundancia diferencial no fueron significativos. Una posibilidad para explicar la ausencia de diferencias significativas es que nuestro diseño experimental tuvo una variabilidad muy alta entre los valores de las muestras, además del bajo número de muestras dentro de cada tratamiento.

## 5.2. Índice de diversidad alfa

El paquete *microeco* calcula la diversidad alfa por distintas métricas. En nuestro caso, usamos el índice de Shannon para ver la presencia o ausencia de diferencias significativas de riqueza y abundancia de las muestras.

En la Figura 16 tenemos representados los cuantiles de los datos de la alfa diversidad según el índice de Shannon (H) en función de los cuantiles teóricos. En ella podemos ver como los datos no siguen una distribución normal. Este supuesto se confirmó con una prueba de Shapiro ( $p\text{-value} < 2,2e-16$ ).

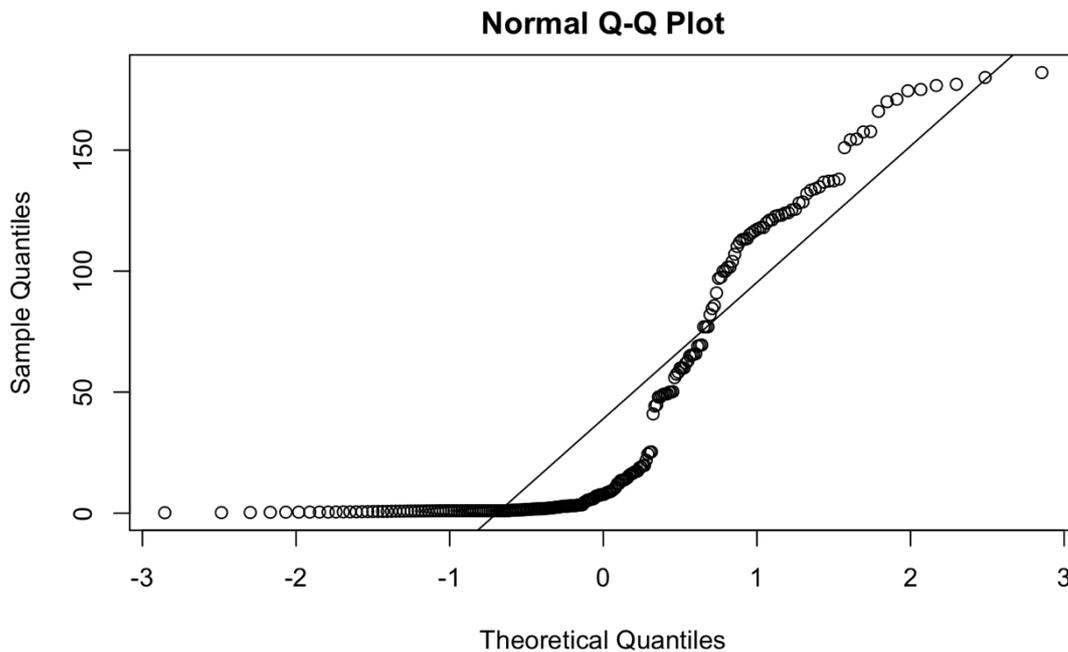


Figura 16. Distribución de los valores de la alfa diversidad

Al no cumplir los supuestos de normalidad, se eligió el método de Kruskal-Wallis para realizar las comparaciones por parejas y ver si existían diferencias significativas de alfa diversidad. Como se puede ver en la Tabla 7, ninguno de los contrastes mostró diferencias significativas ( $p\text{-value} < 0,05$ ) entre los grupos experimentales analizados en el presente estudio, aunque se observa una tendencia no significativa en la disminución de los valores medios de este índice en las dietas experimentales respecto la dieta control.

Tabla 7. *Kruskal-Wallis* de la alfa diversidad con la métrica de Shannon

table_alpha				
Groups	Measure	Test method	p.value	Significance
1	DC vs DP.L	Shannon	KW	0,109
2	DC vs DI	Shannon	KW	0,337
3	DC vs DMIX	Shannon	KW	0,423
4	DC vs DM	Shannon	KW	0,465
5	DP.L vs DI	Shannon	KW	0,200
6	DP.L vs DMIX	Shannon	KW	0,423
7	DP.L vs DM	Shannon	KW	0,100
8	DI vs DMIX	Shannon	KW	0,873
9	DI vs DM	Shannon	KW	0,273
10	DMIX vs DM	Shannon	KW	0,273
11	DC vs DP.L vs DI vs DMIX vs DM	Shannon	KW	0,302

La representación de los valores medios de los índices de alfa diversidad según el índice de Shannon en cada uno de los grupos experimentales se muestra en la Figura 17.

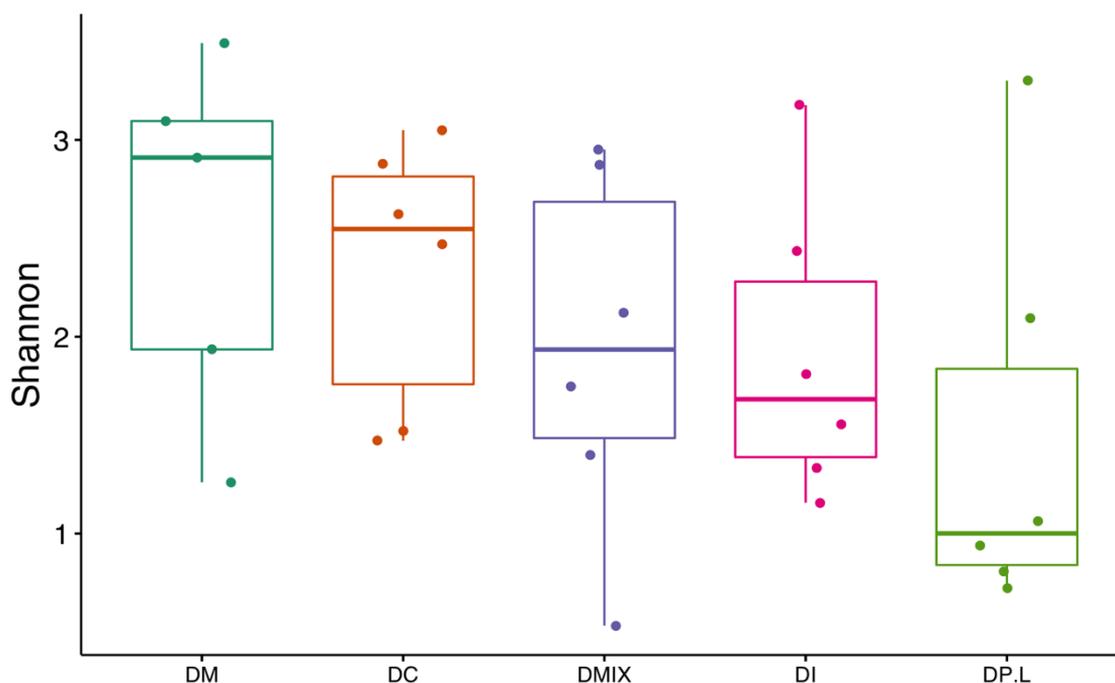


Figura 17. Diagrama de caja y bigotes de la alfa diversidad según la métrica de Shannon

En la Figura 17 vemos en orden descendiente los valores del índice de Shannon en función del grupo experimental. (1) grupo de peces alimentados con microalgas (DM), con un índice de Shannon (H) de  $2,54 \pm 0,91$  (*Mean*  $\pm$  *standard deviation*, SD), (2) grupo de peces control (DC) ( $H = 2,34 \pm 0,68$ ), (3) grupo de animales alimentado con una mezcla de harina de insectos, microalgas y residuos del enlatado del atún (DMix) ( $H = 1,49 \pm 1,02$ ), (4) grupo de peces alimentados con harina de insectos (DI) ( $H = 1,91 \pm 0,76$ ) y (5) grupo, con menos riqueza, de animales alimentados con residuos del enlatado del atún (DP.L) (H

=  $1,48 \pm 1,01$ ). Se observa que el grupo que presenta una mayor riqueza de especies es el grupo de peces alimentados con microalgas (DM).

De los resultados de contrastes de Kruskal-Wallis y el diagrama de cajas y bigotes se concluyó que no existen diferencias significativas de riqueza y abundancia entre las muestras de los intestinos de corvina cuando se considera la dieta como factor de variabilidad ( $p\text{-value} = 0,302$ ).

Los datos obtenidos del índice de Shannon que nos proporcionó la herramienta *microeco* y los representados en la Figura 17 difieren de valores. A pesar de ello, los valores numéricos nos ayudaron a interpretar la Figura 17.

### 5.3. Índice de diversidad beta

Los resultados de la diversidad beta se pueden visualizar a través del gráfico generado por el análisis de coordenadas principales (Figura 18). Este gráfico está formado por distintos elementos, el eje de las X (PCo 1 = 35,4%) que representa el primer eje principal que puede distinguir todas las muestras y puede explicar el 35.4% de las diferencias entre las muestras. Y el eje de las Y, (PCo 2 = 17,%) que representa el segundo eje de coordenadas principal que puede distinguir todas las muestras, lo que puede explicar el 17,5% de la variación. En la Figura 18 podemos observar que las muestras no siguen ningún criterio de ordenación que nos ayude a interpretar la variabilidad de las muestras.

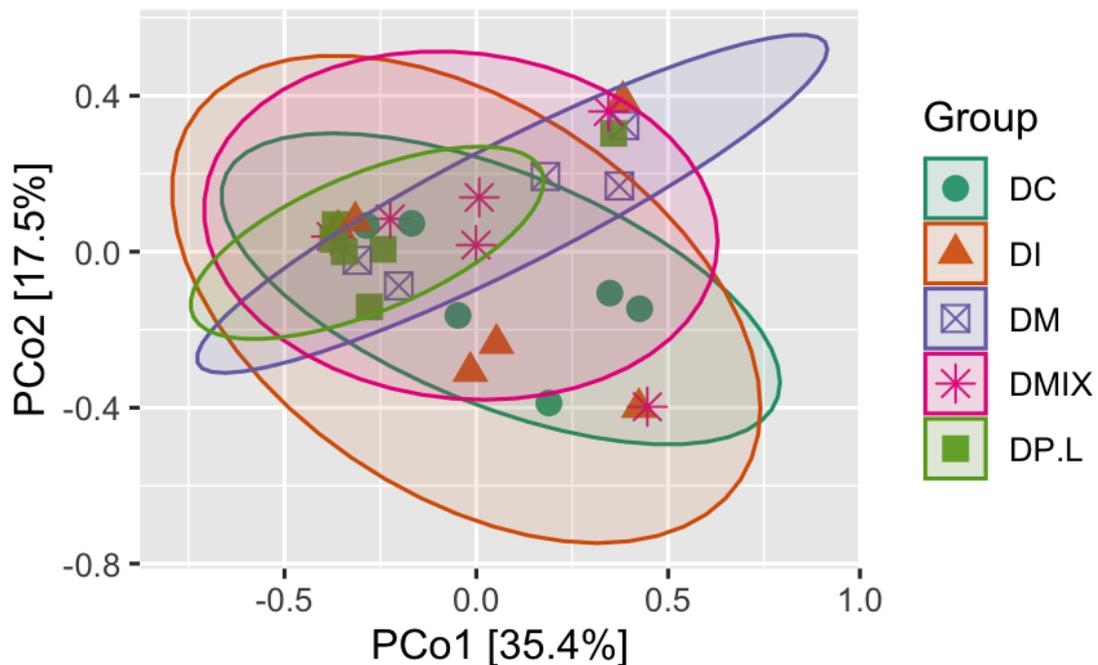


Figura 18. Gráfico PCoA de la beta diversidad con la métrica de Bray-Curtis

En la Figura 19 podemos observar el diagrama de caja y bigotes resultante de la comparación de medias de la matriz de distancia por la métrica de Bray-Curtis. Según la Figura 19 observamos que los grupos de peces alimentados con microalgas (DM) y los peces alimentados con una mezcla de harina de insectos, microalgas y residuos del enlatado del atún (DMix) son los que presentan una mayor distancia respecto al grupo control DC.

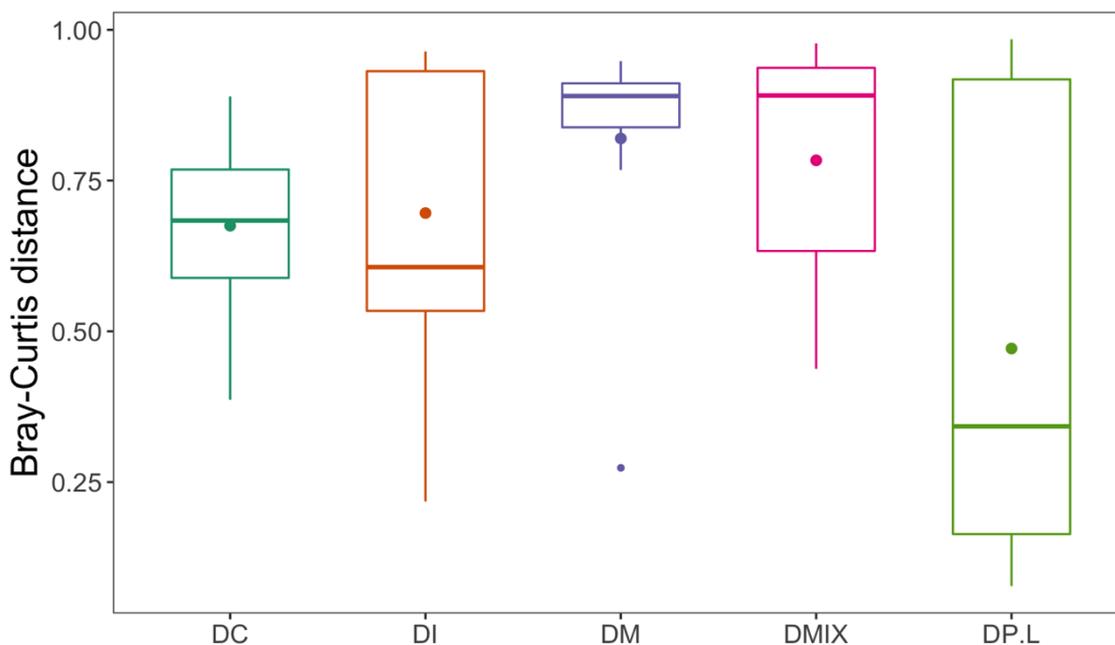


Figura 19. Diagrama de caja y bigotes de la beta diversidad según las distancias medias de Bray-Curtis

Según se muestran los resultados en la Tabla 8 y basándonos en la distancia de Bray-Curtis, la prueba PERMANOVA muestra la ausencia de efectos significativos de la variable grupo. Los grupos alimentados con distintas fuentes de proteína no difieren en la composición de la microbiota ( $p\text{-value} = 0,575$ ), es decir; el grado de diferenciación entre comunidades biológicas entre muestras no es significativo, o lo que sería lo mismo, nos indica que los cinco grupos de interés no presentan diferencias significativas en la estructura de las comunidades biológicas que los forman.

Tabla 8. PERMANOVA de la variable grupo

permanova_group						
	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
Group	4	0,985	0,246	0,915	0,132	0,575
Residuals	24	6,458	0,269	NA	0,868	NA
Total	28	7,443	NA	NA	1,000	NA

El PERMANOVA entre grupos, como se muestra en la Tabla 9, no muestra diferencias significativas.

Tabla 9. PERMANOVA entre grupos

permanova_between_group						
	Groups	measure	permutations	R2	p.value	Significance
1	DC vs DP.L	bray	999	0,162	0,115	
2	DC vs DI	bray	999	0,042	0,950	
3	DC vs DMIX	bray	999	0,066	0,731	
4	DC vs DM	bray	999	0,085	0,594	
5	DP.L vs DI	bray	999	0,117	0,247	
6	DP.L vs DMIX	bray	999	0,093	0,400	
7	DP.L vs DM	bray	999	0,133	0,202	
8	DI vs DMIX	bray	999	0,034	0,897	
9	DI vs DM	bray	999	0,094	0,451	
10	DMIX vs DM	bray	999	0,069	0,881	

La Tabla 10 muestra los resultados de la prueba PERMANOVA del modelo que incluye la variable tanque. Como se puede observar, el *p-value* de la variable tanque (*p-value* = 0,090) mostró una tendencia más evidente de la variabilidad que la correspondiente al tratamiento (*p-value* = 0,398), lo cual sugiere un efecto del hábitat común de los peces sobre la composición de la microbiota.

Tabla 10. PERMANOVA de la variable grupo y tanque

permanova_group_tank						
	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
Group	4	0,985	0,246	1,059	0,132	0,398
Tank	10	3,202	0,320	1,377	0,430	0,090
Residuals	14	3,256	0,233	NA	0,437	NA
Total	28	7,443	NA	NA	1,000	NA

El efecto tanque es una variable complicada de controlar, aunque cada tanque está conectado al mismo sistema de recirculación, es difícil controlar de forma individual la cantidad de oxígeno exacta que llega a cada tanque, el caudal, así como los niveles de los parámetros químicos como son la salinidad, los niveles de nitratos y nitritos, y los *biofilms* bacterianos que pueden crecer en el interior de las tuberías que abastecen con agua a cada uno de los tanques, pudiendo todas estas variables afectar de forma no controlada la comunidad bacteriana intestinal de los peces objeto de estudio.

## 6. Discusión

En el campo de la acuicultura el análisis del microbioma ha ganado una notable importancia en los últimos años. Cada vez existen más estudios que exploran la microbiota de los peces y su relación con caracteres de interés para el sector de la acuicultura. La corvina es una de las nuevas especies que se está introduciendo en los últimos años en el sector acuícola para diversificar el tipo de especies cultivadas. Existen diversos estudios sobre cómo el manejo afecta a su bienestar o sobre el mantenimiento de esta especie en granjas<sup>54-57</sup>, pero su microbiota intestinal todavía no ha sido estudiada. Este estudio constituye una primera investigación de la microbiota intestinal de la corvina. El objetivo principal es la evaluación del efecto modulador de la dieta sobre la composición y diversidad de la microbiota intestinal en corvina en condiciones de cultivo.

En este estudio, el análisis taxonómico del microbioma de la corvina mostró un total de 123 familias de las cuales las más abundantes fueron Mycoplasmataceae, Enterobacteriaceae, Bacillaceae y Flavobacteriaceae. Se observaron un total de 179 géneros siendo los más abundantes los géneros *Mycoplasma*, *Enterobacter*, *Bacillus* y *Flavobacterium*. Las pruebas de abundancia diferencial mostraron que ninguna de las abundancias relativas de los diferentes taxones era significativa para ninguno de los grupos. El género *Mycoplasma*, fue el más predominante en todos los grupos. Concretamente, el grupo de peces alimentados con residuos del enlatado del atún (DP.L) mostró una mayor abundancia (65,7%) en comparación con el resto. Al igual que el género *Enterobacter*, se mostró más abundante al grupo de peces alimentados con harina de insectos (DI) con una abundancia relativa del 21,5%. El género *Bacillus* se representó en todos los grupos, pero el grupo DI fue el que presentó una mayor abundancia de este género (12,4%). Y, por último, el género *Flavobacterium* mostró una mayor abundancia al grupo de peces alimentados con microalgas(DM), con un porcentaje de abundancia relativa del 9,5%.

En cuanto a la alfa diversidad, según el índice de Shannon no se observaron diferencias significativas de riqueza y diversidad en las muestras. A pesar de que el grupo de peces alimentado con microalgas (DM) fue el que mostró un mayor índice de diversidad H, y, por tanto, una mayor riqueza de especies ( $H = 2,54 \pm 0,91$ ). El grupo alimentado con los residuos del enlatado del atún (DP.L) mostró el índice de Shannon más bajo ( $H=1,48 \pm 1,01$ ). La prueba de Kruskal-Wallis manifestó que ninguna de las diferencias entre grupos fue significativa con un  $p\text{-value} = 0,302$ .

En el análisis del índice de beta diversidad según las distancias de Bray-Curtis concluyó que ninguna de las comparaciones entre grupos era estadísticamente significativa según la composición de la estructura microbiana. La variable tanque mostró una mayor significancia sobre la variabilidad de las muestras ( $p$ -value = 0,090) que la propia variable de interés grupo, con un  $p$ -value = 0,398.

En el presente trabajo, el número de muestras por grupo experimental ( $n=6$ ) recogidas fue muy bajo, lo cual unido a la alta variabilidad entre individuos, podrían explicar la ausencia de diferencias estadísticamente significativas en el análisis. No obstante, los resultados sugieren una tendencia de mayor diversidad a favor de la dieta basada en microalgas. Esto se podría explicar debido al efecto de las microalgas y la presencia en las mismas de bacterias simbiotes que podrían aumentar este valor de índice de riqueza (H). En el caso de la dieta a base de harina de grillos era de esperar una mayor abundancia, ya que la harina de insectos contiene grandes cantidades de fibra que ayudan a promover la aparición de bacterias intestinales, según han demostrado algunos estudios<sup>58</sup>. En todo caso y de forma preliminar, estos resultados sugieren que el reemplazo parcial de la harina de pescado, fuente principal de proteína en la dieta, por fuentes alternativas como las harinas de insectos, de microalgas y de subproductos de pesca, no afectó a la composición de la microbiota intestinal de la corvina, resultados que pueden considerarse como positivos con relación al mantenimiento de la diversidad de la comunidad bacteriana en el intestino.

Además, es importante recalcar que los resultados adicionales del estudio anterior, donde se realizó la comparación de los distintos ingredientes alternativos en la dieta para evaluar los efectos de rendimiento, crecimiento, digestibilidad y calidad muscular, indicaron que las dietas con harina del insecto *Archeta domesticus* (DI), con residuos del enlatado del atún (DP.L) y con microalgas marinas (DM) no mostraron efectos negativos sobre los parámetros estudiados. Este dato, junto a los resultados de este estudio, sugieren que el reemplazo parcial de la harina de pescado por estas fuentes alternativas no altera el microbioma intestinal de la corvina y tampoco los índices somáticos.

Estos nuevos ingredientes alternativos estudiados muestran un mayor grado de sostenibilidad, ya que presentan una proporción de pescado requerido: pescado obtenido (FIFO) más baja que la dieta control utilizada en el estudio. Estos resultados son bastante prometedores porque integran la eficiencia zootécnica con la sostenibilidad ambiental.

## 7. Conclusiones

### 7.1. Conclusiones

En el presente estudio la composición, riqueza y diversidad microbioana de la corvina (*Argyrosomus regius*) ha sido caracterizada. La microbiota intestinal de la corvina está dominada por 123 familias, de las cuales, las más abundantes son *Mycoplasmataceae*, *Enterobacteriaceae*, *Bacillaceae* y *Flavobacteriaceae*, con una alta prevalencia de los géneros *Mycoplasma*, *Enterobacter*, *Bacillus* y *Flavobacterium*. En el análisis de la diversidad alfa, según el índice de Shannon, y en el análisis de la diversidad beta, según la métrica de Bray-Curtis, los resultados no mostraron un efecto claro de la dieta sobre la caracterización de la microbiota intestinal.

La ausencia de diferencias significativas de composición, riqueza y diversidad, en los diferentes grupos podría ser debido a un número bajo de muestras recogidas y la variabilidad entre ellas. Lo aconsejable son 10 muestras por tratamiento<sup>59</sup>. Otra causa podría ser que el porcentaje de sustitución de los ingredientes alternativos en la dieta de la corvina no fue lo suficientemente grandes como para modificar su microbiota intestinal. También tiene un papel importante la base de datos usada para la clasificación taxonómica. En este estudio se empleó la base de datos Greengenes, que no se actualiza desde del 2013.

Poco se conoce sobre el microbioma de los peces. Podría probarse si con una base de datos más actualizada se obtienen diferencias entre los tratamientos, ya que en los gráficos de abundancia hay un elevado porcentaje de especies que se clasifican como *Others*, puesto que no se pueden identificar. Una alternativa hubiera sido usar la base de datos SILVA, pero por cuestiones de tiempo no se ha podido probar.

### 7.2. Líneas de futuro

Una estrategia futura, para ver diferencias significativas, en caso de que existieran, podría ser aumentar el tamaño de animales muestreados por grupo, es decir, de cada tanque muestrear entre 10-15 animales. Para calcular el tamaño muestral que garantice una potencia dada, por ejemplo del 95%, podríamos utilizar alguna función de cálculo de tamaño muestral, como puede ser `power.anova.test`. De esta forma conseguiríamos minimizar la variabilidad entre muestras y también reducir el efecto del factor tanque. La variable tanque es una variable muy difícil de controlar y muchas veces es más relevante que la misma variable de interés, como ha pasado en este estudio.

En la estrategia futura sumaría el cambio de base de datos a SILVA para la clasificación taxonómica, puesto que es una base de datos más actualizada.

Otra estrategia interesante complementaria a la anterior podría ser la comparación de la microbiota intestinal de la corvina entre animales de cultivo y animales salvajes de la misma especie. Como la variabilidad en animales salvajes es muy grande, y muchas de las variables no las podemos controlar, se tendrían que simular al máximo las condiciones salvajes en las instalaciones.

Esta comparación se podría conseguir alimentando a los grupos con las dietas incluidas en este estudio y un grupo experimental salvaje alimentado con una dieta salvaje típica de la corvina como crustáceos, moluscos, gusanos y peces. Además, sería importante, secuenciar el microbioma de los animales usados como dieta para diferenciar la microbiota autóctona de la microbiota alóctona.

### 7.3. Seguimiento de la planificación

Se ha seguido de forma constante la planificación y metodología propuesta. Solo un inconveniente puso en riesgo la cronología del estudio. La falta de memoria RAM de la computadora usada hizo que fuese necesario el lanzamiento de uno de los *scripts* en una supercomputadora externa.

## 8. Glosario

**ADN:** ácido desoxirribonucleico, es una molécula que contiene la información genética en todos los seres vivos.

**ANOVA:** modelos estadísticos que analiza la varianza, muy utilizados en el análisis y diseño de experimentos para evaluar el efecto de los tratamientos en la variabilidad de la variable respuesta.

**ARNr:** el ácido ribonucleico ribosómico o ribosomal es un ácido ribonucleico (ARN) que forma parte de los ribosomas y es esencial para la síntesis proteica en todos los seres vivos.

**ARNr 16S:** componente de la subunidad pequeña (30S) del ribosoma procariota codificado por el gen ARNr 16S, muy empleado para la reconstrucción de filogenias debido a sus bajas tasas de evolución<sup>60</sup>.

**ASV:** la variante de secuencia de amplicón es un término para referirse a secuencias de ADN individuales recuperadas de un análisis de genes marcadores de alto rendimiento.

**BLAST:** programa informático de alineamiento de secuencias de tipo local ya sea de secuencias de ADN o de secuencias de aminoácidos.

**DADA2:** complemento de control de calidad de las secuencias para estudiar las variantes exactas de las secuencias de amplicón en los datos de secuenciación de alto rendimiento.

**DC:** dieta control, peces alimentados con pienso comercial.

**DI:** dieta a base de harina de insectos, con un 15% de harina de *Acheta domesticus*.

**DIBAQ:** empresa dedicada a la fabricación de piensos para mascotas y peces de acuicultura.

**DM:** dieta de microalgas (DM) que contenía 10% de una mezcla de cuatro microalgas marinas (*Nannochloropsis gaditana*, *Tisochrysis lutea* (CCAP 927/14), *Rhodomonas lens* (ECC030) y *Isochrysis galbana* (CCAP927/1)).

**DMix:** dieta basada en la inclusión de tres ingredientes. 10% de harina de microalgas (*Nannochloropsis gaditana*, *Tisochrysis lutea* (CCAP 927/14), *Rhodomonas lens* (ECC030) y *Isochrysis galbana* (CCAP927/1)), 15% de harina de insectos (*Acheta domesticus*) y 2% y 9.4% de fracción de proteínas y lípidos del proceso de fabricación del atún.

**DP & L / DP.L:** dieta con un 7% de proteína y 11% de aceite proveniente del de agua de cola (TWC) del enlatado del atún.

**FDR:** la tasa de falsos descubrimientos es una predicción estadística de cuántos resultados se pueden esperar que sean falsos positivos.

**IRTA:** el Instituto de Investigación y Tecnología Agroalimentarias es una empresa pública de la Generalidad de Cataluña. La finalidad del IRTA es contribuir a la modernización, a la mejora y al impulso de la competitividad, al desarrollo sostenible de los sectores agrario, alimentario, agroforestal, acuícola y pesquero, y también de los directa e indirectamente relacionados con el suministro de alimentos sanos y de calidad a los consumidores finales; a la seguridad alimentaria y a la transformación de los alimentos y, en general, a la mejora del bienestar y la salud de la población.

**MAFFT:** es un programa utilizado para crear alineaciones de múltiples secuencias de aminoácidos o nucleótidos.

**MD5:** es un algoritmo de reducción criptográfico de 128 bits ampliamente usado.

**Metastat:** método estadístico para comparar muestras de dos poblaciones de tratamiento sobre el archivo de recuento para detectar características abundantes de forma diferencial. El método emplea la tasa de falso descubrimiento para mejorar la especificidad en entornos de alta complejidad y maneja por separado características menos abundantes usando la prueba exacta de Fisher.

**MS-222:** la triclaína metanosulfanato, es un anestésico ampliamente empleado en acuicultura y que está regulado para su uso en instalaciones de producción en la Unión Europea y Estados Unidos.

**NGS:** la secuenciación de nueva generación (NGS, por sus siglas en inglés) se refiere a maneras nuevas y más rápidas de secuenciar el ADN y el ARN que están revolucionando de manera eficaz la genómica y la biología molecular.

**OTU:** las unidades taxonómicas operativas (OTU) habían sido la unidad estándar para el análisis de genes marcadores. Las OTU se crean mediante la agrupación basada en similitudes entre las lecturas de secuenciación o agrupando bases de datos de referencia para definir y etiquetar una OTU.

**PCoA:** el análisis de coordenadas principales es un método para explorar y visualizar similitudes o diferencias de datos. Comienza con una matriz de similitud o disimilitud y asigna a cada elemento una ubicación en un espacio de baja dimensión.

**PCR:** la reacción en cadena de la polimerasa es una técnica de biología molecular que permite obtener un gran número de copias de un fragmento de ADN particular, partiendo de un mínimo.

**PERMANOVA:** el análisis de varianza multivariado con permutaciones usando matrices de distancia realiza comparaciones de medias entre dos o más grupos (comunidades), es decir, prueba la hipótesis de igualdad de medias multivariada entre grupos y su significancia es evaluado con permutaciones.

**QIIME2:** son las iniciales de *Quantitative Insights Into Microbial Ecology*, se trata de un paquete de análisis de microbioma potente, que permite analizar los datos de secuencias de ADN y obtener cifras de calidad de publicación y resultados estadísticos.

**RAS:** se trata de sistemas de recirculación en acuicultura. Estos sistemas pueden ser muy variados, pero todos ellos permiten el control de ciertos parámetros fisicoquímicos como la temperatura, la salinidad, el oxígeno, etc.

**SILVA:** es una base de datos de ARN ribosomal con más de 4000000 de secuencias de la subunidad pequeña del ribosoma (16S/18S) y 400000 de la subunidad grande (23S/28S).

**TWC:** el agua de cola se trata de un subproducto obtenido de centrifugar el caldo de prensa en la industria de la harina de pescado.

## 9. Bibliografía

1. Estado Mundial, E. LA SOSTENIBILIDAD EN ACCIÓN. doi:10.4060/ca9231es.
2. Stentiford, G. D. *et al.* Sustainable aquaculture through the One Health lens. *Nature Food* vol. 1 468–474 (2020).
3. Gatlin, D. M. *et al.* Expanding the utilization of sustainable plant products in aquafeeds: A review. *Aquaculture Research* vol. 38 551–579 (2007).
4. Bandara, T. & Tharindu Bandara, C. Alternative feed ingredients in aquaculture: Opportunities and challenges. ~ 3087 ~ *Journal of Entomology and Zoology Studies* **6**, (2018).
5. Konar, M. *et al.* Illustrating the hidden economic, social and ecological values of global forage fish resources. *Resources, Conservation and Recycling* **151**, (2019).
6. Naylor, R. L. *et al.* A 20-year retrospective review of global aquaculture. *Nature* vol. 591 551–563 (2021).
7. Kalhor, H. *et al.* Soy protein concentrate as a substitute for fish meal in diets for juvenile *Acanthopagrus schlegelii*: effects on growth, phosphorus discharge and digestive enzyme activity. *Aquaculture Research* **49**, 1896–1906 (2018).
8. Henry, M., Gasco, L., Piccolo, G. & Fountoulaki, E. Review on the use of insects in the diet of farmed fish: Past and future. *Animal Feed Science and Technology* vol. 203 1–22 (2015).
9. Biancarosa, I. *et al.* Replacing fish meal with insect meal in the diet of Atlantic salmon (*Salmo salar*) does not impact the amount of contaminants in the feed and it lowers accumulation of arsenic in the fillet. *Food Additives and Contaminants - Part A Chemistry, Analysis, Control, Exposure and Risk Assessment* **36**, 1191–1205 (2019).
10. Sarker, P. K. *et al.* Microalgae-blend tilapia feed eliminates fishmeal and fish oil, improves growth, and is cost viable. *Scientific Reports* **10**, (2020).
11. Sarker, P. K., Kapuscinski, A. R., Vandenberg, G. W., Proulx, E. & Sitek, A. J. Towards sustainable and ocean-friendly aquafeeds: Evaluating a fish-free feed for rainbow trout (*Oncorhynchus mykiss*) using three marine microalgae species. *Elementa* vol. 8 (2020).
12. Fao. *GLOBEFISH RESEARCH PROGRAMME Food and Agriculture Organization of the United Nations Products By-products of tuna processing.* (2013).
13. Cottrell, R. S., Blanchard, J. L., Halpern, B. S., Metian, M. & Froehlich, H. E. Global adoption of novel aquaculture feeds could substantially reduce forage fish demand by 2030. *Nature Food* **1**, 301–308 (2020).
14. Bates, J. M. *et al.* Distinct signals from the microbiota promote different aspects of zebrafish gut differentiation. *Developmental Biology* **297**, 374–386 (2006).
15. Fraune, S. & Bosch, T. C. G. Why bacteria matter in animal development and evolution. *BioEssays* vol. 32 571–580 (2010).
16. Stephens, W. Z. *et al.* The composition of the zebrafish intestinal microbial community varies across development. *ISME Journal* **10**, 644–654 (2016).
17. Egerton, S., Culloty, S., Whooley, J., Stanton, C. & Ross, R. P. The gut microbiota of marine fish. *Frontiers in Microbiology* vol. 9 (2018).
18. Wang, A. R., Ran, C., Ringø, E. & Zhou, Z. G. Progress in fish gastrointestinal microbiota research. *Reviews in Aquaculture* vol. 10 626–640 (2018).

19. Han, S. *et al.* Analysis of bacterial diversity in the intestine of grass carp (*Ctenopharyngodon idellus*) based on 16S rDNA gene sequences. *Aquaculture Research* **42**, 47–56 (2010).
20. Hansen, G. H. & Olafsen, J. A. Bacterial interactions in early life stages of marine cold water fish. *Microbial Ecology* vol. 38 1–26 (1999).
21. Dehler, C. E., Secombes, C. J. & Martin, S. A. M. Seawater transfer alters the intestinal microbiota profiles of Atlantic salmon (*Salmo salar* L.). *Scientific Reports* **7**, (2017).
22. Bledsoe, J. W., Peterson, B. C., Swanson, K. S. & Small, B. C. Ontogenetic characterization of the intestinal microbiota of channel catfish through 16S rRNA gene sequencing reveals insights on temporal shifts and the influence of environmental microbes. *PLoS ONE* **11**, (2016).
23. Grigorakis, K., Fountoulaki, E., Vasilaki, A., Mittakos, I. & Nathanailides, C. Lipid quality and filleting yield of reared meagre (*Argyrosomus regius*). *International Journal of Food Science and Technology* **46**, 711–716 (2011).
24. Estévez, A., Blanco, B., Fernández, L., Ferreira, M. & Soula, M. Effects of alternative and sustainable ingredients, insect meal, microalgae and protein and lipid from tuna cooking water, on meagre (*Argyrosomus regius*) growth, food conversion and muscle and liver composition. *Aquaculture* **548**, 737549 (2022).
25. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* vol. 37 852–857 (2019).
26. Documento de usuario de QIIME 2. 1 Introducción e instalación (2019.7) - programador clic. <https://programmerclick.com/article/83272105544/>.
27. microeco | Un paquete R para el análisis de datos en ecología de la comunidad microbiana. <https://chiliubio.github.io/microeco/>.
28. Liu, C., Cui, Y., Li, X. & Yao, M. microeco: an R package for data mining in microbial community ecology. *FEMS Microbiology Ecology* **97**, (2021).
29. Estaki, M. *et al.* QIIME 2 Enables Comprehensive End-to-End Analysis of Diverse Microbiome Data and Comparative Studies with Publicly Available Data. *Current Protocols in Bioinformatics* **70**, e100 (2020).
30. QIIME 2 View. <https://view.qiime2.org/>.
31. Bokulich, N. A. *et al.* Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods* **10**, 57–59 (2013).
32. Rideout, J. R. *et al.* Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ* **2014**, e545 (2014).
33. Calculating minimum frequency count - User Support - QIIME 2 Forum. <https://forum.qiime2.org/t/calculating-minimum-frequency-count/6205>.
34. Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, (2017).
35. Training feature classifiers with q2-feature-classifier — QIIME 2 2021.8.0 documentation. <https://docs.qiime2.org/2021.8/tutorials/feature-classifier/>.
36. “Moving Pictures” tutorial — QIIME 2 2021.8.0 documentation. <https://docs.qiime2.org/2021.8/tutorials/moving-pictures/>.
37. What is the difference between Greengenes and SILVA? - User Support - QIIME 2 Forum. <https://forum.qiime2.org/t/what-is-the-difference-between-greengenes-and-silva/1560/3>.
38. Data resources — QIIME 2 2021.8.0 documentation. <https://docs.qiime2.org/2021.8/data-resources/>.

39. Werner, J. J. *et al.* Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. *ISME Journal* vol. 6 94–103 (2012).
40. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research* **47**, W256–W259 (2019).
41. GitHub - ChiLiubio/file2meco: Tranform files to the microtable object in microeco package. <https://github.com/ChiLiubio/file2meco>.
42. Tutorial for R microeco package (v0.5.0). [https://chiliubio.github.io/microeco\\_tutorial/](https://chiliubio.github.io/microeco_tutorial/).
43. El microbioma humano | Harrison. Principios de Medicina Interna, 19e | AccessMedicina | McGraw Hill Medical. <https://accessmedicina.mhmedical.com/content.aspx?bookid=1717&sectionid=114913448#1137919660>.
44. Anderson, M. J. A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26**, 32–46 (2001).
45. Anderson, B. T., Salvucci, G., Ruane, A. C., Roads, J. O. & Kanamitsu, M. NOTES AND CORRESPONDENCE A New Metric for Estimating the Influence of Evaporation on Seasonal Precipitation Rates. (2008) doi:10.1175/2007JHM968.1.
46. adonis function - RDocumentation. <https://www.rdocumentation.org/packages/vegan/versions/2.4-2/topics/adonis>.
47. Introducción a phyloseq y a análisis de diversidad. <http://www.castrolab.org/isme/biodiversity/biodiversity.html#analisis-de-abundancias-y-visualizaciones>.
48. White, J. R., Nagarajan, N. & Pop, M. Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. *PLOS Computational Biology* **5**, e1000352 (2009).
49. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440–9445 (2003).
50. Zahran, E., Mahgoub, H. A., Abdelhamid, F., Sadeyen, J. R. & Risha, E. Experimental pathogenesis and host immune responses of *Enterococcus faecalis* infection in Nile tilapia (*Oreochromis niloticus*). *Aquaculture* **512**, (2019).
51. Kuebutornye, F. K. A. *et al.* Mechanisms and the role of probiotic *Bacillus* in mitigating fish pathogens in aquaculture. *Fish physiology and biochemistry* **46**, 819–841 (2020).
52. Microbiota intestinal en salud y enfermedad. <http://www.revistagastroenterologiamexico.org/es-pdf-S037509061200050X>.
53. Asencio-Alcudia, G. *et al.* Stressors Due to Handling Impair Gut Immunity in Meagre (*Argyrosomus regius*): The Compensatory Role of Dietary L-Tryptophan. *Frontiers in physiology* **10**, (2019).
54. Campoverde, C. *et al.* Gene expression analysis of the innate immune system during early rearing and weaning of meagre (*Argyrosomus regius*). *Fish and Shellfish Immunology* **94**, 819–832 (2019).
55. Guardiola, F. A. *et al.* Humoral and mucosal immune responses in meagre (*Argyrosomus regius*) juveniles fed diets with varying inclusion levels of carob seed germ meal. *Fish and Shellfish Immunology* **79**, 209–217 (2018).
56. Campoverde, C. *et al.* Ontogeny and modulation after PAMPs stimulation of  $\beta$ -defensin, hepcidin, and piscidin antimicrobial peptides in meagre (*Argyrosomus regius*). *Fish and Shellfish Immunology* **69**, 200–210 (2017).

57. Stull, V. J. *et al.* Impact of Edible Cricket Consumption on Gut Microbiota in Healthy Adults, a Double-blind, Randomized Crossover Trial. *Scientific Reports* 2018 8:1 **8**, 1–13 (2018).
58. Predicting sample metadata values with q2-sample-classifier — QIIME 2 2021.11.0 documentation. <https://docs.qiime2.org/2021.11/tutorials/sample-classifier/>.
59. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5088 (1977).

## 10. Anexos

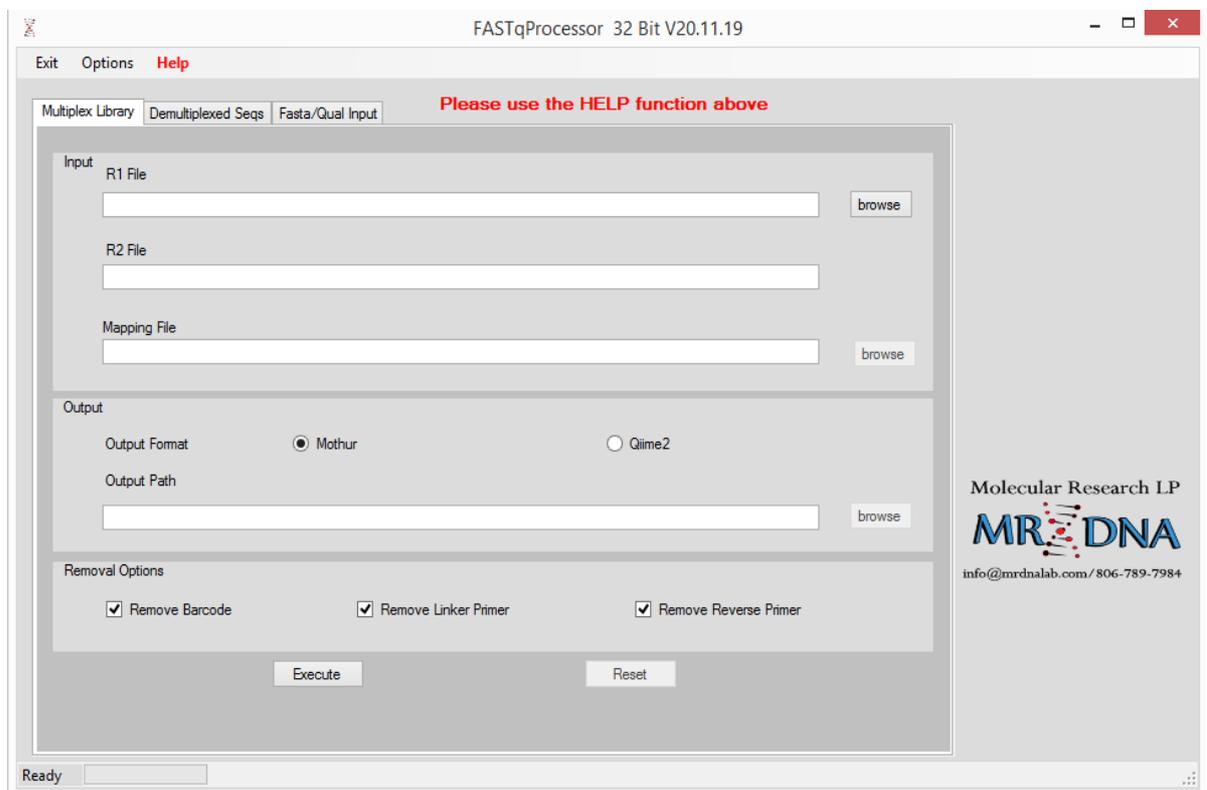
### Anexo 1 – Obtención de los datos con FASTQ Processor

Partimos de los datos crudos que nos proporciona la plataforma online de Illumina®, BaseSpace. Buscamos nuestros datos, que están dentro del proyecto llamado 052421ARwhoi341F-274105858. Encontramos un total de 6 archivos dentro de este proyecto. Contenido de la carpeta del proyecto:

- SAM1-50\_S1\_L001\_R1\_001.fastq
- SAM1-50\_S1\_L001\_R2\_001.fastq
- SAM51-104\_S2\_L001\_R1\_001.fastq
- SAM51-104\_S2\_L001\_R2\_001.fastq
- SAM105-164\_S3\_L001\_R1\_001.fastq
- SAM105-164\_S3\_L001\_R2\_001.fastq

Las muestras de nuestro estudio se encuentran dentro del primer conjunto de muestras (SAM1-50). Vamos a usar los dos primeros archivos, y el archivo que nos envió la empresa de secuenciación MR DNA, SAM1-50-mapping.txt.

Con la ayuda de una herramienta, proporcionada por la misma empresa de secuenciación, **FASTQ processor** (<https://www.mrdnalab.com/mrdnafreesoftware/fastq-processor.html>) y los archivos mencionados anteriormente obtendremos los tres archivos necesarios para empezar a trabajar con QIIME2.



Dentro del procesador FASTQ hay tres ventanas, seleccionamos *Multiplex Library*. En el *Input* tendremos que seleccionar los tres archivos comentados anteriormente, y en el *output* vamos a marcar que el formato de trabajo sea para QIIME2. En *Removal Options* vamos a marcar las tres opciones para que nos elimine la secuencia de los códigos de barras, de los adaptadores y del cebador antisentido.

Finalmente, obtenemos un directorio llamado qiime2 que contiene, la carpeta emp-paired-end-sequences. Dentro de ella se encuentran los siguientes archivos:

- barcodes.fastq.gz
- forward.fastq.gz
- reverse.fastq.gz

Otro archivo muy importante, sample-metadata.tzv. En este archivo está la correlación del nombre nuestras muestras con la identificación individual de cada muestra (BarcodeSequence). Este archivo lo proporciona la misma empresa de secuenciación, pero tendremos que añadir información sobre el diseño experimental para poder ver las diferencias entre los grupos. Se le añaden las columnas Group y Tank:

- Group. Información sobre a que grupo experimental al que pertenecen las muestras:
  - DC (control): peces alimentados con la dieta control
  - DI (insectos): peces alimentados con harina de insectos
  - DM (microalgas): peces alimentados con microalgas
  - DP.L (residuos conservas): peces alimentados con la resta de residuos del proceso de la conserva del atún
  - DMix (mezcla): peces alimentados con una mezcla de los tres anteriores.
- Tank. Unidad experimental, tanque donde se encontraban los animales

## Anexo 2 – Análisis bioinformático con QIIME2

### #Importación de los datos a QIIME2

```
qiime tools import \  
  --type EMPPairedEndSequences \  
  --input-path emp-paired-end-sequences \  
  --output-path emp-paired-end-sequences.qza
```

### #Demultiplexación de las muestras

```
qiime demux emp-paired \  
  --m-barcodes-file sample-metadata.tsv \  
  --m-barcodes-column BarcodeSequence \  
  --i-seqs emp-paired-end-sequences.qza \  
  --o-per-sample-sequences demux-full.qza \  
  --p-no-golay-error-correction \  
  --o-error-correction-details demux-details.qza
```

### #Transformación del artefacto demux-full.qza a demux-full.qzv

```
qiime demux summarize \  
  --i-data demux-full.qza \  
  --o-visualization demux-full.qzv
```

### #Creación del subconjunto de muestras

```
qiime demux filter-samples \  
  --i-demux demux-full.qza \  
  --m-metadata-file per-sample-fastq-counts-filter.tsv \  
  --p-where 'CAST([filter] AS INT) = 1' \  
  --o-filtered-demux demux-filtered.qza
```

### #Control de calidad de las secuencias

```
qiime dada2 denoise-paired \  
  --i-demultiplexed-seqs demux-filtered.qza \  
  --p-trim-left-f 20 \  
  --p-trim-left-r 20 \  
  --p-trunc-len-f 250 \  
  --p-trunc-len-r 240 \  
  --o-table table.qza \  
  --o-representative-sequences rep-seqs.qza \  
  --o-denoising-stats denoising-stats.qza
```

## #Transformación de los artefactos de DADA2 a artefactos visualizables

```
qiime metadata tabulate \  
  --m-input-file denoising-stats.qza \  
  --o-visualization denoising-stats.qzv  
  
qiime feature-table summarize \  
  --i-table table.qza \  
  --o-visualization table.qzv \  
  --m-sample-metadata-file sample-metadata.tsv  
  
qiime feature-table tabulate-seqs \  
  --i-data rep-seqs.qza \  
  --o-visualization rep-seqs.qzv
```

## #Filtraje de los datos

```
qiime feature-table filter-features \  
  --i-table table.qza \  
  --p-min-samples 2 \  
  --p-min-frequency 20 \  
  --o-filtered-table filtered-table_16S.qza  
  
qiime feature-table summarize \  
  --i-table filtered-table_16S.qza \  
  --o-visualization filtered-table_16S.qzv \  
  --m-sample-metadata-file sample-metadata.tsv
```

## #Curvas de rarefacción

```
qiime diversity alpha-rarefaction \  
  --i-table filtered-table_16S.qza \  
  --i-phylogeny rooted-tree.qza \  
  --p-max-depth 65000\  
  --m-metadata-file sample-metadata.tsv \  
  --o-visualization alpha-rarefaction.qzv
```

## #Clasificación taxonómica

```
qiime tools import \  
  --type 'FeatureData[Sequence]' \  
  --input-path 99_otus.fasta \  
  --output-path 99_otus.qza  
  
qiime tools import \  
  --type 'FeatureData[Taxonomy]' \  
  --input-format HeaderlessTSVTaxonomyFormat \  
  --input-path 99_otu_taxonomy.txt \  
  --output-path ref-taxonomy.qza  
  
qiime feature-classifier extract-reads \  
  --i-sequences 99_otus.qza \  
  --p-f-primer CCTACGGGNGGCWGCAG \  
  --p-r-primer GACTACHVGGGTATCTAATCC \  
  --p-max-length 500 \  
  --o-reads ref-seqs.qza  
  
qiime feature-classifier fit-classifier-naive-bayes \  
  --i-reference-reads ref-seqs.qza \  
  --i-reference-taxonomy ref-taxonomy.qza \  
  --o-classifier classifier.qza  
  
qiime feature-classifier classify-sklearn \  
  --i-classifier classifier.qza \  
  --i-reads rep-seqs.qza \  
  --o-classification taxonomy.qza  
  
qiime metadata tabulate \  
  --m-input-file taxonomy.qza \  
  --o-visualization taxonomy.qzv
```

## #Filtraje de la tabla de características, eliminando características asociadas a los taxones cloroplastos y mitocondrias

```
qiime taxa filter-table \  
  --i-table filtered-table_16S.qza \  
  --i-taxonomy taxonomy.qza \  
  --p-exclude mitochondria,chloroplast \  
  --o-filtered-table table-no-mitochondria-no-chloroplast.qza
```

## #Generación de un árbol para el análisis de la diversidad filogenética

```
qiime phylogeny align-to-tree-mafft-fasttree \  
  --i-sequences rep-seqs.qza \  
  --o-alignment aligned-rep-seqs.qza \  
  --o-masked-alignment masked-aligned-rep-seqs.qza \  
  --o-tree unrooted-tree.qza \  
  --o-rooted-tree rooted-tree.qza
```

## #Obtención de los porcentajes de abundancia en función de la variable grupo y exportación de la tabla

```
qiime feature-table group \  
  --i-table table-no-mitochondria-no-chloroplast.qza \  
  --p-axis sample \  
  --m-metadata-file sample-metadata.tsv \  
  --m-metadata-column Group \  
  --p-mode mean-ceiling \  
  --o-grouped-table grouped-table.qza  
  
qiime taxa collapse \  
  --i-table grouped-table.qza \  
  --i-taxonomy taxonomy.qza \  
  --p-level 6 \  
  --o-collapsed-table table-l6.qza  
  
qiime metadata tabulate \  
  --m-input-file table-l6.qza \  
  --o-visualization table-l6.qzv  
  
qiime tools export \  
  --input-path table-l6.qza \  
  --output-path exported-table-l6
```

## Anexo 3 – Análisis bioestadístico con el paquete *microeco* del software R

Cargamos el paquete de interés, y algunos paquetes más que van a ser necesarios para el análisis

```
library(microeco)
library(file2meco)
library(seqinr)
library(dplyr)
```

Definimos el directorio de trabajo, donde se encuentran los archivos obtenidos a partir del análisis bioinformático con QIIME2

```
setwd("/Users/edgar/Master Bioinformatica i Bioestadistica/M0.180-TFM/microeco")
```

Creamos el objeto dataset de la clase *microtable* para poder empezar a trabajar con el paquete *microeco*. Con la función `qiime2meco()` del paquete *file2meco* podemos generar el objeto de la clase *microtable*. Aquí como *output* introducimos la tabla de características.

```
dataset <- file2meco::qiime2meco(ASV_data = "table-no-mitochondria-no-chloroplast.qza", sample_data = "sample-metadata.tsv", taxonomy_data = "taxonomy.qza", phylo_tree = "rooted-tree.qza", rep_fasta = "rep-seq.s.qza")
```

Para que la especie y la información de las muestras sean consistentes en los diferentes archivos del objeto, podemos usar la función `tidy_dataset()` para recortar el conjunto de datos

```
dataset$tidy_dataset()
```

Luego, eliminamos las ASV que no están asignadas al Reino "k\_\_Archaea" o "k\_\_Bacteria"

```
dataset$tax_table %>% base::subset(Kingdom == "k__Archaea" | Kingdom = "k__Bacteria")
```

También nos aseguramos de que se hayan eliminado las ASV con las asignaciones taxonómicas "mitocondrias" o "cloroplasto". Este proceso en principio no tendría que eliminar ninguna taxonomía, ya que se hizo con QIIME2.

```
dataset$filter_pollution(taxa = c("mitochondria", "chloroplast"))
```

```
## Total 0 taxa are removed from tax_table ...
```

Luego, para hacer que las ASV sean iguales en `otu_table`, `tax_table` y `phylo_tree`, usamos `tidy_dataset()` de nuevo.

```
dataset$tidy_dataset()
```

Usamos `sample_sums()` para comprobar el rango del número de lecturas de las muestras:

```
dataset$sample_sums() %>% range
```

```
## [1] 13091 154018
```

A veces, para reducir los efectos de la profundidad de secuenciación en las mediciones de diversidad, necesitamos realizar el remuestreo para que el número de lecturas sea igual para cada muestra. La función `rarefy_samples` puede invocar la función `tidy_dataset` automáticamente antes y después de la normalización. Indicamos que el número de lecturas por muestras sea de 21640, que es aproximadamente el valor que obtuvimos con QIIME2.

```
dataset$rarefy_samples(sample.size = 21640)
```

```
## 1 samples removed because they contained fewer reads than `sample.size`.
```

Con este valor, una de las muestras se elimina porque no contiene el número mínimo de lecturas indicadas.

Verificamos que en nuestro objeto *microtable* este todo correcto, y comprobamos ahora el rango de lecturas de nuestras muestras.

```
dataset
```

```
## microtable class:  
## sample_table have 29 rows and 9 columns  
## otu_table have 575 rows and 29 columns  
## tax_table have 575 rows and 7 columns  
## phylo_tree have 575 tips  
## rep_fasta have 575 sequences
```

```
dataset$sample_sums() %>% range
```

```
## [1] 21640 21640
```

Podemos ver que ahora nuestro objeto solo contiene 29 muestras, pero el número de lecturas en ellas es igual.

Calculamos la abundancia de taxones en cada rango taxonómico usando `cal_abund()`. Esta función devuelve una lista llamada `taxa_abund` que contiene varios marcos de datos de la información de abundancia en cada rango taxonómico. La lista se almacena en el objeto *microtable* automáticamente.

```
dataset$cal_abund()
```

```
## The result is stored in object$taxa_abund ...
```

La función `save_abund()` se puede utilizar para guardar el archivo de abundancia de taxones en un lugar local.

```
dir.create("taxa_abund")
```

```
## Warning in dir.create("taxa_abund"): 'taxa_abund' already exists
```

```
dataset$save_abund(dirpath = "taxa_abund")
```

Calculamos la diversidad alfa. El resultado también se almacena en el objeto *microtable* automáticamente. Indicamos PD = FALSE para no agregar la diversidad filogenética de Faith.

```
dataset$cal_alphadiv(PD = FALSE)

## The result is stored in object$alpha_diversity ...
```

Creamos un directorio para guardar los resultados de la diversidad alfa, y lo guardamos en él

```
dir.create("alpha_diversity")

## Warning in dir.create("alpha_diversity"): 'alpha_diversity' already
exists

dataset$save_alphadiv(dirpath = "alpha_diversity")
```

Calculamos la matriz de distancia de la diversidad beta utilizando la función *cal\_betadiv()*. Proporcionamos cuatro índices más usados: Bray-curtis, Jaccard, Unifrac ponderado y Unifrac no ponderado. Indicamos unifrac = TRUE para calcular las métricas Unifrac.

```
dataset$cal_betadiv(unifrac = TRUE)

## The result is stored in object$beta_diversity ...
```

A continuación, creamos un directorio para guardar los resultados de diversidad beta, y guardamos el output en ese directorio.

```
dir.create("beta_diversity")

## Warning in dir.create("beta_diversity"): 'beta_diversity' already e
xists

dataset$save_betadiv(dirpath = "beta_diversity")
```

Ahora ya tenemos todos los elementos para empezar los distintos análisis. En nuestro caso vamos a calcular la abundancia taxonómica, la diversidad alfa y la diversidad beta.

## Abundancia taxonómica

Vamos a crear el objeto `trans_abund`, que se utiliza para poder transformar los datos de abundancia taxonómica en gráficos de abundancia de taxones con el paquete `ggplot2`. Antes de graficar los resultados tendremos que calcular la lista `taxa_abund` del objeto `microtable` que queremos mostrar. En nuestro caso nos parece interesante representar los taxones por el rango taxonómico de familia y género, y el número de taxones a visualizar en cada caso serán 10. Eliminamos los nombres de muestra en el eje x y agregamos el argumento `facet` para mostrar la abundancia según los grupos.

```
t1 <- trans_abund$new(dataset = dataset, taxrank = "Family", ntaxa = 10)
t1$plot_bar(others_color = "grey70", facet = "Group", xtext_keep = FALSE, legend_text_italic = FALSE)

## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
```

*#No se muestra el resultado (Figura 13)*

```
t1 <- trans_abund$new(dataset = dataset, taxrank = "Genus", ntaxa = 10)
t1$plot_bar(others_color = "grey70", facet = "Group", xtext_keep = FALSE, legend_text_italic = FALSE)

## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
```

*#No se muestra el resultado (Figura 14)*

Puede ser interesante usar el argumento `groupmean` para obtener un barplot con la abundancia media de cada grupo.

```
t1 <- trans_abund$new(dataset = dataset, taxrank = "Genus", ntaxa = 10, groupmean = "Group")

## Group column is used to calculate mean abundance ...

t1$plot_bar(others_color = "grey70", legend_text_italic = FALSE)
```

*#No se muestra el resultado (Figura 15)*

## Alfa diversidad

La diversidad alfa se puede transformar y graficar usando un objeto de la clase `trans_alpha`. Crear el objeto de la clase `trans_alpha` puede invocar los datos `alpha_diversity` del objeto de la clase `microtable` creado. El objeto `trans_alpha` tiene dos marcos de datos: `alpha_data` y `alpha_stat`. En este último es donde se encuentran todas las métricas de alfa diversidad.

```
t1 <- trans_alpha$new(dataset = dataset, group = "Group")
## The group statistics are stored in object$alpha_stat ...
## The transformed diversity data is stored in object$alpha_data ...
```

Ahora que tenemos los datos de alfa diversidad tendremos que comprobar si son significativos o no. Podemos realizar dos pruebas, el ANOVA que es un test paramétrico y el test no paramétrico de Kruskal-Wallis.

Para poder aplicar el ANOVA antes tendremos que verificar si nuestros datos se distribuyen de forma normal. Podemos verificar el supuesto de normalidad con un `qqplot` y una prueba de Shapiro.

```
qqnorm(t1$alpha_data$Value)
qqline(t1$alpha_data$Value)
```

*#No se muestra el resultado (Figura 16)*

```
shapiro.test(t1$alpha_data$Value)
##
## Shapiro-Wilk normality test
##
## data:  t1$alpha_data$Value
## W = 0.75833, p-value < 2.2e-16
```

A la vista de la disposición de los cuantiles y del p-value del test de Shapiro, no podemos aceptar la hipótesis de normalidad.

Realizamos el test no paramétrico de Kruskal-Wallis, que como nuestros datos de abundancia alfa no siguen una distribución normal es el que tenemos que usar para la interpretación de los datos.

```
t1$cal_diff(method = "KW")

## The result is stored in object$res_alpha_diff ...

table_alpha <- filter(t1$res_alpha_diff, Measure %in% "Shannon")

write.csv2(table_alpha, "table_alpha.csv")

filter(t1$alpha_stat, Measure %in% "Shannon")

##   Group Measure N      Mean      SD      SE
## 1    DC Shannon 6  2.335761 0.6804357 0.2777867
## 2    DI Shannon 6  1.911453 0.7639679 0.3118886
## 3    DM Shannon 5  2.538727 0.9157593 0.4095400
## 4  DMIX Shannon 6  1.937423 0.9208584 0.3759389
## 5  DP.L Shannon 6  1.488085 1.0184619 0.4157853
```

Vemos que ninguna de las comparaciones es significativa según el índice de Shannon. Representamos estos resultados en un diagrama de cajas.

```
t1$plot_alpha(add_letter = FALSE, measure = "Shannon", use_boxplot = TRUE,
              boxplot_color = TRUE, order_x_mean = TRUE)
```

*#No se muestra el resultado (Figura 17)*

## Beta diversidad

La matriz de distancia de la diversidad beta se puede transformar y graficar usando el objeto de la clase `trans_beta`. Existen distintos análisis en el paquete *microeco* para la diversidad beta. Nosotros no vamos a decantar por un gráfico PCoA para ver como se distribuyen las muestras según las distancias de disimilitud de Bray-Curtis y el PERMANOVA.

Primero mostramos la ordenación de las muestras utilizando un PCoA. Creamos el objeto `trans_beta` para invocar la matriz de distancia de la métrica Bray-Curtis

```
t1 <- trans_beta$new(dataset = dataset, group = "Group", measure = "bray")
```

```
## Please also cite the original paper: An et al. (2019). Soil bacterial community structure in Chinese wetlands. Geoderma, 337, 290-299.
```

Ordenamos la matriz de distancias de Bray-Curtis por el método PCoA y luego graficamos los resultados en función del grupo.

```
t1$cal_ordination(ordination = "PCoA")
```

```
## The ordination result is stored in object$res_ordination ...
```

```
t1$plot_ordination(plot_color = "Group", plot_shape = "Group", plot_group_ellipse = TRUE)
```

*#No se muestra el resultado (Figura 18)*

A continuación, calculamos y graficamos las distancias entre grupos.

```
t1$cal_group_distance()
```

```
## The result is stored in object$res_group_distance ...
```

```
t1$plot_group_distance(distance_pair_stat = FALSE)
```

*#No se muestra el resultado (Figura 19)*

Realizamos un análisis de permutaciones (PERMANOVA) para ver si las distancias entre grupos son significativas. Primero para el factor grupo y después por comparaciones por parejas.

```
t1$cal_manova(cal_manova_all = TRUE)

## The result is stored in object$res_manova ...

t1$res_manova$aov.tab

## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##           Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## Group      4    0.9851 0.24627 0.91517 0.13234 0.592
## Residuals 24    6.4583 0.26910          0.86766
## Total     28    7.4434          1.00000
```

```
t1$cal_manova(cal_manova_paired = TRUE)

## The result is stored in object$res_manova ...

t1$res_manova
##           Groups measure permutations      R2 p.value
Significance
## 1    DC vs DP.L    bray          999 0.16192187 0.121
## 2    DC vs DI     bray          999 0.04160076 0.946
## 3    DC vs DMIX   bray          999 0.06613891 0.727
## 4    DC vs DM     bray          999 0.08528480 0.589
## 5    DP.L vs DI   bray          999 0.11701598 0.243
## 6    DP.L vs DMIX bray          999 0.09316168 0.353
## 7    DP.L vs DM   bray          999 0.13326018 0.206
## 8    DI vs DMIX   bray          999 0.03432288 0.896
## 9    DI vs DM     bray          999 0.09449064 0.455
## 10   DMIX vs DM   bray          999 0.06889448 0.871
```

También puede ser análisis realizar un análisis de permutaciones con la variable tanque, que en este tipo de experimentos es muy importante

```
t1$cal_manova(cal_manova_set = "Group + Tank")

## The result is stored in object$res_manova ...

t1$res_manova$aov.tab

## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##           Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## Group      4    0.9851 0.24627 1.0589 0.13234 0.407
## Tank     10    3.2024 0.32024 1.3770 0.43023 0.092 .
## Residuals 14    3.2559 0.23257          0.43743
## Total     28    7.4434          1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### Abundancia diferencial

Las pruebas de abundancia diferencial son una parte muy importante en el análisis de datos de microbioma. Se puede utilizar para encontrar taxones significativos en la determinación de las diferencias de la comunidad microbiana entre grupos. Con el paquete *microeco*, el objeto de clase `trans_diff` tiene cuatro enfoques para realizar este tipo de análisis. En nuestro caso nos vamos a decantar por `metastat`. `Metastat` depende de las permutaciones y de la prueba `t` y funciona bien con los datos escasos.

```
t1 <- trans_diff$new(dataset = dataset, method = "metastat", group = "
Group", metastat_taxa_level = "Family")

## Total 10 paired group for calculation ...

## Run 1 : DC vs DP.L ...

## Run 2 : DC vs DI ...

## Run 3 : DC vs DMIX ...

## Run 4 : DC vs DM ...

## Run 5 : DP.L vs DI ...

## Run 6 : DP.L vs DMIX ...

## Run 7 : DP.L vs DM ...

## Run 8 : DI vs DMIX ...

## Run 9 : DI vs DM ...

## Run 10 : DMIX vs DM ...

## The metastat result is stored in object$res_metastat ...

## The metastat group information is stored in object$res_metastat_group_matrix ...
```

```
t1 <- trans_diff$new(dataset = dataset, method = "metastat", group = "
Group", metastat_taxa_level = "Genus")

## Total 10 paired group for calculation ...

## Run 1 : DC vs DP.L ...

## Run 2 : DC vs DI ...

## Run 3 : DC vs DMIX ...

## Run 4 : DC vs DM ...

## Run 5 : DP.L vs DI ...

## Run 6 : DP.L vs DMIX ...

## Run 7 : DP.L vs DM ...
```

```
## Run 8 : DI vs DMIX ...
## Run 9 : DI vs DM ...
## Run 10 : DMIX vs DM ...
## The metastat result is stored in object$res_metastat ...
## The metastat group information is stored in object$res_metastat_group_matrix ...
```

Guardamos los archivos con los datos de abundancia diferencial

```
write.csv2(t1$res_metastat, "trans_diff_family.csv")
write.csv2(t1$res_metastat, "trans_diff_genus.csv")
```

Vemos que ninguna de las comparaciones es significativa, ya que ninguna presenta un *q-value* inferior a 0.05

## 11. Agradecimientos

Agradecer a Enric Gisbert, como responsable del programa y compañero del IRTA, por brindarme la oportunidad de analizar datos del programa de Acuicultura y por la confianza que ha depositado en mí para llevar a cabo este proyecto.

Eternamente agradecido con Maria Ballester y Yulixis Ramayo, compañeros del IRTA de Torre Marimon, por toda su dedicación y tiempo. Por mostrarme que con dedicación, esfuerzo y sacrificio todo es posible. Ellos han sido los guías de este último viaje del máster.

A Dolors Furones, por creer en mí, por apoyarme en cada decisión y por la enseñanza y consejos de vida.

Agradecer a todas aquellas personas que me han animado y apoyado durante estos dos últimos años, ya que la combinación trabajo - estudio no ha sido nada fácil. Esas personas me han mostrado la luz durante los días de oscuridad.

A mi mujer, Sonia.

Muchas gracias