Ciencia Latina Revista Científica Multidisciplinar
Septiembre-Octubre, 2023, Volumen 7, Número 5
https://doi.org/10.37811/cl_rcm.v7i5.8382

Identificación de Microorganismos en Muestras Ambientales: Análisis Bioinformático del Gen 16S RRNA Mediante QIIME2

Valetín Pérez Hernández[1]
vperezhdez@hotmail.com
https://orcid.org/0000-0001-9907-1316
Laboratory of Soil Ecology
Centro de Investigación y Estudios Avanzados (CINVESTAV)
Ciudad de México
México

Mario Hernández Guzmán
hg.maryo@gmail.com
https://orcid.org/0000-0003-1420-6280
Laboratorio de Metagenómica
Centro de Investigación Científica y Educación Superior de Ensenada (CICESE)
Baja California
México

RESUMEN

La diversidad de microorganismos en el suelo es elevada y su identificación mediante el uso de técnicas tradicionales de cultivo resulta inadecuada y limitada para un elevado porcentaje de los mismos. En la actualidad se cuentan con tecnologías de secuenciamiento masivo del ADN que ha permitido, junto a otras técnicas y herramientas, incluyendo la bioinfomática, la identificación de microrganismos sin necesidad del uso de medios de cultivos. Sin embargo, la secuenciación masiva ha generado enormes cantidades de información que requiere ser analizada y por ende demanda un esfuerzo computacional considerable. Existen diversos programas bioinformáticos, basados en uno o varios lenguajes de programación, para el análisis molecular *in silico* de secuencias de ADN, e.g., MOTHUR, QIIME1, DADA2 y QIIME2. De estos, QIIME2, por sus siglas en inglés "*Quantitative insights into microbial ecology*", es una herramienta frecuentemente empleada para el analisis datos de secuenciamiento de marcadores moleculares o genes funcionales, e.g., 16S rRNA, 18S rRNA, ITS, COI, entre otros. Dada la importancia de éstas, y de la necesidad de acceso a este conocimiento en lenguaje español, en esta revisión se describe y detalla el flujo de trabajo para el análisis de secuencias del gen 16S rRNA provenientes de muestras ambientales empleando QIIME2.

Palabras clave: *bioinformática; microbioma; NGS; protocolo*

[1] Autor principal.
Correspondencia: hg.maryo@gmail.com

Microorganisms Identification in Environmental Samples: Bioinformatic Analysis of 16S rRNA Gene with QIIME2

ABSTRACT

The soil harbors a great diversity of microorganisms and their identification using traditional cultivation techniques is not suitable for those that cannot be isolated and cultivated. Currently, there are a some available DNA sequencing technologies which along with bioinformatics tools and techniques allows the identification of microorganisms without the requirement of culture media. Next generation sequencing has generated enormous amounts (gigabytes and terabytes size) of information that required to be analyzed. Currently, some bioinformatics programs are available, e.g., MOTHUR, QIIME1, DADA2 and QIIME2. which are based on different programming languages. Among these, QIIME2 is the most cited pipeline and it's usually employed for marker or functional genes' sequencing data analyses, e.g., 16S rRNA, 18S rRNA, ITS, COI, etc. The aim of this work was to describe into a detailed manner a general pipeline for the 16S rRNA gene sequences from environmental soil samples.

***Keywords*:** *bioinformatics; microbiome; NGS; protocol; QIIME2*

Artículo recibido 19 setiembre 2023
Aceptado para publicación: 28 octubre 2023

INTRODUCCIÓN

Los microorganismos presentes en el ambiente representan cerca del 18% de la biomasa total del planeta tierra y aproximadamente el 5% de la materia orgánica presente en el suelo (Bar-On *et al.,* 2018; Singh *et al.,* 2023). De la diversidad de microorganismos sólo se han podido identificar cerca del 1% mediante técnicas tradicionales de cultivo (Martiny, 2019). Para intentar solucionar esta falta de información y conocimiento, se han empleado tecnologías de secuenciamiento masivo que hoy permiten la identificación de los microorganismos a través del análisis de su material genético, por ejemplo, DNA.

La primera técnica de secuenciación fue mediante el método de Sanger (primera generación), desarrollada en 1970, que permitió la identificación de microorganismos de forma individual (Slatko *et al.,* 2018). La secuenciación de Sanger se usó exitosamente en estudios relacionados con el análisis del genoma del bacteriofago PhiX174 y del genoma humano (Sanger *et al.,* 1977; Schloss *et al.,* 2020). En la actualidad, esta técnica ha caído en cierto desuso, sin embargo, se emplea en estudios particulares, por ejemplo, identificación de bacterias patógenas (Rudkjøbing *et al.,* 2016; Furutani *et al.,* 2022).

La secuenciación de "primera generación" fue suplida cerca de los años 2000. Hoy conocida como la "segunda generación" de secuenciadores, fue dominada por compañías como Roche-454™ (pirosecuenciación) e Illumina®. Estas tecnologías permiten el secuenciamiento de centenas de miles a millones de fragmentos de ADN en forma paralela, aumentando así la cantidad de información y una reducción considerable en los costos de secuenciación (Heather & Chain, 2016). Sin embargo, la limitante de éstos fue el proveer de productos cortos de secuenciación, e.g., Illumina®, y errores en lecturas largas (pirosecuenciación) (Slatko *et al.,* 2018).

En la última década surgió la "tercera generación" de secuenciadores que buscaban corregir las limitaciones de las tecnologías previas, por ejemplo, evitar la limitada longitud de lecturas y aumentar profundidad de secuenciamiento (Heather & Chain, 2016). Las tecnologías de PacBio™ y Oxford Nanopore® son los pioneros de la tercera generación de secuenciadores (Schadt *et al.,* 2010).

El uso de las tecnologías de segunda y tercer generación han generado de gigas- a petabytes de información de secuencias de ADN, lo que a su vez ha provocado la necesidad y creación de programas bioinformáticos que permiten interpretar y manipular este tipo de información (Kanzi *et al.,* 2020). El *software* que se empleará para el análisis de secuencias de ADN dependerá tanto de la tecnología de

secuenciación (primera, segunda o tercera generación) usado y de la metodología usada para tratar las muestras (marcadores molecular, RNA-Seq, metagenómica, etc).

2. Secuenciamiento de amplicones basados en el GEN 16S RRNA

El secuenciamiento de amplicones, que no son más que regiones cortas de ADN producto de la técnica PCR (amplificación mediante reacción en cadena de la polimerasa), permite el uso de *primers* o cebadores específicos según las necesidades del investigador y del objetivo de la investigación. En la literatura se tiene registro de un gran número de cebadores diseñados que son usados para la identificación de diferentes grupos taxonómicos o genes funcionales, e.g., grupos bacterianos, hongos o arqueas (Tabla 1). Los cebadores que se emplean con mayor regularidad para la identificación de comunidades procationtes (Bacteria y Arqueas) son aquellos que amplifican una, dos o más regiones hipervariables (V1 a V9) del marcador filogenético 16S rRNA.

En la actualidad, la tecnología de secuenciación más popular para la secuenciación de amplicones, dada la limitante en producto de secuenciación, es la plataforma Illumina® MiSeq™. Esta plataforma permite diferentes longitudes para el producto de secuenciamiento, por ejemplo, 2×150, 2×250 ó 2×300 con extremos pareados (PE, *paired-end*) (Illumina, 2022). Esta plataforma permite generar hasta 15 GB de información con un costo aproximado de 18 dólares por muestra (Illumina, 2023), basándonos en la versión 2×300 PE Kit V3 (https://www.illumina.com/systems/sequencing-platforms/miseq/specifications.html). Debido a que se maneja menor cantidad de información, en comparación al de datos de metagenomas (*shotgun metagenomics*), los requerimientos computacionales, para el análisis de secuencias producto de amplicón, se reduce de forma relativamente considerable. Sin embargo, los requerimientos computacionales dependen principalmente de la cantidad de datos a ser analizados. Cuando se manejan $\leq 4\times10^6$ lecturas, se requiere una computadora personal con memoria > 8 GB de RAM, 20 GB de espacio de almacenamiento en disco, y procesadores con mediana capacidad (a >2.0 GHz), para su análisis. Como es de esperarse, algunos procesos requerirán mayor esfuerzo computacional comparados con otros.

Los programas más populares que se han empleado con el fin de analizar datos provenientes del secuenciamiento de amplicones son QIIME1 (Caporaso *et al*., 2010), QIIME2 (Bolyen *et al.,* 2019) y MOTHUR (Schloss *et al*., 2009). La herramienta QIIME2 fue introducida en 2018 como actualización

y reemplazo de QIIME1. QIIME2 implementa diversos algoritmos denominados "*plugins*" que hacen eficiente el análisis de secuencias de tipo extremo único (del inglés "*single-end*") o extremos pareados (del inglés "*paired-end*"). Por mencionar algunos, QIIME2 cuenta con *plugins* como las herramientas *cutadapt* (Martin, 2011), *DADA2* (Callahan *et al.,* 2016), *blast+* (Camacho *et al.,* 2009) y *sklearn-classifier* (Pedregosa *et al.,* 2011), entre otros. El listado completo de *plugins* para QIIME2 se encuentra en la documentación de la misma y disponible en https://library.qiime2.org/plugins/.

METODOLOGÍA

QIIME2 ha sido la herramienta empleada en investigaciones relacionadas al estudio de microbiomas en diversos ambientes (Boylen et al. 2018). Hasta la fecha, ha mantenido actualizaciones regularmente nombradas con un código que indica el año y versión, por ejemplo: qiime2-2023.7 indica la séptima actualización durante el año 2023. El flujo de análisis bioinformático implementado permite analizar secuencias de amplicones para la identificación de diferentes grupos microbianos, incluyendo bacterianas, hongos, arqueas o genes funcionales e.g., nifh, amoA, pmoA, etc. En este protocolo se detalla el análisis de secuencias provenientes del secuenciamiento de dos regiones hipervariables del gen 16S rRNA, empleando un subset de datos de una investigación en donde se estudió la comunidad bacteriana en suelos de agricultura de conservación (Hernández-Guzmán et al., 2022). La secuenciación se realizó mediante la tecnología Illumina® MiSeq™ 2×300 PE, y basados en la región V3-V4 del gen 16S rRNA. Los datos crudos del secuenciamiento se encuentran publicados en el NCBI (BioProject PRJNA545497). Las secuencias y el protocolo pueden es de acceso público desde el repositorio en GitHub en https://github.com/MaryoHg/protocolo_qiime2_16S.

Para poder trabajar con con los datos, se sugiere realizar una dos actividades: i) clonar el repositorio de forma local mediante la línea de comandos (ver abajo), ó ii) ir al repositorio, y descargar un archivo comprimido del mismo (zip), para posteriormente descomprimir de forma local y trabajar dentro del directorio nuevo creado.

Para clonar el repositorio, se emplea el siguiente comando:

```
$ git clone https://github.com/MaryoHg/protocolo_qiime2_16S
$ cd protocolo_qiime2_16S
```

RESULTADOS Y DISCUSIÓN

A continuación se detalla la instalación del software requerido para realizar los análisis bioinformáticos, así como el diagrama general de análisis (Fig. 1). El diagrama general inicia con la importación de datos crudos a artefactos de QIIME2, incluye análisis de alfa y beta diversidad, finalizando con la exportación de datos para posterior visualización y análisis estadísticos.

I. Instalación del ambiente QIIME2

La instalación de QIIME2 puede realizarse en versiones recientes de los sistemas operativos Windows®, Linux y Mac (macOS™). QIIME2 está diseñado para correr en sistemas Unix, por lo que los usuarios con sistema Windows® se verán obligados a elegir una de dos opciones:

1. Crear una sistema virtual de Unix a través del WSL (Subsistema de Windows® para Linux, por sus siglas en ingles) ó
2. Instalar el manejador de paquetes ANACONDA para Windows® que posee una interfaz gráfica y que puede ser de mayor facilidad para algunos usuarios (https://www.anaconda.com/download).

En este protocolo se detalla el cómo emular el sistema Unix a través de WSL para sistemas Windows® (https://learn.microsoft.com/es-es/windows/wsl/), y el procedimiento para instalación en los diferentes sistemas operativos se detalla a continuación.

a) Instalación en sistema operativo Windows®

1. Instalación del Subsistema de Windows® para Linux (WSL)

1.1. Abrir la aplicación de WindowsShell o la terminal del sistema Windows® (cmd) con permisos de administrador.

1.2. En la consola, escribir y ejecutar el comando de instalación y esperar que termine el proceso. Durante el proceso se solicitará crear nombre de usuario y contraseña del nuevo sistema Linux.

```
$ wsl --install
```

1.3. Tras reiniciar el sistema, abrimos de nueva cuenta la PowerShell o la terminal del sistema (cmd). Si WSL no aparece activado, ejecutamos el comando para activarlo. Aquí se realizarán los análisis.

```
$ wsl
```

Para mayores detalles de la instalacion de WSL en el sistema Windows®, podemos referirnos a la página web oficial en https://learn.microsoft.com/en-us/windows/wsl/install.

2. Instalación del manejador de ambientes “miniconda” dentro del subsistema de Linux para Windows® (WSL)

Posterior a la istalación de WSL, procedemos a la instalación del manejador de ambiente “miniconda” (https://docs.conda.io/en/latest/miniconda.html). Éste nos permite crear ambientes (aislados) en donde podemos manejar versiones compatibles de software sin problemas de compatibilidad.

2.1. Para instalar “miniconda” requerimos conocer la versión del programa “Python” en el sistema WSL. Para ello, ejecutamos:

```
$ python --version
```

2.2. Tras identificarla, nos dirimos a descargar la versión correspondiente de “miniconda” desde su página web oficial en https://docs.conda.io/projects/miniconda/en/latest/miniconda-other-installer-links.html. Si tienes Python v3.10 instalado, puedes ejecutar lo siguiente:

```
$ wget -c https://repo.anaconda.com/miniconda/Miniconda3-py310_23.3.1-0-Linux-x86_64.sh
$ chmod +x Miniconda3-py310_23.3.1-0-Linux-x86_64.sh
```

2.3. Tras descargar “miniconda” (archivo con extensión *.sh), ejecutamos con el intérprete bash y aceptamos los términos de la licencia del software e instalación por defecto (opción "-b") con el siguiente comando:

```
$ bash Miniconda3-py38_23.3.1-0-Linux-x86_64.sh -b -p $HOME/miniconda3
```

2.4. Tras terminar y reiniciar la consola, verificamos que miniconda haya sido instalado en el sistema. Con el comando a continuación podemos ver los detalles de la instalación:

```
$ conda info
```

3. Instalación de QIIME2

3.1 Para instalar QIIME2, seguiremos las instrucciones de la página oficial, donde se nos indica

a) Descargar un archivo YML para Linux (similar a Windows® (vía WSL)),

b) Crear un ambiente con el archivo YML, activar el ambiente creado y verificar (*qiime --help*).

```
$ wget -c https://data.qiime2.org/distro/core/qiime2-2023.5-py38-linux-conda.yml
$ conda env create -n qiime2-2023.5 --file qiime2-2023.5-py38-linux-conda.yml
$ conda activate qiime2-2023.5
$ qiime --help
```

3.3 El último comando nos permitir ver la ayuda de QIIME2; podremos visualizar la lista de *plugins* y ayuda del programa. Esto significa que la instalación fue exitosa.

b) Instalación en sistemas Linux y macOS

Para la instalación de QIIME2 en sistemas Linux y macOS podemos proceder directamente con la instalación de "miniconda", y posterior creación del ambiente de QIIME2. La instalación se realiza mediante las terminales correspondientes a cada sistema operativo. En macOS pueden presentarse variaciones debido al procesador del equipo (Intel® vs M1/M2 (Apple Silicon)). Se recomienda ver los detalles en la página oficial de documentación (https://docs.qiime2.org/2023.5/install/native/).

Breve repaso de conceptos en QIIME2

QIIME2 almacena y maneja la información en archivos conocidos como "artefactos" (*artifacts*, del idioma inglés) que poseen extensión "qza" ó "qzv"; ambos tipos son archivos comprimidos (gzip) que mantienen organizada los información de entrada o salida de los diferentes comandos del flujo de trabajo. Así, para poder usar QIIME2 es necesario entender y manipular estos nuevos formatos (https://docs.qiime2.org/2023.9/concepts/#data-files-qiime-2-artifacts). Para visualizar los resultados o salidas de QIIME2 se requiere la generación de artefactos o archivos "qzv"; éstos pueden visualizarse en cualquier explorador web a través de los servidores de QIIME2 directamente en https://view.qiime2.org/ ó mediante el comando *qiime tools view*. Éste último requiere la ruta y nombre de un archivo qzv como único argumento.

Análisis bioinformático

a) Verificación de los metadatos (sample metadata)

Los metadatos en un archivo de texto plano que contiene las variables descriptivas de las muestras analizadas, es decir, variables ambientales o fisicoquímicas (variables numéricas), tiempo ó tratamientos (variables categóricas). Esta base de datos debe cumplir con los siguientes requerimientos:

1. Archivo de texto plano (formatos txt, tsv ó csv)
2. La primera columna debe indicar obligatoriamente los nombres (etiquetas únicas) de las muestras (SampleID). Esta columna llamarse como uno de los siguientes: SampleID, id, sampleid, o featureid.

3. Columnas extras a la primera obligatorio, i.e., SampleID, son opcionales. Se recomienda evitar el uso de símbolos o acentos, y registrar tantas variables ambientales o categóricas como sea posible para tener una descripción detalladas de las muestras.

Se recomienda revisar detalles específicos de la creación y manejo de los metadatos desde la documentación del software en https://docs.qiime2.org/2023.7/tutorials/metadata/.

1. Visualización y comprobación del formato de los metadatos: crear el artefacto qzv de los metadatos, nos permite verificar que éste cumple con los requerimientos de QIIME2. Para crearlo corremos el siguiente comando, dentro del directorio de trabajo (repositorio clonado anteriormente):

```
$ cd  protocolo_qiime2_16S
$ qiime metadata tabulate \
--m-input-file sample-metadata.tsv \
--o-visualization sample-metadata.qzv
```

Explicación: --m-input-file indica la ruta del archivo con metadatos (sample-metadata.tsv); --o-visualization indica nombre y ruta de salida del archivo de visualización (*.qzv). El comando no debe indicar ningún tipo de error en la consola.

2. Visualizamos el archivo *sample-metadata.qzv* directamente en el navegador (https://view.qiime2.org/) o mediante comando (*qiime tools view sample-metadata.qzv*) (Fig. 2).

b) Importación de las secuencias a QIIME2

Para importar las secuencias y crear el artefacto de QIIME2, usamos el comando *qiime tools import*. Las secuencias de este ejemplo son secuencias crudas, y en formato fastq con tamaño 2×300 PE. Para poder importar a un artefacto de QIIME2 requerimos que exista un directorio que contenga tres archivos: *forward.fastq.gz*, *reverse.fastq.gz*, y *barcodes.fastq.gz*. Estos corresponden a las lecturas R1, R2 y los indices (barcodes) para realizar el proceso de *demultiplexing.* El nombre de los archivos y la compresión gzip son obligatorios.

1. El directorio de trabajo debe contener un directorio denomanido *seqs/* que debe contener los 3 archivos en formato FastQ. Algunos detalles de este procesos se encuentran en la documentación oficial en https://docs.qiime2.org/2023.5/tutorials/importing/:

```
$ qiime tools import \
--type EMPPairedEndSequences \
--input-path seqs/ \
--output-path raw_seqs.qza
```

Explicación: --type indica el tipo de formato de nuestras secuencias; --input-path indica la carpeta donde se encuentran las secuencias; --output-path indica la dirección y el nombre de salida del archivo ó artefacto *.qza. El archivo de salida (raw_seqs.qza) contendrá las secuencias crudas y lo usaremos este para los análisis posteriores. Este paso demorará alrededor de 10 segundos.

c) Rastreo de muestras por etiquetas o barcodes: *Demultiplexing*

El proceso de demultiplexado (*demultiplexing*) consiste en identificar y agrupar las secuencias con base en una etiqueta única insertada durante la amplificación mediante la técnica de PCR (también conocido como "código de barra", del inglés *barcode*. Los códigos de barra deben ser únicos para cada muestra y/o réplica biológica; esto permite agrupar, identificar y ordenar las secuencias por SampleID. Así se relacionan las secuencias con sus metadatos y se matiene el registro de cada una durante el protocolo bioinformático.

1. Para separar y etiquetar las secuencias crudas por su identificador o SampleID ("*demultiplexing*") empleamos el comando *qiime demux emp-paired* dado que las secuencias son pareadas (PE) y obtenidas con base en el "EMP Protocol". El artefacto de salida del comando contendrá las secuencias demultiplexadas.

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

Explicación: --i-seqs indica el nombre y dirección del archivo qza con las secuencias crudas (raw_seqs.qza); --m-barcodes-file indica nombre y dirección de los metadatos (sample-metadata.tsv); -

-m-barcodes-column indica la columna que contiene los *barcodes* de las muestras dentro de los metadatos; --o-per-sample-sequences indica la dirección y nombre de salida de las secuencias demultiplexadas (qza); --o-error-correction-details indicamos la dirección y nombre de salida del archivo que contiene los detalles de errores durante el demultiplexado; --p-no-golay-error-correction indicamos que no empleamos barcodes tipo Golay, dado que los *barcodes* fueron de 8 nt en este caso.

2. Para visualizar el número de secuencias crudas obtenidas para cada SampleID, es decir, visualizar los resultados del demultiplexado (Fig. 3), debemos de crear el archivo "qzv" del artefacto de salida del comando, para ello empleamos el comando *qiime demux summarize*. El comando *qiime tools view* nos permitirá visualizar los datos en el explorador web por omisión.

```
$ qiime demux summarize \
--i-data demux-seqs.qza \
--o-visualization demux-seqs.qzv
$ qiime tools view demux-seqs.qzv
```

Explicación: --i-data ruta y nombre del archivo de entrada (demux-seqs.qza); --o-visualization ruta y nombre del artefacto qzv (demux-seqs.qzv).

d) Eliminación de adaptadores (*primers*)

La secuencias pueden contener los adaptadores o *primers* usados durante la PCR. Se recomienda eliminarlos dado que la presencia de secuencias no biológicas pueden tener un efecto negativo y significativo en la inferencia de ASVs. Para removerlos empleamos el comando *qiime cutadapt trim-paired* con los siguientes parámetros:

```
$ qiime cutadapt trim-paired \
--i-demultiplexed-sequences demux-seqs.qza \
--p-front-f CCTACGGGNGGCWGCAG \
--p-front-r GACTACHVGGGTATCTAATCC \
--p-cores 4 \
 --p-match-adapter-wildcards \
--o-trimmed-sequences demux-trimmed-seqs.qza
```

Explicación: --i-demultiplexed-sequences ruta y nombre de las secuencias demultiplexadas (demux-seqs.qza); --p-front-f es la secuencia del *primer* empleado para R1 durante la PCR (*forward primer*); --p-front-r es la secuencia del *primer* usado en R2 durante la PCR (*reverse primer*); --p-match-adapter-wildcards indicamos que se reconozca el código IUPAC en los adaptadores (*primers* degenerados); --o-trimmed-sequences indica el nombre y dirección del artefacto de salida que contendrá las secuencias filtradas (demux-trimmed-seqs.qza). El comando se demorará cerca de 30 segundos.

La visualización de los resultados se logra convirtiendo el arfacto de salida del comando anterior, i.e, *demux-trimmed-seqs.qza*, en un artefacto para visualizacion (qzv) con el siguiente comando:

```
$ qiime demux summarize \
--i-data demux-trimmed-seqs.qza \
--o-visualization demux-trimmed-seqs.qzv
$ qiime tools view demux-trimmed-seqs.qzv
```

El artefacto qzv se puede visualizar mediante comando o en el explorador en https://view.qiime2.org/. La visualización es similar a lo mostrado en la Figura 3. Comparando ambos resultados, i.e., *demultiplexing* versus recortados con cutadapt, podremos observar si se presentó la eliminación de secuencias no deseadas y recorte.

e) DADA2: Filtrado por calidad, reducción de ruido (*denoising*), eliminación de quimeras, deduplicación e inferencia de secuencias de amplicón (ASVs)

Para este proceso empleamos el comando *qiime dada2 denoise-paired*. El plugin DADA2 emplea un modelo de corrección que posteriormente es aplicado a las secuencias crudas (R1 y R2) para corrección de errores (Callaham et al., 2016). El algoritmo realiza los siguientes procesos:

- Inspección de los perfiles de calidad (Phred ó Q-score) y posterior filtrado y eliminación de las secuencias que no cumplan con los parámetros configurados (valoración Phred y longitud de las lecturas). No permite nucléotidos ambiguos (N);
- Determinación y aplicación del modelo ("*learn error rates*") para corrección de errores ("*denoising*") sobre las secuencias filtradas;
- Deduplicación de secuencias: compara y combina las secuencias idénticas en una secuencia única, conservando el valor de abundancia correspondiente (*dereplication*);

- Identifica las secuencias únicas presentes en la muestra (inferencia de ASVs, "*amplicon sequence variants*") para cada lectura: R1 y R2, respectivamente y sobrelapa ("*merge overlaping reads*") las lecturas R1 y R2;
- Identifica y elimina las secuencias quiméricas, para finalmente contruir la matriz de valores (*feature-table.qza*) e impresión de los estadísticos del proceso en los artefactos correspondientes (*stat-denoising.qza*).

Para conocer más detalles se recomiendan las lecturas de los protocolos correspondientes de DADA2 (https://docs.qiime2.org/2023.7/plugins/available/dada2/denoise-paired/) en QIIME2 y R (https://benjjneb.github.io/dada2/tutorial_1_8.html).

Dentro de los parámetros requeridos del comando *qiime dada2 denoise-paired* se requiere conocer la longitud a la cual las secuencias R1 y R2 serán recortadas (*trimming*). Esta información lo obtenemos de las secuencias demultiplexadas de manera visual (Fig. 4). Este gráfico nos permite identificar la longitud mínima a la que las secuencias aún mantienen un valor de calidad Phred (Q-score) ≥ 20. Se recomienda tomar la longitud en el valor donde el valor Phred pase por debajo del valor de 20. La longitud puede ser diferente para R1 y R2, y regularmente, la calidad en R2 es menor que en R1. Para este ejemplo la longitud para R1 es 260, mientras que para R2 es 230. Se recomienda ser cuidadoso y evitar recortar las secuencias de tal manera que se pierda la longitud mínima de traslape (*merge paired-en reads*) o unión necesaria, es decir, 12 pares de bases. Después de obtener los datos de longitud de las secuencias, procedemos a ejecutar el comando *qiime dada2 denoise-paired*:

```
$ qiime dada2 denoise-paired \
--i-demultiplexed-seqs demux-trimmed-seqs.qza
--p-trunc-len-f 260 \
--p-trunc-len-r 230 \
--p-n-threads 4 \
--o-table table.qza \
--o-representative-sequences rep-seqs.qza \
--o-denoising-stats stats-denoising.qza
```

Explicación: --i-demultiplexed-seqs indica la dirección y nombre del artefacto qza con de las secuencias demultiplexadas; --p-trim-left-r longitud a la que se recortarán R1; --p-trunc-len-r longitud a la que se recortarán R2; --p-n-threads número de procesadores a usar; --o-table ruta y del artefacto de salida para la tabla de frecuencias (feature-table.qza); --o-representative-sequences ruta y nombre del artefacto de salida para las secuencias representativas (rep-seq.qza); --o-denoising-stats ruta y nombre del artefacto de salida con los estadísticos del proceso (stats-denoising.qza). El comando correrá cerca de 8 minutos.

Tras ejecutar DADA2, obtendremos tres archivos: i) feature-table.qza (tabla con las frecuencias), ii) rep-seqs.qza (secuencias representativas) y iii) stats-denoising.qza (estadísticos del proceso). Para poder visualizar los resultados (Fig. 5), convertimos los artefactos qza a qzv con diferentes comandos. Visualizar el producto de stats-denoising (Fig. 5a) y la feature-table (Fig. 5b) se logra con los siguientes:

```
$ qiime metadata tabulate \
--m-input-file denoising-stats.qza \
--o-visualization denoising-stats.qzv
$ qiime tools view stats-denoising.qzv
$ qiime feature-table summarize \
--i-table table.qza \
--m-sample-metadata-file sample-metadata.tsv \
--o-visualization table.qzv
$ qiime tools view table.qzv
```

Mediante la visualización y análisis de los estadísticos del proceso (stats-denoising.qzv) podremos observar el número de secuencias eliminadas durante el proceso de inferencia por DADA2 (Fig. 5a). En caso de observar una perdida substancial de lecturas, se sugiere modificar los parámetros de filtrado en DADA2, evitando ser más permisibles que la configuración por omisión. Si lo anterior no mejora la retención, se sugiere analizar sólo las lecturas R1, descartando R2. Por su parte, la visualización del *feature-table* nos permite identificar cuántas variantes de secuencias de amplicon fueron obtenidas, i.e., número de ASVs totales y su frecuencia por muestra (Fig. 5b).

f) Entrenamiento de la base de datos para anotación taxonómica

Existen diferentes bases de datos disponibles con fines de anotación taxonómica de amplicones, a saber, Greengenes v13.7 (McDonald et al., 2012), Greengenes2 (McDonald et al., 2023), SILVA 138 (Quast et al., 2012), entre otros. En este protocolo se empleó la base de datos de SILVA v138.1 (Quast et al., 2012). Para descargarla nos dirigimos a la página de recursos de QIIME2 (https://docs.qiime2.org/2023.5/data-resources/). Después de descargar la base de datos (dos artefactos qza), ejecutaremos el comando *qiime feature-classifier extract-reads* con el fin de extraer sólo aquellas secuencias que amplifiquen con los *primers* usados durante la amplificiación mediante PCR. Para conocer detalles se sugiere revisar la ayuda del comando (https://docs.qiime2.org/2023.5/plugins/available/feature-classifier/extract-reads).

```
$ wget -c https://data.qiime2.org/2023.5/common/silva-138-99-seqs.qza
$ wget -c https://data.qiime2.org/2023.5/common/silva-138-99-tax.qza
$ qiime feature-classifier extract-reads \
--i-sequences silva-138-99-seqs.qza \
--p-f-primer CCTACGGGNGGCWGCAG \
--p-r-primer GACTACHVGGGTATCTAATCC \
--p-min-length 50 \
--p-max-length 500 \
--o-reads selected-silva-138-99-seqs.qza
```

Explicación: --i-sequences indica la dirección y nombre de las secuencias de la base de datos; --p-f-primer indicamos el *primer* empleado en R1 durante la PCR (*forward primer*); --p-r-primer indicamos el *primer* R2 empleado durante la PCR (*reverse primer*); --p-min-length indicamos la longitud mínima de las secuencias que queremos extraer de la base de datos; --p-max-length indicamos la longitud máxima de las lecturas de la base de datos; --o-reads ruta y nombre del artefacto de salida para las secuencias extraídas.

Posteriormente realizamos el entrenamiento de la base de datos empleando el comando *qiime feature-classifier fit-classifier-naive-bayes*. Este proceso "reduce", disminuyendo el esfuerzo computacional necesario, la base de datos y el algoritmo mejora la anotación taxonómica (Werner et al., 2012). El

proceso de entrenamiento requiere esfuerzo computacional y no puede realizarse de forma paralela. Se recomienda el uso de una computadora con ≥ 16 GB de RAM. La misma base de datos generada podrá usarse para anotar librerías dentro de la misma versión de QIIME2 empleada para entrenarse

```
$ qiime feature-classifier fit-classifier-naive-bayes \
--i-reference-reads selected-silva-138-99-seqs.qza \
--i-reference-taxonomy silva-138-99-tax.qza \
--o-classifier trained_classifier_DB.qza
```

Explicación: --i-reference-reads ruta y nombre de las secuencias filtradas de la base de datos; --i-reference-taxonomy indicamos la dirección y nombre de la taxonomía de la base de datos; --o-classifier ruta y nombre del artefacto de salida para nuestro clasificador entrenado.

g) Anotación taxonómica

La asignación taxonómica se realiza con el comando *qiime feature-classifier classify-sklearn*, que emplea *machine learning*, que ha demostrado mejorar la precisión de anotación taxonómica (Bokulich et al., 2018).

```
$ qiime feature-classifier classify-sklearn \
--i-classifier trained_classifier_DB.qza \
--i-reads rep-seqs.qza \
--p-confidence 0.97 \
--o-classification taxonomy.qza
```

Explicación: --i-classifier ruta y nombre de la base de datos entrenada; --i-reads ruta y nombre de las secuencias representativas (obtenidas después de DADA2); --p-confidence confidencia en la anotación taxonómica, rango de 0 a 1; --o-classification ruta y nombre del artifactor de salida que contendrá la taxonomía asignada a las secuencias analizadas.

El artefacto de salida *taxonomy.qza* contiene la anotación taxonómica de los ASVs de nuestro estudio, y podremos acceder a ella mediante su visualización mediante gráficos de barras interactivos tras convertilo a un artefacto qzv. Este gráfico se muestra en términos de abundancia relativa de los diferentes grupos identificados a sus distintos niveles taxonómicos (puede ser desde domino a especie, si fuese el caso) (Fig. 6). La visualización nos permite cambiar el ordenamiento de muestras, colores y

niveles taxonómicos.

```
$ qiime taxa barplot \
--i-table table.qza \
--i-taxonomy taxonomy.qza \
--m-metadata-file sample-metadata.tsv \
--o-visualization taxa-barplot.qzv
$ qiime tools view taxa-barplot.qza
```

Explicación: --i-table ruta y nombre de la tabla de frecuencias de frecuencias (feature-table); --i-taxonomy dirección de la taxonomía asignada; --m-metadata-file dirección del sample-metadata; --o-visualization ruta y nombre del artefacto de salida para uso de visualización.

h) Creación de árbol filogenético

El protocolo de QIIME2 incluye la creación de árboles filogenéticos que son empleados en análisis de diversidad filogenética, por ejemplo, el índice de PD_whole tree (Faith's phylogenetic diversity ó en la determinación de matrices de distancia: weighted y unweighted UniFrac. Para su creación requerimos el artefacto con las secuencias de los ASVs (rep-seqs.qza) y un directorio de salida en donde se almacenarán cuatro artefactos que podremos usar posteriormente. El comando con parámetros por omisión es el siguiente:

```
$ qiime phylogeny align-to-tree-mafft-fasttree \
--i-sequences rep-seqs.qza \
--output-dir phyleny/
```

Explicación: --i-sequences ruta y nombre de las secuencias representativas (obtenidas en DADA2); --output-dir es el directorio de salida en donde se almacenarán 4 artefactos de salida: secuencias alineadas (--o-alignment), artefacto con secuencias enmascaradas (--o-masked-alignment), el árbol filogenético (--o-tree), y el árbol filogenético con raíz o ancestro común (--o-rooted-tree).

Al finalizar el comando, obtendremos cuatro archivos. El artefacto con el árbol filogenético con raíz o ancestro común, i.e., artefacto nombrado *rooted-tree.qza*, es el que podremos usar para posteriores análisis, e incluso convertirlo a otros formatos para posterior visualización en herramientas como iTOL (https://itol.embl.de/), e.g., formato Newick.

i) Calculo de diversidad alfa y beta

QIIME2 permite la determinación de diferentes índices tradicionales de alfa-diversidad, e.g., Chao1, Shannon, Simpson, entre otros, así como determinación de beta-diversidad. Este último basado matrices de distancias, por ejemplo *Bray-Curtis*, *Jaccard index*, *Unifrac*, entre otros (https://forum.qiime2.org/t/alpha-and-beta-diversity-explanations-and-commands/2282). Para determinar la diversidad usamos el comando *qiime diversity core-metrics-phylogenetic*; este requeire de un valor numérico que indica el número de secuencias en la muestra con menor profundidad de secuenciación. Con éste se realizará la rarefacción de los datos y podemos obtenerlo del artefacto *feature-table.qzv* (ver proceso de DADA2 – "sección e)").

```
$ qiime diversity core-metrics-phylogenetic \
--i-phylogeny rooted-tree.qza \
--i-table feature-table.qza \
--p-sampling-depth 1114 \
--m-metadata-file sample-metadata.tsv \
--output-dir core-metrics-results
```

Explicación: --i-phylogeny ruta y nombre del árbol filogenético (*rooted-tree.qza*); --i-table ruta y nombre de la tabla de frecuencias (*feature-table*); --p-sampling-depth profundidad mínima a la cual se realizarán algunos cálculos del comando; --m-metadata-file dirección del archivo de metadatos; --output-dir directorio de salida para los resultados del comando. Los resultados obtenidos pueden ser visualizados con el comando *qiime tool view* (Fig. 7).

j) Exportación de datos

La exportación en QIIME2 se puede realizar sobre cualquiera de los resultados obtenidos durante el análisis. Para realizarlo usamos el comando *qiime tools export*: este proceso nos permitirá crear una BIOM-table a partir de la taxonomía (*taxonomy.qza*) y la tabla de frecuencias (table.qza):

a. Crear el archivo taxonomy.tsv

```
$ qiime tools export \
--input-path taxonomy.qza \
--output-path exported/
```

b. Crear la BIOM

```
$ qiime tools export \
--input-path feature-table.qza \
--output-path exported/
```

Con la información de taxonomía y tabla de frecuencias exportados podremos crear una tabla en formato BIOM, que contendra la información taxonómica y su frecuencia en un solo archivo BIOM (http://biom-format.org/). Antes de crear la tabla BIOM, debemos editar los encabezados del archivo *taxonomy.tsv*. Los dos encabezados de este archivo se deben llamar: "#OTUID" y "taxonomy". La edición de los encabezados de *taxonomy.tsv* puede realizar de manera manual abriendo el archivo con un editor de texto plano, sin embargo, se muestran los comandos para realizarlo de forma automática. En caso de no renombrar los encabezados, el comando a emplear (*biom add metadata*) indicará errores:

```
$ sed -i "s/Feature ID/#OTUID/g" exported/taxonomy.txt
$ sed -i "s/Taxon/taxonomy/g" exported/taxonomy.txt
$ biom add-metadata \
-i exported/feature-table.biom \
--observation-metadata-fp exported/taxonomy.tsv \
-o exported/featured-table-with-taxonomy.biom \
--sc-separated taxonomy
```

El archivo BIOM generado puede ser exportado a herramientas como el lenguaje R ó RStudio-Posit™ (Posit™, 2023; R Core Team, 2023) a través de la paquetería Phyloseq (McMurdie & Holmes, 2013) para posteriores análisis o visualización.

ILUSTRACIONES, TABLAS, FIGURAS.

Tabla 1
Resumen de cebadores empleados en el estudio del microbioma en muestras ambientales

Marcador/Gen y cebadores	Población	Muestra	Referencia
16S rRNA 515F: 5′-GTGCCAGCMGCCGCGG-3′ 806R: 5′-GACTACHVGGGTWTCTAAT-3′ 907R: 5′-CCGTCAATTCMTTTRAGTTT-3′	Bacterias	Suelo de agricultura	Liu *et al.,* 2022
		Sedimento marino	Demko *et al.*, 2021
		Agua potable	Jing *et al*., 2021
ITS (*internal transcribed spacer*) ITS3 F: 5′-GCATCGATGAAGAACGCA GC-3′ ITS4 R: 5′-TCCTCCTATTGATATGC-3′ ITS1 F: 5′-CTTGGTCATTTAGAGGAAGTAA-3′ ITS2 R: 5′-GCTGCGTTCTTCATCGATGC-3′ NL1: 5′-GCATATCAATAAGCGGAGGAAAAG-3′ NL4: 5’-GGTCCGTGTTTCAAGACGG-3′	Hongos	Suelo forestal	Delgado *et al*., 2021
		Suelo contaminado	Gil-Martínez *et al.*, 2021
		Pinus densiflora, Pinus koraiensis, Pnus rigida, y Pinus thunbergi	Rim *et al*., 2021
pmoA (metano monooxigenasa) A189f: GGNGACTGGGACTTCTGG A682r: GAASGCNGAGAAGAASGC Cmo182F: TCACGTTGACGCCGATCC Cmo568R: GATGGGGATGGAGTATGTGC	Bacterias oxidadoras de metano	Sedimentos mineros	Baesman *et al*., 2015
		Lodos y sedimentos de aguas residuales	Siniscalchi *et al*., 2022
amoA (amonio monooxigenasa) CamoA-19F: 5′-ATGGTCTGGYTWAGACG-3′ TamoA-629R: 5′-TGGCANTAYMGATGGATGGC-3′	Arqueas oxidadoras de amonio	Suelo del Ártico	Alves *et al*., 2019
nifH (nitrogenasa) nifHF: 5’-TACGGNAARGGSGGNATCGGCAA-3’ nifHR: 5’-AGCATGTCYTCSAGYTCNTCCA-3’	Bacterias fijadoras de nitrógeno	Acacia xanthophloea, Faidherbia albida y Albizia versicolor	Teixeira *et al*., 2016

Figura 1
Diagrama general del análisis de secuencias empleando QIIME2

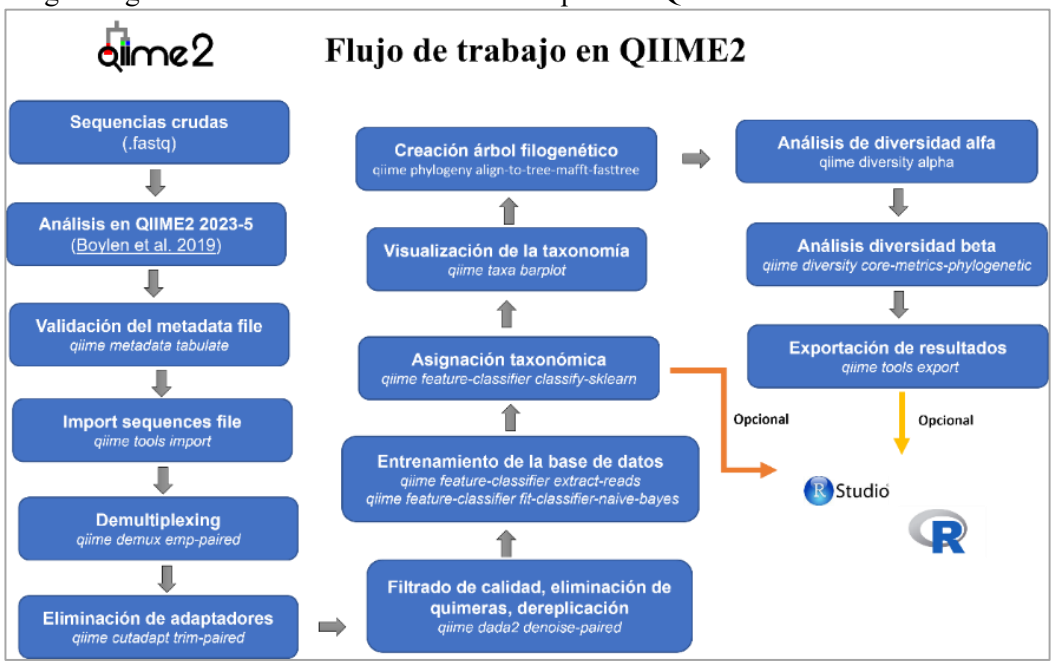
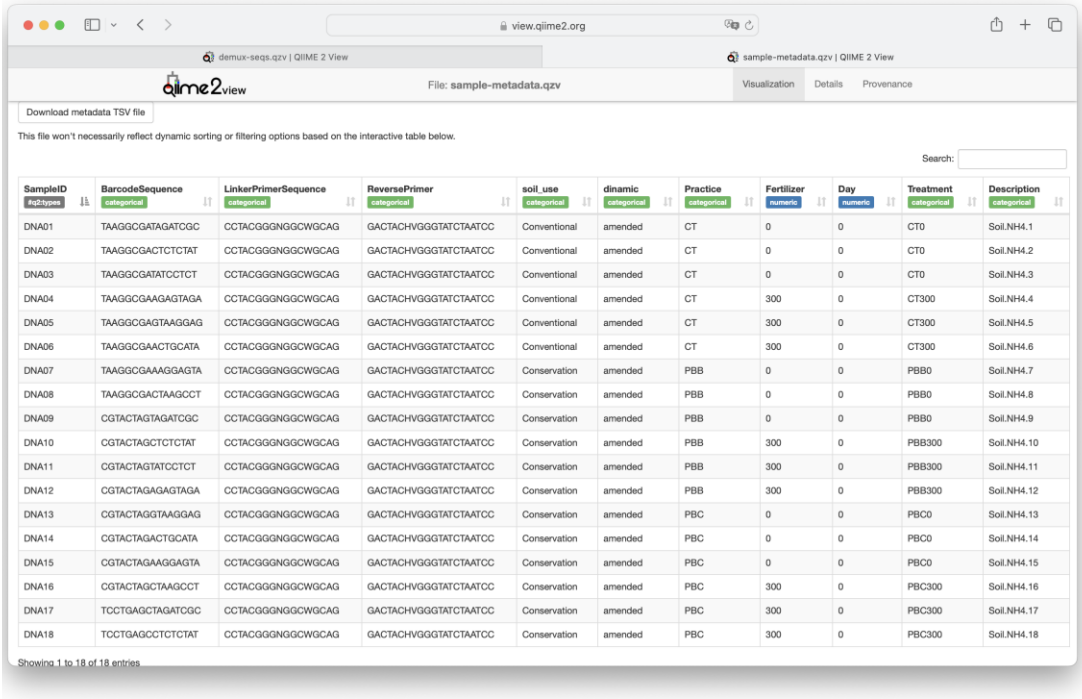


Figura 2
Visualización del sample-metadata con *qiime tool view*



SampleID	BarcodeSequence	LinkerPrimerSequence	ReversePrimer	soil_use	dinamic	Practice	Fertilizer	Day	Treatment	Description
DNA01	TAAGGCGATAGATCGC	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	Conventional	amended	CT	0	0	CT0	Soil.NH4.1
DNA02	TAAGGCGACTCTCTAT	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	Conventional	amended	CT	0	0	CT0	Soil.NH4.2
DNA03	TAAGGCGATATCCTCT	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	Conventional	amended	CT	0	0	CT0	Soil.NH4.3
DNA04	TAAGGCGAAGAGTAGA	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	Conventional	amended	CT	300	0	CT300	Soil.NH4.4
DNA05	TAAGGCGAGTAAGGAG	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	Conventional	amended	CT	300	0	CT300	Soil.NH4.5
DNA06	TAAGGCGAACTGCATA	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	Conventional	amended	CT	300	0	CT300	Soil.NH4.6
DNA07	TAAGGCGAAAGGAGTA	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	Conservation	amended	PBB	0	0	PBB0	Soil.NH4.7
DNA08	TAAGGCGACTAAGCCT	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	Conservation	amended	PBB	0	0	PBB0	Soil.NH4.8
DNA09	CGTACTAGTAGATCGC	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	Conservation	amended	PBB	0	0	PBB0	Soil.NH4.9
DNA10	CGTACTAGCTCTCTAT	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	Conservation	amended	PBB	300	0	PBB300	Soil.NH4.10
DNA11	CGTACTAGTATCCTCT	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	Conservation	amended	PBB	300	0	PBB300	Soil.NH4.11
DNA12	CGTACTAGAGAGTAGA	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	Conservation	amended	PBB	300	0	PBB300	Soil.NH4.12
DNA13	CGTACTAGGTAAGGAG	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	Conservation	amended	PBC	0	0	PBC0	Soil.NH4.13
DNA14	CGTACTAGACTGCATA	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	Conservation	amended	PBC	0	0	PBC0	Soil.NH4.14
DNA15	CGTACTAGAAGGAGTA	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	Conservation	amended	PBC	0	0	PBC0	Soil.NH4.15
DNA16	CGTACTAGCTAAGCCT	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	Conservation	amended	PBC	300	0	PBC300	Soil.NH4.16
DNA17	TCCTGAGCTAGATCGC	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	Conservation	amended	PBC	300	0	PBC300	Soil.NH4.17
DNA18	TCCTGAGCCTCTCTAT	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	Conservation	amended	PBC	300	0	PBC300	Soil.NH4.18

Showing 1 to 18 of 18 entries

.

Figura 3
Visualización de los estadísticos de las muestras demultiplexadas (demux-seqs.qza).

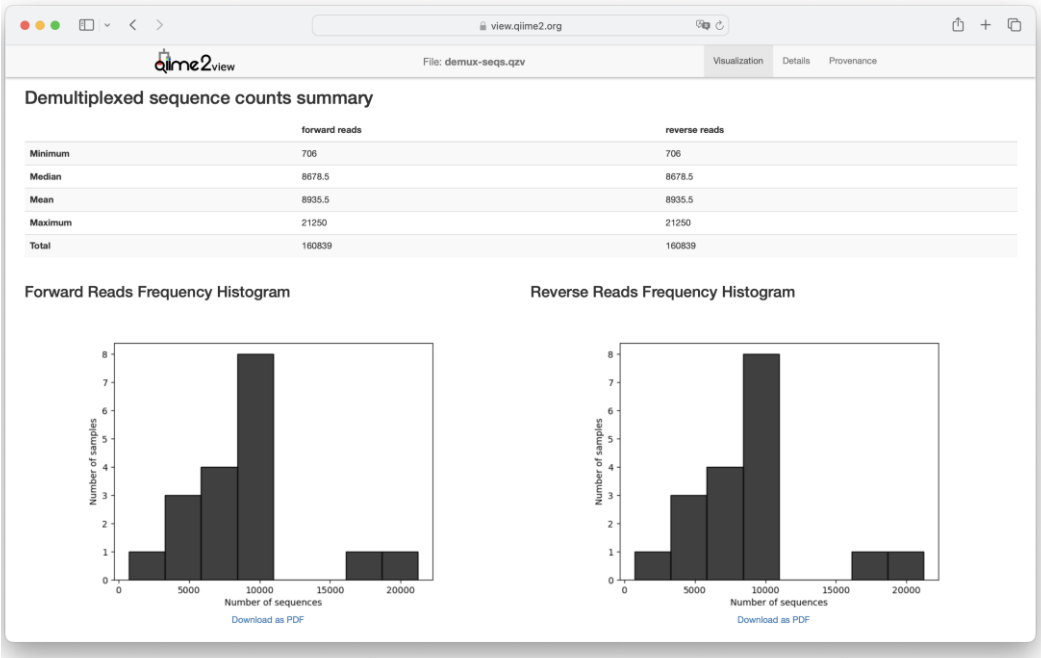
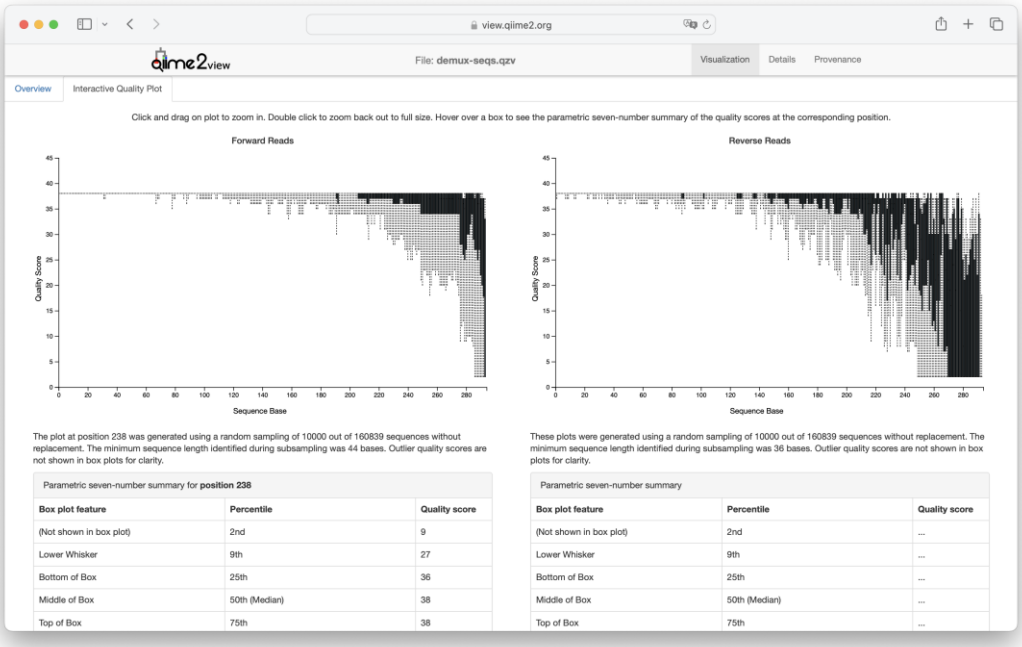


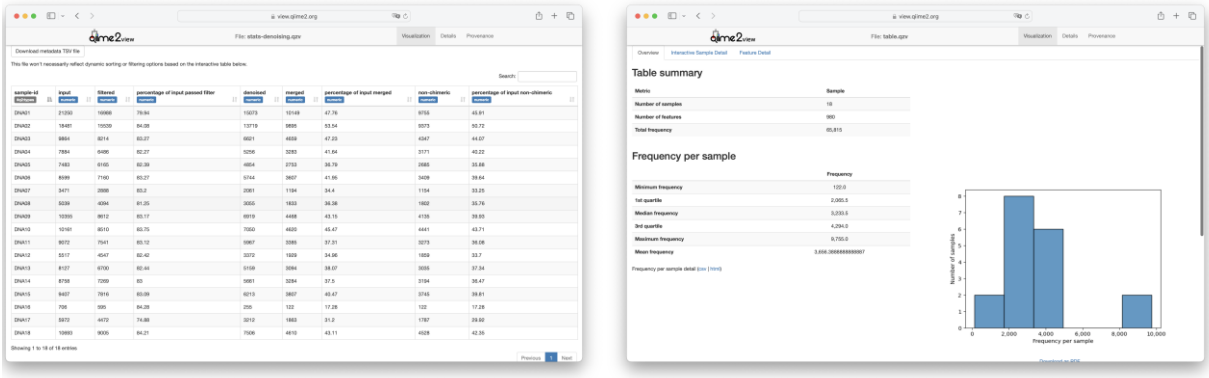
Figura 4



Visualización de la calidad Phred (Q-score) de las secuencias demultiplexadas y sin adapdadores emplaendo el artefacto demux-seqs.qzv. Se observan dos gráficos (de cajas y bigotes): uno para R1 (izquierda) y otro para R2 (derecha), en donde la longitud de las secuencia se dan sobre el eje “x”, y su calidad Phred sobre el eje “y”.

Figura 5

a) b)



Visualizacion de resultados del comando *qiime dada2 denoise-paired*: a) *stats-denoising* que muestra el tabular con el número de secuencias crudas, filtradas, corregidas, unidas y no quiméricas; b) *feature-table*: estadísticos de los ASVs obtenidos después de los filtros para cada muestra.

Figura 6

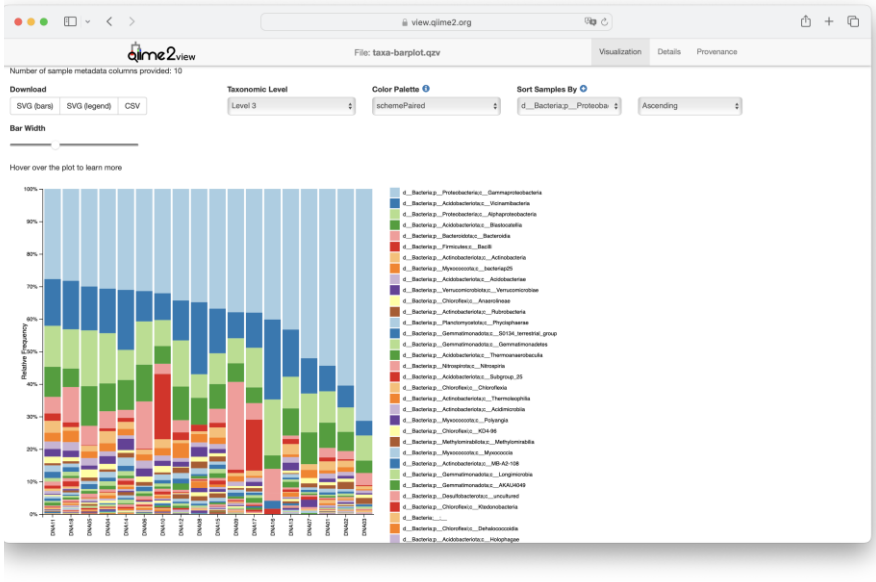
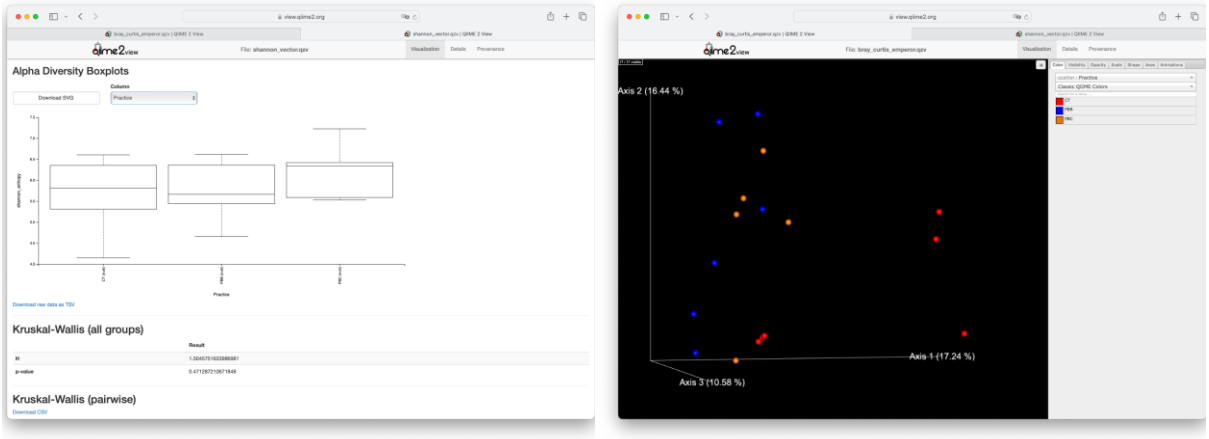


Gráfico de barra en términos de abundancia relativa de los grupos taxonómicos analizados. Nombre de las muestras sobre el eje “x”, y la abundancia relativa (0-100%) de los grupos taxonómicos sobre el eje “y”. Las leyendas/colores corresponden a los grupos taxonómicos identificados.

Figura 7

a) b)



Visualización del análisis de diversidad. a) Gráfico de cajas y bigotes y análisis estadístico (mediante Kruskall-Wallis) de alfa-diversidad (índice de Shannon) por práctica de manejo (*Practice* variable); y b) Diagrama de coordenadas principales (PCoA) basado en la matriz de distancia de *Bray-Curtis*.

CONCLUSIONES

Las herramientas de secuenciación masiva han permitido explorar la diversidad microbiana en los diferentes ambientes de la biósfera. Con la ayuda de herramientas bioinformáticas, que usan diferentes lenguajes de programación para el manejo, manipulación y análisis de secuencias de ADN, este proceso ha sido más fácil día con día. El programa QIIME2 nos facilita y hace más eficiente, y menos demorado, el análisis de los datos provenientes del secuenciamiento del gen 16S rRNA o cualquier otro gen o marcador molecular. En este trabajo se detalló el protocolo de análisis de secuencias provenientes de muestras ambientales que puede ser implementado para estudios con similar tecnología de secuenciación y análisis.

REFERENCIAS BIBLIOGRAFICAS

Alves, R. J. E., Kerou, M., Zappe, A., Bittner, R., Abby, S. S., Schmidt, H. A., Pfeifer, K., & Schleper, C. (2019). Ammonia Oxidation by the Arctic Terrestrial *Thaumarchaeote Candidatus Nitrosocosmicus* arcticus Is Stimulated by Increasing Temperatures. Frontiers in Microbiology, 10. https://www.frontiersin.org/articles/10.3389/fmicb.2019.01571

Baesman, S. M., Miller, L. G., Wei, J. H., Cho, Y., Matys, E. D., Summons, R. E., Welander, P. V., & Oremland, R. S. (2015). Methane Oxidation and Molecular Characterization of

Methanotrophs from a Former Mercury Mine Impoundment. Microorganisms, 3(2), 290–309. https://doi.org/10.3390/microorganisms3020290

Bar-On, Y. M., Phillips, R., & Milo, R. (2018). The biomass distribution on Earth. Proceedings of the National Academy of Sciences, 115(25), 6506–6511. https://doi.org/10.1073/pnas.1711842115

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology, 37(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

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

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. BMC Bioinformatics, 10(1), 421. https://doi.org/10.1186/1471-2105-10-421

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. Nature Methods, 7(5), 335–336. https://doi.org/10.1038/nmeth.f.303

Delgado, E. F., Valdez, A. T., Covarrubias, S. A., Tosi, S., & Nicola, L. (2021). Soil Fungal Diversity of the Aguarongo Andean Forest (Ecuador). Biology, 10(12). https://doi.org/10.3390/biology10121289

Demko, A. M., Patin, N. V., & Jensen, P. R. 2021. Microbial diversity in tropical marine sediments assessed using culture-dependent and culture-independent techniques. Environmental Microbiology, 23(11), 6859–6875. https://doi.org/10.1111/1462-2920.15798

Furutani, S., Furutani, N., Kawai, Y., Nakayama, A., & Nagai, H. (2022). Rapid DNA Sequencing Technology Based on the Sanger Method for Bacterial Identification. Sensors (Basel, Switzerland), 22(6). https://doi.org/10.3390/s22062130

Gil-Martínez, M., López-García, Á., Domínguez, M. T., Kjøller, R., Navarro-Fernández, C. M., Rosendahl, S., & Marañón, T. (2021). Soil fungal diversity and functionality are driven by plant species used in phytoremediation. Soil Biology and Biochemistry, 153, 108102. https://doi.org/10.1016/j.soilbio.2020.108102

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics, 107(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Hernández-Guzmán, M., Pérez-Hernández, V., Navarro-Noya, Y. E., Luna-Guido, M. L., Verhulst, N., Govaerts, B., & Dendooven, L. (2022). Application of ammonium to a N limited arable soil enriches a succession of bacteria typically found in the rhizosphere. Scientific Reports, 12(1), 4110. https://doi.org/10.1038/s41598-022-07623-4

Illumina. (2022). *Specification Sheet: MiSeq System*. https://www.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/miseq-system-data-sheet-m-gl-00006/miseq-data-sheet-m-gl-00006.pdf

Illumina. (2023). Cost of NGS | Comparisons and budget guidance. 2023. https://www.illumina.com/science/technology/next-generation-sequencing/beginners/ngs-cost.html. Consultado 18 de septiembre de 2023

Jing, Z., Lu, Z., Mao, T., Cao, W., Wang, W., Ke, Y., Zhao, Z., Wang, X., & Sun, W. (2021). Microbial composition and diversity of drinking water: A full scale spatial-temporal investigation of a city in northern China. Science of The Total Environment, 776, 145986. https://doi.org/10.1016/j.scitotenv.2021.145986

Kanzi, A. M., San, J. E., Chimukangara, B., Wilkinson, E., Fish, M., Ramsuran, V., & de Oliveira, T. (2020). Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance. Frontiers in Genetics, 11. https://www.frontiersin.org/articles/10.3389/fgene.2020.544162

Liu, S., Sun, Y., Shi, F., Liu, Y., Wang, F., Dong, S., & Li, M. (2022). Composition and Diversity of

Soil Microbial Community Associated With Land Use Types in the Agro–Pastoral Area in the Upper Yellow River Basin. Frontiers in Plant Science, 13. https://www.frontiersin.org/articles/10.3389/fpls.2022.819661

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R., & Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. The ISME Journal, 6(3), 610–618. https://doi.org/10.1038/ismej.2011.139

McDonald, D., Jiang, Y., Balaban, M., Cantrell, K., Zhu, Q., Gonzalez, A., Morton, J. T., Nicolaou, G., Parks, D. H., Karst, S. M., Albertsen, M., Hugenholtz, P., DeSantis, T., Song, S. J., Bartko, A., Havulinna, A. S., Jousilahti, P., Cheng, S., Inouye, M., … Knight, R. (2023). Greengenes2 unifies microbial data in a single reference tree. Nature Biotechnology. https://doi.org/10.1038/s41587-023-01845-1

McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE, 8(4), e61217. https://doi.org/10.1371/journal.pone.0061217

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, 17(1), 10. https://doi.org/10.14806/ej.17.1.200

Martiny, A. C. (2019). High proportions of bacteria are culturable across major biomes. The ISME Journal, 13(8), 2125–2128. https://doi.org/10.1038/s41396-019-0410-3

Sanger, F., Air, G. M., Barrell, B., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A., Slocombe, P. M., & Smith, M. (1977). Nucleotide sequence of bacteriophage φX174 DNA. Nature, 265, 687–695.

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. Human Molecular Genetics, 19(R2), R227–R240. https://doi.org/10.1093/hmg/ddq416

Schloss, J. A., Gibbs, R. A., Makhijani, V. B., & Marziali, A. (2020). Cultivating DNA Sequencing Technology After the Human Genome Project. Annual Review of Genomics and Human Genetics, 21(1), 117–138. https://doi.org/10.1146/annurev-genom-111919-082433

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R.

A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., & Weber, C. F. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Applied and Environmental Microbiology, 75(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

Singh, B., Yeasmin, S., & Sparks, D. L. (2023). Mineral-organic-microbial interactions. In M. J. Goss & M. Oliver (Eds.), Encyclopedia of Soils in the Environment (Second Edition) (pp. 387–406). Academic Press. https://doi.org/10.1016/B978-0-12-822974-3.00128-2

Siniscalchi, L. A. B., Siqueira, J. C., Batista, A. M. M., & Araújo, J. C. (2022). Detection of methanotrophic microorganisms in sludge and sediment samples from sewage treatment systems. Water Practice and Technology, 17(1), 329–335. https://doi.org/10.2166/wpt.2021.101

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. Current Protocols in Molecular Biology, 122(1), e59. https://doi.org/10.1002/cpmb.59

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., and Duchesnay E. (2011). Scikit-learn: machine learning in python. *Journal of machine learning research,* 12:2825–2830.

Posit team (2023). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. URL http://www.posit.co/

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research, 41(D1), D590–D596. https://doi.org/10.1093/nar/gks1219

R Core Team. (2023). R: A Language and Environment for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rim, S. O., Roy, M., Jeon, J., Montecillo, J. A. V., Park, S.-C., & Bae, H. (2021). Diversity and

Communities of Fungal Endophytes from Four Pinus Species in Korea. Forests, 12(3). https://doi.org/10.3390/f12030302

Rudkjøbing, V. B., Thomsen, T. R., Xu, Y., Melton-Kreft, R., Ahmed, A., Eickhardt, S., Bjarnsholt, T., Poulsen, S. S., Nielsen, P. H., Earl, J. P., Ehrlich, G. D., & Moser, C. (2016). Comparing culture and molecular methods for the identification of microorganisms involved in necrotizing soft tissue infections. BMC Infectious Diseases, 16(1), 652. https://doi.org/10.1186/s12879-016-1976-2

Teixeira, H., & Rodríguez-Echeverría, S. (2016). Identification of symbiotic nitrogen-fixing bacteria from three African leguminous trees in Gorongosa National Park. Systematic and Applied Microbiology, 39(5), 350–358. https://doi.org/10.1016/j.syapm.2016.05.004