Dedicado a

...

Índice general

Índice de figuras

Agradecimientos

¡Muchas gracias a todos!

Resumen

En este proyecto se analizan datos del estudio *David et al.* [1] procedentes del microbioma de dos individuos. Se trata de datos de secuenciación de la región V4 del ARN ribosomal 16S de los microorganismos presentes en intestino y saliva, tomados a lo largo de un año. Además, se recopilaron datos del estilo de vida de los donantes tales como dieta, ejercicio o enfermedad. Esto supone un total aproximado de 64,8GB de datos. Los sujetos son interesantes porque el primero de ellos tuvo una infección de *Salmonella* durante el estudio y el segundo realizó un viaje al extranjero, lo que nos permite ver cómo varía la dinámica del microbioma en estas situaciones.

En primer lugar, se comprobó la calidad y longitud de la secuenciación y se hizo un filtro previo que redujo muy poco el set de datos. A continuación, se agruparon las *reads* en OTUs al 97 % de similitud de secuencia y se asignó la taxonomía a nivel de género. Con estos datos se obtuvieron unas tablas que incluyen la abundancia absoluta de cada taxón, en cada día y en cada individuo.

En segundo lugar, se hizo un estudio de variabilidad temporal de los microorganismos presentes a partir de su abundancia. Se busca si estos datos se ajustan a algún modelo y encontramos que siguen la ley de Taylor, lo cual nos permite definir a cada individuo con tan solo dos variables. Además, se comprueba la variabilidad haciendo un análisis ordenando por abundancia total a lo largo de los días y haciendo un *ranking* de los microorganismos.

Por último, se calculan las correlaciones de abundancia relativa entre géneros. Se realiza una clasificación previa en grupos de comportamiento en base a una perturbación y se comprueba la correlación entre y dentro de grupos. Mediante contraste bibliográfico se trata de explicar este comportamiento y se abren distintas perspectivas de futuro.

1 Introducción

A lo largo de la evolución, billones de seres vivos han convivido en simbiosis con el ser humano y son fundamentales para su salud. A este conjunto se le denomina **flora microbiana** o **microbioma humano**. Se estima que tenemos 10 veces más microbios en nuestro organismo que células propias, pero estos huéspedes no ocupan mucho espacio ya que son mucho más pequeños que una célula humana (del 1-3 % de la masa total del cuerpo humano). Adquirimos los microorganismos a partir del nacimiento, durante el parto y la lactancia, y a lo largo de nuestra vida van colonizando nuestra piel, mucosas y, sobre todo, el tubo intestinal con clara preferencia por el intestino grueso. Los obtenemos de los alimentos, el agua y el contacto con otras personas. La importancia del microbioma radica en sus numerosas funciones: regulación de procesos digestivos, producción de sustancias bacteriostáticas y antibióticos naturales contra patógenos, enseña a nuestro sistema inmune a reconocer invasores dañinos, etc. Numerosos estudios demuestran que el cambio en la composición del microbioma está relacionado con estados de enfermedad, planteando la posibilidad de manipular estas comunidades como posible tratamiento. Los conocimientos en este campo han sido escasos hasta el año 2008 cuando se inició el Proyecto Microbioma Humano cuya misión era generar recursos que permitan la caracterización del microbioma humano y el análisis de su papel en la salud y la enfermedad humana.

La forma tradicional de estudiar los microorganismos de cualquier ambiente ha sido mediante técnicas de cultivo. Se obtiene la muestra del medio natural (suelo, agua o heces), se cultiva en un medio previamente definido con las condiciones óptimas y cuando los microorganismos crecen formando colonias, se extrae su material genético para analizarlo. A día de hoy, esta técnica se sigue llevando a cabo en ocasiones en las que se quiere estudiar el genoma de un organismo cultivable concreto ya que es muy sencilla y barata. Sin embargo, no todos los microorganismos presentes en las muestras ambientales son cultivables. Se estima que sólo el 1 % de las bacterias del suelo y entre el 0,1 - 0,01 % de las bacterias marinas. A estos microorganismos se les denomina **no cultivables** y se debe al desconocimiento de los requisitos específicos del cultivo y a la existencia de grupos de microorganismos que deben mantenerse en equilibrio para sobrevivir.

La **metagenómica** surge como alternativa a las técnicas de cultivo para explicar cuáles son los microorganismos presentes en cualquier muestra y estudiar todo el ADN genómico presente en dicha muestra. En este caso, se obtiene la muestra natural, se aislan los microorganismos presentes, se extrae su material genético y se secuencia todo el genoma o la región de interés. Ejemplos históricos en metagenómica son los estudios de

Tyson et al., 2004 [2] y Venter et al., 2004 [3]. El primero se realizó en ambientes extremos donde se encontró un grupo relativamente pequeño de microorganismos y el segundo, sin embargo, se realizó en el océano donde se encontró una enorme variedad de especies. Ambos ponen de manifiesto el esfuerzo que requiere la secuenciación de estas muestras y su posterior análisis. Gracias a la exponencial evolución hacia la secuenciación de siguiente generación, que ha permitido reducir costos y realizar proyectos más robustos, se pueden realizar estos estudios a gran escala. Y gracias al desarrollo de la bioinformática se han obtenido herramientas potentes y sofisticadas para poder analizar esa inmensa cantidad de datos generada.

Cuando se pretende caracterizar la estructura taxonómica de una comunidad microbiana se puede utilizar la secuenciación aleatoria también conocida como *shotgun* o un gen marcador como el ARN ribosómico (ARNr) 16S. La primera aproximación trata de romper un genoma en fragmentos al azar, secuenciar cada uno de ellos y luego organizar estas partes superpuestas para guiar el ensamblaje; se puede realizar por clonación utilizando vectores o por secuenciación directa. Los datos en los que se centra este estudio se obtuvieron por la segunda aproximación ya que el 16S es considerado como el cronómetro molecular. Es un polirribonucleótido de unas 1542 pares de bases, codificado por el gen rss y se encuentra en la subunidad pequeña 30S del ribosoma, orgánulo encargado de la síntesis celular de proteínas. Se utiliza como marcador porque es un gen presente en todas las bacterias y contiene regiones conservadas de forma universal, mientras que otras regiones son variables (Figura 1.1) y esto es lo que hace posible la identificación a un nivel taxonómico suficientemente informativo. Para conseguir una identificación lo suficientemente específica, hay que tener en cuenta la cobertura de secuenciación (para detectar microorganismos que se encuentren en una menor concentración), la longitud de las *reads* (secuencias más largas permiten una asignación taxonómica más precisa) y la tasa de error en la secuenciación (puede generar una asignación incorrecta al enmascarar las regiones variables).

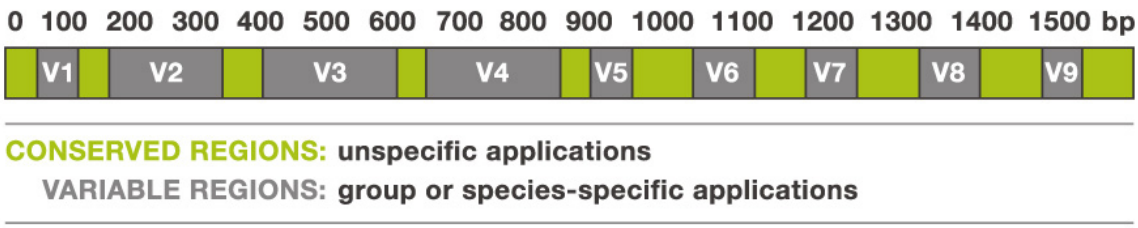

Figura 1.1: Representación esquemática de ARNr 16S. En verde se colorean las regiones conservadas (C) y en gris las variables (V).

1.1. Pasos en un proyecto metagenómico

1. **Extracción del ADN**. Se pueden utilizar distintos kits comerciales cuyo fundamento se basa en la lisis de las células mediante compuestos químicos (EDTA, lisozima, detergentes...), seguida de la eliminación de los componentes celulares y finalmente se precipita el material genético para obtener ADN puro en alta concentración.

2. **Creación una genoteca de 16S para la secuenciación**:

   - Amplificación por reacción en cadena de la polimerasa (*polymerase chain reaction*[PCR]) de la región a ser analizada. Los métodos de secuenciación no son aplicables a moléculas individuales, por lo que es necesario sintetizar múltiples copias para obtener una lectura. En cada ciclo de PCR se llevan a cabo 3 etapas: desnaturalización (donde las cadenas de ADN se separan calentándolas), hibridación (se utilizan cebadores o *primers* que hibridan con su secuencia complementaria por el extremo 3') y extensión (etapa donde actúa la Taq polimerasa sobre el *primer* y agrega bases complementarias para crear cadenas completas de ADN). El tipo de PCR a utilizar puede ser PCR en emulsión o PCR puente.
   - *Barcoding* del ADN para multiplexar el ensayo. Este método permite secuenciar varias muestras en el mismo run de secuenciación, mejorando así el costo-eficacia y el tiempo de obtención de los resultados. En el proceso de PCR, además de los cebadores se añaden por ligación otros trozos de secuencia que no son complementarios a la región diana y por tanto quedan "colgando". Esos trozos incluyen el adaptador para la plataforma de secuenciación (Illumina en este caso), un *linker* y un código de barras (*barcode*) específico para cada secuencia. El producto de amplificación final es el que se muestra en la Figura 1.2. Éste se utiliza para el siguiente paso.

3. **Secuenciación**. El desarrollo de la secuenciación de nueva generación (*next generation sequencing [NGS]*) permite obtener millones de fragmentos de ADN de forma paralela, logrando que el número de bases que se pueden secuenciar por unidad de precio haya crecido exponencialmente en los últimos años. Existen distintas plataformas de segunda generación que se pueden clasificar según su método de secuenciar:

   - Sintesis: se basa en el proceso de síntesis de ADN usando la enzima ADN polimerasa para identificar las bases presentes en la molécula complementaria de ADN. A su vez se agrupa en otros tres tipos atendiendo al sistema de detección: **pirosecuenciación** basada en la detección quimioluminiscente de pirofosfato liberado (Ej.: Roche/454), **fluorescencia** cuando los nucleótidos están marcados fluorescentemente (Ej.: Illumina/MiSeq, HiSeq) y **secuenciación no óptica** que mide la liberación de protones (Ej.: ThermoFisher Scientific/Ion Torrent).

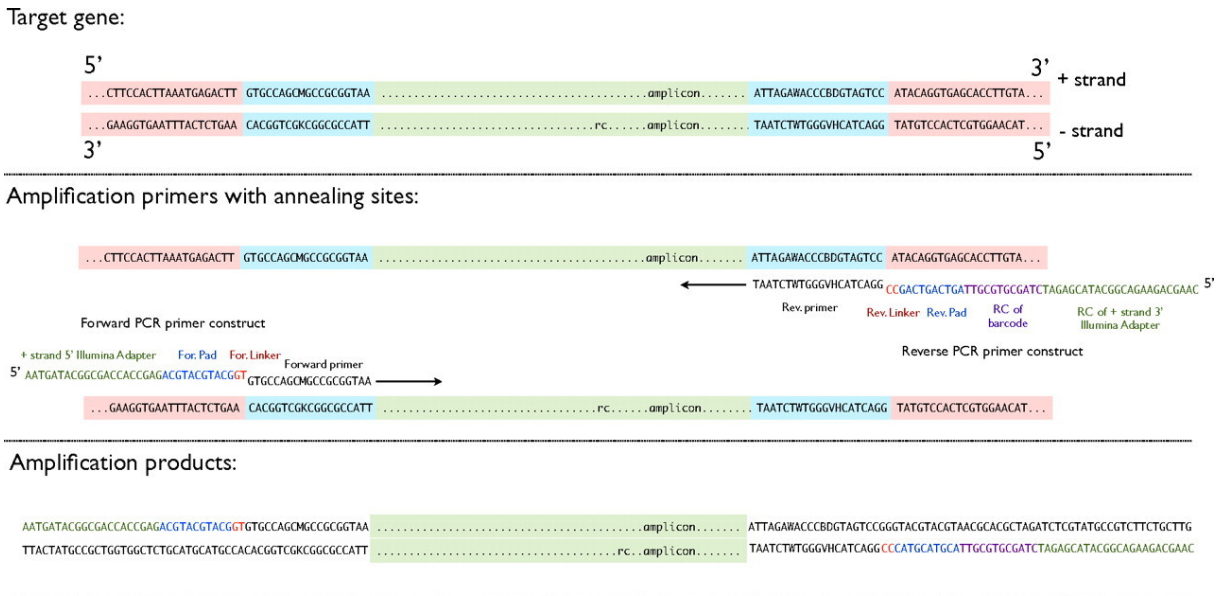

Figura 1.2: Protocolo de secuenciación Illumina con *barcode*. En la primera parte, el gen diana es ARNr 16S en el que se colorean de azul las regiones conservadas y en verde las regiones adecuadas para la clasificación taxonómica. En la segunda parte, se procede a la amplificación por PCR utilizando cebadores de PCR homólogos a la secuencia (negro) que incluyen por ligación un *linker* (rojo), un *barcode* (violeta), y el adaptador Illumina (verde). Lo único que queda unido por hibridación al molde es la parte negra. En la tercera parte, se observa el producto de amplificación.

- Ligación: se emplea una ADN ligasa en lugar de una polimerasa para identificar la secuencia objetivo (Ej.: SOLiD).
- Hibridación: un método no enzimático que utiliza una única muestra de ADN marcada fluorescentemente y se hibrida con una colección de secuencias conocidas (chip de ADN). Si la muestra hibrida fuertemente en un punto dado, se deduce que esa es su secuencia.

En general, los nuevos secuenciadores generan lecturas a partir de los dos extremos de un fragmento de ADN, dando lugar a lecturas apareadas con una distancia conocida entre ellas. Para ello utilizan dos estrategias diferentes (Figura 1.3): *paired end* que proporcionan rangos de tamaños de inserto más estrechos y *mate pair* que cubren tamaños mayores. Las lecturas de tipo *paired end* se generan mediante la fragmentación del ADN en pequeños segmentos de los cuales se secuencia el final de ambos extremos. Por contra, los *mate pairs* se crean a partir de fragmentos de ADN de tamaño conocido, que se circularizan y se ligan usando un adaptador interno. Estos fragmentos circularizados se trocean al azar para luego purificar los segmentos que contienen el adaptador a partir del cual se secuenciará.

Actualmente, se encuentran en desarrollo las tecnologías de tercera generación que tienen como objetivo la secuenciación de moléculas individuales para eliminar el paso de amplificación de ADN, que suele introducir errores y/o sesgos. Además, han conseguido alargar la longitud de secuencia obtenida lo que facilita el ensamblaje. Se predice que

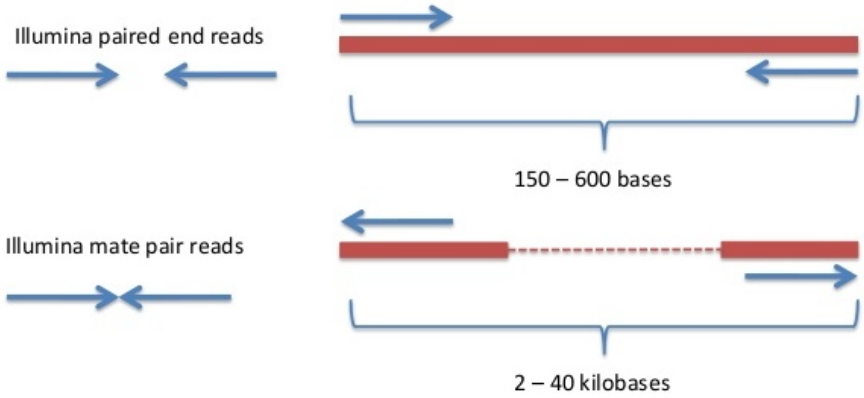

Figura 1.3:
Paired end vs Mate pair.

la plataforma MinION (basada en nanoporos) supondrá una revolución porque incluye las características mencionadas anteriormente y, además, reduce el coste de reactivos y aparatos al tratarse de una técnica no óptica. En julio de 2016 fue utilizado por la NASA para identificar un hongo que está creciendo en la Estación Espacial Internacional (ISS) e investigar la posible existencia de vida fuera del planeta.

4. **Análisis bioinformático**. Existen varias herramientas pero las más utilizadas por su fácil manejo y su amplia documentación son QIIME y mothur. Ambos *pipelines* incorporan algoritmos para llevar a cabo el control de calidad, agrupación de secuencias, asignación de taxonomía, cálculo de diversidad y visualización de resultados.

5. **Interpretación de los resultados**.

Gracias a estas tecnologías se puede conocer la composición microbiana de cualquier ambiente en un determinado momento. E incluso se puede ir más allá analizando series temporales para observar la dinámica de estas poblaciones. En este trabajo se analiza una serie temporal anual que integra la información de cientos de muestras, es una de las más grandes que se han realizado hasta la fecha.

2 Materiales y métodos

2.1. Preprocesado: FastQC, MultiQC y seq_crumbs

FastQC (`http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`) es una herramienta de control de calidad para datos de secuenciación, de código abierto e implementada en Java. Permite un fichero de entrada en distintos formatos (fastq, SAM o BAM) y produce un fichero de salida en formato HTML con gráficos y tablas que permiten evaluar los datos. Proporciona mucha información sobre una única muestra: estadísticas numéricas (codificación de calidad según la plataforma utilizada, número total de secuencias...), *score* de calidad, contenido en GC, distribución de longitud de secuencias, etc. Así se puede detectar rápidamente cualquier problema que hay que tener en cuenta antes de realizar análisis posteriores.

MultiQC (`http://multiqc.info/`) es una herramienta de código abierto, implementada en Python que da soporte a muchas herramientas bioinformáticas, entre ellas FastQC. Produce un reporte HTML muy parecido pero permite un análisis a lo largo de varias muestras. La visualización de las muestras en conjunto permite realizar comparaciones y también recopila estadísticas numéricas de cada muestra para ver cómo se comportan los datos.

seq_crumbs(`https://bioinf.comav.upv.es/seq_crumbs/`) es un software de código abierto implementado en Python que utiliza Biopython e incluye utilidades para procesar secuencias. Toma un fichero de secuencias como entrada y crea un nuevo fichero de salida con las secuencias procesadas. Dentro de sus muchas funciones, caben destacar: filtrado de secuencias por calidad media, filtrado por longitud según un umbral máximo y mínimo, eliminación de regiones de baja calidad en los extremos (*trimming*), conversión de formatos, etc.

2.2. QIIME

QIIME (`http://qiime.org/index.html`) [4] son la siglas en inglés de *Quantitative Insights Into Microbial Ecology*. Es un *pipeline* bioinformático de código abierto para realizar análisis de microbiomas a partir de datos de secuenciación. Fue construido utilizando la herramienta PyCogent con una implementación modular para poder elegir entre las distintas alternativas dentro de todas sus funciones. Todos los análisis se realizan utilizando *scripts* de python (.py). El flujo de trabajo puede observarse en la figura 2.1.

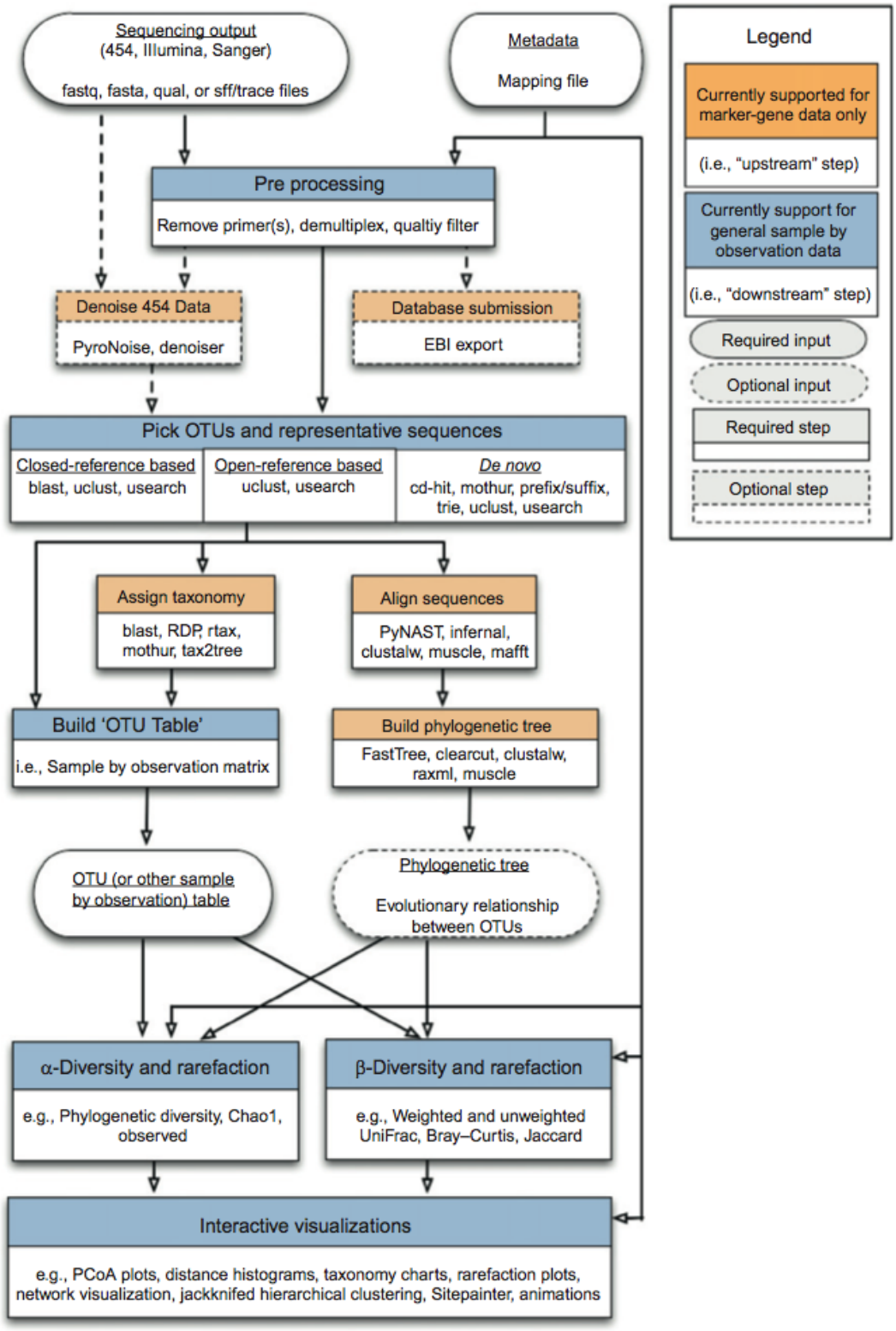

Figura 2.1: Flujo de trabajo de QIIME.

2.2.1. Preprocesado

QIIME incorpora su propio método de preprocesado. Un *script* realiza el filtrado de *reads* por calidad, longitud y el demultiplexado simultáneamente. Para ello utiliza

un fichero "mapa" proporcionado por el usuario que incluye el nombre de la muestra, la secuencia del *barcode*, la secuencia de los *primers* y una descripción como elementos obligatorios (permitiendo añadir más elementos optativos).

2.2.2. Selección de OTUs

OTU (del inglés *Operational Taxonomic Unit*) es una unidad taxonómica operativa, es decir, una unidad de clasificación elegida por el investigador para individualizar los objetos de su estudio sin juzgar si se corresponden a una entidad biológica particular. Se aplica cuando se tienen datos de secuencias de ADN o morfológicos. Puede considerarse OTU un individuo, una población, una especie o cualquier otro taxón. QIIME ofrece tres estrategias de selección diferentes para este paso:

- ***Closed-reference***: Las lecturas son agrupadas contra una colección de secuencias referencia y las que no agrupan son excluidas del análisis. Es el método más rápido, al ser muy paralelizable y se obtienen mejores taxonomías porque son OTUs definidas previamente. Sin embargo, no permite detectar nuevas OTUs así que depende mucho de lo bien caracterizada que esté la base de datos.
- ***De novo***: Las lecturas se agrupan por similitud unas contra otras, sin ningún tipo de referencia externa. El beneficio es que todas las *reads* son agrupadas pero no es paralelizable por lo que sería un proceso muy lento para grandes sets de datos.
- ***Open-reference***: Las lecturas son agrupadas contra la referencia y las que no se encuentran en la referencia son agrupadas posteriormente *de novo*. Presenta la ventaja de que todas las reads son agrupadas y además es paralelizable la mitad del proceso. Suele ser la estrategia preferida aunque no es recomendable utilizarla en datos con pocas referencias porque el proceso puede tardar días en completarse.

De entre los diferentes métodos que incorpora QIIME, se ha elegido **uclust** (`http://drive5.com/usearch/manual/uclust_algo.html`). Es un algoritmo diseñado para agrupar secuencias de nucleótidos o aminoácidos en base a su similitud. Cada grupo o *cluster* está definido por una secuencia representativa conocida como "centroide". Sigue dos criterios de agrupamiento simples, con respecto a un umbral de similitud (T) dado: (1) todas las secuencias dentro de un *cluster* tienen similitud >= T con la secuencia centroide y (2) todos los centroides tienen similitud < T entre ellos. Hay que tener en cuenta que una secuencia puede coincidir con dos centroides diferentes con similitud > T. Idealmente, se asignará al centroide más cercano, pero puede haber dos o más a la misma distancia, en cuyo caso la asignación de *cluster* es ambigua y debe hacerse una elección arbitraria. La similitud se calcula utilizando alineamiento global. Además, se trata de un algoritmo voraz (también conocido como *greedy*) que es aquél que elige la opción óptima en cada paso local esperando llegar a una solución general óptima, por lo que es importante el orden

en que van entrando las secuencias. Si la secuencia entrante coincide con un centroide existente, se asigna a ese grupo y si no coincide, se convierte en el centroide de un nuevo grupo. Esto significa que las secuencias deben estar ordenadas para que los centroides más adecuados tienden a aparecer más temprano. Dado que las lecturas más abundantes tienen más probabilidades de ser secuencias de amplicones correctas, y por tanto son más probables de ser verdaderas secuencias biológicas, considera las secuencias de entrada en orden de disminución de la abundancia.

De todas las bases de datos existentes, se ha elegido como referencia **Greengenes** (`http://greengenes.secondgenome.com`). Contiene taxonomía de 16S de calidad controlada, basada en una filogenia *de novo* que proporciona conjuntos de OTUs estándar. Está bajo la licencia Creative Commons BY-SA 3.0.

Al final del proceso, se obtiene una tabla cuya primera columna es el identificador de OTU y la segunda columna son los conteos que pueden obtenerse en valor absoluto o relativo.

2.2.3. Asignación de taxonomía

Una vez se ha creado la tabla OTU, QIIME permite asignar una taxonomía a cada secuencia representativa. Actualmente los métodos implementados son BLAST, clasificador RDP, RTAX, mothur y uclust. Después se realiza un resumen de la representación de los grupos taxonómicos dentro de cada muestra según un nivel elegido por el usuario. Ese nivel dependerá del formato que se devuelva desde el paso de la asignación de taxonomía. La base de datos de Greengenes utiliza los siguientes niveles:

- Nivel 1 = Reino (por ejemplo, *Bacterias*),
- Nivel 2 = Filo (por ejemplo, *Actinobacteria*),
- Nivel 3 = Clase (por ejemplo, *Actinobacteria*),
- Nivel 4 = Orden (por ejemplo, *Actinomycetales*),
- Nivel 5 = Familia (por ejemplo, *Streptomycetaceae*),
- Nivel 6 = Género (por ejemplo, *Streptomyces*),
- Nivel 7 = Especies (por ejemplo, *mirabilis*).

La salida de este proceso es una tabla donde la primera columna es la taxonomía y en la segunda columna se mantienen los conteos en valor absoluto o relativo.

Un paso adicional que incorpora es el alineamiento de las secuencias para construir un árbol filogenético que incluye las OTUs *de novo*. Estos dos últimos pasos, asignación taxonómica y generación de árboles, son opcionales dejando al usuario la alternativa de desactivarlos si no le resultan necesarios.

2.3. CCruncher

For example, your methods might be contingent upon the ideal gas law

$$PV = nRT. \tag{2.1}$$

The constant R in (2.1) is called the *Boltzmann constant* and has value

$$R \approx 8{,}314\ {}^{\mathrm{J}}/_{\mathrm{K \cdot mol}}.$$

We are measuring the volume V in meters (m), the pressure P in pascals (Pa), the amount of gas n in moles (mol), and the temperature T in kelvin (K).

2.4. Fourier

2.5. LIMITS: Lotka-Volterra

3 Resultados

3.1. Estado de los datos

Los datos utilizados en el presente trabajo proceden del estudio *David et al.* [1] y se encuentran en el repositorio EBI (European Bioinformatics Institute) ENA (European Nucleotide Archive) con el número de acceso ERP006059.

Comprenden un total de 820 ficheros en formato fastq cada uno de los cuales corresponde a un día de toma de muestra y secuenciación. Los donantes fueron dos varones de 26 y 36 años denominados sujeto A y B, respectivamente. Las muestras fueron tomadas de saliva y heces para analizar el microbioma de boca e intestino, generando tres grupos de estudio:

- Boca del donante A: muestras recogidas entre los días 26-364 que comprenden un total de 286 ficheros.
- Intestino del donante A: muestras recogidas entre los días 0-364 que comprenden un total de 342 ficheros.
- Intestino del donante B: muestras recogidas entre los días 0-318 que comprenden un total de 192 ficheros.

No se tomaron muestras de saliva del sujeto B. Como puede observarse en los grupos anteriores, la saliva comenzó a recolectarse más tarde y hay que destacar que en todos los grupos hay algunos días sin muestra (razones sin especificar). Las muestras las tomaban los propios donantes en casa guardándolas temporalmente a -20ºC hasta que se transportaban al laboratorio donde se almacenaban a -80ºC.

También se tomaron metadatos sobre el estilo de vida mediante una aplicación iOS que utiliza una base de datos SQL donde los sujetos anotaban diariamente 13 categorías: alimentación, movimientos intestinales, notas, dieta, ejercicio, aptitud física, cambio de ubicación, medicación, estado de ánimo, higiene bucal, sueño, micción y consumo de vitaminas.

Respecto a la identificación de microorganismos, se decantaron por secuenciar la región V4 del ARN ribosomal 16S con la plataforma Illumina GAIIx. El ADN fue amplificado utilizando *barcoding* y secuenciando lecturas *paired end* de 100 pb. El primer obstáculo con los datos se encuentra aquí, ya que en el repositorio no hay dos archivos por muestra como suele ocurrir cuando se trabaja con *paired end* (Materiales y métodos). Existen 820 archivos únicos (uno por día) con *reads* de longitud <= 100 pb. Se desconoce el procedimiento llevado a cabo por los autores, aunque puede hipotetizarse que o bien

utilizaron *single end* pero han cometido una errata al describir la forma de secuenciación, o bien utilizaron *paired end* pero posteriormente los solaparon creando lecturas *single* de 100 pb con mejor calidad o incluso que solo hayan utilizado uno de los dos pares (5' o 3'). Para este trabajo son necesarias secuencias no apareadas (únicas) para posteriores análisis así que se da por hecho que las secuencias de los ficheros descargados del repositorio son *single* ya que 100 pb son suficientes para una resolución biológicamente significativa si se eligen juiciosamente los cebadores [5].

Se han encontrado otras incidencias en los datos como la falta de metadatos para la muestra "Stool69.1260101.fastq" y la existencia dos muestras para el mismo día (concretamente los días 79, 127, 128, 231, 238, 275, 277, 284 en saliva sujeto A; los días 7, 44, 74, 79, 82, 84, 106, 120, 162, 277 en intestino sujeto A y el día 177 en intestino sujeto B).

Para hacerse una idea de la calidad de los datos, se utilizó FastQC generando un informe para cada uno de los 820 ficheros. Como es muy tedioso ir inspeccionando uno por uno, se utilizó MultiQC para obtener un fichero resumen de todos ellos. El resultado puede observarse en la figura 3.1. El primer plot muestra la calidad medida con *phred score (q)* a lo largo de las bases en las 820 muestras. En general la calidad es buena pues casi todas las bases (a excepción de dos) presentan valores superiores a 20. La calidad tiende a bajar un poco en los extremos de las secuencias, fenómeno que suele darse en secuenciación frecuentemente. El segundo plot muestra la calidad en base al número de secuencias. Se forma algo parecido a una campana de Gauss, mostrando que la mayoría de las secuencias tienen calidad q=35 y hay muy pocas de "mala calidad" (q<30) en su extremo izquierdo.

3.2. Preprocesado

Antes de realizar cualquier análisis es necesario un preprocesado. Del secuenciador se obtiene el fragmento de ADN mencionados en el apartado 1.1.2 – *barcoding* (figura 1.2). A esto se le denomina "multiplex" y a la acción de procesarlo se le denomina "demultiplexar". Además, hemos visto que la plataforma de secuenciación no es perfecta y en ocasiones se obtienen calidades no deseadas.

En este caso no es necesario realizar un demultiplexado ya que los autores han realizado este paso previamente y los datos que se encuentran en el repositorio ya están libres de adaptadores. Se comprobó buscando la secuencia del cebador de PCR directo (GTGCCAGCMGCCGCGGTAA) y el cebador reverso (GGACTACHVGGGTWTCTAAT) en las *reads* pero no fueron hayadas (ni tampoco las secuencias reversa, complementaria y reversa-complementaria a los cebadores). Por ello se deduce que ya fueron eliminados.

En general las calidades de secuenciación eran buenas como se ha visto en el apartado anterior, pero aún pueden eliminarse algunas secuencias que tenían peor calidad. El valor q al cual filtrar es arbitrario, siempre hay que llegar a un compromiso entre quedarse

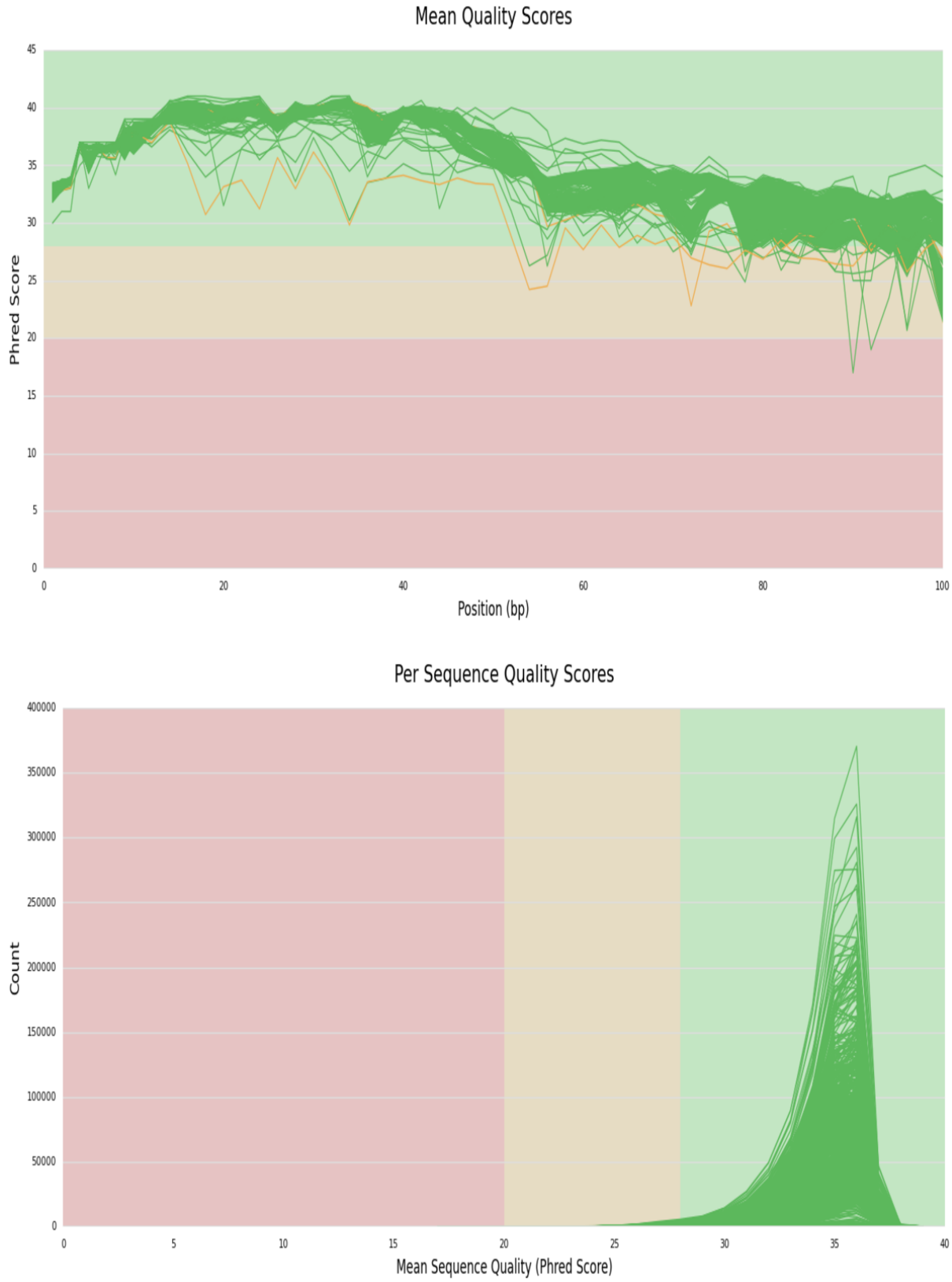

Figura 3.1: Múltiples imágenes

con lecturas de buena calidad pero sin perder demasiada información. En este caso, con un valor de 30 se pierden pocas lecturas y nos quedamos con una buena calidad: habían

211.731.053 *reads* de partida y tras el filtrado quedan 208.266.760 *reads*, así que se ha eliminado el 1,636 % de las lecturas al filtrar. Las secuencias fueron filtradas con seq_crumbs para eliminar todas aquellas con calidad media q<30. Los resultados pueden observarse en la figura 3.2. Como se ha filtrado por calidad media no se observa ningún cambio en el primer plot de calidad a lo largo de las bases. Sin embargo, en el segundo plot se ve muy claro que el programa ha eliminado todas las *reads* por debajo de 30, perdiendo esa cola que rozaba la franja sombreada en naranja.

El número de lecturas que se obtienen cada día también es un factor importante. Si en un fichero aparecen tan solo 2 o 3 *reads* es un indicativo de que algo no salió bien en la secuenciación ese día. Por tanto, se ha realizado un paso más de preprocesado de los datos, eliminando aquellos ficheros que contenían un número de *reads* inferior a 10.000. Este valor también es arbitrario en base al compromiso cantidad-calidad de información, esto es, si elimino muchas secuencias me quedaré sin información pero si incluyo los ficheros con pocas *reads* estaré introduciendo errores a mi análisis. Los días eliminados del estudio aparecen detallados en la tabla 3.1. De 208.266.760 *reads* que venían del paso anterior, ahora obtenemos 208.231.302 *reads* totales con las que se va a realizar todo el análisis.

ID muestra	Número de *reads*	Donante
Stool448.1259730	1	Sujeto B
Stool196.1259770	2	Sujeto A
Stool13.1259916	4	Sujeto A
Saliva267.1260193	5	Sujeto A
Stool85.1260354	8	Sujeto A
Stool217.1260272	8	Sujeto A
Stool63.1259769	29	Sujeto A
Stool120.1259849	31	Sujeto A
Stool147.1260039	39	Sujeto A
Stool36.1259652	54	Sujeto A
Stool453.1260253	1006	Sujeto B
Stool92.1259811	1423	Sujeto A
Stool452.1259809	1738	Sujeto B
Stool384.1259728	2501	Sujeto B
Stool340.1260381	2746	Sujeto A
Stool4.1260013	3553	Sujeto A
Stool343.1259705	4004	Sujeto A
Stool454.1260333	4491	Sujeto B
Stool382.1260123	6395	Sujeto B
Stool345.1259808	7420	Sujeto A
TOTAL	35458	

Tabla 3.1: Tabla de ancho fijo.

Figura 3.2: Múltiples imágenes

3.3. Clasificación taxonómica

Para este paso se utilizó QIIME v1.9.1. Tras el filtro de calidad, se convirtieron los ficheros fastq en fasta que es el formato de entrada en QIIME. A continuación se eliminaron

quimeras, que son combinaciones de dos o más secuencias producidas durante el proceso de PCR como un artefacto. De 208.266.760 *reads* de partida se obtienen 206.928.490 *reads* tras este paso, por lo que el 0,64 % de las secuencias eran quimeras. Para la selección de OTUs se eligió la estrategia *open-reference* al 97 % de similitud con la base de datos Greengenes y con el método UCLUST. Por último, se asignó la taxonomía resumiendo los taxones a nivel de género (L6).

En este procedimiento se parte de un gran número de ficheros que contienen secuencias de ADN tomadas a lo largo de un año y se obtiene una gran tabla que resume la abundancia absoluta de OTUs (filas) que había cada uno de los días de ese año (columnas). Se genera una tabla de abundancia por cada muestra y sujeto con las siguientes dimensiones:

- Saliva del donante A: 573 (OTUs) x 285 (días).
- Heces del donante A: 582 (OTUs) x 329 (días).
- Heces del donante B: 402 (OTUs) x 186 (días).

Todo este proceso queda detallado en el *pipeline* del anexo 1, en el que pueden encontrarse todos los *scripts* de QIIME utilizados con la explicación de cada opción. Es totalmente reproducible e incluye además un *script* de creación propia implementado en Python, que formatea el fichero de salida de QIIME para ser utiilzado por la siguiente herramienta.

3.4. Estudio estadístico

3.5. Correlaciones

4 Discusión y conclusiones

5 Perspectivas de futuro

Bibliografía

[1] **David, L.A., Materna, A.C., Friedman, J., Campos-Baptista, M.I., Blackburn, M.C., Perrotta, A., Erdman S.E., and Alm, E.J.** 2014. Host lifestyle affects human microbiota on daily timescales. Genome biology, **15**(7), p. R89.

[2] **Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin E.M., Rokhsar D.S., and Banfield, J.F.** 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature, **428**(6978), pp.37-43.

[3] **Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W. and Fouts, D.E.** 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science, **304**(5667), pp.66-74.

[4] **Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., and Huttley, G. A.** 2010. QIIME allows analysis of high-throughput community sequencing data. Nature methods, **7**(5), pp.335-336.

[5] **Liu, Z., Lozupone, C., Hamady, M., Bushman, F. D., and Knight, R.** 2007. Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic acids research, **35**(18), p. e120.