Metagenómica

Resumen

La metagenómica es el estudio de la estructura y función de todas las secuencias de nucleótidos aisladas y analizadas de todos los organismos (habitualmente microbios) en una muestra a granel ambiental. Forma parte de las ciencias ómicas, conociéndose como genómica ambiental, ecogenómica o genómica de la comunidad.

Sandra Mingo Ramírez

UAM - 2024/25

7 de abril de 2025 10:53

Universidad Autónoma de Madrid
Bioinformática y Biología Computacional

Código en Github

Índice general

Capítulo I

Introducción a la metagenómica

I.1. Importancia de la microbiología

Los microorganismos desempeñan un papel crucial en la biosfera debido a su participación en el reciclaje de recursos en los ciclos biogeoquímicos y su influencia en eventos histórico-globales, como la primera extinción masiva del planeta (O_2). Sólo las cianobacterias son responsables de 1/2 de la actividad fotosintética del planeta.

A nivel aplicado, su relevancia es evidente en campos como: la medicina, por su capacidad patogénica, o la agricultura, donde una problación microbiana (bacterias, hongos, etc.) equilibrada y sana en el (ecosistema que constituye el) suelo es fundamental para el crecimiento saludable de las plantas (como nosotros con los probióticos y los bífidus). La optimización de cultivos agrícolas puede lograrse mediante la modulación de las relaciones microbianas. Lo que permite reducir la dependencia de agroquímicos, cuyo uso excesivo hipersaliniza el suelo, genera toxicidad y contaminan gran parte de los cuerpos acuíferos.

Los microorganismos también se pueden aplicar en otras tareas como la biorremediación, donde se emplean para restaurar ecosistemas contaminados por hidrocarburos, metales pesados y otros contaminantes. También se explora la producción de biocombustibles a partir de biomasa para reducir la dependencia en los hidrocarburos. Estos ejemplos ilustran solo algunas de las aplicaciones prácticas de la microbiología.

En el cuerpo humano, las bacterias superan 10^1 en número a las células propias y constituyen aproximadamente 2 kg de biomasa. Estos albergan alrededor de 3.3 millones de genes, distribuidos en más de 1000 spp. de bacterias (estudio de 2010), que exceden por mucho los de nuestro genoma (240K genes). Alteraciones en la microbiota, como el desequilibrio entre *Firmicutes* y *Bacteroidetes*, pueden afectar el metabolismo de nutrientes y contribuir a condiciones como la obesidad, como se ha demostrado en estudios con ratones.

Queremos caracterizar e identificar estas comunidades microbianas, sin embargos, caracterizar microorganismos no es sencillo. La aproximación clásica, basada en el cultivo (axénico) y observación microscópica, es limitada, ya que muchas bacterias no son cultivables en laboratorio. Este fenómeno, conocido como la "anomalía del recuento en placa", implica que solo el 15-35 % de la microbiota puede cultivarse

(desconocemos sus condiciones de crecimiento o estos son mutalistas obligados). Por lo que no podemos cubrir la diverversidad ni saber todo lo que están haciendo dentro de nuestro organismo. Esto ha llevado al desarrollo de la metagenómica, que permite estudiar comunidades microbianas directamente en su entorno natural sin necesidad de cultivo.

I.1.1. Conceptos clave: microbiota, microbioma y metagenoma

- **Microbiota**: Conjunto de microorganismos presentes en un entorno específico (punto de extracción), como el colon o la cavidad bucal, cada uno con una composición característica.
- **Microbioma**: Genoma colectivo de una microbiota, es decir, la suma de los genomas de todos los microorganismos presentes. Este término fue acuñado por Hooper y Gordon en 2001. Entiéndase que si consiguiéremos conocer nuestro microbioma, ¡podríamos saber las rutas metabólicas e inferir las condiciones de crecimiento de la microbiota de cada individuo (persona)!
- **Metagenoma**: es visualmente... "el completo": la suma cada genoma individual de cada individuo. Es un concepto reciente nacido de la limitación de los anteriores. La metagenómica es por definición: el conjunto de técnicas (y avances) de biología molecular que permiten analizar y discernir el microbioma sin necesidad de cultivarlo.

I.2. Metagenómica

La **metagenómica**, definida en (K.Chen & L. Pachter, 2005), aplica técnicas genómicas modernas para estudiar comunidades microbianas en su entorno natural, evitando la necesidad de aislar y cultivar especies individuales. No debe confundirse con el *metabarcoding* o metagenómica funcional, que utiliza genes marcadores [1] específicos que permiten discernir los distintos individuos en una población. El metagenoma contiene toda la información del microbioma, pudiendo tener acceso a las rutas metabólicas y las condiciones de crecimiento. De hecho, la Wikipedia abarca esa definición, pero también comprende el análisis mediante genes marcadores, pero esto no está estrictamente incluido y es una incorrección; la metagenómica es solo para (meta)genomas completos, mientras que la secuenciación de amplicones es para marcadores específicos.

[1] Un gen marcador es un gen que aparece en todos los genomas de una población (o en la parte de la población que nos interese) y cuya evolución de cada gen en cada individuo debe permitir discernir a los individuos entre sí. En resumen, debe estar conservado y presentar pequeñas diferencias. Así, se pueden taggear los distintos individuos.

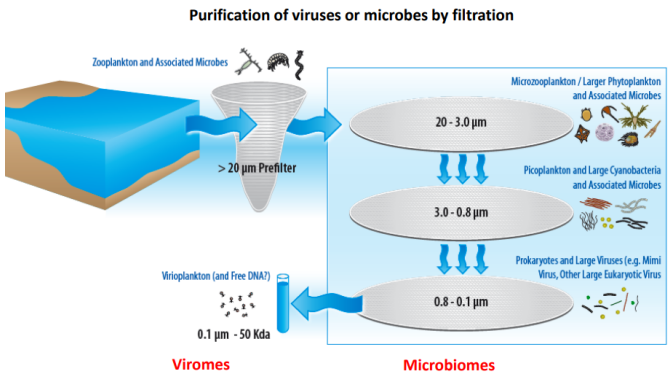
I.2.1. Genómica vs metagenómica

La genómica tradicional requiere el aislamiento y amplificación del genoma de organismos individuales, pero conlleva la pérdida de una gran parte de la población microbiana. En contraste, la metagenómica permite el aislamiento y ensamblaje directo de genomas en contigs, aunque con el riesgo de generar quimeras (genomas que sean mezclas de varios individuos).

La metagenómica no solo se limita a caracterizar la composición microbiana, sino que también busca entender su funcionalidad. Esto incluye la identificación de rutas metabólicas y actividades enzimáticas, lo que es aplicable tanto a bacterias como a hongos. Para ello, pueden utilizanse técnicas como librerías de expresión metagenómicas (analizar un tipo de expresión génica que nos interese, p.ej: la actividad nitrogenasa) o herramientas bioinformáticas: se ensambla el metagenoma y con BLAST se buscan los genes de interés en los contigs. Los resultados de esta segunda ruta deben contrastarse experimentalmente clonando (PCR) y amplificando -> para análisis funcional. No obstante, para ello hay que buscar secuencias concretas de las que se conoce su funcionalidad, pero esto es un proceso lento. Además, algunos organismos pueden haber desarrollado otra secuencia a través de una línea de evolución diferente con una funcionalidad similar, pero que no es conocida (sesgo). Los métodos tradicionales, como la PCR y la clonación, son laboriosos y lentos. Hoy en día, las tecnologías ómicas (metagenómica, metatranscriptómica, metaproteómica y metametabolómica) permiten un análisis/ caracterización más rápido y completo de las comunidades microbianas, de sus hábitats y de sus interacciones/ relaciones ecológicas.

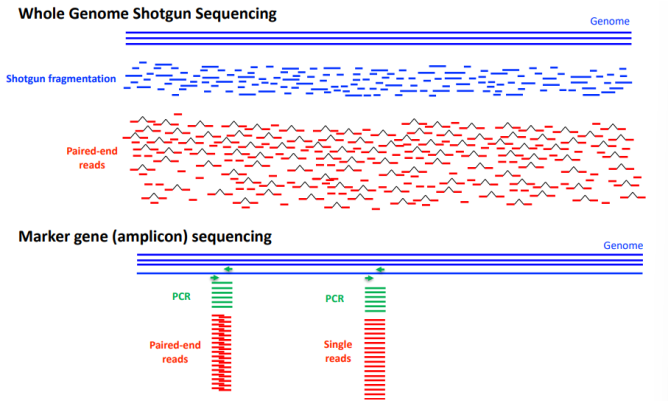
Los procesos en metagenómica son los siguientes:

1. **Diseño experimental y muestreo:** Es crucial seleccionar muestras biológicas adecuadas y diseñar experimentos que permitan obtener resultados representativos. Las muestras pueden provenir de agua, suelo o tejidos humanos, y cada tipo requiere protocolos específicos. Por ejemplo, en el caso de una columna de agua extraída del mar, se pueden utilizar filtros de distinto tamaño, y nos quedamos con el filtro adecuado en función de lo que se desea estudiar. En el caso de bacterias, filtros de 8-9 micras. Y en el caso de cianobacterias y sus microorganismos asociados, de 3 a 8 micras.



2. **Extracción y secuenciación del ADN:** Una vez obtenida la muestra, se extrae el ADN y se prepara para la secuenciación, ya sea mediante fragmentación o amplificación de genes marcadores (como el 16S rRNA).

3. **Ensamblaje y anotación:** Los fragmentos de ADN se ensamblan en contigs, que luego se comparan con bases de datos para identificar genes y asignar taxonomía. Sin embargo, solo se reconstruye alrededor del 50 % del metagenoma debido a la presencia de genes desconocidos (no se sabe anotar la mitad de las lecturas). Además, se utilizan umbrales que descartan contigs de menos de 1000 pares de bases, pudiendo descartar poblaciones minoritarias. Esto se puede salvar utilizando metabarcoding secuenciando un gen (por ejemplo, el gen 16S de la subunidad pequeña del ribosoma) para que las poblaciones minoritarias sí encuentren representación. No obstante, hay que tener en cuenta que con un fragmento tan pequeño no es tan fácil discernir la especie.

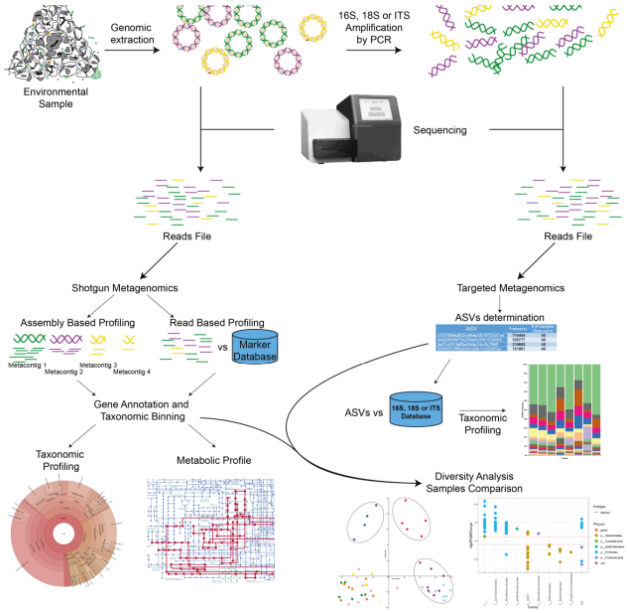


4. **Análisis de diversidad y funcional:** Se utilizan herramientas bioinformáticas para analizar la diversidad microbiana y las rutas metabólicas presentes en un individuo. Esto incluye la creación de perfiles taxonómicos y funcionales. Esto sirve, por ejemplo, para comparar la diversidad microbiana bajo una condición experimental con fármaco. Se ha multiplicado exponencialmente la cantidad de artículos relacionados con el metagenoma y la cantidad de datos en las bases de datos. De hecho, cuando se publican los datos, una de las primeras cosas que piden es subirlo a repositorios como MGnify, MG-Rast, HMP, NCBI-SRA, Metavir, etc. En ellas se explica cómo procesar los datos y clasificarlos para facilitar su posterior búsqueda.

I.2.2. Ribosoma como reloj evolutivo

El ribosoma se ha conocido tradicionalmente como el reloj evolutivo, al ser la maquinaria que permite la traducción del ARNm a proteína. Si hay errores en el ribosoma, la mutación es deletérea, al haber una gran presión evolutiva en el ribosoma. Pero no es constante, se pueden admitir pequeñas variaciones, las cuales se pueden utilizar para discernir los distintos grupos taxonómicos.

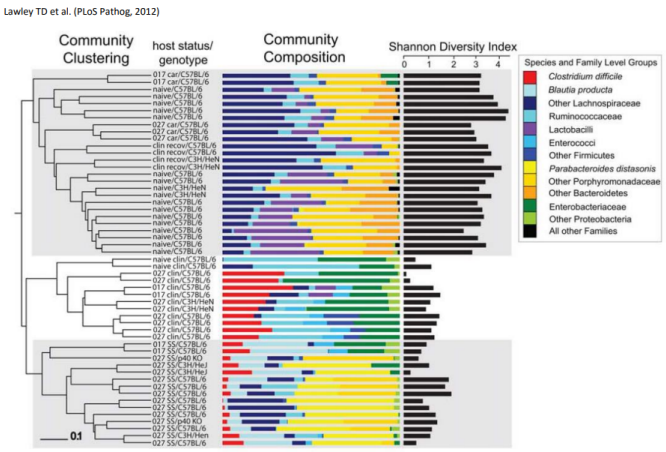
El ARN ribosomal está presente en todos los organismos vivos, y está conservado al tener un papel crucial. Además, se utiliza para generar un árbol filogenético de todas las especies. Esto sirvió para ver que los ribosomas humanos se parecen más a las arqueas que a las bacterias. Y se encontró que el 33 % de la diversidad del árbol corresponde con individuos nuevos que no se habían caracterizado, denominados CPR (*Candidate Phyla Radiation*), con un papel bastante crucial para el establecimiento de poblaciones.

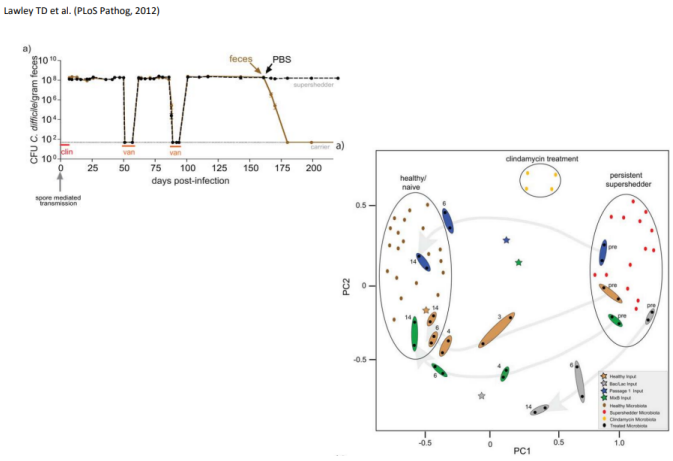


I.3. Aplicaciones de la metagenómica

La metagenómica ha revolucionado el estudio de la microbiota humana y de los ecosistemas. Proyectos como el **Human Microbiome Project** buscan caracterizar el microbioma humano en diferentes condiciones y localizaciones, identificando correlaciones entre la composición microbiana y enfermedades como la obesidad, la psoriasis y las enfermedades cardiovasculares. A este conjunto de enfermedades se le conoce como **disbiosis**, patologías causadas por desregulaciones de las poblaciones microbianas. No se conoce si se provoca primero la enfermedad y en base al microclima cambia también la microbiota o al revés, pero la caracterización del microbiota sirve como marcador de estas enfermedades.

Un ejemplo destacado es el tratamiento de infecciones por *Clostridium difficile* mediante trasplantes de microbiota fecal, que han demostrado restaurar la diversidad microbiana y mejorar la salud del paciente. Se utiliza el índice de Shannon como índice de diversidad (se ve disminuido en el sobrecrecimiento de *Clostridium difficile* y aumentado con el transplante).





I.4. Conclusiones

La metagenómica es una herramienta poderosa para estudiar comunidades microbianas en su entorno natural, superando las limitaciones de los métodos tradicionales. Su aplicación abarca desde la agricultura y la biorremediación hasta la medicina, ofreciendo insights valiosos sobre la diversidad y funcionalidad de los microorganismos. Sin embargo, su éxito depende de un diseño experimental riguroso y del uso adecuado de herramientas bioinformáticas para interpretar los datos generados.

Capítulo II

Análisis de poblaciones ambientales mediante genes marcadores

II.1. Análisis del microbioma mediante metabarcoding - Perfil de las comunidades 16S

El objetivo del metabarcoding es utilizar un gen marcador para caracterizar y cuantificar los miembros de una comunidad microbiana compleja, permitiendo comparar comunidades y evaluar su similitud o disimilitud. El gen marcador más utilizado es la subunidad 16S del ARN ribosomal, debido a su conservación evolutiva y su utilidad como estimador filogenético. Las mutaciones en este gen suelen ser deletéreas, ya que afectan la estructura y función del ribosoma. Sin embargo, ciertas regiones del gen 16S permiten variaciones, lo que las convierte en marcadores ideales para estudios filogenéticos.

El gen 16S completo tiene aproximadamente 1.500 nucleótidos, pero su secuenciación completa solo es posible con tecnologías como Oxford Nanopore y PacBio, que son más costosas y menos eficientes para estudios de diversidad. Por ello, se suelen amplificar y secuenciar regiones hipervariables (como V3-V4 o V4-V5), que son más cortas y adecuadas para plataformas como Illumina. Estas regiones se seleccionan en función del organismo o entorno estudiado. Por ejemplo, en plantas, donde el ADN mitocondrial y cloroplástico puede interferir, se utilizan cebadores específicos (como V3-V5) o se aplican filtros previos para evitar contaminación.

II.1.1. Herramientas para el análisis de metabarcoding

Existen varias herramientas bioinformáticas para analizar secuencias de metabarcoding, como **Mothur** y **QIIME2**, que ofrecen flujos de trabajo similares pero en entornos distintos (como Microsoft Word vs Google Docs). Ambas utilizan algoritmos como **vsearch** o **DADA2** para determinar la diversidad bacteriana. Estos algoritmos, implementados en R, también pueden usarse directamente mediante scripts personalizados. Otros paquetes útiles incluyen:

- **Phyloseq:** Permite almacenar secuencias junto con metadatos para comparar secuencias.
- **Vegan:** Facilita estudios de diversidad.
- **Microbiome:** Combina funcionalidades de Phyloseq y Vegan para analizar diferencias entre poblaciones.
- **Phangorn:** Utilizado para calcular árboles filogenéticos.

II.1.2. Flujo de trabajo básico en metabarcoding

Importar datos Para analizar las lecturas, se necesita un **archivo de metadatos**, que es un archivo de texto tabular donde la primera columna identifica las muestras y las siguientes contienen descriptores relevantes (como condiciones experimentales). Es crucial asegurarse de que el formato del archivo sea correcto, especialmente en Windows, donde los saltos de línea deben ser de un solo bit (LF en lugar de CRLF).

Demultiplexado El primer paso en el flujo de trabajo es separar las lecturas crudas del secuenciador según la muestra a la que pertenecen. Este proceso, llamado demultiplexado, genera archivos individuales para cada muestra.

Denoising y Clustering Las lecturas demultiplexadas pueden contener errores de secuenciación. El proceso de denoising elimina estos errores, mientras que el clustering agrupa las secuencias biológicamente significativas. El resultado es una tabla de abundancias (feature table), que muestra la frecuencia de cada secuencia en cada muestra, y un archivo FASTA con las secuencias representativas. Con estos datos, se pueden realizar análisis posteriores, como:

- Asignación taxonómica en los 7 niveles.
- Alineación de secuencias para establecer relaciones filogenéticas.
- Estimación de diversidad y abundancia diferencial.
- Representaciones gráficas (heatmaps, PCA, etc.).
- Análisis estadístico.

Entre los factores a tener en cuenta está la calidad de las lecturas. Illumina tiene una plataforma de secuenciación (novaseq) que no se recomienda para estos estudios. Esto se debe a que está parametrizado por inteligencia artificial, y para 16S se requieren valores continuos no parametrizados.

En cuanto al denoising y clustering en detalle, se pueden diferenciar los siguientes pasos:

1. **Eliminación de adaptadores**: Se utilizan herramientas como Cutadapt para eliminar adaptadores de secuenciación (Illumina) y de PCR. Las plataformas de secuenciación suelen eliminar solo los adaptadores propios, por lo que este paso es esencial.

2. **Filtrado de calidad**: Se analiza la calidad de las lecturas y se filtran aquellas que no cumplen con los estándares requeridos.

3. **Clustering**: Se agrupan las secuencias en función de su homología. Algoritmos como vsearch, DADA2 o Deblur se utilizan para este fin. El resultado es una tabla de frecuencias y un archivo FASTA con las secuencias representativas.

4. **Eliminación de Singletons y Raretones**: Las secuencias que aparecen una sola vez (singletons) o en baja frecuencia (raretons) se eliminan, ya que suelen ser artefactos. Sin embargo, esto puede llevar a la pérdida de poblaciones minoritarias, por lo que debe hacerse con precaución.

5. **Detección de quimeras**: Las quimeras son secuencias artificiales generadas durante la PCR, donde dos o más secuencias biológicas se fusionan (normalmente cuando la fase de elongación es incompleta). Se detectan comparando las secuencias más abundantes y eliminando aquellas que parecen ser mezclas.

II.1.3. Unidades taxonómicas operativas (OTUs) y Variantes de secuencia de Amplicones (ASVs)

II.1.3.1. OTUs (Operational Taxonomic Units)

Las OTUs son grupos de secuencias con una homología superior al 97 %, lo que permite agrupar secuencias similares en una única unidad taxonómica. Existen tres estrategias principales para asignar OTUs:

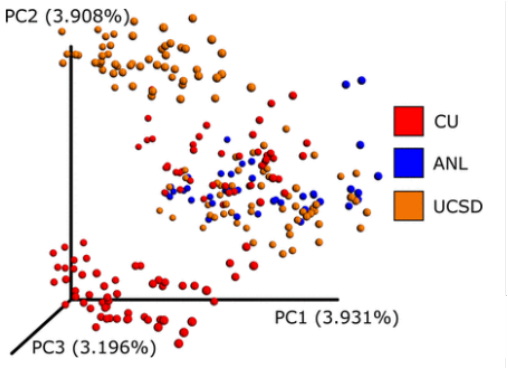
- **OTUs de novo**: Las secuencias se comparan entre sí sin utilizar una base de datos externa. Es útil cuando no hay referencias disponibles, pero es computacionalmente costoso. Esto no es conveniente cuando los amplicones no solapan, pero la ventaja es que todas las lecturas van a ser clasificadas, incluidas aquellas desconocidas que no cuentan con una referencia externa.

- **OTUs de referencia cerrada**: Las secuencias se comparan con una base de datos, por lo que sí se pueden utilizar amplicones grandes. Es más rápido, pero las secuencias no encontradas en la base de datos se descartan.

- **OTUs de referencia abierta**: Combina las dos estrategias anteriores. Primero se compara con una base de datos, y las secuencias no identificadas se agrupan de novo. Esta estrategia tiene todas las ventajas y desventajas de las anteriores: no se pueden usar amplicones largos (para aquellas secuencias que no aparezcan en la base de datos), se depende de una base de datos, pero todas las lecturas serán asignadas y a una velocidad más rápida.

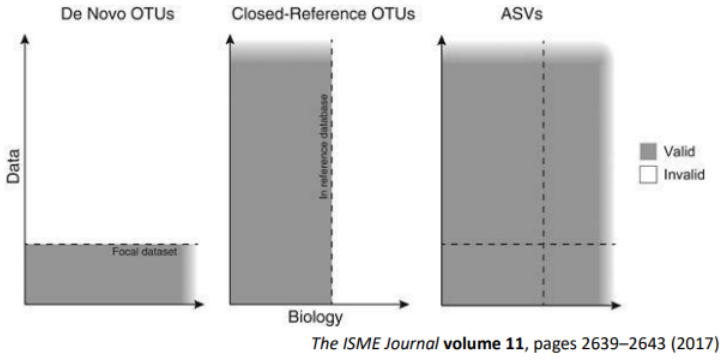
El umbral del 97 % se ha utilizado clásicamente considerando que las lecturas pueden mostrar una probabilidad de error de al menos el 0,1 % por nucleótido. La estrategia de agrupación de OTU reduce ese problema. No obstante, no podemos comparar secuencias de varios experimentos, ya que la OTU (centroide) se calcula cada vez. La secuencia centroide es la secuencia representativa (consenso del alineamiento) que se va a ir modificando conforme se van añadiendo más secuencias miembro al

clúster. Las agrupaciones cambian en función del orden de las secuencias durante la comparación. Sólo son comparables los ensayos que utilizan las estrategias de referencia cerrada y de referencia abierta, aunque existen diferencias.



II.1.3.2. ASVs (Amplicon Sequence Variants)

Las ASVs superan las limitaciones de las OTUs al capturar toda la variación biológica presente en los datos. A diferencia de las OTUs, las ASVs son reproducibles y comparables entre conjuntos de datos, lo que las hace más robustas para estudios a largo plazo.



II.1.3.3. sOTUs

Los Sub-Operational Taxonomic Units (sOTUs) son un enfoque que busca alcanzar una resolución a nivel de nucleótido único mediante métodos estadísticos avanzados. A diferencia de las OTUs tradicionales, que agrupan secuencias con una similitud del 97 %, los sOTUs inferen las secuencias únicas reales trabajando con cada muestra por separado y utilizando distancias de Hamming (que miden las diferencias entre secuencias a nivel de nucleótidos). Este método permite una mayor precisión en la identificación de variantes microbianas.

Sin embargo, este enfoque tiene un costo: se pierden aproximadamente el 50 % de las lecturas debido al filtrado exhaustivo que aplica. A pesar de esto, la resolución obtenida es comparable o incluso superior a la de las OTUs tradicionales. Al igual que en otros métodos, es esencial detectar y eliminar quimeras, que se identifican a partir de las secuencias más abundantes. Los algoritmos más utilizados para este fin son Deblur y DADA2, que han demostrado ser eficaces en la inferencia de sOTUs.

II.1.4. Consideraciones importantes

Correspondencia entre OTUs y especies No se puede asumir una correspondencia directa 1:1 entre las OTUs (o sOTUs) y las especies en una población microbiana. Esto se debe a varias razones:

- Una misma especie puede tener múltiples copias del gen 16S, las cuales no son idénticas debido a variaciones naturales o transferencia horizontal de genes.
- En el mejor de los casos, una OTU puede corresponder a una copia específica del gen 16S, pero no necesariamente a una especie única.

Sesgos experimentales Los resultados pueden verse afectados por errores experimentales, como:

- **Errores de secuenciación:** Las plataformas de secuenciación no son perfectas y pueden introducir errores en las lecturas.
- **Artefactos de PCR:** Durante la amplificación, pueden generarse quimeras o sesgos en la representación de ciertas secuencias.

Estos factores deben tenerse en cuenta al interpretar los resultados, ya que pueden afectar la precisión y la fiabilidad de los análisis de diversidad microbiana.

Capítulo III

Ensamblaje de novo

III.1. Introducción a la metagenómica

La definición de 1998 es que la metagenómica es el estudio del material genético obtenido de muestras ambientales. En el 2005 se actualizó a la aplicación de técnicas genómicas modernas sin la necesidad de aislar y cultivar en el laboratorio las especies individuales.

Las muestas ambientales pueden ser desde agua de mar, tierra, etc, pero también se incluyen muestras de microorganismos asociados a humanos (microbiota intestinal, vaginal, cutánea, etc.). Esto no está restringido a humanos, pudiendo extrapolarlo a plantas y otros animales. En este caso, se denomina muestra asociada a hospedador y no ambiental.

La metagenómica permite estudiar quién está ahí, qué hacen y cómo lo hacen. Las aplicaciones principales incluyen:

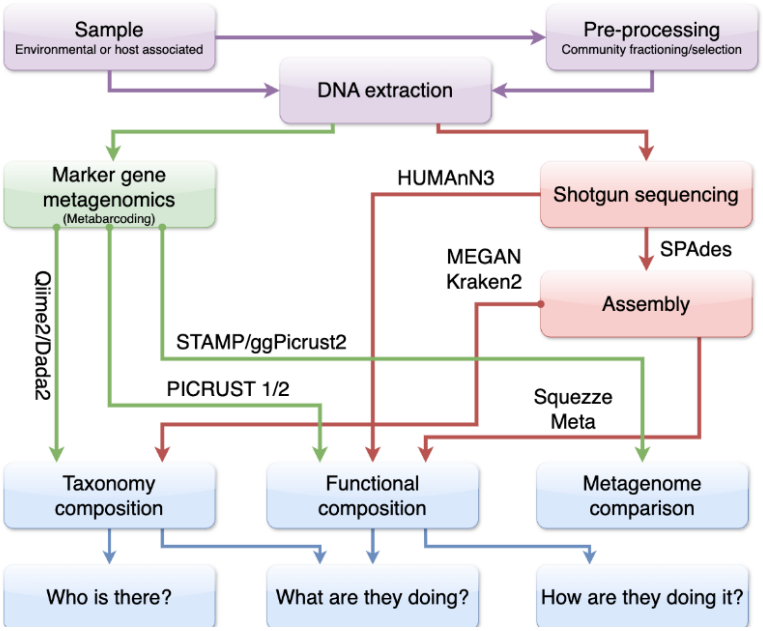
- Biodiversidad
- Evolución de los microorganismos
- Ecología global (ciclos geoquímicos del carbono, producción de oxígeno, etc.)
- Microbiota humana y otros animales (implicación en enfermedades)
- Microbiota de las plantas (simbiosis y mejora de cultivos)
- Bioremediación (reciclaje de aguas residuales y otros residuos)
- Búsqueda de nuevas enzimas para la industria (biocombustibles, etc.)

Así, la definición queda de la siguiente forma: «La metagenómica es el estudio del material genético recuperado directamente de muestras ambientales mediante técnicas genómicas modernas sin necesidad de aislar y cultivar en laboratorio especies individuales». Esto antes se denominaba ecología microbiana.

La ecología microbiana clásica se basaba en la obtención de muestras ambientales, aislamiento y purificación de microorganismos y cultivo de los mismos para su análisis.

Se realizan análisis morfológicos y tinciones, análisis bioquímico, etc. El problema es que la mayoría de los microorganismos no se pueden cultivar.

La metagenómica se impulsó gracias al descubrimiento del 16S como marcador filogenético y la PCR (polymerase chain reaction). Además, surgieron las tecnologías de secuenciación NGS.



III.2. Ensamblajes de novo

Un contig (de contiguous) es un conjunto de segmentos de ADN superpuestos que juntos representan una región consenso de ADN.

Hay dos tipos de ensambladores:

- **Ensambladores basados en OLC (overlap, layout, consensus):**
  - Celera: Ensamblador con el mejor gráfico de solapamiento (CABOG). Diseñado para secuencias Sanger, pero funciona con 454 y lecturas PacBio corregidas de errores.
  - Newbler, también conocido como GS de novo Assembler. Diseñado para secuencias 454, pero funciona con lecturas Sanger.

  Primero encuentra todos los pares de secuencias que solapan. Con eso, se crea un grafo con la información solapante. Se combinan los pares de secuencias que solapan de forma no ambigua y se resuelve encontrando el camino hamiltoniano.

- **Ensambladores basados en DBG (grafo de Bruijn):**
  - EULER (P. Pevzner): el primer ensamblador que utiliza DBG
  - Velvet (D. Zerbino): una opción popular para genomas pequeños

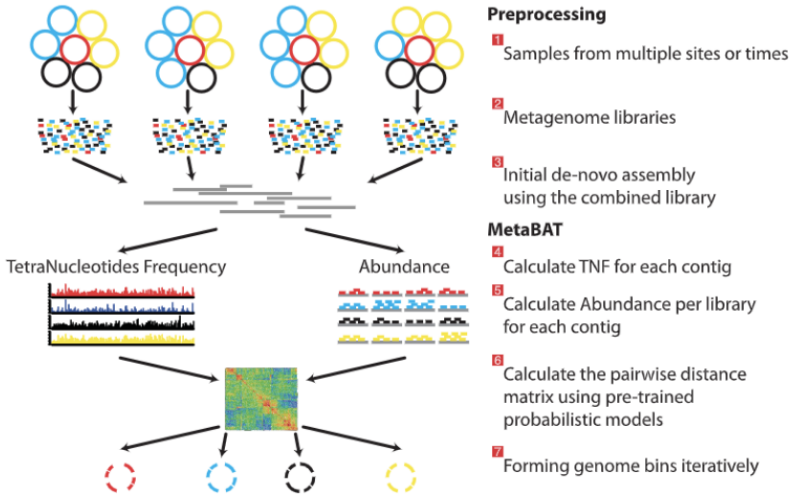
- SOAPdenovo: ampliamente utilizado para ensamblajes relativamente desestructurados
- ALLPATHS-LG: probablemente el ensamblador más fiable para genomas grandes (pero con estrictos requisitos de entrada)
- IDBA: muy popular en metagenómica
- SPAdes: el ensamblador más popular hoy en día
- MegaHit: muy compatible con fasta y poca memoria

Se dividen las lecturas en k-mers y se construye el grafo en el que los bordes son k-mers y los nodos son (k-1)mers. Todo nodo (k-1)-mer está conectado por una arista dirigida a un segundo (k-1)-mer si existe algún kmer cuyo prefijo sea el primero y cuyo sufijo sea el segundo. Esto se resuelve mediante un camino euleriano.

Puede haber problemas en las regiones repetidas por la posibilidad de que formen quimeras.

Se define como binning la agrupación de contigs ensamblados en grupos individuales utilizando varios enfoques diferentes, como la composición nucleotídica, la profundidad de secuenciación, la coabundancia, la taxonomía, etc.

MetaBAT



III.3. Práctica

Esta práctica se realiza en la máquina virtual. El primer paso es crear un entorno de conda. Para ello, utilizamos el comando `conda create -n ngs python=3.11 -y`.

Podemos iniciar el entorno con `conda activate ngs` y es recomendable instalar GDown para poder descargar ficheros desde Google Drive: `pip install gdown`.

Vamos a trabajar con los datos de ECTV. Los descargamos con

```
gdown https://drive.google.com/uc?id=1gtnWLZWdZxn6j-oDvro2sDw3YHPZI2LM
unzip ECTV_reads.zip.
```

III.3.1. Comprobar la integridad

Es muy recomendable comprobar la integridad de los archivos que acabamos de descargar. MD5sum y otros programas más recientes (SHA1sum) son algoritmos que «transforman» el contenido del archivo en una cadena corta de caracteres (hashes). Los hashes no cambian a menos que se modifique el contenido de los archivos (el nombre del archivo no es relevante). Por lo tanto, es muy común que las bases de datos o las instalaciones de secuenciación proporcionen hashes MD5 a los usuarios para permitir la comprobación de la integridad.

- `md5sum ECTV_R1.fastq`: fa3e37e336213d01d927df2a4f0aea12
- `md5sum ECTV_R2.fastq`: 8a569dc04acc87067d33d3d58b26dd6d

Por último, basta con inspeccionar los hashes a ojo para comprobar si hay algún cambio. Por lo general, si los archivos se rompen por cualquier razón, el hash md5 es completamente diferente y sólo mirando a los últimos 5-6 dígitos nos va a mostrar si algo va mal.

III.3.2. Contar número de reads

Ambos archivos deberían tener el mismo número de lecturas (Illumina paried-end reads). Es una buena práctica comprobar el número de lecturas en ambos archivos. A pesar de haber comprobado los hashes MD5, a veces, los archivos subidos a la base de datos son erróneos (el remitente puede haber subido archivos truncados en lugar de los archivos originales). Además, conocer el número de lecturas sería útil para las métricas de calidad de base (lecturas ensambladas/mapeadas o lecturas de pase de calidad).

So, the easy way of counting the number of reads in a file is using wc linux command (word count): `wc -l ECTV_R1.fastq` y `wc -l ECTV_R2.fastq`. Se aplica la opción -l para contar el número de líneas. Sin embargo, tenemos que dividir por 4 para obtener el número de lecturas. Para evitar esto, podemos aplicar algunos piping: `wc -l ECTV_R1.fastq | awk 'print $1/4'`. En este caso, ambos ficheros tienen 50000 líneas.

III.3.3. Comprobar la calidad

Utilizamos FastQC para comprobar la calidad de las secuencias. Se instala con `conda install -c bioconda fastqc -y`.

Ahora se ejecuta el programa: `fastqc ECTV_R1.fastq -o ECTV_Quality/`.

Podemos utilizar las opciones -t para aumentar el número de hilos que utilizará el programa (no es necesario en este pequeño conjunto de datos).

Abra los archivos html para obtener información sobre el número de secuencias, la distribución de longitudes, el contenido de %GCs, la calidad media, etc.

Todos los ticks están en verde salvo "Per base sequence content", pero esto no es preocupante al tratarse de un virus. Esto depende de la naturaleza de la especie de estudio.

Queremos eliminar las lecturas con baja calidad. Para ello, se pueden utilizar herramientas como cutadapt o trimmomatic.

`conda install -c bioconda trimmomatic -y`

Para ejecutar esto, debemos indicar que son secuencias Pair-End,

`trimmomatic PE -phred33 ECTV_R1.fastq ECTV_R2.fastq ECTV_R1_qf_paired.fastq ECTV_R1_qf_unpaired.fastq ECTV_R2_qf_paired.fastq ECTV_R2_qf_unpaired.fastq SLIDINGWINDOW:4:20 MINLEN:700`

Una vez con esto, volvemos a hacer Quality Control para verificar que esté bien.

`mkdir ECTV_QF_Quality fastqc ECTV_R1_qf_paired.fastq -o ECTV_QF_Quality fastqc ECTV_R2_qf_paired.fastq -o ECTV_QF_Quality`

Es importante eliminar las lecturas que alineen al genoma de phiX174. Se utiliza en todas las carreras de Illumina a modo de control interno. Por ello, se deben quitar las lecturas antes de ensamblar. En este caso, también vamos a eliminar lecturas humanas, ya que el virus se ha cultivado en células humanas y podrían quedarse trazas por contaminación. Si se cultivase en células de mono, entonces se deben quitar las lecturas que alineen al genoma de mono. Esto lo haremos con la herramienta Bowtie2, pero también se podría utilizar BWA.

Para la descontaminación del genoma humano, no nos vamos a descargar todo el genoma porque es muy grande y generar el índice tardaría demasiado. Nosotros vamos a utilizar solo las regiones codificantes. Como aun así la creación del índice puede tardar 1 hora, el profesor nos ha dado acceso a los ficheros ya indexados desde su Google Drive:

```
gdown --folder
    https://drive.google.com/drive/folders/1ames4kONYqKlkxObuGbjJLDh2-UwHVdH
for f in human_cds_index/*; do ln -s $f .; done
bowtie2 -x human_cds -1 ECTV_R1_qf_paired.fastq -2
    ECTV_R2_qf_paired.fastq --un-conc ECTV_qf_paired_nohuman_R%.fastq -S
    tmp.sam
wc -l *_qf_paired_nohuman* | awk '{print $1/4}'
```

Y ahora hacemos la decontaminación de phiX.

```
wget
    https://ftp.ncbi.nlm.nih.gov/genomes/refseq/viral/Sinsheimervirus_phiX174/latest_assemb

gunzip GCF_000819615.1_ViralProj14015_genomic.fna.gz
```

Creamos el índice de Bowtie2:

```
bowtie2-build GCF_000819615.1_ViralProj14015_genomic.fna phix

#Aligning human decontaminated reads against PhiX174 index
bowtie2 -x phix -1 ECTV_qf_paired_nohuman_R1.fastq -2
    ECTV_qf_paired_nohuman_R2.fastq --un-conc
    ECTV_qf_paired_nohuman_noPhiX_R%.fastq -S tmp.sam
```

```
#Count decontaminated reads
wc -l ECTV_qf_paired_nohuman_noPhiX_R* | awk '{print $1/4}'}
```

Ahora descargamos Spades, un ensamblador de gráficos de novo. Hay dos protocolos que vamos a comprobar: careful y isolate.

```
spades.py --careful -t 2 -1 ECTV_qf_paired_nohuman_noPhiX_R1.fastq -2
    ECTV_qf_paired_nohuman_noPhiX_R2.fastq -o ECTV_careful

spades.py --isolate -t 2 -1 ECTV_qf_paired_nohuman_noPhiX_R1.fastq -2
    ECTV_qf_paired_nohuman_noPhiX_R2.fastq -o ECTV_careful

grep -c '>' ./ECTV_careful/*.fasta
grep -c '>' ./ECTV_isolate/*.fasta

grep '>' -m 5 ./ECTV_careful/*.fasta
grep '>' -m 5 ./ECTV_isolate/*.fasta

grep '>' ./ECTV_careful/scaffolds.fasta
grep '>' ./ECTV_isolate/scaffolds.fasta
```

Careful tiene menos contigs y scaffolds, por lo que se puede considerar que va mejor. Pero isolate tiene mayor porcentaje de fragmentación, por lo que depende lo que se priorice.

Ahora instalamos quast con `conda install quast -y`. Ahora creamos una carpeta quast y copiamos ahí los ficheros contigs y scaffolds de las carpetas ECTV careful y ECTV isolate (vale con un enlace simbólico). Obtenemos el genoma viral y utilizamos quast:

```
wget
    ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/841/905/GCF_000841905.1_ViralProj14211/GCF_000
gunzip GCF_000841905.1_ViralProj14211_genomic.fna.gz
mv GCF_000841905.1_ViralProj14211_genomic.fna ECTV_reference_genome.fasta
quast.py contigs* scaffolds* -R ECTV_reference_genome.fasta
```

III.4. Ensamblaje de un metagenoma

Para tener un ejemplo más realista de todo el proceso vamos a utilizar lecturas metagenómicas simuladas. Se ha utilizado InSilicoSeq. La ventaja de utilizar reads simuladas en lugar de datos metagenómicos reales reside en que con las reads simuladas podemos tener los genomas originales utilizados para la simulación y la proporción de cada uno en los datos.

Creamos una carpeta nueva y creamos un script de bash llamado virome_script.sh. El contenido es el siguiente:

```
#!/bin/bash
```

```
# We need to add this command to avoid a conda initiation problem when
    running from a script ->
    https://stackoverflow.com/questions/34534513/calling-conda-source-activate-from-bash-sc
eval "$(conda shell.bash hook)"

# Activation ngs environment
conda activate ngs

# Download reads
gdown 1QModYfordyNUOLjnE27-plr-QEftbSi5

# If you are using virtual machine the file is already downloaded
# ln -s /home/metag/Documents/data/viromas/virome_1.tar.gz .
tar -xzf virome_1.tar.gz

# Raw reads quality assessment
mkdir quality
fastqc virome_1_R1.fastq.gz -o quality
fastqc virome_1_R2.fastq.gz -o quality

# Quality filtering
trimmomatic PE -phred33 virome_1_R1.fastq.gz virome_1_R2.fastq.gz \
    virome_1_R1_qf_paired.fq.gz virome_1_R1_qf_unpaired.fq.gz \
    virome_1_R2_qf_paired.fq.gz virome_1_R2_qf_unpaired.fq.gz \
    SLIDINGWINDOW:4:20 MINLEN:150 LEADING:20 TRAILING:20 AVGQUAL:20

# QF reads quality assessment
fastqc virome_1_R1_qf_paired.fq.gz -o quality
fastqc virome_1_R2_qf_paired.fq.gz -o quality

# Decontaminating human reads
bowtie2 -x ../unit_3/human_cds -1 virome_1_R1_qf_paired.fq.gz -2
    virome_1_R2_qf_paired.fq.gz --un-conc-gz
    virome_1_qf_paired_nonHuman_R%.fq.gz -S tmp.sam

# Decontaminating PhiX174 reads
bowtie2 -x ../unit_3/phix -1 virome_1_qf_paired_nonHuman_R1.fq.gz -2
    virome_1_qf_paired_nonHuman_R2.fq.gz --un-conc-gz
    virome_1_qf_paired_nonHuman_nonPhix_R%.fq.gz -S tmp.sam

# Assembly
spades.py -t 4 --careful -1 virome_1_qf_paired_nonHuman_nonPhix_R1.fq.gz
    -2 virome_1_qf_paired_nonHuman_nonPhix_R2.fq.gz -o virome_1_careful
spades.py -t 4 --meta  -1 virome_1_qf_paired_nonHuman_nonPhix_R1.fq.gz -2
    virome_1_qf_paired_nonHuman_nonPhix_R2.fq.gz -o virome_1_meta
spades.py -t 4 --sc    -1 virome_1_qf_paired_nonHuman_nonPhix_R1.fq.gz -2
    virome_1_qf_paired_nonHuman_nonPhix_R2.fq.gz -o virome_1_sc

# Assembly analysis

# Activation quast environment
```

```
mkdir quast
cd quast
ln -s ../virome_1_careful/contigs.fasta virome_1_contigs_careful.fasta
ln -s ../virome_1_careful/scaffolds.fasta
    virome_1_scaffolds_careful.fasta
ln -s ../virome_1_meta/contigs.fasta  virome_1_contigs_meta.fasta
ln -s ../virome_1_meta/scaffolds.fasta virome_1_scaffolds_meta.fasta
ln -s ../virome_1_sc/contigs.fasta    virome_1_contigs_sc.fasta
ln -s ../virome_1_sc/scaffolds.fasta  virome_1_scaffolds_sc.fasta
ln -s ../virome_1_genomes.fasta       virome_1_genomes.fasta
quast.py virome_1_contigs_careful.fasta virome_1_contigs_meta.fasta
    virome_1_contigs_sc.fasta virome_1_scaffolds_careful.fasta
    virome_1_scaffolds_meta.fasta virome_1_scaffolds_sc.fasta -R
    virome_1_genomes.fasta
conda deactivate
```

Damos permisos de ejecución con `chmod +x virome_script.sh` y lo ejecutamos.

El mejor modelo sería mayor número de bases y menor número de trozos. Nunca se consigue alinear el 100 % debido a las regiones repetidas.

Capítulo IV

Clasificación de taxonomía

IV.1. Pipeline de clasificación taxonómica

A la hora de ensamblar un metagenoma, se debe también clasificar la taxonomía. Nosotros, por capacidad computacional, haremos la clasificación taxonómica desde las lecturas, pero dependiendo del programa que se utilice, el camino puede cambiar.

Para una clasificación taxonómica se necesitan las secuencias o contigs, una base de datos con las secuencias y su taxonomía asociada y un alineador. Con las lecturas alienadas, se obtienen las taxonomías.

Hay distintas herramientas online, pero no terminas de controlar lo que hacen:

- MG-Rast: es lento
- EBI-Metagenomics
- IMG/M: integrated microbial genomics and microbiomes
- MetaVir: ya no se mantiene y dejó de funcionar. Esto puede ocurrir con todas.
- iMicrobe: colección de herramientas
- : CyVerse: colección de herramientas

En cuanto a las bases de datos están:

- GenBank del NCBI: la más accesible y que más curada está
- European nucleotide archive (ENA) at EBI: el buscador es más complicado de utilizar, pero es más fácil de descargar secuencias.
- DNA data bank of Japan (DDBJ): está sincronizada con las dos anteriores.
- Uniprot: enfocado en secuencias proteicas
- Uniclust: versión clusterizada de Uniprot con distintos niveles de identidad

Existen los siguientes programas de alineamiento de secuencias:

- Blast
- MMSeqs2
- Diamond
- Centrifuge
- Kraken2

Blast alinea las secuencias localmente en función a la base de datos, por lo que hay que tener en cuenta el tamaño de la misma. El problema de Blast es que es muy lento, por lo que se recomienda utilizar Diamond, que es 2.500 veces más rápido y encuentra más del 94 % de los matches.

IV.2. Práctica

Vamos a utilizar los datos del viroma de la práctica anterior. Podemos copiar los ficheros o crear un enlace simbólico:

```
cd unit_4
ln -s ../unit_3b/virome_1_qf_paired_nonHuman_nonPhix_R1.fq.gz
    virome_1_qf_R1.fq.gz
ln -s ../unit_3b/virome_1_qf_paired_nonHuman_nonPhix_R2.fq.gz
    virome_1_qf_R2.fq.gz
```

Además, vamos a descargar lecturas de calidad filtrada de virome_2 para comparar la clasificación taxonómica de ambos viromas:

```
conda activate ngs
gdown https://drive.google.com/uc?id=11xOf45e5aIIKLTc1pEKsUCHpYyC-84NF
gdown https://drive.google.com/uc?id=1TuTyun2dlmUMvsF6N9LK6zAySI3xvqtx

# MD5
# 7c508583dbda80b948b5f88eb879ae16 virome_2_qf_R1.fq.gz
# d1894eec561128bc29e4a3050e0eafaa virome_2_qf_R2.fq.gz

# Virome_2 is also avaible in Moodle
```

Ahora instalamos Diamond, descargamos la base de datos de proteínas del NCBI y creamos la base de datos de referencia para Diamond:

```
conda install -c bioconda diamond -y
#asegurar que sea la versión .10 y no la .11

wget
    https://ftp.ncbi.nlm.nih.gov/refseq/release/viral/viral.1.protein.faa.gz
gunzip viral.1.protein.faa.gz

diamond makedb --in viral.1.protein.faa -d viralproteins
```

Esto creará un archivo binario de base de datos DIAMOND con el nombre: viralproteins.dmnd Como tenemos lecturas emparejadas, y Diamond no puede manejarlas, podemos ejecutar Diamond para cada par y luego fusionar los resultados o podemos fusionar las lecturas emparejadas en un solo archivo y ejecutar Diamond una sola vez. Sin embargo, aunque el tiempo de ejecución de Diamond no es muy elevado (unas 160k lecturas tarda aproximadamente 4 minutos), MEGAN, el programa que vamos a utilizar para analizar los hits de Diamond y asignar la taxonomía a las lecturas, utiliza una enorme cantidad de memoria RAM y con un conjunto de datos grande muchas veces se bloquea. Para evitar esto, vamos a tomar una submuestra aleatoria utilizando seqtk:

```
conda install -c bioconda seqtk -y

# Virome_1
seqtk sample -s 123 virome_1_qf_R1.fq.gz 5000 > virome_1_10k.fq
seqtk sample -s 123 virome_1_qf_R2.fq.gz 5000 >> virome_1_10k.fq

# Virome_2
seqtk sample -s 123 virome_2_qf_R1.fq.gz 5000 > virome_2_10k.fq
seqtk sample -s 123 virome_2_qf_R2.fq.gz 5000 >> virome_2_10k.fq

# Note the ">>" in the second subsample for each virome

head virome_1_10k.m8
```

Este es el significado de las 12 columnas:

1. qseqid significa Seq-id de consulta
2. sseqid significa código de secuencia del sujeto
3. pident significa Porcentaje de coincidencias idénticas
4. length significa Longitud de la alineación
5. mismatch significa número de coincidencias erróneas
6. gapopen significa Número de huecos abiertos
7. qstart significa Inicio de la alineación en la consulta
8. qend significa Fin de la alineación en la consulta
9. sstart significa Inicio de la alineación en el tema
10. send significa fin de la alineación en el tema
11. evalue significa Valor esperado
12. bitscore significa puntuación de bits

Podemos tomar el número de acceso de uno de los hits (columna 2) y pegarlo en NCBI y mirar la taxonomía de estas secuencias. A continuación, puede repetir este paso uno por uno varios miles de veces para tener un perfil taxonómico de estos metagenomas. Alternativamente, puede utilizar un programa específico que toma la salida de la comparación de Blast o Diamond y devuelve el resultado de todos los matches juntos en un gráfico.

MEGAN6 analiza el contenido taxonómico de un conjunto de lecturas de ADN alineadas con un conjunto de datos del NCBI y asigna las lecturas a un árbol taxonómico.

```
conda install -c bioconda megan -y
gdown https://drive.google.com/uc?id=1330Lx36_mMvylVTUHnDI8iIlSCg0WnA6
MEGAN
```

Con esto se abre el programa de Megan. File > Import from Blast y seleccionamos uno de los viromas en formato BlastTab y modo BlastX. También se debe cargar el la base de datos de MeganMap y se aplica todo. Esto se hace para ambos viromas, y posteriormente File > Compare y seleccionamos ambos.

IV.3. Taxonomía completa con Kraken

Descargamos Kraken y Pavian, el primero en la terminal y el segundo en RStudio. También debemos descargar la base de datos del viroma. Con kraken, le pasamos la base de datos y las lecturas. El output de Kraken se puede visualizar en Pavian. Debemos abrir los ficheros report.txt y podemos ver los resultados de cada viroma e incluso compararlas.

```
conda activate ngs
conda install -c bioconda kraken2

wget https://genome-idx.s3.amazonaws.com/kraken/k2_viral_20221209.tar.gz
mkdir k2_viral
tar -xzf k2_viral_20221209.tar.gz -C k2_viral

# Virome 1
kraken2 -db k2_viral --paired virome_1_qf_R1.fq.gz virome_1_qf_R2.fq.gz
    --report virome_1_report.txt > virome_1_k2_output.txt

# Virome 2
kraken2 -db k2_viral --paired virome_2_qf_R1.fq.gz virome_2_qf_R2.fq.gz
    --report virome_2_report.txt > virome_2_k2_output.txt
```

Para descargar Pavian:

```
if (!require(remotes)) { install.packages("remotes") }
remotes::install_github("fbreitwieser/pavian")

# Run Pavian server
```

```
options(shiny.maxRequestSize=500*1024^2) # Increase max memory available
pavian::runApp(port=5000)
```

Desde aquí podemos subir los ficheros generados de report.txt, ver los resultados en "Results Overview" y comparar ambos reports en "Comparison". Se puede seleccionar que la comparación sea a nivel de filo, clase, orden, familia, género y especie.