Estudio del resistoma del compost: identificación de genes bacterianos resistentes a antibióticos

Aritz Durana Sarria
Master en Bioinformática y Bioestadística
Microbiología, biotecnología y biología molecular

Consultora: Paloma Pizarro Tobías
Profesor responsable de la asignatura: David Merino Arranz

05/01/2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Estudio del resistoma del compost: identificación de genes bacterianos resistentes a antibióticos*
Nombre del autor:	*Aritz Durana Sarria*
Nombre del consultor/a:	*Paloma Pizarro Tobías*
Nombre del PRA:	*David Merino Arranz*
Fecha de entrega (mm/aaaa):	01/2021
Titulación::	*Master en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Microbiología, biotecnología y biología molecular*
Idioma del trabajo:	*Castellano*
Palabras clave	*resistoma, compost, metagenoma*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

El uso indiscriminado de antibióticos durante las últimas décadas, tanto en humanos como en animales, ha promovido la aparición de gran número de genes resistentes a antibiótico (ARGs). Esto se traduce en un preocupante problema sanitario dado que cada vez resulta más complicado tratar infecciones de manera eficiente debido a la proliferación de bacterias con ARGs.

Dentro de las diferentes estrategias para controlar y gestionar la presencia de estos ARGs en el medioambiente se encuentra el compostaje del estiércol de animales tratados con antibióticos. Mediante esta técnica se consigue reducir la proliferación y el impacto de los ARGs por lo que sería interesante conocer la composición genómica del compost.

En este trabajo se estudia la distribución de los ARGs en diferentes muestras de compost usando datos de secuenciación masiva. Para ello primero se ha generado una base de datos de ARGs partiendo de datos bibliográficos. Posteriormente se han seleccionado varios metagenomas de muestras de compost y, por último, los reads presentes en estos metagenomas se han alineado con los genes de la base de datos.

Los resultados preliminares muestran que las resistencias más frecuentes de los genes (aminoglucósidos, tetraciclinas, multirresistencia,…) son similares en los diferentes casos estudiados. Son tipos de resistencias ampliamente conocidos debido a su incidencia clínica.

A pesar de que no se haya podido completar la planificación inicial, ha sido

posible identificar un alto número de ARGs en las muestras de compost, lo que es indicativo de la ubiquidad de genes resistentes en el medioambiente.

Abstract (in English, 250 words or less):

The extensive use of antibiotics during the last decades, both in humans and farm animals, has promoted the appearance of many antibiotic resistant genes (ARGs). This is a worrying health issue because it is increasingly difficult to treat bacterial infections efficiently due to the spread of bacteria with ARGs.

Among the different strategies to control and manage the presence of these ARGs in the environment, composting of manure of animals treated with antibiotics is one of them. Using this technique it is possible to reduce the proliferation and the impact of the ARGs, so it would be interesting to know the genomic composition of the compost.

In this work the distribution of ARGs in different compost samples is studied using next-generation sequencing data. First a database of ARGs was generated based on literature data. Then, several metagenomes from compost samples were selected and lastly, the reads present in these metagenomes were aligned to the genes in the database.

Preliminary results show that the most frequent resistances in genes (aminoglycoside, tetracycline, multidrug,…) are similar in all the different cases studied. They are widely known types of resistance due to their clinical incidence.

Even though it was not possible to complete the initial planning, it was possible to identify a high number of ARGs in compost samples, which is indicative of the ubiquity of the resistant genes in the environment.

Índice

Lista de figuras

Lista de tablas

1. Introducción

1.1 Contexto y justificación del Trabajo

La resistencia a antibióticos se ha convertido en uno de los mayores problemas sanitarios de los últimos años y se prevé una evolución negativa a medio plazo si no se toman medidas preventivas adecuadas [1, 2]. Este tipo de resistencia siempre ha estado presente en la naturaleza como parte del desarrollo evolutivo de las bacterias frente a los antibióticos naturales de mohos, hongos y otros microorganismos pero ha sido a partir del uso, a veces indiscriminado de antibióticos tanto en humanos como en animales, que se ha convertido en un problema de salud muy relevante [3].

Desde la segunda mitad del siglo XX los antibióticos se han utilizado de manera extensiva y han salvado millones de vidas humanas pero este uso a menudo excesivo e innecesario ha promovido la aparición y proliferación de un gran número de bacterias que cuentan con genes resistentes a antibiótico (antibiotic resistant gene, ARG) [4], lo que ha hecho que las infecciones creadas por estas bacterias resistentes sean una de las principales causas de mortandad dentro del ámbito hospitalario al no poder tratar estas infecciones de manera efectiva.

Una parte muy importante de estos ARGs proviene de diferentes entornos medioambientales como suelos, tanto urbanos como agrícolas y ganaderos, plantas de tratamiento de aguas residuales y océanos o ríos [5-10]. De hecho, el concepto de resistoma define la colección de ARGs presentes en la comunidad bacteriana de una muestra medioambiental, incluyendo tanto las bacterias patógenas como las no patógenas. Uno de los resistomas más estudiados es el del estiércol de animales tratados con antibióticos [11-14].

Es conocido que el ganado criado de manera intensiva ha sido sistemáticamente tratado con antibióticos durante las últimas décadas para aumentar su factor de crecimiento al reducir las enfermedades. Este aumento de la productividad ha posibilitado una respuesta adecuada a la demanda de productos de origen animal pero con varias contrapartidas como el bienestar animal o la aparición y proliferación de bacterias resistentes a antibióticos en el medioambiente [15].

El exceso de antibióticos excretados por estos animales puede crear una presión selectiva que favorece la aparición de bacterias con ARGs en su material genético, por lo que sería interesante desarrollar estrategias de control y gestión, como la elaboración de compost a partir del estiércol de estos

animales, para minimizar la presencia de estos genes en el medioambiente (Figura 1) [4].

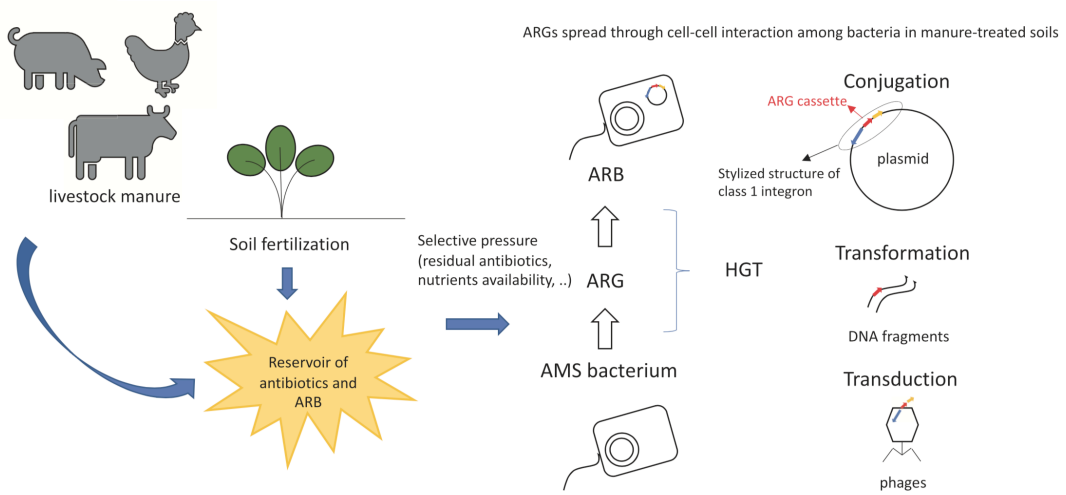


FIGURE 1 | Spread of ARGs and ARB in farm-related environments. ARB, antimicrobial resistant bacteria; ARG, antibiotic resistance gene; AMS, antimicrobial sensitive; HGT, horizontal gene transfer.

Figura 1: Proliferación de ARGs en el medioambiente

El compostaje es sencillamente la descomposición aeróbica de materia orgánica que ocurre espontáneamente bajo ciertas condiciones ambientales de humedad, temperatura, cierta relación carbono/nitrógeno de la muestra y aditivos [16-24]. Durante el proceso de compostaje varía la temperatura de los residuos en varias etapas (mesófila, termófila y de maduración) según la actividad bacteriana y estas temperaturas elevadas junto con la degradación de la materia orgánica se traducen en una reducción de la patogenicidad del compost.

Otro punto a tener en cuenta para desarrollar estrategias de control es que se conoce que la diseminación de los ARGs se da a través de la transferencia genética horizontal (horizontal gene transfer, HGT) [25-28] de elementos genéticos móviles (mobile genetic elements, MGE) como plasmidos o integrones y que condiciones de alta densidad microbiana en presencia de nutrientes y elementos antimicrobianos, como es el caso del estiércol de animales tratados con antibióticos, favorecen la ocurrencia de eventos de HGT.

Esto implica que tanto en el estiércol como parcialmente en el compost se dan las condiciones apropiadas para un aumento de bacterias resistentes y por ello es importante identificar los ARGs presentes en el compost así como su incidencia en la comunidad microbiana y poder evaluar así la efectividad del tratamiento realizado al estiércol. Recientemente diversos estudios han mostrado la presencia de ARGs en el compost y de su diseminación en el medioambiente pero sería interesante realizar un análisis desde un punto de

vista que incluya el uso de datos de secuenciación masiva [15, 29] dado que en la mayoría de los estudios publicados hasta hace pocos años la identificación de las bacterias resistentes se limitaba a las cepas cultivables en el laboratorio [30].

El desarrollo durante estos últimos años de técnicas de ultrasecuenciación ha permitido generar una elevadísima cantidad de datos de muestras medioambientales de todo tipo que se encuentran alojadas en diferentes bases de datos de secuenciación masiva [31]. Técnicas como la secuenciación shotgun o WGS (Whole Genome Sequencing), que se basa en cortar aleatoriamente el ADN en segmentos cortos que tras secuenciarlos se obtienen unas secuencias cortas (reads) que se pueden superponer de manera computacional, permite la secuenciación del genoma sin necesidad de cultivar las bacterias en el laboratorio.

Este tipo de técnicas se han utilizado para el desarrollo de la metagenómica o la secuenciación del ADN total de una muestra medioambiental, que es una técnica muy potente para estudiar la composición de la comunidad microbiana pero que no muestra sus dinámicas de expresión [32]. De modo análogo, se puede utilizar la metatranscriptómica o la secuenciación del ARN total de una muestra medioambiental para conocer la dinámica y la funcionalidad de la comunidad microbiana. Por lo tanto, el estudio de metagenomas de muestras de compost puede resultar interesante para describir con mayor exactitud la composición de la comunidad bacteriana del compost y poder conocer así mejor su relación con la presencia de bacterias resistentes a antibiótico.

En concreto, este Trabajo de Fin de Master (TFM) trata sobre la identificación de ARGs en el compost elaborado con estiércol de animales. Para ello primero se generará una base de datos de ARGs presentes en el compost y después se utilizarán datos de secuenciación masiva para localizar estos genes dentro del proceso dinámico del compostaje. Así mismo sería interesante conocer la incidencia de los ARGs en la comunidad microbiana del compost debido a que su uso como abono para cultivos facilita el acceso de estos genes al medioambiente y su presencia puede estar relacionada con la aparición de bacterias resistentes a antibióticos, cuya proliferación es un serio problema sanitario.

1.2 Objetivos del Trabajo

1.2.1 Objetivos generales

- Identificar genes bacterianos resistentes a antibiótico (ARGs) presentes en muestras de compost
- Definir la incidencia de los ARGs en la comunidad microbiana del compost
- Estudiar la relación entre los ARGs del compost y la presencia de bacterias resistentes a antibiótico en el medioambiente

1.2.2 Objetivos específicos

- Generar una base de datos de ARGs presentes en el compost
- Seleccionar datos de secuenciación masiva de muestras de compost y comparar los metagenomas obtenidos con la base de datos de ARGs
- Identificar las bacterias resistentes más abundantes en los metagenomas seleccionados

1.3 Enfoque del Trabajo

En la mayoría de los estudios publicados sobre la presencia de ARGs en compost, los experimentos se llevan a cabo tomando muestras de explotaciones agrarias y se identifican los genes presentes en la muestra en el laboratorio mediante técnicas de cultivo celular. Esa estrategia es inviable para un trabajo de este tipo por lo que se utilizarán datos de secuenciación masiva de muestras de compost para realizar el estudio [15]. El uso de estos datos junto con la información recopilada en la base de datos de ARGs permitirá identificar los genes resistentes presentes en muestras de compost y estudiar su incidencia sobre la comunidad microbiana del mismo de una manera rápida y directa.

1.4 Materiales y métodos

Este trabajo se ha realizado en un ordenador MacBook Pro de principios de 2015 con el sistema operativo macOS Mojave versión 10.14.6. A lo largo del trabajo se ha utilizado R (versión 4.0.1) [33] y RStudio (versión 1.1.463) [34] además de la línea de comandos de Mac (Terminal, versión 2.9.5). Los ficheros generados durante el trabajo se encuentran en el repositorio público de Github [35] de este TFM (https://github.com/adurana/TFM).

Creación de base de datos de ARGs: A la hora de generar la base de datos de ARGs, las secuencias de cada gen se almacenaron en ficheros individuales en formato FASTA pero para poder usarlas como una base de datos en las siguientes tareas se decidió juntar todas las secuencias en un único archivo multifasta *secuencias.fasta* mediante el siguiente código de la línea de comandos:

```
$ awk 1 *.fasta > secuencias.fasta
```

Alguna de las herramientas de alineamiento que se probaron requiere que la base de datos sea de proteínas homólogas en lugar de nucleótidos por lo que el fichero multifasta *secuencias.fasta* se tradujo usando herramientas web conocidas [36, 37] al fichero *ARGtDB.fasta*. Estos dos ficheros multifasta de secuencias de ARGs se encuentran en el repositorio https://github.com/adurana/TFM.

Selección de metagenomas: Para realizar la tarea de seleccionar y descargar metagenomas de muestras de compost se utilizó el servidor de análisis metagenómico MG-RAST [38, 39]. La selección de los metagenomas, se realizó a partir de dos búsquedas en la herramienta de búsqueda de la web del servidor [40]. En la primera búsqueda se utilizó la palabra *compost* en el campo de búsqueda general y en la ventana de búsqueda avanzada se definió el campo *sequence type* con el término *shotgun metagenome*. Esta primera búsqueda dio lugar a un listado de 147 metagenomas de tipo shotgun de muestras de compost (mgrastweb_1.txt). De manera similar se realizó una segunda búsqueda limitando la primera búsqueda al definir en la ventana de búsqueda avanzada el campo *material* con el término *compost*. Esta segunda búsqueda dio lugar a 139 metagenomas (mgrastweb_2.txt). La información de ambas búsquedas se descargó utilizando el botón *download search results*. Ambos ficheros se encuentran disponibles en el repositorio https://github.com/adurana/TFM. Sorprendentemente, al analizar ambos listados (ver apartado 2.3), se observó que había proyectos (código mgp) que sólo estaban en el segundo listado más restrictivo así como otros que sólo estaban en el primer listado, por lo que se decidió unir ambos listados obteniendo un total de 160 metagenomas (códigos mgm) de 20 proyectos diferentes. Este listado final de metagenomas se puede encontrar en el Apartado 6 de esta memoria (Tabla A1).

Tras seleccionar los metagenomas de interés, el siguiente paso fue la descarga de los ficheros de datos de secuenciación masiva. El servidor MG-RAST tiene varios tipos de ficheros disponibles para la descarga, todos ellos pasos intermedios del pipeline de análisis implementado en el servidor, a los que se accede a través del botón de descarga (una nube con una flecha hacia abajo) de cada metagenoma seleccionado en la pantalla de búsqueda. En este caso,

se utilizaron los ficheros de tipo 350, que son el output del paso 9 del análisis (*Identify putative protein coding features (genecalling)*). Es un tipo de fichero en formato FASTA de las regiones de codificación predichas por el servidor, esto es, de los reads correspondientes a los genes que se han encontrado en el metagenoma. Para la descarga de estos ficheros se utilizó el código de la línea de comandos que se muestra a continuación donde el asterisco representa el código de metagenoma de cada uno de los metagenomas seleccionados. En total se descargaron unos 170 GB de información en 160 ficheros .faa.

```
$ curl "https://api.mg-rast.org/download/mgm*.3?file=350.1" > /MGRAST/mgm*.3.faa
```

Identificación de ARGs: En este apartado se probaron varias herramientas bioinformáticas para la identificación de ARGs en los metagenomas de las muestras de compost. La primera herramienta que se probó fue MetaHMM [41, 42]. Esta herramienta web esta diseñada para encontrar genes en metagenomas usando modelos ocultos de Markov a partir de alineamientos realizados con Clustal Omega [43]. En la descripción se indica que el input debe ser un listado de códigos Uniprot [44] de proteínas homólogas y que, además de ofrecer varios metagenomas disponibles, es posible realizar la búsqueda con metagenomas personalizados. Para acceder a estos metagenomas, la herramienta requiere de una conexión directa *via* http o ftp a los mismos. Primero se probó a establecer una conexión ftp [45-47] y se comprobó que era posible acceder a la carpeta compartida de manera local pero no fue posible acceder de manera remota con un gestor de ficheros ftp como Filezilla [48] a pesar de haber modificado los permisos y accesos a ficheros en las preferencias del sistema del ordenador. A continuación se intentó crear una conexión http primero instalando y configurando el servidor http Apache [49-51] y definiendo el localhost [52-54] y luego modificando el firewall del sistema y el puerto de acceso a través de un port forwarding del router [55-57]. Por desgracia, del mismo modo que con la conexión ftp, no fue posible la conexión remota a la carpeta compartida. Por lo tanto, tras acumular muchos días de retraso con respecto a la planificación inicial y sin ningún avance claro, se decidió dejar de lado esta herramienta.

La segunda herramienta que se probó fue Metalign [58, 59], que es un método diseñado para estimar la composición taxonómica de metagenomas basado en alineamientos eficientes de metagenomas WGS con diferentes bases de datos. Esta herramienta utiliza el motor de alineamiento Minimap2 [60]. Se siguió la documentación [61] para proceder a la instalación a través de Bioconda [62], y tras varias horas se instaló el programa y la extensa base de datos (más de 300 GB) correctamente. Al intentar replicar el *test example* instalado por defecto, el programa dio un error al no poder crear ficheros temporales en las carpetas /var/folders y a pesar de que se modificaron los permisos de escritura

de estas carpetas no se consiguió realizar dicho test de ejemplo [63-66]. Además, revisando la documentación más a fondo no se observó una manera clara y rápida de utilizar una base de datos personalizada [67-68] por lo que se decidió dejar esta herramienta de lado.

La tercera y última herramienta bioinformática que se probó fue DIAMOND [69-71]. Esta herramienta es básicamente un alineador de secuencias para proteínas diseñado y optimizado para analizar datos de secuenciación masiva mucho más rápido y de manera más eficiente que BLAST [72]. Se siguió la documentación para completar la instalación a través de Bioconda y se consiguió definir una base de datos personalizada de manera rápida y sencilla [73]. Hay que mencionar que esta herramienta requiere una base de datos de proteínas homólogas por lo que se utilizó el fichero *ARGtDB.fasta* para definir la base de datos personalizada. Por último, se realizaron los alineamientos de los 160 metagenomas con la base de datos en un margen de tiempo razonable obteniendo los ficheros .tsv con la información generada por defecto por la herramienta. El código de línea de comandos utilizado para llevar a cabo este análisis se muestra a continuación [73]. En este caso el asterisco representa el código de metagenoma de cada uno de los metagenomas seleccionados:

```
#Definición de la base de datos personalizada
$ diamond makedb --in ARGtDB.fasta -d ARGtDB


#Alineamientos de los ficheros .faa con la base de datos
$ diamond blastp -d ARGtDB -q mgm*.faa -o mgm*.tsv --sensitive
```

1.5 Planificación del Trabajo

1.5.1 Tareas

Las principales tareas a realizar para llevar a cabo el estudio son las siguientes:

- T1: Realizar una búsqueda bibliográfica de ARGs en artículos científicos y en bases de datos sobre resistencia a antibióticos
- T2: Generar una base de datos de ARGs presentes en compost
- T3: Seleccionar varios metagenomas de muestras de compost en bases de datos de secuenciación masiva
- T4: Identificar los ARGs presentes en esos metagenomas
- T5: Observar la incidencia de los ARGs en la distribución de la comunidad microbiana de las muestras de compost

- T6: Generar un listado de las bacterias resistentes más abundantes en los metagenomas seleccionados

1.5.2 Calendario

En la Tabla 1 se detalla el tiempo dedicado a las tareas definidas en el apartado 1.5.1 así como a las tareas de redacción de la memoria (T7), preparación de la presentación (T8) y defensa pública del TFM (T9). Alternativamente se ha plasmado esta distribución temporal en un diagrama de Gantt (Figura 2) generada con la herramienta libre Gantt Project [74].

Tabla 1: Cronograma del TFM

Tarea	Descripción	Duración (días)
T1	Búsqueda bibliográfica	12
T2	Creación base de datos	4
T3	Selección metagenomas	4
T4	Identificación ARGs	10
T5	Incidencia ARGs en bacterias	11
T6	Listado bacterias resistentes	3
T7	Redacción memoria	15
T8	Elaboración presentación	4
T9	Defensa pública	6

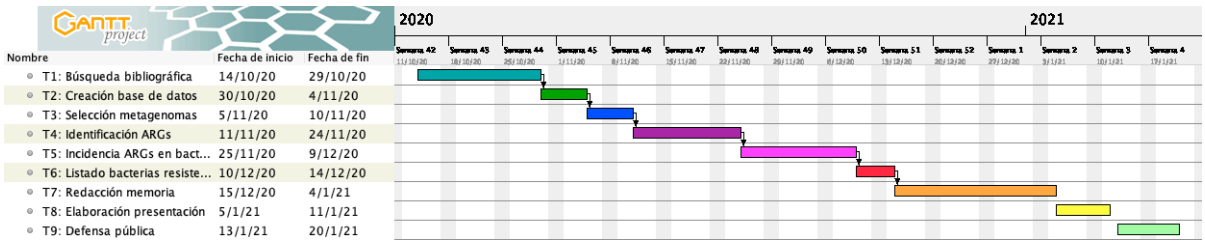


Figura 2: Diagrama de Gantt del TFM

1.5.3 Hitos

Los hitos necesarios para el desarrollo del TFM son los siguientes:

- Generar la base de datos de ARGs
- Seleccionar los metagenomas de las bases de datos de secuenciación masiva
- Identificar los ARGs presentes en esos metagenomas
- Identificar las bacterias resistentes más comunes en las muestras de compost

1.5.4 Análisis de riesgos

Existen varios factores que pueden repercutir negativamente en el desarrollo del plan de trabajo y por lo tanto en la finalización del proyecto. Además de factores generales como pueden ser la falta de tiempo para completar las tareas (debido a una mala estimación de la disponibilidad o la mala organización de las tareas) y la propia extensión del proyecto (debido a una mala planificación de los objetivos y tareas dentro de la extensión descrita en el Plan Docente), para este TFM se han identificado varios factores de riesgos específicos:

- Búsqueda bibliográfica incompleta: Es posible que la búsqueda de ARGs no sea la correcta en términos de extensión por lo que la base de datos generada no sería todo lo completa que debería ser. Si se detecta este problema en etapas tempranas del plan de trabajo se puede volver a realizar una búsqueda más exhaustiva para generar una base de datos de ARGs más completa
- Mala selección de metagenomas: Es posible que los metagenomas seleccionados en las bases de datos de secuenciación masiva no sean los más idóneos para el estudio a realizar por diversas razones (escasez de ARGs en los datos, datos incompletos, origen de datos no apropiado,…). Para evitar este factor de riesgo es importante hacer una buena selección inicial para no tener que repetir análisis posteriores
- Incorrecta identificación de ARGs: Si alguno de los dos factores de riesgo descritos anteriormente ocurren y no se detectan, la identificación de los ARGs será incompleta o incorrecta. Por lo tanto, para minimizar este riesgo es importante generar una buena base de datos y seleccionar los metagenomas apropiados. Adicionalmente, un método de comparación poco apropiado también puede derivar en una incorrecta identificación de ARGs

1.6 Breve sumario de productos obtenidos

Los productos obtenidos de este trabajo son los siguientes:

- La propia memoria del trabajo que incluye entre otros apartados la introducción al tema de estudio, los objetivos a lograr, los resultados obtenidos de manera detallada y las conclusiones extraídas de los mismos así como la bibliografía consultada para la realización del TFM.
- Una base de datos de ARGs con información sobre el tipo de resistencia del gen o el organismo del que se ha aislado, además de las referencias de diferentes bases de datos para caracterizar completamente cada gen.

- Un listado de metagenomas de muestras de compost del servidor web de datos de secuenciación masiva MG-RAST [38] cuyas secuencias se compararán con las de los genes de la base de datos para detectar los genes resistentes presentes en las muestras.
- Los ficheros de alineamiento entre la base de datos y los metagenomas utilizando la herramienta DIAMOND [69] que muestran el porcentaje de identidad entre los genes de la base de datos y las secuencias de los metagenomas junto con otros parámetros de calidad.
- El código en lenguaje R [33] creado para el tratamiento de estos ficheros de alineamiento.

1.7 Breve descripción de los otros capítulos de la memoria

El proyecto se ha estructurado de la manera que se muestra a continuación:

- Capítulo 1 Introducción: Contextualización del tema sobre el que se desarrollará el TFM y justificación del trabajo, definición de los objetivos que se quieren lograr con el TFM, descripción detallada de la metodología utilizada en las diferentes tareas del TFM y la planificación del trabajo incluyendo la descripción de las tareas e hitos a realizar dentro de un marco temporal definido.
- Capítulo 2 Resultados: Presentación detallada de los resultados obtenidos durante el estudio y su discusión crítica.
- Capítulo 3 Conclusiones: Recapitulación de las conclusiones derivadas de los resultados más relevantes del TFM y ajuste de los objetivos y tareas iniciales.
- Capítulo 4 Glosario: Resumen de los acrónimos y de los términos más utilizados durante el TFM.
- Capítulo 5 Bibliografía: Resumen de los materiales y fuentes consultadas para la realización del TFM.
- Capítulo 6 Anexos: Recopilación de información complementaria útil para la comprensión de la memoria.

2. Resultados

2.1 Búsqueda bibliográfica (T1)

Según lo descrito en la primera tarea de la planificación del trabajo (Apartado 1.5), se llevó a cabo una búsqueda bibliográfica con la intención de generar una base de datos de genes resistentes a antibióticos (ARGs) presentes en muestras de compost.

Para comenzar esta búsqueda, la consultora de mi TFM me proporcionó dos artículos [15, 29] muy interesantes que me sirvieron para familiarizarme con el tema así como de punto de partida para poder identificar correctamente los aspectos relevantes del tema del trabajo.

La primera parte de la búsqueda se hizo principalmente utilizando la Biblioteca de la UOC [75] usando palabras claves como *compost*, *resistome*, *antibiotic resistance* o *metagenome* entre otras. Se limitó esta búsqueda a los artículos publicados durante los años 2015-2020 para seleccionar así los trabajos más recientes.

Tras hacer una selección de estos primeros artículos, se amplió la búsqueda revisando varios de los artículos referenciados en esta primera selección de artículos. Asimismo, también se buscaron artículos recientes que referencian los artículos más interesantes de la primera selección. Todo este trabajo de búsqueda se realizó a través de la Biblioteca de la UOC y de las webs de las diferentes editoriales como Elsevier [76], Wiley [77], Springer [78], PNAS [79] o Nature [80] entre otras.

Tras revisar más de cien artículos, se consideró que se había reunido suficiente información como para poder generar una base de datos de ARGs presentes en muestras de compost actualizada y completa utilizando para ello la información obtenida de algunos de esos artículos [1, 15, 21, 29, 81-93].

2.2 Creación de base de datos (T2)

La siguiente tarea descrita en la planificación del trabajo fue la creación de una base de datos de ARGs que servirá para poder identificar los genes resistentes presentes en los metagenomas de muestras de compost. De los artículos mencionados anteriormente se obtuvo un listado con 360 nombres de genes resistentes, pero para poder generar una base de datos completa se decidió añadir mucha más información relativa a dichos genes. Se han utilizado diversas bases de datos y recursos web para completar la información de la

base de datos de ARGs, pero principalmente han sido dos las utilizadas: CARD (Comprehensive Antibiotic Resistance Database) [94, 95] y el catálogo de genes patógenos de NCBI (National Center for Biotechnology Information) [96]. De estas dos bases de datos se ha obtenido prácticamente toda la información necesaria sobre tipo de resistencia, secuencia, organismo, etc. de la mayoría de los genes resistentes. Los pocos genes sobre los que no se ha podido encontrar información en estas dos bases de datos se han buscado en UniProt (Universal Protein Resource) [44] y en ENA (European Nucleotide Archive) [97]. Todas las secuencias de los genes de la base de datos se encuentran recogidas en un fichero en formato FASTA llamado *secuencias.fasta* que se puede encontrar en el repositorio https://github.com/adurana/TFM.

En la figura 3 se muestran los treinta primeros genes de la base de datos en la que se observan los diferentes campos que se han incluido. Hay que mencionar que la información de estos campos se ha conseguido completar para la mayoría de los genes. El fichero *BD_ARGs.csv* con la base de datos completa se puede encontrar en https://github.com/adurana/TFM.

	A	B	C	D	E	F	G	H	I	J	K
1	**Nombre del gen**	**Nombres alternativos**	**Clase**	**Antibiotico**	**Mecanismo**	**Database**	**Organismo**	**RefSeq Sequence**	**Genbank Accession**	**CARD Accession**	**Enlace**
2	aac(3)-II		Aminoglycoside	Gentamicin	Inactivation	NCBI	Sphingopyxis alaskensis RB2256	NG_048573.1	CP000356.1		https://www
3	aac(3)-IV	aacC4, AAC(3)-IVa	Aminoglycoside	Apramycin, Gentamicin, Tobramycin	Inactivation	CARD	Escherichia coli	NG_047253.1	DQ241380.1	ARO:3002539	https://www
4	aac(6')-31		Aminoglycoside	Aminoglycoside	Inactivation	CARD	Pseudomonas putida		AM283489.1	ARO:3002585	https://www
5	aac(6')-Ia	aacA1, aac(6')-I	Aminoglycoside	Aminoglycoside	Inactivation	CARD	Plasmid R		M18967.1	ARO:3002545	https://www
6	aac(6')-Ib	aacA4, AAC(6')-4	Aminoglycoside	Aminoglycoside	Inactivation	CARD	Klebsiella pneumoniae		JQ808129.1	ARO:3002546	https://www
7	aac(6')-Ib-cr		Aminoglycoside,	Aminoglycoside, Fluoroquinolone	Inactivation	CARD	Escherichia coli		DQ303918.1	ARO:3002547	https://www
8	aac(6')-Ie-aph(2'')-Ia	aac(6')-bifunctional aa	Aminoglycoside	Amikacin, Kanamycin, Tobramycin	Inactivation	CARD	Staphylococcus aureus	NG_047055.1	GU565967.1	ARO:3002597	https://www
9	aac(6')-Ii		Aminoglycoside	Aminoglycoside	Inactivation	CARD	Enterococcus faecium	NG_047298.1	L12710.1	ARO:3002556	https://www
10	aac(6')-IIa		Aminoglycoside	Aminoglycoside	Inactivation	CARD	Salmonella enterica subsp. enterica serovar Typhi		AY123251.1	ARO:3002594	https://www
11	aac(6')-Iid		Aminoglycoside	Aminoglycoside	Inactivation	CARD	Enterococcus hirae	NG_047299.1	AJ584700.2	ARO:3002589	https://www
12	aac(3)-Ia	aacC-A1, aacC1	Aminoglycoside	Gentamicin	Inactivation	CARD	Pseudomonas aeruginosa	NG_056001.1	U12338.3	ARO:3002528	https://www
13	aac(3)-Ic	aacC2	Aminoglycoside	Gentamicin	Inactivation	CARD	Escherichia coli	NG_047250.1	X54723.1	ARO:3002535	https://www
14	aad(6)	ant(6)-Ia, aac(6)	Aminoglycoside	Streptomycin	Inactivation	CARD	Streptococcus oralis		AY712687.1	ARO:3002628	https://www
15	ANT(9)-Ia	aad(9), spc, spw, aad9	Aminoglycoside	Spectinomycin	Inactivation	CARD	Staphylococcus aureus	NG_047397.1	X02588.1	ARO:3002630	https://www
16	aadA	ANT(3'')-Ia, aadA1-pm	Aminoglycoside	Streptomycin	Inactivation	CARD	Escherichia coli		AF550679.1	ARO:3002601	https://www
17	aadA2		Aminoglycoside	Streptomycin, Spectinomycin	Inactivation	CARD	Klebsiella pneumoniae		AF156486.1	ARO:3002602	https://www
18	aadA24		Aminoglycoside	Streptomycin, Spectinomycin	Inactivation	CARD	Salmonella enterica subsp. enterica serovar Newport		DQ677333.1	ARO:3002621	https://www
19	aadA3		Aminoglycoside	Streptomycin	Inactivation	CARD	Plasmid NR79	NG_047353.1	AF047479.2	ARO:3002603	https://www
20	aadA4		Aminoglycoside	Streptomycin	Inactivation	CARD	Acinetobacter baumannii	NG_047356.1	AY138986.1	ARO:3002604	https://www
21	aadA5		Aminoglycoside	Streptomycin, Spectinomycin	Inactivation	CARD	Escherichia coli	NG_047357.1	AF137361.1	ARO:3002605	https://www
22	aadA6		Aminoglycoside	Streptomycin	Inactivation	CARD	Pseudomonas aeruginosa		AM087411.1	ARO:3002606	https://www
23	aadA9	ANT(9)-Ia	Aminoglycoside	Streptomycin	Inactivation	CARD	Corynebacterium sp. L2-79-05		DQ390458.1	ARO:3002609	https://www
24	ANT(2'')-Ia	aadB	Aminoglycoside	Gentamicin, Kanamycin, Tobramycin	Inactivation	CARD	Pseudomonas aeruginosa		AF078527.1	ARO:3000230	https://www
25	ANT(4')-Ia	aadD, aadD2	Aminoglycoside	Amikacin, Kanamycin, Tobramycin	Inactivation	CARD	Bacillus clausii	NG_047392.1	EF540343.1	ARO:3002623	https://www
26	ANT(6)-Ia	ant6, aadE	Aminoglycoside	Streptomycin	Inactivation	CARD	Exiguobacterium sp. S3-2		KF648874.1	ARO:3002626	https://www
27	abeM		Fluoroquinolone,	Multidrug	Efflux	CARD	Acinetobacter baumannii		AB204810.2	ARO:3000753	https://www
28	abeS		Macrolide, Amin	Multidrug	Efflux	CARD	Acinetobacter baumannii AB307-0294		CP001172.1	ARO:3000768	https://www
29	acrA		Tetracycline, Pen	Multidrug	Efflux	CARD	Klebsiella pneumoniae		AJ318073.1	ARO:3000207	https://www
30	acrB	ECK0456, JW0451	Tetracycline, Pen	Multidrug	Efflux	CARD	Escherichia coli str. K-12 substr. MG1655		U00096.3	ARO:3000216	https://www
31	acrD		Aminoglycoside	Multidrug	Efflux	CARD	Escherichia coli str. K-12 substr. W3110		AP009048.1	ARO:3000491	https://www

Figura 3: Primeros genes de la base de datos BD_ARGs

Los campos con las que se completó la base de datos son los siguientes:

- Nombre del gen: Nombre más común del gen
- Nombres alternativos: Otros nombres comunes del gen
- Clase: Clase de antibiótico al que es resistente el gen
- Antibiótico: Antibiótico concreto al que es resistente el gen
- Mecanismo: Mecanismo de resistencia a antibiótico del gen
- Database: Base de datos de la que se ha obtenido mayoritariamente la información para completar esta base de datos (CARD, NCBI o UNIPROT/ENA)
- Organismo: Organismo del que se ha aislado la secuencia del gen
- RefSeq Sequence Reference: Número de referencia de RefSeq para el gen

- Genbank Accession Number: Número de referencia de Genbank para el gen
- CARD Accession: Número de referencia de CARD para el gen
- Enlace: Enlace a la web del NCBI donde se encuentra la secuencia del gen

Hay que mencionar que en la base de datos se han incluido tres números de referencia diferentes que son útiles a la hora de caracterizar el gen: Genbank [98], RefSeq [99] y CARD [95]. Genbank es una base de datos anotada de todas las secuencias de nucleótidos y de sus traducciones a proteínas mantenida por el NCBI y de acceso público. RefSeq es una base de datos muy similar también mantenida por el NCBI pero a diferencia de Genbank no es redundante, esto es, solo proporciona una secuencia para cada biomolécula, lo que permite realizar estudios comparativos. Por último CARD es una base de datos rigurosamente revisada y específica para genes resistentes a antibióticos que utiliza un sistema propio de ontología (Antibiotic Resistance Ontology, ARO) para la organización de las secuencias.

Al analizar la base de datos *BD_ARGs.csv* en R [33], se puede obtener información principalmente de tres campos (clase, antibiótico y organismo) que pueden dar una idea general sobre el tipo de genes que contiene la base de datos. Primero se carga la base de datos en R y a continuación se generan el gráfico circular (Figura 4) y la tabla (Tabla 2) correspondientes a la clase de antibiótico de la base de datos.

```
#Cargar el fichero de la base de datos de ARGs (BD_ARGs.csv) en R

BD_ARGs <- read.csv2(file = "~/TFM/BD_ARGs.csv", header = T)
```

```
#Piechart con las clases de antibiótico de la base de datos
labels_cl <- rownames(head(sort(table(BD_ARGs$Clase), decreasing = T),
6))

pie(sort(table(BD_ARGs$Clase), decreasing = T),
    init.angle = 90, clockwise = T, labels = labels_cl)
```

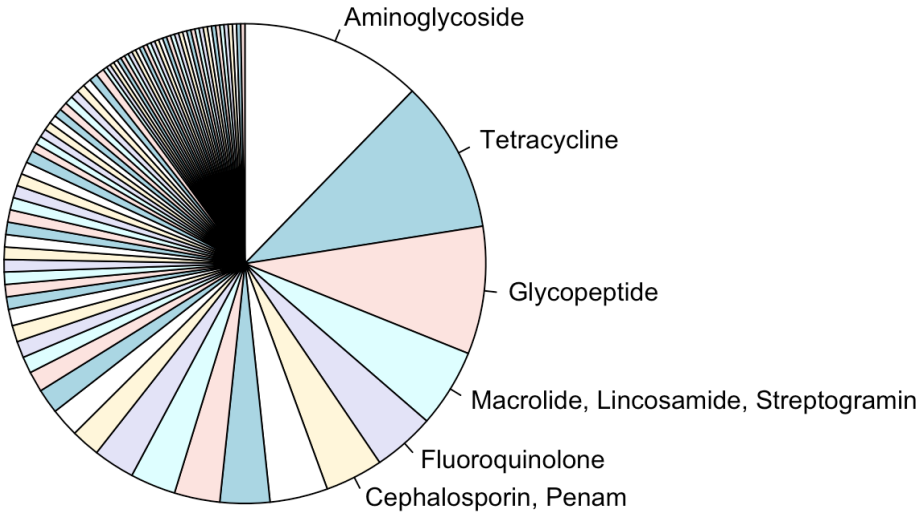


Figura 4: Gráfico de las clases de antibiótico más comunes en la base de datos

Tabla 2: Tabla de las clases de antibiótico más comunes en la base de datos

```
#Tabla con las primeras clases de antibiótico de la base de datos
head(as.data.frame(sort(table(BD_ARGs$Clase), decreasing = T)), 10)
```

```
##                                    Var1 Freq
## 1                        Aminoglycoside   44
## 2                          Tetracycline   37
## 3                          Glycopeptide   31
## 4  Macrolide, Lincosamide, Streptogramin   19
## 5                        Fluoroquinolone   15
## 6                  Cephalosporin, Penam   14
## 7                               Peptide   14
## 8                                 Penam   12
## 9                     Diaminopyrimidine   11
## 10                              Phenicol   11
```

Como se observa tanto en el gráfico como en la tabla, los genes con resistencia a los grupos de antibióticos aminoglucósidos [100], tetraciclinas [101] y glucopéptidos [102] son los más frecuentes en la base de datos. Las tres clases son relevantes desde el punto de vista clínico a la hora de tratar infecciones bacterianas lo que refuerza el concepto de que la resistencia a antibióticos es un problema sanitario muy grave.

Con respecto al antibiótico concreto al que presentan resistencia los genes, tanto en la figura 5 como en la tabla 3 se observa que genes con multirresistencia [103], con resistencia a antibióticos beta-lactámicos [104] y con resistencia a tetraciclina [101] son los más frecuentes. En este caso el resultado también es preocupante debido a que los genes con múltiples resistencias sean los más frecuentes es indicativo de lo extendida que se encuentra la resistencia a antibióticos en el medioambiente.

```r
#Piechart con los antibióticos de la base de datos
labels_ab <- rownames(head(sort(table(BD_ARGs$Antibiotico), decreasing
= T), 8))

pie(sort(table(BD_ARGs$Antibiotico), decreasing = T),
    init.angle = 90, clockwise = T, labels = labels_ab)
```

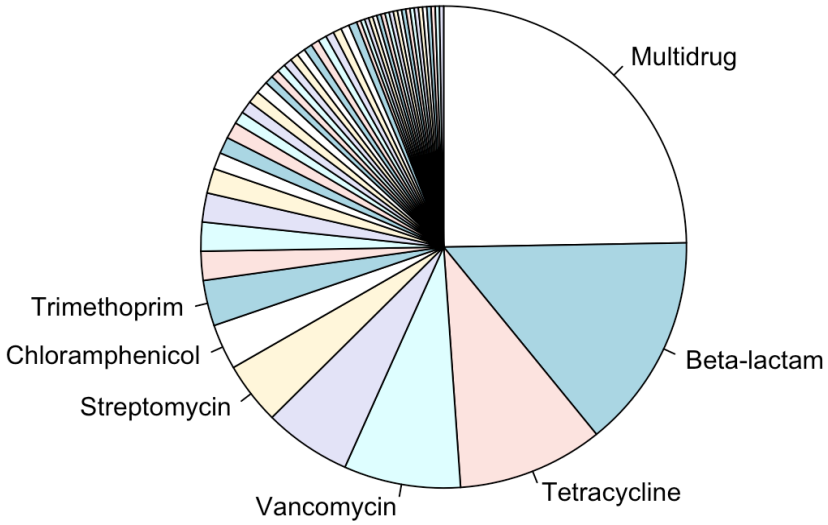


Figura 5: Gráfico de los antibióticos más frecuentes en la base de datos

Tabla 3: Tabla de los antibióticos más frecuentes en la base de datos

```r
#Tabla con los primeros antibióticos de la base de datos
head(as.data.frame(sort(table(BD_ARGs$Antibiotico), decreasing = T)),
10)
```

```
##                Var1 Freq
## 1         Multidrug   89
## 2       Beta-lactam   52
## 3     Tetracycline    35
## 4       Vancomycin    28
## 5                     21
## 6     Streptomycin    15
## 7  Chloramphenicol    11
## 8     Trimethoprim    11
## 9     Erythromycin     7
## 10 Fluoroquinolone     7
```

Por último, en cuanto al organismo del que se han aislado los genes resistentes, se observa que diferentes cepas de *Escherichia coli* [105], *Pseudomonas aeruginosa* [106] y dos bacterias del género *Enterococcus* [107, 108] son las más frecuentes (Figura 6 y Tabla 4). Esto no es ninguna sorpresa porque estas bacterias son muy conocidas por haber desarrollado resistencias a múltiples antibióticos.

```r
#Piechart con los organismos de la base de datos
labels_org <- rownames(head(sort(table(BD_ARGs$Organismo), decreasing
= T), 8))

pie(sort(table(BD_ARGs$Organismo), decreasing = T),
    init.angle = 90, clockwise = T, labels = labels_org)
```

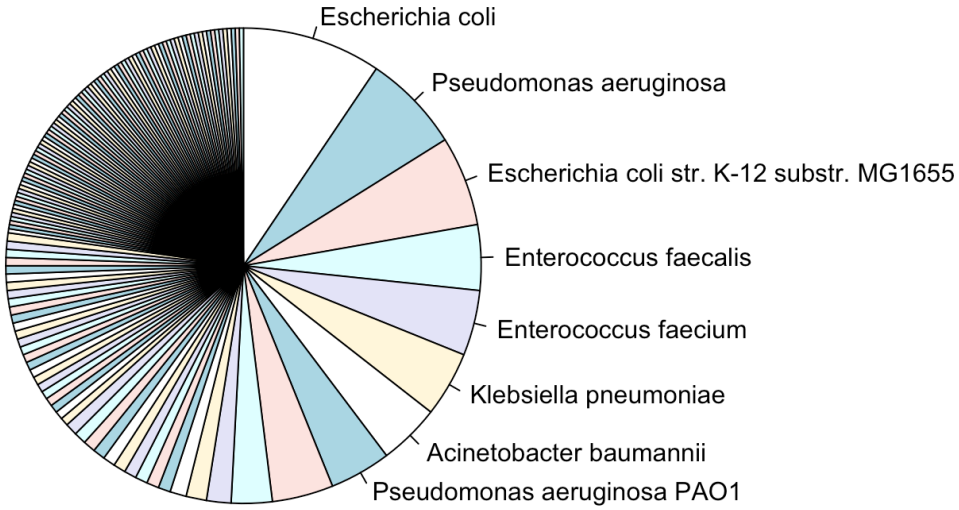


Figura 6: Gráfico de los organismos más frecuentes en la base de datos

Tabla 4: Tabla de los organismos más frecuentes en la base de datos

```r
#Tabla con los primeros organismos de la base de datos
head(as.data.frame(sort(table(BD_ARGs$Organismo), decreasing = T)),
10)
```

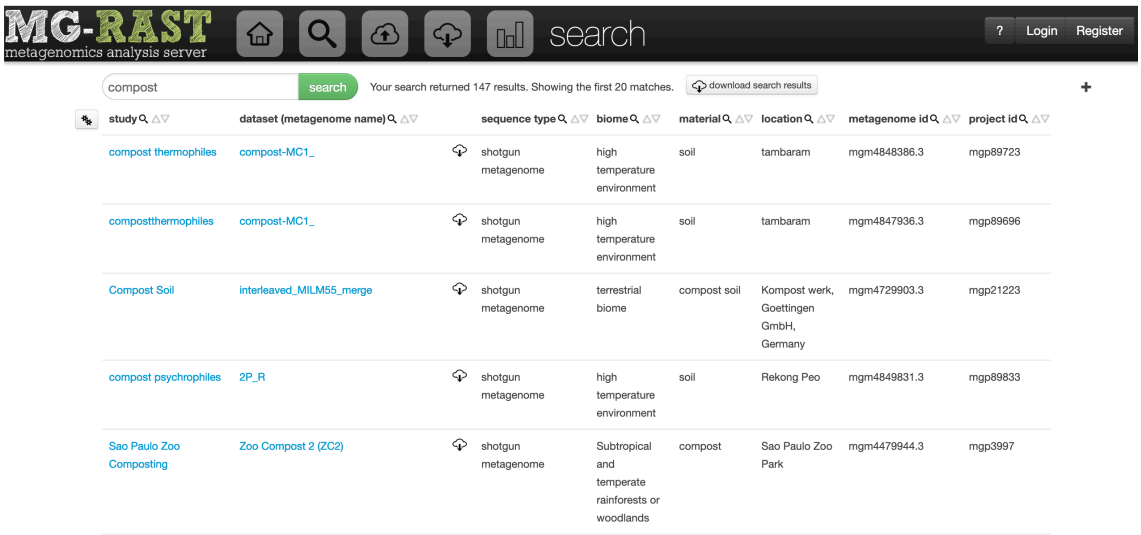
```
##                                          Var1 Freq
## 1                           Escherichia coli   34
## 2                     Pseudomonas aeruginosa   24
## 3  Escherichia coli str. K-12 substr. MG1655   22
## 4                      Enterococcus faecalis   16
## 5                       Enterococcus faecium   16
## 6                      Klebsiella pneumoniae   16
## 7                    Acinetobacter baumannii   15
## 8                Pseudomonas aeruginosa PAO1   15
## 9                      Staphylococcus aureus   15
## 10  Escherichia coli str. K-12 substr. W3110   10
```

Este análisis de la base de datos parece confirmar que la presencia de genes resistentes a antibióticos con relevancia clínica está muy extendido dentro de las muestras de compost. Por lo tanto, se decidió que esta base de datos era lo suficientemente completa y exhaustiva como para poder utilizarla a lo largo del TFM. La generación de esta base de datos permitió completar el primer hito (Apartado 1.5.3) descrito en el plan de trabajo.

2.3 Selección de metagenomas (T3)

Tras la creación de la base de datos de ARGs, la siguiente tarea descrita en la planificación del trabajo fue la selección de metagenomas de datos de secuenciación masiva de muestras de compost. En la actualidad existen multitud de alternativas [31] para la obtención de estos datos como IMG/M [109], METAREP [110], MetaStorm [111] o MetaWRAP [112] pero para este trabajo se eligió utilizar los datos del servidor de análisis metagenómico MG-RAST [38, 39]. Este servidor, además de proporcionar un potente pipeline para el análisis de datos crudos de secuenciación masiva, también permite buscar y descargar los ficheros intermedios generados en cada paso de los análisis depositados en el mismo. El servidor es muy completo y permite realizar diversas opciones de búsqueda, análisis o tratamiento de datos. Existe bastante documentación sobre el uso del servidor pero a veces ésta no es demasiado clara o detallada [113-115]. De cualquier manera no es complicado familiarizarse con el funcionamiento del servidor, al menos para realizar búsquedas y descargar ficheros como es el caso.

El servidor consta de diferentes proyectos (project id, mgp) en los que se incluye toda la información de cada estudio concreto subido al servidor. En la figura 7 se muestran los primeros resultados de la primera búsqueda descrita en el apartado 1.4.



study	dataset (metagenome name)		sequence type	biome	material	location	metagenome id	project id
compost thermophiles	compost-MC1_		shotgun metagenome	high temperature environment	soil	tambaram	mgm4848386.3	mgp89723
compostthermophiles	compost-MC1_		shotgun metagenome	high temperature environment	soil	tambaram	mgm4847936.3	mgp89696
Compost Soil	interleaved_MILM55_merge		shotgun metagenome	terrestrial biome	compost soil	Kompost werk, Goettingen GmbH, Germany	mgm4729903.3	mgp21223
compost psychrophiles	2P_R		shotgun metagenome	high temperature environment	soil	Rekong Peo	mgm4849831.3	mgp89833
Sao Paulo Zoo Composting	Zoo Compost 2 (ZC2)		shotgun metagenome	Subtropical and temperate rainforests or woodlands	compost	Sao Paulo Zoo Park	mgm4479944.3	mgp3997

Figura 7: Primeros resultados de la primera búsqueda realizada en MG-RAST

Dentro de cada proyecto, además de mucha información sobre el diseño de los mismos, se pueden encontrar los datos concretos de cada experimento o muestra y a cada uno se le asigna un código (metagenome id, mgm). Dentro de cada metagenoma existen varios tipos de ficheros dado que cada uno representa el output de cada uno de los pasos del pipeline de tratamiento de los datos crudos. En concreto el tipo de fichero de interés para este estudio es

el del tipo 350, que es un fichero en formato FASTA de las regiones de codificación predichas por el servidor [113], esto es, los reads correspondientes a los genes que se han encontrado en el metagenoma. Estos son los datos que se utilizarán para comparar con la base de datos de ARGs.

Para este trabajo se realizó una primera búsqueda según lo descrito en el apartado 1.4 de esta memoria usando palabras claves como *compost* y *shotgun metagenome* de la que se obtuvieron 147 metagenomas.

```
#Cargar el fichero de la búsqueda 1 en MG-RAST
mgrastweb_1 <- read.delim("~/TFM/mgrastweb_1.txt", header = T)

#Tamaño de la búsqueda 1
length(mgrastweb_1$project_id)

## [1] 147
```

En la tabla 5 se observa el número de metagenomas por proyecto que cumplen los requisitos de esta búsqueda.

Tabla 5: Metagenomas en cada proyecto de la primera búsqueda

```
#Número de metagenomas en cada proyecto en la búsqueda 1
data.frame(sort(table(mgrastweb_1$project_id), decreasing = T))
```

	Var1	Freq
1	mgp5435	104
2	mgp80207	12
3	mgp12953	6
4	mgp6843	6
5	mgp12863	2
6	mgp12864	2
7	mgp18601	2
8	mgp3997	2
9	mgp80101	2
10	mgp89870	2
11	mgp12908	1
12	mgp21223	1
13	mgp3456	1
14	mgp45	1
15	mgp89696	1
16	mgp89723	1
17	mgp89833	1

Como se detalla en el apartado 1.4, se realizó una segunda búsqueda limitando los resultados de la primera al definir el material del que se obtuvieron los datos de secuenciación masiva.

```
#Cargar el fichero de la búsqueda 2 en MG-RAST
mgrastweb_2 <- read.delim("~/TFM/mgrastweb_2.txt", header = T)

#Tamaño de la búsqueda 2
length(mgrastweb_2$project_id)

## [1] 139
```

En esta segunda búsqueda se obtuvieron 139 metagenomas (Tabla 6) pero, como se puede observar al comparar las tablas 5 y 6 existen diferencias entre los listados por lo que se decidió unir los proyectos encontrados en ambas búsquedas (Tabla 7).

Tabla 6: Metagenomas en cada proyecto de la segunda búsqueda

```
#Número de metagenomas en cada proyecto en la búsqueda 2
data.frame(sort(table(mgrastweb_2$project_id), decreasing = T))
```

	Var1	Freq
1	mgp5435	104
2	mgp80207	12
3	mgp6843	6
4	mgp12931	3
5	mgp12932	3
6	mgp12863	2
7	mgp12864	2
8	mgp3997	2
9	mgp80101	2
10	mgp12866	1
11	mgp12908	1
12	mgp21223	1

Tabla 7: Proyectos resultantes de la unión de las dos búsquedas

```
#Proyectos resultantes de la unión de las dos búsquedas
union(mgrastweb_1$project_id, mgrastweb_2$project_id)

##  [1] "mgp89870" "mgp89833" "mgp89723" "mgp89696" "mgp80207" "mgp80101"
##  [7] "mgp6843"  "mgp5435"  "mgp45"    "mgp3997"  "mgp3456"  "mgp21223"
## [13] "mgp18601" "mgp12953" "mgp12908" "mgp12864" "mgp12863" "mgp12932"
## [19] "mgp12931" "mgp12866"
```

Tras esta unión se obtuvieron 160 metagenomas diferentes de muestras de compost de un total de 20 proyectos con los que se generó un listado con los códigos de los proyectos y los metagenomas seleccionados (Tabla A1, Apartado 6). A continuación se descargaron un total de 170 GB de ficheros de tipo 350 en formato FASTA que incluyen la información de los reads de cada uno de los metagenomas del listado (detalles en el apartado 1.4). La selección de estos metagenomas permite completar el segundo hito (Apartado 1.5.3) descrito en el plan de trabajo.

Alternativamente, además de la opción de búsqueda dentro de la web, es posible utilizar la interfaz API que ofrece el propio servidor [116]. El uso de esta interfaz esta muy bien documentado y es más potente que hacerlo *via* web dado que ofrece muchas más posibilidades de búsqueda y análisis. Su uso es sencillo ya que solo hay que introducir la solicitud de la acción que se quiera realizar en el navegador. Por ejemplo, al realizar una búsqueda a través de la API[1] en la que se define el material como compost, el dataset como de acceso público y el tipo de secuencia como WGS, se obtienen 139 metagenomas diferentes presentes en la base de datos de MG-RAST. Hay que mencionar que los datos obtenidos de esta manera son idénticos a los obtenidos a través de la búsqueda en la web.

2.4 Identificación de ARGs (T4)

Tras la creación de la base de datos de ARGs y la selección de metagenomas de muestras de compost en el servidor MG-RAST, la siguiente tarea programada en la planificación del trabajo fue la identificación de los ARGs en estas muestras. Para ello se probaron diferentes herramientas bioinformáticas con resultados dispares. Hay que señalar que los detalles tanto de la instalación como del funcionamiento de las tres herramientas que se mencionan a continuación (MetaHMM, Metalign y DIAMOND) se encuentran recopilados en el apartado 1.4 de esta memoria.

Primero se probó la herramienta web MetaHMM [41]. Esta herramienta está diseñada para encontrar genes en metagenomas usando modelos ocultos de Markov. A primera vista podría ser una herramienta muy útil para realizar la tarea de identificación de ARGs dado que se basa en buscar un listado de genes (en este caso serían los presentes en la base de datos de ARGs) en diferentes metagenomas predeterminados. Lo interesante es que la herramienta da la opción de definir metagenomas personalizados con lo que se podrían utilizar los metagenomas seleccionados con anterioridad. Los ficheros de estos metagenomas deben estar accesibles a través de una conexión directa de ftp o http pero tras varios intentos no se consiguió que la herramienta accediera a los ficheros por lo que, tras acumular un retraso considerable en el cronograma, se decidió dejar de lado esta herramienta para intentar explorar otras alternativas más sencillas y directas.

Como segunda opción se probó Metalign [58] que es un método diseñado para estimar la composición taxonómica de metagenomas basado en alineamientos eficientes de metagenomas WGS con diferentes bases de datos. En principio podría ser una herramienta muy válida que además proporcionaría directamente la composición taxonómica de las muestras, esto es, la

[1] https://api.mg-rast.org/search?material=compost&public=TRUE&limit=1000&sequence_type=WGS

distribución de las bacterias con ARGs para cada metagenoma. Según la documentación [67] es posible definir a través de la línea de comandos una base de datos personalizada para realizar las comparaciones pero, tras una instalación muy extensa (>300 GB) en la que se instala además del programa en sí una base de datos muy completa, fue imposible definir una base de datos personalizada. En este punto, desafortunadamente ya muy por detrás de la planificación temporal inicial, se decidió dejar esta herramienta de lado y se llevó a cabo una búsqueda bibliográfica más extensa con la esperanza de encontrar alguna herramienta que permitiera realizar el análisis deseado de manera rápida y sencilla.

A estas alturas de proyecto ya se había descartado realizar el análisis taxonómico de las muestras por falta de tiempo (tareas T5 y T6 de la planificación del trabajo, apartado 1.5) por lo que la búsqueda se centró en encontrar alguna herramienta específica de alineamiento de metagenomas frente a diferentes bases de datos personalizadas que pudiera ser útil para el desarrollo del proyecto. Para ello se revisó parte de la bibliografía recopilada con anterioridad además de los nuevos artículos encontrados en esta búsqueda.

De entre todos los potenciales candidatos revisados [117-123] se eligió la herramienta DIAMOND [69]. Esta herramienta, un poco más antigua que las anteriores, es básicamente un alineador de secuencias para proteínas diseñado y optimizado para analizar datos de secuenciación masiva mucho más rápido y de manera más eficiente que BLAST [72]. Tras una instalación rápida, se pudo definir la base de datos de ARGs como base de datos personalizada y se pudo realizar la comparación con los metagenomas seleccionados de muestras de compost de manera sencilla a través de la línea de comandos (detalles en el apartado 1.4). Es cierto que alguno de los análisis se alargó más de dos horas pero hay que mencionar que esto ocurrió sólo con los metagenomas más extensos de más de 10 GB. De cada análisis realizado se generó un fichero de alineación .tsv cuya estructura general se muestra a continuación (Figura 8).

	A	B	C	D	E	F	G	H	I	J	K	L
1	M01677:7:000000000-A4LYE:1:1107:16186:3624_1_100_-	ceoB	72.7	33	9	0	1	33	746	778	1.1e-10	55.1
2	M01677:7:000000000-A4LYE:1:1107:16186:3624_1_100_-	adeF	71.9	32	9	0	2	33	745	776	1.9e-10	54.3
3	M01677:7:000000000-A4LYE:1:1107:16186:3624_1_100_-	MexF	68.8	32	10	0	2	33	750	781	2.5e-10	53.9
4	M01677:7:000000000-A4LYE:1:1107:16186:3624_1_100_-	acrD	51.5	33	16	0	1	33	737	769	9.3e-05	35.4
5	M01677:7:000000000-A4LYE:1:1107:16186:3624_1_100_-	mdtF	48.5	33	17	0	1	33	737	769	9.3e-05	35.4
6	M01677:7:000000000-A4LYE:1:1107:16186:3624_1_100_-	TtgB	45.5	33	18	0	1	33	737	769	2.1e-04	34.3
7	M01677:7:000000000-A4LYE:1:1107:16186:3624_1_100_-	acrF	50.0	30	15	0	1	30	738	767	2.7e-04	33.9
8	M01677:7:000000000-A4LYE:1:2109:10905:8766_1_188_-	smeE	87.1	62	8	0	1	62	865	926	1.2e-26	109.0
9	M01677:7:000000000-A4LYE:1:2109:10905:8766_1_188_-	acrD	67.7	62	20	0	1	62	861	922	4.6e-21	90.5
10	M01677:7:000000000-A4LYE:1:2109:10905:8766_1_188_-	MexB	62.9	62	23	0	1	62	862	923	3.0e-20	87.8

Figura 8: Primeras filas del fichero de alineamiento mgm4538732.tsv

Hay que mencionar que es posible modificar el output de estos ficheros pero en este caso se decidió mantener los doce campos definidos por defecto en la herramienta [73]:

1. Código de la consulta: Código del read del metagenoma que se consulta contra la base de datos
2. Código del objetivo: Código del gen de la base de datos contra el que se alinea la consulta
3. Identidad de secuencia: Porcentaje de aminoácidos idénticos alineados en un alineamiento local de las dos secuencias
4. Longitud: Longitud total del alineamiento incluyendo mismatches y gaps en ambas secuencias
5. Mismatches: Número de aminoácidos no idénticos alineados
6. Gaps: Número de gaps en el alineamiento
7. Inicio de la consulta: Residuo inicial en el alineamiento para la secuencia de consulta
8. Fin de la consulta: Residuo final en el alineamiento para la secuencia de consulta
9. Inicio del objetivo: Residuo inicial en el alineamiento para la secuencia objetivo
10. Fin del objetivo: Residuo final en el alineamiento para la secuencia objetivo
11. E-valor: Es el valor esperado del número de alineamientos de similar o mejor calidad que se espera encontrar al realizar esta consulta frente a una base de datos de secuencias aleatorias del mismo tamaño que la base de datos real. Un valor de 0.001 significa que un alineamiento de esta calidad se puede encontrar al azar en promedio en una de cada 1000 consultas
12. Bit score: Es una medida independiente de la matriz de puntuación de la similitud de las dos secuencias alineadas. Número altos significan mayor similitud entre secuencias.

Debido a la gran cantidad de tiempo invertido en encontrar una herramienta que permitiera realizar el alineamiento de los genes de la base de datos de ARGs con los metagenomas seleccionados, el análisis de los resultados de estos alineamientos ha sido bastante reducido.

El análisis se comenzó cargando todos los ficheros de alineamiento .tsv al entorno de R, tras lo cual se generó una lista con todos datos de los ficheros de alineamiento descritos anteriormente.

```
#Configuración del directorio de trabajo que contiene los ficheros
.tsv
setwd("~/TFM/Diamond")

#Listado de los ficheros .tsv en el directorio
myfiles <- list.files(pattern = "*.tsv")

#Carga de los datos de los ficheros .tsv en los objetos R del listado
for (i in 1:length(myfiles)) assign(myfiles[i], read.delim(myfiles[i],
header = F))

#Creación de una lista con toda la información de los ficheros .tsv
(Tamaño muy grande)
mylist <- lapply(ls(pattern="*.tsv"), function(x) get(x))

#Definición del los nombres de los campos de cada fichero
colnames <- c("Read", "Gen", "Identidad", "Longitud", "Mismatch",
"Gap","Inicio read", "Fin read", "Inicio gen", "Fin gen", "E-value",
"Bit score")

#Modificación de los nombres de los campos de cada fichero en la lista
mylistcol <- lapply(mylist, setNames, colnames)
```

A continuación, se seleccionaron aquellos genes con reads en alguno de los metagenomas con una identidad de secuencia superior al 98%. Estos genes se tabularon de manera que se muestran ordenados según la frecuencia con la que aparecen en las muestras de metagenomas.

```
#Creación dataframe vacío
id98 <- NULL

#Selección de reads con identidad de secuencia >98% y campos reducidos
for (i in 1:length(mylistcol)) {id98[[i]] = mylistcol[[i]][c(2, 3, 4,
5, 6, 11, 12)][mylistcol[[i]][c(2, 3, 4, 5, 6, 11,
12)]$Identidad>98.0,]
}

#Creación dataframe vacío
id98unico <- NULL

#Selección genes únicos de la lista anterior
for (i in 1:length(mylistcol)) {id98unico[[i]] = unique(id98[[i]]$Gen)
}

#Tabla de genes únicos de manera individual con identidades >98%
id98tabla <- sort(table(unlist(lapply(id98unico, unique))), decreasing
= T)

#Frecuencia de genes en metagenomas
id98xc <- round(id98tabla/length(myfiles)*100, 2)
```

En la figura 9 se muestran los genes que aparecen al menos en un 25% de los metagenomas con identidades superiores al 98%. El fichero *id98xc25.csv* se ha completado con información obtenida de la base de datos de ARGs para generar el fichero final *id98final.csv*. Este fichero se puede encontrar en el repositorio https://github.com/adurana/TFM.

```
#Selección de genes más frecuentes (>25%)
id98xc25 <- id98xc[id98xc>25.0]

length(id98xc25)

## [1] 24

#Guardar datos para completar con información de la base de datos
write.csv2(id98xc25, file = "~/TFM/id98xc25.csv")

#Lectura del fichero final con información de clase, antibiótico y
organismo de la base de datos
id98final <- read.csv2(file = "~/TFM/id98final.csv", header = T)

id98final
```

	A	B	C	D	E
1	Gen	Frecuencia	Clase	Antibiotico	Organismo
2	rrnS	80,62	Aminoglycoside	Streptomycin	Chlamydomonas reinhardtii
3	aadA	59,38	Aminoglycoside	Streptomycin	Escherichia coli
4	sul2	58,13	Sulfonamide	Sulfonamide	Vibrio cholerae
5	aadA24	56,25	Aminoglycoside	Streptomycin, Spectinomycin	Salmonella enterica subsp. enterica serovar Newport
6	aph(6)-Id	54,37	Aminoglycoside	Streptomycin	Pseudomonas aeruginosa
7	sul1	54,37	Sulfonamide	Sulfonamide	Vibrio fluvialis
8	APH(3'')-Ib	44,38	Aminoglycoside	Streptomycin	Pseudomonas aeruginosa
9	aadA6	40,62	Aminoglycoside	Streptomycin	Pseudomonas aeruginosa
10	acrB	38,12	Tetracycline, Penam, Rifamycin, Glycylcycline, Cephalos	Multidrug	Escherichia coli str. K-12 substr. MG1655
11	acrD	34,38	Aminoglycoside	Multidrug	Escherichia coli str. K-12 substr. W3110
12	acrF	33,75	Fluoroquinolone, Penam, Cephalosporin, Cephamycin	Multidrug	Escherichia coli str. K-12 substr. MG1655
13	mtrA	31,87	Macrolide, Penam	Multidrug	Mycobacterium tuberculosis H37Rv
14	smeE	31,25	Fluoroquinolone, Macrolide, Tetracycline, Phenicol	Multidrug	Stenotrophomonas maltophilia
15	ttgB	29,38	Tetracycline	Tetracycline	Pseudomonas putida
16	adeJ	29,38	Diaminopyrimidine, Rifamycin, Penem, Tetracycline, Ph	Multidrug	Acinetobacter baumannii
17	CRP	27,5	Penam, Macrolide, Fluoroquinolone	Multidrug	Escherichia coli str. K-12 substr. W3110
18	MexB	27,5	Peptide, Penam, Diaminopyrimidine, Sulfonamide, Fluo	Multidrug	Pseudomonas aeruginosa
19	mdtB	27,5	Aminocoumarin	Multidrug	Escherichia coli str. K-12 substr. MG1655
20	TEM-1	26,88	Cephalosporin, Penam, Penem, Monobactam	Beta-lactam	Salmonella enterica subsp. enterica serovar Typhi str. CT18
21	emrB	26,88	Fluoroquinolone	Multidrug	Escherichia coli str. K-12 substr. MG1655
22	cpxA	26,25	Aminoglycoside, Aminocoumarin		Escherichia coli O157:H7 str. Sakai
23	adeF	25,62	Tetracycline, Fluoroquinolone	Multidrug	Acinetobacter baumannii AYE
24	tet(L)	25,62	Tetracycline	Tetracycline	Geobacillus stearothermophilus
25	mdtF	25,62	Fluoroquinolone, Macrolide, Penam	Multidrug	Escherichia coli str. K-12 substr. MG1655

Figura 9: Genes más frecuentes con identidad de secuencia superior al 98%

La columna de frecuencia representa el porcentaje sobre los metagenomas totales analizados en los que se ha encontrado dicho gen, con un valor de identidad de secuencia superior al 98%.

De este modo se encontraron 222 genes únicos de la base de datos de ARGs (360 genes) con al menos una read con una identidad de secuencia superior al 98%. Estos genes corresponden a un 61,7% de los genes totales identificados en los metagenomas de muestras de compost.

```r
#Número de genes y porcentaje sobre la base de datos
length(id98xc)

## [1] 222

round(length(id98xc)/nrow(BD_ARGs)*100, 1)

## [1] 61.7
```

De manera similar a lo realizado para la base de datos, se analizaron los datos para la clase, el antibiótico y el organismo de los genes mas frecuentes con más de un 98% de identidad de secuencia.

Con respecto a la clase de antibióticos se observó (Figura 10 y Tabla 8) que la resistencia más común de los genes analizados es para la familia de los antibióticos de tipo aminoglucósidos del mismo modo que ocurría con la base de datos completa.

```r
#Piechart con las clases de antibiótico de los genes con 98% identidad
labels_cl_98 <- rownames(head(sort(table(id98final$Clase), decreasing
= T), 5))

pie(sort(table(id98final$Clase), decreasing = T),
    init.angle = 90, clockwise = T, labels = labels_cl_98, col =
rainbow(nrow(id98final)))
```

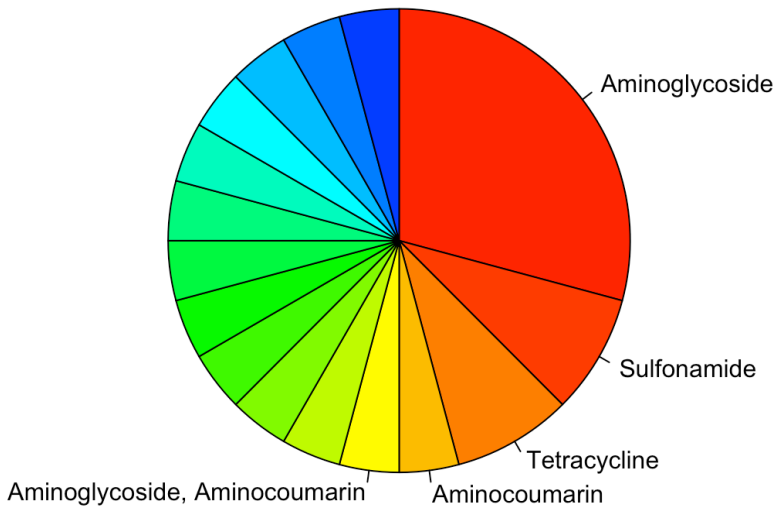


Figura 10: Gráfico de las clases de antibiótico más comunes en genes con identidades de secuencia superiores al 98%

Tabla 8: Tabla de las clases de antibiótico más comunes en genes con identidades de secuencia superiores al 98%

```
#Tabla con las primeras clases de antibiótico de los genes con 98%
identidad
head(as.data.frame(sort(table(id98final$Clase), decreasing = T)), 10)
```

```
##                                                        Var1  Freq
## 1                                            Aminoglycoside     7
## 2                                               Sulfonamide     2
## 3                                              Tetracycline     2
## 4                                             Aminocoumarin     1
## 5                                Aminoglycoside, Aminocoumarin     1
## 6                    Cephalosporin, Penam, Penem, Monobactam     1
## 7   Diaminopyrimidine, Rifamycin, Penem, Tetracycline,
          Phenicol, Carbapenem, Macrolide, Lincosamide,
                              Fluoroquinolone, Cephalosporin     1
## 8                                         Fluoroquinolone     1
## 9                          Fluoroquinolone, Macrolide, Penam     1
## 10 Fluoroquinolone, Macrolide, Tetracycline, Phenicol     1
```

En cuanto a los antibióticos concretos (Figura 11 y Tabla 9), los resultados muestran que los genes más frecuentes son los que tienen múltiples resistencias a antibióticos, del mismo modo que ocurría en la base de datos completa.

```
#Piechart con los antibióticos de los genes con 98% identidad
labels_ab_98 <- rownames(head(sort(table(id98final$Antibiotico),
decreasing = T), 6))

pie(sort(table(id98final$Antibiotico), decreasing = T),
    init.angle = 90, clockwise = T, labels = labels_ab_98, col =
rainbow(nrow(id98final)))
```

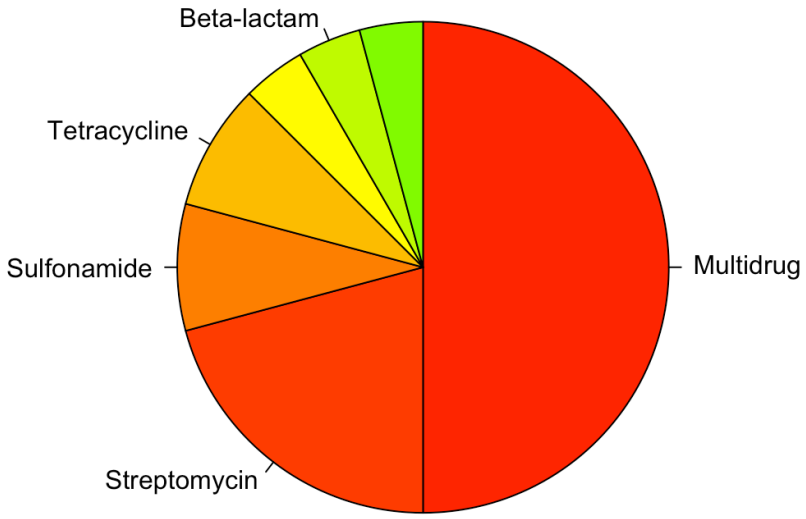


Figura 11: Gráfico de las resistencias a antibiótico más comunes en genes con identidades de secuencia superiores al 98%

Tabla 9: Tabla de las resistencias a antibiótico más comunes en genes con identidades de secuencia superiores al 98%

```
#Tabla con los primeros antibióticos de los genes con 98% identidad
head(as.data.frame(sort(table(id98final$Antibiotico), decreasing =
T)), 10)
```

```
##                                Var1 Freq
## 1                         Multidrug   12
## 2                       Streptomycin    5
## 3                        Sulfonamide    2
## 4                       Tetracycline    2
## 5                                       1
## 6                        Beta-lactam    1
## 7 Streptomycin, Spectinomycin          1
```

Por último, al igual que se observó en la base de datos completa, los organismos más frecuentes en genes con más de un 98% de identidad de secuencia son *Escherichia coli* y *Pseudomonas aeruginosa* (Figura 12 y Tabla 10).

```
#Piechart con los organismos de los genes con 98% identidad
labels_org_98 <- rownames(head(sort(table(id98final$Organismo),
decreasing = T), 5))

pie(sort(table(id98final$Organismo), decreasing = T),
    init.angle = 90, clockwise = T, labels = labels_org_98, col =
rainbow(nrow(id98final)))
```

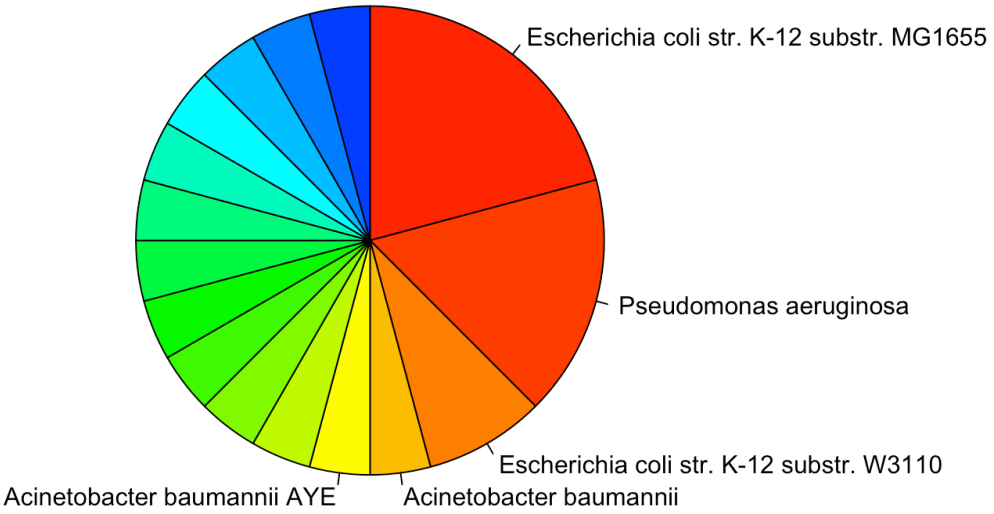


Figura 12: Gráfico de los organismos más comunes en genes con identidades de secuencia superiores al 98%

Tabla 10: Tabla de los organismos más comunes en genes con identidades de secuencia superiores al 98%

```
#Tabla con los primeros organismos de los genes con 98% identidad
head(as.data.frame(sort(table(id98final$Organismo), decreasing = T)),
10)
```

```
##                                          Var1 Freq
## 1  Escherichia coli str. K-12 substr. MG1655    5
## 2                      Pseudomonas aeruginosa    4
## 3   Escherichia coli str. K-12 substr. W3110    2
## 4                    Acinetobacter baumannii    1
## 5                Acinetobacter baumannii AYE    1
## 6                  Chlamydomonas reinhardtii    1
## 7                             Escherichia coli    1
## 8        Escherichia coli O157:H7 str. Sakai    1
## 9             Geobacillus stearothermophilus    1
## 10          Mycobacterium tuberculosis H37Rv    1
```

El mismo tratamiento de los datos de alineamiento llevado a cabo para un porcentaje de identidad superior al 98% se repitió para un porcentaje de identidad de secuencia superior al 90%.

```
#Creación dataframe vacio
id90 <- NULL

#Selección de reads con identidad de secuencia >90% y campos reducidos
for (i in 1:length(mylistcol)) {id90[[i]] = mylistcol[[i]][c(2, 3, 4,
5, 6, 11, 12)][mylistcol[[i]][c(2, 3, 4, 5, 6, 11,
12)]$Identidad>90.0,]
}

#Creación dataframe vacio
id90unico <- NULL

#Selección genes únicos de la lista anterior
for (i in 1:length(mylistcol)) {id90unico[[i]] = unique(id90[[i]]$Gen)
}

#Tabla de genes unicos con identidades >90% individualmente
id90tabla <- sort(table(unlist(lapply(id90unico, unique))), decreasing
= T)

#Frecuencia de genes en metagenomas
id90xc <- round(id90tabla/length(myfiles)*100, 2)
```

En la figura 13, de manera similar al caso anterior, se muestran los treinta genes más frecuentes que aparecen al menos en un 25% de los metagenomas con identidades superiores al 90%. El fichero *id90xc25.csv* se ha completado con información obtenida de la base de datos de ARGs para generar el fichero

final *id90final.csv*. El listado completo se puede encontrar en el repositorio https://github.com/adurana/TFM.

```
#Selección de genes más frecuentes (>25%)
id90xc25 <- id90xc[id90xc>25.0]

length(id90xc25)

## [1] 55

#Guardar datos para completar con información de la base de datos
write.csv2(id90xc25, file = "~/TFM/id90xc25.csv")

#Lectura del fichero final con información de clase, antibiotico y
organismo de la base de datos
id90final <- read.csv2(file = "~/TFM/id90final.csv", header = T)

id90final
```

	A	B	C	D	E
1	Gen	Frecuencia	Clase	Antibiotico	Organismo
2	rrnS	88,75	Aminoglycoside	Streptomycin	Chlamydomonas reinhardtii
3	acrB	80,63	Tetracycline, Penam, Rifamycin, Glycylcycline, Cephalosporin, Phenicol, Triclosan, Fluoroquinolone	Multidrug	Escherichia coli str. K-12 substr. MG1655
4	SdeY	76,88	Multidrug	Multidrug	Serratia marcescens
5	mtrA	76,25	Macrolide, Penam	Multidrug	Mycobacterium tuberculosis H37Rv
6	MexB	75,00	Peptide, Penam, Diaminopyrimidine, Sulfonamide, Fluoroquinolone, Monobactam, Phenicol, Carbap	Multidrug	Pseudomonas aeruginosa
7	acrF	72,50	Fluoroquinolone, Penam, Cephalosporin, Cephamycin	Multidrug	Escherichia coli str. K-12 substr. MG1655
8	smeE	71,25	Fluoroquinolone, Macrolide, Tetracycline, Phenicol	Multidrug	Stenotrophomonas maltophilia
9	adeF	71,25	Tetracycline, Fluoroquinolone	Multidrug	Acinetobacter baumannii AYE
10	MexF	70,00	Diaminopyrimidine, Phenicol, Fluoroquinolone	Multidrug	Pseudomonas aeruginosa PAO1
11	mexK	69,38	Macrolide, Tetracycline, Triclosan	Multidrug	Pseudomonas aeruginosa PAO1
12	acrD	68,13	Aminoglycoside	Multidrug	Escherichia coli str. K-12 substr. W3110
13	ttgb	66,88	Tetracycline	Tetracycline	Pseudomonas putida
14	ceoB	65,00	Aminoglycoside, Fluoroquinolone	Multidrug	Burkholderia cepacia
15	smeB	64,38	Cephalosporin, Cephamycin, Penam, Aminoglycoside	Multidrug	Stenotrophomonas maltophilia
16	mexW	62,50	Tetracycline, Acridine, Macrolide, Phenicol, Fluoroquinolone		Pseudomonas aeruginosa PAO1
17	adeJ	62,50	Diaminopyrimidine, Rifamycin, Penem, Tetracycline, Phenicol, Carbapenem, Macrolide, Lincosamide	Multidrug	Acinetobacter baumannii
18	aadA	61,88	Aminoglycoside	Streptomycin	Escherichia coli
19	aadA24	61,88	Aminoglycoside	Streptomycin, Spectinomycin	Salmonella enterica subsp. enterica serovar Newport
20	sul2	58,13	Sulfonamide	Sulfonamide	Vibrio cholerae
21	mdtF	58,13	Fluoroquinolone, Macrolide, Penam	Multidrug	Escherichia coli str. K-12 substr. MG1655
22	APH(6)-Id	56,25	Aminoglycoside	Streptomycin	Pseudomonas aeruginosa
23	aadA2	56,25	Aminoglycoside	Streptomycin, Spectinomycin	Klebsiella pneumoniae
24	pncA	55,63	Pyrazinamide	Pyrazinamide	Mycobacterium tuberculosis H37Rv
25	aadA3	55,63	Aminoglycoside	Streptomycin	Plasmid NR79
26	sul1	55,63	Sulfonamide	Sulfonamide	Vibrio fluvialis
27	msbA	54,38	Nitroimidazole	Multidrug	Escherichia coli str. K-12 substr. MG1655
28	mdtB	54,38	Aminocoumarin	Multidrug	Escherichia coli str. K-12 substr. MG1655
29	aadA6	53,75	Aminoglycoside	Streptomycin	Pseudomonas aeruginosa
30	macB	50,00	Macrolide		Neisseria gonorrhoeae
31	mdtC	48,75	Aminocoumarin	Multidrug	Escherichia coli str. K-12 substr. MG1655

Figura 13: Primeros genes más frecuentes con identidad de secuencia superior al 90%

En el caso de los genes con valores superiores a un 90% de identidad de secuencia, se encontraron 279 genes únicos de la base de datos de ARGs con al menos una read con una identidad de secuencia superior al 90%. Estos genes corresponden a un 77,5% de los genes totales identificados en los metagenomas de muestras de compost.

```
#Número de genes y porcentaje sobre la base de datos
length(id90xc)

## [1] 279

round(length(id90xc)/nrow(BD_ARGs)*100, 1)

## [1] 77.5
```

De manera similar al caso anterior, se realizó el mismo análisis para observar los datos más frecuentes para la clase, el antibiótico y el organismo de los genes con más de un 90% de identidad de secuencia.

Con respecto a la clase de antibióticos (Figura 14 y Tabla 11) se observó que las resistencias más comunes de los genes analizados son para las familias de antibióticos aminoglucósidos y tetraciclinas, del mismo modo que ocurría con la base de datos completa y con los genes con porcentaje de identidad al 98%.

```
#Piechart con las clases de antibiótico de los genes con 90% identidad
labels_cl_90 <- rownames(head(sort(table(id90final$Clase), decreasing
= T), 8))

pie(sort(table(id90final$Clase), decreasing = T),
    init.angle = 90, clockwise = T, labels = labels_cl_90, col =
rainbow(nrow(id90final)))
```

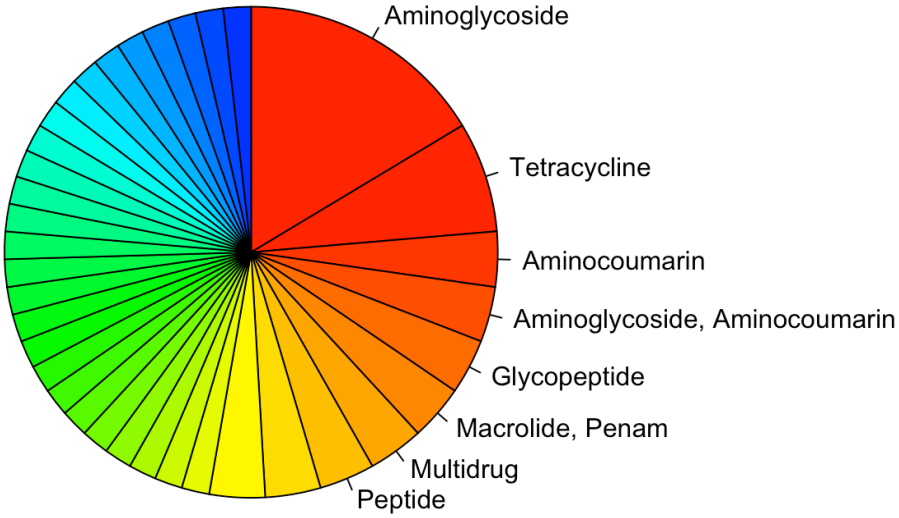


Figura 14: Gráfico de las clases de antibiótico más comunes en genes con identidades de secuencia superiores al 90%

Tabla 11: Tabla de las clases de antibiótico más comunes en genes con identidades de secuencia superiores al 90%

```
#Tabla con las primeras clases de antibiótico de los genes con 90%
identidad
head(as.data.frame(sort(table(id90final$Clase), decreasing = T)), 10)
```

```
##                                Var1 Freq
## 1                     Aminoglycoside    9
## 2                       Tetracycline    4
## 3                     Aminocoumarin     2
## 4  Aminoglycoside, Aminocoumarin      2
## 5                      Glycopeptide     2
## 6                  Macrolide, Penam     2
## 7                         Multidrug     2
## 8                           Peptide     2
```

```
## 9                      Sulfonamide    2
## 10   Tetracycline, Glycylcycline    2
```

En cuanto a los antibióticos concretos, los resultados muestran que los más frecuentes son los genes multiresistentes, del mismo modo que en todos los casos anteriores (Figura 15 y Tabla 12).

```
#Piechart con los antibióticos de los genes con 90% identidad
labels_ab_90 <- rownames(head(sort(table(id90final$Antibiotico),
decreasing = T), 8))

pie(sort(table(id90final$Antibiotico), decreasing = T),
    init.angle = 90, clockwise = T, labels = labels_ab_90, col =
rainbow(nrow(id90final)))
```

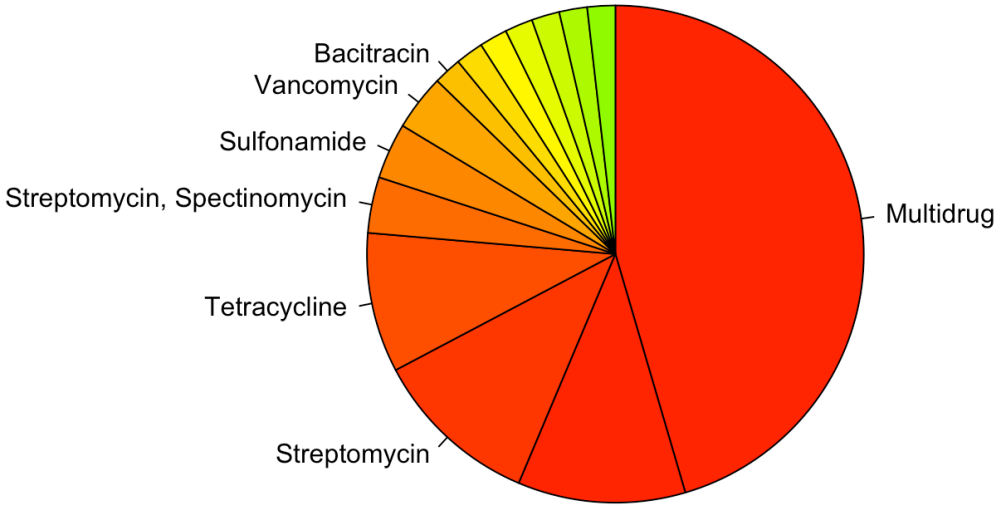


Figura 15: Gráfico de las resistencias a antibiótico más comunes en genes con identidades de secuencia superiores al 90%

Tabla 12: Tabla de las resistencias a antibiótico más comunes en genes con identidades de secuencia superiores al 90%

```
#Tabla con los primeros antibióticos de los genes con 90% identidad
head(as.data.frame(sort(table(id90final$Antibiotico), decreasing =
T)), 10)
```

```
##                          Var1 Freq
## 1                    Multidrug   25
## 2                                6
## 3                 Streptomycin    6
## 4                 Tetracycline    5
## 5  Streptomycin, Spectinomycin    2
## 6                  Sulfonamide    2
## 7                   Vancomycin    2
## 8                   Bacitracin    1
## 9                   Beta-lactam    1
## 10             Chloramphenicol    1
```

Por último, del mismo modo que los casos anteriores, los organismos más frecuentes de los genes con porcentajes de identidad superiores al 90% son diferentes cepas de *Escherichia coli* y *Pseudomonas aeruginosa* (Figura 16 y Tabla 13).

```
#Piechart con los organismos de los genes con 90% identidad
labels_org_90 <- rownames(head(sort(table(id90final$Organismo),
decreasing = T), 5))

pie(sort(table(id90final$Organismo), decreasing = T),
    init.angle = 90, clockwise = T, labels = labels_org_90, col =
rainbow(nrow(id90final)))
```

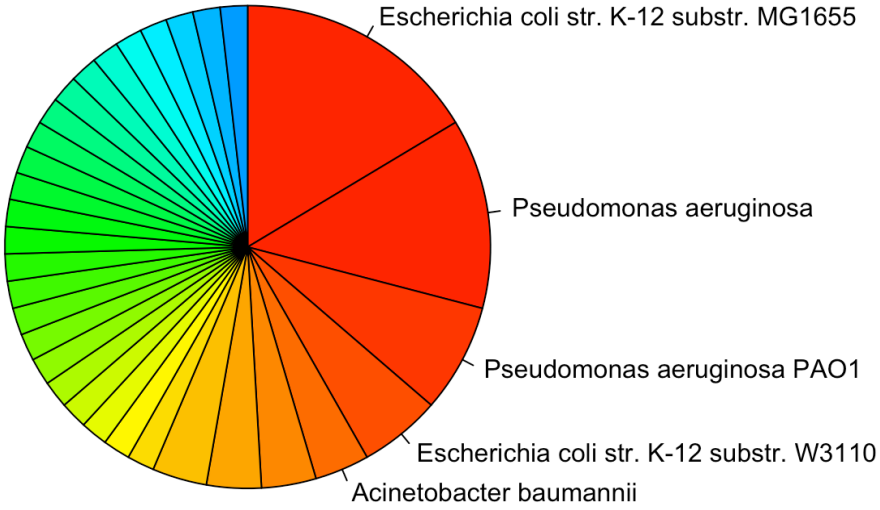


Figura 16: Gráfico de los organismos más comunes en genes con identidades de secuencia superiores al 90%

Tabla 13: Tabla de los organismos más comunes en genes con identidades de secuencia superiores al 90%

```
#Tabla con los primeros organismos de los genes con 90% identidad
head(as.data.frame(sort(table(id90final$Organismo), decreasing = T)),
10)
```

```
##                                        Var1 Freq
## 1  Escherichia coli str. K-12 substr. MG1655    9
## 2                    Pseudomonas aeruginosa    7
## 3                Pseudomonas aeruginosa PAO1    4
## 4   Escherichia coli str. K-12 substr. W3110    3
## 5                    Acinetobacter baumannii    2
## 6                             Escherichia coli    2
## 7           Mycobacterium tuberculosis H37Rv    2
## 8               Stenotrophomonas maltophilia    2
## 9                Acinetobacter baumannii AYE    1
## 10                      Bacteroides fragilis    1
```

De todo este análisis preliminar se puede concluir que las características de los genes más frecuentes tanto en la base de datos completa, como en los casos de los genes identificados en los metagenomas con diferentes porcentajes de identidad de secuencia son bastante similares, lo que podría indicar que la composición de las diferentes muestras de compost son similares, al menos en cuanto al tipo de ARG que se encuentra en las mismas.

Es interesante remarcar que incluso en el caso de genes con más de un 98% de identidad de secuencia (Figura 9), se encontró una cantidad elevada de coincidencias en los diferentes metagenomas. En concreto se encontraron 6 genes presentes en más de la mitad de los metagenomas analizados. Este número aumenta en el caso de genes con identidad de secuencia superior al 90% a un total de 29 genes (Figura 13). La ubiquidad de muchos de estos genes evidencia la gran extensión que tienen los ARGs en las diferentes muestras de compost.

Como era de esperar, a menor porcentaje de identidad de secuencia, la frecuencia con la que se identifican genes en los metagenomas aumenta. En concreto, al revisar los datos individuales de algún gen como por ejemplo *acrB*, se observó una gran variabilidad entre el 80,63% de frecuencia en el caso de una identidad de secuencia del 90% y 38,12% de frecuencia para una identidad del 98%. Esto, *a priori*, parece indicar que el tratamiento de los datos es correcto pero sin un análisis más detallado y completo es complicado obtener conclusiones concretas de los mismos.

Al conseguir identificar los genes resistentes a antibiótico presentes en muestras de compost se consiguió cumplir el tercer hito de la planificación del trabajo descrito en el apartado 1.5. Por desgracia, como se ha comentado anteriormente, resultó imposible realizar un análisis más exhaustivo de estos datos así como ni siquiera comenzar con las tareas T5 y T6 descritas en la planificación del proyecto debido al retraso acumulado en la tarea T4 a la hora de encontrar una herramienta bioinformática válida para realizar la identificación de los ARGs. De todas maneras es interesante reseñar que los ficheros .tsv resultantes de los alineamientos de DIAMOND se pueden acoplar al pipeline del analizador interactivo de metagenomas MEGAN6 [124, 125] por lo que la realización de estas tareas podría llevarse a cabo de una forma bastante directa en el futuro.

3. Conclusiones

La principal conclusión que se puede extraer de este trabajo es que la distribución de los genes resistentes a antibióticos en muestras de compost es bastante similar a la de la base de datos de referencia, al menos en lo relativo al tipo de antibiótico al que son resistentes.

Se ha observado tanto para la base de datos como para los metagenomas que los genes resistentes a familias de antibióticos como aminoglucósidos o tetraciclinas, multirresistentes y provenientes de diferentes cepas de *Escherichia coli* y *Pseudomonas aeruginosa* son los más frecuentes en todos los casos analizados.

También es reseñable que se ha conseguido identificar un porcentaje elevado de genes resistentes a antibióticos en muestras de compost utilizando datos de secuenciación masiva: 77.5% para genes con un porcentaje de identidad de secuencia superior al 90% y 61.7% para genes con un porcentaje de identidad de secuencia superior al 98%. El análisis preliminar de los datos obtenidos indica que a mayor porcentaje de identidad de secuencia entre los reads y los genes se identifican menos genes, como era de esperar. Esta tendencia es indicativa de que el tratamiento de los datos de alineamiento parece correcto. Además se ha encontrado un número elevado de genes presentes en muchos de los metagenomas, lo que evidencia la amplia distribución de los ARGs en el medioambiente.

Siendo realistas, no es posible extraer más conclusiones concretas con respecto a los resultados de este trabajo porque el retraso generado en la identificación de los genes ha impedido la realización de un análisis completo y exhaustivo de los datos de alineamiento. Por el mismo motivo, no ha sido posible lograr algunos de los objetivos propuestos en la planificación del trabajo (Apartado 1.2) así como varias de las tareas descritas inicialmente (Apartado 1.5).

De todas maneras, sí que se han logrado varios de los objetivos planteados. Concretamente, el objetivo general de identificar genes bacterianos con resistencia a antibiótico presentes en muestras de compost se ha completado como se ha mostrado en el apartado de resultados. Además, se han logrado otros objetivos secundarios como la creación de una base de datos de ARGs y la selección de datos de secuenciación masiva y su comparación con dicha base de datos. Por desgracia, no se han logrado el resto de objetivos debido al considerable retraso generado a la hora de encontrar una herramienta eficaz y fiable para el alineamiento de los metagenomas con la base de datos. En la

misma línea, las tareas T1-T4 se han podido completar dentro del marco del proyecto, pero no así las tareas T5-T6.

Hay que mencionar que la planificación original se siguió sin problemas para las tres primeras tareas pero que la tarea T4 se alargó hasta consumir la totalidad de tiempo restante por lo que el cronograma final del proyecto (Figura 17) refleja el aumento de tiempo para completar la tarea T4 y el nulo tiempo asignado para las tareas T5 y T6.

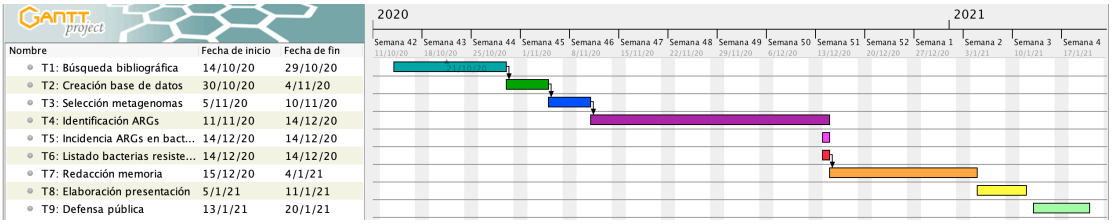


Figura 17: Diagrama de Gantt final del proyecto

Como se ha mencionado anteriormente, no se ha podido realizar un análisis completo y en profundidad de los datos de alineamiento pero sería muy interesante poder llevarlo a cabo dado que es posible que se obtuviera información relevante de los mismos. Además, sería interesante continuar con las tareas T5 y T6 dado que realizando este análisis taxonómico de las muestras, se podría comprender mejor la distribución de bacterias resistentes a antibiótico en muestras de compost. Por lo tanto, este trabajo queda abierto a un posterior desarrollo.

4. Glosario

- ADN: Ácido desoxirribonucleico
- ARG: Gen resistente a antibiótico (Antibiotic Resistant Gene)
- ARN: Ácido ribonucleico
- BLAST: Programa de alineamiento local de secuencias (Basic Local Alignment Search Tool)
- Composición taxonómica: Clasificación ordenada de los organismos de una comunidad bacteriana
- FASTA: Formato de fichero en texto usado para representar secuencias
- FTP: Protocolo de red para la transferencia de archivos (File Transfer Protocol)
- HGT: Transferencia genética horizontal (Horizontal Gene Transfer)
- HTTP: Protocolo de comunicación que permite la transferencia de información a través de archivos (Hypertext Transfer Protocol)
- Metagenoma: Conjunto de genes de una muestra medioambiental concreta
- MGE: Elementos genéticos móviles (Mobile Genetic Elements)
- NGS: Secuenciación de nueva generación, secuenciación masiva o ultrasecuenciación (Next Generation Sequencing)
- Resistoma: Colección de genes resistentes presentes en la comunidad bacteriana de una muestra medioambiental.
- WGS: Secuenciación del genoma completo (Whole Genome Sequencing)

5. Bibliografía

[1] He, Y.; Yuan, Q.; Mathieu, J.; Stadler, L.; Senehi, N.; Sun, R.; Alvarez, P. J. J., 2020. Antibiotic resistance genes from livestock waste: occurrence, dissemination, and treatment. npj Clean Water 3, 4.

[2] Singer, A. C.; Shaw, H.; Rhodes, V.; Hart, A., 2016. Review of Antimicrobial Resistance in the Environment and Its Relevance to Environmental Regulators. Front. Microbiol. 7:1728

[3] Berendonk, T. U.; Manaia, C.M.; Merlin, C. et al., 2015. Tackling antibiotic resistance: the environmental framework. Nat. Rev. Microbiol. 13, 301-317.

[4] Checcucci, A.; Trevisi, P.; Luise, D.; Modesto, M.; Blasioli, S.; Braschi, I.; Mattarelli, P., 2020. Exploring the animal waste resistome: the spread of antimicrobial resistance genes through the use of livestock manure. Front. Microbiol. 11:1416.

[5] Nguyen, B-A. T.; Chen, Q-L.; He, J-Z.; Hu, H-W., 2020. Microbial regulation of natural antibiotic resistance: Understanding the protist-bacteria interactions for evolution of soil resistome. Sci. Total Environ. 705, 135882.

[6] Jechalke, S.; Heuer, H.; Siemens, J.; Amelung, W.; Smalla, K., 2014. Fate and effects of veterinary antibiotics in soil. Trends in Microbiol. 22(9), 536-545.

[7] Yuan, Q-B.; Zhai, Y-F.; Mao, B-Y.; Schwarz, C.; Hu, N., 2019. Fates of antibiotic resistance genes in a distributed swine wastewater treatment plant. Water Environ. Res. 91:1565-1575.

[8] Dungan, R. S.; McKinney, C. W.; Leytem, A. B., 2018. Tracking antibiotic resistance genes in soil irrigated with dairy wastewater. Sci. Total Environ. 635, 1477-1483.

[9] Troiano, E.; Beneduce, L.; Gross, A.; Ronen, Z., 2018. Antibiotic-Resistant Bacteria in Greywater and Greywater-Irrigated Soils. Front. Microbiol. 9:2666.

[10] Fang, P.; Peng, F.; Gao, X.; Xiao, P.; Yang, J., 2019. Decoupling the Dynamics of Bacterial Taxonomy and Antibiotic Resistance Function in a Subtropical Urban Reservoir as Revealed by High-Frequency Sampling. Front. Microbiol. 10:1448.

[11] Wright, G. D., 2007. The antibiotic resistome: the nexus of chemical and genetic diversity. Nat. Rev. Microbiol. 5, 175-186.

[12] Perry, J. A.; Westman, E. L.; Wright, G. D., 2014. The antibiotic resistome: what's new? Curr. Opin. Microbiol. 21:45-50.

[13] Lopatto, E.; Choi, J.; Colina, A.; Ma, L.; Howe, A.; Hinsa-Leasure, S., 2019. Characterizing the soil microbiome and quantifying antibiotic resistance gene dynamics in agricultural soil following swine CAFO manure application. PLoS ONE 14(8): e0220770.

[14] Gao, F-Z.; He, L-Y.; He, L-X.; Zou, H-Y.; Zhang, M.; Wu, D-L.; Liu, Y-S.; Shi, Y-J.; Bai, H.; Ying, G-G., 2020. Untreated swine wastes changed antibiotic resistance and microbial community in the soils and impacted abundances of antibiotic resistance genes in the vegetables. Sci. Total Environ. 741, 140482.

[15] Wang, C.; Dong, D.; Strong, P. J.; Zhu, W.; Ma, Z.; Qin, Y.; Wu, W., 2017. Microbial phylogeny determines transcriptional response of resistome to dynamic composting processes. Microbiome 5:103.

[16] Youngquist, C. P.; Mitchell, S. M.; Cogger, C. G., 2016. Fate of Antibiotics and Antibiotic Resistance during Digestion and Composting: A Review. J. Environ. Qual. 45:537-545.

[17] Qian, X.; Gu, J.; Sun, W.; Wang, X.; Li, H., 2019. Effects of passivators on antibiotic resistance genes and related mechanisms during composting of copper-enriched pig manure. Sci. Total Environ. 674, 383-391.

[18] Sun, Y.; Qiu, T.; Gao, M.; Shi, M.; Zhang, H.; Wang, X., 2019. Inorganic and organic fertilizers application enhanced antibiotic resistome in greenhouse soils growing vegetables. Ecotoxicol. Environ. Saf. 179, 24-30.

[19] Xu, M.; Stedtfeld, R. D.; Wang, F.; Hashsham, S. A.; Song, Y.; Chuang, Y.; Fan, J.; Li, H.; Jiang, X.; Tiedje, J. M., 2019. Composting increased persistence of manure-borne antibiotic resistance genes in soils with different fertilization history. Sci. Total Environ. 689, 1172-1180.

[20] Zhang, M.; He, L-Y.; Liu, Y-S.; Zhao, J-L.; Liu, W-R.; Zhang, J-N.; Chen, J.; He, L-K.; Zhang, Q-Q.; Ying, G-G., 2019. Fate of veterinary antibiotics during animal manure composting. Sci. Total Environ. 650, 1363-1370.

[21] Liu, Y.; Cheng, D.; Xue, J.; Weaver, L.; Wakelin, S. A.; Feng, Y.; Li, Z., 2020. Changes in microbial community structure during pig manure composting and its relationship to the fate of antibiotics and antibiotic resistance genes. J. Hazard. Mater. 389, 122082.

[22] Wei, H.; Ding, S.; Qiao, Z.; Su, Y.; Xie, B., 2020. Insights into factors driving the transmission of antibiotic resistance from sludge compost-amended soil to vegetables under cadmium stress. Sci. Total Environ. 729, 138990.

[23] Guo, W.; Huang, C.; Xi, B.; Tang, Z.; Tan, W.; Li, W.; Zhang, Y.; Li, W., 2021. The maturity period is the main stage of antibiotic resistance genes reduction in aerobic composting process of swine manure in sub-scale farms. Bioresour. Technol. 319, 124139.

[24] Zhao, W.; Gu, J.; Wang, X.; Hu, T.; Wang, J.; Yu, J.; Dia, X.; Lei, L., 2021. Effects of shrimp shell powder on antibiotic resistance genes and the bacterial community during swine manure composting. Sci. Total Environ. 752, 142162.

[25] Thomas, C. M.; Nielsen, K. M., 2005. Mechanisms of, and barrier to, horizontal gene transfer between bacteria. Nat. Rev. Microbiol. 3, 711-721.

[26] Alekshun, M. N.; Levy, S. B., 2007. Molecular Mechanisms of Antibacterial Multidrug Resistance. Cell 128, 1037-1050.

[27] van Hoek, A. H. A. M.; Mevius, D.; Guerra, B.; Mullany, P.; Roberts, A. P.; Aarts, H. J. M., 2011. Acquired antibiotic resistance genes: an overview. Front. Microbiol. 2:203.

[28] Blair, J. M. A.; Webber, M. A.; Baylay, A. J.; Ogbolu, D. O.; Piddock, L. J. V., 2015. Molecular mechanisms of antibiotic resistance. Nat. Rev. Microbiol. 13, 42-51.

[29] Zhang, M.; He, L-Y.; Liu, Y-S.; Zhao, J-L.; Zhang, J-N.; Chen, J.; Zhang, Q-Q.; Ying, G-G., 2020. Variation of antibiotic resistome during commercial livestock manure composting. Environ. Int. 136, 105458.

[30] Gao, Q.; Dong, Q.; Wu, L.; Yang, Y.; Hale, L; Qin, Z.; Xi, C.; Zhang, Q.; Van Nostrand, J. D.; Zhou, J., 2020. Environmental antibiotics drives the genetic functions of resistome dynamics. Environ. Int. 135, 105398.

[31] Dudhagara, P.; Bhavsar, S.; Bhagat, C.; Ghelani, A.; Bhatt, S.; Patel, R., 2015. Web Resources for Metagenomics Studies. Genomics Proteomics Bioinformatics 13, 296-303.

[32] Wang, Y.; Hu Y.; Gao G. F., 2020. Combining metagenomics and metatranscriptomics to study human, animal and environmental resistomes. Medicine in Microecology 100014.

[33] https://www.r-project.org/, 14/10/2020

[34] https://rstudio.com/, 14/10/2020

[35] https://github.com/, 02/11/2020

[36] https://www.bioinformatics.org/sms2/translate.html, 29/11/2020

[37] https://www.ebi.ac.uk/Tools/st/emboss_transeq/, 29/11/2020

[38] Meyer, F.; Paarmann, D.; D'Souza, M.; Olson, R.; Glass, E. M.; Kubal, M.; Paczian, T.; Rodriguez, A.; Stevens, R.; Wilke, A.; Wilkening, J.; Edwards, R. A., 2008. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics, 9:386.

[39] https://www.mg-rast.org/, 04/11/2020

[40] https://www.mg-rast.org/mgmain.html?mgpage=search&search=, 04/11/2020

[41] Szalkai, B.; Grolmusz, V., 2019. MetaHMM: A webserver for identifying novel genes with specified functions in metagenomic samples. Genomics 111, 883-885.

[42] https://pitgroup.org/metahmm/, 13/11/2020

[43] https://www.ebi.ac.uk/Tools/msa/clustalo/, 17/11/2020

[44] https://www.uniprot.org/, 17/11/2020

[45] https://www.solvetic.com/tutoriales/article/5873-como-instalar-ftp-en-macos-mojave/, 18/11/20

[46] https://techristic.com/how-to-install-ftp-on-macos-mojave-high-sierra/, 18/11/2020

[47] https://brew.sh/, 18/11/2020

[48] https://filezilla-project.org/, 19/11/2020

[49] https://httpd.apache.org/, 20/11/2020

[50] https://dyclassroom.com/howto-mac/how-to-install-apache-mysql-php-on-macos-mojave-10-14, 20/11/2020

[51] https://medium.com/better-programming/install-apache-mysql-php-macos-mojave-10-14-b6b5c00b7de, 20/11/2020

[52] https://discussions.apple.com/docs/DOC-13841, 21/11/2020

[53] https://stackoverflow.com/questions/10598926/access-localhost-from-another-computer-not-on-network, 21/11/2020

[54] https://ngrok.com/docs, 21/11/2020

[55] https://setuprouter.com/what-is-port-forwarding/, 22/11/2020

[56] https://superuser.com/questions/30917/how-to-make-a-port-forward-in-mac-os-x, 22/11/2020

[57] https://wilsonmar.github.io/ports-open/, 22/11/2020

[58] LaPierre, N.; Alser, M.; Eskin, E.; Koslicki, D.; Mangul, S., 2020. Metalign: efficient alignment-based metagenomic profiling via containment min hash. Genome Biol. 21, 242.

[59] https://github.com/nlapier2/Metalign, 24/11/2020

[60] Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34(18), 3094-3100.

[61] https://github.com/nlapier2/Metalign/wiki, 24/11/2020

[62] https://bioconda.github.io/index.html, 24/11/2020

[63] https://github.com/ocaml/opam/issues/3576, 25/11/2020

[64] https://apple.stackexchange.com/questions/266821/how-can-i-fix-permissions-for-files-in-the-var-folders-zz, 25/11/2020

[65] https://stackoverflow.com/questions/56695865/unable-to-modify-permissions-of-folders-on-mac-mojave, 25/11/2020

[66] https://apple.stackexchange.com/questions/358687/right-way-to-add-paths-to-path-in-mojave, 25/11/2020

[67] https://github.com/nlapier2/Metalign/issues/22, 26/11/2020

[68] https://gist.github.com/dkoslicki/4e0344080038088bf72d8bfa52626c91, 26/11/2020

[69] Buchfink, B.; Xie, C.; Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 59-60.

[70] https://www.wsi.uni-tuebingen.de/lehrstuehle/algorithms-in-bioinformatics/software/diamond/, 28/11/2020

[71] https://github.com/bbuchfink/diamond, 28/11/2020

[72] https://blast.ncbi.nlm.nih.gov/Blast.cgi, 15/11/2020

[73] https://github.com/bbuchfink/diamond/wiki, 28/11/2020

[74] https://www.ganttproject.biz/, 03/10/2020

[75] http://biblioteca.uoc.edu/es/, 14/10/2020

[76] https://www.sciencedirect.com, 15/10/2020

[77] https://onlinelibrary.wiley.com, 15/10/2020

[78] https://www.springer.com/gp, 15/10/2020

[79] https://www.pnas.org, 15/10/2020

[80] https://www.nature.com, 15/10/2020

[81] Zhu, Y-G.; Johnson, T. A.; Su, J-Q.; Qiao, M.; Guo, G-X.; Stedtfeld, R. D.; Hashsham, S. A.; Tiedje, J. M., 2013. Diverse and abundant antibiotic resistance genes in Chinese swine farms. PNAS 110, 9, 3435-3440.

[82] He, L-Y.; Ying, G-G.; Liu, Y-S.; Su, H-C.; Chen, J.; Liu, S-S.; Zhao, J-L., 2016. Discharge of swine wastes risks water quality and food safety: Antibiotics and antibiotic resistance genes from swine sources to the receiving environments. Environ. Int. 92-93, 210-219.

[83] Xie, W-Y.; Yang, X-P.; Li, Q.; Wu, L-H.; Shen, Q-R.; Zhao, F-J., 2016. Changes in antibiotic concentrations and antibiotic resistome during commercial composting of animal manures. Environ. Pollut. 219, 182-190.

[84] Tien, Y-C.; Li, B.; Zhang, T.; Scott, A.; Murray, R.; Sabourin, L.; Marti, R.; Topp, E., 2017. Impact of dairy manure pre-application treatment on manure composition, soil dynamics of antibiotic resistance genes, and abundance of antibiotic-resistance genes on vegetables at harvest. Sci. Total Environ. 581-582, 32-39.

[85] Tong, P.; Ji, X.; Chen, L.; Liu, J.; Xu, L.; Zhu, L.; Zhou, W.; Liu, G.; Wang, S.; Guo, X.; Feng, S.; Sun, Y., 2017. Metagenome analysis of antibiotic resistance genes in fecal microbiota of chickens. Agri Gene 5, 1-6.

[86] Qian, X.; Gu, J.; Sun, W.; Wang, X-J.; Su, J-Q.; Stedfeld, R., 2018. Diversity, abundance, and persistence of antibiotic resistance genes in various types of animal manure following industrial composting. J. Hazard. Mater. 344, 716-722.

[87] Awasthi, M. K.; Liu, T.; Chen, H.; Verma, S.; Duan, Y.; Awasthi, S. K.; Wang, Q.; Ren, X.; Zhao, J.; Zhang, Z., 2019a. The behavior of antibiotic resistance genes and their associations with bacterial community during poultry manure composting. Bioresour. Technol. 280, 70-78.

[88] Awasthi, M. K.; Chen, H.; Awasthi, S. K.; Duan, Y.; Liu, T.; Pandey, A.; Varjani, S.; Zhang, Z., 2019b. Application of metagenomic analysis for detection of the reduction in the antibiotic resistance genes (ARGs) by the addition of clay during poultry manure composting. Chemosphere 220, 137-145.

[89] Chen, C.; Pankow, C. A.; Oh, M.; Heath, L. S.; Zhang, L.; Du, P.; Xia, K.; Pruden, A., 2019a. Effect of antibiotic use and composting on antibiotic resistance gene abundance and resistome risks of soils receiving manure-derived amendments. Environ. Int. 128, 233-243.

[90] Chen, Z.; Zhang, W.; Yang, L.; Stedtfeld, R. D.; Peng, A.; Gu, C.; Boyd, S. A.; Li, H., 2019b. Antibiotic resistance genes and bacterial communities in cornfield and pasture soils receiving swine and dairy manures. Environ. Pollut. 248, 947-957.

[91] Eckstrom, K.; Barlow, J. W., 2019. Resistome metagenomics from plate to farm: The resistome and microbial composition during food waste feeding and composting on a Vermont poultry farm. PloS ONE 14(11): e0219807.

[92] Zeng, J.; Pan, Y.; Yang, J.; Hou, M.; Zeng, Z.; Xiong, W., 2019. Metagenomic insights into the distribution of antibiotic resistome between the gut-associated environments and the pristine environments. Environ. Int. 126, 346-354.

[93] Zhou, Z.; Yao, H., 2020. Effects of Composting Different Types of Organic Fertilizer on the Microbial Community Structure and Antibiotic Resistance Genes. Microorganisms 8, 268.

[94] Alcock, B. P.; Raphenya, A. R.; Lau, T. T. Y.; Tsang, K. K. et al., 2020. CARD 2020: antibiotic resistome surveillance with the comprehensive resistance database. Nucleic Acids Res. 48, D517-D525.

[95] https://card.mcmaster.ca/, 30/10/2020

[96] https://www.ncbi.nlm.nih.gov/ pathogens/refgene/, 30/10/2020

[97] https://www.ebi.ac.uk/ena/browser/home, 02/11/2020

[98] https://www.ncbi.nlm.nih.gov/genbank/, 01/11/2020

[99] https://www.ncbi.nlm.nih.gov/refseq/, 01/11/2020

[100] https://en.wikipedia.org/wiki/Aminoglycoside, 30/12/2020

[101] https://en.wikipedia.org/wiki/Tetracycline_antibiotics, 30/12/2020

[102] https://en.wikipedia.org/wiki/Glycopeptide_antibiotic, 30/12/2020

[103] https://en.wikipedia.org/wiki/Multiple_drug_resistance, 30/12/2020

[104] https://en.wikipedia.org/wiki/%CE%92-lactam_antibiotic, 30/12/2020

[105] https://en.wikipedia.org/wiki/Escherichia_coli, 30/12/2020

[106] https://en.wikipedia.org/wiki/Pseudomonas_aeruginosa, 30/12/2020

[107] https://en.wikipedia.org/wiki/Enterococcus_faecalis, 30/12/2020

[108] https://en.wikipedia.org/wiki/Enterococcus_faecium, 30/12/2020

[109] Markowitz, V. M.; Ivanova, N. N.; Szeto, E.; Palaniappan, K. et al., 2008. IMG/M: a data management and analysis system for metagenomes. Nucleic Acid Res. 36, D534-D538.

[110] Goll, J.; Rusch, D. B.; Tanenbaum, D. M.; Thiagarajan, M.; Li, K.; Methé, B. A.; Yooseph, S., 2010. METAREP: JCVI metagenomics reports – an open source tool for high-performance comparative metagenomics. Bioinformatics 26(20), 2631-2632.

[111] Arango-Argoty, G.; Singh, G.; Heath, L. S.; Pruden, A.; Xiao, W.; Zhang, L., 2016. MetaStorm: A Public Resource for Customizable Metagenomics Annotation. PLoS ONE 11(9): e0162442.

[112] Uritskiy, G. V.; DiRuggiero, J.; Taylor, J., 2018. MetaWRAP – a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome 6:158.

[113] https://help.mg-rast.org/user_manual.html, 07/11/2020

[114] https://github.com/MG-RAST, 07/11/2020

[115] https://adina-howe.readthedocs.io/en/latest/mgrast/, 07/11/2020

[116] https://api.mg-rast.org/api.html, 08/11/2020

[117] Hyatt, D.; Chen, G-L.; LoCascio, P. F.; Land, M. L.; Larimer, F. W.; Hauser, L. J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics, 11:119.

[118] Zankari, E.; Hasman, H.; Cosentino, S.; Vestergaard, M.; Rasmussen, S.; Lund, O.; Aarestrup, F. M.; Larsen, M. V., 2012. Identification of acquired antimicrobial resistance genes. J. Antimicrob. Chemother. 67:2640-2644.

[119] Gupta, S. K.; Padmamabhan, B. R.; Diene, S. M.; Lopez-Rojas, R.; Kempf, M.; Landraud, L., 2014. ARG-ANNOT, a New Bioinformatic Tool To Discover Antibiotic Resistance Genes in Bacterial Genomes. Antimicrobial Agents and Chemotherapy 58(1), 212-220.

[120] Rowe, W.; Baker, K. S.; Verner-Jeffreys, D.; Baker-Austin, C.; Ryan, J. J.; Maskell, D.; Pearce, G., 2015. Search Engine for Antimicrobial Resistance: A Cloud Compatible Pipeline and Web Interface for Rapidly Detecting Antimicrobial Resistance Genes Directly from Sequence Data. PLoS ONE 10(7):e0133492.

[121] Bortolaia, V.; Kaas, R. S.; Ruppe, E.; Roberts, M. C.; Schwarz, S. et al., 2020. ResFinder 4.0 for predictions of phenotypes from genotypes. J. Antimicrob. Chemother. 75:3491-3500.

[122] Silva, R.; Padovani, K.; Góes, F.; Alves, R.; 2020. geneRFinder: gene finding in distinct metagenomic data complexities. bioRxiv, doi: https://doi.org/10.1101/2020.08.21.262147.

[123] https://groot-documentation.readthedocs.io/en/latest/, 28/11/2020

[124] Huson, D. E.; Beier, S.; Flade, I.; Górska, A.; El-Hadidi, M.; Mitra, S.; Ruscheweyh, H-J.; Tappu, R., 2016. MEGAN Community Edition – Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput. Biol. 12(6).

[125] https://github.com/husonlab/megan-ce, 31/12/2020

6. Anexos

Tabla A1: Listado de metagenomas seleccionados en MG-RAST

Project id	Metagenome id
mgp89870	mgm4850709.3
mgp89870	mgm4850710.3
mgp89833	mgm4849831.3
mgp89723	mgm4848386.3
mgp89696	mgm4847936.3
mgp80207	mgm4742545.3
mgp80207	mgm4742547.3
mgp80207	mgm4742537.3
mgp80207	mgm4742543.3
mgp80207	mgm4742541.3
mgp80207	mgm4742540.3
mgp80207	mgm4742536.3
mgp80207	mgm4743981.3
mgp80207	mgm4743979.3
mgp80207	mgm4749377.3
mgp80207	mgm4743980.3
mgp80207	mgm4742538.3
mgp80207	mgm4749376.3
mgp80207	mgm4749374.3
mgp80207	mgm4749373.3
mgp80207	mgm4749375.3
mgp80207	mgm4742544.3
mgp80207	mgm4742542.3
mgp80101	mgm4741983.3
mgp80101	mgm4741982.3
mgp6843	mgm4544117.3
mgp6843	mgm4544118.3
mgp6843	mgm4544116.3
mgp6843	mgm4544115.3
mgp6843	mgm4559244.3
mgp6843	mgm4559243.3
mgp45	mgm4446153.3
mgp3997	mgm4479944.3
mgp3997	mgm4479361.3
mgp3456	mgm4513787.3
mgp21223	mgm4729903.3
mgp18601	mgm4704716.3
mgp18601	mgm4704721.3
mgp12953	mgm4622704.3

Project id	Metagenome id
mgp5435	mgm4565311.3
mgp5435	mgm4556395.3
mgp5435	mgm4549531.3
mgp5435	mgm4549526.3
mgp5435	mgm4544128.3
mgp5435	mgm4544129.3
mgp5435	mgm4533996.3
mgp5435	mgm4604928.3
mgp5435	mgm4587465.3
mgp5435	mgm4570096.3
mgp5435	mgm4565313.3
mgp5435	mgm4556404.3
mgp5435	mgm4549530.3
mgp5435	mgm4538739.3
mgp5435	mgm4538734.3
mgp5435	mgm4587467.3
mgp5435	mgm4587458.3
mgp5435	mgm4544135.3
mgp5435	mgm4544125.3
mgp5435	mgm4538740.3
mgp5435	mgm4538738.3
mgp5435	mgm4533997.3
mgp5435	mgm4534000.3
mgp5435	mgm4530681.3
mgp5435	mgm4530682.3
mgp5435	mgm4604927.3
mgp5435	mgm4587468.3
mgp5435	mgm4587462.3
mgp5435	mgm4587455.3
mgp5435	mgm4549527.3
mgp5435	mgm4544137.3
mgp5435	mgm4544132.3
mgp5435	mgm4544131.3
mgp5435	mgm4538742.3
mgp5435	mgm4538733.3
mgp5435	mgm4534001.3
mgp5435	mgm4532159.3
mgp5435	mgm4604932.3
mgp5435	mgm4587469.3

mgp12953	mgm4622705.3
mgp12953	mgm4622707.3
mgp12953	mgm4622706.3
mgp12953	mgm4622702.3
mgp12953	mgm4622703.3
mgp12932	mgm4622490.3
mgp12932	mgm4622488.3
mgp12932	mgm4622489.3
mgp12931	mgm4622485.3
mgp12931	mgm4622486.3
mgp12931	mgm4622487.3
mgp12908	mgm4622371.3
mgp12866	mgm4622205.3
mgp12864	mgm4622372.3
mgp12864	mgm4622200.3
mgp12863	mgm4622198.3
mgp12863	mgm4622199.3
mgp5435	mgm4556400.3
mgp5435	mgm4556398.3
mgp5435	mgm4544133.3
mgp5435	mgm4538732.3
mgp5435	mgm4530680.3
mgp5435	mgm4604933.3
mgp5435	mgm4556403.3
mgp5435	mgm4549532.3
mgp5435	mgm4544127.3
mgp5435	mgm4538736.3
mgp5435	mgm4533998.3
mgp5435	mgm4533999.3
mgp5435	mgm4532702.3
mgp5435	mgm4587456.3
mgp5435	mgm4570097.3
mgp5435	mgm4565314.3
mgp5435	mgm4556405.3
mgp5435	mgm4549533.3
mgp5435	mgm4549528.3
mgp5435	mgm4544136.3
mgp5435	mgm4533995.3
mgp5435	mgm4532163.3
mgp5435	mgm4532162.3
mgp5435	mgm4604929.3

mgp5435	mgm4587463.3
mgp5435	mgm4587457.3
mgp5435	mgm4577137.3
mgp5435	mgm4556407.3
mgp5435	mgm4556402.3
mgp5435	mgm4556396.3
mgp5435	mgm4544126.3
mgp5435	mgm4544124.3
mgp5435	mgm4533994.3
mgp5435	mgm4532703.3
mgp5435	mgm4604926.3
mgp5435	mgm4587466.3
mgp5435	mgm4587464.3
mgp5435	mgm4587459.3
mgp5435	mgm4587454.3
mgp5435	mgm4587453.3
mgp5435	mgm4570098.3
mgp5435	mgm4556409.3
mgp5435	mgm4556406.3
mgp5435	mgm4556399.3
mgp5435	mgm4549529.3
mgp5435	mgm4538741.3
mgp5435	mgm4538737.3
mgp5435	mgm4538735.3
mgp5435	mgm4532161.3
mgp5435	mgm4587461.3
mgp5435	mgm4587460.3
mgp5435	mgm4570099.3
mgp5435	mgm4565312.3
mgp5435	mgm4556410.3
mgp5435	mgm4556408.3
mgp5435	mgm4556401.3
mgp5435	mgm4556397.3
mgp5435	mgm4549534.3
mgp5435	mgm4549525.3
mgp5435	mgm4544134.3
mgp5435	mgm4544130.3
mgp5435	mgm4538743.3
mgp5435	mgm4532160.3
mgp5435	mgm4530683.3
mgp5435	mgm4530679.3