

# Comparación de estrategias metagenómicas en muestras ambientales de biofilms

---

*Trabajo Final de Grado (TFG)*

- » **Autora:** Paula Izquierdo Altarejos
- » **Curso académico:** 4º
- » **Tutor empresa:** Fernando González Candelas
- » **Tutora universidad:** Lynne Paula Yenush
- » **Fecha de presentación:** 28-VIII-2015
- » **Localidad:** Valencia (España)
- » **Empresa/institución:** Centro Superior de Investigación en Salud Pública (CSISP) – Fundación para la Investigación Sanitaria y Biomédica en la Comunidad Valenciana (FISABIO)
  - Departamento de Genómica y Salud

**Grado en Biotecnología**

**Universitat Politècnica de València (UPV)**

**Escuela Técnica Superior de Ingeniería Agronómica y del Medio Natural (ETSIAMN)**



UNIVERSITAT  
POLITÈCNICA  
DE VALÈNCIA



CENTRO SUPERIOR DE  
INVESTIGACIÓN EN  
SALUD PÚBLICA



Fundación para el Fomento de la  
Investigación Sanitaria y Biomédica  
de la Comunitat Valenciana

## Formulario para el depósito en RiuNet

AUTORA	1 <sup>er</sup> APELLIDO	2 <sup>o</sup> APELLIDO	NOMBRE	DNI/NIE
	Izquierdo	Altarejos	Paula	21704753K
TUTORA ACADÉMICA	1 <sup>er</sup> APELLIDO	2 <sup>o</sup> APELLIDO	NOMBRE	DNI/NIE
	Yenush		Lynne Paula	X2494558
COTUTOR	1 <sup>er</sup> APELLIDO	2 <sup>o</sup> APELLIDO	NOMBRE	DNI/NIE
	González	Candelas	Fernando	22636751J
UNIVERSIDAD	Universidad Politécnica de Valencia (UPV)			
CENTRO	Escuela Técnica Superior de Ingeniería Agronómica y del Medio Natural (ETSIAMN)			
TITULACIÓN	Grado en Biotecnología			
LUGAR Y FECHA DE PRESENTACIÓN	Valencia, a 28 de agosto de 2015			
TÍTULO DEL TFG	Comparación de estrategias metagenómicas en muestras ambientales de biofilms			
KEY WORDS	Metagenomics, biofilms, 16S gene, ribosomal RNA, V1-V4 hypervariable regions, sequencing, universal primers, bacterial community, taxonomic assignment, next-generation sequencing technologies (NGS), pathogenic agents, <i>Legionella pneumophila</i>			
PALABRAS CLAVE	Metagenómica, biofilms, gen 16S, ARN ribosomal, regiones hipervariables V1-V4, secuenciación, cebadores universales, población bacteriana, asignación taxonómica, tecnologías de secuenciación masiva (NGS), agentes patógenos, <i>Legionella pneumophila</i>			
ABSTRACT	<p>The main purposes of this study are to characterize the bacterial community present in environmental samples of biofilms and compare the results obtained using three different approaches: metagenomics without prior culturing using NGS technologies, active search of 16S ARNr regions within the metagenomes obtained and sequencing of specific regions of the 16S rRNA from colonies grown in culture. The analyzed samples were obtained in natural springs in the region of Alcoy (Alicante, Spain) and they show epidemiological interest due to the known endemism of the pathogenic bacteria <i>Legionella pneumophila</i> in this area.</p> <p>The experimental workflow of the project involved the analysis of individual colonies grown on GVPC medium using DNA extraction, amplification of the V1-V4 hypervariable regions with universal 16S rRNA primers, amplicons purification, Sanger sequencing and sequence analysis, using bioinformatic tools and databases focused on the ribosomal RNA. This methodology allowed assigning most sequences to the genus level, but the resolution was not enough for the accurate identification of species, except for a low number of the colonies.</p> <p>According to the 16S sequencing, the bacterial community studied is made up of 3 phyla, 6 classes, 13 orders, 19 families and 26 genera. Within the main phyla, the community was found to be dominated by <i>Proteobacteria</i>, with a relative abundance greater than 80 %, especially the <i>Gammaproteobacteria</i> class, representing 70 % of the total population. Regarding the order and family levels, <i>Aeromonadales</i> and <i>Aeromonadaceae</i> stood out, respectively, both with a relative abundance greater than 50 %. At the genus level, the abundance of <i>Aeromonas</i> (53'61 %) was remarkable, pointing out the possibility that the culture medium used (GVPC) somehow improves their growth.</p>			

	<p>About the presence of pathogens, the genus <i>Legionella</i> was not detected despite its endemism in the area. Nevertheless, it was found a surprisingly high percentage of pathogenic related genera.</p> <p>In the comparison between the community characterized according to the regions of 16S rRNA and the distribution obtained with the metagenomic approach by NGS only minor similarities or just at general levels were found, but at the genus level the results were very different. According to the data generated from the 16S sequences retrieved from the metagenomes, no strong correlation was found. Differences and biases are likely due to the sample size (in the 16S study only a fraction of the community was analyzed), the step of culturing on GVPC medium and the databases in which each approach rely.</p>
RESUMEN	<p>Los objetivos principales del trabajo son caracterizar la población bacteriana presente en muestras ambientales de biofilms y comparar los resultados obtenidos mediante tres técnicas: secuenciación de metagenomas completos antes del cultivo, búsqueda activa de las regiones del 16S en estos metagenomas y secuenciación de regiones específicas del 16S ARNr a partir de las colonias crecidas en cultivo. Las muestras empleadas proceden de fuentes naturales de la región de Alcoy (Alicante, España) y presentan interés epidemiológico debido al conocido endemismo de la bacteria patógena <i>Legionella pneumophila</i> en esta zona.</p> <p>La parte experimental del proyecto consistió en el análisis individual de las colonias cultivadas en medio GVPC, empleando extracción del material genético, amplificación de las regiones hipervariables V1-V4 del 16S ARNr con cebadores universales, purificación de los amplicones, secuenciación mediante la técnica Sanger y análisis de las secuencias, apoyándose en herramientas bioinformáticas y bases de datos fundamentadas en el ARN ribosomal. Esta metodología permitió asignar la mayoría de las secuencias hasta el nivel de género, pero la resolución no fue suficiente para la identificación precisa de especies, excepto en un porcentaje minoritario.</p> <p>Basándose en la secuenciación del 16S, la población bacteriana estudiada está constituida por 3 filos, 6 clases, 13 órdenes, 19 familias y 26 géneros. Entre los filos encontrados predomina <i>Proteobacteria</i> con una abundancia relativa superior al 80 %, concretamente la clase <i>Gammaproteobacteria</i>, que representa un 70 % de la población total. En cuanto a órdenes y familias, destacan <i>Aeromonadales</i> y <i>Aeromonadaceae</i>, respectivamente, ambos con una abundancia relativa superior al 50 %. Respecto a los géneros, destaca la abundancia de <i>Aeromonas</i> (53'61 %), indicando la posibilidad de que el medio de cultivo empleado (GVPC) favorezca su crecimiento.</p> <p>En cuanto a la presencia de agentes patógenos, no se detectó el género <i>Legionella</i>, endémico de la región de procedencia de las muestras. En cambio, un elevado porcentaje de los géneros encontrados presenta relación con agentes patógenos.</p> <p>Al comparar la población obtenida en función de las regiones del 16S ARNr con la derivada de la estrategia metagenómica mediante NGS solo se encontraron similitudes puntuales o a niveles generales, pero centrándose en la distribución por géneros los datos no resultan comparables. Tampoco se encontró correlación con los datos generados a partir de la extracción de las secuencias de 16S de los metagenomas y su búsqueda en bases de datos. Los sesgos y diferencias encontrados se deben principalmente al tamaño muestral (por 16S se analizó una fracción de la población total de las muestras), la etapa de cultivo de las muestras en medio GVPC y las bases de datos empleadas en cada caso.</p>
TIPO DE LICENCIA	Licencia Creative Commons "Reconocimiento-No Comercial- Sin Obra Derivada"

Gracias a todos los que han contribuido a este trabajo. En primer lugar, a Fernando, mi tutor, por sus útiles comentarios y explicaciones, así como por confiar siempre en mí. También a Leo, ahora ya la Dra. Sánchez, que me ayudó desde el principio y de cuya tesis surgió este trabajo. He de incluir a Paula Ruiz, por enseñarme el funcionamiento del laboratorio y por estar siempre dispuesta a atender mis consultas.

Sin olvidar a mi familia, que ya se está familiarizando con el mundo de la Biotecnología; y a mis amigos, por su apoyo y por ayudarme a que me despejara cuando era necesario.

# ÍNDICE

---

<b>1. INTRODUCCIÓN.....</b>	<b>I</b>
1.1. Metagenómica mediante NGS.....	II
1.2. Metagenómica basada en el gen 16S ARNr.....	II
1.3. Biofilms.....	IV
1.4. Antecedentes.....	IV
1.5. Perspectivas futuras.....	V
<b>2. OBJETIVOS.....</b>	<b>VII</b>
<b>3. MATERIALES Y MÉTODOS.....</b>	<b>VIII</b>
3.1. Protocolo experimental.....	VIII
3.1.1. Naturaleza y preparación de las muestras.....	VIII
3.1.2. Amplificación por PCR y purificación.....	IX
3.1.3. Secuenciación.....	X
3.2. Análisis bioinformático.....	XI
3.2.1. Tratamiento de los cromatogramas.....	XI
3.2.2. Asignación taxonómica.....	XI
3.2.3. Identificación de géneros potencialmente patógenos.....	XI
3.3. Visión general del flujo de trabajo.....	XII
<b>4. RESULTADOS Y DISCUSIÓN.....</b>	<b>XIII</b>
4.1. Asignación taxonómica.....	XIII
4.1.1. Visión global.....	XIII
4.1.2. Identificación de especies.....	XV
4.2. Descripción de la población bacteriana.....	XVII
4.2.1. Taxones presentes y abundancias relativas.....	XVII
4.2.2. Identificación de patógenos.....	XVIII
4.2.3. Distribución de las poblaciones por muestra.....	XX
4.3. Comparación de resultados.....	XXII
4.3.1. Sesgos esperados <i>a priori</i> .....	XXII
4.3.2. Similitudes y diferencias entre la aproximación metagenómica y el análisis basado en el gen 16S.....	XXIII
4.3.3. Correlación estadística.....	XXVIII
4.4. Evaluación del método experimental y propuestas de optimización.....	XXIX
<b>5. CONCLUSIONES.....</b>	<b>XXXI</b>
<b>6. REFERENCIAS BIBLIOGRÁFICAS.....</b>	<b>XXXIII</b>
<b>7. ANEXOS.....</b>	<b>XXXVII</b>

## Índice de tablas

---

### MATERIALES Y MÉTODOS

**Tabla 3.1.** Secuencia y temperatura de fusión o *melting* ( $T_m$ ) de los primers empleados..... IX

**Tabla 3.2.** Volumen de reactivos utilizados por cada reacción de PCR..... IX

**Tabla 3.3.** Programa de temperaturas empleado en las reacciones de PCR..... IX

**Tabla 3.4.** Volumen de reactivos utilizados para cada reacción de secuenciación..... X

**Tabla 3.5.** Ciclos de temperatura empleados para la reacción de secuenciación por Sanger..... X

### RESULTADOS Y DISCUSIÓN

**Tabla 4.1.** Asignación taxonómica obtenida al contrastar las secuencias del 16S ARNr con la base de datos RDP. Se ha incluido el número de colonias totales asignadas a cada género y algunos ejemplos de especies con las que se ha encontrado mayor identidad ( $S_{ab\ score} > 97\ %$ ), indicando entre paréntesis el identificador de acceso a la secuencia en la base donde se encuentra depositada. En caso de existir más especies candidatas que superan el umbral, se ha indicado mediante un “[+]”..... XIII

**Tabla 4.2.** Listado de especies con las que se ha encontrado una sola coincidencia al buscar las secuencias problema, aplicando tres umbrales de identidad basados en el parámetro  $S_{ab\ score}$  (97, 99 y 100 %) y abundancia relativa de las especies identificadas en la población estudiada. Se indican entre paréntesis los identificadores de acceso a las secuencias de la base de datos..... XVI

**Tabla 4.3.** Listado de géneros identificados en las muestras que contienen especies patógenas según la clasificación de la Organización Mundial de la Salud. Se ha indicado la abundancia relativa de cada género en las muestras, las especies consideradas como patógenos humanos y animales dentro de cada género y el grupo de riesgo en el que se engloban..... XIX

**Tabla 4.4.** Géneros con abundancia relativa superior al 1 % detectados en el análisis del 16S, ordenados en función de su abundancia. Se indica la abundancia obtenida en las estrategias de NGS (metagenomas), análisis del 16S y la basada en las secuencias del 16S extraídas de los metagenomas. Se han destacado los géneros con frecuencia de aparición  $> 1\ %$  detectados en dos o más de las estrategias. Se incluye también el número de secuencias depositadas en la base RPD para cada género y el número de éstas que proceden de aislados..... XXV

**Tabla 4.5.** Géneros con abundancia relativa superior al 1 % detectados en el análisis de los metagenomas, ordenados en función de su abundancia. Se indica la abundancia obtenida en las estrategias de NGS (metagenomas), análisis del 16S y la basada en las secuencias del 16S extraídas de los metagenomas. Se han destacado los géneros con frecuencia de aparición  $> 1\ %$  detectados en dos o más de las estrategias. Se incluye también el número de secuencias depositadas en la base RPD para cada género y el número de éstas que proceden de aislados..... XXVII

**Tabla 4.6.** Registro del número de colonias picadas por cada muestra inicial, el número de amplicones tras la PCR de colonia (amplificado I), tras el choque térmico (amplificado II) y totales, así como el número de secuencias consenso finalmente obtenidas. Se incluyen dos rendimientos: el rendimiento I hace referencia al ratio de secuencias obtenidas en función del número de colonias picadas, mientras que el rendimiento II describe el ratio de secuencias

obtenidas en función del número de amplicones generados. Se incluyen en la última fila los valores referidos al cómputo global de las 14 muestras..... XXX

# Índice de figuras

---

## INTRODUCCIÓN

**Figura 1.1.** Diagrama del protocolo básico de un estudio metagenómico, señalando las dos estrategias principales posibles (secuenciación de genomas completos y secuenciación dirigida). *Adaptado de* <http://medicalplanzone.com/bacterial-artificial-chromosome/>..... I

**Figura 1.2.** Representación esquemática del gen de 16S ARNr, indicando las regiones variables (V1-V9), las conservadas (C1-C9) y la invariable, comunes a todas las bacterias. *Elaboración propia*..... III

## MATERIALES Y MÉTODOS

**Figura 3.1.** Distribución geográfica de las fuentes de las que se tomaron las muestras, señaladas con triángulos rojos en el mapa. *Cedido por la Dra. Sánchez Busó*..... VIII

**Figura 3.2.** Diagrama esquemático del procedimiento de trabajo aplicado, destacando las tres estrategias que se han comparado y el diseño experimental realizado en el laboratorio en el análisis del 16 ARNr. *Elaboración propia*..... XII

## RESULTADOS Y DISCUSIÓN

**Figura 4.1.** Ejemplo del resultado de la búsqueda para una de las secuencias (colonia BCol\_10). Se muestran las 20 secuencias de mayor coincidencia encontradas, incluyendo sus identificadores, tres parámetros de la búsqueda – entre ellos el *S<sub>ab</sub> score* (resaltado en naranja)– y los nombres completos de las secuencias..... XV

**Figura 4.2.** Abundancias relativas de los géneros detectados en el conjunto de las 14 muestras. En el apartado A se muestran los géneros con una frecuencia de aparición de al menos el 1 %, así como el porcentaje que no ha podido ser asignado a este nivel, y en el B aparecen los géneros con abundancias relativas inferiores al 1 % (la escala del eje se ha ajustado en cada caso para su correcta visualización)..... XVIII

**Figura 4.3.** Abundancias relativas de las clases (A) y géneros (B) detectados en las 14 muestras..... XX

**Figura 4.4.** Gráficas obtenidas al representar los valores de las abundancias relativas acumuladas para dos muestras – A) N9 y B) N10– frente a los valores del conjunto. Los datos se ajustaron a una recta de regresión lineal, indicando el valor del coeficiente  $R^2$ ..... XXI

**Figura 4.5.** Gráficas obtenidas al representar los valores de las abundancias relativas de los géneros obtenidas por un método frente a los valores obtenidos por otra de las aproximaciones. Los datos se ajustaron a una recta de regresión lineal, indicando el valor del coeficiente  $R^2$ . Se compararon las estrategias dos a dos: A) Metagenomas y secuenciación del 16S; B) Secuenciación del 16S y búsqueda de este gen en los genomas; y C) Metagenomas y búsqueda del 16S en ellos..... XXVIII

## ANEXOS

**Figura 7.1.** Representación de la placa stock correspondiente a las muestras N2, N3 y N4, indicando la nomenclatura empleada para designar las colonias picadas..... XXXVII

**Figura 7.2.** Representación de la placa stock correspondiente a las muestras N1 y N5, indicando la nomenclatura empleada para designar las colonias picadas..... XXXVII



**Figura 7.3.** Representación de la placa stock correspondiente a las muestras N6 y N7, indicando la nomenclatura empleada para designar las colonias picadas..... XXXVII

**Figura 7.4.** Representación de la placa stock correspondiente a las muestras N8 y N9, indicando la nomenclatura empleada para designar las colonias picadas..... XXXVIII

**Figura 7.5.** Representación de la placa stock correspondiente a las muestras N10 y N11, indicando la nomenclatura empleada para designar las colonias picadas..... XXXVIII

**Figura 7.6.** Representación de la placa stock correspondiente a las muestras N23 y N24, indicando la nomenclatura empleada para designar las colonias picadas..... XXXVIII

**Figura 7.7.** Representación de la placa stock correspondiente a la muestra N25, indicando la nomenclatura empleada para designar las colonias picadas..... XXXVIII

**Figura 7.8.** Resultado de la electroforesis de los productos de PCR obtenidos a partir de cinco controles positivos (dos de ellos procedentes de cultivo mixto y los otros tres obtenidos de cultivo puro), incluyendo un marcador de pesos moleculares en el primer carril (GeneRuler 1 kb) y un control negativo en el séptimo carril. Las bandas aparecen a una altura correspondiente a las 750 pb, aproximadamente..... XXXIX

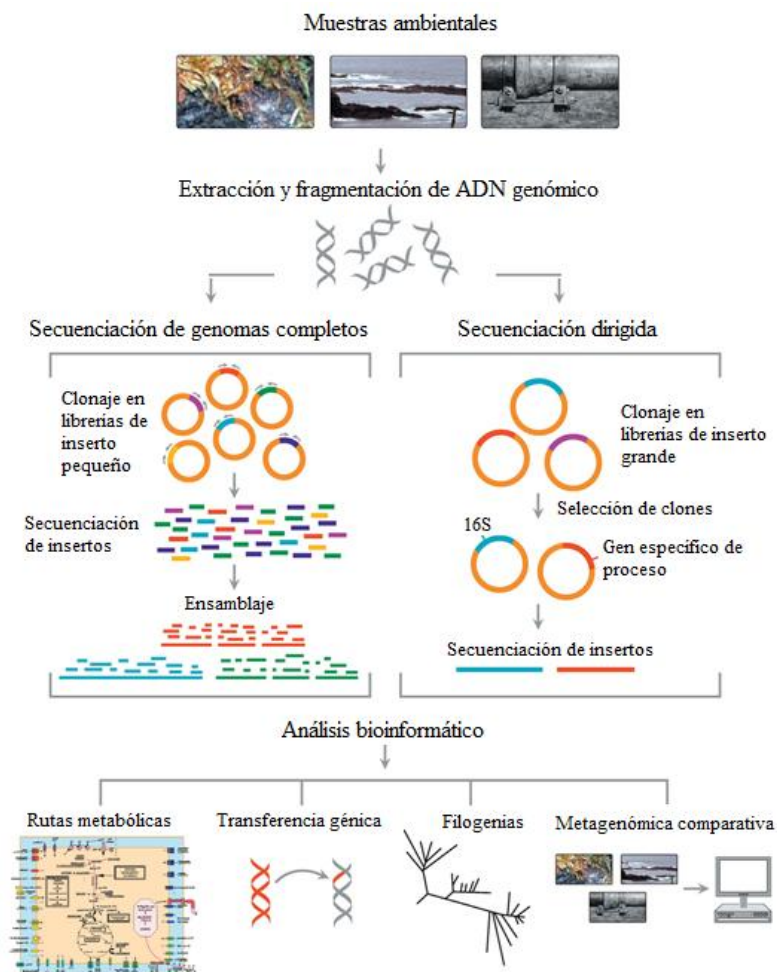
## 1. INTRODUCCIÓN

Tradicionalmente, la **genómica** se ha centrado en secuenciar y analizar los genomas de organismos concretos. Aplicada al estudio de poblaciones microbianas, implica la necesidad de obtener cultivos puros de los microorganismos que se quieren estudiar, para posteriormente extraer y secuenciar el ADN de cada uno de ellos<sup>58</sup>.

La **metagenómica**, en cambio, consiste en estudiar directamente el conjunto de genomas de una muestra, eludiendo el requisito previo del cultivo<sup>63, 65</sup> e identificando las especies presentes *a posteriori*. Se estima que más del 99 % de los microorganismos no son cultivables<sup>54</sup>, por lo que la metagenómica supone una alternativa para estudiar este amplio porcentaje de las comunidades microbianas que ha permanecido fuera del alcance de la investigación durante años.

Existen distintas aproximaciones desde las que abordar un proyecto de metagenómica, que pueden englobarse en: secuenciación de genomas completos (*Whole Genome Sequencing* o WGS) y secuenciación dirigida a zonas concretas del genoma, como se refleja en la Figura 1.1.

En el presente estudio se han comparado los datos metagenómicos obtenidos al aplicar una estrategia de tipo WGS y otra dirigida al gen 16S del ARN ribosomal en muestras ambientales de biofilms procedentes de la región de Alcoy (Alicante, España), de interés epidemiológico por el conocido endemismo de la bacteria patógena *Legionella pneumophila* en esta zona.



**Figura 1.1.** Diagrama del protocolo básico de un estudio metagenómico, señalando las dos estrategias principales posibles (secuenciación de genomas completos y secuenciación dirigida). Adaptado de <http://medicalplanzone.com/bacterial-artificial-chromosome/>

## 1.1 Metagenómica mediante NGS

El objetivo de esta aproximación es la caracterización de la muestra a estudio a través de la identificación de los organismos presentes en la misma gracias a la presencia de secuencias particulares que permitan dicha identificación. Para ello, se extrae el ADN total directamente de la muestra y se fragmenta. Se emplea un abordaje de tipo *shotgun*: se secuencian los fragmentos empleando tecnologías de secuenciación masiva y posteriormente se ensamblan para reconstruir los distintos genomas<sup>58</sup>, obteniéndose generalmente genomas parciales. Antes de la implementación de las técnicas de secuenciación masiva era necesario construir una librería de ADN mediante clonaje de forma previa a la secuenciación; la introducción de las NGS ha permitido evitar la etapa de clonaje, ya que posibilita secuenciar directamente los fragmentos de ADN de la muestra mediante la ligación de adaptadores y las genotecas se generan *in vitro*.

Existen distintos estudios en los que se ha empleado esta técnica para estudiar la composición de especies y la diversidad de poblaciones bacterianas, como los realizados por Tyson *et al.* en 2004<sup>59</sup> y Lu *et al.* en 2015<sup>32</sup>, entre otros.

### *Ventajas y limitaciones*

La principal ventaja de esta metodología es la gran cantidad de información que se puede extraer de los resultados generados, puesto que analiza el conjunto de genomas presentes en la muestra. De esta forma, no sólo permite la identificación taxonómica y la caracterización de la diversidad microbiana, sino que también proporciona información de las funciones potenciales – por ejemplo, rutas metabólicas– presentes en las comunidades microbianas. Además, las técnicas de NGS han permitido reducir los costes de secuenciación, posibilitando proyectos a gran escala como es el caso del *Earth Microbiome Project*<sup>17</sup>, iniciado en 2010, con el objetivo de elaborar un catálogo global de la diversidad microbiana del planeta a partir del estudio de más de 200.000 muestras ambientales.

No obstante, la técnica también presenta varias limitaciones: generalmente, la profundidad de secuenciación (cobertura) no es suficiente para detectar las especies menos representadas en la población<sup>23</sup> y las abundancias relativas que se obtienen se ven altamente influenciadas por los protocolos de extracción y secuenciación del ADN seguidos<sup>35</sup>.

Por otra parte, la calidad y la longitud de las secuencias obtenidas mediante técnicas de secuenciación masiva son menores que empleando el método Sanger, lo que resulta un inconveniente en los proyectos de metagenómica, en los que se desconoce *a priori* qué especies se están secuenciando. Los errores de secuenciación pueden llevar al incorrecto ensamblaje de ciertas secuencias y complicar la identificación de los organismos presentes, mientras que la limitación en la longitud de las secuencias que es posible obtener supone una dificultad a la hora de ensamblarlas informáticamente y realizar una anotación precisa<sup>38</sup>.

## 1.2 Metagenómica basada en el gen 16S ARNr

En la década de los 70, Woese propuso por primera vez que los ARNs de los ribosomas podían actuar como cronómetros moleculares universales<sup>15</sup>. Este planteamiento se desarrolló posteriormente y condujo al establecimiento del ARN ribosomal como el marcador filogenético más frecuente y ampliamente utilizado.

El gen 16S ARNr se ha empleado comúnmente para tal fin desde los 90<sup>36</sup> por las siguientes razones:

- ✓ Se encuentra presente en todas las especies de los dominios *Bacteria* y *Archaea* (para eucariotas se emplea el 18S ARNr).
- ✓ Los genes para ARN ribosomal contienen regiones más conservadas que pueden emplearse para diseñar cebadores de PCR de amplio espectro (teóricamente, universales) y regiones altamente variables útiles para la discriminación entre microorganismos y clasificación taxonómica<sup>26</sup>.

- ✓ Es el único marcador taxonómico lo suficientemente informativo del que se disponen bases de datos con un amplio número de secuencias depositadas (superior a los 4 millones<sup>67</sup>) y de calidad contrastada<sup>39, 66</sup>.

En este tipo de estudios, el ADN total se extrae directamente de la muestra y se amplifican por PCR los genes 16S del conjunto de genomas<sup>67</sup>. A continuación pueden seguirse distintos protocolos, entre los que destaca la clonación de los productos de PCR para la construcción de genotecas y posterior secuenciación de los insertos<sup>62</sup>. Para la secuenciación pueden emplearse técnicas de secuenciación masiva o bien el método Sanger, este último especialmente para proyectos a menor escala.

También existe la posibilidad de estudiar el 16S sin amplificación previa, construyendo una librería inicial con fragmentos aleatorios de los genomas en vectores que permitan un mayor tamaño de inserto, de la que se seleccionan los insertos de interés mediante sondas complementarias a regiones del 16S y se subclonan en vectores de menor tamaño<sup>45</sup>.

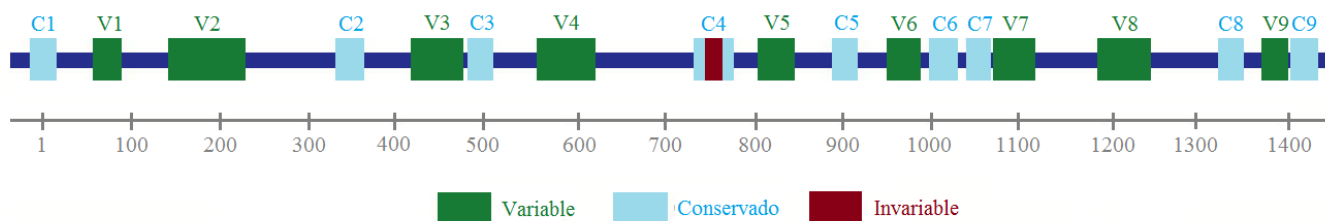
Debe tenerse en cuenta que la secuenciación del 16S puede estar sesgada por la amplificación diferencial de los genes de ARNr de cada organismo. Aunque se empleen cebadores “universales”, éstos se unen con mayor afinidad a las secuencias de ciertas especies, en función de la complementariedad de bases. También pueden existir secuencias flanqueantes que interfieran con el sitio de unión al primer, disminuyendo la eficiencia de la amplificación<sup>20</sup>, por lo que la librería puede ser poco representativa, principalmente a nivel cuantitativo<sup>12</sup>.

En este estudio se ha propuesto una estrategia a medio camino entre la metagenómica y la genómica clásica, consistente en amplificar el 16S a partir de colonias de microorganismos en cultivo, lo cual se espera que corrija el sesgo generado por la amplificación diferencial. De esta forma, aunque una colonia determinada dé lugar a un mayor número de amplicones, no va a resultar en una mayor abundancia relativa de la especie en sí.

La búsqueda de las secuencias obtenidas en bases de datos específicas para el 16S ARNr, como RDP (<http://rdp.cme.msu.edu/>) o SILVA (<http://www.arb-silva.de/>) permite clasificar taxonómicamente los microorganismos presentes en la muestra.

### Regiones del 16S ARNr

La variabilidad genética entre secuencias del ARN ribosomal 16S no se encuentra distribuida de forma homogénea en toda la extensión del gen, sino que se concentra en nueve regiones hipervariables (V1-V9)<sup>50</sup>, como se muestra en la Figura 1.2. Además, algunas de estas regiones presentan mayor variabilidad que otras<sup>68, 69</sup> y permiten obtener una asignación taxonómica más precisa<sup>31, 61</sup>.



**Figura 1.2.** Representación esquemática del gen de 16S ARNr, indicando las regiones variables (V1-V9), las conservadas (C1-C9) y la invariable, comunes a todas las bacterias. *Elaboración propia.*

Según los estudios realizados por Kim y Morrison<sup>25</sup>, las regiones V1-V3 y V1-V4 presentan dos ventajas:

- » Debido a su elevada variabilidad, proporcionan una mayor capacidad resolutive para distinguir entre especies.
- » Existe un mayor número de secuencias parciales correspondientes a las regiones V1-V4 contenidas en bases de datos como la RDP, por lo que analizar estas regiones facilita los análisis taxonómicos y filogenéticos.

Además, dado que sólo el 23 % de las secuencias de 16S ARNr depositadas en las bases tienen una longitud mayor de las 900 pares de bases<sup>67</sup>, secuenciar el gen completo no resulta en una mejora de los resultados tan significativa como se podría pensar en un principio. Yarza *et al.*<sup>67</sup> concluyen que secuenciar los 750 primeros nucleótidos del gen es suficiente para discriminar más del 90 % de la diversidad a nivel de familia, género y especie; mientras que al secuenciar fragmentos de menor tamaño la información es insuficiente incluso para obtener resultados aceptables a nivel de género.

### 1.3 Biofilms

Uno de los tipos de muestra que permite analizar la metagenómica son los **biofilms**, consistentes en comunidades microbianas adheridas a superficies que adquieren unas características biológicas y físicas determinadas. En los biofilms, el desarrollo microbiano se ve favorecido por la mayor captación de nutrientes<sup>46</sup>, que quedan adsorbidos en la superficie de la comunidad; la proximidad de los microorganismos<sup>34</sup>, que facilita interacciones metabólicas y genéticas como la transferencia de ácidos nucleicos<sup>6, 21, 28</sup>; y la mayor resistencia a agentes antimicrobianos, debida a la sinergia entre especies<sup>3</sup>.

Se ha visto que en la mayoría de ambientes naturales, los microorganismos suelen agregarse y crecer en estas estructuras<sup>8, 34</sup>, siendo especialmente comunes en ríos<sup>9</sup> y canalizaciones<sup>33</sup>, ya que la adhesión a la superficie permite a los microorganismos establecerse en estos ambientes a pesar de la tensión creada por el flujo del agua circulante.

### 1.4 Antecedentes

Las investigaciones en metagenómica más relevantes se han centrado principalmente en el estudio de ecosistemas marinos, como es el caso del proyecto a escala global llevado a cabo por el consorcio *TARA Oceans*<sup>24</sup> y del estudio realizado por Venter *et al.* en el Mar de los Sargazos<sup>60</sup>, o bien en comunidades microbianas asociadas al cuerpo humano, como en el Proyecto Microbioma Humano<sup>27</sup>.

Sin embargo, en la literatura también pueden encontrarse diversos estudios enfocados a muestras de biofilms. El primer proyecto a gran escala analizó la población microbiana de muestras ambientales de biofilm en una mina ácida<sup>59</sup>. Este estudio empleó la estrategia WGS mediante secuenciación de tipo *shotgun* por el método de Sanger y generó 78 millones de bases nucleotídicas a partir de las cuales fue posible reconstruir varios genomas de los principales organismos presentes.

Centrándonos en el estudio de biofilms de la red de agua potable (dado que este el tipo de muestra analizado en el trabajo), se mencionan a continuación algunos de los ejemplos encontrados en la literatura.

En el estudio de Schmeisser *et al.*<sup>45</sup> se aplicaron distintas estrategias para caracterizar muestras de biofilm, incluyendo el análisis del 16S ARNr y la secuenciación de fragmentos aleatorios clonando previamente en vectores. Los datos obtenidos indicaban que la comunidad microbiana estaba dominada por microorganismos no cultivables pertenecientes al filo *Proteobacteria*.

En la revisión de Gómez-Álvarez sobre metagenómica comparativa<sup>18</sup> se describen distintos protocolos para el análisis taxonómico y funcional de poblaciones microbianas que crecen en biofilms y están asociadas a la corrosión de canalizaciones de aguas residuales. Se realizó un estudio mediante secuenciación por NGS en este tipo de muestras, centrándose en la detección de microorganismos y rutas metabólicas relacionadas con la formación de compuestos de azufre responsables de la corrosión. Se encontraron diferencias en la población bacteriana presente en diferentes partes de las canalizaciones congruentes con el proceso de formación de ácido sulfúrico, causante de la corrosión.

Otro ejemplo es el de las investigaciones de Chao *et al.*<sup>5</sup> en biofilms crecidos sobre distintos soportes, basadas en la secuenciación del gen 16S ARNr y un abordaje metagenómico de tipo *shotgun*. El filo más abundante encontrado correspondía a *Proteobacteria*, especialmente la clase *Alphaproteobacteria* y fue posible identificar varias rutas funcionales relacionadas con la resistencia a estrés oxidativo y a distintos desinfectantes empleados para controlar el crecimiento de biofilms en los sistemas de distribución de agua.

Tal y como se destaca en el artículo de Shaw<sup>49</sup>, el desarrollo de biofilms en la red de distribución de agua puede causar una disminución de la calidad y las cualidades organolépticas del agua potable, reducir la eficiencia del flujo e incluso se ha relacionado con la transmisión de organismos patógenos<sup>10, 19</sup>, por lo que resulta relevante comprender la diversidad microbiana ligada a estas estructuras.

Por último, respecto a la comparación entre aproximaciones metagenómicas y basadas en el 16S ARNr, estudios previos<sup>23, 48</sup> muestran que las dos aproximaciones proporcionan resultados similares pero no idénticos en cuanto a la distribución taxonómica que reflejan. En 2010 se diseñó un protocolo informático (Phyloshop) que localiza *in silico* los fragmentos del 16S de los datos de secuenciación de metagenomas, los analiza taxonómicamente y compara los resultados con los obtenidos por secuenciación directa del 16S<sup>48</sup>. El estudio mostró que existían diferencias significativas en la diversidad microbiana obtenida mediante cada técnica para una misma muestra y que las conclusiones extraídas de cada método no eran directamente comparables. Aunque en el estudio se apunta la necesidad de análisis posteriores, según los resultados obtenidos la secuenciación directa de 16S describía de forma más detallada la comunidad bacteriana, ya que era capaz de detectar especies que se encontraban menos representadas.

## 1.5 Perspectivas futuras

La aparición de las nuevas tecnologías de secuenciación masiva, comúnmente conocidas como NGS, y los avances implementados en las técnicas de secuenciación y herramientas de análisis han contribuido al progreso de este campo de investigación<sup>47</sup>, ya que han permitido aumentar la cantidad de datos hasta un orden de billones de bases secuenciadas por muestra<sup>26</sup> aunque siguen mostrando limitaciones a la hora de obtener ensamblajes *de novo* de alta calidad<sup>26</sup>, debido a los errores introducidos durante la secuenciación y la restricción en la longitud de los fragmentos que puede ser secuenciada. En consecuencia, es necesario optimizar los métodos actuales de extracción de ADN, construcción de librerías, ensamblaje de secuencias y análisis bioinformáticos para aplicarlos a los proyectos de metagenómica<sup>58</sup>.

La aplicación de las NGS a la secuenciación del gen 16S actualmente está generando cientos de miles de secuencias al año y ha revelado una gran diversidad microbiana que no había sido identificada hasta la fecha, dado que la clasificación de bacterias y arqueas se ha fundamentado durante años en la caracterización genética y fenotípica de especies cultivables, que habitualmente resulta compleja y lenta, especialmente para microorganismos con requisitos metabólicos complejos<sup>67</sup>.

Este tipo de estudios han demostrado tener aplicación no solo en muestras ambientales, con proyectos a escala global como el *Earth Microbiome Project*<sup>17</sup>, sino también en laboratorios de microbiología clínica<sup>40, 55</sup>, en los que la identificación de aislados bacterianos es una tarea rutinaria. Los esquemas de identificación basada en características fenotípicas que se emplean

requieren obtener cultivos puros, pueden depender de criterios subjetivos a la hora de interpretar los resultados de los tests<sup>53</sup> y ser poco precisos en algunos casos. En este sentido, la identificación basada en el 16S supone una alternativa rápida y reproducible, ya que proporciona datos no ambiguos incluso para aislados poco comunes<sup>11</sup>.

Para mejorar la calidad de la información obtenida mediante las aproximaciones basadas en el gen 16S es necesario aumentar la cantidad de secuencias de alta calidad depositadas en las bases de datos específicas existentes y desarrollar herramientas bioinformáticas que faciliten la interpretación de los resultados<sup>11</sup>.



## 2. OBJETIVOS

En este trabajo se pretende analizar la población bacteriana en muestras ambientales de biofilms mediante un enfoque basado en la secuenciación del gen 16S del ARN ribosomal, ampliamente empleado como marcador filogenético.

En primer lugar, se busca obtener la asignación taxonómica lo más completa posible de las colonias analizadas, para poder describir la población microbiana (taxones presentes y abundancia relativa de cada uno de ellos).

- Se prestará especial atención a la posible detección del género patógeno *Legionella*, por ser endémico de la localidad de procedencia de las muestras y haber dado lugar a brotes de forma recurrente en esta zona. Asimismo, se estudiará la presencia de otros agentes que puedan suponer un riesgo para la salud pública.
- Se evaluará el rendimiento y los resultados obtenidos mediante el procedimiento experimental seguido, indicando los aspectos que hayan presentado mayor dificultad o que podrían ser optimizados.

A continuación, se compararán los resultados obtenidos por este método con los generados al aplicar una estrategia metagenómica sobre las mismas muestras. La obtención de estos datos se realizó de forma previa a este trabajo mediante la construcción de genotecas, secuenciación y ensamblaje de los genomas.

- Se procederá a comparar los taxones identificados mediante cada enfoque, así como su abundancia relativa, identificando las principales similitudes y diferencias.
- También se compararán los resultados obtenidos al realizar la asignación taxonómica en función de las secuencias de 16S ARNr extraídas de la secuenciación de los metagenomas.
- Por último, se señalarán y discutirán las posibles razones de los sesgos encontrados en cada estrategia.

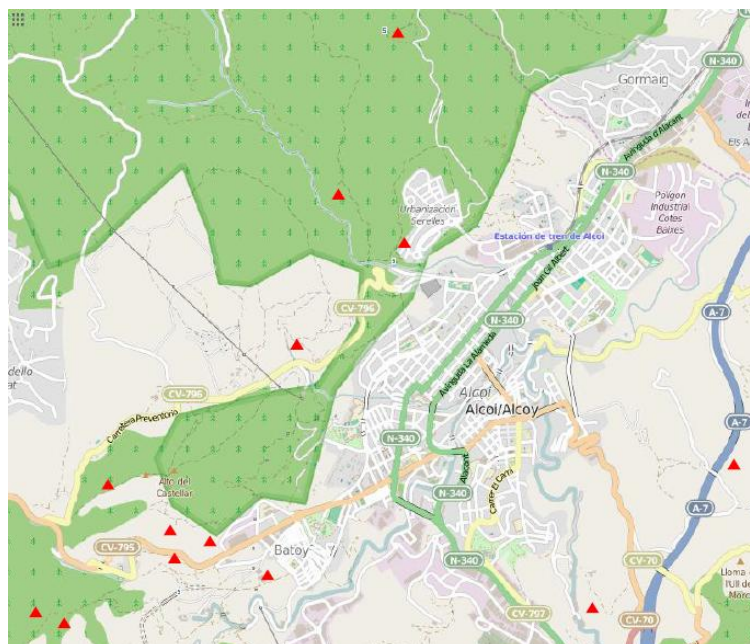


### 3. MATERIALES Y MÉTODOS

#### 3.1 Protocolo experimental

##### 3.1.1 Naturaleza y preparación de las muestras

Se trabajó con 14 muestras de biofilm procedentes de fuentes naturales situadas en los alrededores de la ciudad de Alcoy (Alicante, España), muestreadas en octubre de 2013<sup>44</sup> y registradas con la nomenclatura N1, N2, N3, N4, N5, N6, N7, N8, N9, N10, N11, N23, N24 y N25.



**Figura 3.1.** Distribución geográfica de las fuentes de las que se tomaron las muestras, señaladas con triángulos rojos en el mapa. *Cedido por la Dra. Sánchez Busó.*

En el laboratorio se conservaban cultivos en medio sólido GVPC realizados a partir del sobrenadante de estas muestras (~500 µL) almacenados a -80° C para su conservación.

Las placas fueron descongeladas y se procedió a preparar stocks de colonias individuales de cada muestra (véase *Anexos, Figuras 7.1-7.7*). Para ello, se emplearon placas VWR® 96-Well Flat Plates a las que se añadieron 25 µL de agua autoclavada por pocillo, se picaron colonias individuales y se resuspendieron en cada uno de los pocillos. Se obtuvo un tamaño muestral de 48 colonias por placa de cultivo, excepto para las muestras N2, N3, N4 y N25, con n=50, n=15, n=31 y n=36, respectivamente.

Se barajaron varios métodos de extracción de ADN (choque térmico, tratamiento con proteinasa K, lisozima, etc.), y se llevaron a cabo algunas pruebas previas en el laboratorio aplicando choque térmico y realizando directamente PCR de colonia. Se consideró preferible esta última opción, ya que permite reducir el gasto de reactivos y tiempo.

### 3.1.2 Amplificación por PCR y purificación

Tras consultar la bibliografía, se optó por amplificar las regiones V1-V4 del gen 16S ARNr por las ventajas expuestas en la introducción y por generar amplicones de unos 750 pares de bases, más viables para la secuenciación posterior por el método Sanger que el gen completo. Concretamente, se emplearon los cebadores 27F y 785R de la casa comercial IDT (*Integrated DNA Technologies*), cuyas características se muestran en la Tabla 3.1:

**Tabla 3.1.** Secuencia y temperatura de fusión o *melting* ( $T_m$ ) de los primers empleados.

Primer	Secuencia	$T_m$
<b>27F (forward)</b>	5' AGA GTT TGA TCM TGG CTC AG 3'	53'2° C
<b>785R (reverse)</b>	5' CCV GGG TAT CTA ATC C 3'	46'8° C

Se siguió un protocolo estándar de PCR, con ciertas modificaciones para optimizar la reacción: se aumentó el tiempo de la etapa de desnaturalización inicial a 8 minutos, para provocar la lisis celular; se ajustó la temperatura de anillamiento a 50° C, teniendo en cuenta las  $T_m$  de cada primer indicadas por el fabricante (Tabla 3.1) y se ajustó el tiempo de extensión a 45 segundos, dada la longitud de los fragmentos a amplificar.

En primer lugar se verificó que los cebadores a utilizar funcionaban correctamente (véase *Anexos, Figura 7.8*). A continuación, se realizaron pruebas para testar el protocolo a seguir, obteniéndose amplificado para los controles positivos, pero apareciendo también amplificado en los controles negativos. Por ello, se prepararon stocks de reactivos nuevos y se redujo el número de ciclos de 35 a 30, a fin de reducir el efecto de posibles contaminaciones. Los reactivos y el programa de temperaturas del termociclador finalmente empleados se muestran en las Tablas 3.2 y 3.3, respectivamente.

**Tabla 3.2.** Volumen de reactivos utilizados por cada reacción de PCR.

Reactivos	Volúmenes ( $\mu$ L)
H <sub>2</sub> O ultrapura	18'5
Buffer 10x con Mg <sub>2</sub> Cl	2'5
dNTPs (10 mM)	0'5
Primer <i>forward</i> (10 $\mu$ M)	0'5
Primer <i>reverse</i> (10 $\mu$ M)	0'5
Polimerasa (Taq, 1 U/ $\mu$ L, Biotools)	0'5
ADN molde	2
<b>Volumen total</b>	<b>25</b>

Antes de realizar la transferencia del ADN molde, se centrifugaron las placas stock a 12.000 rpm durante 5 minutos para precipitar restos celulares indeseados. Los 2  $\mu$ L de ADN molde indicados en la Tabla 3.2 se tomaron del sobrenadante obtenido tras este paso de centrifugación.

**Tabla 3.3.** Programa de temperaturas empleado en las reacciones de PCR.

Etapa	Temperatura	Duración
Desnaturalización inicial	95 °C	8 min
Desnaturalización	95 °C	30 s
Anillamiento de primers	50 °C	30 s
Extensión	72 °C	45 s
Extensión final	72 °C	5 min

30 ciclos

Las PCRs se validaron realizando electroforesis de ADN en geles de agarosa al 0'8 %, bajo unas condiciones estándar de 120 V durante 25 minutos, comprobando si se observaba banda en las muestras amplificadas, y anotando aquellas bandas perceptibles pero de menor intensidad (análisis cualitativo y semicuantitativo de los resultados).

A fin de maximizar el número de colonias analizadas, aquellas que inicialmente dieron negativo para la PCR fueron transferidas a tubos Eppendorf de 1 mL para aplicar un método de extracción de ADN mediante choque térmico, consistente en ciclos alternantes de 5 minutos a 99° C y a 0° C. Se cuantificó la concentración de ADN mediante un espectrofotómetro NanoDrop™ 1000 y se ajustó a 50-100 ng/μL para evitar que se saturara la reacción de PCR. Estas colonias se sometieron a una segunda reacción de PCR, empleando los mismos volúmenes de reactivos y programa de temperaturas indicados en las Tablas 3.2 y 3.3, reduciendo el tiempo de desnaturalización inicial a 5 minutos.

Una vez asegurada la presencia de producto amplificado de PCR, se llevó a cabo un método de purificación de los amplicones para eliminar los reactivos sobrantes de la PCR y otros posibles contaminantes. Para ello, se siguió un protocolo basado en el uso de placas con filtro de la casa comercial EdgeBio (*ExcelsaPure™ 96-Well UF PCR Purification Plates*) para retener el ADN, dos etapas de lavado mediante centrifugación (a 1500 rcf durante 5 minutos) con 75 y 100 μL de agua autoclavada, respectivamente, y resuspensión del ADN en 25 μL de agua autoclavada. Por último, se transfirió el producto purificado a placas de 96 pocillos para su conservación a -20° C.

### 3.1.3 Secuenciación

El siguiente paso consistió en la secuenciación por el método de Sanger de los amplicones previamente amplificados y purificados. Se realizaron dos lecturas por cada amplicón, una en sentido *forward* y otra en sentido *reverse*, con los cebadores empleados en las reacciones de PCR. En el laboratorio se realizaron los protocolos de preparación de las placas con los reactivos necesarios y las reacciones de Sanger, mientras que la electroforesis capilar y la lectura de la fluorescencia para la obtención de los cromatogramas fue llevada a cabo en otras instalaciones (Servicio de Secuenciación del Servicio Central de Soporte a la Investigación Experimental de la Universidad de Valencia) por carecer del equipo necesario para ello en el laboratorio. Los reactivos y el programa de temperaturas empleados se muestran en las Tablas 3.4 y 3.5:

**Tabla 3.4.** Volumen de reactivos utilizados para cada reacción de secuenciación.

Reactivos	Volumen (μL)
H <sub>2</sub> O autoclavada	4
Tampón de secuenciación ( <i>BigDye® Terminator v1.1, v1.3 5x Sequencing Buffer</i> )	1'6
Kit de secuenciación <sup>I</sup> ( <i>BigDye™ Terminator v3.0 Ready Reaction Cycle Sequencing Kit, Applied Biosystems</i> )	0'4
Primer <i>forward</i> o <i>reverse</i> (5μM)	1
ADN purificado <sup>II</sup>	1'5

<sup>I</sup> El kit contiene ddNTPs marcados diferencialmente con fluorocromos, dNTPs sin marcaje y polimerasa termoestable Taq.

<sup>II</sup> En el caso de las colonias que habían dado una banda tenue en la electroforesis se añadieron 2'5 μL de ADN a la reacción de secuenciación.

**Tabla 3.5.** Ciclos de temperatura empleados para la reacción de secuenciación por Sanger.

Etap a	Temperatura	Duración
Desnaturalización	96 °C	10 s
Anillamiento de primers	50 °C	5 s
Extensión	60 °C	4 min
<b>Número de ciclos totales</b>		<b>99</b>

## 3.2 Análisis bioinformático

### 3.2.1 Tratamiento de los cromatogramas

Para el procesado de los cromatogramas y la obtención de secuencias consenso de cada amplicón se emplearon los programas Pregap4 y Gap4 del paquete *Staden*<sup>52</sup>.

En primer lugar se ejecutó el Pregap4 sobre los cromatogramas en formato .ab1 para detectar y enmascarar zonas de la secuencia de baja calidad y crear archivos en formato .exp con los que realizar el ensamblaje posterior. Para ello, se seleccionaron las opciones *Estimate Base Accuracies*, *Initialise Experiment Files*, *Quality Clip* e *Interactive Clipping* dentro de la pestaña *Configure Modules* del programa.

Los archivos .exp generados fueron empleados para obtener una secuencia consenso de cada una de las colonias analizadas, ensamblando los pares de lecturas en sentido *forward* y *reverse* mediante el Gap4. Estas secuencias se revisaron manualmente empleando el visor del programa para escindir las bases correspondientes a los primers y corregir errores en la asignación de bases (*base calling*) debidos a ruido de fondo en el cromatograma, picos de baja intensidad o zonas de baja resolución, entre otros. Las secuencias consenso se guardaron en formato .fasta para los análisis posteriores.

A fin de detectar posibles secuencias anómalas o errores en el proceso de eliminar la secuencia de los primers se realizaron alineamientos múltiples de las secuencias obtenidas para cada una de las 14 muestras iniciales empleando el algoritmo MUSCLE del programa MEGA v5.2.2<sup>56</sup>.

### 3.2.2 Asignación taxonómica

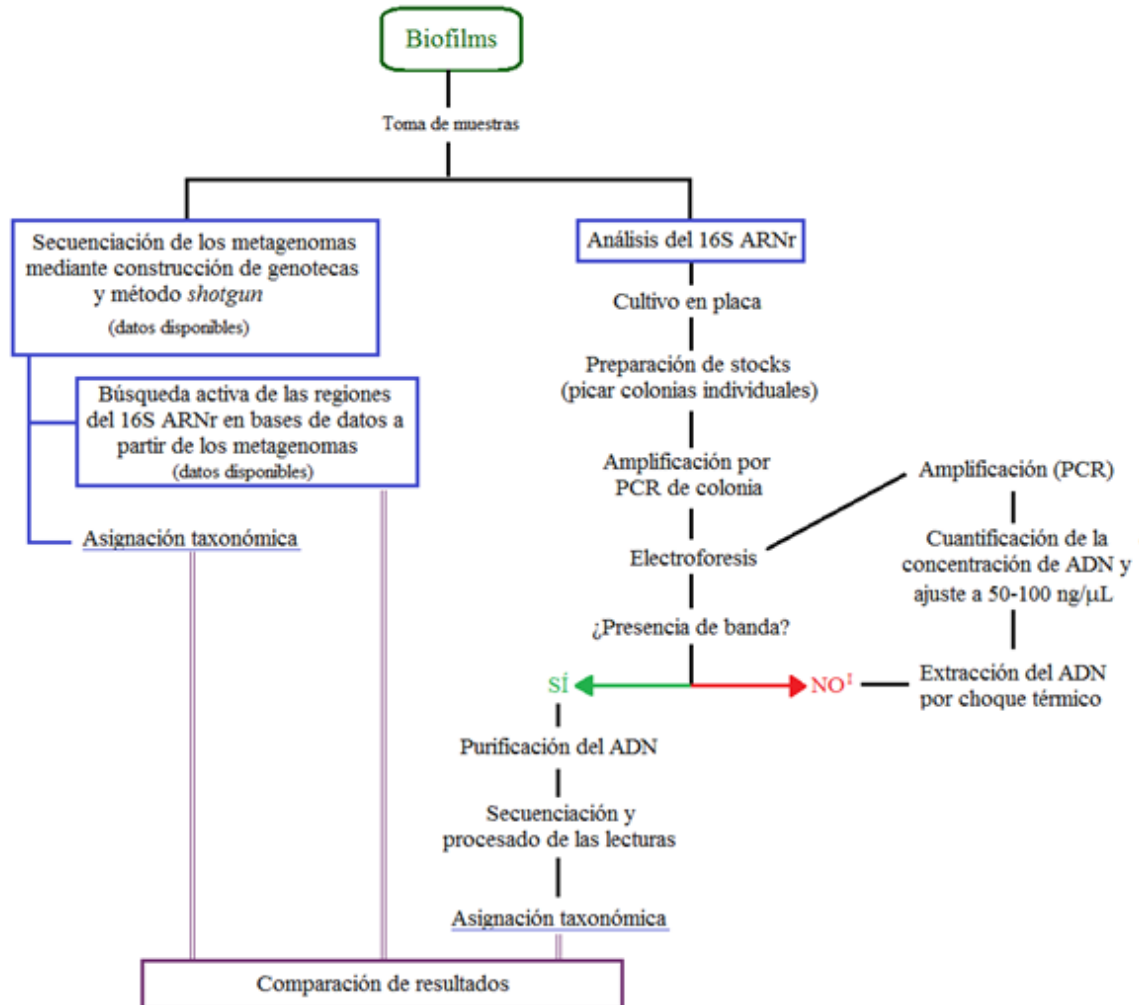
Para la identificación de las secuencias se empleó la herramienta SeqMatch de la base de datos RDP<sup>7</sup> (<https://rdp.cme.msu.edu>), específica para el ARN ribosomal, que contrasta las secuencias buscadas con las depositadas en la base y permite obtener su clasificación taxonómica. El grado de similitud entre cada secuencia buscada y las coincidencias encontradas viene indicado por el parámetro *S<sub>ab</sub> score*, definido como el número de heptámeros compartidos entre dos secuencias, dividido entre el número más bajo de oligos únicos en las secuencias. Este parámetro puede diferir de la similitud de secuencia si se compara base a base, pero se calcula que en el 95 % de los casos las 20 coincidencias mostradas son las que presentan mayor identidad de secuencia, ya sea basándose en el *S<sub>ab</sub> score* o en el porcentaje de similitud.

### 3.2.3 Identificación de géneros potencialmente patógenos

Con objeto de identificar posibles agentes presentes en las muestras que supongan un riesgo para la salud pública, se llevó a cabo una búsqueda en la base de datos de la Asociación Americana de Seguridad Biológica (ABSA - <http://www.absa.org/riskgroups/index.html>), basada en la clasificación de los microorganismos en grupos de riesgo de la Organización Mundial de la Salud.

### 3.3 Visión general del flujo del trabajo

A continuación se muestra, a modo de resumen y a grandes rasgos, un esquema del flujo de trabajo seguido (Figura 3.2):



<sup>1</sup>En el caso de ausencia de banda tras la extracción por choque térmico, se descartó la colonia del estudio.

**Figura 3.2.** Diagrama esquemático del procedimiento de trabajo aplicado, destacando las tres estrategias que se han comparado y el diseño experimental realizado en el laboratorio en el análisis del 16 ARNr. *Elaboración propia.*

## 4. RESULTADOS Y DISCUSIÓN

### 4.1 Asignación taxonómica

#### 4.1.1 Visión global

De las 14 muestras analizadas, tras el procesado de los datos de secuenciación, finalmente se obtuvieron 416 secuencias consenso que se contrastaron con las depositadas en la base de datos RDP, específica para el 16S ribosomal.

Fue posible asignar todas las secuencias a los niveles taxonómicos de filo, clase y orden; 413 (99'28 %) a nivel de familia y 409 (98'31 %) a nivel de género. Por otra parte, la resolución no fue suficiente para realizar la identificación de especies de forma precisa en la mayoría de las colonias, como se detalla más adelante. Los resultados se muestran en la Tabla 4.1, organizados taxonómicamente e indicando el número de colonias asignadas a cada género.

**Tabla 4.1.** (Continúa en la página siguiente) Asignación taxonómica obtenida al contrastar las secuencias del 16S ARNr con la base de datos RDP. Se ha incluido el número de colonias totales asignadas a cada género y algunos ejemplos de especies con las que se ha encontrado mayor identidad ( $S_{ab}$  score > 97 %), indicando entre paréntesis el identificador de acceso a la secuencia en la base donde se encuentra depositada. En caso de existir más especies candidatas que superan el umbral, se ha indicado mediante un “[+]”.

#### FILO: *Proteobacteria* – CLASE: *Gammaproteobacteria*

ORDEN	FAMILIA	GÉNERO	Nº DE COLONIAS	ESPECIES CANDIDATAS
<i>Aeromonadales</i>	<i>Aeromonadaceae</i>	<i>Aeromonas</i>	223	<i>A. allosaccharophila</i> (HF936927); <i>A. hydrophila</i> subsp. <i>null</i> (AF472491); <i>A. media</i> (KC914093); <i>A. salmonicida</i> (AF472493); <i>A. veronii</i> (AB182094) [+] <i>P. fluorescens</i> (KJ534246); <i>P. protegens</i> (HF985318); <i>P. putida</i> (FJ599770); <i>P. stutzeri</i> (AF307873); <i>P. xanthomarina</i> (JQ810936) [+]
<i>Pseudomonadales</i>	<i>Pseudomonadaceae</i>	<i>Pseudomonas</i>	28	<i>A. baumannii</i> (GU415588); <i>A. calcoaceticus</i> (AJ247199); <i>A. junii</i> (AJ247204); <i>A. rhizosphaerae</i> (JX295980) [+]
	<i>Moraxcellaceae</i>	<i>Acinetobacter</i>	16	<i>S. marcescens</i> (AY186053); <i>S. nematodiphila</i> (KC894027); <i>S. fonticola</i> (AY178556) [+]
<i>Enterobacteriales</i>	<i>Enterobacteriaceae</i>	<i>Serratia</i>	13	<i>E. cloacae</i> (AJ417474), <i>E. ludwigii</i> ; (FJ599776); <i>Pantoea agglomerans</i> (JN941307) [+]
		<i>Enterobacter</i>	8	<i>C. freundii</i> (KJ997753); <i>C. gillenii</i> (EU888872); <i>C. murliniae</i> (KC017346) [+]
		<i>Citrobacter</i>	2	<i>K. pneumoniae</i> <sup>1</sup> (AY376437)
		<i>Klebsiella</i>	1	-
		No asignado	1	-
<i>Xanthomonadales</i>	<i>Xanthomonadaceae</i>	<i>Stenotrophomonas</i>	1	<i>S. maltophilia</i> <sup>1</sup> (KJ545901); <i>S. sp. EB35</i> (AB714662)

#### FILO: *Proteobacteria* – CLASE: *Betaproteobacteria*

ORDEN	FAMILIA	GÉNERO	Nº DE COLONIAS	ESPECIES CANDIDATAS
<i>Burkholderiales</i>	<i>Comamonadaceae</i>	<i>Delftia</i>	21	<i>D. acidovorans</i> (AF457648); <i>D. lacustris</i> (HF585312) y <i>D. tsuruhatensis</i> (GU591410) [+]
		<i>Comamonas</i>	2	<i>C. aquatica</i> (AB308304); <i>C. sp. Arv-22-2.10c</i> (KF214953); <i>C. sp. J3-A73</i> (DQ454124) [+]
		No asignado	2	-
	<i>Alcaligenaceae</i>	<i>Achromobacter</i>	1	<i>A. xylosoxidans</i> subsp. <i>null</i> (AY684781); <i>A. insolitus</i> (DQ386150); <i>A. sp. JDM-3-03</i> (GU324708) [+]
		<i>Alcaligenes</i>	3	<i>A. faecalis</i> subsp. <i>faecalis</i> (JN786084); <i>Burkholderia cepacia</i> (AB086822); <i>A. sp. HPC358</i> (AY999031) [+]
<i>Neisseriales</i>	<i>Neisseraceae</i>	<i>Chromobacterium</i>	9	<i>C. haemolyticum</i> (GU195136) y <i>C. violaceum</i> (JQ312052)



**FILO: Proteobacteria – CLASE: Alphaproteobacteria**

ORDEN	FAMILIA	GÉNERO	Nº DE COLONIAS	ESPECIES CANDIDATAS
<i>Caulobacterales</i>	<i>Caulobacteraceae</i>	<i>Brevundimonas</i>	19	<i>Bacterium HPC300</i> (AY854106); <i>B. HPC350</i> (AY999027); <i>B. diminuta</i> (JQ897393) [+]
<i>Sphingomonadales</i>	<i>Sphingomonadaceae</i>	<i>Sphingomonas</i>	6	<i>S. koreensis</i> (KC420522); <i>alpha proteobacterium HA-est-Is33-14</i> (HM368283); <i>S. sp. S4-A16</i> (KC139682)
<i>Rhizobiales</i>	<i>Brucellaceae</i>	<i>Ochrobactrum</i>	2	<i>O. tritici</i> (AY972166); <i>O. anthropi</i> (AJ784809); <i>Brucellaceae bacterium D180a</i> (AB480354) [+]
	<i>Phyllobacteriaceae</i>	<i>Aquamicrobium</i>	1	<i>Defluviobacter sp. RH.Sph.068</i> (DQ099485)
<i>Rhodospirillales</i>	<i>Acetobacteraceae</i>	<i>Roseomonas</i>	1	-

**FILO: Bacteroidetes – CLASE: Flavobacteria**

ORDEN	FAMILIA	GÉNERO	Nº DE COLONIAS	ESPECIES CANDIDTAS
<i>Flavobacteriales</i>	<i>Flavobacteriaceae</i>	<i>Chryseobacterium</i>	27	-
		<i>Empedobacter</i> <sup>II</sup>	2	<i>E. brevis</i> <sup>I</sup> (LN774590)

**FILO: Bacteroidetes – CLASE: Sphingobacteria**

ORDEN	FAMILIA	GÉNERO	Nº DE COLONIAS	ESPECIES CANDIDATAS
<i>Sphingobacteriales</i>	<i>Sphingobacteriaceae</i>	<i>Sphingobacterium</i>	1	<i>S. sp. JDM-3-02</i> (GU324707)

**FILO: Firmicutes – CLASE: Bacilli**

ORDEN	FAMILIA	GÉNERO	Nº DE COLONIAS	ESPECIES CANDIDATAS
<i>Bacillales</i>	<i>Bacillaceae 1</i>	<i>Bacillus</i>	4	<i>B. anthracis</i> (EU982898); <i>B. cereus</i> (GU169024); <i>B. macroides</i> (EF010605); <i>B. simplex</i> (EF010618); <i>B. thuringiensis</i> (GQ505328) [+]
		<i>Bacillus</i> <sup>II</sup>	3	-
	<i>Paenibacillaceae 1</i>	<i>Paenibacillus</i>	6	<i>P. dendritiformis</i> <sup>I</sup> (JQ420900)
	<i>Bacillaceae 2</i>	<i>Oceanobacillus</i>	4	<i>O. polygoni</i> (LN774503); <i>O. profundus</i> (HM032810); <i>Bacillus sp. B118</i> (DQ277990) [+]
		<i>Virgibacillus</i> <sup>II</sup>	1	<i>V. proomii</i> <sup>I</sup> (JQ897425); <i>V. sp. nbri08</i> <sup>I</sup> (GU991544); <i>V. sp. mdu09</i> <sup>I</sup> (HM488994)
	<i>Planococcaceae</i>	<i>Sporosarcina</i>	4	<i>S. koreensis</i> (KM009125); <i>S. luteola</i> ; (KC894026)
		No asignado	1	-
	No asignado	-	3	-

<sup>I</sup>*S<sub>ab</sub> score < 97 %*

<sup>II</sup>Asignación no unívoca, se obtuvieron otros géneros en la asignación pero con un *S<sub>ab</sub> score* significativamente menor.

### 4.1.2 Identificación de especies

Para cada una de las secuencias buscadas se obtuvieron 20 posibles secuencias depositadas en la base con las que presentaban mayor identidad. En la Figura 4.1 se muestra un ejemplo de los resultados obtenidos para una de las secuencias buscadas.

```

+ rootrank Root (0/20/278296) (selected/match/total RDP sequences)
+ domain Bacteria (0/20/200655)
+ phylum "Proteobacteria" (0/20/85024)
+ class Alphaproteobacteria (0/20/19108)
+ order Caulobacterales (0/20/1031)
+ family Caulobacteraceae (0/20/911)
+ genus Brevundimonas (0/20/655)
 S000027591 not_calculated 1.000 0638 Brevundimonas sp. DB3; AY008334
 S000027592 not_calculated 1.000 0633 Brevundimonas sp. DB2; AY008333
 S000029845 not_calculated 0.998 0520 Brevundimonas sp. DB5; AY008336
 S000478290 not_calculated 1.000 0560 bacterium HPC300; AY854106
 S000478299 not_calculated 0.998 0563 bacterium HPC317; AY854116
 S000541240 not_calculated 0.965 0568 Brevundimonas HPC350; AY999027
 S000712409 not_calculated 0.993 0422 Brevundimonas diminuta; ATCC 19146; DQ650706
 S000768951 not_calculated 0.955 0760 Brevundimonas sp. kly76m-1; EF028207
 S000927770 not_calculated 1.000 0578 Brevundimonas sp. YIM C697; EU135613
 S001352757 not_calculated 1.000 0490 Brevundimonas sp. WH293; FJ866657
 S001352758 not_calculated 0.986 0489 Brevundimonas sp. WH296; FJ866658
 S001352781 not_calculated 1.000 0490 Brevundimonas sp. WH029; FJ866681
 S001611323 not_calculated 0.958 0755 Brevundimonas sp. BCAN5-S20; FM998721
 S001745936 not_calculated 1.000 0583 Brevundimonas diminuta; 2T1-1; GU126799
 S002229604 not_calculated 0.972 0648 Brevundimonas diminuta; HT7; HM446258
 S003298414 not_calculated 0.997 0773 Brevundimonas diminuta; 17778_C7; JQ897393
 S003302601 not_calculated 1.000 0604 Brevundimonas diminuta; Nazemi; JX134073
 S003616799 not_calculated 0.967 0919 Brevundimonas diminuta; SEN16N; JX490088
 S003720502 not_calculated 0.959 0659 Brevundimonas sp. DT50-1; KC686389
 S004230195 not_calculated 0.962 0949 Brevundimonas diminuta; TR6; KM009127

```

**Figura 4.1.** Ejemplo del resultado de la búsqueda para una de las secuencias (colonia BCol\_10). Se muestran las 20 secuencias de mayor coincidencia encontradas, incluyendo sus identificadores, tres parámetros de la búsqueda – entre ellos el *S<sub>ab</sub> score* (resaltado en naranja)– y los nombres completos de las secuencias.

Puesto que en un amplio de número de estudios se toma el criterio del 97 % de identidad para considerar que dos secuencias del gen 16S pertenecen a la misma especie bacteriana<sup>50</sup>, se seleccionaron algunas de las posibles especies que superaban este umbral de similitud, mostradas en la Tabla 4.1 como especies candidatas.

En prácticamente ninguna de las colonias se encontró una única especie con la que la identidad fuera igual o superior al 97 %, por lo que solo fue posible realizar la asignación taxonómica a este nivel en un 87.89 % de las colonias.

Cabe mencionar que el umbral del 97 % ha sido cuestionado por otros estudios previos para la identificación basada en el 16S ARNr<sup>16</sup>, recomendando aplicar este umbral para la asignación de géneros y emplear un criterio más estricto del 99 % para la identificación de especies<sup>11</sup>.



En la Tabla 4.2 se muestran las coincidencias encontradas a nivel de especie aplicando distintos umbrales de *S<sub>ab</sub> score*:

**Tabla 4.2.** Listado de especies con las que se ha encontrado una sola coincidencia al buscar las secuencias problema, aplicando tres umbrales de identidad basados en el parámetro *S<sub>ab</sub> score* (97, 99 y 100 %) y abundancia relativa de las especies identificadas en la población estudiada. Se indican entre paréntesis los identificadores de acceso a las secuencias de la base de datos.

Especies identificadas	Umbrales de <i>S<sub>ab</sub> score</i>			Abundancia relativa (%)
	≥ 97 %	≥ 99 %	100 %	
<i>Aeromonas</i> sp. MBRG 4.2 (AJ508692)	✓			0'0024
<i>Aeromonas veronii</i> (AF472504)	✓	✓	✓	0'0024
<i>Bacillus shackletonii</i> (KJ513349)	✓			0'0024
<i>Bacillus</i> sp. A24MTQ (FN293231)	✓	✓	✓	0'0024
<i>Brevundimonas HPC350</i> (AY999027)	✓			0'0024
<i>Citrobacter amalonaticus</i> (KC689293)	✓	✓		0'0024
<i>Comamonas</i> sp. Arv-22-2.10c (KF214953)	✓	✓	✓	0'0024
<i>Defluviobacter</i> sp. RH.Sph.068 (DQ099485)	✓			0'0024
<i>Paenibacillus dendritiformis</i> (KF597219)	✓			0'0024
<i>Sphingobacterium</i> sp. JDM-3-02 (GU324707)	✓			0'0024
<i>Sphingomonas</i> sp. S4-A16 (KC139682)	✓			0'0024
<i>Sporosarcina koreensis</i> (KM009125)	✓			0'0024
<i>Sporosarcina luteola</i> (KC894026)	✓			0'0024
<i>Stenotrophomonas</i> sp. EB35 (AB714662)	✓			0'0024
<i>Acinetobacter junii</i> (AB976099)	✓	✓		0'0024
<i>Aeromonas</i> sp. B92 (AM179877)	✓	✓		0'0024
<i>Aeromonas</i> sp. E251 (DQ159216)	✓	✓		0'0024
<i>Aeromonas veronii</i> (AB182219)	✓	✓		0'0024
<b><i>Aquatic bacterium RI-G3</i> (AB195770)</b>	✓	✓		0'0024
<i>Chromobacterium haemolyticum</i> (GU195136)	✓	✓	✓	0'0216
<i>Delftia</i> sp. WH039 (FJ866687)	✓	✓		0'0024
<i>Aeromonas</i> sp. AE150 (EU723931)	✓	✓	✓	0'0024
<i>Aeromonas</i> sp. BC3 (DQ343216)	✓	✓	✓	0'0048
<i>Bacillus</i> sp. B118 (DQ277990)	✓	✓	✓	0'0048
<i>Bacillus</i> sp. B118 (DQ277990)	✓	✓	✓	0'0024
<i>Comamonas aquatica</i> (AB308304)	✓	✓	✓	0'0024
<i>Pseudomonas</i> sp. AU (AF325153)	✓	✓	✓	0'0024
<b>TOTAL ESPECIES</b>	<b>27</b>	<b>17</b>	<b>10</b>	<b>8'89 %</b>

Entre los problemas encontrados en la identificación, en un 16 % de las búsquedas no se encontraron coincidencias con una identidad de al menos el 97 %, principalmente en secuencias pertenecientes a los filos *Bacteroidetes* y *Firmicutes*.

Es destacable que en muchos de los casos sí que se llegó a obtener una identidad muy elevada, incluso del 100 %, pero como las coincidencias se daban con varias secuencias pertenecientes a distintas especies, siguió sin ser suficiente para identificar la especie de forma inequívoca. Concretamente, un 65 % de las secuencias mostraban varias coincidencias con valores de *S<sub>ab</sub> score* del 100 %. Se ha observado que este problema ocurría de forma

recurrente en el género *Aeromonas*, que por sí solo representa más de la mitad de las secuencias estudiadas.

En todo caso, puede concluirse que la resolución de las regiones V1-V4 del 16S no es suficiente para la asignación precisa a nivel de especie. Estos resultados son congruentes con los obtenidos en investigaciones previas como la de Kuczynski y colaboradores<sup>26</sup>, que ya apuntaban este hecho. Uno de los motivos señalados es que para algunos géneros existen pocas secuencias depositadas en las bases de datos, por lo que es difícil encontrar secuencias con un porcentaje de similitud superior al 97 % cuando se realiza la búsqueda de una secuencia problema<sup>11</sup>. Además, ciertos géneros engloban especies con secuencias del 16S altamente similares, lo que reduce el poder de discriminación de esta estrategia más allá de la identificación de géneros. Por ejemplo, se ha observado que especies del género *Bacillus* poseen idénticas secuencias del 16S, lo que dificulta su identificación<sup>2, 16</sup>. Otros géneros que presentan complicaciones de forma sistemática a la hora de asignar especies mediante la aproximación del 16S son *Enterobacter*<sup>11, 57</sup> y *Pantoea*<sup>11</sup>.

## 4.2. Descripción de la población bacteriana

### 4.2.1 Taxones presentes y abundancias relativas

Todas las asignaciones realizadas se incluyen en el dominio *Bacteria*, como cabía esperar, puesto que los cebadores empleados para la amplificación del ARN ribosomal eran específicos de bacterias (en eucariotas, el componente de ARN ribosomal que forma parte de la subunidad pequeña es el 18S, en vez del 16S; las arqueas sí que poseen el gen 16S pero éste difiere significativamente del de bacterias y no amplifica con los cebadores empleados). Por los resultados obtenidos de los genomas completos se conocía de antemano que la presencia de los dominios *Eukarya* y *Archaea* era minoritaria (< 1 % cada uno), por lo que la población bacteriana representa prácticamente la totalidad de la comunidad microbiana en las muestras.

La población bacteriana estudiada se distribuye en 3 **filos**, con una marcada abundancia de *Proteobacteria* (86'54 %). Curiosamente, este resultado coincide con el obtenido en el estudio de Schmeisser *et al.*<sup>45</sup> en muestras de biofilm de redes de agua potable, en el cual la identificación taxonómica basada en el gen 16S reveló que la mayoría de bacterias presentes (el 86 %) pertenecían a este filo.

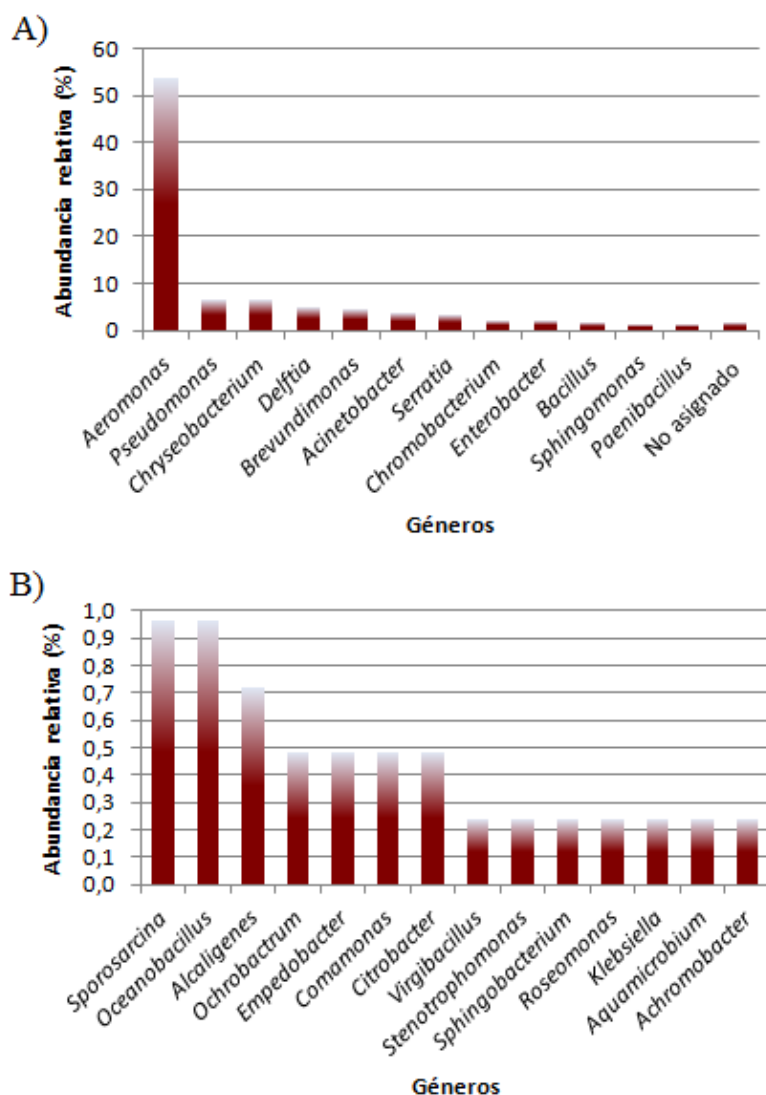
Fijándonos en las **clases**, encontramos 6 divisiones, dominadas por *Gammaproteobacteria* (70'43 %), seguida de otras cuatro clases con abundancias comprendidas entre el 5 y el 10 % y, por último, la clase minoritaria *Sphingobacteria* (< 1%).

En el siguiente nivel taxonómico aparecen 13 **órdenes**, de nuevo con un tipo que representa más de la mitad de la población, *Aeromonadales* (53'61 %). El segundo orden más abundante es el *Pseudomonadales*, aunque con una abundancia significativamente menor (10'58 %). De los órdenes restantes, cuatro de ellos presentan abundancias relativas inferiores al 1 %.

En cuanto a **familias**, la población queda dividida en 19 tipos. Dado que dentro del orden *Aeromonadales* solo se ha detectado una única familia, *Aeromonadaceae*, ésta sigue representando más de la mitad de la comunidad (53'61 %). Del resto de familias, ninguna supera el 10 % de la población y seis de ellas presentan una abundancia relativa inferior al 1 %. Dentro del orden *Bacillales* queda un 0'72 % sin asignar a nivel de familia.

Por último, se han encontrado 26 **géneros**, de los cuales destaca *Aeromonas*, con una abundancia relativa del 53'61 %. Otros 12 géneros presentan frecuencias de aparición entre el 1 y el 10 %, siendo los más abundantes *Pseudomonas* y *Chryseobacterium*, aunque en todo caso se encuentran en porcentajes mucho menores que el género predominante. Los 14 géneros restantes suponen en conjunto solo un 6'25 % de la población, sin que ninguno de ellos alcance el 1 % por sí mismo. El porcentaje de colonias sin asignar a nivel de género se sitúa en el 1'68 %, incluyendo el 0'72 % no adscrito a ninguna familia, un 0'48 % perteneciente a la

familia *Comamonadaceae*, un 0'24 % englobado en la familia *Enterobacteriaceae* y un 0'24 % correspondiente a la familia *Planococcaceae*. Cabe destacar que las secuencias asignadas a los géneros *Empedobacter* y *Virgibacillus*, así como tres de las secuencias del género *Bacillus*, presentaban coincidencias con otros géneros. Sin embargo, al revisar los porcentajes de similitud se vio que las identidades con éstos eran significativamente menores (del orden del 80 % de *S\_ab score*) y se decidió asignar manualmente estas secuencias al género con mayor identidad. Los géneros presentes y sus abundancias relativas pueden observarse en la Figura 4.2.



**Figura 4.2.** Abundancias relativas de los géneros detectados en el conjunto de las 14 muestras. En el apartado A se muestran los géneros con una frecuencia de aparición de al menos el 1 %, así como el porcentaje que no ha podido ser asignado a este nivel, y en el B aparecen los géneros con abundancias relativas inferiores al 1 % (la escala del eje se ha ajustado en cada caso para su correcta visualización).

#### 4.2.2 Identificación de patógenos

En este análisis no se ha detectado la presencia del patógeno *Legionella pneumophila* ni de otras especies del mismo género, a pesar de estar considerada endémica de la región donde fueron tomadas las muestras<sup>43</sup> y habitar comúnmente en aguas superficiales y sistemas relacionados con la red de distribución de agua. Por otra parte, estos resultados son congruentes con los obtenidos en el estudio de los metagenomas completos, donde el género *Legionella* presentaba una abundancia relativa muy baja (0'01-0'07 %).

Por otra parte, resulta remarcable el elevado número de géneros detectados relacionados con especies potencialmente patógenas, así como su abundancia, tal y como puede observarse en la Tabla 4.3.

**Tabla 4.3.** Listado de géneros identificados en las muestras que contienen especies patógenas según la clasificación de la Organización Mundial de la Salud. Se ha indicado la abundancia relativa de cada género en las muestras, las especies consideradas como patógenos humanos y animales dentro de cada género y el grupo de riesgo en el que se engloban.

Género	Abundancia relativa (%)	Patógenos humanos en el género (grupo de riesgo 2)	Patógenos animales en el género (grupo de riesgo 2)
<i>Aeromonas</i>	53'61	<i>A. allosaccharophila</i> , <i>A. bestiarium</i> , <i>A. caviae</i> , <i>A. hydrophilia</i> , <i>A. jandaei</i> , <i>A. punctata</i> , <i>A. salmonicida</i> , <i>A. schubertii</i> , <i>A. sobria</i> , <i>A. trota</i> , <i>A. tructii</i> , <i>A. veronii</i> y <i>A. spp.</i>	<i>A. hydrophilia</i>
<i>Pseudomonas</i>	6'73	<i>P. aeruginosa</i> , <i>P. alcaligenes</i> , <i>P. cepacea</i> , <i>P. luteola</i> , <i>P. mendocina</i> , <i>P. oryzihabitans</i> y <i>P. spp.</i>	<i>P. anguilliseptica</i>
<i>Brevundimonas</i>	4'57	<i>B. diminuta</i>	-
<i>Acinetobacter</i>	3'85	<i>A. baumannii</i> , <i>A. calcoaceticus</i> , <i>A. haemolyticus</i> , <i>A. johnsonii</i> , <i>A. junii</i> , <i>A. lwoffii</i> y <i>A. spp.</i>	-
<i>Serratia</i>	3'13	<i>S. grimesii</i> , <i>S. liquefaciens</i> , <i>S. marcescens</i> y <i>S. proteamaculans</i>	-
<i>Chromobacterium</i>	2'16	<i>C. violaceum</i>	-
<i>Enterobacter</i>	1'92	<i>E. aerogenes</i> , <i>E. agglomerans</i> , <i>E. amnigenus</i> , <i>E. asburiae</i> , <i>E. cancerogenus</i> , <i>E. cloacae</i> , <i>E. gergoviae</i> , <i>E. hormaechei</i> , <i>E. intermedius</i> , <i>E. sakazakii</i> y <i>E. spp.</i>	-
<i>Sphingomonas</i>	1'44	<i>S. paucimobilis</i> y <i>S. parapaucimobilis</i>	-
<i>Paenibacillus</i>	1'44	<i>P. popilliae</i>	<i>P. larvae</i>
<i>Bacillus</i>	0'96	<i>B. anthracis</i> <sup>1</sup> , <i>B. cereus</i> , <i>B. pseudomycoides</i> y <i>B. rhamnosus</i>	<i>B. anthracis</i> <sup>1</sup> , <i>B. cereus</i> , <i>B. lentimorbus</i> , <i>B. popilliae</i> , <i>B. sphaericus</i> y <i>B. thuringiensis</i>
<i>Alcaligenes</i>	0'72	<i>A. faecalis</i> , <i>A. piechaudii</i> , <i>A. denitrificans</i> y <i>A. spp.</i>	-
<i>Citrobacter</i>	0'48	<i>C. amalonaticus</i> , <i>C. braakii</i> , <i>C. farmeri</i> , <i>C. freundii</i> , <i>C. koseri</i> , <i>C. sedlakii</i> , <i>C. werkmanii</i> , <i>C. youngae</i> y <i>C. spp.</i>	<i>C. rodentium</i>
<i>Comamonas</i>	0'48	<i>C. terrigena</i>	-
<i>Ochrobactrum</i>	0'48	<i>O. anthropi</i>	-
<i>Empedobacter</i>	0'48	<i>E. brevis</i>	-
<i>Klebsiella</i>	0'24	<i>K. granulomatis</i> , <i>K. mobilis</i> , <i>K. ornithinolytica</i> , <i>K. oxytoca</i> , <i>K. ozaenae</i> , <i>K. pneumoniae</i> , <i>K. rhinoscleromatis</i> y <i>K. spp.</i>	<i>K. pneumoniae</i> y <i>K. spp.</i>

<sup>1</sup>Considerado también como potencial arma biológica y como agente de riesgo para la agricultura por el Centro para el Control y Prevención de Enfermedades (CDC) y el Departamento de Agricultura de Estados Unidos (USDA), respectivamente.

Todas las especies patógenas encontradas están clasificadas en el grupo de riesgo 2, es decir, pueden causar enfermedad en humanos o animales pero generalmente no suponen una amenaza para los trabajadores de laboratorio, la colectividad o el medio ambiente. La exposición en el laboratorio puede causar una infección grave pero existen medidas preventivas y tratamientos efectivos y el riesgo de propagación de la infección es limitado<sup>37</sup>.

Dada la naturaleza ambiental de las muestras, *a priori* no se esperaba encontrar una presencia tan significativa de agentes patógenos. No obstante, estos datos deben interpretarse con cautela:

- Por una parte, no se ha llegado a asignar de forma inequívoca la población bacteriana a nivel de especie, por lo que los microorganismos presentes en la muestra pueden corresponder a especies no patógenas de los géneros indicados.
- Además, las secuencias depositadas en las bases de datos corresponden frecuentemente a organismos patógenos, por lo que al contrastar las secuencias aparece un sesgo hacia este tipo de bacterias.

Por tanto, serían necesarios estudios más específicos para confirmar la presencia real de especies patógenas en las fuentes donde se tomaron las muestras.

### 4.2.3 Distribución de las poblaciones por muestra

A continuación se presentan los datos poblacionales obtenidos para cada muestra. En la Figura 4.3 se muestra la distribución a nivel de clase y género para las 14 muestras.

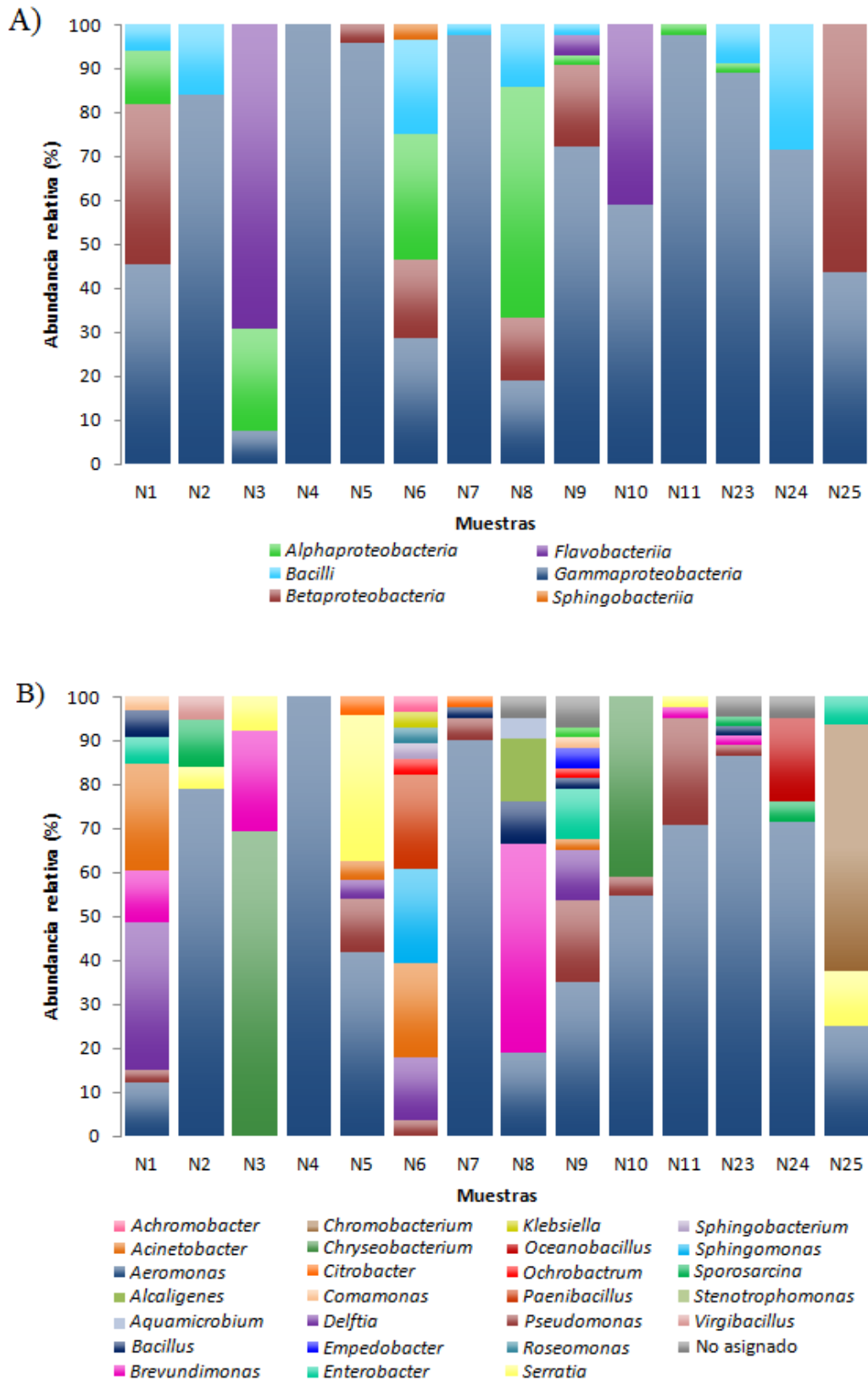


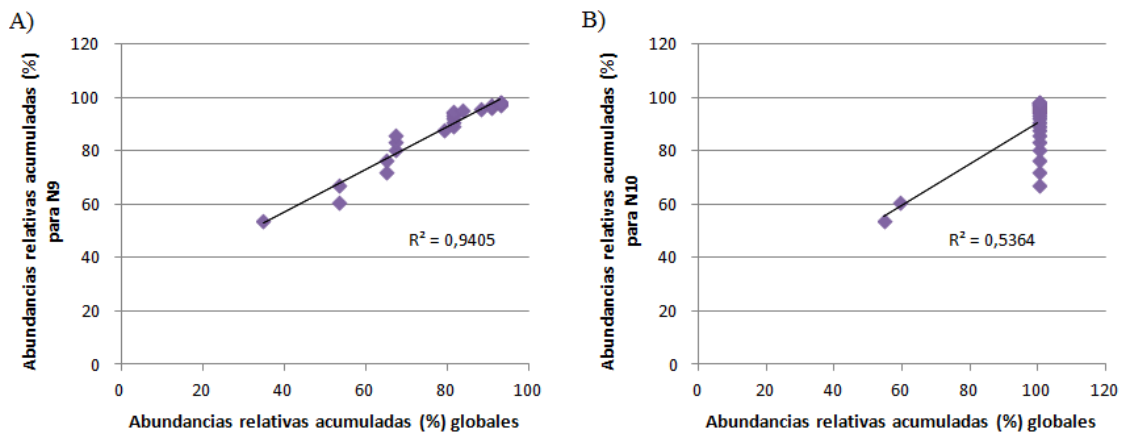
Figura 4.3. Abundancias relativas de las clases (A) y géneros (B) detectados en las 14 muestras.

La clase más abundante, *Gammaproteobacteria*, se encuentra presente en todas las muestras, aunque en un rango de abundancias dispar (7'69-100 %). Las clases restantes no aparecen en todas las muestras, aunque *Bacilli*, *Alphaproteobacteria* y *Betaproteobacteria* se encuentran como mínimo en cinco de ellas. Concretamente, *Bacilli* se detecta en 8 de las muestras en un rango de abundancias variable pero más estrecho (2'50-28'57 %).

En cuanto a la distribución por géneros, *Aeromonas*, pese a ser el predominante en el cómputo global, no aparece en las muestras N3 y N6 y se encuentra presente en el resto en un porcentaje muy variable (12'12-100 %). Algunos géneros aparecen concentrados en una o unas pocas de las muestras, como ocurre con *Chromobacterium* (solo presente en N25 pero con una abundancia del 56'25 %) o *Chryseobacterium* (en N3 y N10, con abundancias del 69'23 y 40'91 %, respectivamente). Las colonias no asignadas pertenecen a las muestras N8, N9, N23 y N24 y en ningún caso superan el 7 % de las secuencias obtenidas por muestra.

Resulta necesario plantearse la significación estadística de estos datos, puesto que de las colonias picadas inicialmente no se consiguió obtener el mismo número de secuencias de cada muestra y, en todo caso, el número de secuencias (entre 13 y 45) resulta reducido si consideramos cada muestra por separado.

Para ello, se calcularon las frecuencias de aparición acumuladas para cada género en cada muestra y en el conjunto de las mismas y se procedió a determinar el coeficiente  $R^2$  entre la distribución de frecuencias de cada muestra y la distribución de frecuencias global. Se obtuvo que las muestras más similares al global en cuanto a distribución de géneros eran N1, N8, N9 y N23, todas ellas con un valor de  $R^2$  superior al 0'9. La muestra N4 presentaba una relación nula con el global, ya que está plenamente dominada por un único género. El resto de muestras dieron coeficientes de correlación intermedios, con valores entre 0'5 y 0'9. En la Figura 4.4 se incluye un ejemplo del tipo de correlaciones obtenidas.



**Figura 4.4.** Gráficas obtenidas al representar los valores de las abundancias relativas acumuladas para dos muestras - A) N9 y B) N10 - frente a los valores del conjunto. Los datos se ajustaron a una recta de regresión lineal, indicando el valor del coeficiente  $R^2$ .

Estos valores confirman que la distribución encontrada en cada una de las muestras no es representativa de la población total, si bien algunas de ellas presentan una correlación significativa. Se comprobó también que el valor de  $R^2$  no presentaba relación con el número de secuencias obtenidas en cada muestra, indicando que las coincidencias entre ciertas muestras y el conjunto se deben más bien a efectos espurios, y que serían necesarias más secuencias por muestra para que la distribución encontrada tuviera una significación comparable con la global.

Teniendo en cuenta que las 14 muestras iniciales son de la misma naturaleza, proceden de la misma región geográfica y fueron tomadas en la misma fecha, se considera recomendable prestar más atención a la descripción de la población bacteriana en el conjunto de las muestras, dado el mayor tamaño muestral.



## 4.3 Comparación de resultados

### 4.3.1 Sesgos esperados *a priori*

Dado el método empleado en la aproximación basada en el 16S ARN ribosomal, debe tenerse en cuenta que los resultados van a presentar un sesgo debido a los siguientes motivos:

- **Cultivo**

En primer lugar, tal y como se mencionó en la introducción, se estima que solo un 1 % de los microorganismos son cultivables, por lo que las estrategias que eluden el requisito de obtener cultivos previos suponen una alternativa para detectar especies no identificadas con aproximaciones dependientes de cultivo. Es preciso tener en consideración que al haber extraído el ADN de colonias cultivadas se está perdiendo parte de la diversidad bacteriana presente en las muestras. No obstante, puede destacarse que las aproximaciones dirigidas a la secuenciación de metagenomas completos no son capaces de identificar todas las especies presentes en una muestra, debido precisamente a que muchas de ellas no han sido descritas con anterioridad y no están depositadas en las bases de datos. De hecho, el porcentaje de secuencias asignadas tras la secuenciación de genomas completos se encuentra actualmente alrededor del 20 %<sup>44</sup>. Esta limitación de la metagenómica, unida a las mejoras producidas en las técnicas de cultivo de microorganismos hasta ahora no cultivables<sup>30</sup>, supone que la diferencia entre el número de especies detectadas por las dos estrategias sea significativa, pero no tan marcada como podría pensarse en un principio.

En segundo lugar, es necesario tener presente que las muestras con las que se trabajó no fueron cultivadas en un medio general, sino en el medio **GVPC**, que favorece el crecimiento del género *Legionella*, debido a que se tomaron en el contexto de una investigación más amplia enfocada en la epidemiología y variabilidad genética de *L. pneumophila*. Este medio contiene tres antibióticos encaminados a disminuir el crecimiento de otros microorganismos: vancomicina, polimixina B y cicloheximida.

- La vancomicina posee efecto bactericida frente a bacterias gram-positivas, siendo especialmente efectivo frente a especies como *Staphylococcus aureus*, *Streptococcus pyogenes*, *Streptococcus viridans*, *Streptococcus pneumoniae* y *Clostridium difficile*.
- La polimixina B actúa contra las bacterias gram-negativas, como *Pseudomonas aeruginosa*, *Escherichia coli*, *Haemophilus influenzae*, *Enterobacter aerogenes*, y *Klebsiella pneumoniae*.
- La cicloheximida resulta tóxica para células eucariotas. En este caso, no influye en los resultados, puesto que no se ha estudiado la población eucariota en las muestras.

Se desconoce el impacto real que puede tener el medio sobre los resultados, ya que experimentalmente se ha comprobado que, a pesar de incluir antibióticos, al realizar los cultivos de bacterias no pertenecientes a género *Legionella* había crecimiento de colonias.

- **Tamaño muestral**

Se intentó analizar un número homogéneo y lo más representativo posible de las muestras empleadas. Sin embargo, una parte de las colonias presentes en las muestras quedaron sin analizar, y también quedó un porcentaje de colonias picadas que no pudieron ser asignadas taxonómicamente, por lo que los resultados podrían presentar cierta desviación respecto de la población total cultivada.

- **Bases de datos empleadas**

Las secuencias del 16S se buscaron en una base de datos específica para el ARN ribosomal, mientras que en la aproximación metagenómica por NGS la asignación se realizó empleando la base de nucleótidos contenida en el NCBI (<http://www.ncbi.nlm.nih.gov/>), de carácter general, lo que puede desviar los resultados hacia los organismos más representados en cada base.

- **Método de extracción del ADN**

Al estudiar una población bacteriana en su conjunto no es posible optimizar un único protocolo de extracción con el mismo rendimiento para todas las bacterias presentes, puesto que los organismos estudiados presentan distinta sensibilidad a la lisis. Los métodos empleados (PCR de colonia con etapa de lisis y choque térmico) no son tan efectivos como otros métodos enzimáticos o kits comerciales disponibles en el mercado, por lo que es probable que bacterias más difíciles de lisar, como las gram positivas, estén menos representadas en los resultados<sup>29</sup>.

- **Amplificación diferencial**

Los primers empleados para amplificar el gen del 16S se consideran teóricamente “universales”, es decir, capaces de amplificar dicho gen en todas las especies bacterianas. Sin embargo, se sabe que en la práctica no amplifica con la misma eficacia en todas las especies, debido al grado de complementariedad entre el cebador y la secuencia de unión, que condiciona la eficiencia del anillamiento de los cebadores, y la influencia de las secuencias flanqueantes a los sitios de unión<sup>20</sup>.

Según el protocolo experimental aplicado, se espera que el efecto de la amplificación diferencial del 16S sea mínimo, puesto que la amplificación no se ha realizado directamente de una mezcla compleja de genomas, sino que se han cultivado previamente y se han picado colonias individuales. De esta forma, aunque se obtengan más copias de una colonia concreta, esto no aumentará la representación de dicha especie.

- **Desviación en la secuenciación mediante NGS**

En las estrategias metagenómicas basadas en la secuenciación masiva del conjunto de genomas presentes en una muestra solo se llega a asignar una fracción de la población total, existiendo un sesgo hacia las especies más abundantes y pudiendo no detectarse las especies menos representadas en la comunidad microbiana. En este estudio concreto, en el abordaje basado en NGS se asignaron entre el 10 y el 30 % de las lecturas obtenidas<sup>44</sup>.

#### **4.3.2 Similitudes y diferencias entre la aproximación metagenómica y el análisis basado en el gen 16S**

##### **Filos**

En la estrategia metagenómica mediante NGS se detectaron 35 fillos, aunque cuatro de ellos explicaban el 97 % de la abundancia relativa acumulada. Dos de estos fillos principales coincidían con los identificados mediante el análisis del 16S, y representaban porcentajes relativamente similares según ambas estrategias: *Proteobacteria* (77'85 % según NGS; 86'54 % en función del 16S) y *Bacteroidetes* (4'48 % según NGS; 7'20 % en función del 16S).

A continuación se exponen posibles motivos por los cuales no se han detectado mediante la secuenciación del 16S los otros dos fillos mayoritarios presentes en la descripción de la población mediante NGS.



- Filo *Cyanobacteria*. Se trata de bacterias capaces de realizar la fotosíntesis oxigénica y presentan dificultades para su aislamiento y cultivo<sup>4,13</sup>. Se ha observado que el agar, empleado comúnmente en los medios de cultivo como agente solidificante, inhibe el crecimiento de las cianobacterias<sup>1</sup>. Por tanto, para su cultivo se emplean medios con bajas concentraciones de agar o se aplican distintos procedimientos para minimizar este efecto. Además, estas bacterias requieren nutrientes específicos, como sales minerales (MgSO<sub>4</sub>, CuSO<sub>4</sub>, etc.) a baja concentración<sup>41</sup>, y deben ser cultivadas en condiciones determinadas de temperatura (entre 18 y 37° C, dependiendo de su procedencia) y de intensidad lumínica, en ciclos de luz y oscuridad de 12 horas.

Debido a estas condiciones de cultivo específicas, es lógico que no se haya detectado su presencia en el análisis del 16S, puesto que los cultivos fueron realizados en medios y condiciones no apropiados para el desarrollo de cianobacterias. En este caso, la aproximación por NGS debería ofrecer resultados más exactos, puesto que elude el paso previo del cultivo y sus dificultades intrínsecas.

- Filo *Actinobacteria*. Es uno de los filos con mayor número y variedad de especies identificadas dentro del dominio *Bacteria*<sup>51</sup>, lo cual lleva a descartar que no haya sido detectado por baja representación en la base de datos. La razón parece ser más bien la misma que en el caso de las cianobacterias: poseen requerimientos específicos de cultivo<sup>64</sup> que les impiden crecer de forma óptima en el medio empleado. Por ejemplo, uno de los sustratos recomendados es el medio complejo 5006, que contiene sacarosa, dextrina, extracto de carne, caldo de triptona y sales minerales, componentes de los que carece el GVPC. Otro motivo es que, al tratarse de bacterias gram-positivas, son más difíciles de lisar y pueden haber quedado en la fracción de colonias de las cuales no se ha obtenido amplificado.

El tercer filo detectado mediante la aproximación del 16S, *Firmicutes*, aparece con una abundancia relativa del 6'25 %, mientras que según la estrategia de NGS no supera el 1 %. En este caso, la frecuencia real de aparición podría estar infravalorada, tal y como se detalla a continuación.

- Filo *Firmicutes*. Se sabe que son un grupo predominante dentro del dominio *Bacteria*; según investigaciones previas basadas en métodos tanto dependientes como independientes de cultivo, es el segundo filo más abundante<sup>22</sup>. Sin embargo, en el análisis basado en el 16S aparece en un porcentaje relativamente bajo (6'25 %) y en la estrategia de metagenómica mediante NGS ni siquiera alcanza el 1 % de la población (0'66 %). Una posible justificación es el hecho de que muchas especies del filo *Firmicutes* son formadoras de endosporas y se conoce que estas estructuras son resistentes a muchos de los métodos tradicionales de extracción de ADN, por lo que pueden no ser detectadas en las estrategias metagenómicas<sup>14</sup>. Por ejemplo, en el estudio previo realizado por Filippidou *et al.*<sup>14</sup> en muestras de biofilm se analizó esta desviación en conjuntos de datos metagenómicos y se probó un método de extracción dirigido especialmente a las endosporas. Al estudiar la diversidad de la muestra mediante la amplificación del 16S y otros genes específicos se detectaron especies formadoras de esporas en elevado porcentaje. Sin embargo, al aplicar el método de secuenciación de genomas completos por *shotgun* solo quedaba reflejada una pequeña proporción de los *Firmicutes* presentes. La explicación que se ofrece a este hecho es que al estudiar la población microbiana basándose en la amplificación y secuenciación del 16S ribosomal se aumenta la proporción de *Firmicutes* detectados, debido a que este filo contiene en promedio unas 7 copias del 16S. Además, se sugiere que las herramientas para el análisis de los metagenomas no están optimizadas para la detección de ciertos taxones, como es el caso de los *Firmicutes*, de forma que la secuenciación de genomas completos da una estimación de estas especies menor a la real.

### Géneros

Respecto a este nivel taxonómico, la estrategia metagenómica mediante NGS reveló 927 géneros, 176 de ellos con una abundancia relativa superior al 1 % en al menos una de las muestras. Aparece una elevada diversidad de géneros minoritarios, ya que, aunque individualmente se encuentren en porcentajes inferiores al 1 %, en conjunto suponen casi la mitad de la población (45'50 %).

En el análisis del 16S se detectaron 26 géneros, 12 de los cuales presentaban abundancias superiores al 1 %. En este caso los géneros mayoritarios explican mayor parte de la diversidad de la población, mientras que los 14 géneros minoritarios representan en conjunto el 6'25 % de la comunidad microbiana.

Se dispone también del perfil taxonómico obtenido a partir de la selección de los genes 16S ARNr de los datos metagenómicos y su búsqueda en bases de datos. Mediante esta aproximación se detectaron 255 géneros, 26 de ellos con frecuencias de aparición superiores al 1 %. De forma similar a los resultados de los metagenomas, dado el elevado número de géneros menos abundantes, éstos representan una parte significativa de la población (37'25 %).

En la Tabla 4.4 se listan los 12 géneros mayoritarios detectados mediante la secuenciación del gen 16S, comparando las abundancias obtenidas mediante cada estrategia.

**Tabla 4.4.** Géneros con abundancia relativa superior al 1 % detectados en el análisis del 16S, ordenados en función de su abundancia. Se indica la abundancia obtenida en las estrategias de NGS (metagenomas), análisis del 16S y la basada en las secuencias del 16S extraídas de los metagenomas. Se han destacado los géneros con frecuencia de aparición > 1% detectados en dos o más de las estrategias. Se incluye también el número de secuencias depositadas en la base RDP para cada género y el número de éstas que proceden de aislados.

Géneros	Abundancia relativa (%)			Secuencias depositadas en RDP	Secuencias de aislados depositadas en RDP
	16S	Metagenomas	16S de los metagenomas		
<i>Aeromonas</i>	53'61	0'12	0'05	5417	1893
<b><i>Pseudomonas</i></b>	<b>6'73</b>	<b>1'44</b>	0'33	64013	27711
<i>Chryseobacterium</i>	6'49	0'15	0'36	1745	629
<i>Delftia</i>	5'05	0'10	0'01	3590	472
<b><i>Brevundimonas</i></b>	<b>4'57</b>	<b>1'36</b>	<b>1'40</b>	2614	655
<i>Acinetobacter</i>	3'85	0'32	0'27	12248	3867
<i>Serratia</i>	3'13	0'06	-	3226	998
<i>Chromobacterium</i>	2'16	0'04	-	133	97
<i>Enterobacter</i>	1'92	0'04	-	4350	1647
<i>Bacillus</i>	1'68	0'08	0'02	38315	24121
<i>Paenibacillus</i>	1'44	0'03	0'01	4218	2325
<b><i>Sphingomonas</i></b>	<b>1'44</b>	<b>5'38</b>	<b>6'43</b>	13691	2422

Como puede observarse en la Tabla 4.4, tan solo dos géneros de los detectados como mayoritarios en el análisis del 16S (*Brevundimonas* y *Sphingomonas*) aparecen en porcentajes superiores al 1 % en las tres estrategias. En el caso de *Pseudomonas*, la coincidencia se da entre el análisis de 16S y la aproximación metagenómica, mientras que en la tercera estrategia empleada es minoritario.

Tal y como se esperaba en un principio, se dan variaciones entre las abundancias relativas detectadas, puesto que en el análisis del 16S ribosomal se ha estudiado una fracción de la población total presente en las muestras y se cuenta con el sesgo producido en la etapa de cultivo.

Algunas de las diferencias encontradas pueden deberse al número de secuencias depositadas en la base de datos empleada (RDP), concretamente al número de secuencias procedentes de aislados cultivados, puesto que las secuencias buscadas proceden también de cultivo. Puede plantearse que los géneros *Pseudomonas*, *Acinetobacter* y *Bacillus*, principalmente, podrían aparecer sobrerrepresentados debido a que existe un elevado número de secuencias de estos géneros depositadas en la RDP. No obstante, serían necesarios otros estudios para confirmar esta hipótesis, ya que las diferencias observadas podrían deberse a otros de los factores mencionados anteriormente. En el caso de las secuencias de ARN ribosomal se emplearon otras bases de datos específicas, por lo que no puede contrastarse este planteamiento con los valores obtenidos mediante esta estrategia.

De entre los 14 géneros restantes identificados por secuenciación del 16S, 9 se han detectado mediante las tres estrategias en porcentajes inferiores al 1 %, uno de ellos (*Roseomonas*) aparece en porcentajes ligeramente superiores en los metagenomas y las secuencias de 16S extraídas de los mismos y el género *Stenotrophomonas* se detecta solo en la secuenciación del 16S y los metagenomas.

Además, se han detectado tres géneros por la estrategia de 16S que no aparecen en el análisis de los metagenomas —*Aquamicrobium*, *Empedobacter* y *Virgibacillus*— aunque cabe destacar que los dos últimos se asignaron con menor precisión, tomando como umbral un *S<sub>ab</sub> score* superior al 95 %.

Por último, resulta destacable la elevada abundancia relativa del género *Aeromonas*, que según la secuenciación del 16S representa más de la mitad de la comunidad microbiana, mientras que analizando los metagenomas aparece en un porcentaje muy inferior (0'12 %). Además de la variación esperable debido a la diferencia en el tamaño muestral empleado, resulta razonable pensar que el medio de cultivo GVPC favorezca el crecimiento de este género, a pesar de no estar diseñado para tal fin, y así aumente su presencia relativa al estudiar las colonias cultivadas. Por otra parte, según la bibliografía se trata de un género común en ambientes acuáticos y redes de agua potable<sup>42</sup>, por lo que quizá cabría esperar una abundancia mayor de la que aparece en los metagenomas.

Fijándonos en los resultados obtenidos de los metagenomas (Tabla 4.5), aparecen 26 géneros con una abundancia relativa total superior al 1 %, de los cuales se han detectado cinco a través del análisis del 16S directo y 22 mediante la búsqueda del 16S de los metagenomas.

**Tabla 4.5.** Géneros con abundancia relativa superior al 1 % detectados en el análisis de los metagenomas, ordenados en función de su abundancia. Se indica la abundancia obtenida en las estrategias de NGS (metagenomas), análisis del 16S y la basada en las secuencias del 16S extraídas de los metagenomas. Se han destacado los géneros con frecuencia de aparición > 1 % detectados en dos o más de las estrategias. Se incluye también el número de secuencias depositadas en la base RDP para cada género y el número de éstas que proceden de aislados.

Géneros	Abundancia relativa (%)			Secuencias depositadas en RDP	Secuencias de aislados depositadas en RDP
	Metagenomas	16S	16S de los metagenomas		
<i>Porphyrobacter</i>	6'69	-	8'66	1180	167
<i>Rhodobacter</i>	5'53	-	4'22	2097	333
<i>Sphingomonas</i>	5'38	1'44	6'43	13691	2422
<i>Xanthomonas</i>	4'05	-	0'60	877	710
<i>Rubrivivax</i>	3'17	-	0'14	167	35
<i>Sandarakinorhabdus</i>	2'68	-	1'07	154	24
<i>Sphingobium</i>	2'03	-	0'20	3060	826
<i>Acidovorax</i>	1'98	-	0'61	7878	622
<i>Flavobacterium</i>	1'55	-	2'16	11703	1597
<i>Methylibium</i>	1'54	-	0'54	46	-
<i>Leptothrix</i>	1'53	-	0'02	134	54
<i>Oscillatoria</i>	1'49	-	-	-	-
<i>Pseudomonas</i>	1'44	6'73	0'33	64013	27711
<i>Brevundimonas</i>	1'36	4'57	1'40	2614	655
<i>Bradyrhizobium</i>	1'32	-	0'16	3993	1292
<i>Burkholderia</i>	1'27	-	-	16958	6490
<i>Sphingopyxis</i>	1'26	-	3'16	1111	369
<i>Methylobacterium</i>	1'26	-	0'02	5566	1213
<i>Roseomonas</i>	1'27	0'24	2'38	766	100
<i>Erythrobacter</i>	1'19	-	0'47	1463	644
<i>Variovorax</i>	1'18	-	0'27	2881	871
<i>Paracoccus</i>	1'13	-	0'23	3735	992
<i>Hydrogenophaga</i>	1'09	-	4'47	1945	125
<i>Microcoleus</i>	1'08	-	-	-	-
<i>Novosphingobium</i>	1'06	-	0'52	5431	591
<i>Stenotrophomonas</i>	1'01	0'24	-	9329	3930

Al contrario de lo que sucede en los resultados derivados de la secuenciación del 16S, la aproximación metagenómica no revela ningún género predominante en la comunidad microbiana, sino que los géneros mayoritarios se encuentran en frecuencias entre el 1 y el 10 %.

Como puede observarse en la Tabla 4.5, para algunos de los géneros detectados mediante NGS no existen secuencias del 16S depositadas en la base de datos, por lo que es lógico que no aparezcan en el análisis. El resto de géneros no detectados se explica por los factores anteriormente señalados.

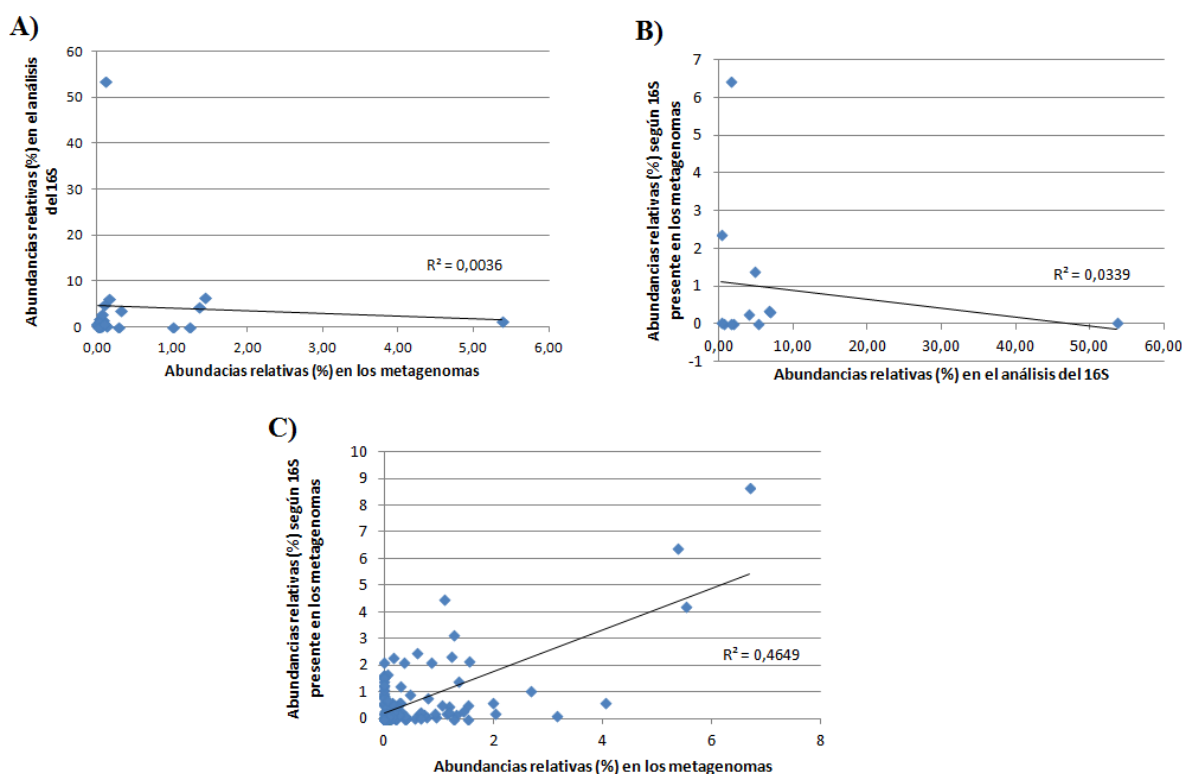
### 4.3.3 Correlación estadística

Además de describir las similitudes y diferencias observadas, se evaluó la correlación entre los resultados obtenidos mediante las tres estrategias: secuenciación de genomas completos, búsqueda activa de los genes 16S ARNr en los genomas secuenciados y análisis del 16S tras el cultivo.

Para ello, se calcularon los coeficientes de correlación  $R^2$  entre los resultados obtenidos en cada método, basándose en los valores de abundancia relativa de los géneros detectados.

Puesto que en el análisis del 16S se ha detectado una fracción de los géneros encontrados en los metagenomas, solo se han empleado para la comparación los géneros detectados por ambas estrategias. Se ha procedido de manera análoga para comparar la secuenciación del 16S con la búsqueda del gen en los metagenomas.

Los resultados pueden observarse en la Figura 4.5.



**Figura 4.5.** Gráficas obtenidas al representar los valores de las abundancias relativas de los géneros obtenidas por un método frente a los valores obtenidos por otra de las aproximaciones. Los datos se ajustaron a una recta de regresión lineal, indicando el valor del coeficiente  $R^2$ . Se compararon las estrategias dos a dos: A) Metagenomas y secuenciación del 16S; B) Secuenciación del 16S y búsqueda de este gen en los genomas; y C) Metagenomas y búsqueda del 16S en ellos.

Como puede observarse, la correlación existente entre el análisis del 16S y las otras dos estrategias es prácticamente nula (apartados A y B de la Figura 4.5), siendo en ambos casos no significativa ( $p$ -valor = 0'11 y 0'21, respectivamente). Se da una mayor correlación entre las dos aproximaciones enfocadas al 16S pero aun así ésta es cercana a cero. Estos resultados son congruentes con los sesgos esperables producidos por la metodología empleada e indican que las poblaciones obtenidas no son comparables a nivel de género.

Por otra parte, la relación entre los datos de los metagenomas y las secuencias del 16S extraídas de los mismos es mayor que en las otras comparaciones y, aunque no se trata de una correlación fuerte, resulta estadísticamente significativa ( $p$ -valor  $< 2'2 \cdot 10^{-16}$ ). Tal y como se apunta en el estudio referente a los metagenomas<sup>50</sup> estas diferencias se deben a la composición de las bases de datos y los algoritmos empleados en cada análisis.

#### 4.4 Evaluación del método experimental y propuestas de optimización

En el protocolo experimental se llegaron a aislar 612 colonias individuales de las 14 muestras, insuficiente para describir de forma precisa la población de cada muestra, pero adecuado al ámbito de este trabajo y los objetivos planteados para estudiar la comunidad microbiana en su conjunto.

Uno de los puntos que presentó más dificultades fue el de picar las colonias para la preparación de los stocks. Por una parte, al haberse mantenido los cultivos congelados a  $-80\text{ }^{\circ}\text{C}$ , algunos de ellos estaban deteriorados y durante el proceso se observó la formación de líquido en los medios, lo que imposibilitó aislar un mayor número de colonias, puesto que algunas se desprendían del medio y quedaban mezcladas entre sí. Se comprobó que la manera de proceder más efectiva era transferir las placas de cultivo almacenadas a  $-80\text{ }^{\circ}\text{C}$  a un congelador a  $-20\text{ }^{\circ}\text{C}$  y posteriormente dejar que se descongelaran a temperatura ambiente en el interior de una cabina de seguridad biológica, revisándolas periódicamente para ir eliminando el excedente de líquido. Por otra parte, al trabajar con cultivos heterogéneos y con una densidad de colonias variable, en algunos casos las colonias habían crecido solapadas y no fue posible incluirlas en el estudio porque no podían picarse de manera aislada.

Hay que tener en cuenta que los cultivos de partida se realizaron previamente a este estudio, por lo que no estaban optimizados para tal fin. En caso de plantear un análisis similar partiendo de nuevas muestras, se recomendaría emplear un medio de cultivo general para evitar el sesgo producido por el medio específico para *Legionella* con el que se ha trabajado. Para facilitar el aislamiento de las colonias sería conveniente cultivar un volumen adecuado para obtener una densidad de siembra tal que las colonias crezcan separadas una de otras, o bien subcultivar hasta obtener cultivos puros de cada microorganismo, aunque esta última opción implicaría mayor consumo tanto de medios como de tiempo. En el estudio de Drancourt *et al.*<sup>44</sup> se indica la necesidad de obtener cultivos puros antes de analizar el gen 16S. Sin embargo, puede argumentarse que en este estudio el empleo de cultivos puros podría haber aumentado el número de colonias identificadas, pero no ha resultado imprescindible.

De las colonias aisladas, se obtuvieron 328 amplicones de las regiones V1-V4 del 16S mediante la PCR de colonia y 193 amplicones tras aplicar la extracción por choque térmico y ajustar la concentración, lo que da un total de 521 secuencias amplificadas, tal y como puede observarse en la Tabla 4.6.

Se considera adecuado haber aplicado los dos métodos, pues la PCR de colonia inicial permitió obtener amplicones de forma más sencilla, mientras que con la segunda amplificación se consiguió incrementar significativamente el número de secuencias obtenidas. Podría haberse realizado un tercer ciclo de amplificación, aplicando un método de extracción para bacterias más resistentes a la lisis, pero se considera que no habría aumentado en gran medida el número de secuencias amplificadas, puesto que resulta más probable que de las colonias picadas inicialmente haya un porcentaje del cual no se haya conseguido extraer suficiente ADN para la amplificación, o bien que éste estuviera degradado.

Como puntualización, cabe destacar que el paso de centrifugación de los stocks previo a la reacción de PCR resultó ser esencial para obtener una correcta amplificación de la secuencia de interés. En pruebas realizadas previamente se comprobó que la omisión de este paso daba lugar a un *smear* en la electroforesis de los productos de PCR, debido a la elevada cantidad de material bacteriano presente.

**Tabla 4.6.** Registro del número de colonias picadas por cada muestra inicial, el número de amplicones tras la PCR de colonia (amplificado I), tras el choque térmico (amplificado II) y totales, así como el número de secuencias consenso finalmente obtenidas. Se incluyen dos rendimientos: el rendimiento I hace referencia al ratio de secuencias obtenidas en función del número de colonias picadas, mientras que el rendimiento II describe el ratio de secuencias obtenidas en función del número de amplicones generados. Se incluyen en la última fila los valores referidos al cómputo global de las 14 muestras.

Muestra	Colonias picadas	Amplificado I	Amplificado II	Amplificado total	Secuencias obtenidas	Rendimiento I (%)	Rendimiento II (%)
N1	48	26	12	38	33	68'75	86'84
N2	50	30	6	36	19	38'00	52'78
N3	15	15	0	15	13	86'67	86'67
N4	31	20	10	30	28	90'32	93'33
N5	48	24	3	27	24	50'00	88'89
N6	48	33	11	44	28	58'33	63'64
N7	48	23	22	45	40	83'33	88'89
N8	48	6	40	46	21	43'75	45'65
N9	48	40	8	48	43	89'58	89'58
N10	48	47	0	47	44	91'67	93'62
N11	48	36	11	47	41	85'42	87'23
N23	48	21	25	46	45	93'75	97'83
N24	48	6	25	31	21	43'75	67'74
N25	36	1	20	21	16	44'44	76'19
<b>14</b>	<b>612</b>	<b>328</b>	<b>193</b>	<b>521</b>	<b>416</b>	<b>67'97</b>	<b>79'85</b>

De las 521 secuencias amplificadas, finalmente se obtuvieron 416 secuencias consenso, lo que supone un 68 % de las colonias picadas inicialmente y un 80 % de las secuencias amplificadas. Las razones principales por las que no se consiguió obtener una secuencia de suficiente calidad en el 20 % de los amplicones restantes son las siguientes:

- Tal y como se mencionó en el apartado de Materiales y Métodos del presente trabajo, se anotaron algunas de las colonias que daban una banda tenue en la electroforesis para posteriormente añadir un mayor volumen de producto de PCR purificado en la reacción de secuenciación. No obstante, al analizar los cromatogramas obtenidos de estas secuencias, la mayoría de ellos resultaron ser de baja calidad y/o presentar una longitud claramente inferior a los 750 nucleótidos que representan las regiones amplificadas, aproximadamente. Por tanto, se considera que habría sido preferible amplificar de nuevo estas secuencias antes de secuenciar.
- Uno de los pasos que se podría optimizar para aumentar el rendimiento de secuencias obtenidas es el de la purificación. Para la eliminación de contaminantes y lavado del ADN se añadió agua autoclavada por ser el protocolo seguido en otros ensayos del grupo de trabajo, pero se podría sustituir por tampón de elución para conseguir mejores resultados. Asimismo, se plantea la conveniencia de aumentar los tiempos de centrifugación, debido a la observación de que quedaba un volumen de producto purificado mayor del esperado, por lo que en algunos casos el ADN podría estar diluido en exceso para la secuenciación posterior.



## 5. CONCLUSIONES

El análisis basado en la secuenciación de las regiones V1-V4 del gen 16S ribosomal a partir de colonias aisladas de cultivos permitió asignar taxonómicamente la mayoría de secuencias (> 98 %) hasta el nivel de género. Sin embargo, se encontraron dificultades en la identificación a nivel de especie, pudiéndose asignar menos del 9 % de las colonias estudiadas. Los resultados obtenidos indican que el análisis de regiones hipervariables del 16S es útil para el estudio de la diversidad bacteriana presente en las muestras hasta el nivel de género, pero su resolución resulta insuficiente para obtener una asignación precisa a nivel de especie.

Basándose en la secuenciación del 16S, la población bacteriana estudiada se encuentra dividida en 3 filos, 6 clases, 13 órdenes, 19 familias y 26 géneros. Entre los filos encontrados predomina *Proteobacteria* con una abundancia relativa superior al 80 %, concretamente la clase *Gammaproteobacteria*, que representa un 70 % de la población total. En cuanto a órdenes y familias, destacan *Aeromonadales* y *Aeromonadaceae*, respectivamente, ambos con una abundancia relativa superior al 50 %. Respecto a los géneros, destaca la abundancia de *Aeromonas* (53'61 %), indicando la posibilidad de que el medio de cultivo empleado (GVPC) favorezca su crecimiento.

La comunidad bacteriana se halla heterogéneamente distribuida en las 14 muestras iniciales y, dado el reducido número de colonias estudiadas por muestra, cada una de ellas resulta poco representativa de la población total, por lo que resulta preferible centrarse en la descripción de la población del conjunto de las muestras.

Respecto a la identificación de patógenos, no se ha detectado la presencia del género *Legionella*, endémico de la región de procedencia de las muestras analizadas. No obstante, resulta sorprendente el elevado número de géneros potencialmente patógenos encontrados, aunque los datos deben interpretarse con precaución, debido al sesgo de las bases de datos hacia organismos patógenos, y serían necesarios estudios más concretos para confirmar la presencia de agentes que supongan un riesgo de salud pública en las fuentes naturales donde se tomaron las muestras.

Comparando los resultados obtenidos con los derivados de la estrategia metagenómica mediante NGS y la búsqueda activa de secuencias del 16S extraídas de los metagenomas, se han encontrado algunas similitudes puntuales, sobre todo en taxones superiores, como es el caso de los filos (por ejemplo, los tres enfoques indican que *Proteobacteria* es el filo predominante). Sin embargo, en taxones más específicos las coincidencias encontradas son de carácter menor y no se ha encontrado una correlación estadística al contrastar los datos referidos a los géneros.

Como principales razones de las diferencias encontradas pueden destacarse las siguientes:

- Etapa de cultivo. Se considera que es una de las fuentes principales del sesgo, tanto por el hecho de que existe un amplio porcentaje de microorganismos no cultivables, como por haber empleado un medio específico para *Legionella*, que puede inhibir el crecimiento de algunas especies.
- Tamaño muestral. Es otro de los factores clave, puesto que en la secuenciación del 16S solo fue posible estudiar una fracción de los organismos analizados mediante NGS.
- Bases de datos empleadas. Las secuencias de ARN ribosomal se buscaron en una base de datos específica, lo que puede desviar los resultados hacia los organismos más representados en la base.
- Método de extracción del ADN. Las bacterias más resistentes a la lisis pueden haber quedado menos representadas en los resultados.



Los datos obtenidos son congruentes con estudios previos en los cuales se señala que la caracterización de poblaciones microbianas obtenida aplicando estrategias metagenómicas mediante NGS no es directamente comparable con el perfil taxonómico obtenido al basarse en el gen 16S, incluso eliminando el sesgo debido al tamaño muestral. De hecho, fijándonos en los datos generados al extraer las secuencias del ARN ribosomal de los metagenomas, la población que se obtiene difiere de la descrita por la aproximación de NGS, a pesar de que en esta comparación la población es la misma.

Por último, los resultados indican la necesidad de tener en cuenta los sesgos presentes en cada estrategia aplicada y de avanzar hacia la optimización de herramientas informáticas, algoritmos y bases de datos con objeto de minimizar las diferencias que se obtienen en función de la aproximación empleada. Asimismo, parecen indicar la conveniencia de complementar distintas aproximaciones para mejorar la comprensión de las poblaciones microbianas presentes en los biofilms.

## 6. REFERENCIAS BIBLIOGRÁFICAS

1. ALLEN EAD, GORHAM PR. Culture of planktonic cyanophytes on agar. En: W. W. Carmichael. The water environment: algal toxins and health. Nueva York; 1981. 185-192. doi: 10.1007/978-1-4613-3267-1\_14
2. ASH C, FARROW JA, DORSCH M, STACKEBRANDT E, COLLINS MD. Comparative analysis of *Bacillus anthracis*, *Bacillus cereus*, and related species on the basis of reverse transcriptase sequencing of 16S rRNA. *Int J Syst Bacteriol*. 1991 Jul; 41 (3):343-346
3. BURMØLLE M, WEBB JS, RAO D, HANSEN LH, SØRENSEN SJ, KJELLEBERG S. Enhanced biofilm formation and increased resistance to antimicrobial agents and bacterial invasion are caused by synergistic interactions in multispecies biofilms. *Appl Environ Microbiol*. 2006 Jun; 72 (6): 3916-3923
4. CASTENHOLZ RW. Culturing methods for cyanobacteria. *Methods Enzymol*. 1988; 167: 68-93
5. CHAO Y, MAO Y, WANG Z, ZHANG T. Diversity and functions of bacterial community in drinking water biofilms revealed by high-throughput sequencing. *Sci Rep*. 2015 Jun 12; 5: 10044. doi: 10.1038/srep10044
6. CHRISTENSEN BB, STERNBERG C, ANDERSEN JB, EBERL L, MOLLER S, GIVSKOV M *et al*. Establishment of new genetic traits in a microbial biofilm community. *Appl Environ Microbiol*. 1998 Jun; 64 (6): 2247-2255
7. COLE JR, CHAI B, FARRIS RJ, WANG Q, KULAM SA, MCGARRELL DM *et al*. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res*. 2005 Jan 1; 33 (Database issue): D294–D296. doi: 10.1093/nar/gki038
8. COSTERTON JW, CHENG KJ, GEESEY GG, LADD TI, NICKEL JC, DASGUPTA M *et al*. Bacterial biofilms in nature and disease. *Annu Rev Microbiol*. 1987; 41: 435-464
9. COSTERTON JW, GEESEY GG, CHENG KJ. How bacteria stick. *Sci Am*. 1978 Jan; 238 (1): 86-95
10. DONLAN RM, COSTERTON JW. Biofilms: survival mechanisms of clinically relevant microorganisms. *Clin Microbiol Rev*. 2002 Apr; 15 (2): 167-193
11. DRANCOURT M, BOLLET C, CARLIOZ A, MARTELIN R, GAYRAL JP, RAOULT D. 16S ribosomal DNA sequence analysis of a large collection of environmental and clinical unidentifiable bacterial isolates. *J Clin Microbiol*. 2000 Oct; 38 (10): 3623-3630
12. FARELLY V, RAINEY FA, STACKEBRANDT E. Effect of genome size and *rrn* gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. *Appl Environ Microbiol*. 1995; 61: 2798-2801
13. FERRIS MJ, HIRSCH CF. Method for isolation and purification of cyanobacteria. *Appl Environ Microbiol*. 1991 May; 57 (5): 1448-1452
14. FILIPPIDOU S, JUNIER T, WUNDERLIN T, LO CC, LI PE, CHAIN PS *et al*. Under-detection of endospore-forming *Firmicutes* in metagenomic data. *Comput Struct Biotechnol J*. 2015 Apr 25; 13: 299-306. doi: 10.1016/j.csbj.2015.04.002
15. FOX GE, PECHMAN KR, WOESE CR. Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to procaryotic systematics. *Int J Syst Bacteriol*. 1977; 27: 44–57
16. FOX GE, WISOTZKEY JD, JURTSCHUK P Jr. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int. J. Syst. Bacteriol*. 1992 Jan; 42 (1) :166–170
17. GILBERT JA, JANSSON JK, KNIGHT R. The Earth Microbiome project: successes and aspirations. *BMC Biol*. 2014 Aug 22; 12: 69. doi: 10.1186/s12915-014-0069-1
18. GÓMEZ-ÁLVAREZ V. Biofilm-growing bacteria involved in the corrosion of concrete wastewater pipes: protocols for comparative metagenomic analyses. *Methods Mol Biol*. 2014; 1147: 323-340. doi: 10.1007/978-1-4939-0467-9\_23

19. HALL-STOODLEY L, COSTERTON JW, STOODLEY P. Bacterial biofilms: from the natural environment to infectious diseases. *Nat Rev Microbiol.* 2004 Feb; 2 (2): 95-108
20. HANSEN MC, TOLKER-NEILSON T, GIVSKOV M, MOLIN S. Biased 16S rDNA PCR amplification caused by interference from DNA flanking template region. *FEMS Microbiol Ecol.* 1998; 26: 141-149
21. HAUSNER M, WUERTZ S. High rates of conjugation in bacterial biofilms as determined by quantitative in situ analysis. *Appl Environ Microbiol.* 1999 Aug; 65 (8):3710-3713
22. HUGENHOLTZ P. Exploring prokaryotic diversity in the genomic era. *Genome Biol.* 2002; 3 (2): REVIEWS0003
23. KALYUZHNYAYA MG, LAPIDUS A, IVANOVA N, COPELAND AC, MCHARDY AC, SZETO E *et al.* High-resolution metagenomics targets specific functional types in complex microbial communities. *Nat Biotechnol.* 2008 Sep; 26(9): 1029-34. doi: 10.1038/nbt.1488
24. KARSENTI E, ACINAS SG, BORK P, BOWLER C, DE VARGAS C, RAES J *et al.* A holistic approach to marine eco-systems biology. *PLoS Biol.* 2011 Oct; 9 (10): e1001177. doi: 10.1371/journal.pbio.1001177
25. KIM M, MORRISON M, YU Z. Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *J Microbiol Meth.* 2011; 84 (1): 81-87. doi:10.1016/j.mimet.2010.10.020
26. KUCZYNSKI J, LAUBER CL, WALTERS WA, PARFREY LW, CLEMENTE JC, GEVERS D *et al.* Experimental and analytical tools for studying the human microbiome. *Nat Rev Genet.* 2011 Dec 16; 13 (1): 47-58. doi: 10.1038/nrg3129
27. LEWIS CM Jr, OBREGÓN-TITO A, TITO RY, FOSTER MW, SPICER PG. The Human Microbiome Project: lessons from human genomics. *Trends Microbiol.* 2012 Jan; 20 (1): 1-4. doi: 10.1016/j.tim.2011.10.004
28. LI YH, LAU PC, LEE JH, ELLEN RP, CVITKOVITCH DG. Natural genetic transformation of *Streptococcus mutans* growing in biofilms. *J Bacteriol.* 2001 Feb; 183 (3): 897-908
29. LILES MR, MANSKE BF, BINTRIM SB, HANDELSMAN J, GOODMAN RM. A census of rRNA genes and linked genomic sequences within a soil metagenomic library. *Appl Environ Microbiol.* 2003 May; 69 (5): 2684-2691
30. LING LL, SCHNEIDER T, PEOPLES AJ, SPOERING AL, ENGELS I, CONLON BP *et al.* A new antibiotic kills pathogens without detectable resistance. *Nature.* 2015 Jan 22; 517 (7535): 455-459. doi: 10.1038/nature14098
31. LIU Z, LOZUPONE C, HAMADY M, BUSHMAN FD, KNIGHT R. Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res.* 2007; 35 (18): e120.
32. LU X, ZHANG XX, WANG Z, HUANG K, WANG Y, LIANG W *et al.* Bacterial Pathogens and Community Composition in Advanced Sewage Treatment Systems Revealed by Metagenomics Analysis Based on High-Throughput Sequencing. *PLoS One.* 2015; 10 (5): e0125549. doi: 10.1371/journal.pone.0125549
33. MCCOY WF, BRYERS JD, ROBBINS J, COSTERTON JW. Observations of fouling biofilm formation. *Can J Microbiol.* 1981 Sep; 27 (9): 910-917
34. MCLEAN RJ, KAKIRDE KS. Enhancing metagenomics investigations of microbial interactions with biofilm technology. *Int J Mol Sci.* 2013 Nov 11; 14 (11): 22246-22257. doi: 10.3390/ijms141122246
35. MORGAN JL, DARLING AE, EISEN JA. Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS One.* 2010 Apr 16; 5 (4): e10209. doi: 10.1371/journal.pone.0010209
36. MUYZER G, DE WAAL EC, UITTERLINDEN AG. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl Environ Microbiol.* 1993 Mar; 59 (3): 695-700
37. ORGANIZACIÓN MUNDIAL DE LA SALUD. Manual de bioseguridad en el laboratorio. 3ª ed. Ginebra: Organización Mundial de la Salud; 2005. [Disponible on-line en:

- [http://www.who.int/csr/resources/publications/biosafety/WHO\\_CDS\\_CSR\\_LYO\\_2004\\_11/en/](http://www.who.int/csr/resources/publications/biosafety/WHO_CDS_CSR_LYO_2004_11/en/)
38. OULAS A, PAVLOUDI C, POLYMENAKOU P, PAVLOPOULOS GA, PAPANIKOLAOU N, KOTOULAS G *et al.* Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies. *Bioinform Biol Insights*. 2015; 9: 75–88. doi: 10.4137/BBLI.S12462
  39. QUAIST C, PRUESSE E, YILMAZ P, GERKEN J, SCHWEER T, YARZA P *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. 2013 Jan; 41 (Database issue): D590-596. doi: 10.1093/nar/gks1219
  40. RELMAN DA, SCHMIDT TM, MACDERMOTT RP, FALKOW S. Identification of the uncultured bacillus of Whipple's disease. *N Engl J Med*. 1992 Jul 30; 327 (5): 293–301
  41. RIPPKA R. Isolation and purification of cyanobacteria. *Methods Enzymol*. 1988; 167: 3-27
  42. ROBERTSON BK, HARDEN C, SELVARAJU SB, PRADHAN S, YADAV JS. Molecular Detection, Quantification, and Toxigenicity Profiling of *Aeromonas* spp. in Source- and Drinking-Water. *Open Microbiol J*. 2014; 8: 32–39. doi: 10.2174/1874285801408010032
  43. SÁNCHEZ-BUSÓ L, COMAS I, JORQUES G, GONZÁLEZ-CANDELAS F. Recombination drives genome evolution in outbreak-related *Legionella pneumophila* isolates. *Nat Genet*. 2014 Nov; 46 (11): 1205-11. doi: 10.1038/ng.3114
  44. SÁNCHEZ-BUSÓ, L. Genetic and genomic variability of *Legionella pneumophila*: applications to molecular epidemiology and public health [Tesis doctoral no publicada]. Valencia: Universidad Politécnica de Valencia; 2015 [Disponible en <http://hdl.handle.net/10251/52854>]
  45. SCHMEISSER C, STÖCKIGT C, RAASCH C, WINGENDER J, TIMMIS KN, WENDEROTH DF *et al.* Metagenome survey of biofilms in drinking-water networks. *Appl Environ Microbiol*. 2003 Dec; 69 (12): 7298-7309
  46. SCHMITT J, NIVENS D, WHITE DC, FLEMMING HC. Changes of biofilm properties in response to sorbed substances: An FTIR-ATR study. *Water Sci. Technol*. 1995; 32: 149–155
  47. SCHUSTER SC. Next-generation sequencing transforms today's biology. *Nat Methods*. 2008 Jan; 5 (1): 16-18. doi: 10.1038/nmeth1156
  48. SHAH N, TANG H, DOAK TG, YE Y. Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics. *Pac Symp Biocomput*. 2011: 165-176
  49. SHAW JL, MONIS P, FABRIS R, HO L, BRAUN K, DRIKAS M *et al.* Assessing the impact of water treatment on bacterial biofilms in drinking water distribution systems using high-throughput DNA sequencing. *Chemosphere*. 2014 Dec; 117: 185-192. doi: 10.1016/j.chemosphere.2014.06.077
  50. STACKEBRANDT E, GOEBEL BM. Taxonomic note: a place for DNA–DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Bacteriol*. 1994; 44 (4): 846–849
  51. STACKEBRANDT E, SPROER C, RAINEY FA, BURGHARDT J, PÄUKER O, HIPPE H. Phylogenetic analysis of the genus *Desulfotomaculum*: evidence for the misclassification of *Desulfotomaculum guttoideum* and description of *Desulfotomaculum orientis* as *Desulfosporosinus orientis* gen. nov., comb. nov. *Int J Syst Bacteriol*. 1997 Oct; 47 (4): 1134-1139
  52. STADEN R. The Staden sequence analysis package. *Mol Biotechnol*. 1996 Jun; 5 (3): 233-241
  53. STAGER CE, DAVIS JR. Automated systems for identification of microorganisms. *Clin Microbiol Rev*. 1992 Jul; 5 (3): 302-327
  54. STREIT WR, SCHMITZ RA. Metagenomics--the key to the uncultured microbes. *Curr Opin Microbiol*. 2004 Oct; 7 (5): 492-498
  55. STROUS M, FUERST JA, KRAMER EH, LOGEMANN S, MUYZER G, VAN DE PAS-SCHOONEN KT *et al.* Missing lithotroph identified as new planctomycete. *Nature*. 1999 Jul 29; 400 (6743): 446-449
  56. TAMURA K, PETERSON D, PETERSON N, STECHER G, NEI M, KUMAR S. MEGA5: molecular evolutionary genetics analysis using maximum

- likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 2011 Oct; 28 (10): 2731-2739. doi: 10.1093/molbev/msr121
57. TANG YW, ELLIS NM, HOPKINS MK, SMITH DH, DODGE DE, PERSING DH. Comparison of phenotypic and genotypic techniques for identification of unusual aerobic pathogenic gram-negative bacilli. *J Clin Microbiol.* 1998 Dec; 36 (12): 3674-3679
58. TRINGE SG, RUBIN EM. Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet.* 2005 Nov; 6 (11): 805-814
59. TYSON GW, CHAPMAN J, HUGENHOLTZ P, ALLEN EE, RAM RJ, RICHARDSON PM *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature.* 2004 Mar 4; 428 (6978): 37-43
60. VENTER JC, REMINGTON K, HEIDELBERG JF, HALPERN AL, RUSCH D, EISEN JA *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science.* 2004 Apr 2; 304 (5667): 66-74
61. WANG Q, GARRITY GM, TIEDJE JM, COLE JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol.* 2007; 73 (16): 5261-5267
62. WEISBURG WG, BARNS SM, PELLETIER DA, LANE DJ. 16S ribosomal DNA amplification for phylogenetic study. *J Bacteriol.* 1991 Jan; 173 (2): 697-703
63. WILLIAMSON SJ, Yooseph S. From bacterial to microbial ecosystems (metagenomics). *Methods Mol Biol.* 2012; 804: 35-55. doi: 10.1007/978-1-61779-361-5\_3
64. WINK J. Compendium of Actinobacteria [On-line]. Disponible en: <https://www.dsmz.de/bacterial-diversity/compendium-of-actinobacteria.html> [Consultado en julio de 2015]
65. WOOLEY JC, YE Y. Metagenomics: Facts and Artifacts, and Computational Challenges. *J Comput Sci Technol.* 2009 Jan; 25 (1): 71-81. doi: 10.1007/s11390-010-9306-4
66. YARZA P, RICHTER M, PEPLIES J, EUZEBY J, AMANN R, SCHLEIFER KH *et al.* The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst Appl Microbiol.* 2008 Sep; 31 (4): 241-250. doi: 10.1016/j.syapm.2008.07.001
67. YARZA P, YILMAZ P, PRUESSE E, GLÖCKNER FO, LUDWIG W, SCHLEIFER KH *et al.* Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol.* 2014 Sep; 12 (9): 635-45. doi: 10.1038/nrmicro3330
68. YOUSSEF N, SHEIK CS, KRUMHOLZ LR, NAJAR FZ, ROE BA, ELSHAHED MS. Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. *Appl Environ Microbiol.* 2009; 75 (16): 5227-5236. doi: 10.1128/AEM.00592-09
69. YU Z, MORRISON M. Comparisons of different hypervariable regions of rrs genes for use in fingerprinting of microbial communities by PCR-denaturing gradient gel electrophoresis. *Appl Environ Microbiol.* 2004; 70 (8): 4800-4806

## **IMÁGENES**

- Figura 1.1: imagen original disponible en <http://medicalplanzone.com/bacterial-artificial-chromosome/>. Consultado el 17/07/2015.
- Figura 3.1: imagen original contenida en la referencia 50, cedida por la Dra. Sánchez Busó.