Estudio metagenómico de muestras de agua superficial y de sedimento procedentes de la laguna hipersalina de Pétrola (Albacete)

Guillermo Sanz Martín
Máster universitario en Bioinformática y Bioestadística UOC-UB
Área del trabajo final: Análisis e integración de datos ómicos

Nombre Consultor: Andreu Paytuví
Nombre Tutor: Juan José Gómez Alday
Nombre Profesor responsable de la asignatura: Ferran Prados Carrasco

Fecha Entrega: 05-06-2019

FICHA DEL TRABAJO FINAL

Título del trabajo:	Estudio metagenómico de muestras de agua superficial y de sedimento procedentes de la laguna hipersalina de Pétrola (Albacete)
Nombre del autor:	*Guillermo Sanz Martín*
Nombre del consultor/a:	*Andreu Paytuví*
Nombre del PRA:	Ferran Prados Carrasco
Fecha de entrega (mm/aaaa):	06/2019
Titulación:	Máster universitario en Bioinformática y Bioestadística UOC-UB
Área del Trabajo Final:	Análisis e integración de datos ómicos
Idioma del trabajo:	*Castellano*
Palabras clave	Metagenómica, Laguna hipersalina, 16S rRNA

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

La finalidad de este trabajo ha sido caracterizar desde un punto de vista microbiológico la laguna hipersalina de la región de Pétrola.

Para ello se extrajo el ADN total de muestras de agua y sedimento de siete puntos representativos de la laguna y se secuenció el gen que codifica para 16S rRNA con la tecnología de secuenciación masiva Ilumina.

El análisis bioinformático para el tratamiento de las secuencias y la asignación taxonómica se llevó a cabo gracias a la herramienta QIIME2. Para llevar a cabo este proceso fue necesario el uso de una supercomputadora. La obtención de la diversidad alfa y beta, y el análisis estadístico de abundancia diferencial fueron llevados a cabo con el software R y los paquetes vegan y DESeq2. También se ha realizado un análisis comparando los resultados obtenidos con 5 estudios de varias lagunas con diferentes salinidades.

Respecto a la diversidad alfa se ha podido observar la diferencia de diversidad que existe entre las muestras de agua y sedimento, siendo los sedimentos mucho más ricos en diversidad que las aguas. La diversidad beta mostro como tres muestras de tres puntos (dos de sedimento y una de agua) diferían respecto del resto de muestras debido a las características de la laguna y a la localización geográfica de estos puntos.

El análisis de abundancia diferencial ha mostrado que existen organismos que tienen una mayor predisposición a crecer en ambientes acuáticos. El estudio comparativo con los artículos ha mostrado que las aguas de la laguna tienen una gran riqueza microbiológica.

The purpose of this work has been to characterize the hypersaline lagoon of the region of Pétrola from a microbiological point of view.

For this, the total DNA was extracted from water and sediment samples from seven representative points of the lagoon and the gene coding for 16S rRNA was sequenced with the Ilumina massive sequencing technology.

The bioinformatic analysis for the treatment of the sequences and the taxonomic assignment was carried out with QIIME2 tool. To carry out this process it was necessary to use a supercomputer. The obtaining of alpha and beta diversity, and the statistical analysis of differential abundance was carried out with the software R. An analysis was also carried out comparing the results obtained with 5 studies of several lagoons with different salinities.

Regarding alpha diversity, it has been possible to observe the difference in diversity that exists between water and sediment samples, with sediments being much richer in microbial diversity than water. The beta diversity has shown that three samples of three points (two sediment and one water) differed with respect to the rest of the samples due to the characteristics of the lagoon and the geographical location of these points.

The analysis of differential abundance has shown that there are organisms that have a greater predisposition to grow in aquatic environments. The comparative study with the articles has shown that the waters of the lagoon have a great microbiological richness.

Abstract (in English, 250 words or less):

Índice

Lista de figuras

Lista de tablas

1. Introducción

1.1 Contexto y justificación del Trabajo

El presente Trabajo de Fin de Máster (TFM) pretende caracterizar desde un punto de vista microbiológico la laguna hipersalina que se encuentra en el término municipal de Pétrola (Albacete), a unos 30 km al este de la Ciudad Albacete. Una descripción detallada de la laguna y su entorno físico se puede encontrar en Gómez-Alday et al (2004, 2008) [1 y 2].

La laguna está sometida a varias presiones ambientales provenientes de fuentes antropogénicas: aguas residuales, actividades agrarias y ganaderas y una antigua salina (actualmente en desuso) [3 y 4]. El impacto que provoca este tipo de presiones se puede observar en la aparición de contaminantes orgánicos como pesticidas en la masa de agua subterránea, e inorgánicos como nitrato encontrados en la propia laguna.

La zona de estudio es la cuenca endorreica de la Laguna salada de Pétrola (Cuenca Hidrográfica del Segura), la laguna en sí tiene una superficie en guas altas de 2 Km^2. Esta laguna tiene oscilaciones en su nivel ocasionadas por las épocas de sequía acusada, pero sin llegar a secarse en ninguna época del año (desde 2008). Con estas condiciones, la laguna adquiere un carácter hipersalino que puede alcanzar valores de conductividad de 129 mS/cm (octubre 2010) asociados con las mayores tasas de evaporación.

Todas estas características hacen de la laguna un punto de ecología microbiana sumamente interesante debido a la necesidad de las especies presentes de adaptarse a un ambiente hostil con altas concentraciones de sales y nitratos y a la posible aparición de contaminantes orgánicos derivados de las actividades agrícolas y de la población.

El interés en la realización de este trabajo reside en poder conocer como la salinidad y las características físico-químicas de los lagos dirige las estructuras de las comunidades microbianas y su diversidad. Adquirir este tipo de conocimientos es importante ya que a) permite obtener una primera visión de cómo la estructura microbiana puede cambiar con los usos de los suelos (presiones antropogénicas) y b) el cambio climático podría acarrear una mayor tasa de evaporación y con esto un aumento en la concentración de sales de lagos y lagunas. Todo esto unido a que el microbioma en ambientes hipersalinos permanece ampliamente desconocido hace que su estudio sea de interés.

1.2 Objetivos del Trabajo

Los objetivos generales del TFM son:

1. Realizar un análisis de la composición microbiológica en muestras de sedimento y agua para poder analizar la relación que existe entre los distintos puntos geográficos de la laguna.

Objetivos específicos:

Dentro de los objetivos generales se especifican los siguientes objetivos específicos:

Objetivos específicos del objetivo general 1:

1.1. Preprocesar las secuencias, eliminando los extremos con baja calidad de secuenciación.

1.2. Obtener una tabla de frecuencias de Operational Taxonomic Unit (OTUs) y asignar el taxón correspondiente a cada OTU.

1.3. Obtener la diversidad alfa y beta de los distintos puntos de muestreo.

Objetivos específicos del objetivo general 2:

2.1 Describir y comparar las características físico-químicas de los distintos puntos de muestreo

2.2 Comparar la composición taxonómica con las características físico-químicas de los distintos puntos de muestreo mediante análisis estadístico como el análisis de correspondencia canónica.

1.3 Enfoque y método seguido

Para llevar a cabo los objetivos propuestos se utilizarán herramientas como IPython notebook [5] para la creación de un script que trabajará sobre la herramienta bioinformática QIIME2 [6]. Este script tendrá como propósito llevar a cabo las tareas necesarias para cumplir con el objetivo 1.

Para llevar a cabo el análisis estadístico del objetivo 2.2 y la creación de graficas se utilizara el entorno y lenguaje de programación R [7]. Se utilizarán paquetes como *vegan* [8] para el análisis estadístico y paquetes como *ggplot2* [9] y *plotly* [10] para la creación de graficas interactivas que faciliten el análisis de los datos.

1.4 Planificación del Trabajo

1.4.1 Planificación inicial:

En la figura 1 se ofrece la planificación temporal del trabajo inicial. En figura 2 se puede ver la planificación que se siguió finalmente.

	TFM	90 días	mié 20/02/19	mar 25/06/19
	Inicio	0 días	mié 20/02/19	mié 20/02/19
	PEC0	9 días	mié 20/02/19	lun 04/03/19
	Entrega PEC0	0 días	lun 04/03/19	lun 04/03/19
	PEC1	10 días	mar 05/03/19	lun 18/03/19
	Entrega PEC1	0 días	lun 18/03/19	lun 18/03/19
	PEC2	**27 días**	**mar 19/03/19**	**mié 24/04/19**
	Preprocesado	7 días	mar 19/03/19	mié 27/03/19
	Tabla de frecuencias	7 días	jue 28/03/19	vie 05/04/19
	Asignacion de Taxones	7 días	jue 28/03/19	vie 05/04/19
	Diversidad alfa y beta	7 días	lun 08/04/19	mar 16/04/19
	Informe PEC2	6 días	mié 17/04/19	mié 24/04/19
	Entrega PEC2	0 días	mié 24/04/19	mié 24/04/19
	PEC3	**18 días**	**jue 25/04/19**	**lun 20/05/19**
	Características fisicoquímicas	6 días	jue 25/04/19	jue 02/05/19
	Comparación canonica	6 días	vie 03/05/19	vie 10/05/19
	Informe PEC3	6 días	lun 13/05/19	lun 20/05/19
	Entrega PEC3	0 días	lun 20/05/19	lun 20/05/19
	PEC4 Redaccion de memoria	12 días	mar 21/05/19	mié 05/06/19
	Entrega PEC4	0 días	mié 05/06/19	mié 05/06/19
	PEC5a Elaboracion presentacion	6 días	jue 06/06/19	jue 13/06/19
	Entrega PEC5a	0 días	jue 13/06/19	jue 13/06/19
	PEC5b Defensa	8 días	vie 14/06/19	mar 25/06/19

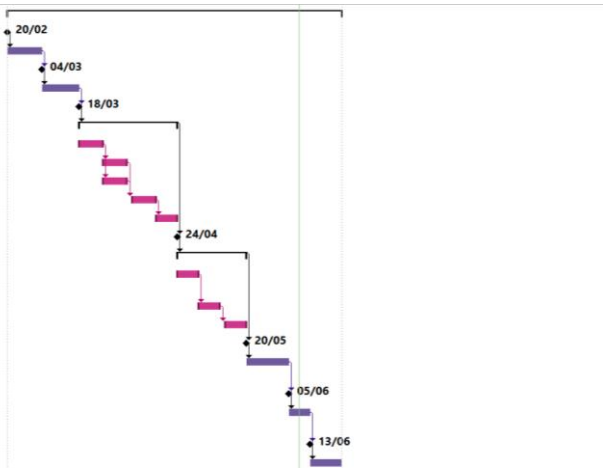

Figura 1: Diagrama de Gannt donde se muestra la planificación que se siguió inicialmente

1.4.2 Planificación final seguida:

	TFM	90 días	mié 20/02/19	mar 25/06/19
	Inicio	0 días	mié 20/02/19	mié 20/02/19
	PEC0	9 días	mié 20/02/19	lun 04/03/19
	Entrega PEC0	0 días	lun 04/03/19	lun 04/03/19
	PEC1	10 días	mar 05/03/19	lun 18/03/19
	Entrega PEC1	0 días	lun 18/03/19	lun 18/03/19
	PEC2	**27 días**	**mar 19/03/19**	**mié 24/04/19**
	Preprocesado	7 días	mar 19/03/19	mié 27/03/19
	Tabla de frecuencias	7 días	jue 28/03/19	vie 05/04/19
	Asignacion de Taxones	7 días	jue 28/03/19	vie 05/04/19
	Diversidad alfa y beta	7 días	lun 08/04/19	mar 16/04/19
	Informe PEC2	6 días	mié 17/04/19	mié 24/04/19
	Entrega PEC2	0 días	mié 24/04/19	mié 24/04/19
	PEC3	**18 días**	**jue 25/04/19**	**lun 20/05/19**
	Análisis de abundancia diferencial	6 días	jue 25/04/19	jue 02/05/19
	Comparación de resultados con otros articulos	6 días	vie 03/05/19	vie 10/05/19
	Informe PEC3	6 días	lun 13/05/19	lun 20/05/19
	Entrega PEC3	0 días	lun 20/05/19	lun 20/05/19
	PEC4 Redaccion de memoria	12 días	mar 21/05/19	mié 05/06/19
	Entrega PEC4	0 días	mié 05/06/19	mié 05/06/19
	PEC5a Elaboracion presentacion	6 días	jue 06/06/19	jue 13/06/19
	Entrega PEC5a	0 días	jue 13/06/19	jue 13/06/19
	PEC5b Defensa	8 días	vie 14/06/19	mar 25/06/19

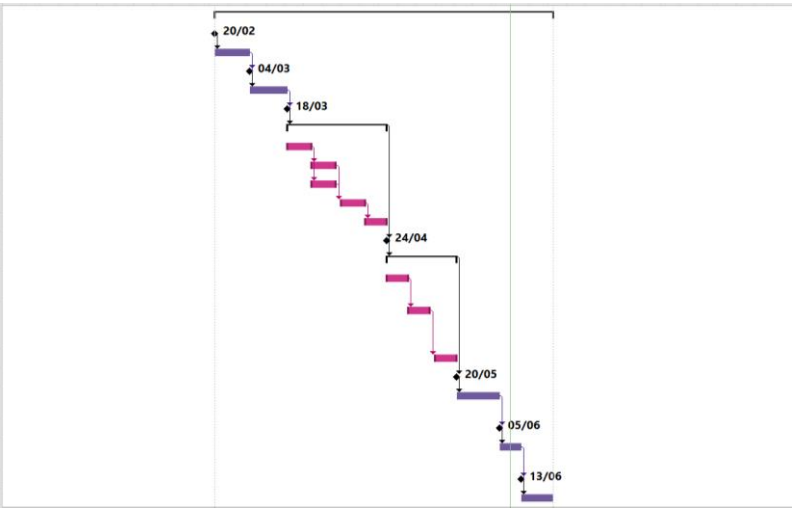

Figura 2: Diagrama de Gannt donde se muestra la planificación que se siguió finalmente

1.5 Breve sumario de productos obtenidos

Los resultados esperados durante la realización de este TFM son:
-Plan de trabajo
-Memoria
-Presentación virtual
-Autoevaluación del proyecto

2. Materiales y métodos

2.1 Obtención de las muestras

Las muestras de sedimento y agua se han obtenido de siete puntos del lago de Pétrola (Figura 3). La selección de estos puntos se ha realizado en base a los usos de los suelos próximos. El punto 2635 es una antigua explotación salina. Los puntos 2643 y 2650 se ven afectados por los vertidos de aguas residuales provenientes de la población. El punto 2648 recibe los vertidos de una granja porcina. A los puntos 2649 y 2652 llegan fertilizantes lixiviados provenientes de la agricultura. El punto 2651 es el punto más céntrico y que menos afectado puede verse por los usos de los suelos. Se han recogido 6 muestras de 1L de agua superficial de los puntos 2635, 2643, 2648, 2649, 2651 y 2652 y siete muestras de unos 5 cm de profundidad de sedimentos de los puntos 2635, 2643, 2648, 2649, 2650, 2651 y 2652. En el punto 2650 no fue posible obtener muestra de agua ya que el nivel de agua era demasiado bajo. Las muestras de agua fueron guardadas en botellas estériles y las muestras de sedimentos fueron guardadas en tubos de centrifuga Falcon™. Todas las muestras fueron almacenadas a -20ºC para su posterior análisis.

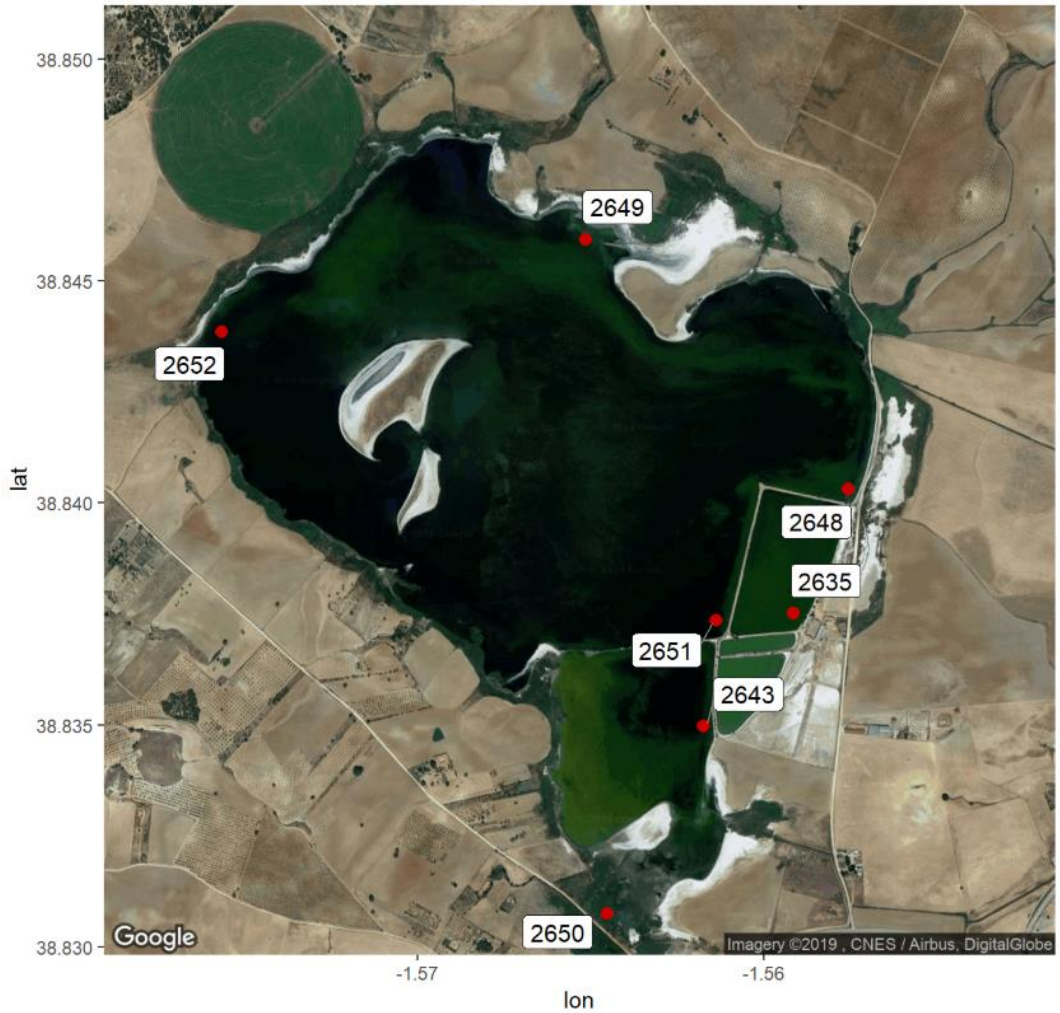

Figura 3: Mapa de la laguna de Pétrola con los puntos de muestreo. Lat.: latitud; lon: Longitud

2.2 Extracción del ADN de muestras ambientales

Las muestras de agua fueron filtradas con filtros con un tamaño de poro de 0.45µm. El ADN total de los filtros y de 0.5g de las muestras de sedimento fue extraído siguiendo las instrucciones del kit de NucleoSpin® Soil DNA, RNA, and protein purification kit (Macherey-Nagel, Düren, Germany). Posteriormente se revisó la integridad del ADN mediante electroforesis en gel de agarosa al 1.5% y la cantidad de ADN mediante espectrofotometría con el sistema de BioTek® Cytation 5. Una vez obtenidas muestras con integridad y cantidad suficiente de ADN (figura 4) se enviaron para su secuenciación a la empresa StabVida®.

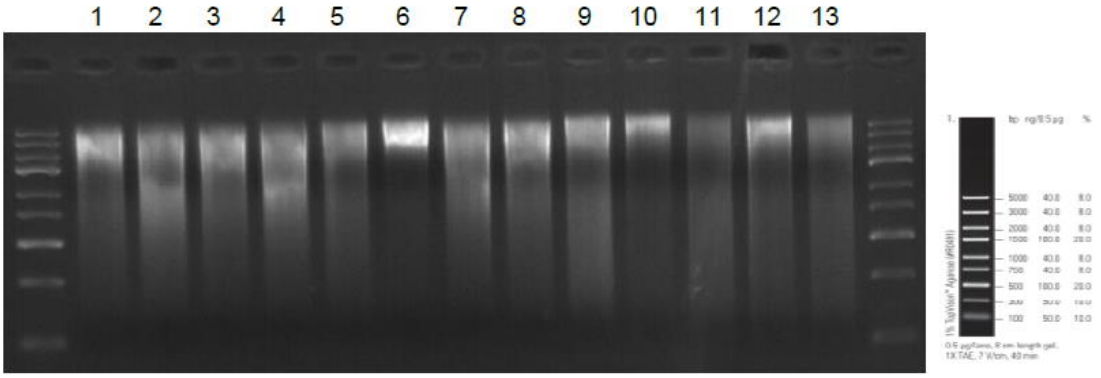

Figura 4: Electroforesis en gel de agarosa al 1.5%.
Las muestras del 1 al 13 son respectivamente 2635-S, 2643-S, 2648-S, 2649-S, 2650-S, 2651-S, 2652-S, 2635-W, 2643-W, 2648-W, 2649-W, 2651-W, 2652-W.

2.3 Secuenciación

La secuenciación se realizó sobre la región hipervariable V3-V4 del gen de RNAr 16S presente en bacterias y archeas. Para ello primero se realizó una amplificación mediante PCR utilizando los primers 341F y 805R junto con el adaptador para la tecnología Ilumina. La secuencia final de los primers fue:

Forward:
5'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG3'
Reverse:
5'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC3'

En rojo la secuencia del adaptador para la tecnología Ilumina y en azul la secuencia especifica de la región V3-V4 del gen RNAr 16S.

2.4 Análisis bioinformático

El análisis bioinformático de las secuencias de las muestras se llevó a cabo usando el software QIIME2 v2019. 1. Para ello se utilizó una máquina virtual utilizando el software Oracle VM VirtualBox donde se instaló QIIME2 mediante una imagen OVF. Para facilitar la escritura de los scripts se ha utilizado el notebook Jupyter. En primer lugar, se revisaron los archivos suministrados por la empresa. Se tratan de archivos fastq con la calidad codificada en la escala

Phred+33, y por cada muestra hay dos archivos correspondientes a las lecturas forward y reverse.

Antes de seguir con el análisis se ha creado un archivo de metadatos (Tabla 1) con la información referente a las muestras de sedimento y agua. Este archivo contendrá una columna (BarcodeSecuence) para su análisis con QIIME2 aunque no aporte ninguna información adicional. Esta columna es necesaria ya que, durante la secuenciación, y para reducir costes, se suelen juntar varias muestras amplificadas con unos primers específicos de cada muestra. A esa secuencia específica de la muestra se denomina Barcode y nos serviría para realizar el procedimiento que se denomina demultiplexado: Este procedimiento consiste en separar cada conjunto de secuencias de cada muestra según la secuencia Barcode.

Tabla 1: Metadatos de las muestras secuenciadas

#SampleID	BarcodeSecuence	Description	Sedimento/Agua
2635-S-N0500	None	2635-S	Sedimento
2635-W-N0501	None	2635-W	Agua
2643-S-N0502	None	2643-S	Sedimento
2643-W-N0503	None	2643-W	Agua
2648-S-N0504	None	2648-S	Sedimento
2648-W-N0505	None	2648-W	Agua
2649-S-N0506	None	2649-S	Sedimento
2649-W-N0507	None	2649-W	Agua
2650-S-N0508	None	2650-S	Sedimento
2651-S-N0509	None	2651-S	Sedimento
2651-W-N05010	None	2651-W	Agua
2652-S-N05011	None	2652-S	Sedimento
2652-W-N05012	None	2652-W	Agua

A continuación, se ha usado el plugin DADA2 [11] con el método denoise-paired. Este método consiste en los siguientes pasos:

-Filtered: se filtraran todas las secuencias que no tuvieran la longitud mínima después del recortado de las partes con baja calidad.

-Denoised: combinación de todas las lecturas idénticas en secuencias únicas asignando la abundancia correspondiente a cada secuencia. En este paso DADA2 mantiene un resumen de la calidad de cada secuencia. Esto será crucial para el siguiente paso.

-Merged: Mediante el uso de un modelo de error paramétrico basado en las tasas de error para cada posible transición (por ejemplo, A-> C, A-> G,...), se fusionan las lecturas forward y reverse. Las lecturas pareadas que no coincidan serán eliminadas.

-Non_chimeric: Eliminación de las posibles quimeras.

Para realizar este procedimiento se han usado los siguientes parámetros:

- trunc_len_f = 240
- trunc_len_r = 240
- trim_left_f = 24
- trim_left_r = 24
- chimera_method = 'consensus'
- n_threads = 0
- n_reads_learn = 1000000

Con los primeros cuatro parámetros se especifica que corte los primeros 24 pares de bases de las lecturas “forward” y “reverse” y no lea más allá de las 240 pares de bases de ambas. Esto se ha hecho para eliminar las partes con peor calidad de secuenciación.

El método que se ha utilizado para eliminar las secuencias quimeras ha sido por consenso. Este método busca quimeras en cada muestra por separado y las secuencias quiméricas encontradas en una fracción suficientemente grande de muestras son eliminadas. El parámetro n_threads = 0 implica que el número de cores utilizado para el proceso será el máximo. El parámetro n_reads_learn = 1000000 son el número de lecturas a utilizar cuando se entrena el modelo de error. Se ha utilizado el designado por defecto. Este parámetro es el cuello de botella por el que se ha tenido que utilizar una supercomputadora para realizar el análisis.

La supercomputadora utilizada ha sido gracias al centro de supercomputación I^3A del instituto de investigación en informática de Albacete de la universidad de Castilla La Mancha. Se concedió el acceso a un nodo de computación de 64 Gigas de RAM con 2 procesadores Intel Xeon E5-2650 2.00GHz 8 cores, con un total de 16 cores por nodo.

El siguiente paso es la asignación taxonómica de la tabla de ASV (*amplicon sequence variant*) para este proceso se ha utilizado la base de datos SILVA [12] y tres métodos distintos, BLAST [13], VSEARCH [14] y el clasificador sklearn [15].

- El método BLAST se basa en el algoritmo del NCBI BLASTn donde se realizan alineamientos de cada secuencia contra la base de datos. Utiliza el método heurístico.

- VSEARCH es una herramienta de código abierto y gratuito para procesar y preparar datos de secuenciación. En su aplicación para la asignación taxonómica VSEARCH utiliza, como BLAST, un método heurístico para la asignación de taxones.
- El método sklearn se ha realizado entrenando un clasificador Naive Bayes con la base de datos. Para ello se ha empleado el plugin feature-classifier con el método extract-reads para obtener unas secuencias de referencia de la base de datos en función de los primers que han sido utilizados en la secuenciación sin el adaptador para Ilumina:

  Forward:
  5’CCTACGGGNGGCWGCAG3’
  Reverse:

5’GACTACHVGGGTATCTAATCC3’

A continuación, se ha empleado el método “fit-classifier-naive-bayes” para entrenar el clasificador con esas secuencias y la taxonomía de referencia. Este paso también ha sido realizado en la supercomputadora. Posteriormente se ha clasificado la tabla de secuencias representativas con el método classify-sklearn del plugin feature-classifier.

Los tres métodos anteriormente descritos han dado como resultado tres archivos que permiten la creación de gráficos de barras interactivos empleando el método de visualización barplot del plugin taxa. Para cada uno de los tres métodos se puede observar las gráficas barplot correspondientes a los taxones filo, clase, orden, familia y género. Para comparar los tres métodos se ha realizado la media de la correlación de Pearson entre los distintos puntos de muestreo a nivel de género.

2.5 Análisis estadístico con el software R

2.5.1 Diversidad alfa

El cálculo de la diversidad alfa se ha realizado respecto de los resultados obtenidos en la clasificación taxonómica a nivel de género de los tres métodos. Se ha optado por este método por tres razones. La primera razón es que existe la posibilidad de que para un mismo microorganismo haya más de una ASV debido a errores de secuenciación o a la convivencia de más de una especie bacteriana por especie y esto hinche la diversidad alfa. La segunda razón es que bajo la categoría taxonómica “unassigned”, en los tres métodos, hay un número muy bajo de lecturas, por lo que el efecto contrario de perdida de diversidad también es muy bajo. La tercera razón es que la diferencia entre las distintas profundidades de lectura entre las muestras no influye en el número de ASVs. Esto lo podemos observar en la gráfica de rarefacción, donde todas las muestras llegan al máximo de ASVs.

Los índices de diversidad alfa calculados para cada uno de los métodos de asignación taxonómica han sido:

- OTUs observadas
- Índice Simpson [16]
- Índice Shannon [17]

2.5.2 Diversidad Beta

Al igual que en la diversidad alfa, la diversidad beta se ha calculado a partir de los resultados de la asignación taxonómica a nivel de género. Para ello se ha realizado una doble estandarización de Wisconsin [18] de la raíz cuadrada de los datos debido a los valores elevados de algunos datos. Se ha utilizado la distancia Bray-Curtis [19] para medir la disimilitud entre las muestras y el método ha sido el análisis multivariante de escalamiento multidimensional no métrico, los

resultados se han obtenido en dos dimensiones para su mejor entendimiento y visualización. Para realizar este procedimiento se ha utilizado la función metaMDS() del paquete vegan.

2.5.3 Abundancia diferencial con DESeq2:

El paquete DESeq2 [20] proporciona métodos para estudiar la abundancia diferencial con el uso de modelos lineales generalizados binomiales negativos. Esto proporciona flexibilidad para analizar diseños complejos. Para analizar la abundancia diferencial tenemos que partir de una matriz donde las columnas sean las muestras y cada fila representa un taxón por ello se ha utilizado los resultados de la asignación taxonómica a nivel de género de cada método.

La visualización de los resultados del estudio de la abundancia diferencial
Se ha realizado mediante una gráfica volcano donde en el eje x tenemos el $\log_2$ del fold change y en el eje y el $-\log_{10}$ del p-value. Al haber mostrado el test tantos resultados diferencialmente abundantes se han marcado con el nombre los 10 organismos más diferencialmente expresados en favor de las muestras de sedimentos que supiéramos su nombre a nivel de género y todos los organismos que conociéramos su nombre a nivel de genero más diferencialmente expresados en favor de las muestras de aguas.

3. Resultados y discusión

3.1 Calidad de los resultados de secuenciación:

La secuenciación de las muestras dio como resultado un total de 26 archivos comprimidos en formato fastq [21]. A cada muestra le corresponden dos archivos, uno contiene las secuencias obtenidas con el primer reverse y el otro las secuencias obtenidas con el primer forward. El número total de secuencias en todas las muestras fue de 3.674.234.

Para ver la calidad de la secuenciación (forward y reverse), se tomó la calidad media de 1.000 secuencias aleatorias, no repetidas, podemos ver el resultado en las figuras 5 y 6:

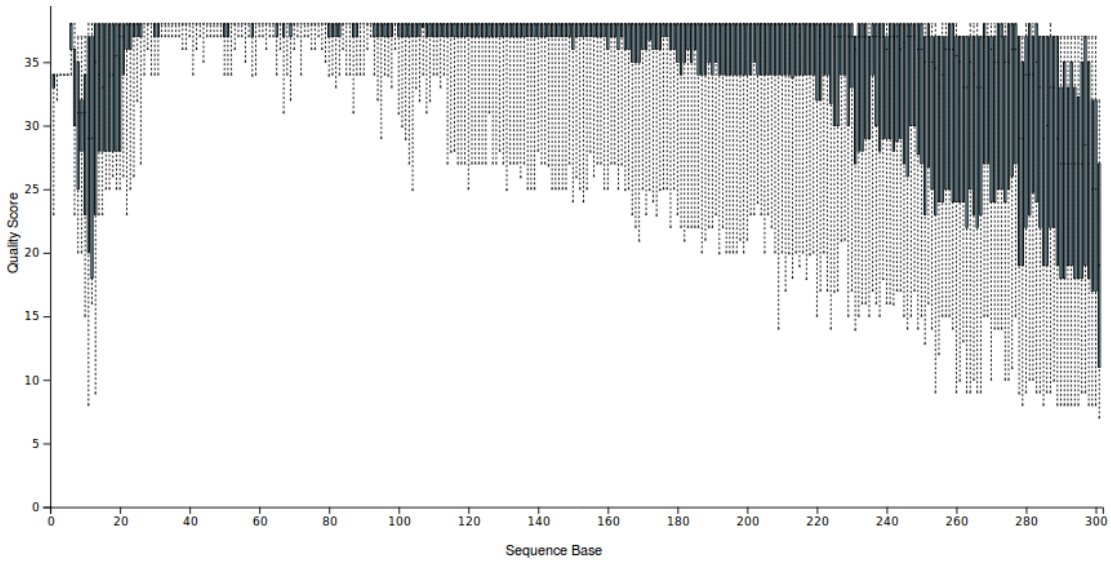

Figura 5: Calidad de secuenciación de las secuencias forward

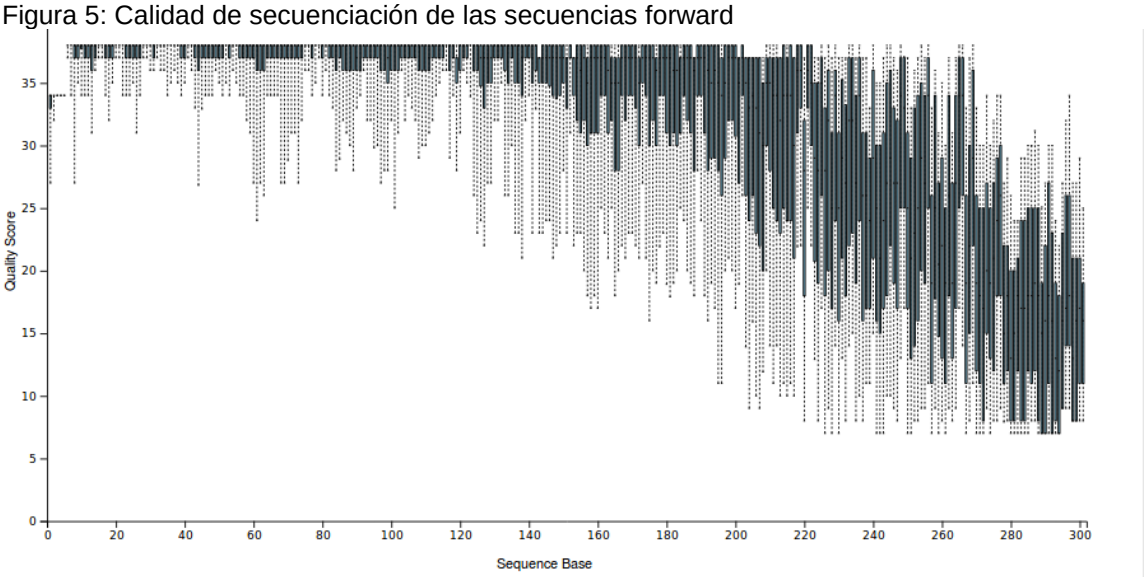

Figura 6: Calidad de secuenciación de las secuencias reverse

Se puede observar que la calidad en las secuencias forward desciende en los primeros 24 pares de bases y que después de las 240 más o menos la calidad desciende en las secuencias forward y reverse. Esto será importante para los

análisis posteriores, ya que se emplearán en el Análisis con el plugin DADA2. La longitud de las secuencias que solapara

3.2 Reducción de secuencias mediante DADA2:

A través del análisis con el plugin DADA2 se puede observar cómo se van reduciendo el número secuencias obtenidas en cada paso (Figura 7). La columna input representa los 3.674.234 de secuencias iniciales obtenidas durante la secuenciación. El resto de columnas representan el número de “reads” aceptadas al final de cada paso.

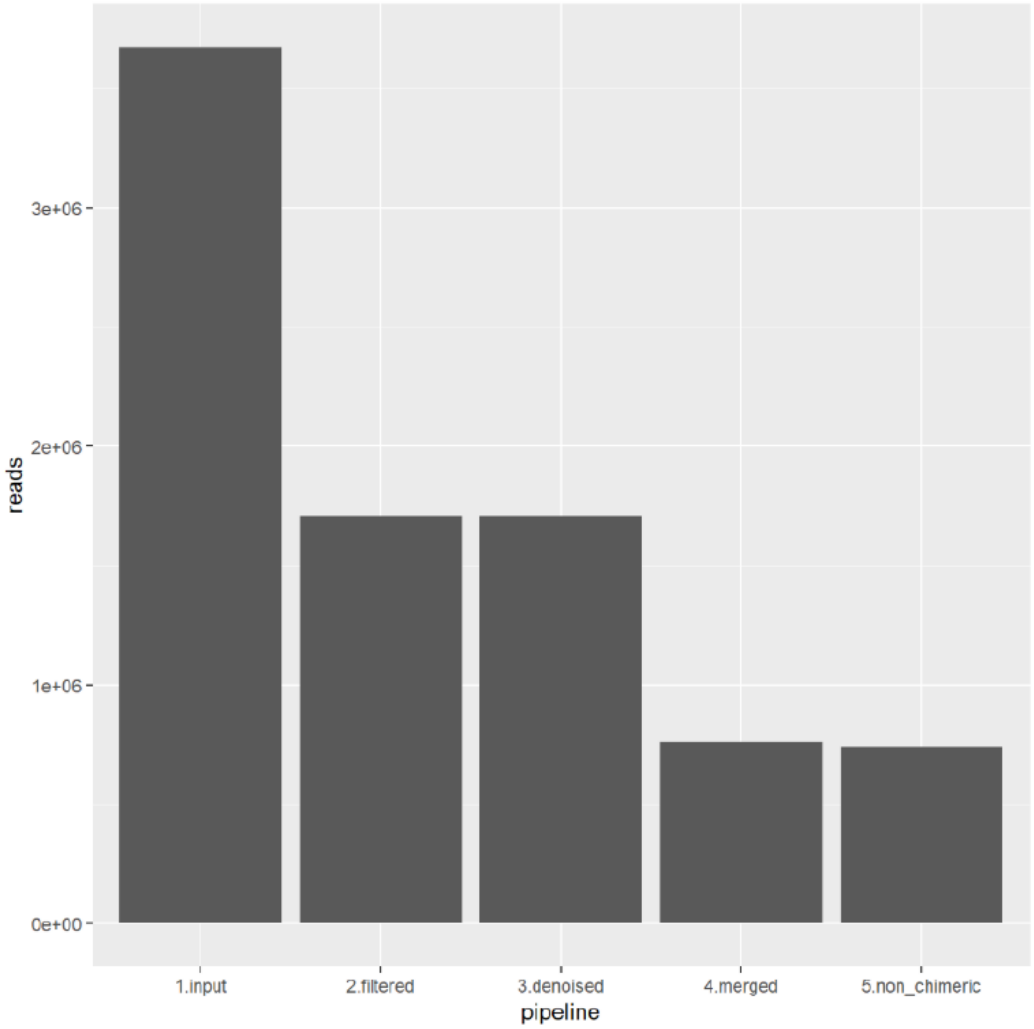

Figura 7: Número de secuencias resultante de cada paso del pipeline del plugin DADA2.

Las curvas de rarefacción que se ofrecen en la Figura 5, obtenidas a través del análisis con el plugin DADA2 nos permiten comprobar si la secuenciación se ha realizado a una profundidad de lectura adecuada. Podemos observar que, aunque no todas las muestras tengan el mismo número de lecturas, todas llegan al máximo número de OTUs, por lo que la profundidad de lectura ha sido adecuada y un aumento en la profundidad de lectura no supondrá un aumento en el número de OTUs observadas.

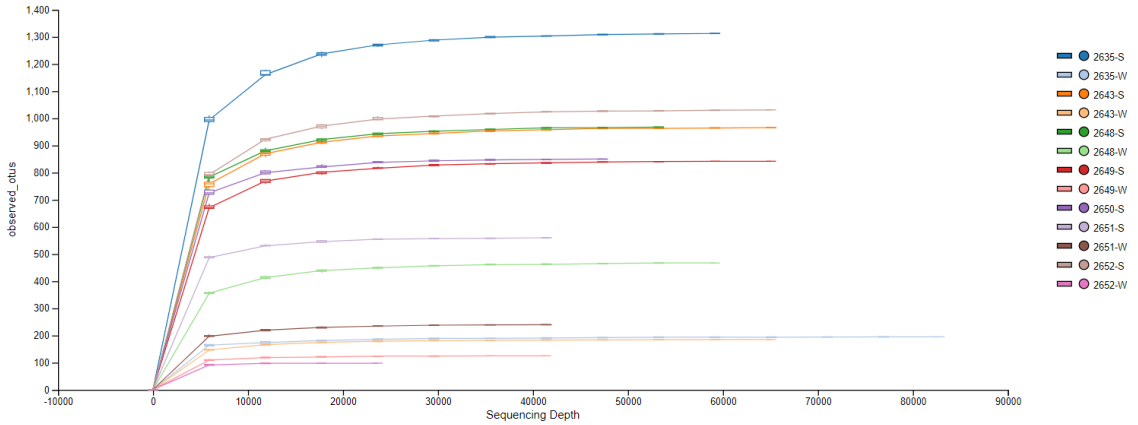

Figura 8: Curvas de rarefacción.

3.3 Diversidad taxonómica en las muestras de agua y sedimento

A continuación, se pueden ver los resultados obtenidos de la asignación taxonómica mediante los tres métodos descritos: BLAST, VSEARCH, y Sklearn.

3.3.1 BLAST

La asignación taxonómica, a nivel de Filo, permite ver que los filos Proteobacteria y Bacteroidetes ocupan el 50% de la abundancia relativa en todas las muestras y que en la muestra de agua del punto 2635 hay una gran abundancia de cianobacterias (Figura 9). Esto puede ser debido a que esta sección de la laguna fue aislada para su explotación como salina. Este hecho, no permite el intercambio de agua con el resto de la laguna (solo en periodos cuando el nivel del agua es muy alto), lo que provoca que los nutrientes que llegan a las aguas de este punto no se mezclen con los del resto de la laguna ni al contrario. Se puede observar que a nivel de género existe una clara diferencia entre las muestras obtenidas de sedimento y de agua que no se observaba a nivel de filo (Figura 10).

En los metodos BLAST y VSEARCH se puede observar que las secuencias que no han sido asignadas a ningún taxon han sido clasificadas dentro de una sección denominada “Unassigned”.

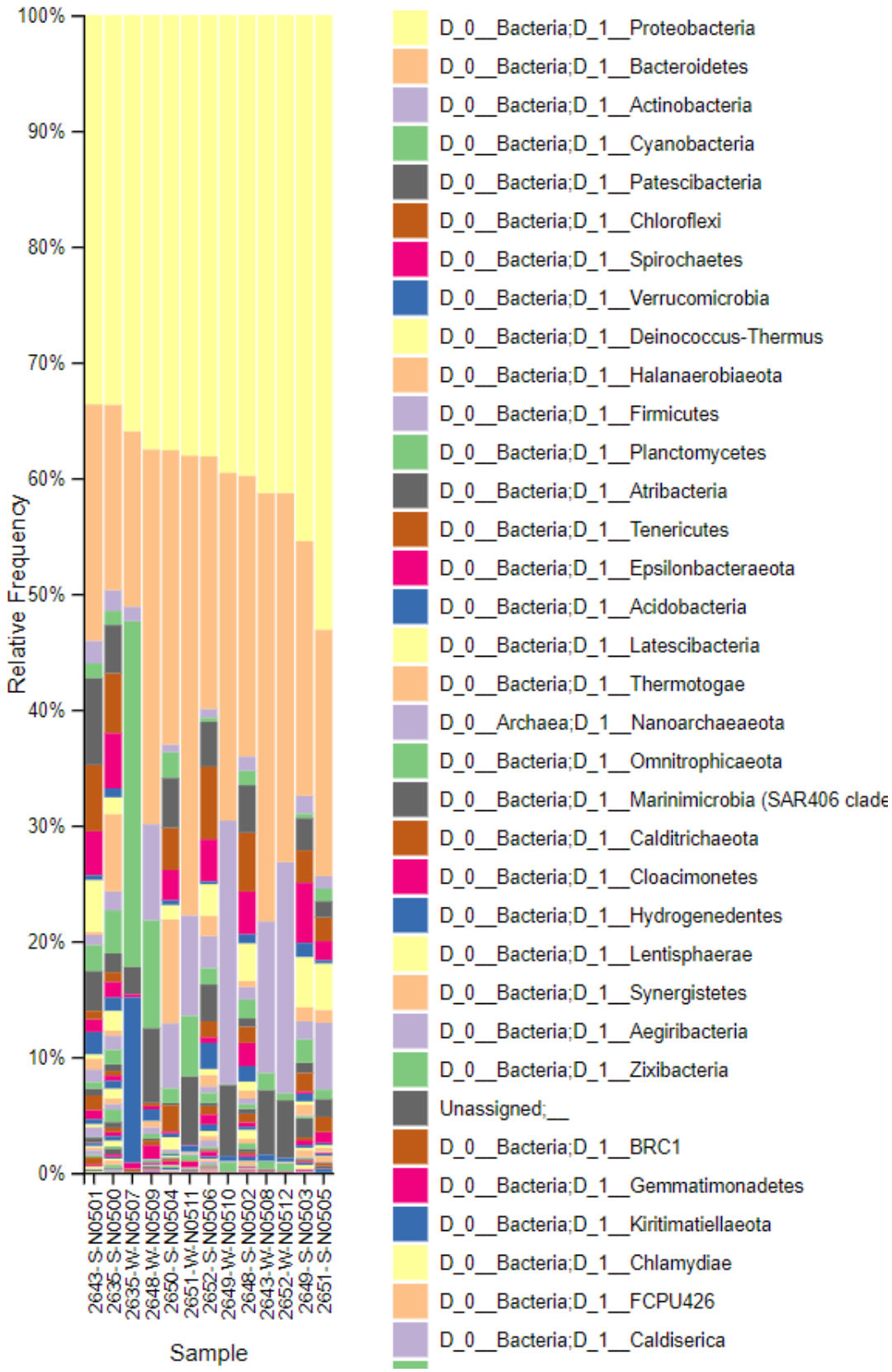

Figura 9: Asignación taxonómica a nivel de filo según el método BLAST

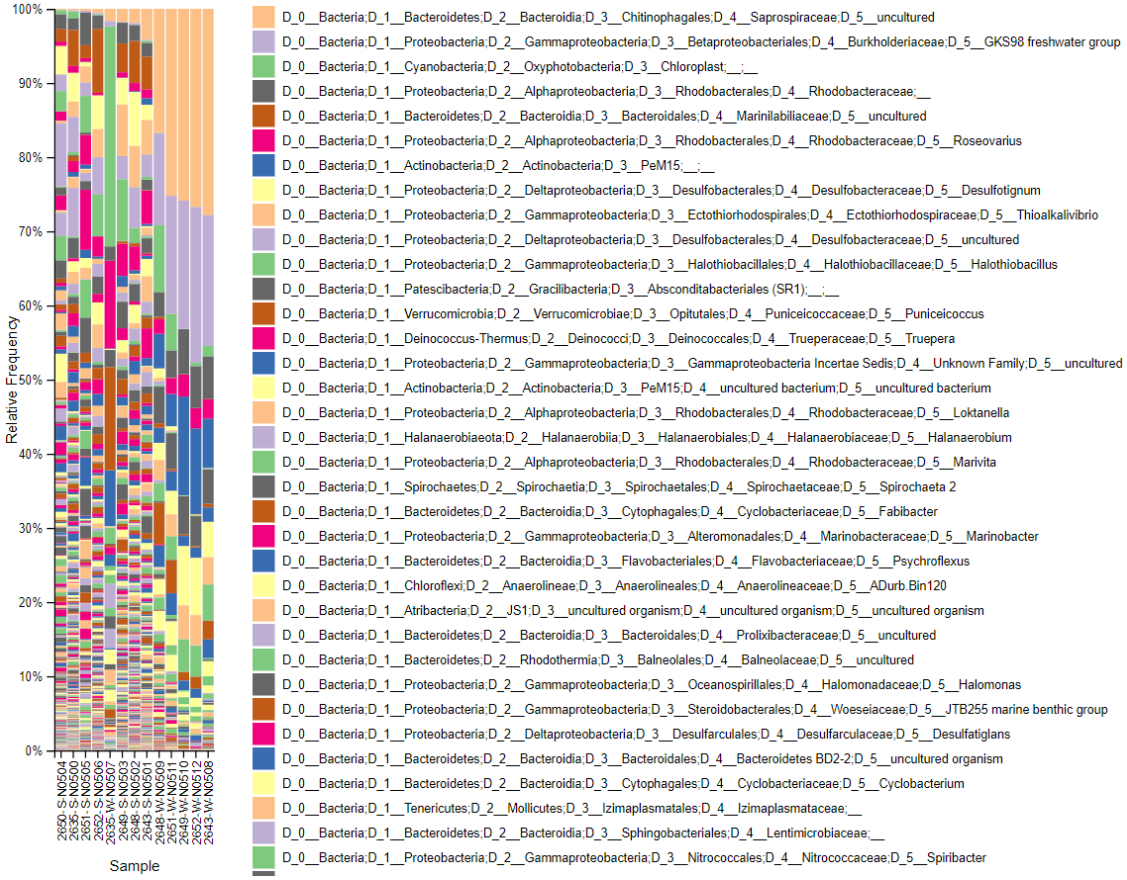

Figura 10: Asignación taxonómica a nivel de género según el método BLAST.

3.3.2 VSEARCH

Los resultados obtenidos, tanto a nivel de Filo como de Género, son similares a los del método BLAST (Figuras 11 y 12).

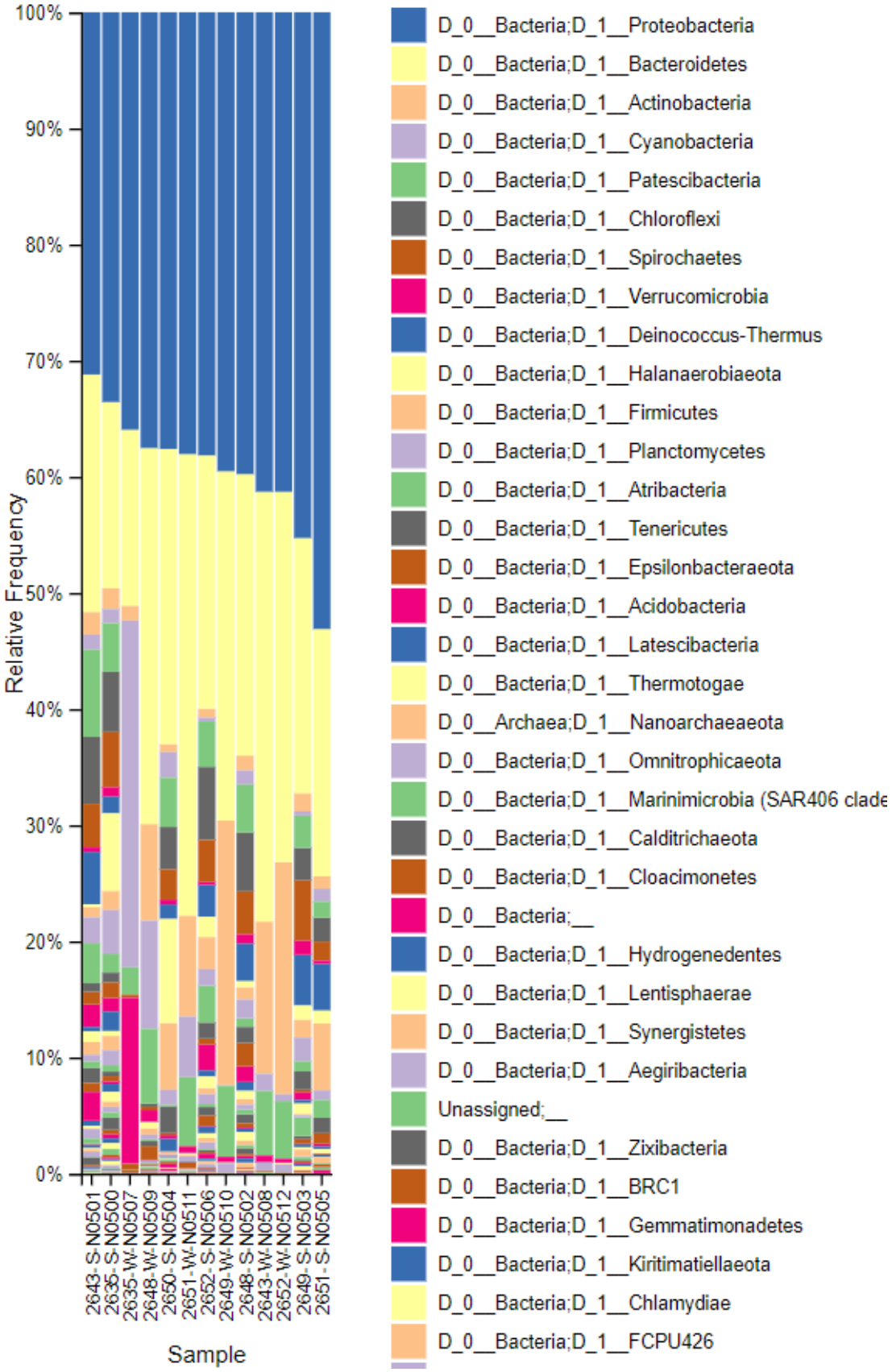

Figura 11: Asignación taxonómica a nivel de filo según el método VSEARCH

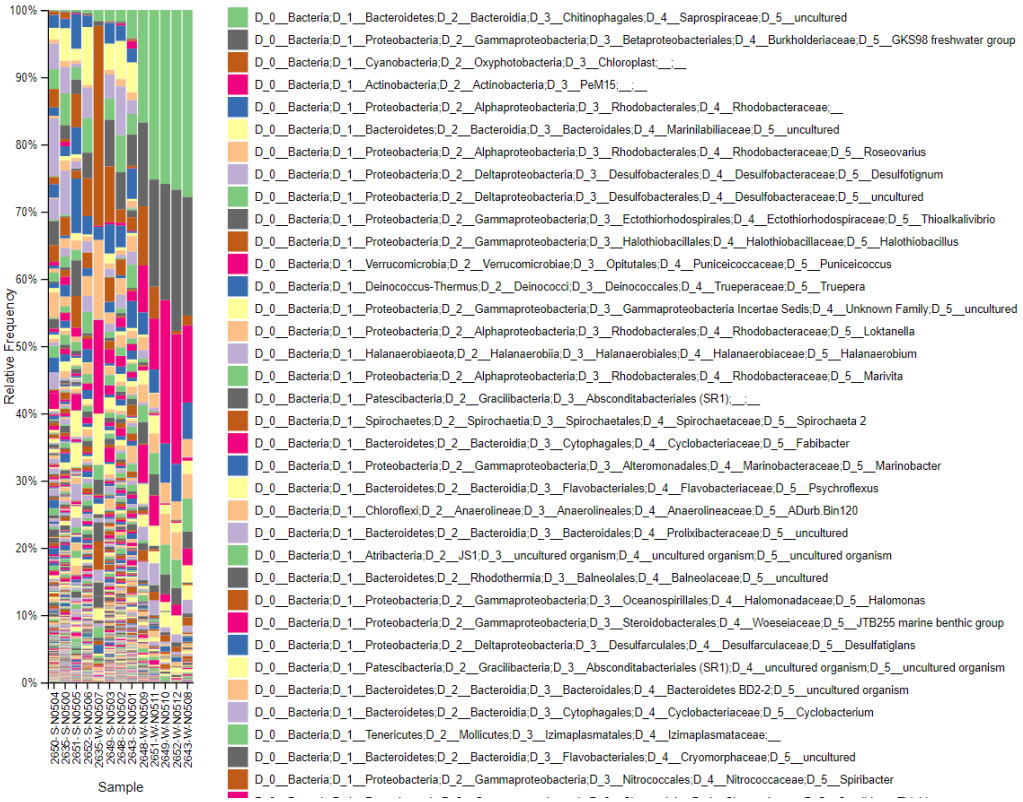

Figura 12: Asignación taxonómica a nivel genero según el método VSEARCH.

3.3.3 Sklearn:

A diferencia con los anteriores métodos no hay ningun apartado denominado “Unassigned”. Es decir, este metodo es capaz de clasificar todas las secuencias en taxones.

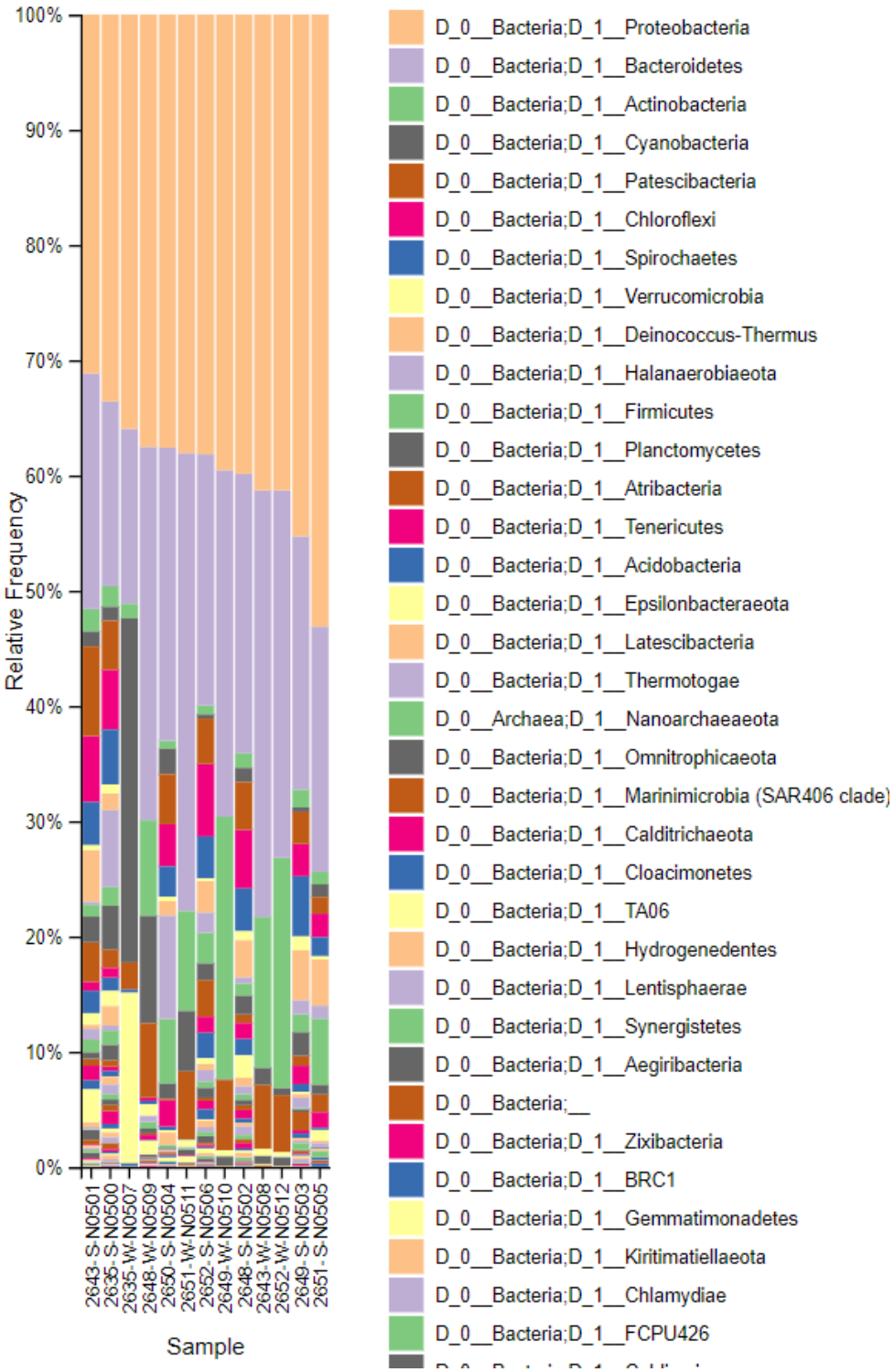

Figura 13: Asignación taxonómica a nivel genero según el método sklearn.

Figura 14: Asignación taxonómica a nivel genero según el método sklearn.

3.3.4 Comparación de métodos

Para comparar los tres métodos se ha realizado la media de la correlación de Pearson como se ha explicado anteriormente. Los resultados obtenidos se dan en la tabla 2:

Tabla 2: Tabla de correlaciones entre los distintos modelos.

	BLAST	VSEARCH	Sklearn
BLAST	1.000	0.981	0.917
VSEARCH	0.981	1.000	0.929
sklearn	0.917	0.929	1.000

Como podemos ver existe una mayor correlación entre los métodos que usan el método heurístico para la clasificación, recordamos que estos son BLAST y VSEARCH que han obtenido una correlación de 0.981, mientras que el método sklearn obtiene un menor grado de correlación con BLAST y VSEARCH. Son respectivamente 0.917 y 0.929.

3.4 Diversidad Alfa: Comparación de la diversidad a nivel de género de entre los puntos de muestro

La diversidad alfa se ha calculado a partir de los tres métodos.

3.4.1 BLAST

Para conocer la diversidad se han calculado los índices de Shannon y Simpson. Se puede observar que hay una clara diferencia entre las muestras de sedimento y las muestras de agua; la diversidad microbiana en las muestras de sedimento es mayor. Podemos recalcar la gran diversidad que tiene las muestra de sedimentos en el punto 2635 y punto 2648, respecto a las otras muestras de agua. Las que presentan una menor diversidad.

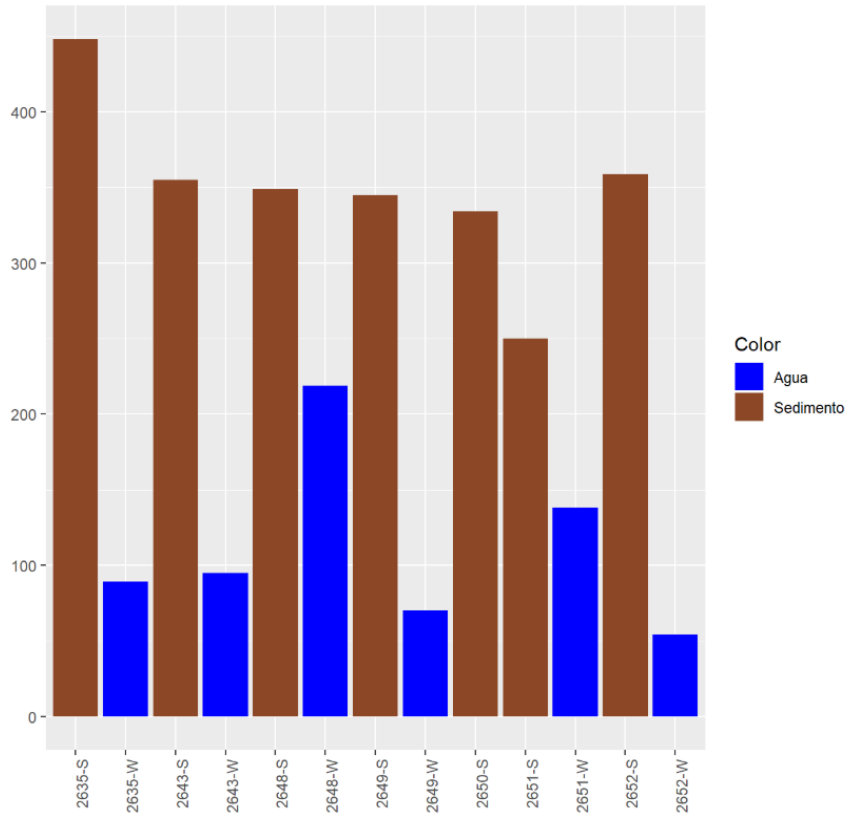

Figura 15: OTUs observadas a nivel de género para el método BLAST.

Tabla 3: Tabla resumen de índices de diversidad alfa con el metodo BLAST

Puntos	OTUs	Índice Shannon	Índice Simpson
2635-S	448	4.902	0.983
2635-W	89	2.741	0.867
2643-S	355	4.651	0.982
2643-W	95	2.692	0.872
2648-S	394	4.642	0.978
2648-W	219	3.465	0.934
2649-S	345	4.481	0.975
2649-W	70	2.486	0.867
2650-S	334	4.668	0.980
2651-S	250	4.264	0.974
2651-W	138	2.963	0.894
2652-S	359	4.517	0.976
2652-W	54	2.404	0.855

3.4.2 VSEARCH

Respecto al indice Simpson podemos ver que es muy alto en todas las muestras (el valor del indice Simpson puede variar entre 1 y 0), esto indica que unos pocos taxones tienen la mayor parte de la abundancia diferencial. Comparando las muestras de agua y las de sedimento podemos ver que las de agua tienen valores ligeramente mas bajos por lo que aunque las muestras de sedimento tengan una mayor diversidad respecto al indice shannon, las muestras de agua tienen las abundancias estan repartidas de forma mas uniforme entre los taxones.

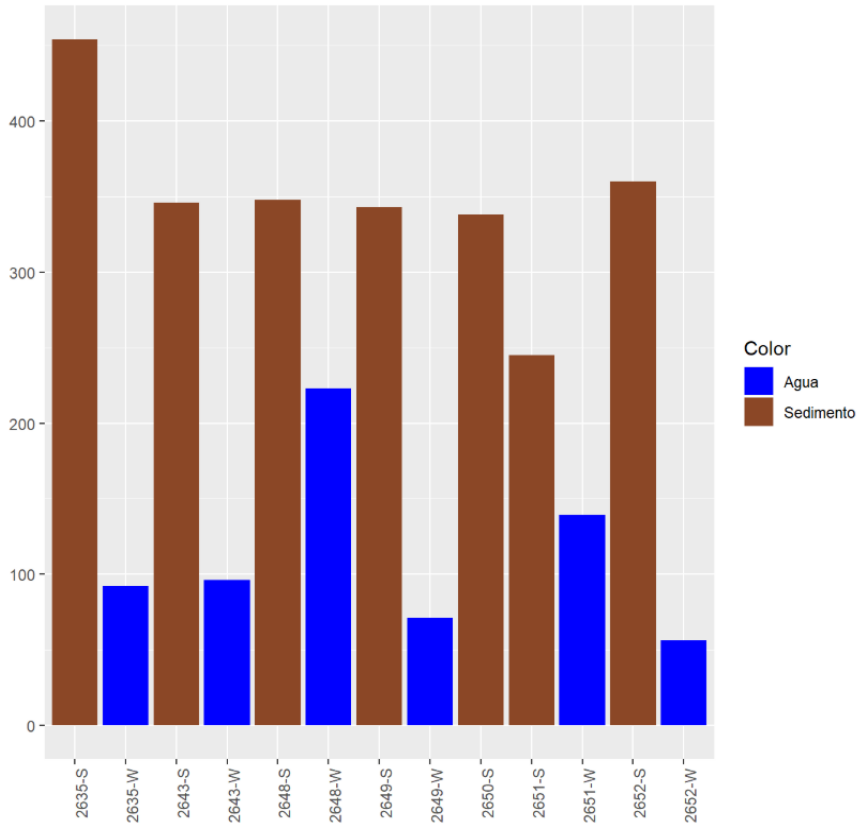

Figura 16: OTUs observadas a nivel de género para el método VSEARCH.

Tabla 4: Tabla resumen de índices de diversidad alfa con el método VSEARCH

Puntos	OTUs	Índice Shannon	Índice Simpson
2635-S	454	4.892	0.983
2635-W	92	2.745	0.867
2643-S	346	4.620	0.982
2643-W	96	2.660	0.867
2648-S	348	4.630	0.978
2648-W	223	3.461	0.933
2649-S	343	4.466	0.975
2649-W	71	2.392	0.847
2650-S	338	4.655	0.980
2651-S	245	4.244	0.973
2651-W	139	2.960	0.892
2652-S	360	4.511	0.975
2652-W	56	2.305	0.838

3.4.3 Sklearn

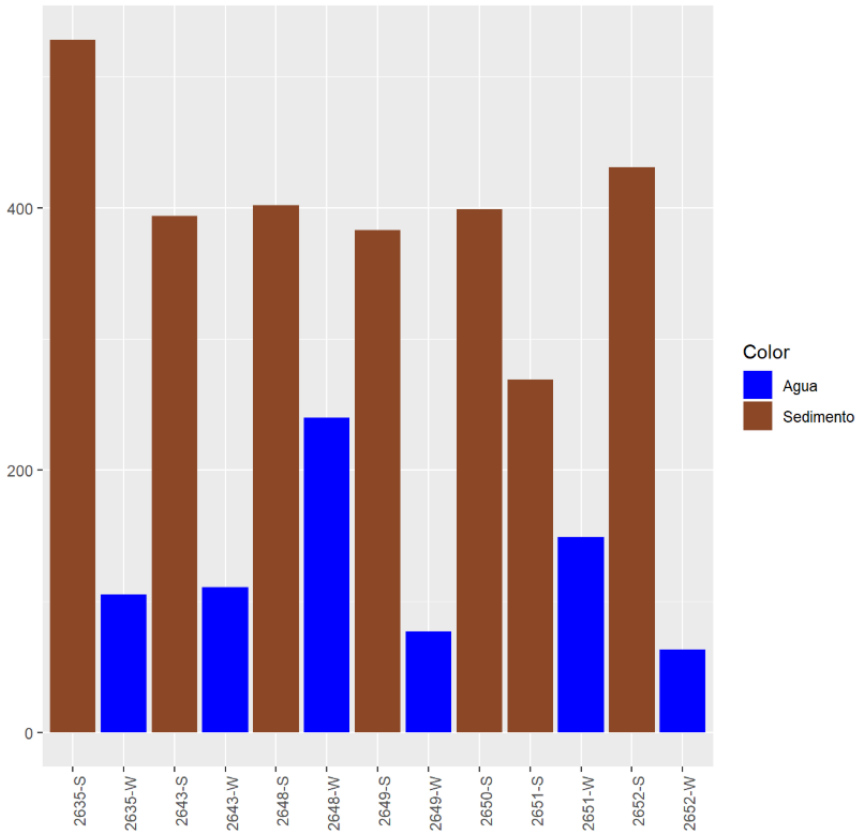

Figura 17: OTUs observadas a nivel de genero para el método sklearn.

Tabla 5: Tabla resumen de índices de diversidad alfa con el método sklearn

Puntos	OTUs observadas	Indice Shannon	Indice Simpson
2635-S	528	5,034	0,985
2635-W	105	2,793	0,870
2643-S	394	4,718	0,982
2643-W	111	2,748	0,870
2648-S	402	4,717	0,978
2648-W	240	3,271	0,937
2649-S	383	4,553	0,976
2649-W	77	2,468	0,854
2650-S	399	4,781	0,981
2651-S	369	4,362	0,975
2651-W	149	3,032	0,895
2652-S	431	4,623	0,976
2652-W	63	2,405	0,845

3.3.4 Comparación de métodos

Se puede observar que la diversidad entre las muestras es parecida entre los distintos metodos; sin embargo es interesante recalcar que el número de OTUs observadas por el metodo sklearn es ligeramente mayor que los otros dos métodos.

3.6 Diversidad Beta: Comparación a nivel de género de entre los puntos de muestro

A continuación, se pueden ver los resultados obtenidos del análisis de diversidad beta calculado mediante escalamiento multidimensional no métrico a partir de las distancias de Bray-Curtis.

3.6.1 BLAST

Los grupos de agua y sedimento se separan muy bien en el primer eje mostrando las diferencias que hay entre estos dos tipos de muestras. Tambien se puede observar que la muestra 2635-W se separa del resto de muestras de agua. Esto puede ser debido a la gran cantidad de cianobacterias presentes en la muestra.

Por otro lado las muestras 2650-S y 2651-S tambien se separan del resto de muestras de sedimento. Mirando las tablas de taxonomia no podemos decir que sea debido a un solo organismo por lo que su diferenciación puede ser debida a la influencia antropogenica que tienen estos puntos ya que estan muy cerca de los vertidos de aguas residuales. Lo que indica que esta influencia tiene consecuencias en la ecologia microbiana de los sedimentos.

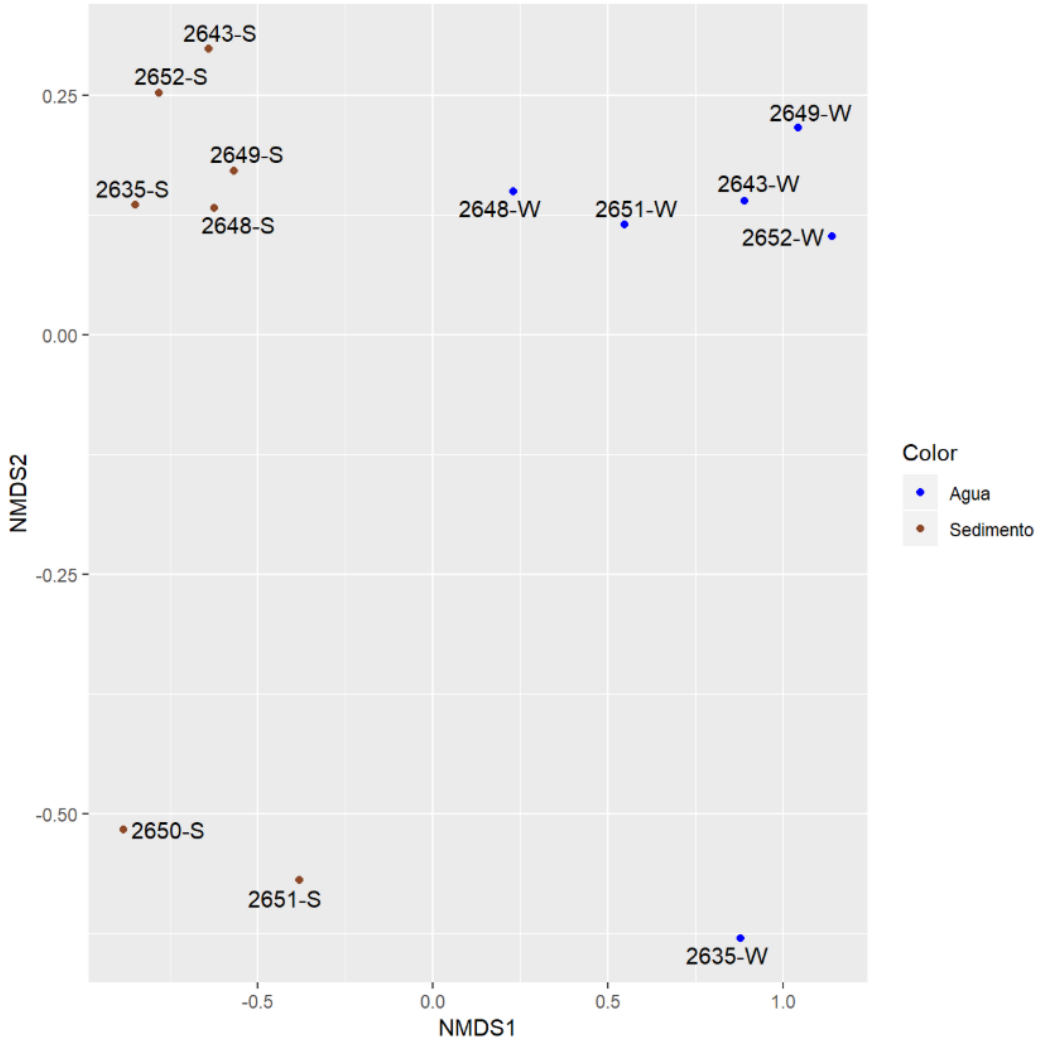

Figura 18: Diversidad beta con distancias de Bray-Curtis y escalamiento multidimensional no métrico a nivel de genero del método BLAST

3.6.2 VSEARCH:

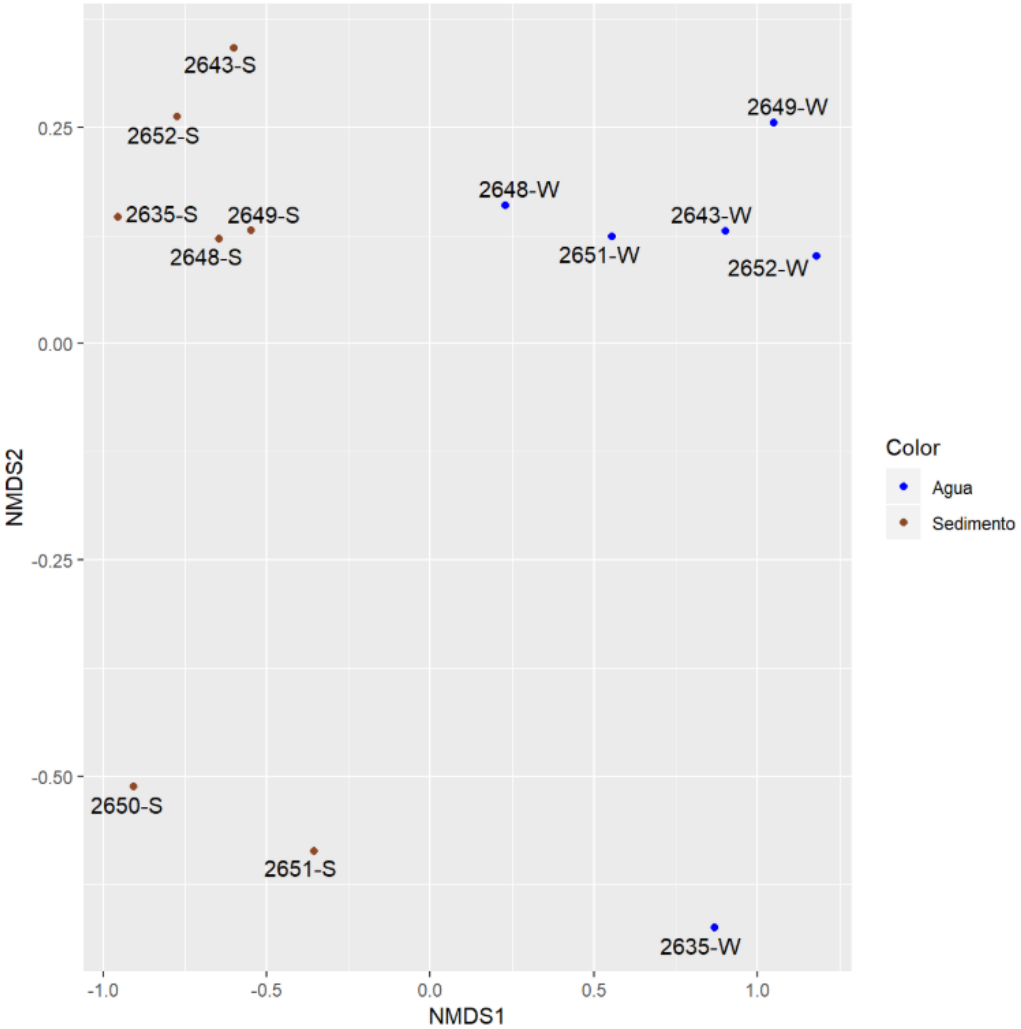

Figura 19: Diversidad beta con distancias de Bray-Curtis y escalamiento multidimensional no métrico a nivel de genero del método VSEARCH

No se pueden observar cambios significativos entre los metodos de BLAST y VSEARCH.

3.6.3 Sklearn:

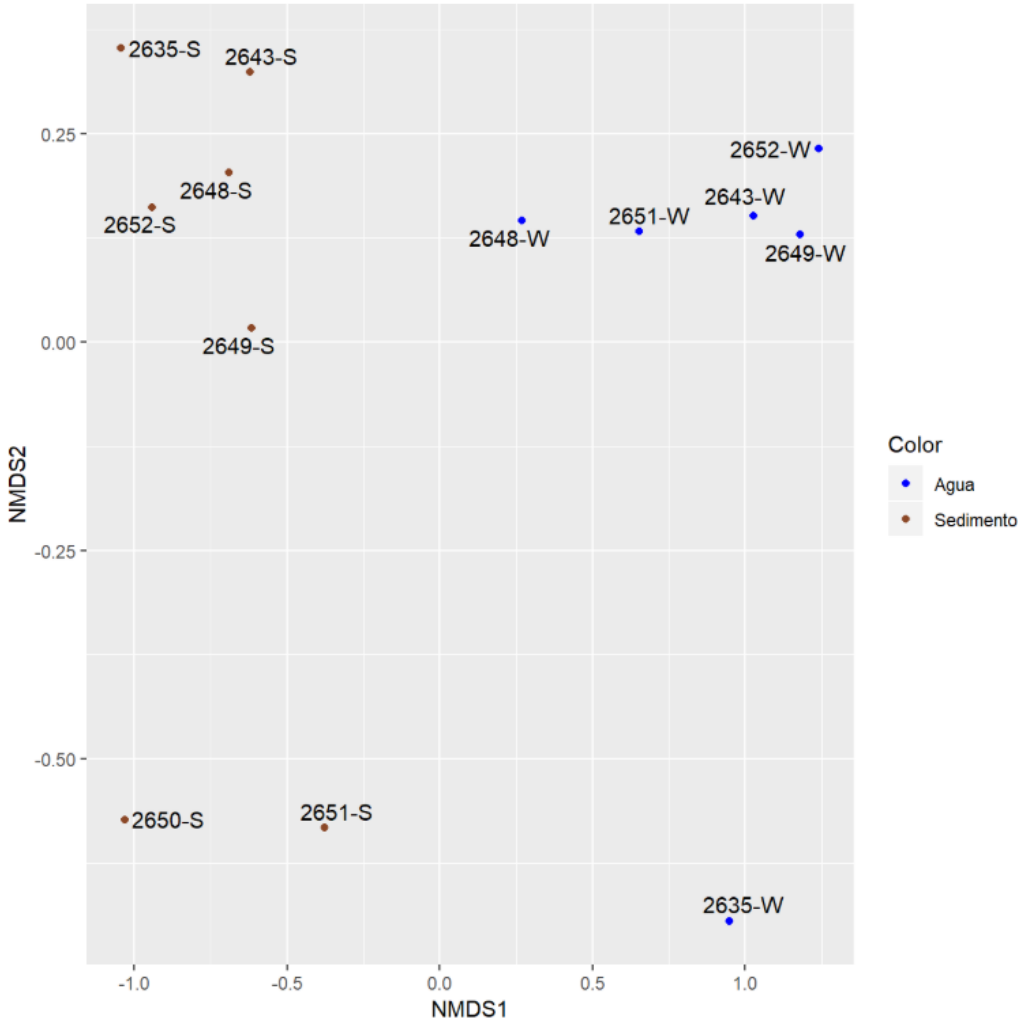

Figura 20: Diversidad beta con distancias de Bray-Curtis y escalamiento multidimensional no métrico a nivel de genero del método sklearn

Se puede observar que los tres metodos obtienen resultados muy parecidos.

3.7 Abundancia diferencial entre las muestras de agua y sedimento con DESeq2:

3.7.1 BLAST

Se ha incluido en el grafico el punto perteneciente a la familia Bacteriovoracaceae aunque no se concrete el género al que pertenece ya que su $-\log_{10}$(p-value) es mucho mayor al resto de puntos diferencialmente expresados de las muestras de aguas.

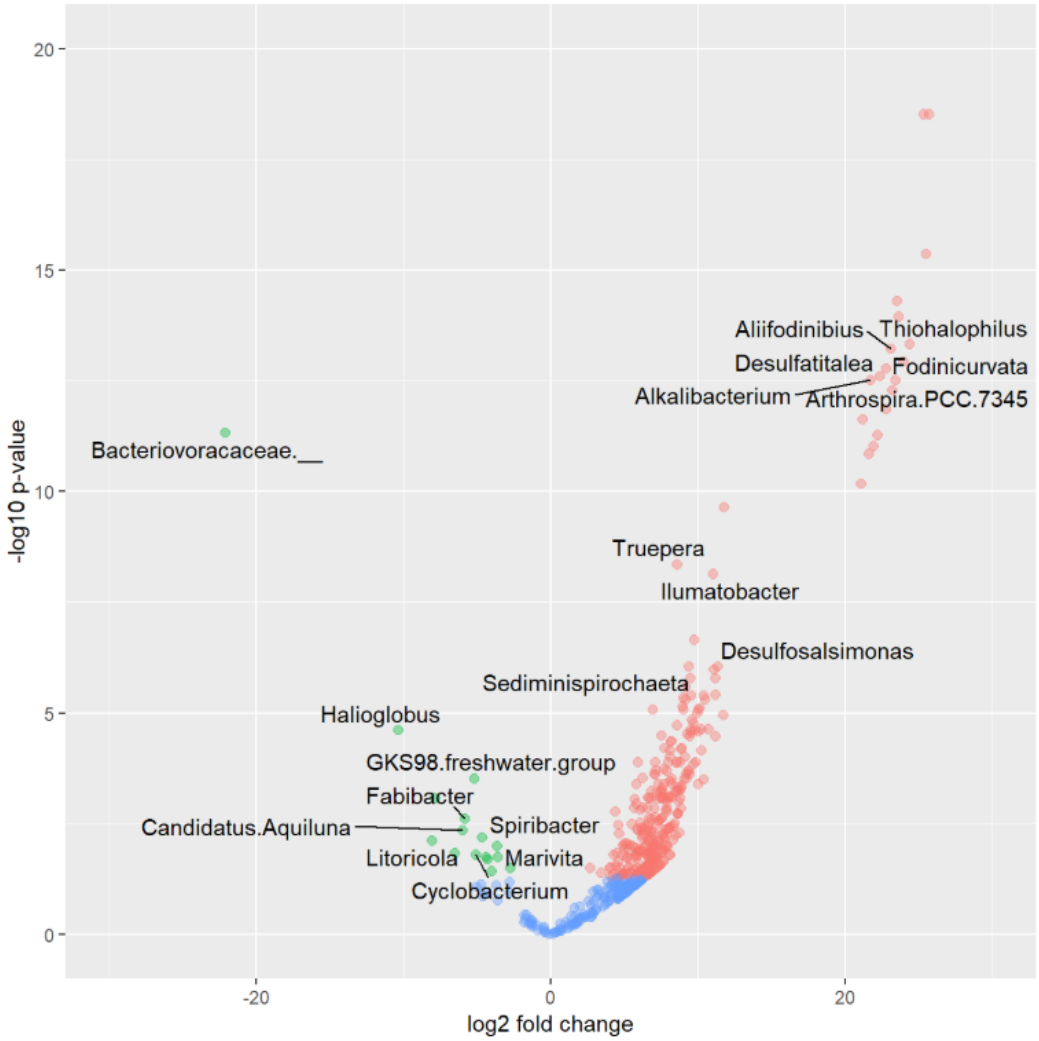

Figura 21: Abundancia diferencial entre las muestras de agua y sedimento a nivel de género según el método BLAST

3.7.2 VSEARCH

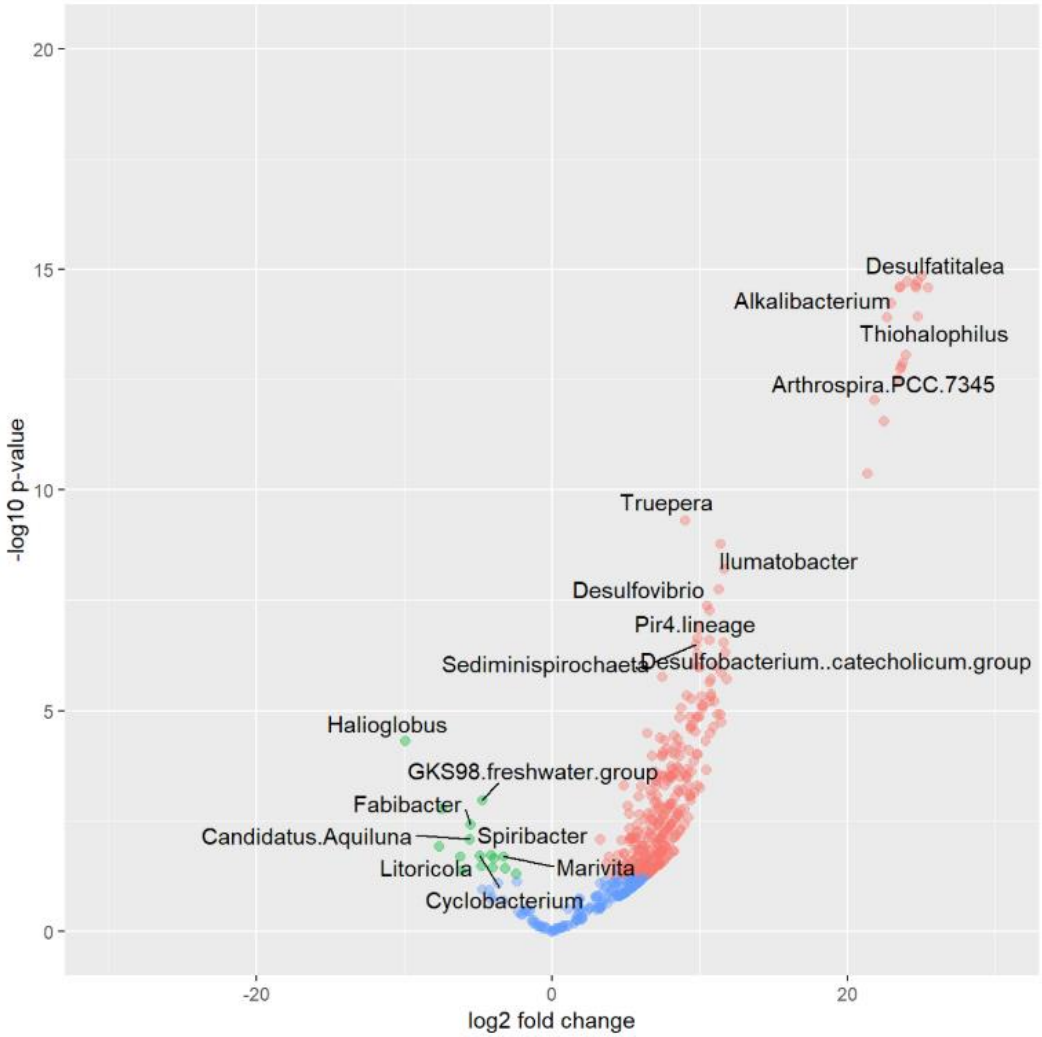

Figura 22: Abundancia diferencial entre las muestras de agua y sedimento a nivel de género según el método VSEARCH

3.7.3 Sklearn:

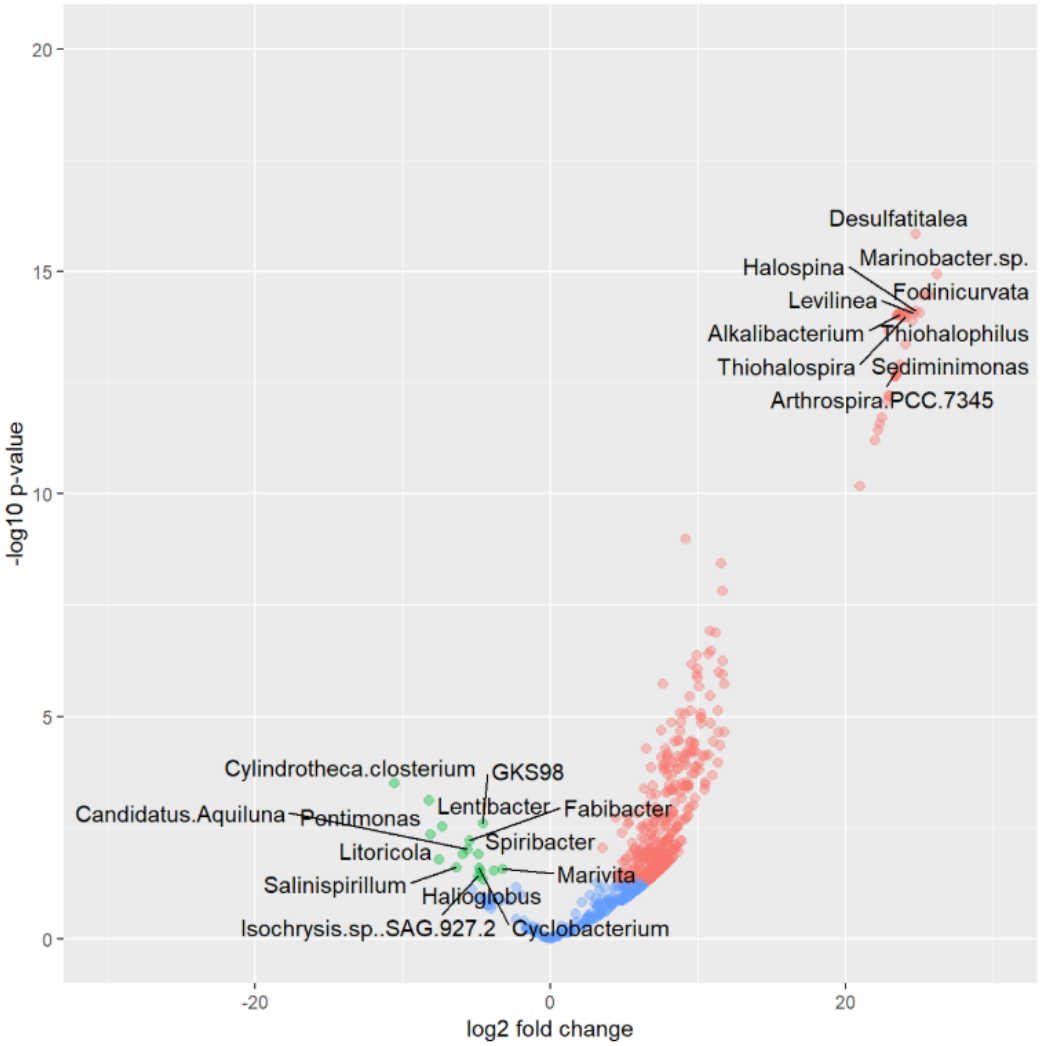

Figura 23: Abundancia diferencial entre las muestras de agua y sedimento a nivel de género según el método sklearn

El numero de taxones diferencialmente abundantes entre agua y sedimentos son, para cada metodo son:

-BLAST: 325
-VSEARCH: 346
-Skleran: 402

Se puede observar que hay muchos mas taxones diferencialmente abundantes con el metodo Sklearn que con los otros dos metodos.

Tambien se puede observar que el metodo sklearn obtiene varios taxones diferencialmente expresados en favor de las muestras de agua que los otros dos metodos, los cuales obtienen exactamente los mismos taxones.

3.8 Comparación de los resultados con otros artículos

Se han revisado un total de cinco artículos donde se estudia la diversidad microbiana de lagunas y sistemas acuáticos en agua y sedimentos con distintos niveles de salinidad. Los puntos clave en los que se ha centrado esta comparación han sido los distintos filos encontrados en cada sitio y los índices de diversidad alfa.

En el artículo **"Spatial distribution of prokaryotic communities in hypersaline soils"** [22] se ha observado que en muestras de sedimentos en ambientes hipersalinos existen los siguientes filos:

Euryarchaeota, Proteobacteria, Bacteroidetes, *Balneolaeota*, R*hodothermaeota, Cyanobacteria, Ca*. Nanohaloarchaeota, *Gemmatimonadetes*, *Planctomycetes*, *Actinobacteria, Firmicutes*, *Verrucomicrobia*, *Chloroflexi* y D*einococcus-Thermus*

El índice Shannon de diversidad alfa tiene un valor medio de 5.5 mientras que el índice Simpson de diversidad alfa tiene un valor medio de 0.98.

En comparación con nuestros datos se puede observar que *Balneolaeota* y *Rhodothermaeota* no se encuentran entre los filos presentes en nuestros datos, esto es debido a que en este estudio han actualizado manualmente la base de datos SILVA para incluir estos filos.

En el artículo **"Microbiome analysis and bacterial isolation from Lejía Lake soil in Atacama Desert"** [23] se ha observado que en muestras de sedimentos en ambientes hipersalinos existen los siguientes filos:

Más del 1%: *Acidobacteria, Actinobacteria, Bacteroidetes, firmicutes, proteobacteria* y *thermi*.

Menor del 1%*: Armatimonadetes*, *Chlorobi*, *Chloroflexi*, *Cyanobacteria*, *Fusobacteria*, *Gemmatimonadetes*, *Planctomycetes*, *Spirochaetes*, *Synergistetes*, *Tenericutes* y *Verrucomicrobia*.

El índice Shannon de diversidad alfa tiene un valor medio de 6.5.

En nuestros resultados, el filo *Chlorobi* ha sido clasificado como una clase dentro del filo B*acteroidetes*

En el artículo **"Distribution of sediment bacterial and archaeal communities in plateau freshwater lakes"** [24] se ha observado que en muestras de sedimentos en ambientes no salinos existen los siguientes filos:

Verrucomicrobia, planctomycetes, nitrospirae, gemmatimonadetes, firmicutes, cyanobacteria, chloroflexi, chlorobi, bacteroidetes, actinobacteria, acidobacteria, epsilonproteobacteria, deltaproteobacteria, gammaproteobacteria, betaproteobacteria, alphaproteobacteria.

El índice Shannon de diversidad alfa tiene un valor de más de 7 en todas las muestras.

En el artículo **"Biodiversity analysis of the unique geothermal microbial ecosystem of the Blue Lagoon (Iceland) using nextgeneration sequencing (NGS)"** [25] se ha observado que en muestras de agua en ambientes geotermales existen los siguientes filos:

Verrucomicrobia, thermotogae, spirochaetae, epsilonproteobacteria, deltaproteobacteria, gammaproteobacteria, betaproteobacteria, alphaproteobacteria, planctomycetes, firmicutes, cyanobacteria, chloroflexi, chlamydiae, bacteroidetes, aquificae, actinobacteria

El índice Shannon de diversidad alfa tiene un valor medio de 2.72 mientras que el índice Simpson de diversidad alfa tiene un valor medio de 0.175.

En el artículo **"Bacterial Communities of Three Saline Meromictic Lakes in Central Asia"** [26] se ha observado que en muestras de agua en ambientes salinos existen los siguientes filos:

Proteobacteria, cyanobacteria, bacteroidetes, actinobacteria, firmicutes, tenericutes, verrucomicrobia y *candidate divisions.*

El índice Shannon de diversidad alfa tiene un valor medio de 1.99 mientras que el índice Simpson de diversidad alfa tiene un valor medio de 0.044.

Para poder comparar los resultados de los distintos artículos se ha calculado la media de los dos índices de diversidad obtenidos según si las muestras corresponden a agua o sedimento:

	Media índice Shannon	Media índice Simpson
Muestras de Agua	2,756	0,877
Muestras de Sedimento	4,612	0,978

Podemos observar que en comparación a todos los artículos donde se ha mirado la diversidad en muestras de sedimentos, nuestras muestras tienen una diversidad menor, sobre todo respecto a las muestras de ambientes no salinos.

Sin embargo las muestras de agua muestran una mayor diversidad en los dos índices respecto a los dos artículos de ambientes geotermales y salinos.

4. Conclusiones

Las conclusiones extraídas de los resultados según los apartados son:

Reducción de secuencias mediante DADA2:

Este ha sido el proceso que más tiempo ha necesitado para realizarse ya que se necesita una gran capacidad de procesamiento, el cual se ha conseguido gracias al uso de la supercomputadora. Se puede considerar como el cuello de botella en este tipo de análisis.

Diversidad taxonómica en las muestras de agua y sedimento:

- El apartado clave para realizar la asignación taxonómica es la elección de la base de datos, ya que si se elige una base de datos desfasada o incompleta podría influir de manera muy negativa en los resultados.
- El método sklearn obtiene mejores resultados ya que presenta un menor número de lecturas asignadas al apartado “unassigned”. Esto explica que haya una mayor diversidad alfa con este método frente a los otros dos.

Diversidad Alfa: Comparación de la diversidad a nivel de género de entre los puntos de muestro:

Se puede ver que las muestras de sedimentos tienen una diversidad mucho mayor que las muestras de agua, sin embargo las muestras de agua tienen repartida la abundancia relativa de cada taxón de una manera más uniforme.

Diversidad Beta: Comparación a nivel de género de entre los puntos de muestro:

Los puntos 2650, 2651 y 2635 se ven afectados por las presiones antropogénicas que existen sobre la laguna, influyendo en las muestras de sedimentos de los puntos 2650 y 2651 y en la muestra de agua del punto 2635. Estas presiones son, respectivamente, los vertidos de aguas residuales y el aislamiento de la laguna para su uso como salinera.

Comparación de los resultados con otros artículos:

La laguna de Pétrola posee una mayor diversidad en las aguas que otros sistemas de características similares, sin embargo esta diversidad se ve mermada debido a la distribución de la abundancia en los organismos, dando lugar a muchos organismos muy poco abundantes y unos pocos que abarcan la mayor parte de la abundancia total.

En futuros trabajos se intentará describir la relación de la composición taxonómica de los distintos puntos con las características físico-químicas (pH, O2, saturación de O2, temperatura, Eh, TDS, Conductividad, NH_4^+, SO_4^{2-}, carbono total, carbono inorgánico y nitrógeno total) que definen cada uno de los puntos

5. Glosario

16S rRNA: Componente de la subunidad 30S de los ribosomas procariontes que se une a la secuencia de Shine-Dalgarno.

ASV: Amplicon Sequence Varian, método para analizar datos de secuenciación de alto rendimiento de genes marcadores, que controla los errores de manera que las variantes de secuencia de amplicón se puedan resolver exactamente, a nivel de las diferencias en un solo nucleótido sobre la región del gen secuenciado.

Clasificado Naive Bayes: en teoría de la probabilidad y minería de datos, es un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales.

Distancia Bray-Curtis: estadístico usado para cuantificar la diferencia de composición entre dos sitios diferentes, en función de los recuentos de cada sitio. Toma valores entre 0 y 1, donde 0 significa que los dos sitios tienen la misma composición (que comparten todas las especies con la misma abundancia), y 1 significa que los dos sitios no comparten ninguna especie.

Diversidad alfa: riqueza de especies de una comunidad particular a la que consideramos homogénea.

Diversidad beta: grado de cambio en la composición de la comunidad o grado de diferenciación de la comunidad, en relación a un ambiente de gradiente complejo o un patrón de ambientes.

Gráfico volcano: en estadística, es un tipo de diagrama de dispersión que se usa para identificar rápidamente los cambios en grandes conjuntos de datos compuestos de datos duplicados.

Metagenoma: conjunto de genes microbianos presentes en un entorno o ecosistema determinado.

Máquina virtual: software que simula un sistema de ordenador. Se puede instalar en el sistema operativo de elección y puede ejecutar programas como si fuese un ordenador real

Primer: es una cadena corta de ARN o ADN (de alrededor de 18-22 bases) que sirve de punto de partida para la síntesis de ADN

Python: lenguaje de programación orientado a objetos, de código abierto e interpretado que no necesita compilar el código fuente para poder ejecutarse.

6. Bibliografía

1- Gómez Alday, J., Castaño Fernández, S. and Sanz Martínez, D. Origen geológico de los contaminantes (Sulfatos) presentes en las aguas subterráneas de la Laguna de Pétrola. (Albacete, España). Resultados preliminares, Geogaceta, 35, pp. 167-170, 2004.

2- Gómez Alday, J., Castaño Fernández, S. and Sanz Martínez, D. Contribución al estudio de la salinización en las aguas subterráneas de la cuenca endorreica de la laguna de Pétrola (Pétrola, Albacete). Sabuco: revista de estudios albacetenses, 6, pp.9-31, 2008.

3- JJ Gómez-Alday, R Carrey, N Valiente, N Otero, A Soler, C Ayora, D Sanz, A Muñoz-Martín, S Castaño, C Recio, A Carnicero, A Cortijo. Denitrifecation in a hypersaline lake-aquifer system (Pétrola Basin, Central Spain): The role of recent organic matter and Cretaceous organic rich sediments. Science of the Total Environment, 497, pp. 594-606, 2014.

4- Valiente, N., Menchen, A., Carrey, R., Otero, N., Soler, A., Sanz, D. and Gomez-Alday, J. (2017). Sulfur Recycling Processes in a Eutrophic Hypersaline System: Pétrola Lake (SE, Spain). Procedia Earth and Planetary Science, 17, pp.201-204.

5- Fernando Pérez, Brian E. Granger, *IPython: A System for Interactive Scientific Computing*, Computing in Science and Engineering, vol. 9, no. 3, pp. 21-29, May/June 2007.

6- Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope E, Da Silva R, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley G, Janssen S, Jarmusch AK, Jiang L, Kaehler B, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MG, Lee J, Ley R, Liu Y, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton J, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson, II MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CH, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. PeerJ 2018.

7- R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018

8- Oksanen, Jari F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs and Helene Wagner. Vegan: Community Ecology Package. R package version 2.5-4. 2019.

9- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. 2016.

10- Carson Sievert (2018) plotly for R.

11- Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. Dada2: high-resolution sample inference from illumina amplicon data. Nature methods, 13(7):581. 2016.

12- Pelin Yilmaz, Laura Wegener Parfrey, Pablo Yarza, Jan Gerken, Elmar Pruesse, Christian Quast, Timmy Schweer, Jörg Peplies, Wolfgang Ludwig, Frank Oliver Glöckner, The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks, *Nucleic Acids Research*, Volume 42, Issue D1, Pages D643–D648, 1 January 2014.

13- Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. Blast+: architecture and applications. *BMC bioinformatics*, 10(1):421, 2009.

14- Torbjørn Rognes, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. Vsearch: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584, 2016.

15- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

16- Shannon, C.E. and W. Weaver. The Mathematical Theory of Communication. University Illinois Press, Urbana, IL. 1949.

17- Simpson, E.H. Measurement of Diversity. Nature, 163: 688. 1949.

18- Cottam G., Goff F.G., Whittaker R.H. Wisconsin Comparative Ordination. In: Whittaker R.H. (eds) Ordination of Plant Communities. Handbook of Vegetation Science, vol 5-2. Springer, Dordrecht, 1978.

19- Bray, J. R. and J. T. Curtis. An ordination of upland forest communities of southern Wisconsin. Ecological Monographs 27:325-349. 1957.

20- Love, M.I., Huber, W., Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 Genome Biology 15(12):550, 2014.

21- Cock, P. J. A.; Fields, C. J.; Goto, N.; Heuer, M. L.; Rice, P. M. "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants". Nucleic Acids Research. 38 (6): 1767–1771. 2009.

22- Blanca Vera-Gargallo, Taniya Roy Chowdhury, Joseph Brown, Sarah J. Fansler, Ana Durán-Viseras, Cristina Sánchez-Porro, Vanessa L. Bailey, Janet K. Jansson & Antonio Ventosa. Spatial distribution of prokaryotic communities in hypersaline soils. Scientific Reports. 9. 2019.

23- Mandakovic, D., Maldonado, J., Pulgar, R. et al. Microbiome analysis and bacterial isolation from Lejía Lake soil in Atacama Desert Extremophiles 22: 665, 2018.

24- Zhang, J., Yang, Y., Zhao, L., Li, Y., Xie, S., & Liu, Y. Distribution of sediment bacterial and archaeal communities in plateau freshwater lakes. Applied Microbiology and Biotechnology, 99(7), 3291-3302. 2014.

25- Palinska, K., Vogt, J., & Surosz, W. Biodiversity analysis of the unique geothermal microbial ecosystem of the Blue Lagoon (Iceland) using next-generation sequencing (NGS). Hydrobiologia, 811(1), 93-102, 2017.

26- Baatar, B., Chiang, P., Rogozin, D., Wu, Y., Tseng, C., & Yang, C. et al. Bacterial Communities of Three Saline Meromictic Lakes in Central Asia. PLOS ONE, 11(3), e0150847, 2016.

7. Agradecimientos

Este trabajo ha sido financiado por los proyectos de investigación FEDER/Ministerio de Ciencia e Innovación y Universidades-Agencia Estatal de Investigación/CICYT-CGL2017-87216-C4-2-R del gobierno de España, PEIC-2014-004-P y SBPLY/17/180501/000296L del gobierno regional de Castilla–La Mancha.

Me gustaría agradecer al centro de supercomputación I^3A del instituto de investigación en informática de Albacete de la universidad de Castilla La Mancha, por cederme los recursos necesarios para realizar este trabajo.

Por último y especialmente, me gustaría agradecerles a todos los que han estado a mi lado durante todo este proceso. A mis tutores, Andreu Paytuví y Juan Jose Gomez, por haberme ayudado tanto. A mis compañeros Nicolas, Bea, Alfonso, Yolanda, Danka y Manu por haberme hecho reír tantísimo y compartir vuestro tiempo conmigo. A Jose Antonio, por organizar unas fiestas tan increíbles y amenizarnos los descansos con tus conocimientos. A toda la gente de mi grupo scout, porque mi vida es mucho más grande con ellos.

A mis padres por estar siempre ahí y nunca rendirse conmigo. A mi hermana por ser mi mayor inspiración. A toda mi familia. A Isabel, por ayudarme a querer ser mejor persona. Y por último a mi abuelita Maruchi y a mi abuelita Encarnita porque las quiero con locura.

Muchas gracias.

8. Anexos

1. Script con las instrucciones de QIIME2 en el notebook jupyter.
2. Documento en HTML con el código y los resultados del análisis de abundancia diferencial y la diversidad alfa y beta.
3. Tabla de secuencias representativas y la asignación de taxones por los métodos de BLAST, VSEARCH y Sklearn