UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS

DE LA INFORMACIÓN

ALINEAMIENTO DE SECUENCIAS PARA CONJUNTO DE DATOS DE MICROBIOMA PARA VAGINOSIS BACTERIANA

Trabajo recepcional bajo la modalidad de Tesis

Que para obtener el grado de:

Maestro en Ciencias de la Computación

Presenta:

Isaí Angulo Jiménez

Directora:

Dra. Juana Canul Reich

Jurado revisor:

Dra. Betania Hernández Ocaña

Dr. Miguel Antonio Wister Ovando

Dr. José Hernández Torruco

Cuerpo Académico:

Inteligencia Artificial

Línea de Generación y Aplicación del Conocimiento:

Representación y Manejo del Conocimiento

Cunduacán, Tabasco, México. Mayo 2022

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
"ESTUDIO EN LA DUDA. ACCIÓN EN LA FE"

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

México 2021

"2021, Año de la Independencia de México"

Cunduacán, Tabasco a 06 de octubre de 2021
Oficio No. 1306/DACYTI/CP/2021

Asunto: Dirección de Tesis

Dra. Juana Canul Reich
Profesor Investigador

De conformidad con lo establecido en el Reglamento de Estudios de Posgrado Vigente, de la Universidad Juárez Autónoma de Tabasco, me permito informarle, que ha sido designado como Director de la tesis titulada **"ALINEAMIENTO DE SECUENCIAS PARA CONJUNTOS DE DATOS DE MICROBIOMA PARA VAGINOSIS BACTERIANA"**, a realizar por el **C. Isaí Angulo Jiménez** matrícula **182H13001**, para obtener el grado de Maestro en Ciencias de la Computación.

Sin otro particular, aprovecho la ocasión para enviarle un afectuoso saludo.

Atentamente

MTE. Óscar Alberto González González
Director

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

C.c.p. Dr. Eddy Arquímedes García Alcocer. Encargado del Despacho de la Coordinación de Posgrado
Alumno. C. Isaí Angulo Jiménez.
Archivo

MTE'OAGG/EAGA

Carretera Cunduacán-Jalpa Km. 1, Colonia Esmeralda. C.P. 86690.
Cunduacán, Tabasco, México.
Tel: (993) 358 1500 ext. 6727; (914) 336 0616; Fax: (914) 336 0870
E-mail: direccion.dacyti@ujat.mx

www.ujat.mx

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

F6: Respuesta de jurado

Cunduacán, Tabasco, a 22 de marzo de 2022.

MTE. Óscar Alberto González González
Director de la División Académica de Ciencias y Tecnologías de la Información
Presente

En atención a los oficios girados por usted, en los que se nos designa como parte del jurado para efectuar la revisión de la tesis titulada ***"Alineamiento de secuencias para conjunto de datos de microbioma para vaginosis bacteriana"***, realizada por el ***C. Isaí Angulo Jiménez***, estudiante de la Maestría en Ciencias de la Computación, nos permitimos informarle que, en virtud de que ha atendido las observaciones realizadas, otorgamos nuestra aprobación para que continúe los trámites para la obtención del grado.

Sin otro particular, aprovechamos la ocasión para enviarle un cordial saludo.

Atentamente integrantes del jurado

Dra. Betania Hernández Ocaña

Dr. Miguel Antonio Wister Ovando

Dr. José Hernández Torruco

c.c.p. Mtro. Eddy Arquímedes García Alcocer. Encargado del despacho de la Coordinación de Posgrado.
Estudiante.

Recibí 22/03/2022 11:31 hrs

RECIBIDO
22 MAR. 2022
D.A.C.Y.T.I.
DIRECCIÓN

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

F8: Cesión de Derechos

El que suscribe, autoriza por medio del presente escrito a la Universidad Juárez Autónoma de Tabasco para que utilice tanto física como digitalmente la Tesis de Maestría titulada **"Alineamiento de secuencias para conjunto de datos de microbioma para vaginosis bacteriana"**, de la cual soy autor y titular de los derechos de autor.

La finalidad del uso por parte de la Universidad Juárez Autónoma de Tabasco de la tesis antes mencionada será única y exclusivamente para difusión, educación, sin fines de lucro; autorización que se hace de manera enunciativa mas no limitativa para subirla a la Red Abierta de Bibliotecas Digitales (RABID) y a cualquier otra red académica con las que la Universidad tenga relación institucional.

Por lo antes mencionado, libero a la Universidad Juárez Autónoma de Tabasco de cualquier reclamación legal que pudiera ejercer respecto al uso y manipulación de la Tesis mencionada y para los fines estipulados en este documento.

Se firma la presente autorización en la ciudad de Cunduacán, Tabasco a los 22 días del mes de marzo de 2022.

Autorizo

C. Isaí Angulo Jiménez

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
"ESTUDIO EN LA DUDA. ACCIÓN EN LA FE"

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

2022 Ricardo Flores Magón
PRECURSOR DE LA REVOLUCIÓN MEXICANA

"2022, Año de Ricardo Flores Magón"

Cunduacán, Tabasco a 22 de marzo de 2022
Oficio No. 0374/DACYTI/CP/2022

Asunto: Autorización de impresión de Tesis

C. Isaí Angulo Jiménez
Matricula: 182H13001

En virtud de que cumple satisfactoriamente los requisitos establecidos en el Reglamento General de Estudios de Posgrado vigente en la Universidad, informo a Usted que se autoriza la impresión del trabajo recepcional **"ALINEAMIENTO DE SECUENCIAS PARA CONJUNTO DE DATOS DE MICROBIOMA PARA VAGINOSIS BACTERIANA"**, para presentar examen y obtener el Grado de Maestro en Ciencias de la Computación.

Sin otro particular, aprovecho la ocasión para enviarle un afectuoso saludo.

Atentamente

MTE. Óscar Alberto González González
Director

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

C.c.p. Dr. Eddy Arquímedes García Alcocer. - Encargado del Despacho de la Coordinación de Posgrado DACYTI
Archivo.
Consecutivo.

MTE´OAGG/EAGA

Carretera Cunduacán-Jalpa Km. 1, Colonia Esmeralda, C.P. 86690.
Cunduacán, Tabasco, México.
Tel: (993) 358 1500 ext. 6727; (914) 336 0616; Fax: (914) 336 0870
E-mail: direccion.dacyti@ujat.mx

www.ujat.mx

Agradecimientos

Agradezco a la máxima casa de estudios de los tabasqueños, la Universidad Juárez Autónoma de Tabasco, por brindarme todo tipo de facilidades a lo largo de mis estudios de maestría, así como la oportunidad de conocer otros estados realizando actividades académicas y permitiéndome crecer mi acervo cultural. A los profesores que dedicaron tiempo y esfuerzo en su labor docente al colaborar conmigo y que en más de una ocasión velaron por mi bien y salir delante de la mejor forma. A mis compañeros de clases, pero sobre todo a los que no conocí en un aula, llegando a compartir más tiempo fuera de una que dentro. Todos fueron pilar fundamental para culminar mi última etapa de la maestría.

Dedicatoria

A Dios, por concederme la dicha de un hogar conformado por mis padres, mi hermana y sus hijos, las chicas que siempre nos han ayudado. También me concedió a mi ayuda idónea junto con sus padres y su hermana y todo lo que ello implica. Me permitió compartir aula con notables profesores, así como con alumnos sobresalientes. Me ha permitido compartir de lo mucho que me ha dado, al igual que me permite presentar este trabajo como culminación de un trecho largo que veía muy distante, pero ahora se vuelve una realidad.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Generalidades

1.1. Introducción

El análisis computacional de los datos de secuencias a menudo implica el uso de distintos programas, códigos o herramientas que no necesariamente se encuentren relacionados entre sí, o pueden estar escritos en diferentes lenguajes de programación, diseñados para plataformas específicas, o en el peor de lo casos poseer una implementación poco intuitiva. Las tareas de preprocesamiento, clasificación taxonómica, alineamiento de múltiples secuencias (MSA) y obtención de árboles filogenéticos emplean métodos matemáticos que están implementados en diferentes herramientas, lo que ocasiona que existan diversos formatos para su almacenamiento, haciendo tediosa la tarea del análisis. R (R Core Team, 2021), a pesar de ser un entorno para estadística, permite análisis filogenético a través de paquetes e implementación de funciones, lo cual ayuda a que el manejo de secuencias y sus respectivos análisis puedan realizarse dentro de una sola plataforma.

Una tarea central para el análisis filogenético es la obtención de un árbol de distancias entre las secuencias, para lo cual se requiere que las secuencias sean alineadas previamente (Daugelaite et al., 2013). Los MSA logran su objetivo mediante la programación dinámica, método que requiere tiempo y espacio en memoria de orden $N * M$, en donde N y M son el ancho de las secuencias a y b, respectivamente. Debido a la complejidad que implica el alineamiento de secuencias largas, heurísticas son utilizadas para acelerar el alineamiento sin impactar negativamente en la precisión.

Esta precisión varía en función del número de secuencias que son añadidas al alineamiento, pues puede ser que la identidad entre secuencias o similitud cada vez sea menor, lo cual implicaría la obtención de un alineamiento impreciso. Este problema se encuentra frecuentemente en muestras de secuencias donde hay gran diversidad bacteriana, como en el caso de la microbiota vaginal.

De acuerdo con Ortiz-Rodríguez et al. (2000) la vaginosis bacteriana (VB), *"de origen polimicrobiano, es una alteración de la ecología vaginal donde la flora normal se ve prácticamente sustituida por gérmenes anaerobios. Muchos microorganismos han sido propuestos como causa de esta enfermedad, como la Gardnerella, Atopobium, Leptotrichia, Sneathia spp"*.

Callahan et al. (2016a) describen que el microbioma está compuesto por las comunidades ecológicas de microorganismos que dominan el mundo de los seres vivos. Actualmente las bacterias pueden ser identificadas a través del uso de tecnología de secuenciación masiva (NGS) aplicada en varios niveles. En toda bacteria está presente un "gen huella dactilar" denominado 16S ácido ribunocleico ribosomal (rRNA) y es útil para la identificación y cuantificación de taxas individuales (o especies). Este gen presenta regiones variables que pueden ser utilizadas para identificar distintas taxas.

Acorde con Callahan et al. (2016a), la secuenciación de alto rendimiento de marcadores taxonómicos amplificados mediante reacciones en cadena polimerasa permite un análisis de comunidades bacterianas complejas conocidas como microbiomas. Muchas herramientas existen para cuantificar y comparar los niveles de abundancia o composición de unidades taxonómicas operacionales de comunidades en condiciones diferentes. Las lecturas de secuenciación deben ser sometidas a tareas de la preparación de datos, tales como limpieza de datos, normalización de datos e identificación de ruido. Posteriormente se aplican tareas de reducción de datos, como selección de instancias para remover el ruido encontrado.

1.2. Planteamiento del problema

1.2.1. Definición del problema

De acuerdo con Aguilera-Arreola y Giono-Cerezo, los métodos Nugent y Amsel se basan en la *cuantificación de la abundancia de morfotipos de bacterias relacionados a Lactobacillus spp., G. vaginalis, Bacteroides spp., y Mobiluncus spp.* Esta cuantificación involucra observaciones en microscopio, presencia de olores en las muestras y, por ejemplo, Nugent es más barato que Amsel pero más compleja en su interpretación.

Aguilera-Arreola y Giono-Cerezo también señalan el uso de AFFIRM VPIII®, *"detección química, enzimática o genética de G. vaginalis, Candida spp. y a Trichomonas vaginalis, sin embargo, sólo detecta a uno de los posibles microorganismos involucrados en la VB y la gardnerela, debería estar en la secreción en una concentración mayor a* $2 * 10^5$ *UFC para poder ser detectada".*

D'Amore et al. (2016) hacen un estudio involucrando a distintas plataformas de secuenciación: MiSeq (Illumina), The Pacific Biosciences RSII, 454 GS-FLX/+ (Roche), and IonTorrent (Life Technologies) y comparan el rendimiento y resultados de secuenciación. Concluyen que la composición resultante siempre presenta un sesgo que depende de la plataforma, región secuenciada y selección de *primers*, pero la secuenciación del gen 16S rRNA es cuantitativa y los cambios en la abundancia de las distintas taxa entre muestras pueden ser recuperados.

El alineamiento que requieren las secuencias de un conjunto de datos de microbioma para vaginosis bacteriana se puede realizar utilizando distintos métodos o algoritmos, los cuales a su vez varían en su rendimiento y costo computacional y están implementados en distintas plataformas (sistemas operativos, equipos especializados).Por esta razón, se dificultan las tareas propias del alineamiento de múltiples secuencias.

1.3. Delimitación de la investigación

1.3.1. Alcances

El alcance de la investigación involucra los siguientes aspectos:

1. Aplicación de métodos de alineamiento de múltiples secuencias a muestras vaginales.
2. Obtención de árboles filogenéticos de cada método MSA.
3. Validación estadística del resultado de cada método MSA.
4. Identificación de la composición bacteriana dentro de las muestras.
5. Agrupamiento de cinco tipos de comunidades presentes en las muestras originales.

1.3.2. Limitaciones

La fuente de los datos es el repositorio público del European Nucleotide Archive (ENA), con número de acceso PRJNA544732 (https://www.ebi.ac.uk/ena/browser/view/PRJNA544732).

Se utilizó un equipo de cómputo de uso personal, y las etapas de la investigación se realizaron únicamente de manera local.

1.4. Preguntas de investigación e hipótesis

Debido a la problemática anterior, surgen los siguientes cuestionamientos:

- ¿Qué métodos se utilizan para el alineamiento de múltiples secuencias para conjunto de datos de microbioma?
- De las herramientas utilizadas ClustalW, Clustal Omega, MUSCLE y DECIPHER, ¿cuáles arrojan resultados con significancia estadística?

Con los métodos de alineamiento de múltiples secuencias utilizados fue posible la identificación de comunidades bacterianas presentes en la vaginosis bacteriana, y con la validación estadística aplicada a los 4 métodos MSA, el método MSA DECIPHER resulta el más viable biológicamente con un $MOS = 0.7$

1.5. Objetivos

1.5.1. Objetivo general

Aplicar métodos de alineamiento de múltiples secuencias a un conjunto de datos de microbioma para encontrar bacterias que predominen en la microbiota de la vaginosis bacteriana.

1.5.2. Objetivos específicos

- Contrastar el alineamiento progresivo contra el enfoque iterativo de los métodos de programación dinámica para la resolución del problema de alineamiento de múltples secuencias para datos de microbioma vaginal.
- Alinear las secuencias a través de los métodos ClustalW, Clustal Omega, MUSCLE y DECIPHER.
- Obtener el árbol filogenético de cada método MSA para el análisis comparativo visual de los cuatro métodos.
- Evaluar estadísticamente de los métodos MSA aplicando las funciones AOS, MOS y OS.

1.6. Justificación

De acuerdo con Delgado Moya (2005), se proponen las siguientes interrogantes referentes al desarrollo de la investigación: ¿por qué es importante?, ¿para qué va a servir?, y ¿a quién va a beneficiar? Es importante evaluar el resultado de los métodos de alineamiento de múltiples secuencias porque implican un costo computacional

proporcional al número de secuencias de entrada. En el presente caso, el número de secuencias, 1,974, supone que el costo sea elevado, pero se requiere utilizar una herramienta que permita obtener resultados en el menor tiempo posible, y a su vez con una significancia estadística relevante. Tal como describen Aguilera-Arreola y Giono-Cerezo *"en México, se han realizado diversos estudios sobre la frecuencia de la VB en diferentes poblaciones. La prevalencia varía de 20 a 60 % dependiendo del tipo de población y la metodología utilizada. La literatura internacional indica que en 50 % de los casos, la VB es asintomática. Los usos de métodos moleculares independientes de cultivo para el diagnóstico de VB incluyen al menos: la PCR de amplio rango, la PCR especie específica, la DGGE del gen 16S rRNA, los RFLPPCR del gen 16S rRNA vía FISH y la PCR en tiempo real cuantitativa, todos estos métodos han revelado un panorama más complejo de lo que hasta ahora se ha reconocido."*

La disponibilidad de los datos de microbioma caracterizado mediante la secuenciación del gen 16S rRNA del estudio por el Centro Médico Universitario de Liubliana, Eslovenia, facilitó la labor de análisis. La presente investigación sirve para conjuntar herramientas computacionales ya existentes en un solo flujo de trabajo con el fin de replicar este análisis con datos de microbioma vaginales de cualquier otra región.

1.7. Metodología utilizada

En la Figura 1.1 se ilustra la metodología propuesta para esta investigación, comprendida por una serie de pasos que se describen a continuación:

1. Obtención de las secuencias a utilizar, en este caso pertenecientes a muestras vaginales.

2. Preprocesamiento de las secuencias con las herramientas *Cutadapt* y *DADA2*.

3. Clasificación taxonómica con *DADA2* y la base de datos *Silva* versión 132.

4. Alineamiento de las secuencias resultantes del preprocesamiento. Los métodos utilizados fueron ClustalW, Clustal Omega, MUSCLE y DECIPHER.

5. Análisis estadístico del resultado de las secuencias alineadas utilzando la herramienta *MUMSA*

Dentro de la etapa del preprocesamiento, se describen las tareas realizadas:

1. Filtrado y recorte de las secuencias originales utilizando *Cutadapt*.
2. Generación del modelo de errores usando *DADA2* para verificar la viabilidad del filtro realizado.
3. Inferencia de la composición muestral indicando el número de posibles bacterias presentes en la muestra.
4. Unión de lecturas y remoción de secuencias quimeras para descartar "bacterias creadas"por la inferencia que biológicamente no existen.

El proceso de alinear las secuencias con cada método consistió en los siguientes pasos:

1. Alineamiento de las secuencias con cada método utilizando los parámetros por defecto.
2. Obtención del árbol filogenético que indica la relación entre las secuencias alineadas.
3. Creación de un objeto que incluye la abundancia de las bacterias, sus secuencias y el árbol filogenético

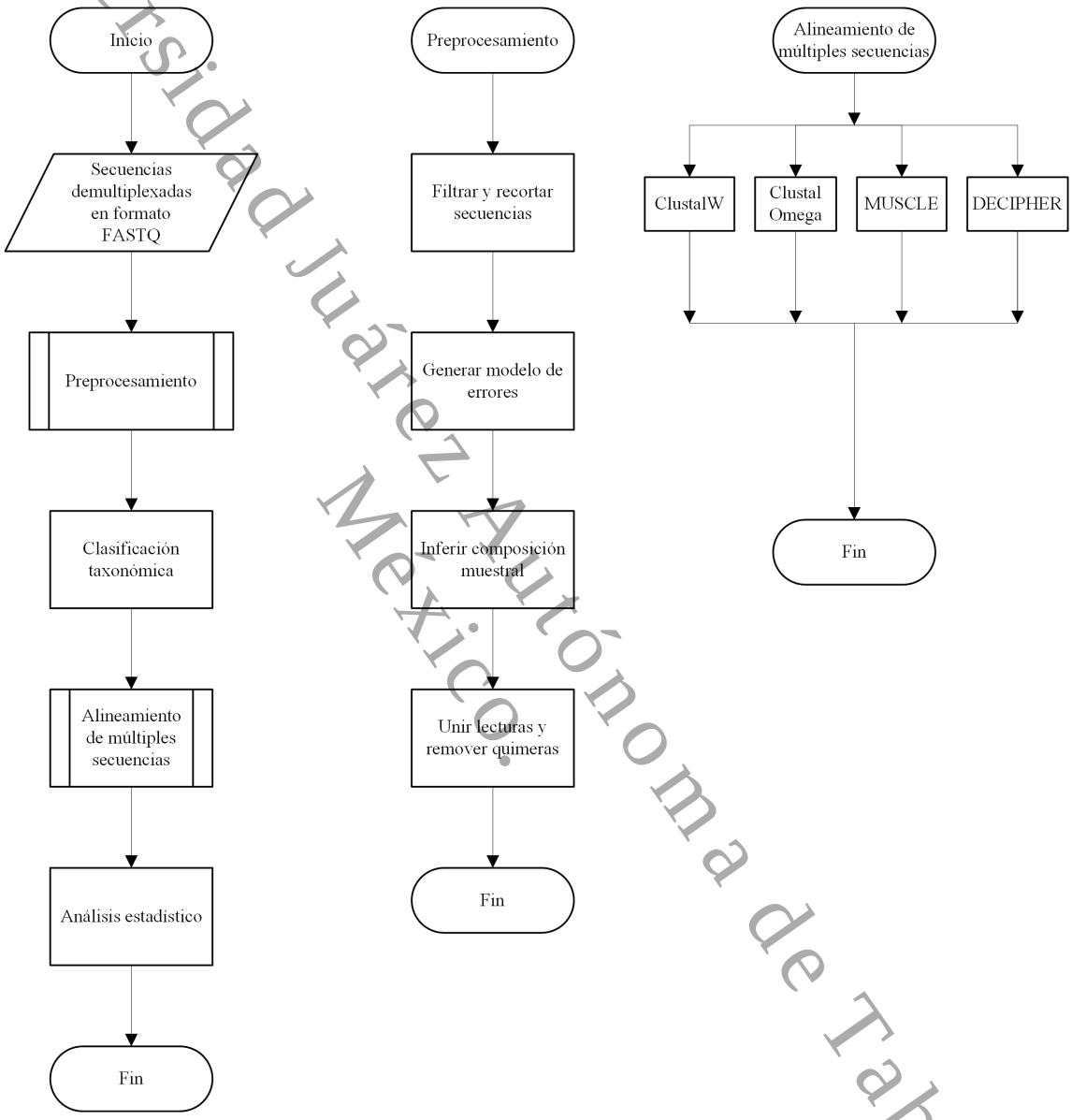


Figura 1.1: Metodología.

Capítulo 2

Marco teórico

2.1. Marco conceptual

A continuación, se definen conceptos que requieren ser comprendidos para la presente investigación, que serán aplicados en el desarrollo computacional que implica alinear las secuencias a través de los algoritmos.

2.1.1. Programación dinámica

Técnica matemática que se utiliza para la solución de problemas, en los cuales se toma una serie de decisiones de manera secuencial. Proporciona un procedimiento sistemático para encontrar la combinación de decisiones que maximice la efectividad total, al descomponer el problema en etapas, parte del problema que posee un conjunto de alternativas mutuamente excluyentes de las cuales se seleccionará la mejor alternativa, que pueden ser completadas de una o más formas o estados, donde se refleja la condición o estado de las restricciones que enlazan las etapas. La programación dinámica provee un alineamiento óptimo para una función objetivo dada y así alinear secuencias pareadas, que consiste en alinear solo dos secuencias (Chowdhury y Garai, 2017).

2.1.2. Alineamiento progresivo

Los métodos progresivos son un enfoque heurístico en donde los problemas de alineamiento de múltiples secuencias son separados en subproblemas. Esto resuelve problemas de alineamiento múltiple con programación dinámica indirectamente. Los métodos progresivos construyen un árbol guía aproximado a partir de distancias calculadas entre todos los pares posibles de secuencias. De la matriz de distancias resultante se construye un árbol usando un método algorítmico rápido. El árbol guía resultante se emplea para construir el alineamiento de manera progresiva. Las dos secuencias más similares se alinean primero usando programación dinámica y una matriz o esquema de ponderación particular. Una vez alineado el primer par, los *gaps* (-) generados ya no se mueven. Este par es tratado como una sola secuencia y es alineado contra la siguiente secuencia o grupo de secuencias más aproximadas en el árbol. Se repite hasta alinear todas las secuencias (Chowdhury y Garai, 2017).

2.1.3. Alineamiento de múltiples secuencias

Solución algorítmica para el alineamiento de más de dos secuencias, fundamental para tareas como búsquedas de homología entre secuencias, anotación genómica, predicción de la estructura de una proteína, áreas de biología evolutiva computacional, redes reguladoras de genes, y genómica funcional. Mediante la ordenación de secuencias de ADN, ARN o proteínas se localizan bloques conservados o regiones de similitud. El alineamiento de múltiples secuencias (MSA, por sus siglas en inglés) está clasificado como un problema NP-completo. Los métodos MSA se dividen en 4 tipos: exactos (programación dinámica), progresivos, basados en consistencia, e iterativos (Issa y Hassanien, 2020).

2.1.4. ClustalW

Este método MSA surge como alternativa para la elección de parámetros fijos, lo cual crea resultados inciertos en distintas etapas del alineamiento. La w hace referencia a "pesos"que indica el esquema que utiliza basado en pesos para minimizar grupos de secuencias sobrerepresentados. Al proponer modificaciones al alineamiento progresivo, (Thompson et al., 1994) crean ClustalW, la cual se convierte en una

herramienta práctica para MSA, siendo su publicación del método una de las más citadas en la literatura. El método se puede caracterizar en 3 grandes etapas:

1. Alineamiento por pares de secuencias por el método k-tupla.
2. Construcción del árbol guía por el método Neighbour joining (NJ).
3. Alineamiento progresivo.

El método k-tupla (Wilbur y Lipman, 1983), heurística basada en la mejor aproximación, se usa para el alineamiento entre pares de secuencias de todos los posibles pares de secuencias. Se usa específicamente cuando el número de secuencias a ser alineadas es grande. Las puntuaciones de similitud son calculadas como el número coincidente de k-tuplas (serie de residuos idénticos, usualmente 2-4 para secuencias nucleótidas) en el alineamiento entre un par de secuencias. La puntuación de similitud se calcula al dividir el número de coincidencias entre el número de residuos pareados de dos secuencias comparadas. Penalidades fijas para cada hueco son sustraídas de la puntuación de similitud, siendo estas últimas puntuaciones convertidas a puntuaciones de distancia al dividir la puntuación de similitud entre 100 y sustrayéndola de 1.0 para proveer el número de diferencias entre sitios.

Posteriormente, la totalidad de k-tuplas entre dos secuencias son almacenadas utilizando una tabla asociativa (hash). Una matriz de puntos entre las dos secuencias se produce, con cada k-tupla coincidente representada por un punto. Las diagonales con el mayor número de coincidencias se encuentran y marcan en cada diagonal superior. Esto indica la región con mayor probabilidad en donde pueda ocurrir la similitud entre dos secuencias. La última fase del método k-tupla es hallar el acomodo completo de las k-tuplas coincidentes

2.1.5. Clustal Omega

Este método MSA es el más reciente de la familia Clustal y en un principio solo alineaba secuencias de proteínas (Sievers et al., 2011); posteriormente se añadió la posibilidad de alinear nucleótidos, por lo cual fue útil para los fines de esta investigación. Las 5 etapas principales son las siguientes:

1. Se produce un alineamiento entre pares con el método k-tupla, mismo método empleado por ClustalW

2. Luego de determinar las puntuaciones de similitud del alineamiento anterior, se hace uso del método mBed de complejidad $O(N \log N)$. mBed embebe cada secuencia en un espacio de n dimensiones, donde n es proporcional a $\log N$. Cada secuencia es reemplazada por un elemento n vector. Cada elemento es la distancia a una de n "secuencias de referencia". Estos vectores son apropiados para ser agrupados rápidamente por métodos k-means o Unweighted Pair Group Method with Arithmetic mean (UPGMA).

3. Clustal Omega hace uso del método k-means++ para agrupar, propuesto por Arthur y Vassilvitskii (2006). El método k-means busca minimizar la distancia cuadrada promedio entre puntos del mismo grupo y k-means++ soluciona el problema de definir centros de grupos para k-means a la vez que mejora la velocidad y precisión del método.

4. Clustal Omega utiliza UPGMA como método para la construcción del árbol guía, método que se basa en un algoritmo de agrupamiento secuencial que identifica homologías locales entre unidad taxonómica operacional (OTU) por orden de similitud. Pares de OTU similares son tratadas como una sola OTU. Subsecuentemente, del nuevo grupo de OTU el par con la mayor similitud se identifica y agrupa, así hasta solo tener un par de OTU.

5. Por último, Clustal Omega utiliza HHalign por Söding (2005), para completar los alineamientos progresivos. Este método alinea dos perfiles de modelos ocultos de Markov (HMM) en lugar de comparar perfiles a perfiles, lo que mejora la sensibilidad y precisión del alineamiento. Todos los métodos de comparación HMM para perfiles-secuencias y secuencias se basan en la puntuación *log-odds*, medida para saber cuánto más probable es que una secuencia sea emitida por un HMM que por un modelo nulo aleatorio.

2.1.6. MUSCLE

MUSCLE (Edgar, 2004) se refiere a "Comparación de MÚltiples Secuencias por Esperanza Logarítmica" y sus principales pasos son:

1. utiliza dos distancias: k-mer para pares de secuencias no alineados y Kimura para pares alineados.

2. UPGMA se utiliza para la construcción de árboles guías.

3. Se construye un alineamiento progresivo basado en el árbol UPGMA lo cuál produce un primer MSA. A partir de aquí, se inicia la segunda fase del método MUSCLE con el fin de mejorar el alineamiento.

4. Se recalcula el árbol guía inicial mediante Kimura, el cuál requiere un alineamiento como parámetro. UPGMA reagrupa las secuencias produciendo así un segundo árbol guía.

5. Se construye un segundo alineamiento progresivo basado en el segundo árbol UPGMA lo cuál produce un segundo MSA.

6. Un tercer MSA se produce a partir del primero y del segundo, mediante al alineamiento de los perfiles. Si la puntuación suma de pares (SPS) se mejora en el segundo MSA, entonces el tercero permanece y el primero es descartado, de lo contrario, se elimina y el primer alineamiento es utilizado.

2.1.7. DECIPHER

Método MSA que toma en cuenta el contexto de las secuencias a través de la predicción de estructuras secundarias en el contexto de una secuencia local, incrementando la precisión del método. Esto permite la generación escalable de alineamientos de secuencias grandes manteniendo una precisión alta aún en conjuntos diversos de secuencias (Wright, 2015). El método se puede generalizar en 4 etapas:

1. Inicializar: Predicción de las probabilidades de estructura secundaria para todas las secuencias de entrada.

2. Preliminar: Cálculo de un árbol guía crudo basado en k-mers compartidos, seguido de un alineamiento progresivo.

3. Iteraciones: Cálculo de un árbol guía UPGMA basado de distancias de pares, seguido de un alineamiento progresivo.

4. Ajustes: División repetida del alineamiento en grupos que son realineados. El alineamiento con la mayor puntuación continúa en el proceso.

2.1.8. Comparación estadística de los métodos MSA

Las tareas tales como las búsquedas de homología entre secuencias, anotación genómica, predicción de la estructura de una proteína, así como áreas de biología evolutiva computacional, redes reguladoras de genes, y genómica funcional dependen del resultado de un método MSA. El resultado obtenido de estas tareas bioinformáticas antes mencionadas tendrá una mayor significancia biológica a mayor precisión del resultado del MSA (Lecompte et al., 2001). Sin embargo, debido a que no existe una función objetivo para medir verdaderamente la precisión o correctividad biológica de un alineamiento, existen métodos basados en distintas suposiciones. La comparación cuantitativa de dos métodos MSA distintos ayuda a tomar decisiones sobre qué regiones están preservadas o cuáles deben ser removidas para tareas posteriores (Lassmann y Sonnhammer, 2005).

Coincidencia (*OS*):

La función refleja la similitud entre dos alineamientos Q_a y Q_b, y está definida como la relación entre la cardinalidad de la intersección de dos conjuntos de residuos alineados y la cardinalidad promedio de cada conjunto:

$$Q_{ab} = \frac{|Q_a \cap Q_b|}{(|Q_a| + |Q_b|)/2}. \tag{2.1}$$

Coincidencia promedio (*AOS*):

Cada alineamiento se representa mediante el concepto de residuos de pares alineados. Cada uno de estos pares son extraídos de todos los alineamientos m de

entrada. La dificultad de un caso de alineamiento está definida por la puntuación de coincidencia promedio entre todos los alineamientos de entrada:

$$AOS = \frac{\sum_{i}^{m-1} \sum_{j=i-1}^{m} O_{ij}}{m\,(m-1)\,/2} \tag{2.2}$$

Esta medida representa qué tan dispersos están los alineamientos en el espacio de todas las soluciones y se seleccionó como medida principal para decidir qué alineamiento utilizar. Para casos simples, un método MSA dará como resultado alineamientos similares y el valor AOS será muy cercano a 1, mientras en casos difíciles su valor será cercano a 0.

Coincidencia múltiple (*MOS*):

Se asignan puntuaciones a cada par de residuos alineados reflejando su proliferación en todos los alineamientos. Sea $n(\sigma)$ el número de los $m-1$ alineamientos que contienen σ. Un par que ocurra en todos los alineamientos es, en consecuencia, asignado con la puntuación mayor $(m-1)$ mientras que un par que ocurre en un solo alineamiento es asignado con la puntuación menor de cero. Estas puntuaciones son sumadas para el alineamiento Q_a para determinar su puntuación de coincidencia múltiple:

$$MOS\,(Q_A) = \frac{\sum n(\sigma) : \sigma \in Q_a}{|Q_a|\,(m-1)} \tag{2.3}$$

El numerador suma las puntuaciones de cada par de residuos alineados presentes en el alineamiento Q_a . El denominador refleja la puntuación máxima posible. Los residuos alineados que son encontrados en varios alineamientos son más confiables, y el alineamiento con el mayor número de tales pares se asume como el más significativo biológicamente.

2.1.9. DADA2

Algoritmo de partición divisiva (DADA) que distingue la variación biológica de los errores aleatorios provenientes del proceso de secuenciación. Su propósito es corregir errores en amplicones (conjunto de moléculas ADN resultantes de una reacción en

cadena de polimerasa, PCR) sin necesidad de construcción de OTU (clasificación de organismos basada en la identidad en el gen marcador 16S rRNA).

DADA2 (Callahan et al., 2016b) es un paquete en R que implementa un modelo estimación de errores de amplicón, lo cual mejora el algoritmo DADA. La composición muestral se infiere al dividir las lecturas de amplicones en partes consistentes con un modelo de error. No requiere secuencias de referencia para su uso y el flujo completo incluido en el paquete se compone de: filtrado, desreplicado, identificación de quimeras y fusión de lecturas pareadas.

2.1.10. Puntuación Phred

Las puntuaciones incluidas en secuencias con puntuación de calidad obtenidas por los productos Illumina, están codificadas mediante la extensión FASTQ, y utilizan una versión modificada del algoritmo de puntuación Phred.

Esta puntuación de calidad Phred es un predictor de la probabilidad de un error en la obtención de una base. Una mayor puntuación implica que una base es más confiable y menos propensa a ser incorrecta. Su escala va del 1 al 40, y se divide en 4 rangos: Q10, Q20, Q30 y Q40. La tasa de error para cada cuartil es de 0.1, 0.01, 0.001 y 0.0001, respectivamente. Es decir, para la obtención de bases con una calidad Q40, una obtención de cada 10,000 es predicha como incorrecta (Illumina, 2014).

2.1.11. Bioinformática

Es un área interdisciplinaria que aplica la informática a la recopilación, almacenamiento, organización, análisis, manipulación, presentación y distribución de información relativa a los datos biológicos o médicos, tales como macromoléculas (por ejemplo, ADN o proteínas). Ha evolucionado para servir de enlace entre las observaciones (datos) y el conocimiento que se deriva (información) sobre, por ejemplo, la función de los procesos y, posteriormente, la aplicación (conocimiento) (Mitra y Acharya, 2005).

2.1.12. rRNA 16S

Valenzuela-González et al. (2015) describen el 16S rRNA como: Secuencia nucleotídica de cadena sencilla de aproximadamente 1500 nucleótidos, contiene nueve regiones (V1-V9) menos conservadas o hipervariables, que son las que aportan la mayor información útil para estudios de filogenética y taxonomía. Las regiones conservadas son de gran ayuda para diseñar sebadores o iniciadores universales que permitan la amplificación de las diversas regiones hipervariables de la gran mayoría de los 16S rRNA de los microorganismos presentes en una comunidad. Todas las células poseen material genético (ADN) el cual sufre mutaciones de manera natural a través del tiempo, las cuales se acumulan en sus secuencias y se transmiten generacionalmente. En la realización de estudios filogenéticos del microbioma cobra importancia el análisis de las secuencias de ADN (Garzón-Pinto, 2017).

2.1.13. Microbiota

Es la totalidad de microbios (bacteria, archaea y fungi) en un ambiente en particular. Se refiere a la taxonomía y abundancia de los microbios miembros de la comunidad (Schlaeppi y Bulgarelli, 2015).

2.1.14. Microbioma

Totalidad de genomas de la microbiota. A menudo utilizado para describir la identidad de los rasgos microbianos (funciones) codificados en la microbiota (Schlaeppi y Bulgarelli, 2015). Moya (2017) describe al microbioma como: "el ecosistema interno constituido por el hombre y los microorganismos que en él conviven, los cuales colonizan y viven de manera fisiológica en el cuerpo humano. Factores fisiológicos, como temperatura, humedad, o la presencia de nutrientes, entre otros, favorecen el desarrollo de una comunidad bacteriana en determinados ecosistemas del cuerpo humano, estableciéndose así diferentes microbiotas: oral, respiratoria, gastrointestinal, de la piel, vaginal o del tracto urogenital".

2.1.15. Vaginosis Bacteriana

La VB es una alteración de la ecología vaginal donde la flora normal se ve prácticamente sustituida por gérmenes anaerobios. Muchos microorganismos han sido propuestos como causa de esta enfermedad, como la Gardnerella vaginalis y los estreptococos anaerobios; actualmente se conoce que es de origen polimicrobiano. En 1984 se reconoce como síndrome y se adopta el nombre de VB. Numerosos estudios la relacionan con la ocurrencia de afecciones tales como la enfermedad inflamatoria pélvica, la rotura prematura de las membranas ovulares, el bajo peso del recién nacido y la displasia cervicouterina. El diagnóstico de VB se realiza sobre la base de criterios bien establecidos a partir del examen de la secreción vaginal: un pH mayor que 4,5; una prueba de aminas positiva; la existencia de células guías y una leucorrea homogénea no adherente. Se da por positivo aquel caso donde se encuentren al menos 3 criterios. La determinación de el o los agentes etiológicos no es imprescindible para el adecuado manejo y curación de la paciente. Este diagnóstico no requiere de recursos costosos, es rápido y poco laborioso (Ortiz-Rodríguez et al., 2000).

2.2. Marco referencial

Al-Farha et al. (2018) realizaron un diagnóstico para identificar y distinguir entre cuatro microplasmas distintos y Acholeplasma laidlawii, cinco microplasmas en total. Asignaron OTU mediante el paquete BioEdit v.7.0.4.1. y BLAST. Las secuencias de nucleótidos de cepas microplasma relevantes fueron utilizadas como cepas de referencia para alineamiento de secuencias de nucleótidos utilizando ClustalW v.2.

Miller et al. (2015) realizaron un diagnóstico para identificar y diferenciar patógenos de roedores en instalaciones animales: biotipo “Jawetz” y “Heyl” de Pasteurella pneumotropica, Actinobacillus muris, y Haemophilus influenzaemurium. Utilizaron 6 cepas de referencia para una búsqueda de homología entre secuencias usando BLAST.

Ajitkumar et al. (2012) identificaron patógenos de mastitis bovina. Utilizaron 9 cepas de referencias. Las secuencias 16S fueron alineadas, comparadas y se creó un cladograma (árbol filogenético) utilizando el programa ClustalW. La literatura coloca al algoritmo ClustalW como uno de los más citados de todos los tiempos, así como al MSA de ser uno de los métodos de modelado en biología más ampliamente

utilizados (Chatzou et al., 2016).

Han surgido trabajos que ejemplifican el uso de una sola herramienta para lograr un análisis filogenético, como por ejemplo Dadasnake, de Weißbecker et al. (2020), el cual es un script en Python que hace uso del Divisive Amplicon Denoising Algorithm (DADA2) (Callahan et al., 2016b) para el preprocesamiento de secuencias y Clustal Omega como método de alineamiento, aunque no es su fin realizar un análisis filogenético. Su uso está orientado a la ejecución en infreaestructuras de cómputo de alto rendimiento, las cuales cuentan con abundantes recursos de hardware, por lo que acceder a ellas en un principio podría implicar un problema al hacer estudios preliminares de este tipo.

De igual manera, Toparslan et al. (2020) realizan un flujo de trabajo para secuencias de ADN mitocondriales escrito en su totalidad en R, pero no hay tareas de preprocesamiento debido a la naturaleza de las secuencias. Se hace uso del método de alineamiento ClustalW con parámetros por defecto.

Rossi-Tamisier et al. (2015) alinean secuencias 16s rRNA para diferenciar bacterias asociadas al género humano y utlizan Clustal Omega con los parámetros por defecto.

Para la utlización de los parámetros por defecto en MUSCLE, se tomó como referencia el trabajo de Sato y Miyazaki (2017), en el cual hacen un estudio filogenético para el *Genus Enterobacter*.

El propósito de esta investigación es proveer un flujo de trabajo unificado en R que incluya implementaciones ya existentes para permitir tareas de análisis filogenético en una única plataforma, así como seleccionar métodos que puedan ser ejecutados en un equipo de cómputo personal. Para ello se hace una comparación entre implementaciones de los siguientes MSA: ClustalW, Clustal Omega, MUSCLE y DECIPHER.

2.3. Marco tecnológico y legal

2.3.1. R

R es un lenguaje y entorno para computación estadística y gráficas. Provee amplia variedad de técnicas estadísticas (modelado linear y no linear, pruebas clásicas

estadísticas, análisis de series de tiempo, clasificación, agrupamiento, entre otras.) y gráficas, y es altamente extensible (R Core Team, 2021). Una fortaleza de R es la facilidad con la que se producen gráficas bien diseñadas y de calidad, incluyendo la utilización de simbología matemática y fórmulas donde sea necesario. Presta mucho cuidado sobre los parámetros por defecto menores en las gráficas, pero el usuario mantiene el control total. Se compila y ejecuta en variedad de plataformas UNIX y sistemas similares (incluyendo FreeBSD y Linux), Windows y MacOS. R es software libre; utiliza paquetes y software asociado, que operan bajo las siguientes licencias: "GNU Affero General Public License" versión 3 (AGPL-3), "Artistic License" versión 2.0, "BSD 2-clause License", "BSD 3-clause License", "GNU General Public License" versión 2, "GNU General Public License" versión 3, "GNU Library General Public License" versión 2, "GNU Lesser General Public License" versión 2.1, "GNU Lesser General Public License" versión 3, "MIT License". Como paquete, R está bajo la licencia GPL v3, que se refiere a la versión 3 de la Licencia Pública General de GNU (agp). La intención de esta licencia es garantizar la libertad de compartir y modificar todas las versiones del programa, para asegurar que permanezca libre a los usuarios.

2.3.2. RStudio

Entorno de desarrollo que incluye también paquetes como shiny, ggvis y dplyr entre otros (RStudio Team, 2021). Este Entorno de Desarrollo Integrado (IDE) está disponible con licencia de código abierto y comercial. La versión de código abierto está bajo la licencia AGPL v3, que se refiere a la versión 3 de la Licencia Pública General Affero de GNU (agp). Incluye todos los beneficios de la licencia GPL-3, además de la implementación de dichos beneficios en los servidores, esto es, que el código fuente modificado esté disponible a los usuarios del servidor. Así, el uso público de una versión modificada, en un servidor accesible públicamente, ofrece acceso público al código fuente de la versión modificada.

Capítulo 3

Preprocesamiento y clasificación taxonómica de secuencias

En este capítulo se describen las tareas realizadas para preprocesar y clasificar taxonómicamente las secuencias, desde su obtención en formato comprimido *FASTQ.gz* del ENA, con el número de acceso PRJNA544732 hasta la anotación de las secuencias con sus bacterias correspondientes.

3.1. Análisis bioinformático

Las secuencias fueron importadas en R, versión 4.0.5, a partir de los archivos FASTQ. El análisis del microbioma se realizó con la paqeutería *Bioconductor* siguiendo el flujo de trabajo de Callahan et al. (2016b). Este análisis consistió en dos tareas principales, cada una con sus respectivos procesos.

1. Preprocesamiento.
2. Asignación taxonómica.

3.2. Preprocesamiento

Para la primera tarea, *DADA2* fue aplicado para el filtrado de lecturas *forward* y *reverse* con una calidad mínima de 20 basada en la puntuación Phred, es decir,

un recorte aplicado cuando la media en la puntuación de calidad sea inferior a Q20. La longitud mínima para las lecturas *forward* fue de 270 bases pareadas y para las lecturas *reverse*, 200 bases pareadas. El número de bases pareadas utilizado va en función de la exploración visual de las muestras y en este estudio, los valores atípicos se identificaron debajo de la calidad Q20.

Se aprecia en las Figuras 3.1 y 3.2 el mapa de calor en escala de grises la frecuencia de cada puntuación de calidad en cada posición de las bases. La media de la puntuación de calidad se indica en la línea color verde y los cuartiles por las líneas color naranja. Se aprecia el ancho de 300 nucleótidos en las muestras, por lo que se recortaron los últimos 30 de las lecturas *forward*, así como los últimos 100 para las lecturas *reverse*. A manera de ejemplo se utilizó la visualización de las primeras dos muestras, SRR9122766 y SRR9122767.

Se realizó un recorte para remover los primers o iniciadores con secuencias ambiguas en los primeros 17 nucleótidos de las lecturas *forward* y 21 para las *reverse* utilizando la herramienta *Cutadapt* (Martin, 2011).

El número de lecturas distribuidas en las 155 muestras tanto *forward* como *reverse* es de 21,326,390. Como resultado del proceso de limpieza, quedaron un total de 10,327,055, lo cual representa un 48.42 % del total de lecturas, luego de un tiempo de cómputo de 734.721 segundos. Con este total de lecturas se continuó con la estimación de las tasas de error.

El algoritmo *DADA2* hace uso de un modelo de error paramétrico de aprendizaje automático, el cual aprende el modelo de error al alternar la estimación de las tasas de error e inferencia de la composición de la muestra hasta que convergen en una solución consistente.

En las Figuras 3.3 y 3.4 se muestran las tasas de error para cada transición posible: A hacia C, A hacia G, A hacia T, etc. Los puntos indican la tasa de error observada para cada puntuación Phred de calidad. La línea en color negro muestra las tasas de error estimadas después de la convergencia del algoritmo de aprendizaje automático. La línea en color rojo representa las tasas de error esperadas bajo la definición de que, a menor calidad se presenta un incremento en la tasa de error, así como a mayor calidad, una disminución. El tiempo de cómputo para cada estimación fue de 71.699 segundos para las lecturas *forward* y 89.829 segundos para las *reverse*.

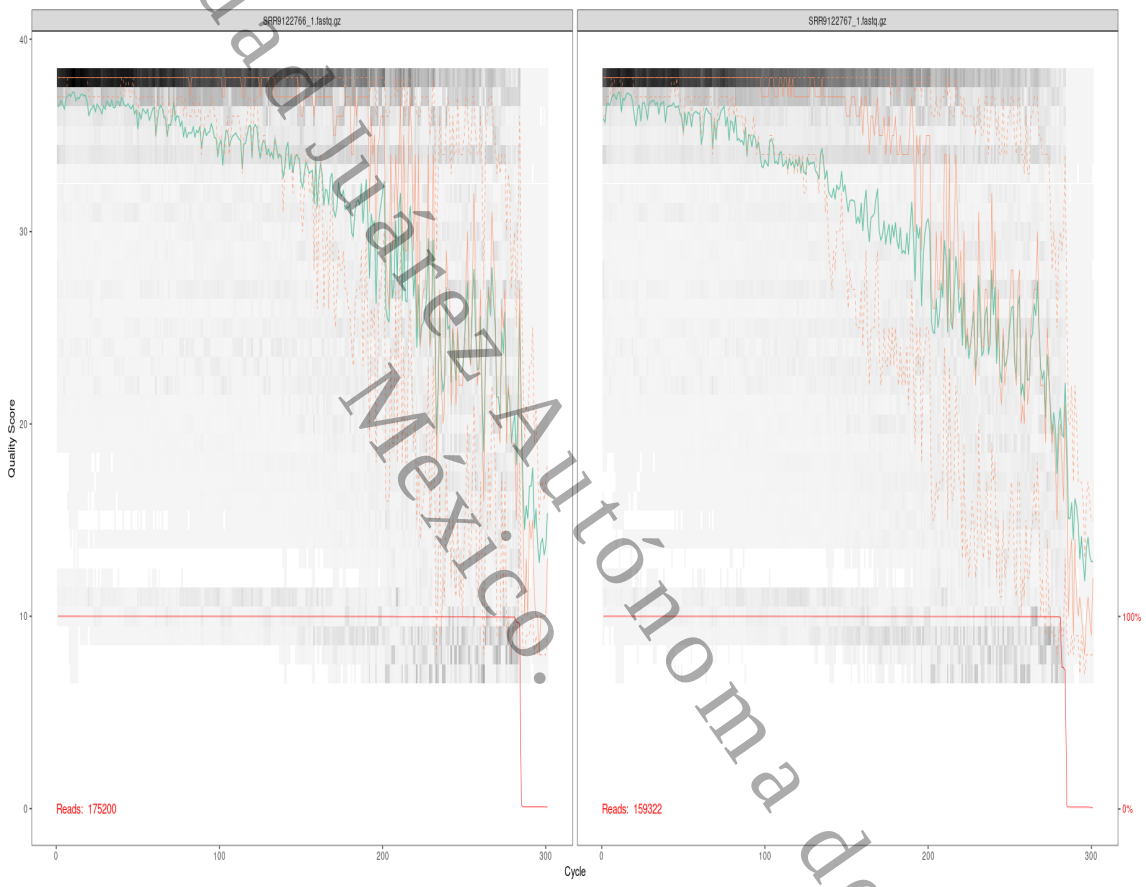


Figura 3.1: Lecturas *forward* de la puntuación de calidad Phred a través del ancho de las bases pareadas.

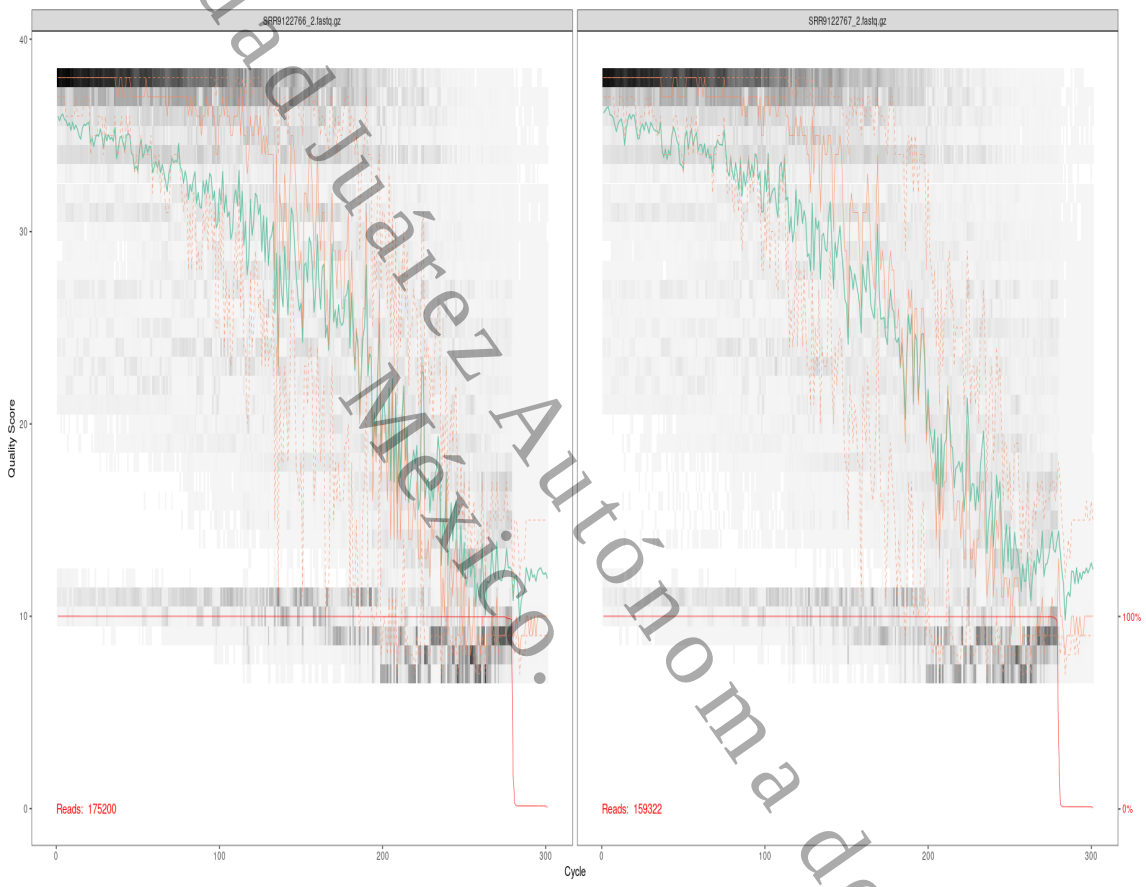


Figura 3.2: Lecturas *reverse* de la puntuación de calidad Phred a través del ancho de las bases pareadas.

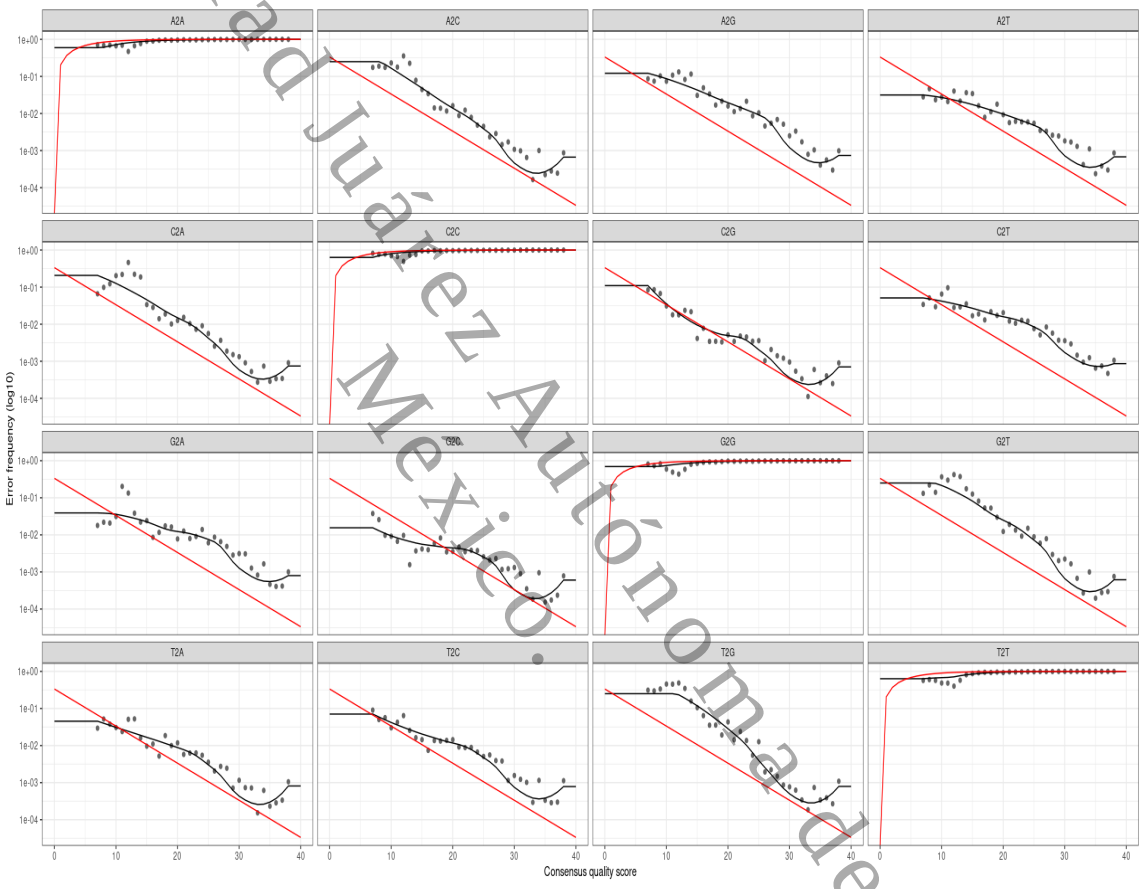


Figura 3.3: Modelo de error para lecturas *forward* obtenido por *DADA2*.

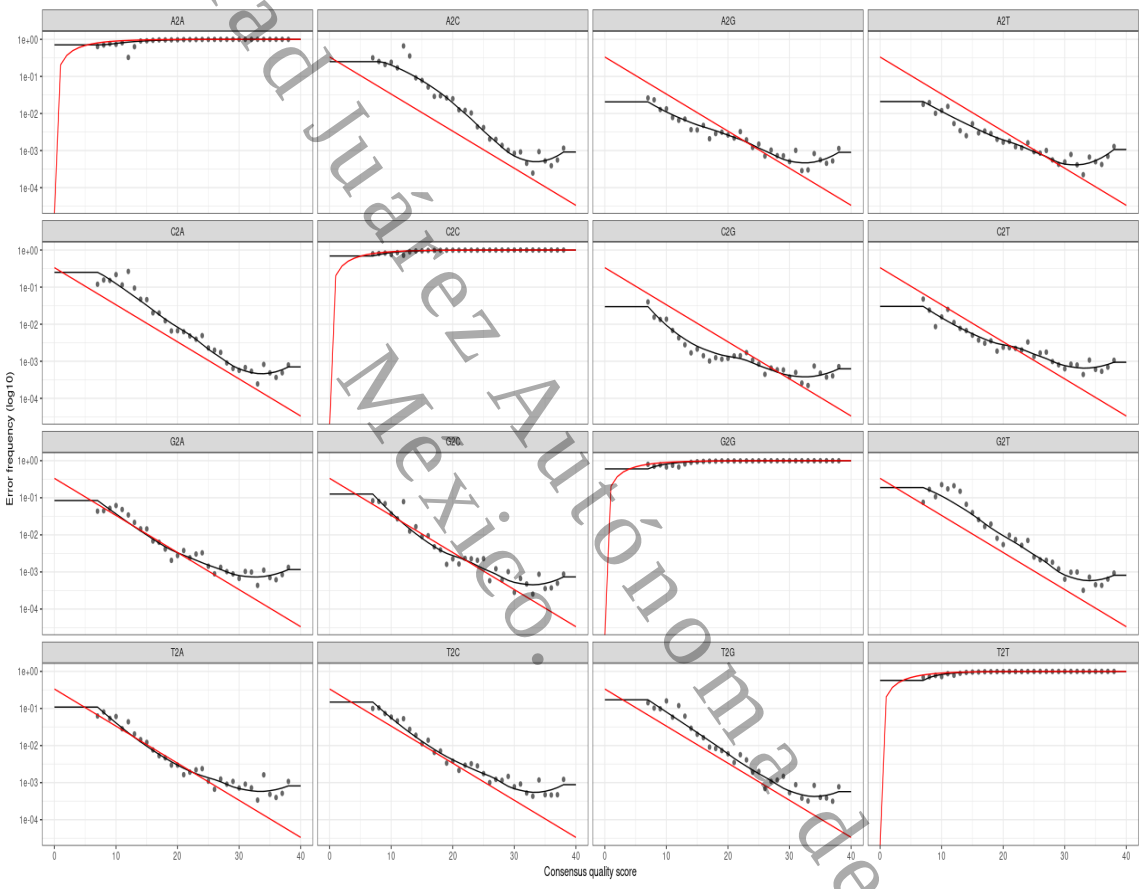


Figura 3.4: Modelo de error para lecturas *reverse* obtenido por *DADA2*

Debido a que se observó una inferencia en las tasas de error acorde a lo esperado en ambas lecturas basándose en la definición anterior, se prosiguió a la tarea de desreplicado de las lecturas, la cual consiste en iterar por bloques a través de las lecturas para tareas de filtrado y recorte de estas. Para las lecturas *forward* el tiempo de cómputo fue de 195.211 segundos y para las *reverse*, 159.189 segundos.

La siguiente etapa fue aplicar el algoritmo *DADA2* a las lecturas ya filtradas, posteriormente inferir variantes de secuencias de amplicón (ASV), para luego fusionar las lecturas *forward* con las *reverse* y así obtener secuencias únicas. Con esta fusión se logró construir una tabla de variantes de secuencia, la cual tiene como filas variantes de secuencia correspondientes con 11,816 secuencias y como columnas el ancho de las secuencias, lo cual se visualiza en la Figura 3.5a.

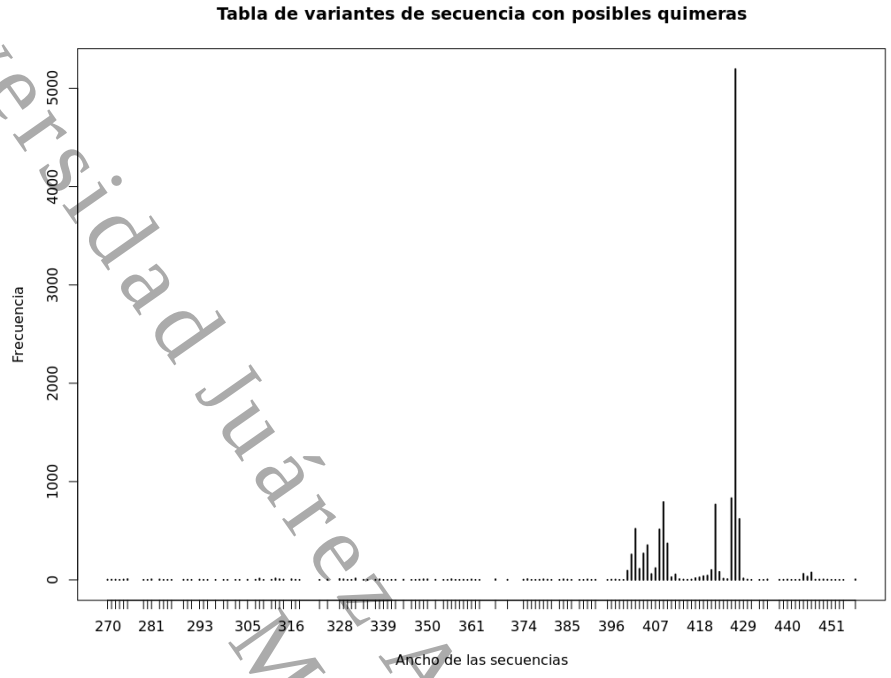
Aún con esta información, una última remoción de secuencias quimeras fue necesaria. Este proceso llevó un tiempo de procesamiento en paralelo de 63.6 segundos, dando como resultado 2,297 secuencias contenidas en las 155 muestras. Cada secuencia tiene un ancho específico, el cual se visualiza en la Figura 3.5b.

El porcentaje de lecturas removidas por este proceso fue de un 4.04 %, aun cuando el número de secuencias decreció de 11,816 a 2,297. El último proceso de esta primera tarea consistió en verificar el cambio en el número de lecturas desde su importación hasta la remoción de quimeras, lo cuál arrojó un promedio de 62.29 % de lecturas conservadas.

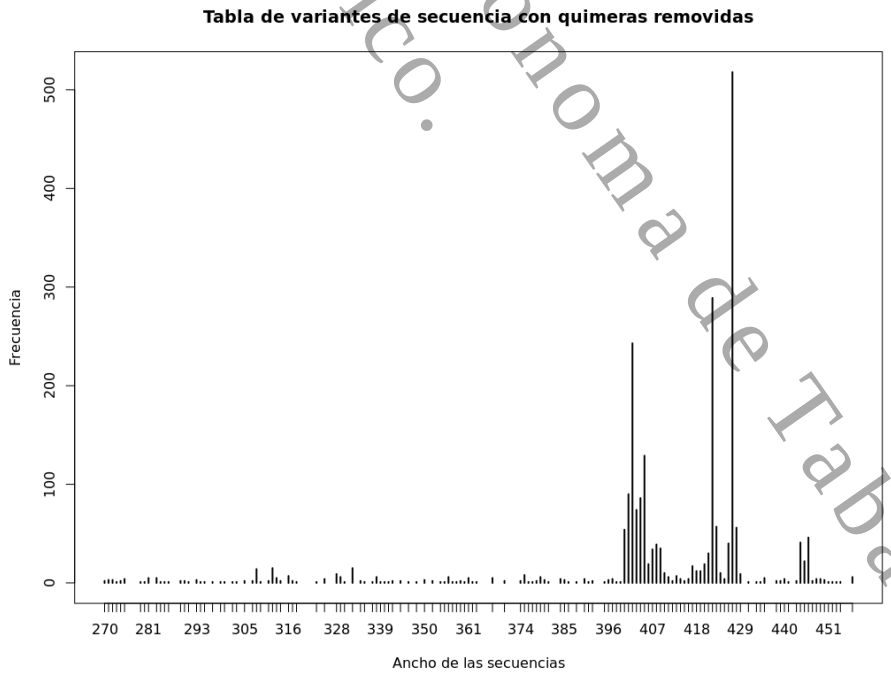
3.3. Asignación taxonómica

La asignación taxonómica microbiana se refiere a la comparación de secuencias con el fin de identificar sus rangos taxonómicos, la cual puede realizarse a través de diferentes enfoques. Para esto, se realizó una clasificación de las secuencias resultantes utilizando la base de datos taxonómica *Silva* versión 132 como referencia (Quast et al., 2012). *Silva* es un proyecto para crear y mantener bases de datos de ARN ribosomal de alta calidad, conteniendo secuencias del 16S rRNA.

La clasificación se realiza a través del método de clasificación naive-bayesiano integrado en *DADA2*, utlizando *Silva* 132 como conjunto de entrenamiento y las 2,297 secuencias como conjunto de prueba. De acuerdo con (Hočevar et al., 2019) se



(a) 11,816 variantes de secuencia con posibles quimeras.



(b) 2,297 variantes de secuencias no quimeras.

Figura 3.5: Variación en la frecuencia de ASV.

filtraron aquellas anotaciones que tuviesen un valor mayor a 0.6 en la estimación de confianza *bootstrap* para seguir utilizándolas.

Adicionalmente, utlizando la misma base de datos, se realizó la asignación de especies ya que se busca minimizar el número de secuencias que contengan anotaciones con valores faltantes en algún rango taxonómico, en este caso *Species*. El resultado de la asignación taxonómica se ejemplifica en la Tabla 3.1

De las 2,297 secuencias, 1,974 fueron clasificadas dentro del reino *Bacteria*, 290 dentro de *Eukaryota*, 1 dentro de *Archaea* y 36 sin clasificar debido a que pertenecen al reino *Fungi*, el cual no está presente en la base taxonómica utilizada. Debido al enfoque de la investigación referente a Vaginosis Bacteriana, se seleccionaron exclusivamente las 1,974 secuencias clasificadas en el reino *Bacteria*, pues son los organismos relevantes para el estudio.

Tabla 3.1: Clasificación taxonómica de las primeras 4 ASV de 2,297 totales.

TAGGGAATCTTCCACAATGGACGCAAGTCTGATGGAGCAACGCCGCGTGAGTGAAGA
AGGGTTTCGGCTCGTAAAGCTCTGTTGTTGGTGAAGAAGGACAGGGGTAGTAACTGA
CCTTTGTTTGACGGTAATCAATTAGAAAGTCACGGCTAACTACGTGCCAGCAGCCGC
GGTAATACGTAGGTGGCAAGCGTTGTCCGGATTTATTGGGCGTAAAGCGAGTGCAGG
CGGCTCGATAAGTCTGATGTGAAAGCCTTCGGCTCAACCGGAGAATTGCATCAGAAA
CTGTCGAGCTTGAGTACAGAAGAGGAGAGTGGAACTCCATGTGTAGCGGTGAAATGC
GTAGATATATGGAAGAACACCGGTGGCGAAGGCGGCTCTCTGGTCTGTTACTGACGC
TGAGGCTCGAAAGCATGGGTAGCGAACA

Kingdom	Phylum	Class	Order	Family	Genus	Species	ASV
Bacteria	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Lactobacillus*	*iners*	ASV1

TAGGGAATCTTCCACAATGGACGCAAGTCTGATGGAGCAACGCCGCGTGAGTGAAGA
AGGTTTTCGGATCGTAAAGCTCTGTTGTTGGTGAAGAAGGATAGAGGTAGTAACTGG
CCTTTATTTGACGGTAATCAACCAGAAAGTCACGGCTAACTACGTGCCAGCAGCCGC
GGTAATACGTAGGTGGCAAGCGTTGTCCGGATTTATTGGGCGTAAAGCGAGCGCAGG
CGGAAGAATAAGTCTGATGTGAAAGCCCTCGGCTTAACCGAGGAACTGCATCGGAAA
CTGTTTTTCTTGAGTGCAGAAGAGGAGAGTGGAACTCCATGTGTAGCGGTGGAATGC
GTAGATATATGGAAGAACACCAGTGGCGAAGGCGGCTCTCTGGTCTGCAACTGACGC
TGAGGCTCGAAAGCATGGGTAGCGAACA

Kingdom	Phylum	Class	Order	Family	Genus	Species	ASV
Bacteria	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Lactobacillus*	NA	ASV2

TAGGGAATCTTCCACAATGGACGCAAGTCTGATGGAGCAACGCCGCGTGAGTGAAGA
AGGGTTTCGGCTCGTAAAGCTCTGTTGGTAGTGAAGAAAGATAGAGGTAGTAACTGG
CCTTTATTTGACGGTAATTACTTAGAAAGTCACGGCTAACTACGTGCCAGCAGCCGC
GGTAATACGTAGGTGGCAAGCGTTGTCCGGATTTATTGGGCGTAAAGCGAGTGCAGG
CGGTTCAATAAGTCTGATGTGAAAGCCTTCGGCTCAACCGGAGAATTGCATCAGAAA
CTGTTGAACTTGAGTGCAGAAGAGGAGAGTGGAACTCCATGTGTAGCGGTGGAATGC
GTAGATATATGGAAGAACACCAGTGGCGAAGGCGGCTCTCTGGTCTGCAACTGACGC
TGAGGCTCGAAAGCATGGGTAGCGAACA

Kingdom	Phylum	Class	Order	Family	Genus	Species	ASV
Bacteria	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Lactobacillus*	NA	ASV3

TGGGGAATATTGCGCAATGGGGGAAACCCTGACGCAGCGACGCCGCGTGCGGGATG
AAGGCCTTCGGGTTGTAAACCGCTTTTGATTGGGAGCAAGCCTTTTGGGTGAGTGTA
CCTTTCGAATAAGCGCCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGCG
CAAGCGTTATCCGGAATTATTGGGCGTAAAGAGCTTGTAGGCGGTTCGTCGCGTCTG
GTGTGAAAGCCCATCGCTTAACGGTGGGTTTGCGCCGGGTACGGGCGGGCTAGAGTG
CAGTAGGGGAGACTGGAATTCTCGGTGTAACGGTGGAATGTGTAGATATCGGGAAG
AACACCAATGGCGAAGGCAGGTCTCTGGGCTGTTACTGACGCTGAGAAGCGAAAGCG
TGGGGAGCGAACA

Kingdom	Phylum	Class	Order	Family	Genus	Species	ASV
Bacteria	*Actinobacteria*	*Actinobacteria*	*Bifidobacteriales*	*Bifidobacteriaceae*	*Gardnerella*	*vaginalis*	ASV4

Capítulo 4

Alineamiento de múltiples secuencias para conjuntos de datos de microbioma

4.1. Descripción del hardware empleado para las pruebas

A continuación, se describe de manera general el equipo de cómputo utilizado para la totalidad de las pruebas de la presente investigación:

- Sistema operativo: Ubuntu 20.04.2 focal LTS.
- Núcleo: x86_64 Linux 5.4.72-microsoft-standard-WSL2.
- CPU: Intel Core i7-8750H @ 12x 2.208GHz.
- RAM: 25,562 MB.
- R versión 4.0.5 (2021-03-31).
- Matrix products: default.
- BLAS: /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.9.0.

- LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.9.0.
- Paquetes base: *stats4*, *parallel*, *stats*, *graphics*, *grDevices*, *utils*, *datasets*, *methods*, *base*.
- Otros paquetes: *ape_5.3*, *seqinr_3.6-1*, *tictoc_1.0*, *DECIPHER_2.18.1*, *RSQLite_2.2.0*, *ShortRead_1.48.0*, *GenomicAlignments_1.22.1*, *SummarizedExperiment_1.16.1*, *DelayedArray_0.12.3*, *matrixStats_0.56.0*, *Biobase_2.46.0*, *Rsamtools_2.2.3*, *GenomicRanges_1.38.0*, *GenomeInfoDb_1.22.1*, *BiocParallel_1.20.1*, *msa_1.22.0*, *Biostrings_2.54.0*, *XVector_0.26.0*, *IRanges_2.20.2*, *S4Vectors_0.24.4*, *BiocGenerics_0.32.0*, *dada2_1.18.0*, *phangorn_2.7.0*.
- *Cutadapt* versión 2.10.
- *Mumsa* version 1.0.

4.2. Experimentación y resultados

4.2.1. Alineamiento con los cuatro métodos MSA

Como principal parámetro de entrada se utilizaron las 1,974 secuencias obtenidas del preprocesamiento para los 4 MSA, cada método ejecutándose con sus parámetros por defecto, los cuáles se exponen en la Tabla 4.1.

Los métodos MSA ClustalW, Clustal Omega y MUSCLE están implementados en el lenguaje de programación C y C++ y fueron ejecutados mediante la función `msa()` del paquete *msa* de R. Esta función sirve de interfaz para que sea posible la ejecución de los métodos dentro de R y el alineamiento sea entendible por R. Por otro lado, *DECIPHER* es un paquete que dentro de sus funciones, mediante `AlignSeqs()` se alinean las secuencias. En la Tabla 4.1 se exponen los parámetros utilizados para cada método MSA

Alineamiento con ClustalW

El alineamiento de las 1,974 secuencias se realizó mediante la función `msa()`, la cual recibió como parámetros el total de las secuencias, el método MSA ClustalW

Tabla 4.1: Parámetros de entrada de los métodos MSA. Los primeros tres se realizan mediante el paquete *msa* y el último mediante *DECIPHER*. Se muestran sus respectivas funciones.

ClustalW			
`msa(seqs3, method=''ClustalW'', type=''dna'', order=''input'', substitutionMatrix = ''clustalw'')`			
Apertura de hueco	Extensión de hueco	Máximo de iteraciones	Orden
15	6.66	3	Entrada
Clustal Omega			
`msa(seqs3, method=''Clustal Omega'', type=''dna'', order=''input'', verbose=TRUE)`			
Apertura de hueco	Extensión de hueco	Máximo de iteraciones	Orden
6	1	Sin límite	Entrada
MUSCLE			
`msa(seqs3, method=''Muscle'', cluster = ''upgma'', type=''dna'', order=''input'', verbose=TRUE)`			
Apertura de hueco	Extensión de hueco	Máximo de iteraciones	Orden
400	0	2	Entrada
DECIPHER			
`AlignSeqs((DNAStringSet(seqs3)), processors = NULL)`			
Apertura de hueco	Extensión de hueco	Máximo de iteraciones	Orden
16	1	2	Entrada

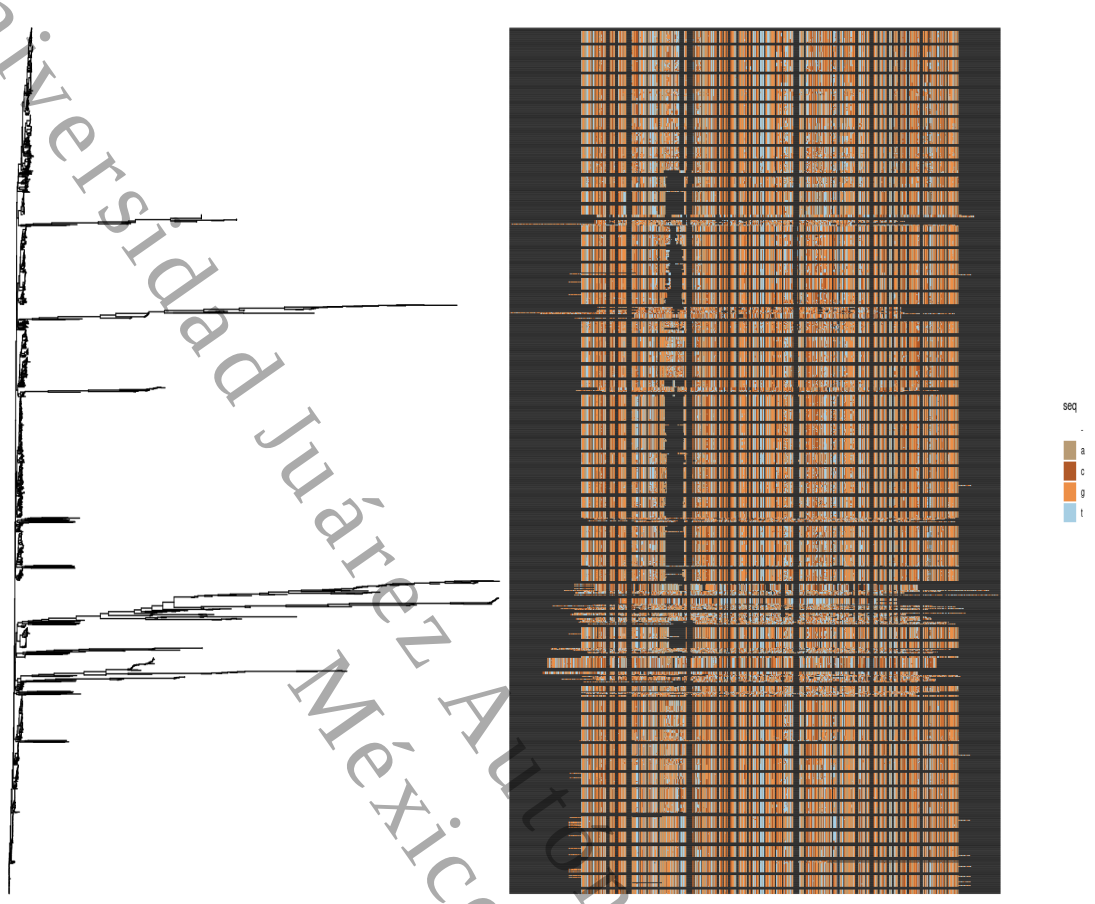


Figura 4.1: ClustalW con 714 columnas.

como método a utilizar, el tipo de información correspondiente a secuencias de ADN y el orden de la salida de las secuencias alineadas de acuerdo con el orden de entrada. El ancho del alineamiento resultante fue de 714 columnas, representadas en la Figura 4.1.

Alineamiento con Clustal Omega

Se utilizó también la función `msa()` y el total de las secuencias, el método MSA Clustal Omega, el tipo de información de secuencias de ADN y el orden de la salida del alineamiento de acuerdo con el orden de entrada. El ancho del alineamiento resultante fue de 908 columnas, representadas en la Figura 4.2.

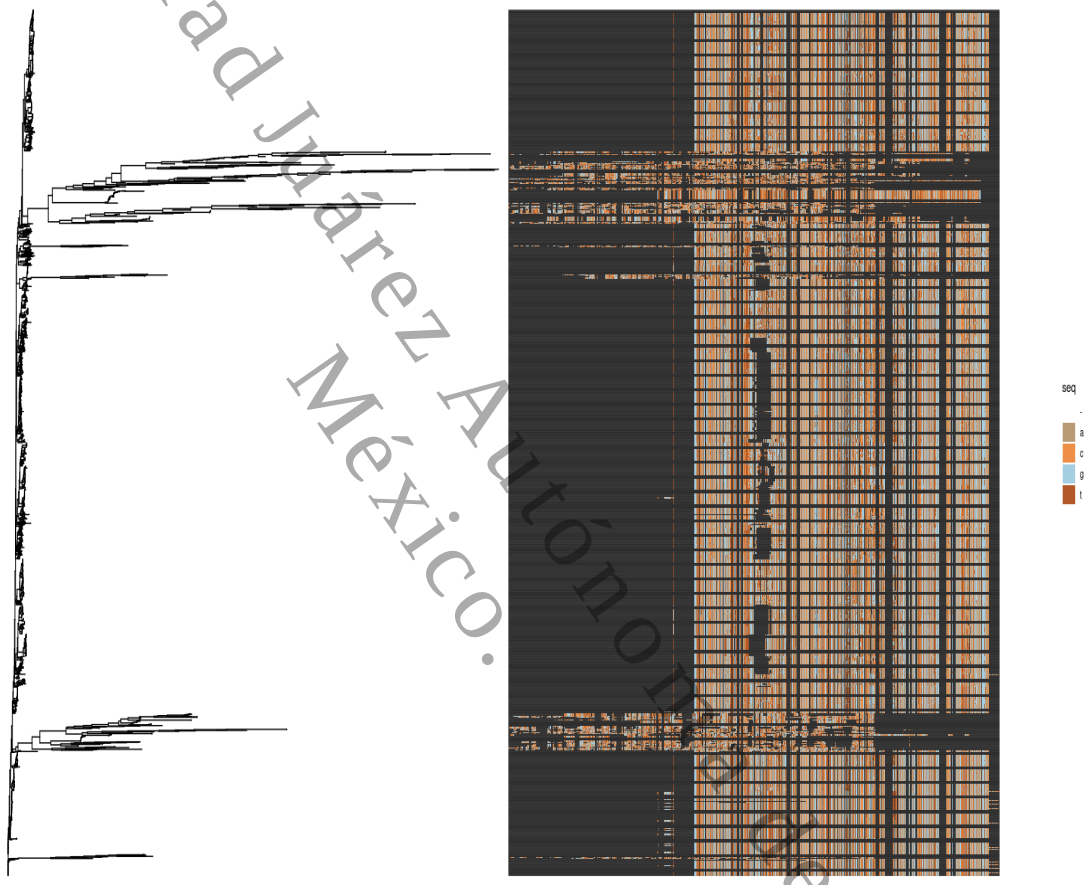


Figura 4.2: Clustal Omega con 908 columnas.

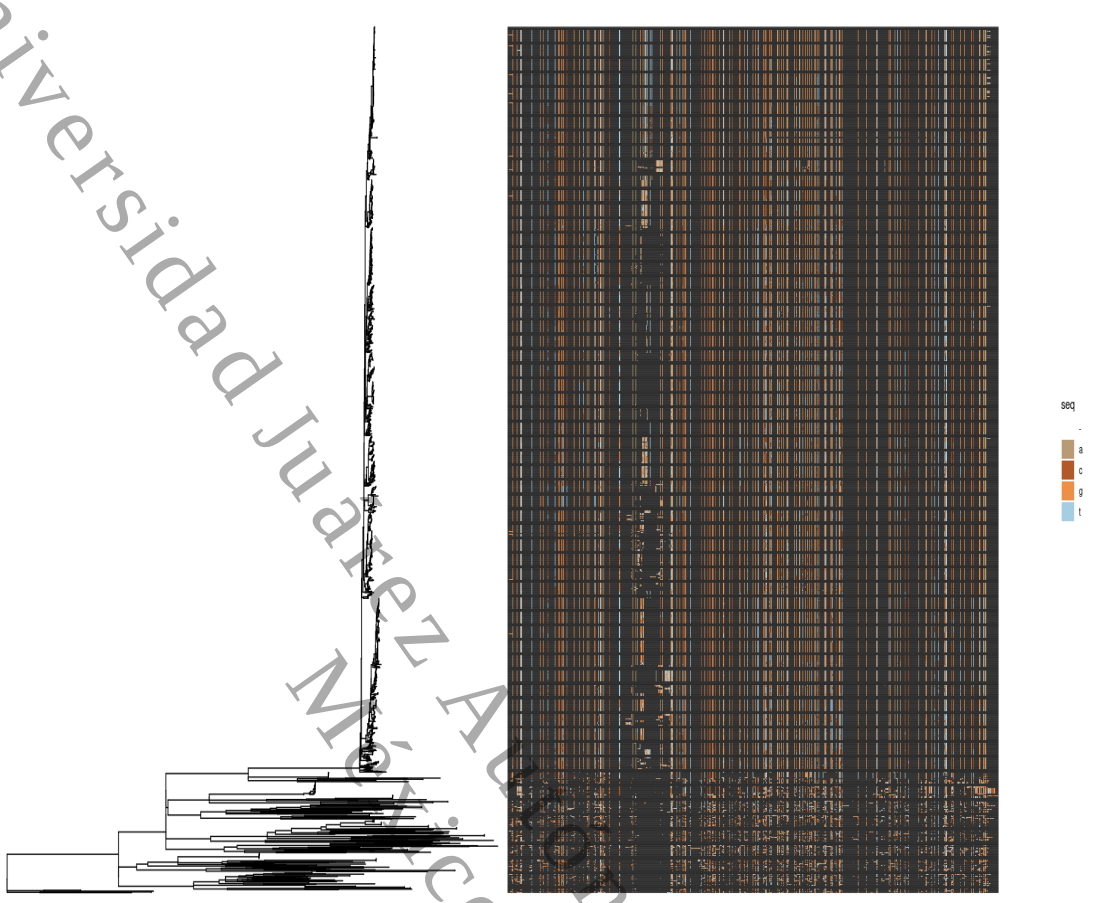


Figura 4.3: MUSCLE con 2,379 columnas.

Alineamiento con MUSCLE

Igualmente se utilizó la función `msa()` y el total de las secuencias, el método MSA MUSCLE, el tipo de información de secuencias de ADN y el orden de la salida del alineamiento de acuerdo con el orden de entrada. El ancho del alineamiento resultante fue de 2,379 columnas, representadas en la Figura 4.3.

Alineamiento con DECIPHER

Por último, con DECIPHER se empleó la función `AlignSeqs()` con el total de las secuencias convertidas mediante la función `DNAStringSet()`, además de utilizar el número total de procesadores disponibles. El ancho del alineamiento resultante

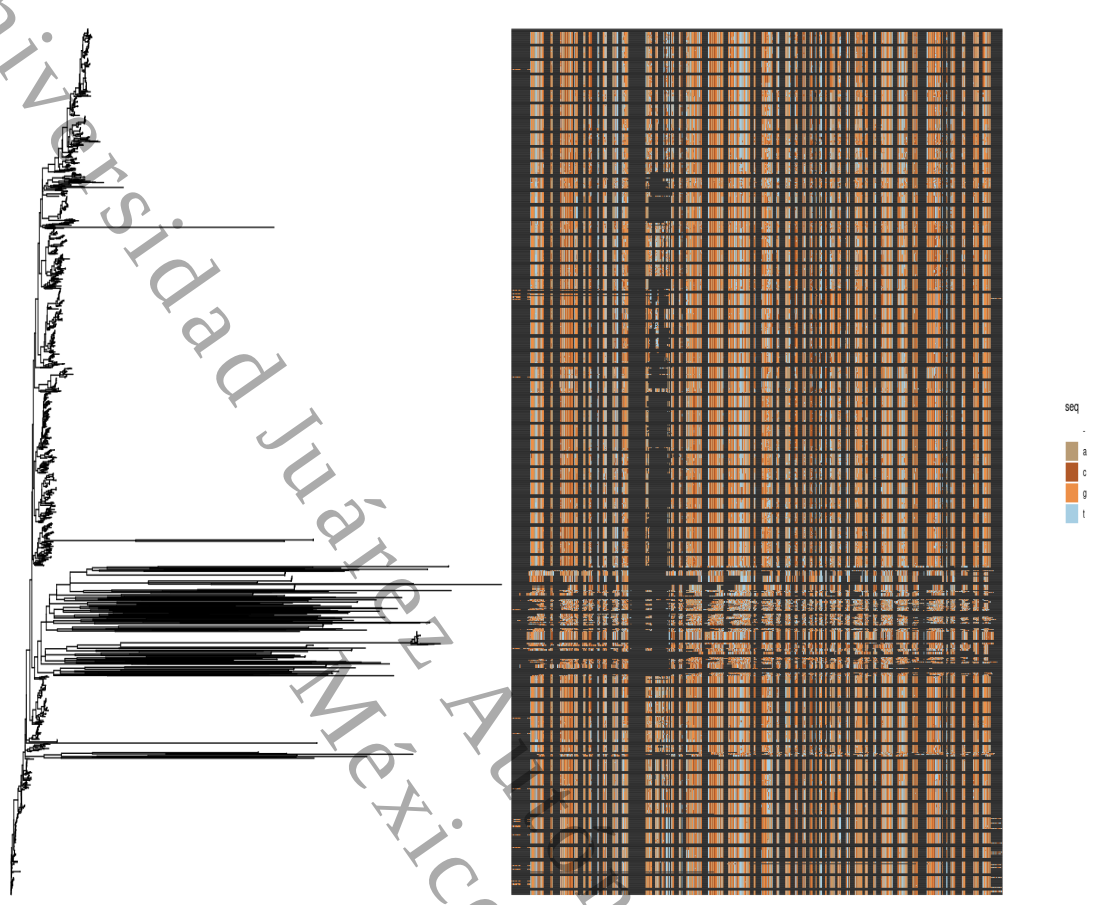


Figura 4.4: DECIPHER con 796 columnas.

fue de 796 columnas, representadas en la Figura 4.2.

Árbol filogenético de los cuatro métodos MSA

Gracias a la obtención de los alineamientos resultantes de los 4 MSA distintos, se calcularon 4 árboles de distancia distintos. El proceso fue el siguiente:

1. El paquete *phangorn* en R se utilizó para construir dos árboles filogenéticos iniciales basados en matrices de distancia. El primer método empleado fue el algoritmo de unión de vecinos (NJ) y el segundo el algoritmo de agrupamiento de pares sin peso mediante media aritmética (UPGMA).

2. Se evalúa qué árbol resultante es mejor mediante la función `parsimony`; la

cual resulta en una puntuación que indica el número de cambios mínimos para describir la información en un árbol dado. Las puntuaciones obtenidas entre los dos métodos para los 4 MSA fueron las siguientes: ClustalW NJ = 43,704; UPGMA = 43,164. Clustal Omega NJ = 44,571; UPGMA = 42,861. MUSCLE NJ = 48,537; UPGMA = 50,107. DECIPHER NJ = 41,449; UPGMA = 42,989.

3. A continuación se calcula la máxima verosimilitud del árbol con menor puntuación *parsimony* mediante la función `pml`, teniendo como diferencia principal con los métodos de distancia que se incluye información completa del alineamiento.

4. Una vez obtenida la máxima verosimilitud de cada árbol, se optimiza la topología del árbol y longitud de rama, ajustando el árbol a un modelo generalizado reversible en el tiempo con distribución Gamma (GTR+G+I).

5. Por último, se vuelven a evaluar los árboles resultantes de cada modelo mediante *parsimony*. Las puntuaciones obtenidas por cada MSA fueron: ClustalW = 42,707; Clustal Omega = 42,094; MUSCLE = 51,363; DECIPHER = 41,056.

El árbol del método DECIPHER resulta ser el de menor puntuación entre los distintos MSA. Los árboles resultantes se visualizan junto con su alineamiento correspondiente en las Figuras 4.1, 4.2, 4.3 y 4.4.

Con toda esta información obtenida, se procedió al uso de *MUMSA* para la validación estadística de los resultados obtenidos por los métodos MSA, con el fin de definir el método MSA que calculara el alineamiento más significativo biológicamente.

4.3. Validación estadística de los resultados

Para todas las tareas realizadas se utilizó una sola semilla de valor 100, incluso para los métodos MSA. Una vez obtenido el resultado de los cuatro métodos el primer uso de *MUMSA* fue predecir la dificultad del alineamiento partiendo de la coincidencia múltiple entre alineamientos y el resultado se observa en la Figura 4.5. Esta primer medida no considera los residuos alineados a los huecos ni los pares de residuos alineados. Al tener como resultado un $AOS = 0.580479$, se está frente a un caso de alineamiento de dificultad cercana a la media, recordando el rango

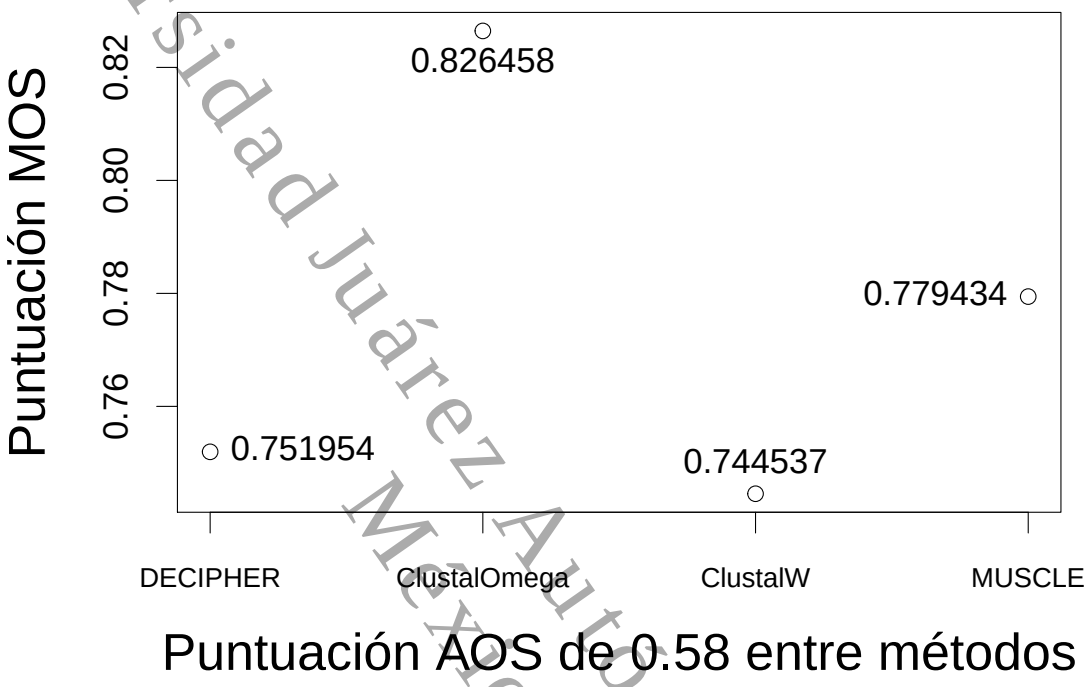


Figura 4.5: Puntuaciones preliminares *MOS* y *AOS* para los cuatro métodos MSA.

de dificultad $[0, 1]$, lo que supondría que el método Clustal Omega pareciese ser el indicado para cuestiones de este flujo de trabajo con un $MOS = 0.826458$, ya que supera al resto de los métodos.

Sin embargo, al realizar la evaluación que sí toma las consideraciones de los residuos se obtiene como resultado un $AOS = 0.5$, indicando una dificultad mayor a la anteriormente obtenida, pero es el método DECIPHER con un valor $MOS = 0.7$ el que refleja una mayor confiabilidad estadística en los pares de residuos alineados a diferencia de los otros tres métodos (ver Figura 4.6).

Partiendo de la premisa de que el alineamiento con el mayor número de estos pares supone el biológicamente más significativo, se calculó la coincidencia *OS* entre métodos métodos respecto a DECIPHER, como se muestra en la Figura 4.7,

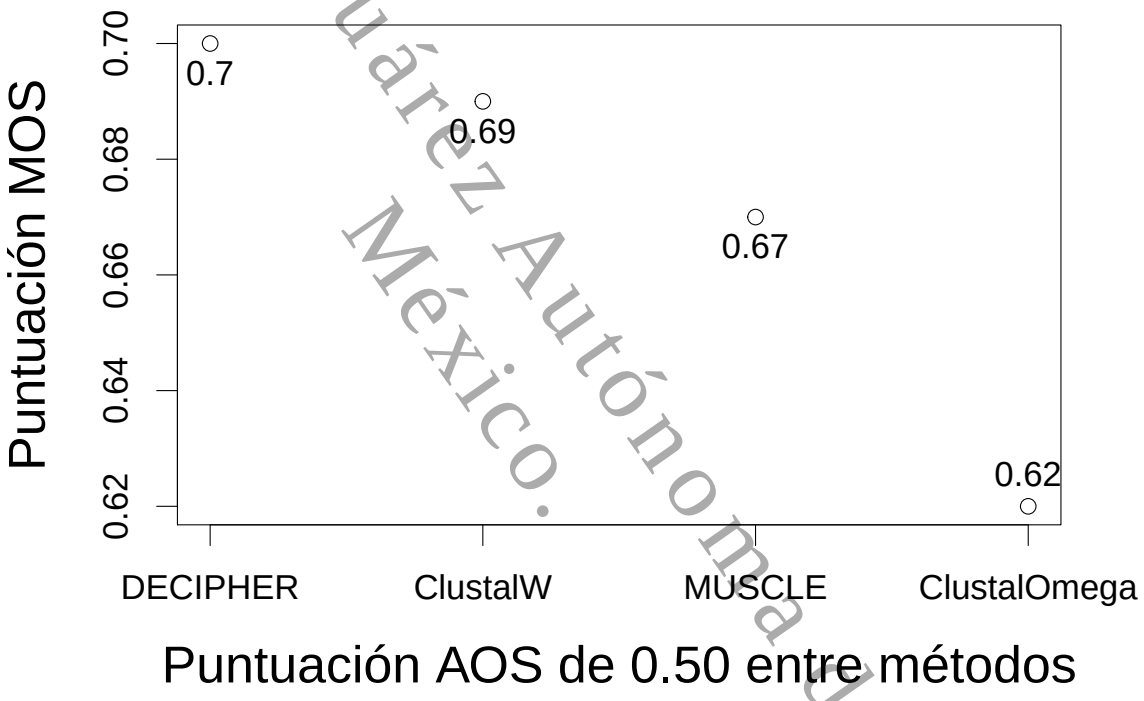


Figura 4.6: Puntuaciones finales *MOS* y *AOS* para los cuatro métodos MSA.

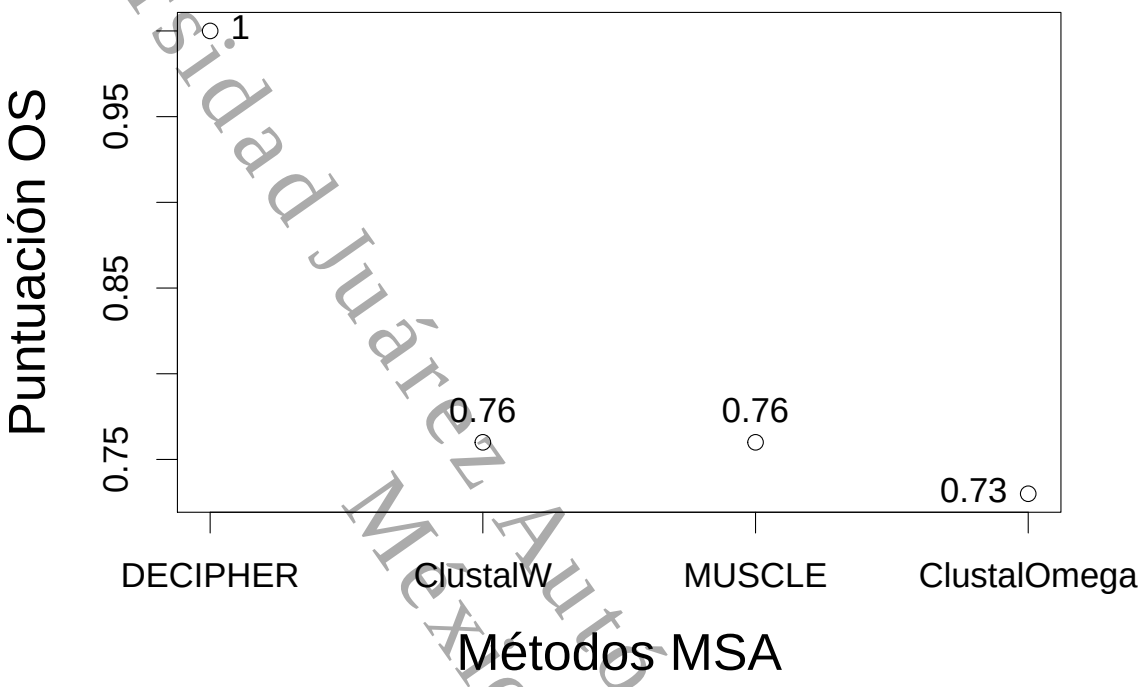


Figura 4.7: Puntuación de coincidencia *OS* respecto al método DECIPHER.

en donde se aprecia que hay una diferencia importante en la intersección entre los demás métodos MSA con DECIPHER, que podría estudiarse a partir del número de columnas: ClustalW = 714, Clustal Omega = 908, MUSCLE = 2,379, DECIPHER = 796, pero eso está fuera del enfoque de la presente investigación.

En la Tabla 4.2 se detalla el tiempo de ejecución de los principales procesos en el flujo de trabajo, así como la disponibilidad de multiprocesamiento de dicho proceso. Si solo se considera el tiempo aquí expuesto, el total sería de aproximadamente 36,813 s, poco más de 10 horas. Habría que considerar que si se utiliza solo DECIPHER, método MSA biológicamente más significativo para este análisis, el flujo de trabajo tomaría un tiempo de ejecución aproximado de 11,800 s, poco más de tres horas.

Tabla 4.2: Tiempo de ejecución de funciones principales.

Preprocesamiento		
Proceso	Tiempo estimado por *tictoc* en segundos (s)	Multihilo (verdadero o falso)
Filtro y recorte para *Cutadapt*	504.795 s	Verdadero
Cutadapt	1,098.482 s	Verdadero
Filtro y recorte al resultado de *Cutadapt*	446.305 s	Verdadero
Modelo de errores (*forward*)	38.051 s	Verdadero
Modelo de errores (*reverse*)	195.136 s	Verdadero
Unión de lecturas e inferencia muestral con *DADA2*	2,637.742 s	Verdadero
Clasificación taxonómica		
Proceso	Tiempo estimado por *tictoc* en segundos (s)	Multihilo (verdadero o falso)
Asignación taxonómica	183.517 s	Verdadero
Asignación de especies	651.717 s	Falso
Métodos MSA		
Proceso	Tiempo estimado por *tictoc* en segundos (s)	Multihilo (verdadero o falso)
ClustalW	2,908.898 s	Falso
Clustal Omega	100.779 s	Falso
MUSCLE	30.558 s	Falso
DECIPHER	61.368 s	Verdadero
Ajuste de árbol GTR+G+I ClustalW	8,703.4 s	Verdadero
Ajuste de árbol GTR+G+I Clustal Omega	6,300.23 s	Verdadero
Ajuste de árbol GTR+G+I MUSCLE	6,220.14 s	Verdadero
Ajuste de árbol GTR+G+I DECIPHER	6,012.964 s	Verdadero
Validación estadística		
Proceso	Tiempo estimado por *tictoc* en segundos (s)	Multihilo (verdadero o falso)
MUMSA $AOS = 0.58$	43.188 s	Falso
MUMSA $AOS = 0.50$	115.923 s	Falso
MUMSA coincidencia *OS*	44.139 s	Falso

4.4. Identificación de bacterias que indican VB en las muestras

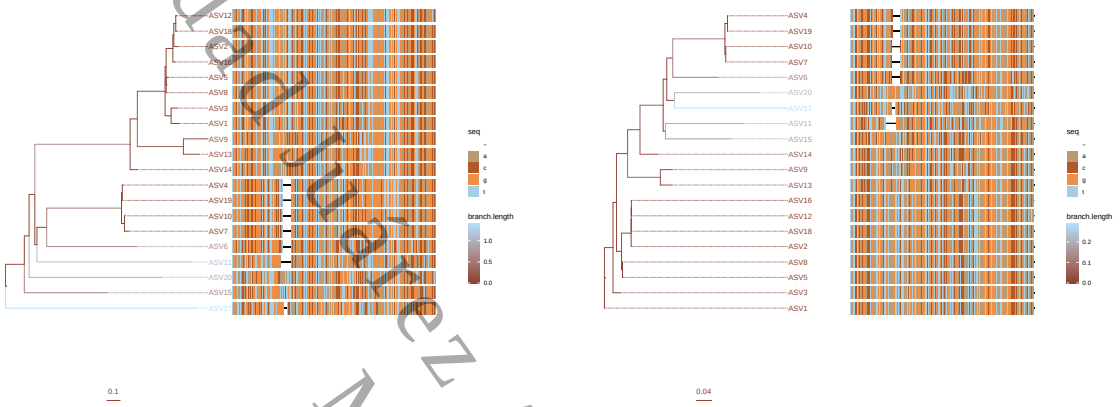
Debido al número de secuencias con las que se trabajó, y para una mejor visualización de los resultados, se creó un objeto por cada MSA obtenido. Este objeto contiene la tabla OTU que indica la abundancia de cada ASV, la tabla taxonómica de cada ASV, el árbol filogenético y las secuencias. Específicamente 1,974 ASV contenidas en las 155 muestras, cada ASV con su secuencia correspondiente.

Con el fin de observar la composición bacteriana de las muestras, se ordenaron todas las 1,974 ASV de mayor a menor y se seleccionaron las 20 más abundantes. De este segundo objeto, se puede visualizar en la Figura 4.8 su árbol filogenético derivado del obtenido de cada MSA, así como el alineamiento correspondiente a cada ASV. Las ASV 7, 10, 4 y 19 corresponden la especie *Gardnerella*. En cambio, la 16, 12, 2, 18, 8, 5, 3 y 1 pertenecen a *Lactobacillus*. En color azul se aprecia la ASV con la mayor distancia a las demás y pertenece a *Prevotella*. *Sneathia*, 11 y *Ureaplasma*, 20, destacan igualmente por sus distancias, es decir, menor relación con las demás.

De este segundo objeto también se identificaron las bacterias a través de sus niveles taxonómicos, y se comprobó que el resultado es independiente del alineamiento, ya que las ASV's no cambian, como sí lo hacen las distancias de los árboles de cada método. En la Figura 4.9 se muestran las bacterias más abundantes para 10 muestras obtenidas de un submuestreo del total de 155 con semilla de valor 100. Los rangos comprendidos son *Genus* y *Species* en donde la mayoría de estas muestras están conformadas en gran parte por *Lactobacillus*, aunque también se aprecian muestras que podrían indicar VB debido a la alta concentración de *Gardnerella*, *Atopobium* o *Sneathia* además de una disminución de *Lactobacillus*.

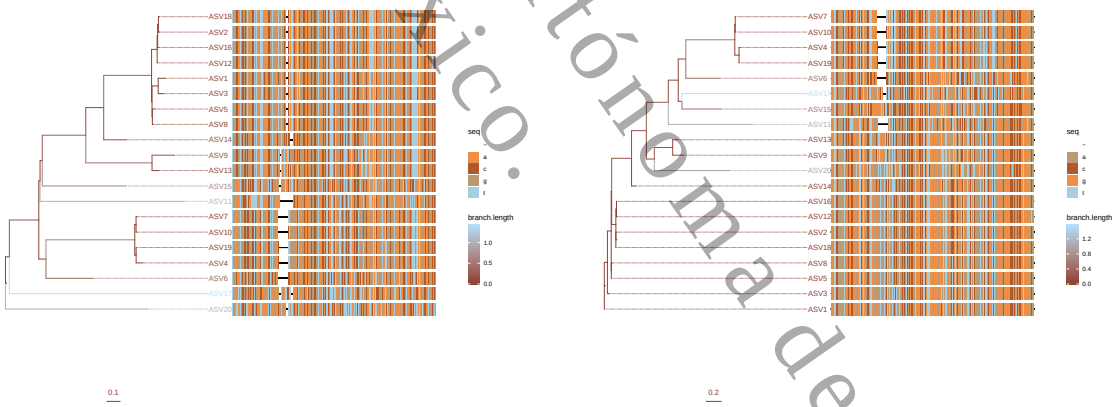
Con los resultados de esta exploración, se procedió a identificar tipos de estado de comunidades presentes en las muestras. De acuerdo con Ma y Li (2017), proponen un criterio cuantitativo para definir estas comunidades y siguiendo la metodología propuesta por DiGiulio et al. (2015), se obtuvieron las siguientes cuatro comunidades dominadas por alguna variante de *Lactobacillus* representadas en la Figura 4.10.

También se identificó que las muestras que conforman el tipo 4, como se muestra la Figura 4.11 presenta una disminución de *Lactobacillus* y un aumento de *Gardnerella*,



(a) ClustalW

(b) Clustal Omega



(c) MUSCLE

(d) DECIPHER

Figura 4.8: Árbol filogenético y alineamiento de las 20 ASV más abundantes de los 4 MSA.

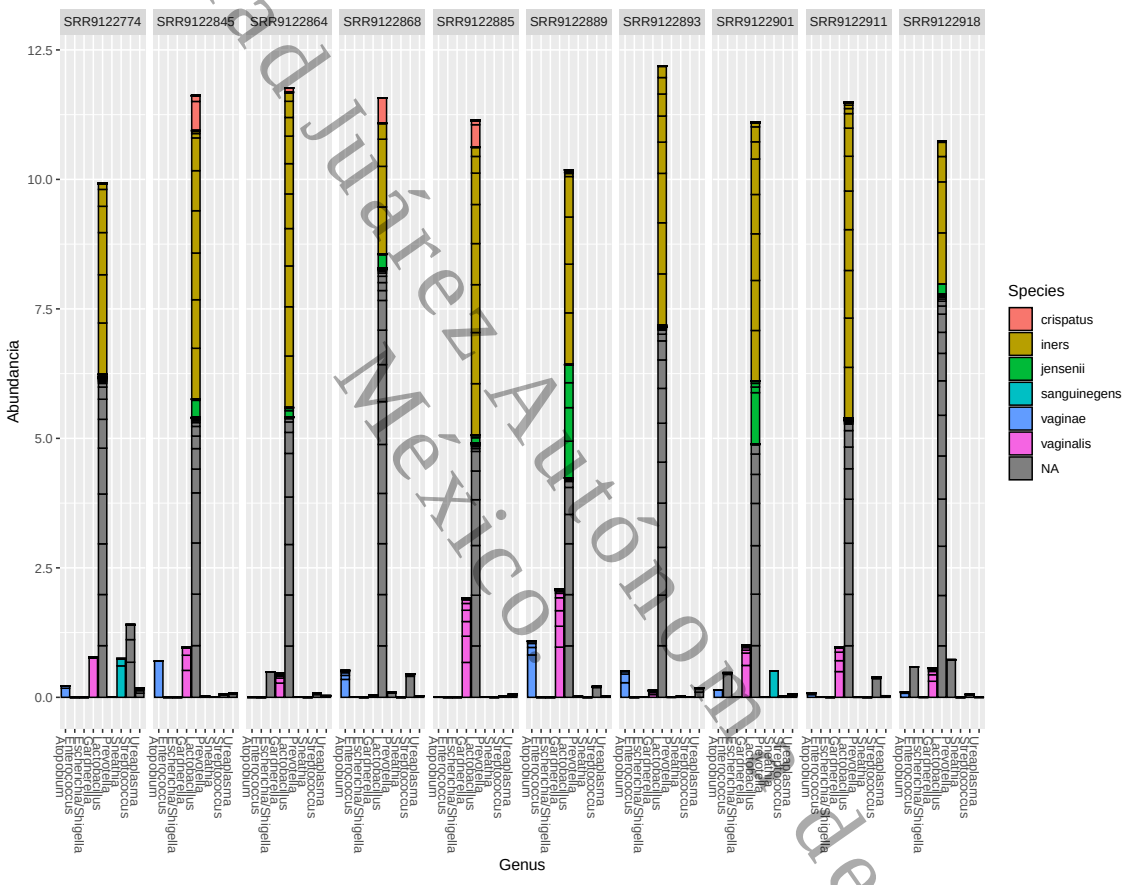
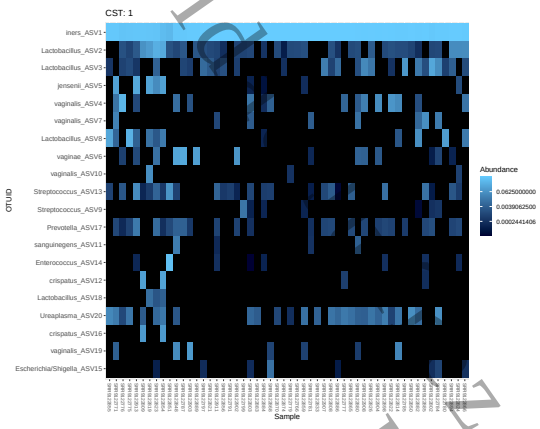
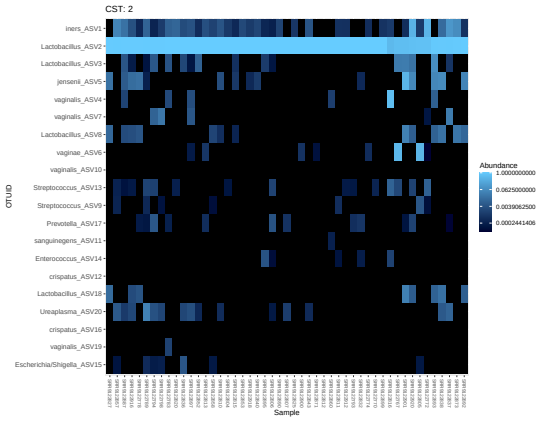


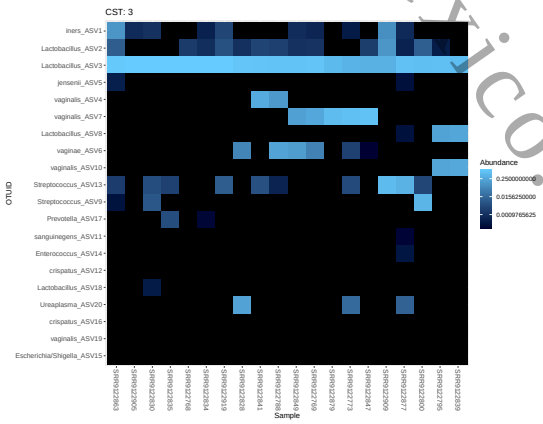
Figura 4.9: Abundancia bacteriana de 10 muestras en rangos *Genus* y *Species*.



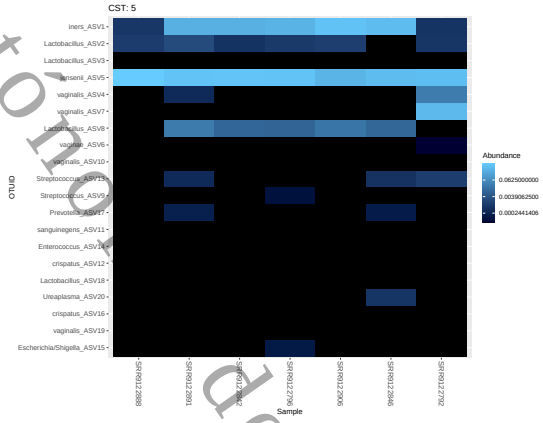
(a) Comunidad dominada por *L. iners*



(b) Comunidad dominada por *L. crispatus*



(c) Comunidad dominada por *L. gasseri*



(d) Comunidad dominada por *L. jensenii y L. iners*

Figura 4.10: Tipos de estado de comunidades con mayor abundancia de alguna variante de *Lactobacillus*.

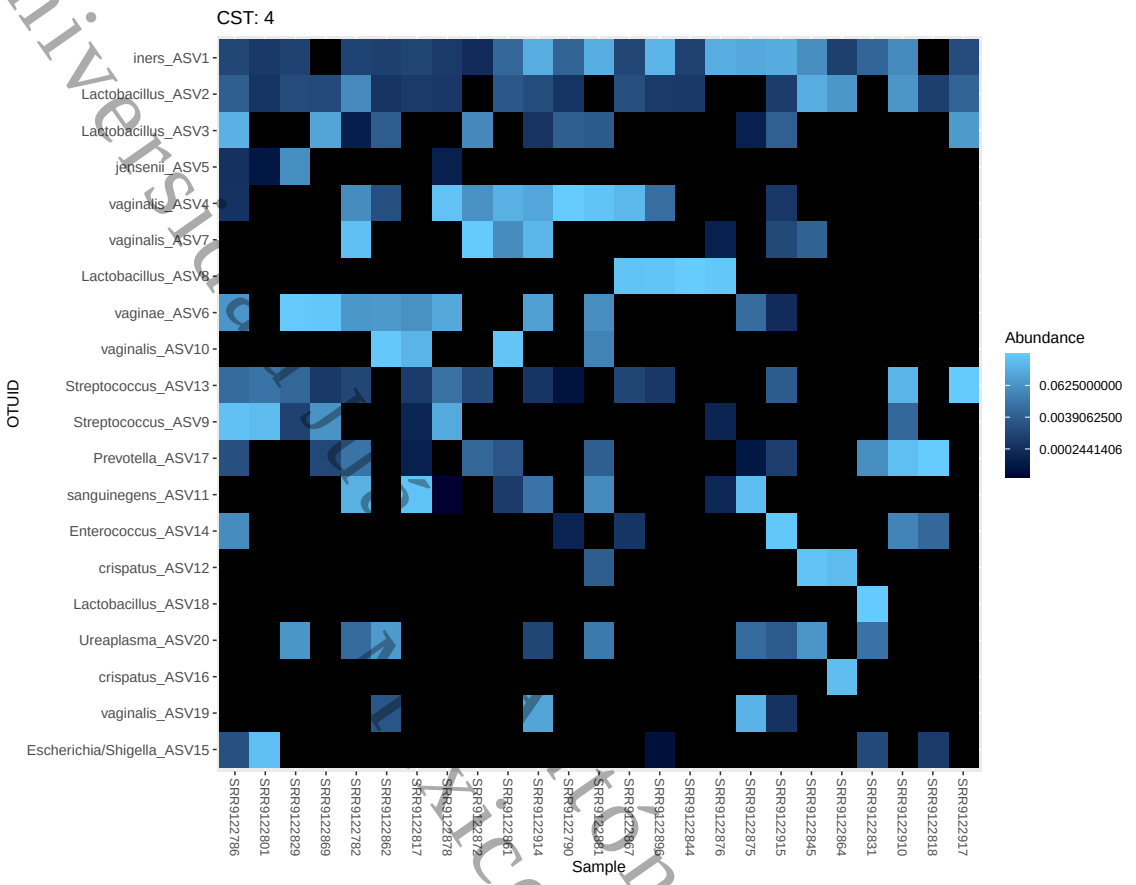


Figura 4.11: Abundancia bacteriana de las muestras en la comunidad de tipo 4.

Atopobium o *Sneathia*, lo cual indica que las muestras sufren una posible alteración en la microbiota vaginal y se da pie a que se puedan realizar pruebas para hallar BV.

Capítulo 5

Conclusiones y trabajos futuros

Conclusiones

El fin de la presente investigación fue contrastar métodos MSA para así crear un flujo de trabajo intuitivo que permita un análisis filogenético en R e incluir validaciones estadísticas para tareas del análisis bioinformático. Para ello se utilizaron herramientas ya desarrolladas para el sistema operativo Ubuntu, *Cutadapt* y *MUMSA*, y se llamaron mediante R para así complementar las tareas de preprocesamiento y validar estadísticamente los resultados de los alineamientos.

El total de lecturas pareadas contenidas en las 155 muestras analizadas fue de 22,154,990. Después de un primer filtrado con *Cutadapt* se obtuvieron 21,326,390, representando un 96.25 % del total. Estas fueron preprocesadas hasta obtener 9,767,545 lecturas, representando un 44.08 % del total. El número total de ASV's resultantes del preprocesamiento fue de 1,974 bacterias.

Estas 1,974 bacterias presentes en las 155 muestras fueron agrupadas en cinco tipos de comunidades, cuatro de ellas conformadas por una mayoria significativa de alguna bacteria *Lactobacillus* como *L. iners*, *L. crispatus*, *L. gasseri* o *L. jensenii*, pero está presente una comunidad en donde además de una disminución de *Lactobacillus* abundan también especies como *Gardnerella*, *Atopobium* o *Sneathia* indicando una posible presencia de Vaginosis Bacteriana en dichas muestras.

Aun cuando los cuatro métodos MSA utilizados se basan en la programación dinámica, ClustalW, ClustalOmega y DECIPHER además hacen uso de alineamientos progresivos. MUSCLE se basa en una aproximación iterativa. DECIPHER des-

tacó del resto por la velocidad en la obtención del alineamiento, debido al uso de heurísticas al buscar *k-mers* de manera ordenada en la creación del árbol guía, con lo cual se cubre el objetivo de contrastar estas aproximaciones distintas y entre estas dos, la aproximación iterativa obtuvo resultados en el menor tiempo de ejecución.

De cada método MSA se obtuvieron los árboles filogenéticos, analizados mediante la función parsimonia para conocer el grado de utilidad de la información en las secuencias. Se compararon las versiones NJ y UPGMA y se realizó un ajuste del mejor árbol a un modelo generalizado reversible en el tiempo con distribución Gamma (GTR+G+I). El árbol NJ del método DECIPHER obtuvo la mejor puntuación parsimonia, resultado congruente con los obtenidos por las funciones AOS y MOS.

Para este estudio se evaluaron estadísticamente los alineamientos obtenidos entre sí, una primera vez sin considerar residuos y una segunda vez tomándolos en cuenta. ClustalOmega resultó el de mejor puntuación MOS para la primera, pero en la segunda, DECIPHER resultó el de mayor valor MOS. Debido a que se definió utilizar la puntuación AOS como parámetro, la segunda prueba tiene mayor significancia estadística, obteniendo un $AOS = 0.5$ contra el $AOS = 0.580479$ de la primera, implicando que existe una mayor dificultad en el alineamiento.

La segunda medida fue la puntuación MOS, siendo DECIPHER el de mayor puntuación con un MOS, lo cual indica una mayor fiabilidad biológica en el resultado del alineamiento. Así se tomó como referencia el resultado de DECIPHER y comparó contra el resto para obtener la tercer y última medida, OS, cuyos valores representan la coincidencia entre métodos respecto a DECIPHER y se visualiza en la Figura 4.7.

Trabajos futuros

Se propone la realización de una interfaz gráfica en R que permita el uso de este flujo de trabajo y sirva como herramienta práctica para el análisis filogenético de secuencias del gen 16S ARN ribosomal, así como la exploración visual de la microbiota en las muestras analizadas. También se propone un enfoque de investigación que involucre la clasificación de los pacientes sin necesidad de poseer datos clínicos para contrastar resultados contra investigaciones que los incluyen. Por último, se motiva al desarrollo de una medida de distancia personalizada para alineamiento con el fin de mejorar los tiempos pero también incrementar la utilidad del resultado

Apéndice A

Glosario

A continuación se hace detallan las ocurrencias de distintas siglas y su significado contenidas en el documento.

A.1. Siglas

A Adenina. 35

AOS puntuación coincidencia promedio. 18

ASV variantes de secuencias de amplicón. 40

C Citosina. 35

DADA Algoritmo de partición divisiva. 28

ENA European Nucleotide Archive. 17

G Guanina. 35

MOS puntuación coincidencia múltiple. 18

MSA alineamiento de múltiples secuencias. 14

NGS tecnología de secuenciación masiva. 15

NJ Neighbour joining. 24

OS puntuación coincidencia. 18

OTU unidad taxonómica operacional. 25

PCR reacción en cadena de polimerasa. 29

R lenguaje de programación R. 14

rRNA ácido ribunocleico ribosomal. 15

T Timina. 35

UFC unidad formadora de colonias. 16

UPGMA Unweighted Pair Group Method with Arithmetic mean. 25

VB vaginosis bacteriana. 15

Bibliografía

Gnu affero general public license. URL https://www.gnu.org/licenses/agpl-3.0.html.

Gnu general public license. URL http://www.gnu.org/licenses/gpl.html.

María Guadalupe Aguilera-Arreola y Silvia Giono-Cerezo. Perspectiva ecológica de la vaginosis bacteriana e infecciones de transmisión sexual en méxico. URL https://www.asieslamedicina.org.mx/perspectiva-ecologica-de-la-vaginosis-bacteriana-e-infecciones-de-transmision-sexual-en-mexico/?pdf=2733. Consultado: 02-20-2020.

Praseeda Ajitkumar, Herman W Barkema, y Jeroen De Buck. Rapid identification of bovine mastitis pathogens by high-resolution melt analysis of 16s rdna sequences. *Veterinary microbiology*, 155(2-4):332–340, 2012.

Abd Al-Bar Al-Farha, Kiro Petrovski, Razi Jozani, Andrew Hoare, y Farhid Hemmatzadeh. Discrimination between some mycoplasma spp. and acholeplasma laidlawii in bovine milk using high resolution melting curve analysis. *BMC research notes*, 11(1):1–4, 2018.

David Arthur y Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Inf. téc., Stanford, 2006.

Ulrich Bodenhofer, Enrico Bonatesta, Christoph Horejš-Kainrath, y Sepp Hochreiter. msa: an r package for multiple sequence alignment. *Bioinformatics*, 31(24):3997–3999, 2015.

Matthew Caesar, Tyson Condie, Jayanthkumar Kannan, Karthik Lakshminarayanan, y Ion Stoica. ROFL: Routing on flat labels. En *ACM SIGCOMM*. 2006.

Ben J Callahan, Kris Sankaran, Julia A Fukuyama, Paul J McMurdie, y Susan P Holmes. Bioconductor workflow for microbiome data analysis: from raw reads to community analyses. *F1000Research*, 5, 2016a.

Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, y Susan P Holmes. Dada2: high-resolution sample inference from illumina amplicon data. *Nature methods*, 13(7):581–583, 2016b.

J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K Goodrich, Jeffrey I Gordon, et al. Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335–336, 2010.

R Carraro, G Dalla Rovere, S Ferraresso, L Carraro, R Franch, A Toffan, F Pascoli, T Patarnello, y L Bargelloni. Development of a real-time pcr assay for rapid detection and quantification of photobacterium damselae subsp. piscicida in fish tissues. *Journal of fish diseases*, 41(2):247–254, 2018.

Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, y Barbara Borges. *shiny: Web Application Framework for R*, 2021. URL https://CRAN.R-project.org/package=shiny. R package version 1.6.0.

Maria Chatzou, Cedrik Magis, Jia-Ming Chang, Carsten Kemena, Giovanni Bussotti, Ionas Erb, y Cedric Notredame. Multiple sequence alignment modeling: methods and applications. *Briefings in bioinformatics*, 17(6):1009–1023, 2016.

Biswanath Chowdhury y Gautam Garai. A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics*, 109(5-6):419–431, 2017.

Patrick W. Daly. *Naural Sciences Citations and References*, 2007.

Jurate Daugelaite, Aisling O'Driscoll, y Roy D Sleator. An overview of multiple sequence alignments and cloud computing in bioinformatics. *International Scholarly Research Notices*, 2013, 2013.

Ruben Delgado Moya. *Elaboración de tesis. Protocolo de investigación*. Sista, México, 2005.

Daniel B DiGiulio, Benjamin J Callahan, Paul J McMurdie, Elizabeth K Costello, Deirdre J Lyell, Anna Robaczewska, Christine L Sun, Daniela SA Goltsman, Ronald J Wong, Gary Shaw, et al. Temporal and spatial variation of the human microbiota during pregnancy. *Proceedings of the National Academy of Sciences*, 112(35):11060–11065, 2015.

Rosalinda D'Amore, Umer Zeeshan Ijaz, Melanie Schirmer, John G Kenny, Richard Gregory, Alistair C Darby, Migun Shakya, Mircea Podar, Christopher Quince, y Neil Hall. A comprehensive benchmarking study of protocols and sequencing platforms for 16s rrna community profiling. *BMC genomics*, 17(1):1–20, 2016.

Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

Carlos Fernández-Collado, Pilar Baptista-Lucio, y Roberto Hernández-Sampieri. Metodología de la investigación. *Editorial McGraw Hill*, 2014.

Nuria Garzón-Pinto. Filogenia de los seres vivos: dominio archaea. 2017.

Keli Hočevar, Aleš Maver, Marijana Vidmar Šimic, Alenka Hodžić, Alexander Haslberger, Tanja Premru Seršen, y Borut Peterlin. Vaginal microbiome signature is associated with spontaneous preterm delivery. *Frontiers in medicine*, 6:201, 2019.

John H Holland. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.

I Illumina. Understanding illumina quality scores. *Technical Note: Informatics*, 23, 2014.

Mohamed Issa y Aboul Ella Hassanien. Multiple sequence alignment optimization using meta-heuristic techniques. En *Data Analytics in Medicine: Concepts, Methodologies, Tools, and Applications*, págs. 565–579. IGI Global, 2020.

Donald E. Knuth. *The TEXbook*. Addison-Wesley, 1984.

Leslie Lamport. *LATEX: A Document Preparation System*. Addison-Wesley, 1986.

Timo Lassmann. Kalign 3: multiple sequence alignment of large datasets. 2020.

Timo Lassmann y Erik LL Sonnhammer. Automatic assessment of alignment quality. *Nucleic acids research*, 33(22):7120–7128, 2005.

Odile Lecompte, Julie D Thompson, Frédéric Plewniak, Jean-Claude Thierry, y Olivier Poch. Multiple alignment of complete sequences (macs) in the post-genomic era. *Gene*, 270(1-2):17–30, 2001.

Zhanshan Sam Ma y Lianwei Li. Quantifying the human vaginal community state types (csts) with the species specificity index. *PeerJ*, 5:e3366, 2017.

Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):10–12, 2011.

Manuel Miller, Julia Zorn, y Markus Brielmeier. High-resolution melting curve analysis for identification of pasteurellaceae species in experimental animal facilities. *PLoS One*, 10(11):e0142560, 2015.

Sushmita Mitra y Tinku Acharya. *Data mining: multimedia, soft computing, and bioinformatics*. John Wiley & Sons, 2005.

Avelina Suárez Moya. Microbioma y secuenciación masiva. *Revista española de quimioterapia*, 30(5):305–311, 2017.

Isaac Newton y Naomi Campbell. A re-formulation of gravity with respect to really cool models. *Jornal of Funny Physics*, 35:39–78, 1997.

Cecilia Ortiz-Rodríguez, Mirta Ley-Ng, Carmen Llorente-Acebo, y Caridad Almanza-Martínez. Vaginosis bacteriana en mujeres con leucorrea. *Revista Cubana de Obstetricia y Ginecología*, 26(2):74–81, 2000.

Fabiano Sviatopolk-Mirsky Pais, Patrícia de Cássia Ruy, Guilherme Oliveira, y Roney Santos Coimbra. Assessing the efficiency of multiple sequence alignment programs. *Algorithms for molecular biology*, 9(1):1–8, 2014.

Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, y Frank Oliver Glöckner. The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic acids research*, 41(D1):D590–D596, 2012.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

Morgane Rossi-Tamisier, Samia Benamar, Didier Raoult, y Pierre-Edouard Fournier. Cautionary tale of using 16s rrna gene sequence similarity values in identification of human-associated bacterial species. *International journal of systematic and evolutionary microbiology*, 65(Pt_6):1929–1934, 2015.

RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC, Boston, MA, 2021. URL http://www.rstudio.com/.

David James Russell. *Multiple sequence alignment methods*. Springer, 2014.

Mitsuharu Sato y Kentaro Miyazaki. Phylogenetic network analysis revealed the occurrence of horizontal gene transfer of 16s rrna in the genus enterobacter. *Frontiers in microbiology*, 8:2225, 2017.

Klaus Schlaeppi y Davide Bulgarelli. The plant microbiome at work. *Molecular Plant-microbe interactions*, 28(3):212–217, 2015.

Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology*, 7(1):539, 2011.

Johannes Söding. Protein homology detection by hmm–hmm comparison. *Bioinformatics*, 21(7):951–960, 2005.

Julie D Thompson, Desmond G Higgins, y Toby J Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22(22):4673–4680, 1994.

Emine Toparslan, Kemal Karabag, y Ugur Bilge. A workflow with r: Phylogenetic analyses and visualizations using mitochondrial cytochrome b gene sequences. *PloS one*, 15(12):e0243927, 2020.

Hideo Umeki. *The geometry package*, 2002.

Fabiola Valenzuela-González, Ramón Casillas-Hernández, Enrique Villalpando, y Francisco Vargas-Albores. El gen arnr 16s en el estudio de comunidades microbianas marinas. *Ciencias marinas*, 41(4):297–313, 2015.

Christina Weißbecker, Beatrix Schnabel, y Anna Heintz-Buschart. Dadasnake, a snakemake implementation of dada2 to process amplicon sequencing data for microbial ecology. *GigaScience*, 9(12):giaa135, 2020.

W John Wilbur y David J Lipman. Rapid similarity searches of nucleic acid and protein data banks. *Proceedings of the National Academy of Sciences*, 80(3):726–730, 1983.

Andreas Wilm, Indra Mainz, y Gerhard Steger. An enhanced rna alignment benchmark for sequence alignment programs. *Algorithms for molecular biology*, 1(1):1–11, 2006.

Erik S Wright. Decipher: harnessing local sequence context to improve protein multiple sequence alignment. *BMC bioinformatics*, 16(1):1–14, 2015.

Erik S Wright. Using decipher v2. 0 to analyze big biological sequence data in r. *R Journal*, 8(1), 2016.

Vahid Zarezade y Ali Veysi. Predicting 3d structure of the neuropeptide s receptor: A candidate for asthma treatment. *Iranian Journal of Allergy, Asthma & Immunology*, 17, 2018.