

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias Biológicas y Ambientales

**Comparación de plataformas de clasificación taxonómica en
muestras provenientes de murciélagos**

Joselyn Katerine Tulcanaz Montesdoca

Ingeniería en Biotecnología

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniera en Biotecnología

Quito, 22 de diciembre de 2022

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias Biológicas y Ambientales

HOJA DE CALIFICACIÓN DE TRABAJO DE FIN DE CARRERA

**Comparación de plataformas de clasificación taxonómica en muestras
provenientes de murciélagos**

Joselyn Katerine Tulcanaz Montesdoca

Nombre del profesor, Título académico

Gabriel Trueba, Ph.D.

Quito, 22 de diciembre de 2022

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en <http://bit.ly/COPETHeses>.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on <http://bit.ly/COPETHeses>.

RESUMEN

Las técnicas de secuenciación de alto rendimiento pueden producir secuencias de múltiples organismos que se encuentran en una muestra (metagenómica). Estos datos deben ser analizados para la comprensión de la dinámica biológica que se encuentra en un nicho determinado, por ejemplo, la determinación de especies microbianas con capacidad de transmisión zoonótica y que están presentes en murciélagos. Entre ellas, la metodología de Oxford Nanopore Technologies para la lectura de secuencias largas. Su interpretación se hace mediante plataformas de bioinformática que clasifican los organismos por la asignación de un taxón conocido. Sin embargo, estas plataformas tienen una sensibilidad variada por la forma de procesamiento, el tiempo de obtención de resultados y su accesibilidad. Así, las plataformas Kaiju, Kraken y BugSeq (rápidas, sensibles y de manejo sencillo). Cada una de ellas con diferentes métodos informáticos de asignación. Se comparó las tres plataformas con el uso de muestras provenientes de murciélagos (saliva, sangre, heces e hisopado orofaríngeo). Estas muestras fueron obtenidas de la amazonía ecuatoriana sector El Edén y Yasuní ITT, mismas que fueron secuenciadas por el Instituto de Microbiología de la Universidad San Francisco de Quito. Se concluyó que BugSeq tiene una detección de lecturas superior en 3% sobre Kraken y 20% sobre Kaiju, (a nivel de filo y especies), sumado a la capacidad de ensamblaje (1% del total). Kaiju, obtuvo una asignación mínima de lecturas, posiblemente por su técnica basada en aminoácidos. Los filios prominentes fueron Chordata, Proteobacteria, Firmicutes, Tenericutes y las especies *Homo sapiens*, *Escherichia coli*, *Ureaplasma parvum*, *Klebsiella pneumoniae* y *Pasteurella multocida*. Además, se determinó la importancia de realizar remoción de secuencias del hospedador para disminuir la detección de contaminación experimental.

Palabras clave: secuenciación, NanoPore, asignación taxonómica, Kaiju, Kraken, BugSeq, zoonosis, microorganismos, metagenómica, murciélagos.

ABSTRACT

High throughput sequencing techniques allow to sequence a complex communities present in a given niche, for example, the determination of zoonotic microbial species in bats. The Oxford Nanopore Technologies reads long sequences. Its interpretation is done through bioinformatics platforms that classify organisms by assigning sequences to a known taxon. However, these platforms have a variable sensitivity due to different sample processing protocols, also differ in the time to obtain results and the accessibility of their use. The Kaiju, Kraken, and BugSeq platforms are fast, sensitive, and easy to use, each with different computer allocation methods. A comparison of the three platforms was made using samples from bats (saliva, blood, feces and oropharyngeal swabs). Samples obtained from the Ecuadorian Amazon (El Edén and Yasuní ITT) were sequenced at the Institute of Microbiology of the Universidad San Francisco de Quito. BugSeq showed 3% higher detection than Kraken and 20% higher than Kaiju (at the phylum and species levels). Moreover, the assembly capacity of genomes around 1% of the total of sequences. Kaiju got minimal read allocation, possibly because of his amino acid-based technique. Prominent phyla were Chordata, Proteobacteria, Firmicutes, Tenericutes, and the species *Homo sapiens*, *Escherichia coli*, *Ureaplasma parvum*, *Klebsiella pneumoniae*, and *Pasteurella multocida*. In addition, the importance of removing the host sequences was determined to reduce the detection of contamination.

Key words: sequencing, NanoPore, Kaiju, Kraken, BugSeq, taxonomic assignment, zoonosis, microorganisms, metagenomics, bats.

TABLA DE CONTENIDO

INTRODUCCIÓN.....	11
MÉTODOS.....	15
RESULTADOS.....	18
DISCUSIÓN.....	21
CONCLUSIONES.....	25
TABLAS.....	26
FIGURAS.....	28
REFERENCIAS BIBLIOGRÁFICAS.....	33
ANEXOS.....	36

ÍNDICE DE TABLAS

Tabla 1. Calidad del secuenciamiento con NanoPlot de datos crudos y hospedero removido.....	25
Tabla 2. Porcentajes de asignación taxonómica dependientes del dominio (bacterias, eucariotas, arqueas y virus), no clasificados, otros y ensamblados.....	26

ÍNDICE DE FIGURAS

Figura 1. Porcentajes de asignación taxonómica general	27
Figura 2. Promedio de porcentajes de asignación taxonómica de las muestras de murciélago correspondientes al Filo.....	28
Figura 3. Asignación taxonómica de 12 muestras provenientes de murciélagos y la remoción del hospedero, datos correspondientes al Filo.....	29
Figura 4. Porcentajes de asignación taxonómica de Filo para la muestra “11-BC14” con el uso de Kaiju	30
Figura 5. Asignación taxonómica de 12 muestras provenientes de murciélagos y la remoción del hospedero, datos correspondientes a Especies.....	31

ÍNDICE DE ANEXOS

Anexo 1. Script de NanoPlot	36
Anexo 2. Script remoción del hospedero	36
Anexo 3. Tabla metadata y otu para microbiome analyst.....	37

1. INTRODUCCIÓN

1.1 Nuevas tecnologías de secuenciación

La caracterización de especies microbianas en muestras biológicas ha sido un reto para los investigadores porque de acuerdo a Reimer (2022), solo el 0,1% de especies existentes se han podido aislar y caracterizar dentro del laboratorio (p. 3) de ahí que el 99,9% de especies microbianas eran desconocidas y fue posible identificarlas mediante técnicas moleculares (Reimer, et al., 2022). Las nuevas tecnologías de secuenciamiento nos permiten generar una gran cantidad de lecturas de ADN de manera masiva y en paralelo, de forma rápida, con alta precisión y a costos menores por base secuenciada, a esto le llamamos Secuenciamiento de Nueva Generación o Next Generation Sequencing (NGS) (Rubio, et al., 2020), entre ellas se caracterizan las ONT (Oxford Nanopore Technologies) que se basan en la utilización de nanoporos anclados a una membrana que emite una señal específica de cada base y es contabilizada por el equipo mediante un algoritmo (Bharagava, et al., 2019). Incluye la lectura de secuencias largas desde 10^4 a 10^6 bases con buena calidad, sin exigencias excesivas de material, brinda resultados en tiempo real y costos accesibles (Bharagava, et al., 2019)

1.2 Importancia de la metagenómica

La metagenómica comprende el análisis de todos los organismos dentro de una muestra específica. Se identifican mediante la elaboración de bibliotecas de ADN o ARN para su posterior análisis bioinformático (NIH, 2022). Se utiliza paquetes estadísticos basados en el sistema Linneano para establecer; reino, filo, clase; orden, familia, género, especie e incluso isotipos reconocibles solo a nivel molecular (Müller, 2022; Rugiero, et al., 2015). La clasificación taxonómica compara cada lectura obtenida por secuenciación con una base de datos, sin embargo, existen programas que realizan la alineación directa de secuencias que tardan un largo tiempo (Wanner, 2022; Menzel, et al., 2016). Por tal razón, se proponen nuevas

plataformas de asignación taxonómica más eficientes en tiempo y uso para identificar microorganismos de interés.

1.3 Los murciélagos y su relevancia en el ambiente

Los murciélagos son mamíferos pertenecientes al orden Chiroptera con más de 1400 especies conocidas. Están presentes en todos los continentes a excepción de regiones polares, montañas de gran altura y océanos (Núñez, 2021). En Ecuador el 39% de los mamíferos son chiropteros, con 178 especies habitantes y 5 especies endémicas pertenecientes a la Costa y Amazonía (Urgiles, et al., 2021). Es uno de los mamíferos más antiguos que ha transcurrido largos periodos de especiación (más de 100 millones de años) (Wang & Anderson, 2019). Su capacidad de volar permite su distribución en diversas ubicaciones geográficas, además de su amplia adaptación fisiológica a temperaturas variadas y diversos hábitos alimenticios como insectos, frutas, néctar, flores, hojas, carroña, sangre, peces, reptiles, anfibios, entre otros (Urgiles, et al., 2021). Son importantes para los ecosistemas por su rol como polinizadores, dispersores de semillas y control de plagas que afectan a la agricultura. Además de ser reconocidos por su predisposición limitada a contraer cáncer y por sus largos tiempos de vida (Wang & Anderson, 2019).

No obstante, se han reconocido como un reservorio natural de virus, bacterias y parásitos, por esta razón son de preocupación sanitaria por el riesgo de transmisión de enfermedades zoonóticas que representan el 75% de enfermedades infecciosas emergentes (Wang & Anderson, 2019). Entre los virus de preocupación se incluyen rabia, sarbecovirus (SARS, MERS), ébola, virus NIPAH y virus Hendra, además de la posibilidad de que existan microorganismos desconocidos emergentes (Urgiles, et al., 2021). Por tal razón, la caracterización de su genoma nos puede brindar información valiosa sobre la microbiota presente en los murciélagos que puede predisponer a los humanos a infecciones. Las

principales vías de transmisión son saliva, heces, orina, sangre y otros fluidos (Wang & Anderson, 2019).

1.4 Plataformas de asignación taxonómica

Kaiju, Kraken y Bugseq son plataformas de asignación taxonómica rápida, de fácil acceso y no se requiere un alto conocimiento para su ejecución, cada uno de ellos tiene diferentes algoritmos de identificación. Por ejemplo, Kraken es un programa que ejecuta su identificación de lecturas de ADN por la asignación de *k-mers*, subcadenas cortas de una coincidencia de lecturas (*k*) de longitud fija y así poder comparar la cadena de nucleótidos obtenidos, utilizando por defecto $k=31$ (Wood, et al., 2019; Wood & Salzberg, 2014). Estos *k-mers* se almacenan uno a uno en una plataforma para una ayuda de búsqueda rápida asignando un taxón que coincida con el ancestro común más bajo o en sus siglas en inglés: lowest common ancestor (LCA) que están almacenados en una base de datos “Genbank”, aquellos *k-mers* que no se encuentran en la base no son registrados (Wood, et al., 2019; Wood & Salzberg, 2014).

Kaiju utiliza esta metodología en comparación a una base de datos de proteínas microbianas que se almacenan como aminoácidos (Patil, et al., 2021). Utiliza coincidencias exactas mínimas (MEM) guardadas en la base de datos de proteínas "NR" de NCBI con genomas microbianos de referencia para su comparación, en cuanto se encuentra un MEM se genera un distintivo del nodo correspondiente, afín a la especie o cepa perteneciente. La técnica de clasificación en bloques Burrows Wheeler (BWT) alinea las MEM con las lecturas en la base de datos (Kempa & Kociumaka, 2022). En adición, Kaiju tiene la suma de una matriz específica BLOSUM62 a través de esta, es posible realizar ciertas sustituciones (Greedy) no mayor de 1 a 5 aminoácidos para una identificación completa y evitar la pérdida de información a causa del desconocimiento del segmento (Menzel, et al., 2016).

Otra plataforma que realiza asignación taxonómica basada en ADN es BugSeq que utiliza la plataforma en línea RefSeq que abarca genomas completos de bacterias, hongos, virus y humanos que es monitoreada y actualizada mensualmente (Chandrakumar, et al., 2022). Este programa realiza una reasignación de lecturas genómicas señaladas y almacenadas en la herramienta “minimap2” por un método bayesiano correspondiente a su LCA que se hizo con el uso de Pathoscope (versión 2.0.7). El LCA es calculado e incluido en una plataforma conocida como Recentrifuge (v1.1.1) para una revisión de los taxones mínimos y concentrar la información para su visualización. Además, este programa realiza un control de calidad usando la herramienta MultiQC (Fan, et al., 2021). Los programas que realizan su clasificación mediante una base de datos de referencia tienen una desventaja porque solo una pequeña parte de metagenomas han sido descifrados hasta el momento (Menzel, et al., 2016). Por lo tanto, los sistemas de *novo* son necesarios para nuevos descubrimientos.

Además, el análisis de la calidad de secuenciamiento es relevante para evitar resultados erróneos por la incorrecta asignación de bases nucleotídicas. Se propuso la utilización NanoPlot mediante el servidor público Bioconda (Anaconda, 2022) fundamentado en Python, utiliza archivos en formatos de entrada FASTA, FASTQ, bam y zip para generar ilustraciones sobre la calidad de la lectura para su mejor comprensión; longitud, rendimiento, calidad, identidad y mapeo de lectura (Wouter, et al., 2018).

En el presente estudio se utilizó ONT para determinar las secuencias microbianas que se encontraban en 12 diferentes muestras de murciélagos provenientes de la Amazonía ecuatoriana Yasuní ITT y El Edén mismas que fueron previamente secuenciadas por el Instituto de Microbiología de la USFQ, esto con el objetivo principal de establecer una base referencial para la comparación de tres plataformas taxonómicas que fueron escogidas por su mayor utilización gracias a su accesibilidad, estas plataformas fueron Kaiju, Bugseq y Kraken.

2. MÉTODOS

2.1 Recolección muestral

Se analizaron muestras provenientes de murciélagos de la Amazonía ecuatoriana, de los sectores “El Edén” y de la reserva protegida “Yasuní” que fueron previamente secuenciadas a través de un proyecto de investigación realizado en el Instituto de Microbiología de la USFQ. Se analizaron 12 muestras provenientes de murciélagos (sangre, saliva, muestra orofaríngea y heces). Se usó la tecnología de ONT (Oxford Nanopore Technologies, Oxford, Reino Unido) de acuerdo a las instrucciones del fabricante con el objetivo de la adquisición de secuencias de lecturas largas. La serie de secuenciación se programó con MinKnow por 18 horas con una base de datos activa en tiempo real, se guardaron en formato FASTQ. Se utilizó el programa Porechop para el agrupamiento de secuencias (Rrwick, 2018) de acuerdo a sus barcodes asignados y adaptadores. A continuación, se retiró los adaptadores. Las 12 muestras fueron demultiplexeadas quitando los barcodes añadidos, para impedir la modificación de lecturas en los microorganismos obtenidos. Por último, se guardaron en formato FASTQ.

2.2 Análisis de calidad de las secuencias

Se determinó la calidad del secuenciamiento mediante el uso de la terminal del sistema operativo Ubuntu para la activación de NanoPlot (Anaconda, 2022), de acuerdo a los comandos especificados en el Anexo 1. Dentro de los resultados obtenidos, se realizó una comparación de la calidad de lectura media, la media de longitud de reads, mediana de longitud de lecturas, mediana de calidad de reads, número total de reads, longitud de read N50 y total de bases secuenciadas. A continuación del análisis de la calidad de secuenciamiento, se resumió en una tabla para determinar si es aceptable, se tomó como valor de referencia superior al 11% (Magi, et al., 2017) siendo el valor estándar para Nanopore Technologies.

2.3 Remoción de lecturas del hospedero

Por otro lado, para evitar un alto grado de lecturas provenientes de organismos sin interés directo como es el genoma del murciélago (Yeoh, 2021) se realizó la eliminación de lecturas del hospedero (*Phyllostomus elongatus*) para las 12 muestras recolectadas de acuerdo al script del Anexo 2 adjunto. Se obtuvo el genoma de referencia de la plataforma NCBI y se descargó en formato FASTA (NIH, 2022). Se requirió la alineación de los genomas de interés con las secuencias del genoma de referencia con la herramienta “BWA MEM”.

Se observó los resultados del alineamiento (número de lecturas mapeadas) con “SAMtools view” (SAMtools, 2012). A continuación, las secuencias alineadas se convirtieron en un archivo binario (.bam) y se ordenaron por posicionamiento de acuerdo al genoma de referencia con ayuda de la herramienta “SAMtools sort”. Se indexó las lecturas con “idxstats” y se observó las estadísticas de que las muestras hayan sido mapeadas y que no se encuentre ningún valor en cero con “SAMtools idxstats”. Finalmente, “BEDtools” para el filtrado que solo conserva aquellas lecturas que no se mapearon con el genoma de referencia obteniendo así, 12 muestras con los datos de origen y 12 muestras con la remoción del hospedador (Aaron, et al., 2022)

2.4 Análisis de clasificación taxonómica usando la plataforma BugSeq

Para el uso de BugSeq, dentro de la página “bugseq.com/academic” se seleccionó las 24 muestras (datos crudos sin secuencias del hospedero) en formato FASTQ. Se envió a la plataforma utilizada “NanoPore” y “MinION/GridION -R9.4.1” (Bugseq, Vancouver, Canadá) y se digitó los porcentajes de lecturas del total de la raíz, tomando en cuenta clasificados y no clasificados de filo y especie. Adicional a esto, se cuantificó aquellos taxones que fueron ensamblados por el programa.

2.5 Análisis de clasificación taxonómica usando la plataforma Kraken

De acuerdo al instructivo de uso, para Kraken se ingresó a Galaxy Europe “<https://galaxyproject.eu/>” se creó una cuenta y se subió todas las secuencias en formato FASTQ al almacenamiento en línea. Para dotar de un código de clasificación taxonómica se colocó “Kraken2 on collection”. Se obtuvo los resultados numéricos de clasificación con la opción “kraken report on collection” y “mpa report”, se anotaron los resultados para filo y especie (Wood, et al., 2019).

2.6 Análisis de clasificación taxonómica usando la plataforma Kaiju

En cuanto a la plataforma Kaiju, se ingresó a la página “<https://kaiju.binf.ku.dk/server>” (Menzel, et al., 2016). Se digitó: nombre de proyecto y correo. Se subió el archivo con las lecturas en formato zip. Dentro de las opciones se escogió la base de datos de referencia “NCBI BLAST nr +euk”, se acepta filtro “SEG” para aquellas secuencias de baja complejidad, modo de corrida “MEM”. Para las opciones de longitud mínima de coincidencia, puntuación mínima y desajustes permitidos se dejó por default del programa y se presiona enviar. Este procedimiento se realizó para cada una de las 24 muestras. Se recopiló datos de filo y especie.

Finalmente, para una mejor visualización de los resultados obtenidos, se utilizó el programa de Microbiome Analyst, donde se redactaron las tablas en formato textual tanto para los datos OTU y la metadata, tomando en cuenta las lecturas >150 o las 5 más representativas a nivel de Filo y Especie (Xia Lab, 2022).

3. RESULTADOS

3.1 Análisis de la calidad de secuenciamiento

En la Tabla 1 se observan los valores correspondientes a la calidad de secuenciamiento obtenido por NanoPlot. Se comparan los valores de los datos crudos y los valores con el hospedador removido. Se evaluó las siguientes categorías; calidad de secuencias, la longitud media de las secuencias, el número total de reads, la longitud de las secuencias N50 y el número total de bases. Los valores obtenidos para la calidad de las secuencias en la Tabla 1, hacen referencia al promedio de la media de puntuaciones de la calidad de las bases que son: para datos crudos 11.1 y 10.26 para remoción del hospedero, acuerdo a los parámetros de NanoPlot son una buena puntuación para los protocolos de ONT (Oxford Nanopore Technologies) con un mínimo de 10, es decir, una precisión de asignación de bases del 90% (Griffiths, 2021). Se excluye la muestra “Phy-sangre-BC17” de la remoción del hospedero por un valor de 8.8.

La longitud media de las lecturas para los datos crudos fue de 657.25 pares de bases y para los datos de la remoción del hospedero 591.17, no se observa una distribución variable. Por lo tanto, se consideran valores aceptables dentro de los parámetros del dispositivo MinION por su capacidad de secuenciación en secuencias de lectura larga entre 200 pb a 150000 (Nanopore, 2021).

El número de lecturas (reads) totales obtenidos del secuenciamiento para los datos crudos fue en promedio de 129,346.75 y después de la remoción del hospedador de 22,088.00 correspondiente al 17% de los datos crudos.

La longitud de lectura N50 para los datos crudos de 887.92 y 849.00 confirmando que el 50% de las lecturas tienen este número de longitudes alineadas para posteriormente realizar la identificación taxonómica. Finalmente, el número total de bases recopiladas del

secuenciamiento para los datos crudos 99,754,402.67 y en la remoción del hospedero 14,774,767.75 correspondiente al 14% de los datos crudos. Todos estos datos se visualizan en la Tabla 1 del análisis de calidad de secuenciamiento, mismo que cumple con los valores requeridos para asignación taxonómica.

3.2 Comparación de datos crudos y remoción del hospedero

En la Figura 1 se observa el resumen de la contabilización de lecturas taxonómicas asignadas de acuerdo al Dominio al que pertenecen: Bacteria y Eucariota son las más representativas, no obstante, Arquea tuvo la menor asignación de lecturas. Se observa en BugSeq y Kraken que en los datos crudos se obtiene una mayor asignación de eucariotas frente a bacterias. Sin embargo, después de la remoción del hospedero este orden se invierte, dejando con mayor porcentaje a Bacterias que a Eucariotas. Kaiju tuvo la menor clasificación y después de la remoción del hospedero tuvo porcentajes de Bacterias de 0.05%, Eucariotas y Arqueas 0% como se observa en la Tabla 2.

También, se observa la clasificación de Virus que de acuerdo a la Tabla 2 tienen valores menores al 0.05% el cual no es un valor estadísticamente representativo. La capacidad de ensamblaje de BugSeq es del 0.096% de los datos crudos y 0.000023% para remoción del hospedero. Además, los grupos de “no clasificados” y “otros” tienden a aumentar después de la remoción del hospedero. Los “no clasificados” superan el 50% del total de las lecturas.

3.3 Asignación taxonómica correspondiente al Filo

3.3.1 Asignación de Filo de acuerdo al promedio general de las muestras.

En la Figura 2 se visualizan los 5 taxones más representativos de la clasificación que son los Filos Chordata, Proteobacterias, Firmicutes, Tenericutes además de los grupos “no clasificados” y “otros”. Con respecto a las plataformas de BugSeq y Kraken, se puede observar una abundancia relativa de las lecturas superior a 0.5 de 1 en ambas. Al comparar ambas

plataformas observamos que el Filo Chordata perteneciente al Reino Metazoa es el taxón más prominente, seguido de Proteobacterias, “otros”, Firmicutes y Tenericutes respectivamente. Se puede visualizar que ambas plataformas siguen patrones similares. Kraken, con abundancia de 0.40 para datos crudos y 0.30 para remoción del hospedero y BugSeq con 0.35 y 0.20 respectivamente. Por otro lado, Kaiju contó con menor clasificación taxonómica que las demás, tomando en cuenta solo los filios de Proteobacteria y “otros”.

3.3.2 Asignación de Filo de acuerdo a cada muestra.

En la Figura 3 se visualizan los Filos más prominentes para cada muestra, siendo los 5 más representativos y ordenados en forma descendente: “no clasificados”, Reino Metazoa, Filo de Proteobacterias, “otros”, Firmicutes y Tenericutes. Siendo las 3 muestras con mayor clasificación “Sample-17-BC19”, “Sample-11-BC14” y “Phy-saliva-BC16”. Kraken cuenta con un mayor número de secuencias asignadas para cada muestra, BugSeq ocupa el segundo lugar y Kaiju el tercero con una asignación taxonómica mínima. En la Figura 4 se realiza una ampliación de la muestra “11-BC14” con el uso de Kaiju donde se suma el Reino Fungi en un 2% del total.

3.4 Asignación taxonómica correspondiente a la Especie

En la Figura 5 se visualiza la clasificación de las 5 especies con mayor número de lecturas además de “no clasificados” y “otros”, en orden descendente son: “no asignados”, *Homo sapiens*, “otros”, *Escherichia coli*, *Ureaplasma parvum*, *Pasteurella multocida* y *Klebsiella pneumoniae*. También se observa que existe un mayor número de lecturas no asignadas de Especies que de acuerdo al Filo. Por otro lado, Kaiju solo tiene valores de “otros” y *Ureaplasma parvum* que después de la remoción del hospedador ya no son visibles.

4. DISCUSIÓN

La metagenómica utiliza técnicas de secuenciación dirigidas a determinar el material genómico completo en una muestra, nos brinda información sobre identificación de nuevos genes, sobre su estructura y relaciones evolutivas de una comunidad biológica (Roumpeka, et al., 2017). Sin embargo, sus principales limitaciones son: a) obtiene los genomas diferentes al del hospedador al que pertenece, por lo que pueden detectarse secuencias adicionales a las de interés directo b) la identificación es inespecífica ya que no se toman en cuenta codones, promotores, sitios de unión, entre otros. c) es necesario mantener estándares de calidad de secuenciación altos (Rompeuka, et al., 2017)

El análisis de la calidad de las lecturas secuenciadas es importante porque nos permite crear parámetros de aceptación para que el investigador se asegure de que el proceso de secuenciación se llevó a cabo correctamente (Ullah, et al., 2022).

En los análisis de metagenomas es posible encontrar diversos contaminantes ambientales por la manipulación experimental de la muestra que puede ocurrir por procedimientos como el almacenamiento, preparación y análisis (Chrisman, et al., 2022). Por tal razón se confirmó que la remoción del hospedero nos permite aumentar la sensibilidad de caracterización taxonómica, al reducir el número de lecturas registradas correspondientes a *Homo sapiens* (Figura 5). Con respecto a Kraken y BugSeq después de la remoción de las secuencias del hospedero, aumentó significativamente las secuencias de bacterias presentes en la muestra. Estas moléculas de ADN sin interés directo generan nuevas secuencias que se presentan como falsos positivos (Menzel, et al., 2016). Por otro lado, Kaiju mostró una menor asignación de lecturas, posiblemente por el uso de una base de datos que utiliza aminoácidos para su identificación, ya que tiene información más limitada que aquellas basadas en nucleótidos (Tang, et al., 2022). Chrisman (et al, 2022) recomienda que todos los análisis deben realizarse en una misma línea

de tiempo, utilizar muestras control y realizar la remoción del hospedero en caso de requerirlo porque elimina aquellas secuencias que interfieran y generen valores atípicos (p. 12)

Adicionalmente, el valor de las lecturas no clasificadas excedió un 50% sobre las demás clasificaciones taxonómicas de las 3 plataformas (Figura 1), esto contrasta con que los métodos de secuenciación arrojan una gran cantidad de datos que los clasificadores taxonómicos no mapean y generalmente no se toman en cuenta, sin embargo, estas secuencias no mapeadas pueden representar: secuencias desconocidas actualmente, genomas de referencia incompletos y secuencias provenientes de contaminación (Chrisman, et al., 2022). Se recomienda el uso de bases de datos actualizadas y completas, además de análisis de *novo*.

Kaiju, Kraken y BugSeq son clasificadores taxonómicos rápidos, sin complejidad de uso y de acceso público. Una desventaja es que grandes cantidades de lecturas genómicas en una corta cantidad de tiempo son limitadas para superar la divergencia evolutiva, dejando así gran parte de las lecturas metagenómicas sin clasificar (Menzel, et al., 2016).

Kaiju por su asignación de aminoácidos supone una mayor sensibilidad al ser secuencias mayormente conservadas, excluyendo errores de la degeneración del código genético, sin embargo, la biblioteca de NCBI “NR” de proteínas contiene menor información que aquella basada en nucleótidos y selecciona aquellas lecturas que tienen una suma acumulativa de sus repeticiones (Ullah, et al., 2022). Además, la técnica de búsqueda Burrows Wheeler de MEM requieren una RAM pequeña menor a Kraken y BugSeq siendo una plataforma más rápida de lecturas, sin embargo, su sensibilidad tiende a reducirse conforme la distancia evolutiva aumenta por la sustitución de aminoácidos que realiza (Menzel, et al., 2016).

En la Figura 1 se observa la clasificación de Kaiju con valores mínimos de asignación taxonómica (<3%), al realizar un acercamiento a la muestra “11-BC14” los filos más prominentes son Proteobacteria, Firmicutes y algunos filos provenientes de hongos, sin

embargo, la abundancia que representan es mínima. Se recalca que se clasificó como “otros” a aquellas asignaciones que contenían menos de 150 lecturas por taxón (Figuras 1, 2, 4, 5), esto porque estos clasificadores taxonómicos pueden suponer resultados falsos positivos por coincidencias aleatorias, así, solo se seleccionan los más repetitivos (Menzel, et al., 2016).

Para Kraken el presente estudio determinó una sensibilidad de lecturas taxonómicas del 70% (Figura 2). Esto se atribuye a que utiliza la base de datos pública RefSeq de NCBI, tiene una colección de secuencias de ADN, ARN y proteínas anotadas y curadas, almacenando una secuencia identificadora por cada biomolécula (Wood, et al., 2019). Sin embargo, Wood (et al, 2019) define que una desventaja del uso de *k-mers* es que aquellas secuencias que no coincidan con un *k-mer* determinado no se tomarán en cuenta. RefSeq no utiliza genomas completos ensamblados por lo cual resulta en información limitada (p. 8). En concordancia con la Figura 2 Kraken supone una clasificación taxonómica superior a BugSeq del 5% en datos crudos y 10% para la remoción del hospedero y para Kaiju del 70% para datos crudos. Se detectó los filos de Chordata, Proteobacteria, Firmicutes y Tenericutes.

BugSeq, por otra parte, representó una clasificación total del 60%. Sin embargo, a nivel de clasificación microbiana supera a Kraken con el 5% y a Kaiju con el 20% (Figura 2). Este programa asigna lecturas genómicas señaladas y almacenadas en la herramienta “minimap2” por un método bayesiano correspondiente a su LCA con una referencia de taxones mínimos. Además, este programa realiza un control de calidad con MultiQC (Fan, et al., 2021). Este programa tiene la ventaja de utilizar un clasificador de lecturas pequeñas que pueden ensamblarse mediante la agrupación de contigs por identidad taxonómica de acuerdo a la base de datos de dominio público Genbank de NCBI que abarca genomas completos e incompletos de bacterias, hongos, virus y humanos (Chandrakumar, et al., 2022). En el presente estudio los genomas ensamblados fueron del 0.096% (Tabla 2). De acuerdo a Chandrakumar (et al, 2022),

este método ha sido catalogado como una alineación y, carece de sensibilidad para caracterizar taxones divergentes (p. 5). Además, BugSeq usa datos de referencia tiene la desventaja de que solo una pequeña parte de genomas completos han sido descifrados hasta el momento (Menzel, et al., 2016).

La clasificación de estas tres plataformas no requiere ensamblaje por lo cual es más rápida y de fácil utilización. No obstante, estas técnicas son limitadas para caracterizar comunidades microbianas complejas (Chandrakumar, et al., 2022). Actualmente, uno de los principales desafíos que representan estos clasificadores taxonómicos es la cantidad de falsos positivos de baja abundancia que informan, por lo cual los investigadores suelen filtrar estos informes con un umbral de abundancias predeterminado. Por esta razón se sugiere que se estandarice este proceso en las plataformas para un mejor rendimiento (Chandrakumar, et al., 2022).

Dentro del presente estudio se determinó que los filos prominentes en murciélagos fueron Chordata (secuencias de murciélagos o humanas), Proteobacteria, Firmicutes y Tenericutes (Figura 4) y las especies bacterianas *Escherichia coli*, *Ureaplasma parvum*, *Klebsiella pneumoniae* y *Pasteurella multocida* (Figura 5). Los microorganismos son parte de la microbiota de estos animales y probablemente no representan ningún riesgo a la población humana (PAHO, 2022).

5. CONCLUSIONES

En las tres plataformas se observó una disminución del dominio de eucariotas y aumento de bacterias después de la remoción de las secuencias del hospedero. Es importante recalcar que para un análisis de metagenomas se deben mantener estándares de calidad de secuenciación óptimos para evitar resultados erróneos. Se concluyó que las plataformas de clasificación taxonómica Kaiju, Kraken y BugSeq son rápidas, de fácil utilización y de acceso público. La sensibilidad de detección de lecturas varía dependiendo el tipo de metodología informática de asignación y la base de datos que utiliza. Kraken obtuvo mejores resultados de asignación taxonómica general del 70%, seguido BugSeq 60% y Kaiju 10%, sin embargo, BugSeq fue la plataforma que brindó mayor porcentaje de asignación microbiana después de la remoción del hospedero con un 30%, seguido de Kraken 27% y Kaiju 5%. Se recomienda el uso de Kraken y BugSeq para análisis metagenómicos. Adicionalmente, es posible que la alta cantidad de no clasificados obtenidos en este estudio se deba a secuencias desconocidas actualmente, por lo que se debería realizar más estudios. Los filos más representativos fueron Chordata, Proteobacteria, Firmicutes y Tenericutes y las especies más representativas fueron *Escherichia coli*, *Ureaplasma parvum*, *Klebsiella pneumoniae* y *Pasteurella multocida*, algunas cepas de estas bacterias pueden ser patógenas para los humanos.

6. TABLAS

Tabla 1. Calidad del secuenciamiento con NanoPlot de datos crudos y hospedero removido

Tipo de Muestra	SampleID	Calidad de secuencias	Longitud media	Número de Reads	Longitud de lectura N50	Total de bases
Datos crudos	C_BC21	10.60	742.00	46,569.00	1,035.00	40,208,917.00
	Sample11_BC14	10.30	550.00	54,360.00	745.00	34,510,251.00
	Sample17_BC19	10.70	673.00	31,076.00	894.00	24,085,526.00
	Art_anal_BC02	11.30	659.00	122,641.00	861.00	91,498,367.00
	Art_saliva_BC19	11.40	614.00	175,348.00	812.00	125,769,904.00
	Car_anal_BC15	11.50	693.00	283,077.00	990.00	237,702,898.00
	Car_saliva_BC13	11.30	621.00	147,949.00	869.00	110,642,794.00
	Car_sangre_BC14	11.20	676.00	152,553.00	899.00	120,587,614.00
	F_BC20	10.70	742.00	38,673.00	965.00	32,255,714.00
	Phy_anal_BC18	11.70	740.00	190,481.00	1,019.00	167,497,722.00
	Phy_saliva_BC16	11.40	616.00	112,782.00	823.00	82,048,381.00
	Phy_sangre_BC17	11.10	561.00	196,652.00	743.00	130,244,732.00
	Prom	11.10	657.25	129,346.75	887.92	99,754,401.67
Host removido	C_BC21	10.20	748.00	7,203.00	1,049.00	6,011,158.00
	Sample11_BC14	10.30	539.00	51,764.00	727.00	32,056,715.00
	Sample17_BC19	10.40	671.00	14,367.00	926.00	11,135,755.00
	Art_anal_BC02	10.50	545.00	27,107.00	792.00	16,610,917.00
	Art_saliva_BC19	10.40	555.00	30,781.00	794.00	19,795,549.00
	Car_anal_BC15	10.40	561.00	48,127.00	850.00	32,259,835.00
	Car_saliva_BC13	10.50	583.00	35,282.00	843.00	24,023,118.00
	Car_sangre_BC14	10.10	574.00	10,121.00	870.00	6,955,894.00
	F_BC20	10.40	763.00	12,880.00	971.00	10,537,381.00
	Phy_anal_BC18	10.30	564.00	8,093.00	871.00	5,292,217.00
	Phy_saliva_BC16	10.80	609.00	15,727.00	830.00	10,857,855.00
	Phy_sangre_BC17	8.80	382.00	3,605.00	665.00	1,760,819.00
	Prom	10.26	591.17	22,088.08	849.00	14,774,767.75

Tabla 2. Porcentajes de asignación taxonómica dependientes del dominio (bacterias, eucariotas, arqueas y virus), no clasificados, otros y ensamblados

	Bacterias	Eucariotas	Arqueas	Virus	No clasificados	Otros	Ensamblados
BUGSEQ	6.031	18.266	2.259	0.0279	56.129	17.219	0.096
BUGSEQ-NH	12.389	3.904	0.0275	0.0569	54.023	29.657	0.000023
KRAKEN	8.512	30.468	0.048	0.0275	11.907	49.066	-
KRAKEN-NH	14.49	8.262	0	0.0283	54.237	23.011	-
KAIJU	0.985	0.0775	0	0.0061	84.643	14.294	-
KAIJU-NH	0.051	0	0	0.0051	98.993	0.955	-

7. FIGURAS

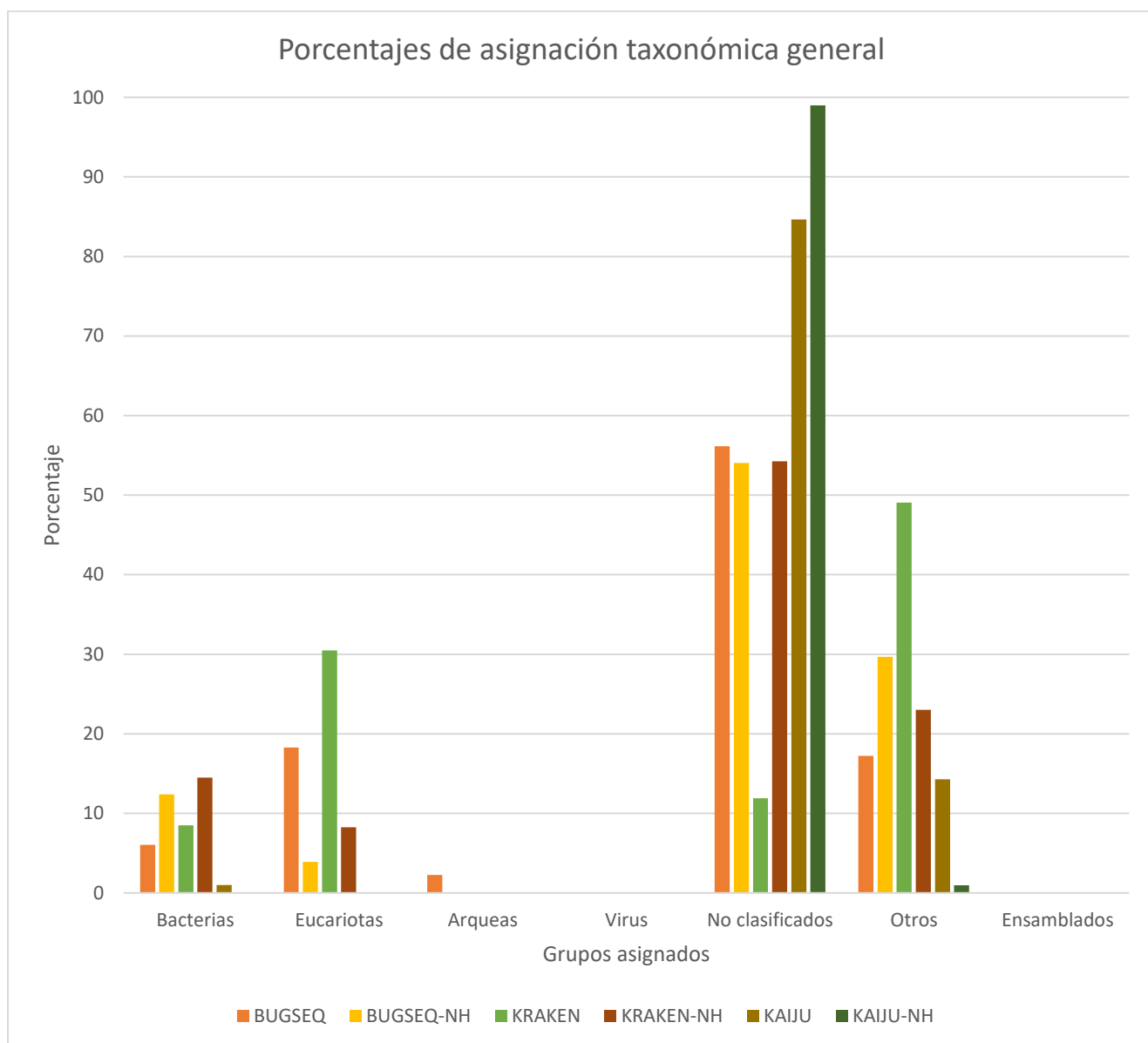


Figura 1. Porcentajes de asignación taxonómica general. El eje vertical representa el porcentaje alcanzado y el eje horizontal la asignación taxonómica correspondiente al dominio al que pertenecen, además de virus, no clasificados, otros y ensamblados por las plataformas (BugSeq, Kaiju y Kraken) (en colores) con su respectivo análisis de remoción del hospedero. Diagrama realizado en Excel.

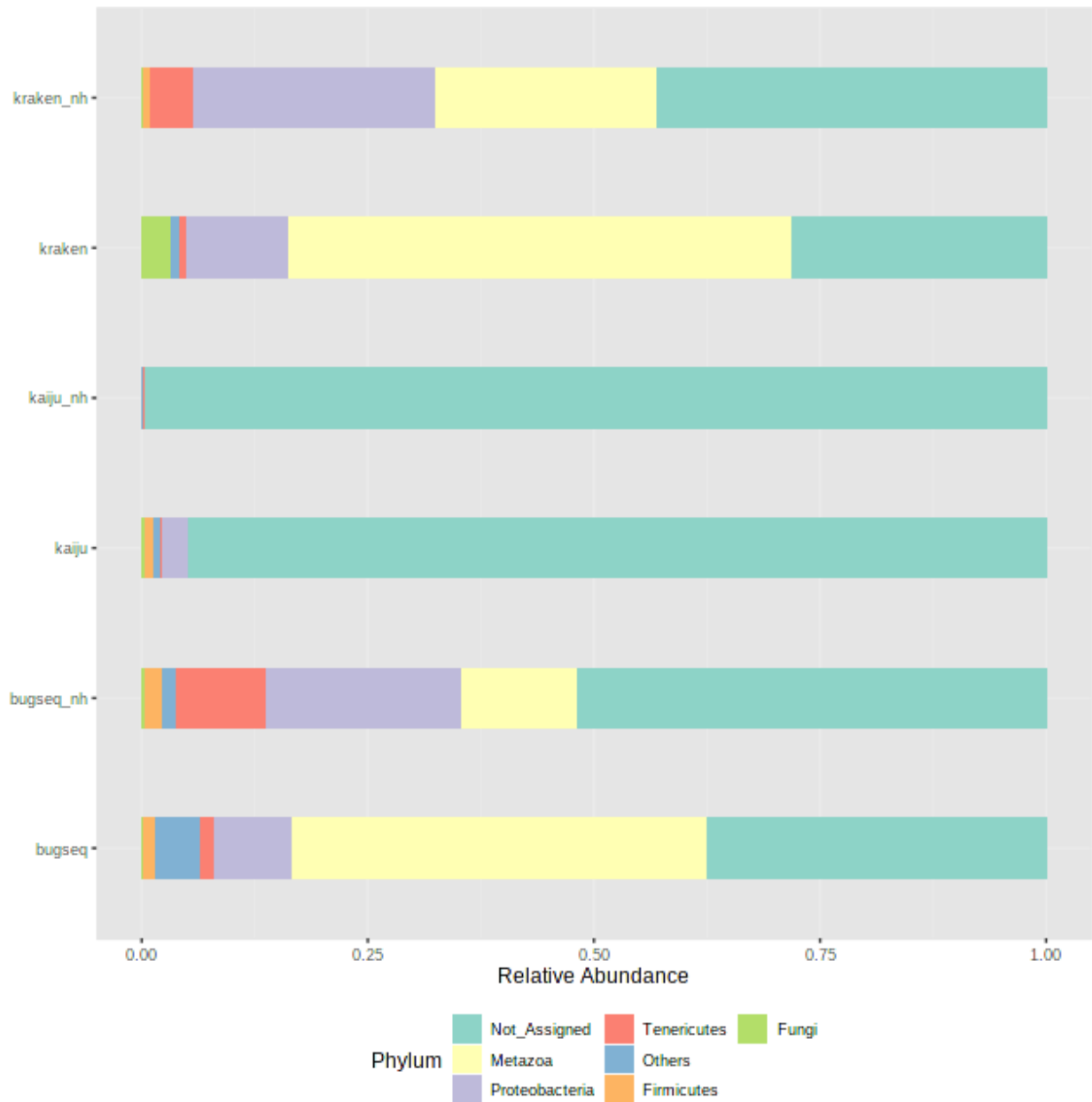


Figura 2. Promedio de porcentajes de asignación taxonómica de las muestras de murciélago correspondientes al Filo. En el eje vertical se ubican las plataformas analizadas (Bugseq, Kaiju, Kraken) además de las muestras con remoción del hospedero (nh). En el eje horizontal está la abundancia relativa en escala =1. Diagrama realizado en Microbiome Analyst.

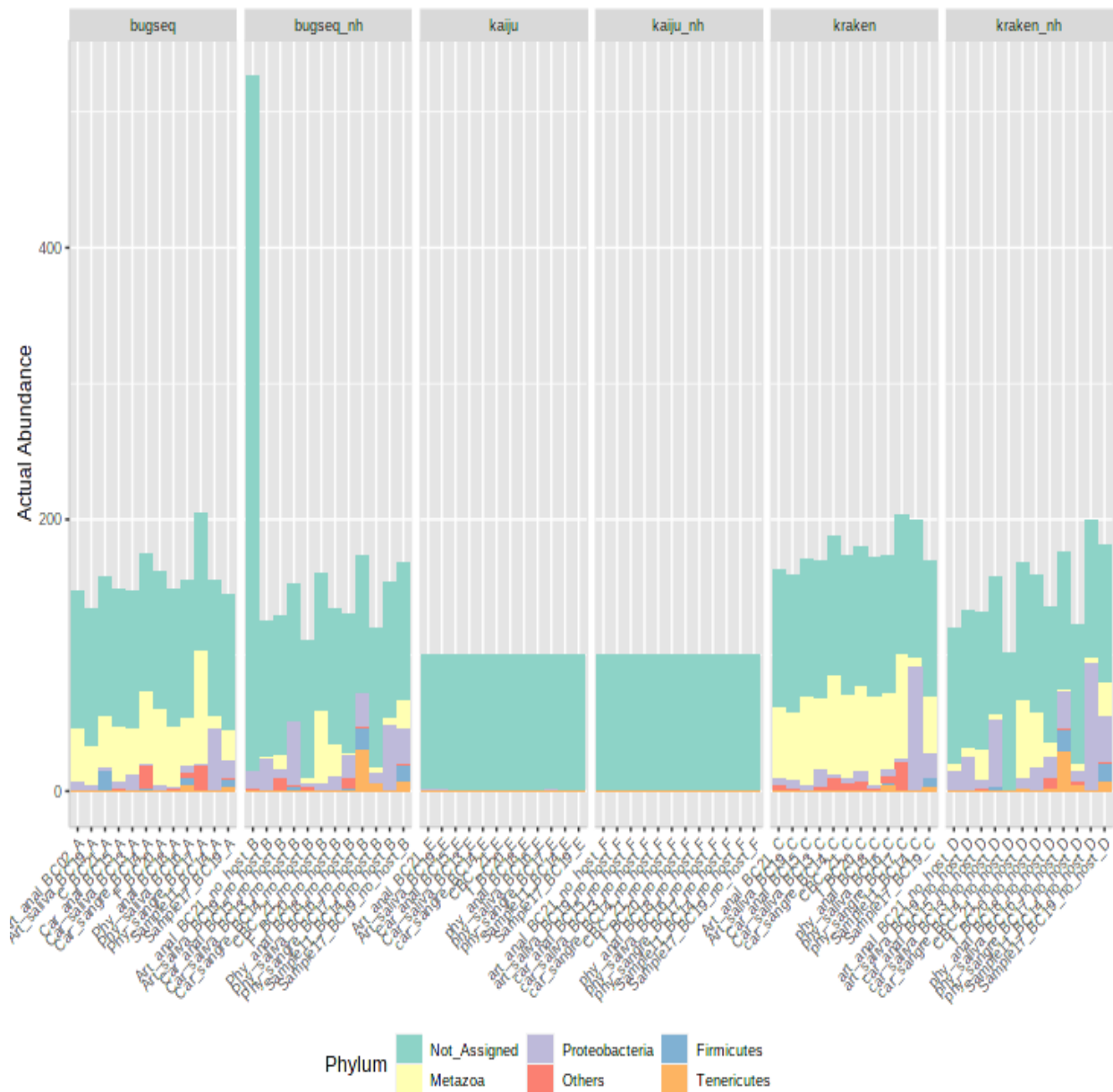


Figura 3. Asignación taxonómica de 12 muestras provenientes de murciélagos y la remoción del hospedero, datos correspondientes al Filo. Presenta la clasificación de cada columna (C) para cada Filo de la siguiente forma: C1 BugSeq datos crudos; C2 BugSeq con remoción del hospedero; C3 Kaiju datos crudos; C4 Kaiju con remoción del hospedero; C5 Kraken datos crudos; C6 Kraken remoción del hospedero. Diagrama realizado en Microbiome Analyst.

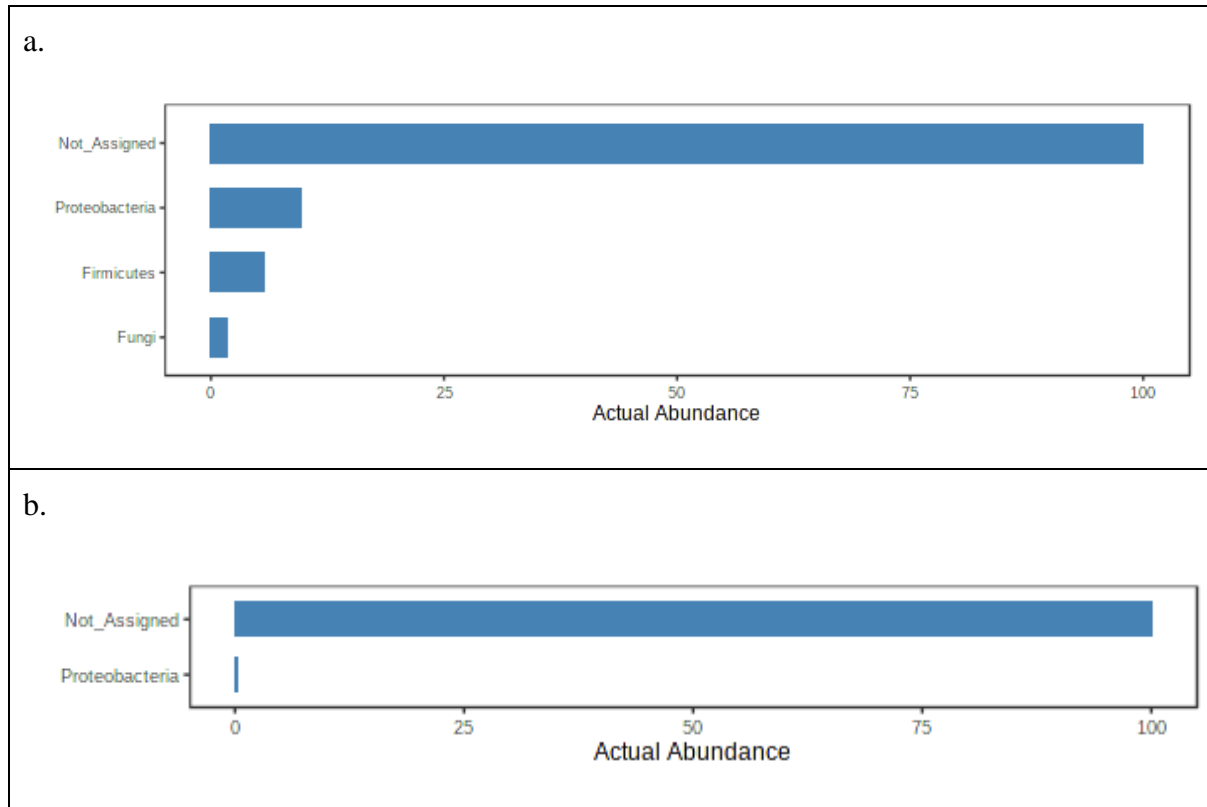


Figura 4. Porcentajes de asignación taxonómica de Filo para la muestra “11-BC14” con el uso de Kaiju. a) Datos crudos de clasificación taxonómica para el Filo de la muestra “11-BC14” b) Datos con hospedador removido de clasificación taxonómica para el Filo de la muestra “11-BC14”. Diagrama realizado en Microbiome Analyst.

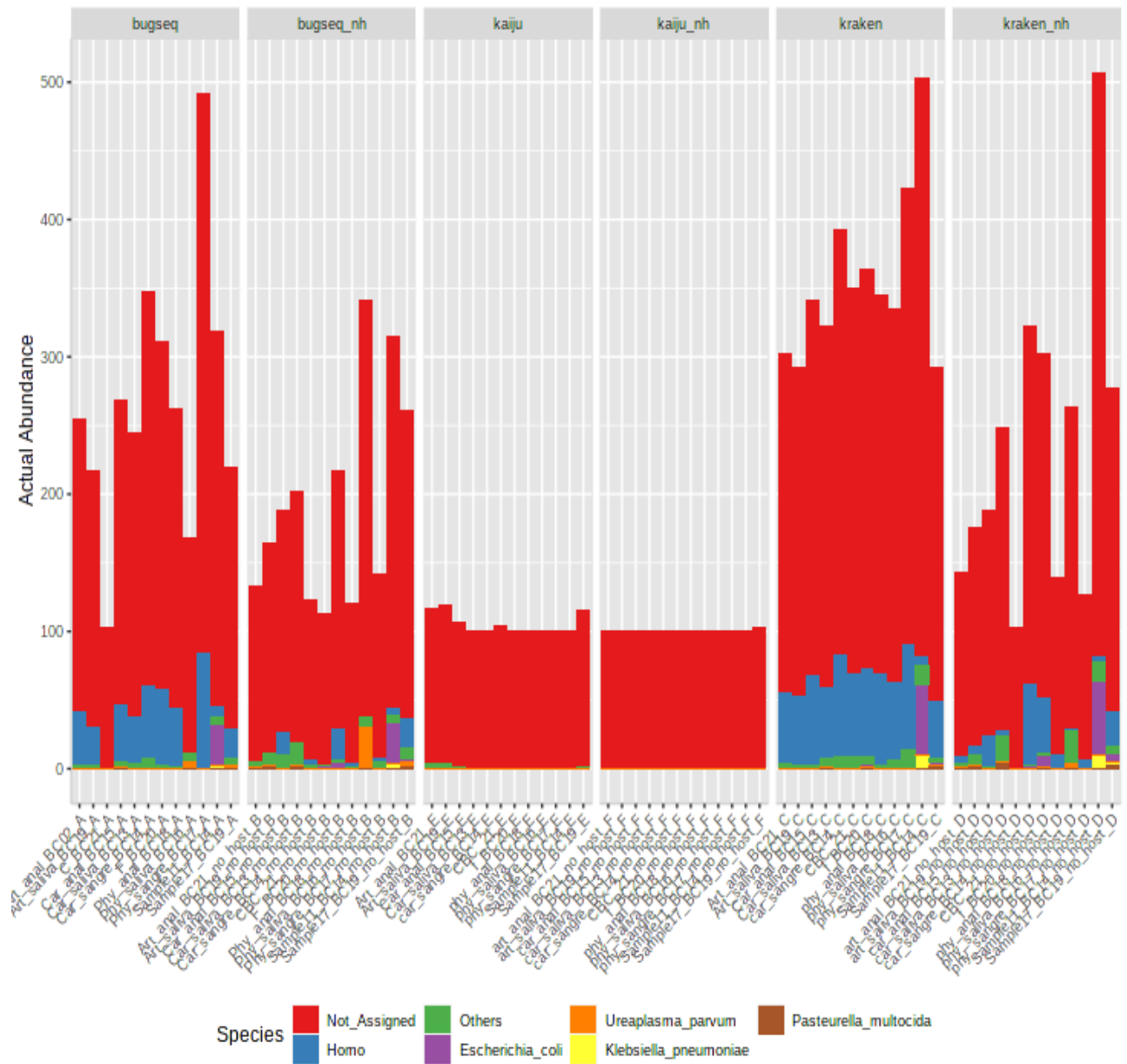


Figura 5. Asignación taxonómica de 12 muestras provenientes de murciélagos y la remoción del hospedero, datos correspondientes a Especies. Presenta la clasificación de cada columna (C) para cada Filo de la siguiente forma: C1 BugSeq datos crudos; C2 BugSeq con remoción del hospedero; C3 Kaiju datos crudos; C4 Kaiju con remoción del hospedero; C5 Kraken datos crudos; C6 Kraken remoción del hospedero. Diagrama realizado en Microbiome Analyst.

8. REFERENCIAS BIBLIOGRÁFICAS

- Anaconda (2022) *Bioconda*. Obtenido de <https://anaconda.org/bioconda>
- Bharagava, R., Purchase, D., Saxena, G., (2019). *Chapter 26 - Applications of Metagenomics in Microbial Bioremediation of Pollutants: From Genomics to Environmental Cleanup*. Microbial Diversity in the Genomic Era. Academic Press. Pages 459-477. Obtenido de <https://doi.org/10.1016/B978-0-12-814849-5.00026-5>.
- Chandrakumar, I., Gauthier, N.P.G., Nelson, C. (2022) *BugSplit enables genome-resolved metagenomics through highly accurate taxonomic binning of metagenomic assemblies*. Commun Biol 5, 151. <https://doi.org/10.1038/s42003-022-03114-4>
- Chrisman, B., He, C., Jung, JY. et al. (2022) *The human “contaminome”: bacterial, viral, and computational contamination in whole genome sequences from 1000 families*. Sci Rep 12 <https://doi.org/10.1038/s41598-022-13269-z>
- Fan, J., Huang, S. & Chorlton, S.D. (2021) *BugSeq: a highly accurate cloud platform for long-read metagenomic analyses*. BMC Bioinformatics 22, 160. <https://doi.org/10.1186/s12859-021-04089-5>
- Griffiths, S. (2021) *Quality Scores And Read Accuracy*. Obtenido de <https://labs.epi2me.io/quality-scores/>
- Illumina Inc. (2022) *Key differences between next-generation sequencing and Sanger sequencing*. Obtenido de <https://www.illumina.com/science/technology/next-generation-sequencing/ngs-vs-sanger-sequencing.html>
- Kempa, D., Kociumaka, T. (2022) *Resolution of the Burrows-Wheeler Transform Conjecture*. Obtenido de <https://cacm.acm.org/magazines/2022/6/261178-resolution-of-the-burrows-wheeler-transform-conjecture/fulltext>
- Kunkler, C., Hulewicz, J, Hickman, S. (2019) Stability of an RNA•DNA–DNA triple helix depends on base triplet composition and length of the RNA third strand, Nucleic Acids Research, Volume 47, Pages 7213–7222, <https://doi.org/10.1093/nar/gkz573>
- Magi A, Giusti B, Tattini L. (noviembre 2017) *Characterization of MinION nanopore data for resequencing analyses*. Brief Bioinform.18(6):940-953. doi: 10.1007/978-1-0716-1040-4_13
- Menzel, P., Ng, K. & Krogh, A. (2016) *Fast and sensitive taxonomic classification for metagenomics with Kaiju*. Nat Commun 7, 11257. Obtenido de <https://doi.org/10.1038/ncomms11257>
- Microbiome Analyst (2022) *Microbiome Analyst*. <https://www.microbiomeanalyst.ca/MicrobiomeAnalyst/upload/OtuUploadView.xhtml>
1
- Müller-Wille, S. (2022, September 12). *Carolus Linnaeus*. Obtenido de Encyclopedia Britannica. <https://www.britannica.com/biography/Carolus-Linnaeus>

- Nanopore Technologies (2021) *Oxford Nanopore releases 'Rev D' flow cells, enabling increase in data yields*. Obtenido de <https://nanoporetech.com/>
- NIH (2022) *Metagenomics*. National Human Genome Research Institute. Obtenido de <https://www.genome.gov/genetics-glossary/Metagenomics>
- Núñez, E. (2021) *Murciélagos*. National Geographic. Obtenido de <https://www.nationalgeographic.com/animals/mammals/facts/bats#:~:text=Classificati on,on%20blood%20rather%20than%20insects>.
- Oxford Nanopore Technologies (2022) *Oxford Nanopore Technologies*. Obtenido de <https://nanoporetech.com/>
- Reimer, L., Carbasse, J., Koblitz, J., Ebeling, C. (2022) *BacDive in 2022: the knowledge base for standardized bacterial and archaeal data*, *Nucleic Acids Research*, Volume 50, Issue D1, 7, Pages D741–D746, <https://doi.org/10.1093/nar/gkab961>
- Rubio, S., Pacheco, R., Gómez, A. (2020) *DNA Next-Generation Sequencing (NGS): Present and Future in Clinical Practice*. Pontificia Universidad Javeriana. Universitas Medica, vol. 61, núm. 2 Obtenido de <https://revistas.javeriana.edu.co/index.php/vnimedica/article/view/27461>
- Ruggiero M., Gordon D., Orrell T., Bailly N., Bourgoin T., Brusca R., et al. (2015) *A Higher Level Classification of All Living Organisms*. PLoS ONE 10(4): e0119248. <https://doi.org/10.1371/journal.pone.0119248>
- Rompeuka, D., Wallace, J., Frank, E., et al. (2017) *A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data*. *Frontiers in Genetics*. Volume 8. Obtenido de <https://www.frontiersin.org/articles/10.3389/fgene.2017.00023>
- Ullah, A., Gao, T., Rahman, W. (2022) *LDBPR: Latest Database of Protein Research*. *Fortune Journals*. <https://www.fortunejournals.com/articles/ldbpr-latest-database-of-protein-research.pdf>
- Urgiles, C., Hiruiois, M., Calero, B. (2021) *Reproductive notes on various species of Phyllostomid bats (Chiroptera: Phyllostomidae) in Ecuador*. *Boletín científico de la Asociación Ecuatoriana de Mastozoología*. ISSN 2697-3286. <https://mamiferosdelecuador.com/mammalia-aequatorialis/index.php/boletin/article/view/39/58>
- Wang L., Anderson D. (2019). *Viruses in bats and potential spillover to animals and humans*, *Current Opinion in Virology*, Volume 34, 2019, Pages 79-89, Obtenido de <https://doi.org/10.1016/j.coviro.2018.12.007>.
- Wanner, M. (2022) *Progress with long read sequencing*. The Jackson Laboratory. Obtenido de <https://www.jax.org/news-and-insights/jax-blog/2022/june/progress-with-long-read-sequencing>
- Wood, D.E., Lu, J. & Langmead, B. (2019). *Improved metagenomic analysis with Kraken 2*. *Genome Biol* 20, 257. <https://doi.org/10.1186/s13059-019-1891-0>

- Wood, D.E., Salzberg, S. (2014). *Kraken: ultrafast metagenomic sequence classification using exact alignments*. *Genome Biol* 15, R46. <https://doi.org/10.1186/gb-2014-15-3-r46>
- Wouter C., D’Hert, S., Schultz, D., Cruts, M., Broeckhoven, C. (2018). *NanoPack: visualizing and processing long-read sequencing data*, *Bioinformatics*, Volume 34, Issue 15, Pages 2666–2669, <https://doi.org/10.1093/bioinformatics/bty149>
- WSU (2021) *Institutional Animal Care and Use Committee*. Obtenido de Washington State University. <https://iacuc.wsu.edu/zoonoses-bats/#:~:text=Imported%20bats%20can%20also%20carry,in%20humans%20and%20other%20animals>.
- Xia Lab (2022) *MicrobiomeAnalyst- comprehensive statistical, visual and meta-analysis of microbiome data*. Obtenido de <https://www.microbiomeanalyst.ca/MicrobiomeAnalyst/home.xhtml>
- Tang, X., Shang, J., Sun, Y. (2022) *RdRp-based sensitive taxonomic classification of RNA viruses for metagenomic data*, *Briefings in Bioinformatics*, Volume 23, Issue 2, <https://doi.org/10.1093/bib/bbac011>
- Yeoh YK. (2021) *Removing Host-derived DNA Sequences from Microbial Metagenomes via Mapping to Reference Genomes*. *Methods Mol Biol*. 2232:147-153. Obtenido de [doi:10.1007/978-1-0716-1040-4_13](https://doi.org/10.1007/978-1-0716-1040-4_13). PMID: 33161546

ANEXO 1: SCRIPT DE NANO PLOT

- Control + T
- conda activate nanoplot
- NanoPlot --fastq Pathhasta/input_porechop.fasq --readtype 1D -t 4 --title "nanoplot+nombremuestra" -o Nanoplot_results_nombremuestra
- NanoPlot --fastq 'ubicación-muestra.fastq' --readtype 1D -t 4 --title "nombremuestra+Nanoplot" -o Nanoplot_nombremuestra/

ANEXO 2: SCRIPT REMOCIÓN DEL HOSPEDERO

- mkdir nombre_archivo
- cd nombre_archivo
- # -t = specify number of threads to use
- bwa mem 'output' -t 8 '/muestra.fastq' > muestra.sam
- samtools view -@ 8 -bS muestra_mapped_host.sam -o Sample_mapped_host.bam
- samtools sort -@ 8 Sample_mapped_host.bam -o Sample_mapped_host_sorted_location.bam
- samtools sort -n -@ 24 sample_mapped_host.bam -o sample_mapped_host_sorted_name.bam
- samtools index -@ 24 sample_mapped_host_sorted_location.bam
- samtools idxstats sample_mapped_host_sorted_location.bam > sample_mapped_host_sorted_location.idxstats

ANEXO 3: TABLA METADATA Y OTU PARA MICROBIOME ANALYST

- Tabla OTU [ESPECIES-TABLA2.0 \(1\).txt](#)
- Metadata [METADATA8 \(1\).txt](#)