PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INGENIERÍA

MAESTRÍA EN BIOLOGÍA COMPUTACIONAL

COMPARACIÓN DE HERRAMIENTAS DE IDENTIFICACIÓN DE PLÁSMIDOS

TRABAJO PREVIO A LA OBTENCIÓN DEL TÍTULO DE MAGÍSTER EN BIOLOGÍA COMPUTACIONAL

PAULINA ISABEL TERÁN AMORES

2023

AGRADECIMIENTOS

- A mis padres y hermana por brindarme su comprensión y apoyo incondicional.
- A mi esposo, Francis, por siempre motivarme a ser una mejor persona y profesional.
- Al PhD. Francisco Flores por su generosa guía y apoyo en la ejecución del presente trabajo de titulación.

RESUMEN

El interés en el mundo microbiano y la emergencia de nuevas herramientas de biología molecular y bioinformática ha favorecido el desarrollo de la metagenómica y la plasmidómica. Estas tecnologías han permitido el descubrimiento de microorganismos no cultivables y secuencias pertenecientes a proteínas de interés con potenciales aplicaciones biotecnológicas. Algunas de estas secuencias que confieren ventajas evolutivas a los microorganismos están presentes fuera del cromosoma, en los plásmidos. Consecuentemente, es importante contar con estudios comparativos de las herramientas existentes para la predicción de los plásmidos en datos metagenómicos. El presente trabajo de titulación pretende explorar y comparar tres herramientas disponibles para la clasificación de secuencias metagenómicas: PlasFlow, SOURCEFINDER Y PlasmidHunter, las cuales se emplearon para la clasificación de contigs resultantes del ensamblado de metagenoma de planta *in vitro* y tumor de corona de planta adulta de *Rosa* sp. En el conjunto de datos analizado, PlasFlow reportó el mayor número de fragmentos de plásmidos (5041), seguido de SOURCEFINDER (2647) y PlasmidHunter (148). En contraste, PlasmidHunter reportó mayores precisiones (hasta 100 %) y exactitudes (99.86 %) mientras que PlasFlow tuvo las mayores sensibilidades (33.28 %). Las diferencias pueden atribuirse a la arquitectura de las herramientas y a los datos que fueron empleados para su entrenamiento.

Palabras clave: Metagenómica, Plásmidos, Herramientas de Clasificación, Machine Learning

ABSTRACT

The interest towards the microbial world and the availability of new molecular biology and bioinformatic tools has promoted the growth of metagenomics and plasmidomics. These technologies have allowed the discovery of unculturable microorganisms and sequences coding for proteins with potential applications in biotechnology. Some of these sequences confer evolutive advantages to their hosts and are located outside the chromosome, in the plasmids. Consequently, it is important to have comparative studies of the available tools for plasmid prediction in metagenomic datasets. The aim of this work was to explore and compare three tools available for metagenomic sequence classification: PlasFlow, SOURCEFINDER and PlasmidHunter, which were used to classify contigs obtained from *in vitro* plants and crown gall from *Rosa* sp. PlasFlow reported the highest amount of plasmid fragments (5041), followed by SOURCEFINDER (2647) and PlasmidHunter (148). On the other hand, PlasmidHunter reported higher precision values (up to 100 %) and accuracies (99.86 %) while PlasFlow had the highest sensitivy (33.28 %). Differences between the three tools might be related to their architecture as well as to the data used for their training.

Key Words: Metagenomics, Plasmids, Classification tools, Machine Learning

ÍNDICE

1. INTRODUCCIÓN

1.1 ANTECEDENTES

El interés en el mundo microbiano y la emergencia de nuevas herramientas de biología molecular y bioinformática ha potenciado el desarrollo de la metagenómica y la plasmidómica (Lapidus & Korobeynikov, 2021). Múltiples algoritmos han sido desarrollados para identificar plásmidos de ensamblajes de metagenómica en los últimos años (Pu & Shamir, 2022). Algunos de ellos están optimizados para la detección de plásmidos de importancia clínica como PlasmidFinder/pMLST. Otros como PLACNET y el clasificador Kraken permiten distinguir contigs de origen plasmídico de aquellos de origen cromosómico (Andreopoulos et al., 2022).

PlasmidSPAdes, cBAR, PlasFlow, Recycler y PlasmidSeeker están optimizados para identificar contigs de plásmidos putativos en ensamblajes genómicos conforme a la topología y cobertura de lectura de un gráfico de ensamblaje o la composición de ADN de los contigs ensamblaje o ADN sin ensamblar. Sin embargo, la mayoría de estas herramientas no han sido probadas para detectar plásmidos validados experimentalmente (Andreopoulos et al., 2022).

1.2 JUSTIFICACIÓN

Los metagenomas representan una fuente potencial de material genético de interés para las ciencias biológicas y aplicadas. Gracias a los avances de secuenciación y creación de repositorios como MGnify, que a la fecha cuenta con alrededor de 33827 metagenomas (Mitchell et al., 2020), hoy más que nunca se dispone de información para planteamiento de investigación *in silico*.

Actualmente, herramientas como PlasForest y MOB-recon identifican pocos contigs como plásmidos en datos de metagenómica (0.18% y 0.24%) mientras que herramientas como PlasFlow, PlasClass y PPR-Meta predicen hasta 25.4%de contigs como plásmidos. No obstante, entre métodos se ha observado que coinciden en un 36.8% a 48.4% en sus predicciones (Pradier et al., 2021).

Dadas las limitaciones de cada técnica de identificación de plásmidos específicamente en datos de metagenómica y, ante la falta de una comparación sistemática de los mismos (Krawczyk et al., 2018), es crucial contar con un estudio comparativo que evalúe la calidad de las predicciones y exponga las ventajas y desventajas de emplear las herramientas.

1.3 PLANTEAMIENTO DEL PROBLEMA

A la fecha, miles de plásmidos han sido secuenciados y ensamblados directamente a partir de bacterias aisladas. Sin embargo, el ensamblaje a partir de muestras de secuencias ambientales, con microorganismos difíciles de cultivar en laboratorio, es más complejo (Gilbert & Dupont, 2011; Pellow et al., 2021). Esto se debe a que los plásmidos representan una pequeña fracción del ADN total de la muestra, la similitud de secuencias con los genomas bacterianos, topología circular, la naturaleza móvil de los plásmidos y el grado de complejidad de los gráficos de ensamblaje de metagenomas (Arredondo-Alonso et al., 2017; Rozov et al., 2017; Pellow et al., 2021).

Consecuentemente, se han desarrollado herramientas optimizadas para la identificación de plásmidos a nivel de genomas y recientemente, de metagenomas. Estas herramientas pueden verse limitadas en cuanto a precisión y sensibilidad, lo cual puede no afectar determinadas aplicaciones. Por ejemplo, los métodos basados en k-méros identifican secuencias de plásmidos en metagenomas y las herramientas basadas en homología funcionan cuando se trabaja con plásmidos de interés. No obstante, no se ha descrito una herramienta para describir con alta precisión y sensibilidad un gran conjunto de secuencias como plásmido o cromosoma (Pradier et al., 2021). Por otra parte, herramientas como plasmidSPAdes y Recycler pueden detectar un buen nivel de nuevos plásmidos, pero reportan una cantidad significativa de falsos positivos (Lapidus & Korobeynikov, 2021).

Cabe mencionar que las predicciones también pueden verse sesgadas al estar optimizadas para un género específico, por lo que no podrían emplearse para conjuntos de datos de metagenómica (Van der Graaf-Van Bloois et al., 2021).

1.4 OBJETIVOS

1.4.1 OBJETIVO GENERAL

Comparar herramientas de identificación de plásmidos en metagenomas

1.4.2 OBJETIVOS ESPECÍFICOS

- Recuperar datos de secuencias metagenómicas de calidad
- Evaluar las herramientas PlasmidHunter, SOURCEFinder y PlasFlow en los metagenomas seleccionados
- Comparar los resultados generados por las herramientas en términos de métricas de evaluación

2. MARCO TEÓRICO

2.1 PLÁSMIDOS

2.1.1 DEFINICIÓN E IMPORTANCIA

Los plásmidos son moléculas circulares de ADN extracromosomal en su mayoría, aunque, ocasionalmente, pueden ser lineares o estar conformados por ARN. Su tamaño oscila entre los 5 y 500 kbp (Ankita et al., 2019). Pueden encontrarse en una o múltiples copias en la célula y son capaces de replicarse de forma autónoma y transmitirse entre células microbianas hospederas (Clark et al., 2019; Andreopoulos et al., 2022).

Se encuentran principalmente en bacterias y, se estima que alrededor del 50% de bacterias encontradas en la naturaleza contienen uno o más plásmidos. También se han reportado en organismos como levaduras y hongos (Clark et al., 2019).

A diferencia de los cromosomas, los plásmidos no son considerados una parte permanente del genoma ya que pueden estar presentes en células de diferentes especies y movilizarse entre hospederos. A su vez, los plásmidos no se encuentran en todos los individuos de la especie hospedera (Clark et al., 2019).

Si bien portan información genética que puede expresarse, los plásmidos son dispensables para el crecimiento y división celular bajo condiciones normales (Noel, 2009; Clark et al., 2019). En contraste, los plásmidos son elementos genéticos móviles que facilitan la evolución y adaptación de sus portadores ante situaciones ambientales

cambiantes. Esto debido a que portan genes accesorios que confieren caracteres favorecedores al hospedero microbiano como la resistencia a antibióticos, virulencia, defensa ante bacteriófagos, tolerancia a metales pesados, o rutas catabólicas exclusivas (Andreopoulos et al., 2022). En el caso de los microorganismos asociados a plantas, los plásmidos que contienen genes de nodulación de los rizobios son clave en la interacción planta-bacteria (Schierstaedt et al., 2019; Andreopoulos et al., 2022). En la Figura 1, se resumen y ejemplifican los caracteres descritos en plásmidos.

Resistencia y defensa	• Resistencia a metales pesados como níquel, cobalto, plomo, cadmio, cromo, bismuto, antimonio, zinc, cobre, plata y mercurio • Resistencia a antibióticos, incluyendo aminoglucósidos, β-lactamas, cloranfenicol, sulfonamidas, trimethoprim, ácido fusídico, tetraciclinas, macrolidas, fosfomicina y quinolinas • Resistencia a aniones tóxicos como arsenato, arsenito, borato, cromato, selenato, telurito. • Resistencia a agentes intercalantes como acridinas y bromuro de etidio • Protección contra radiación UV y rayos X • Resistencia a bacteriófagos
Agresividad y virulencia	• Producción de bacteriocinas y antibióticos • Formación de tumores en cuello y raíces (*Agrobacterium*) • Formación de nódulos en raíces de legumbres (*Rhizobium*) • Síntesis de toxinas • Evasión del sistema inmunológico
Rutas metabólicas	• Consumo de azúcares como lactosa, rafinosa y sacarosa • Degradación de hidrocarburos alifáticos y aromáticos, así como sus derivados • Degradación de hidrocarburos halogenados como bifenilos policlorados • Degradación de proteínas • Síntesis de sulfuro de hidrógeno • Denitrificación (*Alcaligenes*) • Síntesis de pigmentos (*Erwinia*)
Otras	• Transporte de citrato (*E. coli*) • Transporte de hierro • Producción de vacuola de gas (*Halobacterium*)

Figura 1. Resumen de propiedades conferidas por plásmidos (Adaptado de Clark et al., 2019)

2.2 METAGENÓMICA

2.2.1 DEFINICIÓN E IMPORTANCIA

La metagenómica es el estudio de los genomas de comunidades biológicas presentes en un hábitat determinado sin requerir del cultivo, aislamiento o identificación de microorganismos (Fricke et al., 2011; Raza & Shahid, 2020). El término “meta” deriva de “metaanálisis”, el cual es el proceso de combinar análisis separados estadísticamente. A diferencia de la genómica, la metagenómica se enfoca en múltiples organismos, entidades que contienen material genético (p.ej. virus, viroides, plásmidos) y ADN libre (Clark et al., 2019).

Mediante la metagenómica es posible descifrar la composición taxonómica y funcional de las comunidades microbianas de entornos ambientales, agrícolas y clínicos. En la Figura 2, se resumen los principales aportes de los estudios de metagenómica.

Figura 2. Aplicaciones principales de la metagenómica

2.2.2 PROTOCOLO

En la metagenómica el ADN genómico total es extraído de las muestras y este es secuenciado con tecnología Illumina y método shotgun. La asignación de los fragmentos resultantes de ADN, lecturas individuales o contigs ensamblados a grupos taxonómicos o secuencias conocidas se ejecuta mediante herramientas bioinformáticas (Fricke et al., 2011). En la Figura 3 se resumen los pasos principales de un proyecto típico de metagenómica.

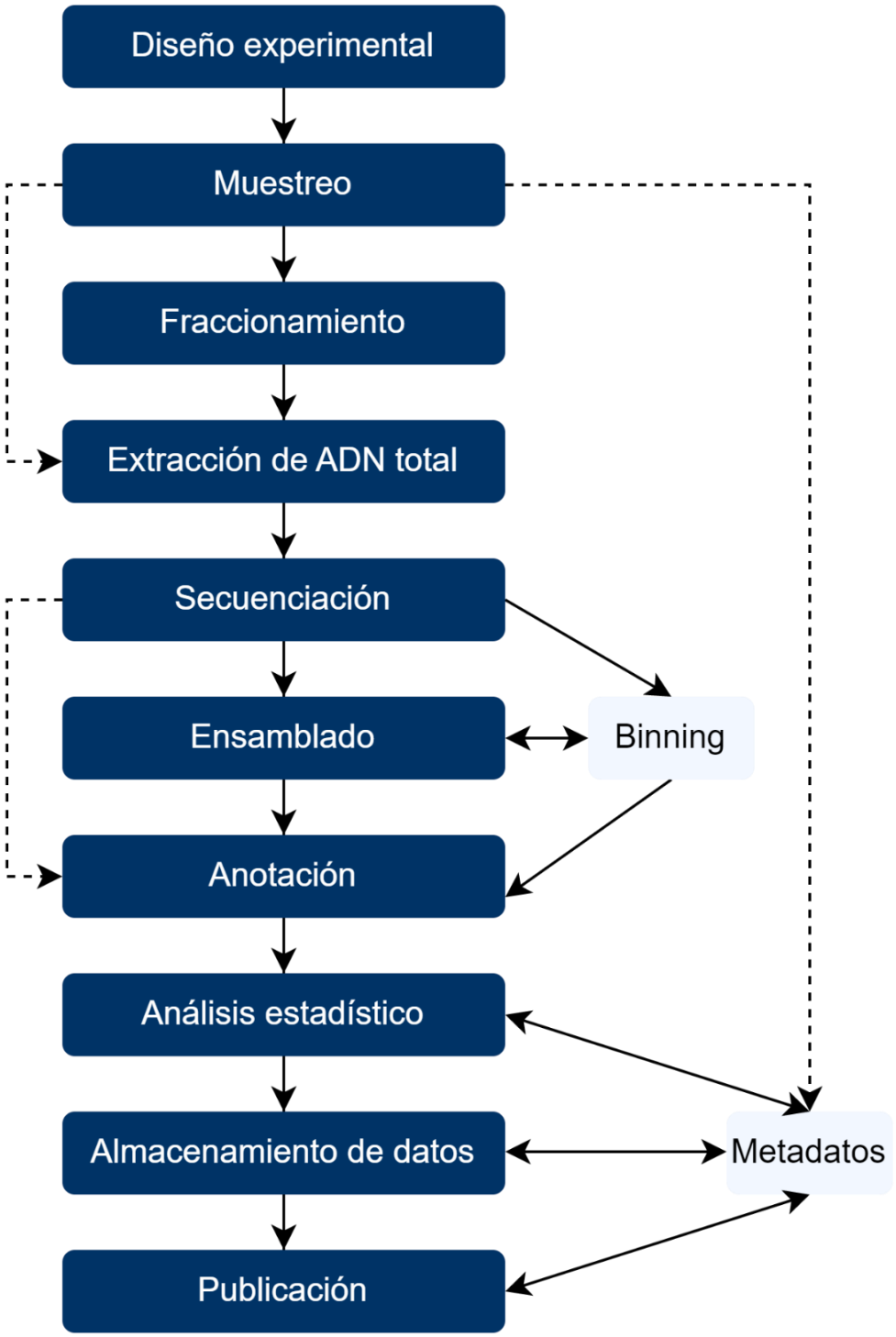

Figura 3. Diagrama de flujo de un proyecto de metagenómica (Adaptado de Thomas et al., 2012)

2.2.2.1 COLECTA DE MUESTRA Y CONSTRUCCIÓN DE LIBRERÍAS

La colecta, preparación y almacenamiento de la muestra es crucial para el proyecto de metagenómica ya que se deben garantizar cantidades óptimas de material genético representativo de las células presentes en la muestra. Es importante evitar los ciclos de congelamiento y descongelamiento (Shuikan et al., 2020). Además, en función del microbioma, se selecciona un protocolo adecuado que garantice la compatibilidad del material genético con los procesos posteriores (Thomas et al., 2012). Cuando la muestra está asociada a un hospedero, es necesario un fraccionamiento o una lisis selectiva (Reddy & Surekha, 2016).

El ADN extraído se emplea para construir las librerías. Generalmente esto se logra mediante la adición de adaptadores específicos a los extremos de los fragmentos de ADN con la finalidad de manejar el pool de muestras y posteriormente conectarlas a la muestra original. La construcción de librerías se logra mediante dos estrategias. La primera, llamada mate-pair emplea fragmentos grandes y la segunda son librerías paired-end con insertos cortos. Los fragmentos cortos suelen clonarse en plásmidos mientras que los fragmentos de hasta 40 Kpb en fósmidos o cósmidos. Si los fragmentos son mayores a 40 Kpb se usan cromosomas bacterianos artificiales (BACs). Finalmente, se remueven los adaptadores para prevenir el ruido de secuenciación (Shuikan et al., 2020). Cabe mencionar que las tecnologías de secuenciación de nueva generación ya no

precisan de la clonación para la preparación de librerías, reduciendo el riesgo de contaminación cruzada de ADN (Barooah et al., 2021).

2.2.2.2 TECNOLOGÍAS DE SECUENCIACIÓN

En las últimas décadas, la secuenciación shotgun de metagenomas ha migrado de la tecnología Sanger a la secuenciación de alto rendimiento, High Throughput Sequencing. La secuenciación tipo Sanger aún sigue siendo un estándar considerando su baja tasa de error, longitud de lecturas (> 700 pb) y tamaño de inserto (>30 Kb). No obstante, la secuenciación Sanger es costosa y laboriosa, además de que los genes tóxicos para el hospedero no son clonados exitosamente (Thomas et al., 2012).

En contraste, la secuenciación de nueva generación ha permitido generar lecturas de mayor longitud y reducir la necesidad de clonación. Estas ventajas han impulsado el desarrollo de múltiples plataformas como Roche 454, Illumina®, SOLiD de Applied Biosystems e Ion Torrent. Dichas tecnologías de secuenciación emplean sensores ópticos o semiconductores capaces de detectar señales de luminiscencia o fluorescencia emitidas durante la incorporación de bases en la síntesis de una nueva cadena (Barooah et al., 2021).

La tecnología NGS también tiene sus desventajas, entre ellas un menor tamaño de lectura, sesgo de PCR y problemas de detección durante la señalización basada en fluorescencia. Ante lo cual se han desarrollado nuevas plataformas de tercera generación

(Third Generation Sequencing, TGS) o de secuenciación de una sola molécula (Single-molecule-sequencing, SMS) que omiten la PCR previa a la secuenciación y que en general ~~y~~ capturan la señal en tiempo real mediante el monitoreo de reacciones enzimáticas. Entre estas tecnologías se incluye al secuenciador PacBio technology/SMRT y Oxford Nanopore (Shuikan et al., 2020; Barooah et al., 2021).

2.2.2.3 ENSAMBLAJE

En proyectos destinados a la recuperación de genomas de organismos no cultivados o de secuencias codificantes complejas, se ensamblan los fragmentos de lecturas cortas para obtener contigs genómicos más extensos. Para el ensamblaje de muestras metagenómicas se pueden emplear referencias o ensamblar *de novo* (Thomas et al., 2012).

El ensamblaje basado en referencias puede lograrse con paquetes como Newbler, AMOS o MIRA los cuales contienen algoritmos rápidos y eficientes en términos de memoria. Por su parte, el ensamblaje *de novo* requiere de mayores recursos computacionales. Consecuentemente se han creado herramientas basadas en gráficos de Bruijn para manipular grandes cantidades de datos como Velvet o SOAP (Thomas et al., 2012).

Considerando que la mayoría de las comunidades microbianas albergan una gran diversidad de cepas que pueden pasar por desapercibido para los algoritmos de

ensamblaje, se han desarrollado ensambladores como MetaVelvet. Este permite ensamblar lecturas cortas de múltiples especies al identificar subgráficos que representan genomas relacionados en el gráfico de Bruijn principal (Namiki et al., 2012)

2.2.2.4 BINNING

El binning es el proceso de ordenar las secuencias de ADN en diferentes bins que representan grupos taxonómicos (Wickramarachchi & Lin, 2022). Para el binning se han desarrollado múltiples algoritmos que emplean dos tipos de información de las secuencias: (i) la composición conservada de nucleótidos (p. ej. Contenido de GC o distribución de k-méros) y (ii) la similitud de las secuencias con las referencias. En el primer caso se usan algoritmos como Phylopythia, S-GSOM, PCAHIER y TACAO, mientras que para similitudes se usa IMG/M, MG-RAST, MEGAN, CARMA, SOrt-ITEMS y MetaPhyler. A la vez hay algoritmos de binning que consideran tanto la composición como la similitud como PhymmBL y MetaCluster (Thomas et al., 2012).

2.2.2.5 ANÁLISIS DE DATOS DE METAGENÓMICA

Se han desarrollado múltiples herramientas para el análisis de metagenomas a nivel molecular, de especie y cepas. Uno de los enfoques para filogenia y taxonomía microbiana es el metabarcoding, que puede lograrse mediante el análisis del gen del ARN ribosomal 16S [RR1](Shuikan et al., 2020). Este gen es utilizado como marcador dada su universalidad, alto grado de conservación y tamaño (Brooks et al., 2015). Para los análisis de ARNr 16S existen diversas herramientas como QIIME, MOTHUR, DADA2,

UPARSE Y MED mientras que, para la identificación de especies en datos metagenómicos, así como su abundancia, se puede usar MetaPhlAn2, Kraken, CLARK, FOCUS, SUPERFOCUS, y MG-RAST (Shuikan et al., 2020).

El análisis de metagenomas también abarca la anotación de genes predichos mediante búsqueda comparativa en bases de datos como SWISSPROT, NCBI o KEGG para proteínas. También se identifican elementos como ARNt, ARNr, péptidos señal, regiones transmembrana, repeticiones CRISPR y localización subcelular. Las anotaciones obtenidas se usan para minado de datos funcionales (Sugitha et al., 2020).

2.2.2.5 ENSAMBLAJE E IDENTIFICACIÓN DE PLÁSMIDOS

Considerando que los plásmidos intercambian material genético con los cromosomas del hospedero y pueden estar presentes de forma lineal o circular, así como tener diferentes tamaños y contenido génico es complejo definirlos computacionalmente para diferenciarlos de los cromosomas. Paralelamente, el ensamblaje de plásmidos se complica por la presencia de repeticiones múltiples intraplasmídicas, interplasmídicas y compartidas entre plásmidos y cromosomas (Antipov et al., 2019). A fin de solventar esta problemática se han implementado herramientas basadas en la diferencia entre las coberturas de los plásmidos y cromosomas como plasmidSPAdes y Recycler. No obstante, estas herramientas pueden reportar múltiples falsos positivos especialmente cuando la cobertura del cromosoma no es uniforme. Antipov et al. (2019) crearon metaplasmidSPAdes, un algoritmo que (i) extrae subgráficos de forma iterativa mediante

el incremento gradual del gráfico de ensamblado del metagenoma, (ii) encuentra plásmidos putativos como ciclos cubiertos uniformemente en los subgráficos y (iii) verifica los plásmidos con la herramienta plasmidVerify. SCAPP es otro algoritmo que acepta como entrada un gráfico de ensamblaje y detecta secuencias cíclicas con ensamblajes de plásmidos (Pellow et al., 2021).

2.2.2.5.1 PLASMIDHUNTER

PlasmidHunter es una herramienta que usa Machine Learning para predecir secuencias plasmídicas a partir de ensamblados basado en el perfil de contenido genético. No considera los datos crudos de secuenciación, topología de secuencia, cobertura o gráfico de ensamblaje por lo que puede trabajar con un archivo de secuencia ensamblada producida por cualquier algoritmo (Tian & Imanian, 2023).

Para la construcción de PlasmidHunter se usaron 25,898 genomas y 0.96 millones de proteínas representativas indexadas como base de datos. De los genomas completos, 15 000 cromosomas y 15 000 plásmidos se usaron para modelamiento y validación (Tian & Imanian, 2023).

Las secuencias de entrada se emplean para predecir secuencias codificantes en cada contig. Estas secuencias son traducidas y usadas para alineamiento con Diamond usando una base de datos personalizada. Seguidamente, el perfil de contenido genético

es filtrado para retener las propiedades génicas usadas en la etapa de modelado y predecir el origen de los contigs con el paquete sklearn de Python. Mediante este esquema de trabajo, PlasmidHunter ha reportado exactitud de hasta el 96.7 % y rapidez en la obtención de resultados (Tian & Imanian, 2023).

2.2.2.5.2 SOURCEFINDER

SourceFinder es un clasificador desarrollado usando Random Forest para identificar el origen de las secuencias, sean cromosomas, plásmidos o bacteriófagos. SourceFinder fue entrenado con una colección de 23,211 secuencias de cromosomas, plásmidos y bacteriófagos de cientos de especies bacterianas. Se emplearon 5-méros y fragmentos de 5 kb con diez rondas de submuestreo (Aytan-Aktug et al., 2022).

SourceFinder está disponible como línea de comandos, requiriendo de paquetes de Anaconda y KMC. También cuenta con un servicio en línea en https://cge.food.dtu.dk/services/SourceFinder/ que corre en software backend. Cuenta con una opción para escoger la cantidad de veces en la que se muestrea un contig y el tiempo de corrida es de 1 minuto por genoma (Aytan-Aktug et al., 2022).

2.2.2.5.3 PLASFLOW

PlasFlow es un conjunto de scripts para la identificación de plásmidos en contigs de metagenómica sin conocimiento previo de la composición taxonómica o funcional de

las muestras. A la vez, puede reconocer secuencias de plásmidos circulares y lineales, así como hacer una clasificación taxonómica inicial de las secuencias. PlasFlow es un método de clasificación basado en k-méros que se construyó empleando marcajes genómicos de 9,565 cromosomas y plásmidos bacterianos, entrenando una red neuronal profunda para separar las secuencias cromosómicas y plasmídicas de diferentes phyla. En datos de prueba, PlasFlow ha demostrado hasta un 96% de exactitud de clasificación y ha sido probado en datos metagenómicos reales (Krawczyk et al., 2018). La herramienta puede descargarse de conda y pypi y está disponible en línea en la plataforma Galaxy (https://usegalaxy.org/).

PlasFlow genera 4 archivos de salida de los cuales el más importante es un archivo tabular que contiene las predicciones y consta de múltiples columnas detalladas en la Tabla 1 (Krawczyk, 2021). Los demás archivos contienen las secuencias filtradas según su clasificación en formato .fasta. Es decir, se genera un archivo .fasta para los contigs clasificados como cromosomas, otro para contigs clasificados como plásmidos y un último para las que no pudieron ser clasificadas.

Tabla 1. Columnas presentes en el archivo de salida tabular de PlasFlow (Krawczyk, 2021)

Columna	Contenido
contig_id	Identificador interno de la secuencia usado para la clasificación
contig_name	Nombre de un contig usado en la clasificación
contig_length	Longitud de la secuencia clasificada
id	Identificador interno de la etiqueta generada
label	Clasificación
…	Columnas que muestran la probabilidad de asignación a cada clase posible, incluyendo: chromosome.Acidobacteria, chromosome.Actinobacteria, chromosome.Bacteroidetes, chromosome.Chlamydiae, chromosome.Chlorobi, chromosome.Chloroflexi, chromosome.Cyanobacteria, chromosome.DeinococcusThermus, chromosome.Firmicutes, chromosome.Fusobacteria, chromosome.Nitrospirae, chromosome.other, chromosome.Planctomycetes, chromosome.Proteobacteria, chromosome.Spirochaetes, chromosome.Tenericutes, chromosome.Thermotogae, chromosome.Verrucomicrobia, plasmid.Actinobacteria, plasmid.Bacteroidetes, plasmid.Chlamydiae, plasmid.Cyanobacteria, plasmid.DeinococcusThermus, plasmid.Firmicutes, plasmid.Fusobacteria, plasmid.other, plasmid.Proteobacteria, plasmid.Spirochaetes

3. METODOLOGÍA

En el diagrama de flujo a continuación se resume la metodología empleada en el presente trabajo.

Figura 4. Diagrama de flujo de la metodología empleada

3.1 OBTENCIÓN DE SECUENCIAS

Las muestras de tumor de corona de *Rosa sp*. (de especímenes *in vitro* y adulto) fueron proporcionadas por el laboratorio IDgen y enviadas a la casa comercial Novogene Co. Ltd. Brevemente, el ADN genómico fue divido en fragmentos cortos al azar. Las terminaciones de los fragmentos fueron reparadas y se les agregó una cola de adenina (A-tailing). Seguidamente, se ligaron con adaptadores de Illumina para la amplificación por PCR, selección basada en tamaño y purificación.

Las librerías fueron cuantificadas mediante fluorometría (Qubit) y PCR en tiempo real. Se empleó el sistema Bioanalyzer para la detección de la distribución de tamaños. Las librerías cuantificadas fueron agrupadas y secuenciadas en la plataforma Illumina. En la Figura 5, se observa la distribución de la calidad de las lecturas, la distribución del porcentaje de error, el contenido de nucleótidos y composición de las lecturas.

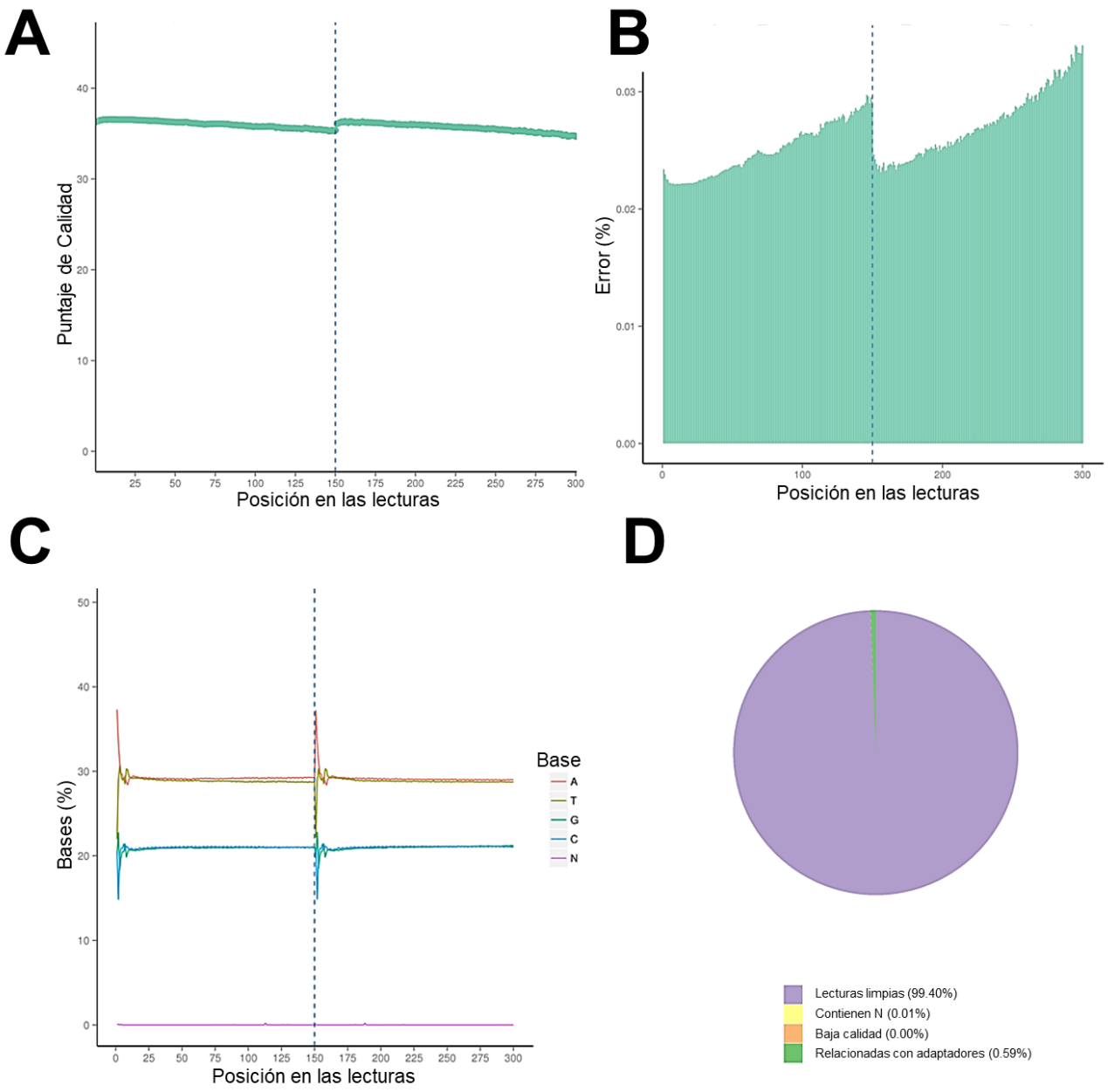

Figura 5. A. Distribución de la calidad de secuencias. El eje horizontal representa la posición de las bases y el eje vertical, la calidad de secuenciación. **B.** Distribución de la tasa de error (%) en relación con la posición de las bases en las lecturas. **C.** Contenido de A, T, G, C y N. **D.** Composición de las lecturas.

3.2 ENSAMBLADO Y BINNING

Se eliminaron las secuencias del hospedero en Linux, en el clúster de CEDIA. Una vez limpias las secuencias, estas fueron ensambladas y clasificadas taxonómicamente empleando la plataforma KBase (Arkin et al., 2018). Los bins procesados se muestran en la Tabla 2.

Tabla 2. Resumen de los bins analizados

Muestra	BINS	Contigs	GC (%)	Longitud (pb)
In vitro	BIN001.fasta.assembly	170	67.73	6,366,530
	BIN002.fasta.assembly	10,679	38.25	19,299,250
	BIN003.fasta.assembly	1,102	52.29	7,597,847
	BIN004.fasta.assembly	1047	66.49	4,383,949
Planta adulta	Bin.002.M402.Pantoea.agglomerans[FJ2]	753	54.46	6,181,523
	Bin.004.M402.Sphingomonas	1561	62.72	4,704,106

3.3 IDENTIFICACIÓN DE PLÁSMIDOS

3.3.1 PLASFLOW

Conforme a las recomendaciones del desarrollador, se filtraron las contigs por longitud, trabajando solo con aquellos mayores a 1000 pb. Este filtrado también puede mejorar las predicciones y evitar un alto consumo de la memoria RAM (Krawczyk, 2021). Para el filtrado se empleó la herramienta 'Filter Assembled Contigs by Length - v1.2.0' disponible en la plataforma KBase (Arkin et al., 2018).

Los archivos filtrados fueron analizados con la herramienta PlasFlow v.1.0 disponible en la plataforma Galaxy (https://usegalaxy.org/) con un umbral de 0.8 (Valentino et al., 2022) como se muestra en la Figura 6.

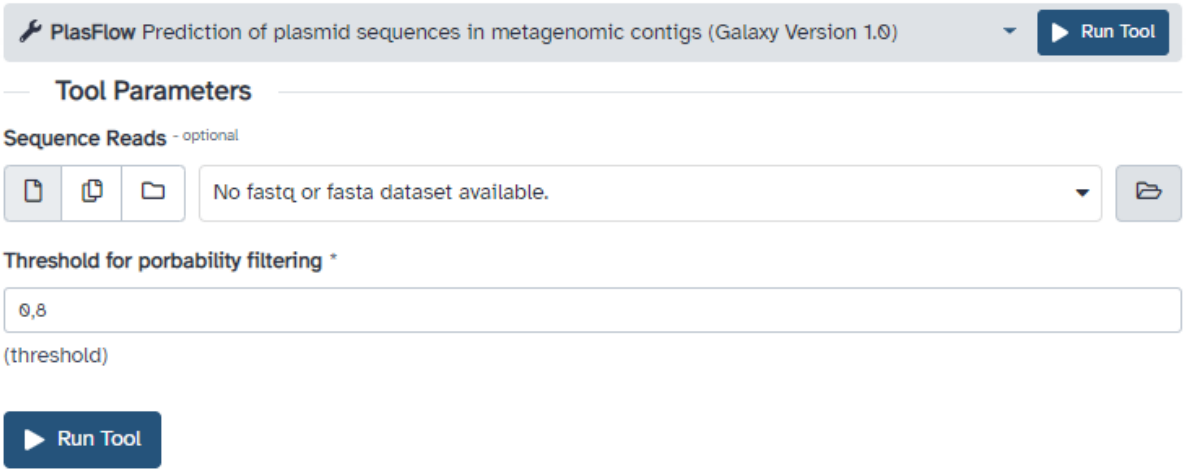

Figura 6. Herramienta PlasFlow disponible en el entorno Galaxy

Las secuencias de los plásmidos predichos se extrajeron directamente a un archivo de formato .fasta (Figura 7). Los nombres de los contigs se recuperaron con el comando:

```
grep ">" 'archivo.fasta'
```

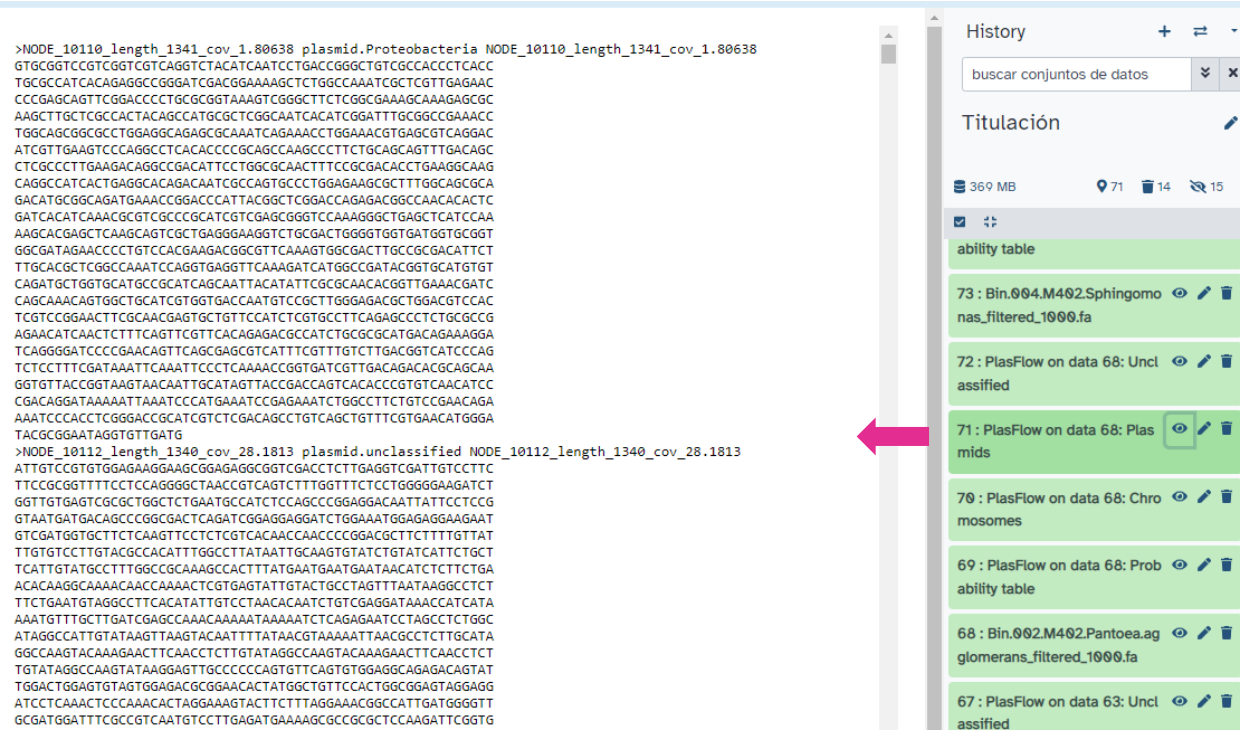

Figura 7. Archivo de salida .fasta con las secuencias de los plásmidos predichos por la herramienta PlasFlow

3.3.2 SOURCEFINDER

Se subieron las secuencias filtradas a la plataforma en línea de SourceFinder v 1.0 disponible en https://cge.food.dtu.dk/services/SourceFinder/ para su procesamiento. Se mantuvieron 5 rondas de muestreo por defecto (Aytan-Aktug et al., 2022) (Figura 8). En la salida generada (Figura 9) se filtraron los contigs clasificados como plásmidos.

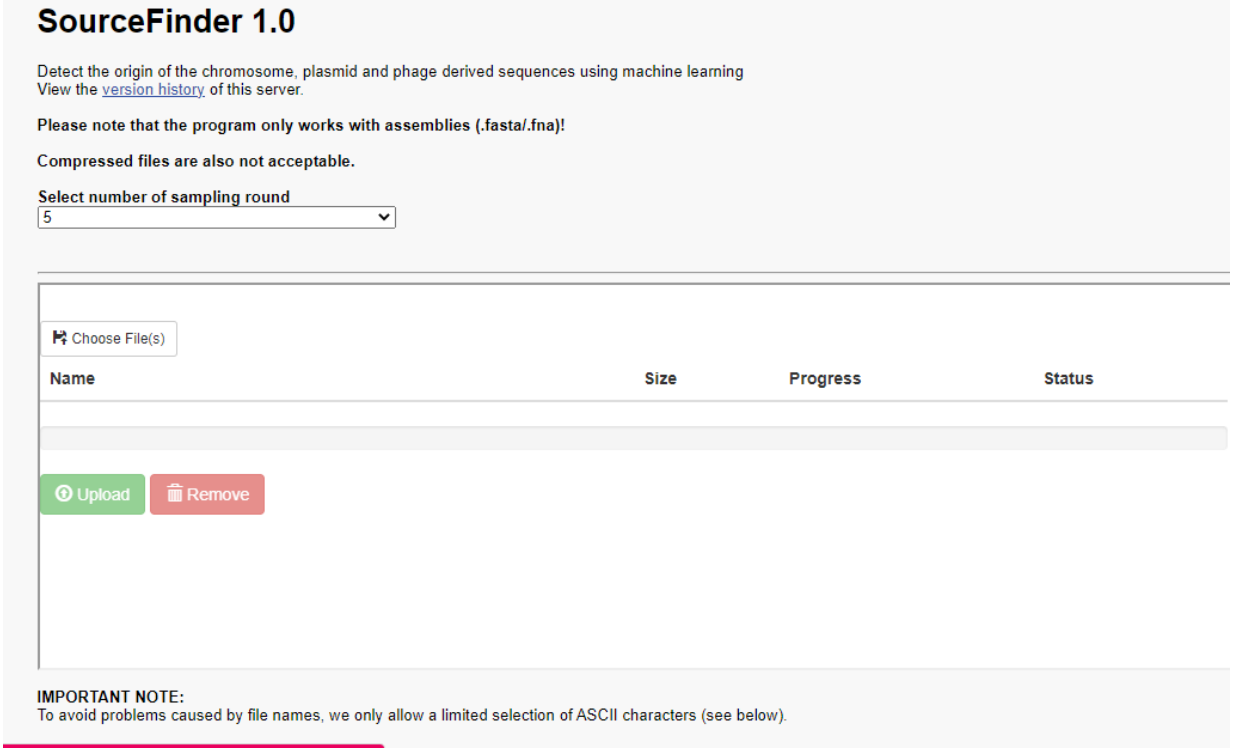

Figura 8. Herramienta SourceFinder con la configuración seleccionada para la clasificación de secuencias

SourceFinder-1.0 Server - Results

Results:

Entry	Predicted Host(s)
>NODE_10044_length_1224_cov_2.11292	Chromosome
>NODE_1010_length_3957_cov_16.2171	Chromosome
>NODE_10158_length_1215_cov_2.50948	Chromosome
>NODE_10184_length_1212_cov_44.6906	Chromosome
>NODE_101_length_18486_cov_15.3481	Chromosome
>NODE_10265_length_1207_cov_2.26128	Phage
>NODE_10286_length_1206_cov_2.14509	Phage
>NODE_10353_length_1201_cov_2.73473	Phage
>NODE_103_length_18085_cov_27.0618	Chromosome
>NODE_10446_length_1195_cov_1.66579	Chromosome
>NODE_1044_length_3892_cov_103.005	Chromosome
>NODE_10476_length_1193_cov_2.53866	Phage
>NODE_10651_length_1182_cov_2.00976	Chromosome
>NODE_10722_length_1177_cov_2.45455	Chromosome
>NODE_107_length_17179_cov_14.9661	Phage
>NODE_10819_length_1172_cov_1.93912	Phage
>NODE_10836_length_1171_cov_3.08244	Chromosome
>NODE_10888_length_1168_cov_2.79874	Chromosome
>NODE_11171_length_1150_cov_1.92329	Chromosome
>NODE_11298_length_1142_cov_2.20147	Phage
>NODE_11329_length_1140_cov_2.10138	Chromosome
>NODE_11330_length_1140_cov_1.90968	Chromosome
>NODE_11356_length_1138_cov_2.63066	Chromosome
>NODE_11415_length_1135_cov_2.28704	Chromosome
>NODE_11416_length_1135_cov_2.11481	Chromosome
>NODE_11539_length_1127_cov_42.9543	Chromosome
>NODE_115_length_15174_cov_26.417	Plasmid

Figura 9. Salida de herramienta SourceFinder

3.3.3 PLASMIDHUNTER

Se empleó la versión de PlasmidHunter v1.1 disponible en https://colab.research.google.com/github/tianrenmaogithub/PlasmidHunter/blob/main/PlasmidHunter.ipynb. Se ejecutó el notebook (Figura 10) instalando las dependencias necesarias (conda, pandas, scikit-learn==1.0.2, bioconda, Biopython, diamond, Prodigal) y PlasmidHunter a partir de su repositorio en GitHub (https://github.com/tianrenmaogithub/PlasmidHunter/raw/main/PlasmidHunter_1.1.zip).

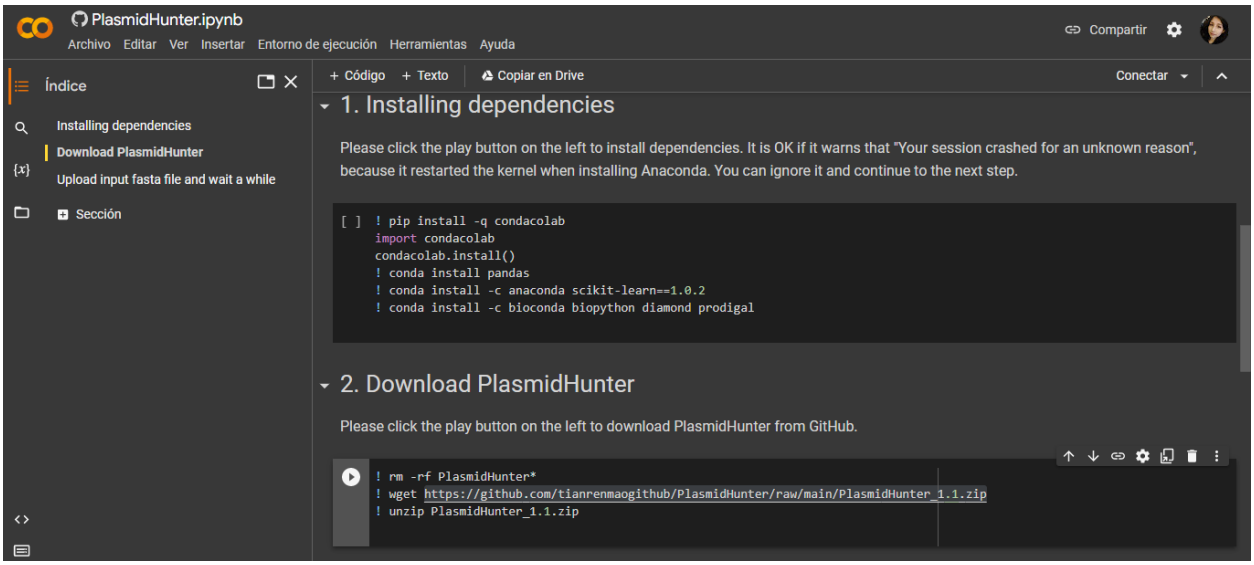

Figura 10. Cuaderno de Jupyter ejecutable de la herramienta PlasmidHunter

Seguidamente, se subieron los archivos filtrados para la predicción. El archivo de salida en formato .tsv se descargó para el análisis. Los contigs predichos como plásmidos fueron marcados en el archivo de salida con 1.0 y como cromosomas como 0.0 como se observa en la Figura 11.

PlasmidHunter

	Prediction (0: chromosome, 1: plasmid)	Probability of 0	Probability of 1
NODE_100_length_11596_cov_4.30101	0.0	1.0	0.0
NODE_101_length_11523_cov_4.24756	0.0	1.0	0.0
NODE_102_length_11520_cov_4.07928	0.0	1.0	0.0
NODE_103_length_11392_cov_4.14404	1.0	0.0	1.0

Figura 11. Formato de archivo de salida .tsv generado por la herramienta PlasmidHunter

3.4 BÚSQUEDA EN BLAST

Los contigs predichos como plásmidos fueron buscados en la base curada de plásmidos de Sistrom (2018) empleando la versión stand-alone de BLAST de nucleótidos (blastn) (Khezri et al., 2021).

El comando empleado para la búsqueda fue:

```
blastn -query 'archivo de entrada.fasta' -db 'base de datos' -out 'archivo de salida.txt'
```

3.5 ANÁLISIS DE DATOS

Se determinaron los parámetros de sensibilidad (R), exactitud (A) y precisión (P) empleando las fórmulas que se muestran a continuación:

$$R(\%) = \left(\frac{VP}{VP + FN}\right) \cdot 100$$

Ecuación 1. Cálculo de sensibilidad

$$A(\%) = \left(\frac{VP + VN}{VP + VN + FP + FN}\right) \cdot 100$$

Ecuación 2. Cálculo de la exactitud

$$P(\%) = \left(\frac{VP}{VP + FP}\right) \cdot 100$$

Ecuación 3. Cálculo de la precisión

- Los verdaderos positivos (VP) se consideraron como las secuencias que correspondieron a plásmidos en la base de datos con un porcentaje de similitud mayor a 70 %.
- Los falsos positivos (FP) fueron secuencias clasificadas como plásmidos pero que al buscarlas en la base curada de plásmidos no dieron ningún resultado (No hits found).
- Los verdaderos negativos (VN) se definieron como secuencias que no fueron clasificadas como plásmidos y que al buscarlas no produjeron resultados en la base de datos plasmídica.
- Los falsos negativos (FN) se definieron como secuencias que erróneamente no fueron clasificadas como plásmidos.

4. RESULTADOS

4.1 CLASIFICACIÓN TAXONÓMICA DE LAS LECTURAS

4.1.1 COMPARACIÓN CON BASE DE DATOS DE GENOMAS DE REFERENCIA

En la planta *in vitro* se identificaron lecturas de 11 especies bacterianas, principalmente de *Mycobacterium marseillense* (16. 029 %), seguidas por lecturas de *Pantoea rwandensis* (5.182 %), *Pantoea vagans* (4.573 %), *Mycobacterium intracellulare* (2.414 %) y *Stenotrophomonas maltophilia* (2.129 %) (Figura 12).

Respecto a la planta adulta, se identificaron lecturas de nueve especies, principalmente de *Pantoea agglomerans* (20.365 %), *Agrobacterium rubi* (2.375 %), *Serratia* sp. 1D1416 (0.780 %), *Sphingomonas insulae* (0.751 %) y *Microbacterium hydrocarbonoxydans* (0.619 %) (Figura 13).

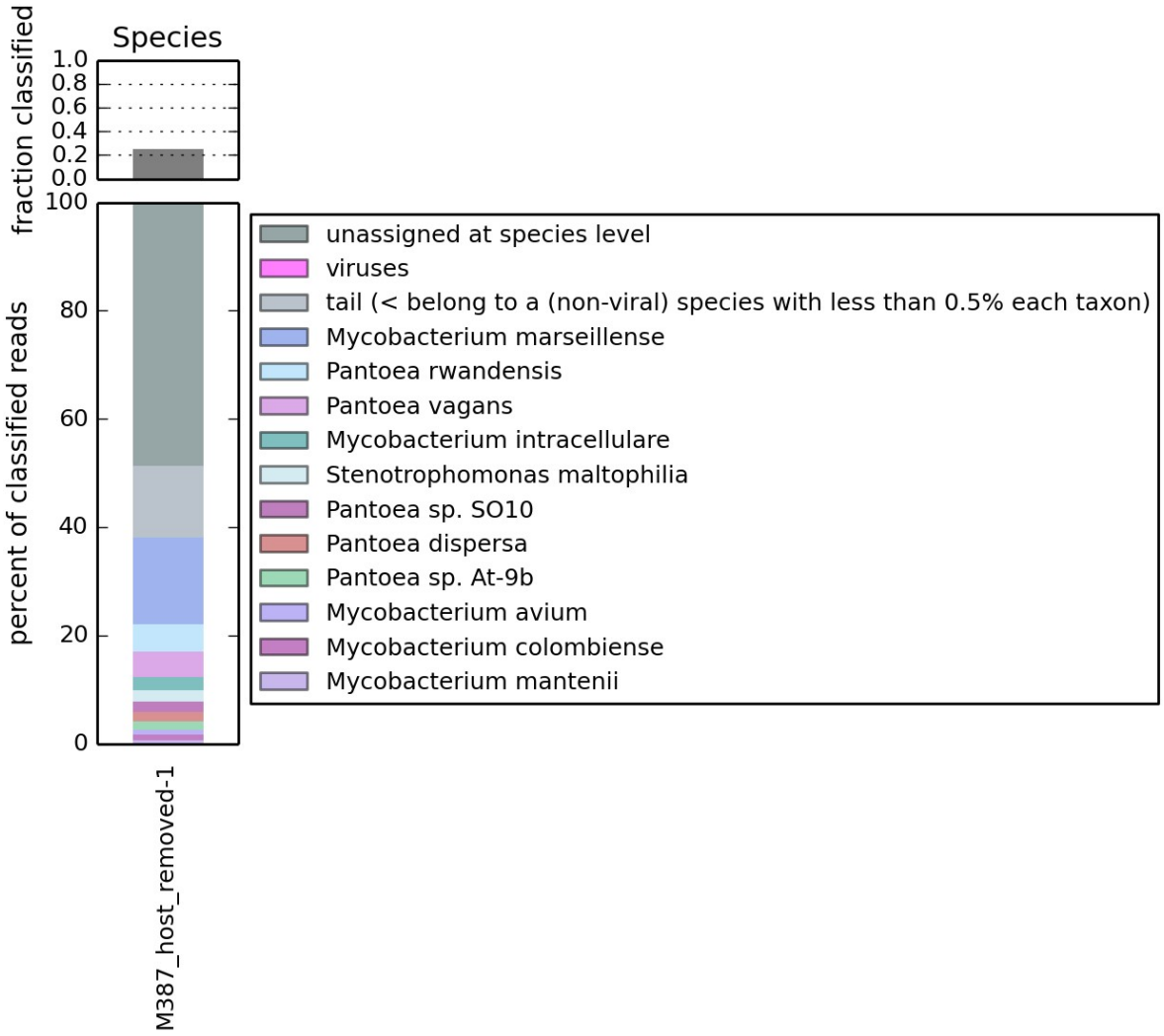

Figura 12. Clasificación taxonómica de las lecturas de la planta *in vitro* empleando la base de datos RefSeq Genomes (no Euks)

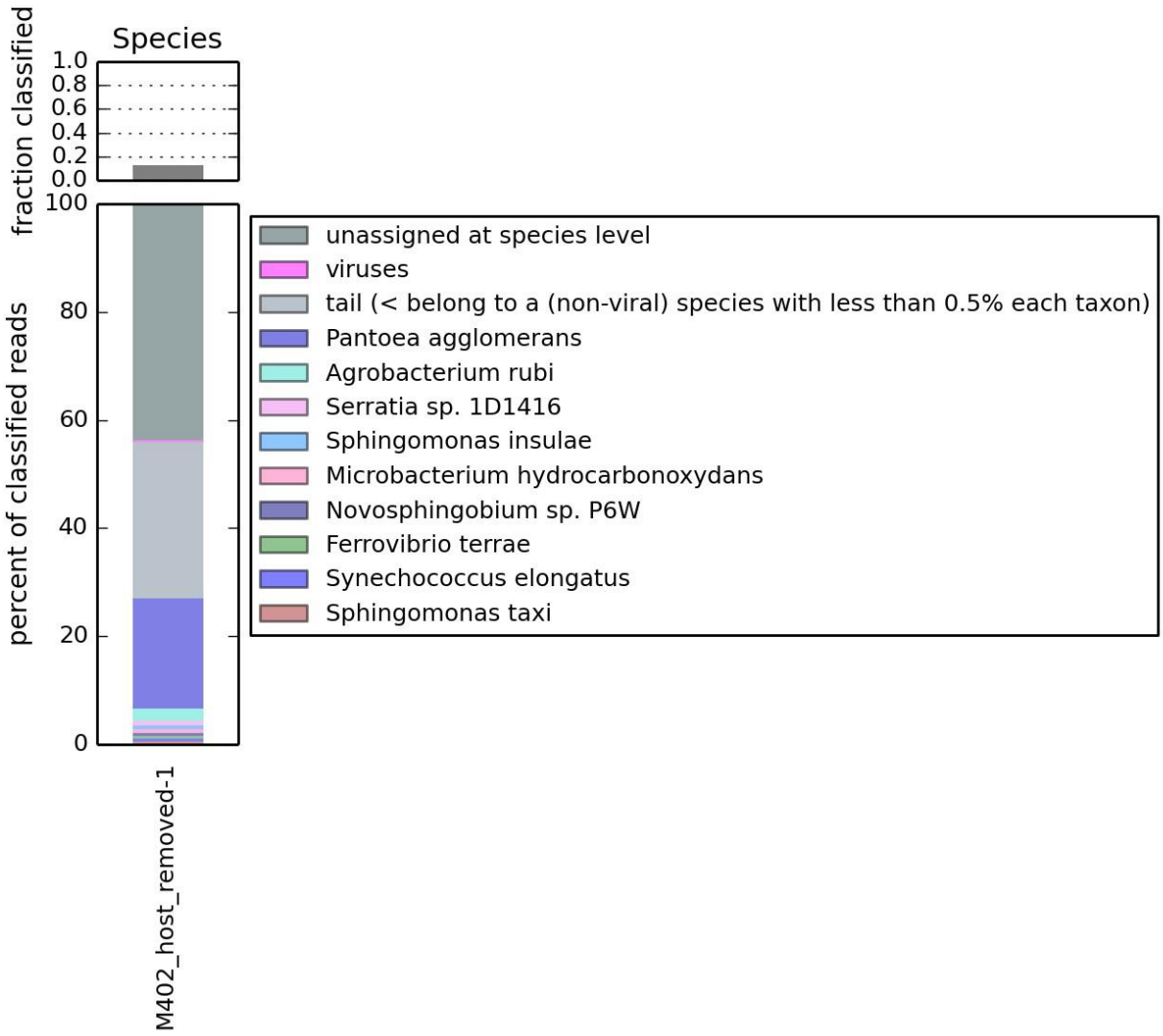

Figura 13. Clasificación taxonómica de las lecturas de la planta adulta empleando la base de datos RefSeq Genomes

4.1.2 COMPARACIÓN CON BASE DE DATOS DE PLÁSMIDOS

En las Figuras 14 y 15 se observa la clasificación taxonómica de las lecturas obtenidas a partir de las plantas *in vitro* y adulta respectivamente usando como referencia la base de datos de plásmidos. Para la planta *in vitro*, se identificaron mayoritariamente secuencias pertenecientes a plásmidos de *Escherichia coli* (8.018 %) y *Tsukaramella tyrosinosolvens* (7.262 %). También se identificaron secuencias de 15 especies del género *Pantoea,* incluyendo *Pantoea* sp. SO10 (2.143 %), *P. dispersa* (1.867 %)*, P. cypripedii* (1.214 %), *P. agglomerans* (1.172 %)*, Pantoea* sp. At-90b (1.102 %)*, Pantoea* sp. MSR2 (1.083 %), *P. vagans* (0.566 %), *P. ananatis* (0.450 %), *P. stewartii* (0.166 %), *P. eucrina* (0.075 %), *Pantoea* sp. PSNIH1 (0.008 %), *P. eucalypti* (0.006 %), *Pantoea* sp. CCBC3-3-1 (0.004 %), candidato de *P. carbekii* (0.002 %) y *P. alhagi* (0.002 %) (Figura 14).

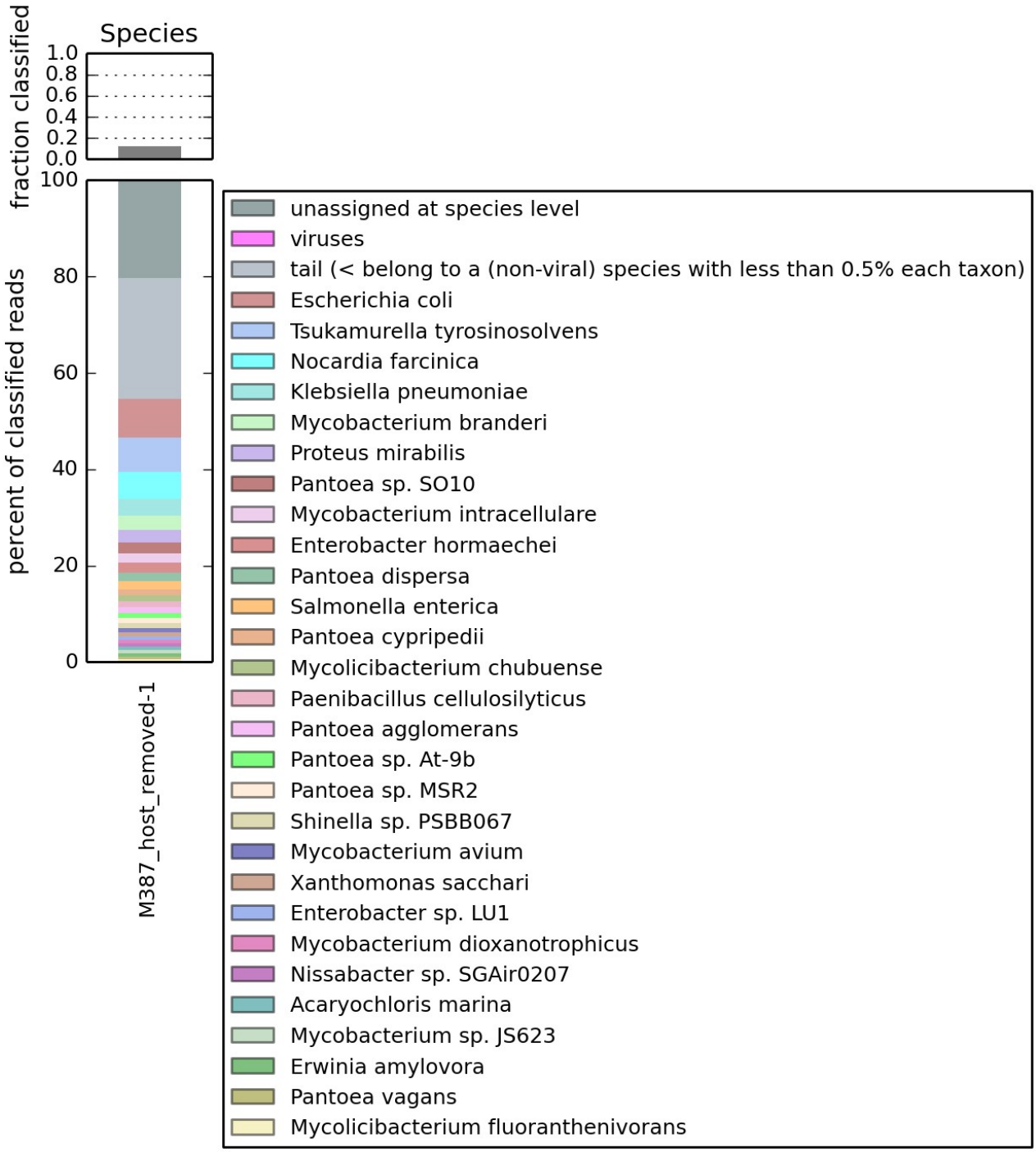

Figura 14. Clasificación taxonómica de las lecturas obtenidas a partir de la planta *in vitro*

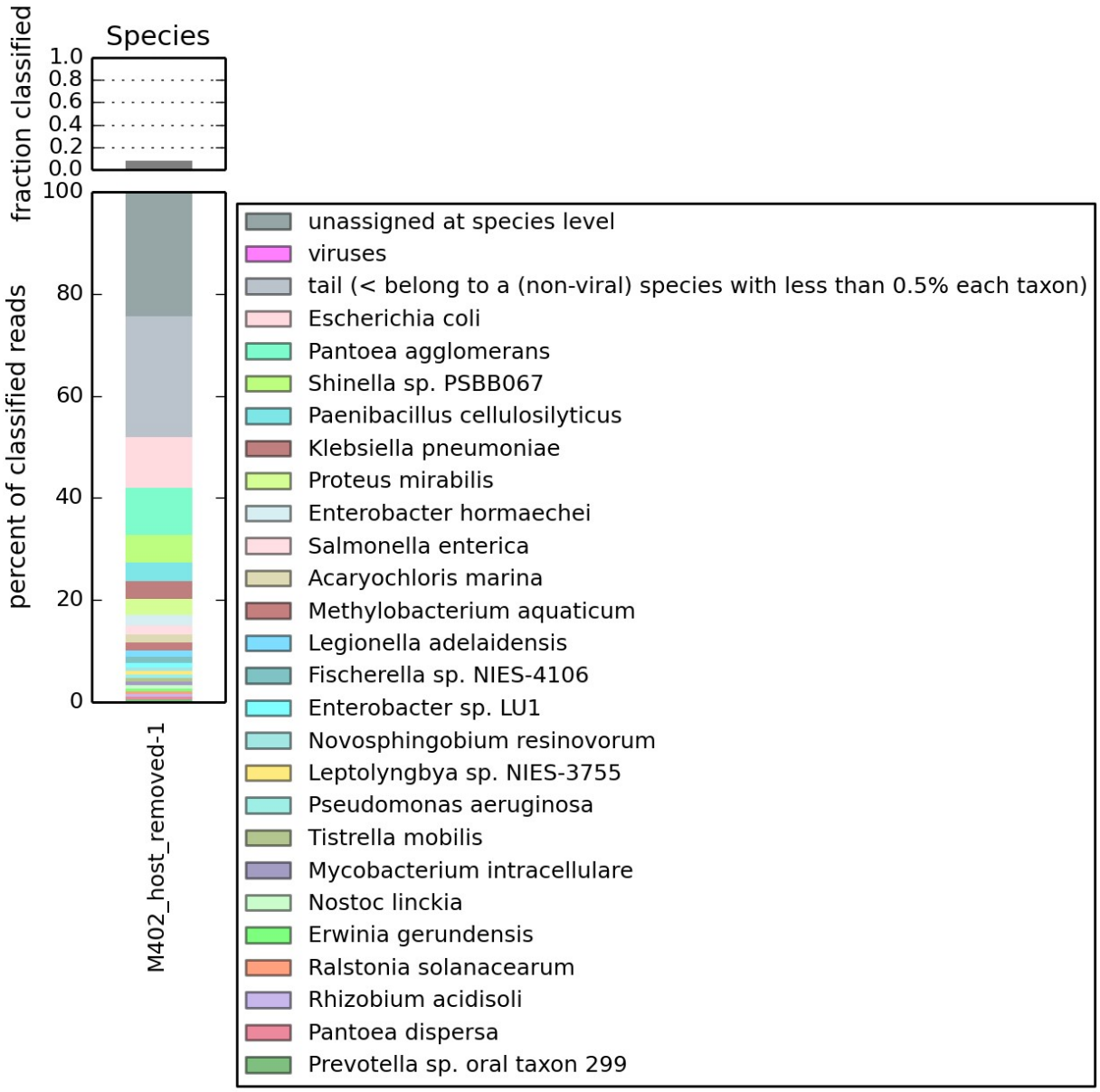

Figura 15. Clasificación taxonómica de las lecturas obtenidas a partir de la planta adulta

De forma similar en la muestra de la planta adulta (Figura 15) se evidencia una mayor abundancia de lecturas de plásmidos de *Escherichia coli* (9.854 %), *P. agglomerans* (9.215 %) y *Shinella* sp. PSBB067 (5.843 %). También se identificaron lecturas de plásmidos de 12 especies adicionales del género *Pantoea*, incluyendo *P. dispersa* (0.506 %), *Pantoea* sp. SO10 (0.341 %), *P. vagans* (0.335 %), *P. cypripedii* (0.312 %), *P. ananatis* (0.273 %), *Pantoea* sp. MSR2 (0.179 %), *Pantoea* sp. At-9b (0.133 %), *P. eucalypti* (0.031 %), *P. eucrina* (0.027 %), *P. stewartii* (0.015 %), *Pantoea* sp. MT58 (0.004 %) y candidato a *P. carbekii* (0.002 %).

4.2 PREDICCIONES

En las Tablas 3 y 4, se muestra el número de fragmentos de plásmidos predichos por cada herramienta. PlasFlow fue la herramienta con mayor número de predicciones en todos los bins analizados (5041), seguida de SOURCEFINDER (2647) y PlasmidHunter (148). El bin con mayor número de plásmidos detectados fue el Bin.002.fasta.assembly, con un rango que varió de 5 a 3921 predicciones según la herramienta empleada. Este fue sucedido por el Bin.004.M402.Sphingomonas, con un rango de 82 a 522 predicciones.

Tabla 3. Resumen de fragmentos de plásmidos predichos por las herramientas estudiadas

Archivo	Herramienta	Fragmentos predichos
BIN001.fasta.assembly	PlasmidHunter	6
	SOURCEFINDER	1
	PlasFlow	17
BIN002.fasta.assembly	PlasmidHunter	5
	SOURCEFINDER	2412
	PlasFlow	3921
BIN003.fasta.assembly	PlasmidHunter	7
	SOURCEFINDER	20
	PlasFlow	282
BIN.004.fasta.assembly	PlasmidHunter	42
	SOURCEFINDER	4
	PlasFlow	93
Bin.002.M402.Pantoea.agglomerans	PlasmidHunter	6
	SOURCEFINDER	55
	PlasFlow	206
Bin.004.M402.Sphingomonas	PlasmidHunter	82
	SOURCEFINDER	155
	PlasFlow	522

Las 3 herramientas entre sí tuvieron 18 coincidencias. Interesantemente, las herramientas SOURCEFINDER y PlasFlow tuvieron mayores coincidencias entre sí (1260) que con PlasmidHunter (Tabla 4).

Tabla 4. Resumen de fragmentos de plásmidos predichos por cada herramienta

Herramienta	Fragmentos Predichos	Coincidencia con otras herramientas		
		PlasmidHunter	SOURCEFINDER	PlasFlow
PlasmidHunter	148	-	9	32
SOURCEFINDER	2647	9	-	1260
PlasFlow	5041	32	1260	-

4.3 BÚSQUEDA EN BASES DE DATOS

4.3.1 GENERAL

Los contigs de todos los bins estudiados fueron sometidos a una búsqueda con la herramienta BLASTn contra la base de datos curada de plásmidos de Sistrom (2018). Se encontraron 1322 coincidencias en total de contigs que se resumen en la Tabla 5.

Tabla 5. Contigs con coincidencia en la base de datos de Sistrom (2018)

Bin	Contigs con coincidencia en la base de datos	Porcentaje
BIN001.fasta.assembly	78	6%
BIN002.fasta.assembly	12	1%
BIN003.fasta.assembly	63	5%
BIN004.fasta.assembly	456	34%
BIN002.fasta.assembly.Pantoea.agglomerans	52	4%
Bin.004.M402.Sphingomonas	661	50%
Total	1322	

El Bin.004.M402.Sphingomonas tuvo el mayor número de coincidencias (661) seguido del BIN004.fasta.assembly (456). En contraste, el BIN002.fasta.assembly tuvo el menor número de coincidencias (Tabla 5 y Figura 16).

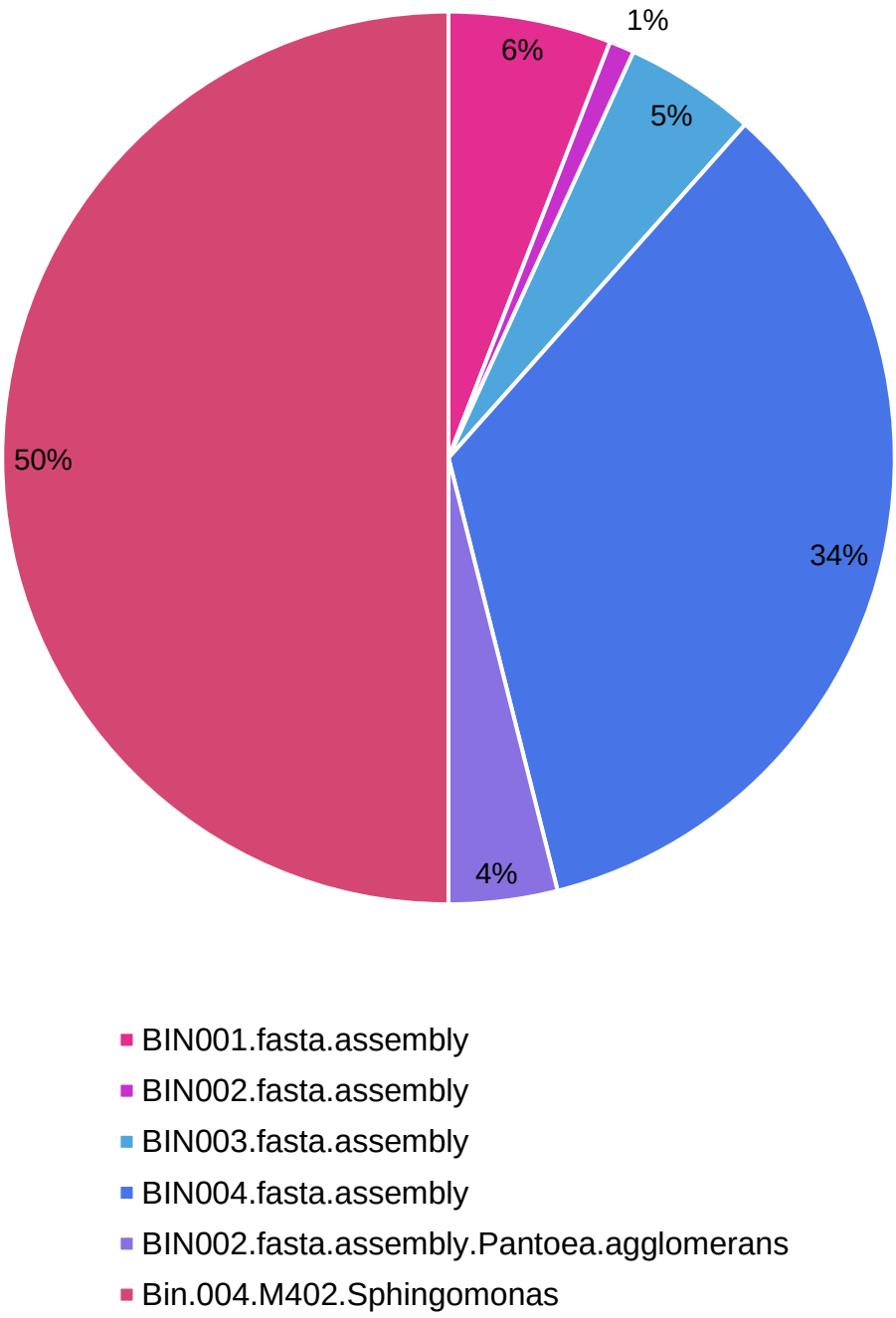

Figura 16. Proporción de contigs con coincidencias en la base de datos de Sistrom (2018)

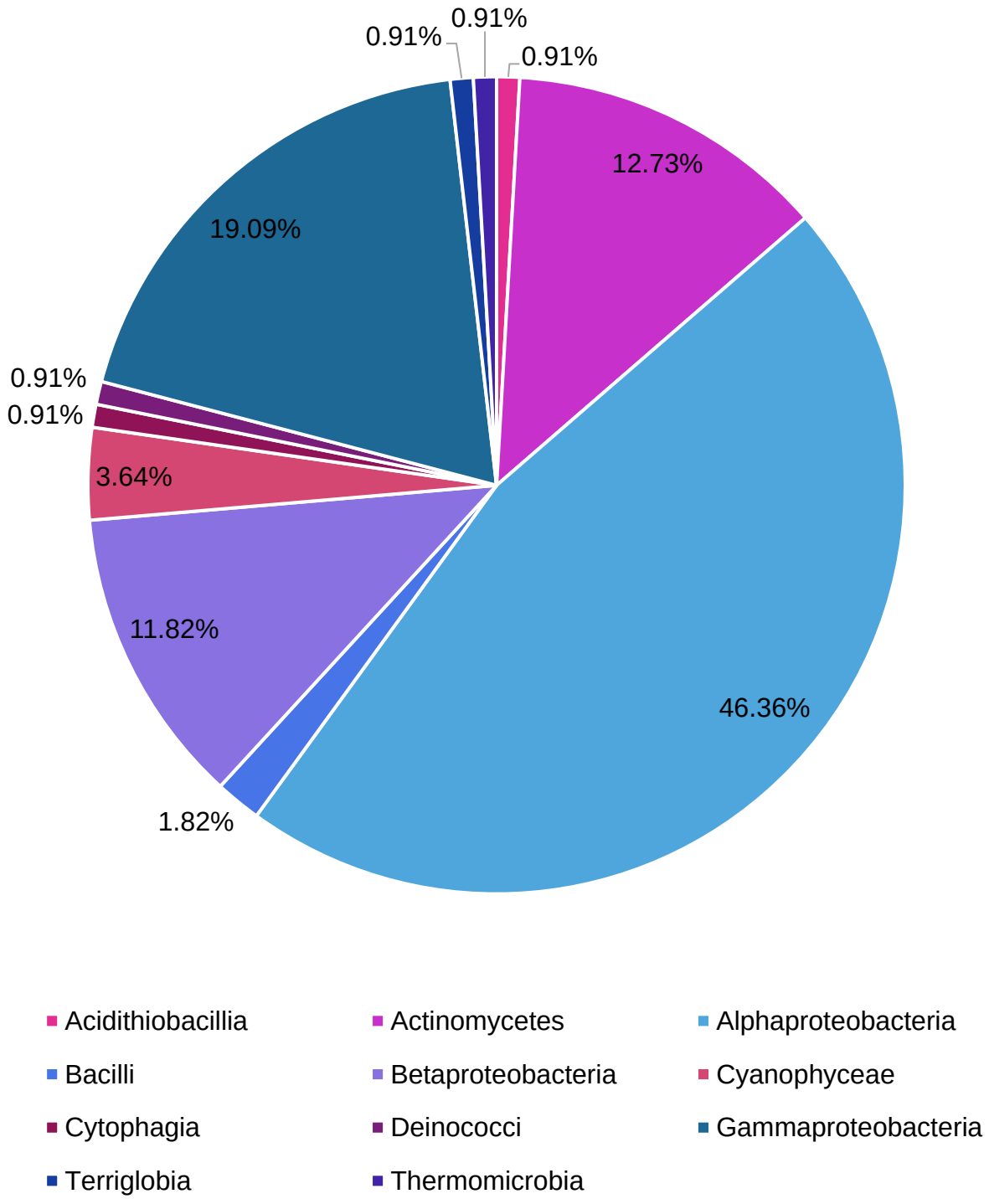

Figura 17. Clases identificadas en la búsqueda con BLASTn en la base de datos de Sistrom (2018)

Respecto a las clases de microorganismos identificadas en la búsqueda con BLASTn en la base de datos, la mayoría de los fragmentos de plásmidos corresponde a la clase Alphaproteobacteria (46.36 %) (Figura 17).

4.3.2 BÚSQUEDA DE RESULTADOS DE PREDICCIONES

Los contigs predichos como fragmentos de plásmidos fueron buscados en la base de datos, encontrándose 468 coincidencias. El resto de las secuencias no tuvieron coincidencia en la base de datos de Sistrom (2018). Principalmente en el caso del BIN002.fasta.assembly para el cual se encontraron 7 coincidencias con la base de datos. De estas la más representativa fue del 80% (907/1131) con *Azospirillum argentinense* strain Az39 plasmid AbAZ39_p1 para el contig NODE_132_length_13385_cov_425.176. En la Figura 18, se resumen las clases de las coincidencias de la búsqueda en BLASTn.

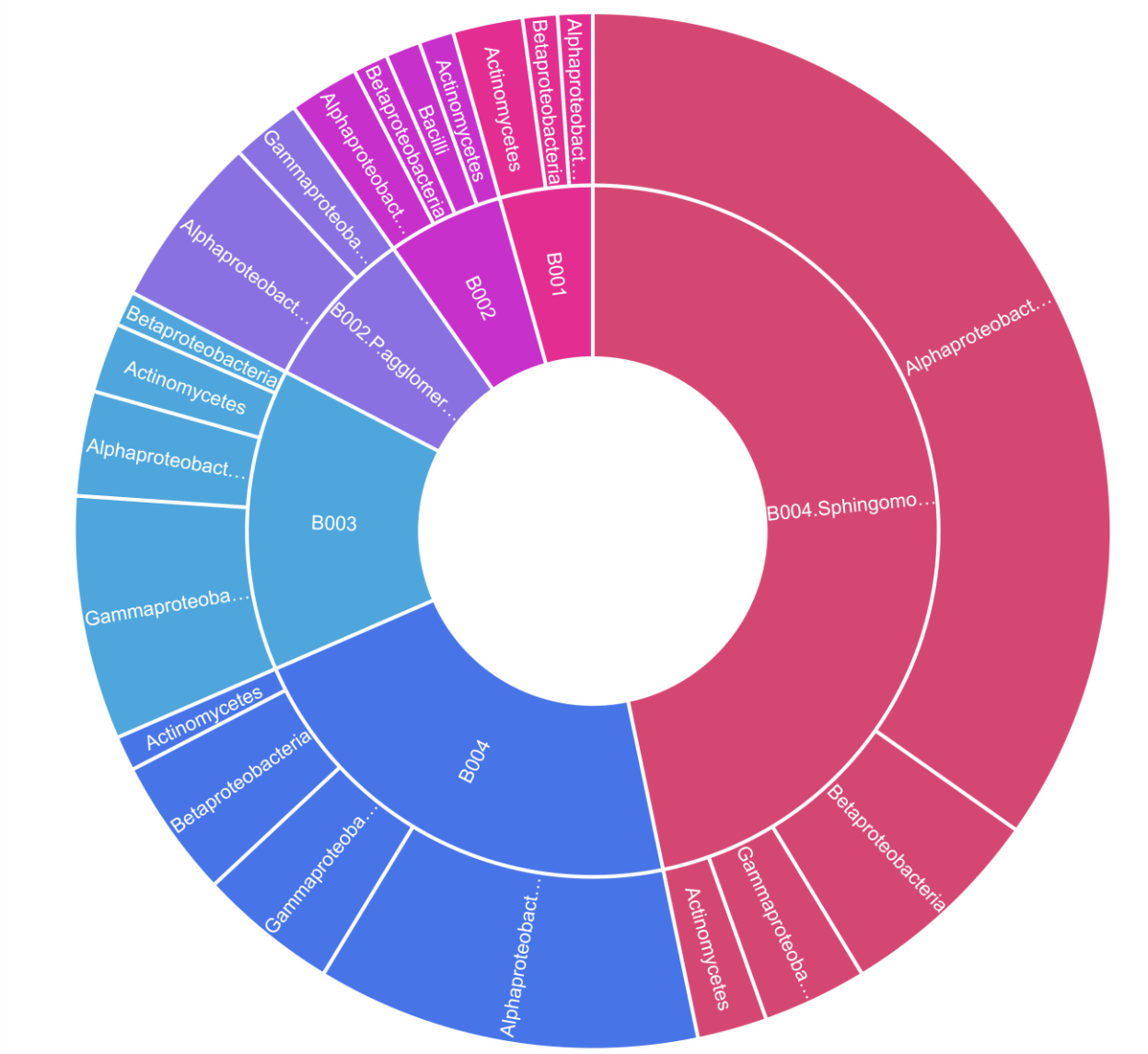

Figura 18. Resumen de coincidencias de búsqueda de los fragmentos predichos por las herramientas PlasmidHunter, SOURCEFINDER y PlasFlow en cada bin analizado

4.4 EVALUACIÓN DEL DESEMPEÑO DE LAS HERRAMIENTAS

En la Tabla 6 se muestran los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos de cada herramienta conforme al bin analizado y, en la Tabla 7, los cálculos de sensibilidad (R%), exactitud (A%) y precisión (P%).

Tabla 6. Resumen de las predicciones de las herramientas

BIN	HERRAMIENTA	VP	VN	FP	FN
BIN001.fasta.assembly	PlasmidHunter	6	92	0	72
	SOURCEFINDER	0	91	1	78
	PlasFlow	7	82	10	71
BIN002.fasta.assembly	PlasmidHunter	1	10663	4	11
	SOURCEFINDER	3	8258	2409	9
	PlasFlow	3	6749	3918	9
BIN003.fasta.assembly	PlasmidHunter	3	1035	4	60
	SOURCEFINDER	6	1025	14	57
	PlasFlow	11	768	271	52
BIN004.fasta.assembly	PlasmidHunter	24	573	18	432
	SOURCEFINDER	0	587	4	456
	PlasFlow	31	529	62	425
Bin.002.M402.Pantoea.agglomerans	PlasmidHunter	3	698	3	49
	SOURCEFINDER	7	653	48	45
	PlasFlow	10	505	196	42
Bin.004.M402.Sphingomonas	PlasmidHunter	61	879	21	600
	SOURCEFINDER	72	817	83	589
	PlasFlow	220	598	302	441

Tabla 7. Cálculo de sensibilidad (R%), exactitud (A%) y precisión (P%).

BIN	HERRAMIENTA	R%	A%	P%
BIN001.fasta.assembly	PlasmidHunter	7.69%	57.65%	100.00%
	SOURCEFINDER	0.00%	53.53%	0.00%
	PlasFlow	8.97%	52.35%	41.18%
BIN002.fasta.assembly	PlasmidHunter	8.33%	99.86%	20.00%
	SOURCEFINDER	25.00%	77.36%	0.12%
	PlasFlow	25.00%	63.23%	0.08%
BIN003.fasta.assembly	PlasmidHunter	4.76%	94.19%	42.86%
	SOURCEFINDER	9.52%	93.56%	30.00%
	PlasFlow	17.46%	70.69%	3.90%
BIN004.fasta.assembly	PlasmidHunter	5.26%	57.02%	57.14%
	SOURCEFINDER	0.00%	56.06%	0.00%
	PlasFlow	6.80%	53.49%	33.33%
Bin.002.M402.Pantoea.agglomerans	PlasmidHunter	5.77%	93.09%	50.00%
	SOURCEFINDER	13.46%	87.65%	12.73%
	PlasFlow	19.23%	68.39%	4.85%
Bin.004.M402.Sphingomonas	PlasmidHunter	9.23%	60.22%	74.39%
	SOURCEFINDER	10.89%	56.95%	46.45%
	PlasFlow	33.28%	52.40%	42.15%

5. DISCUSIÓN

5.1 CLASIFICACIÓN TAXONÓMICA DE LAS LECTURAS

5.1.1 COMPARACIÓN CON BASE DE DATOS DE GENOMAS DE REFERENCIA

En la planta *in vitro* se identificaron cuatro especies del complejo *Mycobacterium avium* (*M. marseillense*, *M. intracellulare*, *M. avium* y *M. colombiense*) (Murcia et al., 2006; Ben Salah et al., 2009; Kim et al., 2017). Estas especies son ubicuas en suelo y agua y pueden causar infecciones intratables y mortales, por lo que es importante monitorear su presencia en tejidos vegetales (Keen et al., 2021). Se han reportado especies de *Mycobacterium* en otras rosáceas como *Fragaria* sp., específicamente en las partes comestibles cercanas al suelo o bajo el mismo que pudieron ingresar por absorción radicular (Hruska & Kaevska, 2012).

Otras especies encontradas son del género *Pantoea*, el cual ha sido reportado como predominante previamente en estudios metagenómicos de *Rosa* sp. (Xia et al., 2020). Salvetti et al. (2016) reportaron una abundancia relativa del 76 % de *Pantoea* en *Vitis vinifera*.

En el tumor de corona de la planta adulta, se reportó una mayor cantidad de lecturas de *Pantoea agglomerans* (20.365 %), de la cual algunas cepas han sido reportadas como formadoras de tumores en gypsophila, remolacha, abeto de Douglas y

arándano rojo debido a la presencia del plásmido pPATH de 150 kb con una región patogénica (Dutkiewicz et al., 2016).

Otro microorganismo tumorigénico presente en la planta adulta fue *Agrobacterium rubi* (2.375 %), reconocido como patógeno de *Rubus* spp., *Rosa* spp., *Vitis vinífera*, *Daphne mezereum* y *Vaccinium* spp. (Abrahamovich et al., 2014). De manera similar a otros miembros del género *Agrobacterium*, se ha reportado que *A. rubi* posee plásmidos Ti que le confieren propiedades patogénicas (Schell et al., 1979; Gordon & Christie, 2014; Crespo-Sempere et al., 2016).

Interesantemente, se encontraron lecturas del género *Serratia*, el cual ha reportado actividad antagonista ante *Agrobacterium* gracias a la presencia de lipasas, proteasas, DNAsas y producción de sideróforos (Asghari et al., 2019).

Por su parte, *Sphingomonas insulae* y *Microbacterium hydrocarbonoxydans*, también presentes en la planta adulta, han sido previamente detectados en asociación a plantas (Samayoa et al., 2020; Zicca et al., 2020).

5.1.2 COMPARACIÓN CON BASE DE DATOS DE PLÁSMIDOS

En la planta *in vitro* se encontraron secuencias de plásmidos para las especies *Pantoea* sp. SO10, *P. dispersa*, *P. cypripedii, P. agglomerans, Pantoea* sp. At-9b*, Pantoea* sp. MSR2, *M. intracellulare*, *M. avium* en concordancia con la base de datos de referencia de genomas. En contraste, en la planta adulta se detectaron secuencias plasmídicas de *P. agglomerans* en paralelo con secuencias genómicas de la misma.

Cabe mencionar que, en las plantas *in vitro* y adulta se evidenció la presencia de lecturas de plásmidos de *E. coli.* Sin embargo, al comparar con las clasificaciones taxonómicas efectuadas con la base de datos de genomas de RefSeq, no se detectaron lecturas de genomas de *E. coli.* Esto puede deberse a la naturaleza de la base de datos de plásmidos ya que RefSeq a la fecha cuenta con 78 943 coincidencias para secuencias plasmídicas de las cuales 17 905 corresponden a *E. coli*. Esto evidencia la necesidad de caracterizar los plásmidos de microorganismos pertenecientes a nichos poco estudiados que no figuran en las bases de datos ya que las bases de datos disponibles están altamente sesgadas a microorganismos aislados de ambientes ampliamente estudiados (Smith et al., 2022). Además, es necesario fortalecer los esfuerzos de curación de bases de datos a fin de evitar repetición y verificar errores.

Otro factor que pudo influir en los resultados es la posible presencia de material genético contaminante en el proceso de preparación de las muestras, la cual puede verificarse repitiendo la extracción con material esterilizado.

5.2 CAPACIDAD DE PREDICCIÓN DE LAS HERRAMIENTAS SELECCIONADAS

5.2.1 PLASFLOW

Las herramientas empleadas tuvieron diferencias marcadas en su capacidad de predicción. Como se observa en las Tablas 3 y 4, la herramienta con mayor número de coincidencias de fragmentos de plásmidos fue PlasFlow. Esta herramienta ha reportado altos niveles de éxito de identificación con una exactitud de hasta el 96 % en metagenomas ensamblados sin conocimiento previo de la composición taxonómica o funcional de las muestras. Además, puede reconocer secuencias circulares y lineares, haciendo una clasificación taxonómica inicial (Krawczyk et al., 2019). En el presente estudio, PlasFlow detectó 5041 fragmentos de plásmidos en los contigs con diferentes rangos de sensibilidad, exactitud y precisión en cada bin, alcanzando su mayor sensibilidad (33.28 %), y precisión (42.15) en el Bin.004.M402.Sphingomonas y, su mayor exactitud en el Bin.003.fasta.assembly (70.69 %) (Tabla 7). En el estudio de Krawczyk et al. (2019), esta herramienta logró predecir correctamente el 90.44% de los filos en el caso de plásmidos.

Las diferencias en los valores obtenidos y las del artículo de referencia pueden atribuirse a múltiples factores, uno de ellos es la base de datos utilizada, ya que la base de datos de Sistrom (2018) es una base de datos curada que cuenta con 10,892 secuencias de plásmidos y PlasFlow se construyó con 7604.

El uso de bases de datos curadas como la de Sistrom (2018) y PLSDB (Schmartz et al., 2022) es altamente recomendado para búsquedas considerando que la colección

de RefSeq está parcialmente incompleta, es inconsistente, carece de funcionalidad y contiene secuencias cromosómicas (Schmartz et al., 2022).

Adicionalmente, el BIN002.fasta.assembly fue el que tuvo mayor número de predicciones con PlasFlow (3921, Tabla 3) pero estas no coincidieron con los resultados de la base de datos. Al buscar algunos de los contigs de dicho bin en el repositorio NCBI, se encontraron coincidencias con secuencias cloroplásticas y mitocondriales de *Rosa* sp. Esto puede atribuirse a que la composición nucleotídica de los plásmidos es similar a la del ADN de cloroplastos y mitocondrias (Jacobs et al., 1992).

Precisamente, para la construcción de PlasFlow, se tomó la diferencia de la composición media de nucleótidos de genomas (%GC = 51.05%) y plásmidos (%GC = 44.31%). Estos valores oscilaron entre 28.50 % (Tenericutes) y 67.18% (*Deinococcus-Thermus*) para genomas y, entre 25.95% (Fusobacteria) y 69.42% (*Deinococcus-Thermus*) para plásmidos según los filos analizados, por lo que Krawczyk et al. los incorporaron para la inferencia de información taxonómica en los resultados finales (2019).

Otra explicación para las diferencias en el caso de PlasFlow es que, de manera similar a otros métodos basados en k-méros, PlasFlow no es apto para procesar secuencias cortas ya que es difícil obtener una cobertura óptima. Por ende, los resultados para secuencias cortas no son confiables (Krawczyk, 2021).

5.2.2 SOURCEFINDER

El clasificador de SOURCEFINDER fue el segundo con mayor número de predicciones (2647). Su mayor sensibilidad (25.00%) se reportó para el BIN002.fasta.assembly mientras que su mayor exactitud (93.56%) para el BIN003.fasta.assembly y mayor precisión (46.45%) para Bin.004.M402.Sphingomonas. En el estudio de Aytan-Aktug (2022), SOURCEFINDER alcanzó una exactitud del 94.4% para plásmidos, cercana a la mayor obtenida en el presente trabajo. Esta herramienta destaca por usar un algoritmo simple de machine-learning independientemente de si las secuencias están completas o de su clasificación taxonómica. Además, se recomienda para la caracterización de ensamblajes *de novo* y ofrece una interfaz intuitiva.

Interesantemente, SOURCEFINDER tuvo 1260 coincidencias de predicciones con PlasFlow. Las diferencias entre las predicciones pueden atribuirse a que se entrenaron los modelos con diferentes bases de datos dado que SOURCEFINDER incluyó 18022 secuencias cromosómicas de bacterias representativas de RefSeq además de 10852 secuencias plasmídicas y 12930 secuencias de bacteriófagos de PATRIC (Aytan-Aktug et al., 2022). Cabe mencionar que, el modelo de fragmento de SOURCEFINDER se recomienda para caracterizar ensamblajes *de novo* altamente fragmentados.

5.2.3 PLASMIDHUNTER

PlasmidHunter fue la herramienta con menor tasa de predicciones (148), con una sensibilidad máxima de 9.23 % para el Bin.004.M402.Sphingomonas, exactitud máxima de 99.86 % para el Bin.002.fasta y precisión máxima del 100% para el Bin.001.fasta. Su

exactitud es próxima a la de 96.7 % reportada por Tian e Imanian (2023). Sin embargo, su sensibilidad es significativamente menor en comparación a las otras herramientas estudiadas. Respecto a su precisión alcanzada es importante mencionar que PlasmidHunter puede presentar una menor precisión cuando se trabaja con contigs cortos (83.3% y 89.1% para contigs de 5 kbp y 10 Kbp, respectivamente), pero por sus valores de verdaderos positivos, se alcanza un balance (Tian & Imanian, 2023).

6. CONCLUSIONES

En el presente estudio se analizaron secuencias metagenómicas de planta *in vitro* y tumor de corona de *Rosa* sp. En la primera se evidencia la predominancia de *Mycobacterium* spp. y *Pantoea* spp. En la segunda, destacan lecturas de *Pantoea agglomerans*, *Agrobacterium rubi*, *Serratia* sp. 1D1416 y *Sphingomonas insulae*.

Respecto a las herramientas seleccionadas para la identificación de fragmentos de plásmidos, PlasmidHunter tiene la mayor exactitud y precisión. Mientras que PlasFlow es la herramienta con mayor sensibilidad. Estas diferencias pueden atribuirse a las bases de datos empleadas para el entrenamiento de las secuencias y su arquitectura. No obstante, cabe destacar que las tres herramientas empleadas en este estudio son amigables con el usuario, ya que están disponibles en línea sin requerimientos de instalaciones adicionales.

7. RECOMENDACIONES

- Emplear el software autónomo de blastn reduce los tiempos de búsqueda significativamente si se emplea una base de datos especializada
- Emplear bases de datos curadas para restringir la búsqueda a plásmidos correctamente anotados
- Realizar aislamientos bacterianos para mejorar la identificación de las especies de interés

8. REFERENCIAS

Abrahamovich, Eliana, López, Ana C, & Alippi, Adriana M. (2014). Diversidad de cepas de *Agrobacterium rubi* aisladas de arándanos. *Revista argentina de microbiología*, *46*(3), 237-241. Recuperado en 18 de julio de 2023, de http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S0325-75412014000400011&lng=es&tlng=es

Andreopoulos, W. B., Geller, A. M., Lucke, M., Balewski, J., Clum, A., Ivanova, N. N., & Levy, A. (2022). Deeplasmid: deep learning accurately separates plasmids from bacterial chromosomes. *Nucleic Acids Research*, *50*(3), e17–e17. https://doi.org/10.1093/nar/gkab1115

Ankita, K., Yu-Wei, W., John-Marc, C., Marimikel, C., Lara, R., M., R. A., C., J. D., C., H. T., W., S. S., & Aindrila, M. (2019). Large Circular Plasmids from Groundwater Plasmidomes Span Multiple Incompatibility Groups and Are Enriched in Multimetal Resistance Genes. *MBio*, *10*(1), e02899-18. https://doi.org/10.1128/mBio.02899-18

Antipov, D., Raiko, M., Lapidus, A., & Pevzner, P. A. (2019). Plasmid detection and assembly in genomic and metagenomic data sets. *Genome research*, *29*(6), 961–968. https://doi.org/10.1101/gr.241299.118

Arkin, A. P., Cottingham, R. W., Henry, C. S., Harris, N. L., Stevens, R. L., Maslov, S., Dehal, P., Ware, D., Perez, F., Canon, S., Sneddon, M. W., Henderson, M. L., Riehl, W. J., Murphy-Olson, D., Chan, S. Y., Kamimura, R. T., Kumari, S., Drake, M. M., Brettin, T. S., … Yu, D. (2018). KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nature Biotechnology*, *36*(7), 566–569. https://doi.org/10.1038/nbt.4163

Arredondo-Alonso, S., Willems, R. J., van Schaik, W., & Schürch, A. C. (2017). On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microbial Genomics*, *3*(10), e000128

Asghari, S., Harighi, B., Mozafari, A.A. *et al.* Screening of endophytic bacteria isolated from domesticated and wild growing grapevines as potential biological control agents against crown gall disease. *BioControl* **64**, 723–735 (2019). https://doi.org/10.1007/s10526-019-09963-z

Aytan-Aktug, D., Grigorjev, V., Szarvas, J., Clausen, P. T. L. C., Munk, P., Nguyen, M., Davis, J. J., Aarestrup, F. M., & Lund, O. (2022). SourceFinder: a Machine-Learning-Based Tool for Identification of Chromosomal, Plasmid, and Bacteriophage Sequences from Assemblies. *Microbiology Spectrum*, *10*(6). https://doi.org/10.1128/SPECTRUM.02641-22/SUPPL_FILE/SPECTRUM.02641-22-S0002.XLSX

Barooah, M., Goswami, G., Hazarika, D. J., & Kangabam, R. (2021). *High-Throughput Analysis to Decipher Bacterial Diversity and their Functional Properties in Freshwater Bodies BT - Microbial Metatranscriptomics Belowground* (M. Nath, D. Bhatt, P. Bhargava, & D. K. Choudhary (eds.); pp. 511–542). Springer Singapore. https://doi.org/10.1007/978-981-15-9758-9_24

Basak, P., Biswas, A., & Bhattacharyya, M. (2020). Exploration of extremophiles genomes through gene study for hidden biotechnological and future potential. *Physiological and Biotechnological Aspects of Extremophiles*, 315–325. https://doi.org/10.1016/B978-0-12-818322-9.00024-1

Ben Salah, I., Cayrou, C., Raoult, D., & Drancourt, M. (2009). *Mycobacterium marseillense* sp. nov., *Mycobacterium timonense* sp. nov. and *Mycobacterium bouchedurhonense* sp. nov., members of the *Mycobacterium avium* complex. *International journal of systematic and evolutionary microbiology*, *59*(Pt 11), 2803–2808. https://doi.org/10.1099/ijs.0.010637-0

Brooks, J. P., Edwards, D. J., Harwich, M. D., Rivera, M. C., Fettweis, J. M., Serrano, M. G., Reris, R. A., Sheth, N. U., Huang, B., Girerd, P., Strauss, J. F., Jefferson, K. K., Buck, G. A., & members), V. M. C. (additional. (2015). The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiology*, *15*(1), 66. https://doi.org/10.1186/s12866-015-0351-6

Clark, D., Pazdernik, N., & McGehee, M. (2019). *Molecular Biology*. Academic Press

Crespo-Sampere, A., Carralero, A., Plomer, M., Cervera, M., & Albiach-Martí, M. (2016). Precisión en el diagnóstico de la patología de “tumores de cuello” provocada por *Agrobacterium tumefaciens*. *Phytoma España*, 280, 68-70.

Dutkiewicz, J., Mackiewicz, B., Kinga Lemieszek, M., Golec, M., Milanowski, J. (2016). Pantoea agglomerans: a mysterious bacterium of evil and good. Part III. Deleterious effects: infections of humans, animals and plants. *Ann Agric Environ Med*., 23(2), 197-205. https://doi.org/10.5604/12321966.1203878

Fang, Z., Tan, J., Wu, S., Li, M., Xu, C., Xie, Z., & Zhu, H. (2019). PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *GigaScience*, *8*(6), giz066

Fang, Z., Tan, J., Wu, S., Li, M., Wang, C., Liu, Y., & Zhu, H. (2020). PlasGUN: Gene prediction in plasmid metagenomic short reads using deep learning. *Bioinformatics*, *36*(10), 3239–3241

Fricke, W., Cebula, T., & Ravel, J. (2011). *Genomics.* En *Microbial Forensics*, 479–492. doi:10.1016/b978-0-12-382006-8.00028-1

Gilbert, J. A., & Dupont, C. L. (2011). Microbial metagenomics: beyond the genome. *Annual review of marine science*, *3*, 347–371

Gordon JE, Christie PJ. The Agrobacterium Ti Plasmids. Microbiol Spectr. 2014 Dec;2(6):10.1128/microbiolspec.PLAS-0010-2013. doi: 10.1128/microbiolspec.PLAS-0010-2013. PMID: 25593788; PMCID: PMC4292801.

Hilpert, C., Bricheux, G., & Debroas, D. (2021). Reconstruction of plasmids by shotgun sequencing from environmental DNA: which bioinformatic workflow? *Briefings in bioinformatics*, *22*(3), bbaa059. https://doi.org/10.1093/bib/bbaa059

Jacobs, J. D., Ludwig, J. R., Hildebrand, M., Kukel, A., Feng, T.-Y., Ord, R. W., & Volcani, B. E. (1992). *Characterization of two circular plasmids from the marine diatomCylindrotheca fusiformis: plasmids hybridize to chloroplast and nuclear DNA. MGG Molecular & General Genetics, 233(1-2), 302–310.* doi:10.1007/bf00587592

Keen, E. C., Choi, J., Wallace, M. A., Azar, M., Mejia-Chew, C. R., Mehta, S. B., Bailey, T. C., Caverly, L. J., Burnham, C. D., & Dantas, G. (2021). Comparative Genomics of Mycobacterium avium Complex Reveals Signatures of Environment-Specific Adaptation and Community Acquisition. *mSystems*, *6*(5), e0119421. https://doi.org/10.1128/mSystems.01194-21

Kim, S. Y., Shin, S. H., Moon, S. M., Yang, B., Kim, H., Kwon, O. J., Huh, H. J., Ki, C. S., Lee, N. Y., Shin, S. J., & Koh, W. J. (2017). Distribution and clinical significance of Mycobacterium avium complex species isolated from respiratory specimens. *Diagnostic microbiology and infectious disease*, *88*(2), 125–137. https://doi.org/10.1016/j.diagmicrobio.2017.02.017

Krawczyk, P. S., Lipinski, L., & Dziembowski, A. (2018). PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Research*, *46*(6), 35. https://doi.org/10.1093/nar/gkx1321

Lapidus, A. L., & Korobeynikov, A. I. (2021). Metagenomic Data Assembly – The Way of Decoding Unknown Microorganisms. *Frontiers in Microbiology*, *12*, 653

Li, A., Li, L., & Zhang, T. (2015). Exploring antibiotic resistance genes and metal resistance genes in plasmid metagenomes from wastewater treatment plants. *Front. Microbiol*. 6:533

Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature communications*, *7*, 11257. https://doi.org/10.1038/ncomms11257

Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M. R., Kale, V., Potter, S. C., Richardson, L. J., Sakharova, E., Scheremetjew, M., Korobeynikov, A., Shlemov, A., Kunyavskaya, O., Lapidus, A., & Finn, R. D. (2020). MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research*, *48*(D1), D570–D578

Murcia, M. I., Tortoli, E., Menendez, M. C., Palenque, E., & Garcia, M. J. (2006). *Mycobacterium colombiense* sp. nov., a novel member of the *Mycobacterium avium* complex and description of MAC-X as a new ITS genetic variant. *International journal of systematic and evolutionary microbiology*, *56*(Pt 9), 2049–2054. https://doi.org/10.1099/ijs.0.64190-0

Niehaus, J. (2017). *Tips for Using BLAST to Verify Plasmids*. Addgene Blog. Obtenido de https://blog.addgene.org/tips-for-using-blast-to-verify-plasmids

Noel, K. D. (2009). *Rhizobia*. En *Encyclopedia of Microbiology*. (pp. 261–277). Academic Press

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity, and predictive values. *Indian journal of ophthalmology*, *56*(1), 45–50

Pellow, D., Zorea, A., Probst, M., Furman, O., Segal, A., Mizrahi, I., & Shamir, R. (2021). SCAPP: an algorithm for improved plasmid assembly in metagenomes. *Microbiome*, *9*(1), 1–12. https://doi.org/10.1186/s40168-021-01068-z

Pradier, L., Tissot, T., Fiston-Lavier, A., & Bedhomme, S. (2021). PlasForest: a homology-based random forest classifier for plasmid detection in genomic datasets. *BMC Bioinformatics*, *22*(1), 349

Pu, L., & Shamir, R. (2022). 3CAC: improving the classification of phages and plasmids in metagenomic assemblies using assembly graphs. *Bioinformatics (Oxford, England)*, *38*(Supplement_2), ii56–ii61

Raza, A., & Shahid, M. S. (2020). Next-generation sequencing technologies and plant molecular virology: a practical perspective. *Applied Plant Virology: Advances, Detection, and Antiviral Strategies*, 131–140. https://doi.org/10.1016/B978-0-12-818654-1.00010-4

Reddy, N., & Surekha, C. (2016). Next-Generation Sequencing and Metagenomics. En *Computational Biology and Bioinformatics* (pp. 344–364). CRC Press

Rozov, R., Brown Kav, A., Bogumil, D., Shterzer, N., Halperin, E., Mizrahi, I., & Shamir, R. (2017). Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics (Oxford, England)*, *33*(4), 475–482

Salvetti, E., Campanaro, S., Campedelli, I., Fracchetti, F., Gobbi, A., Tornielli, G. B., Torriani, S., & Felis, G. E. (2016). Whole-Metagenome-Sequencing-Based Community Profiles of *Vitis vinifera* L. cv. Corvina Berries Withered in Two Post-harvest Conditions. *Frontiers in microbiology*, *7*, 937. https://doi.org/10.3389/fmicb.2016.00937

Samayoa, B., Shen, F., Lai, W., & Chen, W. (2020). Screening and Assessment of Potential Plant Growth-promoting Bacteria Associated with *Allium cepa* Linn. *Microbes and Environments,* 35(2)

Sampieri, H., Collado, C., & Baptista, M. (2014). *Metodología de la Investigación*. McGraw Hill

Schell, J., Van Montagu, M., De Beuckeleer, M., De Block, M., Depicker, A., De Wilde, M., Engler, G., Genetello, C., Hernalsteens, J. P., Holsters, M., Seurinck, J., Silva, B., Van Vliet, F., & Villarroel, R. (1979). Interactions and DNA Transfer between *Agrobacterium tumefacien*s, the Ti-Plasmid and the Plant Host. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *204*(1155), 251–266. http://www.jstor.org/stable/77524

Schierstaedt, J., Bziuk, N., Kuzmanović, N., Blau, K., Smalla, K., & Jechalke, S. (2019). Role of plasmids in plant-bacteria interactions. *Current Issues in Molecular Biology*, *30*, 17–38. https://doi.org/10.21775/CIMB.030.017

Shintani, M., Sanchez, Z., & Kimbara, K. (2015). Genomics of microbial plasmids: Classification and identification based on replication and transfer systems and host taxonomy. *Frontiers in Microbiology*, *6*(MAR), 242

Shuikan, A., Alharbi, S. A., Alkhalifah, D. H. M. & Hozzein, W. N. (2020). High-Throughput Sequencing and Metagenomic Data Analysis. In *Metagenomics - Basics, Methods and Applications*. IntechOpen. https://doi.org/10.5772/INTECHOPEN.78746

Smith, R.H., Glendinning, L., Walker, A.W., &. Watson, M. (2022). Investigating the impact of database choice on the accuracy of metagenomic read classification for

the rumen microbiome. *Animal Microbiome* **4**, 57. https://doi.org/10.1186/s42523-022-00207-7

Stone, D., & Ellis, J. (2020). *Statistics in Analytical Chemistry - Stats (2)*. Obtenido de: https://sites.chem.utoronto.ca/chemistry/coursenotes/analsci/stats/AimStats.html

Sugitha, T., Binodh, A., Ramasamy, K., & Sivakumar, U. (2020). Bioinformatics in Metagenomics. In *Soil Metagenomics*. CRC Press.

Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microbial informatics and experimentation*, *2*(1), 3. https://doi.org/10.1186/2042-5783-2-3

Tian, R., & Imanian, B. (2023). PlasmidHunter: Accurate and fast prediction of plasmid sequences using gene content profile and machine learning. *BioRxiv*, 2023.02.01.526640. https://doi.org/10.1101/2023.02.01.526640

Valentino, V., Sequino, G., Cobo-Díaz, J. F., Álvarez-Ordóñez, A., De Filippis, F., & Ercolini, D. (2022). Evidence of virulence and antibiotic resistance genes from the microbiome mapping in minimally processed vegetables producing facilities. *Food research international (Ottawa, Ont.)*, *162*(Pt B), 112202. https://doi.org/10.1016/j.foodres.2022.112202

Van der Graaf-Van Bloois, L., Wagenaar, J., & Zomer, A. (2021). RFPlasmid: predicting plasmid sequences from short-read assembly data using machine learning. *Microbial Genomics*, *7*(11), 000683

Wickramarachchi, A., & Lin, Y. (2022). Binning long reads in metagenomics datasets using composition and coverage information. *Algorithms for Molecular Biology*, *17*(1), 14. https://doi.org/10.1186/s13015-022-00221-z

Xia, A. N., Liu, J., Kang, D. C., Zhang, H. G., Zhang, R. H., & Liu, Y. G. (2020). Assessment of endophytic bacterial diversity in rose by high-throughput sequencing analysis. *PloS one*, *15*(4), e0230924. https://doi.org/10.1371/journal.pone.0230924

Zhang, L., Chen, F., Zeng, Z., Xu, M., Sun, F., Yang, L., Bi, X., Lin, Y., Gao, Y., Hao, H., Yi, W., Li, M. & Xie, Y. (2021). Advances in Metagenomics and Its Application in Environmental Microorganisms. *Front. Microbiol*. *12*, 766364. doi: 10.3389/fmicb.2021.766364

Zicca, S., de Bellis, P., Masiello, M., Saponari, M., Saldarelli, P., Boscia, D., & Sisto, A. (2020). Antagonistic activity of olive endophytic bacteria and of Bacillus spp. strains against *Xylella fastidiosa*. *Microbiological Research*, *236*, 126467. https://doi.org/https://doi.org/10.1016/j.micres.2020.126467