



UNIVERSIDAD DEL PAPALOAPAN

CAMPUS TUXTEPEC

**“Análisis y procesamiento de datos de secuenciación masiva
de las regiones ITS y 16S de microbiomas de muestras de suelo”**

TESIS

QUE PARA OBTENER EL TÍTULO DE:

Ingeniero en Biotecnología

PRESENTA:

ALEXANDRO SANCHEZ PANAMA

DIRECTOR DE TESIS: MMM. LUIS URIBE ESPEJO GALICIA

SAN JUAN BAUTISTA TUXTEPEC, OAXACA

2023



UNIVERSIDAD DEL PAPALOAPAN

CAMPUS TUXTEPEC

INGENIERÍA EN BIOTECNOLOGÍA


ACTA DE REVISIÓN DE TESIS

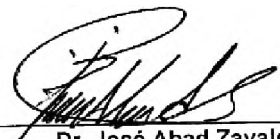
En la ciudad de San Juan Bautista Tuxtepec, Oaxaca, el día 27 de octubre de 2023 a las 11:15 h, los miembros de la comisión revisora de tesis designada por la Jefatura de Carrera de la Ingeniería en Biotecnología se reunieron en la sala de juntas del Instituto de Biotecnología de la Universidad del Papaloapan, con la finalidad de examinar la tesis titulada "Análisis y procesamiento de datos de secuenciación masiva de las regiones ITS Y 16S de microbiomas de muestras de suelo" presentada por el alumno **Alexandro Sánchez Panamá**, con número de matrícula 15090319, aspirante al título de **Licenciatura**.

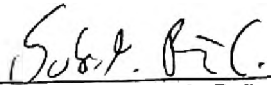
Después de intercambiar opiniones, los miembros de la comisión manifestaron que la tesis **satisface** los requisitos señalados por las disposiciones reglamentarias vigentes, otorgando su aprobación para que el aspirante pueda proceder con el proceso de titulación.


Tuxtepec, Oaxaca, a 27 de octubre de 2023.


ATENTAMENTE
LA COMISIÓN REVISORA



M.M. Luis Uribe Espejo Galicia.
Profesor Investigador Asociado "B"
Universidad del Papaloapan
Director de Tesis


Dr. José Abad Zavaleta
Profesor Investigador Titular "B"
Universidad del Papaloapan
Revisor de Tesis


Dr. Julián Mario Peña Castro
Profesor Investigador Titular "B"
Universidad del Papaloapan
Revisor de Tesis


Dr. Blanca Estela Barrera Figueroa
Profesor Investigador Titular "A"
Universidad del Papaloapan
Revisor de Tesis


Dr. Enrique Milagros Amador
Profesor Investigador Titular "A"
Universidad del Papaloapan
Revisor de Tesis


Dr. Paul Mauricio Sanchez Ocampo
Profesor Investigador Titular "A"
Universidad del Papaloapan
Revisor de Tesis



UNIVERSIDAD DEL PAPALOAPAN
CAMPUS TUXTEPEC

INGENIERÍA EN BIOTECNOLOGÍA

Tuxtepec, Oaxaca, a 31 de octubre de 2023
Oficio No. JCIB/008/10/2023

M.E Yesenia Barrientos Arenal
Jefa de Servicios Escolares
Universidad del Papaloapan

Con base en el dictamen de la comisión revisora, se autoriza la impresión del trabajo de tesis con título final "Análisis y procesamiento de datos de secuenciación masiva de las regiones ITS Y 16S de microbiomas de muestras de suelo" del alumno **Alexandro Sánchez Panamá** con número de matrícula 15090319. Para ser presentado como trabajo de tesis para obtener el título de Licenciado en **Ingeniería en Biotecnología**, toda vez que cumple satisfactoriamente con la reglamentación establecida para tal fin.

El Jurado de Examen Profesional estará compuesto por los siguientes profesores:

Dr. José Abad Zavaleta (Presidente, Universidad del Papaloapan)
Dr. Enrique Villalobos Amador (Vocal, Universidad del Papaloapan)
Dr. Julián Mario Peña Castro (Secretario, Universidad del Papaloapan)
Dra. Blanca Estela Barrera Figueroa (Primer Suplente, Universidad del Papaloapan)
Dr. Paúl Mauricio Sanchez Ocampo (Segundo Suplente, Universidad del Papaloapan)

Sin más por el momento le envío un cordial saludo.

Atentamente

Terra uberrima, mens aperta

Bou Lo-tama, chí jí jú



Jacqueline Capataz Tafur
Dra. Jacqueline Capataz Tafur
Jefe de Carrera de Ingeniería en
Biotecnología
Universidad del Papaloapan

JEFATURA DE INGENIERÍA
EN BIOTECNOLOGÍA
CAMPUS TUXTEPEC

Tania Zuñiga Marroquín
Vo.Bo. Dra. Tania Zuñiga Marroquín
Encargada del Despacho de Vice Rectoría
Académico
Universidad del Papaloapan



VICE-RECTORIA
ACADEMICA

c.c.p. M.M.M. Luis Uribe Espejo Galicia. Director de tesis, Para su conocimiento
c.c.p. Alexandro Sánchez Panamá. Alumno, Para su conocimiento
c.c.p. Archivo

Campus Loma Bonita
Av. Ferrocarril S/N, Col. Ciudad Universitaria, Loma
Bonita, Oaxaca C.P.68400
Tel/Fax: 01 281 872 8230

www.unpa.edu.mx

Campus Tuxtepec
Circuito Central N° 200, Col. Parque Industrial,
Tuxtepec, Oaxaca. C.P. 68301
Tel/Fax: 01 287 875 9240

Hoja de originalidad

El presente trabajo no ha sido aceptado o empleado para el otorgamiento de título o grado diferente o adicional al actual. La tesis es resultado de las investigaciones del autor, excepto donde se indican las fuentes de información consultadas. El autor otorga su consentimiento a la **Universidad del Papaloapan** para la reproducción del documento con el fin del intercambio bibliotecario siempre y cuando se indique la fuente.

Agradecimientos

En primer lugar, deseo expresar mi sincero agradecimiento a la Universidad del Papaloapan por brindarme la oportunidad de adquirir conocimiento y desarrollarme como estudiante. Esta institución me ha proporcionado un entorno de aprendizaje excepcional y un cúmulo de recursos valiosos que han sido fundamentales para mi formación académica.

A mis estimados profesores, el Dr. Abad, el Dr. Aurelio, el Dr. Miguel, el Dr. Enrique, el Dr. Víctor Meza, el Dr. Julián, el Dr. Paul, la Dra. Jacqueline, la Dra. Blanca y la Dra. María de Jesús les estoy enormemente agradecido por compartir su sabiduría, experiencia y dedicación en la enseñanza. Sus lecciones han sido invaluable y me han inspirado a esforzarme constantemente por alcanzar la excelencia académica.

A mi respetado asesor de tesis, el Maestro Luis Uribe, le agradezco profundamente por su orientación experta, su paciencia y su apoyo continuo a lo largo de este proceso. Sus consejos y retroalimentación han sido esenciales para el éxito de mi investigación.

A mis queridos compañeros de clase, Flavio, Itzel, Carolina, José Hilario y a todos los demás que han compartido esta travesía conmigo, les agradezco por el compañerismo, el intercambio de ideas y la colaboración que hemos experimentado juntos.

A mis queridos hermanos, Pablo y Yesica, les estoy enormemente agradecido por su apoyo incondicional a lo largo de este arduo proceso. Su constante aliento, comprensión y disposición para brindarme apoyo han sido un faro de motivación en los momentos de desafío. Sin su inquebrantable respaldo, este logro no habría sido posible. Su contribución ha sido invaluable y estoy profundamente agradecido por tenerlos como pilares en mi vida.

I. INDICE DE CONTENIDO

1. INTRODUCCIÓN	12
2. MARCO TEORICO	13
2.1 Gen ribosomal 16S rARN	13
2.2 Región ITS como “código de barras” para la identificación fúngica.	15
2.3 Secuenciación masiva en la detección de microbiotas.....	17
2.4 Epi2me	18
2.5 Formato FAST 5.....	19
2.6 Formato FASTQ.....	20
2.7 Python.....	21
2.8 Kraken 2.0.....	22
3. JUSTIFICACIÓN	24
4. HIPOTESIS	25
5. OBJETIVOS	25
5.1 General:	25
5.2 Especifico:	25
6. MATERIALES Y METODOS	26
6.1 Diagrama Experimental	26
6.2 Procesamiento de la muestra	27
6.2.1 Recolección de muestra.	27
6.2.2 Extracción de ADN.....	27
6.2.3 Producto de la PCR de la región 16S.....	27
6.2.4 Integridad del producto de PCR	29
6.2.5 Purificación de la PCR 16S.....	29
6.2.6 Producto de la PCR de la región ITS.....	29
6.2.7 Integridad del producto de PCR ITS.....	31
6.2.8 Purificación de la PCR ITS.	31
6.2.9 Incorporación de códigos de barra al producto de la PCR de la región 16S y ITS.....	31
6.2.9 Confirmación del producto de la PCR de la incorporación de código de barra.	33
6.2.10 Purificación de la incorporación de código de barra	34

6.2.11 Secuenciación.	34
6.3 Análisis Bioinformático.....	35
6.3.1 Procesamiento de datos.	35
6.3.2 Clasificación de la muestra.	36
6.3.3 Comparación de bases de datos.	37
6.3.4 Porcentaje de microorganismos	38
7.RESULTADOS	40
7.1 Porcentaje de microorganismos.....	40
7.2 Comparación de bases de datos.....	44
7.2.1 BACTERIAS	44
7.2.2 HONGOS	49
7.2.3 BACTERIAS	52
7.2.4 HONGOS	54
8. DISCUSIÓN	58
9. CONCLUSIONES	64
10.PERSPECTIVAS	65
11. REFERENCIAS	66
12. ANEXO.....	72
REPOSITORIO VIRTUAL	90

II. INDICE DE FIGURAS

Figura 1. Diagrama de la estructura génica de los rRNAs procariotes..	15
Figura 2. Representación esquemática del gen 18S rRNA, gen 5.8 rRNA y gen 28rRNA, que muestra las regiones intergenicas ITS1 e ITS2.	17
Figura 3 . Formato FastQ.	20
Figura 4. Algoritmo de clasificación de secuencias Kraken 2.0.	23
Figura 5. Estrategia experimental usada para el análisis y procesamiento de datos de secuenciación masiva de las regiones 16S y ITS de microbiomas de muestras de suelo.	26
Figura 6. Clasificación taxonómica de bacterias en la muestra 1.	41
Figura 7. Clasificación taxonómica de hongos en la muestra 1.	42
Figura 8. Clasificación taxonómica porcentual de bacterias en la muestra 1.	43
Figura 9. Clasificación taxonómica porcentual de hongos en la muestra 1.	44
Figura 10. Gráfico comparativo de algoritmos de clasificación taxonómica de bacterias.	55
Figura 11. Gráfico comparativo de algoritmos de clasificación taxonómica de Hongos.	56

III. INDICE DE TABLAS

Tabla 1.Composición De la Mezcla Maestra para PCR.	28
Tabla 2.Protocolo de termociclador para PCR.....	28
Tabla 3.Composición De la Mezcla Maestra para PCR.	30
Tabla 4.Protocolo de termociclador para PCR.....	30
Tabla 5. Combinación de códigos de barra por cada producto de 16S y ITS.....	32
Tabla 6.Protocolo de reacción para la incorporación de código de barra tanto para 16S y ITS.....	33
Tabla 7.Protocolo de termociclador para la incorporación de código de barra.....	33
Tabla 8.Comparación de la Concordancia de Géneros Bacterianos entre Muestras (Muestra 1-5)	46
Tabla 9.Comparación de la Concordancia de Géneros Bacterianos entre Muestras (Muestra 6-10)	47
Tabla 10.Concentrado de la Comparación de la Concordancia de Géneros bacterianos entre Muestras.	48
Tabla 11.Comparación de la Concordancia de Géneros Fúngicos entre Muestras (Muestra 1-5).	50
Tabla 12.Comparación de la Concordancia de Géneros Fúngicos entre Muestras (Muestra 6-10).	51
Tabla 13. Concentrado de la Comparación de la Concordancia de Géneros Fúngicos entre Muestras.	52
Tabla 14.Porcentaje de clasificación total de bacterias por muestra.....	45
Tabla 15.Porcentaje de clasificación total de hongos por muestra.....	49
Tabla 16.Porcentaje de clasificación total de especies de bacterias por muestra.....	53
Tabla 17.Porcentaje de clasificación total de especies de Hongos por muestra.	54
Tabla 18.Variabilidad en la Abundancia de Especies Bacterianas según el Algoritmo de Clasificación.....	57
Tabla 19.Variabilidad en la Abundancia de Especies Fúngica según el Algoritmo de Clasificación.	57

IV. Resumen

Los avances tecnológicos en biología molecular han posibilitado estudios más complejos en la identificación de organismos, especialmente en microbiomas y metagenómica (Meneses, 2011). Las compañías de secuenciación genética ofrecen software propio para taxonomía precisa, pero con costos económicos y de infraestructura significativos. Estos programas son versátiles y exigentes computacionalmente, dificultando su uso en equipos convencionales, lo que plantea desafíos para académicos e investigadores. En este proyecto de investigación, se desarrollaron herramientas de software almacenadas en bancos de datos de libre acceso, destinadas a la categorización de lecturas procedentes de secuenciación masiva en las regiones ITS y 16S del ADN. Estos scripts son de uso accesible y emplearon algoritmos bioinformáticos para análisis precisos, facilitando la colaboración científica y permitiendo análisis avanzados en distintos organismos y ambientes. El objetivo es ofrecer una alternativa libre a programas comerciales, promoviendo la colaboración científica y el conocimiento taxonómico. Se compararon dos algoritmos para identificación y clasificación de ITS y 16S (región V3-V4), resaltando la importancia de bases de datos actualizadas. Los resultados fueron dependientes a la calidad de las bases de datos biológicas. Las aportaciones bioinformáticas desarrolladas respondieron a las necesidades de la investigación, demostrando eficacia en la identificación de 16S con los algoritmos epi2me y Kraken 2.0. El análisis comparativo entre el algoritmo comercial y el algoritmo gratuito reveló que el primero exhibe un mayor rendimiento en términos de la capacidad de clasificación, tanto en lo que respecta al número de organismos asignados por muestra, como a la diversidad de especies identificadas por muestra. Sin embargo, hubo diferencias entre microorganismos clasificados en la región ITS, atribuibles a la validación de bases de datos.

V. Abstract

Technological advances in molecular biology have enabled more complex studies in the identification of organisms, especially in microbiomes and metagenomics (Meneses, 2011). Genetic sequencing companies offer their proprietary software for precise taxonomy, but with significant economic and infrastructure costs. These programs are versatile and computationally demanding, making their use on conventional equipment challenging, which poses challenges for academics and researchers. In this research project, software tools were developed and stored in freely accessible databases, intended for categorizing reads from high-throughput sequencing in the ITS and 16S regions of DNA. These scripts are user-friendly and employ bioinformatic algorithms for accurate analysis, facilitating scientific collaboration and enabling advanced analyses in various organisms and environments. The objective is to provide a free alternative to commercial programs, promoting scientific collaboration and taxonomic knowledge. Two algorithms for ITS and 16S (V3-V4 region) identification and classification were compared, highlighting the importance of updated biological databases. The results depended on the quality of the biological databases. The bioinformatic contributions developed met research needs, demonstrating efficiency in 16S identification with the epi2me and Kraken 2.0 algorithms. Comparative analysis between the commercial and free algorithms revealed that the former exhibits higher performance in terms of classification capacity, both in terms of the number of organisms assigned per sample and the diversity of species identified per sample. However, there were differences in the classification of microorganisms in the ITS region, attributable to database validation.

1. INTRODUCCIÓN

El aumento exponencial en la generación de datos biológicos ha planteado desafíos significativos en su procesamiento y análisis. La capacidad insuficiente para gestionar volúmenes masivos de información se ha convertido en un obstáculo común en la investigación biológica. La bioinformática desempeña un papel fundamental al proporcionar herramientas y métodos para abordar esta dificultad (Breitwiese, 2019).

La investigación en biología contemporánea se enfrenta a una creciente demanda debido a la abrumadora cantidad de datos disponibles, lo que requiere tecnologías avanzadas de la información y una capacidad computacional considerable (Wang, 2019). En este contexto, es crucial aprovechar eficazmente la información cruda obtenida de tecnologías de secuenciación de nueva generación y comprender completamente su composición.

En el ámbito de la bioinformática y la secuenciación masiva, se ha observado un progreso significativo en términos de precisión y rendimiento (Álvarez, 2020). Sin embargo, muchas herramientas bioinformáticas actuales son costosas, especialmente las diseñadas para equipos especializados, lo que subraya la importancia de desarrollar metodologías gratuitas para el análisis y procesamiento de datos biológicos, particularmente en la determinación de la composición microbiana.

La combinación de técnicas de biología molecular y bioinformática se ha destacado por su alta especificidad, sensibilidad y velocidad en la determinación de microbiomas en diversos entornos. Esto la sitúa como una metodología destacada para lograr una identificación rápida y precisa de microorganismos.

En resumen, el diseño de un script en Python que aspire a igualar la capacidad de detección y clasificación de microorganismos de programas como EPI2ME representa un paso importante para abordar los desafíos en el procesamiento de datos biológicos. Este enfoque contribuye al avance de la bioinformática y la investigación en biología al proporcionar una herramienta accesible y efectiva para el análisis de datos biológicos.

2. MARCO TEORICO

La principal limitación histórica para el estudio de la diversidad microbiana en suelos radica en la dificultad de aislar los microorganismos. Se estima que solo alrededor del 1% de la población bacteriana presente en los suelos puede ser cultivada utilizando prácticas de laboratorio estándar, pero no se sabe si ese 1% es representativo (Nogales, 2005). En la última década, esta limitación ha sido un obstáculo para investigar los microbiomas en los ecosistemas, ya que la incapacidad de cultivarlos dificultaba su detección. Sin embargo, el desarrollo de nuevas técnicas de Biología Molecular en las últimas décadas ha beneficiado a la Ecología Microbiana, contribuyendo significativamente al estudio y generación de conocimiento sobre la composición de las comunidades microbianas sin la necesidad de cultivar los microorganismos (Valenzuela, 2015). Los marcadores moleculares se han demostrado de gran utilidad para describir y caracterizar las comunidades microbianas, especialmente los organismos no cultivados (Hébert *et al.*, 2003). Las nuevas técnicas de secuenciación han contribuido al incremento exponencial en el registro de secuencias, aunque parciales, de regiones específicas como códigos de barras para microorganismos. Como resultado, ha sido necesario revisar los conceptos y métodos de clasificación taxonómica para estos organismos (Tanabe, 2013). El manejo y análisis de una gran cantidad de información genética han impulsado el desarrollo de bases de datos específicas, algoritmos y herramientas computacionales especializadas para comparar miles de secuencias similares y realizar la asignación taxonómica (Hernández, 2020). Estos avances han permitido un mejor entendimiento de la diversidad microbiana en suelos y otros ambientes, superando las limitaciones impuestas por la dificultad de cultivar los microorganismos y abriendo nuevas oportunidades de investigación en el campo de la Ecología Microbiana.

2.1 Gen ribosomal 16S rARN

La región 16S es un polirribonucleótido de aproximadamente 1500 nucleótidos de longitud, codificado por el gen *rrs*, también conocido como ADN ribosomal 16S (ADNr 16S). La relevancia de esta región ribosomal radica en su disposición de bases nitrogenadas, lo cual permite llevar a cabo una clasificación filogenética y taxonómica de las bacterias (Johnson, 2019).

Al igual que cualquier secuencia de nucleótidos de cadena sencilla, el ARNr 16S se pliega y adquiere una estructura secundaria que incluye segmentos de doble cadena capaces de formar

asas y hélices. Esta molécula se ha reconocido como un marcador universal poderoso, ya que se encuentra presente en todos los organismos conocidos. Su estructura parece mantenerse estable durante largos períodos de tiempo, y dado que su función no ha experimentado cambios, las alteraciones en la secuencia probablemente sean aleatorias. Sin embargo, los ARNr poseen suficiente variabilidad para diferenciar no solo organismos muy divergentes, sino también aquellos más cercanos. En algunos casos, es posible distinguir especies. El tamaño relativamente largo del ARNr 16S (1500 nucleótidos) minimiza las fluctuaciones estadísticas, y su conservación en la estructura secundaria favorece el alineamiento preciso durante la comparación de secuencias (Johnson, 2019). El ARNr 16S contiene nueve regiones (V1-V9) menos conservadas o hipervariables, las cuales proporcionan la mayor información útil para estudios filogenéticos y taxonómicos. Las regiones conservadas son de gran ayuda para diseñar oligonucleótidos que permitan amplificar las diversas regiones hipervariables de la gran mayoría de los ARNr 16S de los microorganismos presentes en una comunidad (Johnson, 2019).

En la Figura 1 se muestra una representación gráfica del gen 16SrRNA (New England Journal of Medicine, 2023) , donde se puede visualizar las dos subunidades que componen el gen 16S, así como las nueve regiones hipervariables que conforman esta estructura genética. Estas regiones hipervariables son de especial relevancia en la investigación microbiológica, ya que son sometidas a amplificación con el propósito de extraer la información genética necesaria para la realización de estudios de índole filogenética y taxonómica. Este enfoque molecular facilita la caracterización y clasificación precisa de las especies bacterianas, contribuyendo significativamente al entendimiento de su diversidad y relaciones evolutivas.

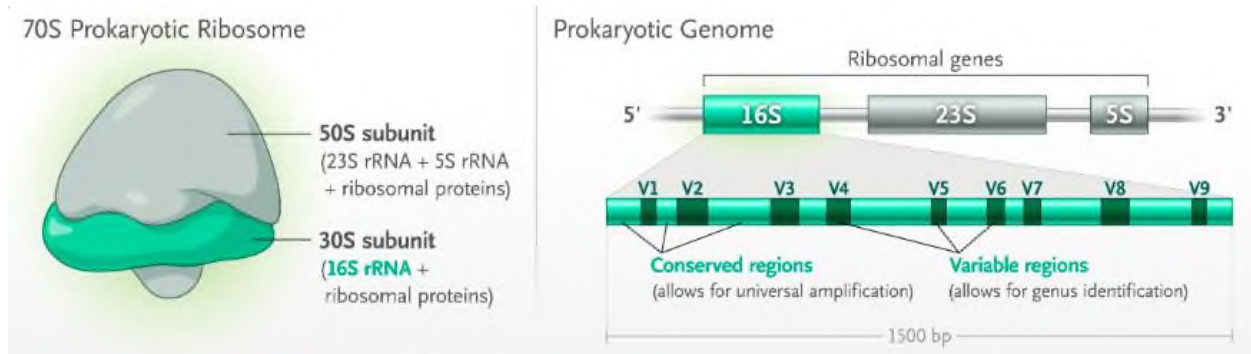


Figura 1 . Diagrama de la estructura génica de los rRNAs procariontes. Donde se muestra las dos subunidades del gen 16S y las nueve regiones hipervariables que lo componen. Tomado de (New England Journal of Medicine, 2023).

El uso de oligonucleótidos ha facilitado la detección y el análisis de secuencias; sin embargo, algunos autores señalan su limitación para detectar un número considerable de especies bacterianas no cultivadas presentes en muestras ambientales (Huws *et al.*, 2007). Existen varios estudios que informan sobre la cobertura de iniciadores universales y sus combinaciones, utilizando secuencias obtenidas de estudios metagenómicos (Liu *et al.*, 2008). La tasa de clasificación y la precisión varían ampliamente entre entornos y regiones de secuencia, debido a varias razones entre las cuales están: la base de datos de referencia proporciona diferentes niveles de cobertura para cada entorno, ningún iniciador es verdaderamente "universal", pero cuanto mayor sea la cobertura de una muestra biológica, mayor será la identificación precisa de los microorganismos presentes en el medio ambiente (Soergel, 2012).

2.2 Región ITS como “código de barras” para la identificación fúngica.

En organismos eucariotas, los genes responsables de la síntesis de los ARN ribosomales están dispuestos en configuraciones genómicas que comprenden unidades transcripcionales repetitivas, a saber, los genes 18S, 5.8S y 28S. Estas unidades, constituyentes fundamentales del ribosoma, son elementos esenciales en la maquinaria molecular encargada de la síntesis de proteínas. En adición a estos componentes estructurales, se hallan presentes dos regiones variables conocidas como espaciadores transcritos internos (ITS1 e ITS2), cuyo carácter dinámico es crucial para la diferenciación de especies y la taxonomía. Además de las regiones ITS, se distinguen dos secuencias adyacentes externas denominadas ETS (External Transcribed Spacer, en

inglés), que desempeñan un papel funcional en la transcripción y procesamiento de los precursores ribosomales, añadiendo una capa de complejidad en la regulación de la expresión génica en organismos eucariotas. (Raja *et al.*, 2017).

Las regiones ITS en los genomas de organismos eucariotas pueden variar considerablemente en tamaño y secuencia (Stajich *et al.*, 2009). Esta variabilidad ha permitido validarlas como "código de barras" (Buée *et al.*, 2009; Wang *et al.*, 2011) para la identificación de hongos, ya que posibilita la identificación a nivel de género y especie (Seifert, 2009; Tedersoo *et al.*, 2010, Schoch *et al.*, 2012).

En contraste, las subunidades 18S, 5.8S y 28S no permiten la identificación a nivel de especie debido a que se consideran regiones conservadas con poca variabilidad en su secuencia, lo que dificulta la clasificación y diferenciación de diferentes tipos de hongos (Raja *et al.*, 2017).

La Figura 2 presenta un esquema de los genes 18S rRNA, 5.8 rRNA y 28 rRNA, fundamentales en la maquinaria molecular ribosomal. Este gráfico destaca las áreas antigénicas denominadas Espaciadores Transcritos Internos 1 (ITS1) y Espaciadores Transcritos Internos 2 (ITS2). Dichas regiones desempeñan un papel de suma importancia en la identificación taxonómica a nivel de género y especie en el ámbito de la biología molecular y la taxonomía. En determinadas instancias, es posible llevar a cabo la identificación de la especie con un grado de certeza notable. Sin embargo, conviene subrayar que la eficacia de este proceso de identificación se encuentra intrínsecamente vinculada a la integridad y la calidad de la información contenida en las bases de datos que albergan secuencias correspondientes a las regiones ITS1 e ITS2. Esta metodología de identificación se basa en el análisis comparativo de las secuencias de dichas regiones con aquellas disponibles en bases de datos taxonómicas de referencia, lo que implica una evaluación rigurosa de la coherencia y la semejanza genética entre las secuencias bajo análisis y sus contrapartes de referencia.



Figura 2. Representación esquemática del gen 18S rRNA, gen 5.8 rRNA y gen 28rRNA, que muestra las regiones intergenicas ITS1 e ITS2. En este esquema visual, se resaltan las zonas denominadas ITS1 y ITS2, las cuales desempeñan un papel crucial en la identificación taxonómica dentro de los campos de biología molecular y taxonomía. Extraído de (White, 1990).

El uso de las regiones ITS variables en el estudio de la diversidad fúngica ha requerido el diseño de oligonucleótidos universales que permitan la amplificación de la región de interés en todos los taxones (Bidartondo, 2005). White et al., (1990) diseñaron los primeros oligonucleótidos universales (ITS1/ITS2 e ITS3/ITS4) para lograr la amplificación de estas regiones, lo cual condujo al diseño de oligonucleótidos universales y específicos para cada taxón de hongos.

Aunque la mayoría de los programas bioinformáticos están orientados al análisis de la diversidad fúngica (Nilsson *et al.*, 2009), actualmente se está mejorando y depurando las bases de datos para lograr una clasificación más precisa de los hongos.

La identificación preliminar de fitopatógenos basada en estudios taxonómicos tradicionales de hongos permite determinar posibles efectos negativos o positivos debido a las interacciones biológicas entre las plantas y los hongos (Aoki *et al.*, 2005), pero requiere tiempo y personal capacitado para realizar una clasificación adecuada. Sin embargo, los métodos moleculares agilizan la identificación al analizar regiones específicas del genoma del organismo. Las secuencias de los espaciadores transcritos internos (ITS) y regiones como el factor de elongación 1-alfa (TEF 1- α) son las más utilizadas para diferenciar y clasificar los hongos (Fernández-Orduño *et al.*, 2010) y, una vez identificados, tomar medidas para controlar su proliferación.

2.3 Secuenciación masiva en la detección de microbiotas

La secuenciación de alto rendimiento ha revolucionado la generación de datos biológicos, produciendo volúmenes masivos que superan la capacidad de las computadoras convencionales (López, 2016). Esto ha elevado la informática a un papel crítico en la gestión y análisis de estos

datos. A diferencia de los sistemas de secuenciación tradicionales, las plataformas modernas pueden generar millones de fragmentos de ADN en un solo proceso, con tiempos récord y costos cada vez más bajos. Sin embargo, la manipulación de estas secuencias aisladas requiere el uso de potentes herramientas informáticas para su alineación y ensamblaje (Rodríguez-Santiago, 2012).

A pesar del avance significativo en la microbiología del suelo gracias a las técnicas de biología molecular, las interrogantes que enfrentan los microbiólogos del suelo siguen siendo complejas. La introducción de técnicas de secuenciación y análisis de genomas ha generado una cantidad masiva de información en los últimos años, superando en magnitud a lo obtenido en las décadas anteriores. La metagenómica, el estudio global de los genomas de microorganismos en comunidades, ha tenido éxito en comunidades microbianas del suelo (Rondón *et al.*, 2000). Estimaciones conservadoras sugieren la presencia de entre 6,000 y 10,000 genomas diferentes en suelos orgánicos no perturbados y un número menor en suelos agrícolas o contaminados con metales pesados (Marcano *et al.*, 2015). Esto refleja la magnitud de la nueva información generada.

En el campo de la microbiología del suelo, la metagenómica se ha centrado en la detección de biocatalizadores, compuestos bioactivos y la caracterización de genomas de microorganismos no cultivados. Estos avances ayudarán a comprender el papel de estos microorganismos en el ecosistema del suelo. Sin embargo, el manejo de las enormes cantidades de secuencias obtenidas es un desafío. Para abordarlo, se han desarrollado diversos programas y paquetes informáticos como BLAST, Bioconductor, NCBI Entrez Utilities, FASTA y Bowtie (Wang, 2019), que permiten el análisis de secuencias, ya sea de un organismo, un grupo de organismos o una o varias muestras.

2.4 Epi2me

Epi2me es una plataforma desarrollada por Oxford Nanopore Technologies (ONT) para el análisis de datos de secuenciación de nanoporos (Kerkhof, 2021). Ofrece herramientas y flujos de trabajo para diversas aplicaciones, como el análisis de genomas, la detección de modificaciones epigenéticas, la identificación de variantes genéticas y la metagenómica. Epi2me facilita la identificación de genes, elementos reguladores y especies en muestras de ADN, y se utiliza en una amplia variedad de investigaciones científicas y aplicaciones de diagnóstico. Su versatilidad lo hace esencial en la genómica y la biología molecular.

Es importante mencionar que, según información proporcionada, el costo de una membresía de uso de Epi2me se estima en aproximadamente 1000 dólares al año, equivalente a unos 20,000 pesos mexicanos al tipo de cambio actual (Oxford Nanopore Technologies,2023). Esta estimación de costos es relevante para aquellos que consideran utilizar esta plataforma en términos de su viabilidad económica y acceso a sus capacidades de análisis de datos de secuenciación de nanoporos. Los costos asociados pueden estar sujetos a variaciones y dependen de factores como la ubicación geográfica y la elección de licencia específica, por lo que se recomienda verificar los precios actuales y las opciones de membresía directamente a través de la fuente oficial o el proveedor de servicios correspondiente.

2.5 Formato FAST 5

El formato FAST5 utilizado por Oxford Nanopore Technologies (ONT) se basa en el sistema HDF5 (Hierarchical Data Format 5), que es un enfoque altamente flexible para el almacenamiento y gestión de datos (Huges, 2017). HDF5 es un formato de archivo binario que permite una estructura jerárquica de datos, dividiéndolos en grupos y conjuntos de datos. Esto se traduce en la categorización y organización de datos similares en categorías cuando se trata de datos en grupo. Por otro lado, cuando se hace referencia a data sets, nos estamos refiriendo a la recopilación y almacenamiento estructurado de datos con un propósito específico para su posterior análisis o procesamiento.

Los grupos en HDF5 actúan como contenedores que pueden alojar otros grupos y data sets, mientras que los data sets son arreglos multidimensionales de información. Esta jerarquía proporciona una eficiente forma de organizar diversos tipos de datos en un solo archivo. Uno de los beneficios destacados de HDF5 radica en su versatilidad para manejar datos de múltiples tipos y tamaños, incluyendo datos numéricos, de texto, imágenes y sonido, en una única estructura de archivo. Además, HDF5 permite la compresión de datos, lo que reduce el tamaño de los archivos sin pérdida de información esencial.

HDF5 también ofrece una interfaz de programación de aplicaciones (API) que facilita el acceso y la manipulación de los datos almacenados en archivos HDF5. Esta API está disponible en varios lenguajes de programación, como C, C++, Python y Java, lo que simplifica su integración en diversas aplicaciones y entornos.

En el caso específico del formato FAST5 de ONT, que utiliza HDF5 como base, esta elección proporciona una gran flexibilidad en la gestión de los datos generados por los secuenciadores de nanoporos de ONT. Los archivos FAST5 pueden almacenar lecturas de secuencias, información de calidad, metadatos experimentales y otros datos relevantes en una única estructura, lo que simplifica significativamente su posterior análisis y procesamiento (Gamaarachchi *et al*, 2022).

2.6 Formato FASTQ

El formato FASTQ es esencial en bioinformática y es un estándar en la secuenciación de nueva generación (NGS). Se basa en texto y tiene la capacidad de almacenar tanto la secuencia biológica, que suele ser una secuencia de nucleótidos, como las puntuaciones de calidad relacionadas con cada posición de la secuencia. Esta combinación de secuencia y puntuaciones de calidad es fundamental para lograr una interpretación precisa y confiable de los datos producidos por los secuenciadores de NGS (Wang *et al.*, 2021).

```
@Secuencia_1
AGCTGATCGATCGTAGCTAGCTAGCAT
+
! ! * ((( (**+))%%%++) (%%%) .1
@Secuencia_2
TGGCAATCGATGCTAGCTGATCGTAG
+
%%%)++) )%%%+**)))+** )%%%1
```

Figura 3 . Formato FastQ. El formato FASTQ es un tipo de archivo de texto utilizado para almacenar secuencias biológicas, como ácidos nucleicos, junto con sus valores de calidad. Tomado de (Wang *et al.*, 2021).

@Secuencia_1 y @Secuencia_2: Estas líneas comienzan con el símbolo "@" seguido de un identificador único para cada secuencia. Esta línea indica a qué secuencia pertenecen las siguientes líneas de datos y calidad.

AGCTGATCGATCGTAGCTAGCTAGCAT y TGGCAATCGATGCTAGCTGATCGTAG: Estas líneas contienen las secuencias nucleotídicas reales obtenidas durante la secuenciación. Son las representaciones de las letras A (adenina), G (guanina), C (citosina) y T (timina) que componen el ADN o ARN.

+: Esta línea comienza con el símbolo "+" y generalmente se utiliza para separar las secuencias de calidad de las secuencias de nucleótidos. A menudo, esta línea se deja en blanco o contiene información adicional sobre la secuencia.

!*(((((*+))%%%+)(%%%).1 y %%%%+))%%%+**)))+**))%%%1:** Estas líneas representan los valores de calidad asociados con cada base de la secuencia correspondiente. Los valores de calidad están codificados en ASCII, donde caracteres diferentes representan diferentes niveles de calidad. Cuanto mayor sea el valor ASCII, mejor será la calidad de la base correspondiente. Por ejemplo, "!" tiene un valor ASCII más bajo que "*", lo que indica una calidad más alta.

Estas puntuaciones de calidad se presentan como caracteres ASCII, donde cada carácter codifica una puntuación numérica específica que refleja la confiabilidad de cada base en la secuencia. La relación entre los caracteres ASCII y las puntuaciones de calidad se establece en una tabla de codificación, como la tabla Phred. El valor numérico de la puntuación de calidad se obtiene restando un valor fijo (generalmente 33 o 64) del valor ASCII correspondiente al carácter (Wang *et al.*, 2021).

El formato FASTQ ofrece una representación compacta y eficiente de los datos de secuenciación al combinar la secuencia y las puntuaciones de calidad en un solo archivo de texto. Esto simplifica el análisis y el procesamiento subsiguientes de los datos mediante herramientas bioinformáticas. Además, permite la capacidad de almacenar múltiples secuencias en un único archivo, lo que facilita la manipulación y transferencia de datos de manera efectiva (Hernández, 2020).

2.7 Python

Python es un lenguaje de programación de alto nivel que goza de una amplia adopción en el ámbito de la informática y la programación. Se destaca por su sintaxis clara y legible, lo que lo convierte en una elección idónea tanto para aquellos que están dando sus primeros pasos en la programación como para programadores con experiencia consolidada. La versatilidad de Python

se manifiesta en su aplicación en una diversidad de campos, abarcando desde el desarrollo web hasta el análisis de datos y el aprendizaje automático. Su capacidad para ser empleado en una variedad tan extensa de aplicaciones lo consolida como una herramienta esencial en el repertorio de programadores y científicos de datos (Prokopyev *et al.*, 2020).

2.8 Kraken 2.0

Kraken 2.0 es un sistema desarrollado para abordar la tarea de asignar etiquetas taxonómicas a secuencias cortas de ADN, especialmente aquellas obtenidas en estudios metagenómicos. El objetivo principal de Kraken 2.0 es realizar una clasificación taxonómica precisa y eficiente de secuencias de ADN.

El sistema utiliza una estrategia basada en la construcción de una base de datos de referencia que contiene información taxonómica y secuencial de diversos organismos, como bacterias, virus, eucariotas y arqueas. Esta base de datos de referencia se compone de secuencias genómicas conocidas y está organizada en una estructura de árbol de taxonomía (Wood & Salzberg, 2014).

El proceso de clasificación con Kraken 2.0 se realiza comparando las secuencias de entrada con la base de datos de referencia utilizando algoritmos eficientes de coincidencia de secuencias. El sistema utiliza un enfoque de mapeo rápido basado en k-mers, que son secuencias cortas de longitud fija. Los k-mers se generan a partir de las secuencias de entrada y se comparan con los k-mers presentes en la base de datos de referencia. La coincidencia de k-mers permite asignar etiquetas taxonómicas a las secuencias de entrada.

En la Figura 4 se muestra el algoritmo de clasificación de secuencias Kraken 2.0 en el cual indica que, para clasificar una secuencia, cada k-mer en la secuencia se asigna al ancestro común más bajo (LCA) de los genomas que contienen ese k-mer en una base de datos. Los taxones asociados con los k-mers de la secuencia, así como los ancestros de los taxones, forman un subárbol podado del árbol de taxonomía general, que se utiliza para la clasificación. En el árbol de clasificación, cada nodo tiene un peso igual al número de k-mers en la secuencia asociada con el taxón del nodo. Cada ruta de raíz a hoja (RTL) en el árbol de clasificación se puntúa sumando todos los pesos de la ruta, y la ruta RTL máxima en el árbol de clasificación es la ruta de clasificación (nodos resaltados en amarillo). La hoja de esta ruta de clasificación (la hoja naranja situada más a la izquierda en el árbol de clasificación) es la clasificación utilizada para la secuencia de consulta.

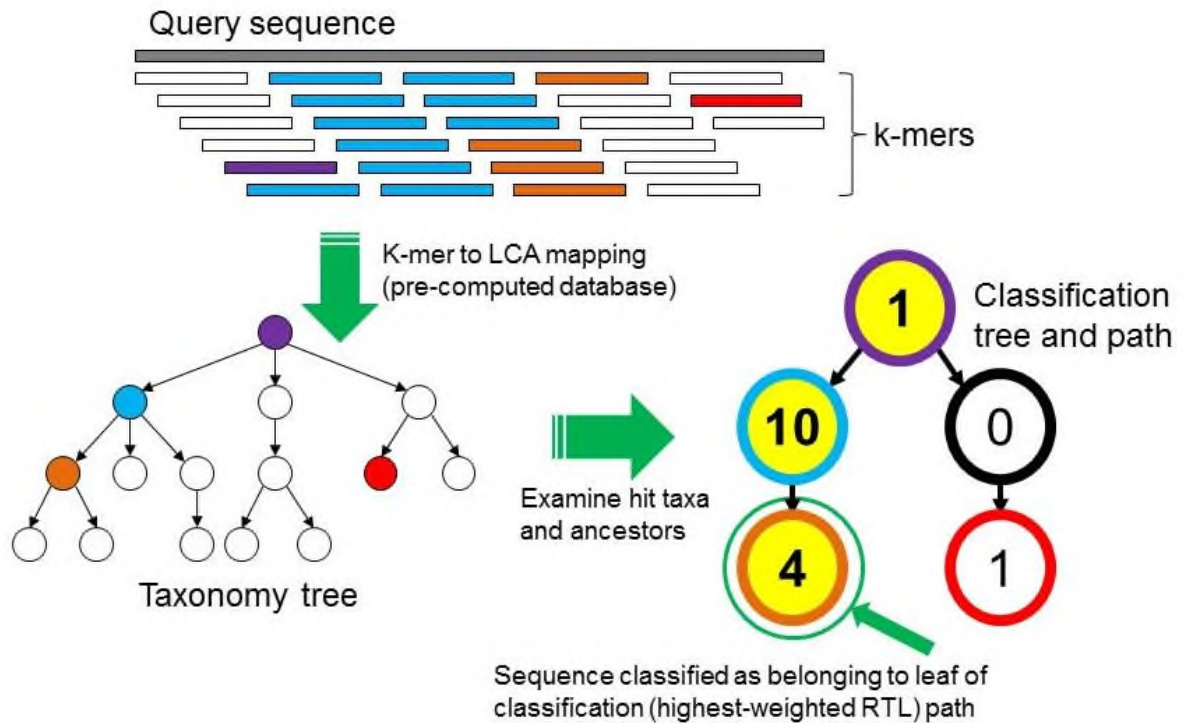


Figura 4. Algoritmo de clasificación de secuencias Kraken 2.0. El algoritmo Kraken 2.0 clasifica secuencias asignando k-mers a ancestros comunes y utiliza el peso de los k-mers para determinar la clasificación final basada en la ruta de raíz a hoja con la mayor puntuación en el árbol de taxonomía. Traducido de (Wood & Salzberg, 2014).

La clasificación taxonómica realizada por Kraken 2.0 es altamente precisa y se basa en la comparación de secuencias específicas en lugar de la alineación global. Esto permite un análisis rápido y escalable de grandes volúmenes de datos metagenómicos.

Además, Kraken 2.0 proporciona una clasificación jerárquica, lo que significa que no solo asigna una única etiqueta taxonómica a una secuencia, sino que también identifica los niveles de taxonomía superiores en el árbol de taxonomía.

La utilidad y eficacia de Kraken 2.0 han sido demostradas en diversos estudios y ha sido ampliamente adoptado por la comunidad científica para el análisis taxonómico en estudios metagenómicos. Además, se han desarrollado versiones mejoradas de Kraken 2.0, como Kraken 2.0, que ofrecen un rendimiento y una precisión aún mayores (Ravelo *et al.*, 2023).

3. JUSTIFICACIÓN

Las tecnologías disponibles en el campo de la biología molecular han abierto nuevas posibilidades para la realización de estudios más complejos en la identificación de organismos, especialmente en el ámbito de los microbiomas y la metagenómica. Cada compañía de secuenciación genética cuenta con su propio software para el procesamiento de datos de secuenciación masiva, con el objetivo de llevar a cabo clasificaciones taxonómicas precisas.

Estos programas informáticos suelen tener un costo, tanto económico como de infraestructura, debido a los derechos de autor y a las amplias características que ofrecen. De hecho, estos programas son tan versátiles y demandantes en términos de recursos computacionales que resulta difícil utilizarlos en equipos informáticos convencionales, lo que representa un desafío tanto para la comunidad académica como para los investigadores.

Por lo tanto, en el presente trabajo de investigación se propone desarrollar una serie de scripts gratuitos que contribuyan a la clasificación taxonómica de las lecturas de las regiones ITS y 16S obtenidas mediante secuenciación masiva. Estos scripts, o programas de computadora, fueron diseñados para ser accesibles y utilizables por la comunidad científica sin restricciones económicas, permitiendo así el acceso a herramientas de clasificación taxonómica de manera más amplia y equitativa.

El objetivo principal de estos scripts gratuitos es facilitar el análisis de datos de secuenciación masiva y proporcionar resultados de clasificación taxonómica confiables. Para lograr esto, se emplearon algoritmos y enfoques bioinformáticos adecuados, asegurando la precisión y la eficiencia en el procesamiento de las secuencias.

Al ofrecer una alternativa gratuita y accesible a los programas comerciales, se espera fomentar la colaboración científica y el avance del conocimiento en el campo de la clasificación taxonómica. Además, la disponibilidad de estos scripts gratuitos permitirá a investigadores con recursos limitados llevar a cabo análisis más sofisticados y realizar estudios taxonómicos en una amplia gama de organismos y ambientes.

4. HIPOTESIS

El diseño de un script en lenguaje Python presentara una capacidad equiparable a la del programa EPI2ME en términos de detección y clasificación de microorganismos.

5. OBJETIVOS

5.1 General:

Diseñar un conjunto de scripts para realizar la clasificación de lecturas de secuenciación masiva en formato fastq de las regiones 16S y ITS provenientes de secuenciación por la tecnología de nanoporos.

5.2 Especifico:

1. Diseño de algoritmo para separar secuencias dependiendo de las muestras.
2. Realizar la clasificación taxonómica de los microorganismos presentes en las lecturas.
3. Comparación entre los resultados obtenidos en base al porcentaje de clasificación taxonómica al emplear EPI2ME y los resultados derivados de la ejecución de los scripts creados en Python.

6. MATERIALES Y METODOS

6.1 Diagrama Experimental

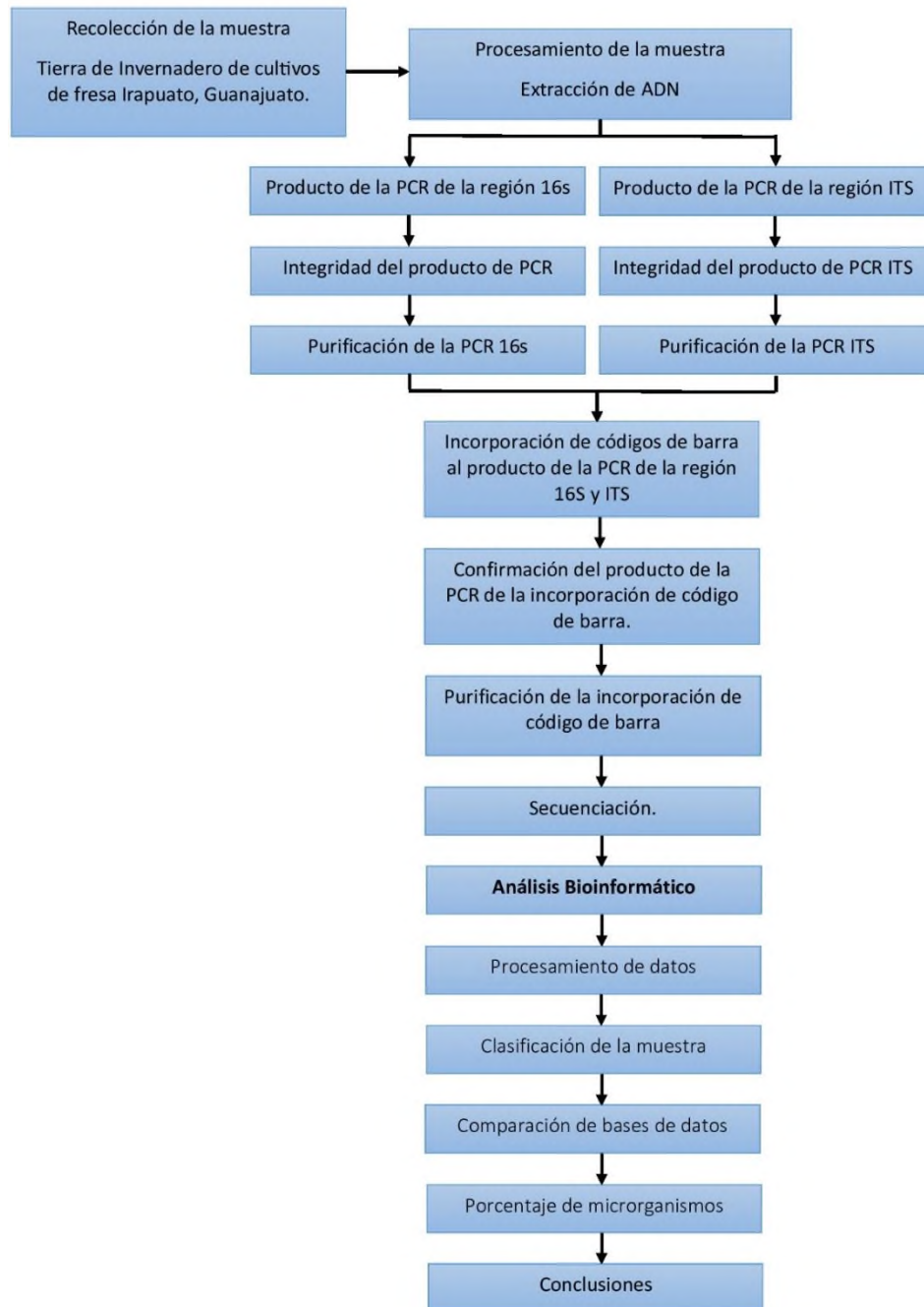


Figura 5. Estrategia experimental usada para el análisis y procesamiento de datos de secuenciación masiva de las regiones 16S y ITS de microbiomas de muestras de suelo.

6.2 Procesamiento de la muestra

6.2.1 Recolección de muestra.

Las muestras se recolectaron en un vivero de fresa localizado en el Irapuato en el estado de Guanajuato, las muestras fueron recolectadas mediante la técnica de zigzag, se tomaron un total de 10 muestras de suelo, cada muestra se tomó a 2 cm de la superficie de suelo, cada muestra se recolectó un total de 30 gramos, recolectada en un tubo falcon de 50 mililitros. Una vez recolectadas las muestras se almacenaron a 4°C para su procesamiento. Las muestras fueron recolectadas el 15 de noviembre del 2019.

6.2.2 Extracción de ADN

Se tomaron 750 mg de cada muestra y se colocaron en un tubo de 2 ml. Se añadieron 850 microlitros de medio de transporte con formulación STUART (Lifeder, 2023) y se mezclaron vigorosamente cada tubo eppendorf durante 10 segundos. Luego, se extrajeron 750 microlitros de la muestra homogeneizada y se transfirieron a un nuevo microtubo de 1.5 mL. La muestra se centrifugó a 1,000 g durante 1 minuto. Se recuperaron 600 µL del sobrenadante y se transfirieron a un nuevo microtubo de 1.5 mL. Este nuevo tubo se centrifugó a 10,000 g durante 5 minutos, y se descartó el sobrenadante. Se resuspendió el pellet resultante con 50 µL de agua libre de nucleasas hasta obtener una mezcla homogénea. El tubo se congeló a -20°C, lo que tomó aproximadamente 10 minutos. Una vez que la muestra estuvo congelada, se sometió a un choque térmico mediante una incubación a 95°C durante 10 minutos en un termociclador. Posteriormente, se centrifugó a 1,000 g durante 3 minutos y se recuperaron 30 µL del sobrenadante en un microtubo de 0.2 mL.

6.2.3 Producto de la PCR de la región 16S

Para la PCR de la región 16S, se empleó el kit Direct PCR Máster de la marca Jena Bioscience con número de catálogo PCR-111S. La reacción de PCR se llevó a cabo en un termociclador de la marca Applied Biosystems. Los oligonucleótidos utilizados fueron el 27F (5'-AGAGTTTGATCCTGGCTCAG-3') y el 1492R (5'-GGTTACCTTGTTACGACTT-3'), diseñados para amplificar las zonas V1-V9 de la región 16S (Sedlar, 2018). Las reacciones se realizaron en un volumen final de 20 µL, según se detalla en la tabla 1. El protocolo para la reacción siguió las pautas propuestas por la casa comercial para productos mayores de 1000 pb, tal como se indica en la tabla 2.

Tabla 1.Composición De la Mezcla Maestra para PCR.

Componentes	Volumen	Concentración
Direct PCR máster	10 μ L	
Oligonucleótidos 27F	1 μ L	10 μ M
Oligonucleótidos 1492R	1 μ L	10 μ M
Muestras	2 μ L	10-30 μ M
Agua libre de ADNasa	6 μ L	
Total	20 μ L	

Tabla 2.Protocolo de termociclador para PCR.

Temperaturas	Tiempo	Ciclos
95°C	300s	1
95°C	20s	35
56.8°C	30s	
72°C	90s	
72°C	120s	1

6.2.4 Integridad del producto de PCR

Se llevó a cabo una electroforesis en gel de agarosa al 1.5%. En cada pozo del gel se cargaron 3 μL de cada muestra junto con 2 μL del buffer de carga azul de bromofenol. Este buffer de carga estaba previamente teñido con bromuro de etidio a una concentración de 10 mg/mL. Luego, el gel se visualizó utilizando luz ultravioleta mediante un fotodocumentador Ingenius.

6.2.5 Purificación de la PCR 16S.

Para eliminar productos secundarios de la PCR con un tamaño inferior a 900 pb, se tomaron 17 μL del producto de la PCR de la región 16S y se les añadieron 13.6 μL de perlas magnéticas Agencourt AMPure XP mediante pipeteo. La mezcla se dejó reposar durante 2 minutos a temperatura ambiente. Luego, se colocó la muestra en una gradilla magnética y se mantuvo allí hasta que la solución se aclaró y se formó un pellet con las perlas magnéticas, un proceso que duró de 3 a 5 minutos.

Una vez que se formó el pellet, se retiró cuidadosamente el sobrenadante desde el lado opuesto al pellet con una pipeta, con precaución para evitar que el pellet se resuspendiera. Se realizaron dos lavados con 200 μL de etanol al 80% en cada lavado, añadiendo el etanol al lado opuesto al pellet. Luego, se realizó un breve giro durante 2 segundos y se dejó la muestra a temperatura ambiente para permitir que el etanol residual se evaporara, lo cual llevó de 3 a 5 minutos.

Después, se añadieron 13 μL de agua libre de nucleasas directamente al pellet y se mezclaron mediante pipeteo, realizando 20 ciclos. Se dejó reposar la muestra a temperatura ambiente durante dos minutos antes de volver a colocarla en la gradilla magnética. Se esperaron otros 2 minutos hasta que la fase acuosa estuvo completamente transparente, momento en el cual se recuperaron 12 μL de la fase acuosa sin tocar el pellet, y se colocaron en un nuevo microtubo de 0.2 mL.

6.2.6 Producto de la PCR de la región ITS

Para la PCR de la región ITS, se empleó el kit Direct PCR Máster de la marca Jena Bioscience con número de catálogo PCR-111S. La reacción de PCR se llevó a cabo en un termociclador de la marca Applied Biosystems modelo veriti. Se utilizaron los oligonucleótidos ITS1 (5'TCCGTAGGTGAACCTGCGG 3') y ITS2 (5'TCCTCCGCTTATTGATATGC3') (White *et al.*, 1990) (Gardes & Bruns, 1993). Las reacciones se realizaron en un volumen final de 20 μL , como se especifica en

la tabla 3. El protocolo para la reacción siguió las pautas propuestas por la casa comercial en el manual para productos con un tamaño menor de 1000 pb, tal como se describe en la tabla 4.

Tabla 3. Composición De la Mezcla Maestra para PCR.

Componentes	Volumen	Concentración
Direct PCR máster	10 μ L	
Oligonucleótidos ITS1	1 μ L	10 μ M
Oligonucleótidos ITS2	1 μ L	10 μ M
Muestras	2 μ L	10-30 μ M
Agua libre de ADNasa	6 μ L	
Total	20 μ L	

Tabla 4. Protocolo de termociclador para PCR.

Temperaturas	Tiempo	Ciclos
95°C	300s	1
95°C	20s	35
56.8°C	30s	
72°C	60s	
72°C	120s	1

6.2.7 Integridad del producto de PCR ITS

Se llevó a cabo una electroforesis en gel de agarosa al 1.5%. En cada pozo del gel se cargaron 3 μL de cada muestra junto con 2 μL del buffer de carga azul de bromofenol. Este buffer de carga fue teñido previamente con bromuro de etidio a una concentración de 10 mg/mL. Luego, el gel se visualizó utilizando luz ultravioleta mediante un fotodocumentador Ingenius.

6.2.8 Purificación de la PCR ITS.

Para eliminar productos secundarios de la PCR con un tamaño inferior a 500 pb, se tomaron 17 μL del producto de la PCR de la región ITS y se les añadieron 9.6 μL de perlas magnéticas Agencourt AMPure XP mediante pipeteo. La mezcla se dejó reposar durante 2 minutos a temperatura ambiente. Posteriormente, se colocó la muestra en una gradilla magnética y se mantuvo allí hasta que la solución se aclaró y se formó un pellet con las perlas magnéticas, un proceso que duró de 3 a 5 minutos.

Una vez que se formó el pellet, se retiró cuidadosamente el sobrenadante desde el lado opuesto al pellet con una pipeta, con precaución para evitar que el pellet se resuspendiera. Se realizaron dos lavados con 200 μL de etanol al 80% en cada lavado, añadiendo el etanol al lado opuesto al pellet. Luego, se realizó un breve giro durante 2 segundos y se dejó la muestra a temperatura ambiente para permitir que el etanol residual se evaporara, lo cual llevó de 3 a 5 minutos.

Después, se añadieron 13 μL de agua libre de nucleasas directamente al pellet y se mezclaron mediante pipeteo, realizando 20 ciclos. Se dejó reposar la muestra a temperatura ambiente durante dos minutos antes de volver a colocarla en la gradilla magnética. Se esperaron otros 2 minutos hasta que la fase acuosa estuvo completamente transparente, momento en el cual se recuperaron 12 μL de la fase acuosa sin tocar el pellet, y se colocaron en un nuevo microtubo de 0.2 mL.

6.2.9 Incorporación de códigos de barra al producto de la PCR de la región 16S y ITS.

Al producto de PCR purificado tanto para la región ITS como 16S, se le añadió un código de barra específico para cada muestra mediante PCR. Los códigos de barra utilizados se detallan en la tabla 3. Para la preparación de la PCR para agregar los códigos de barra a la región 16S, se empleó el kit Direct PCR Máster de la marca Jena Bioscience con número de catálogo PCR-111S. La reacción de PCR se realizó en un termociclador de la marca Applied Biosystems.

Para la región 16S, se utilizaron los oligonucleótidos 27FCBN (5'CBNFAGAGTTTGATCCTGGCTCAG-3') y 1492RCBN (5'CBNRGGTTACCTTGTTACGACTT-3'), donde CBNF y CBNR corresponden al código de barra de cada muestra. Para la región ITS, se usaron ITS1CBNF (5'CBNRTCCGTAGGTGAACCTGCGG 3') y ITS2CBNR (5'CBNRTCCTCCGCTTATTGATATGC3'), y los códigos de barra para las muestras se presentan en la tabla 4.

Las reacciones se llevaron a cabo en un volumen final de 25 µL, como se especifica en la tabla 5. El protocolo para la reacción siguió las pautas propuestas por la casa comercial para productos con un tamaño mayor de 1000 pb, tal como se describe en la tabla 6.

Tabla 5. Combinación de códigos de barra por cada producto de 16S y ITS.

Muestra	Código de barra 5 (CBNF)'	Código de barra 3 (CBNR)'
CB1	GACGCCTAAGCTAACCCGGC	CTGCGGATTCGATTGGGCCG
CB2	ACCCCGACACACCATAGGCG	TGGGGCTGTGTGGTATCCGC
CB3	GACGAGCTCCACACATCGCC	CTGCTCGAGGTGTGTAGCGG
CB4	AAGGGCACCTACCCCATGCG	TTCCCGTGGATGGGGTACGC
CB5	GGGACCTCGTAAGGGCAGAG	CCCTGGAGCATTCCCGTCTC
CB6	GTCCGACTGGGATGGAGCCT	CAGGCTGACCCTACCTCGGA
CB7	CAGAGGAGTGTCGACCGAC	GTCTCCTCACAGGCTGGCTG
CB8	GTCCCGGAAGGGCTTGCACT	CAGGGCCTTCCCGAACGTGA
CB9	GGCTGCAGTTAGGGCACGGA	CCGACGTCAATCCCGTGCCT
CB10	CGGGAATTCTCGGAACGCGG	GCCCTTAAGAGCCTTGCGCC

Tabla 6. Protocolo de reacción para la incorporación de código de barra tanto para 16S y ITS.

Componentes	Volumen	Concentración
Direct PCR máster	12.5 μ L	
CBF (N) [10 μ M]	1 μ L	10 μ M
CBR (N) [10 μ M]	1 μ L	10 μ M
Muestra (purificado de la PCR)	2 μ L	10-50 μ M
Agua libre de ADNasa	8.5 μ L	
Total	25 μ L	

Tabla 7. Protocolo de termociclador para la incorporación de código de barra.

Temperaturas	Tiempo (s)	Ciclos
95°C	300	1
95°C	20	35
56.8°C	30	
72°C	90	
72°C	120	1

6.2.9 Confirmación del producto de la PCR de la incorporación de código de barra.

Se llevó a cabo una electroforesis en gel de agarosa al 1.5%. En cada pozo del gel se cargaron 3 μ L de cada muestra junto con 2 μ L del buffer de carga azul de bromofenol. El buffer de carga se tiñó previamente con bromuro de etidio a una concentración de 10 mg/mL. Luego, el gel se visualizó utilizando luz ultravioleta mediante un fotodocumentador Ingenius.

6.2.10 Purificación de la incorporación de código de barra.

Para eliminar productos secundarios con un tamaño inferior a 900 pb, se tomaron 20 μ L del producto de la PCR de la incorporación de código de barras y se les añadieron 16.0 μ L de perlas magnéticas Agencourt AMPure XP mediante pipeteo. La mezcla se dejó reposar durante 2 minutos a temperatura ambiente. Luego, se colocó la muestra en una gradilla magnética y se mantuvo allí hasta que la solución se aclaró y se formó un pellet con las perlas magnéticas, un proceso que duró de 3 a 5 minutos.

Una vez que se formó el pellet, se retiró cuidadosamente el sobrenadante desde el lado opuesto al pellet con una pipeta, con precaución para evitar que el pellet se resuspendiera. Se realizaron dos lavados con 200 μ L de etanol al 80% en cada lavado, añadiendo el etanol al lado opuesto al pellet. Luego, se agitó durante 2 segundos y se dejó la muestra a temperatura ambiente para permitir que el etanol residual se evaporara, lo cual llevó de 3 a 5 minutos.

Después, se añadieron 20 μ L de agua libre de nucleasas directamente al pellet y se mezclaron mediante pipeteo, realizando 20 ciclos. Se dejó reposar la muestra a temperatura ambiente durante dos minutos antes de volver a colocarla en la gradilla magnética. Se esperaron otros 2 minutos hasta que la fase acuosa estuvo completamente transparente, momento en el cual se recuperaron 20 μ L de la fase acuosa sin tocar el pellet, y se colocaron en un nuevo microtubo de 0.2 mL.

6.2.11 Secuenciación.

Preparación del Producto de PCR y secuenciación.

En un tubo de PCR de 0.2 mL, se mezcló lo siguiente: 47 μ L de producto de PCR purificado, 3.5 μ L de tampón de reparación de ADN NEBNext FFPE, y 2 μ L de NEBNext FFPE DNA Repair Mix. La mezcla se agitó suavemente moviendo el tubo y girando hacia abajo. Luego, utilizando un termociclador, se incubó a 20 °C durante 5 minutos y a 65 °C durante 5 minutos.

La muestra de ADN se transfirió a un tubo Eppendorf limpio de 1.5 mL. Se añadieron 60 μ L de perlas AMPure XP resuspendidas a la reacción de preparación final y se mezcló por pipeteo. La mezcla se incubó en un mezclador rotatorio durante 5 minutos a temperatura ambiente.

Se prepararon 500 μL de etanol fresco al 70% en agua libre de nucleasas. Luego, se colocó el tubo en un imán, se pipeteó el sobrenadante y se mantuvo el imán en su lugar. Se lavaron las cuentas con 200 μL de etanol al 70% recién preparado sin alterar el sedimento. Se retiró el etanol al 70% con una pipeta y se desechó. Se dejó secar durante 30 segundos.

Después, se retiró el tubo de la rejilla magnética y se resuspendió el sedimento en 61 μL de agua libre de nucleasas. La muestra se incubó durante 2 minutos a temperatura ambiente. Luego, se pelletizaron las cuentas en un imán hasta que el eluato se volvió transparente e incoloro. Se retiró el eluato y se transfirieron 61 μL a un tubo limpio Eppendorf de 1.5 mL.

Una vez que se obtuvo el ADN reparado, se utilizó el kit de secuenciación 1D de Nanopore siguiendo el protocolo proporcionado por la empresa (Oxford Nanopore Technologies).

6.3 Análisis Bioinformático.

6.3.1 Procesamiento de datos.

Una vez obtenidas las lecturas de secuenciación por MinION, estas se almacenaron en archivos en formato FAST5 estas lecturas se encuentran almacenadas en el repositorio virtual en el apartado anexos. Cada archivo FAST5 tiene una capacidad limitada de almacenar hasta 40 lecturas. Con el fin de facilitar el procesamiento y análisis de estos datos, se empleó un script diseñado para esta tesis que se encuentra en el repositorio virtual que se generó para este trabajo que se encuentra en el apartado de anexos. El objetivo de este script fue unir los archivos FAST5 individuales y convertirlos en un único archivo en formato FASTQ. Al combinar los archivos, se reduce el número de archivos a analizar, lo que facilita su manipulación y procesamiento en etapas posteriores.

Una vez que se obtuvo el archivo único en formato FASTQ, se procedió a realizar otro script el cual se encuentra su código fuente en el repositorio virtual y aparece con el título de Fastq separar por código de barra utilizando el lenguaje de programación Python. El propósito de este nuevo script fue separar las lecturas en archivos individuales, en función de la presencia de un código de muestra específico. Mediante la aplicación de este segundo script, se logró clasificar las lecturas de secuenciación en diferentes archivos, según la muestra a la que pertenezcan. Esto permitió una organización adecuada de los datos, facilitando el análisis posterior y el estudio de las

diferentes muestras de interés. Con la ejecución de estos scripts, se optimizó el manejo y la organización de los datos de secuenciación masiva obtenidos a través de la tecnología MinION, permitiendo una mayor eficiencia en la posterior interpretación y análisis de los resultados.

6.3.2 Clasificación de la muestra.

En el marco de este estudio, se llevó a cabo la creación de una nueva base de datos utilizando como punto de partida la base de datos existente de Kraken 2.0. Con el fin de ampliar las capacidades de clasificación, se incorporó una carpeta que contiene las secuencias obtenidas de las regiones ITS de hongos registradas desde el año 2000 hasta el 18 de octubre de 2020. obtenidas de la base de datos del NCBI y que se encuentran depositadas en el repositorio virtual con el título lista de microorganismos usados en Kraken 2.0.

Para incorporar datos de la región ITS a Kraken 2.0, se siguieron los siguientes pasos:

Paso 1: Preparación de Datos

- Se aseguró que los datos de secuenciación incluyeran la región ITS.
- Se realizó el control de calidad de las secuencias cuando fue necesario.
- Se ensamblaron los datos si no estaban en secuencias contiguas (contigs).

Paso 2: Construcción de la Base de Datos de Kraken 2.0.

- Se descargó e instaló Kraken 2.0.
- Se descargó una base de datos de referencia de Kraken 2.0 que incluía información taxonómica de la región ITS y se creó una base de datos personalizada que se puede encontrar en el repositorio virtual.
- Se utilizó kraken2-build para construir la base de datos personalizada.
- Se indexó la base de datos personalizada con kraken2-build.

Paso 3: Ejecución de Kraken 2.0 con los Datos ITS

- Se ejecutó Kraken 2.0 con los datos ITS utilizando kraken2.0.
- Se analizaron los resultados para obtener información sobre la taxonomía de las secuencias.
- Se interpretaron los resultados para identificar las especies o grupos taxonómicos presentes en los datos de ITS.

Lo anterior se ha fundamentado en el manual oficial de Kraken 2.0(Oregón State University,2023).

El objetivo principal de esta adición es permitir la clasificación no solo de la región 16S, sino también de la región ITS, ampliando así el alcance de la clasificación taxonómica de los organismos presentes en las muestras analizadas. La decisión de limitar la inclusión de secuencias a los últimos 20 años se basa en la consideración de que existen numerosas secuencias de hongos mal clasificadas durante las décadas anteriores del siglo pasado. Estas secuencias incorrectamente clasificadas podrían introducir un margen de error en la clasificación taxonómica y dar lugar a contradicciones en la arquitectura de los datos. Al seleccionar únicamente las secuencias obtenidas durante los últimos 20 años, se busca mejorar la calidad y la precisión de la clasificación taxonómica de los hongos presentes en las muestras. Con la creación de esta nueva base de datos, se mejoró la capacidad de clasificación taxonómica de Kraken 2.0, debido a que se tiene una mayor cobertura genómica ya que incorpora un conjunto más amplio de secuencias genómicas de diversas especies y taxones, la actualización de datos ya que la creación de una nueva base de datos permite incorporar datos más recientes y actualizados, lo que puede resultar en una mejor clasificación taxonómica al reflejar con mayor precisión la diversidad genómica en el reino biológico, corrección de errores y eliminación de contaminantes, la creación de una nueva base de datos brinda la oportunidad de eliminar datos incorrectos, redundantes o contaminantes que puedan haber estado presentes en versiones anteriores facilitando un análisis más preciso y confiable de las muestras de interés, especialmente en lo que respecta a la identificación y clasificación de las secuencias de la región ITS de hongos.

6.3.3 Comparación de bases de datos.

En el marco de este estudio, se llevó a cabo una comparación bioinformática entre los datos obtenidos mediante el programa modificado de clasificación basado en la base de datos de libre acceso diseñada en esta tesis y la aplicación comercial de Nanopore denominada EPI2ME. El objetivo de esta comparación es determinar si la precisión de la base de datos modificada es comparable a la aplicación comercial de Nanopore en términos de identificación entre otros criterios de una buena identificación taxonómica de organismos a partir de secuencias. Para llevar a cabo esta comparación, se utilizaron conjuntos de datos obtenidos por ambos métodos de identificación. Estos conjuntos de datos consistieron en secuencias provenientes de muestras de

tierra de viveros de fresa en Guanajuato y fueron sometidos a análisis mediante los respectivos programas de clasificación. Se evaluaron diversos parámetros para determinar la precisión de la clasificación en cada caso. Estos parámetros incluyen, entre otros, la sensibilidad, que mide la capacidad de identificar correctamente casos positivos en un conjunto de datos. Una alta sensibilidad indica una buena detección de casos positivos, la especificidad, que mide la capacidad de identificar correctamente casos positivos en un conjunto de datos. Una alta sensibilidad indica una buena detección de casos positivos, y la precisión que mide la proporción de resultados positivos reportados que son verdaderamente positivos. Una alta precisión indica resultados positivos confiables. Además, se realizó un análisis comparativo de los resultados obtenidos por ambos métodos, identificando las similitudes y diferencias en la asignación taxonómica de las secuencias. Se realizó una evaluación de los resultados obtenidos, considerando tanto el rendimiento general de la clasificación como la consistencia y concordancia con las identificaciones conocidas de los organismos presentes en las muestras analizadas. Este análisis comparativo tiene como objetivo determinar si la base de datos modificada diseñada en esta tesis es capaz de proporcionar resultados de clasificación de organismos comparables en términos de identificación a la aplicación comercial EPI2ME de Nanopore. Los resultados de esta comparación permitieron evaluar la utilidad y eficacia de la base de datos modificada para la identificación de secuencias en el contexto de la investigación bioinformática y la biología molecular.

6.3.4 Porcentaje de microorganismos

Con el fin de determinar las proporciones relativas de bacterias y hongos presentes en la muestra, se utilizaron los datos obtenidos a través del proceso de clasificación de cada lectura. Estos datos se analizaron utilizando la herramienta Krona, que se implementa en el lenguaje de programación Ruby. Cabe destacar que esta herramienta no es compatible con el sistema operativo Windows. Aunque es posible ejecutar Krona en Windows con una configuración adicional, esto puede plantear problemas y complicaciones en la instalación y el uso. La incompatibilidad con Windows limita la accesibilidad y la conveniencia de Krona para los usuarios de este sistema operativo, lo que puede resultar en la preferencia por alternativas más compatibles y fáciles de usar en entornos Windows. La herramienta Krona se utilizó para generar un gráfico de círculo dinámico que visualice las proporciones de los diferentes taxones identificados en la muestra. Este tipo de representación gráfica permite una visualización interactiva y detallada de la distribución de las

bacterias y hongos presentes, brindando una representación visual efectiva de las proporciones relativas de cada grupo taxonómico. Se procedió a importar los datos de clasificación obtenidos previamente al software Krona, siguiendo el flujo de trabajo establecido por esta herramienta (Ondov *et al.*, 2015). Además, la naturaleza dinámica del gráfico de círculo de Krona facilitó la exploración y la interacción con los datos, lo que permite una comprensión más profunda de la composición taxonómica de la muestra.

7. RESULTADOS

7.1 Porcentaje de microorganismos

Para determinar las proporciones de bacterias y hongos presentes en la muestra, se utilizaron los datos obtenidos de la clasificación de cada lectura realizada por el programa Epi2me. Estos datos se visualizaron en el diagrama de árbol incluido en el software por defecto, como se muestra en la Figura 6 y 7 para bacterias y hongos respectivamente. Esta representación gráfica permite observar la distribución taxonómica de las secuencias identificadas. Por otro lado, los datos obtenidos de la clasificación de cada lectura mediante el script enlazado a Kraken 2.0 se procesaron utilizando Krona, una herramienta basada en el lenguaje Ruby. Krona permitirá generar un gráfico de círculo dinámico que mostrará la clasificación taxonómica de las secuencias. A continuación, se presenta la Figura 8, la cual ilustra la clasificación taxonómica de las bacterias en la muestra 1, y la Figura 9, que muestra la clasificación taxonómica de los hongos en la muestra 1. Este mismo procedimiento se seguirá para cada una de las 10 muestras analizadas. Estas representaciones gráficas son de gran utilidad para visualizar y comprender la composición taxonómica de las bacterias y hongos presentes en cada muestra.

La muestra número 1 exhibe un alto nivel de identificación de bacterias tanto por el algoritmo Epi2me como por Kraken 2.0. El gráfico de árbol generado por Epi2me muestra la clasificación de organismos hasta el nivel de especie. En el gráfico de árbol, la cantidad de lecturas clasificadas de cada especie se representa con el tamaño de la rama. Ambos gráficos de árbol son generados por el algoritmo Epi2me. La Figura 6 muestra la clasificación de bacterias en la muestra 1, revelando que la especie bacteriana más predominante en dicha muestra es *Massilia niabensis* y la menos predominante es *Lactobacillus acetotolerans*. La Figura 7 muestra la clasificación de hongos en la muestra 1, y destaca que la especie fúngica más predominante es *Malassezia restricta* y la menos predominante es *Sugiyamaella lignohabitans*.

Asimismo, se presentan los resultados obtenidos mediante el algoritmo Kraken 2.0, representados en gráficos circulares porcentuales utilizando Krona. En la Figura 8, se muestra la clasificación porcentual de bacterias, donde cada bacteria se muestra en función de su rango taxonómico, revelando que la especie bacteriana más predominante es *Massilia niabensis* y la

menos predominante es *Vitreoscilla estercolaria*. La Figura 9 exhibe la clasificación porcentual de hongos en la muestra 1, evidenciando que *Malassezia restricta* es la especie fúngica más predominante y la menos predominante es *Rhizomucor pusillus*.

Estos resultados subrayan la similitud entre ambos algoritmos en la identificación y clasificación de microorganismos. Por lo tanto, se puede afirmar que el script gratuito cumple con su función.

El análisis comparativo y la discusión de la predominancia de bacterias y hongos con diferentes niveles de abundancia en cada muestra se presentarán posteriormente en la tabla 18.

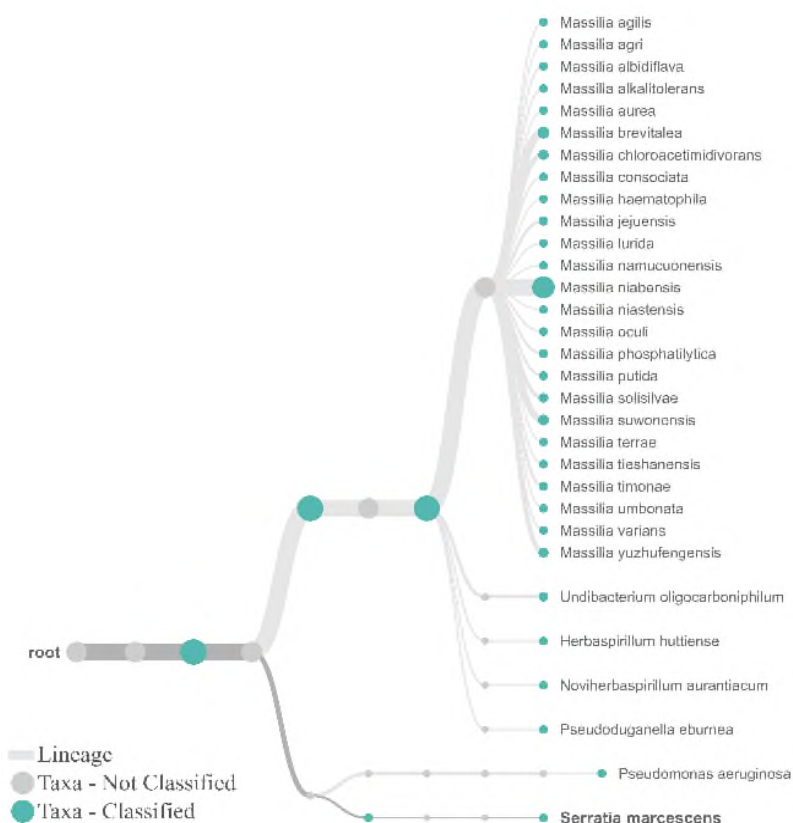


Figura 6. Clasificación taxonómica de bacterias en la muestra 1. Árbol filogenético generado utilizando el software comercial epi2me, en el que cada nodo denota los ancestros comunes compartidos por dos o más taxones. El grosor de las líneas representa el número de lecturas clasificadas a nivel taxonómico, mientras que el tamaño de los círculos refleja la abundancia de un género o especie bacteriana en la muestra.

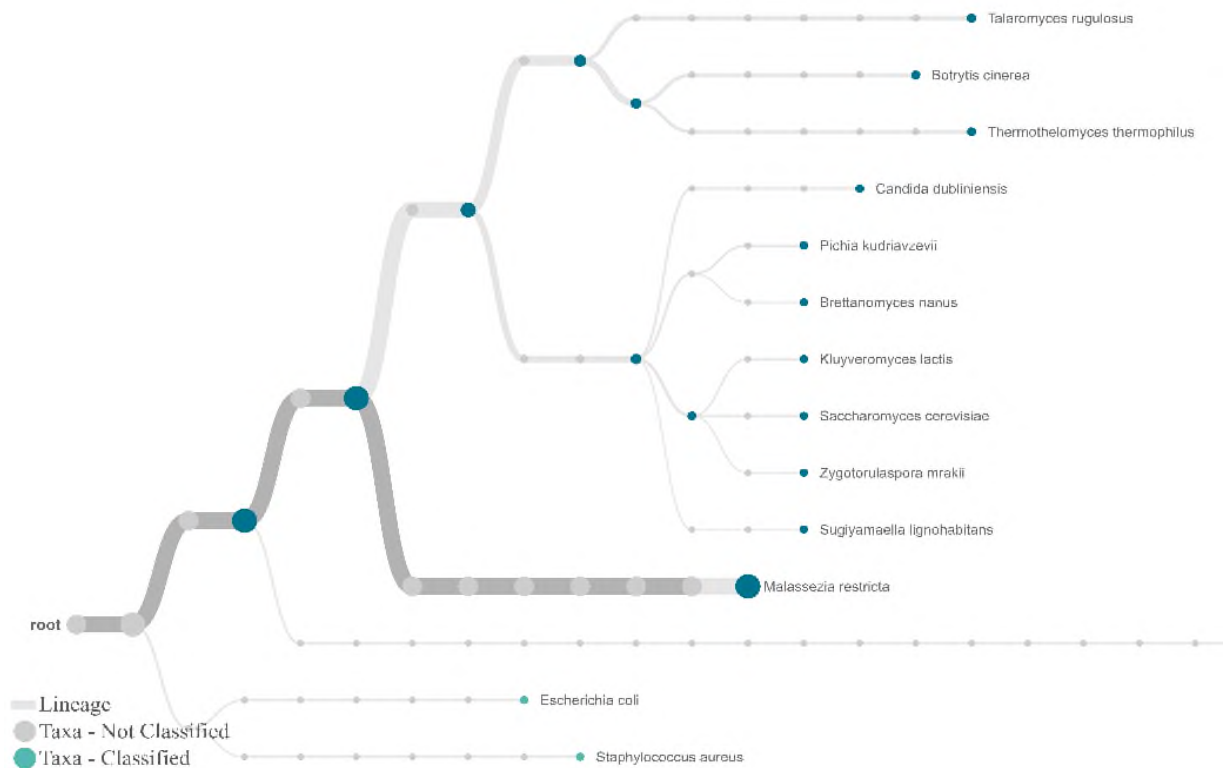


Figura 7. Clasificación taxonómica de hongos en la muestra 1. Árbol filogenético generado utilizando el software comercial epi2me, en el que cada nodo denota los ancestros comunes compartidos por dos o más taxones. El grosor de las líneas representa el número de lecturas clasificadas a nivel taxonómico, mientras que el tamaño de los círculos refleja la abundancia de un género o especie fúngica en la muestra.

clasificación del 79%, siendo el 21% restante no identificado debido a la ausencia de información en las bases de datos utilizadas. La muestra 2 presentó el porcentaje más bajo de clasificación, con un 43%. Es importante destacar que este bajo porcentaje de clasificación en la muestra 2 no implica necesariamente un fallo en el funcionamiento del script de Python, ya que su capacidad de clasificación está vinculada a la disponibilidad de datos en las bases de datos utilizadas.

Tabla 8. Porcentaje de clasificación total de bacterias por muestra.

Número de Muestra.	% clasificación total epi2me	% de clasificación total Kraken 2.0
1	92 %	79 %
2	66 %	43 %
3	67 %	47 %
4	77 %	56 %
5	44 %	49 %
6	68 %	48 %
7	62 %	47 %
8	75 %	51 %
9	77 %	57 %
10	87 %	65 %
Promedio	72%	54%

Es fundamental comprender que la diferencia en los porcentajes de clasificación entre ambos métodos refleja las capacidades y limitaciones de cada uno, especialmente en función de la disponibilidad de datos en las bases de datos utilizadas.

El análisis comparativo y multifacético de la clasificación de géneros bacterianos en diez muestras distintas se realizó utilizando las plataformas EPI2ME y Kraken. Los resultados correspondientes a cada muestra se encuentran detallados en las tablas 9 y 10.

Este análisis permite tomar en consideración más de una característica de los algoritmos y proporciona una comprensión de la variabilidad en la clasificación de géneros bacterianos entre las dos plataformas. Los resultados obtenidos arrojan luz sobre la importancia de la selección de la plataforma analítica más adecuada de acuerdo con los objetivos de investigación y subrayan la necesidad de evaluar la concordancia entre diversas herramientas de clasificación de géneros.

Tabla 9. Comparación de la Concordancia de géneros bacterianos entre muestras (muestra 1-5)

Los datos contenidos en la tabla son representativos de los géneros bacterianos identificados por cada muestra, utilizando los programas EPI2ME y Kraken. En dicha tabla, las celdas resaltadas en color blanco representan el total de géneros identificados en la muestra, mientras que las celdas destacadas en color rojo indican los géneros comunes identificados por ambos algoritmos.

MUESTRA 1		MUESTRA 2		MUESTRA 3		MUESTRA 4		MUESTRA 5	
<i>Massilia</i>	<i>Massilia</i>	<i>Pseudomonas</i>	<i>Klebsiella</i>	<i>Pseudomonas</i>	<i>Klebsiella</i>	<i>Pseudomonas</i>	<i>Klebsiella</i>	<i>Pseudomonas</i>	<i>Klebsiella</i>
<i>Pseudomonas</i>	<i>Collimonas</i>	<i>Massilia</i>	<i>Serratia</i>	<i>Serratia</i>	<i>Serratia</i>	<i>Serratia</i>	<i>Yersinia</i>	<i>Herbaspirillum</i>	<i>Pantoea</i>
<i>Undibacterium</i>	<i>Herbaspirillum</i>	<i>Undibacterium</i>	<i>Pseudomonas</i>	<i>Massilia</i>	<i>Pseudomonas</i>	<i>Undibacterium</i>	<i>Serratia</i>	<i>Massilia</i>	<i>Serratia</i>
<i>Herbaspirillum</i>	<i>Burkholderia</i>	<i>Serratia</i>	<i>Stenotrophomonas</i>	<i>Lysobacter</i>	<i>Hydrogenophaga</i>	<i>Massilia</i>	<i>Pseudomonas</i>	<i>Undibacterium</i>	<i>Pseudomonas</i>
<i>Pseudoduganella</i>	<i>Variovarax</i>	<i>Vicinamibacter</i>	<i>Massilia</i>	<i>Aquicella</i>	<i>Bacillus</i>	<i>Stenotrophomonas</i>	<i>Stenotrophomonas</i>	<i>Ralstonia</i>	<i>Moraxella</i>
<i>Serratia</i>	<i>Erwinia</i>	<i>Stenotrophomonas</i>	<i>Termitoga</i>	<i>Stenotrophomonas</i>	<i>Burkholderia</i>	<i>Hydrogenophaga</i>	<i>Vibrio</i>	<i>Serratia</i>	<i>Stenotrophomonas</i>
<i>Noviherbaspirillum</i>	<i>Klebsiella</i>	<i>Ralstonia</i>		<i>Garritya</i>	<i>Massilia</i>	<i>Pelomonas</i>	<i>Massilia</i>	<i>Akkermansia</i>	<i>Ralstonia</i>
<i>Caulobacter</i>	<i>Yersinia</i>	<i>Ramlibacter</i>		<i>Devosia</i>	<i>Vibrio</i>	<i>Achromobacter</i>	<i>Burkholderia</i>	<i>Bradyrhizobium</i>	<i>Burkholderia</i>
<i>Sphingomonas</i>	<i>Streptococcus</i>	<i>Herbaspirillum</i>		<i>Methylobacterium</i>	<i>Stenotrophomonas</i>	<i>Ralstonia</i>	<i>Herbaspirillum</i>	<i>Sphingomonas</i>	<i>Herbaspirillum</i>
<i>Ralstonia</i>	<i>Lactococcus</i>	<i>Pseudoduganella</i>		<i>Sphingomonas</i>		<i>Herbaspirillum</i>	<i>Janthinobacterium</i>	<i>Hydrogenophaga</i>	<i>Massilia</i>
<i>Pelomonas</i>	<i>Staphylococcus</i>	<i>Pectobacterium</i>		<i>Achromobacter</i>		<i>Noviherbaspirillum</i>	<i>Colimonas</i>	<i>Asinibacterium</i>	<i>Lactobacillus</i>
<i>Janthinobacterium</i>	<i>Paenibacillus</i>			<i>Ralstonia</i>		<i>Sphingomonas</i>	<i>Ralstonia</i>	<i>Caulobacter</i>	<i>Bacillus</i>
<i>Stenotrophomonas</i>	<i>Bacillus</i>			<i>Hydrogenophaga</i>		<i>Morganella</i>	<i>Comamonas</i>	<i>Caedimonas</i>	
<i>Lactobacillus</i>	<i>Thermoanaerobacterium</i>			<i>Herbaspirillum</i>		<i>Bradyrhizobium</i>	<i>Sphigobium</i>	<i>Chelatococcus</i>	
	<i>Halanaerobium</i>			<i>Undibacterium</i>		<i>Ramlibacter</i>	<i>Rickettsia</i>	<i>Cupriavidus</i>	
	<i>Lactobacillus</i>			<i>Providencia</i>		<i>Citrobacter</i>	<i>Akkermansia</i>	<i>Citrobacter</i>	
				<i>Legionella</i>		<i>Klebsiella</i>	<i>Mycoplasma</i>	<i>Rhodanobacter</i>	
				<i>Dialister</i>		<i>Aquicella</i>	<i>Lactococcus</i>	<i>Xanthomonas</i>	
						<i>Haemophilus</i>	<i>Lactobacillus</i>	<i>Paenibacillus</i>	
						<i>Lysobacter</i>	<i>Veillonella</i>	<i>Thermolithobacter</i>	
						<i>Paenibacillus</i>	<i>Bacillus</i>		
						<i>Prevotella</i>	<i>Edwardsiella</i>		
						<i>Alistipes</i>			
						<i>Flavobacterium</i>			
						<i>Gemmatimonas</i>			
						<i>Methylobacterium</i>			
						<i>Enhydrobacter</i>			
						<i>Dexia</i>			
						<i>Burkholderia</i>			
						<i>Cupriavidus</i>			
						<i>Acidovorax</i>			
						<i>Variovarax</i>			
						<i>Zoogloea</i>			
						<i>Microbulbifer</i>			
						<i>Pectobacterium</i>			
						<i>Yersinia</i>			
						<i>Haliangium</i>			
						<i>Enterococcus</i>			
						<i>Lactococcus</i>			
						<i>Streptococcus</i>			
						<i>Anaerobacterium</i>			
						<i>Ruthenobacterium</i>			

Tabla 10. Comparación de la Concordancia de géneros bacterianos entre muestras (Muestra 6-10).

MUESTRA 6		MUESTRA 7		MUESTRA 8		MUESTRA 9		MUESTRA 10	
Hydrogenophaga	Klebsiella	Pseudomonas	Termus	Pseudomonas	Serratia	Undibacterium	Massilia	Undibacterium	Massilia
Pseudomonas	Pseudomonas	Massilia	Paenibacillus	Undibacterium	Klebsiella	Herbaspirillum	Herbaspirillum	Herbaspirillum	Janthinobacterium
Cupriavidus	Serratia	Serratia	Serratia	Serratia	Pseudomonas	Ralstonia	Janthinobacterium	Ralstonia	Herbaspirillum
Herbaspirillum	Stenotrophomonas	Citrobacter	Herbaspirillum	Massilia	Vibrio	Noviherbaspirillum	Ralstonia	Noviherbaspirillum	Cupriavidus
Undibacterium	Hydrogenophaga	Methylobacterium	Stenotrophomonas	Herbaspirillum	Lysobacter	Pseudomonas	Cupriavidus	Pseudomonas	Ralstonia
Serratia	Janthinobacterium	Herbaspirillum	Brachyspira	Ralstonia	Burkholderia	Methylobacterium	Nitrosomonas	Serratia	Burkholderia
Bernardtia	Herbaspirillum	Undibacterium	Yersinia	Ramlibacter	Ralstonia	Stenotrophomonas	Pseudomonas	Massilia	Achromobacter
Gemmatimonas	Burkholderia	Hydrogenophaga	Klebsiella	Methylobacterium	Massilia	Enterobacter	Paenibacillus	Variovorax	Variovorax
Longimicrobium	Hyphomicrobium	Bradyrhizobium	Vibrio	Citrobacter	Herbaspirillum	Paenibacillus	Klebsiella	Stenotrophomonas	Pseudomonas
Reyranelia	Mycoplasma	Morganella	Alteromonas	Nocardioideis	Achromobacter	Cupriavidus	Yersinia	Methylobacterium	Serratia
Sphingomonas	Bacillus	Paenibacillus	Pseudomonas	Akkermansia	Acetobacter	Pelomonas	Vibrio	Gemmata	Klebsiella
Phreatobacter	Sphingomonas	Vicinamibacter	Burkholderia	Bradyrhizobium	Methylobacterium	Serratia	Stenotrophomonas	Bradyrhizobium	Acinetobacter
Limnobacter		Mucilaginibacter	Dechloromonas	Stella	Methylorubrum	Bradyrhizobium	Bradyrhizobium	Sphingomonas	Methylobacterium
Acidovorax		Dongia	Achromobacter	Dongia	Bradyrhizobium	Sphingomonas	Roseomonas	Cupriavidus	Roseomonas
Comamonas		Sphingobium	Hydrogenophaga	Reyranelia	Bacillus	Achromobacter	Methylobacterium	Paucimonas	Paenibacillus
Simplicispira		Achromobacter	Massilia	Burkholderia	Staphylococcus	Hydrogenophaga	Bacillus	Acidovorax	Bacillus
Variovorax		Burkholderia	Dialister	Paraburkholderia	Thermoanaerobacterium	Vogesella	Paenibacillus	Hydrogenophaga	Lactobacillus
Telluria		Triniclia		Comamonas		Citrobacter	Mycoplasma	Duganella	
Chromobacterium		Pelomonas		Hydrogenophaga			Nostoc	Janthinobacterium	
Citrobacter		Ramlibacter		Noviherbaspirillum				Cedecea	
Perleidibaca		Glaciimonas		Tepidimonas				Citrobacter	
Rhodanobacter		Roseateles		Nitrosococcus				Vibrio	
Stenotrophomonas		Vogesella		Tatumella				Dyella	
Hydrogenophilus		Kluyvera		Morganella					
		Shigella		Pectobacterium					
		Pectobacterium		Aquicella					
		Aquicella		Haemophilus					
		Legionella		Stenotrophomonas					
		Haemophilus		Gaiella					
		Stenotrophomonas		Lactococcus					
		Leuconostoc							
		Kineothrix							

En los resultados de las muestras analizadas, se observa una comparación de la identificación taxonómica llevada a cabo por dos plataformas, EPI2ME y Kraken. En la Muestra 1, se identificaron principalmente los géneros *Massilia*, *Pseudomonas* y *Klebsiella* a través de EPI2ME, mientras que Kraken identificó principalmente los géneros *Massilia*, *Pseudomonas* y *Collimonas*. De manera interesante, dos géneros, *Massilia* y *Pseudomonas*, fueron comunes en ambas plataformas. En la Muestra 2, EPI2ME identificó predominantemente los géneros *Serratia*, *Pseudomonas* y *Undibacterium*, en contraste con Kraken, que identificó principalmente los géneros *Serratia*, *Pseudomonas* y *Klebsiella*. Tres géneros, *Pseudomonas*, *Serratia* y *Herbaspirillum*, fueron comunes en ambas plataformas. Finalmente, en la Muestra 3, EPI2ME clasificó principalmente los géneros *Pseudomonas*, *Undibacterium* y *Serratia*, mientras que Kraken clasificó principalmente los géneros *Serratia*, *Pseudomonas* y *Klebsiella*. Notablemente, cinco géneros, a saber, *Undibacterium*, *Pseudomonas*, *Serratia*, *Herbaspirillum* y *Massilia*, se identificaron de manera coincidente en ambas plataformas. Estos resultados resaltan las similitudes y diferencias en la identificación taxonómica entre EPI2ME y Kraken en estas muestras particulares.

Estos patrones se repiten en las muestras restantes. Los resultados resaltan la variabilidad en la clasificación de géneros bacterianos entre las dos plataformas, así como la presencia de géneros en común que pueden considerarse consistentes en las muestras. Estas observaciones subrayan la importancia de seleccionar la plataforma analítica más adecuada según los objetivos de la investigación y resaltan la necesidad de evaluar la concordancia entre diferentes herramientas de clasificación microbiológica.

Los resultados indican que, en las muestras analizadas, la plataforma EPI2ME clasificó un número variable de géneros bacterianos, que oscila entre 11 y 42 géneros, mientras que Kraken identificó entre 7 y 22 géneros. Además, se observó que el número de géneros en común entre ambas plataformas varió de 2 a 10 géneros en las distintas muestras.

Estos hallazgos sugieren diferencias en la capacidad de clasificación de géneros bacterianos entre EPI2ME y Kraken (Tabla 11), así como la presencia de una proporción variable de géneros en común en las distintas muestras.

Tabla 11. Concentrado de la comparación de la Concordancia de géneros bacterianos entre muestras.

Numero de muestra	Géneros Bacterias		
	EPI2ME	Géneros en común	KRAKEN
1	14	2	16
2	11	3	7
3	18	5	9
4	42	10	22
5	20	5	12
6	24	6	12
7	33	9	17
8	31	6	17
9	21	8	22
10	23	9	17

7.2.2 HONGOS

Durante el análisis de las 10 muestras a través del programa Epi2me, se registró un promedio de clasificación del 76.3% en relación con el total del 100% de las muestras evaluadas. Por ejemplo, al observar la muestra 9, se resaltó una destacada clasificación del 89%, lo que implica que únicamente el 11% de la muestra no pudo ser identificado ni categorizado por el programa. De manera similar, tanto la muestra 2 como la muestra 5 lograron una clasificación del 79%, indicando que el 21% de la muestra no pudo ser identificado ni clasificado por el programa.

Por otro lado, el script desarrollado en el lenguaje Python arrojó un promedio de clasificación del 46.75%. En el caso de la muestra 8, se logró una clasificación del 59%, mientras que el 41% restante no pudo ser identificado debido a la falta de información en las bases de datos utilizadas. La muestra 2, por su parte, mostró el porcentaje más bajo de clasificación, con un 39%. Es importante subrayar que este bajo porcentaje de clasificación en la muestra 2 no necesariamente denota un fallo en el funcionamiento del script de Python, ya que su capacidad de clasificación está directamente relacionada con la disponibilidad de datos en las bases de datos empleadas.

Tabla 12. Porcentaje de clasificación total de hongos por muestra.

Numero de muestra	% clasificación total epi2me	% de clasificación total Kraken 2.0
1	87 %	45 %
2	79%	39 %
3	87 %	56 %
4	84 %	57 %
5	87 %	49 %
6	83 %	58 %
7	72 %	53 %
8	88 %	59 %
9	89 %	56 %
10	79 %	55 %
Promedio	76.30 %	46.75 %

La tabla 12 representa el rendimiento de dos métodos, Epi2me y un script de Python, en el análisis de 10 muestras. Epi2me logra un alto promedio de clasificación del 76.3%, con ejemplos destacados como la muestra 9, que alcanza el 89% de clasificación. Las muestras 2 y 5 también tienen una buena clasificación del 79%. En contraste, el script de Python tiene un promedio de clasificación más modesto del 46.75%, siendo la muestra 8 la mejor clasificada con un 59%. La muestra 2 tiene el porcentaje más bajo de clasificación, con un 39%. Se destaca que el bajo rendimiento no necesariamente indica un fallo en el script de Python, sino que depende de la disponibilidad de datos en las bases de datos utilizadas y su confiabilidad.

El análisis comparativo y multifacético de la clasificación de géneros fúngicos en diez muestras distintas se realizó utilizando las plataformas EPI2ME y Kraken. Los resultados correspondientes a cada muestra se encuentran detallados en la tabla 13 y 14.

Tabla 13. Comparación de la concordancia de géneros fúngicos entre muestras (Muestra 1-5).

Los datos contenidos en la tabla son representativos de los géneros fúngicos identificados por cada muestra, utilizando los programas EPI2ME y Kraken. En dicha tabla, las celdas resaltadas en color blanco representan el total de géneros identificados en la muestra, mientras que las celdas destacadas en color rojo indican los géneros comunes identificados por ambos algoritmos.

MUESTRA 1	MUESTRA 2		MUESTRA 3		MUESTRA 4		MUESTRA 5		
Botrytis	Malassezia	Candida	Malassezia	Botrytis	Malassezia	Thermothelomyces	Malassezia	Kluyveromyces	Malassezia
Brettanomyces	Fusarium	Kluyveromyces	Candida	Brettanomyces	Candida	Pichia	Candida	Talaromyces	Candida
Candida	Aspergillus	Brettanomyces	Cryptococcus	Candida	Fusarium	Malassezia	Fusarium	Malassezia	Fusarium
Kluyveromyces		Malassezia	Fusarium	Kluyveromyces	Aspergillus	Sugiyamaella	Aspergillus	Thermothelomyces	Aspergillus
Malassezia		Sugiyamaella	Aspergillus	Malassezia		Brettanomyces	Eremothecium	Botrytis	
Pichia		Botrytis	Eremothecium	Pichia		Zygorulasporea	Kluyveromyces	Candida	
Saccharomyces		Talaromyces	kluyveromyces	Saccharomyces		Botrytis	Saccharomyces	Eremothecium	
Sugiyamaella		Thermothelomyces		Sugiyamaella		Talaromyces			
Talaromyces		Pichia		Talaromyces		Saccharomyces			
Thermothelomyces				Thermothelomyces		Candida			
Zygorulasporea				Zygorulasporea		Kluyveromyces			

Tabla 14. Comparación de la concordancia de géneros fúngicos entre muestras (Muestra 6-10).

MUESTRA 6		MUESTRA 7		MUESTRA 8		MUESTRA 9		MUESTRA 10	
<i>Candida</i>	<i>Malassezia</i>	<i>Candida</i>	<i>Malassezia</i>	<i>Botrytis</i>	<i>Malassezia</i>	<i>Talaromyces</i>	<i>Malassezia</i>	<i>Talaromyces</i>	<i>Malassezia</i>
<i>Eremothecium</i>	<i>Cryptococcus</i>	<i>Kluyveromyces</i>	<i>Cryptococcus</i>	<i>Kluyveromyces</i>	<i>Candida</i>	<i>Brettanomyces</i>	<i>Candida</i>	<i>Malassezia</i>	<i>Cryptococcus</i>
<i>Malassezia</i>	<i>Candida</i>	<i>Talaromyces</i>	<i>Candida</i>	<i>Malassezia</i>	<i>Fusarium</i>	<i>Sugiyamaella</i>	<i>Fusarium</i>	<i>Candida</i>	<i>Candida</i>
<i>Botrytis</i>	<i>Fusarium</i>	<i>Malassezia</i>	<i>Fusarium</i>	<i>Candida</i>	<i>Aspergillus</i>	<i>Malassezia</i>	<i>Aspergillus</i>	<i>Thermothelomyces</i>	<i>Fusarium</i>
<i>Brettanomyces</i>	<i>Aspergillus</i>	<i>Brettanomyces</i>	<i>Aspergillus</i>	<i>Brettanomyces</i>	<i>Eremothecium</i>	<i>Candida</i>	<i>Eremothecium</i>	<i>Botrytis</i>	<i>Aspergillus</i>
<i>Kluyveromyces</i>	<i>Eremothecium</i>	<i>Pichia</i>	<i>Eremothecium</i>	<i>Talaromyces</i>	<i>Kluyveromyces</i>	<i>Thermothelomyces</i>		<i>Thermothielavioides</i>	<i>Kluyveromyces</i>
<i>Sugiyamaella</i>		<i>Botrytis</i>	<i>Kluyveromyces</i>	<i>Thermothelomyces</i>		<i>Botrytis</i>		<i>Sugiyamaella</i>	
<i>Talaromyces</i>		<i>Thermothelomyces</i>		<i>Sugiyamaella</i>		<i>Pichia</i>		<i>Zygorulaspóra</i>	
<i>Thermothelomyces</i>		<i>Sugiyamaella</i>		<i>Pichia</i>		<i>Zygorulaspóra</i>			
<i>Zygorulaspóra</i>		<i>Saccharomyces</i>				<i>Thermothielavioides</i>			

Este estudio consistió en analizar y comparar la clasificación de diversos géneros de hongos en diez muestras diferentes mediante dos plataformas de análisis. A continuación, se presentan de manera concisa los resultados obtenidos tras el análisis taxonómico de diez muestras distintas. En la Muestra 1, los géneros *Botrytis*, *Candida* y *Brettanomyces* sobresalieron como los más prominentes, destacándose la presencia de los géneros *Candida* y *Malassezia*, que fueron compartidos por ambas plataformas de análisis. En la Muestra 2, los géneros *Malassezia*, *Aspergillus* y *Fusarium* figuraron como los géneros de mayor relevancia, y nuevamente se observó la identificación conjunta de *Candida* y *Malassezia* por ambas plataformas. La Muestra 3 exhibió una identificación prominente de los géneros *Candida* y *Brettanomyces*, con la coincidencia de tres géneros, a saber, *Candida*, *Fusarium* y *Malassezia*, identificados por ambas plataformas de análisis. Similarmente, las Muestras 4, 5 y 6 presentaron patrones similares, con diferentes géneros destacados y una variedad de géneros comunes identificados por ambas plataformas. En las Muestras 7, 8, 9 y 10, se detectaron géneros distintos de manera prominente, junto con géneros compartidos entre ambas plataformas de análisis. Estos hallazgos proporcionan una visión comprehensiva de la diversidad taxonómica en las muestras examinadas. Los resultados concentrados de la clasificación de géneros de hongos en diez muestras utilizando las plataformas EPI2ME y Kraken se muestran en la tabla 15. En general, EPI2ME y Kraken difieren en la cantidad de géneros clasificados en cada muestra. Se encontraron géneros comunes entre ambas plataformas en la mayoría de las muestras, lo que sugiere cierta consistencia en la clasificación.

Tabla 15. Concentrado de la comparación de la concordancia de géneros fúngicos entre muestras.

Numero de muestra	Géneros Hongos		
	EPI2ME	Géneros en común	KRAKEN
1	11	1	3
2	9	3	7
3	11	2	4
4	11	4	7
5	7	2	4
6	10	3	6
7	10	3	7
8	9	3	6
9	10	2	5
10	8	2	6

7.2.3 BACTERIAS

La información previamente presentada se refiere a la clasificación general de las muestras, es decir, cuánto lograron clasificar tanto el programa Epi2me como el script de Python de bacterias y hongos dentro del total del 100% de cada muestra. En este contexto, la Tabla 16 proporciona el porcentaje de clasificación total a nivel de especies bacterianas por muestra, es decir, qué parte de la muestra fue clasificada por el programa hasta el nivel taxonómico de especie.

En el caso del programa Epi2me, se destaca que logró clasificar hasta el nivel de especie el 83% de la muestra 1, lo que significa que el 17% restante fue clasificado, pero no alcanzó dicho nivel taxonómico. Por otro lado, el script desarrollado en Python muestra resultados de clasificación bajos, a pesar de que en la clasificación general presentaba resultados favorables con relación al 100% de la muestra. Sin embargo, al evaluar qué porcentaje de microorganismos se clasificó hasta el nivel de especie, se observa que, de las 10 muestras, solo una de ellas logró un nivel de clasificación satisfactorio por encima del 50%. En otras palabras, el resto de las muestras solo alcanzó un bajo porcentaje de clasificación a nivel de especie (Figura 12).

Tabla 16. Porcentaje de clasificación total de especies de bacterias por muestra.

Numero de Muestra.	% lecturas clasificadas de especies epi2me	% lecturas clasificadas de especies Kraken 2.0
1	83 %	25 %
2	46 %	11 %
3	62 %	13 %
4	57 %	5 %
5	68 %	13 %
6	59 %	72 %
7	59 %	50 %
8	60 %	10 %
9	59 %	11 %
10	58 %	11 %
Promedio	55.88 %	21.10 %

7.2.4 HONGOS

Tabla 17. Porcentaje de clasificación total de especies de Hongos por muestra.

Numero de muestra	% de lecturas clasificadas hasta especies	% de clasificación total de especies Kraken 2.0
1	91 %	32 %
2	81 %	25 %
3	91 %	44 %
4	76 %	37 %
5	96 %	39 %
6	90 %	46 %
7	91 %	35 %
8	85 %	44 %
9	90 %	44 %
10	81 %	36 %
Promedio	79.91 %	34.96 %

En el análisis de los resultados, se destaca el desempeño del programa Epi2me en la muestra 5, donde logró clasificar hasta el nivel de especie el 96% de la muestra, lo que implica que solo el 4% restante no pudo ser categorizado a nivel taxonómico de especie. Este resultado indica una capacidad sobresaliente del programa para identificar y clasificar microorganismos en esta muestra específica.

Por otro lado, el script desarrollado en Python, aunque mostró resultados favorables en la clasificación general con relación al 100% de la muestra, presenta resultados de clasificación más bajos a nivel de especie (tabla 17). Al evaluar la clasificación de hongos en estas muestras, se observa que ninguna de las 10 muestras logra superar el 50% de clasificación hasta el nivel de especie. Sin embargo, es relevante destacar que 9 de las 10 muestras exhiben un porcentaje de clasificación hasta el nivel de especie que supera el 30%. Esto sugiere que, aunque el script en

Python no logra un alto grado de clasificación a nivel de especie en ninguna de las muestras, todavía es capaz de identificar y categorizar un porcentaje significativo de microorganismos en la mayoría de los casos.

En la Figura 10, se puede apreciar que ambos algoritmos exhiben una capacidad similar para clasificar e identificar géneros de bacterias en base a muestras. No obstante, en la Figura 14, la asignación específica de cada punto a un número de muestra resulta poco discernible. A pesar de esta limitación, se observa a simple vista una tendencia media en el gráfico, lo cual sugiere que el algoritmo desarrollado en Python y conectado a Kraken 2.0 posee el potencial de servir como una herramienta para la clasificación de bacterias en la región 16S, específicamente en la región Hipervariable v3-v4. Adicionalmente, se aprecian porcentajes de clasificación similares a los obtenidos por el programa epi2me de Oxford Nanopore.

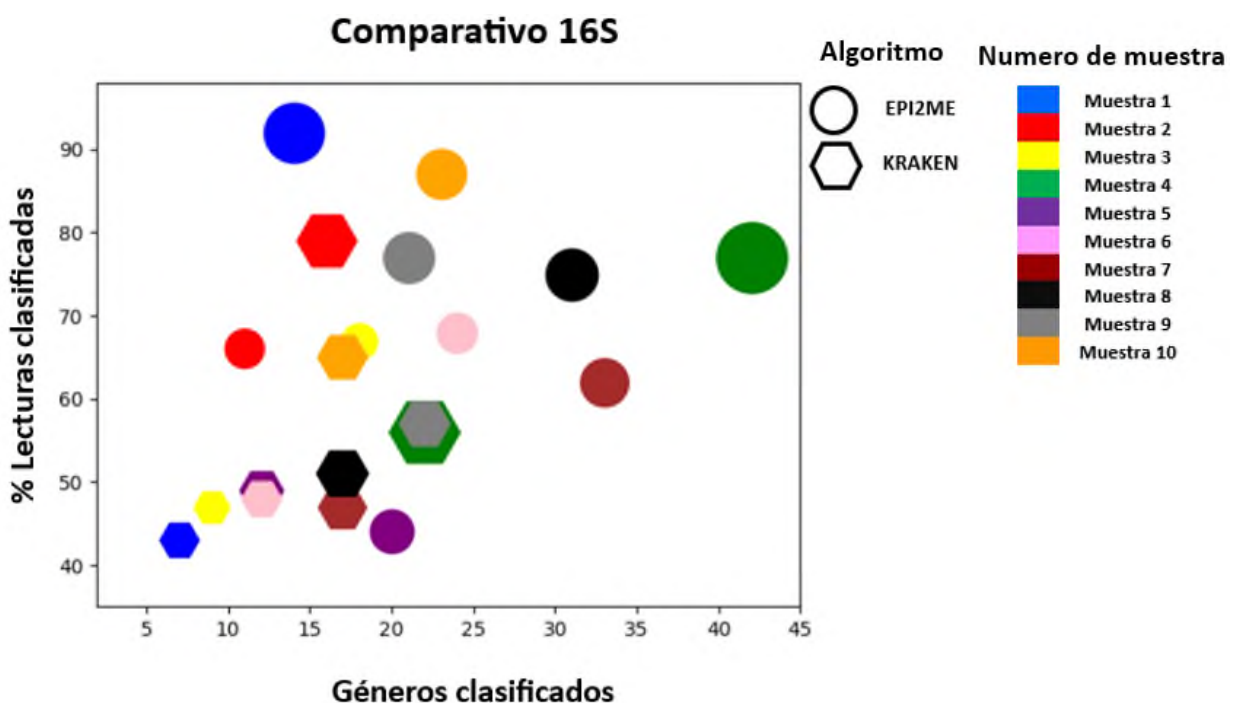


Figura 10. Gráfico comparativo de algoritmos de clasificación taxonómica de bacterias. Representación gráfica que muestra la relación entre el porcentaje de lecturas clasificadas y los géneros bacterianos clasificados, con cada muestra diferenciada por un color distinto y el algoritmo diferenciado por una círculo para epi2me y un hexágono para Kraken.

En contraste, en la Figura 11 se puede observar que el algoritmo Kraken 2.0 logra identificar una mayor cantidad de géneros de hongos por muestra. Sin embargo, se debe destacar que este algoritmo presenta la limitación de no clasificar una gran proporción de las lecturas, lo que resulta en la pérdida de una considerable cantidad de información. Por otro lado, el algoritmo Epi2me realiza un análisis de la mayoría de las lecturas de nuestra muestra, aunque no logre clasificarlas en su mayoría hasta el nivel de especie.



Figura 11. Gráfico comparativo de algoritmos de clasificación taxonómica de Hongos. Representación gráfica que muestra la relación entre el porcentaje de lecturas clasificadas y los géneros bacterianos clasificados, con cada muestra diferenciada por un color distinto y el algoritmo diferenciado por una círculo para epi2me y un hexágono para Kraken.

Los gráficos 10 y 11 previamente presentados ofrecen una representación visual de la clasificación taxonómica de bacterias y hongos en las muestras analizadas. En las tablas 18 y 19 presentan los resultados de un análisis comparativo de la identificación de especies bacterianas y fúngicas predominantes y menos predominantes en diez muestras distintas utilizando dos plataformas de análisis: EPI2ME y Kraken 2.0. La identificación precisa de estas especies es fundamental en estudios microbiológicos, ya que puede ofrecer información valiosa sobre la composición y diversidad de una comunidad en un determinado entorno. La elección de la plataforma de análisis es un paso crítico en la obtención de resultados confiables y precisos. Por lo tanto, esta comparación tiene como objetivo evaluar la concordancia entre las dos plataformas y resaltar las diferencias en la identificación de especies bacterianas predominantes y menos predominantes. Estos resultados proporcionarán información relevante para investigaciones relacionadas con microbiomas y metagenómica, y subrayarán la importancia de considerar múltiples factores al seleccionar la plataforma de análisis más adecuada para los objetivos de investigación.

Tabla 18. Variabilidad en la Abundancia de Especies Bacterianas según el Algoritmo de Clasificación.

NUMERO DE MUESTRA	ESPECIE DE BACTERIA MAS PREDOMINANTE		ESPECIE DE BACTERIA MENOS PREDOMINANTE	
	EPI2ME	KRAKEN 2.0	EPI2ME	KRAKEN 2.0
1	<i>Massilia niabensis</i>	<i>Massilia niabensis</i>	<i>Lactobacillus acetotolerans</i>	<i>Vitreoscilla stercoraria</i>
2	<i>Undibacterium oligocarboniphilum</i>	<i>Photobacterium angustum</i>	<i>Stenotrophomonas rhizophila</i>	<i>Sinorhizobium fredii</i>
3	<i>Undibacterium oligocarboniphilum</i>	<i>Pseudomonas putida</i>	<i>Stenotrophomonas rhizophila</i>	<i>Rhizobium tropici</i>
4	<i>Pseudomonas aeruginosa</i>	<i>Weizmannia coagulans</i>	<i>Ruthenibacterium lactatiformans</i>	<i>Pediococcus acidilactici</i>
5	<i>Pseudomonas aeruginosa</i>	<i>Bdellovibrio bacteriovorus</i>	<i>Thermolithobacter carboxydivorans</i>	<i>Borrelia parkeri</i>
6	<i>Hydrogenophaga taeniospiralis</i>	<i>Ralstonia pickettii</i>	<i>Longimicrobium terrae</i>	<i>Salmonella enterica</i>
7	<i>Serratia marcescens</i>	<i>Pectinatus frisingensis</i>	<i>Mucilagibacter rubeus</i>	<i>Streptococcus equi</i>
8	<i>Pseudomonas aeruginosa</i>	<i>Bacillus benzoovorans</i>	<i>Akkermansia muciniphila</i>	<i>Pelodictyon luteolum</i>
9	<i>Undibacterium oligocarboniphilum</i>	<i>Burkholderia glumae</i>	<i>Citrobacter freundii</i>	<i>Dickeya chrysanthemi</i>
10	<i>Undibacterium oligocarboniphilum</i>	<i>Streptococcus salivarius</i>	<i>Gemmata massiliana</i>	<i>Bordetella pertussis</i>

Tabla 19. Variabilidad en la Abundancia de Especies Fúngica según el Algoritmo de Clasificación.

NUMERO DE MUESTRA	ESPECIE DE HONGO MAS PREDOMINANTE		ESPECIE DE HONGO MENOS PREDOMINANTE	
	EPI2ME	KRAKEN 2.0	EPI2ME	KRAKEN 2.0
1	<i>Malassezia restricta</i>	<i>Penicillium coprophilum</i>	<i>Sugiyamaella lignohabitans</i>	<i>Rhizomucor pusillus</i>
2	<i>Malassezia restricta</i>	<i>Penicillium italicum</i>	<i>Sugiyamaella lignohabitans</i>	<i>Mrakia frigida</i>
3	<i>Malassezia restricta</i>	<i>Candida glabrata</i>	<i>Sugiyamaella lignohabitans</i>	<i>Priceomyces medius</i>
4	<i>Malassezia restricta</i>	<i>Trichophyton violaceum</i>	<i>Zygorulasporea mrakii</i>	<i>Tolyposcladium nubicola</i>
5	<i>Malassezia restricta</i>	<i>Hypoxyton fragiforme</i>	<i>Eremothecium gossypii</i>	<i>Whalleya microplaca</i>
6	<i>Malassezia restricta</i>	<i>Aspergillus japonicus</i>	<i>Brettanomyces nanus</i>	<i>Cladosporium cucumerinum</i>
7	<i>Malassezia restricta</i>	<i>Debaryomyces prosopidis</i>	<i>Saccharomyces cerevisiae</i>	<i>Fusarium oxysporum</i>
8	<i>Malassezia restricta</i>	<i>Candida wancherniae</i>	<i>Brettanomyces nanus</i>	<i>Curvularia eragrostidis</i>
9	<i>Malassezia restricta</i>	<i>Fusarium foetens</i>	<i>Thermothielavioides terrestris</i>	<i>Aspergillus ostianus</i>
10	<i>Malassezia restricta</i>	<i>Lachnellula willkommii</i>	<i>Thermothielavioides terrestris</i>	<i>Taphrina betulina</i>

8. DISCUSIÓN

En la Tabla 18, se detallan las bacterias que exhibieron una predominancia significativa, así como aquellas que mostraron una menor predominancia, basándonos en el número de lecturas detectadas por secuenciación. Es importante destacar que en 4 muestras se identificó repetidamente la presencia de "*Undibacterium oligocarboniphilum*", usando Epi2me. "*Undibacterium oligocarboniphilum*" tiene preferencia por suelos con escasas concentraciones de carbono lo que podría tener implicaciones en la dinámica del suelo como el ciclo del carbono que contribuye a la descomposición de materia orgánica, liberando nutrientes y carbono al suelo. Los nutrientes también se implicarían ya que participan en la mineralización de nitrógeno y fósforo. La comunidad microbiana por su presencia y actividad impactan la estructura y diversidad del suelo, con interacciones competitivas y cooperativas. Finalmente, el ciclo del nitrógeno algunas especies pueden fijar nitrógeno atmosférico, mejorando la fertilidad del suelo y la disponibilidad de nitrógeno para las plantas y en los procesos biogeoquímicos en el contexto de los invernaderos destinados al cultivo de fresas (Eder *et al.*, 2011). Sin embargo, se requiere de investigaciones posteriores para comprender mejor su papel y las posibles interacciones con otros componentes del ecosistema de suelos en estos entornos específicos.

La observación de suelos deficientes en contenido de carbono permite inferir que su carencia impacta de manera adversa en la calidad del suelo, afectando su capacidad para retener nutrientes y agua.

Otra bacteria que se repitió en varias ocasiones y en diferentes muestras es *Pseudomonas aeruginosa*, La presencia de *Pseudomonas aeruginosa* en los cultivos de fresa puede ser motivo de preocupación en la agricultura (Aguado-Santacruz *et al.*, 2012), diversos estudios han demostrado que la bacteria tiene la capacidad de ser patógena tanto para plantas como para humanos, aunque *Pseudomonas aeruginosa* es más conocida por su papel como patógeno en infecciones hospitalarias y en enfermedades en humanos, también puede afectar a las plantas (Paz-Zarza *et al.*, 2019).

En el caso de los cultivos de fresa, *Pseudomonas aeruginosa* puede causar problemas fitopatológicos, como la marchitez bacteriana, que puede resultar en la pérdida de plantas y reducir el rendimiento de la cosecha (Naranjo Feliciano & Martínez Zubiaur, 2013).

El manejo de *Pseudomonas aeruginosa* en los cultivos de fresa implica prácticas agrícolas adecuadas, como la prevención de la contaminación de agua de riego y la promoción de un ambiente de cultivo saludable. También se pueden utilizar productos fitosanitarios específicos para el control de enfermedades bacterianas, aunque es importante seguir las recomendaciones de uso responsable de estos productos.

Una bacteria de impacto que se localizó en algunas muestras es *Stenotrophomonas rhizophila* esta se localiza de manera general en el suelo principalmente en la rizosfera de las plantas, incluyendo el cultivo de fresas. Esta bacteria puede tener efectos beneficiosos en las fresas, como promover su crecimiento, protegerlas contra patógenos y mejorar su resistencia al estrés ambiental (Hagemann *et al.*, 2008). Sin embargo, la respuesta de las plantas a esta bacteria puede variar según la cepa específica y las condiciones locales del suelo. Por lo tanto, es importante realizar investigaciones y pruebas específicas para determinar si *Stenotrophomonas rhizophila* es beneficiosa o perjudicial en un cultivo de fresas particular. En la agricultura, se está explorando el uso de bacterias benéficas como una forma de mejorar la salud y el rendimiento de los cultivos de manera sostenible. De igual modo se encontraron muchas especies de bacterias más incluyendo las benéficas como las fijadoras de nitrógeno y promotoras del crecimiento, así como las patógenas. Estas bacterias desempeñan roles clave en la nutrición y salud de las plantas, pero su efecto puede variar según las condiciones y prácticas agrícolas. La gestión adecuada del suelo es esencial para un cultivo de fresa exitoso.

En la Tabla 19, se detallan las especies de hongos que exhibieron una predominancia significativa, así como aquellas que mostraron una menor predominancia, basándonos en el número de lecturas detectadas por secuenciación. Es importante destacar que en algunas de las muestras se identificó repetidamente la presencia de *Malassezia restricta*.

Malassezia restricta es una levadura que se encuentra comúnmente en la piel humana y otros ambientes. No obstante, el análisis de las secuencias ambientales de *Malassezia* proporciona información respecto a su distribución geográfica a escala global y su prevalencia en múltiples hábitats, que lo sitúa en una categoría de hongos considerados hiperdiversos. (Enmienda *et al.*, 2014), No es un hongo benéfico para las plantas y, en general, se considera una parte natural del suelo agrícola (Theelen *et al.*, 2018).

Por lo tanto, la presencia de *Malassezia restricta* en el suelo de cultivos de fresa no se consideraría una contribución positiva para el crecimiento de las plantas y podría incluso considerarse un contaminante si se introduce en el suelo de manera no deseada. La salud del suelo y el éxito del cultivo de fresa dependen en gran medida de la composición de la microbiota fúngica y otras condiciones específicas del suelo, más que de la presencia de levaduras como *Malassezia restricta* (Gaitanis *et al.*, 2020).

Otra especie fúngica encontrada en las muestras fue *Penicillium italicum* este es un patógeno que podría afectar a las fresas. Este hongo es más conocido por causar problemas en cítricos, pero en ciertas condiciones puede atacar las fresas y provocar podredumbre y daño en los frutos (Kanashiro *et al.*, 2020). Es importante tomar medidas de control para prevenir la propagación de *Penicillium italicum* en los cultivos de fresa, como la implementación de prácticas de manejo integrado de plagas y enfermedades.

Fusarium oxysporum, una cepa fúngica detectable incluso en concentraciones mínimas como se encontró en la muestra 7 esta no pasa inadvertida debido a su categorización como patógeno edáfico, ejerciendo su capacidad de inducir la enfermedad de la marchitez vascular en los cultivos de fresa. Este patógeno, intrincadamente, obstruye el transporte de agua y nutrientes dentro de la planta, precipitando la atrofia progresiva y, en última instancia, el colapso irreversible del sistema vegetal. Las estrategias de control comprenden la elección de cultivares resistentes, la implementación de rotaciones de cultivos, así como la adopción de prácticas de manejo específicas, todas dirigidas a la prevención de la infección (Vega-López & Granados-Montero, 2023).

En la muestra 10 se encontró de manera predominante *Lachnellula willkommii* este es un hongo patógeno que afecta el cultivo de fresas. Este organismo es el agente causal del "moho de la fresa", una enfermedad que se manifiesta en las hojas de las plantas, generando manchas de color blanco a gris en la superficie superior y crecimiento de moho en la inferior. A medida que progresa la enfermedad, las hojas pueden marchitarse, lo que impacta negativamente el desarrollo y la producción de frutos (Giroux *et al.*, 2020).

El ciclo de vida del hongo involucra su supervivencia en el suelo y en restos de plantas infectadas, y su dispersión mediante esporas liberadas en condiciones húmedas y ventosas.

El control y la prevención de esta enfermedad se logran a través de prácticas de manejo cultural, como la eliminación de hojas infectadas, la mejora de la circulación de aire y el mantenimiento de niveles adecuados de humedad en el cultivo. Además, se pueden aplicar fungicidas específicos cuando la enfermedad se torna problemática. La selección de variedades de fresas resistentes y la implementación de buenas prácticas agrícolas son medidas preventivas esenciales.

Los estudios metagenómicos, que se enfocan en el análisis de comunidades microbianas, han surgido como una herramienta poderosa en el campo de las OMICAS. El advenimiento de las tecnologías de secuenciación masiva ha generado una enorme cantidad de datos, los cuales están almacenados en diversos repositorios públicos. Sin embargo, esta gran cantidad de información dificulta el análisis de calidad y la asignación taxonómica precisa. En la tabla 2 podemos observar la disparidad en los resultados de eficiencia de clasificación entre el uso de la plataforma Epi2me y la implementación de un script diseñado para clasificación de microorganismos. La tasa de clasificación del 72% obtenida mediante la plataforma Epi2me destaca su capacidad para identificar con precisión las lecturas de bacterias y asignarlas a categorías taxonómicas específicas. Esta eficiencia es testimonio de la efectividad del enfoque empleado por Epi2me, que posiblemente se base en una combinación de algoritmos sofisticados y una base de datos taxonómica actualizada y bien curada. El alto porcentaje de lecturas correctamente asignadas sugiere que Epi2me ha logrado interpretar las secuencias genómicas con precisión, lo que es crucial en la investigación de microbiomas y metagenómica, donde la identificación precisa es fundamental para comprender la diversidad microbiana en las muestras. Por otro lado, el resultado de una tasa de clasificación del 54% de bacterias en la muestra obtenida a través del script diseñado es motivo de análisis y consideración. Varias razones podrían contribuir a esta discrepancia. La configuración inadecuada de parámetros, la calidad de las bases de datos de referencia o la complejidad del algoritmo de clasificación podrían influir en la eficacia de la clasificación. Además, el proceso de diseño del script podría haber sido desafiante y requerir más ajustes para lograr una clasificación óptima.

Esta divergencia plantea preguntas sobre la viabilidad y efectividad del enfoque de clasificación personalizado, en comparación con herramientas especializadas como Epi2me. Si bien la flexibilidad y adaptabilidad del script son valiosas, los resultados sugieren que la plataforma

especializada puede ofrecer una mayor confiabilidad en términos de clasificación precisa. Sin embargo, es crucial considerar que la eficiencia de clasificación no es el único criterio de evaluación. Factores como la personalización, el control y los costos asociados también deben tenerse en cuenta al decidir el enfoque a utilizar en un estudio específico.

Del mismo modo en la tabla 3 el logro de una eficiencia de clasificación promedio del 76% mediante la plataforma Epi2me evidencia su capacidad para identificar y asignar con éxito las lecturas genómicas a las categorías taxonómicas correspondientes. Este alto nivel de eficiencia sugiere que la plataforma ha implementado un enfoque robusto y preciso para interpretar las secuencias de los hongos en las muestras. La tasa de lecturas correctamente asignadas subraya la capacidad de Epi2me para abordar la diversidad genómica en muestras de hongos, lo que es esencial para estudios en microbiología y metagenómica.

En contraste, la tasa de clasificación del 47% de hongos obtenida a través de la implementación de un script diseñado revela desafíos potenciales en la eficacia del método personalizado. Este resultado más bajo podría ser indicativo de varios factores que influyen en el rendimiento del script, como la selección de bases de datos de referencia, en este caso se usó National Center for Biotechnology Information (NCBI) ya que es ampliamente reconocido y utilizado en la comunidad científica debido a su accesibilidad, disponibilidad y estatus como una de las tres principales plataformas de referencia para bases de datos en el campo de la biología y la biotecnología. Es importante destacar que el desarrollo de un script personalizado puede ser un proceso complejo que requiere una optimización constante para lograr resultados comparables a los obtenidos por plataformas especializadas como Epi2me.

Los resultados de estos estudios pueden presentar bajo porcentaje de clasificación, y esto se debe a varias razones de naturaleza técnica y biológica. En primer lugar, es esencial considerar que los algoritmos utilizados para la clasificación taxonómica se basan en la comparación de las secuencias con bases de datos de referencia disponible, para el caso de la tesis se usaron los disponibles de la página NCBI. Si las secuencias de interés no están representadas o no están bien caracterizadas en estas bases de datos, es probable que se obtenga una baja tasa de clasificación. Además, la calidad de las secuencias obtenidas durante el proceso de secuenciación puede tener un impacto significativo en la capacidad de los algoritmos para asignar una clasificación

taxonómica precisa. Si las secuencias son de baja calidad, contienen errores o fragmentos incompletos, la capacidad de identificar y clasificar correctamente los organismos presentes en la muestra se verá comprometida. La diversidad y complejidad de los organismos presentes en la muestra también pueden influir en la tasa de clasificación. Si la muestra contiene especies poco conocidas, especies que no están bien representadas en las bases de datos o incluso especies desconocidas, es posible que los algoritmos no puedan asignar una clasificación precisa. Es importante destacar que, en las bases de datos utilizadas para la clasificación de secuencias, existen secuencias que no han sido confirmadas mediante otras técnicas complementarias. Esto plantea la posibilidad de que se produzcan errores en la clasificación taxonómica si se basa únicamente en un fragmento de secuencia. La falta de validación independiente de las secuencias presentes en las bases de datos puede aumentar el riesgo de clasificaciones incorrectas o imprecisas. Por lo tanto, es fundamental considerar estas limitaciones al interpretar los resultados obtenidos y se recomienda realizar análisis adicionales para confirmar la clasificación taxonómica de las secuencias de interés. La optimización de la clasificación de meta genomas está en curso, a medida que se depuran las bases de datos y se incorporan mayores números de secuencias reportadas y debidamente clasificadas. Esto permitirá contar con una herramienta mejorada para realizar una clasificación precisa. Es importante destacar que, si bien las nuevas tecnologías de secuenciación masiva desempeñan un papel fundamental en la generación de nuevo conocimiento, también se requiere una mayor regulación para determinar qué se modifica en las bases de datos de secuencias de ADN. No se trata solo de clasificar hongos y bacterias en función de una región, sino de combinar técnicas complementarias para mejorar su clasificación. La hipótesis planteada en este estudio afirmaba que el diseño de un script en lenguaje Python presentaría una capacidad aceptable a la del programa EPI2ME en términos de detección y clasificación de microorganismos. Para evaluar esta hipótesis, se establecieron objetivos específicos relacionados con el diseño de scripts y la comparación de resultados entre ambos enfoques. Luego de llevar a cabo el diseño y la implementación de los scripts en Python, así como la ejecución de EPI2ME en el mismo conjunto de datos, se pudo realizar una comparación de los resultados. Esta comparación consideró varios aspectos, como la precisión de la clasificación taxonómica, la eficacia en la detección de microorganismos y la consistencia en los resultados obtenidos.

9. CONCLUSIONES

Los resultados de esta investigación indican que la hipótesis formulada es, en gran medida, respaldada por la evidencia. Se encontró que los scripts en Python son capaces de lograr una clasificación taxonómica aceptable a la proporcionada por EPI2ME. Esto sugiere que los scripts en Python pueden ser una alternativa efectiva para la detección y clasificación de microorganismos a partir de datos de secuenciación masiva, ofreciendo la ventaja de ser personalizables y adaptables a las necesidades específicas de un proyecto.

Sin embargo, es importante tener en cuenta que la comparación de estos dos enfoques no se limita únicamente a la precisión de la clasificación. Otros factores, como la velocidad de procesamiento, la facilidad de uso y la escalabilidad, también deben considerarse en una evaluación completa.

En resumen, los resultados obtenidos respaldan la hipótesis inicial de que los scripts en Python tienen la capacidad de equipararse a EPI2ME en términos de detección y clasificación de microorganismos a partir de datos de secuenciación masiva. Esto sugiere que el desarrollo de scripts personalizados en Python puede ser una opción prometedora para investigaciones futuras en esta área.

10.PERSPECTIVAS

1. Ampliación de la búsqueda de secuencias: Es importante continuar realizando búsquedas exhaustivas de secuencias para agregar las más recientes y actualizadas a nuestra base de datos. Esto nos permitirá mejorar la precisión y confiabilidad de la clasificación al tener en cuenta la información más reciente sobre las bacterias.
2. Evaluación comparativa con diferentes bases de datos: Además de la base de datos de microorganismos de Illumina BaseSpace Sequence, se debe considerar la comparación con otras bases de datos ampliamente utilizadas que ofrecen servicios de clasificación de ITS y 16S. Esto proporcionará una visión más completa y comparativa de los resultados obtenidos.
3. Comparación exhaustiva de categorías taxonómicas: Para verificar la funcionalidad y desempeño de los algoritmos de identificación utilizados, es esencial realizar una comparación total de todas las categorías taxonómicas, desde el reino hasta la especie. Esto permitirá evaluar la precisión y consistencia de la clasificación en diferentes niveles taxonómicos.
4. Validación con muestras control: Para garantizar la veracidad y confiabilidad de los resultados, se recomienda poner a prueba los algoritmos de clasificación utilizando muestras control. Estas muestras control contienen una composición microbiana conocida y se utilizarán como referencia para evaluar la precisión de los algoritmos.
5. Integración de enfoques complementarios: Además de la clasificación basada en la región V1-V9 del ARNr ribosomal, es importante considerar la integración de enfoques complementarios. Esto podría incluir el uso de otras regiones ribosomales, como V3-V4 o V6-V8, o incluso la combinación de datos de secuenciación metagenómica y metabarcoding. Estos enfoques complementarios pueden brindar una visión más completa de la diversidad microbiana y mejorar la precisión de la clasificación.

11. REFERENCIAS

1. Aguado-Santacruz, G. A., Moreno-Gómez, B., Jiménez-Francisco, B., García-Moya, E. y Preciado-Ortiz, R. E. (2012). Impacto de los sideróforos microbianos y fitosidéforos en la asimilación de hierro por las plantas: una síntesis. *Revista fitotecnia mexicana*, 35(1), 9–21. https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0187-73802012000100004.
2. Alcalde-Alvites, M. A. (2014). Bioinformática: Tecnologías de la información al servicio de la biología y otras ciencias [Bioinformatics: Information Technology in the Service of Biology and other Sciences]. Páginas del artículo. Recibido: 4-10-2014.
3. Álvarez-Díaz, D. A., Laiton-Donato, K., Franco-Muñoz, C., & Mercado-Reyes, M. (2020). SARS-CoV-2 sequencing: The technological initiative to strengthen early warning systems for public health emergencies in Latin America and the Caribbean. *Biomedica*, 40, 188–197. doi: 10.7705/biomedica.5841.
4. Aoki T., O'Donnell K., Scandiani M.M., 2005. Sudden death syndrome of soybean in South America is caused by four species of Fusarium: *Fusarium brasiliense* sp. Nov., *F. cuneirostrum* sp. Nov., *F. tucumaniae*, and *F. virguliforme*. *Mycoscience*. 46: 162-183.
5. Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2019). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, 20(4), 1125–1136. <https://doi.org/10.1093/bib/bbx120>.
6. Bidartondo, M. I. (2005). The Evolutionary Ecology of Myco-heterotrophy. *New Phytologist*, 167(2), 335-352. <https://doi.org/10.1111/j.1469-8137.2005.01429.x>.
7. Eder, W., Wanner, G., Ludwig, W., Busse, H.-J., Ziemke-Kägeler, F., & Lang, E. (2011). Description of *Undibacterium oligocarboniphilum* sp. nov., isolated from purified water, and *Undibacterium pigrum* strain CCUG 49012 as the type strain of *Undibacterium parvum* sp. nov., and emended descriptions of the genus *Undibacterium* and the species *Undibacterium pigrum*. *International Journal of Systematic and Evolutionary Microbiology*, 61(2), 384–391. <https://doi.org/10.1099/ij.s.0.018648-0>.
8. Enmendar, A. (2014). Desde la caspa hasta los respiraderos de aguas profundas: los hongos similares a *Malassezia* son ecológicamente hiperdiversos. *Patógenos PLoS*, 10(8), e1004277. <https://doi.org/10.1371/journal.ppat.1004277>.

9. Gardes, M., & Bruns, T. D. (1993). Imprimaciones ITS con especificidad mejorada para basidiomicetos: aplicación a la identificación de micorrizas y royas. *Ecología molecular*, 2(2), 113–118. <https://doi.org/10.1111/j.1365-294x.1993.tb00005.x>.
10. Gallo, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2009). El formato de archivo Sanger FASTQ para secuencias con puntajes de calidad y las variantes Solexa/Illumina FASTQ. *Investigación de ácidos nucleicos*, 38(6), 1767-1771. doi:10.1093/nar/gkp1137.
11. Gamaarachchi, H., Samarakoon, H., Jenner, S. P., Ferguson, J. M., Amos, T. G., Hammond, J. M., Saadat, H., Smith, M. A., Parameswaran, S., & Deveson, I. W. (2022). Fast nanopore sequencing data analysis with SLOW5. *Nature Biotechnology*, 40(7), 1026–1029. <https://doi.org/10.1038/s41587-021-01147-4>.
12. Giroux, E., & Bilodeau, G. J. (2020). Whole genome sequencing resource of the European larch canker pathogen *Lachnellula willkommii* for molecular diagnostic marker development. *Phytopathology*, 110(7), 1255–1259. <https://doi.org/10.1094/phyto-09-19-0350-a>
13. Hagemann, M., Ribbeck-Busch, K., Klähn, S., Hasse, D., Steinbruch, R., & Berg, G. (2008). The plant-associated bacterium *Stenotrophomonas rhizophila* expresses a new enzyme for the synthesis of the compatible solute glucosylglycerol. *Journal of Bacteriology*, 190(17), 5898–5906. <https://doi.org/10.1128/jb.00643-08>.
14. Hebert PDN, Cywinska A, Ball SL, deWaard JR. 2003. Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. (B)* 270: 313-321. <http://dx.doi.org/10.1098/rspb.2002.2218>.
15. Hernández, M., Quijada, N. M., Rodríguez-Lázaro, D., & Eiros, J. M. (2020). Aplicación de la secuenciación masiva y la bioinformática al diagnóstico microbiológico clínico. *Revista Argentina de microbiología*, 52(2), 150–161. <https://doi.org/10.1016/j.ram.2019.06.003>
16. Jeewon, R. & Hyde, K. D. (2007) Detection and diversity of fungi from environmental samples: traditional versus molecular approaches. In: Varma A. & Oelmüller, R. (ed.) *Advanced Techniques in Soil Microbiology*. Springer-Verlag.
17. Johnson, J. S., Spakowicz, D. J., Hong, B. Y., Petersen, L. M., Demkowicz, P., Chen, L., ... Weinstock, G. M. (2019). Evaluación de la secuenciación del gen 16S rRNA para el análisis de microbiomas a nivel de especie y cepa. *Nature Communications*, 10(1), 5029. DOI: 10.1038/S41467-019-13036-1. PMID: 31695033; PMCID: PMC6834636.

18. Kanashiro, A. M., Akiyama, D. Y., Kupper, K. C., & Fill, T. P. (2020). *Penicillium italicum*: An Underexplored Postharvest Pathogen. *Frontiers in microbiology*, 11. <https://doi.org/10.3389/fmicb.2020.606852>.
19. Kerkhof, L. J. (2021). ¿Está preparada la secuenciación de nanoporos de Oxford para analizar microbiomas complejos? *FEMS Microbiología Ecología*, 97(3). <https://doi.org/10.1093/femsec/fiab001>.
20. Kirk J., Beaudette L., Hart M., Moutoglis P., Klironomos J., Lee H. y Trevors J. T. (2004). Methods of studying soil microbial diversity. *Journal of Microbiological Methods*, 58 (2), 169-188.
21. Kress WJ, Erickson DL. 2012. DNA Barcodes: Methods and Protocols. *Methods in Molecular Biology* 858. Humana Press/Springer, New York, 470 pp.
22. Korabecna, M. (2007) The variability in the fungal ribosomal DNA (ITS1, ITS2, and 5.8 S rRNA gene): its biological meaning and application in medical mycology. *Communicating Current Research and Educational Topics and Trends in Applied Microbiology*.
23. Lai, X., Cao, L., Tan, H., Fang, S., Huang, Y. & Zhou, S. (2007) Fungal communities from methane hydrate-bearing deep-sea marine sediments in South China Sea. *The ISME journal*, 1 (8), 756-762.
24. Liferder. (28 de junio de 2023). Medio Stuart. Recuperado de: <https://www.liferder.com/medio-stuart/>.
25. Liu Z, DeSantis TZ, Andersen GL, Knight R. 2008. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res.* 36: e120-e120. <http://dx.doi.org/10.1093/nar/gkn491>.
26. Lopez de Heredia, U. (2016). Las técnicas de secuenciación masiva en el estudio de la diversidad biológica. *Munibe Ciencias Naturales*, 64, 10.21630/mcn.2016.64.07.
27. Marcano Godoy, K. A., & Delvasto Angarita, P. L. (2015). Contaminación de suelos por metales pesados debido a la presencia de pilas gastadas. *Revista de Investigación*, 40(88). Universidad Pedagógica Experimental Libertador. Instituto Pedagógico de Caracas, Venezuela, & Escuela de Ingeniería Metalúrgica y Ciencia de los Materiales, Universidad Industrial de Santander, Colombia.

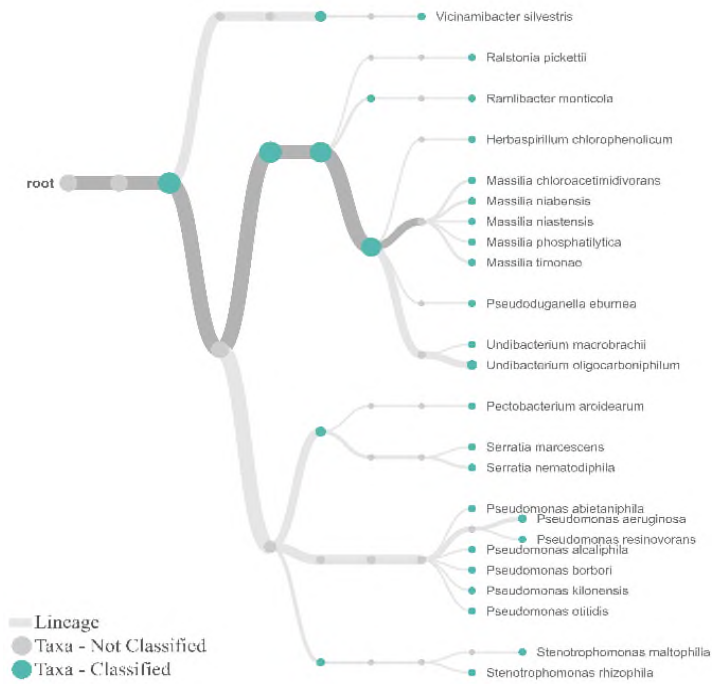
28. McBain, A. J., O'Neill, C. A., & Oates, A. (2016). Skin Microbiology☆. En Reference Module in Biomedical Sciences. Elsevier.
29. Meneses Escobar, C. A. (2011). Bioinformatics Technologies for the Analysis of DNA sequences. *Scientia et Technical*, 16(49), 56-66. ISSN 0122-1701.
30. Monard, C., Gantner, S. & Stenlid, J. (2012) Utilizing ITS1 and ITS2 to study environmental fungal diversity using pyrosequencing. *FEMS microbiology ecology*, 84, 165-175.
31. Naranjo Feliciano, E., & Martínez Zubiaur, Y. (2013). Avances en el diagnóstico de la marchitez bacteriana (*Ralstonia solanacearum*): situación actual y perspectivas en Cuba. *Revista de protección vegetal*, 28(3), 160–170.
32. Nejm. (2023, Agosto). Illustrated Glossary - NEJM. *New England Journal of Medicine*. Recuperado 5 de noviembre de 2023, de https://illustrated-glossary.nejm.org/term/16S_ribosomal_rna_gene_sequencing.
33. Nilsson, R. H., Ryberg, M., Abarenkov, K., Sjökvist, E., & Kristiansson, E. (2009). The ITS region as a target for characterization of fungal communities using emerging sequencing technologies. *FEMS Microbiology Letters*, 296(1), 97-101.
34. Nogales, B., (2005). La microbiología del suelo en la era de la biología molecular: descubriendo la punta del iceberg. *Ecosistemas*, 14(2),41-50. [fecha de Consulta 19 de septiembre de 2023]. ISSN: 1132-6344. Recuperado de: <https://www.redalyc.org/articulo.oa?id=54014206>.
35. Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2015). Krona: Interactive Metagenomic visualization in a web browser. En Springer eBooks (pp. 339-346). https://doi.org/10.1007/978-1-4899-7478-5_802.
36. Oregon State University. (2023). Manual de Kraken 2.0. <https://software.cqls.oregonstate.edu/updates/docs/kraken2/MANUAL.html>.
37. Paz-Zarza, V. M., Mangwani-Mordani, S., Martínez-Maldonado, A., Álvarez-Hernández, D., Solano-Gálvez, S. G., & Vázquez-López, R. (2019). *Pseudomonas aeruginosa*: patogenicidad y resistencia antimicrobiana en la infección urinaria. *Revista Chilena de Infectología: Organo Oficial de La Sociedad Chilena de Infectología*, 36(2), 180–189. <https://doi.org/10.4067/s0716-10182019000200180>.

38. Prokopyev, M. S., Vlasova, E. Z., Tretyakova, T. V., Sorochinsky, M. A., & Solovyeva, R. A. (2020). Development of a programming course for students of a teacher training higher education institution using the programming language Python. *Propósitos y Representaciones*, 8(3). <https://doi.org/10.20511/pyr2020.v8n3.484>.
39. Quaiser, A., Ochsenreiter, T., Klenk, H. P., Kletzin, A., Treusch, A. H., Meurer, G., *et al.* (2002). First insight into the genome of an uncultivated chrenarchaeote from soil. *Environmental Microbiology*, 4, 603-611.
40. Quaiser, A., Ochsenreiter, T., Lanz, C., Schuster, S. C., Treusch, A. H., Eck, J., & Schleper, C. (2003). Acidobacteria form a coherent but highly diverse group within the bacterial domain: evidence from environmental genomics. *Molecular Microbiology*, 50, 563-575.
41. Ravelo-Méndez, W., Escobar-Velásquez, C., & Linares-Vásquez, M. (2023). Kraken 2.0: A platform-agnostic and cross-device interaction testing tool. *Science of Computer Programming*, 225(102897), 102897. <https://doi.org/10.1016/j.scico.2022.102897>.
42. Rodríguez-Santiago, B., & Armengol, L. (2012). Tecnologías de secuenciación de nueva generación en diagnóstico genético pre y postnatal. *Diagnóstico Prenatal*, 23(2), 56-66.
43. Rondon, M. R., August, P. R., Bettermann, A. D., Brady, S. F., Grossman, T. H., Liles, M. R., *et al.* 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* 66: 2541-2547 Saiki, R. K., GELFAND, D.H., STOFFEL, S., SCHARF, S.J., HIGUCHI, R., HORN, G.T., MULLIS, K.B. y Erlich, H.A. (1988). Primer-directed enzymatic amplification of DNA with thermostable DNA polymerase. *Sc.* 239, 487-494.
44. Sedlar, Karel. (2018). *Methods for Comparative Analysis of Metagenomic Data*. 10.13140/RG.2.2.12388.42882.
45. Soergel, DA, Dey, N., Knight, R. y Brenner, SE (2012). Selección de cebadores para la clasificación taxonómica óptima de las secuencias del gen 16S rRNA ambiental. *The ISME journal*, 6 (7), 1440–1444. <https://doi.org/10.1038/ismej.2011.208>.
46. Schoch, C. L., Seifert, K. A., Huhndorf, S., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, A., Chen, W. & Consortium, Fungal Barcoding. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences*. 109 (16), 6241-6246.

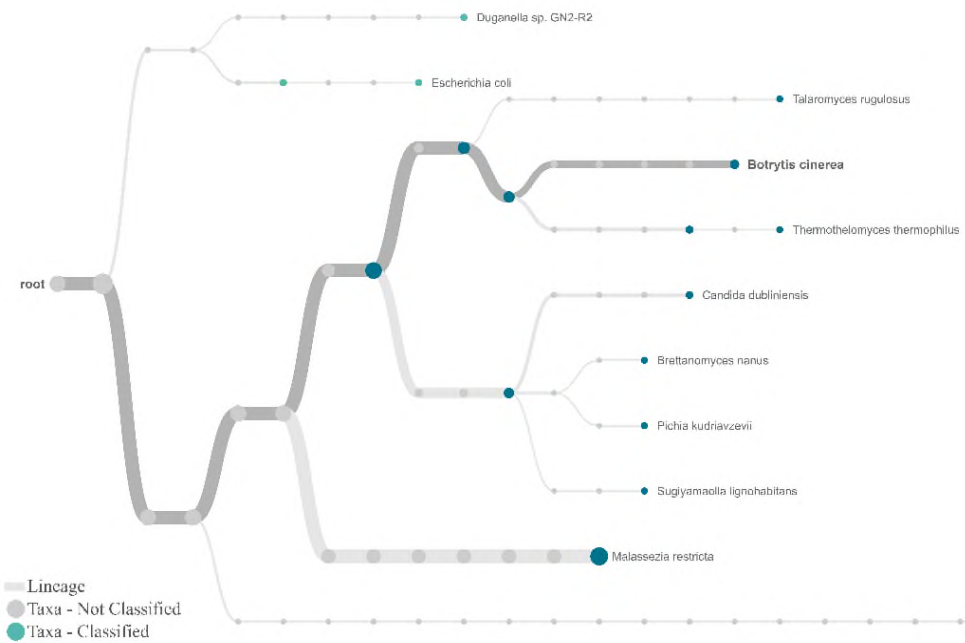
47. Stajich, J. E., Berbee, M. L., Blackwell, M., Hibbett, D. S., James, T. Y., Spatafora, J. W., & Taylor, J. W. (2009). The fungi. *Current Biology*, 19(18), R840-R845.
48. Tanabe AS, Toju H. 2013. Two new computational methods for universal DNA barcoding: A benchmark using barcode sequences of bacteria, archaea, animals, fungi, and land plants. *PLOS ONE* 8: e76910. <http://dx.doi.org/10.1371/journal.pone.0076910>.
49. Tedersoo, L., Nilsson, R. H., Abarenkov, K., Jairus, T., Sadam, A., Saar, I., Bahram, M., Bechem, E., Chuyong, G. & Kõljalg, U. (2010) 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytologist*, 188, 291-301.
50. Theelen, B., Cafarchia, C., Gaitanis, G., Bassukas, I. D., Boekhout, T., & Dawson, T. L., Jr. (2018). *Malassezia* ecology, pathophysiology, and treatment. *Medical Mycology: Official Publication of the International Society for Human and Animal Mycology*. <https://doi.org/10.1093/mmy/myx134>.
51. Valenzuela-González, Fabiola, Casillas-Hernández, Ramón, Villalpando, Enrique, Vargas-Albores, Francisco, & Harris, Christine. (2015). El gen ARNr 16S en el estudio de comunidades microbianas marinas. *Ciencias marinas*, 41(4), 297-313. <https://doi.org/10.7773/cm.v41i4.2492>.
52. Vega-López, M., & Granados-Montero, M. del M. (2023). Eficacia de benomil y folpet sobre *Fusarium oxysporum* patógeno de la fresa. *Revista mexicana de ciencias agrícolas*, 14(3), 485–490. <https://doi.org/10.29312/remexca.v14i3.3253>.
53. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. y Au, K. F. (2021). Tecnología de secuenciación de nanoporos, bioinformática y aplicaciones. *Nature Biotechnology*, 39(11), 1348–1365. <https://doi.org/10.1038/s41587-021-01108-x>.
54. Wang, G., & Peng, B. (2019). Script of Scripts: Un Sistema de flujo de trabajo pragmático para la investigación computacional diaria [Script of Scripts: A pragmatic workflow system for everyday computational research]. *Nombre de la revista*, 15(2), e1006843. doi: 10.1371/journal.pcbi.1006843. PMID: 30811390; PMCID: PMC6411228.
55. Wang, Z., Nilsson, R. H., Lopez-Giraldez, F., Zhuang, W., Dai, Y., Johnston, P. R. & Townsend, J. P. (2011) Tasting soil fungal diversity with earth tongues: phylogenetic test of SATé alignments for environmental ITS data. *PloS one*, 6 (4), 1-12.

56. W. R. Pearson, «Flexible Sequence Similarity Searching with the FASTA3 Program Package», en *Bioinformatics Methods and Protocols*, New Jersey: Humana Press, 1999, pp. 185-219.
57. White, T. J., Bruns, T., Lee, S. & Taylor, J. (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis, Gelfand, M. A., D. H., Sninsky, J. J. & White, J. T., *PCR protocols: a guide to methods and applications*, (pp. 315-322). California, Academic Press.
58. Wood & Salzberg, 2014) Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3). <https://doi.org/10.1186/gb-2014-15-3-r46>.

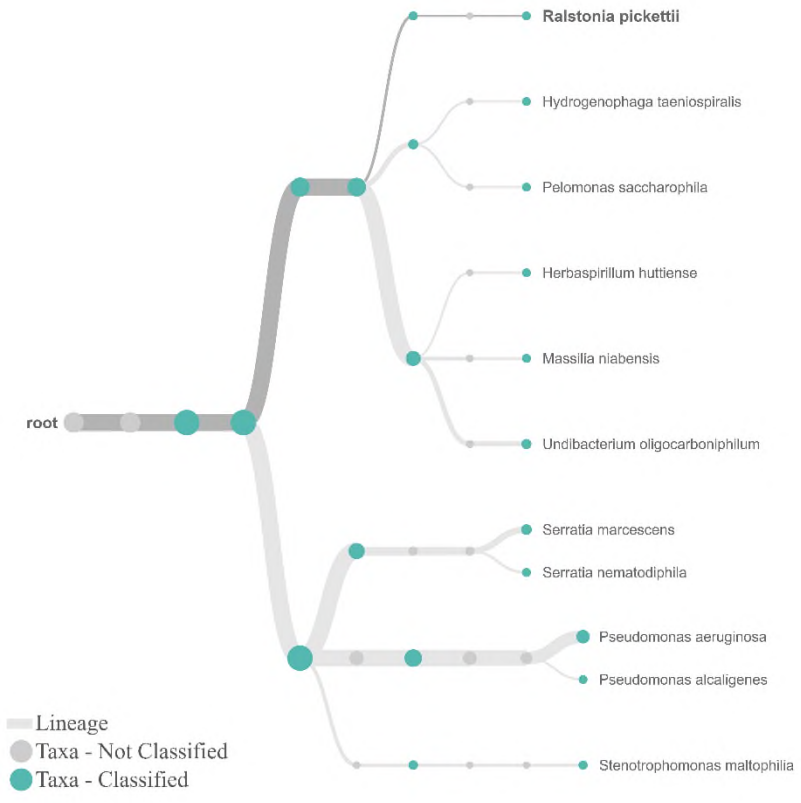
11. ANEXO



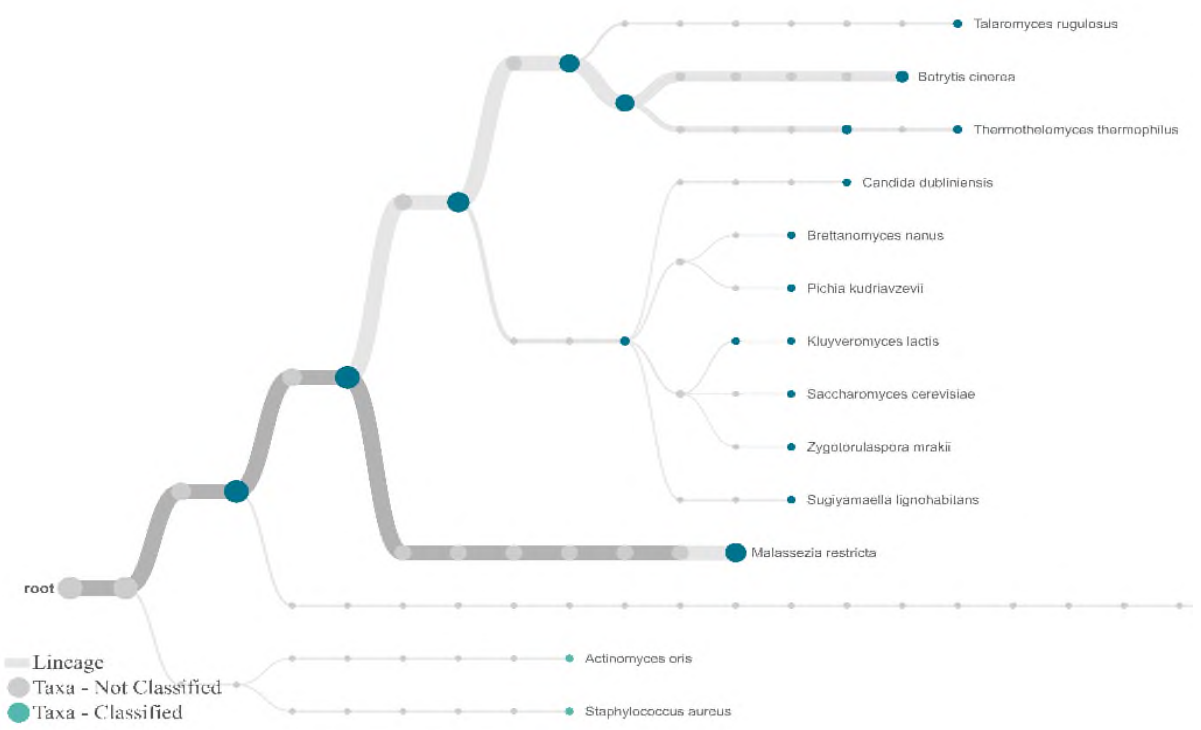
Clasificación taxonómica de bacterias en la muestra 2.



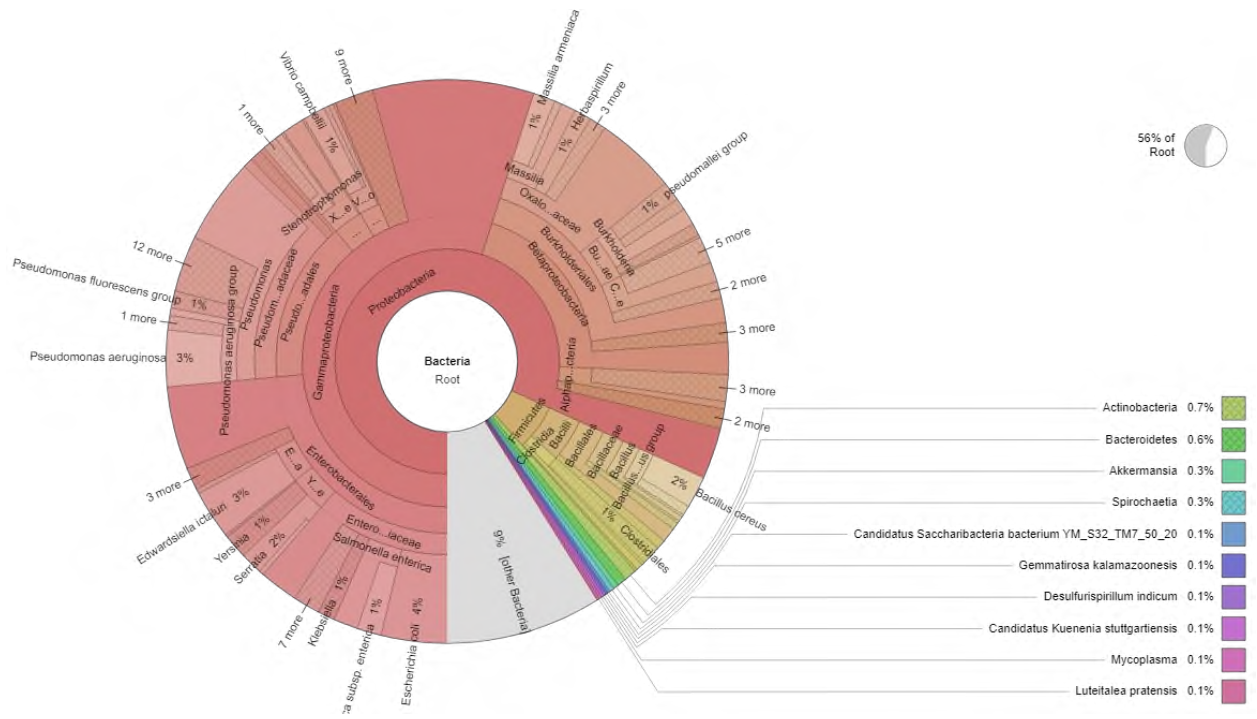
Clasificación taxonómica de hongos en la muestra 2.



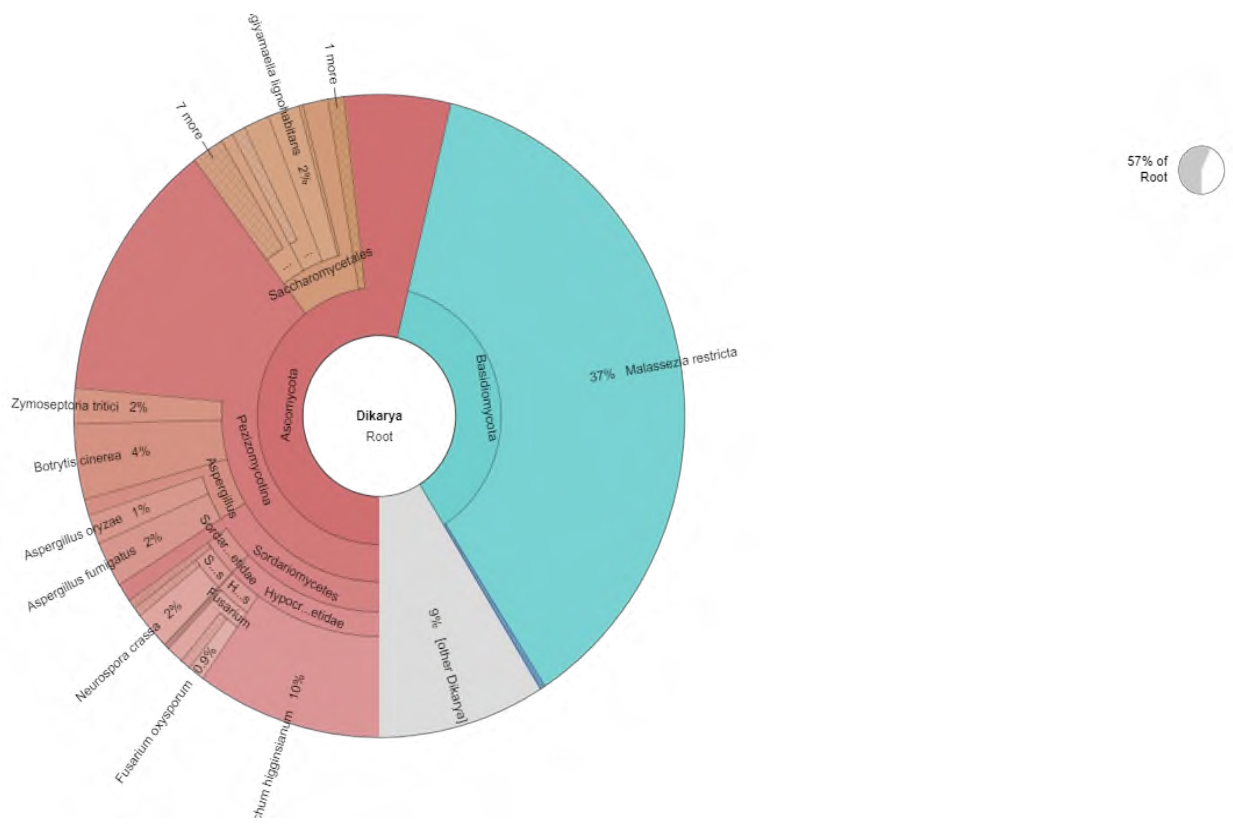
Clasificación taxonómica de bacterias en la muestra 4.



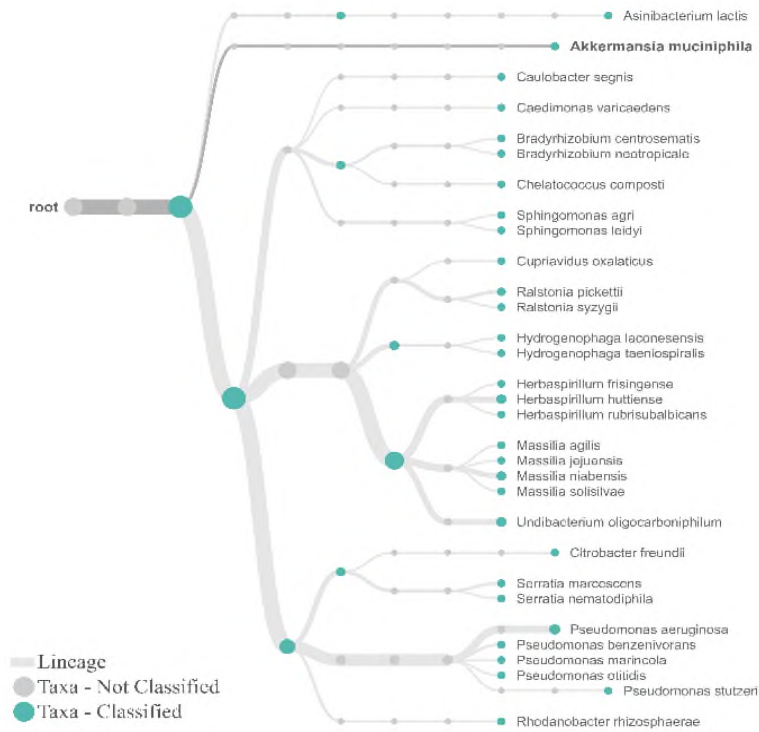
Clasificación taxonómica de hongos en la muestra 4.



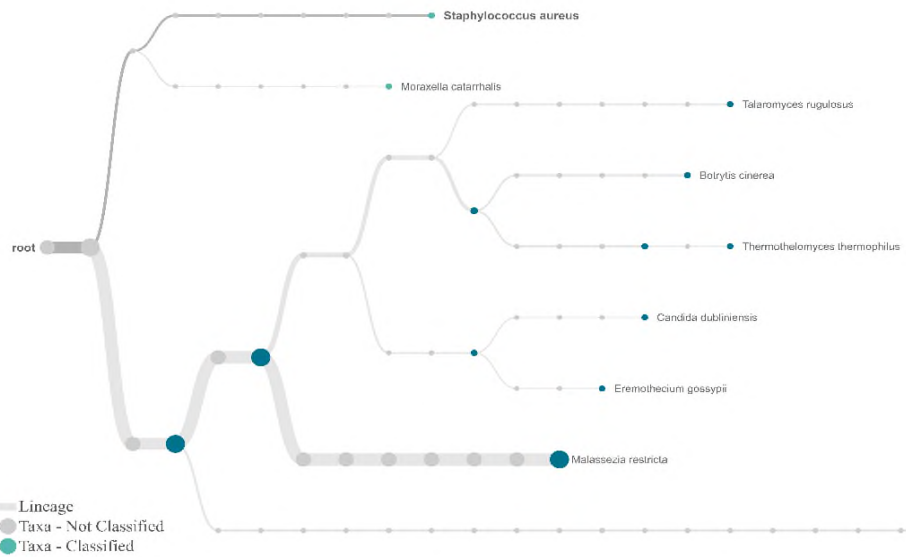
Clasificación taxonómica porcentual de bacterias en la muestra 4.



Clasificación taxonómica porcentual de hongos en la muestra 4.

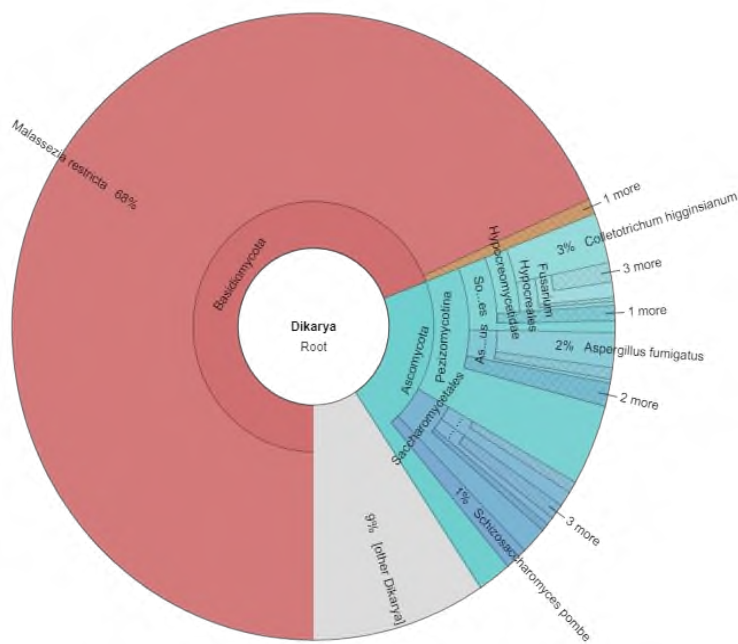
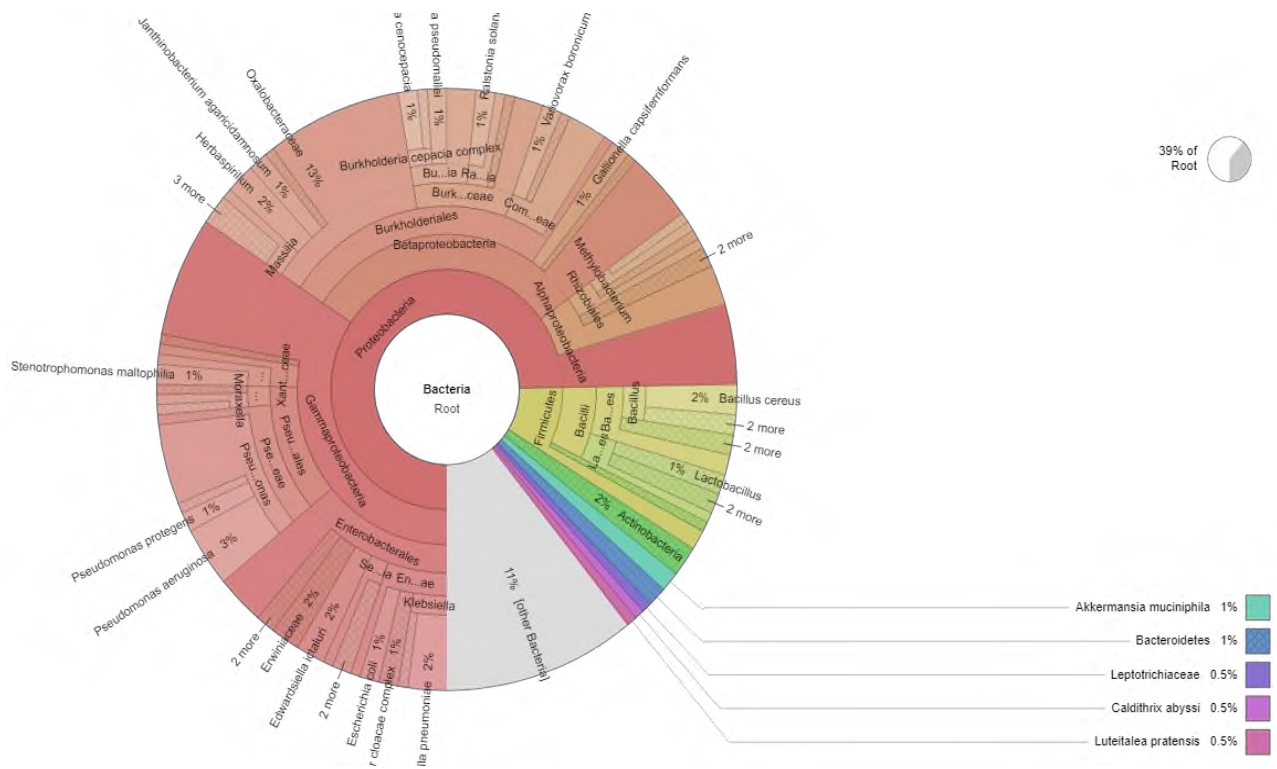


Clasificación taxonómica de bacterias en la muestra 5.

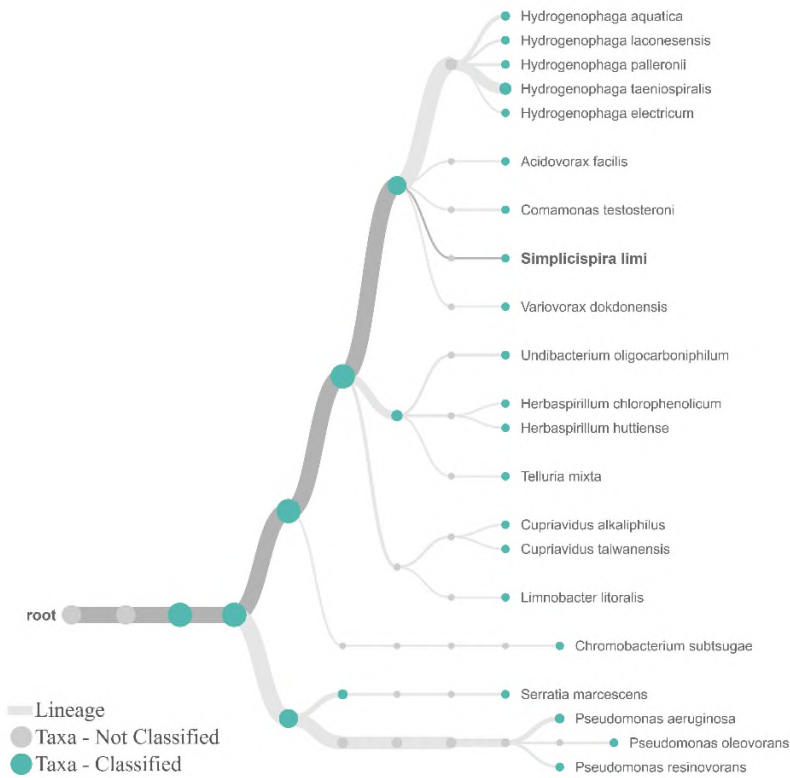


Clasificación taxonómica de hongos en la muestra 5.

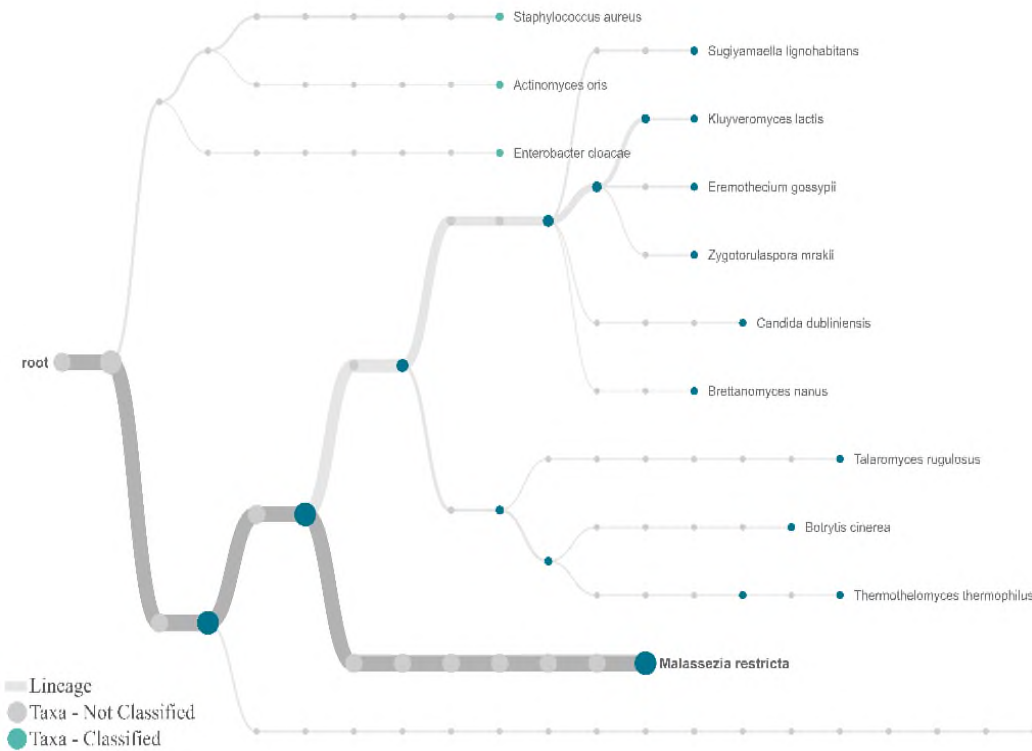
Clasificación taxonómica porcentual de bacterias en la muestra 5.



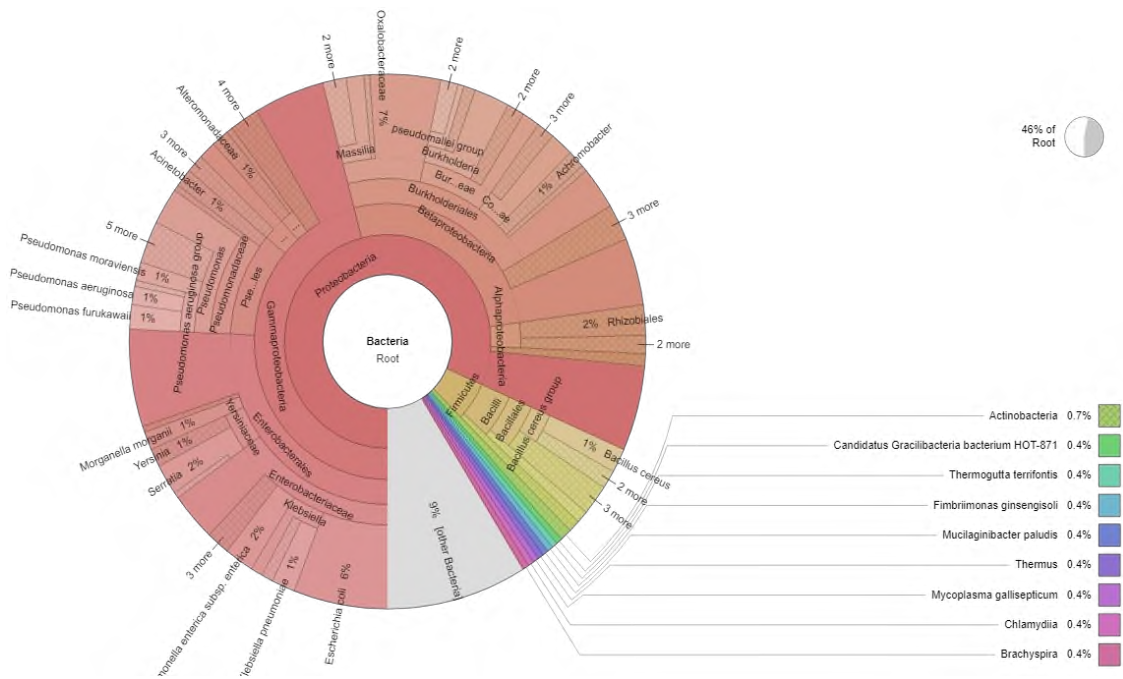
Clasificación taxonómica porcentual de hongos en la muestra 5.



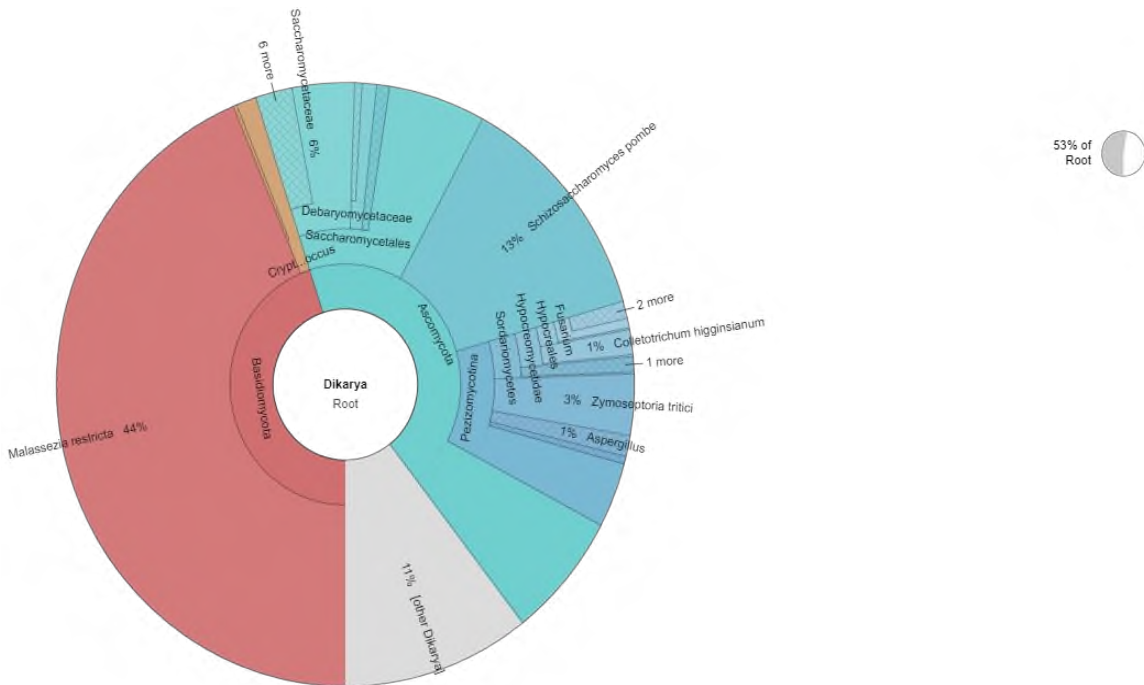
Clasificación taxonómica de bacterias en la muestra 6.



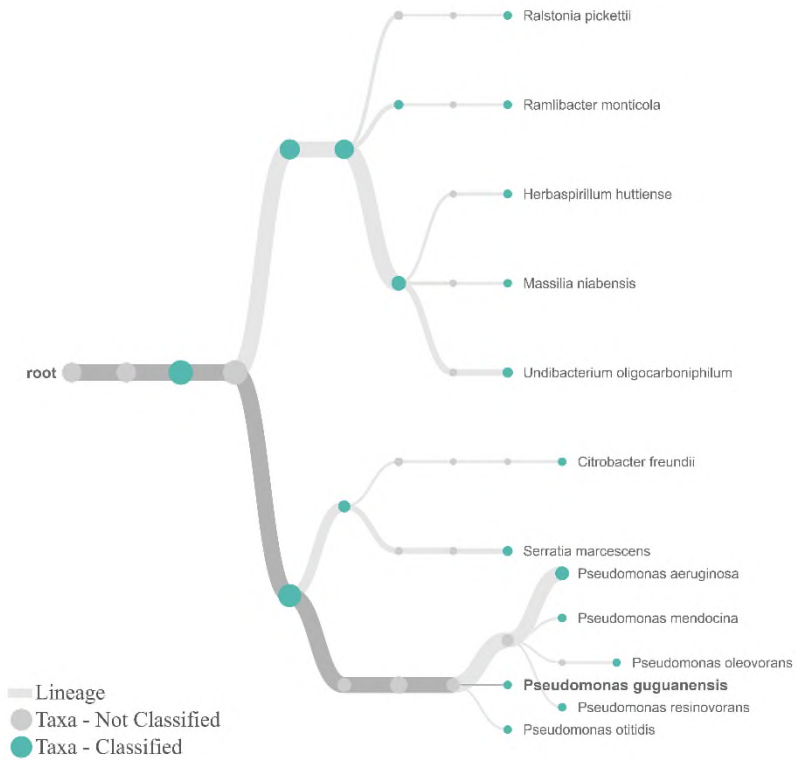
Clasificación taxonómica de hongos en la muestra 6.



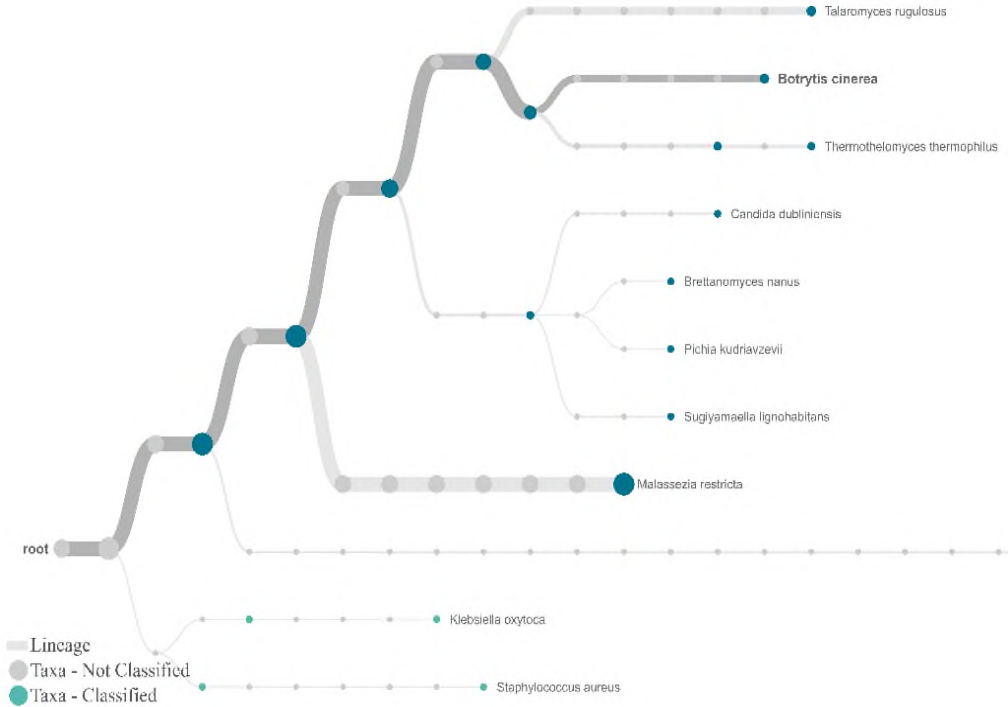
Clasificación taxonómica porcentual de bacterias en la muestra 7.



Clasificación taxonómica porcentual de hongos en la muestra 7.



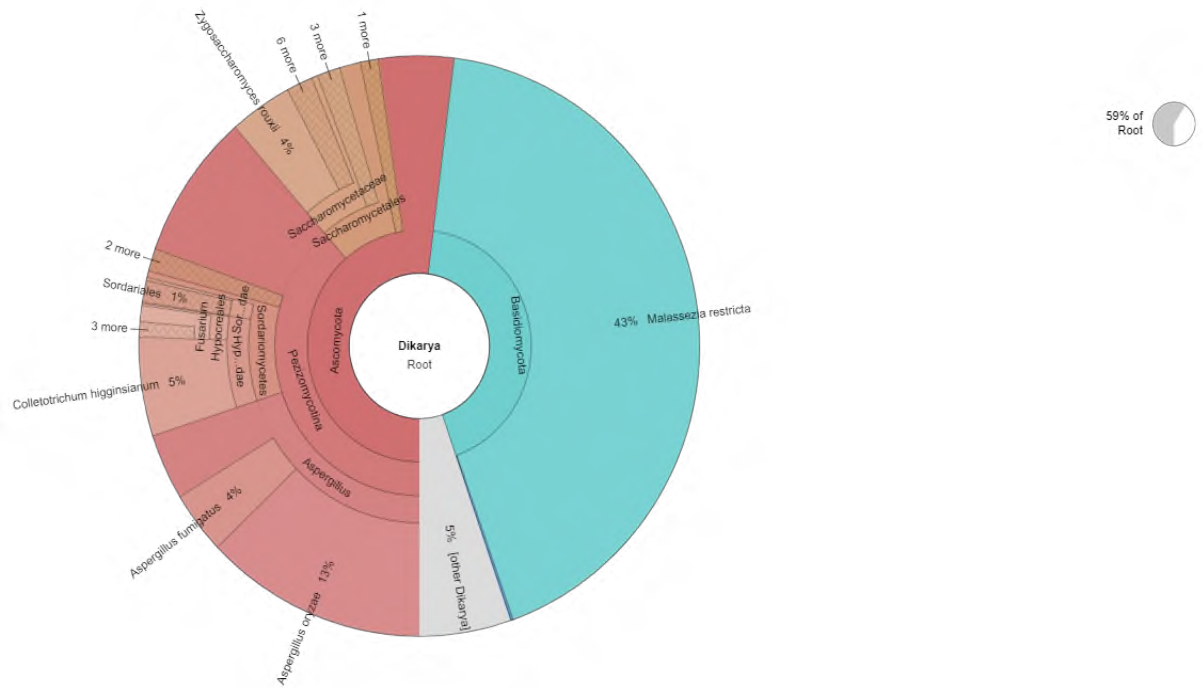
Clasificación taxonómica de bacterias en la muestra 8.



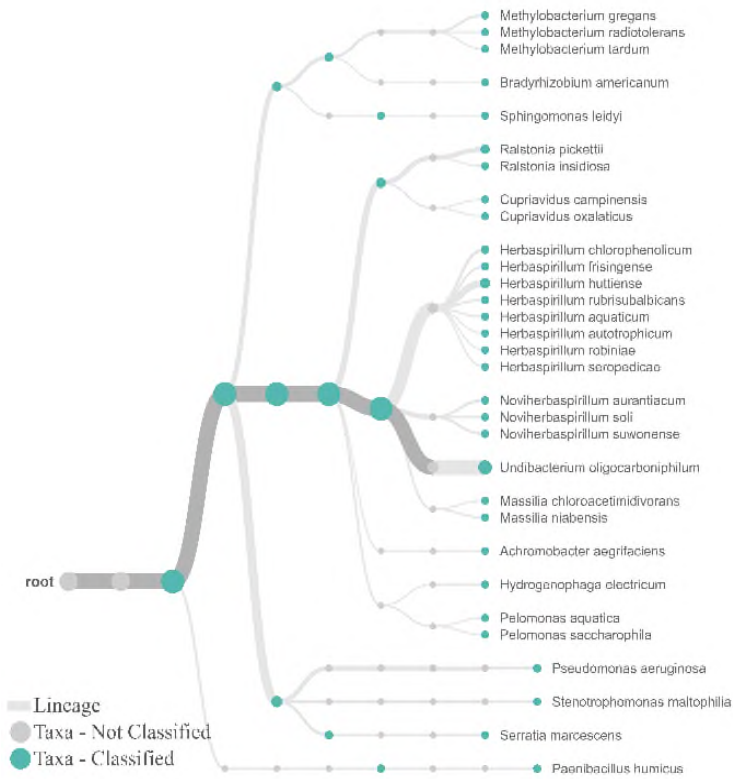
Clasificación taxonómica de hongos en la muestra 8.



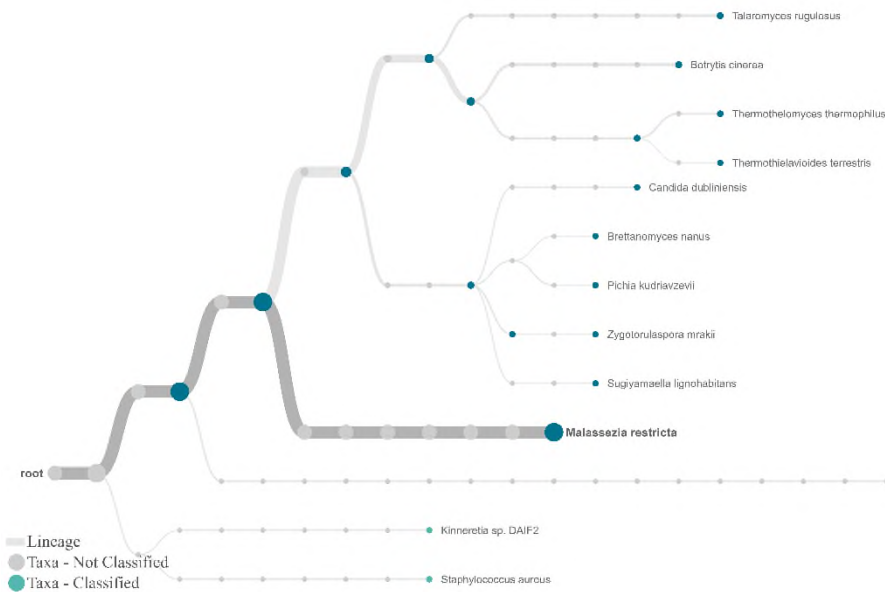
Clasificación taxonómica porcentual de bacterias en la muestra 8.



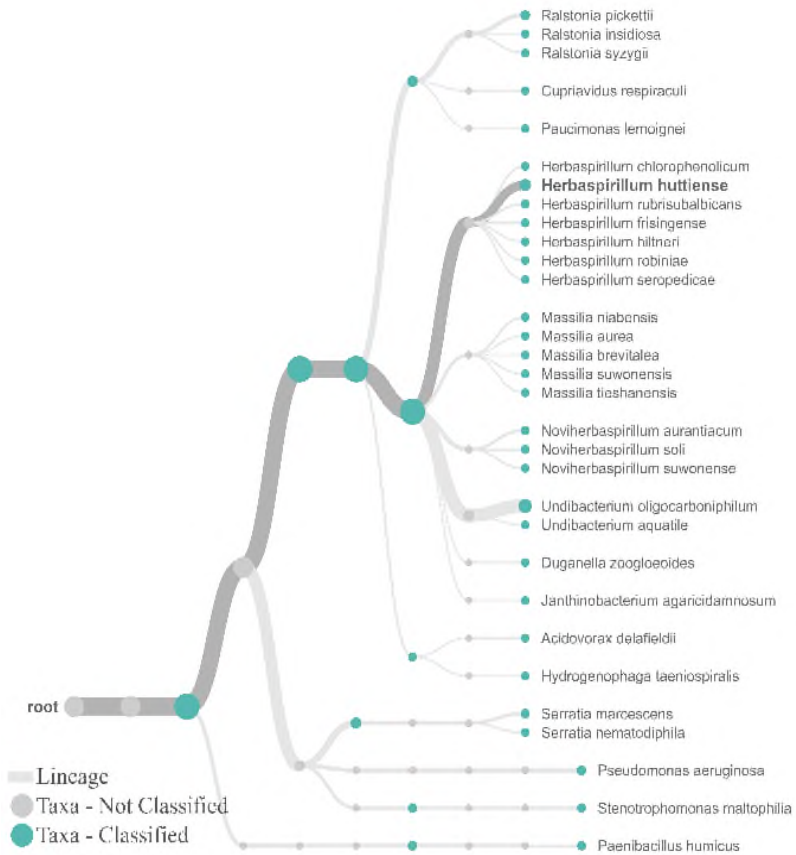
Clasificación taxonómica porcentual de hongos en la muestra 8.



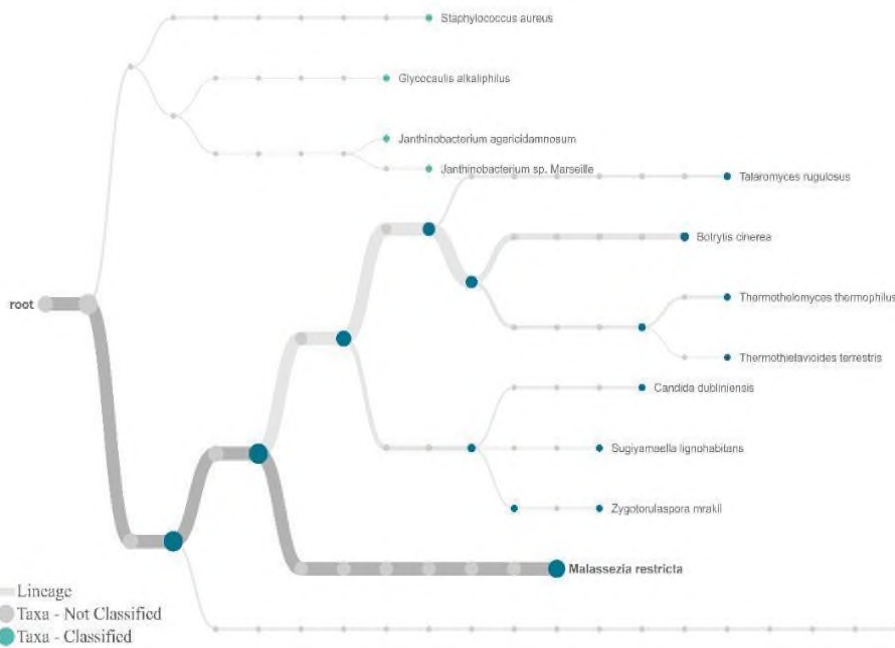
Clasificación taxonómica de bacterias en la muestra 9.



Clasificación taxonómica de hongos en la muestra 9.



Clasificación taxonómica de bacterias en la muestra 10.



Clasificación taxonómica de hongos en la muestra 10.

REPOSITORIO VIRTUAL

Datos	Enlace
Fast5 a Fastq.	https://drive.google.com/drive/folders/1qKJ2ITGGHnEI72ueCT8UgvdGAlzFL5H?usp=sharing
Script de Fast5 a Fastq.	https://drive.google.com/drive/folders/1F1dgezWP5N0u3kjpFi41eilkgXGA5wVR?usp=sharing
Fastq separ ar por código o de barra.	https://drive.google.com/drive/folders/1N8ZiHoy6WIUcrpw4cOuogDU9p8QoSZDE?usp=sharing
Script para separ ar por Código o.	https://drive.google.com/drive/folders/1F1dgezWP5N0u3kjpFi41eilkgXGA5wVR?usp=sharing

Repositorio de las secuencias usadas para el análisis.

Repositorio	Dirección
Google drive	https://drive.google.com/drive/folders/1Ah3SxStaPD_S7SHw1_zpknweAIEh0528?usp=share link

Repositorio de las secuencias por muestra.

Google drive	https://drive.google.com/drive/folders/1cXhYv_ScU6BOPPoVMYp_FBSFYqXrqKt-?usp=sharing
---------------------	---

Repositorio de las muestras de 16S, con su análisis de Epi2me.

Google drive	https://drive.google.com/drive/folders/1Ah3SxStaPD_S7SHw1_zpknweAIEh0528?usp=share link
---------------------	---

Repositorio de las muestras de ITS, con su análisis de Epi2me.

Google drive	https://drive.google.com/drive/folders/1xwNzUQH8wx20tAOWUBwjS1S0yMzJ4YS1?usp=share link
---------------------	---

Lista de microorganismo usados en Kraken 2.0.

Repositorio	Dirección
Google drive	https://drive.google.com/drive/folders/1C2BE7UZKqv3U2Fz1alvuWTIBuyHkMFpj?usp=share link