UNEMI
UNIVERSIDAD ESTATAL DE MILAGRO

REPÚBLICA DEL ECUADOR

UNIVERSIDAD ESTATAL DE MILAGRO

VICERRECTORADO DE INVESTIGACIÓN Y POSGRADO

FACULTAD DE POSGRADOS

INFORME DE INVESTIGACIÓN

PREVIO A LA OBTENCIÓN DEL TÍTULO DE:

MAGÍSTER EN BIOTECNOLOGÍA

TEMA:

Desarrollo y evaluación de un pipeline en R para análisis metagenómico de secuencias ITS de Illumina.

AUTOR:

Blgo. Ricardo Tamayo

TUTOR:

Ing. Juan Valenzuela Cobos, PhD.

MILAGRO, 2026

Derechos de Autor

Sr. Dr.

Fabricio Guevara Viejó

Rector de la Universidad Estatal de Milagro

Presente.

Yo, **Ricardo Andrés Tamayo Cevallos,** en calidad de autor y titular de los derechos morales y patrimoniales de este informe de investigación, mediante el presente documento, libre y voluntariamente cedo los derechos de Autor de este proyecto de desarrollo, que fue realizada como requisito previo para la obtención de mi Grado, de **Magíster en Biotecnología**, como aporte a la Línea de Investigación **Generación de estrategias para la Biorremediación** de conformidad con el Art. 114 del Código Orgánico de la Economía Social de los Conocimientos, Creatividad e Innovación, concedo a favor de la Universidad Estatal de Milagro una licencia gratuita, intransferible y no exclusiva para el uso no comercial de la obra, con fines estrictamente académicos. Conservo a mi favor todos los derechos de autor sobre la obra, establecidos en la normativa citada.

Así mismo, autorizo a la Universidad Estatal de Milagro para que realice la digitalización y publicación de este Proyecto de Investigación en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

El autor declara que la obra objeto de la presente autorización es original en su forma de expresión y no infringe el derecho de autor de terceros, asumiendo la responsabilidad por cualquier reclamación que pudiera presentarse por esta causa y liberando a la Universidad de toda responsabilidad.

Milagro, **13 de abril del 2026**

_____

Ricardo Andres Tamayo Cevallos

C.I.: 0931991095

Aprobación del Tutor del Trabajo de Titulación

Yo, **Juan Valenzuela Cobos**, en mi calidad de tutor del trabajo de titulación, elaborado por **Ricardo Andrés Tamayo Cevallos**, cuyo tema es **Desarrollo y evaluación de un pipeline en R para análisis metagenómico de secuencias ITS de Illumina.**, que aporta a la Línea de Investigación **Generación de estrategias para la Biorremediación**, previo a la obtención del Grado **Magíster en Biotecnología**. Trabajo de titulación que consiste en una propuesta innovadora que contiene, como mínimo, una investigación exploratoria y diagnóstica, base conceptual, conclusiones y fuentes de consulta, considero que el mismo reúne los requisitos y méritos necesarios para ser sometido a la evaluación por parte del tribunal calificador que se designe, por lo que lo **APRUEBO**, a fin de que el trabajo sea habilitado para continuar con el proceso de titulación de la alternativa de Informe de Investigación de la Universidad Estatal de Milagro.

Milagro, 18 de mayo del 2025

Validar únicamente en FirmaEC.
Firmado electrónicamente por:
JUAN DIEGO VALENZUELA COBOS

PhD. Juan Valenzuela Cobos

C.I.:

ACTA DE SUSTENTACIÓN

En la Facultad de Posgrado de la Universidad Estatal de Milagro, a los trece días del mes de abril del dos mil veintiseis, siendo las 08:00 horas, de forma VIRTUAL comparece el/la maestrante, BIOL. TAMAYO CEVALLOS RICARDO ANDRES, a defender el Trabajo de Titulación denominado **" DESARROLLO Y EVALUACIÓN DE UN PIPELINE EN R PARA ANÁLISIS METAGENÓMICO DE SECUENCIAS ITS DE ILLUMINA ."**, ante el Tribunal de Calificación integrado por: CHENCHE LOPEZ OSCAR MAURICIO, Presidente(a), Ph.D. VALENZUELA COBOS JUAN DIEGO en calidad de Vocal; y, Ing. SANCHEZ VASQUEZ VIVIANA LORENA que actúa como Secretario/a.

Una vez defendido el trabajo de titulación; examinado por los integrantes del Tribunal de Calificación, escuchada la defensa y las preguntas formuladas sobre el contenido del mismo al maestrante compareciente, durante el tiempo reglamentario, obtuvo las siguientes calificaciones:

TRABAJO DE TITULACION	59.67
DEFENSA ORAL	39.00
PROMEDIO	**98.67**
EQUIVALENTE	**EXCELENTE**

Para constancia de lo actuado firman en unidad de acto el Tribunal de Calificación, siendo las 09:00 horas.

FIRMADO ELECTRONICAMENTE POR:
OSCAR MAURICIO CHENCHE LOPEZ

CHENCHE LOPEZ OSCAR MAURICIO
PRESIDENTE/A DEL TRIBUNAL

JUAN DIEGO VALENZUELA COBOS

Ph.D. VALENZUELA COBOS JUAN DIEGO
VOCAL

VIVIANA LORENA SANCHEZ VASQUEZ

Ing. SANCHEZ VASQUEZ VIVIANA LORENA
SECRETARIO/A DEL TRIBUNAL

Ricardo Andres Tamayo Cevallos

BIOL. TAMAYO CEVALLOS RICARDO ANDRES
MAGÍSTER

Dedicatoria

¡A Megan, por supuesto!

Agradecimientos

En cada etapa de nuestras vidas hay tantas personas a las que, sin lugar a duda, tenemos que agradecer. Por tanto, mencionar a cada una de ellas aquí sería, al menos hasta el momento, interminable. Pero ello no significa que no merezcan un lugar aquí.

Agradezco a mis padres, que han confiado plenamente en mí en cada decisión que he tomado, quienes me han acompañado a lo largo de mi vida académica y me han orientado en el buen camino. Les debo tanto. También, a mis hermanos, quienes en todo momento han celebrado cada uno de mis logros como si fueran suyos.

A mis amigos de toda la vida Esteban, Caicedo, Dennis y Kevin quienes siempre han estado atentos a cómo me iba en esta nueva etapa. ¡¡Gracias!!

Resumen

La caracterización de la diversidad fúngica mediante métodos tradicionales presenta limitaciones significativas, acentuadas por la barrera técnica que supone el análisis bioinformático de datos de secuenciación masiva. Este estudio tuvo como objetivo desarrollar y validar un pipeline bioinformático reproducible en R para el análisis metagenómico de secuencias de la región ITS fúngica generadas por Illumina. La metodología implementada integró herramientas estandarizadas (DADA2, phyloseq) para el control de calidad, inferencia de Variantes de Secuencia de Amplicón (ASVs), asignación taxonómica utilizando la base de datos UNITE, y análisis de diversidad alfa y beta. Los resultados demostraron la eficacia del pipeline, obteniéndose perfiles de calidad optimizados, matrices de ASVs y visualizaciones que revelaron la composición de las comunidades fúngicas. Los análisis ecológicos identificaron a Ascomycota y Basidiomycota como los filos dominantes y permitieron discernir patrones significativos en la composición de las comunidades entre muestras. Se concluye que el pipeline desarrollado supera las limitaciones de los métodos dependientes de cultivo, caracterizando eficientemente la "materia oscura microbiana". Su diseño accesible y reproducible democratiza el acceso a análisis metagenómicos robustos, constituyendo una herramienta valiosa para aplicaciones en biotecnología y ecología.

Palabras clave: ITS, Metagenómica, Ecología, Diversidad, Bioinformática.

Abstract

Characterizing fungal diversity through traditional methods faces significant limitations, exacerbated by the technical barrier posed by the bioinformatic analysis of massive sequencing data. This study aimed to develop and validate a reproducible bioinformatic pipeline in R for the metagenomic analysis of fungal Internal Transcribed Spacer (ITS) region sequences generated by Illumina. The implemented methodology integrated standard tools (DADA2, phyloseq) for quality control, Amplicon Sequence Variant (ASV) inference, taxonomic assignment using the UNITE database, and alpha and beta diversity analysis. The results demonstrated the pipeline's efficacy, yielding optimized quality profiles, ASV matrices, and visualizations that revealed fungal community composition. Ecological analyses identified Ascomycota and Basidiomycota as the dominant phyla and allowed for the discernment of significant patterns in community composition among samples. In conclusion, the developed pipeline overcomes the limitations of culture-dependent methods by efficiently characterizing the microbial "dark matter". Its accessible and reproducible design democratizes access to robust metagenomic analyses, establishing a valuable tool for applications in biotechnology and ecology.

Keyword: ITS, Metagenomics, Ecology, Diversity, Bioinformatics.

Lista de Figuras

Lista de Tablas

Índice / Sumario

Contenido

Introducción

Los microorganismos son omnipresentes en casi todos los ecosistemas, desde bosques tropicales, hasta ambientes extremos, como fumarolas submarinas o áreas hipersalinas; en general, este grupo incluye arqueas, bacterias, virus, protistas y hongos. La gran diversidad de organismos microscópicos que coexisten de forma intra e interespecífica, permite el correcto funcionamiento de los biomas del planeta tierra. Es así como, al conjunto de todos los microorganismos como a sus genes en un área o hospedero específico se le denomina Microbioma (Bai et al. 2025, Xia et al. 2023). Las comunidades microbianas (Microbioma) cumplen roles importantes dentro de las zonas en las qué se encuentran como el reciclaje de nutrientes, tanto como productores primarios o moviendo la materia orgánica a niveles tróficos superiores (Gosh et al. 2022).

Esta constante interacción permite el mantenimiento de la productividad de los ecosistemas, sin embargo, esto se ve limitado en mayor o menor medida por parámetros ambientales a las que se ven sometidos a causa de la naturaleza del hábitat, variables como salinidad, presencia de iones metálicos, pH entre otros, regulan la abundancia y dominancia de ciertos grupos frente a otros. En este sentido, es de carácter imperativo conocer el microbioma de varios habitas de interés, tanto para sectores como la agricultura (Nagvire et al. 2022), salud humana (Ma et al 2024), así como en la industria biotecnológica (Bai et al. 2025) para la búsqueda de nuevos metabolitos de interés.

No obstante, se estima que la mayoría de los microorganismos en los ecosistemas son no cultivables en laboratorio (Li et al., 2023). Esta limitación de las técnicas tradicionales genera un vacío en el conocimiento, denominado "*Dark matter*

microbiana", que contiene una gran porción de la diversidad y funcionalidad de los microbiomas. Para solucionar esta problemática han emergido una diversidad de técnicas que no dependen de métodos de cultivo, lo que ha permitido un avance fundamental en el estudio de los microbiomas, abriendo un abanico de posibilidades para el descubrimiento de metabolitos de interés, así como cepas con actividad catalítica con aplicaciones industriales(New & Brito, 2020).

Para hacer frente a esta enorme brecha que significa el desconocimiento de la real diversidad de microorganismos que no son cultivables, las técnicas omicas han emergido como una nueva ruta para la exploración del microbioma, permitiendo describir aspectos ecológicos, químicos y genéticos de una gran variedad de microorganismos (Fadiji & Babalola, 2020). Recientemente, una de las técnicas que han ganado aceptación dentro de la comunidad científica para el estudio de la diversidad de los microbiomas es la metagenómica. De acuerdo con el trabajo pionero de Handelsmanl et al. (1998), se define esta técnica como el clonado y análisis funcional de genomas que se encuentran en una muestra ambiental. La implementación de la metagenómica en estudios de diversidad microbiana ha permitido tener una perspectiva más amplia del complejo microbioma, así como también entender como este se relaciona con el hospedero, tanto en especies de plantas de interés comercial y en el microbioma intestinal de los seres humanos (Molefe et al.2021; Navgire et al., 2022).

En metagenómica se aísla directamente ADN a partir de un conjunto de microorganismos que provienen en una muestra ambiental (suelo, agua, biopsias), permitiendo así una caracterización funcional de las comunidades microbianas. Lo que difiere con el método tradicional de secuenciación Sanger, que solo permite la

extracción y posterior secuenciación de una única muestra (Hu et al., 2021), lo que, comparado con el complejo microbioma, que puede tener desde miles a millones de secuencias, obtener una sola, genera resultados pobres sobre la real diversidad microbiana (Nayfach et al., 2021).

Inicialmente, la metagenómica empleaba la secuenciación de genes marcadores específicos, como el gen 16S rRNA para procariotas y las regiones ITS *(Internal Transcribed Spacer*) para hongos. A este enfoque basado en regiones específicas se le denomino metagenómica basada en amplicones, lo que permitió la asignación taxonómica y ecológica al lograr la identificación de nuevos linajes y la caracterización del microbioma (biodiversidad) en una variedad de muestras ambientales (suelo y agua), así como el microbiota humano (Caporaso et al., 2011; Fierer et al., 2012).

Por lo tanto, esta técnica como pilar fundamental en el análisis de microbiomas ha sido empleado en varios campos como la biomedicina, en donde ha permitido conocer la influencia del microbioma en la salud y la patogénesis de enfermedades. En la agronomía, por otra parte, es fundamental para la identificación de microorganismos con propiedades promotoras del crecimiento vegetal, interacciones planta-patógeno, así como el uso adecuado de agentes de biocontrol. En la biorremediación, ha propiciado el descubrimiento de microorganismos con rutas metabólicas capaces de degradar derivados de hidrocarburos y una variedad de contaminantes industriales (Caruso, 2015 ;Gilbert et al., 2018).

A pesar de esto, como se mencionó anteriormente, la metagenómica genera miles o millones de secuencias a partir de una sola muestra, lo que se traduce en terabytes de información, a medida que se aumenta el número de muestras, lo que claramente requiere un alto poder de análisis informático (procesamiento y análisis) y

almacenamiento. Puesto que al contar con millones de secuencias cortas (*short-reads*), su ensamblaje e identificación taxonómica no puede ser realizado con los métodos analíticos convencionales (Zhu et al. 2022;Tas et al. 2021).

En este contexto, es perentorio la especialización en el manejo de herramientas de programación como R y Python. Estos lenguajes permiten acceder a toda la capacidad computacional requeridas para gestionar grandes conjuntos de datos, emplear algoritmos complejos y realizar tareas iterativas. R como lenguaje de programación de acceso libre posee una amplia biblioteca de herramientas diseñadas específicamente para el análisis de secuencias (phyloseq, dada2), lo qué es fundamental para el análisis estadístico y la visualización de datos metagenómicos, mientras que Python, con sus librerías para manipulación de secuencias y diseño de flujos de trabajo, es invaluable para el preprocesamiento y la automatización (McMurdie & Holmes, 2013; Van der Auwera & Goodwin, 2020).

A pesar de la creciente demanda de análisis metagenómicos, se observa una carencia notable de profesionales con una formación interdisciplinaria que combine microbiología, secuenciación y bioinformática (Lewis & Bartlett, 2013). Esta disparidad entre las habilidades requeridas y la disponibilidad de personal cualificado crea un obstáculo considerable para el avance de la investigación y la aplicación práctica de la metagenómica, especialmente en países en desarrollo. Por consiguiente, el desarrollo de *scripts* y *pipelines* robustos que racionalicen el análisis de datos metagenómicos es de suma utilidad y pertinencia. Estos *workflows* predefinidos pueden ayudar enormemente en el estudio de los microbiomas ya que establecen una ruta lógica para la correcta limpieza, remoción de *primers*, asignación taxonómica (OTUs o ASV) lo que faculta a investigadores sin una especialización

profunda en esta área a ejecutar análisis sofisticados de manera práctica y reproducible (Lewis & Bartlett, 2013).

En conclusión, la creación de *workflows* que emplean software libre no solo permite alcanzar e implementar el uso de herramientas de análisis genómicos actuales, sino que tambiénagiliza la caracterización de la diversidad microbiana, permitiendo que un mayor número de científicos aprovechen la vasta cantidad de datos generados y transformen el potencial inherente de los microorganismos en soluciones tangibles para los desafíos globales (Sholder et al., 2020; Oulas et al., 2015).

CAPÍTULO I: El Problema de la Investigación

1.1 Planteamiento del problema

Los hongos junto a las bacterias contribuyen en gran medida a la biodiversidad de los microbiomas, así como también a la funcionalidad de los ecosistemas, actuando como promotores de crecimiento en plantas, proveyendo de ciertas vitaminas y cofactores, de generar asociaciones simbióticas con plantas, patógenos y degradadores de materia orgánica, permitiendo así al reciclaje de nutrientes (Berendsen et al., 2012; Schmidt et al., 2019; Bahram et al., 2021). Considerando que los hongos corresponden a uno de los más diversos clados en eucariotas, con estimados de alrededor de 2.2 – 3.8 millones de especies, han tenido una baja tasa de descripción cuando se compara con su contraparte eucariota (Wang, Kirk & Yao, 2020).

En ese sentido, la mayoría de las técnicas tradicionales de aislamiento y cultivo, mismas que han sido ampliamente usadas desde hace décadas, solo han permitido el reconocimiento de una pequeña fracción de este grupo, lo que ha limitado el conocimiento sobre la real biodiversidad de los hongos en los ecosistemas. Dada está clara limitante para el desarrollo de varias áreas, el advenimiento de herramientas moleculares permitió un gran avance en el estudio de muestras ambientales, generando un abanico de nuevos descubrimientos, desde entender la genética de estas especies, hasta reportar nuevas especies basándose únicamente en las secuencias genéticas de estas especies sin reportar el análisis morfológico correspondiente (Xia et al., 2023; Li et al. 2024).

Dentro de este último contexto, el uso de hongos ha sido clave para desarrollar estrategias que fomentan y explotan los recursos naturales. Géneros como *Aspergillus*, *Penicillium*, *Trichoderma*, *Saccharomyces* y *Talaromyces* son bien conocidos por sus capacidades metabólicas de interés industrial, biotecnológico y

medicinal, facilitando el desarrollo de numerosas aplicaciones al ser relativamente sencillos de aislar y cultivar en laboratorio. Sin embargo, la plasticidad metabólica de este grupo no se restringe a estos géneros; muchos otros, como *Rhizopus*, *Candida* y *Cephalosporium*, también producen metabolitos de gran interés. Por consiguiente, comprender su presencia y abundancia no solo en ambientes convencionales, sino también en áreas extremófilas, es fundamental para la búsqueda y descubrimiento de nuevas especies con potencial biotecnológico (Cuadros-Orellana et al., 2013).

A pesar de la plasticidad metabólica de los hongos, la facilidad de cultivo de ciertos géneros en laboratorio, aún la biodiversidad de este grupo sigue siendo la principal limitante para el desarrollo de estrategias orientadas al uso de recursos naturales. Puesto que en áreas tropicales que se consideran megadiversas, las investigaciones que exploran la diversidad de este reino siguen manteniendo una gran brecha del conocimiento y aún otras siguen sin ser exploradas en su totalidad, generándose año a año descubrimientos de nuevas especies. Lo que claramente hace un llamado a una investigación continúa y profunda de estas áreas (Hawksworth & Lücking, 2017)

Esta brecha se ve aún más grande cuando ciertas especies de hongos no pueden ser cultivadas en laboratorio y requieren únicamente de observaciones *in situ* para confirmar su presencia al interactuar con otras especies al observar patologías asociadas a la infecciones o procesos ecológicos que de forma indirecta hacen referencia a la presencia de estas especies. Sumado a esto, debido a los requerimientos nutricionales y condiciones ambientales específicas que dificultan el registro y estudio *in vitro* de este grupo de hongos, el uso de herramientas moleculares (secuenciación de nueva generación) ha permitido tener un conocimiento más

profundo de la real diversidad de hongos en un ecosistema dado, lo que, claramente permite explorar los metabolitos que podrían encontrarse en estas especies.

Por lo tanto, el presente trabajo pretende desarrollar un pipeline enfocado en el análisis metagenómico de secuencias ITS, marcador molecular característicos de hongos, para realizar análisis de *short-reads* provenientes de Illumina.

1.2 Delimitación del problema

La importancia ecológica de los hongos en los ecosistemas es más que conocida, sin embargo, mantener enfoques dependientes de cultivo, tiene claras limitaciones al momento de establecer roles de interacciones e inducciones metabólicas que los hongos generan con algunos de sus hospederos. Para superar esta clara limitante, el empleo de técnicas genómicas que permitan establecer la real diversidad en un área determinada es de carácter imperativo (Tas et al., 2021). A pesar de la variedad de plataformas que permiten tener este acercamiento a los roles ecológicos que este grupo cumple en las zonas donde se encuentran, la falta de personal especializado que realicen funciones interdisciplinarias entre la biología y informática es bastante notoria (Tedersoo et al. 2022).

Sumado a que, la mayoría de software que se han desarrollado para estudiar estos microbiomas se realizan a través de línea de comando (Bash) en Sistemas operativos de acceso abierto, que en la mayoría de los casos no es el más empleado por investigadores. Esto genera un cuello de botella y una limitante para áreas de interés agronómica como lo son países tropicales como el Ecuador.

Por lo tanto, el presente trabajo propone el desarrollo de un *pipeline* que sea de acceso libre y de fácil empleabilidad para investigadores que requieran un análisis ecológico de alto nivel estadístico y que permita la inclusión de procesos iterativos de

forma fácil. Además, que sea manipulable a los datos del investigador, sin perder la confiabilidad del análisis, sumado a que todo esto debe desarrollarse en un entorno de software que sea estandarizado para la mayoría de Los investigadores. Permitiéndoles así, describir los microbiomas con gráficos que permitan observar e interpretar de mejor manera las interacciones ecológicas.

1.3 Formulación del problema

Desarrollar un *pipeline* de bioinformática accesible y eficiente para el análisis de datos metagenómicos de la región ITS, que permita superar las limitaciones del cultivo en laboratorio y la barrera técnica en el análisis computacional, facilitando así la caracterización de la diversidad microbiana en países en desarrollo como Ecuador

1.4 Preguntas de investigación

- ¿Cuál es la eficacia de diferentes algoritmos de preprocesamiento y bases de datos taxonómicas para la identificación precisa de ASVs (Variantes de Secuencia Amplicón) fúngicas a partir de datos de secuenciación Illumina?
- ¿En qué medida la aplicación del pipeline desarrollado revela una mayor diversidad filogenética de hongos en comparación con los métodos tradicionales de aislamiento y cultivo en un caso de estudio de suelo tropical ecuatoriano?
- ¿Cómo la implementación de una interfaz gráfica de usuario (GUI) o de *scripts* altamente automatizados en R reduce la barrera de entrada para que investigadores microbiológicos sin experiencia en bioinformática realicen análisis metagenómicos ecológicamente robustos?

1.5 Objetivos

1.5.1 Objetivo general

Desarrollar y validar un pipeline de bioinformática accesible y reproducible en R para el análisis de datos metagenómicos del espaciador transcrito interno (ITS) fúngico, que permita caracterizar la diversidad y composición de comunidades microbianas en muestras ambientales.

1.5.2 Objetivos específicos

- Diseñar e implementar un pipeline automatizado que integre las prácticas bioinformáticas estandarizadas para el preprocesamiento, filtrado de calidad, desduplicación (ASV) y asignación taxonómica de secuencias de la región ITS.
- Validar la funcionalidad del *pipeline* en archivos .fastaq provenientes de secuencias de Illumina.
- Caracterizar la composición taxonómica y la diversidad filogenética de las comunidades fúngicas presentes en las muestras ambientales mediante el análisis de secuencias de la región ITS.

1.6 Hipótesis

El desarrollo de un pipeline de bioinformática reproducible y accesible, implementado en R, permitirá caracterizar de manera significativamente más eficiente y completa la diversidad fúngica de muestras ambientales complejas, superando las limitaciones de los métodos cultivo-dependientes y revelando la presencia de táxones crípticos con potencial biotecnológico.

1.7 Justificación

El estudio del microbioma es fundamental para la comprensión de los sistemas ecológicos y biológicos complejos. Como se destacó previamente, los microorganismos son un componente esencial de los procesos biogeoquímicos y su diversidad funcional es crucial para la estabilidad ecosistémica. La metagenómica ha revelado que la gran mayoría de estos microorganismos son "no cultivables", dejando un vasto *dark matter* microbiano inexplorado. Por lo tanto, la capacidad de analizar datos metagenómicos de forma eficiente es un imperativo científico para determinar la composición y función de estas comunidades inobservables y aprovechar su potencial.

De igual forma, la identificación de nuevos linajes filogenéticos y la inferencia de capacidades metabólicas en hongos no cultivables valida la importancia de una herramienta que permita a más investigadores acceder a esta información. El desarrollo de un pipeline accesible democratiza el acceso a metodologías avanzadas, alineándose con los principios de ciencia abierta y colaborativa, permitiendo que más grupos de investigación contribuyan al conocimiento del microbioma fúngico. La secuenciación Illumina de ITS, aunque potente, produce volúmenes de datos que sobrepasan las capacidades de análisis manual o con software comercial de licencia

restrictiva. La escasez de personal especializado en bioinformática dentro de los laboratorios de biología y ecología es un cuello de botella bien documentado

Este trabajo pretende desarrollar un pipeline de fácil ejecución para usuarios no familiarizados con lenguaje de programación, dado que, en la actualidad existe una gran cantidad de pipelines desarrollados para análisis metagenómicos, sin embargo, la vasta mayoría está enfocada el uso de línea de comando en sistemas operativos basados en Unix y otros en plataformas de acceso libre. Además, Si bien existen pipelines metagenómicos, pocos están específicamente optimizados para ITS fúngicos en R y diseñados con una interfaz simplificada para usuarios no expertos en bioinformática.

Por lo tanto, se priorizaría el desarrollo de *scripts* donde el usuario solo necesite especificar la ruta de sus archivos .fastq y su tabla de metadatos. El *script* manejaría la mayor parte de las configuraciones internas, con parámetros predeterminados optimizados para ITS.

1.8 Declaración de las variables (Operacionalización)

Tabla 3. Operacionalización de las Variables

Variable	Dimensión	Definición	Indicadores	Instrumento de medición
VI[1]: *pipeline* para secuencias ITS de Illumina	Configuración técnica	Conjunto de herramientas para el filtrado, *Trimming* y demás pasos de análisis de secuencia ITS	Algoritmos para análisis de secuencias y bases de datos para la asignación taxonómica	Scripts en R
	Reproducibilidad	Capacidad de obtener resultados consistentes en diferentes entornos	Coherencia en resultados al ejecutar el pipeline en múltiples muestras	Comparación estadística
VD[2]: Eficacia del Pipeline	Calidad del procesamiento	Efectividad en la filtración y limpieza de secuencias	Porcentaje de lecturas retenidas tras el filtrado	métricas de DADA2 y scripts de biodiversidad
VD: Diversidad Fúngica	Diversidad Alfa	Riqueza y abundancia de especies dentro de cada muestra	Riqueza de especies, Índice de Shannon y Simpson	Paquete phyloseq y Vegan en R
	Diversidad Beta	Diferencias en la composición de comunidades entre muestras	Distancia de Bray-curtis Agrupamiento de PCoA y NMDS	Análisis de PCoA, NMDS.
	Composición Taxonómica	Abundancia relativa entre cada grupo de hongos	Abundancia relativa de géneros/Phylum (%)	visualización de datos usando ggplot

[1]Variable independiente. [2]Variable dependiente

-CAPÍTULO II: Marco Teórico Referencial

2.1 Antecedentes Referenciales

El estudio sistemático y exhaustivo de la diversidad biológica en el planeta, especialmente en reinos relativamente menos explorados, como los hongos, ha sido históricamente limitado por técnicas dependientes de cultivo. Los métodos convencionales de micología, basados en el aislamiento, cultivo y clasificación morfológica de los hongos, a menudo solo logran identificar una fracción menor de las especies presentes en un ecosistema dado, infraestimando significativamente la verdadera riqueza y diversidad de este grupo (Ravin et al., 2018). Esta limitación se debe a que una gran porción de la micobiota es no cultivable o requiere condiciones de crecimiento muy específicas que son difíciles de replicar, dejando un vacío sin estudiar. El advenimiento de las técnicas moleculares de *High-Throughput Sequencing* (HTS) ha revolucionado este campo, permitiendo una exploración mucho más profunda y completa de las comunidades fúngicas ambientales (Cuadros-Orellana et al. 2013).

El *metabarcoding* ha emergido como la principal herramienta para estudios ecológicos de comunidades microbianas complejas. Esta metodología se basa en la secuenciación masiva de una región genética corta y estandarizada, también llamado código de barras (*barcode*), que es amplificada directamente a partir del ADN ambiental (eDNA) de una muestra. La principal ventaja de este enfoque es su capacidad para identificar la composición taxonómica de múltiples organismos de manera simultánea, permitiendo superar la limitante necesidad de aislar y cultivar cada especie individualmente (Aranguren et al., 2023).

A partir del trabajo de White et al. (1990) el marcador genético más aceptado y ampliamente utilizado a nivel mundial para el estudio de hongos es el espaciador transcrito interno (ITS). El ITS es una región de ADN no codificante que forma parte del *operón* de ARN ribosomal, anidada entre los genes que codifican para los ARN ribosomales 18S y 28S. Su valor reside en que presenta una variación considerable, lo que permite la discriminación a nivel de especie, mientras que sus regiones flanqueantes son lo suficientemente conservadas para permitir el diseño de *primers* universales que pueden amplificar el ADN fúngico de una amplia gama de taxones. Las dos subregiones más comunes utilizadas en el *metabarcoding* son el ITS1 y el ITS2 (Schoch et al. 2012).

El proceso de *metabarcoding* implica una serie de pasos secuenciales y rigurosos. Comienza con la recolección del eDNA, que puede ser tan diverso como el suelo en regiones áridas o el material genético de especímenes de museos. A esto le sigue la extracción eficiente del ADN total y la amplificación por PCR de la región ITS con los *primers* universales. Posteriormente, se realiza la secuenciación de alto rendimiento, que genera millones de lecturas de secuencias cortas. La etapa crucial que le sigue es el análisis bioinformático, donde las secuencias se limpian de errores y quimeras, se agrupan en Unidades Taxonómicas Operacionales (OTUs) o Secuencias de Variantes de Amplicon (ASVs), y finalmente se comparan con bases de datos de referencia como UNITE o GenBank para la asignación taxonómica (Tedersoo et al., 2022).

La caracterización de las comunidades fúngicas del suelo es fundamental para comprender el funcionamiento de los ecosistemas terrestres, especialmente en regiones tropicales donde la biodiversidad alcanza sus niveles más altos a escala global (Landinez-Torres et al., 2019). Los hongos del suelo desempeñan roles

esenciales en los procesos ecosistémicos, incluyendo el ciclaje de nutrientes, la formación del suelo, la descomposición de materia orgánica y el establecimiento de simbiosis con plantas (Urbina et al., 2016). A pesar de su importancia ecológica, el conocimiento sobre la diversidad fúngica en suelos tropicales americanos sigue siendo fragmentario e incompleto, existiendo vacíos significativos en la caracterización de estas comunidades microbianas en diversas regiones neotropicales.

Recientemente, el desarrollo de técnicas de Next-Generation Sequencing (NGS) ha revolucionado el estudio de la diversidad fúngica edáfica, permitiendo la identificación de taxones no cultivables y proporcionando una visión más comprehensiva de la estructura comunitaria (Landinez-Torres et al., 2019). Entre estas técnicas, el *metabarcoding* de la región ITS (*Internal Transcribed Spacer*) del ADN ribosomal se ha establecido como el estándar para la identificación de hongos, ofreciendo una resolución taxonómica adecuada para estudios ecológicos y de diversidad (Aranguren et al., 2023). La aplicación de estos enfoques moleculares en suelos tropicales ha revelado una diversidad fúngica sustancialmente mayor a la previamente documentada mediante métodos tradicionales basados en cultivo.

2.1.2 Avances Metodológicos en el Estudio de Comunidades Fúngicas

La implementación de protocolos estandarizados de metabarcoding ha permitido comparaciones robustas entre estudios realizados en diferentes regiones tropicales. Urbina et al. (2016) desarrollaron un enfoque metodológico comprehensivo para caracterizar la diversidad de hongos del suelo en Puerto Rico, utilizando primers degenerados para minimizar el sesgo de amplificación y variando temperaturas de annealing para capturar una mayor diversidad fúngica. Su protocolo incluyó la

extracción de ADN *in situ*, amplificación del espaciador transcrito interno 2 (ITS2) y secuenciación mediante tecnología Ion Torrent, procesando un total de 566.613 secuencias después de filtros de calidad que se agruparon en 4.140 unidades taxonómicas operativas moleculares (MOTUs).

De manera similar, Aranguren et al. (2023) aplicaron técnicas de *metabarcoding* del ITS2 en suelos Andosoles del sureste de Antioquia, Colombia, reteniendo 353.312 secuencias de alta calidad y identificando 494 OTUs pertenecientes a 9 filos, 29 clases, 53 órdenes, 117 familias, 158 géneros y 174 especies.

En Ecuador, Portalanza et al. (2025) emplearon un enfoque metagenómico para caracterizar comunidades fúngicas asociadas a *Cyperus rotundus* en el ecosistema de manglar de Isla Santay, utilizando secuenciación Illumina MiSeq de la región ITS y procesando las secuencias mediante herramientas bioinformáticas como QIIME2 y DADA2. Su metodología incluyó análisis de redes de co-ocurrencia y medidas de diversidad alfa para comparar comunidades fúngicas entre sitios con diferente presión antrópica.

Por otro lado, Urbina et al. (2016) en Puerto Rico representó uno de los primeros esfuerzos comprehensivos para caracterizar la diversidad fúngica del suelo en una isla tropical del Caribe. Sus resultados demostraron que las comunidades fúngicas del suelo en Puerto Rico están estructuradas por el tipo de ecosistema, con Ascomycota seguido por Basidiomycota dominando la diversidad fúngica edáfica. Entre los Ascomycota, la clase recientemente descrita Archaeorhizomycetes estuvo presente en todas las localidades muestreadas, y los taxones dentro de esta clase se encontraban entre los MOTUs más comúnmente observados.

Una contribución significativa de este estudio fue la identificación de patrones de distribución diferenciales entre los filos fúngicos principales. Mientras que la comunidad de Basidiomycota estuvo dominada por descomponedores del suelo y hongos ectomicorrízicos con una distribución fuertemente afectada por la variación local, los Ascomycota mostraron una distribución más homogénea a través de los diferentes ecosistemas muestreados. Solo 26 MOTUs fueron detectados en las siete localidades estudiadas, todos ellos ascomicetos descomponedores comunes del suelo, lo que sugiere la existencia de un núcleo de especies generalistas junto con una gran proporción de especies especialistas con distribución restringida.

Asimismo, Los ecosistemas andinos de Colombia han sido escenario de múltiples investigaciones que han revelado la extraordinaria diversidad fúngica de estos suelos tropicales de montaña. Landinez-Torres et al. (2019) realizaron un análisis de metabarcoding en agroecosistemas del alto Andino en el departamento de Boyacá, identificando más de 150 especies pertenecientes a 5 filos, con Ascomycota como taxón dominante. Basidiomycota y Zygomycota también estuvieron bien representados, dominados por especies de los géneros *Sebacina* y *Mortierella* respectivamente, principalmente distribuidas en parcelas seminaturales (bosque y pastizal).

Un hallazgo notable de este estudio fue la identificación de *Geoglossum fallax* como especie dominante en la parcela de pastizal no mejorado, un taxón considerado bioindicador de hábitats con valor de conservación. Esta investigación destacó la influencia de las actividades agrícolas y el manejo del suelo en la composición de las comunidades microbianas, con implicaciones para el manejo sostenible de estos ecosistemas.

Complementariamente, Aranguren et al. (2023) evaluaron el impacto de diferentes usos del suelo en la diversidad fúngica de Andosoles del sureste de Antioquia, encontrando que los usos del suelo explican importantes variaciones en los índices de diversidad alfa (41% de variación en áreas de bosque natural vs. actividades agrícolas y 75% en áreas de bosque natural vs. actividades mineras). Estos resultados mostraron que la riqueza fúngica fue mayor en áreas no degradadas y disminuyó con el grado de impacto antrópico, con parámetros como la temperatura del suelo y el contenido de materia orgánica identificados como factores determinantes más importantes que otros factores edáficos.

La sensibilidad de las comunidades fúngicas a las perturbaciones antrópicas las convierte en valiosos bioindicadores de la salud del suelo. Aranguren et al. (2023) cuantificaron el tamaño del efecto del uso del suelo sobre la biodiversidad mediante el cálculo de ratios de respuesta logarítmica (LogRR), encontrando que los índices de diversidad alfa, particularmente el índice de Fisher (S'), mostraron los mayores tamaños de efecto entre los usos de suelo contrastantes. Las abundancias de los órdenes Leucosporidiales, Trechisporales y Malasseziales fueron significativamente afectadas por las actividades agrícolas y mineras, mientras que taxones relevantes como las especies prevalentes y el orden Hypocreales dominante demostraron ser buenos indicadores de los efectos del uso del suelo sobre la comunidad fúngica del suelo.

Por otra parte, Las interacciones entre plantas y hongos constituyen un eje fundamental en el funcionamiento de los ecosistemas terrestres. Portalanza et al. (2025) investigaron las dinámicas de las comunidades fúngicas asociadas a *Cyperus rotundus* y sus implicaciones para *Rhizophora mangle* en un ecosistema de manglar, revelando diferencias significativas en los ensamblajes microbianos entre áreas con

presión antrópica y sitios no impactados. En el área impactada, la rizósfera exhibió menor riqueza fúngica y estuvo dominada por Magnaporthaceae (9%) y Aureobasidium melanogenum (5%), ambos asociados con rasgos de tolerancia al estrés. Por el contrario, la rizósfera del sitio no impactado mostró mayor diversidad de especies, con *Cladosporium dominicanum* (62%) y *Talaromyces* (11%) como taxones endofíticos dominantes.

El análisis de redes de co-ocurrencia realizado en este estudio reveló correlaciones positivas entre taxones fúngicos, descubriendo interacciones ecológicas dentro de estas comunidades. El clúster principal de la red contenía 18 nodos y 45 aristas, con taxones como *Magnaporthaceae* sp y *Psathyrella luteopallida* coexistiendo predominantemente en raíces del área antropizada. Estos patrones sugieren partición de nicho e interacciones cooperativas, destacando los roles ecológicos de taxones clave dentro de estos ambientes.

A pesar de los avances significativos en la caracterización de la diversidad fúngica en suelos tropicales americanos, persisten importantes vacíos de conocimiento. Urbina et al. (2016) señalaron que la mayoría de los MOTUs identificados en Puerto Rico fueron asignados solamente a niveles taxonómicos superiores (familia o niveles más altos), atribuyendo esta limitación a la falta de secuencias de referencia ITS2 de hongos de regiones tropicales y al posible endemismo fúngico en la isla.

De manera similar, Landinez-Torres et al. (2019) identificaron numerosos taxones que permanecen sin identificar, principalmente en el orden Trechisporales, y en los filos Chytridiomycota y Glomeromycota, revelando que la biodiversidad fúngica del agroambiente del alto Andino colombiano permanece en gran medida sin develar. La limitada resolución taxonómica para ciertos grupos fúngicos, particularmente

Glomeromycota, usando primers universales que amplifican el código de barras ITS1, sugiere la necesidad de utilizar secuencias de otros genes para obtener una mayor resolución de estos taxones.

Aranguren et al. (2023) destacaron la escasez de conjuntos de datos de secuencias fúngicas, junto con sus metadatos de muestras de estas áreas, señalando la importancia de expandir los esfuerzos de muestreo y caracterización en suelos tropicales americanos. La integración de variables ambientales y propiedades del suelo en los análisis de diversidad fúngica emerge como una aproximación promisoria para comprender los factores que determinan la estructura y composición de estas comunidades.

El estudio realizado por Portalanza et al. (2025) en los manglares de Isla Santay, Ecuador, representa una contribución significativa al entendimiento de las dinámicas microbianas en ecosistemas costeros tropicales sujetos a presión antrópica. Esta investigación demuestra cómo una especie invasora como Cyperus rotundus puede albergar un conjunto de microbios que potencialmente remodelan el delicado equilibrio ecológico de los sistemas de manglar. Los resultados revelan la intricada interconexión entre plantas, microbios y hábitats, destacando cómo las acciones humanas pueden reverberar a través de esta red de vida.

En conclusión, la investigación de Portalanza et al. (2025) en Ecuador amplía significativamente nuestra comprensión de las complejas interacciones entre plantas invasoras, comunidades fúngicas y ecosistemas de manglar, proporcionando una base sólida para el desarrollo de estrategias de manejo sostenible que preserven la biodiversidad y funcionalidad de estos vitales ecosistemas costeros tropicales.

2.2 Marco Conceptual

La capacidad de descifrar el código genético, o ADN, subyace en casi todos los avances de la biología y la medicina modernas. La secuenciación del ADN es el proceso de determinar el orden preciso de los nucleótidos (adenina, guanina, citosina y timina) dentro de una molécula de ADN (Moti, 2022). Durante décadas, el estándar de oro fue la secuenciación de Sanger, un método desarrollado en 1977 que utiliza la terminación de la cadena didesoxi (Hu et al., 2021). Aunque fue la tecnología fundamental que permitió la finalización del Proyecto Genoma Humano, el método Sanger es inherentemente de bajo rendimiento, costoso y laborioso, limitando su aplicación a escala masiva (Pareek et al., 2011).

La necesidad imperiosa de obtener datos genómicos de manera más rápida y económica impulsó una revolución tecnológica, dando lugar a la Secuenciación de Siguiente Generación (NGS), que redefinió por completo el alcance de la investigación biológica (Pareek et al., 2011). El término NGS, también conocido como secuenciación de alto rendimiento (HTS), no se refiere a una sola tecnología, sino a un conjunto de enfoques innovadores que superaron las limitaciones de Sanger (Hu et al., 2021; Porter & Hajibabaei, 2022). El principio fundamental de la NGS es la secuenciación masivamente paralela, donde millones, e incluso miles de millones, de fragmentos de ADN se secuencian simultáneamente en un solo ciclo (Fasolo et al., 2024). Este cambio de paradigma redujo drásticamente el costo por base y el tiempo de secuenciación, permitiendo que el objetivo de secuenciar genomas individuales a bajo costo se volviera una realidad (Pareek et al., 2011).

Si bien surgieron varias plataformas NGS de segunda generación que compitieron en el mercado, como la pirosecuenciación 454 de Roche y la secuenciación SOLiD de Applied Biosystems, la tecnología que llegó a dominar abrumadoramente el mercado fue la desarrollada por Solexa, una compañía posteriormente adquirida por Illumina Illumina se consolidó como el líder indiscutible en el espacio de la secuenciación de segunda generación, capturando la mayor parte de la cuota de mercado global. El éxito de la plataforma se basa en su robusta tecnología central: la Secuenciación por Síntesis (SBS), la cual es reconocida por su altísima precisión (tasas de error muy bajas) y un rendimiento de datos masivo, capaz de generar terabases (miles de gigabases) de datos en una sola ejecución (Emiyu & Lelisa, 2022).

El flujo de trabajo de la secuenciación Illumina comienza con la preparación de la biblioteca genómica. El ADN genómico de alto peso molecular que se extrae de una muestra primero debe ser fragmentado en trozos más pequeños y manejables, generalmente de unos pocos cientos de pares de bases. A los extremos de estos fragmentos se les ligan adaptadores de ADN especializados (conocidos como P5 y P7), que son secuencias sintéticas cortas. Estos adaptadores son cruciales, ya que contienen los sitios de unión necesarios para la amplificación, el anclaje a la celda de flujo y el inicio de la reacción de secuenciación (Emiyu & Lelisa, 2022; Pareek et al., 2011).

Una vez preparada la biblioteca, el siguiente paso es la generación de clústeres, un proceso que ocurre dentro de una celda de flujo (flow cell) de vidrio. La superficie de esta celda está densamente cubierta por millones de oligonucleótidos (oligos) que son complementarios a las secuencias de los adaptadores P5 y P7 ligados a la biblioteca Los fragmentos de la biblioteca se hibridan aleatoriamente a estos oligos en la superficie. A continuación, se inicia un proceso isotérmico llamado amplificación en

puente (Bridge Amplification): el fragmento de ADN se dobla, formando un "puente" al hibridarse con un oligo adyacente, y una polimerasa crea la hebra complementaria (Emiyu & Lelisa, 2022; Pareek et al., 2011). Este ciclo se repite miles de veces, creando un "clúster" clonal que contiene millones de copias idénticas del fragmento original en una ubicación física discreta (Emiyu & Lelisa, 2022).

Con los clústeres generados y las hebras inversas eliminadas, comienza la Secuenciación por Síntesis (SBS) (Emiyu & Lelisa, 2022). La máquina introduce los reactivos de secuenciación: una polimerasa y una mezcla de cuatro nucleótidos (A, C, T, G). Cada uno de estos nucleótidos está modificado de dos maneras cruciales: posee un tinte fluorescente único (un color diferente para cada base) y un terminador reversible que bloquea químicamente el extremo 3', impidiendo la adición de más bases (Moti, 2022; Emiyu & Lelisa, 2022). En cada ciclo de secuenciación, la polimerasa añade solo un nucleótido terminador a cada hebra en crecimiento dentro de cada clúster.

El impacto de la tecnología Illumina ha sido profundo, abriendo campos de estudio que antes eran inviables. Es la tecnología dominante para la secuenciación del genoma completo (WGS), la secuenciación del exoma (WES) y el análisis de la expresión génica a gran escala a través de RNA-Seq (Pareek et al., 2011). En la genómica del cáncer, por ejemplo, ha sido fundamental para la identificación de mutaciones somáticas, variaciones en el número de copias (CNVs) y reordenamientos estructurales complejos en los genomas tumorales (Pareek et al., 2011). Su alta precisión y rendimiento la han convertido en la plataforma de elección para la investigación genómica y el avance de la medicina personalizada (Hu et al., 2021).

A pesar de su dominio, la principal limitación de Illumina y otras tecnologías de segunda generación es su longitud de lectura corta (generalmente 50-300 pares de bases) (Hu et al., 2021; Moti, 2022). Estas lecturas cortas hacen que el ensamblaje de genomas de novo (es decir, construir un genoma desde cero sin una referencia) sea extremadamente difícil. Esto es particularmente problemático en regiones complejas del genoma que contienen ADN repetitivo (Hu et al., 2021). Además, las lecturas cortas no logran resolver variantes estructurales grandes (SVs), como inversiones o translocaciones largas, ni determinar la fase de los haplotipos (es decir, qué variantes están juntas en el mismo cromosoma) (Hu et al., 2021; Pareek et al., 2011).

Esta limitación impulsó el desarrollo de la Secuenciación de Tercera Generación (TGS), también conocida como secuenciación de lecturas largas (Hu et al., 2021; Moti, 2022). La TGS se diferencia fundamentalmente de la NGS porque se centra en la secuenciación de moléculas únicas en tiempo real, eliminando la necesidad de la amplificación por PCR (como la amplificación en puente) que puede introducir sesgos (Hu et al., 2021). Las dos plataformas principales de TGS que producen lecturas largas son Pacific Biosciences (PacBio) y Oxford Nanopore Technologies (ONT) (Hu et al., 2021; Moti, 2022).

Las tecnologías TGS operan con principios distintos. PacBio utiliza la Secuenciación en Tiempo Real de Molécula Única (SMRT), que observa una polimerasa de ADN individual mientras sintetiza una hebra de ADN en tiempo real dentro de una nanoestructura llamada guía de onda de modo cero (ZMW) (Hu et al., 2021). ONT, por otro lado, utiliza nanoporos proteicos: una corriente eléctrica pasa a través del poro y, a medida que una hebra de ADN lo atraviesa, cada base interrumpe la corriente de manera característica, permitiendo su identificación (Hu et al., 2021; Moti,

2022). Ambas tecnologías son capaces de generar lecturas de miles, decenas de miles, o incluso millones de pares de bases (Hu et al., 2021).

2.2.2. La importancia del *metabarcoding* en el análisis ecológico

El metabarcoding (MB) ha revolucionado la ecología microbiana, consolidándose como la herramienta principal para el estudio de comunidades biológicas complejas, especialmente de hongos y otros eucariotas. Antes de su desarrollo, la ecología microbiana dependía en gran medida de métodos clásicos como el aislamiento en cultivo y la microscopía. Sin embargo, estos enfoques son conocidos por ser extremadamente laboriosos, inherentemente sesgados (ya que solo una pequeña fracción de los microorganismos es cultivable) y requieren un alto nivel de experiencia taxonómica especializada. El metabarcoding, al aplicar la secuenciación de alto rendimiento (HTS) a marcadores genéticos cortos (códigos de barras) obtenidos directamente de muestras ambientales (eDNA), evita estos sesgos de cultivo y permite a los investigadores acceder a la biodiversidad en microhábitats previamente inaccesibles y estudiar taxones crípticos (Tedersoo et al., 2022).

La importancia fundamental de esta técnica radica en su capacidad para proporcionar una instantánea completa y de alto rendimiento de la composición de la comunidad y la diversidad taxonómica. Mediante el análisis de eDNA de sustratos complejos como el suelo, el agua, el aire o los tejidos de organismos, el metabarcoding permite a los ecólogos abordar preguntas a una escala sin precedentes. Esta capacidad ha posicionado al MB como una metodología esencial no solo en la ecología, sino también en campos aplicados como la biogeografía, la biología de la conservación, la ciencia del suelo y la bioseguridad (Tedersoo et al., 2022). Permite, por ejemplo, el

monitoreo rápido de especies invasoras o patógenos, y la evaluación de cómo la diversidad microbiana responde a los cambios ambientales.

La aplicación del metabarcoding permite a los ecólogos investigar patrones biogeográficos a gran escala y las dinámicas temporales de las comunidades microbianas, revelando cómo la diversidad oculta responde a los impulsores globales. Además, aunque el MB se centra en la identificación taxonómica, es el primer paso crítico hacia la ecología. Al identificar con precisión los taxones en una comunidad, los investigadores pueden inferir sus roles ecológicos, como la descomposición, la simbiosis micorrízica o la patogenicidad, conectando así la estructura de la comunidad con la función del ecosistema (Fasolo et al., 2024).

2.2.3 ASVs (Amplicon sequences variants) vs OTUs (Operational taxonomics Units)

En el campo de la ecología molecular y los estudios de microbiomas, el metabarcoding genera millones de secuencias de ADN que deben ser procesadas para reducir su complejidad y separar la señal biológica del ruido técnico (Fasolo et al., 2024; Porter & Hajibabaei, 2022). Históricamente, el método estándar para esta tarea ha sido el agrupamiento (clustering) de secuencias en Unidades Taxonómicas Operacionales (OTUs). Este enfoque agrupa las secuencias basándose en un umbral de similitud predefinido, que comúnmente se ha fijado en el 97% (Fasolo et al., 2024). El objetivo de este agrupamiento es colapsar en una sola unidad de consenso tanto la variabilidad biológica menor (por ejemplo, variantes intragenómicas) como los errores producidos durante la amplificación por PCR y la secuenciación (Porter & Hajibabaei, 2022). Si bien este método puede mitigar el "ruido" y ser computacionalmente eficiente, su principal desventaja es que el umbral del 97% es

arbitrario y puede ocultar diversidad biológica real, agrupando especies crípticas que difieren por menos del 3% (Porter & Hajibabaei, 2022).

Más recientemente, ha surgido un nuevo paradigma centrado en las Variantes de Secuencia de Amplicón (ASVs), también conocidas como Secuencias de Variantes Exactas (ESVs) o ZOTUs (OTUs de radio cero) (Fasolo et al., 2024; Porter & Hajibabaei, 2022). A diferencia del agrupamiento por similitud, los métodos de ASV, implementados en algoritmos como DADA2 o UNOISE3, utilizan un enfoque de *denoising* (eliminación de ruido) para modelar y corregir los errores de secuenciación. Este proceso infiere las secuencias biológicas "verdaderas" presentes en la muestra, resolviendo la diversidad hasta el nivel de un solo nucleótido (Porter & Hajibabaei, 2022). Esta alta resolución es una ventaja clave, ya que permite detectar una diversidad críptica que las OTUs pasarían por alto (Fasolo et al., 2024).

La diferencia más significativa entre ambos métodos radica en su comparabilidad y reproducibilidad (Tabla 2). Las OTUs son inherentemente "específicas del estudio" (Fasolo et al., 2024). Dado que el resultado del agrupamiento depende del conjunto total de secuencias en una corrida de análisis específica, una OTU generada en un estudio no es directamente comparable con una OTU de otro estudio, lo que dificulta enormemente los metaanálisis (Porter & Hajibabaei, 2022). Por el contrario, las ASVs son "reutilizables" (Fasolo et al., 2024) y consistentes. Una secuencia ASV exacta es una entidad biológica inferida que puede ser comparada directamente a través de diferentes experimentos, investigadores y laboratorios, facilitando la síntesis de conocimiento (Fasolo et al., 2024; Porter & Hajibabaei, 2022).

La tendencia actual en el campo se inclina decididamente hacia el uso de ASVs, un cambio que ha sido descrito como un "avance significativo" (Fasolo et al., 2024). Los

estudios que comparan ambos métodos, como el de Fasolo et al. (2024), encuentran que los análisis basados en ASV reportan sistemáticamente una mayor diversidad alfa (riqueza de especies) que los análisis de OTU. Curiosamente, ese mismo estudio también observó que, a pesar de las diferencias en la riqueza detectada, los patrones generales de diversidad beta (la diferencia en la composición entre comunidades) no mostraron diferencias importantes entre los dos métodos (Fasolo et al., 2024). Sin embargo, debido a su resolución superior y, lo que es más importante, a su reproducibilidad y comparabilidad entre estudios, las ASVs se consideran ahora el estándar preferido para las evaluaciones de biodiversidad de alta fidelidad (Porter & Hajibabaei, 2022).

Tabla 4. Comparación entre ASVs y OTUs

Característica	ASVs (Variantes de Secuencia de Amplicón)	OTUs (Unidades Taxonómicas Operacionales)
Definición / Método	Utilizan denoising (eliminación de ruido) para inferir secuencias biológicas "verdaderas" (Porter & Hajibabaei, 2022).	Agrupan secuencias basándose en un umbral de similitud fijo (generalmente 97%) (Fasolo et al., 2024).
Resolución	Máxima resolución posible, a nivel de un solo nucleótido (Fasolo et al., 2024; Porter & Hajibabaei, 2022).	Resolución limitada por el umbral Agrupa variantes biológicas menores (Porter & Hajibabaei, 2022).
Ventajas	**1. Reutilizables**: Son independientes del estudio y directamente comparables, facilitando metaanálisis (Fasolo et al., 2024; Porter & Hajibabaei, 2022). 2. **Alta Resolución:** Detectan diversidad críptica y se considera que reflejan mejor la diversidad biológica real (Fasolo et al., 2024; Porter & Hajibabaei, 2022). 3. **Mayor Riqueza:** Reportan estimaciones más altas de diversidad alfa (riqueza) (Fasolo et al., 2024).	**1. Mitigación de Ruido:** El agrupamiento puede "amortiguar" el ruido de errores de secuenciación (Fasolo et al., 2024). **2. Computacional:** Pueden ser más rápidos computacionalmente al reducir la complejidad de los datos (Porter & Hajibabaei, 2022). 3. **Método Válido:** considerados un método válido para muchos análisis (Fasolo et al., 2024).
Desventajas	1**. Inflación Potencial:** La alta resolución podría, en teoría, confundir errores residuales con diversidad real si el denoising no es perfecto (Fasolo et al., 2024).	1. **No Comparables:** Son "específicas del estudio" y no pueden compararse directamente entre diferentes experimentos (Fasolo et al., 2024; Porter & Hajibabaei, 2022). **2. Umbral Arbitrario:** El 97% es un umbral arbitrario que no tiene una base biológica universal (Porter & Hajibabaei, 2022). **3. Pérdida de Resolución:** Ocultan diversidad biológica real y pueden agrupar erróneamente especies crípticas (Porter & Hajibabaei, 2022).
Tendencia Actual	Preferido. Considerado un "avance significativo" (Fasolo et al., 2024). Es el estándar preferido por su reproducibilidad y alta fidelidad (Porter & Hajibabaei, 2022).	En desuso (para alta resolución). Si bien sigue siendo válido, la tendencia es migrar hacia las ASVs (Fasolo et al., 2024).

2.3 Marco Teórico

La Secuenciación de Próxima Generación (*Next-Generation Sequencing*, NGS), también denominada secuenciación de alto rendimiento (*high-throughput sequencing*), constituye un pilar tecnológico fundamental en las ciencias ómicas contemporáneas. Esta tecnología ha redefinido las fronteras de la genómica, la transcriptómica y, de manera crucial, la metagenómica, al permitir la obtención de volúmenes masivos de datos de secuencias de ácidos nucleicos a un costo y tiempo significativamente reducidos en comparación con los métodos tradicionales (Taş et al., 2021). El advenimiento de la NGS ha sido una fuerza impulsora que ha facilitado la transición de la biología molecular, desde el análisis de genes individuales hacia el estudio de sistemas biológicos complejos y comunidades microbianas completas (Taş et al., 2021). El entendimiento cabal de la NGS es esencial para el progreso en campos tan diversos como la salud humana, la biotecnología ambiental y la genómica comparado vegetal.

La trascendencia de la Secuenciación de Próxima Generación (NGS) radica en su capacidad para ofrecer una perspectiva molecular no sesgada y con una resolución taxonómica elevada de la diversidad biológica y el potencial funcional de sistemas complejos. Históricamente, la microbiología se ha visto limitada por el paradigma cultivo-dependiente, donde se estima que la vasta mayoría de la diversidad microbiana es incultivable bajo condiciones de laboratorio. La NGS ha solventado esta restricción mediante el desarrollo de técnicas cultivo-independientes, permitiendo la secuenciación directa del ADN extraído de muestras ambientales o clínicas (Taş et al., 2021).

En el ámbito de la microbiología ecológica, la NGS, particularmente a través de la secuenciación metagenómica de escopeta (*shotgun metagenomic sequencing*), ha catalizado una "revolución" en la caracterización de comunidades microbianas (New & Brito, 2020). Este enfoque proporciona una cuantificación relativa de la abundancia y el potencial funcional del microbioma completo (el metagenoma). Este nivel de análisis es indispensable para comprender ecosistemas críticos, como los manglares, donde las comunidades microbianas ejercen funciones vitales en el ciclo de nutrientes (Allard et al., 2020; Muwawa et al., 2021). La NGS permite a los investigadores ir más allá de la taxonomía descriptiva para elucidar las vías metabólicas y los genes codificantes de enzimas clave en el contexto *in situ*.

Adicionalmente, la NGS facilita la resolución genómica a nivel de especie y cepa mediante el ensamblaje de Genomas Ensamblados Metagenómicamente (MAGs) (Bai et al., 2025; New & Brito, 2020), lo que ofrece una comprensión detallada de la diversidad genómica de los miembros de una comunidad. La capacidad de discernir entre especies y variantes de cepa es crucial para el seguimiento epidemiológico y la manipulación dirigida de comunidades (New & Brito, 2020). La información generada por la NGS sirve como el insumo primario para sofisticados análisis bioinformáticos gen-céntricos y genoma-céntricos (Taş et al., 2021), que permiten la inferencia de interacciones ecológicas, la filogenia evolutiva y la identificación de nuevos *taxa* y funciones bioquímicas. Por lo tanto, la NGS no solo es una herramienta de detección, sino un habilitador epistemológico para la investigación biológica moderna (Fadiji & Babalola, 2020; Taş et al., 2021).

2.3.1 Evolución Histórica y Técnica de la Secuenciación Genómica

La evolución de las técnicas de secuenciación de ácidos nucleicos se describe mediante una progresión generacional marcada por incrementos exponenciales en el rendimiento y reducciones logarítmicas en el costo por base secuenciada. La primera generación fue establecida por el método de Sanger (Sanger et al., 1977), un método de alto costo y bajo rendimiento que fue inadecuado para el análisis de genomas completos o metagenomas de alta complejidad debido a sus limitaciones en el procesamiento masivo y la capacidad de análisis simultáneo.

El punto de inflexión fue la llegada de la Secuenciación de Próxima Generación (NGS), catalogada como la segunda generación de secuenciación. Esta transición implicó un cambio fundamental hacia la secuenciación masiva en paralelo (*massively parallel sequencing*). La tecnología NGS se caracteriza por la miniaturización de las reacciones y la capacidad de procesar millones de lecturas de secuencias simultáneamente, generando lecturas cortas (típicamente entre 50 y 300 pares de bases). Este avance transformó el *throughput* genómico, abriendo la puerta a los estudios metagenómicos de escopeta (Fadiji & Babalola, 2020; Taş et al., 2021). Dentro de esta generación, se optimizó la secuenciación de amplicones (p. ej., el gen 16S rRNA en microbiología), evolucionando hacia la identificación de Variantes de Secuencia de Amplicones (ASVs). Las ASVs representan una mejora significativa en la resolución taxonómica (New & Brito, 2020).

La evolución continuó con la aparición de la tercera generación de secuenciación (TGS), introduciendo las plataformas de lectura larga (*long-read*). Estas tecnologías emergieron para subsanar los desafíos de ensamblaje inherentes a las lecturas cortas de NGS, especialmente en regiones genómicas repetitivas o en la reconstrucción de

MAGs complejos (Taş et al., 2021). Si bien las lecturas largas ofrecen una continuidad de ensamblaje superior, históricamente exhibían una tasa de error más alta y un rendimiento por corrida inferior en comparación con la NGS tradicional. Por lo tanto, el flujo de trabajo moderno ha evolucionado hacia la implementación de estrategias de secuenciación híbrida, donde la precisión y el alto rendimiento de las lecturas cortas de segunda generación se combinan con la longitud de la tercera generación para la corrección de errores y la obtención de ensamblajes de calidad superior (Taş et al., 2021). Esta evolución constante subraya la naturaleza dinámica de la tecnología genómica, con una trayectoria que apunta a la integración de múltiples *ómicas* y el desarrollo de herramientas bioinformáticas flexibles y amigables con el usuario para el manejo de la creciente complejidad y volumen de los datos (Bai et al., 2025; Wen et al., 2023).

CAPÍTULO III: Diseño Metodológico

3.1 Tipo y diseño de investigación

3.1.2 Tipo de Investigación

El presente trabajo es de tipo aplicada ya que el objetivo general es desarrollar un *pipeline* que permita realizar análisis bioinformáticos a archivos .fastaq provenientes de la plataforma de secuenciación Ilumina para realizar estudios metagenómicos robustos a microbiomas. Asimismo, es de tipo descriptiva porque permite analizar la diversidad de especies de hongos empleando métricas estadísticas como índices de Shannon y Simpson, así como también, análisis estadísticos multivariados como Análisis de Componentes principales (PCoA) y escalamiento multidimensional no métrico (NMDS).

3.1.3 Diseño de investigación

El presente diseño es cuasiexperimental dado que no se manipulan variables directamente en los organismos, si no que se fundamenta en experimentación computacional, ya que evalúa distintas scripts para establecer la manera más viable de realizar análisis metagenómicos ecológicos robustos de secuencias ITS.

3.1.4 La población y la muestra

La población de estudio corresponde a las secuencias de ADN de la región ITS de hongos obtenidas de base de datos de acceso libre (Unite). Las muestras corresponden a los archivos .fastaq de Ilumina provenientes de muestras de suelo de Manglar.

3.1.5 Los métodos y las técnicas

En la presente metodología se expondrá los pasos a seguir durante todo el *pipeline* para el análisis metagenómica de secuencias ITS de Illumina con énfasis en la descripción del microbioma. El siguiente desarrollo incluye la carga de datos crudos

al entorno de Rstudio, su control de calidad, detección y remoción de bases ambiguas, remoción de primers, descarga de base de datos de referencia para hongos y los análisis ecológicos respectivos. Se describe paso a paso la construcción de este con el objetivo que sea personalizable a las necesidades técnicas de los investigadores. Una vista general del *pipeline* se encuentra en la figura 1.

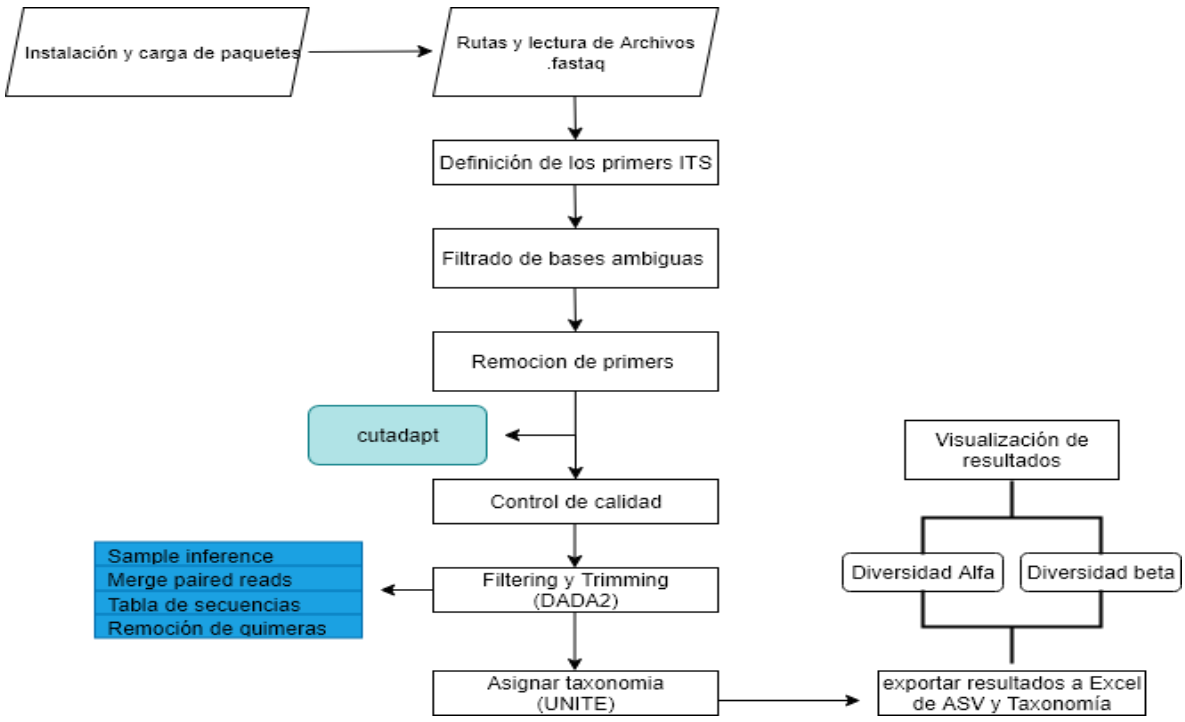

Figura 1. *Workflow* del análisis metagenómico de secuencias ITS.

Considerando que se trata de análisis metagenómico, los archivos .fastaq representa las muestras *demultiplexed* en donde cada uno corresponde a las lecturas del *forward* y el *reverse* con un peso total de 1.3 GB. Para la ejecución del *pipeline* se está empleando una computadora con un procesador Intel core I5 de 11va generación, 16 GB de memoria RAM y 500 GB de disco duro tipo SSD.

3.2.1 Carga de los datos al entorno de R e instalación y llamada de paquetes

Para cargar los datos al entorno de R bajo la interfaz gráfica (Rstudio) es necesario especificar la ubicación de la carpeta de trabajo en donde están ubicados los archivos.

Los archivos tienen el siguiente patrón de nombre “L001_R1_001” para el *Forward* y “L001_R2_001” para el *Reverse*. Para adjuntar la ruta al entorno de Rstudio se emplea la siguiente línea de código:

```
## Cargar los archivos fastq
path <- "C:/Users/Ricardo98/Desktop/Maestria/Tesis_Maestría/Tesis_Maestría/Datos_META/META_4-6"
list.files(path)
```

Listing 1.1: Carga de archivos .fastq al entorno Rstudio

Una vez cargados los archivos .fastaq al entorno, se procedió a la instalación y la llamada a toda la paquetería necesaria para el análisis del microbioma. Dado que algunos de los paquetes que se emplearon provienen del repositorio Bioconductor (Huber et al., 2015), este se instaló en primer lugar. Para ello se emplea la siguiente línea de comando:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
# Instalar paquetes necesarios
BiocManager::install("dada2")
BiocManager::install("ShortRead")
```

Listing 1.2: Instalación del repositorio Bioconductor, paquetes DADA2 y ShortRead

De igual manera instalaremos dos paquetes provenientes del repositorio de Bioconductor que son DADA2 (Callahan et al., 2016) para el análisis de secuencias y *Shortread* (Morgan et al., 2009) para la manipulación de estas; la instalación de estos paquetes se encuentra en el Listing 1.2. Una vez instalados los paquetes se hizo el llamado de cada uno de ellos usando la función library de la siguiente manera:

```
# Cargar paquetes
library(dada2)
library(ShortRead)
```

Listing 1.3: Llamada a paquetes.

3.2.2 Identificación de primers

A continuación, se comienza con la identificación de los primers empleados para la amplificación de la región ITS. Los primers pueden variar de acuerdo con cada región amplificada. En el caso del presente estudio los primer fueron los siguientes: ITS81 “GTGAATCATCGAATCTTTGAA” y el ITS4 “TCCTCCGCTTATTGATATGC” para el *forward* y el *reverse* respectivamente. El *reverse* corresponde a la orientación reverso complemento. Éstas secuencias de nucleótidos serán asignadas a un vector de acuerdo con la siguiente línea de comando:

```
31 # --------------------------
32 FWD <- "GTGAATCATCGAATCTTTGAA"   # ITS 81
33 REV <- "TCCTCCGCTTATTGATATGC"     # ITS4
```

Listing 1.4: Secuencia de primers empleadas en el presente estudio.

3.2.3 Uso del paquete DADA2 y ShortRead para el filtrado y trimming de las secuencias.

El uso del paquete DADA2 tiene una serie de pasos que son característicos para la región ITS. El primer paso es revisar la orientación y la ubicación de los primers empleados y asignados a un vector anteriormente. Para esto se empleó la siguiente línea de comando:

```
35 # Función para generar orientaciones de primers
36 allOrients <- function(primer) {
37   dna <- DNAString(primer)
38   orients <- c(Forward = dna,
39                Complement = complement(dna),
40                Reverse = reverse(dna),
41                RevComp = reverseComplement(dna))
42   return(sapply(orients, toString))
43 }
44
```

Listing 1.5: Ubicación y orientación de los primers.

Una vez realizado este procedimiento se procedió con el primer filtrado, que corresponde a la identificación y remoción de bases ambiguas “Ns”, mismo que se realizó de la siguiente forma:

```
48  # Filtrar bases ambiguas
49
50  fnFs.filtN <- file.path(path, "filtN", basename(fnFs))
51  fnRs.filtN <- file.path(path, "filtN", basename(fnRs))
52  filterAndTrim(fnFs, fnFs.filtN, fnRs, fnRs.filtN, maxN = 0, multithread = FALSE)
```

Listing 1.6: Filtrado de bases ambiguas

Posteriormente se continuo con la remoción de los primers, para ello, el paquete DADA2 recomienda la instalación de una herramienta de línea de comando llamada cutadapt. Esta herramienta no es un paquete nativo de R, por lo que se lo instaló de diferente manera, la descripción de la instalación de cutadapt se describe en el Anexo 1.

Una vez instalado la herramienta, la remoción de los primers se realizó de acuerdo con lo siguiente:

```
54  # Remoción de primers con cutadapt
55  cutadapt <- "cutadapt" # Asegurar que está en PATH
56
57  FWD.RC <- dada2:::rc(FWD)
58  REV.RC <- dada2:::rc(REV)
59  R1.flags <- paste("-g", FWD, "-a", REV.RC)
60  R2.flags <- paste("-G", REV, "-A", FWD.RC)
61
62  fnFs.cut <- file.path(path.cut, basename(fnFs))
63  fnRs.cut <- file.path(path.cut, basename(fnRs))
64
65  for(i in seq_along(fnFs)) {
66    system2(cutadapt, args = c(R1.flags, R2.flags, "-n", 2,
67                               "-o", fnFs.cut[i], "-p", fnRs.cut[i],
68                               fnFs.filtN[i], fnRs.filtN[i]))
69  }
```

Listing 1.7: Remoción de primers de las secuencias ITS.

Finalizado este proceso se procede a realizar un control de calidad de las secuencias obtenidas. Para ello se empleó la función *plotQualityProfile* para cada secuencia (*Forward* y *Reverse*). Esta función genera un gráfico que permite inspeccionar la calidad de las secuencias para realizar el posterior proceso de trimming. En general, el criterio para este proceso corresponde a eliminar las secuencias que se encuentran con un QS (Quality Score) por debajo de 30 – 20. Esto dependerá de las secuencias y del criterio a emplearse por el investigador.

Una vez seleccionado el lugar de corte, que está de acuerdo con el gráfico de QS generado por la función *plotQualityProfile*, se realizó el proceso de *trimming* de la siguiente forma:

```
## Filtrado y cortado

out <- filterAndTrim(cutFs, filtFs, cutRs, filtRs, maxN = 0, maxEE = c(3, 5), truncQ = 10,
                     truncLen = c(240,160),
                     minLen = 50,
                     rm.phix = TRUE,
                     compress = TRUE,
                     multithread = FALSE)
```

Listing 1.8: Proceso de trimming de las secuencias.

Cada una de las secuencias se cortó siguiendo los siguientes criterios de acuerdo con los gráficos de QS. Ademas, la opción multithread se mantuvo en FALSE durante todos los análisis dado que se empleó Windows como Sistema Operativo. Los demás criterios de corte se describen a continuación:

TruncQ = Trunca reads donde la calidad cae por debajo de Q10.

TruncLen = Ajuste importante para reads pareados, manteniendo la superposición necesaria para fusionar reads.

minLen = Descarta reads demasiado cortos.

rm.phix = Elimina contaminación con fago PhiX, común en secuenciación Illumina.

Una vez realizado este proceso, se procede con la parte final del proceso de filtrado y *trimming* previo a la asignación taxonómica de cada una de las secuencias obtenidas. Estos pasos corresponden a la asignación de variantes de secuencias, aprendizaje de tasas de error, la fusión de las secuencias pareadas, la construcción de tabla de secuencias y la remoción de quimeras. Para este proceso se siguió con el código recomendado por DADA2 para secuencias ITS (https://benjjneb.github.io/dada2/ITS_workflow.html) Este procedimiento se realizó con las siguientes líneas de comando:

```
# 5. INFERENCIA DE VARIANTES DE SECUENCIA
# --------------------------
# Aprender tasas de error
errF <- learnErrors(filtFs_10, multithread = FALSE)
errR <- learnErrors(filtRs_10, multithread = FALSE)

# Inferir variantes de secuencia
dadaFs <- dada(filtFs_10, err = errF, multithread = FALSE)
dadaRs <- dada(filtRs_10, err = errR, multithread = FALSE)

# Fusionar reads pareados
mergers <- mergePairs(dadaFs, filtFs_10, dadaRs, filtRs_10, verbose = TRUE)

# --------------------------
# 6. CONSTRUIR TABLA DE SECUENCIAS Y ELIMINAR QUIMERAS
# --------------------------
# Construir tabla de secuencias
seqtab <- makeSequenceTable(mergers)

# Eliminar quimeras
seqtab.nochim <- removeBimeraDenovo(seqtab, method = "consensus", multithread = FALSE)

# Verificar porcentaje de reads no quiméricos
cat("Porcentaje de reads no quiméricos:", sum(seqtab.nochim) / sum(seqtab) * 100, "%\n")
```

Listing 1.9: Inferencia de variantes de secuencia (ASV).

Una vez finalizado este proceso, procedemos con la asignación taxonómica de las secuencias. Para realizar esto se empleó la base de datos UNITE (Abarenkov et al., 2025) que contiene secuencias de la región ITS de Hongos. La base de datos es de acceso público y puede ser descargada a través de este enlace: https://unite.ut.ee/repository.php. Esta asignación taxonómica se realizó con la siguiente línea de comando:

```
# 7. ASIGNACIÓN TAXONÓMICA
# -------------------------
# Asignar taxonomía (ajustar ruta de la base de datos)
taxa <- assignTaxonomy(seqtab.nochim,
                       "C:/Users/Ricardo98/Desktop/Maestria/Tesis_Maestría/Tesis_Maestría/Base de datos",
                       multithread = FALSE)
```

Listing 1.10: Asignación taxonómica empleando la base de datos UNITE.

Es importante mencionar que, durante todo el *pipeline* esta es la acción que más capacidad de computo requiere, por lo que la duración del proceso dependerá del tamaño de datos a analizar y el hardware del equipo empleado.

3.2.4 Procesamiento estadístico de la información

Una vez culminada la asignación taxonómica, se realizaron análisis de diversidad alfa y beta a partir de la matriz de ASV obtenida del paquete de bioconductor DADA2. En primer lugar, estos datos fueron exportados a una matriz de Excel de la siguiente forma:

```
# 9. EXPORTAR RESULTADOS
# ----------------------------
# Instalar y cargar paquetes para exportación
if (!require("writexl")) install.packages("writexl")
if (!require("dplyr")) install.packages("dplyr")
library(writexl)
library(dplyr)

# Preparar datos para exportación
asv_table <- as.data.frame(t(seqtab.nochim))
taxa_table <- as.data.frame(taxa)
combined_data <- cbind(ASV_ID = rownames(asv_table), asv_table, taxa_table)

# Ordenar datos por taxonomía
ordered_data <- combined_data %>%
  arrange(Order, Family, Species)

# Exportar a Excel
write_xlsx(ordered_data, "resultados_taxonomia_ordenados_sample_4.xlsx")
cat("Datos exportados exitosamente a 'resultados_taxonomia_ordenados.xlsx'\n")

```

Listing 1.11: Exportación de matriz de ASV a una hoja de cálculo Excel.

3.2.5 Análisis de diversidad alfa y beta

El análisis de diversidad se realizó empleando índices sinecológicos como Shannon y Simpson de la siguiente forma:

Índice de Shannon, H´ = $\sum(pi \ln pi)$

Índice de Simpson, D´= $\frac{1}{\sum(p_i)^2}$

En donde p es la proporción de individuos de especies i que contribuyen al número total de individuos, de acuerdo con Magurran (1988). Estos índices fueron calculados empleando el paquete *phyloseq* (McMurdie & Holmes, 2013). Los resultados obtenidos fueron graficados usando el paquete *ggplot2* (Wickham , 2016) del ecosistema *tidyverse* (Wickham et al., 2019) de la siguiente foma: En primer lugar, para poder emplear los paquetes de diversidad, se transformó la matriz de asignación taxonómica realizada en el Listing 1.10 a un archivo phyloseq. Esto permitió una manipulación eficiente de los datos obtenidos. Posteriormente se calculan los índices sinecológicos y se realiza un gráfico para una mejor visualización de acuerdo con el Listing 1.12.

```
# 3. ANÁLISIS DE DIVERSIDAD ALFA
# --------------------------
# Calcular índices de diversidad alfa
alpha_diversity <- estimate_richness(physeq, measures = c("Observed", "Shannon", "Simpson"))

# Mostrar resultados
print("Índices de diversidad alfa:")
print(alpha_diversity)

# Graficar diversidad alfa
alpha_plot <- plot_richness(physeq, measures = c("Observed", "Shannon", "Simpson")) +
  geom_point(size = 3, alpha = 0.7) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  ggtitle("Diversidad Alfa")
```

Listing 1.12: Análisis de diversidad alfa de los datos obtenidos.

Después se realizó un análisis de abundancia por taxa Phyla y género, esto permite una exploración visual de la composición relativa de cada taxón de acuerdo con la taxa seleccionada de cada muestra analizada. Con el objetivo de prevenir erros se usó la función tryCatch, posteriormente se usó la siguiente línea de comando para graficar los valores:

```
tryCatch({
  # 11.1 Composición a nivel de Filo
  physeq_phylum <- tax_glom(physeq, taxrank = "Phylum")
  physeq_phylum_rel <- transform_sample_counts(physeq_phylum, function(x) x / sum(x) * 100)

  # Crear paleta de colores
  n_taxa <- nrow(tax_table(physeq_phylum_rel))
  phylum_colors <- colorRampPalette(brewer.pal(12, "Paired"))(n_taxa)

  # Graficar barras de abundancia relativa a nivel de Filo
  abundance_plot_phylum <- plot_bar(physeq_phylum_rel, fill = "Phylum") +
    geom_bar(aes(fill = Phylum), stat = "identity", position = "stack") +
    scale_fill_manual(values = phylum_colors) +
    labs(y = "Abundancia Relativa (%)", title = "Composición a Nivel de Filo") +
    theme_bw() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

  print(abundance_plot_phylum)

  # 11.2 Composición a nivel de Género (Top 10)
  physeq_genus <- tax_glom(physeq, taxrank = "Genus")
  physeq_genus_rel <- transform_sample_counts(physeq_genus, function(x) x / sum(x) * 100)

  # Seleccionar top géneros
  top_genera <- names(sort(taxa_sums(physeq_genus), decreasing = TRUE))[1:min(10, ntaxa(physeq_genus))]
  physeq_genus_top <- prune_taxa(top_genera, physeq_genus_rel)

  genus_colors <- colorRampPalette(brewer.pal(12, "Set3"))(length(top_genera))

  abundance_plot_genus <- plot_bar(physeq_genus_top, fill = "Genus") +
    geom_bar(aes(fill = Genus), stat = "identity", position = "stack") +
    scale_fill_manual(values = genus_colors) +
    labs(y = "Abundancia Relativa (%)", title = "Top 10 Géneros") +
    theme_bw() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

  print(abundance_plot_genus)

}, error = function(e) {
  cat("Error en barplots:", e$message, "\n")
})
```

Listing 1.13. Barplot del análisis de abundancia relativa (%) por taxa.

Asimismo, como parte del análisis de diversidad alfa, se realizó una curva de rarefacsión para establecer el número de taxas identificadas por cada lectura obtenida en la secuenciación para ello se usó el siguiente código:

```r
# Calcular el número de reads por muestra
reads_por_muestra <- rowSums(seqtab.nochim)
cat("Reads por muestra:\n")
print(reads_por_muestra)

# Función para calcular curvas de rarefacción
calcular_rarefaccion <- function(abundancia_muestra, nombre_muestra, pasos = 50) {
  total_reads <- sum(abundancia_muestra)
  # Crear secuencia de tamaños de muestreo
  tamanos_muestreo <- round(seq(1, total_reads, length.out = pasos))

  resultados <- data.frame()

  for (tamano in tamanos_muestreo) {
    # Rarefacción: muestrear sin reemplazo
    if (tamano <= total_reads) {
      # Repetir varias veces para promediar
      riquezas <- replicate(10, {
        muestra_rarefacta <- rrarefy(abundancia_muestra, tamano)
        sum(muestra_rarefacta > 0)  # Número de ASVs
      })
      riqueza_promedio <- mean(riquezas)

      resultados <- rbind(resultados,
                          data.frame(Muestra = nombre_muestra,
                                     Reads = tamano,
                                     ASVs = riqueza_promedio))
    }
  }
  return(resultados)
}

# Calcular curvas para cada muestra
curvas_rarefaccion <- data.frame()

for (i in 1:nrow(seqtab.nochim)) {
  nombre_muestra <- rownames(seqtab.nochim)[i]
  abundancia_muestra <- as.numeric(seqtab.nochim[i, ])

  curva_muestra <- calcular_rarefaccion(abundancia_muestra, nombre_muestra)
  curvas_rarefaccion <- rbind(curvas_rarefaccion, curva_muestra)
}

```

Listing 1.14. Obtención de la curva de rarefacción para cada muestra.

Una vez se generó la matriz de ASV (Amplicon sequences variants) y se realizó el análisis de diversidad alfa, se continuó con el análisis de diversidad Beta. Para esto se empleó el paquete Vegan (Oksaken et al., 2025).

Para el análisis de diversidad de diversidad beta se usó el índice de Bray-curtis. Para el cálculo de la matriz de distancia de Bray-curtis se la calculo a partir de la matriz de ASV de la siguiente forma:

```r
# Calcular distancia de Bray-Curtis
dist_bray <- vegdist(seqtab.nochim, method = "bray")

# Realizar NMDS
nmds_results <- metaMDS(dist_bray, k = 2, trymax = 100)

# Verificar la estructura del objeto NMDS
print(nmds_results)
cat("Estructura del objeto NMDS:\n")
str(nmds_results)

# Extraer scores
nmds_scores <- as.data.frame(nmds_results$points)
colnames(nmds_scores) <- c("NMDS1", "NMDS2")

# Agregar nombres de muestra
nmds_scores$SampleID <- rownames(seqtab.nochim)
```

Listing 1.13. Analisis del índice de Bray-curtis y construcción de los scores.

Una vez calculada la matriz, aplico el cálculo del score y se procedió a realizar el análisis NMDS (*Non-metric multidimension scaling*) y el PCoA (*Principal Coordinates analysis).* Para este último se lo realizo empleando el paquete ape (Paradise & Schliep, 2019) con el siguiente código:

```r
# Calcular PCoA
pcoa_results <- pcoa(dist_bray)

# Extraer scores de PCoA
pcoa_scores <- as.data.frame(pcoa_results$vectors[, 1:2])
colnames(pcoa_scores) <- c("PCoA1", "PCoA2")
pcoa_scores$SampleID <- rownames(seqtab.nochim)
```

Listing 1.14. Cálculo del PCoA usando ape.

Una vez calculado los valores necesarios para cada análisis ambos datos fueron ploteados empleando el paquete ggplot. De igual forma se realizó un heatmap para observar la relación de cada una de las especies identificadas.

CAPÍTULO IV: Análisis e Interpretación de Resultados

4.1 Análisis e Interpretación de Resultados

A partir del pipeline implementado usando el paquete DADA2 específico para la región ITS de hongos, se pudo realizar un análisis de diversidad y una correcta asignación taxonómica de cada uno de los archivos fastq. Uno de los primeros pasos es el control de calidad de cada uno de los archivos, el paquete DADA2 establece una funcionalidad para realizar ese análisis, podemos observar en la figura 2 y 3 como se obtiene un mejor resultado en el corte de las secuencias para poder mejorar la calidad antes de realizar el proceso de *merged*.

Figura 2. Gráfico de *Quality Score* previo al proceso de Filtrado para el *Forward* de la muestra 2.

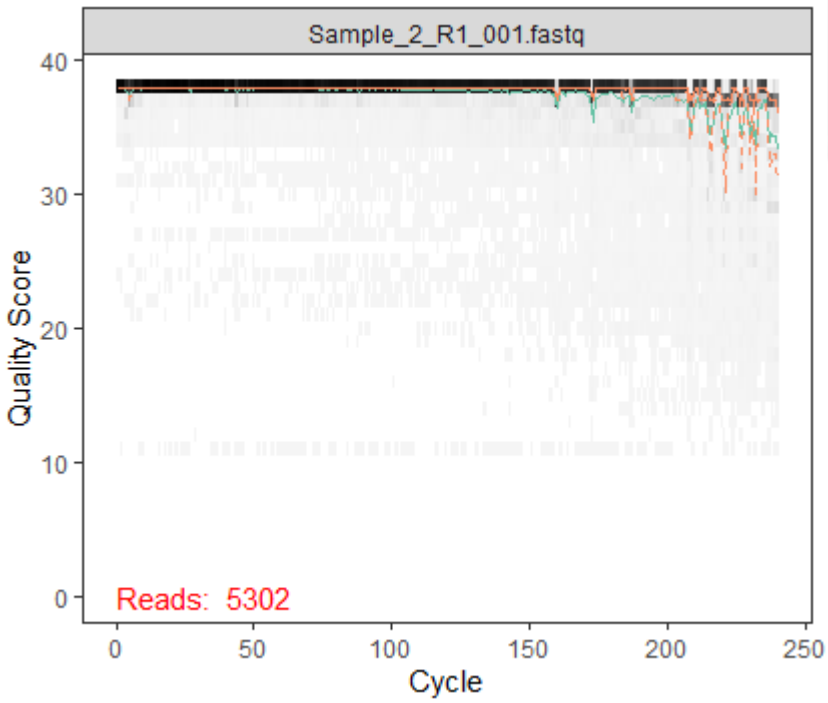

Figura 3 Gráfico de *Quality Score* después del proceso de Filtrado para el *Forward* de la muestra 2.

Después de realizar el proceso de control de calidad para cada uno de los archivos fastq, y usando la base de datos UNITE para la asignación taxonómica de las secuencias se obtuvo una matriz en Excel. Esta matriz de Excel puede usarse para realizar análisis de diversidad empleando otro software (Figura 4).

ASV_ID	ASV_sequence	Sample_1	Sample_2	Sample_3	Kingdom	Phylum	Class	Order	Family	Genus	Species
g__Calophoma_1	CGCACATTGCGC	26810	2780	12861	k__Fungi	p__Ascom	c__Dothid	o__Pleosp	f__Didym	g__Calophoma	
g__Antrodia s__neotropica_1	CGCACCTTGCGC	13396	1311	6320	k__Fungi	p__Basidi	c__Agaric	o__Polypc	f__Fomitc	g__Antroc	s__neotropica
g__Talaromyces	CGCACATTGCGC	4318	443	2321	k__Fungi	p__Ascom	c__Eurotic	o__Eurotii	f__Trichoc	g__Talaromyces	
g__Ceratobasidium_1	CGCACCTTGCGC	1136	121	476	k__Fungi	p__Basidi	c__Agaric	o__Canthi	f__Cerato	g__Ceratobasidium	
g__Trichomerium	CGCACATTGCGC	666	72	366	k__Fungi	p__Ascom	c__Eurotic	o__Chaet	f__Trichor	g__Trichomerium	
g__Calophoma_2	CGCACATTGCGC	672	70	342	k__Fungi	p__Ascom	c__Dothid	o__Pleosp	f__Didym	g__Calophoma	
g__Geastraceae_gen_Incertae_sedis	CGCATCTTGCGC	572	51	235	k__Fungi	p__Basidi	c__Agaric	o__Geastr	f__Geastri	g__Geastraceae_gen_Incertae_sedis	
f__Aspergillaceae	CGCACATTGCGC	269	48	165	k__Fungi	p__Ascom	c__Eurotic	o__Eurotii	f__Aspergillaceae		
g__Ceratobasidium_2	CGCACCTTGCGC	160	15	88	k__Fungi	p__Basidi	c__Agaric	o__Canthi	f__Cerato	g__Ceratobasidium	
g__Malassezia s__restricta	CGCACCTTGCGC	169	12	55	k__Fungi	p__Basidi	c__Malass	o__Malass	f__Malass	g__Malass	s__restricta
g__Thielaviopsis s__ethacetica	CGCACATTGCAC	128	7	44	k__Fungi	p__Ascom	c__Sordar	o__Microa	f__Cerato	g__Thiela	s__ethacetica
g__Calophoma_3	CGCACATTGTGC	147	0	0	k__Fungi	p__Ascom	c__Dothid	o__Pleosp	f__Didym	g__Calophoma	
Unassigned_1	GTGAATCATCGA	63	11	29	k__Fungi						
Unassigned_2	CGCAAGTTGCGC	53	0	41	k__Fungi						
g__Saccharomyces s__cerevisiae	CGCACATTGCGC	51	0	30	k__Fungi	p__Ascom	c__Saccha	o__Saccha	f__Saccha	g__Saccha	s__cerevisiae
g__Fusarium s__brevicaudatum	CGCACATTGCGC	63	0	12	k__Fungi	p__Ascom	c__Sordar	o__Hypoc	f__Nectria	g__Fusari	s__brevicaudatum
g__Diplodia	CGCACATTGCGC	42	6	17	k__Fungi	p__Ascom	c__Dothid	o__Botryc	f__Botryo	g__Diplodia	
g__Ilyonectria s__vredehoekensis	CGCACATTGCGC	46	3	16	k__Fungi	p__Ascom	c__Sordar	o__Hypoc	f__Nectria	g__Ilyone	s__vredehoekensis

Figura 4. Matriz en Excel obtenida utilizando el paquete DADA2 para la asignación taxonómica.

Los análisis de índices sinecológicos también se pudieron calcular y fueron graficados para un análisis más visual. En general podemos observas que la muestra 1 es la más diversa de las demás (Figura 5). Teniendo un valor de para Shannon de 1.38 y un total de 80 taxones identificados.

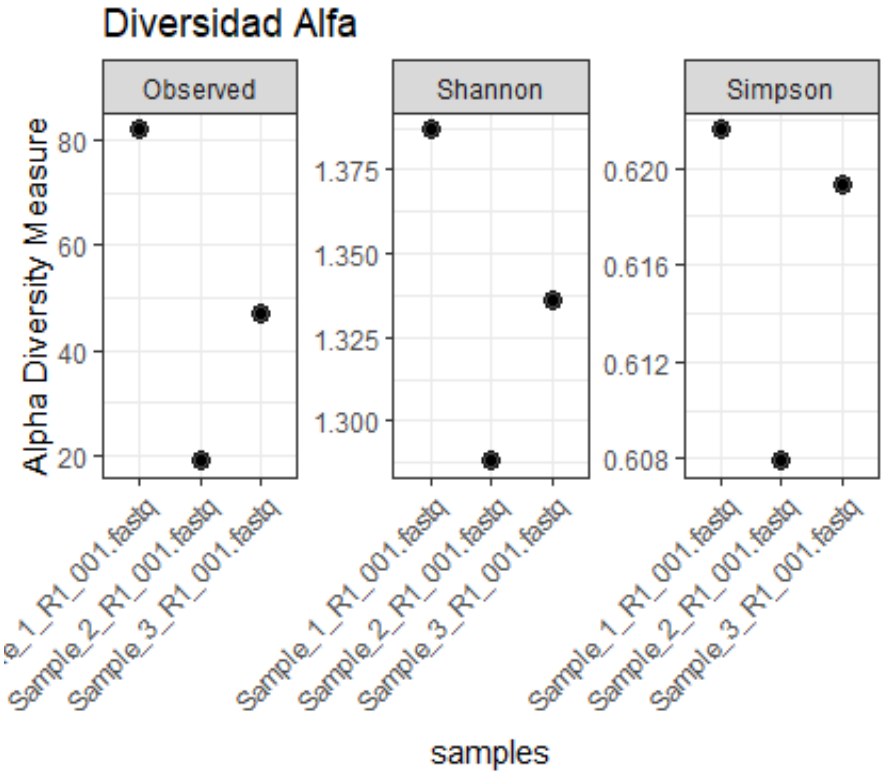

Figura 5. Análisis de diversidad Alfa para cada una de las muestras.

El análisis de curva de rarefacción mostro que cada una de las muestras alcanzaron una asíntota, indicando que el esfuerzo de secuenciación fue el óptimo para obtener el mayor número de ASV (Figura 7 y 8). Por otro lado, en el análisis de abundancia relativa observamos que los fila Ascomycota y Basidiomycota son claramente los más abundantes (Figura 6) De igual forma en el análisis de diversidad beta observamos una clara distinción en cada una de las muestras analizadas, esto se observa en el análisis de NMDS y PCoA (Figura 9 y 10). El heatmap de las abundancias relativas mostro una clara abundancia de tres especies en las tres muestras. (Figura 11).

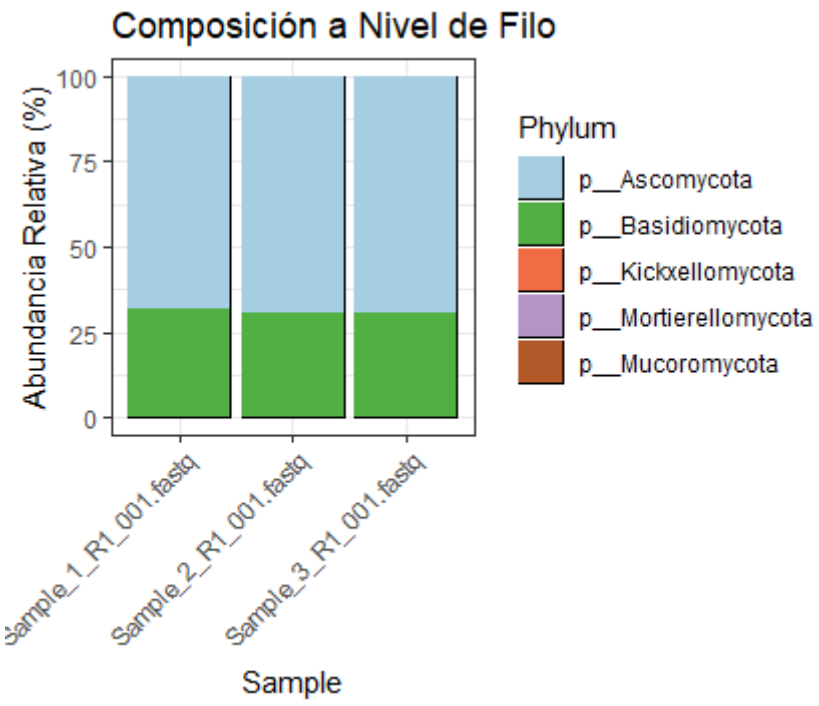

Figura 6. Abundancia relativa de cada uno de los *phyla* más abundantes.

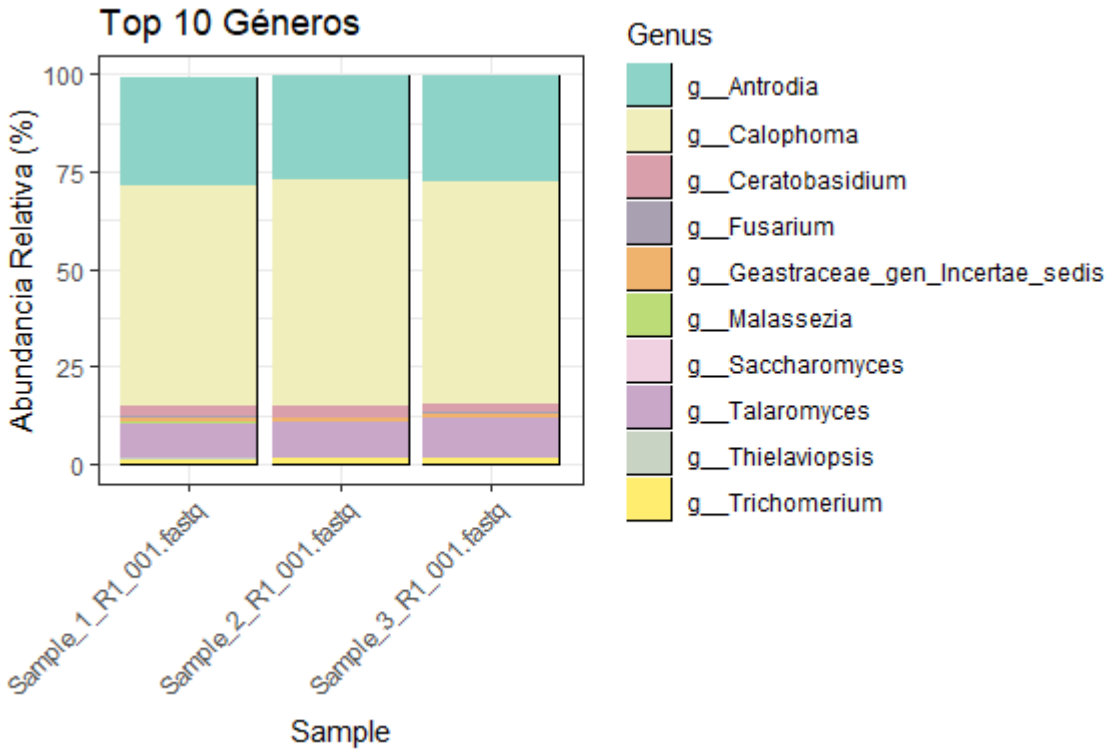

Figura 7. Abundancia relativa de los 10 géneros más abundantes de cada una de las muestras.

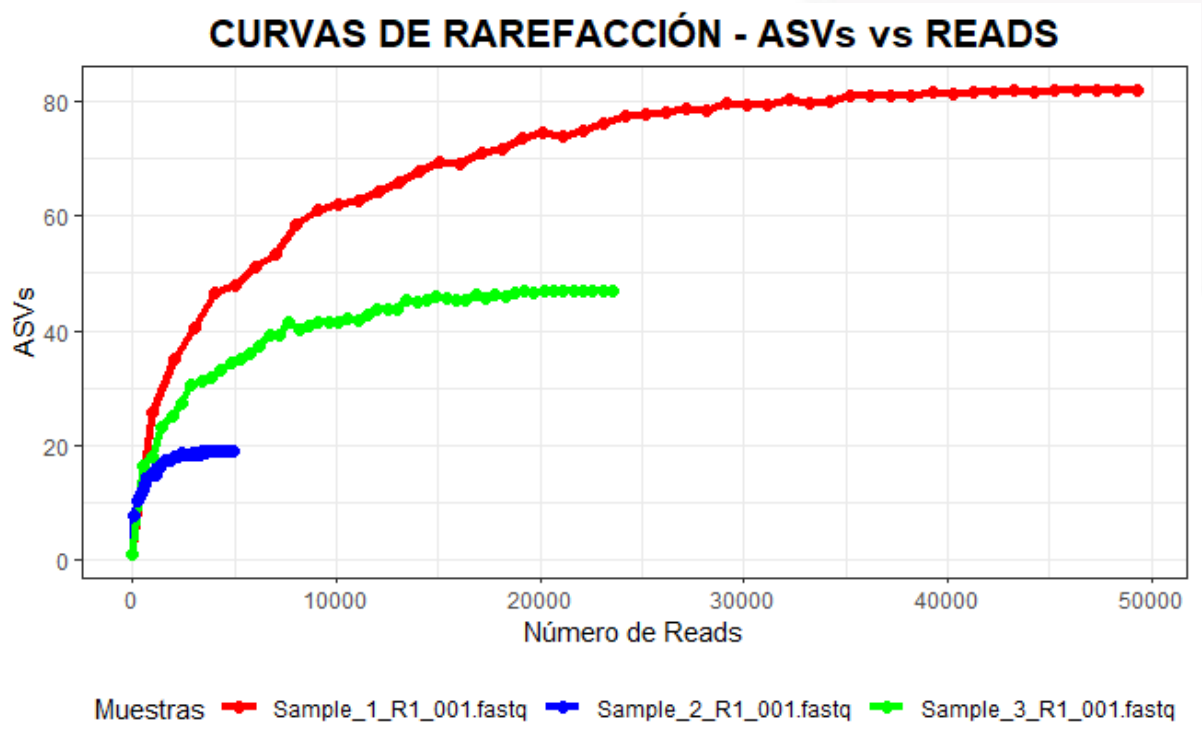

Figura 8. Curva de Rarefacción de las muestras analizadas.

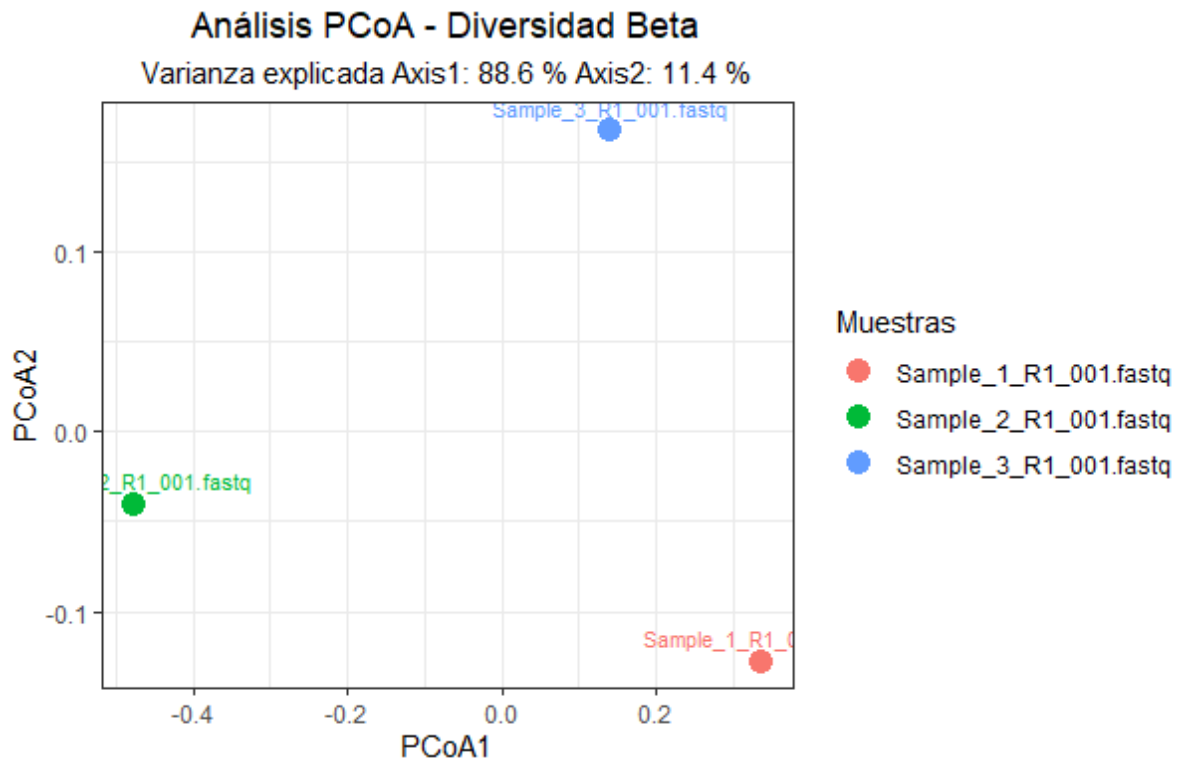

Figura 9. Análisis de PCoA de cada una de las muestras.

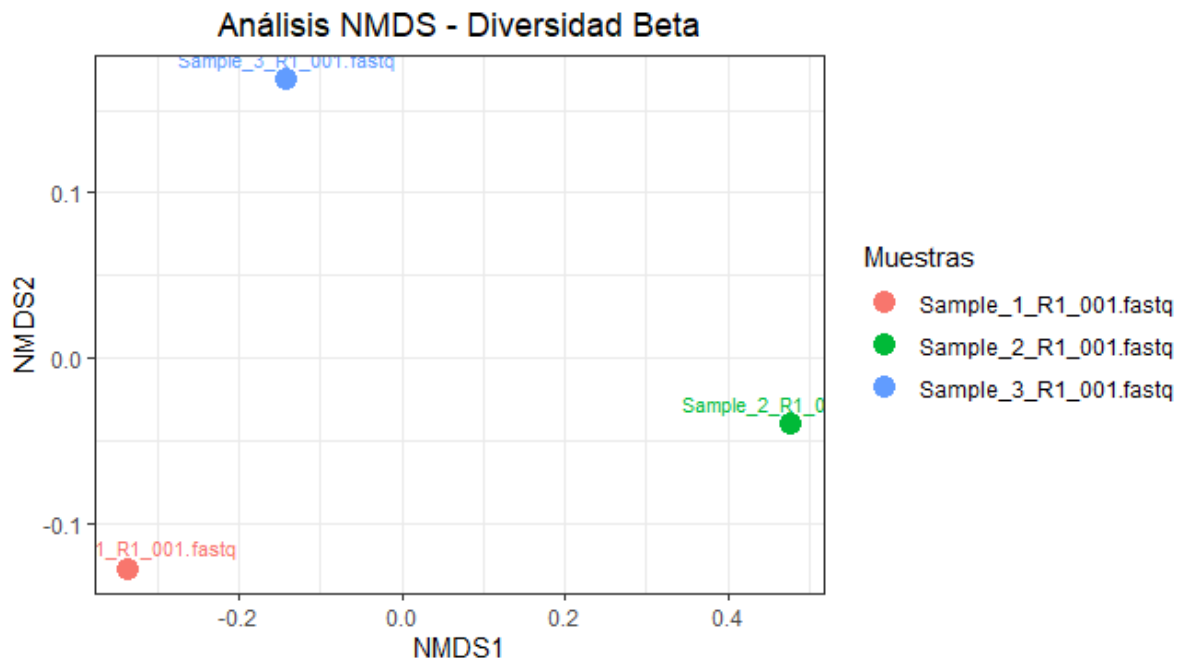

Figura 10. Análisis de NMDS de cada una de las muestras.

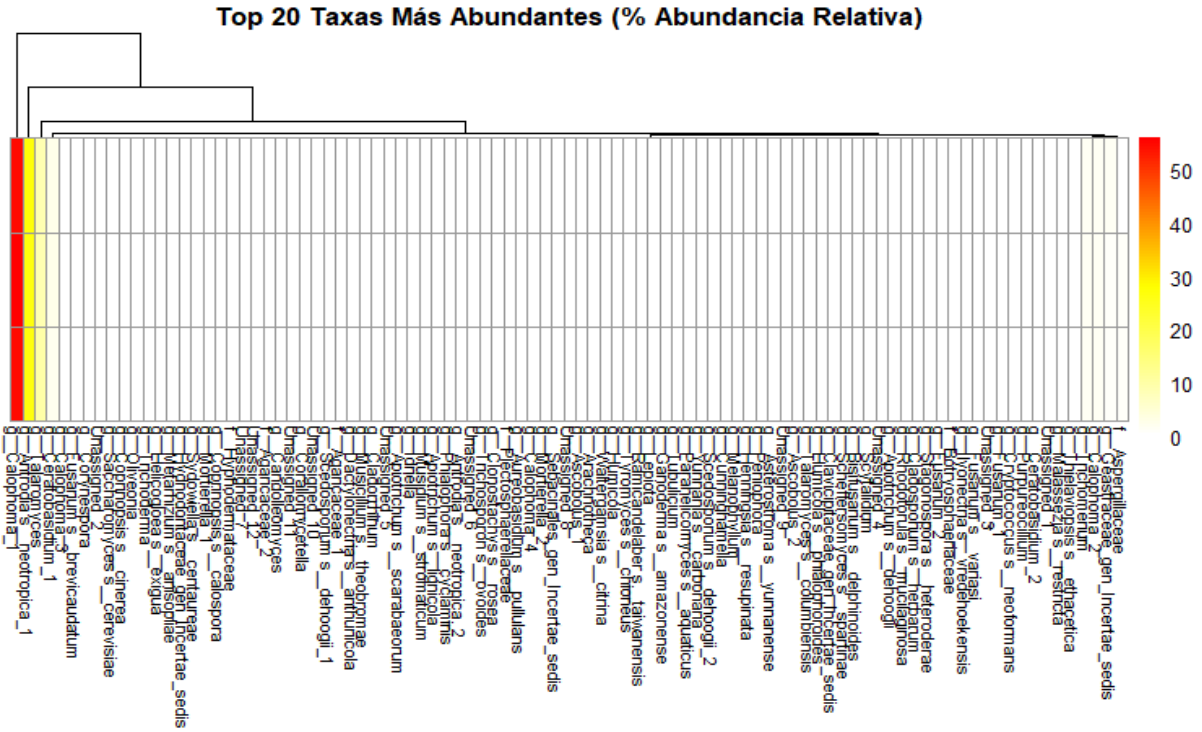

Figura 11. Heatmap de las 20 taxas más abundantes.

CAPÍTULO V: Conclusiones, Discusión y Recomendaciones

5.1 Discusión

El presente estudio tuvo como objetivo central el desarrollo y la validación de un pipeline de bioinformática en R, diseñado para ser accesible y reproducible en el análisis de datos metagenómicos del marcador fúngico ITS provenientes de la plataforma Illumina. Los resultados obtenidos confirman de manera contundente la hipótesis planteada: el pipeline desarrollado no solo fue funcional, sino que permitió una caracterización eficiente y completa de la diversidad fúngica en muestras ambientales. La implementación exitosa de este flujo de trabajo representa una contribución significativa para abordar el "cuello de botella" bioinformático que a menudo limita el potencial de la investigación metagenómica, especialmente en contextos con recursos computacionales y de personal especializado limitados.

El diseño del pipeline integró exitosamente prácticas estandarizadas y herramientas de alto rendimiento. La elección del paquete DADA2 (Callahan et al., 2016) para la inferencia de Variantes de Secuencia de Amplicón (ASV) fue fundamental, ya que este método ofrece una resolución taxonómica superior al agrupamiento tradicional en OTUs, permitiendo diferenciar variantes que difieren en un solo nucleótido. El preprocesamiento de los datos crudos, que incluyó el filtrado de calidad, la remoción de bases ambiguas y la eliminación de *primers* mediante la herramienta *cutadapt*, demostró ser un paso crítico. Como se observó en los perfiles de calidad, este filtrado riguroso mejoró sustancialmente la calidad de las secuencias retenidas, un requisito indispensable para evitar la propagación de errores en los pasos subsecuentes del análisis y para manejar los enormes volúmenes de datos que caracterizan a la

secuenciación de nueva generación (Nayfach et al., 2021; Probst et al., 2021). La capacidad del pipeline para procesar archivos .fastq de Illumina y generar una matriz de ASV final, limpia y anotada taxonómicamente, valida su funcionalidad y eficacia como herramienta analítica.

Desde una perspectiva biológica, los resultados del pipeline son consistentes con el conocimiento actual de la ecología fúngica. La identificación de los filos Ascomycota y Basidiomycota como los más abundantes en las muestras de suelo de manglar se alinea con numerosos estudios que los reportan como los grupos dominantes en la mayoría de los ecosistemas terrestres (Schmidt et al., 2018; Bahram et al., 2021). Este hallazgo no solo sirve como una validación biológica del flujo de trabajo, sino que también subraya la capacidad de las técnicas metagenómicas para revelar la composición de la "dark matter microbiana" (Li et al., 2023), es decir, aquella vasta porción de la diversidad que los métodos de cultivo tradicionales no logran capturar (Wang, Kirk & Yao, 2019). Las curvas de rarefacción, que alcanzaron una asíntota, indicaron que el esfuerzo de secuenciación fue adecuado para capturar la mayor parte de la diversidad de ASVs presentes en las muestras, un nivel de profundidad inalcanzable mediante el aislamiento en placa.

El poder del pipeline para la investigación ecológica se manifestó en los análisis de diversidad. Los índices de diversidad alfa (Shannon y Simpson) permitieron cuantificar y comparar la riqueza y equitatividad dentro de cada comunidad fúngica, revelando diferencias claras entre las muestras analizadas. Aún más revelador fue el análisis de diversidad beta, donde tanto el Análisis de Coordenadas Principales (PCoA) como el Escalamiento Multidimensional No Métrico (NMDS), basados en la distancia de Bray-Curtis, mostraron una separación visualmente evidente entre los grupos de muestras. Esta capacidad de discernir patrones en la composición de las comunidades es crucial

para responder preguntas ecológicas sobre cómo los factores abióticos o bióticos estructuran los microbiomas (Fierer et al., 2012). La integración de paquetes robustos como phyloseq (McMurdie & Holmes, 2013) y vegan (Oksaken et al., 2025) dentro de un único script asegura que estos análisis no solo sean posibles, sino que se realicen bajo un marco estadístico sólido y reproducible.

Una de las justificaciones centrales de este trabajo fue la necesidad de democratizar el análisis metagenómico. La creciente brecha entre la capacidad de generar datos de secuenciación y la disponibilidad de personal con formación interdisciplinaria para analizarlos representa un obstáculo significativo para la ciencia (Lewis & Bartlett, 2013). Este pipeline aborda directamente dicho problema al ofrecer un flujo de trabajo automatizado en R, un lenguaje de programación de acceso libre y ampliamente extendido en la comunidad académica. Al minimizar la necesidad de una profunda experiencia en línea de comandos o en la configuración de múltiples programas, se reduce la barrera de entrada para ecólogos, microbiólogos y otros investigadores de las ciencias de la vida. Esta iniciativa se alinea con los principios de la ciencia abierta, promoviendo la creación de herramientas que no solo sean potentes, sino también accesibles y transparentes (Erlich & Narayanan, 2020; Piro et al., 2021), permitiendo así que un mayor número de científicos pueda explorar la biodiversidad y funcionalidad de los microbiomas.

A pesar de los resultados, la precisión de la asignación taxonómica depende intrínsecamente de la exhaustividad de la base de datos de referencia. Aunque UNITE (Abarenkov et al., 2025) es la principal base de datos para la región ITS en hongos, siempre existirán taxones que no estén representados o que estén pobremente anotados, lo que puede resultar en secuencias asignadas solo a niveles taxonómicos altos (e.g., Reino o Filo) o incorrectamente clasificadas. En segundo lugar, este

estudio se basa en el análisis de ADN (metagenómica de amplicón), que identifica la presencia y abundancia relativa de los taxones, pero no distingue entre organismos vivos, inactivos o esporas, ni informa sobre su actividad metabólica. Para ello, serían necesarios enfoques metatranscriptómicos o metaproteómicos. Finalmente, el pipeline fue validado con un conjunto de datos específico; su aplicación a muestras de diferentes orígenes (e.g., agua, microbioma intestinal) podría requerir una optimización de ciertos parámetros, como los umbrales de filtrado o la longitud de corte de las secuencias.

En conclusión, este trabajo ha culminado con la creación de una herramienta bioinformática valiosa y funcional que simplifica el análisis de la diversidad fúngica a partir de datos de secuenciación masiva. Las implicaciones de este desarrollo son particularmente relevantes para países megadiversos como Ecuador, donde la exploración del potencial biotecnológico de los microorganismos es una frontera de investigación prometedora (Bai et al., 2025). Futuras mejoras podrían incluir el empaquetado del script en una aplicación con interfaz gráfica de usuario (GUI) a través de R Shiny, lo que aumentaría aún más su accesibilidad. Asimismo, el pipeline podría expandirse para incluir análisis funcionales predictivos o ser comparado con otros flujos de trabajo para establecer un *benchmark* de rendimiento. Al proporcionar una solución práctica y de código abierto, este proyecto contribuye a capacitar a la comunidad científica local para que pueda transformar el vasto potencial genético de los microbiomas en conocimiento y soluciones tangibles para los desafíos en salud, agricultura y biotecnología (Nagvire et al., 2022; Ma et al., 2024).

5.2 Conclusiones

1. El pipeline de bioinformática desarrollado en R demostró ser una herramienta eficaz, reproducible y accesible para el análisis de datos metagenómicos de la región ITS fúngica. Cumplió con su objetivo de superar las barreras técnicas y las limitaciones de los métodos dependientes de cultivo, permitiendo una caracterización completa de la diversidad fúngica a partir de datos crudos de secuenciación Illumina.
2. Los resultados biológicos generados por el pipeline son consistentes con el conocimiento ecológico actual, validando su precisión y relevancia. La identificación de los filos Ascomycota y Basidiomycota como dominantes y la obtención de curvas de rarefacción que alcanzan la asíntota confirman que la herramienta no solo es computacionalmente funcional, sino que también produce resultados ecológicamente coherentes y robustos.
3. El desarrollo de este flujo de trabajo contribuye significativamente a la democratización del análisis metagenómico. Al proporcionar una solución automatizada y de código abierto en un entorno de programación ampliamente utilizado como R, se reduce la brecha de conocimiento técnico y se capacita a investigadores de diversas disciplinas para que puedan analizar de manera autónoma la "dark matter microbiana", fomentando los principios de la ciencia abierta.

5.3 Recomendaciones

1.- Se sugiere expandir la funcionalidad del pipeline para incluir análisis predictivos de los roles ecológicos de los hongos identificados. La integración de herramientas

como FUNGuild o bases de datos similares permitiría asignar gremios funcionales (e.g., saprótrofos, patógenos, simbiontes) a las ASVs, pasando de una caracterización de "quién está ahí" a una hipótesis de "qué están haciendo" en el ecosistema.

2.- Realizar una validación y optimización del pipeline utilizando conjuntos de datos de diversos orígenes ambientales (e.g., muestras acuáticas, microbioma intestinal, filosfera) para evaluar su rendimiento y versatilidad.

3.-Adicionalmente, se recomienda llevar a cabo un análisis comparativo (benchmarking) con otros flujos de trabajo establecidos para cuantificar formalmente su eficiencia, precisión y consumo de recursos computacionales.

4.- Expandir la profundidad del análisis de *metabarcoding* a un análisis de *metagenoma* completo, para explorar diversidad metabólica y funcional con el objetivo de encontrar microorganismos con potencial biotecnológico.

REFERENCIAS BIBLIOGRÁFICAS

Lista de referencias en formato APA 7ma edición (sin numerar):

Abarenkov, K., Nilsson, R. H., Larsson, K.-H., Põlme, S., ... & Kõljalg, U. (2025). The UNITE database (version 10.0) for fungal ITS DNA barcoding. *Nucleic Acids Research*, *53*(D1), D123–D128. https://doi.org/10.1093/nar/gkad1020

Aranguren, R., Voyron, S., Ungaro, F., Cañón, J., & Lumini, E. (2023). Metabarcoding reveals impact of different land uses on fungal diversity in the south-eastern region of Antioquia, Colombia. *Plants*, *12*(5), 1126. https://doi.org/10.3390/plants12051126

Bai, D., Chen, T., Xun, J., Ma, C., Luo, H., Yang, H., ... & Liu, Y. X. (2025). EasyMetagenome: A user-friendly and flexible pipeline for shotgun metagenomic analysis in microbiome research. *Imeta*, *4*(1), e70001.

Bahram, M., Netherway, T., Frioux, C., Ferretti, P., Coelho, L. P., Geisen, S., ... & Hildebrand, F. (2021). Metagenomic assessment of the global diversity and distribution of bacteria and fungi. *Environmental Microbiology*, *23*(1), 316–326.

Berendsen, R. L., Pieterse, C. M., & Bakker, P. A. (2022). The rhizosphere microbiome and plant health. *Trends in Plant Science*, *17*(8), 478–486.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., ... & Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences*, *108*(supplement_1), 4516–4522.

Caruso, G. (2015). Plastic degrading microorganisms as a tool for bioremediation of plastic contamination in aquatic environments. *Journal of Pollution Effects & Control*, *3*(3), 1–2.

Cuadros-Orellana, S., Leite, L. R., Smith, A., Medeiros, J. D., Badotti, F., Fonseca, P. L., ... & Góes-Neto, A. (2013). Assessment of fungal diversity in the environment using metagenomics: a decade in review. *Fungal Genomics & Biology*, *3*(2), 1.

Emiyu, K., & Lelisa, K. (2022). Review on Illumina Sequencing Technology. *Austin Journal of Veterinary Science & Animal Husbandry*, *9*(1), 1080.

Fadiji, A. E., & Babalola, O. O. (2020). Metagenomics methods for the study of plant-associated microbial communities: a review. *Journal of Microbiological Methods*, *170*, 105860.

Fasolo, A., Deb, S., Stevanato, P., Concheri, G., & Squartini, A. (2024). ASV vs OTUs clustering: Effects on alpha, beta, and gamma diversities in microbiome metabarcoding studies. *PLOS ONE*, *19*(10), e0309065. https://doi.org/10.1371/journal.pone.0309065

Fierer, N., Leff, J. W., Adams, B. J., Nielsen, U. N., Bates, S. T., Lauber, C. L., ... & Caporaso, J. G. (2012). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences*, *109*(52), 21390–21395.

Ghosh, A., Saha, R., & Bhadury, P. (2022). Metagenomic insights into surface water microbial communities of a South Asian mangrove ecosystem. *PeerJ*, *10*, e13169.

Gilbert, J. A., Jansson, J. K., & Knight, R. (2018). Earth microbiome project and global systems biology. *MSystems*, *3*(3), e00271-17.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, *5*(10), R245–R249.

Hawksworth, D. L., & Lücking, R. (2017). Fungal diversity revisited: 2.2 to 3.8 million species. *Microbiology Spectrum*, *5*(4), FUNK-0052-2016.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oles, A. K., Pagès, H., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12*(2), 115–121. https://doi.org/10.1038/nmeth.3252

Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*, *82*(11), 801–811. https://doi.org/10.1016/j.humimm.2021.02.012

Landinez-Torres, A., Panelli, S., Picco, A. M., Comandatore, F., Tosi, S., & Capelli, E. (2019). A meta-barcoding analysis of soil mycobiota of the upper Andean Colombian agro-environment. *Scientific Reports*, *9*, 10085. https://doi.org/10.1038/s41598-019-46485-1

Lewis, J., & Bartlett, A. (2013). Inscribing a discipline: Tensions in the field of bioinformatics. *New Genetics and Society*, *32*(3), 243–263.

Li, S., Lian, W. H., Han, J. R., Ali, M., Lin, Z. L., Liu, Y. H., ... & Dong, L. (2023). Capturing the microbial dark matter in desert soils using culturomics-based metagenomics and high-resolution analysis. *NPJ Biofilms and Microbiomes*, *9*(1), 67.

Li, W., Wang, M. M., Wang, X. G., Cheng, X. L., Guo, J. J., Bian, X. M., & Cai, L. (2016). Fungal communities in sediments of subtropical Chinese seas as estimated by DNA metabarcoding. *Scientific Reports*, *6*(1), 26528.

Li, Z., Guo, X., Liu, B., Huang, T., Liu, R., & Liu, X. (2024). Metagenome sequencing reveals shifts in phage-associated antibiotic resistance genes from influent to effluent in wastewater treatment plants. *Water Research*, *253*, 121289.

Ma, Y., Wu, N., Zhang, T., Li, Y., Cao, L., Zhang, P., Zhang, Z., Zhu, T., & Zhang, C. (2024). The microbiome, resistome, and their co-evolution in sewage at a hospital for infectious diseases in Shanghai, China. *Microbiology Spectrum*, *12*(3), e03900-23.

Magurran, A. E. (1988). *Ecological diversity and its measurement*. Princeton University Press.

McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLOS ONE*, *8*(4), e61217. https://doi.org/10.1371/journal.pone.0061217

Molefe, R. R., Amoo, A. E., & Babalola, O. O. (2021). Metagenomic insights into the bacterial community structure and functional potentials in the rhizosphere soil of maize plants. *Journal of Plant Interactions*, *16*(1), 258–269.

Morgan, M., Anders, S., Lawrence, M., Aboyoun, P., Pagès, H., & Gentleman, R. (2009). ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*, *25*(19), 2607–2608. https://doi.org/10.1093/bioinformatics/btp450

Moti, T. B. (2022). Illumina sequencing technology review. *Microbiology Research International*, *10*(3), 25–31. https://doi.org/10.30918/MRI.103.22.022

Navgire, G. S., Goel, N., Sawhney, G., Sharma, M., Kaushik, P., Mohanta, Y. K., ... & Al-Harrasi, A. (2022). Analysis and Interpretation of metagenomics data: an approach. *Biological Procedures Online*, *24*(1), 18.

Nayfach, S., Camargo, A. P., Schulz, F., Eloe-Fadrosh, E., Roux, S., & Kyrpides, N. C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology*, *39*(5), 578–585.

New, F. N., & Brito, I. L. (2020). What is metagenomics teaching us, and what is missed?. *Annual Review of Microbiology*, *74*(1), 117–135.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szöcs, E., & Wagner, H. (2025). vegan: Community Ecology Package (Versión 2.6-6) [Software]. Comprehensive R Archive Network. https://CRAN.R-project.org/package=vegan

Oulas, A., Pavloudi, C., Polymenakou, P., Pavlopoulos, G. A., Papanikolaou, N., Kotoulas, G., ... & Iliopoulos, L. (2015). Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinformatics and Biology Insights*, *9*, 75–88.

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Pareek, C. S., Smoczynski, R., & Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, *52*, 413–435. https://doi.org/10.1007/s13353-011-0057-x

Portalanza, D., Acosta-Mejillones, A., Alcívar, J., Colorado, T., Guaita, J., Montero, L., Villao-Uzho, L., & Santos-Ordoñez, E. (2025). Fungal community dynamics in *Cyperus rotundus*: Implications for *Rhizophora mangle* in a mangrove ecosystem. *International Journal of Plant Biology*, *16*(1), 23. https://doi.org/10.3390/ijpb16010023

Porter, T. M., & Hajibabaei, M. (2022). MetaWorks: A flexible, scalable bioinformatic pipeline for high-throughput multi-marker biodiversity assessments. *PLOS ONE*, *17*(9), e0274260. https://doi.org/10.1371/journal.pone.0274260

Probst, M., Ascher-Jenull, J., Insam, H., & Gómez-Brandón, M. (2021). The molecular information about deadwood bacteriomes partly depends on the targeted environmental DNA. *Frontiers in Microbiology*, *12*, 640386.

Ravin, N. V., Rakitin, A. L., Ivanova, A. A., Beletsky, A. V., Kulichevskaya, I. S., Mardanov, A. V., & Dedysh, S. N. (2018). Genome analysis of *Fimbriiglobus ruber* SP5T, a planctomycete with confirmed chitinolytic capability. *Applied and Environmental Microbiology*, *84*(7), e02645-17.

Roy, S. (2011). Next-generation sequencing in oncology: genetic diagnosis, risk prediction and personalized medicine. *Expert Review of Molecular Diagnostics*, *11*(5), 507–518.

Sardar, S. K., & Gupta, D. (2018). Review on Illumina Sequencing Technology. *International Journal of Current Microbiology and Applied Sciences*, *7*(11), 3076-3081.

Schmidt, P. A., Schmitt, I., Otte, J., Bandow, C., Römbke, J., Bálint, M., & Rolshausen, G. (2018). Season-long experimental drought alters fungal community composition but not diversity in a grassland soil. *Microbial Ecology*, *75*(2), 468–478.

Schmidt, R., Mitchell, J., & Scow, K. (2019). Cover cropping and no-till increase diversity and symbiotroph: saprotroph ratios of soil fungal communities. *Soil Biology and Biochemistry*, *129*, 99–109.

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., ... & White, M. M. (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.

Sholder, G., Lanz, T. A., Moccia, R., Quan, J., Aparicio-Prat, E., Stanton, R., & Xi, H. S. (2020). 3’Pool-seq: an optimized cost-efficient and scalable method of whole-transcriptome gene expression profiling. *BMC Genomics*, *21*(1), 64.

Taş, N., de Jong, A. E., Li, Y., Trubl, G., Xue, Y., & Dove, N. C. (2021). Metagenomic tools in microbial ecology research. *Current Opinion in Biotechnology*, *67*, 184–191.

Tedersoo, L., Bahram, M., Kennedy, P. G., Yang, T., Anslan, S., Zinger, L., Nilsson, R. H., & Mikryukov, V. (2022). Best practices in metabarcoding of fungi: From experimental design to results. *Molecular Ecology*, *31*(11), 3019–3047. https://doi.org/10.1111/mec.16460

The evolution of DNA sequencing. (2022). *Nature Methods*, *19*, 649. https://doi.org/10.1038/s41592-022-01562-8

Urbina, H., Scofield, D. G., Cafaro, M., & Rosling, A. (2016). DNA-metabarcoding uncovers the diversity of soil-inhabiting fungi in the tropical island of Puerto Rico. *Mycoscience*, *57*(4), 217–227. https://doi.org/10.1016/j.myc.2016.02.001

Van der Auwera, G. A., & O'Connor, B. D. (2020). *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. O'Reilly Media.

Wang, K., Kirk, P. M., & Yao, Y. J. (2020). Development trends in taxonomy, with special reference to fungi. *Journal of Systematics and Evolution*, *58*(4), 406–412.

White, T. J., Bruns, T., Lee, S., & Taylor, J. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In M. A. Innis, D. H. Gelfand, J. J. Sninsky, & T. J. White (Eds.), *PCR Protocols: A Guide to Methods and Applications* (pp. 315–322). Academic Press.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer.

Xia, H., Zhang, Z., Luo, C., Wei, K., Li, X., Mu, X., Duan, M., et al. (2023). MultiPrime: A Reliable and Efficient Tool for Targeted Next-Generation Sequencing. *iMeta*, *2*, e143. https://doi.org/10.1002/imt2.143

Xia, M., Suseela, V., McCormack, M. L., Kennedy, P. G., & Tharayil, N. (2023). Common and lifestyle-specific traits of mycorrhizal root metabolome reflect ecological strategies of plant–mycorrhizal interactions. *Journal of Ecology*, *111*(3), 601–616.

Zhu, Q., Huang, S., Gonzalez, A., McGrath, I., McDonald, D., Haiminen, N., ... & Knight, R. (2022). Phylogeny-aware analysis of metagenome community ecology

based on matched reference genomes while bypassing taxonomy. *MSystems*, *7*(2), e00167-22.

ANEXOS

Anexo 1. Código para la instalación de la herramienta cutadap de conda en Windows.

```
install_cutadapt <- function() {
  # Verificar si Python está instalado
  python_check <- system("python --version", intern = TRUE)
  if(length(python_check) == 0) {
    stop("Python no está instalado.")
  }
  # Instalar cutadapt
  system("pip install cutadapt")
  version <- system("cutadapt --version", intern = TRUE)
 cat("Cutadapt versión:", version, "\n")}
```

Anexo 2. Enlace para la descargar el pipeline en formato R file

https://drive.google.com/drive/folders/1yOFEeN-tWDVagGShBQTvU4jyp-w_dvpW?usp=sharing