UNIVERSIDAD DE MURCIA

Escuela de Doctorado

TESIS DOCTORAL

Aproximaciones moleculares para el estudio de la comunidad microbiana en suelos seminaturales y agrarios

Molecular approaches to the study of the microbial community in semi-natural and agricultural soils

AUTOR/A María Belén Barquero Martínez

DIRECTOR/ES Felipe Bastida López

Rubén López Mondéjar

2025

UNIVERSIDAD DE MURCIA Escuela de Doctorado

TESIS DOCTORAL

Aproximaciones moleculares para el estudio de la comunidad microbiana en suelos seminaturales y agrarios

Molecular approaches to the study of the microbial community in semi-natural and agricultural soils

AUTOR/A María Belén Barquero Martínez

DIRECTOR/ES Felipe Bastida López
Rubén López Mondéjar

2025

UNIVERSIDAD DE MURCIA

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA EN MODALIDAD DE COMPENDIO O ARTÍCULOS PARA OBTENER EL TITULO DE DOCTOR/A

Aprobado por la Comisión General de Doctorado el 19 de octubre de 2022.

Yo, Dña. María Belén Barquero Martínez, habiendo cursado el Programa de Doctorado en Biodiversidad y Gestión Ambiental de la Escuela Internacional de Doctorado de la Universidad de Murcia (EIDUM), como autor/a de la tesis presentada para la obtención del título de Doctor/a titulada:

Aproximaciones moleculares para el estudio de la comunidad microbiana en suelos seminaturales y agrarios

y dirigida por:

D.: Felipe Bastida López
D.: Rubén López Mondéjar

DECLARO QUE:

La tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la Ley de Propiedad Intelectual (R.D. legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, modificado por la Ley 2/2019, de 1 de marzo, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita, cuando se han utilizado sus resultados o publicaciones.

Además, al haber sido autorizada como prevé el artículo 29.8 del reglamento cuenta con:

- *La aceptación por escrito de los coautores de las publicaciones de que el doctorando las presente como parte de la tesis.*
- *En su caso, la renuncia por escrito de los coautores no doctores de dichos trabajos a presentarlos como parte de otras tesis doctorales en la Universidad de Murcia o en cualquier otra universidad.*

Del mismo modo, asumo ante la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad del contenido de la tesis presentada, en caso de plagio, de conformidad con el ordenamiento jurídico vigente.

Murcia, a 2 de mayo de 2025

Información básica sobre protección de sus datos personales aportados:	
Responsable	Universidad de Murcia. Avenida teniente Flomesta, 5. Edificio de la Convalecencia. 30003; Murcia. Delegado de Protección de Datos: dpd@um.es
Legitimación	La Universidad de Murcia se encuentra legitimada para el tratamiento de sus datos por ser necesario para el cumplimiento de una obligación legal aplicable al responsable del tratamiento. art. 6.1.c) del Reglamento General de Protección de Datos
Finalidad	Gestionar su declaración de autoría y originalidad
Destinatarios	No se prevén comunicaciones de datos
Derechos	Los interesados pueden ejercer sus derechos de acceso, rectificación, cancelación, oposición, limitación del tratamiento, olvido y portabilidad a través del procedimiento establecido a tal efecto en el Registro Electrónico o mediante la presentación de la correspondiente solicitud en las Oficinas de Asistencia en Materia de Registro de la Universidad de Murcia

Esta DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD debe ser insertada en la quinta hoja, después de la portada de la tesis presentada para la obtención del título de Doctor/a.

This doctoral thesis is part of the R&D project / grant PID2020-114942RB-I00, funded by MCIN/ AEI/10.13039/501100011033/ and part of the project CNS2023-144692 funded by the Spanish Ministry of Science, Innovation and Universities and by the European Union "NextGenerationEU/PRTR”.

La vida no persiste porque evite los peligros,

sino porque los supera.

René Dubos

Desde pequeña, mis abuelos y mis padres me enseñaron a esforzarme para conseguir lo que me proponía, a ser humilde, a valorar lo que tengo y lo que logro, a ser buena con quienes me rodean. Pero, sobre todo, me enseñaron a ser agradecida con las personas que me ayudan y que están a mi lado. Es por eso, no puedo comenzar esta tesis doctoral sin dedicar unas palabras a quienes han hecho posible este camino.

Quiero empezar agradeciendo a todo mi grupo de investigación, a mis compañeros, por su ayuda, su apoyo y por ser personas maravillosas a las que siempre estaré agradecida. A María, por ser esa compañera multitarea que siempre está dispuesta a echar una mano a todos. Al Dr. José Siles, por su profesionalidad, por compartir sus conocimientos con todos y por estar siempre dispuesto a ayudar. Ojalá algún día pueda escribir un paper tan bien como tú. Al Dr. Alfonso, por estar siempre ahí para echarme un cable, no solo con la tesis, sino también con todas las gestiones que han ido surgiendo. Muchas gracias por mostrarte siempre dispuesto a ayudarme. A Marina, muchas gracias por ser tan cercana como profesional, siguiendo la tónica de este grupo. Y, por supuesto, a Celia, que además de ser la caña, ha dedicado tanto tiempo al trabajo de campo, esencial para recopilar la información que sustenta esta tesis. Gracias de corazón.

Puedo decir, sin dudarlo, que este grupo no solo está formado por personas inteligentes y preparadas, sino que destaca por su calidad humana. Esto, que a veces o no se valora lo suficiente, para mí es esencial. Gracias, de verdad, por ser tan buena gente.

Más allá del equipo, hay dos personas que han sido claves en este camino: el Dr. Felipe Bastida López y el Dr. Rubén López Mondéjar. No solo han sido mis directores de tesis, han sido mis maestros. Les debo todo lo que he aprendido en estos años, y mucho más. Gracias por vuestra guía, vuestra paciencia infinita y por tratarme siempre con tanto respeto y cariño. Ha sido un auténtico privilegio teneros como mentores.

Quiero mostrar también mi agradecimiento a todas las personas que conforman el grupo de Enzimología y Biorremediación de Suelos y Residuos Orgánicos del CEBAS-CSIC, especialmente a José Luis Moreno y Carlos García. Y al CEBAS, por brindarme la oportunidad de trabajar en un centro puntero en edafología, que es el corazón y sustento de esta tesis.

También me gustaría hacer una mención especial al personal de ITAP, y en particular a Pachi, por su colaboración en el establecimiento de las parcelas que han dado lugar a los estudios de campo incluidos en dos de los capítulos de esta tesis. Gracias por hacer posible algo tan importante.

Quiero expresar también mi más sincero agradecimiento al Instituto de Microbiología de la Academia Checa de Ciencias y, en especial, al grupo de Environmental Microbiology, liderado por el Dr. Petr Baldrian, por brindarme la oportunidad de utilizar sus servidores, una pieza clave para el desarrollo de todos los análisis bioinformáticos llevados a cabo en esta tesis doctoral. Agradezco a la Dra. Priscila Thiago Dobbler su ayuda durante los análisis y por mostrarse siempre tan colaborativa y amable. Y, por último, muchas gracias a la Dra. Camelia Algora por su trabajo en la conformación de los datos empleados en el tercer capítulo de esta tesis. Mi más profundo agradecimiento.

A mis amigos, gracias por estar. A Pedro, Guillermo, a "los Pablos", Josema, Maido, Fran, y en especial a Carmen, que es como una hermana para mí. Muchas gracias.

Me gustaría dar las gracias en especial a mis padres, María José y Antonio, por enseñarme el valor de la familia, de las personas, y por inculcarme siempre que el esfuerzo tiene su recompensa. Todo lo que he hecho siempre ha sido para haceros sentir orgullosos e intentar devolveros un poco de todo lo que habéis hecho por mí. Porque sé lo que os habéis esforzado

para darnos a mí y a mi hermana lo mejor. Y porque sé que si estoy aquí, si he tenido la oportunidad de estudiar, es gracias a vosotros. Sois los padres que todo hijo querría tener.

Muchas gracias a mi hermana, Gema, que le dio sentido a mi vida el día que nació al convertirme en hermana mayor. Y a mis abuelos, Juan y Josefa, que me enseñaron que lo más importante en la vida era ser buena persona.

No podia finalizar sin darle las gracias a mi pareja y compañero de vida. El amor es lo que mueve al mundo, y una de las pocas cosas que realmente merecen la pena. Gracias por existir, Jose.

Esta tesis doctoral va dedicada a Juan y a Julieta, por supuesto.

Gracias.

RESUMEN

Esta tesis doctoral aborda en profundidad la ecología microbiana del suelo, centrándose en los procesos que regulan los ciclos biogeoquímicos del fósforo (P), nitrógeno (N) y carbono (C) en suelos agrícolas y naturales. A través de la integración de enfoques multi-ómicos, que incluyen metagenómica, metaproteómica, metatranscriptómica y genomas ensamblados a partir de metagenomas (MAGs), se analiza la composición funcional y taxonómica del microbioma del suelo, su especialización ecológica y su respuesta a factores ambientales clave como las prácticas de fertilización, la fenología del cultivo y la descomposición de materia orgánica. Los resultados de esta tesis no solo confirman las hipótesis iniciales, sino que revelan dimensiones hasta ahora poco exploradas del funcionamiento microbiano en ecosistemas terrestres, aportando conocimientos aplicables tanto al avance de la ciencia básica como al diseño de estrategias para una agricultura sostenible.

Uno de los hallazgos más relevantes de este estudio es la identificación de gremios microbianos altamente especializados dentro de los ciclos del fósforo, nitrógeno y carbono. En el caso del fósforo, se observa una clara diferenciación funcional entre los microorganismos implicados en la solubilización del fósforo inorgánico y aquellos responsables de la mineralización del fósforo orgánico. Esta separación se refleja en la composición taxonómica: las Actinobacteria albergan genes implicados en la solubilización de fósforo inorgánico, pero no en la mineralización del fósforo orgánico, lo que sugiere una adaptación evolutiva a fuentes específicas de este nutriente. De este modo, se pone de manifiesto la importancia de considerar no solo la disponibilidad total de fósforo en el suelo, sino también la diversidad funcional de los microorganismos involucrados en su transformación. A ello se suma el descubrimiento del papel relevante de las arqueas en el ciclo del fósforo, cuya contribución ha sido históricamente subestimada. Esta tesis demuestra que diversos taxones arqueales contienen genes relacionados con el metabolismo del fósforo, lo que sugiere que desempeñan funciones complementarias o incluso esenciales junto a las bacterias en la regulación de este ciclo. Asimismo, el análisis metaproteómico permitió identificar enzimas clave como la fosfatasa alcalina codificada por el gen *phoX*, la cual se encontró abundantemente expresada en agroecosistemas de maíz y puede emplearse como biomarcador para monitorizar la disponibilidad de fósforo en suelos agrícolas.

Este hallazgo no solo aporta información sobre la funcionalidad del microbioma edáfico, sino que también sugiere posibles aplicaciones prácticas. El desarrollo de biosensores basados en la actividad de *phoX* podría convertirse en una herramienta útil para la gestión del fósforo en sistemas agrícolas, facilitando la toma de decisiones sobre el momento óptimo para aplicar fertilizantes o enmiendas orgánicas. Además, el papel de las arqueas en la regulación de este ciclo reabre el debate sobre su inclusión en modelos de predicción de la dinámica de nutrientes, históricamente centrados en bacterias y hongos.

En cuanto al ciclo del nitrógeno, los datos muestran que los nichos funcionales microbianos están muy bien definidos, con poca superposición entre grupos funcionales. Las Nitrososphaeraceae, por ejemplo, se presentan como nitrificantes especializados que carecen de genes relacionados con la fijación de

nitrógeno molecular o con rutas de transporte de nitrógeno, lo que refuerza la existencia de funciones ecológicas diferenciadas. Este patrón se repite en los desnitrificantes, que forman gremios agrupados taxonómicamente asociados a rutas metabólicas específicas. Las técnicas metagenómicas y metaproteómicas permitieron no solo detectar la presencia de estos gremios, sino también vincularlos con enzimas funcionales como la glutamina sintetasa (GlnA), cuya relevancia en la asimilación de nitrógeno ha sido escasamente reconocida en estudios anteriores. Además, mediante reconstrucciones de MAGs, se confirmó la implicación de determinadas familias bacterianas, como Propionibacteriaceae, en procesos de desnitrificación, aportando una visión más detallada de los actores microbianos que regulan el ciclo del nitrógeno.

Un aspecto clave de este trabajo es la demostración de que la fenología del cultivo es el principal factor que determina la abundancia y actividad de los genes implicados en el ciclo del nitrógeno, superando incluso el efecto de la fertilización. Las diferentes fases del desarrollo vegetal condicionan la expresión de genes responsables de procesos como la nitrificación y la desnitrificación, lo que sugiere que las estrategias de fertilización deberían adaptarse no solo a la composición del suelo, sino también al momento fenológico del cultivo para optimizar el uso del nitrógeno y reducir pérdidas ambientales. No obstante, los distintos tratamientos de fertilización también muestran efectos significativos. Los fertilizantes minerales, como NPK y estruvita, tienden a estimular rutas metabólicas como la nitrificación y la reducción disimilatoria de nitrato a amonio (DNRA), mientras que las enmiendas orgánicas promueven una mayor diversidad microbiana y una actividad funcional más diversa, aunque su eficacia depende de una gestión adecuada que garantice la sincronía entre la liberación de nutrientes y la demanda del cultivo.

El hallazgo de que la fenología modula de manera más decisiva que la fertilización la expresión de genes clave del nitrógeno tiene profundas implicaciones agronómicas. Sugiere, por ejemplo, que la eficiencia de uso del nitrógeno podría mejorarse no solo ajustando dosis de aplicación, sino también modificando el calendario de fertilización en función de los momentos de máxima actividad microbiana, en estrecha sincronía con las necesidades fisiológicas del cultivo.

En el ciclo del carbono, los resultados de esta tesis cuestionan la visión tradicional que otorga un papel predominante a los hongos en la descomposición de la materia orgánica compleja. Se identificaron gremios bacterianos altamente especializados en la degradación de biopolímeros como la quitina, los β-1,3-glucanos y la celulosa. Entre ellos, destacan géneros como *Chitinophaga* y *Pedobacter*, especializados en la degradación de quitina, *Terriglobus* en la degradación de β-1,3-glucanos y *Asticcacaulis* en la degradación de celulosa. Estos resultados evidencian que las bacterias desempeñan un papel clave, y en algunos contextos dominante, en la descomposición de compuestos complejos de origen vegetal y microbiano. Este hallazgo se vio reforzado por análisis metatranscriptómicos, que demostraron que la actividad metabólica de estos descomponedores bacterianos era alta y que, en muchos casos, superaba a la de grupos tradicionalmente considerados como los principales descomponedores. A pesar de que las Proteobacteria presentaban la mayor abundancia de genes en

términos generales, no eran las más activas desde el punto de vista transcripcional, lo que subraya la necesidad de adoptar una perspectiva funcional al estudiar el papel ecológico de los microorganismos, más allá de su mera presencia o abundancia relativa.

Este cambio de paradigma en el papel relativo de bacterias y hongos en la descomposición de materia orgánica puede tener efectos importantes en el diseño de estrategias de manejo del carbono en suelo. Por ejemplo, podría promover el uso de inoculantes bacterianos especializados en la degradación de compuestos recalcitrantes, o incluso inspirar prácticas agrícolas que favorezcan las condiciones de actividad óptima para dichos grupos.

En conjunto, los resultados obtenidos permiten afirmar que la fenología del cultivo es el principal impulsor de los cambios funcionales en las comunidades microbianas del suelo, condicionando la expresión de genes clave en los ciclos de fósforo y nitrógeno. Esta observación tiene importantes implicaciones prácticas, ya que indica que las estrategias de manejo de nutrientes deben estar estrechamente alineadas con el desarrollo fenológico de los cultivos. Por ejemplo, genes relacionados con la solubilización del fósforo y con la nitrificación presentan variaciones significativas a lo largo de las etapas de crecimiento de las plantas, lo que sugiere que la aplicación de fertilizantes o biofertilizantes debería ajustarse temporalmente para maximizar su eficacia. Asimismo, aunque los efectos de la fertilización son secundarios respecto a la fenología, siguen siendo relevantes: los insumos minerales estimulan rutas metabólicas específicas, mientras que las enmiendas orgánicas requieren una mayor planificación para evitar pérdidas por mineralización temprana o inmovilización de nutrientes. Este enfoque permite diseñar esquemas de fertilización más eficientes, que equilibran el suministro de nutrientes con la dinámica microbiana del suelo.

Desde el punto de vista metodológico, esta tesis aporta importantes innovaciones al demostrar la utilidad de integrar múltiples capas de datos ómicos en el estudio de la ecología microbiana del suelo. La metagenómica permitió establecer un marco general sobre la diversidad y el potencial funcional de las comunidades microbianas. La metaproteómica y la metatranscriptómica ofrecieron información detallada sobre la actividad enzimática y la expresión génica bajo diferentes condiciones ambientales y prácticas de manejo. Por su parte, la reconstrucción de genomas a partir de metagenomas permitió identificar nuevos taxones con funciones específicas, ampliando el repertorio de microorganismos relevantes para el funcionamiento de los ciclos biogeoquímicos. Esta integración ha permitido descubrir nuevos actores microbianos implicados en procesos clave como la degradación de celulosa o quitina, así como enzimas subestimadas como la GlnA en la asimilación de nitrógeno. La profundidad y resolución de este enfoque multi-ómico ofrece nuevas posibilidades para el desarrollo de herramientas diagnósticas y estrategias de intervención en la agricultura.

Además de su contribución al conocimiento básico, esta tesis ofrece importantes oportunidades para el diseño de nuevas herramientas biotecnológicas aplicadas al manejo sostenible de los suelos agrícolas.

El descubrimiento de taxones microbianos altamente especializados, junto con la identificación de biomarcadores funcionales como *phoX* y GlnA, abre la posibilidad de desarrollar kits moleculares para el diagnóstico rápido del estado funcional del suelo. Estos biomarcadores podrían integrarse en sistemas de agricultura de precisión, facilitando la toma de decisiones sobre el momento y tipo de fertilización, la aplicación de biofertilizantes o la rotación de cultivos, en función del estado real de la microbiota edáfica y no solo de parámetros químicos convencionales.

Igualmente, la capacidad predictiva derivada de la integración de datos ómicos con algoritmos de aprendizaje automático permite anticipar respuestas funcionales del microbioma del suelo frente a distintos escenarios de manejo, como cambios en el tipo de cultivo, eventos climáticos extremos o la introducción de nuevas prácticas agronómicas. Estos modelos podrían implementarse en plataformas digitales de apoyo a la gestión agrícola, convirtiendo el conocimiento generado por esta investigación en una herramienta efectiva para productores, técnicos y responsables de políticas públicas.

Por otro lado, la constatación del impacto de la fenología sobre la expresión funcional de los ciclos biogeoquímicos plantea nuevos desafíos para la investigación ecológica, particularmente en la necesidad de realizar muestreos más ajustados temporalmente. Esta perspectiva también puede ser útil para estudios de cambio climático, ya que las alteraciones fenológicas inducidas por el aumento de temperaturas pueden desincronizar los procesos microbiológicos del suelo respecto a las necesidades del cultivo, generando desequilibrios que comprometan la productividad y la eficiencia en el uso de nutrientes.

A la luz de los resultados, se destaca la importancia de fomentar líneas de investigación interdisciplinarias que integren microbiología del suelo, agronomía, ciencia de datos y ecología funcional. Además, resulta necesario impulsar estudios de largo plazo en parcelas experimentales para validar la estabilidad de los gremios microbianos identificados y su respuesta sostenida ante diferentes regímenes de manejo. La evolución de estos sistemas bajo un contexto de intensificación sostenible debe ser evaluada no solo en términos de productividad, sino también de estabilidad funcional, diversidad y servicios ecosistémicos.

Más allá del ámbito científico y técnico, los resultados de esta tesis también tienen importantes implicaciones sociales y políticas. La comprensión profunda del papel que desempeñan los microorganismos del suelo en los ciclos de nutrientes puede contribuir a transformar los modelos agrícolas actuales, excesivamente dependientes de insumos externos, hacia sistemas más autosuficientes y basados en procesos ecológicos. Esta transición es fundamental en un contexto global marcado por la crisis climática, la pérdida de biodiversidad y la necesidad urgente de garantizar la seguridad alimentaria a largo plazo.

Asimismo, los conocimientos generados por esta investigación pueden servir de base para diseñar políticas públicas que reconozcan explícitamente la importancia de la salud microbiana del suelo como

componente esencial de la sostenibilidad agrícola. Por ejemplo, los programas de incentivos agroambientales podrían incorporar indicadores funcionales del microbioma como criterios de elegibilidad o éxito, incentivando prácticas que favorezcan su diversidad y funcionalidad.

En el ámbito de la educación y la formación, esta tesis resalta la necesidad de integrar de manera transversal la ecología microbiana en los programas de estudios agronómicos, biotecnológicos y ambientales. La capacidad de interpretar datos ómicos y traducirlos en recomendaciones prácticas será cada vez más demandada, tanto en el sector académico como en el profesional. De este modo, se propone fomentar una nueva generación de investigadores y técnicos con competencias interdisciplinares, capaces de abordar la complejidad de los agroecosistemas desde una perspectiva sistémica e informada por los avances de la biología molecular y la bioinformática.

Finalmente, esta investigación constituye una aportación concreta al paradigma emergente de la gestión regenerativa de suelos, que no solo se limita a conservar los recursos existentes, sino que busca restaurar sus funciones ecológicas a través de una intervención inteligente basada en evidencia científica. El reconocimiento del suelo como un sistema biológicamente activo y dinámico, cuya funcionalidad depende en gran medida de sus comunidades microbianas, representa un cambio de enfoque fundamental con profundas repercusiones ecológicas, económicas y sociales.

ABSTRACT

This doctoral thesis provides an in-depth exploration of soil microbial ecology, focusing on the processes that regulate the biogeochemical cycles of phosphorus (P), nitrogen (N), and carbon (C) in both agricultural and natural soils. By integrating multi-omics approaches—including metagenomics, metaproteomics, metatranscriptomics, and metagenome-assembled genomes (MAGs)—this study analyzes the functional and taxonomic composition of the soil microbiome, its ecological specialization, and its responses to key environmental factors such as fertilization practices, crop phenology, and the decomposition of organic matter. The findings of this thesis not only support the initial hypotheses but also uncover previously underexplored dimensions of microbial functioning in terrestrial ecosystems, contributing knowledge that is applicable both to basic science and to the development of strategies for sustainable agriculture.

One of the most significant findings of this study is the identification of highly specialized microbial guilds involved in the phosphorus, nitrogen, and carbon cycles. In the case of phosphorus, a clear functional differentiation is observed between microorganisms involved in the solubilization of inorganic phosphorus and those responsible for the mineralization of organic phosphorus. This separation is mirrored in taxonomic composition: Actinobacteria harbor genes associated with inorganic phosphorus solubilization but not with organic phosphorus mineralization, suggesting an evolutionary adaptation to specific nutrient sources. This highlights the importance of considering not only total phosphorus availability in soil but also the functional diversity of the microorganisms involved in its transformation. Additionally, the study uncovers the previously underestimated role of archaea in the phosphorus cycle. It demonstrates that various archaeal taxa possess genes related to phosphorus metabolism, suggesting that they may perform complementary or even essential functions alongside bacteria in regulating this cycle. Metaproteomic analysis also identified key enzymes such as the alkaline phosphatase encoded by the *phoX* gene, which was abundantly expressed in maize agroecosystems and may serve as a biomarker for monitoring phosphorus availability in agricultural soils.

This discovery not only enhances our understanding of the soil microbiome's functional capabilities but also suggests potential practical applications. The development of biosensors based on *phoX* activity could become a useful tool for phosphorus management in agricultural systems, informing decisions about optimal timing for fertilizer or organic amendment applications. Furthermore, the role of archaea in regulating this cycle reopens the debate about their inclusion in nutrient dynamics models, which have historically focused on bacteria and fungi.

Regarding the nitrogen cycle, the data reveal that microbial functional niches are well-defined, with minimal overlap between functional groups. For example, Nitrososphaeraceae appear as specialized nitrifiers lacking genes associated with nitrogen fixation or nitrogen transport pathways, reinforcing the existence of distinct ecological roles. This pattern is also evident in denitrifiers, which form taxonomically clustered guilds associated with specific metabolic pathways. Metagenomic and metaproteomic techniques not only

detected the presence of these guilds but also linked them to functional enzymes such as glutamine synthetase (GlnA), whose relevance in nitrogen assimilation has been largely overlooked in previous studies. Moreover, MAG reconstruction confirmed the involvement of specific bacterial families, such as Propionibacteriaceae, in denitrification processes, offering a more detailed understanding of the microbial actors that govern the nitrogen cycle.

A key aspect of this work is the demonstration that crop phenology is the primary factor influencing the abundance and activity of genes involved in the nitrogen cycle, surpassing even the effects of fertilization. The different stages of plant development shape the expression of genes responsible for processes such as nitrification and denitrification, suggesting that fertilization strategies should be tailored not only to soil composition but also to the phenological stage of the crop. This would optimize nitrogen use efficiency and minimize environmental losses. Nonetheless, the type of fertilization also exerts significant effects. Mineral fertilizers such as NPK and struvite tend to stimulate metabolic pathways including nitrification and dissimilatory nitrate reduction to ammonium (DNRA), whereas organic amendments promote greater microbial diversity and broader functional activity. However, their effectiveness depends on proper management to synchronize nutrient release with crop demand.

The finding that phenology more decisively modulates the expression of key nitrogen-related genes than fertilization carries substantial agronomic implications. For instance, nitrogen use efficiency could be improved not only by adjusting application rates but also by modifying the timing of fertilization to align with periods of peak microbial activity and the physiological demands of the crop.

With respect to the carbon cycle, this thesis challenges the traditional view that fungi play a predominant role in the decomposition of complex organic matter. Highly specialized bacterial guilds were identified as key degraders of biopolymers such as chitin, β-1,3-glucans, and cellulose. Notable examples include the genera *Chitinophaga* and *Pedobacter* for chitin degradation, *Terriglobus* for β-1,3-glucans, and *Asticcacaulis* for cellulose breakdown. These findings underscore the crucial—and in some contexts dominant—role of bacteria in the decomposition of complex plant- and microbe-derived compounds. This conclusion was further supported by metatranscriptomic analyses, which revealed that the metabolic activity of these bacterial decomposers was high and often exceeded that of groups traditionally regarded as the principal decomposers. Although Proteobacteria exhibited the highest gene abundance overall, they were not the most transcriptionally active, emphasizing the need to adopt a functional perspective when assessing the ecological role of microorganisms, beyond their mere presence or relative abundance.

This paradigm shift in the relative roles of bacteria and fungi in organic matter decomposition could significantly influence the design of soil carbon management strategies. For example, it may promote the use of bacterial inoculants specialized in the degradation of recalcitrant compounds, or even inspire agricultural practices that favor optimal activity conditions for these microbial groups.

Overall, the results obtained confirm that crop phenology is the primary driver of functional changes in soil microbial communities, shaping the expression of key genes involved in phosphorus and nitrogen cycling. This finding has important practical implications, as it suggests that nutrient management strategies should be closely aligned with the phenological development of crops. For instance, genes related to phosphorus solubilization and nitrification exhibit significant temporal variation across plant growth stages, indicating that the timing of fertilizer or biofertilizer application should be carefully adjusted to maximize efficacy. Although the effects of fertilization are secondary to those of phenology, they remain relevant: mineral inputs stimulate specific metabolic pathways, while organic amendments require more precise planning to avoid early mineralization losses or nutrient immobilization. This approach enables the design of more efficient fertilization schemes that balance nutrient supply with soil microbial dynamics.

From a methodological standpoint, this thesis introduces significant innovations by demonstrating the value of integrating multiple layers of omics data in the study of soil microbial ecology. Metagenomics provided a broad framework for assessing microbial diversity and functional potential. Metaproteomics and metatranscriptomics offered detailed insights into enzymatic activity and gene expression under different environmental conditions and management practices. Genome reconstruction from metagenomes enabled the identification of new taxa with specific functions, thereby expanding the repertoire of microorganisms relevant to the functioning of biogeochemical cycles. This integrative approach revealed novel microbial players involved in key processes such as cellulose or chitin degradation, as well as underestimated enzymes like GlnA in nitrogen assimilation. The depth and resolution afforded by the multi-omics framework open up new possibilities for developing diagnostic tools and targeted intervention strategies in agriculture.

In addition to advancing basic knowledge, this thesis also provides key opportunities for designing new biotechnological tools to support sustainable soil management in agricultural systems. The discovery of highly specialized microbial taxa, along with the identification of functional biomarkers such as *phoX* and GlnA, paves the way for the development of molecular kits for rapid diagnosis of soil functional status. These biomarkers could be integrated into precision agriculture systems, helping guide decisions on the timing and type of fertilization, the application of biofertilizers, or crop rotation, based on the actual condition of the soil microbiota rather than conventional chemical parameters alone.

Furthermore, the predictive capacity derived from integrating omics data with machine learning algorithms allows for the anticipation of functional responses of the soil microbiome to various management scenarios, such as changes in crop type, extreme weather events, or the introduction of new agronomic practices. These models could be implemented in digital decision-support platforms for agricultural management, transforming the knowledge generated by this research into a practical tool for farmers, technical advisors, and policy makers alike.

On the other hand, the demonstrated impact of phenology on the functional expression of biogeochemical cycles presents new challenges for ecological research, particularly in the need for temporally refined sampling strategies. This perspective may also prove valuable in climate change studies, as phenological shifts driven by rising temperatures could desynchronize soil microbial processes from crop requirements, leading to imbalances that compromise both productivity and nutrient use efficiency.

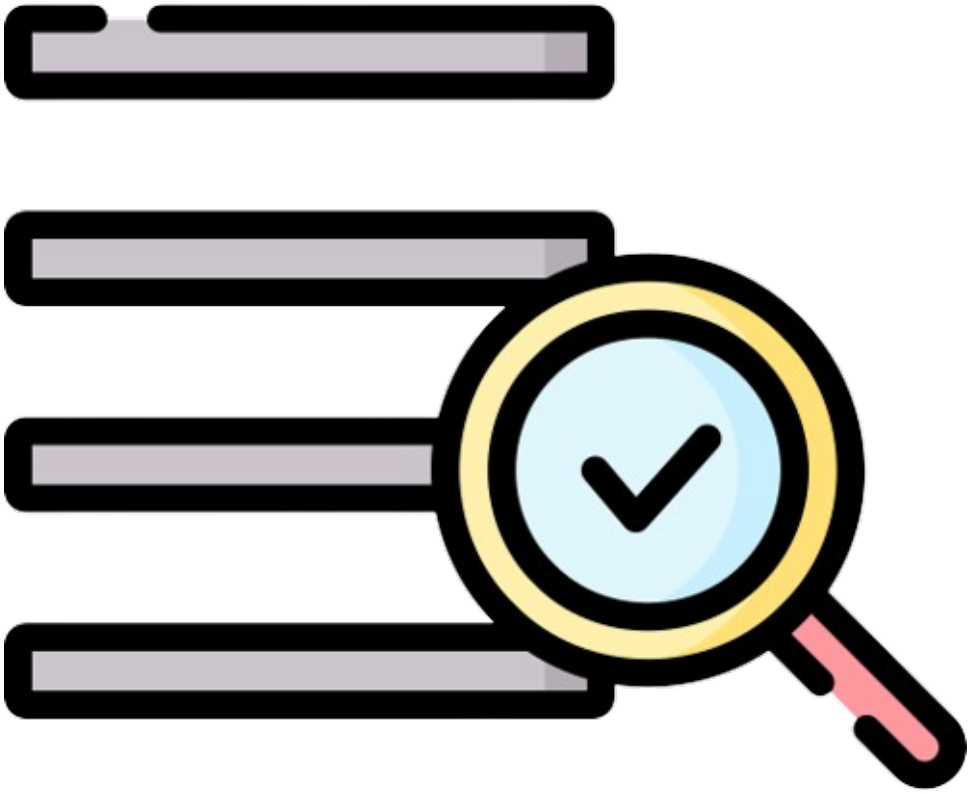
In light of these findings, the importance of promoting interdisciplinary research lines that integrate soil microbiology, agronomy, data science, and functional ecology becomes evident. Additionally, long-term studies in experimental plots are needed to validate the stability of the identified microbial guilds and their sustained responses to different management regimes. The development of these systems within a framework of sustainable intensification must be evaluated not only in terms of productivity but also in relation to functional stability, biodiversity, and ecosystem services.

Beyond the scientific and technical spheres, the results of this thesis also hold important social and political implications. A deep understanding of the role of soil microorganisms in nutrient cycling may help shift current agricultural models—overly reliant on external inputs—towards more self-sufficient systems grounded in ecological processes. This transition is critical in a global context marked by climate crisis, biodiversity loss, and the urgent need to ensure long-term food security.

Moreover, the knowledge generated by this research can serve as a foundation for public policies that explicitly acknowledge the importance of soil microbial health as a key component of agricultural sustainability. For instance, agri-environmental incentive programs could incorporate functional indicators of the microbiome as eligibility or performance criteria, thereby encouraging practices that support microbial diversity and functionality.

In the realm of education and training, this thesis underscores the need to mainstream microbial ecology within agronomic, biotechnological, and environmental curricula. The ability to interpret omics data and translate it into practical recommendations will become increasingly valuable in both academic and professional sectors. In this sense, fostering a new generation of interdisciplinary researchers and practitioners capable of addressing the complexity of agroecosystems from a systems-based perspective—grounded in molecular biology and bioinformatics—is a priority.

Ultimately, this research contributes concretely to the emerging paradigm of regenerative soil management, which goes beyond the conservation of existing resources to actively restore ecological functions through scientifically informed and intelligent intervention. Recognizing soil as a biologically active and dynamic system—whose functionality largely depends on its microbial communities—represents a fundamental shift in perspective with profound ecological, economic, and social implications.



INDEX

INDEX

GENERAL INTRODUCTION

GENERAL INTRODUCTION

1. Soil: A fundamental resource and its functions

Soil is an essential resource for agriculture, as it provides a physical and chemical medium through which plants acquire the nutrients required for their growth and development (Delgado & Gómez, 2024). This medium is primarily composed of particles of sand, clay, silt, and organic matter, which interact in complex ways to determine the structure and fertility of the soil. The proportion of these components influences the soil's ability to store water, nutrients, and support microbial life, which is crucial for the sustainability of agronomic systems (Kome et al., 2019; Russell et al., 1997). Additionally, soil regulates water availability, participates in the storage and release of essential nutrients, and acts as a reservoir of biodiversity (Havlicek & Mitchell, 2014; Nielsen et al., 2015; Turbé et al., 2010). It also functions as a buffer, regulating temperature, retaining moisture, and protecting plant roots from extreme climatic variations (Saco et al., 2021).

Beyond its ecological significance, soil performs critical functions that are essential for ecosystem health and agricultural productivity. These functions are summarized in Figure 1. These functions include carbon sequestration, water filtration, nutrient cycling, climate regulation, and biodiversity support (Smith et al., 2015). Soil facilitates the decomposition of organic matter and the recycling of essential nutrients, enabling plant growth and maintaining ecosystem balance (Horwath, 2007). By storing organic carbon, soil contributes to mitigating climate change and regulating greenhouse gas emissions (Rosenzweig & Hillel, 2000). Furthermore, soil acts as a natural filter, improving water quality and controlling water flow within the environment (Pierzynski et al., 2005). Simultaneously, it serves as a habitat for countless microorganisms that sustain biodiversity, which underpins all other functions (Prasad et al., 2021). Soil also plays a key role in climate stabilization through its involvement in regulating the cycles of phosphorus, nitrogen, carbon, and water (Goh, 2004). Further, it provides a stable foundation for crops, infrastructure, and ecosystems, ensuring human and environmental resilience (F. Shah & Wu, 2019).

The ability of soil to deliver these ecosystem services is integral to sustainable agricultural practices and is closely linked to global and European environmental and agricultural policies, such as the EU Green Deal, the EU Soil Mission, and the Directive on Soil Health. These initiatives emphasize the importance of preserving and enhancing soil functionality to combat climate change, ensure food security, and protect biodiversity across Europe (Creamer et al., 2010; Montaldo, 2022; Panagos et al., 2024). For instance, the EU Green Deal aims to achieve carbon neutrality by 2050, with soil's role in carbon sequestration being pivotal to this goal (Panagos et al., 2022). Similarly, the EU Soil Mission focuses on restoring soil health to ensure that 75% of European soils are in good condition by 2030 (Efthimiou, 2025). Meanwhile, the Directive on Soil Health establishes guidelines to maintain soil functionality and prevent its degradation (Lehmann et al., 2020). Understanding and prioritizing the multifaceted roles of soil enables the development of strategies aligned with these policies, thereby contributing to a more sustainable and resilient future.

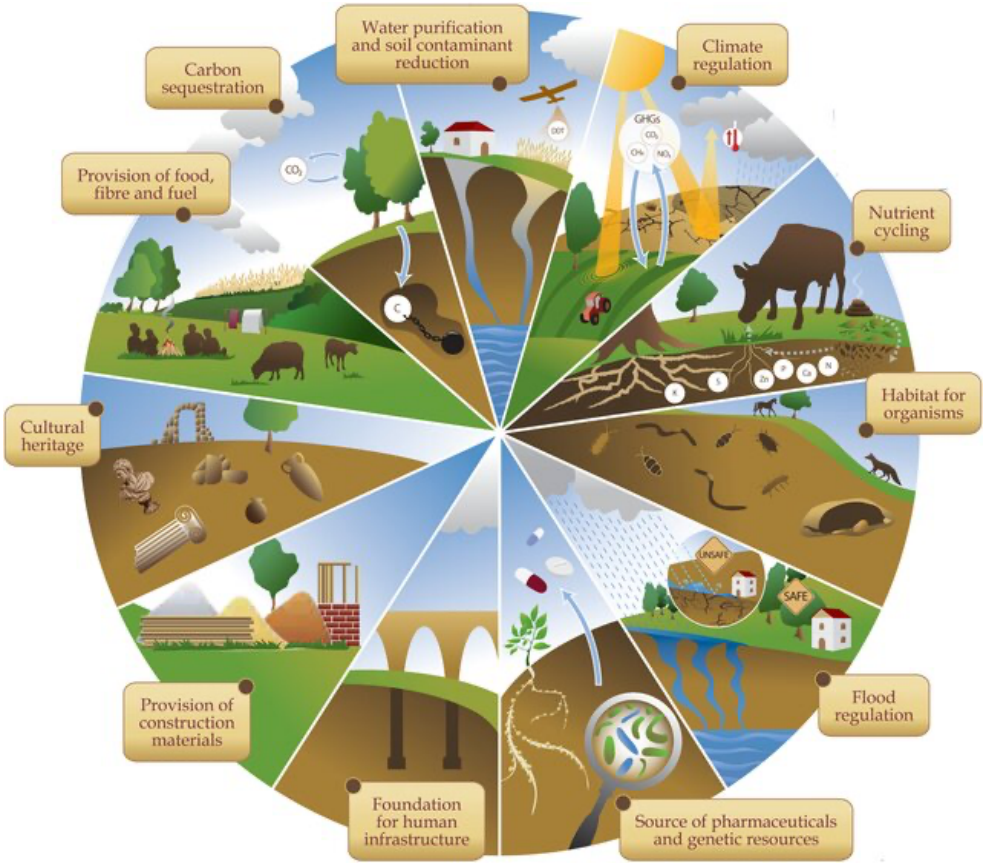


Figure 1: Schematic diagram of soil functions, (Baveye et al., 2020).

Soil health is a critical factor that directly influences productivity (T. Yang et al., 2020). The term "soil health" refers to the soil's capacity to function as a living ecosystem, maintaining its biological, chemical, and physical properties to support the productivity of plants and animals, while also improving air and water quality and promoting biodiversity (Lehmann et al., 2020). Healthy soil not only ensures an efficient supply of nutrients and water but also mitigates environmental impacts, reduces greenhouse gas emissions, and supports sustainable agricultural systems. However, intensive agricultural practices, such as the excessive use of fertilizers and heavy machinery, often negatively impact these functions, leading to issues like compaction, loss of organic matter, and erosion (Gavrilescu, 2021; Kibblewhite et al., 2007). A fundamental pillar of soil health is its biological component, which includes biodiversity and microbial activity. Microorganisms are essential for soil processes such as nutrient cycling, organic matter decomposition, and soil aggregate stabilization (Condron et al., 2010). Microbial communities, among other factors, acts as an indicator of soil resilience and its ability to recover from disturbances (Philippot et al., 2021). Soil degradation, caused by factors such as erosion, salinization, or contamination, not only affects its physical and chemical properties but also leads to a loss of microbial biodiversity (Haj-Amor et al., 2022). This decline in microbial diversity poses a threat to the long-term sustainability of agriculture (Haj-Amor et al., 2022; Hossain et al., 2020). For instance, a reduction in the abundance and diversity of nitrogen-fixing bacteria or phosphorus-solubilizing microorganisms affects nutrient availability for crops, exacerbating challenges to maintaining optimal productivity. In this context, the study of soil microbiomes becomes especially relevant, as they not only constitute an essential biological component of the soil but

are also responsible for many of the processes that determine its sustainability and functionality. Understanding the role of microorganisms in soil is key to addressing current challenges in soil degradation and ensuring long-term sustainable agricultural and forest systems.

2. The importance of the soil microbiome

Soil microbial biodiversity is an essential component of agricultural ecosystems, playing a critical role in the sustainability and functionality of these systems. This term refers to the variety and complexity of microorganisms present in the soil, including bacteria, fungi, archaea, viruses, microscopic algae and protozoa. These microorganisms not only coexist but also interact with one another and with plants, forming a network of symbiotic and competitive relationships that sustain fundamental biological processes vital to terrestrial life (Figure 2) (Hartmann & Six, 2023).

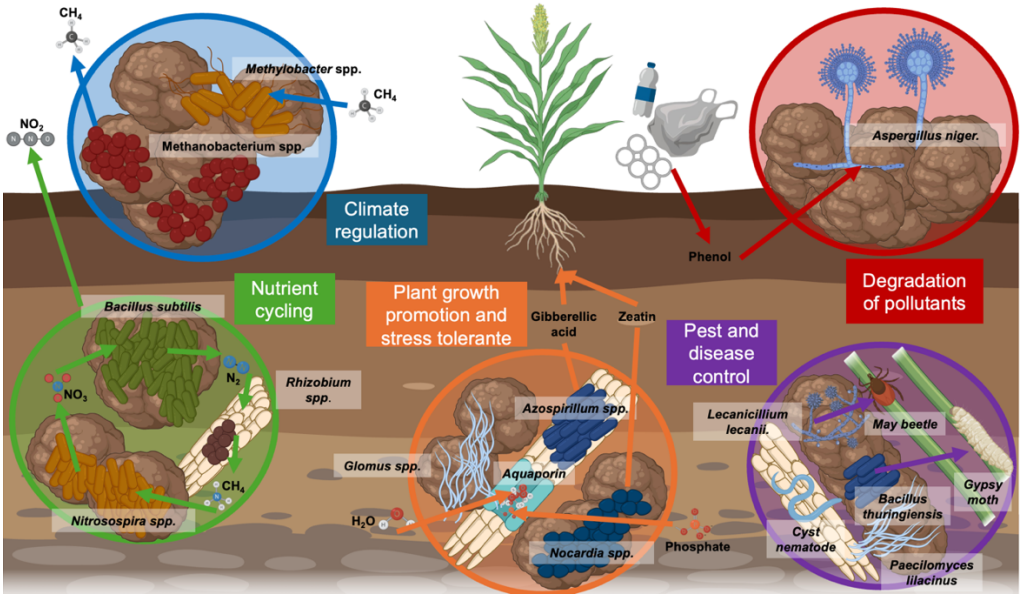


Figure 2: Microbial key functions in the plant-soil system. Image created at app.biorender.com.

One of the most significant contributions of soil microorganisms is their involvement in nutrient cycling, a process essential for maintaining soil fertility. These microorganisms decompose organic matter, releasing key nutrients such as nitrogen, phosphorus, and carbon-rich compounds that are subsequently reused by plants and microbes. This process ensures a constant nutrient dynamic within natural and agricultural ecosystems, thereby promoting their sustainability and productivity (Bhowmik et al., 2017; Jacoby et al., 2017). Without soil microbiomes, soil biogeochemical cycles would be severely disrupted. In addition to their chemical impact, soil microorganisms enhance the physical properties of soil. Microbial exudates, including numerous metabolites such as sugars, amino acids and organic acids, contribute to the formation of soil aggregates, which improve soil structure and promote water retention and aeration (Bronick & Lal, 2005; Costa et al., 2018) and to soil regeneration. These improvements benefit roots development and, consequently, plant growth (Gabasawa et al., 2024). On a biological level, microbial biodiversity also

regulates pathogen populations and promotes plant growth through the production of phytohormones and antimicrobial compounds (Ortíz-Castro et al., 2009).

3. The phosphorus cycle and the contribution of microorganisms to its dynamics

Phosphorus is an indispensable macronutrient for plants, playing a crucial role in key metabolic processes such as photosynthesis, energy transfer, and nucleic acid synthesis. However, its availability in soils is extremely limited, as the majority is present in insoluble forms or chemically bound to minerals and organic compounds that are inaccessible to plants (García-Díaz et al., 2024). Moreover, natural sources of phosphorus, such as phosphate rocks, are non-renewable resources that are being rapidly depleted due to intensive extraction for fertilizer production. This issue is further compounded by European regulations that restrict the exploitation of natural phosphates to reduce associated environmental impacts, thereby encouraging the transition to more sustainable strategies such as recycling and improving phosphorus use efficiency (Bastida et al., 2023; Brownlie et al., 2021; García-Díaz et al., 2024).

Microorganisms play a critical role in the phosphorus cycle by mediating its mineralization, solubilization, and transport (J. A. Siles et al., 2022). The process of phosphorus mineralization is primarily carried out by bacteria and fungi, which decompose phosphorus-rich organic compounds, such as phytates, proteins, and nucleic acids, through the secretion of phosphatase enzymes (D. L. Jones & Oburger, 2011; Ramos Cabrera et al., 2024). These enzymes act on organic forms of phosphorus, converting them into inorganic forms, such as phosphates ($H_2PO_4^-$ and HPO_4^{2-}) (Figure 3), which plants can assimilate (Nannipieri et al., 2011). These processes are regulated by genes such as *phoA*, *phoD*, *phoX* and *phyA* (alkaline phosphatase, alkaline phosphodiesterase, calcium-dependent alkaline phosphatase, and acid phosphatase/phytase, respectively), which encode enzymes responsible for hydrolyzing phosphate esters and phytates, releasing plant-available inorganic phosphate (Garaycochea et al., 2023). However, in soils, phosphorus is often present in insoluble forms, such as calcium, iron, and aluminum phosphates, which are not directly accessible to plants (Barroso & Nahas, 2005). In this context, bacteria and archaea play an essential role in phosphorus solubilization, facilitating its release. Bacteria such as certain *Pseudomonas*, *Bacillus*, and *Enterobacter* species, along with some archaeal species, secrete organic acids such as citric acid and acetic acid, which lower soil pH and solubilize insoluble phosphate compounds, thereby releasing phosphorus in a form that plant roots can absorb (Ayangbenro & Babalola, 2021). This process is regulated by genes such as *gcd* and *pqqC (*glucose dehydrogenase and pyrroloquinoline quinone synthase C*)*, which are involved in the synthesis of acids such as gluconic acid, which is crucial for mobilizing insoluble phosphorus (L. Pan & Cai, 2023). Through this process, bacteria and archaea enhance phosphorus availability for plants, particularly in soils with limited access to this nutrient (Richardson & Simpson, 2011). Additionally, mycorrhizal fungi also contribute to phosphorus solubilization. Through their hyphal networks, these fungi secrete organic acids that dissolve phosphate minerals, further increasing phosphorus availability in the soil (Andrino et al., 2021).

In the context of phosphorus transport, phosphorus-solubilizing bacteria play a central role by colonizing the rhizosphere—the soil zone surrounding plant roots—and employing active mechanisms to transfer phosphorus from the soil to the vicinity of plant roots (Rawat et al., 2021). This process is regulated by genes such as *pstSCAB* and *phoU* (phosphate-specific transport system and PhoU negative regulator), which encode a high-affinity transport system for phosphates, and by genes such as *phoR* and *phoB* (PhoR histidine kinase sensor and PhoB response regulator), which adjust the expression of transport genes depending on the availability of phosphorus in the environment (Lubin et al., 2015). This integrated system of phosphorus mineralization, solubilization, and transport, mediated by bacteria and archaea, not only enhances phosphorus availability in soils but also contributes to the sustainability of agroecosystems by reducing the need for external inputs such as chemical fertilizers (Bargaz et al., 2018).

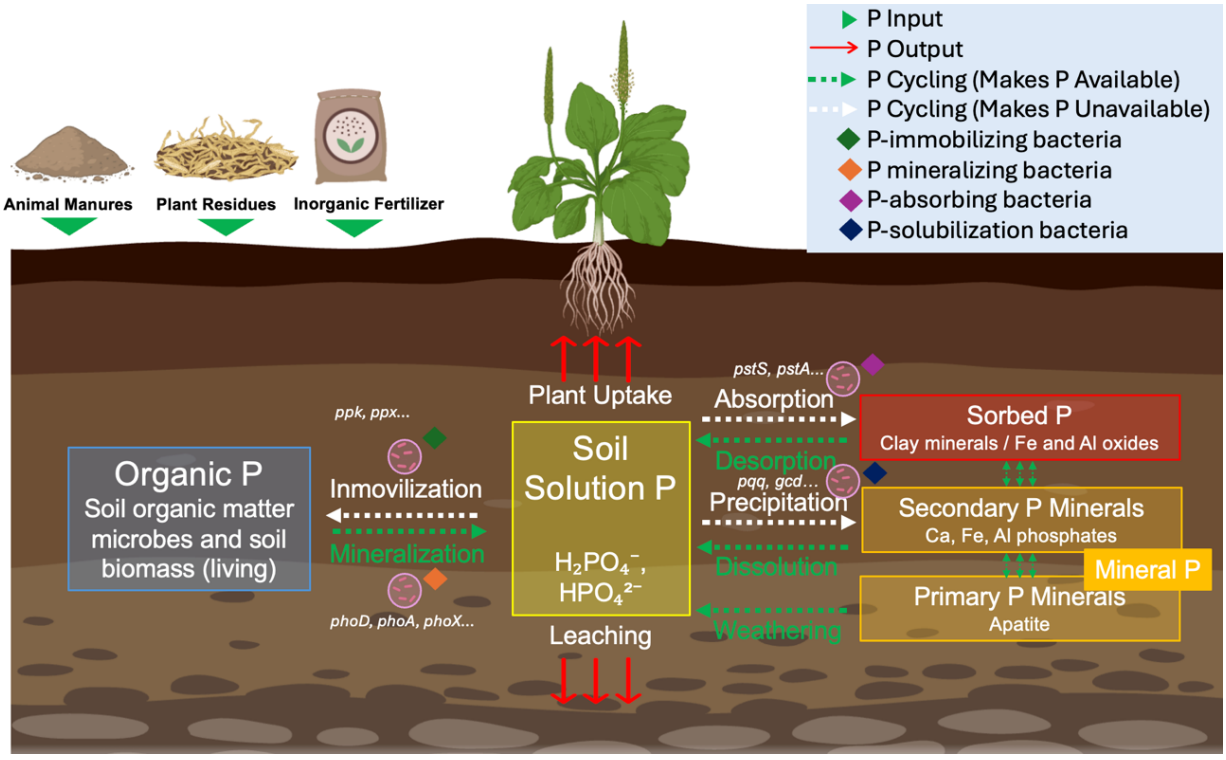


Figure 3: Phosphorus cycle in the soil. This figure illustrates the sources of phosphorus input into the soil, the pathways through which phosphorus becomes available or unavailable for plant uptake, and the pathways of phosphorus loss or removal. It also includes the role of microorganisms in the key processes of the phosphorus cycle and some examples of the genes involved in each of these processes. Image created at app.biorender.com.

4. The nitrogen cycle and the contribution of microorganisms to its dynamics

Nitrogen is an essential nutrient for the growth and development of plants, as it is a key component of proteins, nucleic acids, and other vital compounds (Bastida et al., 2009a). However, despite nitrogen making up approximately 78% of the Earth's atmosphere in the form of nitrogen gas (N_2), plants cannot utilize it directly, making nitrogen one of the most limiting nutrients for plant growth (Cocking, 2000).

Consequently, it is crucial for nitrogen to undergo a series of transformations to become available in forms that plants can absorb and use. This process is governed by a series of chemical transformations in the nitrogen cycle, in which microorganisms play a fundamental role.

In the nitrogen cycle, nitrogen-fixing bacteria, such as those of the *Rhizobium* genus, establish symbiosis with the roots of leguminous plants, forming root nodules where atmospheric nitrogen (N_2) is converted into ammonium (NH_4^+) through the enzyme nitrogenase (Brochado et al., 2023). In addition to these root-associated nitrogen fixers, free-living nitrogen-fixing bacteria, which are not associated with plant roots, also play a critical role in the nitrogen cycle. These bacteria, including genera such as *Azotobacter* and *Clostridium*, are capable of fixing atmospheric nitrogen in soil and aquatic environments. Free-living nitrogen fixers contribute to nitrogen availability in ecosystems by converting N_2 into forms that can be utilized by plants and other organisms (Reed et al., 2011). This process is mediated by genes such as *nifH*, *nifD*, and *nifK*, which encode the structural components of nitrogenase, the enzyme responsible for nitrogen fixation (Dos Santos et al., 2012). Additionally, regulatory genes like *nifA* play a crucial role in controlling the expression of nitrogenase in response to environmental conditions, such as oxygen levels and nitrogen availability (Martinez-Argudo et al., 2004). This ammonium is then incorporated into the soil, where it can be directly utilized by plants or further transformed by other microorganisms into more accessible forms.

Nitrifying microorganisms, including ammonia-oxidizing bacteria (AOB), ammonia-oxidizing archaea (AOA), and complete ammonia oxidizers (comammox), play a central role in the nitrification process. These microorganisms oxidize ammonium (NH_4^+) into nitrites (NO_2^-) and subsequently into nitrates (NO_3^-), which plants absorb as a primary nitrogen source (Fenice, 2021; Norton, 2008). The first step of nitrification, the oxidation of ammonium to nitrite, is primarily carried out by ammonia-oxidizing bacteria (AOB) and archaea (AOA) through the enzyme ammonia monooxygenase (AMO), which is encoded by genes such as *amoA*, *amoB*, and *amoC* (Martikainen, 2022). In the second step, nitrite is oxidized to nitrate by the enzyme nitrite oxidoreductase, encoded by genes like *nxrA* and *nxrB*, which are crucial for sustaining nitrification in both bacterial and archaeal communities (Pester et al., 2014). Moreover, complete ammonia oxidizers (comammox) can perform both steps of the nitrification process, from ammonium to nitrate, highlighting the diverse metabolic pathways involved. These genes are highly conserved among nitrifying microorganisms and are essential for maintaining nitrogen cycling in agricultural and natural ecosystems (Qin et al., 2024; Stein & Klotz, 2016).

On the other hand, denitrifying bacteria, such as *Pseudomonas* and *Clostridium species*, play a crucial role in returning nitrogen to the atmosphere in the form of nitrogen gas (N_2) or nitrous oxide (N_2O), thus closing the cycle and regulating nitrogen concentrations in the soil (Figure 4). Denitrification involves a series of enzymes encoded by specific genes. The reduction of nitrates to nitrites is mediated by nitrate reductase, encoded by the genes *narG*, *narH*, and *narI (Hamada & Soliman, 2023)*. Nitrite is further reduced to nitric oxide (NO) by nitrite reductase, encoded by *nirS* or *nirK* (Pold et al., 2024). The final

steps, the reduction of nitric oxide to nitrous oxide and nitrogen gas, are mediated by nitric oxide reductase (*norB*) and nitrous oxide reductase (*nosZ*), respectively (Torres et al., 2016). These genes are tightly regulated in response to environmental oxygen levels and nitrate availability.

In addition to these processes, ammonification, which converts organic nitrogen compounds into ammonium, is another critical step in the nitrogen cycle. This process is facilitated by enzymes such as glutamate dehydrogenase, encoded by the *gdh* gene, and urease, encoded by *ureC*, which hydrolyze organic nitrogen compounds into ammonium (Jin, 2017). Other genes, such as *asnB* (asparagine synthetase B), play a role in the breakdown of nitrogen-containing organic molecules like asparagine. Regarding nitrogen assimilation, where plants and microorganisms incorporate inorganic nitrogen into organic molecules, genes such as *nasA*, *nirB*, and *nirD* (R. Hu et al., 2022) encoding nitrate and nitrite reductases, are mainly involved in the reduction of nitrates and nitrites into ammonium for incorporation into amino acids and other organic compounds.

The nitrogen cycle is complex and includes a variety of transformations beyond those discussed. Recent studies have highlighted additional pathways such as Anammox (anaerobic ammonia oxidation) and DNRA (dissimilatory nitrate reduction to ammonium), which play important roles under specific environmental conditions. For example, anammox occurs in waterlogged soils like rice paddies where oxygen is limited, while DNRA dominates in carbon-rich anaerobic environments such as fertilized agricultural soils, helping retain nitrogen as ammonium. These processes influence nitrogen availability and greenhouse gas emissions in terrestrial ecosystems (Hamada & Soliman, 2023; P. Wu et al., 2022). Furthermore, the discovery of complete ammonia oxidizers (comammox) has expanded our understanding of nitrification, as these microorganisms can perform both the oxidation of ammonium to nitrite and the subsequent oxidation of nitrite to nitrate (Qin et al., 2024; Stein & Klotz, 2016). These transformations are not only essential for plant nutrition but also regulate environmental contamination by minimizing the accumulation of reactive nitrogen species, such as nitrates, which can leach into groundwater and contribute to issues such as eutrophication (Khan & Mohammad, 2014).

5. The carbon cycle and the role of microorganisms in its dynamics

Carbon is another essential element for plants and represents a fundamental component of organic matter, ranging from carbohydrates to proteins and nucleic acids (Paul, 2016). Indeed, carbon forms the basis of all biomass and is the primary constituent of living organisms (Senesi & Loffredo, 1998). While carbon in the form of CO_2 is directly available to plants through photosynthesis, the transformation of carbon into organic forms via biological and biogeochemical processes is essential to sustain its cycle and availability within ecosystems (Cole et al., 2021). The carbon cycle is the process by which carbon is exchanged among the atmosphere, living organisms, and soil, regulating the availability of this element in the biosphere (Cole et al., 2021).

This cycle is critical for maintaining ecosystem balance, as carbon is a key element for life, but it can also become a challenge when accumulated in excess, as observed with the rising levels of atmospheric CO_2 caused by the burning of fossil fuels (Fu et al., 2022).

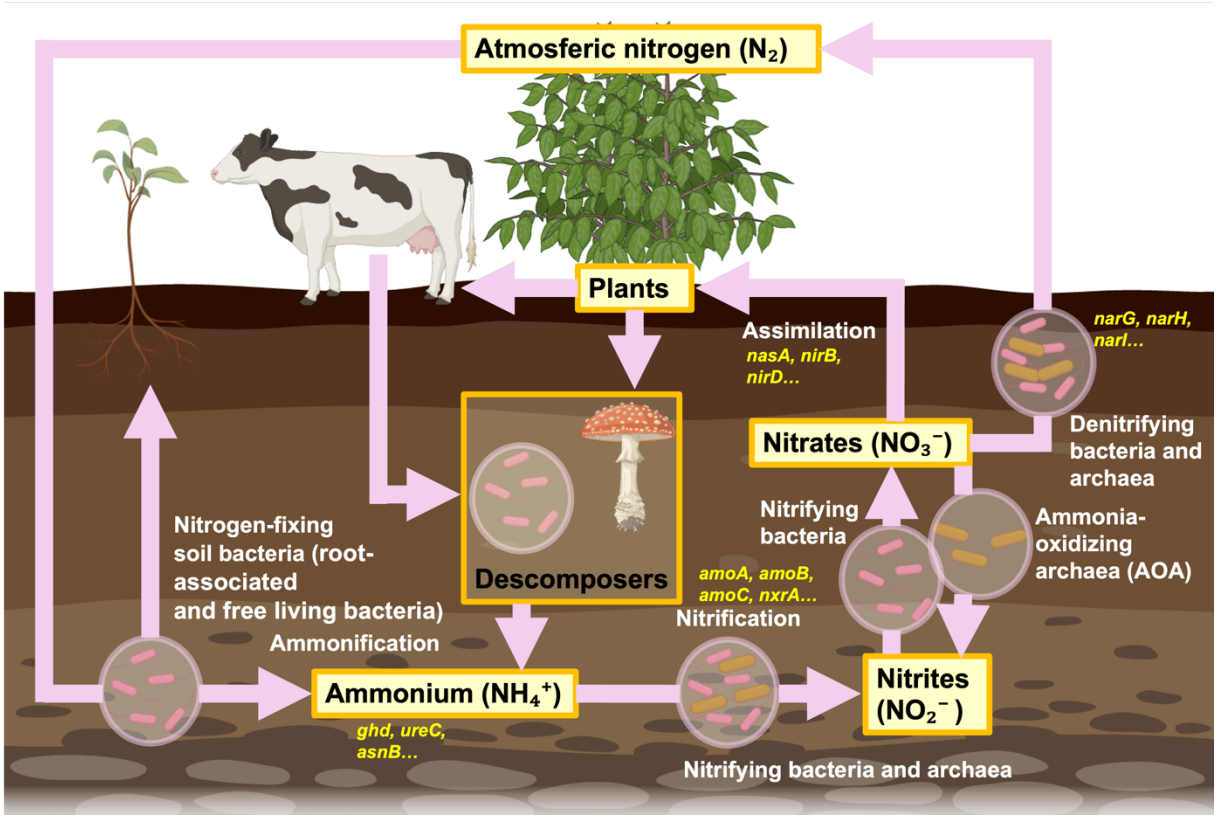


Figure 4: Nitrogen cycle in the soil. This figure illustrates the main nitrogen transformations in the soil and their interactions with plants, animals, and microorganisms. It includes processes such as nitrogen fixation from the atmosphere by bacteria in soil, ammonification by decomposers, nitrification of ammonium (NH_4^+) into nitrites (NO_2^-) and nitrates (NO_3^-), and the assimilation of these compounds by plants. Denitrification is also depicted, where bacteria and archaea convert nitrates into gaseous nitrogen (N_2), completing the cycle. It also includes some genes involved in each of these processes. Image created at app.biorender.com.

The carbon cycle comprises several stages that involve both biological and geological processes. Carbon is absorbed by plants through photosynthesis, a process in which plants take up CO_2 from the atmosphere and, using solar energy, convert it into organic compounds, primarily sugars, which serve as the foundation of the food chain (Janssen et al., 2014). These organic compounds are transferred to primary consumers (herbivores), which are then consumed by carnivores, continuing through the trophic chain. When organisms die, the carbon contained in their biomass is returned to the environment as CO_2, primarily through the decomposition of organic matter by decomposer microorganisms such as bacteria and fungi (Condron et al., 2010). These microorganisms play a crucial role in carbon mineralization, transforming complex organic compounds into CO_2 and other minor products (Horwath, 2007). Additionally, carbon can be stored long-term in the soil as organic matter through biological carbon capture, forming what is known as carbon sinks (Farrelly et al., 2013) (Figure 5). Some of the carbon in the soil is converted into long-term organic carbon, which remains stored for centuries, contributing to the regulation of atmospheric CO_2 levels (Eglin et al., 2010). Processes such as humus formation and the accumulation of soil organic matter,

also referred to as carbon sequestration, help mitigate the effects of climate change by reducing atmospheric CO_2 concentrations (Gerke, 2022).

Microorganisms play an essential role in the carbon cycle, not only in the decomposition of organic matter but also in the processes of carbon fixation and release. During decomposition, bacteria and fungi break down organic residues and release CO_2 into the atmosphere, a process known as microbial respiration (Abatenh et al., 2018). Furthermore, some microorganisms can fix carbon into organic forms through biosynthesis processes, thereby contributing to biomass production.

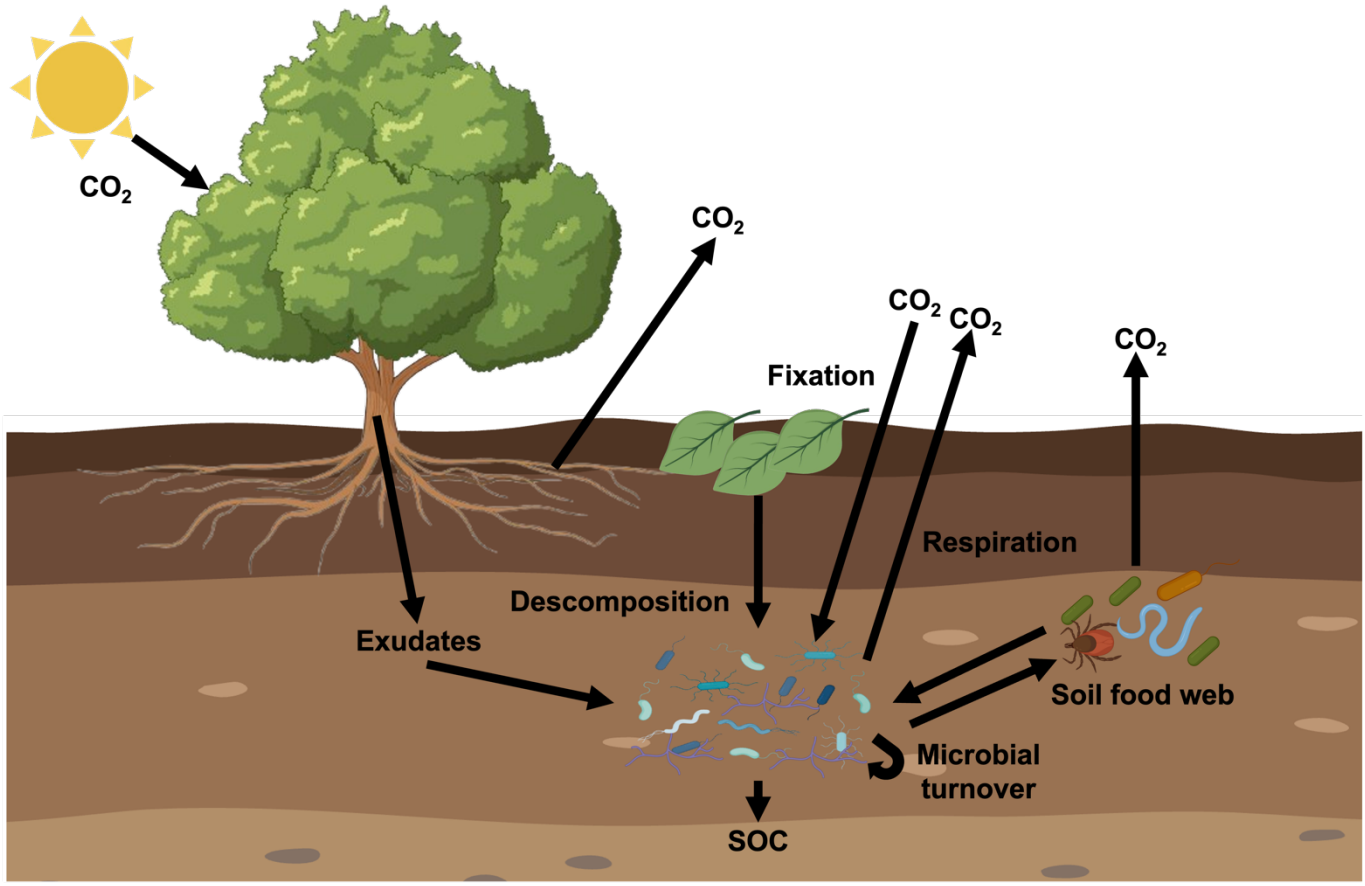


Figure 5: Carbon cycle in the soil. It shows CO_2 absorption by trees during photosynthesis, the release of root exudates, and the decomposition of leaf litter by microorganisms. These processes contribute to the formation of Soil Organic Carbon (SOC), microbial respiration, and microbial turnover. The figure also highlights the interactions within the soil food web, where carbon is exchanged between microorganisms and other soil organisms. Image created at app.biorender.com.

One of the primary tools utilized by microorganisms to break down complex organic materials is a specialized group of enzymes known as CAZymes (Carbohydrate-Active enzymes). These enzymes are crucial to the carbon cycle due to their ability to degrade, modify, and synthesize carbohydrates, enabling the decomposition of organic matter as well as the storage or release of carbon (López-Mondéjar et al., 2022).

CAZymes play a pivotal role in terrestrial ecosystems by facilitating the recycling of carbon in various forms, thereby contributing to the balance of this element in the biosphere (Andrade et al., 2017). CAZymes encompass several classes and families of enzymes, each with specific functions and substrates (Lombard et al., 2014), forming a highly coordinated system for breaking down and transforming the diverse components of organic matter.

Glycoside hydrolases (GHs) are a prominent class of CAZymes that hydrolyze glycosidic bonds in carbohydrates. These enzymes are responsible for degrading structural polysaccharides such as cellulose, hemicellulose, and starch, which are abundant in plant material. For example, cellulases hydrolyze cellulose into glucose, while xylanases target hemicellulose, breaking it down into xylose (Sime et al., 2024). Auxiliary activities (AAs) represent another essential class of CAZymes. This group includes oxidoreductases such as laccases and peroxidases, which act on lignin and other complex plant polymers. These enzymes play a pivotal role in oxidizing or modifying highly recalcitrant compounds, enabling their breakdown into smaller, more accessible molecules that can be recycled within the ecosystem (Chirania et al., 2022). Complementing these activities, carbohydrate esterases (CEs) remove ester groups from polysaccharides, thereby increasing their accessibility for enzymatic attack. For instance, acetylxylan esterases and pectin esterases specifically target acetylated xylan and pectin, respectively (Armendáriz-Ruiz et al., 2018). Polysaccharide lyases (PLs) are another key group of CAZymes, responsible for cleaving polysaccharides such as pectin and alginate via non-hydrolytic mechanisms. This action facilitates the breakdown of complex plant residues into smaller components (Q. Lyu et al., 2018). Glycosyltransferases (GTs) are enzymes that catalyze the transfer of sugar moieties to form glycosidic bonds, contributing to the biosynthesis of structural polysaccharides like cellulose and chitin (Guidi et al., 2023) (Figure 6). Finally, another important group of CAZymes are the carbohydrate-binding modules (CBMs), which are non-catalytic domains that bind specifically to carbohydrates such as cellulose or chitin. By stabilizing the interaction between the enzyme and its substrate, CBMs enhance enzymatic efficiency and facilitate the degradation process (Q. Shi et al., 2023). Together, these families of CAZymes form an integrated enzymatic network that allows microorganisms to degrade, modify, and recycle the diverse array of carbohydrates present in organic matter. This process not only supports the decomposition of plant and microbial residues but also contributes to carbon recycling and sequestration within ecosystems.

In soils, carbohydrate-active enzymes (CAZymes) play a crucial role in the decomposition of organic matter, facilitating carbon recycling and soil sustainability. In forests soils in particular, where there is a high accumulation of organic matter from plant origin, these enzymes are actively involved in breaking down complex plant residues, such as leaves, branches, and roots, through the action of glycoside hydrolases, such as cellulases and xylanases, which degrade structural components like cellulose and xylans. This process releases simple sugars that are subsequently utilized by soil microorganisms as an energy source (Algora et al., 2022). Additionally, lignin-modifying enzymes, including laccases, oxidases and peroxidases, are essential for the degradation of lignin, a highly recalcitrant polymer found in plant

cell walls. This degradation allows the carbon trapped in lignin to be released and reintegrated into the biogeochemical cycle (Algora et al., 2022).

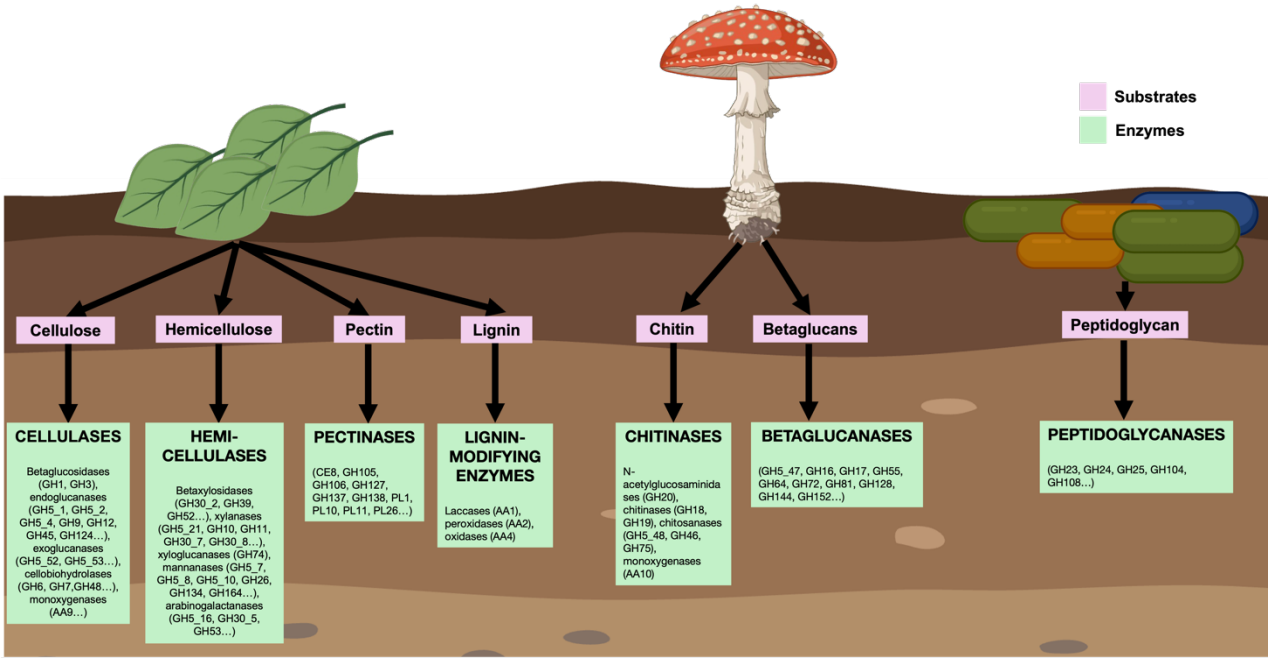


Figure 6: Enzymatic degradation of plant and microbial polymers in soil. This figure highlights the breakdown of key plant components—cellulose, hemicelluloses, pectin, and lignin—by specific enzyme families. Additionally, it includes the degradation of fungal-derived polymers (chitin and β-glucans) and bacterial-derived peptidoglycan. The figure identifies major enzyme activities involved in these processes, such as cellulases, hemicellulases, pectinases, lignin-modifying enzymes, chitinases, β-glucanases and peptidoglycanases, and also the main CAZyme families including these activities. Image created at app.biorender.com.

Moreover, forest soils also present a high amount of organic matter from microbial origin (dead mycelium from mycorrhizal and saprotrophic fungi and dead bacterial biomass), that is degraded by the action of chitinases, betaglucanases and peptidoglycanases contributing to the recycling of carbon in the ecosystem (López-Mondéjar et al., 2020).Collectively, these enzymatic activities not only break down organic residues but also promote the transformation and utilization of modified structural carbohydrates, as facilitated by transferases and esterases, which in turn support microbial growth and biomass production (López-Mondéjar et al., 2022; Sime et al., 2024). This microbial metabolism is fundamental not only for residue degradation but also for the formation of humus, a key component for soil fertility and long-term carbon storage (L. Chen et al., 2023). In this way, CAZymes act as a critical link between biological and geological processes through carbon recycling while also contributing to primary productivity in forests by returning nutrients to the soil. In the context of climate change, these enzymes gain additional significance, as soil microorganisms, through their enzymatic actions, regulate carbon fluxes by either releasing it or sequestering it in stable forms depending on environmental conditions. This underscores their key role in climate change mitigation and global carbon dynamics (Yuan et al., 2023).

6. Agroecosystems and fertilization: Challenges and opportunities

Modern agriculture is a fundamental pillar in ensuring global food security, with fertilizers playing a critical role in sustaining it. Without fertilizers, agriculture systems would be unable to meet the growing food demands of an ever-increasing global population (Mbene et al., 2023). However, conventional fertilizers, such as those based on nitrogen (N), phosphorus (P), and potassium (K) (NPK), face limitations that threaten their long-term sustainability. On one hand, their production relies on non-renewable resources like phosphate rock; on the other, their manufacturing and transportation entail significant economic costs, making them less accessible to small-scale farmers (Chojnacka et al., 2019; Cordell et al., 2009). Beyond economic and availability challenges, the intensive use of chemical fertilizers has led to severe environmental and ecological issues (Chandini et al., 2019). Excessive application has resulted in soil degradation, affecting not only its physical and chemical properties but also its biological component: the soil microbiome (Chandini et al., 2019; Hartmann & Six, 2023). These microbial communities, essential for processes such as nutrient cycling (i.e., N and P), organic matter decomposition, and soil aggregate stabilization, are disrupted by the accumulation of salts, soil acidification, and nutrient leaching, which diminish both their biodiversity and functionality (Rath & Rousk, 2015; Silva et al., 2022). Therefore, while conventional fertilizers remain essential for agricultural productivity, there is an urgent need to identify alternative sources that are economically viable, environmentally responsible, and have a positive or neutral impact on the functionality of soil microbiomes. In this sense, the application of organo-mineral fertilizers emerges as a sustainable and effective alternative to address the challenges posed by soil degradation while simultaneously enhancing soil functionality (García-Díaz et al., 2024a). These fertilizers combine the benefits of organic matter—such as improved soil structure, water retention, and microbial stimulation—with the targeted nutrient supply provided by mineral fertilizers, thereby boosting microbial activity. Furthermore, practices such as reduced tillage, the use of cover crops, and the integration of compost and biochar can complement these efforts by promoting a more balanced and resilient soil microbiome (Lehmann et al., 2020).

These fertilizers combine organic materials, such as plant residues, composted manure, or sewage sludge, with inorganic nutrient sources or minerals like struvite, thereby efficiently providing essential elements like phosphorus. Fertilizers derived from sludge integrate nutrients recovered from waste generated during wastewater treatment processes (García-Díaz et al., 2024; Solon et al., 2019). When applied to soil, sludge not only enhances phosphorus recycling but also increases organic carbon and nitrogen content, as well as the biomass and activity of beneficial soil microbiota, ultimately improving soil fertility and functionality (Bastida et al., 2008). However, the use of sewage sludge presents certain challenges that necessitate regulatory oversight. The Council Directive of June 12, 1986 (86/278/EEC), aimed at protecting the environment—particularly soil safety in agricultural applications of sewage sludge—addresses these concerns. Sewage sludge often contains elevated levels of heavy metals and pathogens, which may pose risks to human, animal, and plant health (García-Díaz et al., 2024a; Usman et al., 2012). Therefore, it is essential to subject sludge to treatment processes that ensure safe utilization under appropriate conditions. Treatments such as thermal stabilization and anaerobic digestion are

effective in significantly reducing pathogen levels (Frost et al., 2022; García-Díaz et al., 2024a; Shao et al., 2019).

Struvite ($MgNH_4PO_4 \cdot 6H_2O$) is another promising alternative, consisting of a magnesium ammonium phosphate salt recovered through precipitation and crystallization processes from waste water and sewage sludge (Ruiz-Navarro et al., 2023). It plays a crucial role in the phosphorus cycle by enhancing the availability of this essential nutrient to plants more efficiently than conventional fertilizers (Krishnamoorthy et al., 2021). As a sustainable source of phosphorus, struvite reduces dependence on non-renewable reserves such as phosphate rock (Bastida et al., 2023; Ruiz-Navarro et al., 2023). Phosphate rock, a natural source of phosphorus, is becoming increasingly scarce. Global phosphate rock resources are estimated at approximately 40,000 Mt of phosphorus (equivalent to 300,000 Mt of phosphate rock) (Bastida et al., 2023). The production of conventional phosphorus fertilizers derived from phosphate rock faces significant challenges due to the anticipated depletion of this mineral reserve and the fact that many phosphate rock mines are situated in regions of geopolitical instability (Bastida et al., 2023). Additionally, the scarcity of alternative phosphorus sources, coupled with stricter European Union regulations—such as limits on cadmium content in fertilizers—emphasizes the urgent need to develop sustainable phosphorus sources to ensure the continued maintenance of crop production (Bastida et al., 2023; García-Díaz et al., 2024).

Alternative phosphorus fertilizers like sludge and struvite offer advantages due to their gradual phosphorus release, minimizing the risk of leaching and water contamination (Yesigat et al., 2022). Their incorporation into agricultural systems promotes the efficient use of phosphorus, contributing to soil sustainability and the responsible management of finite resources. In this context, mineral and organic fertilizers, as those derived from sludge or struvite present promising alternatives. These approaches minimize the depletion of non-renewable resources, improve soil structure, and contribute to long-term regeneration (García-Díaz et al., 2024a). By promoting microbial diversity and enhancing critical biogeochemical cycles, they offer a more sustainable solution for maintaining soil health and productivity without the negative environmental impacts associated with conventional fertilizers (García-Díaz et al., 2024a). These strategies align with the Sustainable Development Goals (SDGs), particularly those addressing food security and environmental sustainability. By balancing agricultural productivity with resource conservation, they provide a pathway to sustainable farming systems that preserve natural ecosystems while ensuring global food security.

7. Natural soils: influence on the carbon cycle

Natural soils, such as those found in forests, grasslands, and undisturbed ecosystems, play a fundamental role in the global carbon cycle. Unlike agricultural soils, which are subject to intensive management practices, natural soils act as long-term carbon sinks, storing large amounts of organic carbon in the form of stable organic matter and humus (Fageria, 2012; Farrelly et al., 2013; Goh, 2004). Forest soils, for example, are among the largest carbon (C) reservoirs on Earth, playing a critical role in global

biogeochemical cycles. Covering over 40 million km², forests store vast amounts of organic matter, with soils alone holding 44% of the estimated 861 Pg of C in these ecosystems (Y. Pan et al., 2011). Forest soils act as major carbon sinks, receiving tons of litter annually and supporting microbial processes essential for decomposition and nutrient cycling. This storage not only contributes to soil fertility but also plays a crucial role in mitigating climate change by reducing atmospheric CO_2 concentrations.

In natural soils, carbon dynamics are closely linked to interactions among vegetation, microorganisms, and environmental conditions. The decomposition of organic matter, including leaves, branches, and roots, is a key process in which soil microorganisms—such as bacteria, and fungi—break down complex compounds such as cellulose, hemicellulose, and lignin (Bani et al., 2018; Condron et al., 2010). This process releases carbon in the form of CO_2 while transforming a fraction into stable compounds that accumulate in the soil. In addition to recycling essential nutrients for plant growth, this decomposition process contributes to humus formation, a critical component for soil carbon stability (Horwath, 2007).

The biodiversity of forest soils is a determining factor in their capacity to store carbon. A diverse array of plant and microbial species enhances carbon stability by increasing decomposition efficiency and the formation of stable organic compounds (Lange et al., 2015). For instance, mycorrhizal fungi, which establish symbiotic associations with plant roots, not only facilitate nutrient uptake but also contribute to soil aggregate formation, protecting organic carbon from microbial degradation (Frey, 2019). Furthermore, the diversity of decomposer microorganisms, such as saprotrophic bacteria and fungi, ensures efficient breakdown of plant residues, promoting the accumulation of carbon in stable forms (Bani et al., 2018; Condron et al., 2010).

However, global change threatens the stability of natural soils. Human activities such as deforestation, forest management, and land-use changes, along with climate-driven stressors like rising temperatures, prolonged droughts, increased fire frequency, and invasive pests, are altering soil carbon dynamics (Baldrian et al., 2023). These activities not only reduce the capacity of soils to store carbon but also release large amounts of CO_2 into the atmosphere, and risk turning forest soils from carbon sinks into net sources of CO_2, with profound implications for climate feedback loops (Smith et al., 2016). For example, the conversion of forests into agricultural land or pastures drastically reduces soil organic carbon content due to vegetation loss and the disruption of microbial communities that regulate the carbon cycle (Verchot, 2010).

Microbial communities, including bacteria and fungi, are fundamental to forest soil function. They regulate carbon turnover, decomposition, and nutrient availability, shaping ecosystem resilience. Understanding how these microbial-driven processes respond to environmental changes is crucial for predicting the future health of forest soils and their role in global carbon and nutrient cycles. Additionally, studying the biological and ecological mechanisms that regulate the carbon cycle in natural soils provides valuable insights for

developing sustainable management strategies applicable to agricultural and forest systems (Gower, 2003).

8. Omics approaches: A new era in the study of soil microorganisms

The study of microorganisms involved in the phosphorus, nitrogen, and carbon cycles has traditionally been challenging due to the immense complexity and diversity of soil microbial communities. The advent of Meta-omics approaches has revolutionized the way soil microorganisms are studied, offering unprecedented insights into their structure, function, and dynamics (V. Kumar et al., 2021). These cutting-edge analytical tools enable the comprehensive investigation of whole microbial communities and their interactions with the environment, providing critical information about their role in biogeochemical cycles and ecosystem functioning. Advances in next-generation sequencing (NGS), mass spectrometry, and bioinformatics tools capable of analyzing large datasets have driven the rise of meta-omics approaches (Kulski, 2016).

Meta-omics disciplines encompass various methodologies, including genomics, transcriptomics, proteomics, metabolomics, phenomics, and ionomics, each targeting specific molecular layers of biological systems (Figure 7). The term "meta" in meta-omics refers to the study of the collective set of all genes, transcripts, proteins, metabolites, etc., present in a particular sample, rather than focusing on a single organism or genome. For example, metagenomics involves analyzing the entire genetic material present in an environmental sample, providing insights into the diversity and functional potential of microbial communities. This contrasts with traditional omics approaches, which typically examine specific, individual molecular components (e.g., the genome, transcriptome, or proteome) of a single organism or biological system.

By employing approaches such as metagenomics, metatranscriptomics, metaproteomics, and the analysis of metagenome-assembled genomes (MAGs), researchers can explore microbial communities in their entirety, including uncultivable microorganisms that were previously inaccessible (Blakeley-Ruiz et al., 2019). These techniques allow scientists to link microbial taxonomy with functional capabilities, revealing the mechanisms that regulate critical processes like nutrient cycling, organic matter decomposition, and greenhouse gas fluxes (Vailati-Riboni et al., 2017). This integrative view of soil ecosystems is particularly valuable in agriculture, where understanding how microorganisms contribute to nutrient availability, soil health, and sustainability is essential (Van Emon, 2016).

Meta-omics approaches have proven indispensable for addressing key sustainability challenges in agricultural systems, such as soil fertility loss, greenhouse gas emissions, and inefficient fertilizer use (Wallace et al., 2017). By providing detailed insights into the structure and function of soil microbiomes, these techniques bridge the gap between microbial diversity and their functional roles, enabling the design of microbiota-based strategies to enhance productivity while minimizing environmental impacts (Djemiel

et al., 2022). Furthermore, they support the identification of potential bioindicators of soil health, the development of biotechnological solutions like biofertilizers and biopesticides, and the creation of real-time monitoring tools to assess soil quality and functionality.

The integration of taxonomy with function underscores the transformative potential of meta-omics for advancing both scientific understanding and practical applications, marking a new era in soil microbiology.

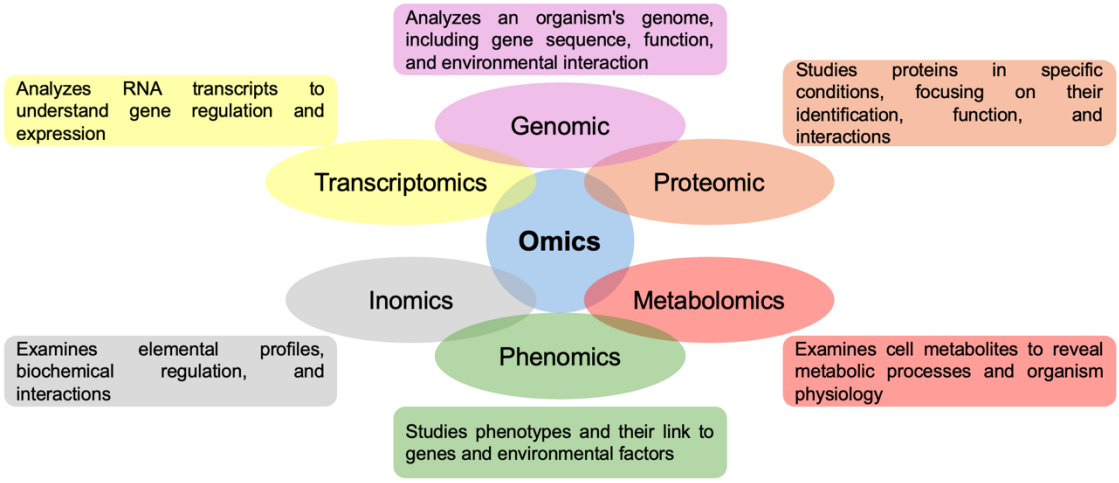


Figure 7: Overview of omics disciplines: This diagram highlights the key omics approaches—genomics, transcriptomics, proteomics, metabolomics, phenomics, and ionomics—and their roles in studying DNA, RNA, proteins, metabolites, phenotypes, and essential ions to understand microbial communities and their functions.

8.1. Metagenomics: Exploring genetic potential

Metagenomics involves the direct extraction and sequencing of DNA from microbial communities present in environmental samples. This approach provides access to the metagenome, which represents the complete set of genetic material within a community, including both cultivable and non-cultivable microorganisms (Garza & Dutilh, 2015). This technique not only identifies which microorganisms are present in a soil sample but also predicts their functional capabilities through the annotation of genes associated with key metabolic processes (Prakash & Taylor, 2012). Metagenomic analysis begins with the collection of environmental samples, such as soil, which harbor diverse microbial communities. Genomic DNA is extracted using specialized protocols to ensure its quality. The DNA is then fragmented and prepared into a library compatible with sequencing technologies such as Illumina (short-reads, where the DNA is fragmented into smaller pieces, typically 2x150bp), PacBio (long-reads, where the DNA is not fragmented but the desired size is selected, typically 15,000-20,000bp), or Oxford Nanopore (long-reads, where the DNA is not fragmented and is passed directly through a nanopore, resulting in ultra-long reads, ranging from 100,000-300,000bp) (Thomas et al., 2012). After sequencing, the raw data undergo bioinformatics processing to transform the billions of short or long reads into meaningful biological information. This includes quality filtering, assembly (using tools such as MEGAHIT or metaSPAdes for short reads), gene prediction (using Prodigal or MetaGeneMark) and taxonomic/functional annotation (using databases such as NCBI, KEGG or COG) (Aramaki et al., 2020; Grünberger et al., 2022; Kanehisa

et al., 2016; Kang et al., 2019; D. Li et al., 2015; Love et al., 2014; Mikheenko et al., 2016; Pruitt et al., 2009; Uritskiy et al., 2018). This metagenomic analysis process is illustrated in Figure 8.

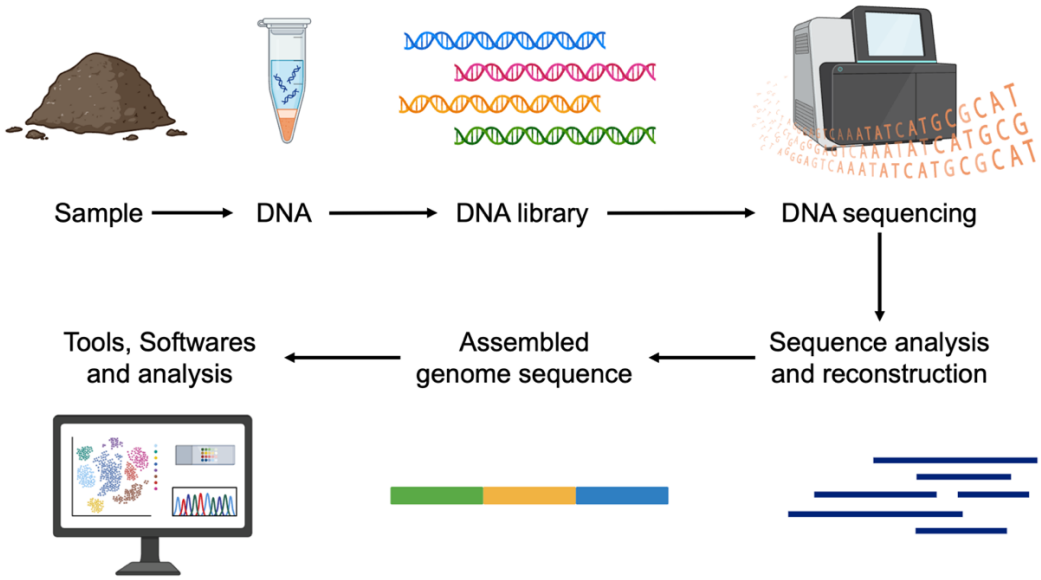


Figure 8: Overview of the metagenomic analysis workflow: From soil sample collection to DNA extraction, followed by library preparation, sequencing, and analysis. The process culminates in the assembly of genomic sequences and the use of specialized tools and software for further analysis and reconstruction of microbial genomes. Image created at app.biorender.com.

In the context of agricultural soils, metagenomics allows the identification of genes linked to nutrient cycles, such those involved in nitrogen fixation (e.g., the *nifH* gene) (Wolińska et al., 2017), and phosphorus mineralization, solubilization or transport (e.g., genes encoding phosphatases) (Liao et al., 2023), and the decomposition of organic matter (e.g., genes coding for cellulases and lignin-modifying enzymes) (C. Wang et al., 2016). Furthermore, it can track genes associated with agriculturally relevant processes, such as pest and disease resistance, or identify microorganisms involved in carbon sequestration within the soil (Jagadesh et al., 2024). Thus, metagenomics enables the construction of functional gene catalogs regulating the carbon, nitrogen, and phosphorus cycles, providing deep insights into how agricultural practices influence the soil microbiota and, consequently, its fertility and sustainability (G. C. Kumar et al., 2021; J. A. Siles et al., 2022).

Compared to other molecular tools like amplicon sequencing or metabarcoding, which target specific marker genes, metagenomics offers significant advantages when studying soil microbial communities. Amplicon sequencing focuses on amplifying and sequencing specific genetic markers, such as the 16S rRNA gene for bacteria and archaea or the internal transcribed spacer (ITS) for eukaryotes, allowing for taxonomic identification. This approach provides high-resolution taxonomic information but is limited in terms of functional diversity and does not offer insights into the metabolic functions of microorganisms (Ramazzotti & Bacci, 2018). This method offers relatively low resolution in terms of functional diversity, as

it can only provide taxonomic information about the microorganisms present in the sample. (M. Liu et al., 2020).

In contrast, metagenomics goes beyond marker gene sequencing by directly sequencing the entire genetic material from environmental samples, without the need for amplification. This approach enables a comprehensive analysis of the genetic and functional potential of all microorganisms, including bacteria, archaea, eukaryotes, and viruses. By sequencing all the DNA in a sample, metagenomics offers deeper insights into both the taxonomic composition and the metabolic functions of microbial communities (Pérez-Cobas et al., 2020).

While marker gene sequencing can infer taxonomic composition, it does not provide direct information on the metabolic functions of microorganisms (Langille et al., 2013). Moreover, amplification biases can occur due to the use of specific primers, potentially limiting the detection of certain microorganisms (Abellan-Schneyder et al., 2021). Metagenomics, by analyzing the entire DNA content, avoids these biases and can detect non-cultivable or low-abundance microorganisms that might be overlooked by 16S rRNA-based techniques (Ramazzotti & Bacci, 2018). Furthermore, metagenomics provides access to the full functional potential of the community by directly identifying genes associated with specific metabolic pathways, such as nitrogen fixation, cellulose degradation, or phosphorus mobilization (De Filippo et al., 2012). This is particularly valuable in agricultural and natural soil studies, as it connects microbial composition with functional activity, revealing how these communities regulate essential processes for soil fertility and the balance of biogeochemical cycles (Liao et al., 2023). Metagenomics also facilitates the study of symbiotic interactions and metabolic networks within microbial communities, which are fundamental for understanding the dynamics influencing soil health and sustainability (Jagadesh et al., 2024). Overall, while amplicon sequencing is useful for obtaining a general taxonomic overview, metagenomics enables a comprehensive analysis that encompasses both microbial diversity and functionality.

8.2. Metatranscriptomics: Analyzing gene expression

Metagenomics provides information about the genetic potential of microbial communities. Metatranscriptomics takes this a step further by analyzing the active transcriptome—that is, the set of RNA molecules transcribed in an environmental sample at a given point in time (Shakya et al., 2019). This is crucial for understanding which genes are actively expressed and how microorganisms respond to environmental stimuli, offering a dynamic perspective on microbial function. Unlike metagenomics, which provides a static view of genetic diversity, metatranscriptomics captures real-time functional activity within microbial communities, linking genetic potential to metabolic activity (A. Kumar & Yadav, 2024).

The metatranscriptomic analysis begins with RNA extraction from environmental samples, such as soil. Due to the labile nature of RNA, this step requires rigorous protocols to preserve its integrity and prevent degradation (Reck et al., 2015). Once extracted, RNA is purified to remove contaminants, including DNA molecules that could interfere with subsequent analyses. The purified RNA is then converted into

complementary DNA (cDNA) through reverse transcription. This cDNA serves as a stable template for sequencing, which is typically performed using high-throughput platforms such as Illumina or Oxford Nanopore (Grünberger et al., 2022; A. Kumar & Yadav, 2024). The obtained sequences undergo bioinformatic processing, which includes quality control, mapping to reference genomes or metagenomes, and functional annotation. This workflow enables the identification of actively transcribed genes and the quantification of their expression levels, providing insights into microbial responses to environmental conditions (Shakya et al., 2019). Figure 9 presents a schematic overview of the entire metatranscriptomic analysis process.

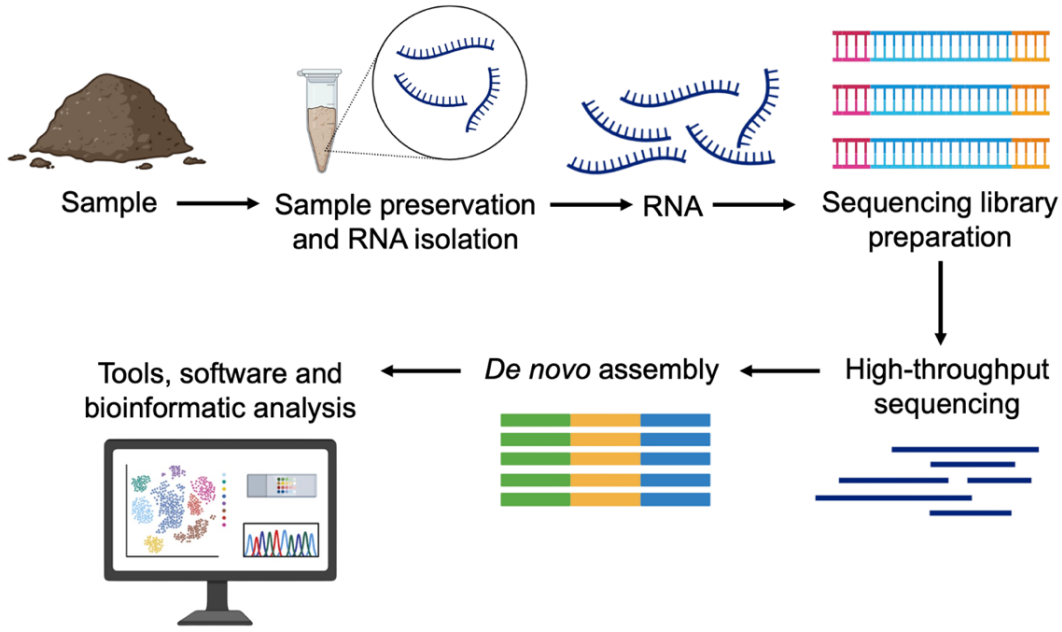


Figure 9: Overview of the metatranscriptomic analysis workflow: From sample collection and preservation, followed by RNA isolation and the preparation of a sequencing library. This is succeeded by high-throughput sequencing, de novo assembly, and bioinformatic processing using specialized tools and software. Image created at app.biorender.com.

Metatranscriptomics plays a fundamental role in elucidating microbial contributions to biogeochemical cycles (A. Kumar & Yadav, 2024). For instance, in the nitrogen cycle, the expression of genes involved in nitrification (e.g., *amoA*, encoding ammonia monooxygenase) and denitrification (e.g., *nirS*, encoding nitrite reductase) can be quantified, offering a snapshot of microbial activity under specific conditions (K. Yu & Zhang, 2012).. In the phosphorus cycle, this technique can detect the expression of genes encoding phosphatases, which mobilize inorganic phosphorus and are often among the most highly expressed in nutrient-limited soils (Xu et al., 2020). Similarly, in the carbon cycle, transcripts encoding cellulases, hemicellulases, and other enzymes involved in the degradation of complex or simple organic matter can be identified (López-Mondéjar et al., 2019). In forest ecosystems, metatranscriptomics has emerged as an essential tool for understanding microbial responses to environmental changes and ecosystem dynamics, such as organic matter decomposition, nutrient cycling, and climate stress (Sharuddin et al., 2022). By linking microbial gene expression to soil conditions, this technique enables the identification of key

biogeochemical processes that sustain soil fertility and contribute to carbon storage. Furthermore, it provides insights into the activity of enzymes involved in the degradation of plant and microbial biomass, helping to reveal the functional roles of specific microorganisms (Žifčáková et al., 2017). Metatranscriptomics also facilitates the detection of bioindicators of soil health, opening new possibilities for adaptive monitoring and the sustainable management of ecosystems (Sharuddin et al., 2022)

While metagenomics remains the most widely used omics approach due to its lower cost and well-established workflows, metatranscriptomics offers unique advantages by directly linking microbial activity to environmental processes. However, its application requires advanced laboratory techniques, bioinformatics expertise, and careful data interpretation, as RNA profiles can vary significantly depending on environmental conditions and sampling timing (Shakya et al., 2019). Furthermore, similar to metaproteomics (see below), this approach represents an additional analytical step that relies on prior metagenomic data, making it even more expensive. Additionally, challenges such as RNA instability, the complexity of soil matrices, and the dominance of ribosomal RNA (rRNA) in total RNA extracts must be addressed to achieve reliable results (R. K. Yadav et al., 2016).

8.3. Metaproteomics: direct functional analysis

As in the case of metatranscriptomics, metaproteomics also takes it a step further than metagenomics by analyzing the proteins expressed in an environmental sample (T. Schneider & Riedel, 2010). This is crucial for understanding which genes are being translated into functional proteins at a given moment, offering a more direct functional perspective. Metaproteomics relies on the analysis of proteins present in the soil using advanced mass spectrometry techniques, such as LC-MS/MS analysis. Figure 10 illustrates the workflow of a metaproteomic analysis, which begins with the extraction of environmental samples, such as soil, followed by protein extraction, where proteins are carefully isolated from the complex matrix. This is typically followed by protein digestion, where enzymes like trypsin are used to break down proteins into smaller peptides, which are more suitable for mass spectrometry analysis (Nebauer et al., 2024). The peptides are then separated using techniques like liquid chromatography (LC), which ensures that the peptides are resolved according to their properties, such as size and charge (Nebauer et al., 2024). Following separation, the peptides are identified and quantified using mass spectrometry (e.g., LC-MS/MS), a powerful technique that measures the mass-to-charge ratio of ions and generates a spectrum that can be used to infer the protein composition of the sample (Nebauer et al., 2024). Finally, the resulting data undergoes bioinformatics analysis, where computational tools are applied to match the peptide spectra with known protein databases, enabling the identification of proteins and the quantification of their relative abundance (Nebauer et al., 2024). These tools allow the identification of specific proteins associated with key functions in biogeochemical cycles (Tartaglia et al., 2020). For instance, in the nitrogen cycle, proteins involved in nitrification (such as ammonia monooxygenase) or denitrification (such as nitrite reductases) can be detected and quantified, providing insight into the metabolic activity of the microorganisms involved (Jose et al., 2020). Similarly, in the carbon cycle, enzymes that degrade complex organic polymers, such as cellulose and lignin, can be identified (Chirania et al., 2022), while in the

phosphorus cycle, phosphatases and other proteins responsible for mobilizing immobilized phosphorus can be tracked (Islam et al., 2024).

While metagenomics offers an accessible and cost-effective approach for studying microbial communities, metaproteomics holds a clear advantage in linking phylogenetic structure with functional activity (Starke et al., 2019). However, metaproteomics requires complex sample extraction protocols—still under development—and sensitive, expensive equipment that is accessible only to a limited number of laboratories worldwide. Additionally, metaproteomics also relies on prior metagenomic analysis for accurate protein annotation (Starke et al., 2019). In the context of agricultural soils, metaproteomics enables the correlation of microbial activity with soil conditions (Bastida & Jehmlich, 2016).

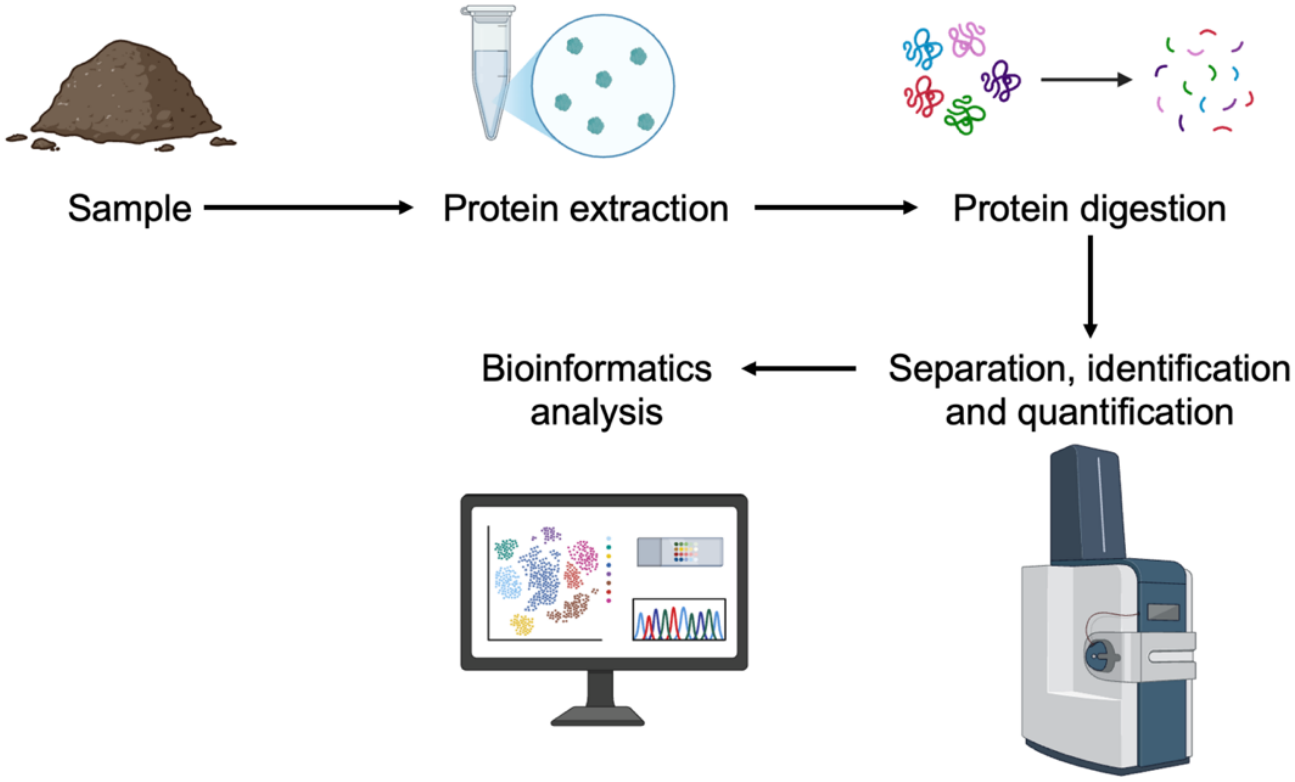


Figure 10: Overview of the metaproteomic analysis workflow: From protein extraction from soil samples to enzymatic digestion, followed by mass spectrometry analysis. The workflow concludes with data processing for protein identification and quantification, bioinformatic mapping of proteins to metabolic pathways. Image created at app.biorender.com.

8.4. Assembly and analysis of MAGs: Reconstruction of microbial genomes

The assembly of MAGs (Metagenome-Assembled Genomes) is an advanced technique that enables the in silico reconstruction of individual bacterial and archaeal genomes from metagenomic data (C. Yang et al., 2021). This process employs bioinformatic algorithms to cluster contigs—assembled DNA fragments—into bins based on sequence composition, coverage, and taxonomic markers. These contigs are then assembled into draft genomes or MAGs, which primarily represent the genetic composition of bacterial and archaeal communities. Reconstruction of prokaryotic genomes is generally more

straightforward due to the absence of introns, which simplifies assembly (Thomas et al., 2012). In contrast, the assembly of eukaryotic genomes poses greater challenges because of their complex genomic architecture, including introns. However, recent advances in long-read sequencing technologies and the expansion of eukaryotic genome databases are improving the annotation of fungal and other eukaryotic sequences, despite persistent limitations (Saraiva et al., 2023). In recent years, this approach has been instrumental in the discovery of novel taxa and previously unrecognized phylogenetic lineages in soil (Ma et al., 2023; Nayfach et al., 2021).

The process begins with high-quality metagenomic sequencing data, which are processed using various bioinformatic tools to filter, assemble, and bin DNA fragments into individual genomes (C. Yang et al., 2021). The resulting MAGs range from partial to nearly complete genome sequences, depending on sequencing depth and computational methods (L.-X. Chen et al., 2020). A major advantage of MAG assembly is its ability to recover genomic information from uncultivable or difficult-to-culture organisms, such as many environmental microbes that play essential roles in biogeochemical cycles (Grossart et al., 2020). Unlike traditional shotgun sequencing, which produces fragmented data that are often difficult to interpret, MAGs provide a more coherent representation of individual microbial genomes, facilitating the association of specific metabolic functions with particular taxa (C. Yang et al., 2021).

In the study of agricultural soils, the analysis of MAGs has been essential for identifying key microorganisms that regulate the cycles of phosphorus, nitrogen, and carbon (X. Hu et al., 2025). For instance, genomes of nitrogen-fixing bacteria, such as those of the genus *Rhizobium*, can be reconstructed. This not only provides information about their metabolic potential but also sheds light on their adaptation to soil conditions, their interactions with other organisms, and their response to agricultural practices (Zhu et al., 2024). Moreover, the ability to reconstruct MAGs allows researchers to delve deeper into the functional attributes of these microorganisms, such as the presence of key enzymes involved in nutrient cycling or their capacity to degrade pollutants (Nagar et al., 2023). Additionally, MAG assembly enables the tracking of the evolution of genes of interest, such as those related to antibiotic resistance or tolerance to environmental stress (Nagar et al., 2023). This is particularly relevant in agricultural soils, where the intensive use of fertilizers and pesticides can impose selective pressure on microbial communities, impacting their diversity and functionality.

OBJETIVES AND HYPHOTESES

OBJECTIVES AND HYPOTHESES

This Doctoral Thesis employs innovative meta-omics approaches to investigate soil microbial communities and their roles in carbon, nitrogen, and phosphorus cycling in natural and agricultural soils. It examines molecular processes such as phosphorus solubilization, nitrogen transformation, and carbon turnover, linking microbial taxonomy with function through bioinformatics. The study highlights key microbial contributors, their ecological roles—particularly in forest soils—and their responses to agricultural practices. The findings support sustainable strategies to enhance soil fertility, reduce nutrient deficiencies, and minimize agriculture's environmental impact, informing advances in biofertilizers and soil health management (Figure 11).

The specific objectives of this PhD Thesis are:

i) *Chapter 1:* To evaluate the responses of microbial communities in agricultural soils to conventional and alternative fertilization strategies, with a particular focus on key processes in the phosphorus (P) cycle, utilizing metagenomics and metaproteomics to investigate community composition and functional dynamics.

ii) *Chapter 2:* To evaluate the responses of microbial communities in agricultural soils to conventional and alternative fertilization strategies, with a particular focus on key processes in the nitrogen (N) cycle, utilizing metagenomics and metaproteomics to investigate community composition and functional dynamics.

iii) *Chapter 3:* To analyze decomposer microbes in forest soils, this study uses metagenomics, metatranscriptomics and Metagenome-Assembled Genomes (MAGs) to identify polymer-specialized guilds, distinguishing generalists from specialists and their carbon cycling roles.

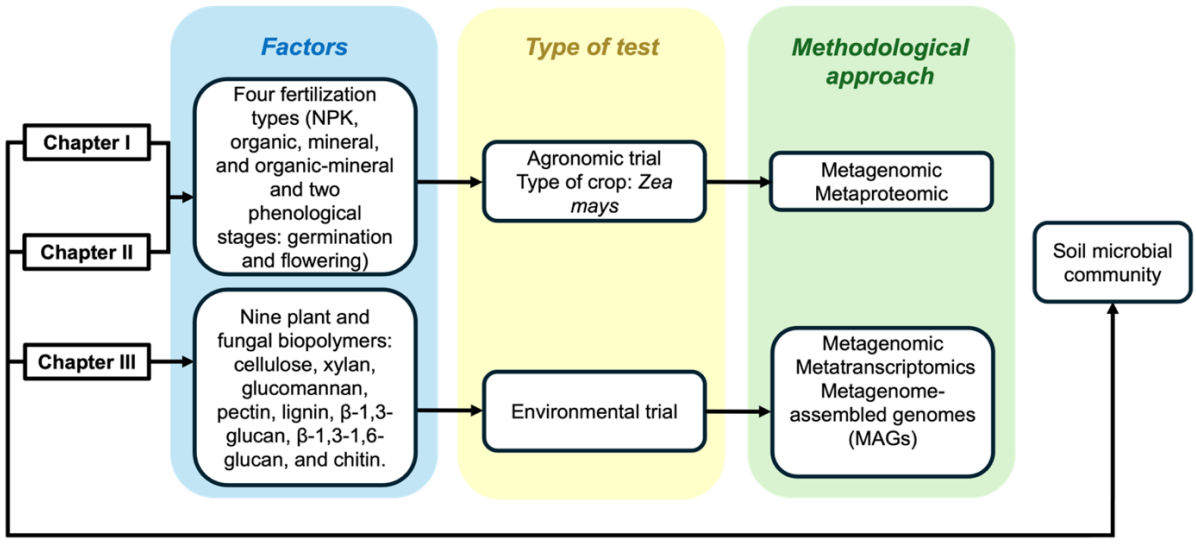
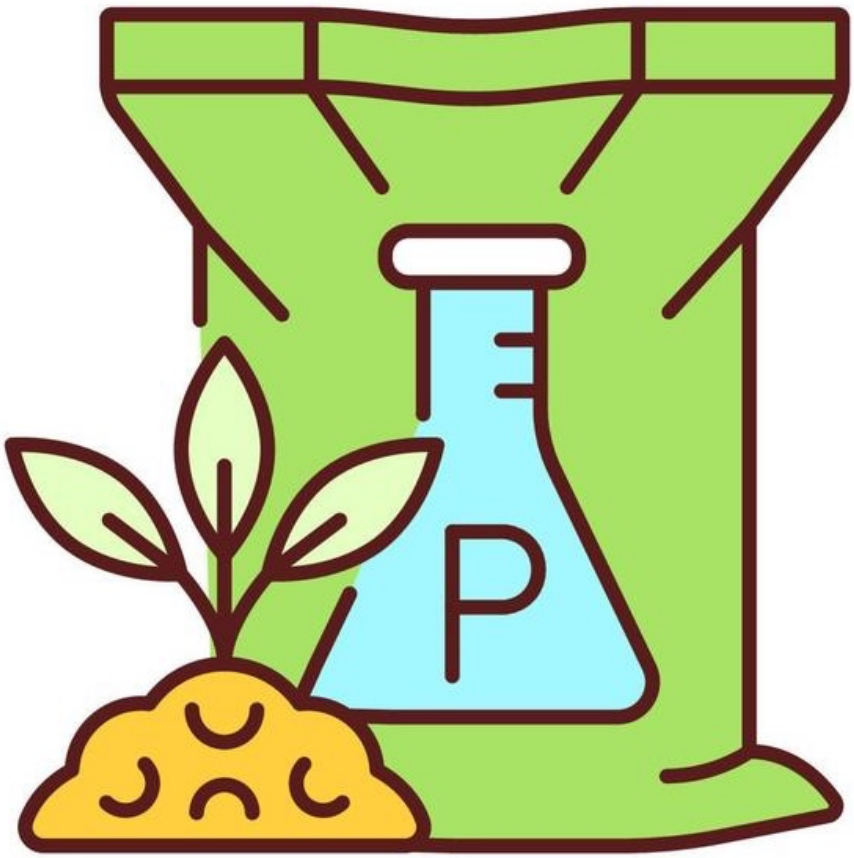


Figure 11: Conceptual figure of the thesis and its chapters.

CHAPTER 1

Contrasting fertilization and phenological stages shape microbial-mediated phosphorus cycling in a maize agroecosystem.



CHAPTER 1

1. INTRODUCTION

Phosphorus (P) is a crucial macronutrient that plays a central role in ecosystem productivity and agricultural yields. Despite its essential function, phosphorus is widely recognized as a limiting nutrient in terrestrial ecosystems (George et al., 2016). To address this limitation, significant amounts of phosphorus are added to agro-ecosystems, predominantly sourced from phosphate rock deposits. However, the scarcity of such resources (Brownlie et al., 2021) and the implementation of stricter regulations within the European Union, such as limits on cadmium content, underscore the urgency of identifying sustainable alternatives to maintain crop productivity. Among these alternatives, P-rich byproducts such as sewage sludge and struvite have garnered attention as potential sources of P, as well as other key nutrients like carbon (C) and nitrogen (N). Struvite, a mineral comprising ammonium and magnesium (Mg) phosphate that can be recovered from wastewater treatment facilities, offers a promising substitute for conventional phosphorus fertilizers (Bastida et al., 2019a). Additionally, struvite provides supplementary nutrients such as Mg and N (Bastida et al., 2019a). Studies have shown that plants fertilized with struvite exhibit improved biomass and phosphorus uptake compared to traditional phosphorus sources (Hertzberger et al., 2020). Similarly, nutrient-rich sewage sludge has demonstrated efficacy comparable to superphosphate in promoting crop production, with the potential to replace soluble phosphate fertilizers (Figueiredo et al., 2021).

The predicted global scarcity of phosphorus fertilizers in the coming decades highlights the necessity of better understanding the microbial processes that regulate soil phosphorus availability (Chowdhury et al., 2017). Although inorganic orthophosphate ions are directly accessible to plants, various soil processes can reduce phosphorus availability. Fresh inorganic phosphorus is only partially absorbed by plants, with the remainder immobilized in forms that are insoluble and inaccessible (J. A. Siles et al., 2022). Organic phosphorus, on the other hand, exists in diverse molecular forms but is often not bioavailable due to its high molecular weight, requiring enzymatic hydrolysis for plant uptake (Kafle et al., 2019). Soil microorganisms play a pivotal role in this process, facilitating inorganic phosphorus solubilization via the release of organic acids and mineralizing organic phosphorus through enzymatic activity. These microbial processes release plant-available phosphorus into the soil. However, phosphorus availability in soil is influenced not only by the chemical properties of fertilizers (e.g., organic vs. mineral) and edaphic factors such as texture, pH, and cation exchange capacity (Kafle et al., 2019; Ruiz-Navarro et al., 2023; J. A. Siles et al., 2022) but also by the plant's growth stage and its interactions with soil microbial communities. Plants exhibit varying phosphorus demands during different phenological stages. In maize, a staple crop of global importance (Soto-Gómez & Pérez-Rodríguez, 2022), phosphorus requirements are especially high during early developmental stages such as germination and flowering (Barry & Miller, 1989). To meet these demands, plants employ dynamic strategies to acquire P, often involving symbiotic relationships with soil microbes (Richardson et al., 2011; J. A. Siles et al., 2022). In response to phosphorus limitation, certain

soil microorganisms activate specific genes that mediate organic phosphorus mineralization, inorganic phosphorus solubilization, phosphorus transport, and adaptation to phosphorus starvation (Dai et al., 2020). For example, microbes can mineralize organic phosphorus through phosphatase enzymes encoded by genes such as *phoA*, *phoD*, and *phoX*, or solubilize inorganic phosphorus using genes like *gcd* and *pqqC* (J. A. Siles et al., 2022).

Organo-mineral fertilization has been shown to influence the abundance of microbial populations carrying genes involved in the phosphorus cycle, including *phoC*, *phoD*, *phnX*, and *gcd* (Wan et al., 2020). Advances in multi-omic technologies, such as metagenomics and metaproteomics, offer powerful tools for exploring these microbial processes in detail (Bastida et al., 2021a; Liang et al., 2020; Miller et al., 2023a; Starke et al., 2019a). These approaches provide insights into the functional roles of soil microbial communities in the phosphorus cycle, allowing researchers to identify and quantify both genes and proteins that are key to phosphorus cycling (Starke et al., 2019a). Importantly, while byproducts like struvite and sewage sludge can influence microbial mechanisms of phosphorus provision by altering the abundance of functional genes and microbial taxa, the plant phenological stage may also play a critical role in shaping these processes. However, the relative contributions of fertilization and plant phenology to the dynamics of soil microbial communities remain poorly understood.

In this study, we aim to investigate the effects of struvite, sewage sludge, and their combination, alongside different phenological stages of maize, on the taxonomic composition and abundance of genes and proteins associated with organic phosphorus mineralization, inorganic phosphorus solubilization, and phosphorus starvation responses. Considering the distinct chemical composition of these fertilizers, with sewage sludge containing a fraction of organic phosphorus not present in struvite, we hypothesize the following: (i) these materials will differentially influence the abundance of genes involved in the phosphorus cycle, as well as the microbial populations harboring these genes; and (ii) the phenological stage of maize will modulate the abundance of specific genes and microbial taxa associated with soil phosphorus cycling. Additionally, the study includes a novel focus on the archaeal community involved in the phosphorus cycle, which has often been overlooked in favor of bacterial communities, despite the recognized abundance and diversity of archaea in phosphorus-deficient soils (J.-T. Wang et al., 2022).

2. MATERIALS AND METHODS

2.1. Site description, experimental design and sampling

The study was conducted at an experimental field located at ITAP (Santa Ana, Albacete, SE Spain) (38°53′39.8″N 1°59′18.0″W), situated in a semi-arid Mediterranean region. The pre-experimental physical and chemical properties of the soil are detailed in Table 1. The experiment consisted of 16 plots, each measuring 18.75 m², separated by aisles 1 m in width (Figure 12). Maize (var. P0937) was planted on May 18, 2022 (Figure 13). The objective of this study was to assess the impact of partially replacing conventional NPK mineral fertilization with struvite, sludge, or their combination (Barquero et al., 2024). A

randomized block design was used, with four treatments and four replicates per treatment. The treatments were: (i) conventional NPK mineral fertilization (NPK); (ii) organic fertilization using thermostabilized sludge (SLU); (iii) mineral fertilization with struvite (STR); and (iv) combined organo-mineral fertilization with both struvite and sludge (STRSLU) (Barquero et al., 2024). The concentrations of organic carbon, total nitrogen, and phosphorus in the struvite were 0.13, 5.80, and 16.30 g 100 g^{-1}, respectively, while those in the sludge were 29.08, 4.92, and 4.14 g 100 g^{-1}, respectively. A full chemical profile of both materials is provided in Table 2.

Table 1: Characteristics of the initial field soil before starting the assay.

		Average	Standard deviation
Sand	%	33.8	2.97
Silt	%	39.8	3.05
Clay	%	26.4	1.77
pH		8.5	0.27
Electrical conductivity	µS/cm	131.3	14.84
Organic matter	%	2.90	0.21
Total C	g/kg	57.96	1.67
Total N	g/kg	1.20	0.084
Total P	g/kg	0.64	0.04
Available P (Olsen)	mg/kg	20.24	3.01
Water-Soluble N	mg/kg	13.27	2.88

Figure 12: General view of the trial carried out in Santa Ana (Albacete, Murcia, SE Spain).

Figure 13: Maize planted in Santa Ana (Albacete, Murcia, SE Spain). Vegetative stage. Day 44.

Table 2: Characterization of struvite and sludge used in fertilizers.

	Struvite	**Sludge**
Organic C (g/100g)	0.13	29.08
Total N (g/100g)	5.80	4.92
Al (mg/Kg)	9,08	12629,32
As (mg/Kg)	0,16	5,79
Be (mg/Kg)	0,05	0,35
B (mg/Kg)	<0,01	<0,01
Ca (g/100g)	0,16	4,08
Cd (mg/Kg)	<0,01	0,94
Co (mg/Kg)	<0,01	7,31
Cr (mg/Kg)	4,17	59,65
Cu (mg/Kg)	0,11	260,74
Fe (mg/Kg)	529,89	21521,25
K (g/100g)	0,05	0,44
Mg (g/100g)	11,31	1,97
Mn (mg/Kg)	692,37	373,91
Mo (mg/Kg)	0,46	9,25
Na (g/100g)	0,002	0,12
Ni (mg/Kg)	0,16	23,45
Pb (mg/Kg)	0,62	63,55
P (g/100g)	16,30	4,13
Si (mg/Kg)	23,64	673,60
S (g/100g)	0,005	1,98
V (mg/Kg)	<0,01	26,27
Zn (mg/Kg)	0,63	770,34

Struvite, sludge, or their mixture were incorporated during the basal fertilization phase, as outlined in Table 3. To achieve the target fertilization rates for maize (≈192 UFN, 225 UFP, and 281 UFK), all forms of phosphorus, nitrogen, and potassium were applied during the initial fertilization phase, and additional nitrogen fertilization (using ammonium nitrate, NAC27) was provided during subsequent dressing stages (Table 3) (Barquero et al., 2024). To meet the crop's nutrient demands, supplemental superphosphate, potassium sulfate, and calcium ammonium nitrate (NAC27) were applied (Table 3).

Basal fertilization was performed on May 13, 2022, while additional applications of nitrogen were conducted on June 20, 2022, and July 8, 2022, coinciding with the V4 and V8 growth stages (Ritchie, S.W & J.J. Hanway, 1982). Irrigation was supplied as needed throughout the growing season (Barquero et al., 2024).

Table 3: Detail of the fertilization carried out in the evaluated treatments.

Treatment	Product	Application	Doses (Kg ha^{-1})	N (Kg ha^{-1})	P (Kg ha^{-1})	K (Kg ha^{-1})
NPK conventional	9-23-30	Basal	712			
	Superphosphate	Basal	338			
	Potassium sulfate	Basal	135			
	NAC27	Dressing	474			
				192,06	*98,82*	*233,313*
SLUDGE	Sludge	Basal	1306			
	Superphosphate	Basal	560			
	Potassium sulfate	Basal	549			
	NAC27	Dressing	474			
				192,23	98,82	233,55
STRUVITE	Struvite	Basal	602			
	Potassium sulfate	Basal	562			
	NSA26	Basal	111			
	NAC27	Dressing	474			
				191,75	98,85	233,52
STRUVITE+SLUDGE	Struvite	Basal	301			
	Sludge	Basal	653			
	Superphosphate	Basal	71			
	Potassium sulfate	Basal	455			
	9-23-30	Basal	165			
	NAC27	Dressing	474			
				192,41	98,98	232,92

Sampling was conducted at two distinct time points. The maize growth stages were classified using the phenological scale described by (Ritchie, S.W & J.J. Hanway, 1982). The selected sampling points corresponded to the phenological stages with the highest phosphorus demand: germination (V1) and flowering (R1) (Ritchie, S.W & J.J. Hanway, 1982). Soil samples from the rhizosphere were collected directly from the root zones by combining soil from five plants within each plot. This composite sample was obtained by shaking the roots of each plant in a plastic zip-lock bag. The collected soil samples were

sieved to 2 mm, with a portion stored at −20 °C for DNA extraction and the remainder air-dried for ent chemical analyses (Barquero et al., 2024).

Olsen phosphorus was determined using ICP-OES following extraction with 0.5 M $NaHCO_3$ (Olsen & Sommers, 1982). The chemical composition of dried sludge and struvite samples was analyzed through inductively coupled plasma optical emission spectroscopy (ICP-OES) (Barquero et al., 2024).

2.2. DNA Extraction and Shotgun Sequencing

DNA was extracted from 32 composite soil samples collected during two phenological stages (germination and flowering) using the DNeasy PowerSoil kit (Qiagen), following the protocol provided by the manufacturer. The extracted soil DNA was then utilized to prepare metagenomic libraries with the NEBNext® Ultra™ DNA Library Prep Kit (New England BioLabs), optimized for Illumina platforms, in accordance with the manufacturer's guidelines.

DNA fragmentation was achieved via acoustic shearing using a Covaris S220 instrument. The fragmented DNA underwent purification, end-repair, and adenylation at the 3′ ends, followed by adapter ligation and enrichment via limited-cycle PCR (Barquero et al., 2024).

To verify the quality of the DNA libraries, an Agilent 5300 Fragment Analyzer (Agilent Technologies) equipped with an NGS Kit was used, while quantification was performed with a Qubit 4.0 Fluorometer (Invitrogen). The libraries were multiplexed before loading onto the Illumina NovaSeq 6000 platform, adhering to the manufacturer's instructions.

Sequencing was performed using a paired-end (PE) configuration of 2 × 150 bp (Barquero et al., 2024). After sequencing, image analysis and base calling were conducted using NovaSeq Control Software version 1.7. The raw sequencing data generated in the form of .bcl files were converted to .fastq files and demultiplexed using Illumina bcl2fastq software version 2.20 (J. A. Siles, De la Rosa, et al., 2024). Library preparation and sequencing were carried out at Genewiz Europe (Leipzig, Germany).

2.3. Metagenomic analysis

The metagenomic libraries were processed following the methodology initially described by Žifčáková et al. (2016), with certain adaptations. The complete code used for the metagenomic analysis is available in the repository at https://github.com/mariabelen-bm/Doctoral_Thesis/blob/78fb7483aff71e570633d6b3a8f0bb666fe368de/metagenomics_pipeline_chapters_1_2.sh. The complete pipeline is also shown in Annex 1. Additionally, the Python scripts employed during the metagenomic pipeline are included in Annex 2. Raw sequencing data were analyzed using bioinformatics tools in a command-line environment (Shell), employing specialized software suites and Python scripts (v3.6).

The metagenomic workflow began with environment setup via Conda, which facilitated the installation of essential tools, including Khmer (https://anaconda.org/bioconda/khmer) and FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Quality control of raw sequences was carried out using FastQC (v0.12.0), with reads exhibiting quality scores below 30 or lengths shorter than 50 bases excluded from further analysis. Subsequently, median normalization was applied to minimize noise and retain sequences of interest, using a k-mer size of 20 and a minimum coverage threshold of 20 for sequence retention (Barquero et al., 2024).

Assembly of the processed data was conducted using MEGAHIT (v1.2.9) (D. Li et al., 2015), and the quality of assemblies was assessed with MetaQUAST (v5.2.0) (Mikheenko et al., 2016). Gene prediction was performed with FragGeneScan (Rho et al., 2010), and alignment of reads was carried out using Bowtie2 (v2.4.1) (https://bowtie-bio.sourceforge.net/bowtie2/index.shtml). During the alignment process, base-level normalization was implemented, and functional annotations were generated using multiple reference databases.

The NCBI nr database (https://www.ncbi.nlm.nih.gov/) was primarily used for taxonomic identification. For functional categorization, the KOG, KEGG, and dbCAN databases were employed: KOG facilitated the classification of sequences into conserved eukaryotic orthologous groups, KEGG enabled the assignment of sequences to metabolic pathways and biological processes, and dbCAN was specifically utilized for identifying carbohydrate-active enzymes, such as hydrolases and lyases (L. Huang et al., 2018; Kanehisa et al., 2016; Tatusov et al., 2003). This approach, following the methodology outlined by Žifčáková et al. (2016) and Žifčáková. (2017).

The results were organized into functional categories for improved interpretation. Specifically, the functional group associated with organic phosphorus mineralization encompassed genes encoding enzymes such as phytases (*3-phytase*), C-P lyases (*phnG*, *phnH*, *phnI*, *phnJ*, *phnK*, *phnL*, and *phnM*), alkaline phosphatases (*phoA*, *phoD*, *phoX*), and acid phosphatases (*aphA* and *phoN*) (Barquero et al., 2024). The group related to inorganic phosphorus solubilization included genes for quinoprotein glucose dehydrogenases (*gcd*), inorganic pyrophosphatases (*ppa*), polyphosphate kinases (*ppk1*), and pyrroloquinoline quinone synthases (*pqqC*). Finally, the starvation phosphorus regulon comprised genes coding for the phosphate regulon response regulator (*phoB* and *phoR*) and its inhibitor protein (*phoU*) (J. A. Siles et al., 2022). A comprehensive summary of these genes and their classification is provided in Table 4.

Given the complexity of the dataset, for the taxonomical analyses of each gene, we plotted the top dominant microbial populations for those genes that showed significant influence of fertilization treatment and/or phenological stage. Raw sequence data have been deposited in NCBI under BioProject accession number PRJNA1118481 (Barquero et al., 2024).

Table 4: Details on the 20 functional genes studied in the present work related to the phosphorus cycle.

Pathway	Gene	Enzyme	KEGG ID
Organic P mineralization	*phoA*	Alkaline phosphatase A	K01077
	phoD	Alkaline phosphatase D	K01113
	phoX	Alkaline phosphatase X	K07093
	phoN	Acid phosphatase class A	K09474
	aphA	Acid phosphatase class B	K03788
	3-phytase	3-phytase	K01083
	phnG	C–P lyase multienzyme complex	K06166
	phnH	C–P lyase multienzyme complex	K06165
	phnI	C–P lyase multienzyme complex	K06164
	phnJ	C–P lyase multienzyme complex	K06163
	phnK	C–P lyase multienzyme complex	K05781
	phnL	C–P lyase multienzyme complex	K05780
	phnM	C–P lyase multienzyme complex	K06162
Inorganic P solubilization	*ppa*	Inorganic pyrophosphatase	K01507
	ppk1	Polyphosphate kinase	K00937
	gcd	Quinoprotein glucose dehydrogenase	K00117
	pqqC	Pyrroloquinoline quinone synthase C	K06137
P-starvation response regulation	*phoB*	Phosphate regulon response regulator	K07657
	phoR	Phosphate regulon sensor histidine kinase	K07636
	phoU	PhoR/phoB inhibitor protein	K02039

2.4. Protein extraction and Mass spectrometry analyses

Proteins were extracted following established methodologies (Bastida et al., 2014; Chourey et al., 2010). To lyse cells and disrupt soil aggregates, samples were boiled for 10 minutes at 100 °C in a sodium dodecyl sulfate (SDS) buffer. Protein separation was achieved using 12 % SDS-PAGE, and gels were subsequently stained with colloidal Coomassie brilliant blue to visualize the proteins. The protein mixture for each sample was excised as a single gel slice. Further processing included the reduction and alkylation of cysteine residues, in-gel tryptic digestion, peptide elution, and desalting (Bastida et al., 2016). Prior to LC-MS analysis, peptide lysates were reconstituted in 0.1 % formic acid (Barquero et al., 2024).

For LC-MS/MS analysis, 5 µL of peptide lysate, equivalent to 1 µg of peptides, was injected into a nanoHPLC system (UltiMate 3000 RSLCnano, Dionex, Thermo Fisher Scientific). Initial trapping was performed on a C18-reverse phase trapping column (C18 PepMap100, 300 µm × 5 mm, 3 µm particle size, Thermo Fisher Scientific), followed by separation on an analytical C18-reverse phase column (Acclaim PepMap100, 75 µm × 25 cm, 3 µm particle size, nanoViper, Thermo Fisher Scientific). Peptides were ionized with a Nanospray Flex™ Ion Source and analyzed on an Orbitrap Exploris™ 480 mass spectrometer (Thermo Fisher Scientific) as outlined by Castañeda-Monsalve et al. (2024).

For further nanoLC-MS measurements, 1 µg of peptides was injected into a Vanquish Neo nanoHPLC system (Thermo Fisher Scientific). The peptide trapping and separation utilized a two-column setup: a trapping column (Acclaim PepMap 100, 75 µm × 2 cm, 3 µM particle size, Thermo Fisher Scientific) and an analytical column (Double nanoViper™ PepMap™ Neo, 75 µm × 150 mm, 2 µM particle size, Thermo Fisher Scientific). Separation was achieved using a two-phase gradient with mobile phases A (0.01 % formic acid in water) and B (80 % acetonitrile, 0.01 % formic acid in water). The gradient consisted of an initial increase in phase B from 4 % to 30 % over 95 minutes, followed by a rise to 55 % over 40 minutes, maintaining a flow rate of 300 nL/min (Barquero et al., 2024).

Mass spectrometric analysis employed an Orbitrap Exploris™ 480 instrument. MS parameters included a scan range of 350–1550 m/z, resolution of 120,000, automatic gain control (AGC) target of 3,000,000, and maximum injection time of 100 ms. For MS/MS, the 10 most intense precursor ions were selected with an isolation window of 1.4 m/z, resolution of 15,000, AGC target of 200,000, and maximum injection time of 100 ms. Dynamic exclusion was applied for 20 seconds, with a minimum intensity threshold of 8,000 ions (Castañeda-Monsalve et al., 2024).

Data processing was carried out using Proteome Discoverer software (v2.5.0.400, Thermo Fisher Scientific) and the SequestHT search engine. The search settings included trypsin specificity (allowing up to two missed cleavages), precursor mass tolerance of 10 ppm, and fragment mass tolerance of 0.02 Da. Carbamidomethylation of cysteine was set as a fixed modification. False discovery rates (FDR) were controlled at 1 % using Percolator (Käll et al., 2007). Protein identification was performed using a database created from the soil metagenome (Barquero et al., 2024). The raw data have been deposited in PRIDE (https://www.ebi.ac.uk/pride/) under the accession number PXD052073.

Proteins associated with phosphorus and nitrogen cycle genes were analyzed using the same approach previously described for metagenomics, based on the KEGG database. The relative abundance of proteins was calculated by normalizing their frequency (i.e., detection count) in the sample, considering contig length, read length, and the number of reads per sample. Normalized values were scaled by multiplying by a factor of 100,000 to enhance data interpretability and facilitate comparisons between samples (Barquero et al., 2024).

2.5. Statistical Analysis

The impact of various fertilization treatments and phenological stages on the abundance of functional genes, as well as the microbial populations carrying these genes, was assessed using ANOVA implemented in R (R-Core-Team, 2023) with the "stats" package. Prior to analysis, the data were tested to ensure normality and homoscedasticity. The same statistical procedure was applied to evaluate the effects of fertilization treatments and phenological stages on the availability of phosphorus (Olsen phosphorus), total nitrogen and water-soluble nitrogen (WSN) (Barquero et al., 2024).

To visualize how treatments and phenological stages influenced the functional composition of bacterial and archaeal communities, including genes involved in organic phosphorus mineralization, inorganic phosphorus solubilization and phosphorus starvation response regulation, non-metric multidimensional scaling (NMDS) was conducted. The dataset of gene abundances underwent a base-10 logarithmic transformation before ordination, which was performed using the Bray-Curtis dissimilarity index via the metaMDS() function in the "vegan" package (Oksanen et al., 2019) in R (R-Core-Team, 2023).

Data visualization was facilitated by generating heatmaps and bar charts using the "ggplot2" package (Wickham, 2016). To examine potential linear relationships among genes involved in the phosphorus cycle, as well as their associations with Olsen phosphorus, Pearson correlation analysis was performed. The correlation matrix was calculated using the cor() function from the "corrplot" package (Taiyun, 2017) in R, and statistical significance was determined using the cor.mtest() function. A 95% confidence level was applied for significance testing.

3. RESULTS

3.1. Olsen phosphorus

The soil Olsen phosphorus content demonstrated a statistically significant relationship with the phenological stage (*P* = 3e-06), as well as with the interaction between fertilization and phenological stage (*P* = 3e-06) (Figure 14). However, no significant effects were observed among the fertilizer treatments themselves. During the germination stage, the combination of struvite and sludge resulted in the lowest Olsen phosphorus content. In contrast, during the flowering stage, the traditional NPK treatment exhibited the lowest Olsen phosphorus content, whereas both the sludge treatment and the struvite plus sludge combination showed higher Olsen phosphorus levels (Figure 14) (Barquero et al., 2024).

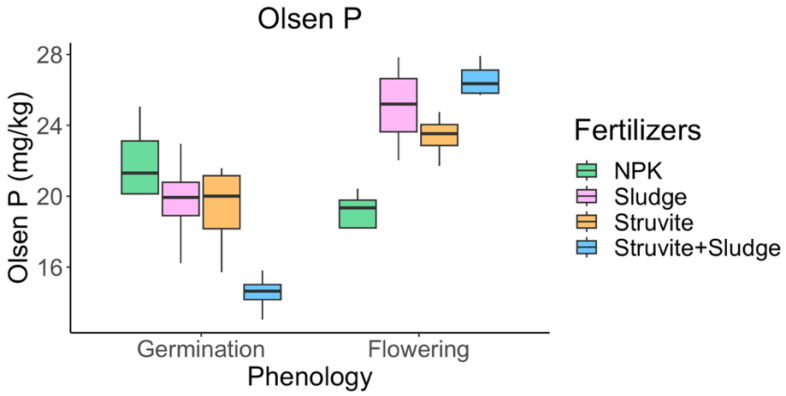


Figure 14. Phosphorus Olsen content during germination and flowering in soils supplemented with the four fertilizers: NPK, Sludge, Struvite and Struvite + Sludge. The ANOVA test carried out to check if there were significant (*P* < 0.05) effects between fertilizers (F), phenology (P) and the interaction between fertilization and phenology (F:P) has been added to the boxplot.

3.2. The abundance of phosphorus genes in the bacterial and archaeal community

A total of 813,089 reads were annotated with genes, out of which only 208 were linked to fungi. Of these, merely three were associated with phosphorus-related genes, so this domain was not further explored. Within the remaining reads, 9998 were attributed to bacteria harboring phosphorus-cycling genes, while archaea accounted for 188 reads (Barquero et al., 2024). The identified phosphorus genes were grouped into three functional categories: organic phosphorus mineralization, inorganic phosphorus solubilization, and phosphorus-starvation response regulation (Figure 15). Fertilization treatments influenced the abundance of genes in the organic phosphorus mineralization category exclusively for archaea, while the bacterial community showed no such effect (Figure 15). However, the interaction between fertilization and phenological stage had an impact on the abundance of bacterial genes involved in organic phosphorus mineralization. For instance, in the case of sludge fertilization, the abundance of these genes peaked during germination but decreased during flowering. In contrast, for the combined treatment (struvite plus sludge), the opposite trend was observed (Barquero et al., 2024).

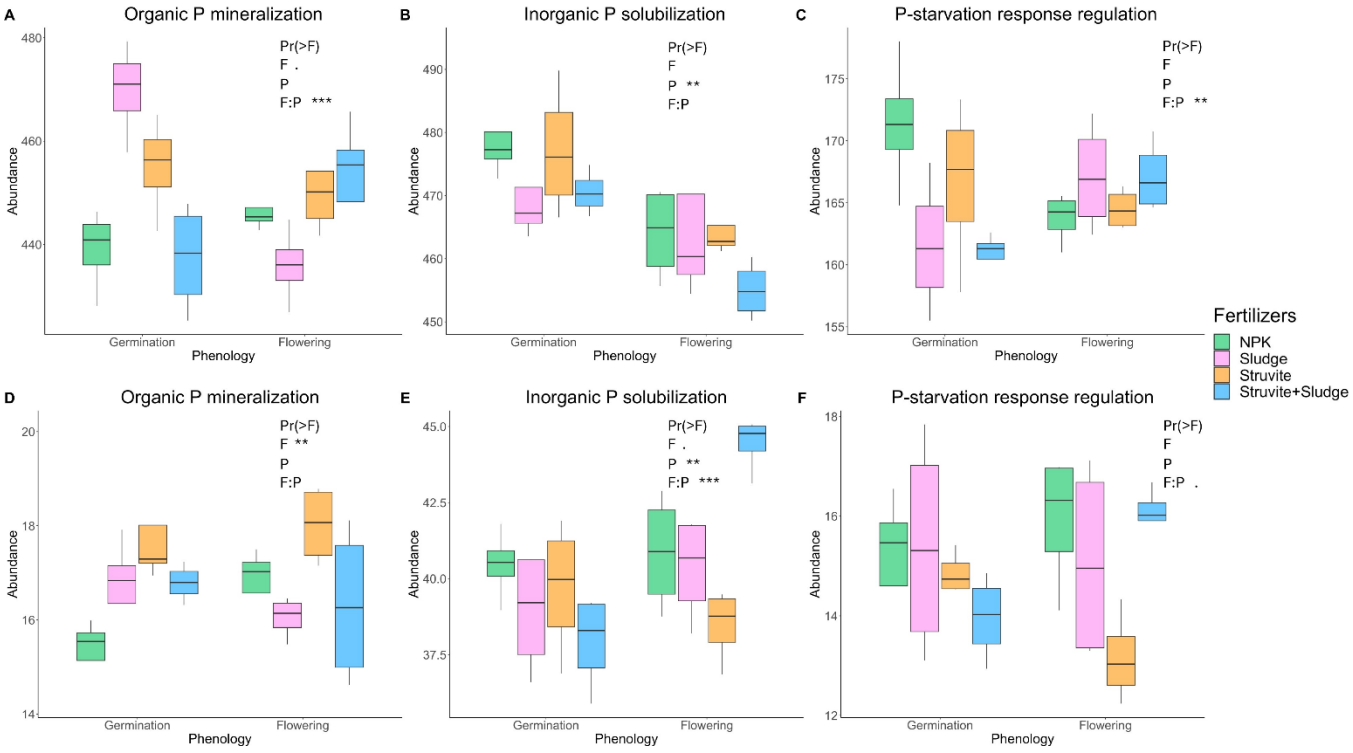


Figure 15: Phosphorus genes abundance of organic phosphorus mineralization, inorganic phosphorus solubilization and phosphorus -starvation response regulation categories in soils supplemented with the four fertilizers (NPK, Sludge, Struvite and Struvite + Sludge) during germination and flowering. Figs. A, B and C correspond to gene abundance in the bacterial community, while Figs. D, E and F correspond to gene abundance in the archaeal community. The ANOVA test carried out to check if there were significant ($P < 0.05$) effects between fertilizers (F), phenology (P) and the interaction between fertilization and phenology (F:P) has been added to the boxplot.

The phenological stage of maize, rather than fertilization, influenced the abundance of genes associated with inorganic phosphorus solubilization in both bacterial and archaeal communities. Specifically, within the bacterial community, the abundance of these genes was consistently higher during germination compared to flowering across all fertilization treatments. Genes involved in phosphorus-starvation response regulation were not significantly affected by individual factors, but their abundance was influenced by the interaction between fertilization and phenological stage in the bacterial community (Figure 15). For example, during flowering, fertilization treatments including sludge—either alone or in combination with struvite—resulted in an increased abundance of these genes, whereas the opposite trend was noted during germination (Barquero et al., 2024).

Non-metric multidimensional scaling (NMDS) analysis indicated that neither fertilization treatments nor phenological stages significantly altered the functional structure of the bacterial community associated with the phosphorus cycle (Figure 16) (Barquero et al., 2024). However, the functional structure of the archaeal community was significantly influenced by phenological stage (Figure 16).

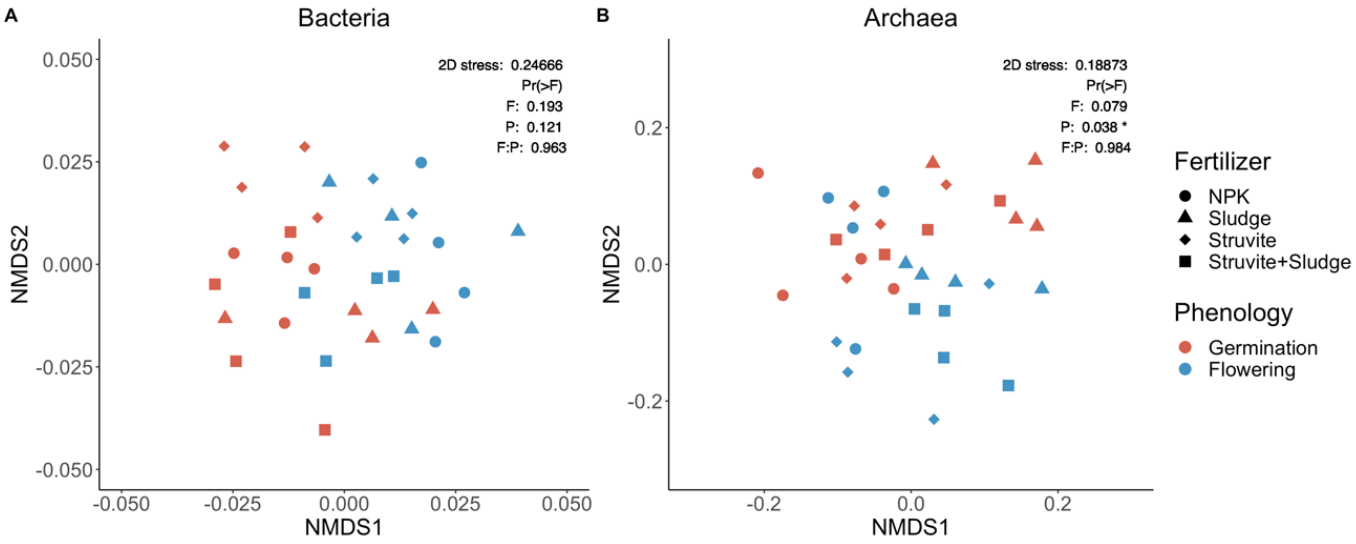


Figure 16: NMDS analysis based on Bray-Curtis dissimilarities of phosphorus gene abundance between soils supplemented with the four fertilizers (NPK, Sludge, Struvite and Struvite+Sludge) during germination and germination. A) Bacteria; B) Archaea. The ANOVA test carried out to check if there were significant (P<0.05) effects between fertilizers (F), phenology (P) and the interaction between fertilization and phenology (F:P) has been added to the boxplot.

Overall, in both bacterial and archaeal communities, genes involved in inorganic phosphorus solubilization and phosphorus-starvation response regulation were more abundant than those related to organic phosphorus mineralization (Barquero et al., 2024). An exception to this was the *phoD* gene, which was highly abundant in archaea (Figure 17). Within bacterial genes associated with organic phosphorus mineralization, several Carbon-Phosphorus lyases, such as *phnG*, *phnH*, *phnI*, *phnJ*, and *phnL*, were significantly affected by fertilization treatments (P < 0.05). Moreover, the abundance of *phnG*, *phnJ*, and *phnL* genes was significantly influenced by phenological stage (P < 0.05), displaying contrasting patterns.

For instance, the abundance of *phnL* was notably lower in the combined struvite and sludge treatment, a trend that was also apparent for the *phnG* gene during germination. In contrast, genes encoding phosphatases, such as *phoD* and *phoX*, were not affected by fertilization treatments (Barquero et al., 2024).

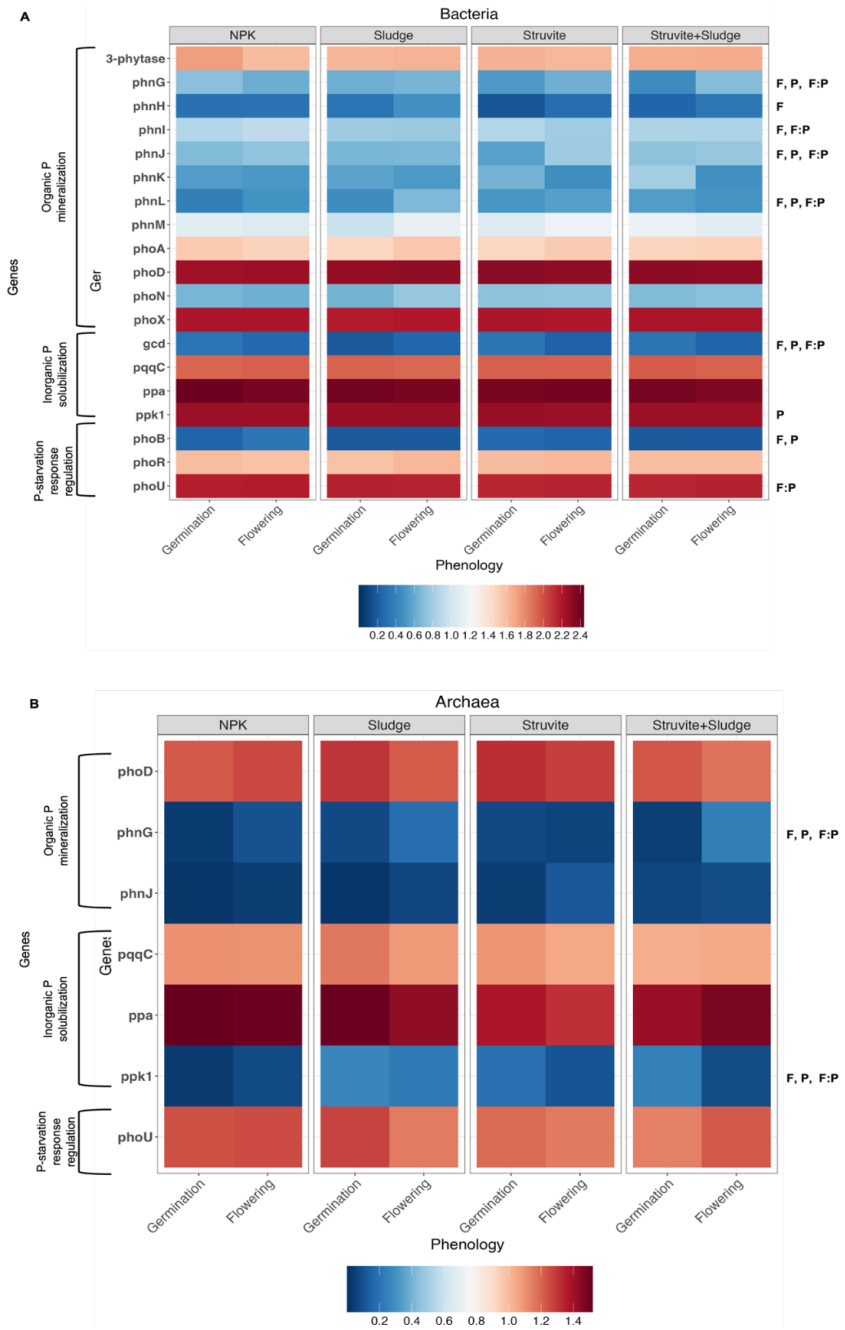


Figure 17: Heatmap of the logarithm in base 10 of the abundance of the different phosphorus genes grouped in the categories of organic phosphorus mineralization, inorganic phosphorus solubilization and phosphorus -starvation response regulation in soils supplemented with the 4 fertilizers (NPK, Sludge, Struvite, Struvite + Sludge) during germination and flowering. A) Bacterial community; B) Archaeal community. The ANOVA test carried out to check if there were significant ($P < 0.05$) effects between fertilizers (F), phenology (P) and the interaction between fertilization and phenology (F:P) has been added to the boxplot.

Among bacterial genes involved in inorganic phosphorus solubilization, the *gcd* gene, which encodes quinoprotein glucose dehydrogenase, was significantly influenced by fertilization, phenology, and their interaction ($P < 0.05$) (Figure 17). This gene exhibited the highest abundance in the NPK treatment during germination but dropped sharply under sludge fertilization. Additionally, the abundance of *gcd* decreased during flowering compared to germination for treatments involving struvite, the combined organo-mineral application, and NPK. Furthermore, genes related to phosphorus-starvation response regulation, such as *phoB* and *phoU*, were the only ones influenced by fertilization. The *phoB* gene, encoding the phosphate regulon response regulator, showed the highest abundance under NPK fertilization during flowering, while *phoU*, encoding the PhoR/phoB inhibitory protein, was most abundant in soils fertilized with struvite plus sludge during germination (Barquero et al., 2024).

Within the archaeal community, only the genes *ppk1* and *phnG* were significantly influenced by fertilization, phenology, and their interaction. The *ppk1* gene, which encodes polyphosphate kinase involved in inorganic phosphorus solubilization, displayed its highest abundance in soils treated with sludge during germination. Furthermore, the abundance of *ppk1* was consistently greater in soils amended with sludge, struvite, or their combination compared to NPK during germination. Meanwhile, *phnG*, a gene encoding a Carbon-Phosphorus lyase, was more abundant in soils fertilized with NPK, sludge, or the combined struvite plus sludge treatment, particularly during flowering (Figure 17) (Barquero et al., 2024).

3.3. Taxonomic distribution of phosphorus cycle genes in bacterial communities across treatments and phenology

Genes related to the phosphorus cycle in soil were assigned to taxa belonging to 17 bacterial phyla, being particularly abundant in Acidobacteria and Actinobacteria. Figure 18 shows the top-dominant taxa harboring bacterial phosphorus-cycle genes. This figure is limited to those genes that showed significant differences in their abundances across treatment, phenology, and/or their interaction in Figure 17. Importantly, we observed a relevant functional clustering across taxa. Thus, there was a pattern in which top-dominant populations harboring genes for inorganic phosphorus solubilization and phosphorus starvation regulon were not dominant (or even absent) in harboring genes of organic phosphorus mineralization. Nevertheless, there were some exceptions. For instance, some populations, such as *Microvirga* and *Hyphomicrobium*, contained genes for phosphorus solubilization and phosphorus starvation regulon, but also genes involved in organic phosphorus mineralization (Barquero et al., 2024).

Regarding inorganic phosphorus solubilization and phosphorus-starvation response genes, *Luteitalea* (Acidobacteria) and *Nocardioides*, *Solirubrobacter*, *Blastococcus*, *Rhodocytophaga* and *Streptomyces* (Actinobacteria) were the predominant microorganisms harboring *ppk1* and *phoU*. *ppk1* was particularly abundant in *Ilumatobacter* and *Microvirga* in NPK and organo-mineral treatments, respectively. The *gcd* gene was only found in three populations (*Brevundimonas*, *Rhodocytophaga* and *Luteitalea*). The abundance of the *gcd* gene in *Luteitalea* peaked in the struvite treatment at the flowering stage of maize.

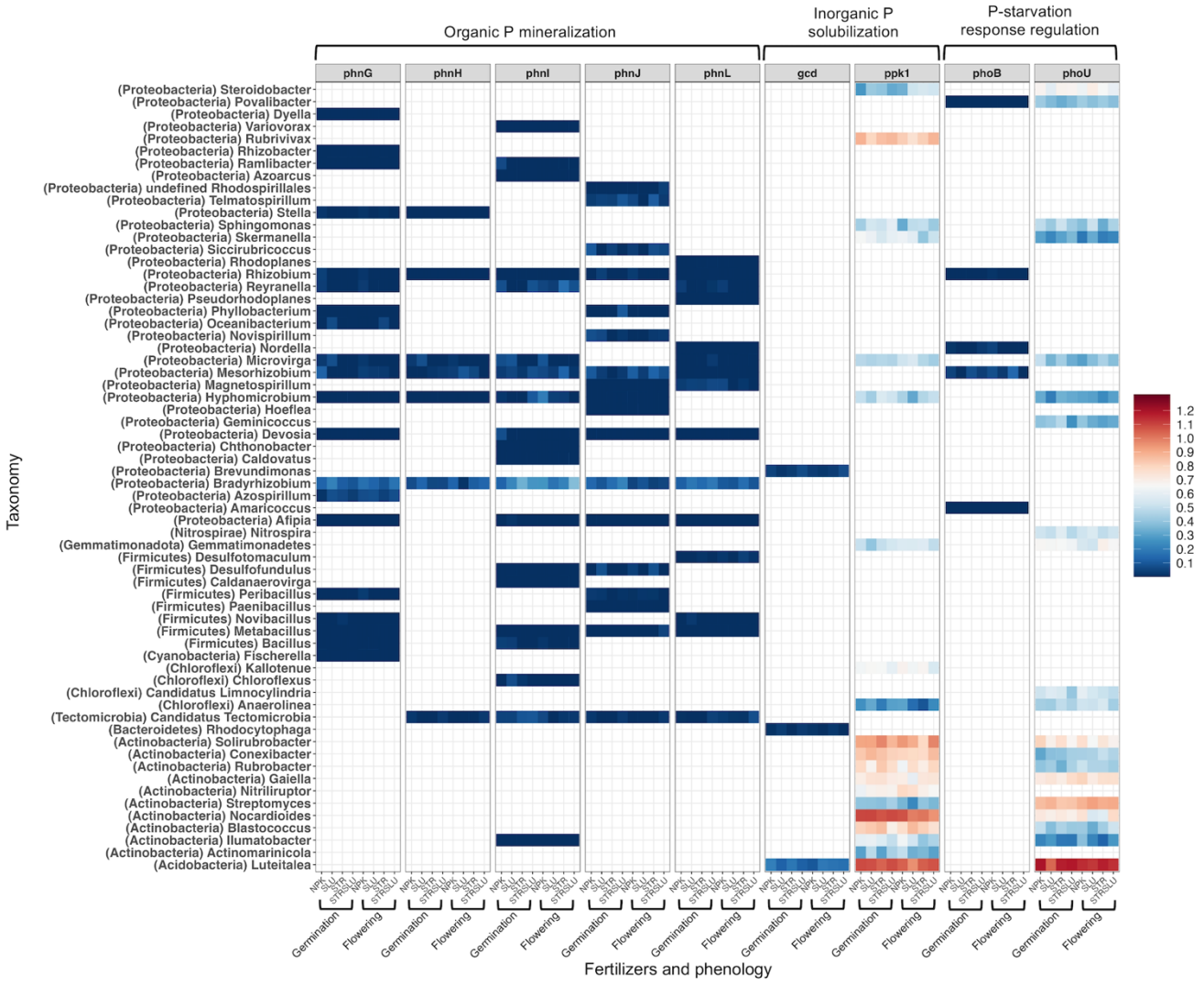


Figure 18: Heatmap of the logarithm in base 10 of bacterial taxonomic abundance in phosphorus genes grouped into organic phosphorus mineralization, inorganic phosphorus solubilization and phosphorus-starvation response regulation categories in soils supplemented with the four fertilizers during germination and flowering. Fertilizers on the x-axis are abbreviated: SLU (Sludge), STR (Struvite) and STRSLU (Struvite+Sludge). Taxonomy has been grouped into phylum, class and genus. The image shows the genes that showed significant differences between fertilizers or between phenology. Only the 15 most abundant bacterial genera for each gene are displayed.

Further, the relative abundance of *phoB* and *phoU* genes, involved in the phosphorus-starvation response regulation, was higher in *Mesorhizobium* and *Ilumatobacter*, respectively. The abundance of *phoB* in *Mesorhizobium* was also phenology-dependent, with greater values in sludge treatments during the flowering stage. Additionally, the abundance of *phoB* in *Rhizobium* responded to treatment and phenology, and was greater in the sludge treatment at the flowering stage (Barquero et al., 2024). Regarding genes associated with organic phosphorus mineralization, in the case of *phnG*, *Peribacillus*, *Reyranella*, and *Rhizobium* were the most abundant bacteria harboring this gene under the NPK treatment. The *phnI* gene was associated with *Devosia*, and the *phnL* gene was mainly linked to *Nordella*, and both peaked in the flowering stage (Figure 15) (Barquero et al., 2024).

3.4. Taxonomic distribution of phosphorus cycle genes in archaeal communities across treatments and phenology

Despite the limited number of genes associated with archaea, we were able to assign them to several taxa that revealed specific patterns. Genes related to the phosphorus cycle in soil were assigned to taxa belonging to 3 archaeal phyla, being particularly abundant in *Thaumarchaeota* (Figure 19). Figure 19 shows the abundance of phosphorus cycle genes in the top dominant archaeal populations for each gene. The *paa* gene, involved in inorganic phosphorus solubilization, was widely represented in different archaeal populations, including members of *Crenarchaeota*, *Euryarchaeota* and *Thaumarchaeota*. These results contrasted with other genes of inorganic phosphorus solubilization such as *ppk1* (exclusively associated to *Methanosphaerula*) and *pqqC*, which was found only in *Nitrosocosmicus* and *Nitrososphaera*. The microorganism mainly associated with *phoD* was *Nitrosopumilus*, which was more represented in the struvite treatments. The abundance of this gene in this microorganism was much higher than that found in the *Euryarchaeota* populations harboring it. In fact, the abundance of *phoD*, *phnG* and *phnJ* was low in the *Euryarchaeota* populations harboring it. Regarding the effects of fertilization treatment and phenology, the abundance of *Natrialba* harboring the *phnG* gene was influenced by the combination of treatment and phenology, being more abundant during germination in the organo-mineral treatment. The abundance of *Halococcus* harboring the *phoD* gene was affected solely by the treatment (Figure 19). Conversely, concerning genes associated with inorganic phosphorus solubilization, the abundance of *Methanosphaerula* harboring the *ppk1* gene was the highest in the organo-mineral treatment, particularly during germination (Barquero et al., 2024).

3.5. Abundance and microbial origin of identified proteins by metaproteomics

Metaproteomics allowed the detection and quantification of 311 proteins, with only 0.96% related to the phosphorus cycle. The phosphorus-related enzymes identified were *phoR*, *phoX*, and *3-phytase*. Among these, only the abundance of *3-phytase* was significantly influenced by fertilizer ($P = 0.007$), as well as by phenology ($P = 0.001$) and the interaction of fertilizer and phenology ($P = 0.011$) (Figure 17).

Regarding the taxonomic distribution of these enzymes, we observed that *phoR* was predominantly harbored by *Nitrospirae*, a genus also highly represented in the metagenome. Its abundance was found to be statistically influenced by fertilizer ($P = 0.001$) and the combination of fertilizer and phenology ($P = 0.0024$), being more abundant in the organo-mineral fertilizer during germination. In contrast, *phoX*, an alkaline phosphatase, showed higher abundance in *Phytohabitans*, *Skermanella*, and *Solirubrobacter*. Notably, *Phytohabitans*, unlike the other two genera, was not among the most abundant in the metagenome (Figure 20). Finally, regarding *3-phytase*, which was exclusively harbored by the genus *Nonomuraea*, it did not feature among the most abundant genera in the metagenome.

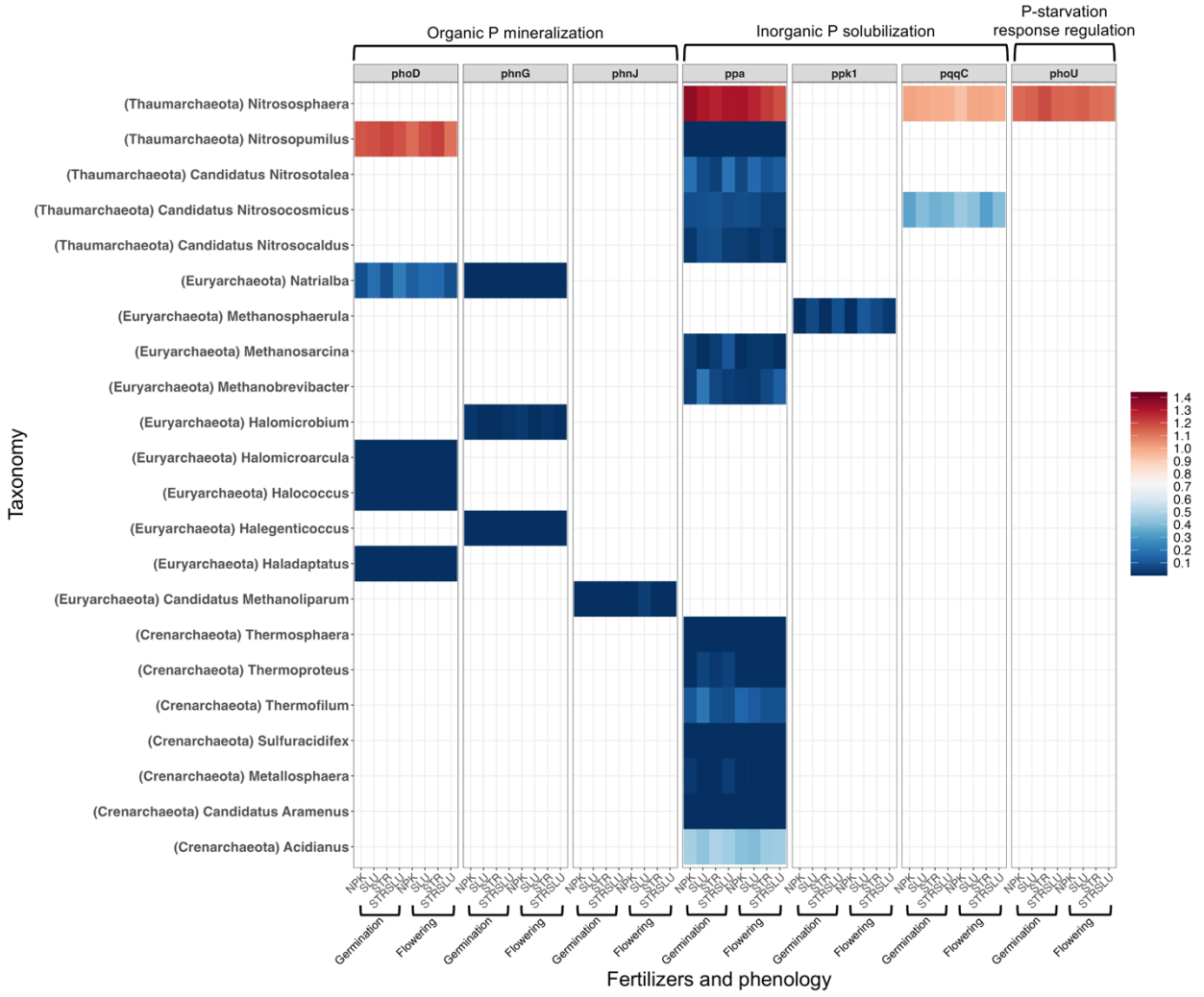


Figure 19: Heatmap of the logarithm in base 10 of archaeal taxonomic abundance in phosphorus genes grouped into organic phosphorus mineralization, phosphorus inorganic solubilization and phosphorus -starvation response regulation categories in soils supplemented with the four fertilizers during germination and flowering. Fertilizers on the x-axis are abbreviated: SLU (Sludge), STR (Struvite) and STRSLU (Struvite+Sludge). Taxonomy has been grouped into phylum, class and genus.

3.6. Correlations between Olsen phosphorus content and the relative abundance of genes

With respect to the bacterial community (Figure 21A), we observed a significant positive correlation between Olsen phosphorus and the *phnG* gene, and between the *phnM* and *phnL* genes, all corresponding to the organic phosphorus mineralization category. In addition, both genes (*phnM* and *phnL*) showed a significant negative correlation with the *gcd* gene, which belongs to the inorganic phosphorus solubilization category. The *phnL* gene also showed a significant negative correlation with the *phoU* gene, which is involved in the category of regulation of the response to phosphorus starvation. As for the archaeal community (Figure 21B), we similarly found a positive correlation between the *phnG* gene and Olsen phosphorus, as well as a strong significant positive correlation between the *ppa* and *phoU* genes, which

are associated with inorganic phosphorus solubilization and regulation of the response to phosphorus stress, respectively.

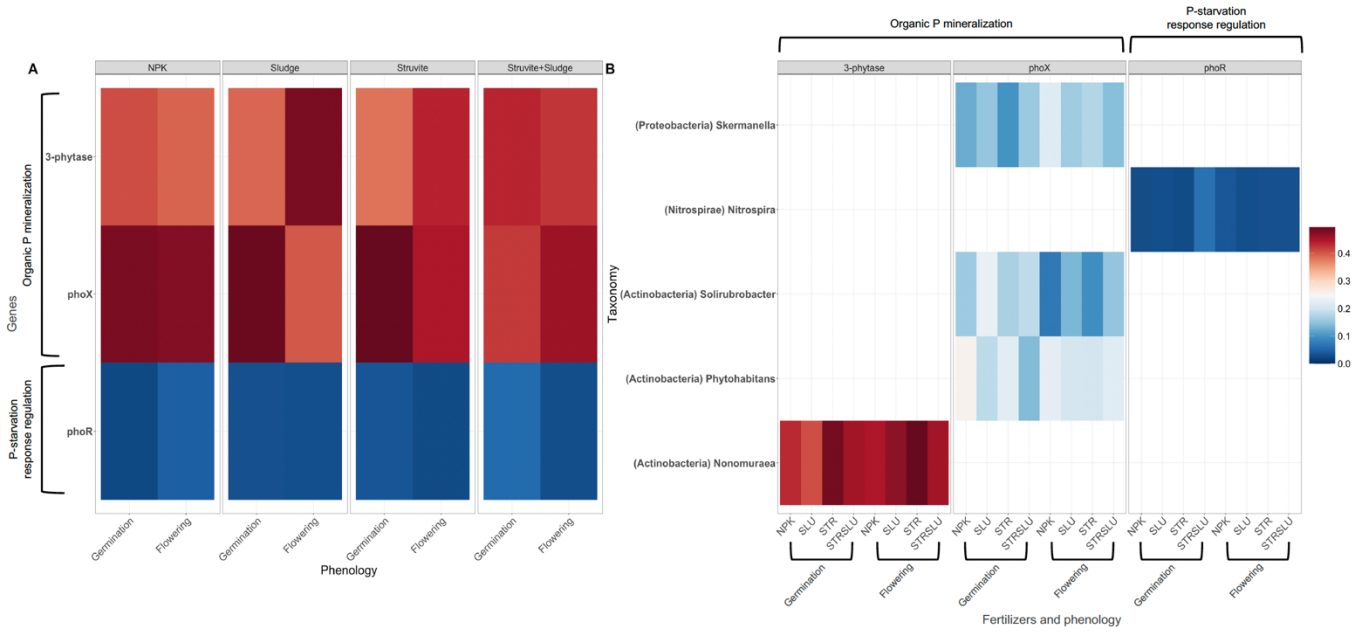


Figure 20: Heatmap of the logarithm in base 10 of the results obtained in proteomics. A) Abundance of phosphorus genes identified in metaproteomics grouped in the categories of organic phosphorus mineralization, inorganic phosphorus solubilization and phosphorus-starvation response regulation in the different soils supplemented with the four treatments (NPK, Sludge, Struvite and Struvite+Sludge) during germination and flowering; B) Abundance of bacterial populations harboring different phosphorus genes grouped into organic phosphorus mineralization, inorganic phosphorus solubilization and phosphorus-starvation response regulation categories in the different soils supplemented with the four fertilizer treatments during germination and flowering. Taxonomy has been grouped into phylum, class and genus. The ANOVA test carried out to check if there were significant ($P < 0.05$) effects between fertilizers (F), phenology (P) and the interaction between fertilization and phenology (F:P) has been added to the boxplot.

4. DISCUSSION

4.1. Abundance of phosphorus-related genes and enzymes and the associated microbiome

Our findings revealed that both bacterial and archaeal communities displayed a higher prevalence of genes associated with inorganic phosphorus solubilization and the regulation of phosphorus starvation responses compared to those involved in the mineralization of organic phosphorus. This result contrasts with prior studies that reported a higher abundance of genes linked to organic phosphorus mineralization (L. Liu et al., 2023; Wan et al., 2020). These discrepancies can be attributed to variations in soil characteristics across different investigations. Specifically, in the semiarid soils of our study, the elevated pH promotes phosphorus immobilization within minerals (Ruiz-Navarro et al., 2023), which differs from the environmental conditions described in the aforementioned research. As a result, our study highlighted

genetic elements primarily associated with mineral phosphorus solubilization, reflecting the limited bioavailability of phosphorus in these soils. This finding aligns with the long history (>20 years) of conventional fertilization in the studied field, during which phosphorus has likely accumulated in soil mineral particles (Bastida et al., 2023).

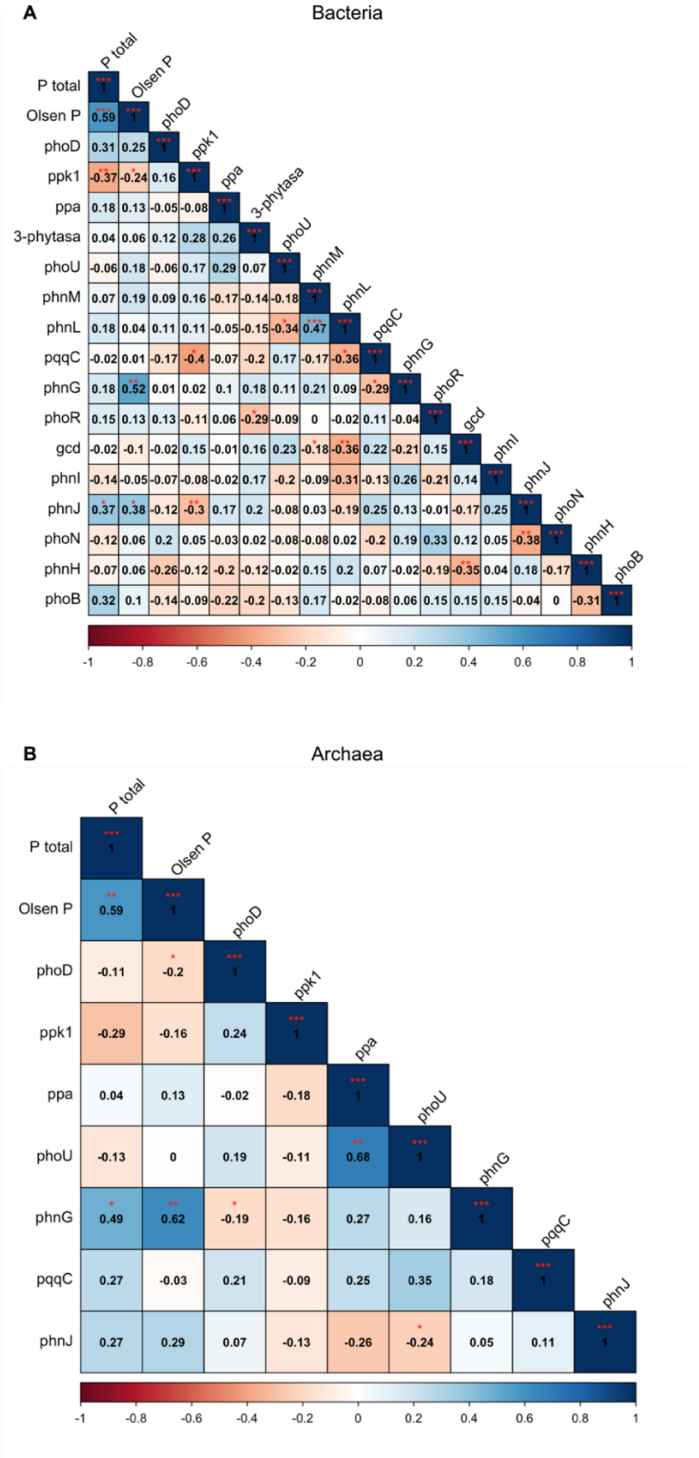


Figure 21: Correlation analysis of identified phosphorus genes and total phosphorus and phosphoruss Olsen. A) Bacteria; B) Archaea. Negative correlations are represented in blue, while positive ones, in red. The asterisks represent those positive or negative correlations that were found to be statistically significant in the correlation test.

Interestingly, the results from our metagenomic analysis differed from the metaproteomic approach, which identified the most abundant proteins as those involved in organic phosphorus mineralization and the regulation of phosphorus starvation responses. This discrepancy suggests a decoupling between the abundance of genes and the expression of proteins (Starke et al., 2019), consistent with prior studies demonstrating that a greater abundance of genes does not necessarily correlate with higher expression levels (Fierer et al., 2012). Notably, while several alkaline phosphatases involved in organic phosphorus mineralization were detected, existing literature has predominantly focused on *phoD*, a gene known to be present across many bacterial phyla and widely distributed in environmental samples (Ragot et al., 2015). Nevertheless, despite the inherent limitations of current metaproteomic techniques—such as reduced protein extraction efficiency and low identification rates—our approach utilized an *ad hoc* metagenome for each sample to improve protein annotation. The predominance of the phosphatase encoded by *phoX* in our soils underscores its critical role in the phosphorus cycle of Mediterranean agroecosystems (Ragot et al., 2017).

From a taxonomic perspective, our study highlights the existence of functional niches related to phosphorus transformation processes that are clustered within distinct microbial groups. Specifically, microorganisms involved in inorganic phosphorus solubilization did not typically carry genes for organic phosphorus mineralization or phosphorus starvation response regulation. Conversely, microorganisms associated with the latter two categories tended to share genes, suggesting the organization of functional guilds or niches linked to the soil phosphorus cycle. For example, genera belonging to Acidobacteria and Actinobacteria were predominant in harboring *ppk1* and *phoU*, genes associated with inorganic phosphorus solubilization and phosphorus starvation regulation, respectively. The *ppk1* gene encodes an enzyme central to the synthesis and degradation of polyphosphates (Achbergerová & Nahálka, 2011). Moreover, our metaproteomic data revealed that *phoR*, a gene associated with phosphorus starvation response regulation, was exclusively identified in members of the phylum Nitrospirae —a group highly represented in the metagenome. This phylum, known for its pivotal role in the nitrogen cycle (Al-Ajeel et al., 2022), has recently been implicated in sensing phosphorus deficiency (Han et al., 2018), where *phoR* may serve as a key regulatory gene.

Our findings also demonstrated that while bacterial communities harbored a relatively higher abundance of genes linked to mineral phosphorus solubilization compared to organic phosphorus mineralization, the phylogenetic distribution of dominant phosphorus solubilizers was constrained. This contrasts with the broader taxonomic distribution of *ppk1*, *phoU*, and other genes related to organic phosphorus mineralization. For instance, *Luteitalea*, a genus recognized for its role in phosphorus solubilization (Valenzuela et al., 2022), was one of only three major genera harboring the *gcd* gene in our study. Similarly, *Brevundimonas*, which also contained *gcd*, emerged as a notable contributor to phosphorus solubilization and enhanced crop productivity (Zaim & Bekkar, 2023). Overall, the phylum Proteobacteria included numerous members encoding phosphorus solubilization genes. Furthermore, comparative studies on phosphorus availability in reforested versus agricultural soils have reported a taxonomic

distribution of the *gcd* gene, similar to that observed in our study, primarily encoded by members of Acidobacteria and Bacteroidetes (X. Wu et al., 2022).

Our metaproteomic data also identified the *phoX* protein in genera such as *Phytohabitants*, *Skermanella*, and *Solirubrobacter*. However, little is known about the roles of *Phytohabitants* and *Skermanella* in the phosphorus cycle. Notably, *Phytohabitants* was not among the most abundant genera in the metagenomic data, reinforcing the notion that microorganisms with lower genetic abundance for specific genes can still exhibit high protein expression levels. Additionally, *Solirubrobacter* has been previously studied as a key genus involved in phosphorus mobilization (X. Yang et al., 2024).

In the case of archaea, research on their role in the phosphorus cycle remains limited, though recent studies have emphasized their importance in modulating soil stoichiometry under phosphorus-deficient conditions (J.-T. Wang et al., 2022). Our findings showed that bacterial reads of phosphorus-related genes significantly outnumbered those from archaea, a disparity that reflects the lower sequencing recovery of phosphorus-related archaeal genes. Similar to bacteria, archaeal genes associated with inorganic phosphorus solubilization and phosphorus starvation regulation were more abundant than those related to organic phosphorus mineralization. The phylogenetic distribution of the *ppa* gene, encoding an inorganic pyrophosphatase involved in phosphorus solubilization, was widespread among archaea, including members of *Crenarchaeota*, *Euryarchaeota*, and *Thaumarchaeota*. In contrast, *ppk1* was confined to the genus *Methanosphaerula* within the phylum Euryarchaeota, and *pqqC* was only detected in *Nitrosocosmicus* and *Nitrososphaera* of the phylum Thaumarchaeota. These genera, known for their roles in the nitrogen cycle (Rodriguez et al., 2021), may require phosphorus for general metabolic processes, as demonstrated by recent findings on their diverse nutrient acquisition strategies under low phosphorus availability (J. Zhao et al., 2023).

4.2. Influence of fertilization on the functionality of the phosphorus-associated microbiome

Our results demonstrate that different types of fertilizers, in conjunction with crop phenology, interacted with the abundance of bacterial genes associated with organic phosphorus mineralization and phosphorus starvation. In the case of archaea, fertilization treatments significantly affect the abundance of genes involved in organic phosphorus mineralization, underscoring the sensitivity of this relatively unexplored microbial group in agroecosystems to fertilization strategies. Overall, our results demonstrated that, despite the extensive history of fertilizer inputs, prokaryotic communities rapidly respond to contrasting fertilizers with high phosphorus content (Y. Wang et al., 2016).

Among the bacterial genes involved in inorganic phosphorus solubilization, the abundance of *gcd* was significantly influenced by fertilization treatments, particularly those supplemented with struvite. Here, we observed a higher abundance of the *gcd* gene in soils fertilized with NPK, struvite, and organo-mineral fertilizers, but not in the sludge treatment. Likely, treatments including a highly mineral source

of phosphorus would induce a greater abundance of this gene, potentially enhancing phosphorus solubilization. These results contrast with previous research in which *gcd* showed higher abundance in soils fertilized with organic manure (J.-T. Wang et al., 2022). *Luteitalea* was the predominant genus harboring *ppk1* and *phoU*, being more abundant in organo-mineral and struvite treatments, respectively. This genus has been identified as important in the decomposition of organic matter, in the carbon cycle, and in nitrogen fixation in soils (C. Lyu et al., 2023), and our findings also highlight a potential role of this genus in the phosphorus cycle.

Consistent with previous studies (Y. Shi et al., 2020) that observed a decrease in the abundance of ammonia-oxidizing archaea such as *Nitrososphaera* or *Nitrosopumilus* with organic amendments, we found a higher abundance of *phoU* in *Nitrososphaera* and of *phoD* in *Nitrosopumilus* in struvite-amended soils. This pattern can be attributed to struvite's ammonium-rich nature (Martines et al., 2010), which makes it a favorable substrate for ammonia-oxidizing archaea.

4.3. Influence of phenology on the functionality of the phosphorus-associated microbiome

According to our hypothesis, phenological stage modulated the overall abundance of genes involved in inorganic phosphorus solubilization in both bacteria and archaea. In contrast, the overall abundance of genes involved in organic phosphorus mineralization was not influenced by crop phenology. Specifically, the abundance of genes related to inorganic phosphorus solubilization was lower during flowering compared to germination in the bacterial community, while the opposite trend was observed in archaea. Both phenological stages, germination and flowering, are highly demanding in phosphorus (Barry & Miller, 1989). It is possible that the plant modulates the bacterial community in the stages prior to flowering (i.e., germination) to increase the release of phosphorus from minerals and organic matter (Fontaine et al., 2024). Thus, afterwards the rest of phosphorus is less soluble, producing a greater accumulation of phosphorus in the flowering stage. Besides the overall abundance of genes, our findings suggest that the functional composition of archaeal genes involved in phosphorus cycling was more sensitive to phenological stages than that of bacterial genes, as demonstrated by NMDS analysis. This sensitivity may arise from a lower level of functional redundancy within the phosphorus cycle among archaea. However, previous research has emphasized that despite bacteria possessing a larger repertoire of phosphorus cycle-related genes compared to archaea, both microbial communities demonstrate notable redundancy, particularly within the genes of the *Pho* regulon (K. D. Schneider et al., 2019; J. A. Siles et al., 2022).

When examining genes related to inorganic phosphorus solubilization, we noted significant differences in *ppk1* gene abundance in *Nocardioides* and *Sphingomonas* across phenological stages and fertilization treatments. The abundance of *ppk1* in *Nocardioides* exhibited higher abundance during germination, which might highlight its role in phosphorus mobilization for plant growth (Shen et al., 2023). Additionally, *Sphingomonas* was more abundant in sludge-treated soils during germination, contrasting with studies linking its presence to lower soil phosphorus content (Lagos et al., 2016). Further, in the case of archaea,

the increased abundance of the *ppk1* gene with alternative fertilization treatments during germination might indicate that archaea have the potential to mobilize inorganic phosphate supplies, as energy and nutrients are required for germination and growth. Indeed, some rhizospheric archaea possess attributes that promote plant growth and aid in nutrient management, such as the solubilization of phosphorus, potassium (*K*), and zinc (*Zn*) (Naitam & Kaushik, 2021). Variations were also observed in microbes with regulatory genes for the response to phosphorus starvation. Specifically, *Mesorhizobium* and *Rhizobium* (both harboring *phoB*) showed higher abundance during flowering in sludge-treated soils. *Mesorhizobium*, known for its phosphate solubilization capabilities and its role as a plant growth promoter (Walia et al., 2017), along with *Rhizobium*, are crucial organisms that facilitate both nitrogen (*N*) fixation and phosphorus solubilization (Muleta et al., 2021).

The abundances of genes involved in organic phosphorus mineralization, such as *phnG*, *phnJ*, and *phnL* for bacteria and *phnG* for archaea, varied between germination and flowering stages across fertilization treatments. Notably, *phnG* showed temporal variations, especially during flowering, aligning with greater Olsen phosphorus levels, suggesting that this gene might play a key role in available phosphorus provision in this agroecosystem. This finding is supported by the increased presence of *phnG* gene in soils fertilized with sludge and sludge plus struvite during flowering. These observations reinforce the hypothesis that the phenological stage of maize could influence the abundance of specific uptake-related phosphorus cycle genes related to nutrient uptake (Duchin et al., 2020; Yep et al., 2020). In the taxonomic distribution within the bacterial community, we observed that certain microorganisms hosting genes implicated in organic phosphorus mineralization, including various *C*-single bond-*P* lyases such as *phnI*, *phnJ*, *phnH*, and *phnL*, exhibited phenological variations. For example, *Devosia* had greater *phnI* abundance during germination in *NPK*-treated soils, consistent with prior studies indicating its enrichment in soils with limited phosphorus availability (Gao et al., 2024). In the archaeal community, interestingly, we highlight that, although our study identified *Natrialba* as a host of genes related to organic phosphorus mineralization, previous research has highlighted its role as one of the most significant phosphate-solubilizing archaea (A. N. Yadav et al., 2015), suggesting that this genus may contain genes associated with inorganic phosphorus solubilization that were not identified in our study.

5. CONCLUSIONS

Our research offers fresh insights into the critical microbial taxa, genes, and proteins associated with the phosphorus cycle in Mediterranean agroecosystems, utilizing an integrated approach that combines metagenomics and metaproteomics. Aligning with our hypotheses, we found that the crop's phenological stage plays a more pivotal role than fertilization practices in shaping the relative abundance of phosphorus cycle-related genes, particularly those involved in the solubilization of inorganic phosphorus by bacterial and archaeal communities. While prior studies have acknowledged the influence of phenology on soil microbial communities, our findings stand out in demonstrating that the phenological stage exerts a greater

impact than fertilization practices. This knowledge paves the way for the development of phosphorus fertilization strategies that account more thoroughly for plant phenology.

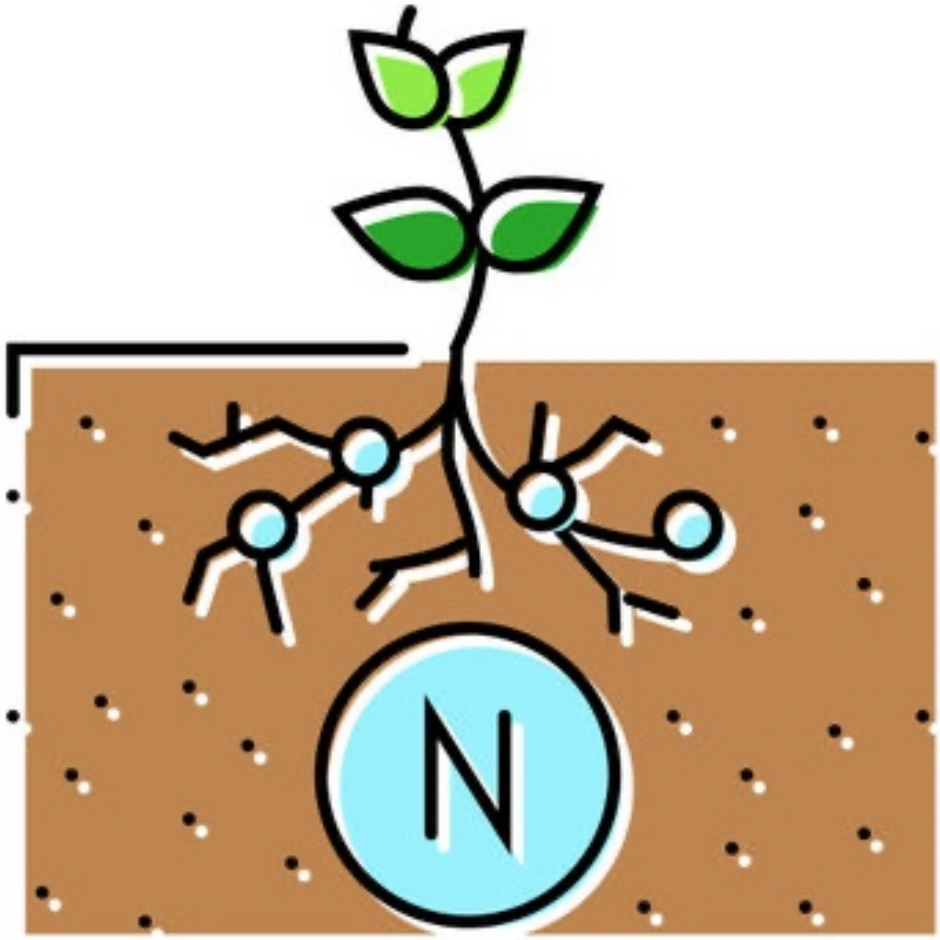
Additionally, our results reveal notable taxonomic clustering of functional processes related to the phosphorus cycle, with plant growth stages exerting a significant influence. For example, within the dominant bacterial populations carrying phosphorus-related genes, microbes harboring genes for the solubilization of inorganic phosphorus were typically distinct from those carrying genes for the mineralization of organic phosphorus. This distinction was particularly evident in members of the phylum Actinobacteria. Moreover, our study underscores the significant impact of phenology on archaeal communities and the associated phosphorus-related genes—a subject that has received relatively little attention in the scientific literature. These findings highlight the promising potential of archaea to contribute to the phosphorus cycle in agroecosystems. Nonetheless, we acknowledge the need to address certain methodological challenges in metaproteomics and the importance of refining metagenome database curation for more accurate insights in multi-omic approaches.

Our work also confirms the initial hypothesis regarding the impact of distinct fertilization treatments on the abundance of critical genes regulating phosphorus cycling. We observed markedly different responses to fertilizers in terms of the abundance of genes linked to organic phosphorus mineralization in both bacterial and archaeal communities. Additionally, fertilizer types influenced genes regulating phosphorus-starvation responses, with these effects being particularly pronounced in bacterial taxa. Importantly, our findings demonstrate complex interactions between fertilizer type and crop phenology, which collectively drive nuanced shifts in the abundance of phosphorus cycling genes across various functional categories in bacterial and archaeal populations.

Lastly, the integration of metaproteomics with metagenomics significantly enhances our understanding of the phosphorus cycle by identifying abundant enzymes that may otherwise be overlooked. For example, the enzyme alkaline phosphatase, encoded by the *phoX* gene, emerged as a key player in phosphorus cycling in this maize agroecosystem. This enzyme, although often disregarded in previous studies, appears to play an essential role in the microbial phosphorus cycle.

CHAPTER 2

Phenology shapes nitrogen cycling more than fertilization: multi-omic evidence of microbial guild specialization in a maize agroecosystem.



CHAPTER 2

1. INTRODUCTION

Nitrogen (N) is an essential macronutrient for plant growth, development and yield (Sun et al., 2020). In plants, nitrogen is a highly required element, as it is used to produce amino acids and nucleotides, the building blocks of proteins and nucleic acids, respectively (Bloom, 2015). In fact, the availability of nitrogen nitrogen from the environment often limits the productivity of agroecosystems. Due to the physicochemical properties of soil, plants often have limited access to this resource, leading to nutrient deficiencies (B. Zhao et al., 2024). Given its importance, large amounts are usually applied as fertilizers, which can lead to problems for ecosystem health. Besides conventional mineral sources of nitrogen, whose prices are increasing, recycled materials can be used therefore contributing to circular economy. Among these sources, byproducts of mineral (i.e., struvite) and organic (i.e., sludges) character are important sources of nitrogen for agroecosystems. In the case of struvite, it is a crystalline mineral compound ($MgNH_4PO_4 \cdot 6H_2O$) composed of magnesium, ammonium, and phosphate (Bastida et al., 2019b). It typically forms in wastewater treatment plants, particularly in the presence of high concentrations of these ions. Struvite contains an important amount of nitrogen (nearly 5%) that can be source of these macronutrient in agroecosystems. Similarly, the content of nitrogen nitrogen in sludge is variable but can be around 5%, with a more organic nature. These byproducts contain not only nitrogen nitrogen but also other essential nutrients and elements, including magnesium (Mg) and phosphorus (P) (Chojnacka et al., 2023; Ha et al., 2023). It has been shown that struvite can be used as a slow-releasing nitrogen-fertilizer for plants (L. Wang et al., 2023), while sludge fertilization leads to an increase in biomass nitrogen (Petersen et al., 2003). Further, the high content of organic carbon in sludges can benefit soil health and promote microbial activity and biomass and the soil organic matter in semiarid soils (J. A. Siles, Gómez-Pérez, et al., 2024).

In addition to the chemical nature of nitrogen fertilizers (i.e., organic vs. mineral) and soil edaphic properties, soil N availability can be influenced by plant growth stages and associated soil microbes (Legay et al., 2020). Plants have different nutritional demands depending on their phenological stage. This prompts plants to develop various dynamic strategies to acquire nitrogen, one of which is forming associations with soil microbes that can modulate the nitrogen cycle (Tao et al., 2019; B. Zhao et al., 2024). In the face of nitrogen scarcity, soil microbial communities respond by regulating the expression of genes involved in denitrification, nitrification, dissimilatory nitrate reduction to ammonium (DNRA), assimilatory nitrate reduction to ammonium (ANRA), N_2 fixation and N transport (Kelly et al., 2021). In this context, soil microbes exhibit their nitrifying potential through the action of genes such as *pmoA-amoA* or *pmoB-amoB* and denitrifying nitrates or nitrites by regulating genes such as *nosZ* or *nosC* (Mosley et al., 2022). It is therefore crucial to emphasize how various byproducts rich in nitrogen, such as struvite and sludge, can modulate soil microbial communities and their contribution to the overall nitrogen cycle (Kelly et al., 2021). For instance, the application of organic amendments has been shown to increase the abundance of denitrification genes, such as *nirK*, which are associated with nitrous oxide (N_2O) emissions (Y. Yang et al.,

2022). Additionally, depending on the environment, organic amendments have also been observed to increase the abundance of nitrification genes, highlighting their role in influencing both nitrification and denitrification processes (Bastida et al., 2009). Similarly, fertilization with struvite has caused changes in the relative abundance of *amoA* and *nosZ* genes, suggesting that struvite can affect both nitrification and denitrification processes (Carreras-Sempere et al., 2024). However, the interaction between the effects of these products on gene abundance and plant growth stages is not adequately considered and warrants further attention, given the complex plant-soil-microbe interactions and the varying plant nitrogen demands throughout the growth cycle. In this sense, employing multi-omics approaches such as metagenomics and metaproteomics, together with the recovery of metagenome-assembled genomes (MAGs), represents a significant advancement in understanding the microbial contribution to soil nitrogen cycling (Bastida et al., 2021; Miller et al., 2023; Starke et al., 2019).

Among these, the analysis of MAGs has emerged as a cutting-edge tool, providing a more comprehensive and high-resolution perspective on microbial community structure and functional potential. MAGs allow the reconstruction of near-complete genomes from complex microbial communities (L.-X. Chen et al., 2020), enabling the identification of specific microbial taxa and their functional roles in the nitrogen cycle with unprecedented precision. The combination of these Meta-omics approaches helps in the identification and quantification of genes and key microbial players potentially involved in the nitrogen cycle, along with proteins that are directly responsible for functional processes (Starke et al., 2019).

In this study, we partially substituted conventional mineral fertilizers with struvite, sludge, and their organo-mineral combination, and investigated the effects of these fertilization strategies on the abundance and taxonomic origin of key functional genes and proteins related to the nitrogen cycle, such as those involved in nitrification, denitrification, DNRA, ANRA, N_2 fixation, and N transport. We further explore how phenological stage of maize interact with soil microorganisms and the potential functionality related to nitrogen cycle. This study was carried out in a maize field trial, given the crop's global importance as the top cereal in terms of production volume and its anticipated role as the most widely cultivated and traded crop in the coming decade (Erenstein et al., 2022). Our research also focused in archaea given their critical role on the N cycle. (Offre et al., 2013).

Given the distinct chemical nature of fertilizers, where struvite contains ammonium, a plant-available form of nitrogen, and sewage sludge contains both ammonium and nitrate and organic nitrogen forms, we hypothesize that: i) these materials will differently impact the abundance of genes involved in nitrogen cycling and the bacterial and archaeal populations that harbor these genes, and ii) the phenological stage of maize will shape the abundance of specific genes and microbes related to nitrogen cycling in the soil, thus influencing the taxonomic clustering of nitrogen-functional processes.

2. MATERIALS AND METHODS

2.1. Experimental setup and sampling

The experiment was carried out during the 2022 maize-growing season in the experimental fields of ITAP (Santa Ana, Albacete, SE Spain; 38°53'39.8" N, 1°59'18.0" W), a semi-arid Mediterranean region, as described in Barquero et al. (2024) (Figure 12), as outlined in Chapter 1. Soil properties prior to the experiment are described in detail in Table 1. A total of 16 plots, each measuring 18.75 m² and separated by 1-meter-wide paths, were established. The maize crop (var. P0937) was sown on May 18, 2022. The experimental design follows a completely randomized block design with four fertilization treatments, each replicated four times. The treatments included: i) mineral NPK fertilization (NPK); ii) organic fertilization with thermostabilized sludge (SLU); iii) mineral fertilization using struvite (STR); and iv) a combination of struvite and sludge (STRSLU). Nutrient compositions of both struvite and sludge are detailed in Table 2. These treatments were applied to partially substitute conventional mineral fertilization. Briefly, struvite contained 0.13, 5.80, and 16.30 g $100g^{-1}$ of organic carbon, total nitrogen, and total phosphorus, respectively, while sludge contained 29.08, 4.92, and 4.14 g $100g^{-1}$ of these nutrients. Fertilizers were applied to satisfy maize's nutritional requirements, approximately 192 kg ha^{-1} of nitrogen (N), 225 kg ha^{-1} of phosphorus (P), and 281 kg ha^{-1} of potassium (K) in all treatments. All nutrients were incorporated during the initial fertilization stage on May 13, 2022, with subsequent nitrogen applications at phenological stages V4 and V8 (June 20 and July 8, 2022, respectively) as per Ritchie, S.W & J.J. Hanway, 1982, classification (see Table 3 for details). For this purpose, superphosphate, potassium sulfate, and calcium ammonium nitrate (NAC27) were applied as required to satisfy plant demands. Irrigation was conducted as needed throughout the crop cycle. Soil samples from rhizosphere were collected during two phenological stages: germination (V1) and flowering (R1) (Ritchie, S.W & J.J. Hanway, 1982), selected as these represent the periods of greatest nitrogen demand. Rhizospheric soil samples were collected by pooling soil from five plants per plot to generate representative samples. Afterwards, samples were sieved to 2 mm and stored at -80ºC for DNA extraction or air-dried for chemical analyses.

2.2. Soil analyses

Water soluble nitrogen (WSN) was measured in a C/N analyzed (Multi N/C 3100, Analytic Jena, Germany). Total nitrogen was assessed using Elemental Analyzer (C/N Flash EA 112 Seres-Leco Truspec) (García-Díaz et al., 2024). Urease activity (E.C. 3.5.1.5) was determined following the method described by Kandeler and Gerber (1988). An aqueous solution of urea (0.48%) was used as the substrate, combined with a 0.06 M borate buffer at pH 10. The NH4+ generated was extracted using a 7.4% KCl solution and quantified through a modified indophenol reaction (Kandeler & Gerber, 1988).

2.3. DNA extraction and shot-gun sequencing

Soil DNA was extracted using the DNeasy PowerSoil kit (Qiagen), following the procedures outlined by the manufacturer. For the preparation of metagenomic libraries, the extracted DNA was fragmented via acoustic shearing using a Covaris S220 ultrasonicator. The fragmented DNA underwent purification, 3'

adenylation, end-repair, and adapter ligation. A limited-cycle PCR was subsequently performed to enrich the libraries. Library validation was achieved using an NGS Kit with an Agilent 5300 Fragment Analyzer (Agilent Technologies), while quantification was carried out using a Qubit 4.0 Fluorometer (Invitrogen). Prior to sequencing, DNA libraries were multiplexed and loaded into the Illumina NovaSeq 6000 platform (Illumina), which was configured for paired-end sequencing (2 × 150 bp). This step was performed in strict accordance with the manufacturer's protocol. Image processing and base calling were managed using NovaSeq Control Software (v1.7), while .bcl files were transformed into demultiplexed .fastq files using the Illumina bcl2fastq software (v2.20) (J. Siles et al., 2024). The sequencing procedures were carried out at Genewiz Europe (Leipzig, Germany), as previously described in Barquero et al. (2024).

2.4. Metagenomic analysis

The metagenomic libraries were processed following the methodology initially described by Žifčáková et al. (2016), with certain adaptations. The complete code used for the metagenomic analysis is available in the repository at https://github.com/mariabelen-bm/Doctoral_Thesis/blob/78fb7483aff71e570633d6b3a8f0bb666fe368de/metagenomics_pipeline_chapters_1_2.sh. The complete pipeline is also shown in Annex 1. Additionally, the Python scripts employed during the metagenomic pipeline are included in Annex 2.

Briefly, quality control of raw reads was performed using FastQC (v0.12.0) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), discarding sequences with a quality score below 30 or length shorter than 50 bp. Normalization using k-mer counting (k = 20) with minimum coverage of 20 was conducted prior to assembly using MEGAHIT (v1.2.9) (D. Li et al., 2015). Assembly quality was evaluated with MetaQUAST (v5.2.0) (Mikheenko et al., 2016), while gene prediction was carried out using FragGeneScan (Rho et al., 2010), and alignment was conducted using Bowtie2 (v2.4.1) (https://bowtie-bio.sourceforge.net/bowtie2/index.shtml). Taxonomic identification was primarily performed using the NCBI nr database (https://www.ncbi.nlm.nih.gov/). For functional annotation, the KEGG database and was employed: KEGG enabled the assignment of sequences to metabolic pathways and biological processes, while dbCAN was specifically utilized for identifying carbohydrate-active enzymes, such as hydrolases and lyases (L. Huang et al., 2018; Kanehisa et al., 2016; Tatusov et al., 2003). Additionally, annotations were linked to nitrogen cycle genes as described in Žifčáková et al. (2016) and (Žifčáková, 2017). The results are presented within functional groups related to the nitrogen cycle. For instance, nitrification includes genes encoding ammonia monooxygenases (*pmoA-amoA*, *pmoB-amoB*, *pmoC-amoC*) and hydroxylamine oxidoreductase (*hao*). Denitrification encompasses genes encoding nitrate reductases (*narI*, *narV*, *narJ*, *narW*, *napA*, *napB*, *napC*), nitrite reductases (*nirK*, *nirS*), nitric oxide reductases (*norB*, *norC*), and nitrous oxide reductase (*nosZ*). Combined nitrification and denitrification category includes genes encoding both nitrate and nitrite oxidoreductases (*narG*, *narZ*, *nxrA*, *narH*, *narY*, *nxrB*). These genes are included in the denitrification/nitrification category because they play crucial roles in both processes. Specifically, *narG* and *narZ* encode the alpha subunits of nitrate reductase, which catalyzes the reduction of nitrate (NO_3^-) to nitrite (NO_2^-), a key step in denitrification. On the other

hand, *nxrA* encodes the alpha subunit of nitrite oxidoreductase, which is involved in the oxidation of nitrite (NO_2^-) to nitrate (NO_3^-), a central reaction in nitrification. Similarly, *narH*, *narY*, and *nxrB* encode additional subunits of these enzymes, supporting their functionality in both pathways. Dissimilatory nitrate reduction to ammonium (DNRA) features genes encoding cytochrome c nitrite reductase (*nrfA*, *nrfH*) and assimilatory nitrite reductases (*nirB*, *nirD*). Assimilatory nitrate reduction to ammonium (ANRA) includes genes encoding nitrate and nitrite reductases (*nasC*, *nasA*, *narB*, *nirA*). N_2 fixation encompasses genes encoding nitrogenase components (*nifH*, *nifD*, *nifK*)**.** N transport involves genes encoding nitrate/nitrite transporters (*nrtA*, *nasF*, *cynA*, *nrtB*, *nasE*, *cynB*, *nrtC*, *nasD*, *nrtD*, *cynD*). Lastly, general N metabolism includes genes encoding glutamine synthetase (*glnA*, *GLUL*), glutamate dehydrogenase (*GLUD1_2*, *gdhA*), and nitric oxide dioxygenase (*NAO*) (Kelly et al., 2021). The details of each gene can be seen in Table 5. Due to the intricate nature of the dataset, the taxonomic analyses focused on the dominant microbial populations associated with genes significantly influenced by fertilization treatments and/or phenological stages. The raw sequencing data have been uploaded to NCBI and are available under the BioProject accession number PRJNA1118481.

2.5. Analysis of metagenome-assembled genomes (MAGs)

The analysis of MAGs was carried out using a pipeline specifically designed for reconstructing microbial genomes from environmental samples, following methodologies adapted from several established approaches. The complete code used for the analysis of MAGs is available in the repository at https://github.com/mariabelen-bm/Doctoral_Thesis/blob/78fb7483aff71e570633d6b3a8f0bb666fe368de/MAGs_pipeline_chapter_2.sh. A detailed explanation of the steps is also provided in Annex 3.

The analysis was performed in a command-line environment (Shell), utilizing a combination of specialized bioinformatics tools and software for MAG reconstruction, quality assessment, taxonomic and functional annotation, and relative abundance quantification. The workflow began with the grouping of assembled contigs into bins, representing individual microbial genomes, using the MetaWRAP pipeline (Uritskiy et al., 2018), which combines algorithms such as MetaBAT2, MaxBin2, and CONCOCT (Alneberg et al., 2013; Kang et al., 2019; Y.-W. Wu et al., 2014). To ensure high-quality results, bins were refined by selecting those with a completeness score of ≥50% and contamination of ≤10%. Quality assessment was further carried out with CheckM2 (Chklovski et al., 2023), which validated the completeness and contamination of the refined MAGs. Taxonomic classification was performed using GTDB-Tk (Chaumeil et al., 2022) with the Genome Taxonomy Database (GTDB) version 2.4.0, providing a standardized taxonomic framework. To verify the genomic consistency of the MAGs and rule out contamination, the Genomic UNcertainty Calculator (GUNC) was employed (Orakov et al., 2021). The relative abundance of MAGs in each sample was determined by mapping metagenomic reads to the reconstructed genomes using CoverM with the minimap2-sr algorithm (H. Li, 2018). Additionally, functional annotation of the MAGs was conducted with

the DRAM pipeline (Shaffer et al., 2020), which assigned metabolic functions based on multiple databases, including KOfam and dbCAN2 (Aramaki et al., 2020; H. Zhang et al., 2018).

2.6. Protein extraction and LC-MS analysis

Protein extraction and mass spectrometry followed protocols previously optimized for similar studies (Barquero et al., 2024; Bastida et al., 2014b; Chourey et al., 2010b). Proteins were extracted from soil samples via SDS buffer boiling, resolved by 12% SDS-PAGE, and visualized using colloidal Coomassie staining. Proteins were subsequently reduced, alkylated, and digested with trypsin, followed by peptide desalting. Peptides were analyzed using nanoHPLC-MS/MS (Thermo Fisher Scientific) as described in Bastida et al. (2016). Briefly, 1 µg of peptides was injected into a Vanquish Neo nanoHPLC coupled to an Orbitrap Exploris™ 480 mass spectrometer. Peptides were trapped and separated on C18 reverse-phase columns with a two-step gradient (4%-30% B over 95 minutes, followed by 30%-55% B over 40 minutes; mobile phase B = 80% acetonitrile with 0.01% formic acid). Data were acquired with a resolution of 120,000 for MS1 and 15,000 for MS/MS. Processing was performed using Proteome Discoverer (v2.5.0.400) with SequestHT and FDR-controlled at 1%. Results were annotated using KEGG, and protein abundances were normalized as previously described in Barquero et al. (2024). Raw proteomics data have been deposited under PRIDE accession PXD052073.

2.7. Statistical analysis

To evaluate the impact of fertilizer treatments and phenological stages on the abundance of functional genes and the populations that harbor them, an ANOVA was performed in R (R-Core-Team, 2023) using the “stats” package. Previously, the normality and heteroscedasticity of the data were checked. To evaluate the effects of fertilization treatments and phenological stages on the functional structure of bacterial and archaeal communities-including genes involved in denitrification/nitrification, denitrification, nitrification, DNRA, ANRA, N2 fixation, N transport, and general N metabolism, the abundance data were subjected to non-metric non-restricted multidimensional scaling (NMDS) (Borcard et al., 2018). Prior to NMDS analysis, abundances were log-transformed to base 10, and ordination was performed using the Bray-Curtis dissimilarity index via the metaMDS() function of the “vegan” package (Oksanen et al., 2019) in R. Pearson correlation analysis was conducted to explore linear relationships among genes involved in the N cycle. The Pearson correlation matrix was computed using the cor() function from the “corrplot” package (Taiyun, 2017) in R. Subsequently, the significance of the observed correlations was assessed using the cor.mtest() function. A 95% confidence level was applied to evaluate statistical significance. For the MAGs analysis, statistical tests were conducted to determine significant differences between the abundance of different treatments. Initially, the Shapiro-Wilk test was applied to assess the normality of the abundance data for each genomic bin and treatment combination (Hanusz et al., 2016). A Kruskal-Wallis test was performed to evaluate significant differences in abundance across treatments for each genomic bin (McKight & Najab, 2010). To further explore these differences, a post-hoc Dunn’s test (Dinno, 2024) was conducted on the bins with significant Kruskal-Wallis results. Adjusted p-values were calculated

using the Bonferroni correction to account for multiple comparisons. Heatmaps and bar charts were generated for data visualization using the "ggplot2" package (Wickham, 2016).

Table 5: Details on the 35 functional genes studied in the present work related to the nitrogen cycle.

Pathway	Gene	Enzyme	KEGG ID
Nitrification/Denitrification	*narG, narZ, nxrA*	Nitrate/nitrite oxidoreductase	K00370
	narH, narY, nxrB	Nitrate/nitrite oxidoreductase	K00371
Denitrification	*napA*	Periplasmic nitrate reductase	K02567
	napB	Periplasmic nitrate reductase	K02568
	napC	Periplasmic nitrate reductase	K02569
	narJ, narW	Nitrate reductase assembly protein	K00373
	narI, narV	Nitrate reductase catalytic subunit	K00374
	nirK	Copper-containing nitrite reductase	K00368
	nirS	Cytochrome cd1 nitrite reductase	K15864
	norB	Nitric oxide reductase subunit B	K04561
	norC	Nitric oxide reductase subunit C	K02305
	nosZ	Nitrous oxide reductase	K00376
Nitrification	*hao*	Hydroxylamine oxidoreductase	K10535
	pmoA-amoA	Particulate methane/ammonia monooxygenase subunit A	K10944
	pmoB-amoB	Particulate methane/ammonia monooxygenase subunit B	K10945
	pmoC-amoC	Particulate methane/ammonia monooxygenase subunit C	K10946
DNRA	*nirB*	Assimilatory nitrite reductase large subunit	K00362
	nirD	Assimilatory nitrite reductase small subunit	K00363
	nrfA	Cytochrome c nitrite reductase	K03385
	nrfH	Cytochrome c nitrite reductase	K15876
ANRA	*narB*	Assimilatory nitrate reductase	K00367
	nasB	Assimilatory nitrate reductase	K00360
	nasC, nasA	Nitrate transporter	K00372
	nirA	Ferredoxin-nitrite reductase	K00366
N_2 Fixation	*nifD*	Nitrogenase molybdenum-iron protein alpha chain	K02586
	nifH	Nitrogenase iron protein	K02591
	nifK	Nitrogenase molybdenum-iron protein beta chain	K02588
N Transport	*nrtB, nasE, cynB*	Nitrate/nitrite transporter subunit	K15577
	nrtA, nasF, cynA	Nitrate/nitrite transporter subunit	K15576
	nrtC, nasD	Nitrate/nitrite transporter subunit	K15578
	nrtD, cynD	Nitrate/nitrite transporter subunit	K15579
General N Metabolism	*gdhA*	Glutamate dehydrogenase	K00261
	glnA, GLUL	Glutamine synthetase	K01915
	GLUD1_2, gdhA	Glutamate dehydrogenase	K00262
	nao	Nitric oxide dioxygenase	K03384

3. RESULTS

3.1. Total nitrogen, WSN, and Urease activity

Total nitrogen content did not show statistically significant differences in relation to fertilization treatments, phenological stages, or their interaction. In contrast, WSN exhibited statistically significant differences among fertilization treatments (F = 12.3, P = 0.001). For urease activity, phenology (F = 4.18, P = 0.043) and the interaction between fertilization and phenology (F = 6.11, P = 0.005) were found to be statistically significant (Figure 22). Regarding WSN, the NPK treatment displayed the highest WSN content in both phenological stages, while the struvite plus sludge treatment showed the lowest values of WSN (Figure 22). In the case of urease activity, the struvite plus sludge treatment during germination exhibited the highest urease activity, whereas the same fertilization treatment during flowering recorded the lowest urease activity.

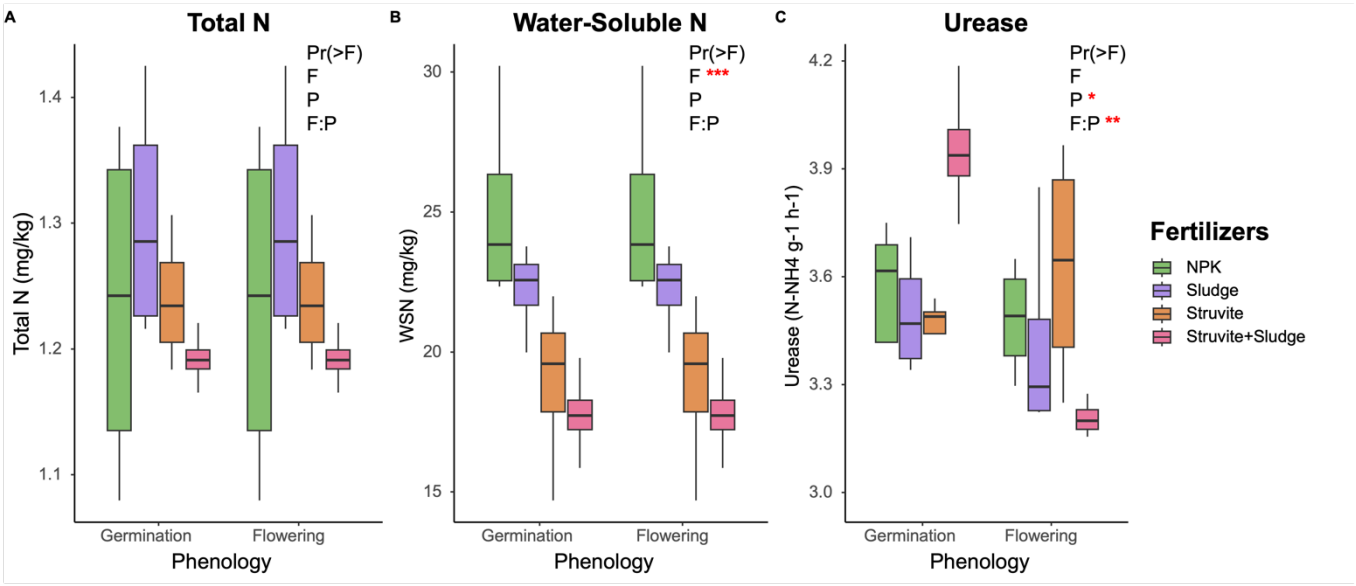


Figure 22: Total nitrogen content, WSN and urease activity during germination and flowering in soils supplemented with the four fertilizers: NPK, Sludge, Struvite and Struvite + Sludge. A) Total nitrogen, B) WSN, C) Urease. The ANOVA test carried out to check if there were significant (P < 0.05) effects between fertilizers (F), phenology (P) and the interaction between fertilization and phenology (F:P) has been added to the boxplot.

3.2. The abundance of nitrogen genes in the bacterial and archaeal community

The total number of reads with annotated genes was 813,089. Of these, only 152 were associated with fungi, and only 4 of them corresponded to nitrogen-related genes, so no further inquiry was made in this domain. Of the total reads, 15,083 were attributed to bacteria containing nitrogen-cycling-related genes, while the count for archaea was 302. N gene readings in the bacterial community were classified into eight categories: denitrification/nitrification, denitrification, nitrification, DNRA, ANRA, N_2 fixation, N transport,

and general N metabolism. In contrast, the archaeal community's nitrogen gene readings were solely classified into two categories: denitrification and nitrification (Figure 23).

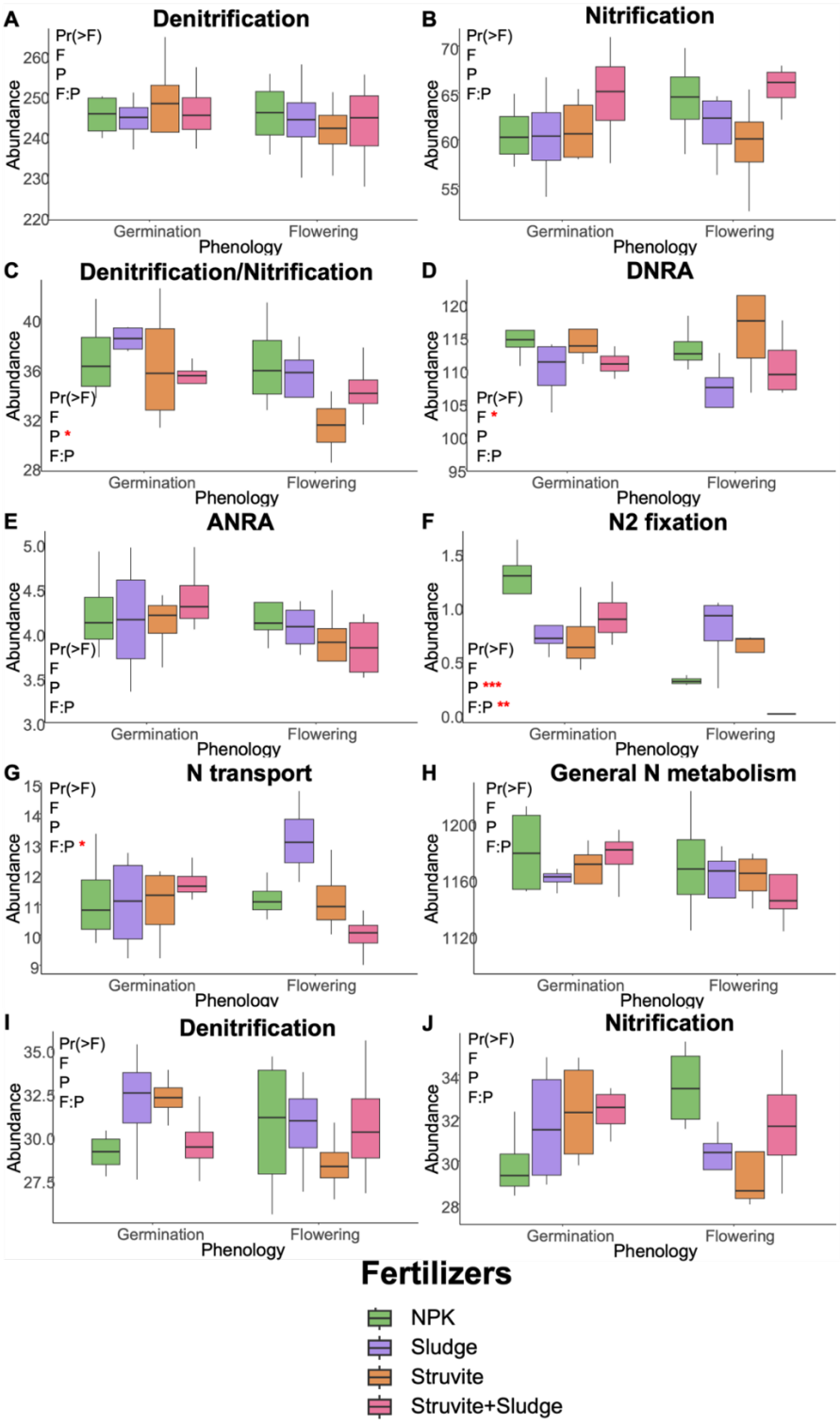


Figure 23: Abundance of nitrogen-related genes in soil bacteria and archaea under different fertilization treatments (NPK, Sludge, Struvite, and Struvite + Sludge) during germination and flowering. Panels (A) to (H) represent bacterial gene abundances for denitrification (A), nitrification (B), denitrification/nitrification (C), DNRA (D), ANRA (E), N_2 fixation (F), N transport (G), and general N metabolism (H). Panels (I) and (J) show archaeal gene abundances for denitrification (I) and nitrification (J). Results from an ANOVA test assessing the significant effects ($P < 0.05$) of fertilization (F), phenological stage (P), and their interaction (F:P) are included within the boxplots.

In the bacterial community, fertilization treatments only influenced the abundance of genes in the DNRA category (Figure 23D). For instance, the abundance of genes involved in this category was higher in the NPK and struvite fertilization treatments compared to the other organic fertilization, both during germination and flowering. Additionally, in the bacterial community, maize phenology influenced the abundance of genes in the denitrification/nitrification and N_2 fixation categories (Figure 23C and Figure 23F). A decrease in the abundance of genes in the denitrification/nitrification category was observed in the sludge, struvite, and struvite-plus-sludge fertilizer treatments during flowering. In contrast, gene abundances in the NPK treatment remained stable across both phenological stages. Further, the abundance of genes involved in N_2 fixation decreased significantly in flowering (Figure 23F). Finally, the interaction between fertilizer treatments and phenology influenced the abundance of genes in the N_2 fixation and N transport categories in the bacterial community (Figure 23F and Figure 23G). In the N_2 fixation category, a decrease in gene abundance was observed during flowering, particularly in the NPK and struvite plus sludge fertilizer treatments. Conversely, in the sludge fertilizer treatment, gene abundance was higher during flowering than during germination, contrary to the pattern observed in the other fertilizer treatments. In the N transport category, gene abundance in the sludge fertilizer treatment was higher during flowering than during germination, whereas in the struvite plus sludge fertilizer treatment, gene abundance was higher during germination than during flowering. In the archaeal community, no significant differences were found in the abundance of genes in the two studied categories, neither between fertilizer treatments nor between phenological stages.

Overall, within the bacterial community, genes involved in denitrification/nitrification, denitrification, nitrification, and DNRA were observed to be more abundant than those involved in N_2 fixation, N transport, and general N metabolism. However, the *glnA, GLUL* gene, which falls under general N metabolism, exhibited a notably high abundance (Figure 24). This gene encodes glutamine synthetase, an enzyme that catalyzes the ATP-dependent conversion of glutamate and ammonia into glutamine, playing a central role in N assimilation and recycling (Eisenberg et al., 2000). In contrast, within the archaeal community, genes related to general N metabolism were more prevalent, with the *glnA, GLUL* gene again being the most abundant (Figure 24).

Among the bacterial genes associated with denitrification/nitrification, the *narH, narY, nxrB* genes were significantly influenced by phenology ($P < 0.05$), showing higher abundance during germination compared to flowering. Within the denitrification category, the *napA* gene was significantly affected by phenology and by the interaction between fertilizer treatments and phenology, while the *nisS* and *nosZ* genes were significantly influenced by fertilizer treatments ($P < 0.05$). A higher abundance of the *napA* gene was observed in the NPK, sludge, and struvite plus sludge fertilizer treatments during germination. For the *nisS* gene, higher abundance was noted in the struvite plus sludge fertilizer treatment, particularly during flowering. In contrast, the nosZ gene showed higher abundance in the NPK fertilizer treatment during germination. In the nitrification category, the *pmoA-amoA* gene was significantly influenced by fertilizer treatments, while the *pmoB-amoB* gene was influenced by phenology ($P < 0.05$).

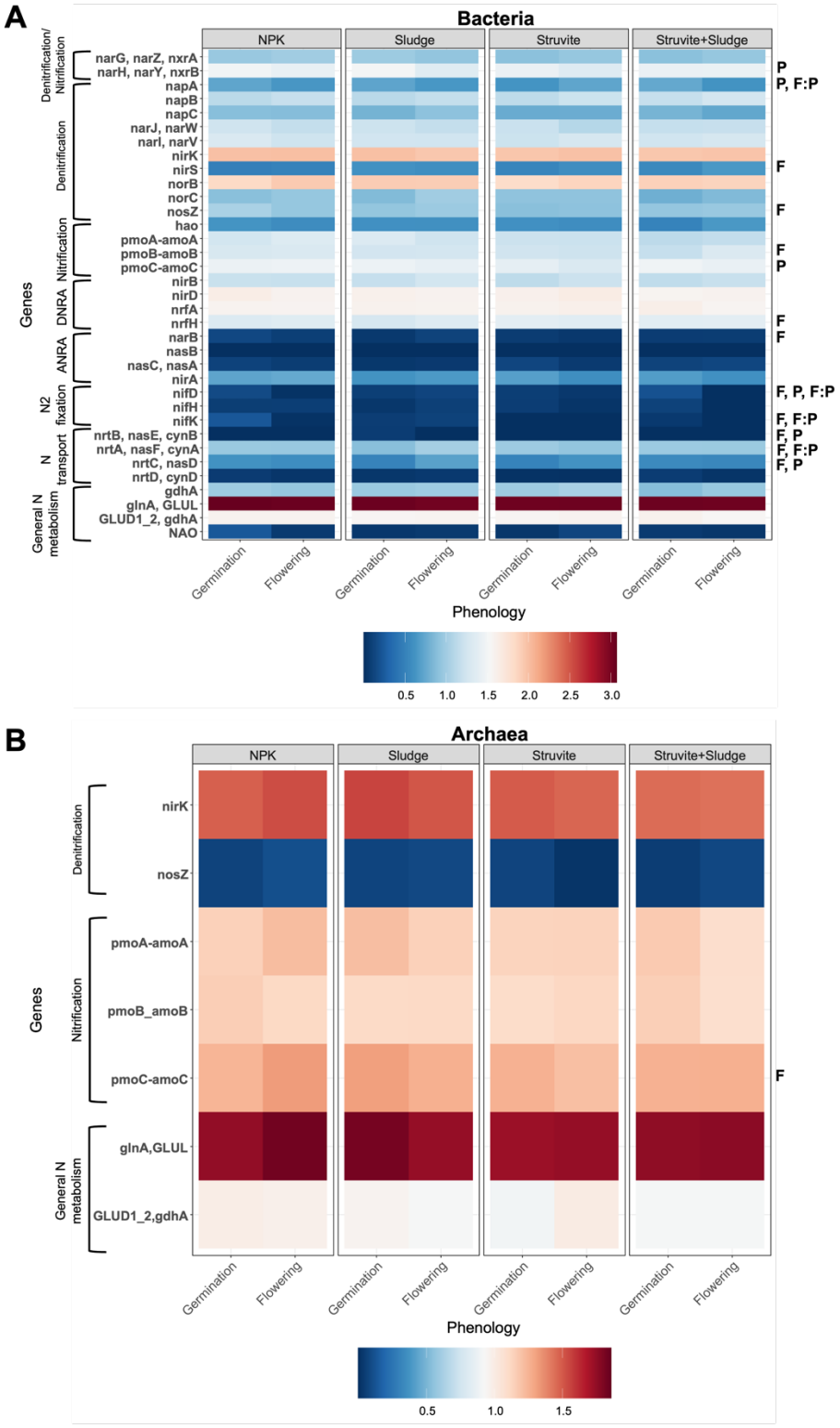


Figure 24: Heatmap showing the base-10 logarithm of nitrogen-related gene abundances, grouped into functional categories: denitrification/nitrification, denitrification, nitrification, DNRA, ANRA, N_2 fixation, N transport, and general N metabolism in bacteria, as well as denitrification, nitrification, and general N metabolism in archaea. Soils were supplemented with four fertilizers (NPK, Sludge, Struvite, and Struvite + Sludge) during germination and flowering. Panel (A) represents the bacterial community, while panel (B) corresponds to the archaeal community. Results from an ANOVA test evaluating the significant effects ($P < 0.05$) of fertilization (F), phenological stage (P), and their interaction (F:P) are included within the boxplots.

The *pmoA-amoA* gene exhibited higher abundance in the NPK, sludge, and struvite treatments, whereas the *pmoB-amoB* gene was more abundant in the sludge, struvite, and struvite plus sludge fertilizer treatments during germination. In the DNRA category, the *nrfH* gene was significantly influenced by fertilizer treatments ($P < 0.05$), showing higher abundance in the NPK, struvite, and struvite plus sludge treatments. In the ANRA category, the *narB* gene was significantly influenced by fertilizer treatments ($P < 0.05$), with higher abundance observed in the NPK, sludge, and struvite fertilizer treatments. Regarding N_2 fixation, the *nifD* gene was significantly influenced by fertilizer treatments, phenology, and the interaction between fertilization and phenology, while the *nifK* gene was significantly affected by fertilizer treatments and the interaction between fertilization and phenology ($P < 0.05$). The *nifD* gene displayed higher abundance in the NPK, struvite, and struvite plus sludge treatments during germination, whereas the *nifK* gene was more abundant during germination in the NPK and struvite plus sludge treatments. In the N transport category, the *nrtB, nasE, cynB* gene was significantly influenced by fertilizer treatments and phenology, while the *nrtC, nasD* gene was influenced by fertilizer treatments and phenology. Additionally, the *nrtA, nasF, cynA* genes were significantly impacted by fertilizer treatments and the interaction between fertilization and phenology ($P < 0.05$). The *nrtB, nasE, cynB* gene exhibited higher abundance in the sludge fertilizer treatment during germination, while the *nrtC, nasD* gene was more abundant in the sludge and struvite fertilizer treatments during flowering. The *nrtA, nasF, cynA* genes showed higher abundance in the sludge fertilizer treatment during flowering. Within the archaeal community, in the nitrification category, the *pmoB-amoB* gene was significantly influenced by fertilizer treatments, with higher abundance observed in the sludge and struvite fertilizer treatments ($P < 0.05$).

3.3. Taxonomic distribution of nitrogen cycle genes in bacterial communities across treatments and phenology

Genes associated with the nitrogen cycle in soil were linked to taxa across 19 bacterial phyla, with Acidobacteria, Actinobacteria, and Proteobacteria being particularly abundant (Figure 25). Figure 25 highlights the most dominant taxa containing bacterial nitrogen-cycle genes. This figure focuses on genes that exhibited significant differences in abundance across fertilization, phenology, and/or their interaction, as shown in Figure 24. Notably, we found significant functional clustering among the taxa. A pattern emerged where dominant populations carried nitrification genes but were not prevalent (or even absent) when it came to denitrification genes. However, a few exceptions were observed. For instance, certain populations like *Bradyrhizobium* (Proteobacteria) and *Ornithinibacter* (Actinobacteria) possessed both denitrification and nitrification genes (Figure 25).

In the denitrification/nitrification category, we found that bacterial genera such as *Solirubrobacter*, *Rubrobacter*, *Ornithinibacter*, *Nocardioides* (Actinobacteria), and *Bacillus* (Firmicutes) were predominant in harboring the gene *narH, narY, nxrB*, with *Nocardioides* being the most abundant genus, particularly prevalent during germination for these genes. In the denitrification category, bacterial genera such as *Steroidobacter* and *Thalassospira* (Proteobacteria) were identified as carriers of the *napA* and *nirS* genes, while the genera harboring *nosZ* were more diverse, including members from the phyla Verrucomicrobia

(*Opitutus*), Proteobacteria (*Microvirga*), Ignavibacteriae (*Melioribacter*), Gemmatimonadetes (*Gemmatimonas*), Firmicutes (*Planifilum*), Chloroflexi (*Anaerolinea*), Bacteroidetes (*Flavitalea*), and Acidobacteria (*Luteitalea*). In the nitrification category, the bacterial genera *Nitrosospira*, *Nitrosovibrio*, and *Nitrosomonas* (Proteobacteria) were the most abundant, harboring the *pmoB-amoB* and *pmoC-amoC* genes. Specifically, the abundance of the bacterial genus *Nitrosospira* in the *pmoB-amoB* gene was higher in the NPK and sludge fertilization treatments across both phenological stages. In contrast, the *pmoC-amoC* gene was more abundant in the struvite and struvite plus sludge treatments, regardless of phenology. In the DNRA category, a wide diversity of bacterial phyla harboring the *nrfH* gene was observed, with *Sorangium*, *Vitiosangium*, and *Myxococcus* (Proteobacteria) being the most abundant. The *narB* gene, belonging to the ANRA category, was found exclusively in a single bacterial population, *Ohtaekwangia*, from the phylum Bacteroidetes. Similarly, the *nifD* and *nifK* genes, part of the N_2 fixation category, were harbored by *Skermanella* in the case of *nifD*, and by *Methylohalobius* and *Azotobacter* in the case of nifK, with all three populations belonging to the phylum Proteobacteria. In the N transport category, we found that the genera *Nitrospira* (*Nitrospirae*), *Azospirillum*, and *Rhodoplanes* (Proteobacteria) were the most abundant carriers of the n*rtA, nasF, cynA* gene, with *Nitrospira* being more abundant in the NPK treatment during flowering. The most abundant bacterial genus harboring the *nrtC*, *nasD* gene was *Piscinibacter* (Proteobacteria). Lastly, it is noteworthy that the *nrtB, nasE, cynB* gene was exclusively found in the genus *Pirellula* (Plantomycetes).

3.4. Taxonomic distribution of nitrogen cycle genes in archaeal communities across treatments and phenology

Despite the relatively small number of genes associated with archaea, we successfully assigned them to various taxa, revealing distinct patterns. Genes involved in the nitrogen cycle in soil were attributed to taxa across three archaeal phyla, with Thaumarchaeota being particularly abundant (Figure 26). Figure 26 illustrates the abundance of nitrogen cycle genes in the most dominant archaeal populations for each gene. The *nirK* gene, involved in denitrification, was exclusively found in taxa belonging to the phylum Thaumarchaeota, while the *nosZ* gene, also associated with denitrification, was represented solely by the archaeal genus *Halogranum*, from the phylum Euryarchaeota, with higher abundance in the struvite fertilization treatment during germination. The nitrification genes *pmoA-amoA*, *pmoB-amoB*, and *pmoC-amoC* in the archaeal community were harbored by the archaeal genera *Nitrososphaera*, *Nitrosopumilus*, and *Candidatus Nitrosocosmicus*, all belonging to the phylum Thaumarchaeota. Additionally, the *pmoA-amoA* gene was also found in an archaeal genus from the phylum Crenarchaeota.

Notably, the genus *Candidatus Nitrosocosmicus*, which harbored the *pmoC-amoC* gene, showed greater abundance during the flowering stage. On the other hand, the *glnA, GLUL* gene, associated with general N metabolism, was widely represented across different archaeal populations, including members of Crenarchaeota and Euryarchaeota.

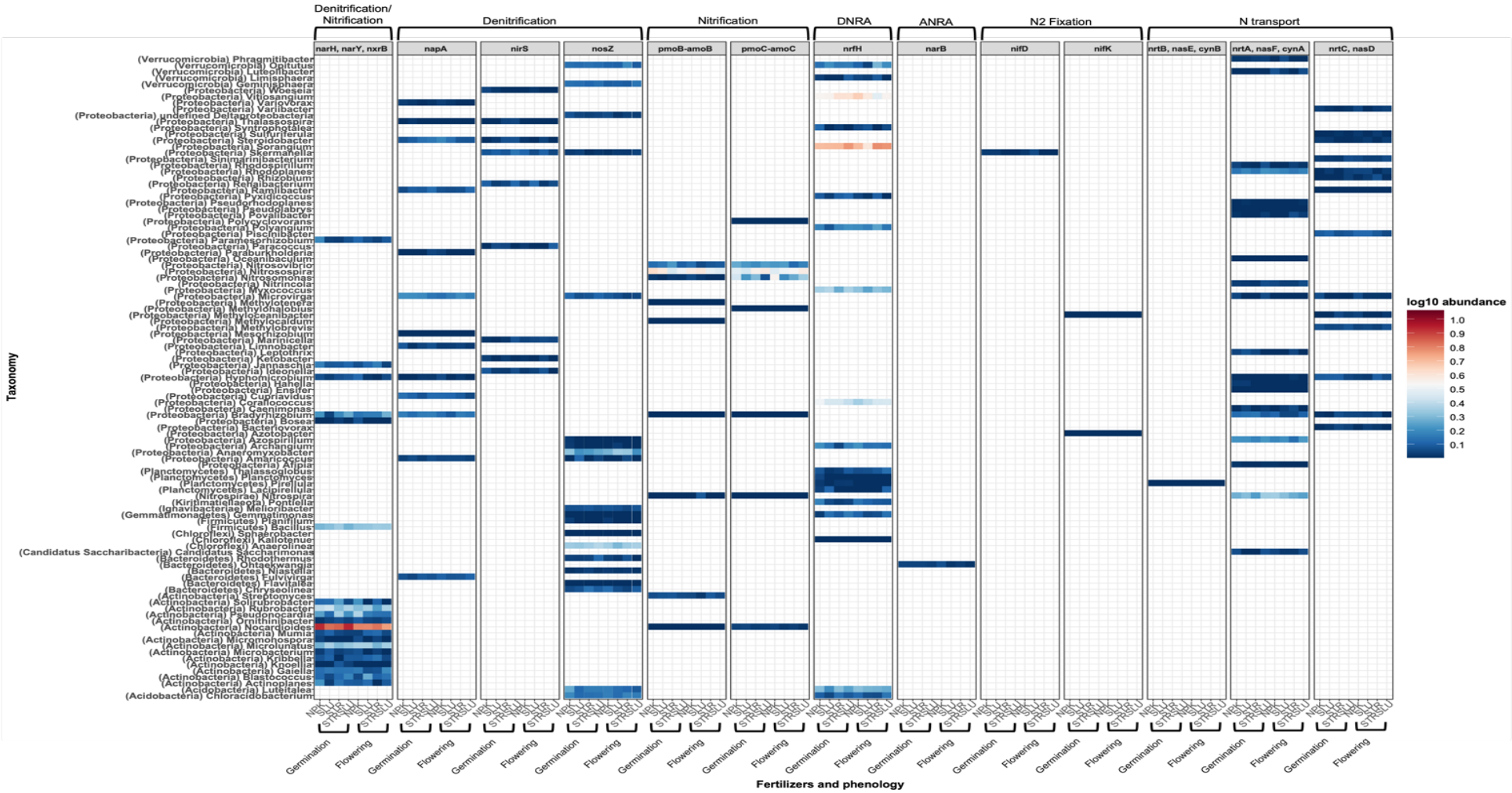


Figure 25: Heatmap of the logarithm in base 10 of bacterial taxonomic abundance in nitrogen genes grouped into denitrification/nitrification, denitrification, nitrification, DNRA, ANRA, N_2 fixation, N transport, and general N metabolism categories in soils supplemented with the four fertilizers during germination and flowering. Fertilizers on the x-axis are abbreviated: NPK, SLU (Sludge), STR (Struvite) and STRSLU (Struvite+Sludge). Taxonomy has been grouped into phylum and genus. The image shows the genes that showed significant differences between fertilizers or between phenology.

These findings contrast with another gene in this category, *GLUD1_2, gdhA*, which was exclusively harbored by *Nitrososphaera* and *Candidatus Nitrosocosmicus*, both from the phylum Thaumarchaeota. The genus *Candidatus Nitrosocosmicus*, harboring the *glnA, GLUL* gene, was more abundant in the struvite fertilization treatment at both phenological stages, while *Nitrososphaera*, which carried the *GLUD1_2, gdhA*, was more abundant during flowering.

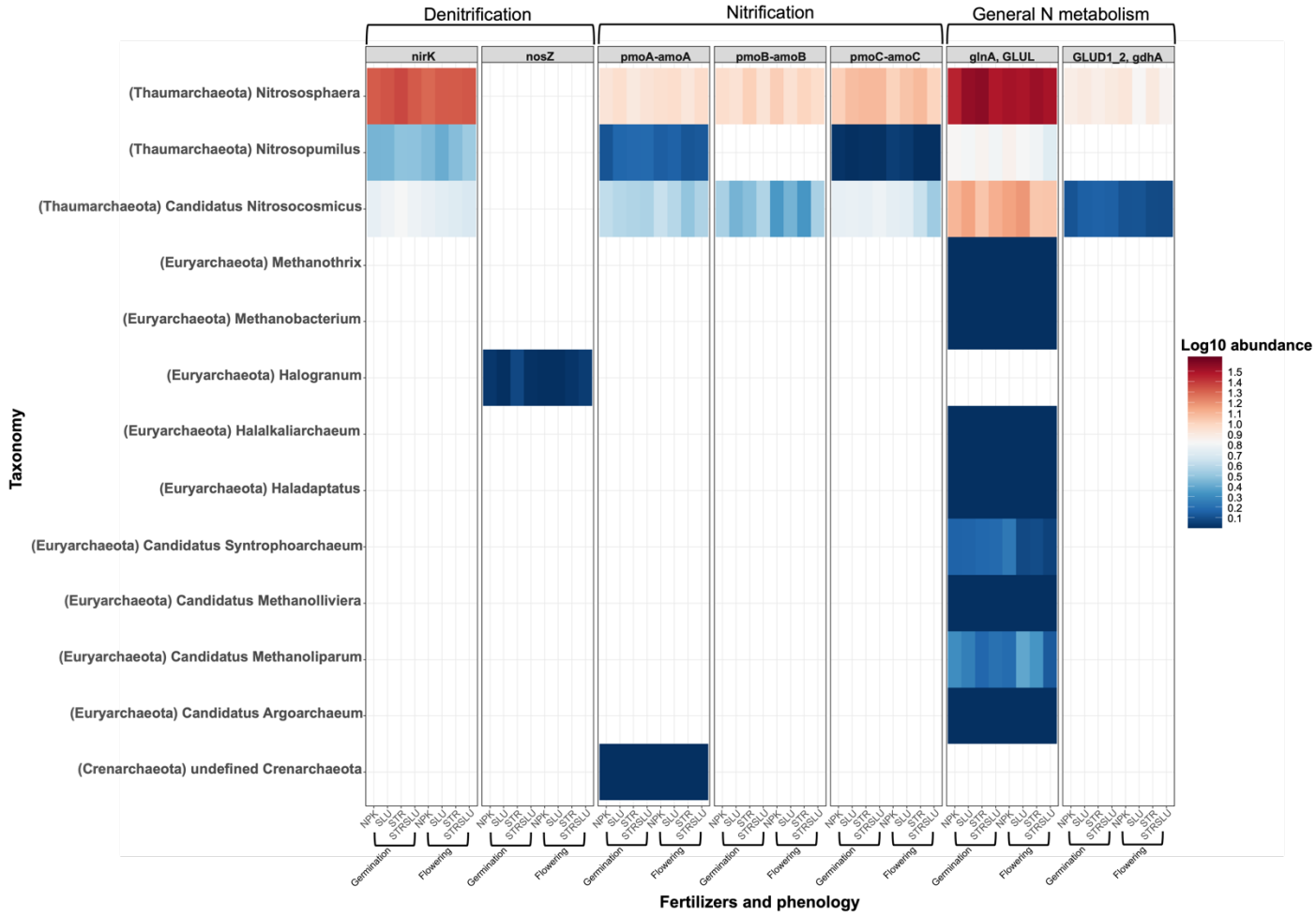


Figure 26: Heatmap of the logarithm in base 10 of archaeal taxonomic abundance in nitrogen genes grouped into denitrification/nitrification, denitrification, nitrification, DNRA, ANRA, N_2 fixation, N transport, and general N metabolism categories in soils supplemented with the four fertilizers during germination and flowering. Fertilizers on the x-axis are abbreviated: NPK, SLU (Sludge), STR (Struvite) and STRSLU (Struvite+Sludge). Taxonomy has been grouped into phylum and genus. The image shows the genes that showed significant differences between fertilizers or between phenology. Only the 15 most abundant bacterial genera for each gene are displayed.

3.5. Abundance and microbial origin of identified proteins

Metaproteomics enabled the detection and quantification of 311 proteins, with only 1.01% related to the nitrogen cycle. The identified nitrogen-related enzymes were *nirK*, *pmoB-amoB*, and *glnA, GLUL*. Among these, no significant influence on abundance was observed due to fertilization treatments, phenology, or the interaction between fertilization and phenology (Figure 27).

Regarding the taxonomic distribution of these enzymes, we observed that *nirK* was predominantly hosted by two archaeal genera, *Nitrososphaera* and *Candidatus Nitrosocosmicus* (Nitrososphaerales), as well as by a bacterial genus, *Nitrosospira (*Nitrosomonadales*)*. These three genera are highly represented in the metagenome, with the two archaeal genera harboring the *nirK* gene in the metagenome (Figure 27).

The enzyme *pmoB-amoB* was found in *Nitrosospira* and *Nitrosomonas* (Nitrosomonadales), both genera belonging to the family Nitrosomonadaceae (phylum Proteobacteria), which also appeared in the metagenome, harboring the *pmoB-amoB* gene (Figure 25). Lastly, the *glnA, GLUL* enzyme was found across a wide diversity of bacterial and archaeal genera. The abundance of this enzyme in the proteobacterial genera *Reyranella* (Hyphomicrobiales) and *Sphingomonas* (Sphingomonadales) was significantly influenced by fertilization treatments (P = 0.001), phenology (P = 0.001), and the interaction between fertilization and phenology (P = 0.001). The abundance of the proteobacterial genus *Skermanella* (Rhodospirillales) was also significantly affected by fertilization treatments (P = 0.041) (Figure 27).

3.6. Taxonomic and functional characterization of reconstructed microbial genomes (MAGs)

A total of 35 bins were recovered, representing the reconstructed genomes of 35 microorganisms. Among these, three bins were discarded due to exceeding thresholds of contamination and completeness. In the following link, a table can be found with all the MAGs' information (https://github.com/mariabelen-bm/Doctoral_Thesis/blob/fdb9e4d131bdc1006035d00eb8cfd327566e7665/table_MAGs_chapter_2.xlsx).

From the remaining 32 bins of interest, taxonomic classification at the genus level was achieved for 27 MAGs. Six bins showed statistically significant differences between fertilization treatments, corresponding to microorganisms from the following classes: Gammaproteobacteria (bin 2), Acidimicrobiia (bin 2), UBA4738 (bin 9), Blastocatellia (bin 15), Actinomycetes (bin 21), and Binatia (bin 35). Specifically, bins bin.2_Gammaproteobacteria, bin.3_ Acidimicrobiia, bin.9_ UBA4738, and bin.21_ Actinomycetes were more abundant under the Struvite+Sludge treatment, while bins bin.15_ Pyrinomonadaceae and bin.35_UBA9968 were more abundant under the Struvite-only treatment (Figure 28).

Among the reconstructed genomes, archaeal taxa such as the genus *Nitrososphaera* and bacterial genera including *Nitrosospira*, *Luteolibacter*, *Rubrobacter*, *Nitrospira*, and *Arthrobacter* were identified as being determinant in the metagenome. Additionally, the archaeal family Nitrososphaeraceae and bacterial families including Nitrosomonadaceae, Akkermansiaceae, Rubrobacteraceae, Nitrospiraceae, Propionibacteriaceae, Bacillaceae, Steroidobacteraceae, Solirubrobacteraceae, Micrococcaceae, and Pyrinomonadaceae were detected in the metagenome (Figure 29).

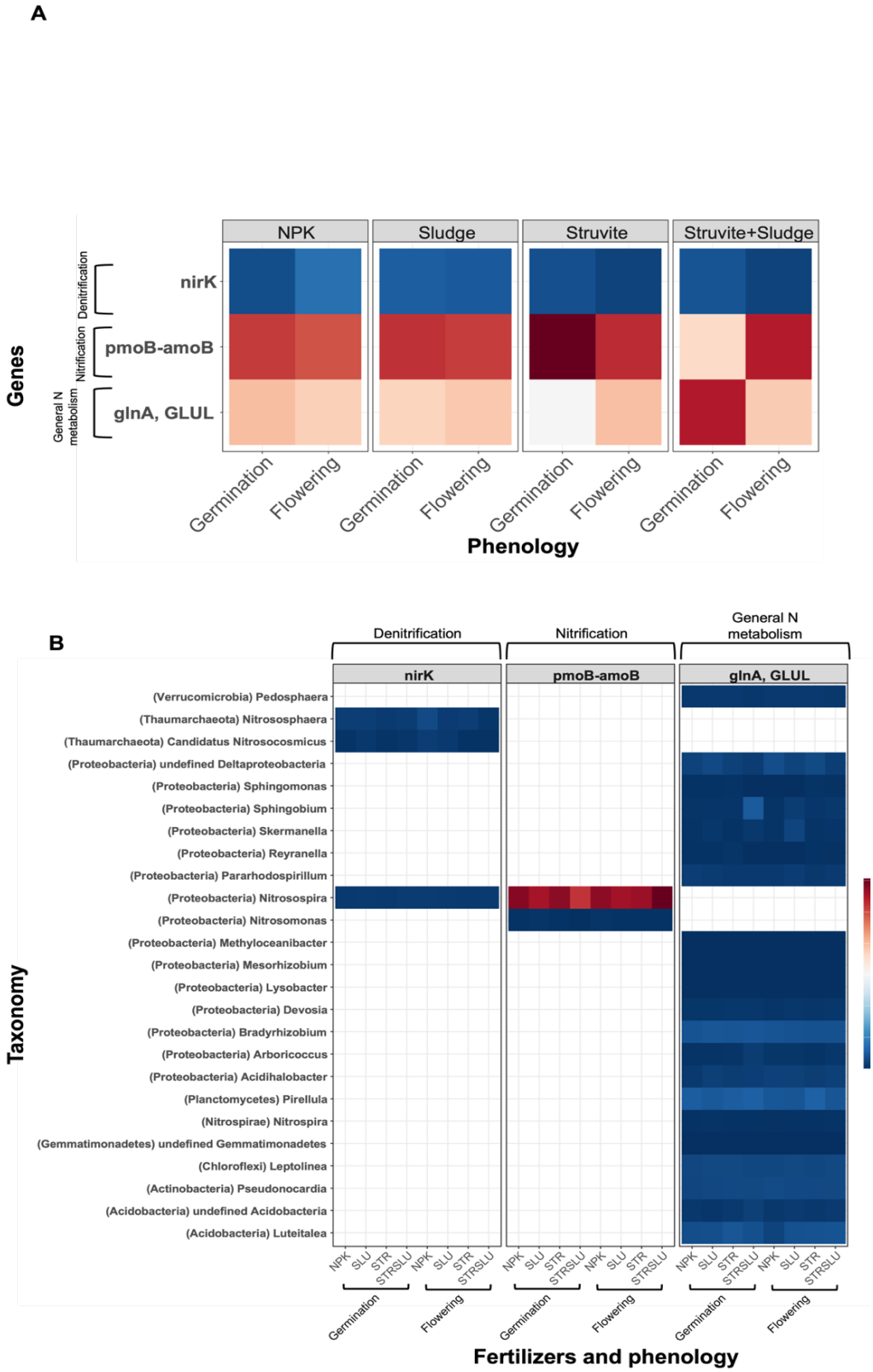


Figure 27: Heatmap of the logarithm in base 10 of the results obtained in proteomics. A) Abundance of nitrogen genes identified in metaproteomics grouped in the categories of denitrification, nitrification and general N metabolism in the different soils supplemented with the four treatments (NPK, Sludge, Struvite and Struvite+Sludge) during germination and flowering; B) Abundance of bacterial and archaeal populations harboring different nitrogen genes grouped into denitrification, nitrification and general N metabolism categories in the different soils supplemented with the four fertilizer treatments during germination and flowering. Taxonomy has been grouped into phylum and genus.

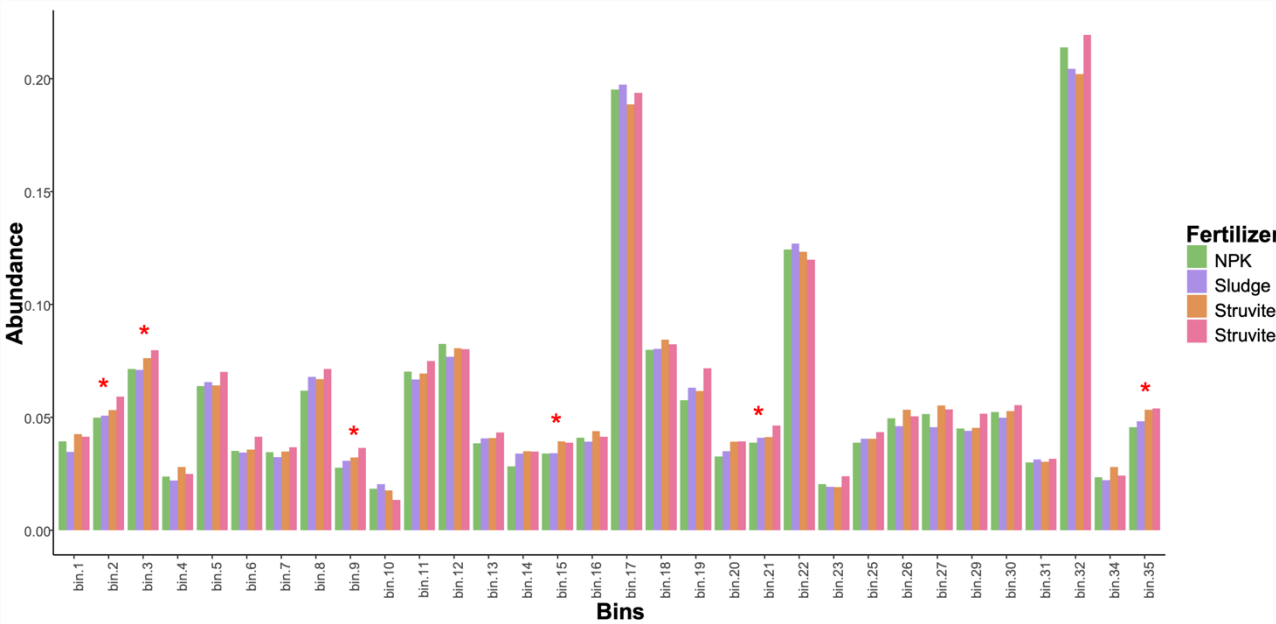


Figure 28: Bar plot of the results obtained from MAGs extraction: The x-axis represents the recovered bins, while the y-axis indicates the abundance of these bins. The plot illustrates the abundance of each fertilizer treatment (NPK, Sludge, Struvite, and Struvite+Sludge) across the different bins. An asterisk (*) above a bin denotes those bins where the fertilizer treatments exhibited statistically significant differences.

Focusing on the reconstructed genomes, specific bins were identified as containing nitrogen-related genes that matched those observed in the overall metagenome. These bins, along with their associated taxonomic families, Included 2 Acidobacteriota (bin.1_Pyrinomonadaceae, bin.15_Pyrinomonadaceae), 1 Verrucomicrobiota (bin.16_Akkermansiaceae), 3 Nitrospirota (bin.18_Nitrospiraceae, bin.20_Nitrospiraceae, bin.7_Nitrosomonadaceae), 5 Thermoproteota (bin.4_Nitrososphaeraceae, bin.11_Nitrososphaeraceae, bin.12_Nitrososphaeraceae, bin.13_Nitrososphaeraceae, bin.32_Nitrososphaeraceae), 1 Actinomycetota (bin.22_Propionibacteriaceae), 1 Bacillota (bin.23_Bacillaceae_H), 2 Pseudomonadota (bin.26_Steroidobacteraceae, bin.27_Steroidobacteraceae), 1 Actinomycetota (bin.30_Micrococcaceae) (Figure 27). For instance, the archaeal family Nitrososphaeraceae, (bin.4_ Nitrososphaeraceae, bin.11_Nitrososphaeraceae, bin.12_ Nitrososphaeraceae, bin.13_ Nitrososphaeraceae and bin.32_ Nitrososphaeraceae), harbored genes such as *GLUD1_2*, *gdhA* , *pmoC-amoC*, *pmoA-amoA*, *pmoB-amoB*, *nirK* and *glnA*, which matched those detected in the metagenome and were aligned with the contributions of these families to the nitrogen cycle as shown in Figure 29. Similarly, bin.22_Propionibacteriaceae contained genes involved with nitrate and nitrite reduction, such as *narH*, *narY*, *nxrB*, *narG*, *narZ*, *nxrA*, *narJ*, *narW*, *narI*, *narV*, *gla_GLUL*, and *glnE*, as shown in Figure 29, which also matched the metagenomic data. Furthermore, the genus *Nitrosospira*, from the Nitrososphaeraceae family, was found to possess the *pmoC-amoC* gene, which was consistent with its presence in the metagenome.

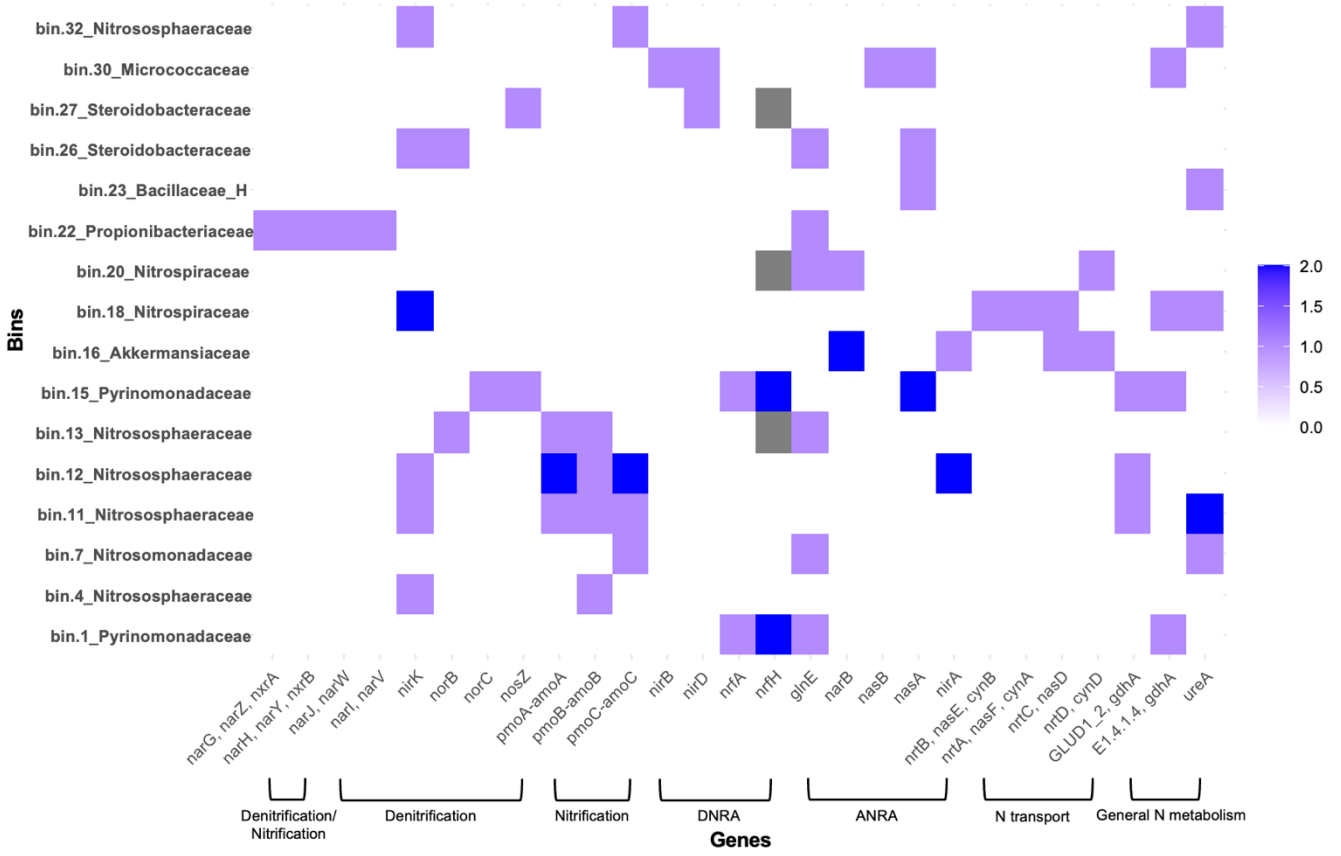


Figure 29: Heatmap of the abundance of nitrogen cycle-related genes recovered in the bins: The x-axis represents nitrogen-related genes categorized into functional groups, including Denitrification/Nitrification, Denitrification, Nitrification, DNRA, ANRA, N transport, and general N metabolism. The y-axis represents the bins containing these nitrogen cycle-related genes. The heatmap displays the copy number of each gene identified in the recovered bins.

3.7. Correlations between nitrogen content, WSN, urease activity and N functional groups and relative gene abundance.

With respect to the bacterial community (Figure 30), significant negative correlations were found between WSN and *napC*, which is involved in denitrification, *narB*, which belongs to ANRA, and *nifH*, which is associated with N_2 fixation. Significant positive correlations were observed between WSN and *nirS*, which is involved in denitrification (Figure 30B). Regarding urease activity, significant negative correlations were detected with *napA*, which belongs to denitrification, *pmoC-amoC*, which is associated with nitrification, *nirA*, which is involved in ANRA, *nifD*, which is associated with N_2 fixation, and *nao*, which belongs to general N metabolism.

Conversely, significant positive correlations were found between urease activity and *napB*, which is associated with denitrification, *nirD*, which belongs to DNRA, and *GLUD1_2* and *gdhA*, which are involved in general N metabolism (Figure 30B). As for the archaeal community, WSN showed a significant positive correlation with *pmoC-amoC*, which is associated with nitrification (Figure 31).

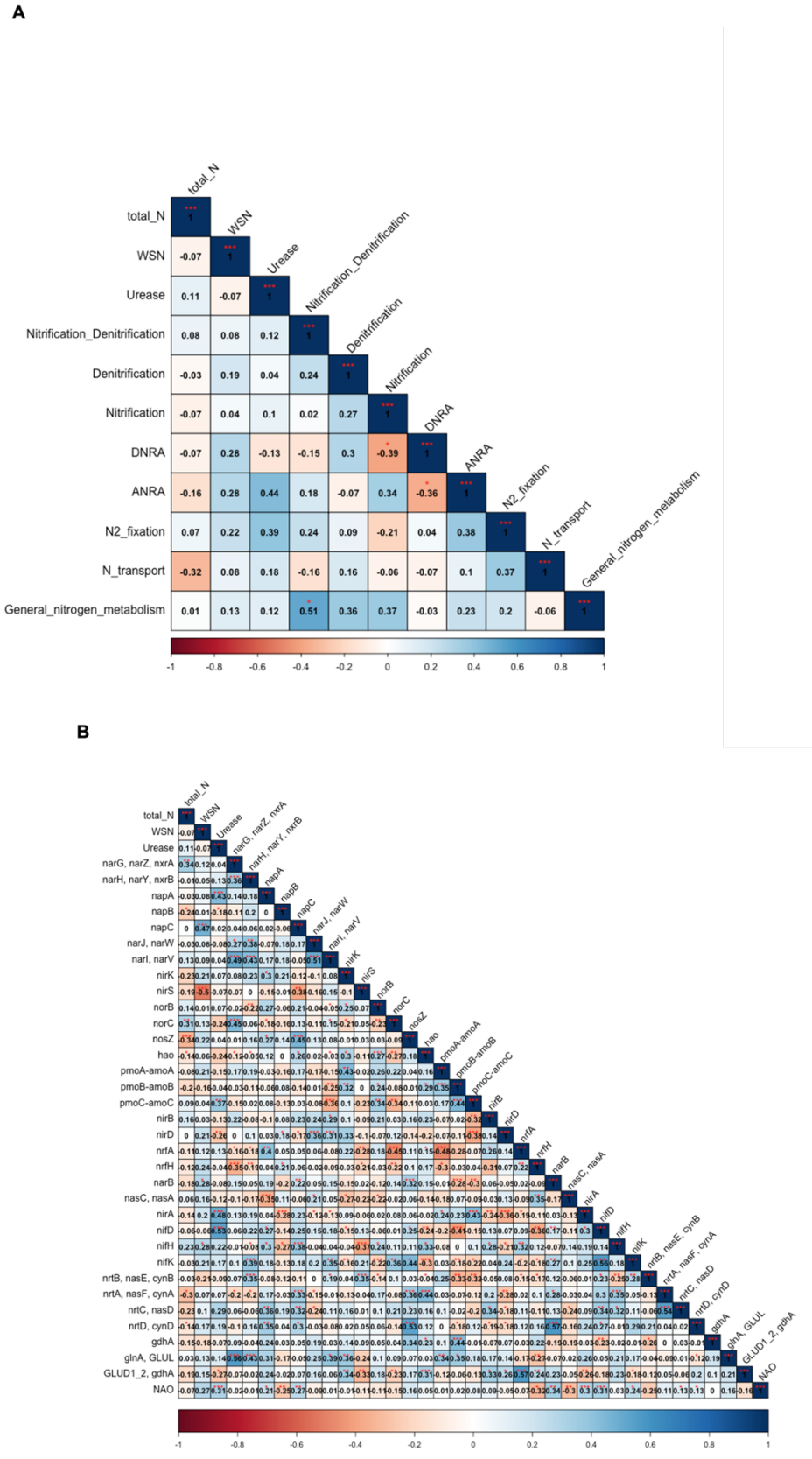


Figure 30: Correlation analysis of nitrogen-related categories, genes, and their relationship with total nitrogen, water-soluble nitrogen (WSN), and urease activity in bacteria. A) Categories, B) Genes. Negative correlations are depicted in blue, while positive correlations are shown in red, with asterisks (*) indicating statistically significant correlations (positive or negative) as determined by the correlation test.

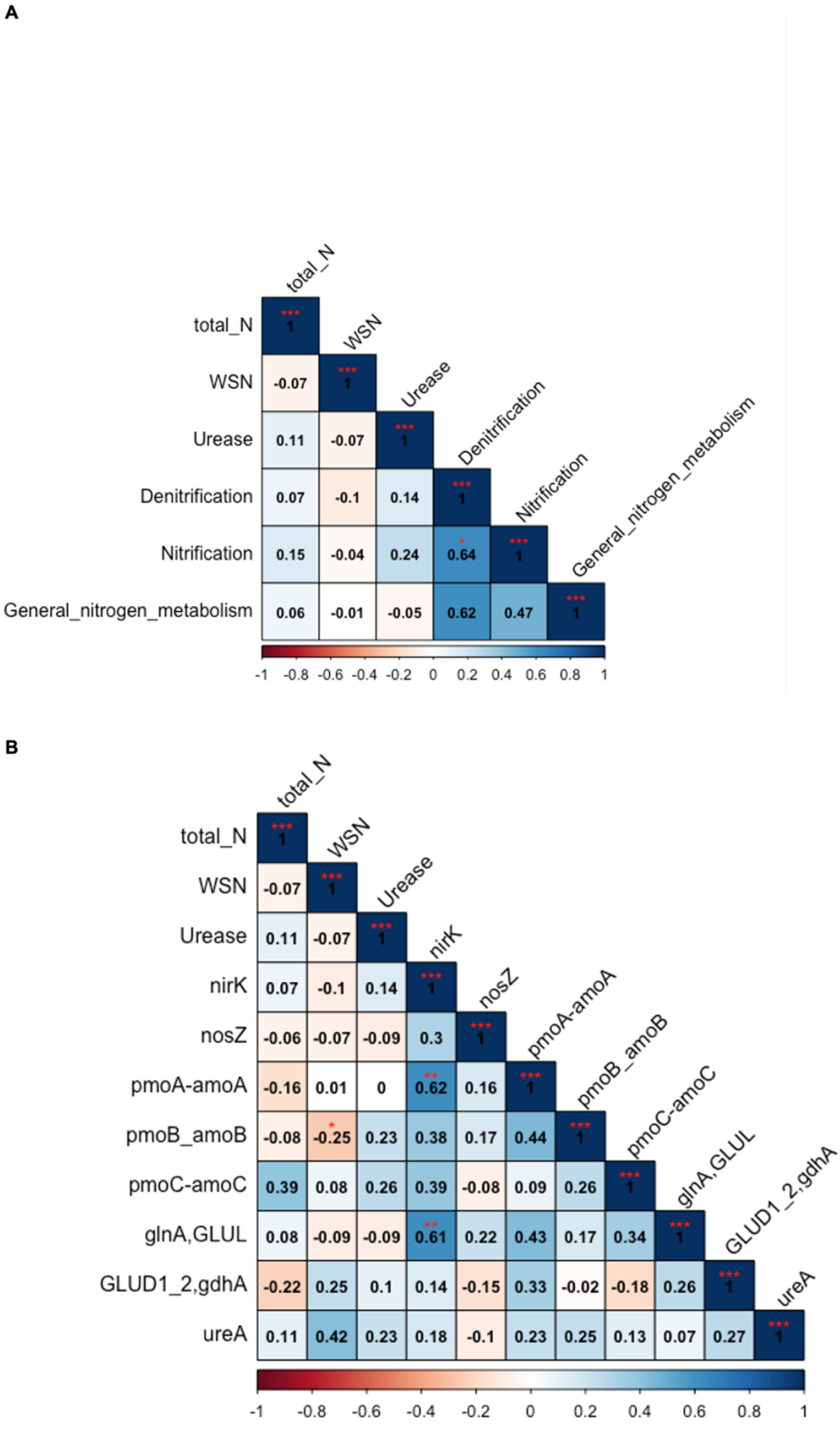


Figure 31: Correlation analysis of nitrogen-related categories, genes, and their relationship with total nitrogen, water-soluble nitrogen (WSN), and urease activity in archaea. A) Categories, B) Genes. Negative correlations are depicted in blue, while positive correlations are shown in red, with asterisks (*) indicating statistically significant correlations (positive or negative) as determined by the correlation test.

4. DISCUSSION

4.1. Abundance of nitrogen-related genes and enzymes and the associated microbiome

Bacterial community had a higher abundance of genes associated with denitrification, nitrification and DNRA compared to genes involved in N_2 fixation, N transport and general N metabolism. These results differ from previous studies reporting a higher abundance of N_2 fixation genes in natural ecosystems, where nitrogen availability is typically limited (Reed et al., 2011). The observed differences may be attributed to the fact that this is a long-term agricultural assay which has received mineral fertilization during years. This likely increased the availability of inorganic nitrogen forms, thereby favoring microbial pathways such as denitrification and nitrification. The results align with studies highlighting the dominance of denitrification and nitrification genes in agricultural soils under conventional fertilization regimes (Raglin et al., 2022; F. Wang et al., 2022). Further, our metagenomic results revealed a notable abundance of genes involved in DNRA, particularly within the bacterial community. This suggests that microbial communities at our study site may play a critical role in nitrogen retention in the soil by converting nitrate into ammonium, a less mobile form of nitrogen (Putz, 2018). However, compare to metagenomic results, the most abundant proteins were those involved in general nitrogen metabolism, such as the enzyme *glnA (GLUL)*. This observation indicates a decoupling between genetic abundance and protein expression levels (Starke et al., 2019) and is consistent with prior studies highlighting that greater genetic abundance does not necessarily correlate with higher expression levels (Fierer et al., 2012).

Regarding taxonomic distribution, we identified the existence of functional niches associated with nitrogen transformation processes, clustered within specific microbial groups. The taxonomic distribution of nitrogen-cycling genes revealed distinct functional clustering among bacterial taxa. For instance, the dominance of Acidobacteria, Actinobacteria, and Proteobacteria in harboring nitrogen-cycling genes aligns with their known roles in soil N transformations (Fierer et al., 2012). The co-occurrence of nitrification and denitrification genes in genera like *Bradyrhizobium* and *Ornithinibacter* suggests a potential of these populations to adapt to fluctuating nitrogen availability (Kuypers et al., 2018). In contrast, the limited diversity of nitrogen-cycling genes in archaea, primarily associated with *Thaumarchaeota*, reflects their specialized role in nitrification and general N metabolism (Stahl & de la Torre, 2012). Our findings highlight that microorganisms involved in denitrification and nitrification typically do not harbor genes related to N_2 fixation or N transport. In contrast, microorganisms associated with general N metabolism, such as those harboring the *glnA_GLUL* gene, tend to exhibit a shared abundance of genes, suggesting an organization into guilds or functional niches linked to soil nitrogen cycling. For example, we observed that genera within the Proteobacteria phylum, including *Nitrosospira*, *Nitrosomonas* and *Nitrosovibrio*, were the predominant microorganisms carrying nitrification-related genes, consistent with their well-established roles as ammonia oxidizers (W. Huang et al., 2020). Similarly, our metaproteomic data revealed that the enzyme *glnA_GLUL* was widely represented across diverse bacterial and archaeal genera, underscoring its central role in nitrogen assimilation. This study also demonstrates that, within the bacterial community, although

the abundance of genes involved in denitrification and nitrification is relatively higher than those associated with N_2 fixation, the functional guilds of denitrifiers and nitrifiers exhibit a relatively narrow phylogenetic distribution among the dominant populations. This highlights the highly restricted phylogenetic distribution of dominant denitrifiers and nitrifiers in this agroecosystem, in contrast to the genes in the general nitrogen metabolism category, which were more broadly distributed taxonomically. For example, *Nitrospira* (Nitrospirae) and *Azospirillum* (Proteobacteria) were identified as key genera harboring N transport genes such as *nrtA*, *nasF*, *cynA*. Additionally, *Skermanella* (Proteobacteria), which contained the *nifD* gene associated with N_2 fixation, emerged as a significant contributor to the N cycle in agricultural soils (Song et al., 2025). Overall, the phylum Proteobacteria encompassed many members carrying denitrification and nitrification genes, consistent with its well-established role in nitrogen transformation processes (C. M. Jones et al., 2013).

Additionally, the results highlight the specialization of microbial guilds in specific N transformation pathways, emphasizing the organization of microbial communities into functional niches. For instance, genera such as *Sorangium* and *Myxococcus* (Proteobacteria) were predominant carriers of DNRA-related genes, while *Skermanella* and *Azotobacter* (Proteobacteria) were key carriers of N_2 fixation genes. This functional specialization suggests that microbial communities in agricultural soils are organized into guilds that optimize resource utilization and minimize competition, as previously proposed (C. Liu et al., 2022). However, the limited phylogenetic distribution of certain functional genes, such as *narB* (ANRA) in *Ohtaekwangia* (Bacteroidetes), underscores the potential vulnerability of this functional process to environmental changes, such as shifts in fertilization practices or soil pH. The dominance of denitrification and nitrification genes, alongside the low abundance of N fixation genes, suggests that conventional fertilization practices may promote nitrogen losses through gaseous emissions and leaching, while limiting biological nitrogen inputs (X. Zhang et al., 2015). This is particularly relevant in the context of sustainable agriculture, where reducing nitrogen losses and improving nitrogen use efficiency are critical objectives. The identification of key microbial taxa involved in nitrogen transformation processes, such as *Nitrosospira* and *Skermanella*, provides valuable insights for developing targeted strategies to manipulate microbial communities and optimize nitrogen cycling in agroecosystems.

4.2. Influence of fertilization on nitrogen dynamics and microbial functional genes

Our results demonstrate that different fertilization strategies significantly influence the categories of nitrogen cycle-related genes and the functional potential of soil microbial communities in a maize agroecosystem. These findings align with previous research emphasizing the importance of fertilization in modulating soil microbial functionality and nutrient cycling (Luo et al., 2018). WSN content exhibited significant variation among treatments, with NPK consistently showing the highest levels. This may be attributed to the readily available inorganic nitrogen forms in synthetic fertilizers (Geisseler & Scow, 2014). In contrast, the struvite plus sludge treatment resulted in the lowest WSN values, likely due to slower mineralization rates of organic nitrogen sources and the long-term mineral character of struvite (Marzi

et al., 2020). The observed correlations between nitrogen content, WSN, urease activity, and gene abundance may provide insights into the linkages between soil nitrogen dynamics and microbial processes. For instance, the positive correlations between WSN and denitrification genes (e.g., *nirS*) highlight the role of soluble nitrogen forms in driving denitrification activity, as highlighted by Bastida et al. (2009). Furthermore, the higher abundance of DNRA-related genes in NPK and struvite treatments aligns with the presence of mineral nitrogen sources, which are known to promote DNRA activity (Pandey et al., 2019). In contrast, organic treatments with slower nitrogen release may limit DNRA activity (Rütting et al., 2011).

In archaeal communities, functional potential was less affected by fertilization compared to bacteria, indicating a more stable functional nitrogen pattern regardless fertilization treatment. However, treatments with organic amendments, such as sludge combined with struvite, showed a higher abundance of archaeal nitrification genes (e.g., *pmoB-amoB*), consistent with the sensitivity of ammonia-oxidizing archaea (AOA) to slow-release nitrogen sources. The predominance of the *glnA (GLUL)* gene in both bacterial and archaeal communities highlights its central role in nitrogen metabolism under different fertilization strategies (Rütting et al., 2011).

4.3. The role of phenology in the dynamics of nitrogen cycling and microbial activity

Crop phenology played a crucial role in modulating microbial activity and the expression of nitrogen cycling-related genes. Changes in phenological stages, particularly between germination and flowering, significantly influenced the relative abundance of genes associated with key nitrogen cycling processes, such as denitrification/nitrification, N_2 fixation, and N transport. These findings underscore the close interaction between plant development and the functionality of the soil microbial community, suggesting that plant nutrient demands at different growth stages can shape microbial activity and functional composition. Moreover, urease activity, a key enzyme in nitrogen mineralization, was significantly influenced by phenology, with higher activity observed during germination compared to flowering. This pattern likely reflects the greater nitrogen demand during the early stages of plant growth, which stimulates microbial activity to release nitrogen from organic sources. During germination, the availability of easily degradable organic substrates, such as those supplied by organic fertilization treatments, may enhance urease activity, consistent with previous findings (Gianfreda & Ruggiero, 2006). Genes related to denitrification/nitrification also showed higher abundance during germination compared to flowering. This pattern may be associated with increased inorganic nitrogen availability in the early stages of plant growth, favoring nitrification processes. The subsequent decline in these genes' abundance during flowering could reflect a shift in microbial metabolic strategies, potentially driven by increased resource competition or changes in soil redox conditions as the plant matures (Kuzyakov & Xu, 2013). Additionally, the influence of phenology on N_2 fixation genes was particularly notable, with a significant reduction in their abundance during flowering. This trend may relate to a reallocation of microbial resources toward other metabolic processes, such as N assimilation, in response to the plant's increased nitrogen demand during flowering

(Houlton et al., 2008). Phenology also influenced the abundance of N transport-related genes, suggesting that microorganisms adjust their capacity to acquire and mobilize nitrogen based on plant needs. For instance, a higher abundance of N transport genes was observed during flowering in the sludge treatment, potentially reflecting the increased nitrogen demand required to sustain reproductive growth. This finding highlights the capacity of organic amendments, such as sludge, to supply nitrogen at critical stages of crop development (Larney & Angers, 2012). The delayed peak in N transport gene abundance during flowering suggests a gradual release of nitrogen from organic sources, aligning with the plant's nutrient requirements at later phenological stages. This molecular adjustment underscores the importance of tailoring fertilization strategies to synchronize nutrient release with plant demands throughout the growth cycle (Fontaine et al., 2024).

In the case of the archaeal community, a lower sensitivity to phenological changes was observed compared to the bacterial community. However, certain genes associated with nitrification, such as *pmoB-amoB* and *pmoC-amoC*, exhibited significant variations between germination and flowering. This indicates that while archaea may possess a lower level of functional redundancy compared to bacteria, they still respond to soil condition changes driven by plant development (Aller & Kemp, 2008). Notably, the *pmoC-amoC* gene showed higher abundance during flowering within the genus *Candidatus Nitrosocosmicus*, suggesting a more active role for this taxon in nitrification during flowering.

4.4. Functional and taxonomic insights from microbial genome reconstruction (MAGs) in nitrogen cycling

The reconstruction of microbial genomes (MAGs) provided profound insights into the taxonomic and functional diversity of the microbial community, particularly highlighting the role of the archaeal family Nitrososphaeraceae in nitrogen transformation processes. Among the 32 high-quality MAGs recovered, five bins associated with Nitrososphaeraceae (bin.4_*Nitrososphaera*, bin.11_*JAJNBK01*, bin.12_*JAJNBK01*, bin.13_*Nitrosocosmicus*, and bin.32_*Nitrososphaera*) harbored key nitrogen-cycling genes, including *pmoC-amoC, pmoA-amoA*, *pmoB-amoB*, and *nirK*, which are critical for nitrification and denitrification. The presence of these genes aligns with the well-documented role of Nitrososphaeraceae, particularly the genus *Nitrososphaera*, as key ammonia oxidizers in terrestrial environments (Lehtovirta-Morley et al., 2024; Tourna et al., 2011). The detection of *nirK*, a gene associated with denitrification, in these bins is particularly intriguing, as it suggests a potential dual functionality in nitrogen cycling. While Nitrososphaeraceae are primarily known for their role in nitrification, the presence of *nirK* indicates a possible adaptation to fluctuating nitrogen conditions, enabling these archaea to contribute to both nitrification and denitrification under specific environmental conditions (Clark et al., 2021; Hetz & Horn, 2021). This metabolic versatility could provide a competitive advantage in agricultural soils, where nitrogen availability is highly dynamic due to fertilization practices. The recovery of multiple Nitrososphaeraceae MAGs further underscores their ecological importance in this agroecosystem, consistent with their ubiquitous presence in soils and their dominance in ammonia oxidation processes (C. Wang et al., 2024). Additionally, the identification of bin.22_Propionibacteriaceae, which contained genes such as *narG, narZ,*

nxrA, and *glnA*, involved in denitrification and general nitrogen metabolism, highlights the potential for non-canonical taxa to contribute to nitrogen transformations in agricultural soils (Q. Yang et al., 2024).

The consistency between the MAGs and the overall metagenomic data underscores the robustness of our approach in reconstructing and characterizing the functional potential of the microbial community. Furthermore, bins associated with the family Nitrospiraceae (bin.18_Nitrospiraceae and bin.20_Nitrospiraceae) contained genes such as *glnA* and *GLUL*, which are involved in nitrogen assimilation and were more abundant in treatments with slow-release nitrogen sources (X. Yang et al., 2021). This observation aligns with the higher abundance of nitrification-related genes in organic treatments, as indicated by the overall metagenomic data. The consistency between the MAGs and metaproteomic analyses, particularly for genes such as *nirK* and *glnA*, reinforces the contribution of these taxa to nitrogen cycling under varying fertilization regimes.

The influence of fertilization on microbial community structure and function was further evidenced by the differential abundance of specific MAGs across treatments. Among the 32 high-quality MAGs recovered, six bins (bin.2_JACCYU01, bin.3_ZC4RG35, bin.9_WHSQ01, bin.15_Pyrinomonadaceae, bin.21_JAKEEW01 and bin.35_UBA9968) showed statistically significant differences between fertilization treatments. Bins bin.2_JACCYU01, bin.3_ZC4RG35, bin.9_WHSQ01 and bin.21_JAKEEW01, which were more abundant under the struvite plus sludge treatment, corresponded to taxa such as Gammaproteobacteria and Actinomycetes, known for their roles in organic matter decomposition and nitrogen cycling (Javed et al., 2021; S. Liu & Liu, 2020). These findings suggest that organic amendments, such as sludge, promote the abundance of microbial taxa involved in complex organic nitrogen transformations. In contrast, bins bin.15_Pyrinomonadaceae and bin.35_UBA9968 -associated with Blastocatellia and Binatia, respectively- were more abundant under the struvite-only treatment. Blastocatellia, a class within the Acidobacteria phylum, has been linked to organic matter degradation and nitrogen assimilation in nutrient-poor soils (Fierer et al., 2012). Similarly, Binatia, a recently described class within the Gemmatimonadetes phylum, may contribute to nitrogen mineralization, releasing ammonium from organic matter (Yao et al., 2021). Their increased abundance under struvite fertilization highlights their potential role in modulating nitrogen availability in semiarid agroecosystems, warranting further investigation into their functional contributions to the nitrogen cycle.

5. CONCLUSIONS

Our study provides new insights into the key microbial players, genes, and enzymes associated with the nitrogen (N) cycle in semiarid agroecosystems through the use of combined metagenomics, metaproteomics, and reconstructed microbial genomes (MAGs). Consistent with our hypotheses, the phenological stage proves to be a more critical factor than fertilizer treatments in influencing the relative abundance of nitrogen cycling genes, particularly regarding denitrification, nitrification, and N_2 fixation potential of bacterial taxa. While the effects of phenology on soil microbial communities were previously

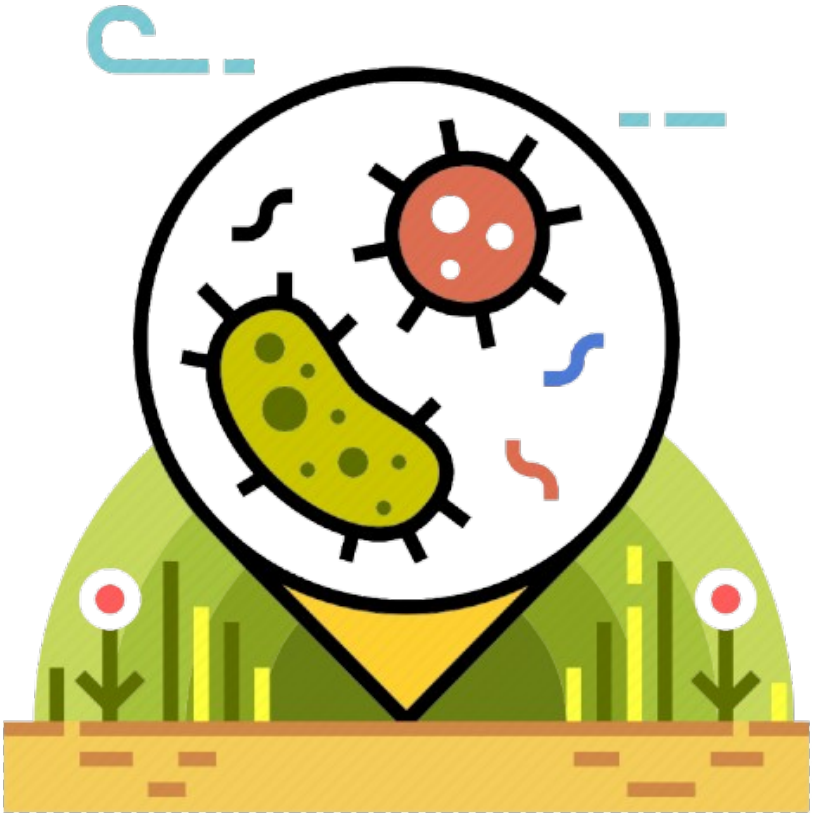
recognized, this finding is novel in highlighting the greater importance of the phenological stage in modulating nitrogen cycle and associated microbiomes compared to fertilization practices. This insight opens the door to new nitrogen fertilization strategies that more deeply consider the phenological stage.

Furthermore, our results suggest notable taxonomic clustering of functional processes related to the nitrogen cycle, profoundly influenced by plant growth stages. Importantly, our findings highlight that microorganisms involved in denitrification and nitrification typically do not harbor genes related to N_2 fixation or N transport, indicating functional specialization among microbial guilds. Among the reconstructed microbial genomes, specific MAGs revealed significant taxonomic and functional insights, such as the identification of bins harboring key nitrogen-cycling genes. These MAGs confirmed the functional roles of taxa like Nitrososphaeraceae in nitrification and Propionibacteriaceae in denitrification. Moreover, metaproteomic approaches enhance the potential of metagenomes by identifying abundant enzymes, such as glutamine synthetase coded by *glnA*, which has been overlooked in numerous studies but appears essential for nitrogen cycling in this maize agroecosystem.

Besides a major role of phenology, our study also demonstrates the influence of contrasting fertilizers on the abundance of key genes governing nitrogen cycling, thereby confirming the initial hypothesis. We observed significantly distinct responses among fertilizers regarding genes involved in DNRA and nitrification across both bacterial and archaeal communities, as well as genes associated with N_2 fixation and N transport, predominantly within bacterial taxa. Our findings demonstrate that fertilization strategies influence both nitrogen dynamics and microbial community functionality in semi-arid agroecosystems. While mineral fertilizers like NPK and struvite enhance microbial processes such as DNRA and nitrification, organic amendments support a more diverse microbial community but require careful management to optimize nitrogen release and minimize losses.

CHAPTER 3

Linking genomic potential and functional activity: Microbial specialization and niche partitioning in the decomposition process revealed by multi-omics approaches.



CHAPTER 3

1. INTRODUCTION

Forest soils are among the most critical carbon (C) sinks on Earth, playing a pivotal role in the global carbon cycle by sequestering atmospheric CO_2 and storing it in both biomass and soil organic matter (Lal, 2005). The capacity of forest soils to act as long-term carbon reservoirs is largely driven by the continuous input of organic carbon from plant and fungal sources, which are subsequently decomposed and transformed by soil microbial communities (De Deyn et al., 2008; H. Li et al., 2024; Mäki et al., 2017). Understanding the dynamics of carbon cycling in forest soils is essential for predicting the responses of these ecosystems to environmental changes and for developing strategies to mitigate climate change.

The organic carbon in forest soils originates primarily from two sources: plant-derived compounds, such as cellulose, hemicelluloses, and lignin, and fungal-derived compounds, including chitin and β-glucans (Ekblad et al., 2013; X. Zhao et al., 2024). Plant biomass, derived from the photosynthetic activity of trees and understory vegetation, constitutes the bulk of organic matter input into forest soils. However, fungal biomass, particularly in forest litter, also represents a significant carbon source, often exceeding microbial biomass in the underlying soil by an order of magnitude (Baldrian et al., 2010). The decomposition of this diverse pool of organic matter is a critical process that not only recycles nutrients but also supports complex decomposer food webs, which are essential for maintaining soil fertility and regulating carbon fluxes (Khatoon et al., 2017).

The decomposition of organic matter in forest soils is mediated by a diverse array of microbial communities, including both fungi and bacteria (Algora et al., 2022; Baldrian, 2017; Bani et al., 2018; Tláskal et al., 2021). Traditionally, fungi have been regarded as the primary decomposers due to their ability to produce a wide range of extracellular enzymes, such as carbohydrate-active enzymes (CAZymes), that break down complex biopolymers like lignin and cellulose (Žifčáková, 2017). However, recent research has highlighted the significant role of bacteria in this process, particularly in the decomposition of fungal-derived biomass (Brabcová et al., 2016; López-Mondéjar et al., 2020). Bacteria are now recognized as key players in the decomposition of complex biopolymers, with certain bacterial groups specializing in the breakdown of fungal-derived materials, such as chitin and β-glucans (López-Mondéjar et al., 2018). Despite this progress, the specific substrate preferences of individual microbial taxa and their functional roles in decomposition remain poorly understood.

Current knowledge about the preferences of microbial taxa for specific biopolymers is contradictory. Some studies suggest that many bacteria are versatile and can utilize carbon from a wide range of sources (López-Mondéjar et al., 2018, 2020), while others indicate the existence of specialized microbial guilds that target specific biopolymers (Bhatnagar et al., 2018; Brabcová et al., 2016). For example, Urbanová et al. (2015) demonstrated that litter of different quality supports distinct bacterial communities, suggesting a degree of specialization. Recently, Algora et al. (2022) provided evidence for the existence of substrate-

specific microbial guilds in forest soils, although their findings also highlighted limitations. While certain microbial taxa were enriched on specific biopolymers, this does not necessarily prove their role as primary decomposers, as bacteria can act as "cheaters" by exploiting the decomposition products generated by other taxa (Berlemont & Martiny, 2013; Tláskal et al., 2021). To address these limitations, this study combines metagenomics and metatranscriptomics to analyze not only the genetic potential of microbial communities but also the expression of key genes involved in biopolymer decomposition, providing a more comprehensive understanding of the functional roles of microbial taxa in situ.

In this study, we investigate the microbial decomposer guilds responsible for breaking down the biopolymers that constitute dead biomass in forest soils. Our primary objective is to associate microbial community composition with function, a critical step for understanding the decomposition process and its role in the carbon cycle. Specifically, we aim to determine whether these guilds are composed of specialist taxa with narrow substrate preferences or generalists capable of degrading multiple biopolymers. We also explore whether specialization occurs for individual polymers or for groups of polymers of similar origin. For example, López-Mondéjar et al. (2016) observed that bacteria growing on cellulose expressed a diverse array of enzymes, including xylanases and glucanases, rather than just cellulases, reflecting the complex nature of natural substrates where cellulose rarely occurs in isolation. This suggests that even specialists may exhibit some degree of metabolic flexibility. By classifying microbial taxa as generalists or specialists and assigning them to substrate-specific guilds, we aim to elucidate the ecological strategies that underpin the decomposition of dead biomass in forest soils.

Furthermore, we examine the differences in genetic potential for decomposition between generalist and specialist taxa, focusing on their enzymatic systems and how gene expression varies across different substrates. We hypothesize that the pool of key carbohydrate-active enzymes (CAZymes) encoded by soil bacteria is quantitatively similar to that encoded by fungi but qualitatively different. We also hypothesize that the contribution of bacteria to decomposition processes is comparable to that of fungi for most substrates and is expected to be higher for fungal-derived biopolymers, as suggested by Brabcová et al. (2016). Additionally, we propose that the pool of CAZymes expressed by bacterial communities is specific to each substrate due to differences in community composition but is similar for substrates of plant origin. We further hypothesize that the bacterial community of decomposers is primarily composed of specialists, with key taxa specializing in the decomposition of different carbon sources in forest soil, as indicated by Algora et al. (2022). These specialists are expected to contain and express specific key genes for the decomposition of a particular carbon source. Finally, we propose that a proportion of the bacterial community consists of generalists or "cheaters" that contain and express genes coding for enzymes involved in the degradation of low-molecular-mass substrates, rather than complex biopolymers.

By addressing these hypotheses, this study aims to provide a deeper understanding of the microbial mechanisms driving carbon cycling in forest soils. Through the integration of metagenomics,metatranscriptomics and metagenome-assembled genomes (MAGs), we aim to

characterize the genetic potential and functional activity of microbial decomposer guilds, shedding light on their ecological roles and contributions to carbon dynamics. This work will contribute to more accurate models of carbon cycling in forest ecosystems and inform strategies for their sustainable management.

2. MATERIALS AND METHODS

2.1. Site description, experimental design and sampling

The research was conducted in a temperate forest situated within the Xaverovský Háj Natural Reserve, located in northeastern Prague, Czech Republic. The complete methodology applied in this study is detailed in Algora et al. (2022). This forest is primarily composed of oak trees (*Quercus petraea*), accompanied by other species such as *Carpinus betulus*, *Tilia* spp., *Acer* spp., and *Picea abies*, which contribute to litter accumulation on the forest floor (Algora et al., 2022) (Figure 32).

Figure 32: Forest of the Xaverovský Háj Nature Reserve.

The soil in this area is classified as an acidic cambisol, characterized by well-defined litter, organic, and mineral horizons. The litter has a pH of 4.3, with carbon and nitrogen contents of 46.2% and 1.76%, respectively. These site characteristics, along with prior studies on litter decomposition, enzymatic activities, fungal and bacterial diversity in soil and litter (López-Mondéjar et al., 2015), and the decomposition of plant, fungal, and bacterial biomass, have been extensively documented (Algora et al., 2022; Algora Gallardo et al., 2021; Baldrian et al., 2010; López-Mondéjar et al., 2018; Šnajdr et al., 2008;

Větrovský & Baldrian, 2013; Voříšková et al., 2014). To capture the forest's environmental variability, four sampling sites were randomly selected: Site 1 (N 50° 5′ 39″, E 14° 37′ 8″), Site 2 (N 50° 5′ 40″, E 14° 36′ 58″), Site 3 (N 50° 5′ 38″, E 14° 36′ 42″), and Site 4 (N 50° 5′ 41″, E 14° 36′ 4″). These locations shared similar topographical conditions, with oak dominating the canopy, except for Site 4, where additional species such as *Tilia* spp. were present.

The experiment utilized mesh bags containing different biopolymers, which were placed in situ to allow microbial colonization. The fungal communities colonizing these mesh bags have been previously studied (Algora Gallardo et al., 2021). Each mesh bag (5 × 5 cm, made of nylon with a 0.05 mm mesh size) was filled with 4 g of a pure, finely powdered substrate (Figure 33). The substrates included cellulose (microcrystalline ca. 0.05 mm, SERVA, Germany), xylan (from beech wood, SERVA, Germany), glucomannan (from Konjac, Natural Nutrition, USA), β-1,3-glucan (from yeast cell wall concentrate, SOLGAR, USA), β-1,3-1,6-glucan (from *Pleurotus ostreatus* oyster mushroom, NATURES, Slovakia), lignin (alkali lignin, Sigma–Aldrich, USA), pectin (from citrus, Alfa Aesar, Germany), and chitin (from shrimp shells, Sigma–Aldrich, USA). All substrates were sterilized using gamma irradiation. While cellulose, xylan, glucomannan, lignin, and pectin are of plant origin, β-glucans and chitin are characteristic of fungal biomass, with chitin also being present in the exoskeletons of arthropods (Algora et al., 2022).

Figure 33: Mesh bags with the different biopolymers in the forest.

Previous studies using this experimental setup have analyzed the fungal and bacterial communities colonizing the mesh bags through amplicon sequencing (Algora et al., 2022; Algora Gallardo et al., 2021). In this study, we extend these findings by employing metagenomics and metatranscriptomics and

recovering MAGs to investigate the functional potential and activity of microbial communities involved in the decomposition of these biopolymers. This approach allows us to identify key genes and pathways associated with biopolymer degradation and to assess the ecological roles of specific microbial taxa in situ. Four mesh bag replicates per substrate were incubated in the bottom layer of the litter horizon across each study site for three weeks during August and September 2018. To promote contact with the surrounding litter, mesh bags were moistened with sterile water at the time of placement (Algora et al., 2022). After incubation, the contents of the mesh bags were carefully retrieved (Figure 34), homogenized, frozen immediately using liquid nitrogen, and stored at −80 °C for DNA and RNA extraction. Additionally, litter samples from each site were collected, cut, frozen with liquid nitrogen, and stored at −80 °C for subsequent analysis (Algora et al., 2022).

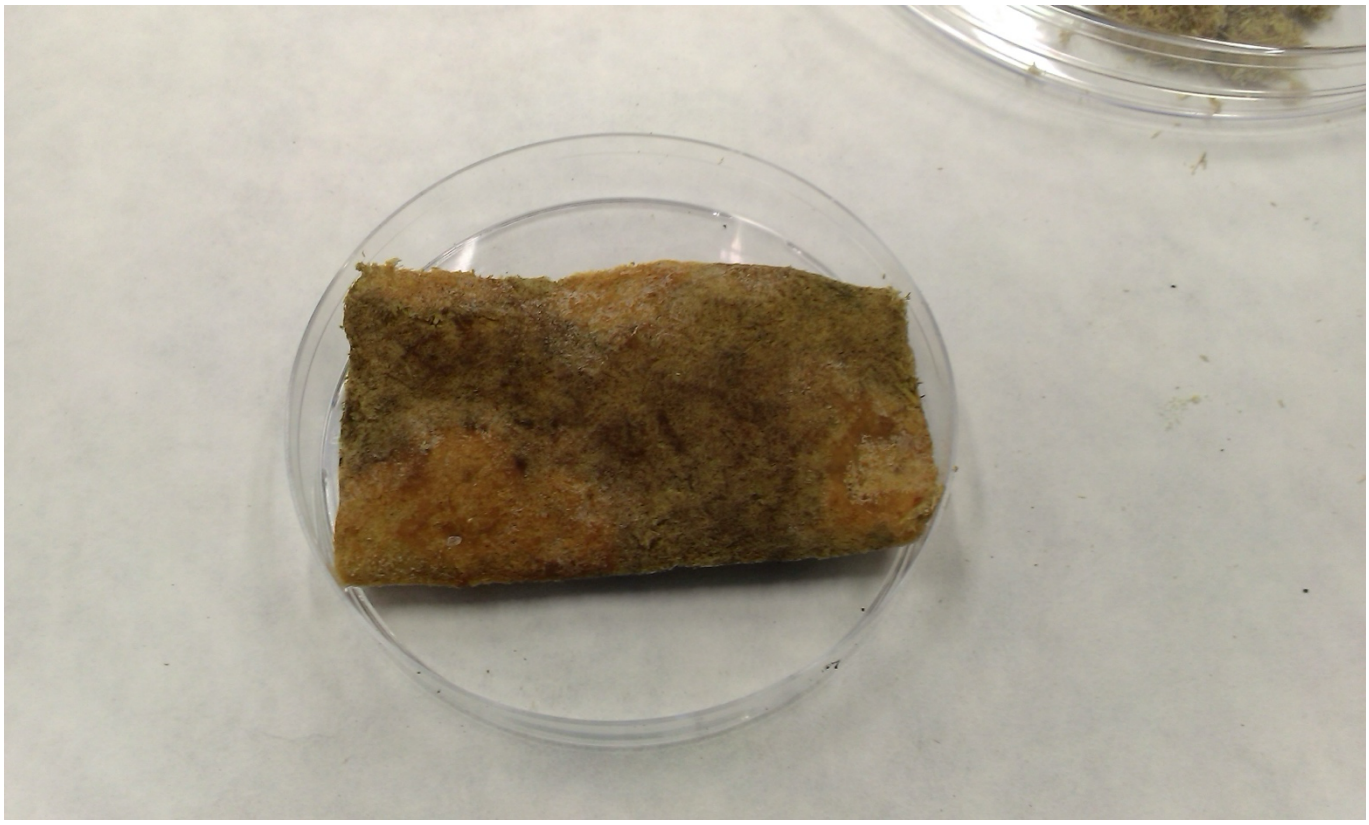


Figure 34: Mesh bag colonized by microorganisms, extracted from the forest.

2.2. Metagenomic analysis

As described in Algora et al. (2022), before proceeding with DNA extraction, subsamples from the mesh bags were subjected to freeze-drying. Total DNA was extracted from 0.15 g of the freeze-dried material, in duplicate, using the FastDNA Spin Kit for Soil (MP Biochemicals, USA), following the protocol provided by the manufacturer. The DNA extracted from duplicate samples was subsequently pooled and used for metagenomic sequencing to analyze the functional potential of the microbial communities involved in biopolymer decomposition. The Truseq Free LT kit was used to generate metagenome libraries and metagenome was sequenced on an Illumina NovaSeq6000 with a 2 x 250 paired-end runs. To complement these analyses, metagenomic sequencing was performed to obtain a comprehensive understanding of the

microbial functional potential and taxonomic composition associated with the different biopolymers. The methodology originally described by Žifčáková et al. (2016) was followed, with specific adaptations. The complete code utilized for the metagenomic analysis is available at https://github.com/mariabelen-bm/Doctoral_Thesis/blob/78fb7483aff71e570633d6b3a8f0bb666fe368de/metagenomics_pipeline_chapter_3.sh and in Annex 4. Additionally, the Python scripts employed during the metagenomic pipeline are included in Annex 5.

Raw sequencing data underwent quality control using FastQC (v0.12.0), with reads filtered to exclude those with quality scores below 30 or lengths shorter than 50 bases. Following quality control, median normalization was applied to reduce noise, using a k-mer size of 20 and a minimum coverage threshold of 20, as described by Barquero et al. (2024). Processed reads were assembled using MEGAHIT (v1.2.9) (D. Li et al., 2015), and the resulting assemblies were assessed for quality using MetaQUAST (v5.2.0) (Mikheenko et al., 2016). Gene prediction was performed with FragGeneScan (Rho et al., 2010), while alignment of reads to assemblies was carried out using Bowtie2 (v2.4.1). Functional and taxonomic annotations were generated using a comprehensive set of reference databases, following a rigorous multi-step approach to ensure accuracy and reliability. For taxonomic annotation, we utilized the NCBI non-redundant (nr) protein database (Pruitt et al., 2009), complemented by fungal-specific databases from the Joint Genome Institute (JGI) to improve the identification of fungal taxa (Grigoriev et al., 2012). Functional annotation was performed using the KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa et al., 2016) and KOG (Eukaryotic Orthologous Groups) (Tatusov et al., 2003) databases to assign genes to metabolic pathways and functional categories . Additionally, carbohydrate-active enzymes (CAZymes) were annotated using the dbCAN database, which specializes in the identification of enzymes involved in carbohydrate metabolism (L. Huang et al., 2018). To ensure high-confidence annotations, we applied stringent filtering criteria: only annotations with E-values lower than $10e^{-30}$ were retained, as values above this threshold were considered unreliable and disregarded, following the approach described by Tláskal et al. (2021). This step was critical to minimize false positives and ensure the accuracy of our functional and taxonomic assignments.

2.3. Metatranscriptomic analysis

Metatranscriptomic sequencing was conducted to gain a comprehensive understanding of the functional activity and taxonomic composition of microbial communities involved in the decomposition of different biopolymers in forest soil. The methodology was adapted from the approach described by Tláskal et al., (2021), with specific modifications to suit the experimental conditions. The complete pipeline used for the metatranscriptomic analysis, including all scripts and detailed steps, is available at https://github.com/mariabelen-bm/Doctoral_Thesis/blob/78fb7483aff71e570633d6b3a8f0bb666fe368de/metatranscriptomic_pipeline_chapter_3.sh and in Annex 6.

Total RNA was isolated using a NucleoSpin RNA plant kit (Macherey-Nagel) according to the manufactirer's protocol. We successfully extracted RNA from meshbags containing three polymers (cellulose, chitin and beta-1,3-glucan) and litter, due to the difficulties of extracting RNA of the rest of polymers. The quality and concentration of the extracted RNA were assessed using a Bioanalyzer (Agilent Technologies) and a Qubit fluorometer (Thermo Fisher Scientific). Ribosomal RNA was subsequently removed using the Ribo-Zero rRNA Removal Kit (Illumina), and the remaining messenger RNA (mRNA) was reverse-transcribed into complementary DNA (cDNA) with the NEBNext Ultra II RNA Library Prep Kit (New England Biolabs). Sequencing libraries were constructed using the TruSeq Stranded RNA kit according to the manufacturer's guidelines and sequenced on the Illumina NovaSeq platform, generating paired-end reads of 150 base pairs.

The raw sequencing data underwent preprocessing to eliminate adapter sequences and low-quality reads using Trimmomatic (Bolger et al., 2014). The cleaned reads were then aligned to a curated database comprising reference genomes and metagenome-assembled genomes (MAGs) using Bowtie2 (Langdon, 2015). Gene expression levels were normalized as transcripts per million (TPM) to account for variations in sequencing depth and gene length. Differential gene expression analysis was conducted with DESeq2 (Love et al., 2014) to pinpoint genes and metabolic pathways exhibiting significant changes in expression. Taxonomic classification of the sequences was conducted using the NCBI database (Pruitt et al., 2009) and publicly available fungal genomes from the Joint Genome Institute (JGI) (Grigoriev et al., 2012). Functional characterization of the expressed genes was performed using the KEGG and CAZy databases (L. Huang et al., 2018; Kanehisa et al., 2016), focusing on identifying critical metabolic pathways and carbohydrate-active enzymes (CAZymes) involved in the breakdown of plant and fungal material. A false discovery rate (FDR) threshold of <0.05 was applied to determine statistical significance.

2.4. Analysis of Metagenome-Assembled Genomes (MAGs)

The reconstruction and analysis of metagenomes-assembled genomes (MAGs) were conducted using a comprehensive pipeline tailored for environmental metagenomic samples. This workflow integrated multiple tools and methodologies to ensure accurate genome recovery, quality assessment, and functional annotation. The full pipeline, along with the source code, is accessible at https://github.com/mariabelen-bm/Doctoral_Thesis/blob/78fb7483aff71e570633d6b3a8f0bb666fe368de/MAGs_pipeline_chapter_3.s, with detailed steps provided in Annex 7.

The initial stage involved binning assembled contigs into genome-like clusters using the MetaWRAP pipeline (Uritskiy et al., 2018), which combines advanced algorithms, including MetaBAT2, MaxBin2, and CONCOCT. Refinement criteria were applied to retain only high-quality MAGs with a completeness of at least 70% and contamination below 10%. The quality of the selected MAGs was further validated with CheckM2 (Chklovski et al., 2023) to confirm genomic integrity. Taxonomic classification was performed using GTDB-Tk (Chaumeil et al., 2022) based on the Genome Taxonomy Database (GTDB) v2.4.0,

ensuring a consistent and robust taxonomic framework. To assess potential contamination and confirm genomic reliability, the Genomic UNcertainty Calculator (GUNC) was applied (Orakov et al., 2021).

To quantify the relative abundance of each MAG across samples, metagenomic reads were mapped to the reconstructed genomes using CoverM with the minimap2-sr algorithm (H. Li, 2018). Functional annotation was performed through the DRAM pipeline (Shaffer et al., 2020), which utilized diverse reference databases, such as GTDB, KOfam, and dbCAN2, to infer the metabolic potential of the MAGs. Particular emphasis was placed on identifying and annotating CAZyme families associated with the decomposition of distinct biopolymers present in the samples. This approach allowed us to link the functional potential of microbial communities to the breakdown of biopolymers that constitute dead biomass. By characterizing the distribution and diversity of CAZyme families across the reconstructed MAGs, we aimed to assess whether microbial taxa exhibit specific enzymatic repertoires tailored to individual biopolymers or broader capabilities targeting groups of biopolymers with shared origins.

2.5. Biopolymer guild classification

As described Algora et al. (2022), metagenome-assembled genomes (MAGs) were assigned to specific guilds based on their colonization patterns of individual biopolymers. A MAG was classified as part of a guild if it exhibited a relative abundance greater than 2% in at least one mesh bag containing a specific biopolymer, or if its relative abundance exceeded the maximum observed in litter samples in at least five mesh bags of the same biopolymer. Furthermore, MAGs were categorized into three groups: “broad-range generalists,” if they were associated with 5–8 guilds; “narrow-range generalists,” if they were present in 3–4 guilds; and “specialists,” if they were linked to only 1 or 2 guilds.

2.6. Statistical analyses and phylogenetic

The non-metric multidimensional scaling (NMDS) method was employed to explore the clustering patterns of biopolymers and sites based on the relative abundance of CAZyme families, specifically focusing on Auxiliary Activities (AAs) and Glycoside Hydrolases (GHs). This analysis aimed to reveal how biopolymers are distributed according to their functional CAZyme profiles. Before conducting the NMDS analysis, relative abundances were transformed using the square root method (Legendre & Legendre, 2012). Ordination was carried out based on the Bray-Curtis dissimilarity index, utilizing the metaMDS() function from the vegan package (Oksanen et al., 2019) in R (R-Core-Team, 2023).

The phylogenetic analysis of MAGs was performed to infer their evolutionary relationships. Taxonomic classification of MAGs was conducted using GTDB-Tk v2.3.2 with the *classify_wf* command and the reference GTDB database version r214 (Chaumeil et al., 2022; Parks et al., 2022). The phylogenetic tree was constructed employing IQ-TREE v2.2.6, implemented in GToTree v1.8.4 (Lee, 2019). The model LG+F+I+R9 was selected based on the Bayesian Information Criterion (BIC), and branch support was assessed with 1,000 bootstrap replicates (Kalyaanamoorthy et al., 2017; Nguyen et al., 2015). The

resulting tree was visualized and annotated using iTOL (Interactive Tree of Life) (Letunic & Bork, 2021) and ggtree v3.8.2 (G. Yu et al., 2017) in R version 4.3.1 (R-Core-Team, 2023), utilizing additional packages from the tidyverse suite, including tidylog v1.0.2 and tidyverse v2.0.0 (Wickham et al., 2019).

To determine whether there were statistically significant differences in the relative abundance of CAZyme families across the studied biopolymers, a one-way analysis of variance (ANOVA) was performed. Post hoc pairwise comparisons were conducted using the Tukey HSD test to identify specific differences between groups. Both analyses were implemented in R version 4.3.1 (R-Core-Team, 2023), utilizing the stats package.

3. RESULTS

3.1. Main characteristics of the metagenome and the metatranscriptome of meshbags

The metagenome sequencing yielded a total of 3,989,347,944 reads, with an average of 22.5 ± 7.2 million reads per sample, which were assembled into 17,936,557 contigs over 200 bp in length. In contrast, the metatranscriptome sequencing produced a total of 620,946,372 reads, with an average of 31.3 ± 9.1 million reads per sample, assembled into 1,332,519 contigs over 200 bp in length.

In Figure 35, the distribution of read counts associated with Bacteria, Fungi, and Other Eukaryota is shown. As observed, the metagenome exhibits a higher number of reads associated with the domain Bacteria compared to the metatranscriptome, particularly for the phylum Proteobacteria. In the metagenome, 4,229 contigs were assigned to Bacteria, 1,882 contigs to Fungi, and 3,719 contigs to Other Eukaryota, while 1,516 contigs did not match any of these classifications. In contrast, the metatranscriptome shows a higher read count for Fungi, with Basidiomycota being the most represented phylum, whereas Ascomycota dominates in the metagenome. In the metatranscriptome, 794,439 contigs were assigned to Bacteria, 389,225 contigs to Fungi, and 52,469 contigs to Eukaryota, while 427,304 contigs remained unassigned.

3.2. Microbial community composition in biopolymer-containing mesh bags

The taxonomic composition and activity of microbial communities associated with the degradation of biopolymers revealed distinct patterns of dominance across different substrates.

In the metagenome, the bacterial community showed significant variations in abundance depending on the biopolymer (Figure 36). For instance, Firmicutes were particularly dominant in substrates such as glucomannan, while Bacteroidetes exhibited a marked increase in abundance in chitin.

Notably, the microbial community associated with litter differed substantially from those found in the biopolymer-enriched mesh bags.

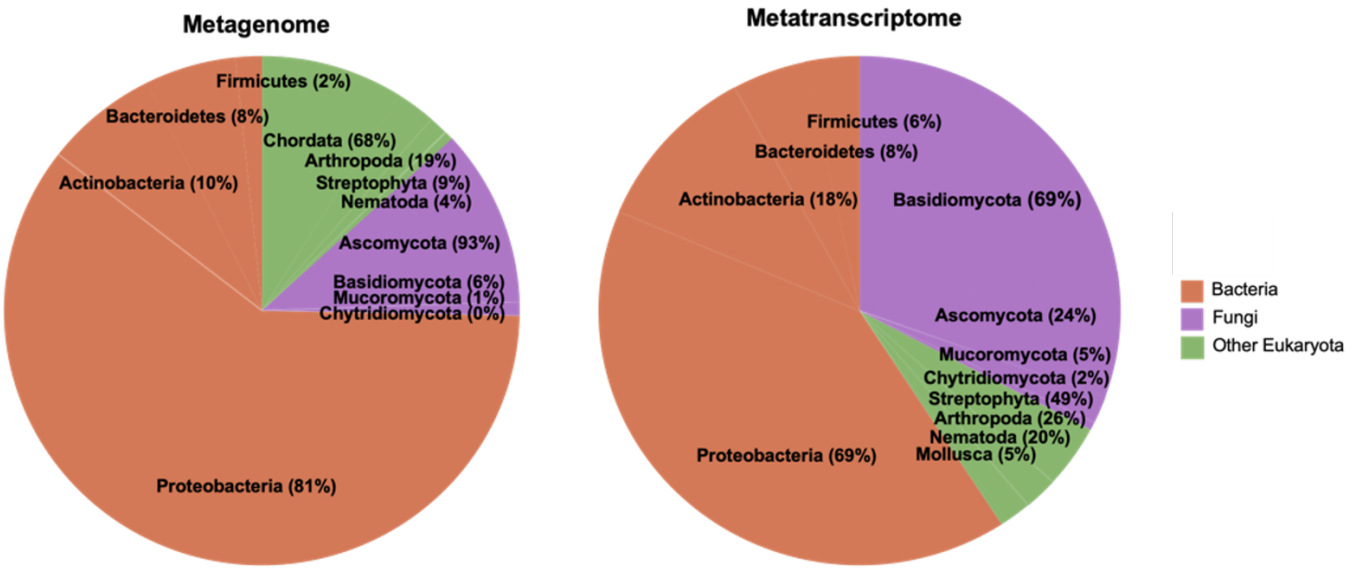


Figure 35: Taxonomic composition of microbial communities based on metagenomic (left) and metatranscriptomic (right) analyses: The abundance of bacterial, fungal, and eukaryotic taxa is shown, highlighting differences in taxonomic representation between the DNA (metagenome) and RNA (metatranscriptome) datasets.

The litter community was characterized by a higher abundance of Actinobacteria, which were less prevalent in the mesh bags. However, Actinobacteria were also found to be highly abundant in pectin. Furthermore, the distribution of Acidobacteria varied significantly across substrates. While this phylum was nearly absent in chitin, it was highly abundant in cellulose, xylan, and beta-1,3-glucan.

Further, we selected the 15 most abundant taxa at the genus level from each biopolymer to analyze the distribution and specialization of microbial communities. The bacterial microbiome (Figure 37) exhibited a diverse distribution across the studied biopolymers, with specific genera showing strong substrate preferences.

For instance, the genera *Mucilaginibacter* and *Pedobacter*, belonging to the phylum Bacteroidetes, were found to colonize chitin exclusively, with no presence detected in other substrates. In contrast, *Terriglobus*, a genus of Acidobacteria, was highly abundant in beta-1,3-glucan but absent in other substrates. Among Actinobacteria, the most abundant genera were exclusively associated with litter. Additionally, the Cyanobacteria *Nostoc* was uniquely identified in lignin. Alphaproteobacteria were particularly significant in cellulose, with *Rhizobium* and *Sphingomonas* being the most prominent genera. These genera were also detected in xylan, beta-1,3-glucan, beta-1,3-1,6-glucan, chitin, lignin, and pectin but were absent in glucomannan. Betaproteobacteria were notably abundant in pectin and were present in all substrates except glucomannan. Gammaproteobacteria, while present across all substrates, showed limited genus-level specialization in cellulose and pectin, with *Pseudomonas* being the most significant genus. In contrast, Gammaproteobacteria were the most dominant bacterial family in glucomannan.

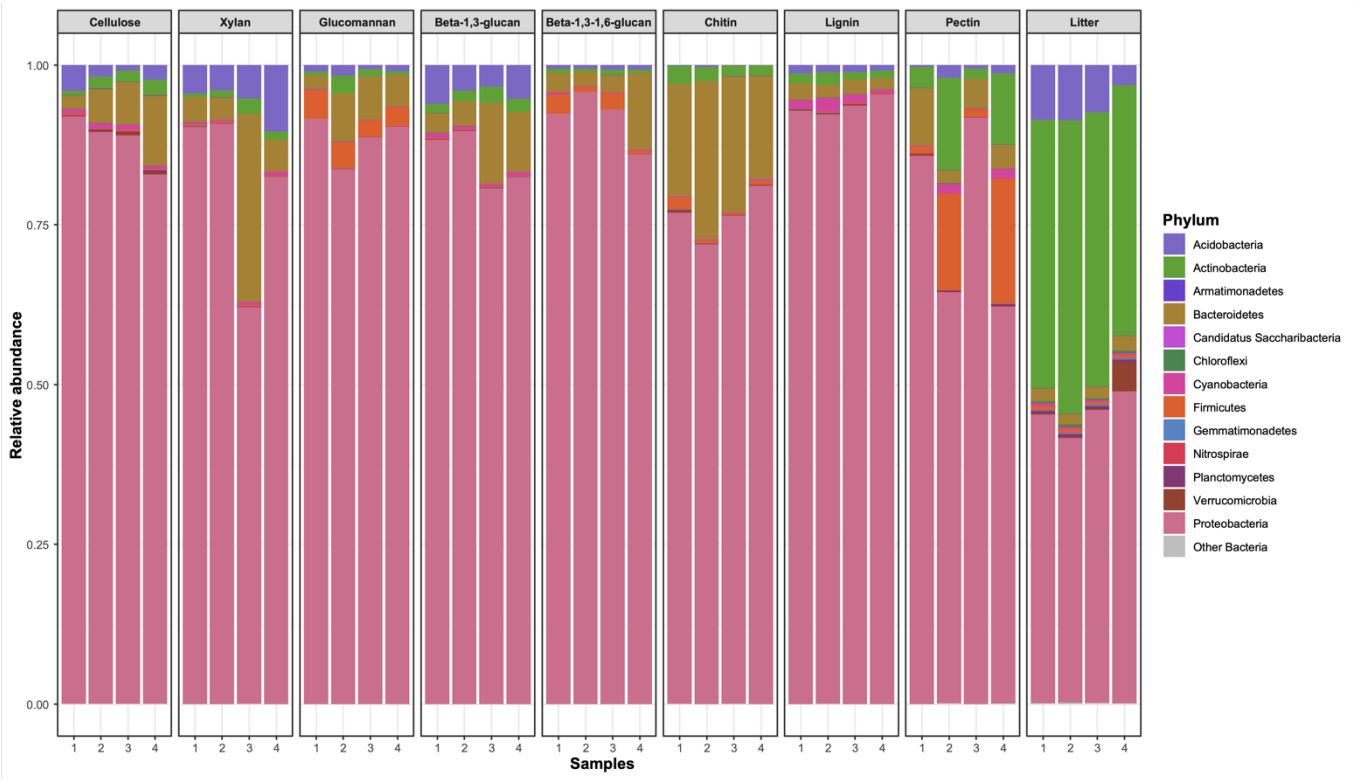


Figure 36: Taxonomic distribution of the bacterial community in the metagenome at the phylum level: The relative abundance of bacterial phyla is shown for each biopolymer: cellulose, xylan, glucomannan, beta-1,3-glucan, beta-1,3-1,6-glucan, chitin, lignin. On the x-axis the different sites are distinguished.

Regarding the taxonomic composition and activity of the fungal communities in the metagenome, we found equally different patterns of dominance on the different substrates (Figure 38).

Ascomycota proved to be the most abundant phylum on all substrates. Notably, Basidiomycota showed a strong preference for specific substrates, being particularly dominant on cellulose and beta-1,3-glucan, while their presence was lower on lignin and completely absent on pectin. In contrast, the Mucoromycota showed a marked specialization for glucomannan, beta-1,3-glucan and chitin, with lower presence in cellulose, beta-1,3-1,6-glucan, lignin and pectin. Chytridiomycota, on the other hand, were exclusively associated with cellulose. A striking observation was the almost exclusive colonization of pectin by Ascomycota.

In addition, we selected the 15 most abundant fungal taxa at the genus level from each biopolymer to analyze the distribution and specialization of the fungal communities. The fungal microbiome (Figure 39) showed a diverse distribution in the biopolymers studied, with specific genera and families showing strong substrate preferences.

For example, the fungal class Dothideomycetes, belonging to the phylum Ascomycota, colonized xylan, glucomannan, beta-1,3-glucan and lignin, while it was absent on cellulose, beta-1,3-1,6-glucan, chitin and pectin. The genus *Alternaria*, a key representative of this class, was particularly significant in beta-1,3-1,6-glucan, highlighting its specialized role in the degradation of this substrate.

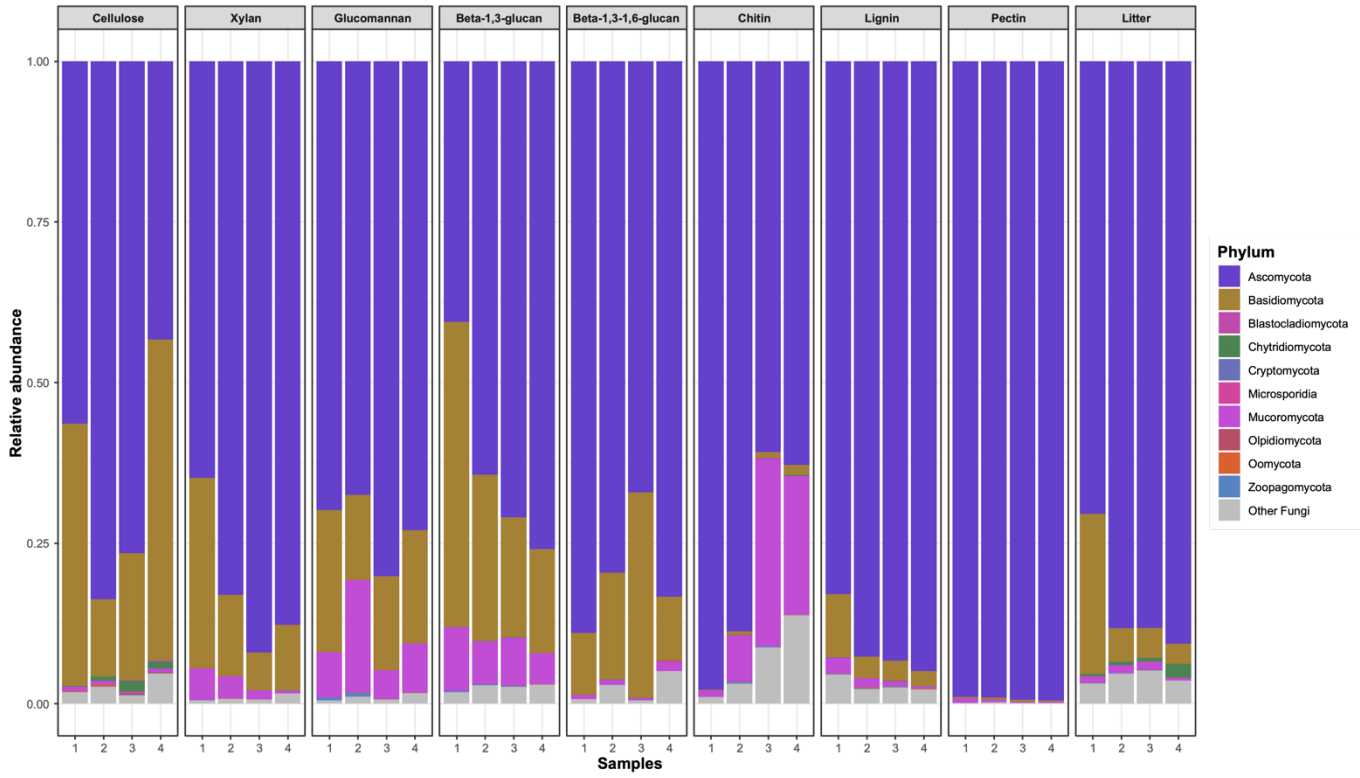


Figure 37: Taxonomic distribution of the fungal community in the metagenome at the phylum level: The relative abundance of fungal phyla is shown for each biopolymer: cellulose, xylan, glucomannan, beta-1,3-glucan, beta-1,3-1,6-glucan, chitin, lignin, pectin and litter. On the x-axis the different sites are distinguished.

The class Eurotiomycetes (Ascomycota) was highly dominant in chitin, lignin and pectin, with only a residual presence in cellulose, glucomannan and xylan, and total absence in beta-1,3-1,6-glucan. Notably, the genus *Penicillium*, a member of this class, colonized almost exclusively pectin and was also highly significant in chitin and lignin, underscoring its versatile but specialized role in the degradation of these biopolymers. The class Sordariomycetes (Ascomycota) was particularly dominant in beta-1,3-1,6-glucan, chitin, lignin and glucomannan, while it was absent in xylan and beta-1,3-glucan. The genera *Trichoderma* and *Chaetomium*, key representatives of this class, were mainly associated with beta-1,3-1,6-glucan and cellulose, respectively.

3.3. Transcriptional profiles of microbial communities in different biopolymers

The structure and activity of bacterial communities involved in biopolymer degradation, as analyzed through metatranscriptomics, revealed clear patterns of dominance and specialization linked to substrate type.

At the phylum level, bacterial abundance varied significantly depending on the biopolymer (Figure 40), reflecting distinct functional roles within each substrate.

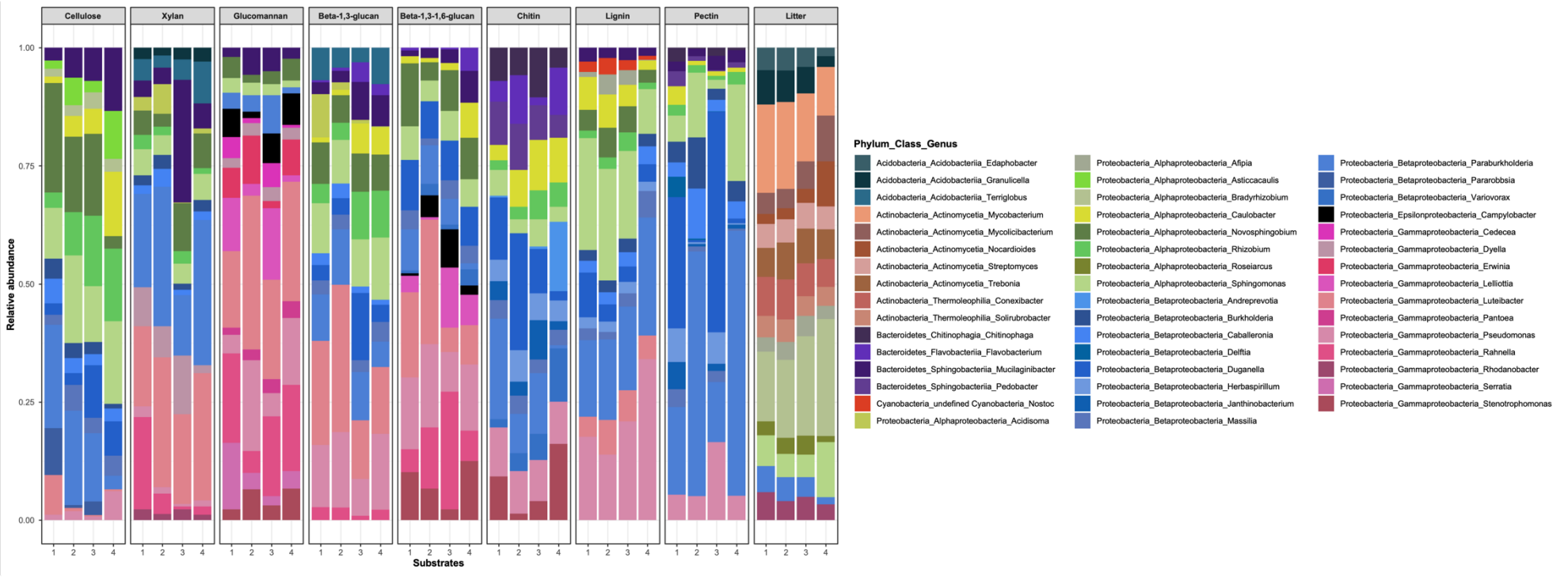


Figure 38: Taxonomic distribution of the bacterial community in the metagenome at the genus level: The relative abundance of bacterial genus is shown for each biopolymer: cellulose, xylan, glucomannan, beta-1,3-glucan, beta-1,3-1,6-glucan, chitin, lignin, pectin and litter. On the x-axis the different sites are distinguished.

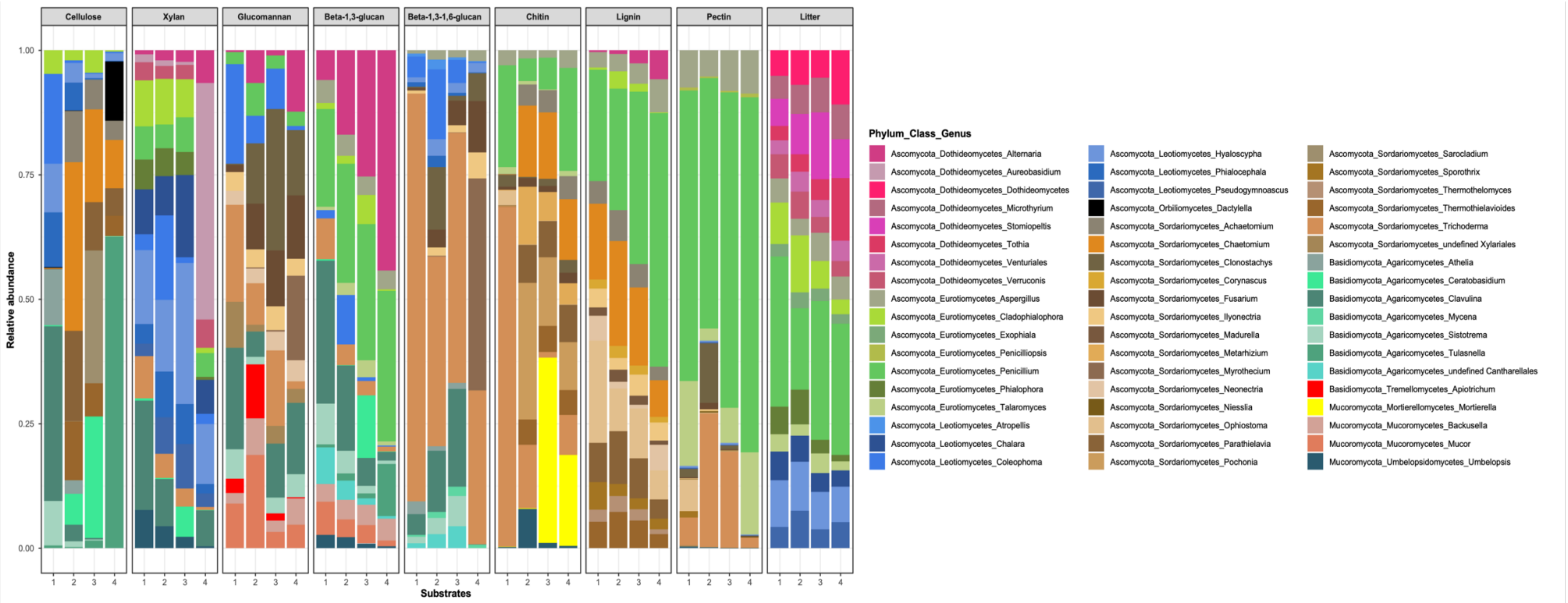


Figure 39: Taxonomic distribution of the fungal community in the metagenome at the genus level: The relative abundance of fungal genus is shown for each biopolymer: cellulose, xylan, glucomannan, beta-1,3-glucan, beta-1,3-1,6-glucan, chitin, lignin, pectin and litter. On the x-axis the different sites are distinguished.

Acidobacteria were prominently active in cellulose and beta-1,3-glucan but were less active in chitin. Proteobacteria dominated across most substrates. In contrast, the litter-associated community showed a high prevalence of Actinobacteria, which were less active in the biopolymer-enriched substrates. Chitin exhibited a distinct microbial profile, with Bacteroidetes and Firmicutes being significantly more active compared to other substrates. Additionally, Cyanobacteria were notably active in chitin, with low or absent activity in other substrates. Cellulose, however, was uniquely characterized by the presence of Planctomycetes, a phylum that was either absent or with low activity in other substrates.

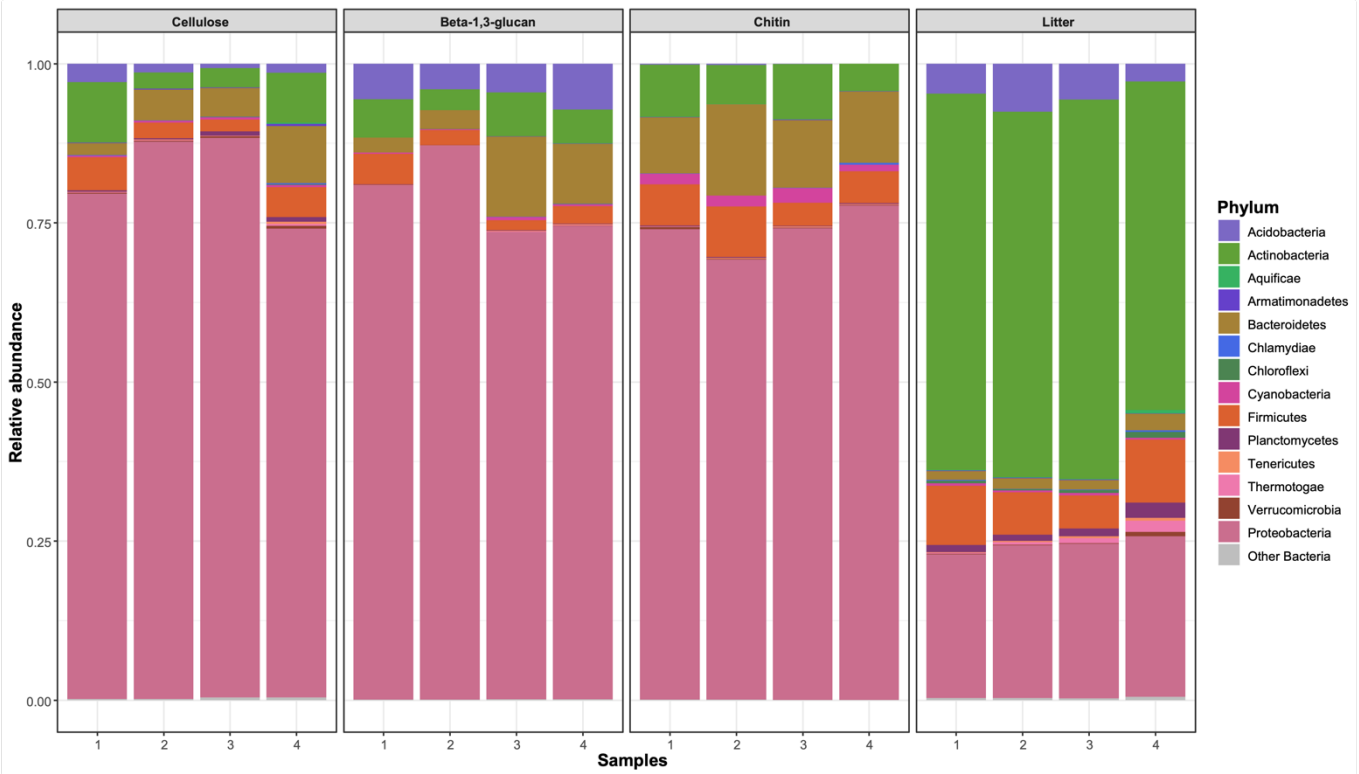


Figure 40: Taxonomic distribution of the bacterial community in the metatranscriptome at the phylum level: The relative abundance of bacterial phyla is shown for each biopolymer: cellulose, beta-1,3-glucan, chitin and litter. On the x-axis the different sites are distinguished.

Furthermore, we investigated the distribution and specialization of bacterial communities by identifying the 15 most active genera within each biopolymer. The bacterial microbiome (Figure 41) demonstrated a diverse presence across the analyzed substrates, with specific genera and phyla exhibiting distinct substrate preferences.

For instance, Actinobacteria were highly active in chitin, with *Streptomyces* emerging as the most characteristic and dominant genus within this phylum, specifically colonizing chitin. Likewise, the genera *Chitinophaga* and *Mucilaginibacter*, both belonging to the phylum Bacteroidetes, were exclusively associated with chitin. Similarly, the genus *Nostoc* (phylum Cyanobacteria) was uniquely found in chitin, further emphasizing its niche specialization. Conversely, the phyla Acidobacteria, Firmicutes, and Planctomycetes were predominantly found in the transcripts from litter. Within Proteobacteria, Alphaproteobacteria played a key role in the degradation of cellulose and beta-1,3-glucan, while their

activity in chitin was minimal. Among them, *Rhizobium* stood out as the most representative genus. In contrast, Betaproteobacteria were widely distributed across all substrates, demonstrating high adaptability, whereas Gammaproteobacteria were particularly abundant in beta-1,3-glucan and chitin but completely absent in cellulose.

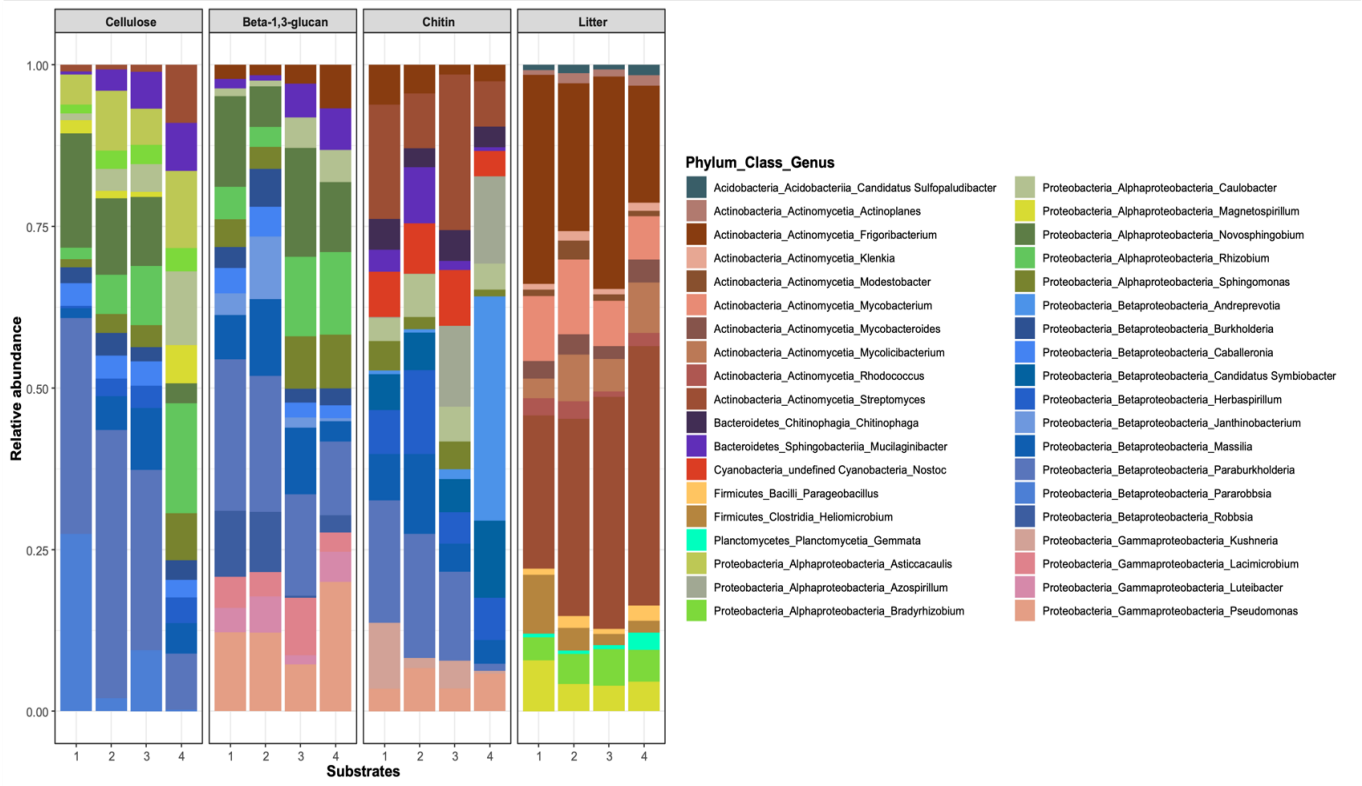


Figure 41: Taxonomic distribution of the bacterial community in the metatranscriptome at the genus level: The relative abundance of bacterial genus is shown for each biopolymer: cellulose, beta-1,3-glucan, chitin and litter. On the x-axis the different sites are distinguished.

As shown in Figure 42, the fungal community also exhibited distinct patterns of substrate colonization at the phylum level in the metatranscriptome.

Ascomycota predominantly colonized chitin, whereas Basidiomycota was the most abundant phylum in cellulose. Notably, Mucoromycota was also significantly active in chitin, though its abundance was minimal or residual in cellulose and beta-1,3-glucan. Similarly, Oomycota was detected in chitin but appeared only marginally in cellulose and beta-1,3-glucan. Additionally, Chytridiomycota was more active in chitin than in other substrates. The phylums Ascomycota and Basidiomycota were the dominant colonizers of chitin and cellulose, respectively. The presence of Mucoromycota, Oomycota, and Chytridiomycota in chitin, albeit at varying levels, suggests a complex interplay of fungal taxa in the degradation of this substrate.

Further, we examined the distribution and specialization of fungal communities by identifying the 15 most active genera within each biopolymer.

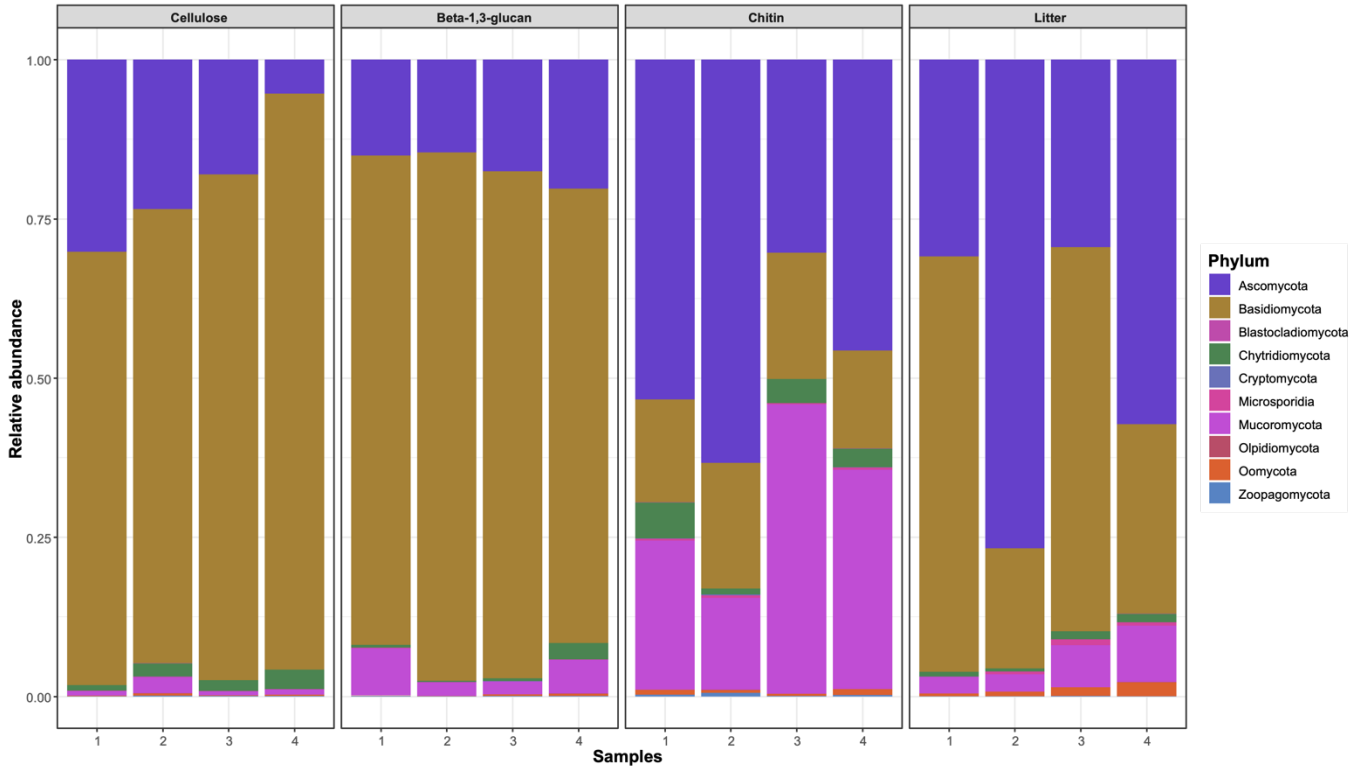


Figure 42: Taxonomic distribution of the fungal community in the metatranscriptome at the phylum level: The relative abundance of fungal phyla is shown for each biopolymer: cellulose, beta-1,3-glucan, chitin and litter. On the x-axis the different sites are distinguished.

The fungal microbiome (Figure 43) revealed a diverse and substrate-specific distribution, with certain genera and classes showing strong preferences for particular biopolymers.

Within the phylum Ascomycota, the genus *Lophium* was detected in both cellulose and chitin but was absent in beta-1,3-glucan. Notably, its presence was more pronounced in chitin than in cellulose. Similarly, within the class Eurotiomycetes, *Aspergillus* emerged as a key colonizer of cellulose and chitin, while it was completely absent in beta-1,3-glucan, highlighting its selective role in these biopolymers. The class Leotiomycetes exhibited a strong association with chitin, yet it was not detected in any other substrate. Likewise, members of the class Sordariomycetes were significantly enriched in chitin but showed minimal presence in cellulose or beta-1,3-glucan. In contrast, the class Agaricomycetes (Basidiomycota) played a major role in colonizing cellulose and beta-1,3-glucan, while its presence in chitin was minimal.

Within this class, *Tulasnella* was particularly abundant in cellulose. Meanwhile, the genus *Slooffia* (Microbotryomycetes) was exclusively found in beta-1,3-glucan. The class Tremellomycetes was also present in cellulose and beta-1,3-glucan but was entirely absent in chitin, reflecting its distinct substrate preferences. Regarding Mucoromycota, the genus *Mortierella* (Mortierellomycetes) was exclusively associated with chitin. Additionally, members of the class Mucoromycetes, including *Mucor*, were notably enriched in chitin but completely absent in cellulose.

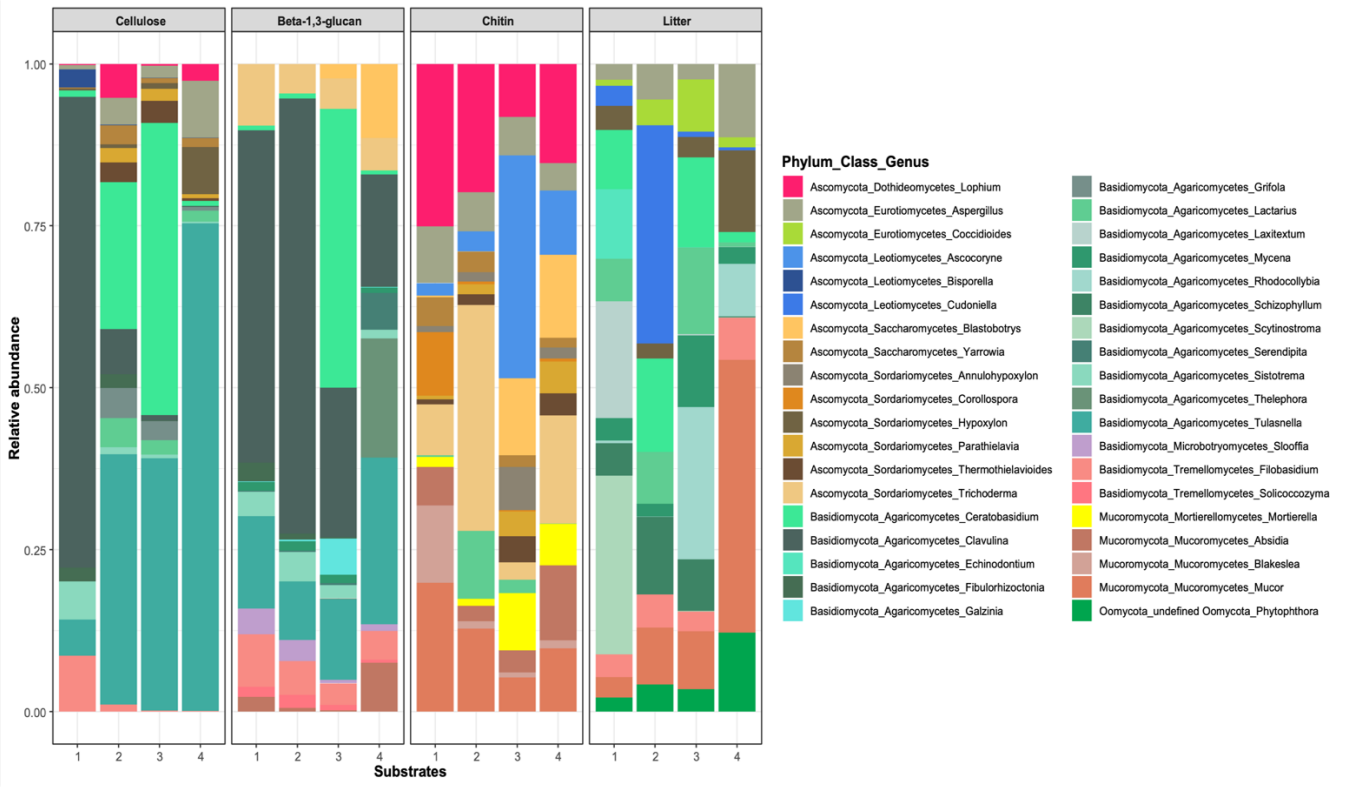


Figure 43: Taxonomic distribution of the fungal community in the metatranscriptome at the genus level: The relative abundance of fungal genus is shown for each biopolymer: cellulose, beta-1,3-glucan, chitin and litter. On the x-axis the different sites are distinguished.

3.4. Functional diversity of enzymes involved in the decomposition of polymers of plant and fungal origin.

The ordination of samples based on CAZyme composition clearly showed a separation between substrates for both fungi and bacteria, in both the metagenome and metatranscriptome.

Separation was stronger in metatranscriptome samples, where the CAZyme pool of microbial communities growing in cellulose and beta-1,3-glucans were clustered closer together than those being expressed in chitin and litter. This pattern was confirmed by NMDS analysis, which evaluated the functional distribution of bacterial and fungal communities in relation to biopolymer degradation, considering both metagenomic and metatranscriptomic data (Figure 44).

The analysis was based on the grouping of CAZyme families into AAs (Auxiliary Activities) and GHs (Glycosyl Hydrolases), allowing for the assessment of differences in microbial functionality based on biopolymers and sampling sites. The substrates analyzed in the metagenome included cellulose, xylan, glucomannan, beta-1,3-glucan, beta-1,3-1,6-glucan, chitin, lignin, pectin, and litter, while in the metatranscriptome, cellulose, beta-1,3-glucan, chitin, and litter were considered. The sampling sites were four, identified as 1, 2, 3, and 4.

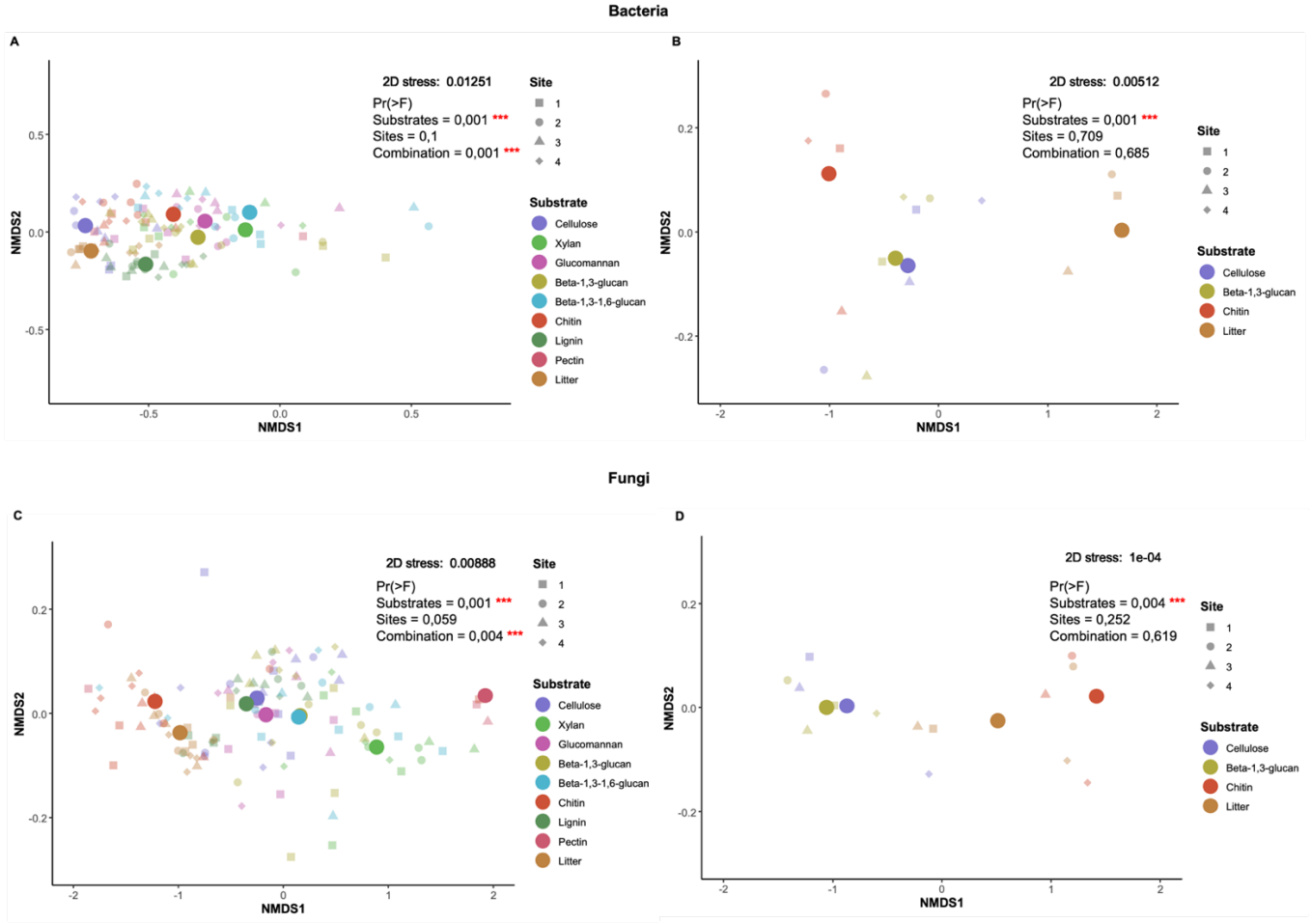


Figure 44: Non-metric multidimensional scaling (NMDS) analysis of functional distribution in bacterial and fungal communities as a function of CAZyme families (AAs and GHs). Panel A represents the metagenome information in the bacterial community. Panel B represents the metatranscriptome information in the bacterial community. Panel C represents the metagenome information in the fungal community. Panel D represents the metatranscriptome information in the fungal community. The substrates analyzed in the metagenome include cellulose, xylan, glucomannan, beta-1,3-glucan, beta-1,3-1,6-glucan, chitin, lignin, pectin, and litter, while the metatranscriptome analysis includes cellulose, beta-1,3-glucan, chitin, and leaf litter. Sampling sites are 1, 2, 3 and 4.

In the case of bacterial communities, the NMDS analysis of the metagenome (Figure 44A) showed significant differences in functionality based on substrates (P = 0.001) and the combination of substrates and sites (P = 0.001), but no significant differences were observed between sites (P = 0.1). Furthermore, we highlight that the pectin substrate is not represented in the bacterial metagenome. This is due to its marked dissimilarity compared to other substrates in bacteria, as well as the extremely low abundance of associated genes, which is likely a result of fungal dominance in pectin degradation. Consequently, pectin appears as a distant outlier in the NMDS ordination, reflecting its minimal contribution to the bacterial functional profile. On the other hand, in the analysis of the bacterial metatranscriptome (Figure 44B), a significant influence of substrates was observed (P = 0.001), but no significant differences were detected either between sites (P = 0.709) or in the combination of substrates and sites (P = 0.685).

Regarding fungal communities, the NMDS analysis of the metagenome (Figure 44C) revealed significant differences in functionality based on substrates (P = 0.001) and the combination of substrates and sites (P = 0.004), but no significant differences were observed between sites (P = 0.059). In the analysis of the fungal metatranscriptome (Figure 44D), a significant influence of substrates was observed (P = 0.004), but no significant differences were detected either between sites (P = 0.252) or in the combination of substrates and sites (P = 0.619).

In the analysis of the functional diversity of enzymes involved in the decomposition of plant- and fungal-origin polymers, it was observed that in the metagenome, gene assignment to CAZyme families ranged from 0.875% to 13.9% in fungi and from 0.183% to 21.2% in bacteria. In the metatranscriptome, the percentages of genes assigned to CAZymes in fungi ranged from 1.02% to 20.5%, while in bacteria, they varied from 0.576% to 43.7%.

In fungi, the most abundant families in the metagenome were "oxidoreductases" (13.9%) and "betaglucanases" (12.8%), whereas in the metatranscriptome, "oxidoreductases" (20.5%) and "betaglucanases" (16.1%) predominated. In bacteria, the most prominent families in the metagenome were "other hemicellulases" (21.2%) and "alphaglucanases" (18.2%), while in the metatranscriptome, "alphaglucanases" (43.7%) and "peptidoglycanases" (23.9%) stood out. These results reflect the relative importance of different CAZyme families in the decomposition of polymers, both in fungi and bacteria, and their differential contributions in the metagenome and metatranscriptome.

Figure 45 illustrates the distribution of genes (Figure 45A) and transcriptional expression (Figure 45B) of CAZyme families associated with the degradation of the studied biopolymers in fungi. In Figure 45A, the proportion of genes is distributed across the substrates, reaching values of up to approximately 15%. In Figure 45B, the proportion of transcription exhibits greater variability, with some substrates showing higher levels of expression. Notably, the transcription of genes related to the degradation of cellulose and beta-1,3-glucan is significantly higher compared to other substrates, while chitin displays lower transcriptional activity.

In fungi, CAZyme families such as cellulases and betaglucanases show high transcriptional activity, particularly in the degradation of cellulose and beta-1,3-glucan, respectively. This suggests that fungi play a significant role in breaking down these substrates. In contrast, families like chitinases exhibit lower transcriptional activity, indicating a reduced role in chitin degradation under the studied conditions. Interestingly, in the metagenome (Figure 45A), a higher presence of genes from CAZyme families involved in pectin degradation is observed, unlike in bacteria, where their presence is minimal.

Some CAZyme families, such as betaglucanases, cellulases, and oxidoreductases, are highly represented in terms of transcription, highlighting their active role in the degradation of cellulose and beta-1,3-glucan. Conversely, families like peptidoglycanases, peroxidases, and pectinases show lower representation in

both the metatranscriptome and metagenome. A notable observation is that, while CAZyme families involved in cellulose and beta-1,3-glucan degradation are highly active in the metatranscriptome, their presence in the metagenome is relatively limited (Figure 45A).

This underscores the dynamic transcriptional response of fungi to specific substrates, particularly cellulose and beta-1,3-glucan, while also revealing their limited engagement in chitin degradation under these conditions.

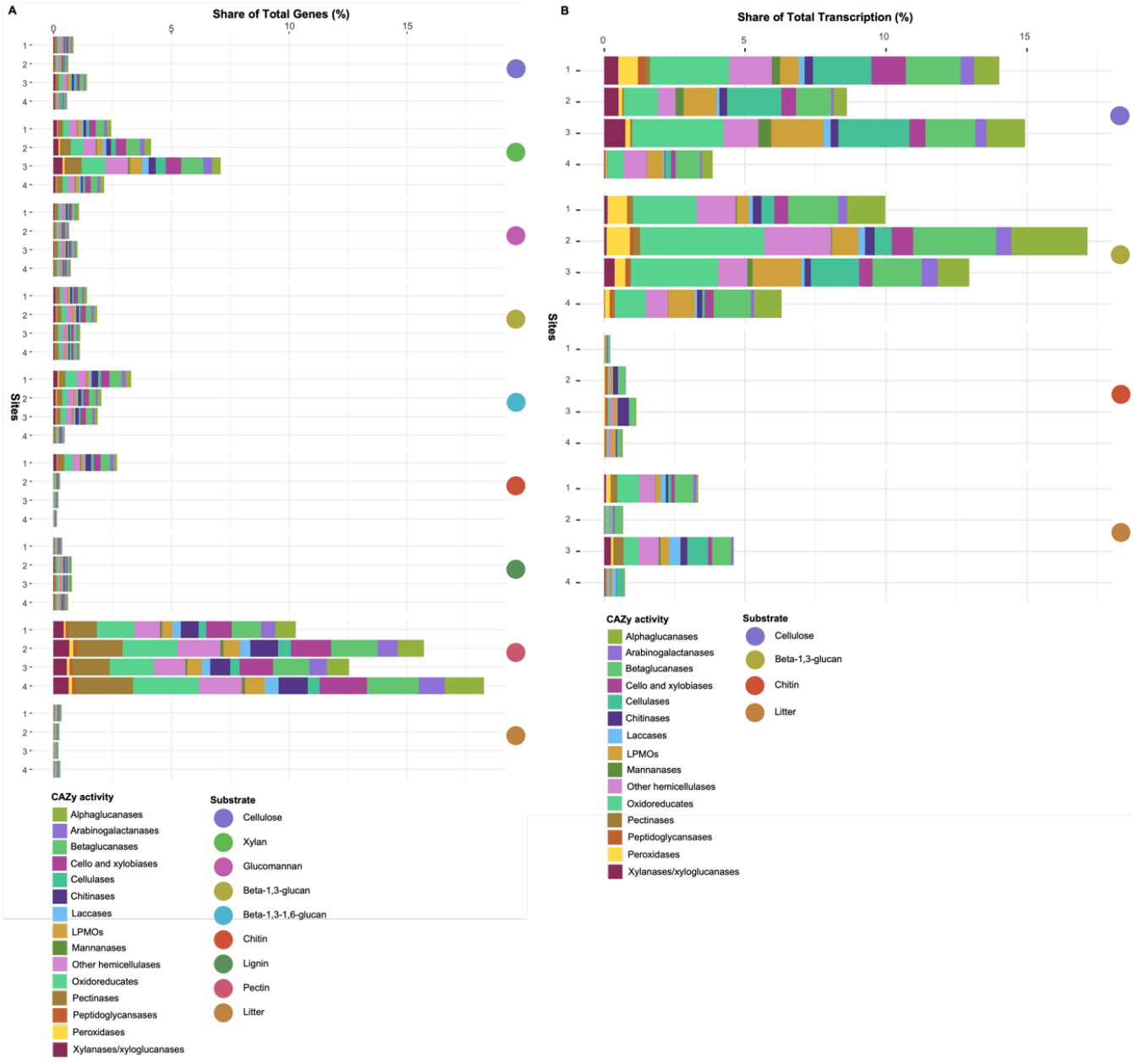


Figure 45: Distribution of genes (A) and transcriptional expression (B) of CAZyme families associated with the degradation of biopolymers in fungi. Panel A shows the proportion of genes distributed across substrates: cellulose, xylan, glucomannan, beta-1,3-glucan, beta-1,3-1,6-glucan, chitin, lignin, pectin, and litter. Panel B represents the proportion of transcription across the substrates: cellulose, beta-1,3-glucan, chitin and litter, highlighting significant variations among them. The x-axis distinguishes the different sampling sites.

Figure 46 illustrates the distribution of genes (Figure 46A) and transcriptional expression (Figure 46B) of CAZyme families associated with the degradation of the studied biopolymers in bacteria.

In Figure 46A, the proportion of genes is relatively evenly distributed across the substrates, with a maximum of around 4%. In contrast, Figure 46B reveals a distinct pattern: the proportion of transcription varies significantly, reaching up to 20% for substrates such as chitin. Notably, the high expression of CAZyme families related to the degradation of cellulose and chitin stands out, particularly in the case of chitin. This highlights the significant role of bacteria in chitin degradation, with families such as chitinases and peptidoglycanases being highly overrepresented in terms of transcription.

Additionally, differences are observed among sampling sites. Some sites show a higher proportion of genes associated with the degradation of certain substrates in the metagenome, while others exhibit higher levels of transcription for different biopolymers. For example, in the case of site 2 for the cellulose biopolymer, the genes of the mannanases CAZyme family were more abundant than in the other sites where this biopolymer was sampled. It is also noted that certain CAZyme families, such as alphaglucanases, chitinases, and peptidoglycanases, are overrepresented in terms of transcription, while others, such as pectinases and mannanases, show lower relative activity despite being present in the metagenome.

When trying to associate the CAZymes with specific phyla, we observed differences between metagenome and metatranscriptome (Figure 47). As shown in the metagenomic data (Figure 47A), on the pectin substrate, the phylum Ascomycota entirely accounted for all CAZyme families. On the chitin substrate, the phylum Mucoromycota predominantly harbored peptidoglycanases and a significant portion of peroxidases, while Ascomycota almost exclusively possessed laccases, mannanases, and xylanases/xyloglucanases. On the beta-1,3-glucan substrate, the phylum Basidiomycota contained the majority of cellulases, alpha-glucanases, and xylanases/xyloglucanases, whereas Ascomycota almost entirely accounted for peroxidases, oxidoreductases, and laccases. On the cellulose substrate, Basidiomycota was found to possess the majority of cellulases, xylanases/xyloglucanases, and mannanases, while Ascomycota almost exclusively harbored peroxidases, oxidoreductases, and laccases.

In the metatranscriptomic data (Figure 47B), Basidiomycota was observed to express the vast majority of beta-glucanases, peptidoglycanases, laccases, oxidoreductases, and peroxidases. On the beta-1,3-glucan and cellulose substrates, Basidiomycota accounted for nearly all the studied CAZyme families, with the exception of peptidoglycanases, which were also partially expressed by the phylum Mucoromycota.

Similarly, when linking CAZyme content and expression with specific bacterial phyla, we found that in the metagenomic data (Figure 47A), on the pectin substrate, the bacterial phylum Actinobacteria predominantly harbored cellulases, mannanases, other hemicellulases, and chitinases, while the phylum

Proteobacteria contained the majority of peptidoglycanases and arabinogalactanases. On the chitin substrate, the phylum Bacteroidetes exhibited the highest abundance of beta-glucanases, pectinases, arabinogalactanases, other hemicellulases, mannanases, xylanases/xyloglucanases, and cellulases. On the beta-1,3-glucan substrate, the phylum Acidobacteria was found to possess a significant number of pectinases, while Bacteroidetes showed a high abundance of mannanases and pectinases. On the cellulose substrate, Proteobacteria were identified as the primary carriers of cellulases.

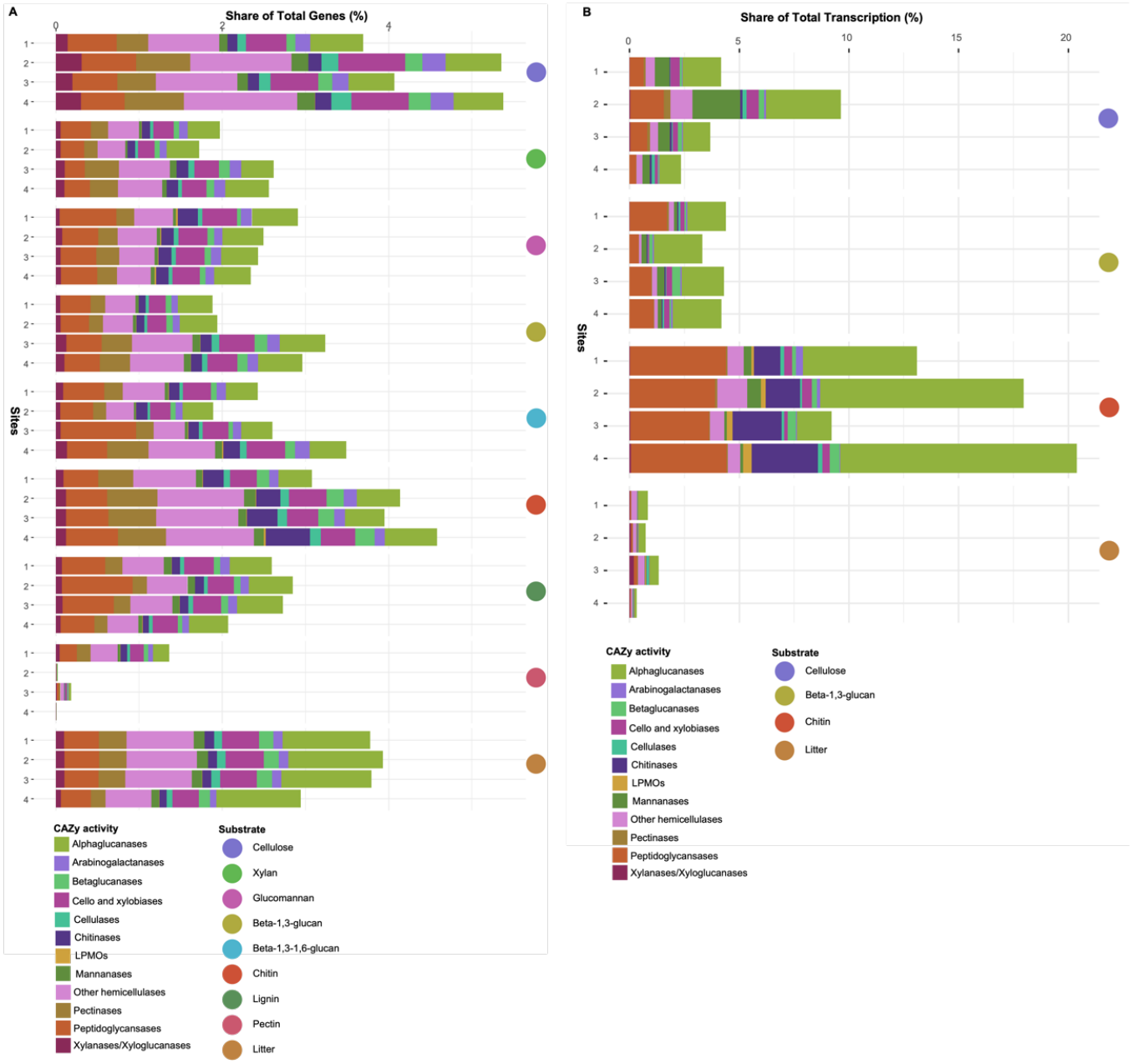


Figure 46: Distribution of genes (A) and transcriptional expression (B) of CAZyme families associated with the degradation of biopolymers in bacteria. Panel A shows the proportion of genes distributed across substrates: cellulose, xylan, glucomannan, beta-1,3-glucan, beta-1,3-1,6-glucan, chitin, lignin, pectin, and litter. Panel B represents the proportion of transcription across the substrates: cellulose, beta-1,3-glucan, chitin and litter, highlighting significant variations among them. The x-axis distinguishes the different sampling sites.

In the metatranscriptomic data (Figure 48B), on the chitin substrate, the phylum Bacteroidetes displayed higher expression of xylanases/xyloglucanases, mannanases, pectinases, and beta-glucanases, whereas Firmicutes were associated with the expression of cellulases. On the beta-1,3-glucan substrate, Bacteroidetes were found to express cello- and xylobiases, cellulases, and arabinogalactanases, while Proteobacteria accounted for the entirety of pectinase expression. On the cellulose substrate, Proteobacteria dominated the expression of alpha-glucanases, cello- and xylobiases, cellulases, mannanases, pectinases, chitinases, other hemicellulases, beta-glucanases, and peptidoglycanases, whereas Bacteroidetes were primarily responsible for the expression of arabinogalactanases.

3.5. Phylogenetic and functional diversity of the MAGs recovered.

In the present study, a total of 209 metagenome-assembled genomes (MAGs) of high-quality (>70% completeness and <10% contamination) were recovered, all belonging exclusively to the domain Bacteria (Figure 49). These MAGs were distributed across 9 phyla, including Pseudomonadota (formerly Proteobacteria), Acidobacteriota (formerly Acidobacteria), Bacteroidota (formerly Bacteroidetes), Patescibacteria, Bdellovibrionota, Myxococcota, Bacillota (formerly Firmicutes), Eremiobacterota, and Actinobacteriota (formerly Actinobacteria). Particularly noteworthy is the recovery of MAGs belonging to less common phyla in soil, such as Patescibacteria (class Saccharimonadia), Myxococcota, and Eremiobacterota.

Regarding the distribution and abundance of the analyzed bins across different biopolymer substrates, a classification into guilds (generalists and specialists) and non-guilds was observed, as described in (Algora et al., 2022) (Figure 49). This classification was based on the relative abundance of bins in the substrates rather than their CAZyme content. The generalists were further divided into two groups: broad-generalists and narrow-generalists. The classification of these taxa as specialists or generalists was determined not only by their abundance in a given substrate but also by their ability to encode and express the enzymes required for substrate degradation. Figure 50 summarize the CAZyme activities expressed by key bins, categorized into complex plant biomass (LPMOs, cellulases and xylanases, mannanases), easy plant biomass (cellobiases, other hemicellulases, arabinogalactanases and pectinases), microbial biomass (chitinases, betaglucanases and peptidoglycanases), and reserve compounds (alphaglucanases).

Among the broad-generalists, bins belonging to the bacterial genus *Rhizobium* (e.g., bin.146, bin.159, bin.33, bin.6) and Sphingobium (bin.172) within the class Alphaproteobacteria were notable for their presence across all studied biopolymers and their activity in the three biopolymers and the litter analyzed in the metatranscriptome. Similarly, bins belonging to the genus *Paraburkolderia* (e.g. bin.209, bin.4) in Gammaproteobacteria also were higly active in the all the biopolymers. As for the narrow-generalists, bins associated with the genus *Flavobacterium* (e.g., bin.101, bin.108, bin.64, bin.57, bin.94, bin.129, bin.143, bin.79, bin.109, bin.89) primarily colonized the substrates chitin, litter, and beta-1,3-glucan, while other *Flavobacterium*-associated bins were also found in cellulose and beta-1,3-1,6-glucan.

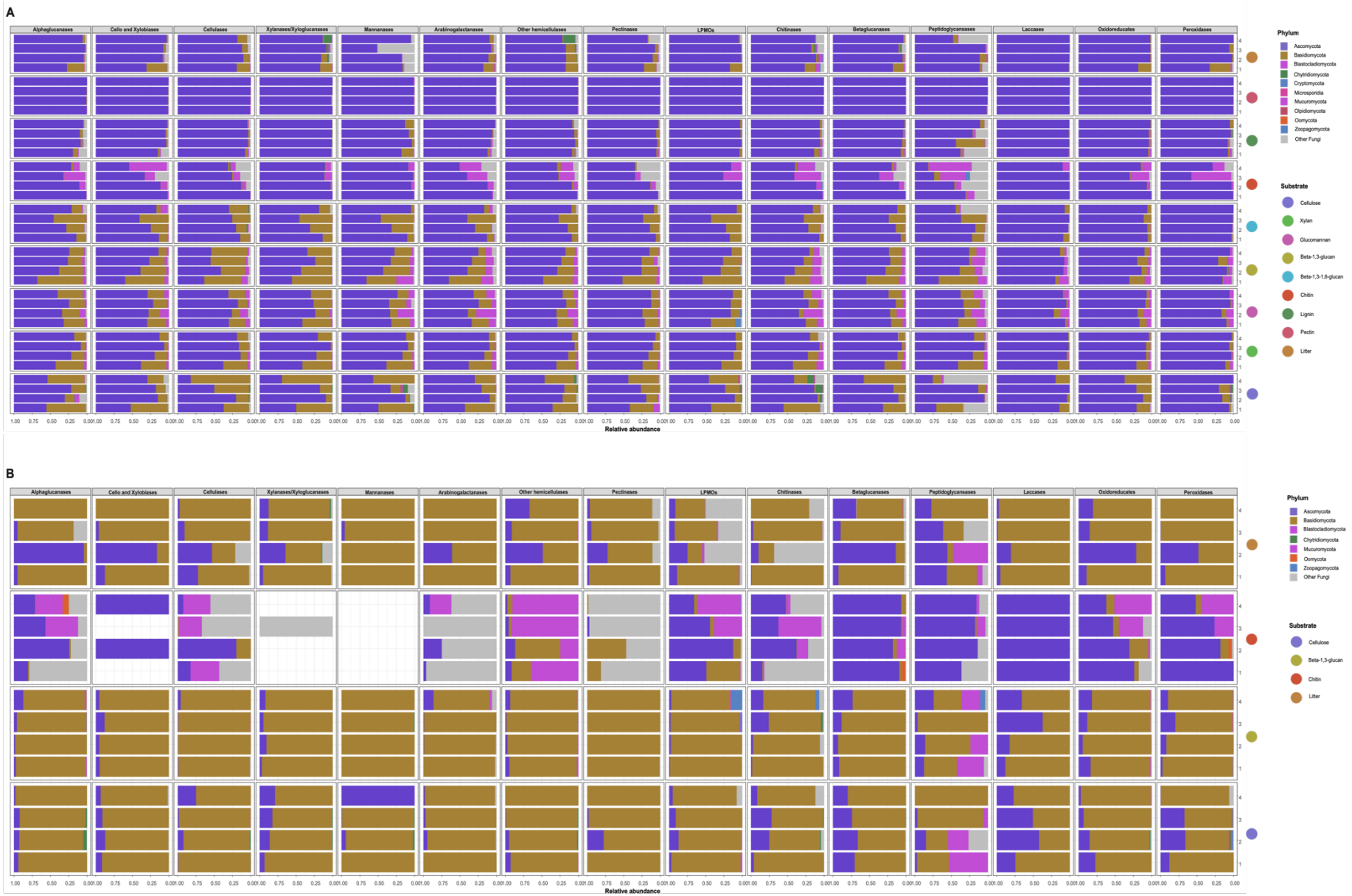


Figure 47: Relative abundance of each CAZyme family associated with fungal phyla across different substrates in the metagenome (Panel A) and the metatranscriptome (Panel B).

Figure 48: Relative abundance of each CAZyme family associated with bacterial phyla across different substrates in the metagenome (Panel A) and the metatranscriptome (Panel B).

The phylogenetic distribution of MAGs revealed a remarkable concentration of specialists in certain phyla, especially in Proteobacteria, Bacteroidetes and Acidobacteria. Figure 50 shows the expression profiles of CAZymes in specialist bins of three key substrates: cellulose, beta-glucan and chitin.

For cellulose, bins bin.151 and bin.207 from the genus *Asticcacaulis* (Alphaproteobacteria) were identified as specialists, expressing various cellulases belonging to the GH5, GH130 and GH26 families (Figure 50A and 50B), all of which fall into the category of complex plant biomass. In the case of beta-glucan, bins bin.57 and bin.94 of the genus *Flavobacterium* (Bacteroidetes) showed elevated expression of beta-glucanases, including enzymes of the GH81 and GH144 families (Figure 50C and 50D), both belonging to the microbial biomass category. Regarding chitin degradation, bin.70 of the genus *Pedobacter*, together with bin.152 of the genus *Cellvibrio* (Gammaproteobacteria), expressed multiple chitinases (GH18 and GH20) and peptidoglycanases (GH23 and GH73) (Figure 50E and 50F), all belonging to the microbial biomass category.

Additionally, other specialists were observed that, although not shown in the figure, presented outstanding activity on some of the substrates tested. For example, bin.188 (*Andreprevotia*) expressed chitinases of the GH18 family associated with chitin degradation, while bin.93 (*Terriglobus*, Acidobacteria) expressed GH55 and GH144, both associated with beta-glucan degradation.

In contrast, other bins such as those of the phylum Eremiobacterota (e.g., bin.15_Tumulicola, bin.140_Cybelea) showed activity in the degradation of litter, expressing enzymes such as GH130, GH20 or CE1, but did not show relevant expression in the specific substrates analyzed (cellulose, beta-glucan or chitin).

Several bins with low similarity to previously described taxa were recovered. Among these, bin.21 *JACMQM01* within Bacteroidota, associated with chitin degradation, and bin.155 in Bacteroidia, specialist for cellulose degradation, as well as bin.84 *UBA1573*, bin.114 *QFOX01*, and bin.177 *JAJPHO01* in Alphaproteobacteria, are particularly noteworthy. These bins not only represent novel taxa within known phyla but may also play key functional roles in the decomposition of specific biopolymers, such as chitin and other components of microbial and plant biomass.

4. DISCUSSION

4.1. Microbial decomposers preferences for different components of dead biomass confirms the existence of decomposers guilds

Our experiment confimed the existence of decomposer guilds in the microbial community of forest soil, as previously revealed by Algora et al. (2021, 2022). The specialization patterns observed in the microbial community suggest that the decomposition of dead biomass is not a homogeneous process but is structured based on the presence of functional guilds both in fungi and bacteria.

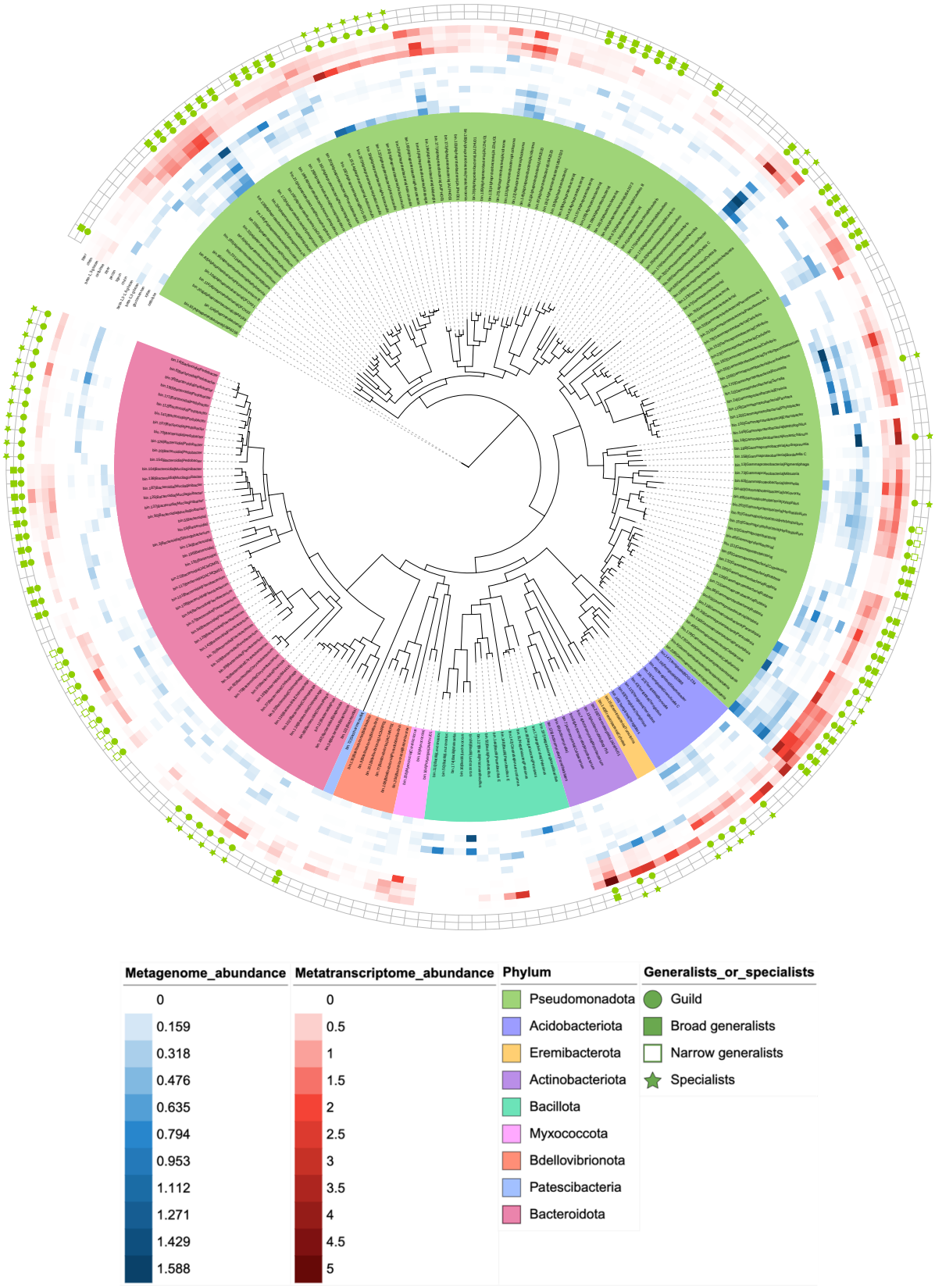


Figure 49: Phylogenetic tree of bacterial MAGs and their functional roles in biopolymer degradation. The tree is differentiated by phyla: The first ring (blue) represents the abundance of genes associated with the degradation of substrates: cellulose, xylan, glucomannan, beta-1,3-glucan, beta-1,3-1,6-glucan, chitin, lignin, pectin, and leaf litter. The second ring (red) indicates the transcriptional activity of these genes in the substrates: cellulose, beta-1,3-glucan, chitin, and litter. The symbols denote guild classification: broad generalists (filled green square), narrow generalists (empty green square), and specialists (star).

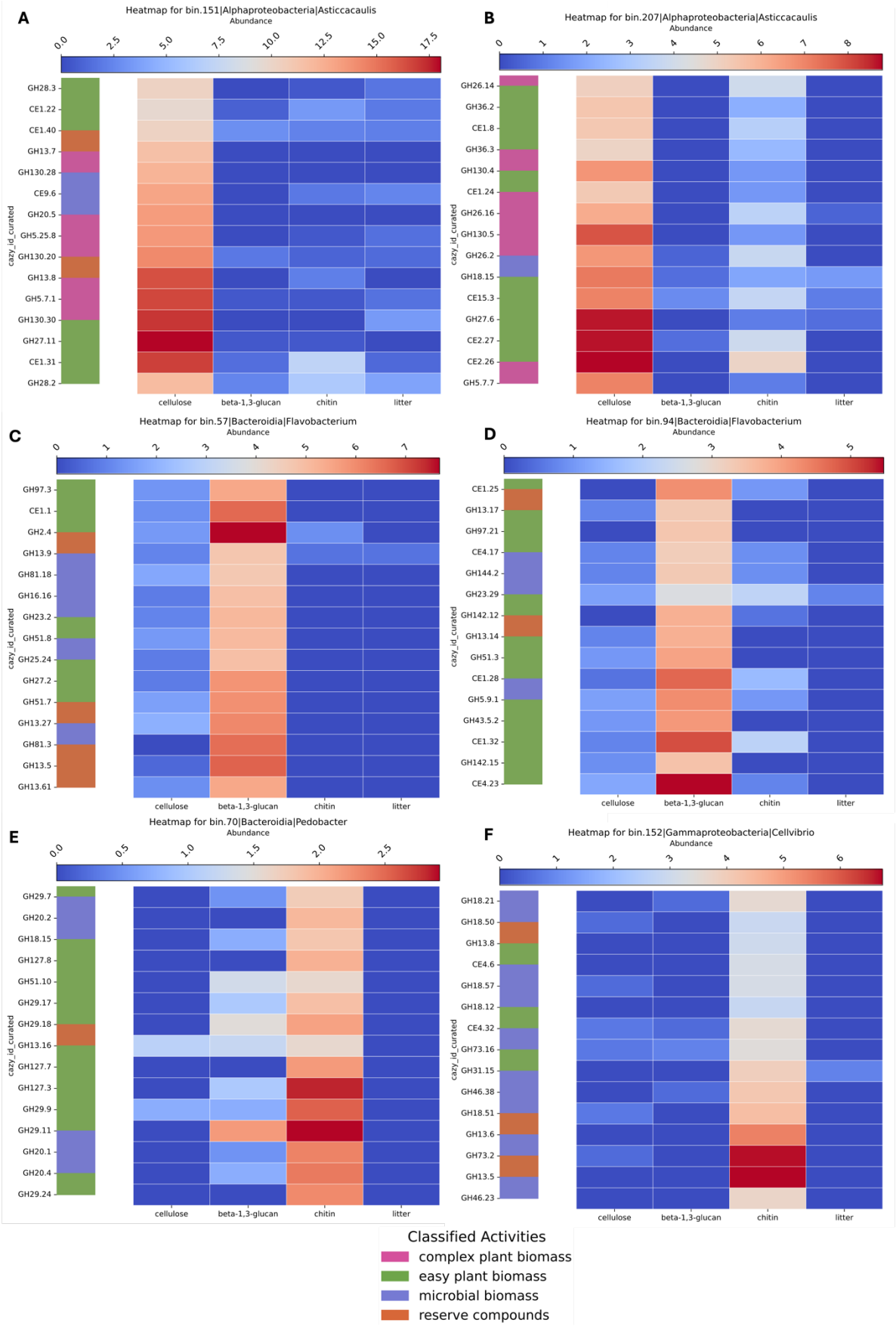


Figure 50: Heatmaps showing the differential expression of CAZymes in selected bins. The Y-axis represents the CAZyme IDs, and the X-axis represents the substrates: cellulose, beta-1,3-glucan, chitin, and the presence of litter.

The identification of bacterial and fungal genera with distinct biopolymer preferences reveals that competition and complementarity among microorganisms are regulated by substrate specificity, indicating that the degradation of organic matter in soils is governed by complex ecological interactions (C. Wang & Kuzyakov, 2024). The exclusive association of *Mucilaginibacter* and *Pedobacter* (phylum Bacteroidetes) with chitin, along with the specialization of *Terriglobus* (phylum Acidobacteria) in beta-1,3-glucan, suggests that certain bacterial groups have evolved to exploit specific niches within the organic matter cycle. The ability of Bacteroidetes to efficiently decompose chitin aligns with their life strategy as degraders of complex polymers, providing them with a competitive advantage in environments rich in this biopolymer (J. Huang et al., 2023; Wieczorek et al., 2019). Conversely, the specialization of Acidobacteria in beta-1,3-glucan suggests that these organisms may play a key role in the degradation of secondary microbial biomass (Ivanova et al., 2016), as this polysaccharide is an important structural component of fungal and some protist cell walls (Ruiz-Herrera & Ortiz-Castellanos, 2019). This guild-based structuring not only reflects differences in enzymatic capacity but also suggests that the coexistence of these groups is driven by resource partitioning, reducing direct competition and promoting the coexistence of diverse species within the same ecosystem (Nuccio et al., 2020). In fungi, the strong specialization of *Penicillium* (Ascomycota) in chitin and pectin, in contrast with the dominance of *Tulasnella* (Basidiomycota) in cellulose, reinforces the idea that fungal taxonomic distribution is closely linked to substrate composition (Bahram et al., 2021; Ye et al., 2019). This pattern aligns with the differential decomposition strategies of saprotrophs in Ascomycota and Basidiomycota: while the former are typically associated with the rapid degradation of more accessible polymers during the initial stages of decomposition, the latter play a key role in the breakdown of more recalcitrant materials such as cellulose and lignin (Brazkova et al., 2022; Manici, Caputo, De Sabata, et al., 2024), facilitating carbon mineralization in the later phases of organic matter decomposition.

Metagenomic and metatranscriptomic data support this functional differentiation, revealing that the expression of genes involved in biopolymer degradation is highly regulated and specific to each guild. The high transcription of chitinases and beta-glucanases in Bacteroidetes, compared to the greater expression of cellulases and xylanases in Proteobacteria, suggests that bacterial guilds are not only structured at the taxonomic level but that their metabolic activity is also defined by the availability of specific substrates (Nunan et al., 2020). This reinforces the idea that gene expression in soil microorganisms is tightly regulated in response to resource availability (Jansson & Hofmockel, 2018; Saleh-Lakha et al., 2005). The same principle is observed in fungi, where the overexpression of cellulose-associated enzymes in Basidiomycota and chitinases in Ascomycota confirms the functional specialization of these groups. The fact that these associations are reflected not only in taxonomic composition but also in transcriptional activity suggests that decomposer guilds are not merely static ecological assemblages but are actively engaged in the degradation of their preferred substrates (Beidler et al., 2020; López-Mondéjar et al., 2018).

The existence of decomposer guilds has important implications for carbon cycle dynamics in soils. By specializing in the degradation of specific biopolymers, these guilds determine the metabolic rates and pathways through which organic matter is processed and recycled in ecosystems (Ferreira et al., 2020; Žifčáková, 2017). For instance, an increase in the proportion of chitin in biomass would favor the development of chitinolytic microorganisms, whereas greater cellulose availability would stimulate the activity of cellulolytic guilds (Algora Gallardo et al., 2021; Hui et al., 2020). This suggests that changes in vegetation and litter composition, whether driven by climate variations, wildfires, or land-use modifications, can alter the structure and function of soil microbial communities, affecting the efficiency of nutrient recycling. A shift in the composition of available substrates could lead to a reconfiguration of decomposer guilds, with potential cascading effects on ecosystem stability and soil resilience to environmental disturbances (Philippot et al., 2021). Additionally, the presence of microorganisms specialized in poorly degradable biopolymers, such as certain Basidiomycota in lignin, may have a direct impact on the formation and stability of soil organic matter in the long term (Manici et al., 2024). The differential activity of decomposer guilds could influence carbon accumulation or loss in soils, modulating their role as either carbon sinks or sources of CO_2 in the context of climate change (Santorufo et al., 2024).

4.2. Decomposer guilds show different functional diversity of CAZymes for each polymer

The results demonstrate a pronounced functional specialization among decomposer guilds based on the biopolymers they degrade, reflected in the diversity and differential expression of CAZyme families in fungi and bacteria. This specialization not only suggests the existence of well-defined functional niches within soil microbial communities but also indicates an optimization of metabolic strategies in response to substrate availability. In fungi, CAZyme activity varies considerably depending on the substrate. The high expression of cellulases and beta-glucanases in cellulose and beta-1,3-glucan, respectively, confirms their essential role in the decomposition of these plant-derived structural polysaccharides (Pradeep & Edison, 2022; Selvaraj et al., 2024). This pattern suggests a strong dependence of fungal metabolism on substrate chemistry, where efficient cellulose degradation is mediated by the coordinated expression of cellobiohydrolases, endoglucanases, and beta-glucosidases (Zang et al., 2018). In contrast, the relatively low transcriptional expression of chitinases, despite their presence in the metagenome, implies that fungi play a secondary role in chitin degradation under the studied conditions. This finding is notable because, although chitin is a key structural component of fungal cell walls and arthropod exoskeletons, its degradation appears to be more strongly driven by bacteria, possibly reflecting an evolutionary resource-partitioning strategy among decomposer groups (Wieczorek et al., 2019). Enzymatic specialization in fungi also aligns with well-defined taxonomic patterns. Basidiomycota saprotrophs dominate cellulose decomposition, whereas Ascomycota shows a greater affinity for more labile polymers such as chitin and pectin, as previously highlighted. This differentiation may be linked to the ecological traits of each phylum. Basidiomycota, often associated with wood and lignocellulosic material degradation, have evolved highly efficient enzymatic machinery for breaking down cellulose and lignin in later decomposition stages. In contrast, Ascomycota, with a greater capacity to degrade accessible polymers, play a key role in the early

phases of biomass recycling, allowing them to rapidly colonize decomposing plant matter (Brazkova et al., 2022; Manici et al., 2024). Bacterial enzymatic specialization follows a distinct dynamic compared to fungi. The overexpression of chitinases and peptidoglycanases in chitin degradation highlights the key role of bacterial groups such as Bacteroidetes (*Mucilaginibacter*, *Pedobacter*) and Actinobacteria (*Streptomyces*), which specialize in decomposing fungal-derived polymers. The high expression of these enzymes in the metatranscriptome, despite their lower representation in the metagenome, suggests that transcriptional regulation of chitin degradation is highly substrate-dependent (Middelboe et al., 2025).

Another notable aspect is the coexistence of specialists and generalists within bacterial communities. While some taxa, such as *Chitinophaga* (Bacteroidetes phylum), exhibit marked specialization for chitin, others, like Alphaproteobacteria (*Rhizobium*), display greater functional versatility, modulating their enzymatic profiles based on substrate availability. This suggests contrasting ecological strategies among decomposer bacteria: specialists possess highly efficient enzymatic systems for specific substrates, whereas generalists adjust their metabolic machinery to exploit diverse biopolymers depending on environmental conditions (Algora et al., 2022). The coexistence of these strategies may confer ecological advantages in environments with heterogeneous carbon sources, ensuring microbial community activity across varying resource landscapes. Additionally, the ability of certain taxa to dynamically respond to substrate shifts could enhance the resilience of soil microbial ecosystems to environmental disturbances (Philippot et al., 2021). NMDS analysis supports these observations, revealing clear segregation of microbial communities based on substrate type in both metagenomic and metatranscriptomic data. The stronger separation in the metatranscriptome indicates that, although genetic potential for polymer degradation is widespread among soil microorganisms, the activation of these metabolic pathways is finely tuned by substrate composition (C.-C. Chen et al., 2020). Notably, CAZyme expression clusters cellulose and beta-1,3-glucan separately from chitin, reinforcing the concept of functionally distinct decomposer guilds for biopolymers of different origin. The limited overlap in expression profiles suggests reduced direct competition, likely due to resource partitioning and functional complementarity among microbial groups.

The observed enzymatic specialization has broader implications for carbon cycling in ecosystems. The functional complementarity between fungi and bacteria indicates that organic matter decomposition is a highly structured process, with different organisms playing specialized roles depending on substrate chemistry and origin (Condron et al., 2010; Khatoon et al., 2017). Moreover, the coexistence of specialist and generalist bacteria may enhance the stability and resilience of decomposition systems. In environments with dynamic and heterogeneous biopolymer composition, metabolically flexible organisms ensure continuous carbon recycling, even under fluctuating resource availability (Schniete et al., 2024). Understanding these interactions is crucial for predicting ecosystem responses to environmental changes and for elucidating how microbial biodiversity contributes to long-term biogeochemical cycling.

4.3. Guilds are composed of specialist for components of plant biomass and fungal biomass

Our findings demonstrate that microbial guilds are structured by specialization in the degradation of specific components of plant and fungal biomass, aligning with the "division of labor" hypothesis. This functional segmentation not only optimizes decomposition efficiency but also reduces competition among taxa and fosters synergistic interactions, contributing to the stability and resilience of the soil ecosystem (Nizamani et al., 2024; Z. Zhang et al., 2021). The observed substrate specialization in bacteria and fungi suggests that niche partitioning is a key mechanism in regulating the carbon cycle. *Streptomyces* and *Chitinophaga*, by focusing on chitin degradation, facilitate the mobilization of carbon derived from fungal biomass (McKee et al., 2019), whereas *Rhizobium*, acting as a generalist, maintains functional flexibility within the bacterial guild (Taylor et al., 2020). This balance between specialists and generalists allows microbial guilds to adapt to changes in resource availability, promoting community stability (Y.-J. Chen et al., 2021a). Similarly, the specialization of *Terriglobus* in beta-1,3-glucans suggests that certain taxa have evolved to exploit specific metabolic niches, minimizing competition with other polysaccharide degraders. In fungi, the functional segregation between *Aspergillus* and *Tulasnella* supports the hypothesis that different taxa have developed distinct enzymatic strategies to maximize cellulose and chitin decomposition (D. Li et al., 2023). The co-occurrence of species with complementary metabolic capabilities on mixed substrates (e.g., *Tulasnella* and *Mortierella* in cellulose and chitin degradation) suggests that cooperation within guilds may be a key factor in decomposition efficiency (Albornoz et al., 2022).

Metatranscriptomic analysis indicates that enzymatic specialization is crucial for guild efficiency. The dominance of *Chitinophaga* and *Pedobacter* in chitinase expression and the central role of *Terriglobus* and *Asticcacaulis* in beta-glucan and cellulose degradation demonstrate that degradative activity is not evenly distributed among taxa but is instead concentrated in highly efficient specialists (Y.-J. Chen et al., 2021b). This finding has direct implications for modeling decomposition dynamics, as it allows predictions of which taxa will be most active under different environmental conditions and substrate availability. The evidence provided by metagenome-assembled genomes (MAGs) reinforces the existence of guilds structured by functional specialization.

The presence of generalist bins, such as those of *Rhizobium*, confirms the importance of functional redundancy in maintaining ecosystem stability (Eisenhauer et al., 2023). In contrast, the presence of highly specialized bins, such as those of *Asticcacaulis* and *Chitinophaga*, suggests that certain taxa have evolved to exploit specific resources with high efficiency. Furthermore, the detection of novel bins with low taxonomic similarity underscores the still-unexplored diversity of microorganisms involved in biopolymer degradation

4.4. Digging into MAGs offers a clearer view of the role of specific bacterial taxa in biopolymer degradation

Our MAG-based analysis provided a high-resolution perspective on the functional roles of bacterial taxa—both known and unknown—in biopolymer decomposition. The recovery of 209 high-quality MAGs (completeness >70%, contamination <10%) spanning nine bacterial phyla highlights the vast bacterial diversity contributing to carbon cycling, as previously found (López-Mondéjar et al., 2022). One of the most striking findings was the clear functional dichotomy observed among MAGs, with generalists capable of degrading multiple substrates and specialists adapted to specific polymers. For instance, bins affiliated with *Rhizobium* (Alphaproteobacteria) and *Flavobacterium* (Bacteroidetes) displayed broad CAZyme repertoires, enabling activity across diverse biopolymers. Conversely, specialists such as *Asticcacaulis* (Alphaproteobacteria) and *Terriglobus* (Acidobacteria) exhibited high expression of enzymes targeting specific polymers like cellulose and beta-1,3-glucans, while *Chitinophaga* and *Pedobacter* (Bacteroidetes) and Cellvibrio (Gammaproteobacteria) were almost exclusively associated with chitin degradation. This specialization likely reflects niche adaptation strategies that enhance metabolic efficiency within microbial communities (Malard & Guisan, 2023; Pacciani-Mori et al., 2020). The observed partitioning of MAGs into generalists and specialists aligns with the "division of labor" hypothesis, where functional differentiation optimizes decomposition efficiency. Generalists, with their metabolic flexibility, contribute to functional redundancy and ecosystem stability (Y.-J. Chen et al., 2021b). In contrast, specialists drive the breakdown of recalcitrant polymers, playing essential roles in later stages of decomposition (Blair et al., 2021).

A particularly intriguing aspect of our findings is the presence of MAGs with low taxonomic similarity to reference genomes, suggesting the existence of novel bacterial lineages with putative roles in decomposition in forest soil. For example, bin.21 *JACMQM01* (Bacteroidota) was strongly linked to chitin degradation, potentially representing a new genus with specialized chitinolytic machinery. Additionally, bin.84 (*UBA1573*), bin.114 (*QFOX01*), and bin.177 (*JAJPHO01*) (Alphaproteobacteria) lacked close genomic relatives yet encoded enzymes indicative of roles in microbial or plant biomass breakdown. In addition, we found a MAG affiliated with Patescibacteria (class Saccharimonadia), suggesting novel enzymatic pathways involved in biopolymer decomposition. Given the small size of their genomes and their possible symbiotic lifestyles, Patescibacteria may play an indirect role in decomposition through interactions with other microbial taxa (H. Hu et al., 2024). The metabolic pathways encoded by these enigmatic taxa may represent evolutionary adaptations to very specific ecological niches, further highlighting the uncultivated microbial diversity of soils.

Future research should focus on isolating or employing single-cell genomics to further characterize these novel taxa, particularly those within Patescibacteria and other poorly studied phyla.

5. CONCLUSIONS

Our study provides compelling evidence that soil microbial communities are structured into specialized functional guilds based on distinct substrate preferences, fundamentally shaping organic matter decomposition in forest ecosystems. Through a MAG-based approach, complemented by metagenomic and metatranscriptomic analyses, we demonstrate how this guild-based organization enhances decomposition efficiency while maintaining ecosystem resilience through niche partitioning and functional complementarity.

The findings reveal clear patterns of substrate specialization among microbial taxa, with bacteria such as *Chitinophaga* and *Pedobacter* specializing in chitin degradation, *Terriglobus* in beta-1,3-glucans, and fungi such as *Tulasnella* and *Penicillium* exhibiting distinct preferences for cellulose and chitin/pectin, respectively. Moreover, our analysis highlights the coexistence of generalist and specialist strategies within these guilds, with taxa such as *Rhizobium* maintaining broad metabolic flexibility, whereas others have evolved highly efficient, specialized degradation pathways. Importantly, we uncover novel microbial diversity, including previously unrecognized taxa such as bin.21 JACMQM01 and rare phyla such as Patescibacteria, which contribute to decomposition processes, expanding our understanding of microbial players involved in carbon cycling. These findings have significant implications, as they establish a framework for classifying microbial decomposers into substrate-specific guilds based on both genomic potential and expressed activity, moving beyond phylogenetic classifications to functional ecological traits. By linking specific taxa to their functional roles in biopolymer degradation, our study provides a valuable resource for interpreting environmental surveys and inferring decomposition processes from microbial community composition data.

Furthermore, the identification of key specialist taxa and their associated CAZyme profiles offers potential biomarkers for monitoring specific decomposition pathways in response to environmental change. This guild-based classification system enhances the accuracy of decomposition models under varying substrate availability scenarios, including those driven by climate change, land use shifts, or vegetation dynamics. Moreover, the discovery of novel taxa and rare phyla involved in decomposition highlights critical gaps in our current understanding of soil microbial diversity and function, pointing to key targets for future cultivation efforts and bioprospecting. Ultimately, by bridging the gap between microbial taxonomy and ecosystem function, this study provides both conceptual and practical tools for advancing soil ecology. The functional classification scheme developed here facilitates more mechanistic interpretations of microbial community dynamics, supporting efforts to manage soil ecosystems for carbon sequestration, nutrient cycling, and climate change mitigation. Future research should build upon this foundation by exploring interactions among these guilds across diverse ecosystems and environmental gradients, as well as investigating how guild structure modulates ecosystem responses to global change.

GENERAL CONCLUSIONS

GENERAL CONCLUSIONS

This Thesis provides a deep and multifaceted understanding of microbial ecology in agricultural and natural soils, focusing on the biogeochemical cycles of phosphorus (P), nitrogen (N), and carbon (C). Through the integration of multi-omics approaches (metagenomics, metaproteomics, metatranscriptomics, and metagenome-assembled genomes (MAGs)), we have inferred the complexity of the soil microbiome and their role in ecosystem processes. These investigations have confirmed our initial hypotheses and revealed new insights into the functional and taxonomic dynamics of microorganisms in response to environmental factors, different fertilizers, plant phenology, and the decomposition of biopolymers from diverse origin in soil.

Functional niches in the phosphorus, nitrogen, and carbon cycles

One of the most significant findings of this thesis is the identification of clearly defined functional niches within microbial communities associated with the phosphorus, nitrogen, and carbon cycles. Our results demonstrate that microbial guilds within these biogeochemical cycles exhibit a high degree of functional specialization, which is strongly influenced by the phenological stage of the crop. In the phosphorus cycle, we observed a distinct taxonomic separation between microorganisms involved in the solubilization of inorganic phosphorus and those responsible for the mineralization of organic phosphorus. This distinction was particularly evident in Actinobacteria, whose members harbor genes related to inorganic phosphorus solubilization but not organic phosphorus mineralization. This pattern suggests an evolutionary adaptation to specific phosphorus pools and highlights the need to consider functional diversity when designing fertilization strategies. Our results also emphasize the underappreciated role of archaea in phosphorus cycling, revealing that archaeal taxa harbor genes involved in phosphorus metabolism and may play critical roles alongside bacteria in regulating phosphorus availability in agroecosystems. Moreover, the integration of metaproteomics allowed us to identify key phosphorus-cycling enzymes such as alkaline phosphatase, encoded by *phoX*, which is abundant in maize agroecosystems and may serve as a crucial biomarker for phosphorus availability.

Similarly, our findings regarding the nitrogen cycle underscore the importance of functional specialization within microbial communities. We observed that microorganisms involved in nitrification, such as Nitrososphaeraceae, typically lack genes related to N_2 fixation or nitrogen transport, reinforcing the idea of distinct ecological roles within nitrogen-cycling guilds. Metagenomic and metaproteomic analyses further revealed that denitrification and nitrification processes are taxonomically clustered, with microbial guilds displaying strong associations with specific functional pathways. Importantly, we found that the phenological stage of the crop is a stronger driver of nitrogen-cycling gene abundance than fertilization treatments. This suggests that nitrogen transformations in agroecosystems are primarily modulated by plant development rather than external nutrient inputs, a finding that could inform more efficient fertilization strategies. Additionally, we identified significant taxonomic and functional insights through the

reconstruction of microbial genomes (MAGs), confirming the roles of Nitrososphaeraceae in nitrification and Propionibacteriaceae in denitrification. The application of metaproteomics further refined these insights by identifying key nitrogen-associated enzymes such as glutamine synthetase (GlnA), which plays a central role in nitrogen assimilation and has been overlooked in previous studies. Our results also demonstrate that different fertilization strategies influence nitrogen-cycling genes, particularly those involved in DNRA and nitrification. We observed distinct responses among fertilizers, with mineral fertilizers such as NPK and struvite enhancing DNRA and nitrification, while organic amendments promoted microbial diversity but required careful management to optimize nitrogen release.

In the carbon cycle, our results provide compelling evidence for the existence of specialized bacterial decomposer guilds, challenging the long-held assumption that bacteria play a minor role in the decomposition process. While fungi, particularly Basidiomycota, have traditionally been considered the primary decomposers of complex organic matter, our findings reveal that bacterial taxa also exhibit functional specialization in the decomposition process. Specifically, we identified distinct bacterial guilds associated with the degradation of key biopolymers, such as *Chitinophaga* and *Pedobacter* specializing in chitin degradation, *Terriglobus* in β-1,3-glucans and *Asticcacaulis* in cellulose. These results suggest that bacteria contribute significantly to the breakdown of complex carbon substrates, complementing fungal activity and expanding our understanding of microbial interactions in decomposition dynamics. Moreover, our metagenomic and metatranscriptomic analyses reveal that while Proteobacteria dominate in terms of gene abundance, the transcriptional activity of bacterial decomposers suggests a more active role than previously recognized. This highlights the need for a functional perspective when assessing microbial contributions to carbon cycling, as taxonomic dominance does not necessarily translate into ecological relevance.

Impact of plant phenology and fertilization practices

A key aspect of our studies is the demonstration that crop phenology has a greater impact than fertilization practices on the relative abundance of genes associated with the phosphorus and nitrogen cycles. This finding is particularly relevant, as it suggests that nutrient management strategies must consider not only the type of fertilizer used but also the growth stage of the crop. For example, in the phosphorus cycle, phenology influences the expression of genes related to solubilization and mineralization, which could affect phosphorus availability for plants at different growth stages. Similarly, we observed that the abundance of genes related to denitrification and nitrification varies significantly depending on the phenological stage, which could influence nitrogen use efficiency and losses through gas emissions or leaching. Nevertheless, we also concluded that fertilization practices have a notable impact on microbial functionality. For instance, mineral fertilizers, such as NPK and struvite, promote processes like nitrification and DNRA (dissimilatory nitrate reduction to ammonium), while organic fertilizers favor greater microbial diversity but require careful management to optimize nutrient release. These results highlight the need to develop fertilization strategies that balance crop nutrient demands with soil microbial dynamics.

Meta-Omics methodologies and their contribution to the study of microbial ecology

The application of multi-omics approaches has been fundamental to advancing our knowledge and understanding of microbial ecology in agricultural and environmental soils. Metagenomics allowed us to identify the taxonomic and functional diversity of microbial communities, while metaproteomics and metatranscriptomics provided valuable insights into enzymatic activity and gene expression. For example, the integration of metagenomics and metaproteomics revealed that enzymes such as alkaline phosphatase (encoded by *phoX*) and glutamine synthetase (encoded by *glnA*) play crucial roles in the phosphorus and nitrogen cycles, respectively, despite being overlooked in previous studies.

Additionally, the reconstruction of microbial genomes (MAGs) enabled us to explore in greater detail the phylogenetic and functional diversity of microbial communities. We identified MAGs specialized in the degradation of specific biopolymers from plant and microbial origin, such as *Asticcacaulis* and *Pararobbsia* for cellulose, and *Pedobacter* for chitin, highlighting the importance of functional specialization in the efficiency of organic matter decomposition. We also uncovered new MAGs involved in the nitrogen cycle, as well as previously undescribed taxa, expanding our understanding of key players in these processes. These integrated approaches have not only enhanced our ability to identify key taxa and genes but have also provided insight into how their activity is influenced by environmental factors, such as nutrient availability and management practices.

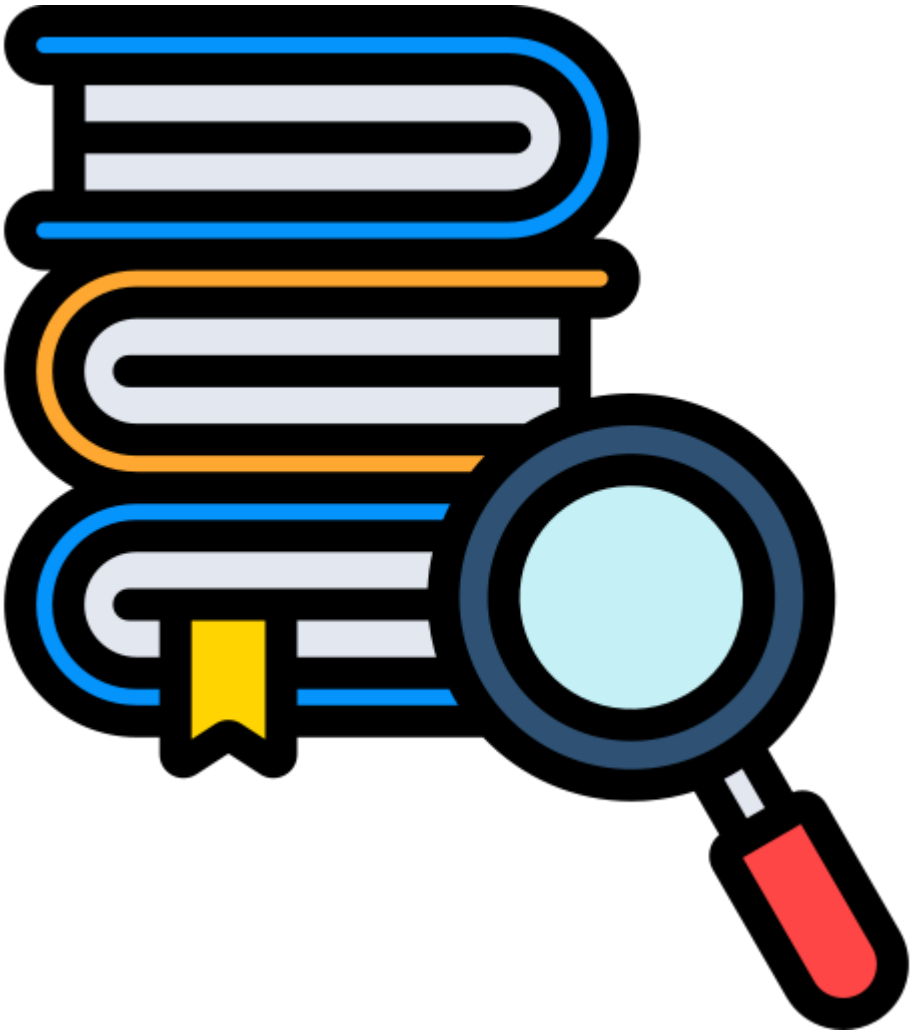
Future perspectives

This doctoral thesis has demonstrated that microbial ecology is a key component for understanding and managing biogeochemical cycles in agricultural and natural soils. The methodological advances and knowledge generated in this work contribute to basic science and have practical implications for the development of more sustainable and resilient agricultural practices. Thus, the results of this thesis open new avenues for research in the field of microbial ecology and sustainable soil management. First, it is necessary to deepen the study of archaea in agricultural soils and their contribution to biogeochemical cycles, as their role has been underestimated compared to bacteria. Additionally, long-term studies are required to evaluate how management practices, such those utilized in this Thesis, influence the stability and resilience of microbial communities as well as their interaction with climate change factors.

Another promising area is the targeted manipulation of microbial communities to optimize nutrient cycling in agroecosystems. For example, the identification of taxa specialized in biopolymer degradation or phosphorus solubilization could be used to select specific microorganisms and develop biofertilizers that improve nutrient use efficiency and reduce environmental losses, being this a timely topic for biotechnological companies nowadays. Moreover, the relevance of crop phenology in controlling soil microbial communities reflect the need of considering this factor when designing and applying biofertilizers.

It is likely that isolation of microbes with a particular capacity in nitrogen or phosphorus cycles (i.e., release of phosphorus for plants) would better work using samples from the most prone phenological stage. Similarly, the application of microbes (i.e., biofertilizers) can be done in the most appropriate phenological stage where they are needed by the plant and conditions would maximize their survival.

The integration of multi-omics approaches with machine learning techniques and predictive models could revolutionize our ability to predict and manage microbial functionality in response to environmental and management changes. These tools could help design more precise and sustainable management strategies, contributing to food security and climate change mitigation.



BIBLIOGRAPHY

BIBLIOGRAPHY

Abatenh, E., Gizaw, B., Tsegaye, Z., Tefera, G., Abatenh, E., Gizaw, B., Tsegaye, Z., & Tefera, G. (2018). Microbial Function on Climate Change – A Review. *Open Journal of Environmental Biology*, *3*(1), 001-007. https://doi.org/10.17352/ojeb.000008

Abellan-Schneyder, I., Matchado, M. S., Reitmeier, S., Sommer, A., Sewald, Z., Baumbach, J., List, M., & Neuhaus, K. (2021). Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing. *mSphere*, *6*(1), 10.1128/msphere.01202-20. https://doi.org/10.1128/msphere.01202-20

Achbergerová, L., & Nahálka, J. (2011). Polyphosphate—An ancient energy source and active metabolic regulator. *Microbial Cell Factories*, *10*(1), 63. https://doi.org/10.1186/1475-2859-10-63

Al-Ajeel, S., Spasov, E., Sauder, L. A., McKnight, M. M., & Neufeld, J. D. (2022). Ammonia-oxidizing archaea and complete ammonia-oxidizing *Nitrospira* in water treatment systems. *Water Research X*, *15*, 100131. https://doi.org/10.1016/j.wroa.2022.100131

Albornoz, F. E., Prober, S. M., Ryan, M. H., & Standish, R. J. (2022). Ecological interactions among microbial functional guilds in the plant-soil system and implications for ecosystem function. *Plant and Soil*, *476*(1), 301-313. https://doi.org/10.1007/s11104-022-05479-1

Algora, C., Odriozola, iñaki, Human, Z., Awokunle Holla, S., Baldrian, P., & López-Mondéjar, R. (2022). Specific utilization of biopolymers of plant and fungal origin reveals the existence of substrate-specific guilds for bacteria in temperate forest soils. *Soil Biology and Biochemistry*, *171*, 108696. https://doi.org/10.1016/j.soilbio.2022.108696

Algora Gallardo, C., Baldrian, P., & López-Mondéjar, R. (2021). Litter-inhabiting fungi show high level of specialization towards biopolymers composing plant and fungal biomass. *Biology and Fertility of Soils*, *57*(1), 77-88. https://doi.org/10.1007/s00374-020-01507-3

Aller, J. Y., & Kemp, P. F. (2008). Are Archaea inherently less diverse than Bacteria in the same environments? *FEMS Microbiology Ecology*, *65*(1), 74-87. https://doi.org/10.1111/j.1574-6941.2008.00498.x

Alneberg, J., Bjarnason, B. S., Bruijn, I. de, Schirmer, M., Quick, J., Ijaz, U. Z., Loman, N. J., Andersson, A. F., & Quince, C. (2013). *CONCOCT: Clustering cONtigs on COverage and ComposiTion* (arXiv:1312.4038). arXiv. https://doi.org/10.48550/arXiv.1312.4038

Andrade, A. C., Fróes, A., Lopes, F. Á. C., Thompson, F. L., Krüger, R. H., Dinsdale, E., & Bruce, T. (2017). Diversity of Microbial Carbohydrate-Active enZYmes (CAZYmes) Associated with Freshwater and Soil Samples from Caatinga Biome. *Microbial Ecology*, *74*(1), 89-105. https://doi.org/10.1007/s00248-016-0911-9

Andrino, A., Guggenberger, G., Kernchen, S., Mikutta, R., Sauheitl, L., & Boy, J. (2021). Production of Organic Acids by Arbuscular Mycorrhizal Fungi and Their Contribution in the Mobilization of Phosphorus Bound to Iron Oxides. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.661842

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., & Ogata, H. (2020). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, *36*(7), 2251-2252. https://doi.org/10.1093/bioinformatics/btz859

Armendáriz-Ruiz, M., Rodríguez-González, J. A., Camacho-Ruíz, R. M., & Mateos-Díaz, J. C. (2018). Carbohydrate Esterases: An Overview. En G. Sandoval (Ed.), *Lipases and Phospholipases: Methods and Protocols* (pp. 39-68). Springer. https://doi.org/10.1007/978-1-4939-8672-9_2

Ayangbenro, A. S., & Babalola, O. O. (2021). Reclamation of arid and semi-arid soils: The role of plant growth-promoting archaea and bacteria. *Current Plant Biology*, *25*, 100173. https://doi.org/10.1016/j.cpb.2020.100173

Bahram, M., Netherway, T., Frioux, C., Ferretti, P., Coelho, L. P., Geisen, S., Bork, P., & Hildebrand, F. (2021). Metagenomic assessment of the global diversity and distribution of bacteria and fungi. *Environmental Microbiology*, *23*(1), 316-326. https://doi.org/10.1111/1462-2920.15314

Baldrian, P. (2017). Forest microbiome: Diversity, complexity and dynamics. *FEMS Microbiology Reviews*, *41*(2), 109-130. https://doi.org/10.1093/femsre/fuw040

Baldrian, P., López-Mondéjar, R., & Kohout, P. (2023). Forest microbiome and global change. *Nature Reviews Microbiology*, *21*(8), 487-501. https://doi.org/10.1038/s41579-023-00876-4

Baldrian, P., Merhautová, V., Cajthaml, T., Petránková, M., & Šnajdr, J. (2010). Small-scale distribution

of extracellular enzymes, fungal, and bacterial biomass in Quercus petraea forest topsoil. *Biology and Fertility of Soils*, *46*(7), 717-726. https://doi.org/10.1007/s00374-010-0478-4
Bani, A., Pioli, S., Ventura, M., Panzacchi, P., Borruso, L., Tognetti, R., Tonon, G., & Brusetti, L. (2018). The role of microbial community in the decomposition of leaf litter and deadwood. *Applied Soil Ecology*, *126*, 75-84. https://doi.org/10.1016/j.apsoil.2018.02.017
Bargaz, A., Lyamlouli, K., Chtouki, M., Zeroual, Y., & Dhiba, D. (2018). Soil Microbial Resources for Improving Fertilizers Efficiency in an Integrated Plant Nutrient Management System. *Frontiers in Microbiology*, *9*. https://doi.org/10.3389/fmicb.2018.01606
Barquero, M. B., García-Díaz, C., Dobbler, P. T., Jehmlich, N., Moreno, J. L., López-Mondéjar, R., & Bastida, F. (2024a). Contrasting fertilization and phenological stages shape microbial-mediated phosphorus cycling in a maize agroecosystem. *Science of The Total Environment*, *951*, 175571. https://doi.org/10.1016/j.scitotenv.2024.175571
Barquero, M. B., García-Díaz, C., Dobbler, P. T., Jehmlich, N., Moreno, J. L., López-Mondéjar, R., & Bastida, F. (2024b). Contrasting fertilization and phenological stages shape microbial-mediated phosphorus cycling in a maize agroecosystem. *Science of The Total Environment*, *951*, 175571. https://doi.org/10.1016/j.scitotenv.2024.175571
Barroso, C. B., & Nahas, E. (2005). The status of soil phosphate fractions and the ability of fungi to dissolve hardly soluble phosphates. *Applied Soil Ecology*, *29*(1), 73-83. https://doi.org/10.1016/j.apsoil.2004.09.005
Barry, D. a. J., & Miller, M. H. (1989). Phosphorus Nutritional Requirement of Maize Seedlings for Maximum Yield. *Agronomy Journal*, *81*(1), 95-99. https://doi.org/10.2134/agronj1989.00021962008100010017x
Bastida, F., Hernández, T., & García, C. (2014a). Metaproteomics of soils from semiarid environment: Functional and phylogenetic information obtained with different protein extraction methods. *Journal of Proteomics*, *101*, 31-42. https://doi.org/10.1016/j.jprot.2014.02.006
Bastida, F., Hernández, T., & García, C. (2014b). Metaproteomics of soils from semiarid environment: Functional and phylogenetic information obtained with different protein extraction methods. *Journal of Proteomics*, *101*, 31-42. https://doi.org/10.1016/j.jprot.2014.02.006
Bastida, F., & Jehmlich, N. (2016). It's all about functionality: How can metaproteomics help us to discuss the attributes of ecological relevance in soil? *Journal of Proteomics*, *144*, 159-161. https://doi.org/10.1016/j.jprot.2016.06.002
Bastida, F., Jehmlich, N., Lima, K., Morris, B. E. L., Richnow, H. H., Hernández, T., von Bergen, M., & García, C. (2016). The ecological and physiological responses of the microbial community from a semiarid soil to hydrocarbon contamination and its bioremediation using compost amendment. *Journal of Proteomics*, *135*, 162-169. https://doi.org/10.1016/j.jprot.2015.07.023
Bastida, F., Jehmlich, N., Martínez-Navarro, J., Bayona, V., García, C., & Moreno, J. L. (2019a). The effects of struvite and sewage sludge on plant yield and the microbial community of a semiarid Mediterranean soil. *Geoderma*, *337*, 1051-1057. https://doi.org/10.1016/j.geoderma.2018.10.046
Bastida, F., Jehmlich, N., Martínez-Navarro, J., Bayona, V., García, C., & Moreno, J. L. (2019b). The effects of struvite and sewage sludge on plant yield and the microbial community of a semiarid Mediterranean soil. *Geoderma*, *337*, 1051-1057. https://doi.org/10.1016/j.geoderma.2018.10.046
Bastida, F., Jehmlich, N., Starke, R., Schallert, K., Benndorf, D., López-Mondéjar, R., Plaza, C., Freixino, Z., Ramírez-Ortuño, C., Ruiz-Navarro, A., Díaz-López, M., Vera, A., Moreno, J. L., Eldridge, D. J., García, C., & Delgado-Baquerizo, M. (2021a). Structure and function of bacterial metaproteomes across biomes. *Soil Biology and Biochemistry*, *160*, 108331. https://doi.org/10.1016/j.soilbio.2021.108331
Bastida, F., Jehmlich, N., Starke, R., Schallert, K., Benndorf, D., López-Mondéjar, R., Plaza, C., Freixino, Z., Ramírez-Ortuño, C., Ruiz-Navarro, A., Díaz-López, M., Vera, A., Moreno, J. L., Eldridge, D. J., García, C., & Delgado-Baquerizo, M. (2021b). Structure and function of bacterial metaproteomes across biomes. *Soil Biology and Biochemistry*, *160*, 108331. https://doi.org/10.1016/j.soilbio.2021.108331
Bastida, F., Kandeler, E., Moreno, J. L., Ros, M., García, C., & Hernández, T. (2008). Application of fresh and composted organic wastes modifies structure, size and activity of soil microbial community under semiarid climate. *Applied Soil Ecology*, *40*(2), 318-329. https://doi.org/10.1016/j.apsoil.2008.05.007
Bastida, F., Pérez-de-Mora, A., Babic, K., Hai, B., Hernández, T., García, C., & Schloter, M. (2009a). Role of amendments on N cycling in Mediterranean abandoned semiarid soils. *Applied Soil Ecology*, *41*(2), 195-

205. https://doi.org/10.1016/j.apsoil.2008.10.009
Bastida, F., Pérez-de-Mora, A., Babic, K., Hai, B., Hernández, T., García, C., & Schloter, M. (2009b). Role of amendments on N cycling in Mediterranean abandoned semiarid soils. *Applied Soil Ecology*, *41*(2), 195-205. https://doi.org/10.1016/j.apsoil.2008.10.009
Bastida, F., Siles, J. A., García, C., García-Díaz, C., & Moreno, J. L. (2023). Shifting the paradigm for phosphorus fertilization in the advent of the fertilizer crisis. *Journal of Sustainable Agriculture and Environment*, *2*(2), 153-156. https://doi.org/10.1002/sae2.12040
Baveye, P., Schnee, L., Boivin, P., Laba, M., & Radulovich, R. (2020). Soil Organic Matter Research and Climate Change: Merely Re-storing Carbon Versus Restoring Soil Functions. *Frontiers in Environmental Science*, *8*, 579904. https://doi.org/10.3389/fenvs.2020.579904
Beidler, K. V., Phillips, R. P., Andrews, E., Maillard, F., Mushinski, R. M., & Kennedy, P. G. (2020). Substrate quality drives fungal necromass decay and decomposer community structure under contrasting vegetation types. *Journal of Ecology*, *108*(5), 1845-1859. https://doi.org/10.1111/1365-2745.13385
Berlemont, R., & Martiny, A. C. (2013). Phylogenetic distribution of potential cellulases in bacteria. *Applied and Environmental Microbiology*, *79*(5), 1545-1554. https://doi.org/10.1128/AEM.03305-12
Bhatnagar, J. M., Peay, K. G., & Treseder, K. K. (2018). Litter chemistry influences decomposition through activity of specific microbial functional guilds. *Ecological Monographs*, *88*(3), 429-444. https://doi.org/10.1002/ecm.1303
Bhowmik, A., Cloutier, M., Ball, E., & Bruns, M. A. (2017). Underexplored microbial metabolisms for enhanced nutrient recycling in agricultural soils. *AIMS Microbiology*, *3*(4), 826-845. https://doi.org/10.3934/microbiol.2017.4.826
Blair, E. M., Dickson, K. L., & O'Malley, M. A. (2021). Microbial communities and their enzymes facilitate degradation of recalcitrant polymers in anaerobic digestion. *Current Opinion in Microbiology*, *64*, 100-108. https://doi.org/10.1016/j.mib.2021.09.008
Blakeley-Ruiz, J. A., Erickson, A. R., Cantarel, B. L., Xiong, W., Adams, R., Jansson, J. K., Fraser, C. M., & Hettich, R. L. (2019). Metaproteomics reveals persistent and phylum-redundant metabolic functional stability in adult human gut microbiomes of Crohn's remission patients despite temporal variations in microbial taxa, genomes, and proteomes. *Microbiome*, *7*(1), 18. https://doi.org/10.1186/s40168-019-0631-8
Bloom, A. J. (2015). The increasing importance of distinguishing among plant nitrogen sources. *Current Opinion in Plant Biology*, *25*, 10-16. https://doi.org/10.1016/j.pbi.2015.03.002
Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120. https://doi.org/10.1093/bioinformatics/btu170
Brabcová, V., Nováková, M., Davidová, A., & Baldrian, P. (2016). Dead fungal mycelium in forest soil represents a decomposition hotspot and a habitat for a specific microbial community. *New Phytologist*, *210*(4), 1369-1381. https://doi.org/10.1111/nph.13849
Brazkova, M., Koleva, R., Angelova, G. V., & Yemendzhiev, H. (2022). Ligninolytic enzymes in Basidiomycetes and their application in xenobiotics degradation. *BIO Web of Conferences*, *45*, 02009. https://doi.org/10.1051/bioconf/20224502009
Brochado, M. G. da S., Silva, L. B. X. da, Lima, A. da C., Guidi, Y. M., & Mendes, K. F. (2023). Herbicides versus Nitrogen Cycle: Assessing the Trade-Offs for Soil Integrity and Crop Yield—An In-Depth Systematic Review. *Nitrogen*, *4*(3), Article 3. https://doi.org/10.3390/nitrogen4030022
Bronick, C. J., & Lal, R. (2005). Soil structure and management: A review. *Geoderma*, *124*(1), 3-22. https://doi.org/10.1016/j.geoderma.2004.03.005
Brownlie, W. J., Sutton, M. A., Reay, D. S., Heal, K. V., Hermann, L., Kabbe, C., & Spears, B. M. (2021). Global actions for a sustainable phosphorus future. *Nature Food*, *2*(2), 71-74. https://doi.org/10.1038/s43016-021-00232-w
Carreras-Sempere, M., Guivernau, M., Caceres, R., Biel, C., Noguerol, J., & Viñas, M. (2024). Effect of Fertigation with Struvite and Ammonium Nitrate on Substrate Microbiota and N2O Emissions in a Tomato Crop on Soilless Culture System. *Agronomy*, *14*(1), Article 1. https://doi.org/10.3390/agronomy14010119
Castañeda-Monsalve, V., Fröhlich, L.-F., Haange, S.-B., Homsi, M. N., Rolle-Kampczyk, U., Fu, Q., von Bergen, M., & Jehmlich, N. (2024). High-throughput screening of the effects of 90 xenobiotics on the simplified human gut microbiota model (SIHUMIx): A metaproteomic and metabolomic study. *Frontiers*

in Microbiology, *15*. https://doi.org/10.3389/fmicb.2024.1349367
Chandini, Kumar, R., Kumar, R., & Prakash, O. (2019). *The Impact of Chemical Fertilizers on our Environment and Ecosystem* (pp. 69-86).
Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2022). GTDB-Tk v2: Memory friendly classification with the genome taxonomy database. *Bioinformatics*, *38*(23), 5315-5316. https://doi.org/10.1093/bioinformatics/btac672
Chen, C.-C., Dai, L., Ma, L., & Guo, R.-T. (2020). Enzymatic degradation of plant biomass and synthetic polymers. *Nature Reviews Chemistry*, *4*(3), 114-126. https://doi.org/10.1038/s41570-020-0163-6
Chen, L., Wang, J., He, L., Xu, X., Wang, J., Ren, C., Guo, Y., & Zhao, F. (2023). Metagenomic highlight contrasting elevational pattern of bacteria- and fungi-derived compound decompositions in forest soils. *Plant and Soil*, *490*(1), 617-629. https://doi.org/10.1007/s11104-023-06104-5
Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M., & Banfield, J. F. (2020a). Accurate and complete genomes from metagenomes. *Genome Research*, *30*(3), 315-333. https://doi.org/10.1101/gr.258640.119
Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M., & Banfield, J. F. (2020b). Accurate and complete genomes from metagenomes. *Genome Research*, *30*(3), 315-333. https://doi.org/10.1101/gr.258640.119
Chen, Y.-J., Leung, P. M., Wood, J. L., Bay, S. K., Hugenholtz, P., Kessler, A. J., Shelley, G., Waite, D. W., Franks, A. E., Cook, P. L. M., & Greening, C. (2021a). Metabolic flexibility allows bacterial habitat generalists to become dominant in a frequently disturbed ecosystem. *The ISME Journal*, *15*(10), 2986-3004. https://doi.org/10.1038/s41396-021-00988-w
Chen, Y.-J., Leung, P. M., Wood, J. L., Bay, S. K., Hugenholtz, P., Kessler, A. J., Shelley, G., Waite, D. W., Franks, A. E., Cook, P. L. M., & Greening, C. (2021b). Metabolic flexibility allows bacterial habitat generalists to become dominant in a frequently disturbed ecosystem. *The ISME Journal*, *15*(10), 2986-3004. https://doi.org/10.1038/s41396-021-00988-w
Chirania, P., Holwerda, E. K., Giannone, R. J., Liang, X., Poudel, S., Ellis, J. C., Bomble, Y. J., Hettich, R. L., & Lynd, L. R. (2022). Metaproteomics reveals enzymatic strategies deployed by anaerobic microbiomes to maintain lignocellulose deconstruction at high solids. *Nature Communications*, *13*(1), 3870. https://doi.org/10.1038/s41467-022-31433-x
Chklovski, A., Parks, D. H., Woodcroft, B. J., & Tyson, G. W. (2023). CheckM2: A rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nature Methods*, *20*(8), 1203-1212. https://doi.org/10.1038/s41592-023-01940-w
Chojnacka, K., Gorazda, K., Witek-Krowiak, A., & Moustakas, K. (2019). Recovery of fertilizer nutrients from materials—Contradictions, mistakes and future trends. *Renewable and Sustainable Energy Reviews*, *110*, 485-498. https://doi.org/10.1016/j.rser.2019.04.063
Chojnacka, K., Skrzypczak, D., Szopa, D., Izydorczyk, G., Moustakas, K., & Witek-Krowiak, A. (2023). Management of biological sewage sludge: Fertilizer nitrogen recovery as the solution to fertilizer crisis. *Journal of Environmental Management*, *326*, 116602. https://doi.org/10.1016/j.jenvman.2022.116602
Chourey, K., Jansson, J., VerBerkmoes, N., Shah, M., Chavarria, K. L., Tom, L. M., Brodie, E. L., & Hettich, R. L. (2010a). Direct Cellular Lysis/Protein Extraction Protocol for Soil Metaproteomics. *Journal of Proteome Research*, *9*(12), 6615-6622. https://doi.org/10.1021/pr100787q
Chourey, K., Jansson, J., VerBerkmoes, N., Shah, M., Chavarria, K. L., Tom, L. M., Brodie, E. L., & Hettich, R. L. (2010b). Direct cellular lysis/protein extraction protocol for soil metaproteomics. *Journal of Proteome Research*, *9*(12), 6615-6622. https://doi.org/10.1021/pr100787q
Chowdhury, R. B., Moore, G. A., Weatherley, A. J., & Arora, M. (2017). Key sustainability challenges for the global phosphorus resource, their implications for global food security, and options for mitigation. *Journal of Cleaner Production*, *140*, 945-963. https://doi.org/10.1016/j.jclepro.2016.07.012
Clark, I. M., Hughes, D. J., Fu, Q., Abadie, M., & Hirsch, P. R. (2021). Metagenomic approaches reveal differences in genetic diversity and relative abundance of nitrifying bacteria and archaea in contrasting soils. *Scientific Reports*, *11*(1), 15905. https://doi.org/10.1038/s41598-021-95100-9
Cocking, E. C. (2000). Helping plants get more nitrogen from the air. *European Review*, *8*(2), 193-200. https://doi.org/10.1017/S1062798700004762
Cole, J. J., Hararuk, O., & Solomon, C. T. (2021). Chapter 7 - The Carbon Cycle: With a Brief Introduction to Global Biogeochemistry. En K. C. Weathers, D. L. Strayer, & G. E. Likens (Eds.), *Fundamentals of Ecosystem Science (Second Edition)* (pp. 131-160). Academic Press. https://doi.org/10.1016/B978-0-12-

812762-9.00007-1
Condron, L., Stark, C., O'Callaghan, M., Clinton, P., & Huang, Z. (2010). The Role of Microbial Communities in the Formation and Decomposition of Soil Organic Matter. En G. R. Dixon & E. L. Tilston (Eds.), *Soil Microbiology and Sustainable Crop Production* (pp. 81-118). Springer Netherlands. https://doi.org/10.1007/978-90-481-9479-7_4
Cordell, D., Drangert, J.-O., & White, S. (2009). The story of phosphorus: Global food security and food for thought. *Global Environmental Change*, *19*(2), 292-305. https://doi.org/10.1016/j.gloenvcha.2008.10.009
Costa, O. Y. A., Raaijmakers, J. M., & Kuramae, E. E. (2018). Microbial Extracellular Polymeric Substances: Ecological Function and Impact on Soil Aggregation. *Frontiers in Microbiology*, *9*. https://doi.org/10.3389/fmicb.2018.01636
Creamer, R. E., Brennan, F., Fenton, O., Healy, M. G., Lalor, S. T. J., Lanigan, G. J., Regan, J. T., & Griffiths, B. S. (2010). Implications of the proposed Soil Framework Directive on agricultural systems in Atlantic Europe – a review. *Soil Use and Management*, *26*(3), 198-211. https://doi.org/10.1111/j.1475-2743.2010.00288.x
Dai, Z., Liu, G., Chen, H., Chen, C., Wang, J., Ai, S., Wei, D., Li, D., Ma, B., Tang, C., Brookes, P. C., & Xu, J. (2020). Long-term nutrient inputs shift soil microbial functional profiles of phosphorus cycling in diverse agroecosystems. *The ISME Journal*, *14*(3), 757-770. https://doi.org/10.1038/s41396-019-0567-9
De Deyn, G. B., Cornelissen, J. H. C., & Bardgett, R. D. (2008). Plant functional traits and soil carbon sequestration in contrasting biomes. *Ecology Letters*, *11*(5), 516-531. https://doi.org/10.1111/j.1461-0248.2008.01164.x
De Filippo, C., Ramazzotti, M., Fontana, P., & Cavalieri, D. (2012). Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. *Briefings in Bioinformatics*, *13*(6), 696-710. https://doi.org/10.1093/bib/bbs070
Delgado, A., & Gómez, J. A. (2024). The Soil: Physical, Chemical, and Biological Properties. En F. J. Villalobos & E. Fereres (Eds.), *Principles of Agronomy for Sustainable Agriculture* (pp. 15-30). Springer International Publishing. https://doi.org/10.1007/978-3-031-69150-8_2
Dinno, A. (2024). *dunn.test: Dunn's Test of Multiple Comparisons Using Rank Sums* (Versión 1.3.6) [Software]. https://cran.r-project.org/web/packages/dunn.test/index.html
Djemiel, C., Dequiedt, S., Karimi, B., Cottin, A., Horrigue, W., Bailly, A., Boutaleb, A., Sadet-Bourgeteau, S., Maron, P.-A., Chemidlin Prévost-Bouré, N., Ranjard, L., & Terrat, S. (2022). Potential of Meta-Omics to Provide Modern Microbial Indicators for Monitoring Soil Quality and Securing Food Production. *Frontiers in Microbiology*, *13*. https://doi.org/10.3389/fmicb.2022.889788
Dos Santos, P. C., Fang, Z., Mason, S. W., Setubal, J. C., & Dixon, R. (2012). Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. *BMC Genomics*, *13*(1), 162. https://doi.org/10.1186/1471-2164-13-162
Duchin, S., Nirit, B., Kamenetsky-Goldstein, R., & Spitzer-Rimon, B. (2020). New insights on flowering of Cannabis sativa. *Acta Horticulturae*, 17-20. https://doi.org/10.17660/ActaHortic.2020.1283.3
Efthimiou, N. (2025). Governance and degradation of soil in the EU. An overview of policies with a focus on soil erosion. *Soil and Tillage Research*, *245*, 106308. https://doi.org/10.1016/j.still.2024.106308
Eglin, T., Ciais, P., Piao, S. L., Barre, P., Bellassen, V., Cadule, P., Chenu, C., Gasser, T., Koven, C., Reichstein, M., & Smith, P. (2010). Historical and future perspectives of global soil carbon response to climate and land-use changes. *Tellus B: Chemical and Physical Meteorology*, *62*(5), 700-718. https://doi.org/10.1111/j.1600-0889.2010.00499.x
Eisenberg, D., Gill, H. S., Pfluegl, G. M. U., & Rotstein, S. H. (2000). Structure–function relationships of glutamine synthetases1. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, *1477*(1), 122-145. https://doi.org/10.1016/S0167-4838(99)00270-8
Eisenhauer, N., Hines, J., Maestre, F. T., & Rillig, M. C. (2023). Reconsidering functional redundancy in biodiversity research. *Npj Biodiversity*, *2*(1), 1-4. https://doi.org/10.1038/s44185-023-00015-5
Ekblad, A., Wallander, H., Godbold, D. L., Cruz, C., Johnson, D., Baldrian, P., Björk, R. G., Epron, D., Kieliszewska-Rokicka, B., Kjøller, R., Kraigher, H., Matzner, E., Neumann, J., & Plassard, C. (2013). The production and turnover of extramatrical mycelium of ectomycorrhizal fungi in forest soils: Role in carbon cycling. *Plant and Soil*, *366*(1), 1-27. https://doi.org/10.1007/s11104-013-1630-3

Erenstein, O., Jaleta, M., Sonder, K., Mottaleb, K., & Prasanna, B. M. (2022). Global maize production, consumption and trade: Trends and R&D implications. *Food Security*, *14*(5), 1295-1319. https://doi.org/10.1007/s12571-022-01288-7
Fageria, N. K. (2012). Role of Soil Organic Matter in Maintaining Sustainability of Cropping Systems. *Communications in Soil Science and Plant Analysis*, *43*(16), 2063-2113. https://doi.org/10.1080/00103624.2012.697234
Farrelly, D. J., Everard, C. D., Fagan, C. C., & McDonnell, K. P. (2013). Carbon sequestration and the role of biological carbon mitigation: A review. *Renewable and Sustainable Energy Reviews*, *21*, 712-727. https://doi.org/10.1016/j.rser.2012.12.038
Fenice, M. (2021). The Nitrogen Cycle: An Overview. En *Nitrogen Cycle*. CRC Press.
Ferreira, V., Elosegi, A., D. Tiegs, S., von Schiller, D., & Young, R. (2020). Organic Matter Decomposition and Ecosystem Metabolism as Tools to Assess the Functional Integrity of Streams and Rivers–A Systematic Review. *Water*, *12*(12), Article 12. https://doi.org/10.3390/w12123523
Fierer, N., Lauber, C. L., Ramirez, K. S., Zaneveld, J., Bradford, M. A., & Knight, R. (2012). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *The ISME Journal*, *6*(5), 1007-1017. https://doi.org/10.1038/ismej.2011.159
Fierer, N., Leff, J. W., Adams, B. J., Nielsen, U. N., Bates, S. T., Lauber, C. L., Owens, S., Gilbert, J. A., Wall, D. H., & Caporaso, J. G. (2012a). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences*, *109*(52), 21390-21395. https://doi.org/10.1073/pnas.1215210110
Fierer, N., Leff, J. W., Adams, B. J., Nielsen, U. N., Bates, S. T., Lauber, C. L., Owens, S., Gilbert, J. A., Wall, D. H., & Caporaso, J. G. (2012b). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences*, *109*(52), 21390-21395. https://doi.org/10.1073/pnas.1215210110
Figueiredo, C. C. de, Reis, A. de S. P. J., Araujo, A. S. de, Blum, L. E. B., Shah, K., & Paz-Ferreiro, J. (2021). Assessing the potential of sewage sludge-derived biochar as a novel phosphorus fertilizer: Influence of extractant solutions and pyrolysis temperatures. *Waste Management*, *124*, 144-153. https://doi.org/10.1016/j.wasman.2021.01.044
Fontaine, S., Abbadie, L., Aubert, M., Barot, S., Bloor, J. M. G., Derrien, D., Duchene, O., Gross, N., Henneron, L., Le Roux, X., Loeuille, N., Michel, J., Recous, S., Wipf, D., & Alvarez, G. (2024). Plant–soil synchrony in nutrient cycles: Learning from ecosystems to design sustainable agrosystems. *Global Change Biology*, *30*(1), e17034. https://doi.org/10.1111/gcb.17034
Frey, S. D. (2019). Mycorrhizal Fungi as Mediators of Soil Organic Matter Dynamics. *Annual Review of Ecology, Evolution, and Systematics*, *50*(Volume 50, 2019), 237-259. https://doi.org/10.1146/annurev-ecolsys-110617-062331
Frost, H., Bond, T., Sizmur, T., & Felipe-Sotelo, M. (2022). A review of microplastic fibres: Generation, transport, and vectors for metal(loid)s in terrestrial environments. *Environmental Science: Processes & Impacts*, *24*(4), 504-524. https://doi.org/10.1039/D1EM00541C
Fu, J., Li, P., Lin, Y., Du, H., Liu, H., Zhu, W., & Ren, H. (2022). Fight for carbon neutrality with state-of-the-art negative carbon emission technologies. *Eco-Environment & Health*, *1*(4), 259-279. https://doi.org/10.1016/j.eehl.2022.11.005
Gabasawa, A. I., Abubakar, G. A., & Obemah, D. N. (2024). Soil Regeneration and Microbial Community on Terrestrial Food Chain. En S. A. Aransiola, B. R. Babaniyi, A. B. Aransiola, & N. R. Maddela (Eds.), *Prospects for Soil Regeneration and Its Impact on Environmental Protection* (pp. 243-267). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-53270-2_11
Gao, Y., Tariq, A., Zeng, F., Sardans, J., Graciano, C., Li, X., Wang, W., & Peñuelas, J. (2024). Soil microbial functional profiles of P-cycling reveal drought-induced constraints on P-transformation in a hyper-arid desert ecosystem. *Science of The Total Environment*, *925*, 171767. https://doi.org/10.1016/j.scitotenv.2024.171767
Garaycochea, S., Altier, N. A., Leoni, C., Neal, A. L., & Romero, H. (2023). Abundance and phylogenetic distribution of eight key enzymes of the phosphorus biogeochemical cycle in grassland soils. *Environmental Microbiology Reports*, *15*(5), 352-369. https://doi.org/10.1111/1758-2229.13159
García-Díaz, C., Siles, J. A., Moreno, J. L., García, C., Ruiz-Navarro, A., & Bastida, F. (2024a).

Phenological stages of wheat modulate effects of phosphorus fertilization in plant-soil microbial interactions. *Plant and Soil*. https://doi.org/10.1007/s11104-024-06880-8
García-Díaz, C., Siles, J. A., Moreno, J. L., García, C., Ruiz-Navarro, A., & Bastida, F. (2024b). Phenological stages of wheat modulate effects of phosphorus fertilization in plant-soil microbial interactions. *Plant and Soil*. https://doi.org/10.1007/s11104-024-06880-8
Garza, D. R., & Dutilh, B. E. (2015). From cultured to uncultured genome sequences: Metagenomics and modeling microbial ecosystems. *Cellular and Molecular Life Sciences*, *72*(22), 4287-4308. https://doi.org/10.1007/s00018-015-2004-1
Gavrilescu, M. (2021). Water, Soil, and Plants Interactions in a Threatened Environment. *Water*, *13*(19), Article 19. https://doi.org/10.3390/w13192746
Geisseler, D., & Scow, K. M. (2014). Long-term effects of mineral fertilizers on soil microorganisms – A review. *Soil Biology and Biochemistry*, *75*, 54-63. https://doi.org/10.1016/j.soilbio.2014.03.023
George, T. S., Hinsinger, P., & Turner, B. L. (2016). Phosphorus in soils and plants – facing phosphorus scarcity. *Plant and Soil*, *401*(1), 1-6. https://doi.org/10.1007/s11104-016-2846-9
Gerke, J. (2022). The Central Role of Soil Organic Matter in Soil Fertility and Carbon Storage. *Soil Systems*, *6*(2), Article 2. https://doi.org/10.3390/soilsystems6020033
Gianfreda, L., & Ruggiero, P. (2006). Enzyme Activities in Soil. En P. Nannipieri & K. Smalla (Eds.), *Nucleic Acids and Proteins in Soil* (pp. 257-311). Springer. https://doi.org/10.1007/3-540-29449-X_12
Goh, K. M. (2004). Carbon sequestration and stabilization in soils: Implications for soil productivity and climate change. *Soil Science and Plant Nutrition*, *50*(4), 467-476. https://doi.org/10.1080/00380768.2004.10408502
Gower, S. T. (2003). Patterns and Mechanisms of the Forest Carbon Cycle1. *Annual Review of Environment and Resources*, *28*(Volume 28, 2003), 169-204. https://doi.org/10.1146/annurev.energy.28.050302.105515
Grigoriev, I. V., Nordberg, H., Shabalov, I., Aerts, A., Cantor, M., Goodstein, D., Kuo, A., Minovitsky, S., Nikitin, R., Ohm, R. A., Otillar, R., Poliakov, A., Ratnere, I., Riley, R., Smirnova, T., Rokhsar, D., & Dubchak, I. (2012). The Genome Portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Research*, *40*(D1), D26-D32. https://doi.org/10.1093/nar/gkr947
Grossart, H.-P., Massana, R., McMahon, K. D., & Walsh, D. A. (2020). Linking metagenomics to aquatic microbial ecology and biogeochemical cycles. *Limnology and Oceanography*, *65*(S1), S2-S20. https://doi.org/10.1002/lno.11382
Grünberger, F., Ferreira-Cerca, S., & Grohmann, D. (2022). Nanopore sequencing of RNA and cDNA molecules in Escherichia coli. *RNA*, *28*(3), 400-417. https://doi.org/10.1261/rna.078937.121
Guidi, C., Biarnés, X., Planas, A., & De Mey, M. (2023). Controlled processivity in glycosyltransferases: A way to expand the enzymatic toolbox. *Biotechnology Advances*, *63*, 108081. https://doi.org/10.1016/j.biotechadv.2022.108081
Ha, T.-H., Mahasti, N. N. N., Lu, M.-C., & Huang, Y.-H. (2023). Ammonium-nitrogen recovery as struvite from swine wastewater using various magnesium sources. *Separation and Purification Technology*, *308*, 122870. https://doi.org/10.1016/j.seppur.2022.122870
Haj-Amor, Z., Araya, T., Kim, D.-G., Bouri, S., Lee, J., Ghiloufi, W., Yang, Y., Kang, H., Jhariya, M. K., Banerjee, A., & Lal, R. (2022). Soil salinity and its associated effects on soil microorganisms, greenhouse gas emissions, crop yield, biodiversity and desertification: A review. *Science of The Total Environment*, *843*, 156946. https://doi.org/10.1016/j.scitotenv.2022.156946
Hamada, M. A., & Soliman, E. R. S. (2023). Characterization and genomics identification of key genes involved in denitrification-DNRA-nitrification pathway of plant growth-promoting rhizobacteria (Serratia marcescens OK482790). *BMC Microbiology*, *23*(1), 210. https://doi.org/10.1186/s12866-023-02941-7
Han, S., Zeng, L., Luo, X., Xiong, X., Wen, S., Wang, B., Chen, W., & Huang, Q. (2018). Shifts in *Nitrobacter*- and *Nitrospira*-like nitrite-oxidizing bacterial communities under long-term fertilization practices. *Soil Biology and Biochemistry*, *124*, 118-125. https://doi.org/10.1016/j.soilbio.2018.05.033
Hanusz, Z., Tarasinska, J., & Zielinski, W. (2016). Shapiro–Wilk Test with Known Mean. *REVSTAT-Statistical Journal*, *14*(1), Article 1. https://doi.org/10.57805/revstat.v14i1.180
Hartmann, M., & Six, J. (2023). Soil structure and microbiome functions in agroecosystems. *Nature Reviews Earth & Environment*, *4*(1), 4-18. https://doi.org/10.1038/s43017-022-00366-w
Havlicek, E., & Mitchell, E. A. D. (2014). Soils Supporting Biodiversity. En J. Dighton & J. A. Krumins

(Eds.), *Interactions in Soil: Promoting Plant Growth* (pp. 27-58). Springer Netherlands. https://doi.org/10.1007/978-94-017-8890-8_2
Hertzberger, A. J., Cusick, R. D., & Margenot, A. J. (2020). A review and meta-analysis of the agricultural potential of struvite as a phosphorus fertilizer. *Soil Science Society of America Journal*, *84*(3), 653-671. https://doi.org/10.1002/saj2.20065
Hetz, S. A., & Horn, M. A. (2021). Burkholderiaceae Are Key Acetate Assimilators During Complete Denitrification in Acidic Cryoturbated Peat Circles of the Arctic Tundra. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/fmicb.2021.628269
Horwath, W. (2007). 12—CARBON CYCLING AND FORMATION OF SOIL ORGANIC MATTER. En E. A. Paul (Ed.), *Soil Microbiology, Ecology and Biochemistry (Third Edition)* (pp. 303-339). Academic Press. https://doi.org/10.1016/B978-0-08-047514-1.50016-0
Hossain, A., Krupnik, T. J., Timsina, J., Mahboob, M. G., Chaki, A. K., Farooq, M., Bhatt, R., Fahad, S., & Hasanuzzaman, M. (2020). Agricultural Land Degradation: Processes and Problems Undermining Future Food Security. En S. Fahad, M. Hasanuzzaman, M. Alam, H. Ullah, M. Saeed, I. Ali Khan, & M. Adnan (Eds.), *Environment, Climate, Plant and Vegetation Growth* (pp. 17-61). Springer International Publishing. https://doi.org/10.1007/978-3-030-49732-3_2
Houlton, B. Z., Wang, Y.-P., Vitousek, P. M., & Field, C. B. (2008). A unifying framework for dinitrogen fixation in the terrestrial biosphere. *Nature*, *454*(7202), 327-330. https://doi.org/10.1038/nature07028
Hu, H., Kristensen, J. M., Herbold, C. W., Pjevac, P., Kitzinger, K., Hausmann, B., Dueholm, M. K. D., Nielsen, P. H., & Wagner, M. (2024). Global abundance patterns, diversity, and ecology of Patescibacteria in wastewater treatment plants. *Microbiome*, *12*(1), 55. https://doi.org/10.1186/s40168-024-01769-1
Hu, R., Liu, S., Huang, W., Nan, Q., Strong, P. J., Saleem, M., Zhou, Z., Luo, Z., Shu, F., Yan, Q., He, Z., & Wang, C. (2022). Evidence for Assimilatory Nitrate Reduction as a Previously Overlooked Pathway of Reactive Nitrogen Transformation in Estuarine Suspended Particulate Matter. *Environmental Science & Technology*, *56*(20), 14852-14866. https://doi.org/10.1021/acs.est.2c04390
Hu, X., Liu, J., Liang, A., Gu, H., Liu, Z., Jin, J., & Wang, G. (2025). Soil metagenomics reveals reduced tillage improves soil functional profiles of carbon, nitrogen, and phosphorus cycling in bulk and rhizosphere soils. *Agriculture, Ecosystems & Environment*, *379*, 109371. https://doi.org/10.1016/j.agee.2024.109371
Huang, J., Gao, K., Yang, L., & Lu, Y. (2023). Successional action of Bacteroidota and Firmicutes in decomposing straw polymers in a paddy soil. *Environmental Microbiome*, *18*(1), 76. https://doi.org/10.1186/s40793-023-00533-6
Huang, L., Zhang, H., Wu, P., Entwistle, S., Li, X., Yohe, T., Yi, H., Yang, Z., & Yin, Y. (2018). dbCAN-seq: A database of carbohydrate-active enzyme (CAZyme) sequence and annotation. *Nucleic Acids Research*, *46*(D1), D516-D521. https://doi.org/10.1093/nar/gkx894
Huang, W., Gong, B., He, L., Wang, Y., & Zhou, J. (2020). Intensified nutrients removal in a modified sequencing batch reactor at low temperature: Metagenomic approach reveals the microbial community structure and mechanisms. *Chemosphere*, *244*, 125513. https://doi.org/10.1016/j.chemosphere.2019.125513
Hui, C., Jiang, H., Liu, B., Wei, R., Zhang, Y., Zhang, Q., Liang, Y., & Zhao, Y. (2020). Chitin degradation and the temporary response of bacterial chitinolytic communities to chitin amendment in soil under different fertilization regimes. *Science of The Total Environment*, *705*, 136003. https://doi.org/10.1016/j.scitotenv.2019.136003
Islam, M. M., Jana, S. K., Sengupta, S., & Mandal, S. (2024). Impact of Rhizospheric Microbiome on Rice Cultivation. *Current Microbiology*, *81*(7), 188. https://doi.org/10.1007/s00284-024-03703-y
Ivanova, A. A., Wegner, C.-E., Kim, Y., Liesack, W., & Dedysh, S. N. (2016). Identification of microbial populations driving biopolymer degradation in acidic peatlands by metatranscriptomic analysis. *Molecular Ecology*, *25*(19), 4818-4835. https://doi.org/10.1111/mec.13806
Jacoby, R., Peukert, M., Succurro, A., Koprivova, A., & Kopriva, S. (2017). The Role of Soil Microorganisms in Plant Mineral Nutrition—Current Knowledge and Future Directions. *Frontiers in Plant Science*, *8*. https://doi.org/10.3389/fpls.2017.01617
Jagadesh, M., Dash, M., Kumari, A., Singh, S. K., Verma, K. K., Kumar, P., Bhatt, R., & Sharma, S. K. (2024). Revealing the hidden world of soil microbes: Metagenomic insights into plant, bacteria, and fungi interactions for sustainable agriculture and ecosystem restoration. *Microbiological Research*, *285*, 127764.

https://doi.org/10.1016/j.micres.2024.127764
Janssen, P. J. D., Lambreva, M. D., Plumeré, N., Bartolucci, C., Antonacci, A., Buonasera, K., Frese, R. N., Scognamiglio, V., & Rea, G. (2014). Photosynthesis at the forefront of a sustainable life. *Frontiers in Chemistry*, *2*. https://doi.org/10.3389/fchem.2014.00036
Jansson, J. K., & Hofmockel, K. S. (2018). The soil microbiome—From metagenomics to metaphenomics. *Current Opinion in Microbiology*, *43*, 162-168. https://doi.org/10.1016/j.mib.2018.01.013
Javed, Z., Tripathi, G. D., Mishra, M., & Dashora, K. (2021). Actinomycetes – The microbial machinery for the organic-cycling, plant growth, and sustainable soil health. *Biocatalysis and Agricultural Biotechnology*, *31*, 101893. https://doi.org/10.1016/j.bcab.2020.101893
Jin, D. (2017). *Omics insights into rumen ureolytic bacterial community and urea metabolism in dairy cows*. https://orbi.uliege.be/handle/2268/212411
Jones, C. M., Graf, D. R. H., Bru, D., Philippot, L., & Hallin, S. (2013). The unaccounted yet abundant nitrous oxide-reducing microbial community: A potential nitrous oxide sink. *The ISME Journal*, *7*(2), 417-426. https://doi.org/10.1038/ismej.2012.125
Jones, D. L., & Oburger, E. (2011). Solubilization of Phosphorus by Soil Microorganisms. En E. Bünemann, A. Oberson, & E. Frossard (Eds.), *Phosphorus in Action: Biological Processes in Soil Phosphorus Cycling* (pp. 169-198). Springer. https://doi.org/10.1007/978-3-642-15271-9_7
Jose, D., Preena, P. G., Kumar, V. J. R., Philip, R., & Singh, I. S. B. (2020). Metaproteomic insights into ammonia oxidising bacterial consortium developed for bioaugmenting nitrification in aquaculture systems. *Biologia*, *75*(10), 1751-1757. https://doi.org/10.2478/s11756-020-00481-3
Kafle, A., Cope, K. R., Raths, R., Krishna Yakha, J., Subramanian, S., Bücking, H., & Garcia, K. (2019). Harnessing Soil Microbes to Improve Plant Phosphate Efficiency in Cropping Systems. *Agronomy*, *9*(3), Article 3. https://doi.org/10.3390/agronomy9030127
Käll, L., Canterbury, J. D., Weston, J., Noble, W. S., & MacCoss, M. J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, *4*(11), 923-925. https://doi.org/10.1038/nmeth1113
Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587-589. https://doi.org/10.1038/nmeth.4285
Kandeler, E., & Gerber, H. (1988). Short-term assay of soil urease activity using colorimetric determination of ammonium. *Biology and Fertility of Soils*, *6*(1), 68-72. https://doi.org/10.1007/BF00257924
Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, *44*(D1), D457-D462. https://doi.org/10.1093/nar/gkv1070
Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, *7*, e7359. https://doi.org/10.7717/peerj.7359
Kelly, C. N., Schwaner, G. W., Cumming, J. R., & Driscoll, T. P. (2021). Metagenomic reconstruction of nitrogen and carbon cycling pathways in forest soil: Influence of different hardwood tree species. *Soil Biology and Biochemistry*, *156*, 108226. https://doi.org/10.1016/j.soilbio.2021.108226
Khan, M. N., & Mohammad, F. (2014). Eutrophication: Challenges and Solutions. En A. A. Ansari & S. S. Gill (Eds.), *Eutrophication: Causes, Consequences and Control: Volume 2* (pp. 1-15). Springer Netherlands. https://doi.org/10.1007/978-94-007-7814-6_1
Khatoon, H., Solanki, P., Narayan, M., Tewari, L., & Rai, J. P. N. (2017). Role of microbes in organic carbon decomposition and maintenance of soil ecosystem. *International Journal of Chemical Studies*, *5*(6), 1648-1656.
Kibblewhite, M. g, Ritz, K., & Swift, M. j. (2007). Soil health in agricultural systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1492), 685-701. https://doi.org/10.1098/rstb.2007.2178
Kome, G. K., Enang, R. K., Tabi, F. O., & Yerima, B. P. K. (2019). Influence of Clay Minerals on Some Soil Fertility Attributes: A Review. *Open Journal of Soil Science*, *9*(9), Article 9. https://doi.org/10.4236/ojss.2019.99010
Krishnamoorthy, N., Dey, B., Unpaprom, Y., Ramaraj, R., Maniam, G. P., Govindan, N., Jayaraman, S.,

Arunachalam, T., & Paramasivan, B. (2021). Engineering principles and process designs for phosphorus recovery as struvite: A comprehensive review. *Journal of Environmental Chemical Engineering*, *9*(5), 105579. https://doi.org/10.1016/j.jece.2021.105579
Kulski, J. (2016). *Next Generation Sequencing: Advances, Applications and Challenges*. BoD – Books on Demand.
Kumar, A., & Yadav, A. (2024). Next Generation Sequencing in Metagenomics and Metatranscriptomics. En I. Mani & V. Singh (Eds.), *Multi-Omics Analysis of the Human Microbiome: From Technology to Clinical Applications* (pp. 49-75). Springer Nature. https://doi.org/10.1007/978-981-97-1844-3_3
Kumar, G. C., Chaudhary, J., Meena, L. K., Meena, A. L., & Kumar, A. (2021). 18—Function-driven microbial genomics for ecofriendly agriculture. En J. S. Singh, S. Tiwari, C. Singh, & A. K. Singh (Eds.), *Microbes in Land Use Change Management* (pp. 389-431). Elsevier. https://doi.org/10.1016/B978-0-12-824448-7.00021-8
Kumar, V., Singh, K., Shah, M. P., Singh, A. K., Kumar, A., & Kumar, Y. (2021). Chapter 1—Application of Omics Technologies for Microbial Community Structure and Function Analysis in Contaminated Environment. En M. P. Shah, A. Sarkar, & S. Mandal (Eds.), *Wastewater Treatment* (pp. 1-40). Elsevier. https://doi.org/10.1016/B978-0-12-821881-5.00001-5
Kuypers, M. M. M., Marchant, H. K., & Kartal, B. (2018). The microbial nitrogen-cycling network. *Nature Reviews Microbiology*, *16*(5), 263-276. https://doi.org/10.1038/nrmicro.2018.9
Kuzyakov, Y., & Xu, X. (2013). Competition between roots and microorganisms for nitrogen: Mechanisms and ecological relevance. *New Phytologist*, *198*(3), 656-669. https://doi.org/10.1111/nph.12235
Lagos, L. M., Acuña, J. J., Maruyama, F., Ogram, A., de la Luz Mora, M., & Jorquera, M. A. (2016). Effect of phosphorus addition on total and alkaline phosphomonoesterase-harboring bacterial populations in ryegrass rhizosphere microsites. *Biology and Fertility of Soils*, *52*(7), 1007-1019. https://doi.org/10.1007/s00374-016-1137-1
Lal, R. (2005). Forest soils and carbon sequestration. *Forest Ecology and Management*, *220*(1), 242-258. https://doi.org/10.1016/j.foreco.2005.08.015
Langdon, W. B. (2015). Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Mining*, *8*(1), 1. https://doi.org/10.1186/s13040-014-0034-0
Lange, M., Eisenhauer, N., Sierra, C. A., Bessler, H., Engels, C., Griffiths, R. I., Mellado-Vázquez, P. G., Malik, A. A., Roy, J., Scheu, S., Steinbeiss, S., Thomson, B. C., Trumbore, S. E., & Gleixner, G. (2015). Plant diversity increases soil microbial activity and soil carbon storage. *Nature Communications*, *6*(1), 6707. https://doi.org/10.1038/ncomms7707
Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Clemente, J. C., Burkepile, D. E., Vega Thurber, R. L., Knight, R., Beiko, R. G., & Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, *31*(9), 814-821. https://doi.org/10.1038/nbt.2676
Larney, F. J., & Angers, D. A. (2012). The role of organic amendments in soil reclamation: A review. *Canadian Journal of Soil Science*, *92*(1), 19-38. https://doi.org/10.4141/cjss2010-064
Lee, M. D. (2019). GToTree: A user-friendly workflow for phylogenomics. *Bioinformatics*, *35*(20), 4162-4164. https://doi.org/10.1093/bioinformatics/btz188
Legay, N., Clément, J. C., Grassein, F., Lavorel, S., Lemauviel-Lavenant, S., Personeni, E., Poly, F., Pommier, T., Robson, T. M., Mouhamadou, B., & Binet, M. N. (2020). Plant growth drives soil nitrogen cycling and N-related microbial activity through changing root traits. *Fungal Ecology*, *44*, 100910. https://doi.org/10.1016/j.funeco.2019.100910
Legendre, P., & Legendre, L. (2012). *Numerical Ecology*. Elsevier.
Lehmann, J., Bossio, D. A., Kögel-Knabner, I., & Rillig, M. C. (2020). The concept and future prospects of soil health. *Nature Reviews Earth & Environment*, *1*(10), 544-553. https://doi.org/10.1038/s43017-020-0080-8
Lehtovirta-Morley, L. E., Ge, C., Ross, J., Yao, H., Hazard, C., Gubry-Rangin, C., Prosser, J. I., & Nicol, G. W. (2024). Nitrosotalea devaniterrae gen. Nov., sp. Nov. And Nitrosotalea sinensis sp. Nov., two acidophilic ammonia oxidising archaea isolated from acidic soil, and proposal of the new order Nitrosotaleales ord. Nov. Within the class Nitrososphaeria of the phylum Nitrososphaerota. *International*

Journal of Systematic and Evolutionary Microbiology, *74*(9), 006387. https://doi.org/10.1099/ijsem.0.006387
Leigh, J. A., & Dodsworth, J. A. (2007). Nitrogen regulation in bacteria and archaea. *Annual Review of Microbiology*, *61*, 349-377. https://doi.org/10.1146/annurev.micro.61.080706.093409
Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293-W296. https://doi.org/10.1093/nar/gkab301
Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, *31*(10), 1674-1676. https://doi.org/10.1093/bioinformatics/btv033
Li, D., Zhang, Z., Wang, J., Zhang, P., Liu, Y., & Li, Y. (2023). Estimate of the degradation potentials of cellulose, xylan, and chitin across global prokaryotic communities. *Environmental Microbiology*, *25*(2), 397-409. https://doi.org/10.1111/1462-2920.16290
Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094-3100. https://doi.org/10.1093/bioinformatics/bty191
Li, H., Chang, L., Liu, H., & Li, Y. (2024). Diverse factors influence the amounts of carbon input to soils via rhizodeposition in plants: A review. *Science of The Total Environment*, *948*, 174858. https://doi.org/10.1016/j.scitotenv.2024.174858
Liang, J.-L., Liu, J., Jia, P., Yang, T., Zeng, Q., Zhang, S., Liao, B., Shu, W., & Li, J. (2020). Novel phosphate-solubilizing bacteria enhance soil phosphorus cycling following ecological restoration of land degraded by mining. *The ISME Journal*, *14*(6), 1600-1613. https://doi.org/10.1038/s41396-020-0632-4
Liao, X., Zhao, J., Yi, Q., Li, J., Li, Z., Wu, S., Zhang, W., & Wang, K. (2023). Metagenomic insights into the effects of organic and inorganic agricultural managements on soil phosphorus cycling. *Agriculture, Ecosystems & Environment*, *343*, 108281. https://doi.org/10.1016/j.agee.2022.108281
Liu, C., Li, X., Mansoldo, F. R. P., An, J., Kou, Y., Zhang, X., Wang, J., Zeng, J., Vermelho, A. B., & Yao, M. (2022). Microbial habitat specificity largely affects microbial co-occurrence patterns and functional profiles in wetland soils. *Geoderma*, *418*, 115866. https://doi.org/10.1016/j.geoderma.2022.115866
Liu, L., Gao, Z., Yang, Y., Gao, Y., Mahmood, M., Jiao, H., Wang, Z., & Liu, J. (2023). Long-term high-P fertilizer input shifts soil P cycle genes and microorganism communities in dryland wheat production systems. *Agriculture, Ecosystems & Environment*, *342*, 108226. https://doi.org/10.1016/j.agee.2022.108226
Liu, M., Clarke, L. J., Baker, S. C., Jordan, G. J., & Burridge, C. P. (2020). A practical guide to DNA metabarcoding for entomological ecologists. *Ecological Entomology*, *45*(3), 373-385. https://doi.org/10.1111/een.12831
Liu, S., & Liu, Z. (2020). Distinct capabilities of different Gammaproteobacterial strains on utilizing small peptides in seawater. *Scientific Reports*, *10*, 464. https://doi.org/10.1038/s41598-019-57189-x
Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research*, *42*(Database issue), D490-495. https://doi.org/10.1093/nar/gkt1178
López-Mondéjar, R., Algora, C., & Baldrian, P. (2019). Lignocellulolytic systems of soil bacteria: A vast and diverse toolbox for biotechnological conversion processes. *Biotechnology Advances*, *37*(6), 107374. https://doi.org/10.1016/j.biotechadv.2019.03.013
López-Mondéjar, R., Brabcová, V., Štursová, M., Davidová, A., Jansa, J., Cajthaml, T., & Baldrian, P. (2018). Decomposer food web in a deciduous forest shows high share of generalist microorganisms and importance of microbial biomass recycling. *The ISME Journal*, *12*(7), 1768-1778. https://doi.org/10.1038/s41396-018-0084-2
López-Mondéjar, R., Tláskal, V., da Rocha, U. N., & Baldrian, P. (2022). Global Distribution of Carbohydrate Utilization Potential in the Prokaryotic Tree of Life. *mSystems*, *7*(6), e00829-22. https://doi.org/10.1128/msystems.00829-22
López-Mondéjar, R., Tláskal, V., Větrovský, T., Štursová, M., Toscan, R., Nunes da Rocha, U., & Baldrian, P. (2020). Metagenomics and stable isotope probing reveal the complementary contribution of fungal and bacterial communities in the recycling of dead biomass in forest soil. *Soil Biology and Biochemistry*, *148*, 107875. https://doi.org/10.1016/j.soilbio.2020.107875

López-Mondéjar, R., Voříšková, J., Větrovský, T., & Baldrian, P. (2015). The bacterial community inhabiting temperate deciduous forests is vertically stratified and undergoes seasonal dynamics. *Soil Biology and Biochemistry*, *87*, 43-50. https://doi.org/10.1016/j.soilbio.2015.04.008
López-Mondéjar, R., Zühlke, D., Becher, D., Riedel, K., & Baldrian, P. (2016). Cellulose and hemicellulose decomposition by forest soil bacteria proceeds by the action of structurally variable enzymatic systems. *Scientific Reports*, *6*(1), 25279. https://doi.org/10.1038/srep25279
Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8
Lubin, E. A., Henry, J. T., Fiebig, A., Crosson, S., & Laub, M. T. (2015). Identification of the PhoB Regulon and Role of PhoU in the Phosphate Starvation Response of Caulobacter crescentus. *Journal of Bacteriology*, *198*(1), 187-200. https://doi.org/10.1128/jb.00658-15
Luo, G., Rensing, C., Chen, H., Liu, M., Wang, M., Guo, S., Ling, N., & Shen, Q. (2018). Deciphering the associations between soil microbial diversity and ecosystem multifunctionality driven by long-term fertilization management. *Functional Ecology*, *32*(4), 1103-1116. https://doi.org/10.1111/1365-2435.13039
Lyu, C., Li, X., Yu, H., Song, Y., Gao, H., & Yuan, P. (2023). Insight into the microbial nitrogen cycle in riparian soils in an agricultural region. *Environmental Research*, *231*, 116100. https://doi.org/10.1016/j.envres.2023.116100
Lyu, Q., Zhang, K., Zhu, Q., Li, Z., Liu, Y., Fitzek, E., Yohe, T., Zhao, L., Li, W., Liu, T., Yin, Y., & Liu, W. (2018). Structural and biochemical characterization of a multidomain alginate lyase reveals a novel role of CBM32 in CAZymes. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1862*(9), 1862-1869. https://doi.org/10.1016/j.bbagen.2018.05.024
Ma, B., Lu, C., Wang, Y., Yu, J., Zhao, K., Xue, R., Ren, H., Lv, X., Pan, R., Zhang, J., Zhu, Y., & Xu, J. (2023). A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources. *Nature Communications*, *14*(1), 7318. https://doi.org/10.1038/s41467-023-43000-z
Mäki, M., Heinonsalo, J., Hellén, H., & Bäck, J. (2017). Contribution of understorey vegetation and soil processes to boreal forest isoprenoid exchange. *Biogeosciences*, *14*(5), 1055-1073. https://doi.org/10.5194/bg-14-1055-2017
Malard, L. A., & Guisan, A. (2023). Into the microbial niche. *Trends in Ecology & Evolution*, *38*(10), 936-945. https://doi.org/10.1016/j.tree.2023.04.015
Manici, L. M., Caputo, F., De Sabata, D., & Fornasier, F. (2024). The enzyme patterns of Ascomycota and Basidiomycota fungi reveal their different functions in soil. *Applied Soil Ecology*, *196*, 105323. https://doi.org/10.1016/j.apsoil.2024.105323
Manici, L. M., Caputo, F., Fornasier, F., Paletto, A., Ceotto, E., & De Meo, I. (2024). Ascomycota and Basidiomycota fungal phyla as indicators of land use efficiency for soil organic carbon accrual with woody plantations. *Ecological Indicators*, *160*, 111796. https://doi.org/10.1016/j.ecolind.2024.111796
Martikainen, P. J. (2022). Heterotrophic nitrification – An eternal mystery in the nitrogen cycle. *Soil Biology and Biochemistry*, *168*, 108611. https://doi.org/10.1016/j.soilbio.2022.108611
Martines, A. M., Nogueira, M. A., Santos, C. A., Nakatani, A. S., Andrade, C. A., Coscione, A. R., Cantarella, H., Sousa, J. P., & Cardoso, E. J. B. N. (2010). Ammonia volatilization in soil treated with tannery sludge. *Bioresource Technology*, *101*(12), 4690-4696. https://doi.org/10.1016/j.biortech.2010.01.104
Martinez-Argudo, I., Little, R., Shearer, N., Johnson, P., & Dixon, R. (2004). The NifL-NifA System: A Multidomain Transcriptional Regulatory Complex That Integrates Environmental Signals. *Journal of Bacteriology*, *186*(3), 601-610. https://doi.org/10.1128/jb.186.3.601-610.2004
Marzi, M., Shahbazi, K., Kharazi, N., & Rezaei, M. (2020). The Influence of Organic Amendment Source on Carbon and Nitrogen Mineralization in Different Soils. *Journal of Soil Science and Plant Nutrition*, *20*(1), 177-191. https://doi.org/10.1007/s42729-019-00116-w
Mbene, K., Alakeh, N. M., Asongwe, G. A., Fomenky, N. N., Nkenganang, M. C. A., & Tening, A. S. (2023). Assessment of the Soil Resources of Sub-Saharan Africa in Relation to Food Security: Perspectives Past, Present, and Future. En *Soil Constraints and Productivity*. CRC Press.
McKee, L. S., Martínez-Abad, A., Ruthes, A. C., Vilaplana, F., & Brumer, H. (2019). Focused Metabolism of β-Glucans by the Soil Bacteroidetes Species Chitinophaga pinensis. *Applied and Environmental*

Microbiology, *85*(2), e02231-18. https://doi.org/10.1128/AEM.02231-18
McKight, P. E., & Najab, J. (2010). Kruskal-Wallis Test. En *The Corsini Encyclopedia of Psychology* (pp. 1-1). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470479216.corpsy0491
Middelboe, M., Traving, S. J., Castillo, D., Kalatzis, P. G., & Glud, R. N. (2025). Prophage-encoded chitinase gene supports growth of its bacterial host isolated from deep-sea sediments. *The ISME Journal*, *19*(1), wraf004. https://doi.org/10.1093/ismejo/wraf004
Mikheenko, A., Saveliev, V., & Gurevich, A. (2016). MetaQUAST: Evaluation of metagenome assemblies. *Bioinformatics*, *32*(7), 1088-1090. https://doi.org/10.1093/bioinformatics/btv697
Miller, S. E., Colman, A. S., & Waldbauer, J. R. (2023a). Metaproteomics reveals functional partitioning and vegetational variation among permafrost-affected Arctic soil bacterial communities. *mSystems*, *8*(3), e01238-22. https://doi.org/10.1128/msystems.01238-22
Miller, S. E., Colman, A. S., & Waldbauer, J. R. (2023b). Metaproteomics reveals functional partitioning and vegetational variation among permafrost-affected Arctic soil bacterial communities. *mSystems*, *8*(3), e01238-22. https://doi.org/10.1128/msystems.01238-22
Montaldo, S. (2022). *The Green Deal and the Case for a Soil Health Framework Directive*. https://iris.unito.it/handle/2318/1871262
Mosley, O. E., Gios, E., Close, M., Weaver, L., Daughney, C., & Handley, K. M. (2022). Nitrogen cycling and microbial cooperation in the terrestrial subsurface. *The ISME Journal*, *16*(11), 2561-2573. https://doi.org/10.1038/s41396-022-01300-0
Muleta, A., Tesfaye, K., Haile Selassie, T. H., Cook, D. R., & Assefa, F. (2021). Phosphate solubilization and multiple plant growth promoting properties of Mesorhizobium species nodulating chickpea from acidic soils of Ethiopia. *Archives of Microbiology*, *203*(5), 2129-2137. https://doi.org/10.1007/s00203-021-02189-7
Nagar, S., Bharti, M., & Negi, R. K. (2023). Genome-resolved metagenomics revealed metal-resistance, geochemical cycles in a Himalayan hot spring. *Applied Microbiology and Biotechnology*, *107*(10), 3273-3289. https://doi.org/10.1007/s00253-023-12503-6
Naitam, M. G., & Kaushik, R. (2021). Archaea: An Agro-Ecological Perspective. *Current Microbiology*, *78*(7), 2510-2521. https://doi.org/10.1007/s00284-021-02537-2
Nannipieri, P., Giagnoni, L., Landi, L., & Renella, G. (2011). Role of Phosphatase Enzymes in Soil. En E. Bünemann, A. Oberson, & E. Frossard (Eds.), *Phosphorus in Action: Biological Processes in Soil Phosphorus Cycling* (pp. 215-243). Springer. https://doi.org/10.1007/978-3-642-15271-9_9
Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.-M., Huntemann, M., Palaniappan, K., Ladau, J., Mukherjee, S., Reddy, T. B. K., Nielsen, T., Kirton, E., Faria, J. P., Edirisinghe, J. N., Henry, C. S., … Eloe-Fadrosh, E. A. (2021). A genomic catalog of Earth's microbiomes. *Nature Biotechnology*, *39*(4), 499-509. https://doi.org/10.1038/s41587-020-0718-6
Nebauer, D. J., Pearson, L. A., & Neilan, B. A. (2024). Critical steps in an environmental metaproteomics workflow. *Environmental Microbiology*, *26*(5), e16637. https://doi.org/10.1111/1462-2920.16637
Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268-274. https://doi.org/10.1093/molbev/msu300
Nielsen, U. N., Wall, D. H., & Six, J. (2015). Soil Biodiversity and the Environment. *Annual Review of Environment and Resources*, *40*(Volume 40, 2015), 63-90. https://doi.org/10.1146/annurev-environ-102014-021257
Nizamani, M. M., Hughes, A. C., Qureshi, S., Zhang, Q., Tarafder, E., Das, D., Acharya, K., Wang, Y., & Zhang, Z.-G. (2024). Microbial biodiversity and plant functional trait interactions in multifunctional ecosystems. *Applied Soil Ecology*, *201*, 105515. https://doi.org/10.1016/j.apsoil.2024.105515
Norton, J. M. (2008). Nitrification in Agricultural Soils. En *Nitrogen in Agricultural Systems* (pp. 173-199). John Wiley & Sons, Ltd. https://doi.org/10.2134/agronmonogr49.c6
Nuccio, E. E., Starr, E., Karaoz, U., Brodie, E. L., Zhou, J., Tringe, S. G., Malmstrom, R. R., Woyke, T., Banfield, J. F., Firestone, M. K., & Pett-Ridge, J. (2020). Niche differentiation is spatially and temporally regulated in the rhizosphere. *The ISME Journal*, *14*(4), 999-1014. https://doi.org/10.1038/s41396-019-0582-x
Nunan, N., Schmidt, H., & Raynaud, X. (2020). The ecology of heterogeneity: Soil bacterial communities

and C dynamics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*. https://doi.org/10.1098/rstb.2019.0249
Offre, P., Spang, A., & Schleper, C. (2013). Archaea in Biogeochemical Cycles. *Annual Review of Microbiology*, *67*(Volume 67, 2013), 437-457. https://doi.org/10.1146/annurev-micro-092412-155614
Oksanen, Jari and Blanchet, F Guillaume and Friendly, Michael and Kindt, Roeland and Legendre, Pierre and McGlinn, Dan and Minchin, Peter R and O'hara, RB and Simpson, Gavin L and Solymos, & Peter and others. (2019). *Package 'vegan'* [Software]. Community ecology package, version 2(9).
Olsen, S. r., & Sommers, L. e. (1982). Phosphorus. En *Methods of Soil Analysis* (pp. 403-430). John Wiley & Sons, Ltd. https://doi.org/10.2134/agronmonogr9.2.2ed.c24
Orakov, A., Fullam, A., Coelho, L. P., Khedkar, S., Szklarczyk, D., Mende, D. R., Schmidt, T. S. B., & Bork, P. (2021). GUNC: Detection of chimerism and contamination in prokaryotic genomes. *Genome Biology*, *22*(1), 178. https://doi.org/10.1186/s13059-021-02393-0
Ortíz-Castro, R., Contreras-Cornejo, H. A., Macías-Rodríguez, L., & López-Bucio, J. (2009). The role of microbial signals in plant growth and development. *Plant Signaling & Behavior*, *4*(8), 701-712. https://doi.org/10.4161/psb.4.8.9047
Pacciani-Mori, L., Giometto, A., Suweis, S., & Maritan, A. (2020). Dynamic metabolic adaptation can promote species coexistence in competitive microbial communities. *PLOS Computational Biology*, *16*(5), e1007896. https://doi.org/10.1371/journal.pcbi.1007896
Pan, L., & Cai, B. (2023). Phosphate-Solubilizing Bacteria: Advances in Their Physiology, Molecular Mechanisms and Microbial Community Effects. *Microorganisms*, *11*(12), Article 12. https://doi.org/10.3390/microorganisms11122904
Pan, Y., Birdsey, R. A., Fang, J., Houghton, R., Kauppi, P. E., Kurz, W. A., Phillips, O. L., Shvidenko, A., Lewis, S. L., Canadell, J. G., Ciais, P., Jackson, R. B., Pacala, S. W., McGuire, A. D., Piao, S., Rautiainen, A., Sitch, S., & Hayes, D. (2011). A Large and Persistent Carbon Sink in the World's Forests. *Science*, *333*(6045), 988-993. https://doi.org/10.1126/science.1201609
Panagos, P., Borrelli, P., Jones, A., & Robinson, D. A. (2024). A 1 billion euro mission: A Soil Deal for Europe. *European Journal of Soil Science*, *75*(1), e13466. https://doi.org/10.1111/ejss.13466
Panagos, P., Montanarella, L., Barbero, M., Schneegans, A., Aguglia, L., & Jones, A. (2022). Soil priorities in the European Union. *Geoderma Regional*, *29*, e00510. https://doi.org/10.1016/j.geodrs.2022.e00510
Pandey, A., Suter, H., He, J.-Z., Hu, H.-W., & Chen, D. (2019). Dissimilatory nitrate reduction to ammonium dominates nitrate reduction in long-term low nitrogen fertilized rice paddies. *Soil Biology and Biochemistry*, *131*, 149-156. https://doi.org/10.1016/j.soilbio.2019.01.007
Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P.-A., & Hugenholtz, P. (2022). GTDB: An ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, *50*(D1), D785-D794. https://doi.org/10.1093/nar/gkab776
Paul, E. A. (2016). The nature and dynamics of soil organic matter: Plant inputs, microbial transformations, and organic matter stabilization. *Soil Biology and Biochemistry*, *98*, 109-126. https://doi.org/10.1016/j.soilbio.2016.04.001
Pérez-Cobas, A. E., Gomez-Valero, L., & Buchrieser, C. (2020). Metagenomic approaches in microbial ecology: An update on whole-genome and marker gene sequencing analyses. *Microbial Genomics*, *6*(8), e000409. https://doi.org/10.1099/mgen.0.000409
Pester, M., Maixner, F., Berry, D., Rattei, T., Koch, H., Lücker, S., Nowka, B., Richter, A., Spieck, E., Lebedeva, E., Loy, A., Wagner, M., & Daims, H. (2014). Encoding the beta subunit of nitrite oxidoreductase as functional and phylogenetic marker for nitrite-oxidizing itrospira. *Environmental Microbiology*, *16*(10), 3055-3071. https://doi.org/10.1111/1462-2920.12300
Petersen, S. O., Petersen, J., & Rubæk, G. H. (2003). Dynamics and plant uptake of nitrogen and phosphorus in soil amended with sewage sludge. *Applied Soil Ecology*, *24*(2), 187-195. https://doi.org/10.1016/S0929-1393(03)00087-8
Philippot, L., Griffiths, B. S., & Langenheder, S. (2021). Microbial Community Resilience across Ecosystems and Multiple Disturbances. *Microbiology and Molecular Biology Reviews*, *85*(2), 10.1128/mmbr.00026-20. https://doi.org/10.1128/mmbr.00026-20
Pierzynski, G. M., Vance, G. F., & Sims, J. T. (2005). *Soils and Environmental Quality*. CRC Press.

Pold, G., Bonilla-Rosso, G., Saghaï, A., Strous, M., Jones, C. M., & Hallin, S. (2024). Phylogenetics and environmental distribution of nitric oxide-forming nitrite reductases reveal their distinct functional and ecological roles. *ISME Communications*, *4*(1), ycae020. https://doi.org/10.1093/ismeco/ycae020
Pradeep, N. S., & Edison, L. K. (2022). *Microbial Beta Glucanases: Molecular Structure, Functions and Applications*. Springer Nature.
Prakash, T., & Taylor, T. D. (2012). Functional assignment of metagenomic data: Challenges and applications. *Briefings in Bioinformatics*, *13*(6), 711-727. https://doi.org/10.1093/bib/bbs033
Prasad, S., Malav, L. C., Choudhary, J., Kannojiya, S., Kundu, M., Kumar, S., & Yadav, A. N. (2021). Soil Microbiomes for Healthy Nutrient Recycling. En A. N. Yadav, J. Singh, C. Singh, & N. Yadav (Eds.), *Current Trends in Microbial Biotechnology for Sustainable Agriculture* (pp. 1-21). Springer. https://doi.org/10.1007/978-981-15-6949-4_1
Pruitt, K. D., Tatusova, T., Klimke, W., & Maglott, D. R. (2009). NCBI Reference Sequences: Current status, policy and new initiatives. *Nucleic Acids Research*, *37*(suppl_1), D32-D36. https://doi.org/10.1093/nar/gkn721
Putz, T. (2018). Imprint of management on microbial communities in arable soil. *Acta Universitatis Agriculturae Sueciae*, *2018:35*. https://res.slu.se/id/publ/104198
Qin, W., Wei, S. P., Zheng, Y., Choi, E., Li, X., Johnston, J., Wan, X., Abrahamson, B., Flinkstrom, Z., Wang, B., Li, H., Hou, L., Tao, Q., Chlouber, W. W., Sun, X., Wells, M., Ngo, L., Hunt, K. A., Urakawa, H., … Winkler, M.-K. H. (2024). Ammonia-oxidizing bacteria and archaea exhibit differential nitrogen source preferences. *Nature Microbiology*, *9*(2), 524-536. https://doi.org/10.1038/s41564-023-01593-7
Raglin, S. S., Soman, C., Ma, Y., & Kent, A. D. (2022). Long Term Influence of Fertility and Rotation on Soil Nitrification Potential and Nitrifier Communities. *Frontiers in Soil Science*, *2*. https://doi.org/10.3389/fsoil.2022.838497
Ragot, S. A., Kertesz, M. A., & Bünemann, E. K. (2015). phoD Alkaline Phosphatase Gene Diversity in Soil. *Applied and Environmental Microbiology*, *81*(20), 7281-7289. https://doi.org/10.1128/AEM.01823-15
Ragot, S. A., Kertesz, M. A., Mészáros, É., Frossard, E., & Bünemann, E. K. (2017). Soil phoD and phoX alkaline phosphatase gene diversity responds to multiple environmental factors. *FEMS Microbiology Ecology*, *93*(1), fiw212. https://doi.org/10.1093/femsec/fiw212
Ramazzotti, M., & Bacci, G. (2018). Chapter 5—16S rRNA-Based Taxonomy Profiling in the Metagenomics Era. En M. Nagarajan (Ed.), *Metagenomics* (pp. 103-119). Academic Press. https://doi.org/10.1016/B978-0-08-102268-9.00005-7
Ramos Cabrera, E. V., Delgado Espinosa, Z. Y., & Solis Pino, A. F. (2024). Use of Phosphorus-Solubilizing Microorganisms as a Biotechnological Alternative: A Review. *Microorganisms*, *12*(8), Article 8. https://doi.org/10.3390/microorganisms12081591
Rath, K. M., & Rousk, J. (2015). Salt effects on the soil microbial decomposer community and their role in organic carbon cycling: A review. *Soil Biology and Biochemistry*, *81*, 108-123. https://doi.org/10.1016/j.soilbio.2014.11.001
Rawat, P., Das, S., Shankhdhar, D., & Shankhdhar, S. C. (2021). Phosphate-Solubilizing Microorganisms: Mechanism and Their Role in Phosphate Solubilization and Uptake. *Journal of Soil Science and Plant Nutrition*, *21*(1), 49-68. https://doi.org/10.1007/s42729-020-00342-7
R-Core-Team. (2023). *R: The R Project for Statistical Computing*. https://www.r-project.org/
Reck, M., Tomasch, J., Deng, Z., Jarek, M., Husemann, P., Wagner-Döbler, I., & On behalf of COMBACTE consortium. (2015). Stool metatranscriptomics: A technical guideline for mRNA stabilisation and isolation. *BMC Genomics*, *16*(1), 494. https://doi.org/10.1186/s12864-015-1694-y
Reed, S. C., Cleveland, C. C., & Townsend, A. R. (2011). Functional Ecology of Free-Living Nitrogen Fixation: A Contemporary Perspective. *Annual Review of Ecology, Evolution, and Systematics*, *42*(Volume 42, 2011), 489-512. https://doi.org/10.1146/annurev-ecolsys-102710-145034
Rho, M., Tang, H., & Ye, Y. (2010). FragGeneScan: Predicting genes in short and error-prone reads. *Nucleic Acids Research*, *38*(20), e191. https://doi.org/10.1093/nar/gkq747
Richardson, A. E., Lynch, J. P., Ryan, P. R., Delhaize, E., Smith, F. A., Smith, S. E., Harvey, P. R., Ryan, M. H., Veneklaas, E. J., Lambers, H., Oberson, A., Culvenor, R. A., & Simpson, R. J. (2011). Plant and microbial strategies to improve the phosphorus efficiency of agriculture. *Plant and Soil*, *349*(1), 121-156.

https://doi.org/10.1007/s11104-011-0950-4
Richardson, A. E., & Simpson, R. J. (2011). Soil Microorganisms Mediating Phosphorus Availability Update on Microbial Phosphorus. *Plant Physiology*, *156*(3), 989-996. https://doi.org/10.1104/pp.111.175448
Ritchie, S.W & J.J. Hanway. (1982a). How a Corn Plant Develops. *Iowa State Univ. of Science and Technol*.
Ritchie, S.W & J.J. Hanway. (1982b). How a Corn Plant Develops. *Iowa State Univ. of Science and Technol*.
Rodriguez, J., Chakrabarti, S., Choi, E., Shehadeh, N., Sierra-Martinez, S., Zhao, J., & Martens-Habbena, W. (2021). Nutrient-Limited Enrichments of Nitrifiers From Soil Yield Consortia of Nitrosocosmicus-Affiliated AOA and Nitrospira-Affiliated NOB. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/fmicb.2021.671480
Rosenzweig, C., & Hillel, D. (2000). SOILS AND GLOBAL CLIMATE CHANGE: CHALLENGES AND OPPORTUNITIES. *Soil Science*, *165*(1), 47.
Ruiz-Herrera, J., & Ortiz-Castellanos, L. (2019). Cell wall glucans of fungi. A review. *The Cell Surface*, *5*, 100022. https://doi.org/10.1016/j.tcsw.2019.100022
Ruiz-navarro, A., Delgado-baquerizo, M., Cano-díaz, C., García, C., & Bastida, F. (2023). Abiotic and biotic drivers of struvite solubilization in contrasting soils. *Pedosphere*, *33*(6), 828-837. https://doi.org/10.1016/j.pedsph.2023.03.014
Russell, E. W., Cooke, G. W., Pirie, N. W., & Bell, G. D. H. (1997). The rôle of organic matter in soil fertility. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *281*(980), 209-219. https://doi.org/10.1098/rstb.1977.0134
Rütting, T., Boeckx, P., Müller, C., & Klemedtsson, L. (2011). Assessment of the importance of dissimilatory nitrate reduction to ammonium for the terrestrial nitrogen cycle. *Biogeosciences*, *8*(7), 1779-1791. https://doi.org/10.5194/bg-8-1779-2011
Saco, P. M., McDonough, K. R., Rodriguez, J. F., Rivera-Zayas, J., & Sandi, S. G. (2021). The role of soils in the regulation of hazards and extreme events. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1834), 20200178. https://doi.org/10.1098/rstb.2020.0178
Saleh-Lakha, S., Miller, M., Campbell, R. G., Schneider, K., Elahimanesh, P., Hart, M. M., & Trevors, J. T. (2005). Microbial gene expression in soil: Methods, applications and challenges. *Journal of Microbiological Methods*, *63*(1), 1-19. https://doi.org/10.1016/j.mimet.2005.03.007
Santorufo, L., Panico, S. C., Zarrelli, A., De Marco, A., Santini, G., Memoli, V., & Maisto, G. (2024). Examining litter and soil characteristics impact on decomposer communities, detritivores and carbon accumulation in the Mediterranean area. *Plant and Soil*, *505*(1), 381-396. https://doi.org/10.1007/s11104-024-06683-x
Saraiva, J. P., Bartholomäus, A., Toscan, R. B., Baldrian, P., & Nunes da Rocha, U. (2023). Recovery of 197 eukaryotic bins reveals major challenges for eukaryote genome reconstruction from terrestrial metagenomes. *Molecular Ecology Resources*, *23*(5), 1066-1076. https://doi.org/10.1111/1755-0998.13776
Schneider, K. D., Thiessen Martens, J. R., Zvomuya, F., Reid, D. K., Fraser, T. D., Lynch, D. H., O'Halloran, I. P., & Wilson, H. F. (2019). Options for Improved Phosphorus Cycling and Use in Agriculture at the Field and Regional Scales. *Journal of Environmental Quality*, *48*(5), 1247-1264. https://doi.org/10.2134/jeq2019.02.0070
Schneider, T., & Riedel, K. (2010). Environmental proteomics: Analysis of structure and function of microbial communities. *PROTEOMICS*, *10*(4), 785-798. https://doi.org/10.1002/pmic.200900450
Schniete, J. K., Brüser, T., Horn, M. A., & Tschowri, N. (2024). Specialized biopolymers: Versatile tools for microbial resilience. *Current Opinion in Microbiology*, *77*, 102405. https://doi.org/10.1016/j.mib.2023.102405
Selvaraj, S., Chauhan, A., Dutta, V., Verma, R., Rao, S. K., Radhakrishnan, A., & Ghotekar, S. (2024). A state-of-the-art review on plant-derived cellulose-based green hydrogels and their multifunctional role in advanced biomedical applications. *International Journal of Biological Macromolecules*, *265*, 130991. https://doi.org/10.1016/j.ijbiomac.2024.130991
Senesi, N., & Loffredo, E. (1998). The Chemistry of Soil Organic Matter. En *Soil Physical Chemistry* (2.ª ed.). CRC Press.

Shaffer, M., Borton, M. A., McGivern, B. B., Zayed, A. A., La Rosa, S. L., Solden, L. M., Liu, P., Narrowe, A. B., Rodríguez-Ramos, J., Bolduc, B., Gazitúa, M. C., Daly, R. A., Smith, G. J., Vik, D. R., Pope, P. B., Sullivan, M. B., Roux, S., & Wrighton, K. C. (2020). DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Research*, *48*(16), 8883-8900. https://doi.org/10.1093/nar/gkaa621

Shah, F., & Wu, W. (2019). Soil and Crop Management Strategies to Ensure Higher Crop Productivity within Sustainable Environments. *Sustainability*, *11*(5), Article 5. https://doi.org/10.3390/su11051485

Shakya, M., Lo, C.-C., & Chain, P. S. G. (2019). Advances and Challenges in Metatranscriptomic Analysis. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00904

Shao, Q., Ju, Y., Guo, W., Xia, X., Bian, R., Li, L., Li, W., Liu, X., Zheng, J., & Pan, G. (2019). Pyrolyzed municipal sewage sludge ensured safe grain production while reduced C emissions in a paddy soil under rice and wheat rotation. *Environmental Science and Pollution Research*, *26*(9), 9244-9256. https://doi.org/10.1007/s11356-019-04417-6

Sharuddin, S. S., Ramli, N., Yusoff, M. Z. M., Muhammad, N. A. N., Ho, L. S., & Maeda, T. (2022). Advancement of Metatranscriptomics towards Productive Agriculture and Sustainable Environment: A Review. *International Journal of Molecular Sciences*, *23*(7), Article 7. https://doi.org/10.3390/ijms23073737

Shen, K., Liu, Q., Yang, G., Chen, H., Liang, C., Lai, P., Li, C., Wang, X., & Hu, Y. (2023). Effects of Phosphorus Reduction on Soil Phosphorus Pool Composition and Phosphorus Solubilizing Microorganisms. *Scientia Agricultura Sinica*, *56*(15), 2941-2953. https://doi.org/10.3864/j.issn.0578-1752.2023.15.009

Shi, Q., Abdel-Hamid, A. M., Sun, Z., Cheng, Y., Tu, T., Cann, I., Yao, B., & Zhu, W. (2023). Carbohydrate-binding modules facilitate the enzymatic hydrolysis of lignocellulosic biomass: Releasing reducing sugars and dissociative lignin available for producing biofuels and chemicals. *Biotechnology Advances*, *65*, 108126. https://doi.org/10.1016/j.biotechadv.2023.108126

Shi, Y., Liu, X., Zhang, Q., Gao, P., & Ren, J. (2020). Biochar and organic fertilizer changed the ammonia-oxidizing bacteria and archaea community structure of saline–alkali soil in the North China Plain. *Journal of Soils and Sediments*, *20*(1), 12-23. https://doi.org/10.1007/s11368-019-02364-w

Siles, J. A., De la Rosa, J. M., González-Pérez, J. A., Fernández-Pérez, V., García-Díaz, C., Moreno, J. L., García, C., & Bastida, F. (2024). Long-term restoration with organic amendments is clearer evidenced by soil organic matter composition than by changes in microbial taxonomy and functionality. *Applied Soil Ecology*, *198*, 105383. https://doi.org/10.1016/j.apsoil.2024.105383

Siles, J. A., Gómez-Pérez, R., Vera, A., García, C., & Bastida, F. (2024). A comparison among EL-FAME, PLFA, and quantitative PCR methods to detect changes in the abundance of soil bacteria and fungi. *Soil Biology and Biochemistry*, *198*, 109557. https://doi.org/10.1016/j.soilbio.2024.109557

Siles, J. A., Starke, R., Martinovic, T., Parente Fernandes, M. L., Orgiazzi, A., & Bastida, F. (2022). Distribution of phosphorus cycling genes across land uses and microbial taxonomic groups based on metagenome and genome mining. *Soil Biology and Biochemistry*, *174*, 108826. https://doi.org/10.1016/j.soilbio.2022.108826

Siles, J., De la Rosa, J., González-Pérez, J., Fernández-Pérez, V., García-Díaz, C., Moreno, J., Garcia, C., & Bastida, F. (2024). Long-term restoration with organic amendments is clearer evidenced by soil organic matter composition than by changes in microbial taxonomy and functionality. *Applied Soil Ecology*, *198*, 105383. https://doi.org/10.1016/j.apsoil.2024.105383

Silva, G. O. A. da, Southam, G., & Gagen, E. J. (2022). Accelerating soil aggregate formation: A review on microbial processes as the critical step in a post-mining rehabilitation context. *Soil Research*, *61*(3), 209-223. https://doi.org/10.1071/SR22092

Sime, A. M., Kifle, B. A., Woldesemayat, A. A., & Gemeda, M. T. (2024). Microbial carbohydrate active enzyme (CAZyme) genes and diversity from Menagesha Suba natural forest soils of Ethiopia as revealed by shotgun metagenomic sequencing. *BMC Microbiology*, *24*(1), 285. https://doi.org/10.1186/s12866-024-03436-9

Smith, P., Cotrufo, M. F., Rumpel, C., Paustian, K., Kuikman, P. J., Elliott, J. A., McDowell, R., Griffiths, R. I., Asakawa, S., Bustamante, M., House, J. I., Sobocká, J., Harper, R., Pan, G., West, P. C., Gerber, J. S., Clark, J. M., Adhya, T., Scholes, R. J., & Scholes, M. C. (2015). Biogeochemical cycles and biodiversity

as key drivers of ecosystem services provided by soils. *SOIL*, *1*(2), 665-685. https://doi.org/10.5194/soil-1-665-2015
Smith, P., House, J. I., Bustamante, M., Sobocká, J., Harper, R., Pan, G., West, P. C., Clark, J. M., Adhya, T., Rumpel, C., Paustian, K., Kuikman, P., Cotrufo, M. F., Elliott, J. A., McDowell, R., Griffiths, R. I., Asakawa, S., Bondeau, A., Jain, A. K., … Pugh, T. A. M. (2016). Global change pressures on soils from land use and management. *Global Change Biology*, *22*(3), 1008-1028. https://doi.org/10.1111/gcb.13068
Šnajdr, J., Valášková, V., Merhautová, V., Herinková, J., Cajthaml, T., & Baldrian, P. (2008). Spatial variability of enzyme activities and microbial biomass in the upper layers of *Quercus petraea* forest soil. *Soil Biology and Biochemistry*, *40*(9), 2068-2075. https://doi.org/10.1016/j.soilbio.2008.01.015
Solon, K., P. Volcke, E. I., Spérandio, M., & Loosdrecht, M. C. M. van. (2019). Resource recovery and wastewater treatment modelling. *Environmental Science: Water Research & Technology*, *5*(4), 631-642. https://doi.org/10.1039/C8EW00765A
Song, J., Haider, S., Song, J., Zhang, D., Chang, S., Bai, J., Hao, J., Yang, G., Ren, G., Han, X., Wang, X., Ren, C., Feng, Y., & Wang, X. (2025). Regulation of soil microbial nitrogen limitation by soybean rhizosphere diazotrophs under long-term no-till mulching. *Applied Soil Ecology*, *206*, 105873. https://doi.org/10.1016/j.apsoil.2025.105873
Soto-Gómez, D., & Pérez-Rodríguez, P. (2022). Sustainable agriculture through perennial grains: Wheat, rice, maize, and other species. A review. *Agriculture, Ecosystems & Environment*, *325*, 107747. https://doi.org/10.1016/j.agee.2021.107747
Stahl, D. A., & de la Torre, J. R. (2012). Physiology and diversity of ammonia-oxidizing archaea. *Annual Review of Microbiology*, *66*, 83-101. https://doi.org/10.1146/annurev-micro-092611-150128
Starke, R., Jehmlich, N., & Bastida, F. (2019a). Using proteins to study how microbes contribute to soil ecosystem services: The current state and future perspectives of soil metaproteomics. *Journal of Proteomics*, *198*, 50-58. https://doi.org/10.1016/j.jprot.2018.11.011
Starke, R., Jehmlich, N., & Bastida, F. (2019b). Using proteins to study how microbes contribute to soil ecosystem services: The current state and future perspectives of soil metaproteomics. *Journal of Proteomics*, *198*, 50-58. https://doi.org/10.1016/j.jprot.2018.11.011
Stein, L. Y., & Klotz, M. G. (2016). The nitrogen cycle. *Current Biology: CB*, *26*(3), R94-98. https://doi.org/10.1016/j.cub.2015.12.021
Sun, Y., Wang, M., Mur, L. A. J., Shen, Q., & Guo, S. (2020). Unravelling the Roles of Nitrogen Nutrition in Plant Disease Defences. *International Journal of Molecular Sciences*, *21*(2), Article 2. https://doi.org/10.3390/ijms21020572
Taiyun. (2017). *Taiyun/corrplot* [R]. https://github.com/taiyun/corrplot
Tao, K., Kelly, S., & Radutoiu, S. (2019). Microbial associations enabling nitrogen acquisition in plants. *Current Opinion in Microbiology*, *49*, 83-89. https://doi.org/10.1016/j.mib.2019.10.005
Tartaglia, M., Bastida, F., Sciarrillo, R., & Guarino, C. (2020). Soil Metaproteomics for the Study of the Relationships Between Microorganisms and Plants: A Review of Extraction Protocols and Ecological Insights. *International Journal of Molecular Sciences*, *21*(22), Article 22. https://doi.org/10.3390/ijms21228455
Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., & Natale, D. A. (2003). The COG database: An updated version includes eukaryotes. *BMC Bioinformatics*, *4*(1), 41. https://doi.org/10.1186/1471-2105-4-41
Taylor, B. N., Simms, E. L., & Komatsu, K. J. (2020). More Than a Functional Group: Diversity within the Legume–Rhizobia Mutualism and Its Relationship with Ecosystem Function. *Diversity*, *12*(2), Article 2. https://doi.org/10.3390/d12020050
Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics—A guide from sampling to data analysis. *Microbial Informatics and Experimentation*, *2*(1), 3. https://doi.org/10.1186/2042-5783-2-3
Tláskal, V., & Baldrian, P. (2021). Deadwood-Inhabiting Bacteria Show Adaptations to Changing Carbon and Nitrogen Availability During Decomposition. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/fmicb.2021.685303
Tláskal, V., Brabcová, V., Větrovský, T., Jomura, M., López-Mondéjar, R., Oliveira Monteiro, L. M., Saraiva, J. P., Human, Z. R., Cajthaml, T., Nunes da Rocha, U., & Baldrian, P. (2021). Complementary

Roles of Wood-Inhabiting Fungi and Bacteria Facilitate Deadwood Decomposition. *mSystems*, *6*(1), 10.1128/msystems.01078-20. https://doi.org/10.1128/msystems.01078-20
Torres, M. J., Simon, J., Rowley, G., Bedmar, E. J., Richardson, D. J., Gates, A. J., & Delgado, M. J. (2016). Chapter Seven - Nitrous Oxide Metabolism in Nitrate-Reducing Bacteria: Physiology and Regulatory Mechanisms. En R. K. Poole (Ed.), *Advances in Microbial Physiology* (Vol. 68, pp. 353-432). Academic Press. https://doi.org/10.1016/bs.ampbs.2016.02.007
Tourna, M., Stieglmeier, M., Spang, A., Könneke, M., Schintlmeister, A., Urich, T., Engel, M., Schloter, M., Wagner, M., Richter, A., & Schleper, C. (2011). Nitrososphaera viennensis, an ammonia oxidizing archaeon from soil. *Proceedings of the National Academy of Sciences*, *108*(20), 8420-8425. https://doi.org/10.1073/pnas.1013488108
Turbé, A., Toni, A. de, Benito, P., Lavelle, P., Lavelle, P., Camacho, N. R., Putten, W. H. van D., Labouze, E., & Mudgal, S. (2010). *Soil biodiversity: Functions, threats and tools for policy makers*. https://hal-bioemco.ccsd.cnrs.fr/bioemco-00560420
Urbanová, M., Šnajdr, J., & Baldrian, P. (2015). Composition of fungal and bacterial communities in forest litter and soil is largely determined by dominant trees. *Soil Biology and Biochemistry*, *84*, 53-64. https://doi.org/10.1016/j.soilbio.2015.02.011
Uritskiy, G. V., DiRuggiero, J., & Taylor, J. (2018). MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*, *6*(1), 158. https://doi.org/10.1186/s40168-018-0541-1
Usman, K., Khan, S., Ghulam, S., Khan, M. U., Khan, N., Khan, M. A., & Khalil, S. K. (2012). *Sewage Sludge: An Important Biological Resource for Sustainable Agriculture and Its Environmental Implications*. *2012*. https://doi.org/10.4236/ajps.2012.312209
Vailati-Riboni, M., Palombo, V., & Loor, J. J. (2017). What Are Omics Sciences? En B. N. Ametaj (Ed.), *Periparturient Diseases of Dairy Cows: A Systems Biology Approach* (pp. 1-7). Springer International Publishing. https://doi.org/10.1007/978-3-319-43033-1_1
Valenzuela, C., Leiva, D., Carú, M., & Orlando, J. (2022). Prediction of the Metabolic Functions of Nitrogen, Phosphorus, and Sulfur Cycling Bacteria Associated with the Lichen Peltigera frigida. *Microbiology*, *91*(5), 604-610. https://doi.org/10.1134/S0026261721102117
Van Emon, J. M. (2016). The Omics Revolution in Agricultural Research. *Journal of Agricultural and Food Chemistry*, *64*(1), 36-44. https://doi.org/10.1021/acs.jafc.5b04515
Verchot, L. V. (2010). Impacts of Forest Conversion to Agriculture on Microbial Communities and Microbial Function. En P. Dion (Ed.), *Soil Biology and Agriculture in the Tropics* (pp. 45-63). Springer. https://doi.org/10.1007/978-3-642-05076-3_3
Větrovský, T., & Baldrian, P. (2013). The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLOS ONE*, *8*(2), e57923. https://doi.org/10.1371/journal.pone.0057923
Voříšková, J., Brabcová, V., Cajthaml, T., & Baldrian, P. (2014). Seasonal dynamics of fungal communities in a temperate oak forest soil. *New Phytologist*, *201*(1), 269-278. https://doi.org/10.1111/nph.12481
Walia, A., Guleria, S., Chauhan, A., & Mehta, P. (2017). Endophytic Bacteria: Role in Phosphate Solubilization. En D. K. Maheshwari & K. Annapurna (Eds.), *Endophytes: Crop Productivity and Protection: Volume 2* (pp. 61-93). Springer International Publishing. https://doi.org/10.1007/978-3-319-66544-3_4
Wallace, R. J., Snelling, T. J., McCartney, C. A., Tapio, I., & Strozzi, F. (2017). Application of meta-omics techniques to understand greenhouse gas emissions originating from ruminal metabolism. *Genetics Selection Evolution*, *49*(1), 9. https://doi.org/10.1186/s12711-017-0285-6
Wan, W., Li, X., Han, S., Wang, L., Luo, X., Chen, W., & Huang, Q. (2020). Soil aggregate fractionation and phosphorus fraction driven by long-term fertilization regimes affect the abundance and composition of P-cycling-related bacteria. *Soil and Tillage Research*, *196*, 104475. https://doi.org/10.1016/j.still.2019.104475
Wang, C., Dong, D., Wang, H., Müller, K., Qin, Y., Wang, H., & Wu, W. (2016). Metagenomic analysis of microbial consortia enriched from compost: New insights into the role of Actinobacteria in lignocellulose decomposition. *Biotechnology for Biofuels*, *9*(1), 22. https://doi.org/10.1186/s13068-016-0440-2
Wang, C., Jiang, Y., Shao, Y., Chen, Z., Gao, Y., Liang, J., Gao, J., Fang, F., & Guo, J. (2024). The influence and risk assessment of multiple pollutants on the bacterial and archaeal communities in

agricultural lands with different climates and soil properties. *Applied Soil Ecology*, *193*, 105130. https://doi.org/10.1016/j.apsoil.2023.105130
Wang, C., & Kuzyakov, Y. (2024). Mechanisms and implications of bacterial–fungal competition for soil resources. *The ISME Journal*, *18*(1), wrae073. https://doi.org/10.1093/ismejo/wrae073
Wang, F., Liang, X., Ding, F., Ren, L., Liang, M., An, T., Li, S., Wang, J., & Liu, L. (2022). The active functional microbes contribute differently to soil nitrification and denitrification potential under long-term fertilizer regimes in North-East China. *Frontiers in Microbiology*, *13*. https://doi.org/10.3389/fmicb.2022.1021080
Wang, J.-T., Zhang, Y.-B., Xiao, Q., & Zhang, L.-M. (2022). Archaea is more important than bacteria in driving soil stoichiometry in phosphorus deficient habitats. *Science of The Total Environment*, *827*, 154417. https://doi.org/10.1016/j.scitotenv.2022.154417
Wang, L., Ye, C., Gao, B., Wang, X., Li, Y., Ding, K., Li, H., Ren, K., Chen, S., Wang, W., & Ye, X. (2023). Applying struvite as a N-fertilizer to mitigate N2O emissions in agriculture: Feasibility and mechanism. *Journal of Environmental Management*, *330*, 117143. https://doi.org/10.1016/j.jenvman.2022.117143
Wang, Y., Ji, H., & Gao, C. (2016). Differential responses of soil bacterial taxa to long-term P, N, and organic manure application. *Journal of Soils and Sediments*, *16*(3), 1046-1058. https://doi.org/10.1007/s11368-015-1320-2
Wickham, H. (2016). Programming with ggplot2. En H. Wickham (Ed.), *Ggplot2: Elegant Graphics for Data Analysis* (pp. 241-253). Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4_12
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686
Wieczorek, A. S., Schmidt, O., Chatzinotas, A., von Bergen, M., Gorissen, A., & Kolb, S. (2019). Ecological Functions of Agricultural Soil Bacteria and Microeukaryotes in Chitin Degradation: A Case Study. *Frontiers in Microbiology*, *10*. https://doi.org/10.3389/fmicb.2019.01293
Wolińska, A., Kuźniar, A., Zielenkiewicz, U., Banach, A., Izak, D., Stępniewska, Z., & Błaszczyk, M. (2017). Metagenomic Analysis of Some Potential Nitrogen-Fixing Bacteria in Arable Soils at Different Formation Processes. *Microbial Ecology*, *73*(1), 162-176. https://doi.org/10.1007/s00248-016-0837-2
Wu, P., Chen, J., Garlapati, V. K., Zhang, X., Wani Victor Jenario, F., Li, X., Liu, W., Chen, C., Aminabhavi, T. M., & Zhang, X. (2022). Novel insights into Anammox-based processes: A critical review. *Chemical Engineering Journal*, *444*, 136534. https://doi.org/10.1016/j.cej.2022.136534
Wu, X., Rensing, C., Han, D., Xiao, K.-Q., Dai, Y., Tang, Z., Liesack, W., Peng, J., Cui, Z., & Zhang, F. (2022). Genome-Resolved Metagenomics Reveals Distinct Phosphorus Acquisition Strategies between Soil Microbiomes. *mSystems*, *7*(1), e01107-21. https://doi.org/10.1128/msystems.01107-21
Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A., & Singer, S. W. (2014). MaxBin: An automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, *2*(1), 26. https://doi.org/10.1186/2049-2618-2-26
Xu, X., Yu, Z., He, L., Cao, X., Chen, N., & Song, X. (2020). Metabolic analyses by metatranscriptomics highlight plasticity in phosphorus acquisition during monospecific and multispecies algal blooms. *Hydrobiologia*, *847*(4), 1071-1085. https://doi.org/10.1007/s10750-019-04169-x
Yadav, A. N., Sharma, D., Gulati, S., Singh, S., Dey, R., Pal, K. K., Kaushik, R., & Saxena, A. K. (2015). Haloarchaea Endowed with Phosphorus Solubilization Attribute Implicated in Phosphorus Cycle. *Scientific Reports*, *5*(1), 12293. https://doi.org/10.1038/srep12293
Yadav, R. K., Bragalini, C., Fraissinet-Tachet, L., Marmeisse, R., & Luis, P. (2016). Metatranscriptomics of Soil Eukaryotic Communities. En F. Martin & S. Uroz (Eds.), *Microbial Environmental Genomics (MEG)* (pp. 273-287). Springer. https://doi.org/10.1007/978-1-4939-3369-3_16
Yang, C., Chowdhury, D., Zhang, Z., Cheung, W. K., Lu, A., Bian, Z., & Zhang, L. (2021). A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Computational and Structural Biotechnology Journal*, *19*, 6301-6314. https://doi.org/10.1016/j.csbj.2021.11.028

Yang, Q., Peng, J., Ni, S., Zhang, C., Wang, J., & Cai, C. (2024). Erosion and deposition significantly affect the microbial diversity, co-occurrence network, and multifunctionality in agricultural soils of Northeast China. *Journal of Soils and Sediments*, *24*(2), 888-900. https://doi.org/10.1007/s11368-023-03687-5
Yang, T., Siddique, K. H. M., & Liu, K. (2020). Cropping systems in agriculture and their impact on soil health-A review. *Global Ecology and Conservation*, *23*, e01118. https://doi.org/10.1016/j.gecco.2020.e01118
Yang, X., Wang, Y., Wang, X., Niu, T., Abid, A. A., Aioub, A. A. A., & Zhang, Q. (2024). Contrasting fertilization response of soil phosphorus forms and functional bacteria in two newly reclaimed vegetable soils. *Science of The Total Environment*, *912*, 169479. https://doi.org/10.1016/j.scitotenv.2023.169479
Yang, X., Zhang, C., Ma, X., Liu, Q., An, J., Xu, S., Xie, X., & Geng, J. (2021). Combining Organic Fertilizer With Controlled-Release Urea to Reduce Nitrogen Leaching and Promote Wheat Yields. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.802137
Yang, Y., Liu, H., & Lv, J. (2022). Response of N2O emission and denitrification genes to different inorganic and organic amendments. *Scientific Reports*, *12*(1), 3940. https://doi.org/10.1038/s41598-022-07753-9
Yao, R., Yang, J., Wang, X., Xie, W., Zheng, F., Li, H., Tang, C., & Zhu, H. (2021). Response of soil characteristics and bacterial communities to nitrogen fertilization gradients in a coastal salt-affected agroecosystem. *Land Degradation & Development*, *32*(1), 338-353. https://doi.org/10.1002/ldr.3705
Ye, L., Li, H., Mortimer, P. E., Xu, J., Gui, H., Karunarathna, S. C., Kumar, A., Hyde, K. D., & Shi, L. (2019). Substrate Preference Determines Macrofungal Biogeography in the Greater Mekong Sub-Region. *Forests*, *10*(10), Article 10. https://doi.org/10.3390/f10100824
Yep, B., Gale, N. V., & Zheng, Y. (2020). Comparing hydroponic and aquaponic rootzones on the growth of two drug-type *Cannabis sativa* L. cultivars during the flowering stage. *Industrial Crops and Products*, *157*, 112881. https://doi.org/10.1016/j.indcrop.2020.112881
Yesigat, A., Worku, A., Mekonnen, A., Bae, W., Feyisa, G. L., Gatew, S., Han, J.-L., Liu, W., Wang, A., & Guadie, A. (2022). Phosphorus recovery as K-struvite from a waste stream: A review of influencing factors, advantages, disadvantages and challenges. *Environmental Research*, *214*, 114086. https://doi.org/10.1016/j.envres.2022.114086
Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28-36. https://doi.org/10.1111/2041-210X.12628
Yu, K., & Zhang, T. (2012). Metagenomic and Metatranscriptomic Analysis of Microbial Community Structure and Gene Expression of Activated Sludge. *PLOS ONE*, *7*(5), e38183. https://doi.org/10.1371/journal.pone.0038183
Yuan, Y., Chen, L., Wang, J., Liu, Y., Ren, C., Guo, Y., Wang, J., Wang, N., Zhao, F., & Wang, W. (2023). Different Response of Plant- and Microbial-Derived Carbon Decomposition Potential between Alpine Steppes and Meadows on the Tibetan Plateau. *Forests*, *14*(8), Article 8. https://doi.org/10.3390/f14081580
Zaim, S., & Bekkar, A. A. (2023). Advances in research on the use of Brevundimonas spp. To improve crop and soil fertility and for soil bioremediation. *Algerian Journal of Biosciences*, *4*(1), 045-051. https://doi.org/10.57056/ajb.v4i1.109
Zang, X., Liu, M., Fan, Y., Xu, J., Xu, X., & Li, H. (2018). The structural and functional contributions of β-glucosidase-producing microbial communities to cellulose degradation in composting. *Biotechnology for Biofuels*, *11*(1), 51. https://doi.org/10.1186/s13068-018-1045-8
Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., Busk, P. K., Xu, Y., & Yin, Y. (2018). dbCAN2: A meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*, *46*(W1), W95-W101. https://doi.org/10.1093/nar/gky418
Zhang, X., Davidson, E. A., Mauzerall, D. L., Searchinger, T. D., Dumas, P., & Shen, Y. (2015). Managing nitrogen for sustainable development. *Nature*, *528*(7580), 51-59. https://doi.org/10.1038/nature15743
Zhang, Z., Shi, Z., Yang, J., Hao, B., Hao, L., Diao, F., Wang, L., Bao, Z., & Guo, W. (2021). A new strategy for evaluating the improvement effectiveness of degraded soil based on the synergy and diversity of microbial ecological function. *Ecological Indicators*, *120*, 106917. https://doi.org/10.1016/j.ecolind.2020.106917
Zhao, B., Jia, X., Yu, N., Murray, J. D., Yi, K., & Wang, E. (2024). Microbe-dependent and independent

nitrogen and phosphate acquisition and regulation in plants. *New Phytologist*, *242*(4), 1507-1522. https://doi.org/10.1111/nph.19263
Zhao, J., Huang, L., Chakrabarti, S., Cooper, J., Choi, E., Ganan, C., Tolchinsky, B., Triplett, E. W., Daroub, S. H., & Martens-Habbena, W. (2023). Nitrogen and phosphorous acquisition strategies drive coexistence patterns among archaeal lineages in soil. *The ISME Journal*, *17*(11), 1839-1850. https://doi.org/10.1038/s41396-023-01493-y
Zhao, X., Tian, P., Zhang, W., Wang, Q., Guo, P., & Wang, Q. (2024). Nitrogen deposition caused higher increases in plant-derived organic carbon than microbial-derived organic carbon in forest soils. *Science of The Total Environment*, *925*, 171752. https://doi.org/10.1016/j.scitotenv.2024.171752
Zhu, L., Li, W., Huang, C., Tian, Y., & Xi, B. (2024). Functional redundancy is the key mechanism used by microorganisms for nitrogen and sulfur metabolism during manure composting. *Science of The Total Environment*, *912*, 169389. https://doi.org/10.1016/j.scitotenv.2023.169389
Žifčáková, L. (2017). *Molecular biology and ecology of microbial decomposition of plant-derived biopolymers in forest ecosystems*. https://dspace.cuni.cz/handle/20.500.11956/16935
Žifčáková, L., Větrovský, T., Howe, A., & Baldrian, P. (2016). Microbial activity in forest soil reflects the changes in ecosystem properties between summer and winter. *Environmental Microbiology*, *18*(1), 288-301. https://doi.org/10.1111/1462-2920.13026
Žifčáková, L., Větrovský, T., Lombard, V., Henrissat, B., Howe, A., & Baldrian, P. (2017). Feed in summer, rest in winter: Microbial carbon utilization in forest topsoil. *Microbiome*, *5*(1), 122. https://doi.org/10.1186/s40168-017-0340-0

ANNEXES

ANNEXES

Annex 1: Complete pipeline of the metagenomic analysis carried out in Chapters 1 and 2.

```
#########################
## METAGENOMIC ANALYSIS ##
#########################

# Download function gdrive_download:
function gdrive_download () {
 CONFIRM=$(wget --quiet --save-cookies /tmp/cookies.txt
--keep-session-cookies --no-check-certificate
"https://docs.google.com/uc?export=download&id=$1" -O- | sed -rn
's/.*confirm=([0-9A-Za-z_]+).*/\1\n/p')
 wget --load-cookies /tmp/cookies.txt
"https://docs.google.com/uc?export=download&confirm=$CONFIRM&id=$1" -O $2
 rm -rf /tmp/cookies.txt
}

# Installation of khmer:
sudo yum install -y python3-devel gcc-c++ make
conda create --name khmerEnv python=3.6

# Open the terminal and activate the conda environment:
conda activate base

# Create a new conda environment for Khmer:
conda create --name khmerEnv

# Activate the conda environment
conda install -c bioconda khmer

#######
## 1 ##
#######

######################
## QUALITY ANALYSIS ##
######################

# A FastQC analysis is used to assess the quality of genomic sequencing data,
such as those generated by platforms like Illumina. It evaluates base
quality,
# base composition, the presence of adapters, sequence duplication levels,
read length distribution, and the overrepresentation of sequences.

# Installation of FASTQC
conda install -c bioconda fastqc

# Quality analysis
conda activate khmerEnv
fastqc *.gz -o ~/metagenomic_analysis/1_FASTQC_RESULTS
conda deactivate

#######
```

```
## 2 ##
#######

################
## INTERLEAVE ##
################

# Interleaving in metagenomics is the process of combining two paired-end
read files into a single file. In this interleaved file,
# the forward and reverse reads of each pair are arranged alternately (i.e.,
the forward read of the pair is followed by its corresponding reverse read).
# Interleaving is primarily used to facilitate data processing by
bioinformatics tools that require paired-end reads to be stored in a single
file.

# Unzip the fastq.gz in fastq
for i in {1..32}
do
    gunzip -c ${i}_R1_001.fastq.gz > ${i}_R1_001.fastq
    gunzip -c ${i}_R2_001.fastq.gz > ${i}_R2_001.fastq
done

# Interleaved
for file in *_R1_001.fastq
do
   sample=${file%%_R1_001.fastq}
   echo “interleave-reads.py ${sample}_R1_001.fastq ${sample}_R2_001.fastq -
or ${sample}.pe.fq”
done > interleave.sh

cat interleave.sh | parallel

# Remove unnecessary files and organize them
rm -rf *.fastq
cd ...
mkdir 2_INTERLEAVED
cd 0_SAMPLES
mv *.pe.fq ../2_INTERLEAVED
cd ../2_INTERLEAVED

#######
## 3 ##
#######

######################
## QUALITY FILTERING ##
######################

# The purpose of this step is to remove low-quality reads from sequencing
data,
# thereby improving the reliability of downstream analyses such as assemblies
or annotations.
# This ensures that the reads used meet a minimum quality standard, reducing
errors and artifacts in the final results.

# The filtering process employs the following parameters:
```

```
# -Q33: Specifies that the quality scores are encoded in the Phred+33 format,
commonly used in Illumina sequencing platforms.
# -q 30: Filters out reads where the average base quality is below 30,
corresponding to high-quality bases.
# -p 50: Retains only reads in which at least 50% of the bases meet or exceed
the specified quality threshold.

for file in *.pe.fq
do
  newfile=${file%%.pe.fq}
  echo "fastq_quality_filter -i ${file} -Q33 -q 30 -p 50 -o
${newfile}.pe.qc.fq"
done > qual_filter.sh

cat qual_filter.sh | parallel

#######
## 4 ##
#######

############################
## REMOVE SHORT SEQUENCES ##
############################

# Download function gdrive_download:
function gdrive_download () {
 CONFIRM=$(wget --quiet --save-cookies /tmp/cookies.txt --keep-session-
cookies --no-check-certificate
"https://docs.google.com/uc?export=download&id=$1" -O- | sed -rn
's/.*confirm=([0-9A-Za-z_]+).*/\1\n/p')
 wget --load-cookies /tmp/cookies.txt
"https://docs.google.com/uc?export=download&confirm=$CONFIRM&id=$1" -O $2
 rm -rf /tmp/cookies.txt
}

# Downlowad filter_fastq_by_length.py script
gdrive_download 1w-OyfdEuMi38utz4cN9g_ng-S9kNeOj9 filter_fastq_by_length.py

# Remove short sequences
for file in *pe.qc.fq
do
  echo "python2.7 filter_fastq_by_length.py ${file} ${file}.cut 50"
done > remove_short.sh

cat remove_short.sh | parallel

#######
## 5 ##
#######

#########################################################
## EXTRACT PAIRED ENDS, RENAME FILES AND MERGE FILES ##
#########################################################

# In this step, paired-end sequence files are processed after quality
cleaning. It includes three main steps: extracting paired reads,
```

```
# removing unnecessary files, and renaming and organizing the output files to
facilitate subsequent analysis.

# Extracting paired-ends
for file in *.pe.qc.fq.cut
do
   echo "extract-paired-reads.py ${file}"
done > extract_command.sh

cat extract_command.sh | parallel

# Remove unnecessary files
rm -rf *.tr.qc.fq.cut

# Rename files and merging pe and se files
for file in *.pe
do
   sample=${file%%.pe.qc.fq.cut.pe}
   mv ${file} ${sample}.pe.qc.fq
done

for file in *.se
do
   sample=${file%%.pe.qc.fq.cut.se}
   mv ${file} ${sample}.se.qc.fq
done

#######
## 6 ##
#######

################################
## PREPARATION FOR ASSEMBLY ##
################################

# In this step, paired-end sequencing data are prepared for assembly. The
script split-paired-reads.py is used to split each paired-end file
(*.pe.qc.fq)
# into two separate files: one containing the forward reads (R1) and the
other containing the reverse reads (R2). This step is necessary for assembly
tools
# such as MEGAHIT, which require paired-end reads to be provided in
individual files.

for file in *.pe.qc.fq
do
   echo "split-paired-reads.py ${file}"
done > split_command.sh

cat split_command.sh | parallel

#######
## 7 ##
#######

##############
## ASSEMBLY ##
```

```
##############

# The assembly step aims to combine sequencing reads (forward, reverse, and
unpaired) to reconstruct complete or contiguous genomic sequences
# from smaller fragments (reads).

mkdir 3_FOR_ASSEMBLY

# Create a file with all forward sequences
cat *.1 > 3_FOR_ASSEMBLY/all.pe.qc.fq.1

# Create a file with all reverse sequences
cat *.2 > 3_FOR_ASSEMBLY/all.pe.qc.fq.2

# Create a file with all unpaired sequences
cat *.se.qc.fq > 3_FOR_ASSEMBLY/all.se.qc.fq

# Assembly using MEGAHIT
megahit -m 0.75 -t 120 -1 all.pe.qc.fq.1 -2 all.pe.qc.fq.2 -r all.se.qc.fq
-o all.Megahit.assembly

#######
## 8 ##
#######

############################
## ASSEMBLY QUALITY CHECK ##
############################

# In this step, MetaQUAST is used, a tool designed to assess the quality of
genomic assemblies.
# The command takes the final contigs generated by the MEGAHIT assembly
(final.contigs.fa) as input. Several evaluation options are specified:
# --rna-finding identifies potential RNA regions.
# --conserved-genes-finding searches for conserved genes
# --max-ref-number 20 limits the maximum number of references for comparison.

# This analysis allows for the verification of the assembly's quality and
integrity.

conda activate quast

# Assembly quality check
metaquast
/mnt/DATA/belen/3_FOR_ASSEMBLY/all.Megahit.assembly/final.contigs.fa -t 120 -
-rna-finding --conserved-genes-finding --max-ref-number 20

# This is the quality report of the samples:

# contigs  5259264
# contigs (>= 0 bp)  12175284
# contigs (>= 1000 bp)     1377527
# contigs (>= 5000 bp)     46151
# contigs (>= 10000 bp)    10406
# contigs (>= 25000 bp)    1303
# contigs (>= 50000 bp)    229
# Largest contig213010
```

```
# Total length  5241019665
# Total length (>= 0 bp)   7719201062
# Total length (>= 1000 bp)     2613105109
# Total length (>= 5000 bp)     417454729
# Total length (>= 10000 bp)    180508676
# Total length (>= 25000 bp)    52451342
# Total length (>= 50000 bp)    16828383
# N50 997
# N90 566
# auN 2360.7
# L50 1384946
# L90 4272456
# GC (%)   ...

#######
## 9 ##
#######

########################
## GENECALLING - FGS ##
########################

# FragGeneScan is a program used to predict genes in DNA sequences.
# The main objective of FragGeneScan is to identify coding sequences (CDS) in
DNA sequences that may be fragmented or incomplete.

mkdir 4_FGS
cd 4_FGS

# Link creation
ln -s
/home/kdanielmorais/bioinformatics/tools/fraggenescan/FragGeneScan1.31/train/
./

# FragGeneScan
FragGeneScan -s
~/metagenomic_analysis/3_FOR_ASSEMBLY/all.Megahit.assembly/final.contigs.fa -
w 1 -o belen_MG_Megahit_genecalling_fgs -t complete -p 120

########
## 10 ##
########

#############
## MAPPING ##
#############

# Mapping is a key step in metagenomic analysis. It consists of mapping the
DNA or RNA sequences obtained from the sample to a reference database,
# usually a database of known sequences.

# In this analysis, sequencing reads were mapped against a reference
assembly. The process involved several steps,
# starting with the preparation of the reference and culminating in the
generation of alignment statistics.
# First, an index was created for the reference contigs file
(final.contigs.fa) using Bowtie2, optimizing the alignment process.
```

```
# Next, paired-end (.pe.qc.fq) and unpaired (.se.qc.fq) reads were combined
into a single file for each sample to facilitate joint mapping.
# The combined reads were then aligned against the reference using Bowtie2,
producing alignment files in SAM format,
# which were subsequently converted to compressed BAM format using Samtools.
# Mapped and unmapped reads were counted to assess the quality and efficiency
of the alignment. BAM files were sorted by reference position,
# indexed, and finally, statistics on read distribution across contigs were
generated.

REF=final.contigs.fa
reference=${REF%%.fa}
echo "reference is" ${reference}
mkdir ${reference}_build
bowtie2-build
~/metagenomic_analysis/3_FOR_ASSEMBLY/all.Megahit.assembly/${REF}
${reference}_build/${reference}.build

conda activate khmerEnv

for file in *.pe.qc.fq
do
 sample=${file%%.pe.qc.fq}
 cat ${sample}.pe.qc.fq ${sample}.se.qc.fq >
~/metagenomic_analysis/5_SAMPLE_MAPPING/${sample}.all.qc.fq
 echo "processing ${sample}...}"

 bowtie2 -p 70 -x
~/metagenomic_analysis/5_SAMPLE_MAPPING/final.contigs_build/final.contigs.bui
ld -q ~/metagenomic_analysis/5_SAMPLE_MAPPING/${sample}.all.qc.fq -S
~/metagenomic_analysis/5_SAMPLE_MAPPING/${sample}.sam
 echo "sam file is done..."

 rm -rf ~/metagenomic_analysis/5_SAMPLE_MAPPING/${sample}.all.qc.fq

 samtools view -Sb ~/metagenomic_analysis/5_SAMPLE_MAPPING/${sample}.sam >
~/metagenomic_analysis/5_SAMPLE_MAPPING/${sample}.bam
 echo "bam file is done..."

 rm -rf ~/metagenomic_analysis/5_SAMPLE_MAPPING/${sample}.sam

 samtools view -c -f 4 ~/metagenomic_analysis/5_SAMPLE_MAPPING/${sample}.bam
> ~/metagenomic_analysis/5_SAMPLE_MAPPING/${sample}.reads-unmapped.count.txt
 echo "unmapped reads info done..."

 samtools view -c -F 4 ~/metagenomic_analysis/5_SAMPLE_MAPPING/${sample}.bam
> ~/metagenomic_analysis/5_SAMPLE_MAPPING/${sample}.reads-mapped.count.txt
 echo "mapped reads info done..."

 samtools sort -o
~/metagenomic_analysis/5_SAMPLE_MAPPING/${sample}.sorted.bam
~/metagenomic_analysis/5_SAMPLE_MAPPING/${sample}.bam
 echo "bam file was sorted..."

 rm -rf ~/metagenomic_analysis/5_SAMPLE_MAPPING/${sample}.bam

 samtools index ~/metagenomic_analysis/5_SAMPLE_MAPPING/${sample}.sorted.bam
```

```
 echo "soerted bam file was indexed..."

 samtools idxstats
~/metagenomic_analysis/5_SAMPLE_MAPPING/${sample}.sorted.bam >
~/metagenomic_analysis/5_SAMPLE_MAPPING/${sample}.reads.by.contigs.txt

 echo "${sample} is done..."
done

# Download count-up-mapped-from-results-txt-with-ctg-length.py script
gdrive_download 1HDB2EF-pq-EJxQxsI1uv1-iVjTl6tAlo count-up-mapped-from-
results-txt-with-ctg-length.py

python2.7 count-up-mapped-from-results-txt-with-ctg-length.py
*.reads.by.contigs.txt

# Validate the consistency between the assembled contigs and the data
generated from the mapping
wc -l summary-count-mapped.tsv
12175286 summary-count-mapped.tsv

grep '>' /mnt/DATA/belen/4_FOR_ASSEMBLY/all.Megahit.assembly/final.contigs.fa
| wc -l
12175284

# Coverage is a key metric in genomics, as it indicates how many times a
genomic region has been sequenced,
# providing insights into the reliability of the assembly and the relative
abundance of the contigs.
# The purpose of this step is to generate a coverage file that associates
each assembled contig with its average coverage.

# Download function gdrive_download:
function gdrive_download () {
 CONFIRM=$(wget --quiet --save-cookies /tmp/cookies.txt --keep-session-
cookies --no-check-certificate
"https://docs.google.com/uc?export=download&id=$1" -O- | sed -rn
's/.*confirm=([0-9A-Za-z_]+).*/\1\n/p')
 wget --load-cookies /tmp/cookies.txt
"https://docs.google.com/uc?export=download&confirm=$CONFIRM&id=$1" -O $2
 rm -rf /tmp/cookies.txt
}

# Download get_assembly_coverage.py script
gdrive_download 1S2AQHd2YIjnxZz2kIa2avo1RSAuj-pWT get_assembly_coverage.py

# Obtain coverage of our data
python get_assembly_coverage.py summary-count-mapped.tsv 151
belen_MG_Megahit_assembly_DN_coverage.txt

########
## 11 ##
########

#######################################
## NORMALISE MAPPING TABLE PER BASE ##
#######################################
```

```
# In this step, the aim is to normalize the mapping table to adjust coverage based on the length of sequencing reads and contigs.
# The script normalize-mapping-table-by-read-length-and-ctg-length.py takes as input the mapping count file (summary-count-mapped.tsv) and the average read length
# (in this case, 151 bases). It generates an output file (TABLE_normalised.txt) where the mapping values are adjusted to provide a more accurate comparative
# measure of relative coverage, regardless of differences in contig or read lengths.
# This procedure is essential to correct potential biases arising from variations in lengths and allows for fair comparisons between different contigs or
# genomic regions.

# Download normalize-mapping-table-by-read-length-and-ctg-length.py script
gdrive_download 1w0bfttjXFZ64NHD8bP7UDaQcS1yd20qR normalize-mapping-table-by-read-length-and-ctg-length.py

python2.7 normalize-mapping-table-by-read-length-and-ctg-length.py summary-count-mapped.tsv 151 TABLE_normalised.txt

# In this step, an additional normalization is performed on the previously normalized table to adjust coverage values based on a predefined scale by columns.
# The script normalize_table_by_columns.py takes the previously generated file (TABLE_normalised.txt) as input, selects a specific column (in this case, column 2),
# and applies a normalization factor (1,000,000) to scale the values per sample. The output is saved in a file named TABLE_normalised_per_sample.txt.

# Download normalize_table_by_columns.py script
gdrive_download 1c_fD520xtrCNlUIq9VqqqSvY2OryMXTU normalize_table_by_columns.py

python2.7 normalize_table_by_columns.py TABLE_normalised.txt 2 1000000 TABLE_normalised_per_sample.txt

########
## 12 ##
########

################
## ANNOTATION ##
################

# Functional and taxonomic annotation are processes used to characterize genetic sequences by assigning biological information and classification.
# - Functional annotation involves identifying the roles or functions of genes and proteins, such as their involvement in specific pathways, cellular processes,
# or molecular interactions.
# - Taxonomic annotation, on the other hand, assigns sequences to their corresponding organisms or taxonomic groups,
# providing insights into the evolutionary and ecological context of the data.
```

```
# Together, these annotations allow researchers to understand both the
biological role and the origin of the sequences,
# which is critical in fields such as genomics, metagenomics, and molecular
biology.

# In this step, gene annotation tasks are performed by integrating alignment
results with fungal protein and NCBI databases.
# First, sample information and genomic sequence data are prepared and
organized.
# Then, these sequences are aligned with fungal proteins and NCBI proteins to
obtain the best matches.
# Subsequently, taxonomic information is added to the alignment results
through the download and processing of taxonomy files.
# The results are formatted, taxonomy tables are combined, and the best
matches are selected based on bit score among the annotations.
# Finally, a table is generated containing taxonomic data and KOG functions
of the annotated proteins, completing the annotation and classification
process.

cd ..
mkdir 7_ANNOTATION
cp ./6_NORMALISE_MAPPING/TABLE_normalised_per_sample.txt ./7_ANNOTATION/
cd ./7_ANNOTATION

# Download contig_mapping_to_genecall_mapping.py script
gdrive_download 1DakO7roc9C2GJ-SkZuy8AZKTV3QHan14
contig_mapping_to_genecall_mapping.py

python2.7 contig_mapping_to_genecall_mapping.py
~/metagenomic_analysis/4_FGS/belen_MG_Megahit_genecalling_fgs.faa
TABLE_normalised_per_sample.txt

# Add “#” to the name of the samples:
head -1 TABLE_normalised_per_sample.txt_genecall.txt| awk -F'\t' '{printf
$1"\t"$2 ;for(i=3; i<=NF; ++i) printf "\t%s", "#"$i }' |  awk -F '\t' '{print
$0}' > header.txt

tail -n +2  TABLE_normalised_per_sample.txt_genecall.txt > table.txt

cat header.txt table.txt > TABLE_NORM_SAMPLES_GENECALL.txt

#####################################
## JGI FUNGAL PROTEINS - BIOCEV PC ##
#####################################

cd /mnt/DATA/DATABASES/FUNGAL_PROTEINS_JGI/
cp JGI_FUNGAL_PROTEINS_ANNOTATED_20210312.faa.zip
/mnt/DATA/belen/7_ANNOTATION/
cd /mnt/DATA/belen/7_ANNOTATION/
unzip JGI_FUNGAL_PROTEINS_ANNOTATED_20210312.faa.zip

diamond blastp -d
~/metagenomic_analysis/7_ANNOTATION/JGI_FUNGAL_PROTEINS_ANNOTATED_20210312.fa
a -q ~/metagenomic_analysis/4_FGS/belen_MG_Megahit_genecalling_fgs.faa -e 1E-
5 -o genecalling_JGI_FUN_20210312.txt -f 6 -p 120 -b12 -c1

###########################################################
```

```
## Total time = 1406.2s                                          ##
## Reported 72585274 pairwise alignments, 72585274 HSPs  ##
## 4541138 queries aligned.                                      ##
#################################################################

export LANG=en_US.UTF-8
export LC_ALL=en_US.UTF-8

sort -t$'\t' -k1,1 -k12,12gr -k11,11g -k3,3gr
genecalling_JGI_FUN_20210312.txt | sort -u -k1,1 --merge >
genecalling_JGI_FUN_20210312_best.txt

# GENERA DEFINED
diamond blastp -d
/mnt/DATA/DATABASES/NCBI_nr_DIAMOND/NCBI_nr_20210225_diamond_GENERA -q
/mnt/DATA/belen/metagenomic_analysis/4_FGS/belen_MG_Megahit_genecalling_fgs.f
aa -e 1E-5 -o belen_MG_genecalling_NCBI_nr_PROTEINS_GENERA.txt -f 6 -p 120 -
b12 -c1

#################################################################
## Total time = 57412.9s                                        ##
## Reported 280564387 pairwise alignments, 280564387 HSPs. ##
## 12605946 queries aligned.                                     ##
#################################################################

export LANG=en_US.UTF-8
export LC_ALL=en_US.UTF-8
sort -t$'\t' -k1,1 -k12,12gr -k11,11g -k3,3gr
belen_MG_genecalling_NCBI_nr_PROTEINS_GENERA.txt | sort -u -k1,1 --merge >
genecalling_NCBI_nr_PROTEINS_best.txt

#  ADD TAXONOMY TO BLAST RESULTS

function gdrive_download () {
 CONFIRM=$(wget --quiet --save-cookies /tmp/cookies.txt --keep-session-
cookies --no-check-certificate
"https://docs.google.com/uc?export=download&id=$1" -O- | sed -rn
's/.*confirm=([0-9A-Za-z_]+).*/\1\n/p')
 wget --load-cookies /tmp/cookies.txt
"https://docs.google.com/uc?export=download&confirm=$CONFIRM&id=$1" -O $2
 rm -rf /tmp/cookies.txt
}

# Download jgi_abr_org_list.txt
gdrive_download 12c28kgIw4mPBIhQutNGladdAXwNLtvlR jgi_abr_org_list.txt

# Download replace_fungal_annot_by_taxname.py script
gdrive_download 1XBTtiC1JYl2rzeV7idN2WrveEZknmnQi
replace_fungal_annot_by_taxname.py

python2.7 replace_fungal_annot_by_taxname.py
genecalling_JGI_FUN_20210312_best.txt jgi_abr_org_list.txt
genecalling_JGI_FUNGAL_PROTEINS_best_reformate.txt

# FUNGAL
awk -F'\t' '{print $2}' genecalling_JGI_FUNGAL_PROTEINS_best_reformate.txt |
sort | uniq > FUNGAL_NAMES.txt
```

```
# NCBI
awk -F'\t' '{print $2}' genecalling_NCBI_nr_PROTEINS_best.txt | sort | uniq >
ALL_ACCESSIONS.txt

# Download get_taxonomy_offline.py script
gdrive_download 1o8KmSbwzOsjjeouK3dR0RNmWkMjdfFow get_taxonomy_offline.py

python2.7 get_taxonomy_offline.py ALL_ACCESSIONS.txt
/mnt/DATA/DATABASES/NCBI_nr_DIAMOND/ACC2TAXID_nr_current.txt
/mnt/DATA/DATABASES/NCBI_nr_DIAMOND/TAXONOMY_TAXID_ALL_fixed.txt
taxa_all_accessions.txt

#########################################################################
## accession list loaded...(4106547)                                   ##
## taxonomy loaded...(907158)                                          ##
## taxonomy retrieved... 4106547 vs acc (4106547) - should be equal!   ##
## Done :]                                                             ##
#########################################################################

# REFORMAT

# Download replace_acc_by_sp_from_taxonomy.py script
gdrive_download 1jQ3F3ZuA0sBJVy3eJiRhwAaxh31LKqSF
replace_acc_by_sp_from_taxonomy.py

python2.7 replace_acc_by_sp_from_taxonomy.py
genecalling_NCBI_nr_PROTEINS_best.txt taxa_all_accessions.txt
genecalling_NCBI_nr_PROTEINS_best_reformat.txt

#######################################################
## number of taxa: 4106547 (4106547)                 ##
## DONE :) Processed blast: 12605946 - NOT FOUND 0   ##
#######################################################

# COMBINE TAXONOMY TABLES

# Download JGI_TAXA_TAB_2021.txt
gdrive_download 1VtSyy7OutKZ6fAZMTYY2HkZUDNcpfY6V JGI_TAXA_TAB_2021.txt

# Download combine_taxonomy_tables.py script
gdrive_download 1F5p28LpaHrSYWwNI_82V9eHoKIb63y8G combine_taxonomy_tables.py

python2.7 combine_taxonomy_tables.py FUNGAL_NAMES.txt JGI_TAXA_TAB_2021.txt
taxa_all_accessions.txt TAX_TAB.tab

#########################################################################
## names loaded...                                                     ##
## FUNGAL TAXONOMY PROCESSED - NOT FOUND 0 vs. FOUND 1498              ##
## OTHER TAXONOMY PROCESSED - REDUCING TO 34416 vs. ORIGINAL 4106548   ##
## DONE :)                                                             ##
#########################################################################

# Download get_best_hit_by_bitscore_multi.py script
gdrive_download 1-3XE5Le8I1_HzQdWbaAHev4ZlrSs6lUi
get_best_hit_by_bitscore_multi.py
```

```
python2.7 get_best_hit_by_bitscore_multi.py
genecalling_NCBI_nr_PROTEINS_best_reformat.txt
genecalling_JGI_FUNGAL_PROTEINS_best_reformate.txt

##############################################################################
##
## FILE: genecalling_NCBI_nr_PROTEINS_best_reformat.txt - HITS: 12605946
##
## NEW ANNOTATIONS: 12605946 - REPLACED: 0 - CURRENT BEST HITS: 12605946
##
##
##
## FILE: genecalling_JGI_FUNGAL_PROTEINS_best_reformate.txt - HITS: 4541138
##
## NEW ANNOTATIONS: 1400 - REPLACED: 7597 - CURRENT BEST HITS: 12607346
##
##
##
## done :)
##
##############################################################################
##

awk -F'\t' '{print $2}' best_of_the_blast.txt | sort | uniq >
ALL_TAXA_NAMES.txt

# Download get_taxonomy_basedonnames.py script
gdrive_download 1XruvN2qGN2-dUHZNSn0jXOYmUxoJ3Uz0
get_taxonomy_basedonnames.py

python2.7 get_taxonomy_basedonnames.py ALL_TAXA_NAMES.txt TAX_TAB.tab
TAX_TAB_FINAL.tab

#########################################################
## names loaded...                                     ##
## TAXONOMY PROCESSED - NOT FOUND 0 vs. FOUND 35600  ##
## DONE :)                                             ##
#########################################################

awk -F'\t' '{print $1"\t"$12"\t"$2""}' best_of_the_blast.txt >
TAXONOMY_BEST_OF_SIMPLE.txt

# KOGG FROM JGI-MYCO-GENOMES

awk -F'[|\t]' '{print $1"\t"$15"\t["$5"]"}'
genecalling_JGI_FUN_20210312_best.txt > FUNCTION_JGI_KOG_SIMPLE.txt

########
## 13 ##
########

##############
## SPLIT IT ##
##############

# The annotated genomic sequences in the FASTA file are divided into smaller
groups.
```

```
# This is performed using a script that fragments the file
belen_MG_Megahit_genecalling_fgs.faa into defined-sized parts (in this case,
83,000 sequences per group).
# This segmentation facilitates the management and processing of large
volumes of genomic data.

# Download split_fasta_by_group_size.py script
gdrive_download 1mGbdx3OBumymosW24WaYfZT9nq_a1z1z
split_fasta_by_group_size.py

python2.7 split_fasta_by_group_size.py
/mnt/DATA/belen/metagenomic_analysis_chapter_4/4_FGS/belen_MG_Megahit_genecal
ling_fgs.faa 83000

cd ..
mkdir 8_SPLIT
cd ./7_ANNOTATION
mv *.fas ../8_SPLIT
cd ../8_SPLIT/

########
## 14 ##
########

######################
## dbCAN ANNOTATION ##
######################

# In this step, the annotation of CAZy (Carbohydrate-Active Enzymes) is
performed using the local dbCAN database.
# First, the FASTA files are processed with the script run_dbcan.py, which
searches for Hidden Markov Model (HMM) profiles within the local dbCAN
database.
# The analysis is executed in parallel to optimize processing. The results
are consolidated into a single file (all_dbCAN.txt),
# from which the best matches are selected based on e-value to generate a
filtered file (all_dbCAN_best.txt).
# Finally, unique gene names are extracted, and a simplified table
(CAZy_BEST_SIMPLE.txt) is created,
# containing the best annotations and identifying carbohydrate-active enzymes
present in the samples.

# dbCAN local database
conda activate run_dbcan

for file in *.fas
do
 sample=${file%%.fas}
 mkdir ${sample}
done

for file in *.fas
do
  sample=${file%%.fas}
  echo "run_dbcan.py ${file} protein --db_dir
/mnt/DATA/DATABASES/run_dbcan_master/db/ -t hmmer --out_dir ${sample} --
hmm_cpu 1 --dia_cpu 1"
```

```
done > dbcan.sh

cat dbcan.sh | parallel

echo "" > all_dbCAN.txt
for file in *.fas
do
 sample=${file%%.fas}
 wc -l ${sample}/hmmer.out
 cat ${sample}/hmmer.out >> all_dbCAN.txt
done

# dbCAN annotation
export LC_ALL=en_US.UTF-8
export LANG=en_US.UTF-8

sort -t$'\t' -k3,3 -k5,5g all_dbCAN.txt | sort -u -k3,3 --merge >
all_dbCAN_best.txt

awk -F'[.\t]' '{print $1}' all_dbCAN_best.txt | sort | uniq >
hmm_names_uniq.txt

awk -F'[.\t]' '{print $1}' all_dbCAN_best.txt > hmm_names.txt

awk -F'\t' '{print $3"\t"$5}' all_dbCAN_best.txt >
all_dbCAN_best_gene_eval.txt

paste -d"\t" all_dbCAN_best_gene_eval.txt hmm_names.txt >
CAZy_BEST_SIMPLE.txt

########
## 15 ##
########

#################
## KOFAM - KOs ##
#################

# In this step, functional annotation is performed using the KOfam database,
which assigns KEGG Orthology (KO) functions to genes based on HMM profiles.
# The workflow begins by organizing directories for results (ko_tbl) and
temporary files (tmp).
# The hmmsearch tool is then used to compare KOfam HMM profiles against
genomic FASTA sequences, with tasks executed in parallel for efficiency.
# All results are merged into a single file (KOFAM_all.out.txt).
# A Python script is then used to filter the results based on predefined
thresholds and e-values,
# producing a final table (hmmsearch_KOFAM_multi_best.txt) containing the
most reliable functional annotations, linking genes to their corresponding
biological roles.

mkdir ko_tbl
mkdir tmp

for i in /mnt/DATA1/priscila/kofamKOALA/db/profiles/*.hmm
do
  file=${i##*/}
```

```
  ko=${file%%.hmm}
  echo "hmmsearch --tblout ko_tbl/${ko}.out.txt --noali --cpu 1 -E 1e-5 ${i}
~/metagenomic_analysis/4_FGS/belen_MG_Megahit_genecalling_fgs.faa >/dev/null
2>&1"
done > hmmsearch_kofam.sh

cat hmmsearch_kofam.sh | parallel -j 70 --tmpdir tmp

cat ko_tbl/*.out.txt > KOFAM_all.out.txt

python2.7 kegg_multi_from_kofamkoala_raw_filterby_thresholds_evalues.py
KOFAM_all.out.txt /mnt/DATA1/priscila/kofamKOALA/db/ko_list
hmmsearch_KOFAM_multi_best.txt

########
## 16 ##
########

#########################
## KEGG AND dbCAN tree ##
#########################

# In this step, the KEGG ontology tree is generated and filtered for unique
KOs (KEGG Orthologies) based on the functional annotations from the previous
step.
# First, a script is used to extract unique KOs from the KEGG annotations
(KO_UNIQUE_from_KO_simple.py).
# Then, the KEGG ontology table (kegg_tab.txt) is processed to retain only
the KOs present in the data, creating a subtable with relevant KOs
(KOFAM_KOs_tree.tab).

# Similarly, a tree for CAZy (Carbohydrate-Active Enzymes) is generated.
# Unique CAZy identifiers are extracted from the annotations and a script
(get_CAZy_tree.py) is used to create a CAZy-specific ontology tree
(CAZy_tree.tab),
# which provides a structured representation of the identified carbohydrate-
active enzymes in the data.

# Download kegg_tab.txt
gdrive_download 11CVgwqy6O2mJ5rc4vxevYQVp04oJrQsl kegg_tab.txt

# Download GET_KEGG_ontology_subtable.py script
gdrive_download 1AmOiMqLHE8nberbvYEDPT12JShAfSW_w
GET_KEGG_ontology_subtable.py

# Download KO_UNIQUE_from_KO_simple.py
gdrive_download 1aqENUDPh3dCoxnO8YfBZaUjwwxkmlMM1 KO_UNIQUE_from_KO_simple.py

python2.7 KO_UNIQUE_from_KO_simple.py hmmsearch_KOFAM_multi_best.txt

python2.7 GET_KEGG_ontology_subtable.py kegg_tab.txt
hmmsearch_KOFAM_multi_best.txt.unique.txt KOFAM_KOs_tree.tab

awk -F'\t' '{print $3}' ~/metagenomic_analysis/8_SPLIT/CAZy_BEST_SIMPLE.txt |
sort | uniq > CAZy_BEST_unique.txt

# Download get_CAZy_tree.py script
```

```
gdrive_download 1SGVK2cqWCLozEPGNLvPRs0YG-ckrF_CZ get_CAZy_tree.py

python2.7 get_CAZy_tree.py CAZy_BEST_unique.txt CAZy_tree.tab

########
## 17 ##
########

##############################
## LINK ANNOTATION TO TABLE ##
##############################

# In this step, the annotation results are linked to the sample table,
integrating multiple sources of functional and taxonomic data.
# We obtain the final tables of the metagenomic analysis, where we can
appreciate the abundance of each sample, the taxonomy and associated
functionality.

# Download link_simple_table_to_mapping_table.py script
gdrive_download 198TDGsV1cBfLEZorb5znFHysG47XEj5t
link_simple_table_to_mapping_table.py

cd ../7_ANNOTATION/
mv TABLE_NORM_SAMPLES_GENECALL.txt ~/metagenomic_analysis/8_SPLIT/
cp TAXONOMY_BEST_OF_SIMPLE.txt ~/metagenomic_analysis/belen/8_SPLIT/
cd ../8_SPLIT

TABLE="~/metagenomic_analysis/8_SPLIT/TABLE_NORM_SAMPLES_GENECALL.txt"

echo "${TABLE}"

TABLE_BASE=${TABLE%%.${TABLE##*.}}

echo "${TABLE_BASE}"

python2.7 link_simple_table_to_mapping_table.py ${TABLE}
TAXONOMY_BEST_OF_SIMPLE.txt TAX_BEST bitscore ${TABLE_BASE}_TAX.txt

python2.7 link_simple_table_to_mapping_table.py ${TABLE_BASE}_TAX.txt
CAZy_BEST_SIMPLE.txt CAZy e-val ${TABLE_BASE}_TAX_CAZy.txt

python2.7 link_simple_table_to_mapping_table.py ${TABLE_BASE}_TAX_CAZy.txt
../7_ANNOTATION/FUNCTION_JGI_KOG_SIMPLE.txt KOG e-val
${TABLE_BASE}_TAX_CAZy_KOG.tab

python2.7 link_simple_table_to_mapping_table.py
${TABLE_BASE}_TAX_CAZy_KOG.tab hmmsearch_KOFAM_multi_best.txt KEGG e-val
${TABLE_BASE}_TAX_CAZy_KOG_KEGG.tab

cd ../7_ANNOTATION/
cp TAX_TAB_FINAL.tab ../8_SPLIT/
cd ../8_SPLIT

# Download add_higher_taxonomy.py script
gdrive_download 1AWuqqPaP2rUMpF_uMHOGEs8aE3Iy7JS2 add_higher_taxonomy.py
```

```
python2.7 add_higher_taxonomy.py ${TABLE_BASE}_TAX_CAZy_KOG_KEGG.tab
TAX_TAB_FINAL.tab TAX_BEST ${TABLE_BASE}_TAX2_CAZy_KOG_KEGG.tab
TAX_tree_genus.tab

cd ..
mkdir 9_FINAL_TABLES
cp ./8_SPLIT/CAZy_tree.tab ./9_FINAL_TABLES
cp ./8_SPLIT/TABLE_NORM_SAMPLES_GENECALL_TAX_CAZy_KOG_KEGG.tab
./9_FINAL_TABLES
cp ./8_SPLIT/TAX_TAB_FINAL.tab ./9_FINAL_TABLES
cp ./8_SPLIT/TABLE_NORM_SAMPLES_GENECALL_TAX2_CAZy_KOG_KEGG.tab
./9_FINAL_TABLES

#########################
## ANALYSIS COMPLETED ##
#########################
```

Annex 2: Scripts used in the metagenomic analysis of Chapters 1 and 2

->filter_fastq_by_length.py

```
import sys

FASTQ_file = sys.argv[1]
OUT_file = sys.argv[2]
length = int(sys.argv[3])

r1_0 = ''
r1_1 = ''
r1_2 = ''
r1_3 = ''
filled = False
open(OUT_file, "w")

def save_by_tag(r1_0,r1_1,r1_2,r1_3,length):
    max_len = len(r1_1)
    if length <= max_len:
        with open(OUT_file, "a") as OUTfile:
            OUTfile.write('%s\n' % r1_0)
            OUTfile.write('%s\n' % r1_1)
            OUTfile.write('%s\n' % r1_2)
            OUTfile.write('%s\n' % r1_3)
            OUTfile.close()
    return;

for n, line in enumerate(open(FASTQ_file)):
    if n % 40000 == 0:
        print n / 4
    if n % 4 == 0:
        r1_0 = line.rstrip()
        #print "line1 %s" % line1
        #print "line2 %s" % line2
    else:
        if n % 4 == 1:
            r1_1 = line.rstrip()
        if n % 4 == 2:
            r1_2 = line.rstrip()
        if n % 4 == 3:
            r1_3 = line.rstrip()
            filled = True

    if filled:
        save_by_tag(r1_0,r1_1,r1_2,r1_3,length)
        filled = False

print "Done."
```

->count-up-mapped-from-results-txt-with-ctg-length.py

```
import sys

d_gene = {}

for f in sys.argv[1:]:
    for line in open(f):
        ch = line[0]
        if ch != '@':
            mg_id = f.split('.txt')[0]
            gene_name = line.rstrip().split('\t')[0]
            gene_length = line.rstrip().split('\t')[1]
            gene = gene_name+"\t"+gene_length
            count = int(line.rstrip().split('\t')[2]) #mapped
            #count = int(dat[3]) #unmapped

            if d_gene.has_key(gene):
                d_gene[gene][mg_id] = count
            else:
                d_gene[gene] = {}
                d_gene[gene][mg_id] = count

fp = open('summary-count-mapped.tsv', 'w')

sorted_samples = sys.argv[1:]

fp.write('ctg_name\tctg_length')

for x in sorted_samples:
    fp.write('\t%s' % x.split('.')[0])

fp.write('\n')

for gene in d_gene:
    fp.write('%s\t' % gene)
    for x in sorted_samples:
        x1 = x.split('.txt')[0]
        if d_gene[gene].has_key(x1):
            fp.write('%s\t' % d_gene[gene][x1])
        else:
            fp.write('0\t')
    fp.write('\n')
```

->get_assembly_coverage.py

```
import sys
import os

summary = sys.argv[1]
seqlen = int(sys.argv[2])
ass_cov_file = sys.argv[3]

min_cov = 10000.0

fp = open(ass_cov_file, 'w')
fp.write('ID\tAvg_fold\n')
for n, line in enumerate(open(summary)):
    if n > 0:
        dat = line.rstrip().split('\t')
        i = 0
        sum = 0
        len = 0
        for x in dat:
            #print('x '+str(i)+' '+x)
            if i==1:
                len = int(x)
                if len == 0:
                    break
            if i>1:
                sum += int(x)
            i += 1
        #print('cover '+str(cov))
        if len > 0:
            cov = (sum * seqlen) / float(len)
            if cov < min_cov:
                min_cov = cov
            fp.write(dat[0]+'\t'+str(cov)+'\n')
        else:
            print('len == 0 for '+dat[0])
fp.close()

print('done :] min cov '+str(min_cov))
```

->normalize-mapping-table-by-read-length-and-ctg-length.py

```
import sys

in_file = sys.argv[1]
read_size = int(sys.argv[2])
out_file = sys.argv[3]

fp = open(out_file, 'w')
for n, line in enumerate(open(in_file)):
    if n>0:
        gene_name = line.rstrip().split('\t')[0]
        gene_length = int(line.rstrip().split('\t')[1])
        if gene_length>0:
            new_line = gene_name+'\t'+line.rstrip().split('\t')[1]
            for x in range(2, len(line.rstrip().split('\t'))):
                reads_count = float(line.rstrip().split('\t')[x])
                norm_val = (read_size * reads_count)/gene_length
                #print gene_name+" "+str(x)+"
"+line.rstrip().split('\t')[x]+"   %.5f" %(norm_val)
                new_line = new_line +'\t'+ str(norm_val)
            fp.write('%s\n' % new_line)
        else:
            print "WARNING: gene length is 0 bp - "+gene_name
    else:
        fp.write('%s\n' % line.rstrip())
print "done..."
fp.close()
```

->normalize_table_by_columns.py

```
import sys

table_file = sys.argv[1]
fixed_columns = int(sys.argv[2])    #2    1)ctg_name 2)ctg_length
multi_const = int(sys.argv[3])      #100  for %
out_file = sys.argv[4]


#get col sums....
sums = []
for n, line in enumerate(open(table_file)):
    if n ==0:
        i=0
        vals = line.strip().split("\t")
        for val in vals:
            if i>=fixed_columns:
                #print str(i-fixed_columns)
                sums.append(0)#[i-fixed_columns]=0
            i=i+1
    else:
        i=0
        vals = line.strip().split("\t")
        for val in vals:
            if i>=fixed_columns:
                sums[i-fixed_columns]=sums[i-fixed_columns]+float(val)
            i=i+1

#show sums...
for sum in sums:
    print str(sum)
#normalise table and save...
fp = open(out_file, 'w')
header = ""
for n, line in enumerate(open(table_file)):
    if n ==0:
        fp.write(line.strip()+"\n")
    else:
        vals = line.strip().split("\t")
        new_line = ''
        i=0
        for val in vals:
            if (i>=fixed_columns)and(sums[i-fixed_columns]>0):
                new_line = new_line+str(float(val)/sums[i-
fixed_columns]*multi_const)+"\t"
            else:
                new_line = new_line+val+"\t"
            i=i+1
        fp.write(new_line.strip()+"\n")
fp.close()

print "Done :)"
```

->contig_mapping_to_genecall_mapping.py

```
import sys
import os

gene_call_fasta = sys.argv[1]
##title:
#>k141_20_1_453_+
ctg_mapping_tab = sys.argv[2]
##contig
#k141_43039
header = ''
abundances = {}
for n, line in enumerate(open(ctg_mapping_tab)):
    if n == 0:
        header = line.rstrip()
    else:
        vals = line.rstrip().split('\t')
        line_vals = ''
        for x in range(1,len(vals)):
            line_vals = line_vals + '\t'+vals[x]
        abundances[vals[0]] = line_vals

print("mapping table read...")

title = ''
sequence = ''
filled = False
genes_names = {}
for n, line in enumerate(open(gene_call_fasta)):
    if n % 20000 == 0:
        print(n / 2)
    if n % 2 == 0:
        title = line.rstrip()
        #print title
        if title[0] != '>':
            print("fasta format error...")
            break
    else:
        if n % 2 == 1:
            sequence = line.rstrip()
            filled = True
    if filled:
        tp = title[1:].rsplit('_',3)
        genes_names[title[1:]] = tp[0]
        filled = False

print("genecall fasta read...")

fp = open(ctg_mapping_tab + "_genecall.txt", 'w')
fp.write(header + "\n")
for name in genes_names:
    new_line = name + abundances[genes_names[name]]
    fp.write(new_line + "\n")
fp.close()
```

```
print("Done :)")
```

->replace_fungal_annot_by_taxname.py

```
import sys
import os

annotation = sys.argv[1]
short_to_tax = sys.argv[2]
output_file = sys.argv[3]

#fwd_k141_10000352_1_282_-       jgi|Dacma1|778102|KOG1368|4.1.2.5       52.5
80      38      0       13      92      20      99      5.1e-15 87.0

names = {}
with open(short_to_tax) as file:
    for line in file:
        vals = line.rstrip().split('\t')
        names[vals[0]] = vals[1]

print('Taxon pairs were loaded - ' + str(len(names)))

with open(output_file, "w") as fp:
    with open(annotation) as file:
        for line in file:
            vals = line.rstrip().split('\t')
            new_line = vals[0]
            for i in range(1, len(vals)):
                if i == 1:
                    tax = vals[i].split('|')[1]
                    if '['+tax+']' in names:
                        new_line = new_line + '\t' + names['['+tax+']']
                    else:
                        print('Taxa abbreviation was not found - ' + tax)
                        new_line = new_line + '\t' + vals[i]
                else:
                    new_line = new_line + '\t' + vals[i]
            #print(new_line)
            fp.write(new_line + '\n')

print('Done :]')
```

->get_taxonomy_offline.py

```
import sys

acc_list = sys.argv[1]
acc2taxid = sys.argv[2]
tax_taxid = sys.argv[3]
tax_out = sys.argv[4]


accs = {}
for line in open(acc_list):
    accs[line.rstrip()] = 0

print("accession list loaded...("+str(len(accs))+")")

i = 0
taxonomy = {}
for line in open(tax_taxid):
    parts = line.rstrip().split('\t')
    taxonomy[parts[7]] = parts[0] + '\t' + parts[1] + '\t' + parts[2] + '\t'
+ parts[3] + '\t' + parts[4] + '\t' + parts[5] + '\t' + parts[6]
    i += 1

print("taxonomy loaded...("+str(i)+")")

i = 0
fp = open(tax_out, 'w')
fp.write('domain\tphylum\tclass\torder\tfamily\tgenus\torganism\ttax_key\n')
for line in open(acc2taxid):
    parts = line.rstrip().split('\t')
    if accs.has_key(parts[0]):
        fp.write(taxonomy['['+parts[1]+']'] + '\t[' + parts[0] + ']\n')
        accs[parts[0]] = 1
        i += 1
fp.close()

print("taxonomy retrieved... "+str(i)+" vs acc ("+str(len(accs))+") - should
be equal!")

fp = open('missing_acc.txt', 'w')
for acc in accs:
    if accs[acc] == 0:
        fp.write(acc + '\n')
fp.close()


print("Done :]")
```

->replace_acc_by_sp_from_taxonomy.py

```
import sys
import os

blast_out6 = sys.argv[1]
taxonomy = sys.argv[2]
blast_reformat = sys.argv[3]


# load taxons
i = 0
acc_to_sp = {}
for n, line in enumerate(open(taxonomy)):
    if n > 0:
        val = line.rstrip().split("\t")
        acc_to_sp[val[7]] = '[' + val[6] + ']'
        i = i + 1
print("number of taxa: "+str(i)+" ("+str(len(acc_to_sp))+")")

# load blast
i = 0
n = 0
fp = open(blast_reformat, "w")
for line in open(blast_out6):
        val = line.rstrip().split("\t")
        acc = '[' + val[1] + ']'
        if acc_to_sp.has_key(acc):
            val[1] = acc_to_sp[acc]
        else:
            print("ERROR ACCESSION "+acc+" NOT FOUND!")
            n += 1
        fp.write("\t".join(val) + '\n')
        i += 1
fp.close()

print("DONE :) Processed blast: "+str(i)+" - NOT FOUND "+str(n))
```

->combine_taxonomy_tables.py

```
import sys

fungi_names = sys.argv[1]
fungi_taxa = sys.argv[2]
other_taxa = sys.argv[3]
final_taxa_out = sys.argv[4]

fungal_names = {}
for line in open(fungi_names):
    vals = line.rstrip().split('\t')
    fungal_names[vals[0]] = 0

print("names loaded...")


fungal_taxonomy = {}
for line in open(fungi_taxa):
    vals = line.rstrip().split('\t')
    if fungal_names.has_key(vals[6]):
        fungal_taxonomy[vals[6]] = line.rstrip()
        fungal_names[vals[6]] = 1

n = 0
k = 0
for name in fungal_names:
    if fungal_names[name] == 1:
        k += 1
    else:
        n += 1
        print("ERROR name ("+name+") was not found...")

print("FUNGAL TAXONOMY PROCESSED - NOT FOUND "+str(n)+" vs. FOUND "+str(k))

if n>0:
    print("THERE ARE ERRORS - TERMINATING SCRIPT...")
    exit()


n = 0
other_taxonomy = {}
for line in open(other_taxa ):
    vals = line.rstrip().split('\t')
    n += 1
    if len(vals)>6:
        vals[6] = "["+vals[6]+"]"
        if not other_taxonomy.has_key(vals[6]):
            if not fungal_taxonomy.has_key(vals[6]):
                other_taxonomy[vals[6]] =
vals[0]+"\t"+vals[1]+"\t"+vals[2]+"\t"+vals[3]+"\t"+vals[4]+"\t"+vals[5]+"\t"
+vals[6]
    else:
        print("PROBLEMATIC: "+line)
```

```
print("OTHER TAXONOMY PROCESSED - REDUCING TO "+str(len(other_taxonomy))+"
vs. ORIGINAL "+str(n))

fp = open(final_taxa_out, 'w')
fp.write("domain\tphylum\tclass\torder\tfamily\tgenus\tkey\n")
for name in other_taxonomy:
    fp.write(other_taxonomy[name]+"\n")
for name in fungal_taxonomy:
    fp.write(fungal_taxonomy[name]+"\n")
fp.close()

print("DONE :)")
```

->get_best_hit_by_bitscore_multi.py

```
import sys

def choose_the_best(ctgs, in_file):
    size = len(ctgs)
    i = 0
    r = 0
    for line in open(in_file):
        line = line.strip()
        vals = line.split("\t")
        # check formate
        #if i<3:
        #    print(line.strip())
        # check formate
        if len(vals)>11: #if len(vals)>11 and i>0:
                bitscore = float(vals[11])
                eval = float(vals[10])
                sim = float(vals[2])
                ctg = vals[0]
                #print("bitscore "+str(bitscore)+" eval "+str(eval)+" sim
"+str(sim))
                if ctgs.has_key(ctg):
                    vals_old = ctgs[ctg].split("\t")
                    if bitscore > float(vals_old[11]):
                        ctgs[ctg] = line
                        r += 1
                    else:
                        if bitscore == float(vals_old[11]):
                            if eval < float(vals_old[10]):
                                ctgs[ctg] = line
                                r += 1
                            else:
                                if eval == float(vals_old[10]) and sim >
float(vals_old[2]):
                                    ctgs[ctg] = line
                                    r += 1
                else:
                    ctgs[ctg] = line
        i += 1
    print("FILE: " + in_file + " - HITS: " + str(i))
    print("NEW ANNOTATIONS: " + str(len(ctgs) - size)+" - REPLACED: " +
str(r) + " - CURRENT BEST HITS: " + str(len(ctgs)))
    print("")

ctgs_best = {}

for f in sys.argv[1:]:
    choose_the_best(ctgs_best, f)

fp = open('best_of_the_blast.txt', 'w')
for ctg in ctgs_best:
    fp.write(ctgs_best[ctg]+'\n')
fp.close()
```

```
print("done :)")
```

->get_taxonomy_basedonnames.py

```
import sys

taxa_names = sys.argv[1]
taxonomy = sys.argv[2]
taxonomy_filtered = sys.argv[3]

names = {}
for line in open(taxa_names):
    vals = line.rstrip().split('\t')
    names[vals[0]] = 0

print("names loaded...")

selected_taxonomy = {}
for line in open(taxonomy):
    vals = line.rstrip().split('\t')
    if len(vals) < 7:
        print(line.rstrip())
    else:
        if names.has_key(vals[6]):
            selected_taxonomy[vals[6]] = line.rstrip()
            names[vals[6]] = 1

n = 0
k = 0
for name in names:
    if names[name] == 1:
        k += 1
    else:
        n += 1
        print("ERROR name ("+name+") was not found...")

print("TAXONOMY PROCESSED - NOT FOUND "+str(n)+" vs. FOUND "+str(k))

fp = open(taxonomy_filtered, 'w')
fp.write("domain\tphylum\tclass\torder\tfamily\tgenus\tkey\n")
for name in selected_taxonomy:
    fp.write(selected_taxonomy[name]+"\n")
fp.close()

print("DONE :)")
```

->split_fasta_by_group_size.py

```
import sys
import ntpath

in_file = sys.argv[1]
group_size = int(sys.argv[2])

part = 1
k = 1

file_name = ntpath.basename(in_file).rstrip().split('.')[0]
fp = open(file_name+str(part)+'.fas', 'w')

for n, line in enumerate(open(in_file)):
    ch = line[0]
    if ch == '>':
        if ((k % group_size)==0):
            part = part+1
            fp.close()
            print "Part: "+str(part)
            fp = open(file_name+str(part)+'.fas', 'w')
        k=k+1
    fp.write(line)
fp.close()
print "Done."
```

->KO_UNIQUE_from_KO_simple.py

```
import sys

simple_tab_file = sys.argv[1]

ko_vars = {}
for n, line in enumerate(open(simple_tab_file)):
    #print line
    vals = line.strip().split("\t")
    if len(vals)>2:
        kk = vals[2].split(";")
        for k in kk:
            if ko_vars.has_key(k):
                ko_vars[k] = ko_vars[k] + 1
            else:
                ko_vars[k] = 1
    else:
        print line

#write unique KO
fp = open(simple_tab_file+'.unique.txt', 'w')
for result in ko_vars:
    fp.write(result+"\t"+str(ko_vars[result])+"\n")
fp.close()

print "Done :)"
```

->GET_KEGG_ontology_subtable.py

```
import sys

kegg_tab_file = sys.argv[1]
unique_KO_file = sys.argv[2]
out_file = sys.argv[3]

kos = {}
for n, line in enumerate(open(unique_KO_file)):
    vals = line.strip().split("\t")
    kos[vals[0]] = ''

fp = open(out_file, 'w')
for n, line in enumerate(open(kegg_tab_file)):
    if n==0:
        fp.write(line.strip()+"\n")
    else:
        vals = line.strip().split("\t")
        if vals[len(vals)-1] in kos:
            fp.write(line.strip()+"\n")
fp.close()

print "Done :)"
```

->get_CAZy_tree.py

```
import sys
import re

input_unique = sys.argv[1]
tree_output = sys.argv[2]

values = []
for line in open(input_unique):
    val = line.rstrip()
    if len(val)>0:
        if not val == "HMM Profile":
            values.append(val)

values.sort(reverse=True)

fp = open(tree_output, 'w')
fp.write("class\tfamily\tmodel\n")
for val in values:
    print(val)
    v = val.split('_')
    ################
    match = re.match(r"([a-z]+)([0-9]+)", v[0], re.I)
    if match:
        cl = match.groups()[0]
    else:
        cl = v[0]
    #################
    fp.write(cl+"\t"+v[0]+"\t"+val+"\n")
fp.close()
```

->link_simple_table_to_mapping_table.py

```
import sys
import os

table = sys.argv[1]
best_function = sys.argv[2]
function_name = sys.argv[3]
identity_var = sys.argv[4]
#k141_1000000_1_442_-   3.30E-22    [K00074]
linked_tab = sys.argv[5]

#read functions...
fun = {}
for n, line in enumerate(open(best_function)):
    vals = line.rstrip().split('\t')
    if len(vals) == 3:
        fun[vals[0]] = vals[1] +'\t' + vals[2]

print 'Functions processed... '+str(len(fun))

#link it...
nfun = 0
fp = open(linked_tab, 'w')
for n, line in enumerate(open(table)):
    line = line.rstrip()
    line_new = ''
    if n == 0:
        #header
        line_new = line +'\t' + identity_var + '\t' +function_name
    else:
        vals = line.rstrip().split('\t')
        fun_line = 'NaN'+ '\t' +'-'
        if fun.has_key(vals[0]):
            nfun = nfun + 1
            fun_line = fun[vals[0]]
        line_new = line + '\t'+fun_line
    fp.write(line_new + "\n")
fp.close()

print 'Done... used functions: '+str(nfun)+'/'+str(len(fun))
```

->add_higher_taxonomy.py

```
import sys
import re

big_table = sys.argv[1]
tax_tree = sys.argv[2]
tax_column = sys.argv[3]
big_table_new = sys.argv[4]
tax_tree_new = sys.argv[5]

taxons = []
tax_pair = {}
fp = open(tax_tree_new, 'w')
fp.write("domain\tphylum\tclass\torder\tfamily\tgenus_key\n")
for n,line in enumerate(open(tax_tree)):
    val = line.rstrip().split('\t')
    if n > 0:
        new_key = "["+val[5]+"]"

taxons.append(val[0]+"\t"+val[1]+"\t"+val[2]+"\t"+val[3]+"\t"+val[4]+"\t"+new
_key)
        tax_pair[val[6]] = new_key

myset = set(taxons)
mylist = list(myset)
mylist.sort(reverse=True)

key_check = {}
for l in mylist:
    key = l.split('\t')[5]
    if key_check.has_key(key):
        print(">>>duplicate<<<")
        print("new: "+l)
        print("old: "+key_check[key])
    else:
        key_check[key] = l
        fp.write(l + "\n")

fp.close()

# add nes taxonomy column
tax_column_index = -1
fp = open(big_table_new, 'w')
for n,line in enumerate(open(big_table)):
    val = line.rstrip().split('\t')
    if n == 0:
        l = val[0]
        i=0
        for v in val:
            if i>0:
                l += "\t"+val[i]
                if tax_column == v:
                    tax_column_index = i
```

```
                    l += "\t" + tax_column+"_genus"
            i+=1
        fp.write(l + "\n")
    else:
        i=0
        l = val[0]
        for v in val:
            if i>0:
                l += "\t"+val[i]
                if i == tax_column_index:
                    if val[i] == '-':
                        l += "\t-"
                    else:
                        l += "\t" + tax_pair[val[i]]
            i += 1
        fp.write(l + "\n")
fp.close()

print("done :]")
```

Annex 3: Complete pipeline of the MAGs (Metagenome-Assembled Genomes) analysis carried out in Chapter 2.

```
###################
## MAGs ANALYSIS ##
###################

# The analysis of MAGs (Metagenome-Assembled Genomes) is an approach used in metagenomics to reconstruct complete genomes of microorganisms present in an environmental
# sample, without the need for prior isolation in culture. Using raw metagenomic sequences, MAGs are obtained by an assembly and binning process,
# in which contigs (DNA fragments) are grouped into bins representing individual genomes.
# These genomes can come from bacteria, archaea or other microbes present in the sample.

#######
## 1 ##
#######

#############
## BINNING ##
#############

# Binning is a key step in metagenomic analysis. The main objective of binning is to group contigs (assembled DNA fragments) into bins,
# where each bin represents a possible individual genome.

conda activate metawrap

metawrap binning -a /mnt/DATA/belen/MAGS_assembly/final.contigs.fa -o binning_metawrap -t 120 -m 1000 --metabat2 --maxbin2 --concoct --universal --run-checkm --interleaved /mnt/DATA/belen/MAGS_assembly/*.pe.qc.fq.gz

#######
## 2 ##
#######

####################
## BIN REFINEMENT ##
####################

# In this step you pick the best version of each bin. You can be more or less stringent in this step by lowering the completeness a bit.

metawrap bin_refinement -o /mnt/DATA/belen/MAGS_assembly/metawrap_refined_bins/ -A /mnt/DATA/belen/MAGS_assembly/binning_metawrap/metabat2_bins/ -B /mnt/DATA/belen/MAGS_assembly/binning_metawrap_concot/concoct_bins/ -C /mnt/DATA/belen/MAGS_assembly/binning_metawrap/maxbin2_bins/ -m 1000 -t 120 -c 50 -x 10

#######
## 3 ##
#######
```

```
##################
## CheckM2 STEP ##
##################

# CheckM2 is used to evaluate the quality of the refined bins (MAGs)
obtained. CheckM2 is a tool that estimates the completeness and contamination
of MAGs.

conda activate checkm2

checkm2 predict -i
/mnt/DATA/belen/MAGS_assembly/metawrap_refined_bins/metawrap_50_10_bins/ -x
fa --output-directory refinded_checkm2 --database_path
/mnt/DATA1/priscila/checkm2/database/CheckM2_database/uniref100.KO.1.dmnd --
tmpdir ./ --threads 240

awk '$2 >= 50 && $3 <=10' refinded_checkm2/quality_report.tsv >
good_bins_checkm2.tsv

awk '$2 >= 50 && $3 <=10' refinded_checkm2/quality_report.tsv | cut -f1 >
bins_list

for i in $(cat bins_list); do cp metawrap_50_10_bins/$i.fa selected_bins/
;done

#######
## 4 ##
#######

############################
## TAXONOMIC ANNOTATION ##
############################

# GTDB-Tk (Genome Taxonomy Database Toolkit) is used to assign taxonomy to
refined MAGs using the GTDB database version 2.4.0 (v220).
# This tool classifies microbial genomes from complete genomic data and
provides a standardized taxonomy based on the phylogenetic tree proposed by
GTDB.

conda activate /mnt/DATA1/priscila/condaenvs/gtdbtk220

gtdbtk classify_wf --genome_dir
/mnt/DATA/belen/MAGS_assembly/metawrap_refined_bins/metawrap_50_10_bins/ --
out_dir refined_checkm2_gtdb220 --cpus 240 --pplacer_cpus 60 -x .fa --tmpdir
./ --skip_ani_screen

#######
## 5 ##
#######

##########
## GUNC ##
##########

# GUNC (Genomic UNcertainty Calculator), a tool designed to assess the
taxonomic contamination and consistency of MAGs, is used.
```

```
# This analysis is crucial to verify the quality of the refined MAGs and
ensure that they represent unique and consistent genomes rather than
# mixtures of genetic material from different organisms.

conda activate gunc

export TMPDIR="/mnt/DATA/projects/priscila/tmp/"
echo $TMPDIR

mkdir selected_bins_gunc

gunc run --input_dir
/mnt/DATA/belen/MAGS_assembly/metawrap_refined_bins/metawrap_50_10_bins/ --
detailed_output --contig_taxonomy_output --use_species_level --out_dir
selected_bins_gunc --threads 120 --db_file
/mnt/DATA1/priscila/database/gunc_db_progenomes2.1.dmnd --file_suffix .fa

#######
## 6 ##
#######

#####################################
## RELATIVE QUANTIFICATION OF MAGs ##
#####################################

# The relative quantification of MAGs is performed using the Minimap2 and
CoverM tools.
# In this step, the relative abundance of each MAG in the microbial community
is determined based on the mapping of MAGs.

mkdir mags_bams

conda activate coverm

export TMPDIR="/mnt/DATA/projects/priscila/tmp/"
export TMPDIR="/mnt/DATA/belen/MAGS_assembly"

echo $TMPDIR

coverm genome --mapper minimap2-sr --methods relative_abundance -o
coverm_relative_abundance_selected.txt --bam-file-cache-directory mags_bams -
-interleaved /mnt/DATA/belen/MAGS_assembly/*.pe.qc.fq.gz --genome-fasta-
directory
/mnt/DATA/belen/MAGS_assembly/metawrap_refined_bins/metawrap_50_10_bins/ -x
fa --threads 240

#######
## 7 ##
#######

############################
## FUNCTIONAL ANNOTATION ##
############################

conda activate DRAM

# We have
```

```
lrwxrwxrwx. 1 belen belen   32 Dec 11 17:39 gtdbtk.ar53.summary.tsv ->
classify/gtdbtk.ar53.summary.tsv
lrwxrwxrwx. 1 belen belen   34 Dec 11 18:29 gtdbtk.bac120.summary.tsv ->
classify/gtdbtk.bac120.summary.tsv

# Combine both files
head -n 1
/mnt/DATA/belen/MAGS_assembly/metawrap_refined_bins/refined_checkm2_gtdb220/g
tdbtk.bac120.summary.tsv > gtdbtk_summary_bac_arch.tsv
tail -n +2
/mnt/DATA/belen/MAGS_assembly/metawrap_refined_bins/refined_checkm2_gtdb220/g
tdbtk.bac120.summary.tsv >> gtdbtk_summary_bac_arch.tsv
tail -n +2
/mnt/DATA/belen/MAGS_assembly/metawrap_refined_bins/refined_checkm2_gtdb220/g
tdbtk.ar53.summary.tsv >> gtdbtk_summary_bac_arch.tsv

# Functional annotation
DRAM.py annotate -i
'/mnt/DATA/belen/MAGS_assembly/metawrap_refined_bins/metawrap_50_10_bins/*.fa
' \
-o dram_annotation/ \
--min_contig_size 2000 \
--gtdb_taxonomy
/mnt/DATA/belen/MAGS_assembly/metawrap_refined_bins/refined_checkm2_gtdb220/g
tdbtk_summary_bac_arch.tsv \
--checkm_quality
/mnt/DATA/belen/MAGS_assembly/refinded_checkm2/quality_report.tsv \
--threads 512 \
--verbose \
--kofam_use_dbcan2_thresholds \
--keep_tmp_dir

cd /mnt/DATA/belen/MAGS_assembly/dram_annotation

DRAM.py distill -i
/mnt/DATA/belen/MAGS_assembly/dram_annotation/annotations.tsv -o
genome_summaries --trna_path
/mnt/DATA/belen/MAGS_assembly/dram_annotation/trnas.tsv --rrna_path
/mnt/DATA/belen/MAGS_assembly/dram_annotation/rrnas.tsv

#######
## 8 ##
#######

################################
## RELATIVE ABUBDANCE OF MAGs ##
################################

# In this step, quantification of the relative abundance of MAGs in each
sample is performed using the Salmon tool.
# The aim is to determine how many reads from each sample map to the
different MAGs, which provides information on the relative abundance of the
assembled genomes
# in the different samples.

conda activate metawrap
```

```
metawrap quant_bins2 -a /mnt/DATA/belen/MAGS_assembly/final.contigs.fa -o
quantified_bins -t 240 -b
/mnt/DATA/belen/MAGS_assembly/metawrap_refined_bins/metawrap_50_10_bins/
/mnt/DATA/belen/MAGS_assembly/binning_metawrap/work_files/*.bam

############################
## MAGs ANALYSIS COMPLETED ##
############################
```

Annex 4: Complete pipeline of the metagenomic analysis carried out in Chapter 3.

```
#########################
## METAGENOMIC ANALYSIS ##
#########################

# Download function gdrive_download:
function gdrive_download () {
 CONFIRM=$(wget --quiet --save-cookies /tmp/cookies.txt
--keep-session-cookies --no-check-certificate
"https://docs.google.com/uc?export=download&id=$1" -O- | sed -rn
's/.*confirm=([0-9A-Za-z_]+).*/\1\n/p')
 wget --load-cookies /tmp/cookies.txt
"https://docs.google.com/uc?export=download&confirm=$CONFIRM&id=$1" -O $2
 rm -rf /tmp/cookies.txt
}

# Installation of khmer:
sudo yum install -y python3-devel gcc-c++ make
conda create --name khmerEnv python=3.6

# Open the terminal and activate the conda environment:
conda activate base

# Create a new conda environment for Khmer:
conda create --name khmerEnv

# Activate the conda environment
conda install -c bioconda khmer

#######
## 1 ##
#######

######################
## QUALITY ANALYSIS ##
######################

# A FastQC analysis is used to assess the quality of genomic sequencing data,
such as those generated by platforms like Illumina. It evaluates base
quality,
# base composition, the presence of adapters, sequence duplication levels,
read length distribution, and the overrepresentation of sequences.

# Installation of FASTQC
conda install -c bioconda fastqc

# Quality analysis
conda activate khmerEnv
fastqc *.gz -o ~/metagenomic_analysis/1_FASTQC_RESULTS
conda deactivate

#######
## 2 ##
```

```
#######

################
## INTERLEAVE ##
################

# Interleaving in metagenomics is the process of combining two paired-end
read files into a single file. In this interleaved file,
# the forward and reverse reads of each pair are arranged alternately (i.e.,
the forward read of the pair is followed by its corresponding reverse read).
# Interleaving is primarily used to facilitate data processing by
bioinformatics tools that require paired-end reads to be stored in a single
file.

# Unzip the fastq.gz in fastq
for i in {1..32}
do
    gunzip -c ${i}_R1_001.fastq.gz > ${i}_R1_001.fastq
    gunzip -c ${i}_R2_001.fastq.gz > ${i}_R2_001.fastq
done

# Interleaved
for file in *_R1_001.fastq
do
   sample=${file%%_R1_001.fastq}
   echo "interleave-reads.py ${sample}_R1_001.fastq ${sample}_R2_001.fastq -
or ${sample}.pe.fq"
done > interleave.sh

cat interleave.sh | parallel

# Remove unnecessary files and organize them
rm -rf *.fastq
cd ...
mkdir 2_INTERLEAVED
cd 0_SAMPLES
mv *.pe.fq ../2_INTERLEAVED
cd ../2_INTERLEAVED

#######
## 3 ##
#######

##############
## TRIMMING ##
##############

# Trimming refers to the process of cleaning up and preparing DNA, RNA or
protein sequences by removing unwanted parts of the raw sequences
# obtained by techniques such as next generation sequencing (NGS).

gdrive_download 1G9G8XbGdOuajMzJGxdBq0dovF0oior_6 remove_adapters.py
gdrive_download 1TasxvzYEym3iBxMe4k8nnKd2bCVnO_Us adapters.txt
```

```
for file in *.fq
do
   echo "python2.7 remove_adapters.py adapters.txt ${file} 1"
done > trimm.sh

cat trimm.sh | parallel

mkdir 3_TRIMMED
mv *.trim ./3_TRIMMED
cd ./3_TRIMMED

#######
## 4 ##
#######

########################
## QUALITY FILTERING ##
########################

# The purpose of this step is to remove low-quality reads from sequencing
data,
# thereby improving the reliability of downstream analyses such as assemblies
or annotations.
# This ensures that the reads used meet a minimum quality standard, reducing
errors and artifacts in the final results.

# The filtering process employs the following parameters:
# -Q33: Specifies that the quality scores are encoded in the Phred+33 format,
commonly used in Illumina sequencing platforms.
# -q 30: Filters out reads where the average base quality is below 30,
corresponding to high-quality bases.
# -p 50: Retains only reads in which at least 50% of the bases meet or exceed
the specified quality threshold.

for file in *.pe.fq
do
  newfile=${file%%.pe.fq}
  echo "fastq_quality_filter -i ${file} -Q33 -q 30 -p 50 -o
${newfile}.pe.qc.fq"
done > qual_filter.sh

cat qual_filter.sh | parallel

#######
## 4 ##
#######

############################
## REMOVE SHORT SEQUENCES ##
############################

# Download function gdrive_download:
function gdrive_download () {
```

```
 CONFIRM=$(wget --quiet --save-cookies /tmp/cookies.txt --keep-session-
cookies --no-check-certificate
"https://docs.google.com/uc?export=download&id=$1" -O- | sed -rn
's/.*confirm=([0-9A-Za-z_]+).*/\1\n/p')
 wget --load-cookies /tmp/cookies.txt
"https://docs.google.com/uc?export=download&confirm=$CONFIRM&id=$1" -O $2
 rm -rf /tmp/cookies.txt
}

# Downlowad filter_fastq_by_length.py script
gdrive_download 1w-OyfdEuMi38utz4cN9g_ng-S9kNeOj9 filter_fastq_by_length.py

# Remove short sequences
for file in *pe.qc.fq
do
  echo "python2.7 filter_fastq_by_length.py ${file} ${file}.cut 50"
done > remove_short.sh

cat remove_short.sh | parallel

#######
## 5 ##
#######

############################################################
## EXTRACT PAIRED ENDS, RENAME FILES AND MERGE FILES ##
############################################################

# In this step, paired-end sequence files are processed after quality
cleaning. It includes three main steps: extracting paired reads,
# removing unnecessary files, and renaming and organizing the output files to
facilitate subsequent analysis.

# Extracting paired-ends
for file in *.pe.qc.fq.cut
do
   echo "extract-paired-reads.py ${file}"
done > extract_command.sh

cat extract_command.sh | parallel

# Remove unnecessary files
rm -rf *.tr.qc.fq.cut

# Rename files and merging pe and se files
for file in *.pe
do
   sample=${file%%.pe.qc.fq.cut.pe}
   mv ${file} ${sample}.pe.qc.fq
done

for file in *.se
do
   sample=${file%%.pe.qc.fq.cut.se}
```

```
   mv ${file} ${sample}.se.qc.fq
done

#######
## 6 ##
#######

##############################
## PREPARATION FOR ASSEMBLY ##
##############################

# In this step, paired-end sequencing data are prepared for assembly. The
script split-paired-reads.py is used to split each paired-end file
(*.pe.qc.fq)
# into two separate files: one containing the forward reads (R1) and the
other containing the reverse reads (R2). This step is necessary for assembly
tools
# such as MEGAHIT, which require paired-end reads to be provided in
individual files.

for file in *.pe.qc.fq
do
   echo "split-paired-reads.py ${file}"
done > split_command.sh

cat split_command.sh | parallel

#######
## 7 ##
#######

##############
## ASSEMBLY ##
##############

# The assembly step aims to combine sequencing reads (forward, reverse, and
unpaired) to reconstruct complete or contiguous genomic sequences
# from smaller fragments (reads).

mkdir 4_FOR_ASSEMBLY

# Create a file with all forward sequences
cat *.1 > 4_FOR_ASSEMBLY/all.pe.qc.fq.1

# Create a file with all reverse sequences
cat *.2 > 4_FOR_ASSEMBLY/all.pe.qc.fq.2

# Create a file with all unpaired sequences
cat *.se.qc.fq > 4_FOR_ASSEMBLY/all.se.qc.fq

# Assembly using MEGAHIT
megahit -m 0.75 -t 120 -1 all.pe.qc.fq.1 -2 all.pe.qc.fq.2 -r all.se.qc.fq -o
all.Megahit.assembly
```

```
#######
## 8 ##
#######

############################
## ASSEMBLY QUALITY CHECK ##
############################

# In this step, MetaQUAST is used, a tool designed to assess the quality of
genomic assemblies.
# The command takes the final contigs generated by the MEGAHIT assembly
(final.contigs.fa) as input. Several evaluation options are specified:
# --rna-finding identifies potential RNA regions.
# --conserved-genes-finding searches for conserved genes
# --max-ref-number 20 limits the maximum number of references for comparison.

# This analysis allows for the verification of the assembly's quality and
integrity.

conda activate quast

# Assembly quality check
metaquast
/mnt/DATA/belen/4_FOR_ASSEMBLY/all.Megahit.assembly/final.contigs.fa -t 120 -
-rna-finding --conserved-genes-finding --max-ref-number 20

# This is the quality report of the samples:

# contigs  5259264
# contigs (>= 0 bp)  12175284
# contigs (>= 1000 bp)     1377527
# contigs (>= 5000 bp)     46151
# contigs (>= 10000 bp)    10406
# contigs (>= 25000 bp)    1303
# contigs (>= 50000 bp)    229
# Largest contig213010
# Total length  5241019665
# Total length (>= 0 bp)   7719201062
# Total length (>= 1000 bp)     2613105109
# Total length (>= 5000 bp)     417454729
# Total length (>= 10000 bp)    180508676
# Total length (>= 25000 bp)    52451342
# Total length (>= 50000 bp)    16828383
# N50 997
# N90 566
# auN 2360.7
# L50 1384946
# L90 4272456
# GC (%)   ...

#######
## 9 ##
#######
```

```
######################
## GENECALLING - FGS ##
######################

# FragGeneScan is a program used to predict genes in DNA sequences.
# The main objective of FragGeneScan is to identify coding sequences (CDS) in
DNA sequences that may be fragmented or incomplete.

mkdir 5_FGS
cd 5_FGS

# Link creation
ln -s
/home/kdanielmorais/bioinformatics/tools/fraggenescan/FragGeneScan1.31/train/
./

# FragGeneScan
FragGeneScan -s
~/metagenomic_analysis_chapter_4/4_FOR_ASSEMBLY/all.Megahit.assembly/final.co
ntigs.fa -w 1 -o belen_MG_Megahit_genecalling_fgs -t complete -p 120

########
## 10 ##
########

#############
## MAPPING ##
#############

# Mapping is a key step in metagenomic analysis. It consists of mapping the
DNA or RNA sequences obtained from the sample to a reference database,
# usually a database of known sequences.

# In this analysis, sequencing reads were mapped against a reference
assembly. The process involved several steps,
# starting with the preparation of the reference and culminating in the
generation of alignment statistics.
# First, an index was created for the reference contigs file
(final.contigs.fa) using Bowtie2, optimizing the alignment process.
# Next, paired-end (.pe.qc.fq) and unpaired (.se.qc.fq) reads were combined
into a single file for each sample to facilitate joint mapping.
# The combined reads were then aligned against the reference using Bowtie2,
producing alignment files in SAM format,
# which were subsequently converted to compressed BAM format using Samtools.
# Mapped and unmapped reads were counted to assess the quality and efficiency
of the alignment. BAM files were sorted by reference position,
# indexed, and finally, statistics on read distribution across contigs were
generated.

REF=final.contigs.fa
reference=${REF%%.fa}
echo "reference is" ${reference}
mkdir ${reference}_build
```

```
bowtie2-build
~/metagenomic_analysis/4_FOR_ASSEMBLY/all.Megahit.assembly/${REF}
${reference}_build/${reference}.build

conda activate khmerEnv

for file in *.pe.qc.fq
do
 sample=${file%%.pe.qc.fq}
 cat ${sample}.pe.qc.fq ${sample}.se.qc.fq >
~/metagenomic_analysis_chapter_4/6_SAMPLE_MAPPING/${sample}.all.qc.fq
 echo "processing ${sample}...}"

 bowtie2 -p 70 -x
~/metagenomic_analysis_chapter_4/6_SAMPLE_MAPPING/final.contigs_build/final.c
ontigs.build -q
~/metagenomic_analysis_chapter_4/6_SAMPLE_MAPPING/${sample}.all.qc.fq -S
~/metagenomic_analysis_chapter_4/6_SAMPLE_MAPPING/${sample}.sam
 echo "sam file is done..."

 rm -rf ~/metagenomic_analysis_chapter_4/6_SAMPLE_MAPPING/${sample}.all.qc.fq

 samtools view -Sb
~/metagenomic_analysis_chapter_4/6_SAMPLE_MAPPING/${sample}.sam >
~/metagenomic_analysis_chapter_4/6_SAMPLE_MAPPING/${sample}.bam
 echo "bam file is done..."

 rm -rf ~/metagenomic_analysis_chapter_4/6_SAMPLE_MAPPING/${sample}.sam

 samtools view -c -f 4
~/metagenomic_analysis_chapter_4/6_SAMPLE_MAPPING/${sample}.bam >
~/metagenomic_analysis_chapter_4/6_SAMPLE_MAPPING/${sample}.reads-
unmapped.count.txt
 echo "unmapped reads info done..."

 samtools view -c -F 4
~/metagenomic_analysis_chapter_4/6_SAMPLE_MAPPING/${sample}.bam >
~/metagenomic_analysis_chapter_4/6_SAMPLE_MAPPING/${sample}.reads-
mapped.count.txt
 echo "mapped reads info done..."

 samtools sort -o
~/metagenomic_analysis_chapter_4/6_SAMPLE_MAPPING/${sample}.sorted.bam
~/metagenomic_analysis_chapter_4/6_SAMPLE_MAPPING/${sample}.bam
 echo "bam file was sorted..."

 rm -rf ~/metagenomic_analysis_chapter_4/6_SAMPLE_MAPPING/${sample}.bam

 samtools index
~/metagenomic_analysis_chapter_4/6_SAMPLE_MAPPING/${sample}.sorted.bam
 echo "soerted bam file was indexed..."

 samtools idxstats
~/metagenomic_analysis_chapter_4/6_SAMPLE_MAPPING/${sample}.sorted.bam >
```

```
~/metagenomic_analysis_chapter_4/6_SAMPLE_MAPPING/${sample}.reads.by.contigs.
txt

 echo "${sample} is done..."
done

# Download count-up-mapped-from-results-txt-with-ctg-length.py script
gdrive_download 1HDB2EF-pq-EJxQxsI1uv1-iVjTl6tAlo count-up-mapped-from-
results-txt-with-ctg-length.py

python2.7 count-up-mapped-from-results-txt-with-ctg-length.py
*.reads.by.contigs.txt

# Validate the consistency between the assembled contigs and the data
generated from the mapping
wc -l summary-count-mapped.tsv
12175286 summary-count-mapped.tsv

grep '>' /mnt/DATA/belen/4_FOR_ASSEMBLY/all.Megahit.assembly/final.contigs.fa
| wc -l
12175284

# Coverage is a key metric in genomics, as it indicates how many times a
genomic region has been sequenced,
# providing insights into the reliability of the assembly and the relative
abundance of the contigs.
# The purpose of this step is to generate a coverage file that associates
each assembled contig with its average coverage.

# Download function gdrive_download:
function gdrive_download () {
 CONFIRM=$(wget --quiet --save-cookies /tmp/cookies.txt --keep-session-
cookies --no-check-certificate
"https://docs.google.com/uc?export=download&id=$1" -O- | sed -rn
's/.*confirm=([0-9A-Za-z_]+).*/\1\n/p')
 wget --load-cookies /tmp/cookies.txt
"https://docs.google.com/uc?export=download&confirm=$CONFIRM&id=$1" -O $2
 rm -rf /tmp/cookies.txt
}

# Download get_assembly_coverage.py script
gdrive_download 1S2AQHd2YIjnxZz2kIa2avo1RSAuj-pWT get_assembly_coverage.py

# Obtain coverage of our data
python get_assembly_coverage.py summary-count-mapped.tsv 151
belen_MG_Megahit_assembly_DN_coverage.txt

########
## 11 ##
########

#######################################
## NORMALISE MAPPING TABLE PER BASE ##
#######################################
```

```
# In this step, the aim is to normalize the mapping table to adjust coverage based on the length of sequencing reads and contigs.
# The script normalize-mapping-table-by-read-length-and-ctg-length.py takes as input the mapping count file (summary-count-mapped.tsv) and the average read length
# (in this case, 151 bases). It generates an output file (TABLE_normalised.txt) where the mapping values are adjusted to provide a more accurate comparative
# measure of relative coverage, regardless of differences in contig or read lengths.
# This procedure is essential to correct potential biases arising from variations in lengths and allows for fair comparisons between different contigs or
# genomic regions.

# Download normalize-mapping-table-by-read-length-and-ctg-length.py script
gdrive_download 1w0bfttjXFZ64NHD8bP7UDaQcS1yd20qR normalize-mapping-table-by-read-length-and-ctg-length.py

python2.7 normalize-mapping-table-by-read-length-and-ctg-length.py summary-count-mapped.tsv 151 TABLE_normalised.txt

# In this step, an additional normalization is performed on the previously normalized table to adjust coverage values based on a predefined scale by columns.
# The script normalize_table_by_columns.py takes the previously generated file (TABLE_normalised.txt) as input, selects a specific column (in this case, column 2),
# and applies a normalization factor (1,000,000) to scale the values per sample. The output is saved in a file named TABLE_normalised_per_sample.txt.

# Download normalize_table_by_columns.py script
gdrive_download 1c_fD520xtrCNlUIq9VqqqSvY2OryMXTU normalize_table_by_columns.py

python2.7 normalize_table_by_columns.py TABLE_normalised.txt 2 1000000 TABLE_normalised_per_sample.txt

########
## 12 ##
########

################
## ANNOTATION ##
################

# Functional and taxonomic annotation are processes used to characterize genetic sequences by assigning biological information and classification.
# - Functional annotation involves identifying the roles or functions of genes and proteins, such as their involvement in specific pathways, cellular processes,
# or molecular interactions.
```

```
# - Taxonomic annotation, on the other hand, assigns sequences to their corresponding organisms or taxonomic groups,
# providing insights into the evolutionary and ecological context of the data.
# Together, these annotations allow researchers to understand both the biological role and the origin of the sequences,
# which is critical in fields such as genomics, metagenomics, and molecular biology.

# In this step, gene annotation tasks are performed by integrating alignment results with fungal protein and NCBI databases.
# First, sample information and genomic sequence data are prepared and organized.
# Then, these sequences are aligned with fungal proteins and NCBI proteins to obtain the best matches.
# Subsequently, taxonomic information is added to the alignment results through the download and processing of taxonomy files.
# The results are formatted, taxonomy tables are combined, and the best matches are selected based on bit score among the annotations.
# Finally, a table is generated containing taxonomic data and KOG functions of the annotated proteins, completing the annotation and classification process.

cd ..
mkdir 8_ANNOTATION
cp ./7_NORMALISE_MAPPING/TABLE_normalised_per_sample.txt ./8_ANNOTATION/
cd ./8_ANNOTATION

# Download contig_mapping_to_genecall_mapping.py script
gdrive_download 1DakO7roc9C2GJ-SkZuy8AZKTV3QHan14
contig_mapping_to_genecall_mapping.py

python2.7 contig_mapping_to_genecall_mapping.py ~/metagenomic_analysis_chapter_4/5_FGS/belen_MG_Megahit_genecalling_fgs.faa TABLE_normalised_per_sample.txt

# Add “#” to the name of the samples:
head -1 TABLE_normalised_per_sample.txt_genecall.txt| awk -F'\t' '{printf $1"\t"$2 ;for(i=3; i<=NF; ++i) printf "\t%s", "#"$i }' |  awk -F '\t' '{print $0}' > header.txt

tail -n +2  TABLE_normalised_per_sample.txt_genecall.txt > table.txt

cat header.txt table.txt > TABLE_NORM_SAMPLES_GENECALL.txt

#####################################
## JGI FUNGAL PROTEINS - BIOCEV PC ##
#####################################

cd /mnt/DATA/DATABASES/FUNGAL_PROTEINS_JGI/
cp JGI_FUNGAL_PROTEINS_ANNOTATED_20210312.faa.zip /mnt/DATA/belen/8_ANNOTATION/
cd /mnt/DATA/belen/8_ANNOTATION/
unzip JGI_FUNGAL_PROTEINS_ANNOTATED_20210312.faa.zip
```

```
diamond blastp -d
~/metagenomic_analysis_chapter_4/8_ANNOTATION/JGI_FUNGAL_PROTEINS_ANNOTATED_2
0210312.faa -q
~/metagenomic_analysis_chapter_4/5_FGS/belen_MG_Megahit_genecalling_fgs.faa -
e 1E-5 -o genecalling_JGI_FUN_20210312.txt -f 6 -p 120 -b12 -c1

############################################################
## Total time = 1406.2s                                   ##
## Reported 72585274 pairwise alignments, 72585274 HSPs  ##
## 4541138 queries aligned.                               ##
############################################################

export LANG=en_US.UTF-8
export LC_ALL=en_US.UTF-8

sort -t$'\t' -k1,1 -k12,12gr -k11,11g -k3,3gr
genecalling_JGI_FUN_20210312.txt | sort -u -k1,1 --merge >
genecalling_JGI_FUN_20210312_best.txt

# GENERA DEFINED
diamond blastp -d
/mnt/DATA/DATABASES/NCBI_nr_DIAMOND/NCBI_nr_20210225_diamond_GENERA -q
/mnt/DATA/belen/metagenomic_analysis_chapter_4/5_FGS/belen_MG_Megahit_genecal
ling_fgs.faa -e 1E-5 -o belen_MG_genecalling_NCBI_nr_PROTEINS_GENERA.txt -f 6
-p 120 -b12 -c1

############################################################
## Total time = 57412.9s                                  ##
## Reported 280564387 pairwise alignments, 280564387 HSPs. ##
## 12605946 queries aligned.                              ##
############################################################

export LANG=en_US.UTF-8
export LC_ALL=en_US.UTF-8
sort -t$'\t' -k1,1 -k12,12gr -k11,11g -k3,3gr
belen_MG_genecalling_NCBI_nr_PROTEINS_GENERA.txt | sort -u -k1,1 --merge >
genecalling_NCBI_nr_PROTEINS_best.txt

#  ADD TAXONOMY TO BLAST RESULTS

function gdrive_download () {
 CONFIRM=$(wget --quiet --save-cookies /tmp/cookies.txt --keep-session-
cookies --no-check-certificate
"https://docs.google.com/uc?export=download&id=$1" -O- | sed -rn
's/.*confirm=([0-9A-Za-z_]+).*/\1\n/p')
 wget --load-cookies /tmp/cookies.txt
"https://docs.google.com/uc?export=download&confirm=$CONFIRM&id=$1" -O $2
 rm -rf /tmp/cookies.txt
}

# Download jgi_abr_org_list.txt
gdrive_download 12c28kgIw4mPBIhQutNGladdAXwNLtvlR jgi_abr_org_list.txt
```

```
# Download replace_fungal_annot_by_taxname.py script
gdrive_download 1XBTtiC1JYl2rzeV7idN2WrveEZknmnQi
replace_fungal_annot_by_taxname.py

python2.7 replace_fungal_annot_by_taxname.py
genecalling_JGI_FUN_20210312_best.txt jgi_abr_org_list.txt
genecalling_JGI_FUNGAL_PROTEINS_best_reformate.txt

# FUNGAL
awk -F'\t' '{print $2}' genecalling_JGI_FUNGAL_PROTEINS_best_reformate.txt |
sort | uniq > FUNGAL_NAMES.txt

# NCBI
awk -F'\t' '{print $2}' genecalling_NCBI_nr_PROTEINS_best.txt | sort | uniq >
ALL_ACCESSIONS.txt

# Download get_taxonomy_offline.py script
gdrive_download 1o8KmSbwzOsjjeouK3dR0RNmWkMjdfFow get_taxonomy_offline.py

python2.7 get_taxonomy_offline.py ALL_ACCESSIONS.txt
/mnt/DATA/DATABASES/NCBI_nr_DIAMOND/ACC2TAXID_nr_current.txt
/mnt/DATA/DATABASES/NCBI_nr_DIAMOND/TAXONOMY_TAXID_ALL_fixed.txt
taxa_all_accessions.txt

#########################################################################
## accession list loaded...(4106547)                                   ##
## taxonomy loaded...(907158)                                          ##
## taxonomy retrieved... 4106547 vs acc (4106547) - should be equal!  ##
## Done :]                                                             ##
#########################################################################

# REFORMAT

# Download replace_acc_by_sp_from_taxonomy.py script
gdrive_download 1jQ3F3ZuA0sBJVy3eJiRhwAaxh31LKqSF
replace_acc_by_sp_from_taxonomy.py

python2.7 replace_acc_by_sp_from_taxonomy.py
genecalling_NCBI_nr_PROTEINS_best.txt taxa_all_accessions.txt
genecalling_NCBI_nr_PROTEINS_best_reformat.txt

#######################################################
## number of taxa: 4106547 (4106547)                ##
## DONE :) Processed blast: 12605946 - NOT FOUND 0  ##
#######################################################

# COMBINE TAXONOMY TABLES

# Download JGI_TAXA_TAB_2021.txt
gdrive_download 1VtSyy7OutKZ6fAZMTYY2HkZUDNcpfY6V JGI_TAXA_TAB_2021.txt

# Download combine_taxonomy_tables.py script
gdrive_download 1F5p28LpaHrSYWwNI_82V9eHoKIb63y8G combine_taxonomy_tables.py
```

```
python2.7 combine_taxonomy_tables.py FUNGAL_NAMES.txt JGI_TAXA_TAB_2021.txt
taxa_all_accessions.txt TAX_TAB.tab

#########################################################################
## names loaded...                                                     ##
## FUNGAL TAXONOMY PROCESSED - NOT FOUND 0 vs. FOUND 1498              ##
## OTHER TAXONOMY PROCESSED - REDUCING TO 34416 vs. ORIGINAL 4106548   ##
## DONE :)                                                             ##
#########################################################################

# Download get_best_hit_by_bitscore_multi.py script
gdrive_download 1-3XE5Le8I1_HzQdWbaAHev4ZlrSs6lUi
get_best_hit_by_bitscore_multi.py

python2.7 get_best_hit_by_bitscore_multi.py
genecalling_NCBI_nr_PROTEINS_best_reformat.txt
genecalling_JGI_FUNGAL_PROTEINS_best_reformate.txt

#########################################################################
##
## FILE: genecalling_NCBI_nr_PROTEINS_best_reformat.txt - HITS: 12605946
##
## NEW ANNOTATIONS: 12605946 - REPLACED: 0 - CURRENT BEST HITS: 12605946
##
##
##
## FILE: genecalling_JGI_FUNGAL_PROTEINS_best_reformate.txt - HITS: 4541138
##
## NEW ANNOTATIONS: 1400 - REPLACED: 7597 - CURRENT BEST HITS: 12607346
##
##
##
## done :)
##
#########################################################################
##

awk -F'\t' '{print $2}' best_of_the_blast.txt | sort | uniq >
ALL_TAXA_NAMES.txt

# Download get_taxonomy_basedonnames.py script
gdrive_download 1XruvN2qGN2-dUHZNSn0jXOYmUxoJ3Uz0
get_taxonomy_basedonnames.py

python2.7 get_taxonomy_basedonnames.py ALL_TAXA_NAMES.txt TAX_TAB.tab
TAX_TAB_FINAL.tab

#########################################################
## names loaded...                                     ##
## TAXONOMY PROCESSED - NOT FOUND 0 vs. FOUND 35600    ##
## DONE :)                                             ##
#########################################################
```

```
awk -F'\t' '{print $1"\t"$12"\t"$2""}' best_of_the_blast.txt >
TAXONOMY_BEST_OF_SIMPLE.txt

# KOGG FROM JGI-MYCO-GENOMES

awk -F'[|\t]' '{print $1"\t"$15"\t["$5"]"}'
genecalling_JGI_FUN_20210312_best.txt > FUNCTION_JGI_KOG_SIMPLE.txt

########
## 13 ##
########

##############
## SPLIT IT ##
##############

# The annotated genomic sequences in the FASTA file are divided into smaller
groups.
# This is performed using a script that fragments the file
belen_MG_Megahit_genecalling_fgs.faa into defined-sized parts (in this case,
83,000 sequences per group).
# This segmentation facilitates the management and processing of large
volumes of genomic data.

# Download split_fasta_by_group_size.py script
gdrive_download 1mGbdx3OBumymosW24WaYfZT9nq_a1z1z
split_fasta_by_group_size.py

python2.7 split_fasta_by_group_size.py
/mnt/DATA/belen/metagenomic_analysis_chapter_4/5_FGS/belen_MG_Megahit_genecal
ling_fgs.faa 83000

cd ..
mkdir 9_SPLIT
cd ./8_ANNOTATION
mv *.fas ../9_SPLIT
cd ../9_SPLIT/

########
## 14 ##
########

######################
## dbCAN ANNOTATION ##
######################

# In this step, the annotation of CAZy (Carbohydrate-Active Enzymes) is
performed using the local dbCAN database.
# First, the FASTA files are processed with the script run_dbcan.py, which
searches for Hidden Markov Model (HMM) profiles within the local dbCAN
database.
# The analysis is executed in parallel to optimize processing. The results
are consolidated into a single file (all_dbCAN.txt),
```

```
# from which the best matches are selected based on e-value to generate a
filtered file (all_dbCAN_best.txt).
# Finally, unique gene names are extracted, and a simplified table
(CAZy_BEST_SIMPLE.txt) is created,
# containing the best annotations and identifying carbohydrate-active enzymes
present in the samples.

# dbCAN local database
conda activate run_dbcan

for file in *.fas
do
 sample=${file%%.fas}
 mkdir ${sample}
done

for file in *.fas
do
  sample=${file%%.fas}
  echo "run_dbcan.py ${file} protein --db_dir
/mnt/DATA/DATABASES/run_dbcan_master/db/ -t hmmer --out_dir ${sample} --
hmm_cpu 1 --dia_cpu 1"
done > dbcan.sh

cat dbcan.sh | parallel

echo "" > all_dbCAN.txt
for file in *.fas
do
 sample=${file%%.fas}
 wc -l ${sample}/hmmer.out
 cat ${sample}/hmmer.out >> all_dbCAN.txt
done

# dbCAN annotation
export LC_ALL=en_US.UTF-8
export LANG=en_US.UTF-8

sort -t$'\t' -k3,3 -k5,5g all_dbCAN.txt | sort -u -k3,3 --merge >
all_dbCAN_best.txt

awk -F'[.\t]' '{print $1}' all_dbCAN_best.txt | sort | uniq >
hmm_names_uniq.txt

awk -F'[.\t]' '{print $1}' all_dbCAN_best.txt > hmm_names.txt

awk -F'\t' '{print $3"\t"$5}' all_dbCAN_best.txt >
all_dbCAN_best_gene_eval.txt

paste -d"\t" all_dbCAN_best_gene_eval.txt hmm_names.txt >
CAZy_BEST_SIMPLE.txt

########
## 15 ##
```

```
########

#################
## KOFAM - KOs ##
#################

# In this step, functional annotation is performed using the KOfam database,
which assigns KEGG Orthology (KO) functions to genes based on HMM profiles.
# The workflow begins by organizing directories for results (ko_tbl) and
temporary files (tmp).
# The hmmsearch tool is then used to compare KOfam HMM profiles against
genomic FASTA sequences, with tasks executed in parallel for efficiency.
# All results are merged into a single file (KOFAM_all.out.txt).
# A Python script is then used to filter the results based on predefined
thresholds and e-values,
# producing a final table (hmmsearch_KOFAM_multi_best.txt) containing the
most reliable functional annotations, linking genes to their corresponding
biological roles.

mkdir ko_tbl
mkdir tmp

for i in /mnt/DATA1/priscila/kofamKOALA/db/profiles/*.hmm
do
  file=${i##*/}
  ko=${file%%.hmm}
  echo "hmmsearch --tblout ko_tbl/${ko}.out.txt --noali --cpu 1 -E 1e-5 ${i}
~/metagenomic_analysis_chapter_4/5_FGS/belen_MG_Megahit_genecalling_fgs.faa
>/dev/null 2>&1"
done > hmmsearch_kofam.sh

cat hmmsearch_kofam.sh | parallel -j 70 --tmpdir tmp

cat ko_tbl/*.out.txt > KOFAM_all.out.txt

python2.7 kegg_multi_from_kofamkoala_raw_filterby_thresholds_evalues.py
KOFAM_all.out.txt /mnt/DATA1/priscila/kofamKOALA/db/ko_list
hmmsearch_KOFAM_multi_best.txt

########
## 16 ##
########

#########################
## KEGG AND dbCAN tree ##
#########################

# In this step, the KEGG ontology tree is generated and filtered for unique
KOs (KEGG Orthologies) based on the functional annotations from the previous
step.
# First, a script is used to extract unique KOs from the KEGG annotations
(KO_UNIQUE_from_KO_simple.py).
```

```
# Then, the KEGG ontology table (kegg_tab.txt) is processed to retain only
the KOs present in the data, creating a subtable with relevant KOs
(KOFAM_KOs_tree.tab).

# Similarly, a tree for CAZy (Carbohydrate-Active Enzymes) is generated.
# Unique CAZy identifiers are extracted from the annotations and a script
(get_CAZy_tree.py) is used to create a CAZy-specific ontology tree
(CAZy_tree.tab),
# which provides a structured representation of the identified carbohydrate-
active enzymes in the data.

# Download kegg_tab.txt
gdrive_download 11CVgwqy6O2mJ5rc4vxevYQVp04oJrQsl kegg_tab.txt

# Download GET_KEGG_ontology_subtable.py script
gdrive_download 1AmOiMqLHE8nberbvYEDPT12JShAfSW_w
GET_KEGG_ontology_subtable.py

# Download KO_UNIQUE_from_KO_simple.py
gdrive_download 1aqENUDPh3dCoxnO8YfBZaUjwwxkmlMM1 KO_UNIQUE_from_KO_simple.py

python2.7 KO_UNIQUE_from_KO_simple.py hmmsearch_KOFAM_multi_best.txt

python2.7 GET_KEGG_ontology_subtable.py kegg_tab.txt
hmmsearch_KOFAM_multi_best.txt.unique.txt KOFAM_KOs_tree.tab

awk -F'\t' '{print $3}'
~/metagenomic_analysis_chapter_4/9_SPLIT/CAZy_BEST_SIMPLE.txt | sort | uniq >
CAZy_BEST_unique.txt

# Download get_CAZy_tree.py script
gdrive_download 1SGVK2cqWCLozEPGNLvPRs0YG-ckrF_CZ get_CAZy_tree.py

python2.7 get_CAZy_tree.py CAZy_BEST_unique.txt CAZy_tree.tab

########
## 17 ##
########

##############################
## LINK ANNOTATION TO TABLE ##
##############################

# In this step, the annotation results are linked to the sample table,
integrating multiple sources of functional and taxonomic data.
# We obtain the final tables of the metagenomic analysis, where we can
appreciate the abundance of each sample, the taxonomy and associated
functionality.

# Download link_simple_table_to_mapping_table.py script
gdrive_download 198TDGsV1cBfLEZorb5znFHysG47XEj5t
link_simple_table_to_mapping_table.py

cd ../8_ANNOTATION/
```

```
mv TABLE_NORM_SAMPLES_GENECALL.txt ~/metagenomic_analysis/9_SPLIT/
cp TAXONOMY_BEST_OF_SIMPLE.txt ~/metagenomic_analysis/belen/9_SPLIT/
cd ../9_SPLIT

TABLE="~/metagenomic_analysis_chapter_4/9_SPLIT/TABLE_NORM_SAMPLES_GENECALL.t
xt"

echo "${TABLE}"

TABLE_BASE=${TABLE%%.${TABLE##*.}}

echo "${TABLE_BASE}"

python2.7 link_simple_table_to_mapping_table.py ${TABLE}
TAXONOMY_BEST_OF_SIMPLE.txt TAX_BEST bitscore ${TABLE_BASE}_TAX.txt

python2.7 link_simple_table_to_mapping_table.py ${TABLE_BASE}_TAX.txt
CAZy_BEST_SIMPLE.txt CAZy e-val ${TABLE_BASE}_TAX_CAZy.txt

python2.7 link_simple_table_to_mapping_table.py ${TABLE_BASE}_TAX_CAZy.txt
../8_ANNOTATION/FUNCTION_JGI_KOG_SIMPLE.txt KOG e-val
${TABLE_BASE}_TAX_CAZy_KOG.tab

python2.7 link_simple_table_to_mapping_table.py
${TABLE_BASE}_TAX_CAZy_KOG.tab hmmsearch_KOFAM_multi_best.txt KEGG e-val
${TABLE_BASE}_TAX_CAZy_KOG_KEGG.tab

cd ../8_ANNOTATION/
cp TAX_TAB_FINAL.tab ../9_SPLIT/
cd ../9_SPLIT

# Download add_higher_taxonomy.py script
gdrive_download 1AWuqqPaP2rUMpF_uMHOGEs8aE3Iy7JS2 add_higher_taxonomy.py

python2.7 add_higher_taxonomy.py ${TABLE_BASE}_TAX_CAZy_KOG_KEGG.tab
TAX_TAB_FINAL.tab TAX_BEST ${TABLE_BASE}_TAX2_CAZy_KOG_KEGG.tab
TAX_tree_genus.tab

cd ..
mkdir 10_FINAL_TABLES
cp ./9_SPLIT/CAZy_tree.tab ./10_FINAL_TABLES
cp ./9_SPLIT/TABLE_NORM_SAMPLES_GENECALL_TAX_CAZy_KOG_KEGG.tab
./10_FINAL_TABLES
cp ./9_SPLIT/TAX_TAB_FINAL.tab ./10_FINAL_TABLES
cp ./9_SPLIT/TABLE_NORM_SAMPLES_GENECALL_TAX2_CAZy_KOG_KEGG.tab
./10_FINAL_TABLES

########################
## ANALYSIS COMPLETED ##
########################
```

Annex 5: Scripts used in the metagenomic analysis of Chapter 3.

->filter_fastq_by_length.py

```
import sys

FASTQ_file = sys.argv[1]
OUT_file = sys.argv[2]
length = int(sys.argv[3])

r1_0 = ''
r1_1 = ''
r1_2 = ''
r1_3 = ''
filled = False
open(OUT_file, "w")

def save_by_tag(r1_0,r1_1,r1_2,r1_3,length):
    max_len = len(r1_1)
    if length <= max_len:
        with open(OUT_file, "a") as OUTfile:
            OUTfile.write('%s\n' % r1_0)
            OUTfile.write('%s\n' % r1_1)
            OUTfile.write('%s\n' % r1_2)
            OUTfile.write('%s\n' % r1_3)
            OUTfile.close()
    return;

for n, line in enumerate(open(FASTQ_file)):
    if n % 40000 == 0:
        print n / 4
    if n % 4 == 0:
        r1_0 = line.rstrip()
        #print "line1 %s" % line1
        #print "line2 %s" % line2
    else:
        if n % 4 == 1:
            r1_1 = line.rstrip()
        if n % 4 == 2:
            r1_2 = line.rstrip()
        if n % 4 == 3:
            r1_3 = line.rstrip()
            filled = True

    if filled:
        save_by_tag(r1_0,r1_1,r1_2,r1_3,length)
        filled = False

print "Done."
```

->count-up-mapped-from-results-txt-with-ctg-length.py

```
import sys

d_gene = {}

for f in sys.argv[1:]:
    for line in open(f):
        ch = line[0]
        if ch != '@':
            mg_id = f.split('.txt')[0]
            gene_name = line.rstrip().split('\t')[0]
            gene_length = line.rstrip().split('\t')[1]
            gene = gene_name+"\t"+gene_length
            count = int(line.rstrip().split('\t')[2]) #mapped
            #count = int(dat[3]) #unmapped

            if d_gene.has_key(gene):
                d_gene[gene][mg_id] = count
            else:
                d_gene[gene] = {}
                d_gene[gene][mg_id] = count

fp = open('summary-count-mapped.tsv', 'w')

sorted_samples = sys.argv[1:]

fp.write('ctg_name\tctg_length')

for x in sorted_samples:
    fp.write('\t%s' % x.split('.')[0])

fp.write('\n')

for gene in d_gene:
    fp.write('%s\t' % gene)
    for x in sorted_samples:
        x1 = x.split('.txt')[0]
        if d_gene[gene].has_key(x1):
            fp.write('%s\t' % d_gene[gene][x1])
        else:
            fp.write('0\t')
    fp.write('\n')
```

->get_assembly_coverage.py

```
import sys
import os

summary = sys.argv[1]
seqlen = int(sys.argv[2])
ass_cov_file = sys.argv[3]

min_cov = 10000.0

fp = open(ass_cov_file, 'w')
fp.write('ID\tAvg_fold\n')
for n, line in enumerate(open(summary)):
    if n > 0:
        dat = line.rstrip().split('\t')
        i = 0
        sum = 0
        len = 0
        for x in dat:
            #print('x '+str(i)+' '+x)
            if i==1:
                len = int(x)
                if len == 0:
                    break
            if i>1:
                sum += int(x)
            i += 1
        #print('cover '+str(cov))
        if len > 0:
            cov = (sum * seqlen) / float(len)
            if cov < min_cov:
                min_cov = cov
            fp.write(dat[0]+'\t'+str(cov)+'\n')
        else:
            print('len == 0 for '+dat[0])
fp.close()

print('done :] min cov '+str(min_cov))
```

->normalize-mapping-table-by-read-length-and-ctg-length.py

```
import sys

in_file = sys.argv[1]
read_size = int(sys.argv[2])
out_file = sys.argv[3]

fp = open(out_file, 'w')
for n, line in enumerate(open(in_file)):
    if n>0:
        gene_name = line.rstrip().split('\t')[0]
        gene_length = int(line.rstrip().split('\t')[1])
        if gene_length>0:
            new_line = gene_name+'\t'+line.rstrip().split('\t')[1]
            for x in range(2, len(line.rstrip().split('\t'))):
                reads_count = float(line.rstrip().split('\t')[x])
                norm_val = (read_size * reads_count)/gene_length
                #print gene_name+" "+str(x)+"
"+line.rstrip().split('\t')[x]+"   %.5f" %(norm_val)
                new_line = new_line +'\t'+ str(norm_val)
            fp.write('%s\n' % new_line)
        else:
            print "WARNING: gene length is 0 bp - "+gene_name
    else:
        fp.write('%s\n' % line.rstrip())
print "done..."
fp.close()
```

->normalize_table_by_columns.py

```
import sys

table_file = sys.argv[1]
fixed_columns = int(sys.argv[2])    #2    1)ctg_name 2)ctg_length
multi_const = int(sys.argv[3])      #100  for %
out_file = sys.argv[4]


#get col sums....
sums = []
for n, line in enumerate(open(table_file)):
    if n ==0:
        i=0
        vals = line.strip().split("\t")
        for val in vals:
            if i>=fixed_columns:
                #print str(i-fixed_columns)
                sums.append(0)#[i-fixed_columns]=0
            i=i+1
    else:
        i=0
        vals = line.strip().split("\t")
        for val in vals:
            if i>=fixed_columns:
                sums[i-fixed_columns]=sums[i-fixed_columns]+float(val)
            i=i+1

#show sums...
for sum in sums:
    print str(sum)
#normalise table and save...
fp = open(out_file, 'w')
header = ""
for n, line in enumerate(open(table_file)):
    if n ==0:
        fp.write(line.strip()+"\n")
    else:
        vals = line.strip().split("\t")
        new_line = ''
        i=0
        for val in vals:
            if (i>=fixed_columns)and(sums[i-fixed_columns]>0):
                new_line = new_line+str(float(val)/sums[i-
fixed_columns]*multi_const)+"\t"
            else:
                new_line = new_line+val+"\t"
            i=i+1
        fp.write(new_line.strip()+"\n")
fp.close()

print "Done :)"
```

->contig_mapping_to_genecall_mapping.py

```
import sys
import os

gene_call_fasta = sys.argv[1]
##title:
#>k141_20_1_453_+
ctg_mapping_tab = sys.argv[2]
##contig
#k141_43039
header = ''
abundances = {}
for n, line in enumerate(open(ctg_mapping_tab)):
    if n == 0:
        header = line.rstrip()
    else:
        vals = line.rstrip().split('\t')
        line_vals = ''
        for x in range(1,len(vals)):
            line_vals = line_vals + '\t'+vals[x]
        abundances[vals[0]] = line_vals

print("mapping table read...")

title = ''
sequence = ''
filled = False
genes_names = {}
for n, line in enumerate(open(gene_call_fasta)):
    if n % 20000 == 0:
        print(n / 2)
    if n % 2 == 0:
        title = line.rstrip()
        #print title
        if title[0] != '>':
            print("fasta format error...")
            break
    else:
        if n % 2 == 1:
            sequence = line.rstrip()
            filled = True
    if filled:
        tp = title[1:].rsplit('_',3)
        genes_names[title[1:]] = tp[0]
        filled = False

print("genecall fasta read...")

fp = open(ctg_mapping_tab + "_genecall.txt", 'w')
fp.write(header + "\n")
for name in genes_names:
    new_line = name + abundances[genes_names[name]]
    fp.write(new_line + "\n")
fp.close()
```

```
print("Done :)")
```
->replace_fungal_annot_by_taxname.py

```
import sys
import os

annotation = sys.argv[1]
short_to_tax = sys.argv[2]
output_file = sys.argv[3]

#fwd_k141_10000352_1_282_-       jgi|Dacma1|778102|KOG1368|4.1.2.5       52.5
80      38      0       13      92      20      99      5.1e-15 87.0

names = {}
with open(short_to_tax) as file:
    for line in file:
        vals = line.rstrip().split('\t')
        names[vals[0]] = vals[1]

print('Taxon pairs were loaded - ' + str(len(names)))

with open(output_file, "w") as fp:
    with open(annotation) as file:
        for line in file:
            vals = line.rstrip().split('\t')
            new_line = vals[0]
            for i in range(1, len(vals)):
                if i == 1:
                    tax = vals[i].split('|')[1]
                    if '['+tax+']' in names:
                        new_line = new_line + '\t' + names['['+tax+']']
                    else:
                        print('Taxa abbreviation was not found - ' + tax)
                        new_line = new_line + '\t' + vals[i]
                else:
                    new_line = new_line + '\t' + vals[i]
            #print(new_line)
            fp.write(new_line + '\n')

print('Done :]')
```

->get_taxonomy_offline.py

```
import sys

acc_list = sys.argv[1]
acc2taxid = sys.argv[2]
tax_taxid = sys.argv[3]
tax_out = sys.argv[4]


accs = {}
for line in open(acc_list):
    accs[line.rstrip()] = 0

print("accession list loaded...("+str(len(accs))+")")

i = 0
taxonomy = {}
for line in open(tax_taxid):
    parts = line.rstrip().split('\t')
    taxonomy[parts[7]] = parts[0] + '\t' + parts[1] + '\t' + parts[2] + '\t'
+ parts[3] + '\t' + parts[4] + '\t' + parts[5] + '\t' + parts[6]
    i += 1

print("taxonomy loaded...("+str(i)+")")

i = 0
fp = open(tax_out, 'w')
fp.write('domain\tphylum\tclass\torder\tfamily\tgenus\torganism\ttax_key\n')
for line in open(acc2taxid):
    parts = line.rstrip().split('\t')
    if accs.has_key(parts[0]):
        fp.write(taxonomy['['+parts[1]+']'] + '\t[' + parts[0] + ']\n')
        accs[parts[0]] = 1
        i += 1
fp.close()

print("taxonomy retrieved... "+str(i)+" vs acc ("+str(len(accs))+") - should
be equal!")

fp = open('missing_acc.txt', 'w')
for acc in accs:
    if accs[acc] == 0:
        fp.write(acc + '\n')
fp.close()


print("Done :]")
```

->replace_acc_by_sp_from_taxonomy.py

```
import sys
import os

blast_out6 = sys.argv[1]
taxonomy = sys.argv[2]
blast_reformat = sys.argv[3]


# load taxons
i = 0
acc_to_sp = {}
for n, line in enumerate(open(taxonomy)):
    if n > 0:
        val = line.rstrip().split("\t")
        acc_to_sp[val[7]] = '[' + val[6] + ']'
        i = i + 1
print("number of taxa: "+str(i)+" ("+str(len(acc_to_sp))+")")

# load blast
i = 0
n = 0
fp = open(blast_reformat, "w")
for line in open(blast_out6):
        val = line.rstrip().split("\t")
        acc = '[' + val[1] + ']'
        if acc_to_sp.has_key(acc):
            val[1] = acc_to_sp[acc]
        else:
            print("ERROR ACCESSION "+acc+" NOT FOUND!")
            n += 1
        fp.write("\t".join(val) + '\n')
        i += 1
fp.close()

print("DONE :) Processed blast: "+str(i)+" - NOT FOUND "+str(n))
```

->combine_taxonomy_tables.py

```
import sys

fungi_names = sys.argv[1]
fungi_taxa = sys.argv[2]
other_taxa = sys.argv[3]
final_taxa_out = sys.argv[4]

fungal_names = {}
for line in open(fungi_names):
    vals = line.rstrip().split('\t')
    fungal_names[vals[0]] = 0

print("names loaded...")


fungal_taxonomy = {}
for line in open(fungi_taxa):
    vals = line.rstrip().split('\t')
    if fungal_names.has_key(vals[6]):
        fungal_taxonomy[vals[6]] = line.rstrip()
        fungal_names[vals[6]] = 1

n = 0
k = 0
for name in fungal_names:
    if fungal_names[name] == 1:
        k += 1
    else:
        n += 1
        print("ERROR name ("+name+") was not found...")

print("FUNGAL TAXONOMY PROCESSED - NOT FOUND "+str(n)+" vs. FOUND "+str(k))

if n>0:
    print("THERE ARE ERRORS - TERMINATING SCRIPT...")
    exit()


n = 0
other_taxonomy = {}
for line in open(other_taxa ):
    vals = line.rstrip().split('\t')
    n += 1
    if len(vals)>6:
        vals[6] = "["+vals[6]+"]"
        if not other_taxonomy.has_key(vals[6]):
            if not fungal_taxonomy.has_key(vals[6]):
                other_taxonomy[vals[6]] =
vals[0]+"\t"+vals[1]+"\t"+vals[2]+"\t"+vals[3]+"\t"+vals[4]+"\t"+vals[5]+"\t"
+vals[6]
    else:
        print("PROBLEMATIC: "+line)
```

```
print("OTHER TAXONOMY PROCESSED - REDUCING TO "+str(len(other_taxonomy))+"
vs. ORIGINAL "+str(n))

fp = open(final_taxa_out, 'w')
fp.write("domain\tphylum\tclass\torder\tfamily\tgenus\tkey\n")
for name in other_taxonomy:
    fp.write(other_taxonomy[name]+"\n")
for name in fungal_taxonomy:
    fp.write(fungal_taxonomy[name]+"\n")
fp.close()

print("DONE :)")
```

->get_best_hit_by_bitscore_multi.py

```
import sys

def choose_the_best(ctgs, in_file):
    size = len(ctgs)
    i = 0
    r = 0
    for line in open(in_file):
        line = line.strip()
        vals = line.split("\t")
        # check formate
        #if i<3:
        #    print(line.strip())
        # check formate
        if len(vals)>11: #if len(vals)>11 and i>0:
                bitscore = float(vals[11])
                eval = float(vals[10])
                sim = float(vals[2])
                ctg = vals[0]
                #print("bitscore "+str(bitscore)+" eval "+str(eval)+" sim
"+str(sim))
                if ctgs.has_key(ctg):
                    vals_old = ctgs[ctg].split("\t")
                    if bitscore > float(vals_old[11]):
                        ctgs[ctg] = line
                        r += 1
                    else:
                        if bitscore == float(vals_old[11]):
                            if eval < float(vals_old[10]):
                                ctgs[ctg] = line
                                r += 1
                            else:
                                if eval == float(vals_old[10]) and sim >
float(vals_old[2]):
                                    ctgs[ctg] = line
                                    r += 1
                else:
                    ctgs[ctg] = line
        i += 1
    print("FILE: " + in_file + " - HITS: " + str(i))
    print("NEW ANNOTATIONS: " + str(len(ctgs) - size)+" - REPLACED: " +
str(r) + " - CURRENT BEST HITS: " + str(len(ctgs)))
    print("")

ctgs_best = {}

for f in sys.argv[1:]:
    choose_the_best(ctgs_best, f)

fp = open('best_of_the_blast.txt', 'w')
for ctg in ctgs_best:
    fp.write(ctgs_best[ctg]+'\n')
fp.close()
```

```
print("done :)")
```

->get_taxonomy_basedonnames.py

```
import sys

taxa_names = sys.argv[1]
taxonomy = sys.argv[2]
taxonomy_filtered = sys.argv[3]

names = {}
for line in open(taxa_names):
    vals = line.rstrip().split('\t')
    names[vals[0]] = 0

print("names loaded...")

selected_taxonomy = {}
for line in open(taxonomy):
    vals = line.rstrip().split('\t')
    if len(vals) < 7:
        print(line.rstrip())
    else:
        if names.has_key(vals[6]):
            selected_taxonomy[vals[6]] = line.rstrip()
            names[vals[6]] = 1

n = 0
k = 0
for name in names:
    if names[name] == 1:
        k += 1
    else:
        n += 1
        print("ERROR name ("+name+") was not found...")

print("TAXONOMY PROCESSED - NOT FOUND "+str(n)+" vs. FOUND "+str(k))

fp = open(taxonomy_filtered, 'w')
fp.write("domain\tphylum\tclass\torder\tfamily\tgenus\tkey\n")
for name in selected_taxonomy:
    fp.write(selected_taxonomy[name]+"\n")
fp.close()

print("DONE :)")
```

->split_fasta_by_group_size.py

```
import sys
import ntpath

in_file = sys.argv[1]
group_size = int(sys.argv[2])

part = 1
k = 1

file_name = ntpath.basename(in_file).rstrip().split('.')[0]
fp = open(file_name+str(part)+'.fas', 'w')

for n, line in enumerate(open(in_file)):
    ch = line[0]
    if ch == '>':
        if ((k % group_size)==0):
            part = part+1
            fp.close()
            print "Part: "+str(part)
            fp = open(file_name+str(part)+'.fas', 'w')
        k=k+1
    fp.write(line)
fp.close()
print "Done."
```

->KO_UNIQUE_from_KO_simple.py

```
import sys

simple_tab_file = sys.argv[1]

ko_vars = {}
for n, line in enumerate(open(simple_tab_file)):
    #print line
    vals = line.strip().split("\t")
    if len(vals)>2:
        kk = vals[2].split(";")
        for k in kk:
            if ko_vars.has_key(k):
                ko_vars[k] = ko_vars[k] + 1
            else:
                ko_vars[k] = 1
    else:
        print line

#write unique KO
fp = open(simple_tab_file+'.unique.txt', 'w')
for result in ko_vars:
    fp.write(result+"\t"+str(ko_vars[result])+"\n")
fp.close()

print "Done :)"
```

->GET_KEGG_ontology_subtable.py

```
import sys

kegg_tab_file = sys.argv[1]
unique_KO_file = sys.argv[2]
out_file = sys.argv[3]

kos = {}
for n, line in enumerate(open(unique_KO_file)):
    vals = line.strip().split("\t")
    kos[vals[0]] = ''

fp = open(out_file, 'w')
for n, line in enumerate(open(kegg_tab_file)):
    if n==0:
        fp.write(line.strip()+"\n")
    else:
        vals = line.strip().split("\t")
        if vals[len(vals)-1] in kos:
            fp.write(line.strip()+"\n")
fp.close()

print "Done :)"
```

->get_CAZy_tree.py

```
import sys
import re

input_unique = sys.argv[1]
tree_output = sys.argv[2]

values = []
for line in open(input_unique):
    val = line.rstrip()
    if len(val)>0:
        if not val == "HMM Profile":
            values.append(val)

values.sort(reverse=True)

fp = open(tree_output, 'w')
fp.write("class\tfamily\tmodel\n")
for val in values:
    print(val)
    v = val.split('_')
    ################
    match = re.match(r"([a-z]+)([0-9]+)", v[0], re.I)
    if match:
        cl = match.groups()[0]
    else:
        cl = v[0]
    #################
    fp.write(cl+"\t"+v[0]+"\t"+val+"\n")
fp.close()
```

->link_simple_table_to_mapping_table.py

```
import sys
import os

table = sys.argv[1]
best_function = sys.argv[2]
function_name = sys.argv[3]
identity_var = sys.argv[4]
#k141_1000000_1_442_-   3.30E-22    [K00074]
linked_tab = sys.argv[5]

#read functions...
fun = {}
for n, line in enumerate(open(best_function)):
    vals = line.rstrip().split('\t')
    if len(vals) == 3:
        fun[vals[0]] = vals[1] +'\t' + vals[2]

print 'Functions processed... '+str(len(fun))

#link it...
nfun = 0
fp = open(linked_tab, 'w')
for n, line in enumerate(open(table)):
    line = line.rstrip()
    line_new = ''
    if n == 0:
        #header
        line_new = line +'\t' + identity_var + '\t' +function_name
    else:
        vals = line.rstrip().split('\t')
        fun_line = 'NaN'+ '\t' +'-'
        if fun.has_key(vals[0]):
            nfun = nfun + 1
            fun_line = fun[vals[0]]
        line_new = line + '\t'+fun_line
    fp.write(line_new + "\n")
fp.close()

print 'Done... used functions: '+str(nfun)+'/'+str(len(fun))
```

->add_higher_taxonomy.py

```
import sys
import re

big_table = sys.argv[1]
tax_tree = sys.argv[2]
tax_column = sys.argv[3]
big_table_new = sys.argv[4]
tax_tree_new = sys.argv[5]

taxons = []
tax_pair = {}
fp = open(tax_tree_new, 'w')
fp.write("domain\tphylum\tclass\torder\tfamily\tgenus_key\n")
for n,line in enumerate(open(tax_tree)):
    val = line.rstrip().split('\t')
    if n > 0:
        new_key = "["+val[5]+"]"

taxons.append(val[0]+"\t"+val[1]+"\t"+val[2]+"\t"+val[3]+"\t"+val[4]+"\t"+new
_key)
        tax_pair[val[6]] = new_key

myset = set(taxons)
mylist = list(myset)
mylist.sort(reverse=True)

key_check = {}
for l in mylist:
    key = l.split('\t')[5]
    if key_check.has_key(key):
        print(">>>duplicate<<<")
        print("new: "+l)
        print("old: "+key_check[key])
    else:
        key_check[key] = l
        fp.write(l + "\n")

fp.close()

# add nes taxonomy column
tax_column_index = -1
fp = open(big_table_new, 'w')
for n,line in enumerate(open(big_table)):
    val = line.rstrip().split('\t')
    if n == 0:
        l = val[0]
        i=0
        for v in val:
            if i>0:
                l += "\t"+val[i]
                if tax_column == v:
                    tax_column_index = i
```

```
                    l += "\t" + tax_column+"_genus"
            i+=1
        fp.write(l + "\n")
    else:
        i=0
        l = val[0]
        for v in val:
            if i>0:
                l += "\t"+val[i]
                if i == tax_column_index:
                    if val[i] == '-':
                        l += "\t-"
                    else:
                        l += "\t" + tax_pair[val[i]]
            i += 1
        fp.write(l + "\n")
fp.close()

print("done :]")
```

ANNEX 6: Complete pipeline of the metatranscriptomic analysis carried out in Chapter 3.

```
conda activate fastp


for file in *_R1.fq.gz
do
   sample=${file%%_R1.fq.gz}
   fastp --detect_adapter_for_pe --adapter_sequence=AGATCGGAAGAG
--adapter_sequence_r2=AGATCGGAAGAG -W 1 -M 3 -5 -3 -g -q 30 -u 50
-l 50 -h ${sample}.html --thread=16 --dont_eval_duplication  -i
${sample}_R1.fq.gz -I ${sample}_R2.fq.gz --
unpaired1=filtered/${sample}.se.fq --
unpaired2=filtered/${sample}.se.fq  --stdout >
filtered/${sample}.pe.trim.qc.fq
done

# cat se reads into same file as pe

for file in *.pe.trim.qc.fq
do
   sample=${file%%.pe.trim.qc.fq}
cat ${sample}.se.fq >> ${file}
done


##REMOVE rRNAs WITH - bbduk.sh

bbdir='/home/kdanielmorais/bioinformatics/tools/BBtools/'
echo ${bbdir}
```

```
for file in *.pe.trim.qc.fq
do
    sample=${file%%.pe.trim.qc.fq}
    echo "${bbdir}bbmap/bbduk.sh ordered k=31
ref=${bbdir}ribokmers.fa.gz ow=true in=${file}
out=rRNA_remove/${sample}_NO_rRNA.pe.fq
outm=rRNA_remove/${sample}_rRNA.pe.fq"
done > remove_rrna.sh

cat remove_rrna.sh | parallel

# fix paired R1 and R2 files

for file in *_NO_rRNA.pe.fq
do
   sample=${file%%_NO_rRNA.pe.fq}
   echo
"/home/kdanielmorais/bioinformatics/tools/BBtools/bbmap/repair.sh
in=${file} out1=for_assembly/${sample}.pe.fq.1
out2=for_assembly/${sample}.pe.fq.2
outsingle=for_assembly/${sample}.se.fq repair"
done > extract_command.sh

cat extract_command.sh | parallel

#rm -rf *.tr.qc.fq.cut

# trinity assembly
## obs about rnaSeq libs: the kit used normaly here are stranded
rna preps, this generates strand-specific transcripts and should
be assembled a bit different. Trinity has an option for this --
SS_lib_type (can be RF or FR) the kit "TruSeq" uses dUTP method
(FR according to Trinity documents). Should try to compare this
data assebled normally and considering strand-specifi option as
well????

## put all reads together

cat *.1 > all.qc.fq.1
cat *.2 > all.qc.fq.2
cat *.se.fq > all.se.qc.fq
#put unpaired reads into file .1
cat all.se.qc.fq >> all.qc.fq.1

### TRINITY ASSEMBLY
mkdir trinity_assembly
#trinity can't use more than 200G of ram
conda activate TrinityEnv
```

```
Trinity --NO_SEQTK  --seqType fq  --left all.qc.fq.1 --right
all.qc.fq.2 --CPU 120 --max_memory 200G --output
trinity_assembly/


## SAMPLE MAPPING
#find a way for this loop to work across different folders!!!!!

REF=../trinity_assembly.Trinity.fasta
reference=${REF%%.fasta}
echo "reference is" ${reference}
mkdir ${reference}_build
bowtie2-build ${REF} ${reference}_build/${reference}.build


ref=trinity_assembly.Trinity_build
reference=${ref%%_build}

for file in *_NO_rRNA.pe.fq
 do
 sample=${file%%_NO_rRNA.pe.fq}
 echo "processing ${sample}... reference ${reference}"

 bowtie2 -p 70 -x
for_assembly/trinity_assembly.Trinity_build/trinity_assembly.Trin
ity.build -q ${file} -S ${sample}.sam
 echo "sam file is done..."

 #rm -rf ${sample}.pe.fq

 samtools view -Sb ${sample}.sam > ${sample}.bam
 echo "bam file is done..."

 rm -rf ${sample}.sam

 samtools view -c -f 4 ${sample}.bam > ${sample}.reads-
unmapped.count.txt
 echo "unmapped reads info done..."

 samtools view -c -F 4 ${sample}.bam > ${sample}.reads-
mapped.count.txt
 echo "mapped reads info done..."

 samtools sort -o ${sample}.sorted.bam ${sample}.bam
 echo "bam file was sorted..."

 rm -rf ${sample}.bam

 samtools index ${sample}.sorted.bam
```

```
 echo "soerted bam file was indexed..."
 samtools idxstats ${sample}.sorted.bam >
${sample}.reads.by.contigs.txt

 echo "sample ${sample} is done..."
done

######GETTING COUNT MAPPING TABLE CONTIG/SAMPLE
gdrive_download 1HDB2EF-pq-EJxQxsI1uv1-iVjTl6tAlo count-up-
mapped-from-results-txt-with-ctg-length.py

python2.7 count-up-mapped-from-results-txt-with-ctg-length.py
*.reads.by.contigs.txt

#####################
# get coverage file #
#####################

https://drive.google.com/file/d/1S2AQHd2YIjnxZz2kIa2avo1RSAuj-
pWT/view?usp=sharing

gdrive_download 1S2AQHd2YIjnxZz2kIa2avo1RSAuj-pWT
get_assembly_coverage.py

# adjusted to 145bp because used the trimmed and filtered reads
_NO_rRNA files

python2.7 get_assembly_coverage.py summary-count-mapped.tsv 145
Ruben_substrateMT_assembly_coverage.txt


###GENE CALLING - FragGeneScan

#####################
# GENECALLING - FGS #
#####################

# Link creation
ln -s
/home/kdanielmorais/bioinformatics/tools/fraggenescan/FragGeneSca
n1.31/train/ ./

# Command
FragGeneScan -s  ../for_assembly/trinity_assembly.Trinity.fasta -
w 1 -o Ruben_substrateMT_trinity_genecalling -t complete -p 120


###FALTA FAZER A PARTE FINAL DO MAPPING E NORMALIZACAO
#####################################
```

```
# 2a NORMALISE MAPPING TABLE PER BASE
#####################################
gdrive_download 1w0bfttjXFZ64NHD8bP7UDaQcS1yd20qR normalize-
mapping-table-by-read-length-and-ctg-length.py

python2.7 normalize-mapping-table-by-read-length-and-ctg-
length.py summary-count-mapped.tsv 145 TABLE_normalised_145.txt

WARNING: gene length is 0 bp - *
done...


#########################################
# 2b NORMALISE MAPPING TABLE PER SAMPLE
#########################################
gdrive_download 1c_fD520xtrCNlUIq9VqqqSvY2OryMXTU
normalize_table_by_columns.py

python2.7 normalize_table_by_columns.py TABLE_normalised_145.txt
2 1000000 TABLE_normalised_per_sample.txt



############################
# 3 MULTIPLY BY GENECALL # -->
############################

#k141_43039
gdrive_download 1DakO7roc9C2GJ-SkZuy8AZKTV3QHan14
contig_mapping_to_genecall_mapping.py

python2.7 contig_mapping_to_genecall_mapping.py
genecalling/Ruben_substrateMT_trinity_genecalling.faa
TABLE_normalised_per_sample.txt

ls
#############################
# ADD "#" TO SAMPLE NAMES  #
#############################

head -1 TABLE_normalised_per_sample.txt_genecall.txt | awk -F'\t'
'{printf $1"\t"$2 ;for(i=3; i<=NF; ++i) printf "\t%s", "#"$i }' |
awk -F '\t' '{print $0}' > header.txt

tail -n +2  TABLE_normalised_per_sample.txt_genecall.txt >
table.txt

cat header.txt table.txt > TABLE_NORM_SAMPLES_GENECALL.txt
```

```
#####################
## DRAM ANNOTATION ##
#####################

#error in description can be fixed with
DRAM-setup.py update_description_db
# runs for a few hours and takes a few hundred Gb disk space
conda activate DRAM
DRAM.py annotate_genes -i
genecalling/Ruben_substrateMT_trinity_genecalling.faa -o
annotation_DRAM --threads 240 --verbose --use_uniref

###obs for this step
https://github.com/WrightonLabCSU/DRAM/issues/62
dbcan-CAZy uses filtering indicated by them - dbCAN2 suggestions
for thresholds" suggested by dbcan2:
(see http://bcb.unl.edu/dbCAN2/blast.php)
E-Value < 1e-15, coverage > 0.35



#summarize results
DRAM.py distill -i annotation_DRAM2/annotations.tsv -o
annotation_DRAM2/distilled


# next time run with FOAM-hmm_rel1a.hmm database



#########################----TAXONOMY NCBI-JGI----
#################################


####################################
# JGI FUNGAL PROTEINS - BIOCEV PC #
####################################


/home/kdanielmorais/bioinformatics/tools/diamond blastp -d
/mnt/DATA/DATABASES/FUNGAL_PROTEINS_JGI/JGI_FUNGAL_PROTEINS_ANNOT
ATED_20210312 -q
genecalling/Ruben_substrateMT_trinity_genecalling.faa -e 1E-5 -o
taxonomy/genecalling_JGI_FUN_20210312.txt -f 6 -p 256 -b12 -c1

export LANG=en_US.UTF-8
export LC_ALL=en_US.UTF-8
```

```
sort -t$'\t' -k1,1 -k12,12gr -k11,11g -k3,3gr
genecalling_JGI_FUN_20210312.txt | sort -u -k1,1 --merge >
genecalling_JGI_FUN_20210312_best.txt


# GENERA DEFINED   ##########RUNNING THIS -- started on
20.07.2022 at
/home/kdanielmorais/bioinformatics/tools/diamond blastp -d
/mnt/DATA/DATABASES/NCBI_nr_DIAMOND/NCBI_nr_20210225_diamond_GENE
RA -q CLEMENTINE_MT_Trinity_genecalling_fgs.faa -e 1E-5 -o
genecalling_NCBI_nr_PROTEINS_GENERA.txt -f 6 -p 512 -b12 -c1
##
Total time = 7076.34s
Reported 27886590 pairwise alignments, 27886590 HSPs.
1229412 queries aligned.
##

export LANG=en_US.UTF-8
export LC_ALL=en_US.UTF-8
sort -t$'\t' -k1,1 -k12,12gr -k11,11g -k3,3gr
genecalling_NCBI_nr_PROTEINS_GENERA.txt | sort -u -k1,1 --merge >
genecalling_NCBI_nr_GENERA_PROTEINS_best.txt



##################################
# ADD TAXONOMY TO BLAST RESULTS #
##################################

# JGI
# latest 20210406
gdrive_download 12c28kgIw4mPBIhQutNGladdAXwNLtvlR
jgi_abr_org_list.txt

gdrive_download 1XBTtiC1JYl2rzeV7idN2WrveEZknmnQi
replace_fungal_annot_by_taxname.py

python2.7 replace_fungal_annot_by_taxname.py
genecalling_JGI_FUN_20210312_best.txt jgi_abr_org_list.txt
genecalling_JGI_FUNGAL_PROTEINS_best_reformate.txt

awk -F'\t' '{print $2}'
genecalling_JGI_FUNGAL_PROTEINS_best_reformate.txt | sort | uniq
> FUNGAL_NAMES.txt


# NCBI
```

```
awk -F'\t' '{print $2}'
genecalling_NCBI_nr_GENERA_PROTEINS_best.txt | sort | uniq >
ALL_ACCESSIONS.txt

#gdrive_download 1FQdQ2Oh3sgyc2IKymYz_B6mKQwlklia_
retrieve_taxonomy_by_accession_with_taxid_library.py
#python2.7 retrieve_taxonomy_by_accession_with_taxid_library.py
ALL_ACCESSIONS.txt
/mnt/DATA/DATABASES/ACC2TAXID/ACC2TAXID_nr_current.txt
taxa_all_accessions.txt

https://drive.google.com/file/d/1o8KmSbwzOsjjeouK3dR0RNmWkMjdfFow
/view?usp=sharing

gdrive_download 1o8KmSbwzOsjjeouK3dR0RNmWkMjdfFow
get_taxonomy_offline.py

python2.7 get_taxonomy_offline.py ALL_ACCESSIONS.txt
/mnt/DATA/DATABASES/NCBI_nr_DIAMOND/ACC2TAXID_nr_current.txt
/mnt/DATA/DATABASES/NCBI_nr_DIAMOND/TAXONOMY_TAXID_ALL_fixed.txt
taxa_all_accessions.txt

accession list loaded...(779978)
taxonomy loaded...(907158)
taxonomy retrieved... 779978 vs acc (779978) - should be equal!
Done :]


# REFORMAT
https://drive.google.com/file/d/1jQ3F3ZuA0sBJVy3eJiRhwAaxh31LKqSF
/view?usp=sharing

gdrive_download 1jQ3F3ZuA0sBJVy3eJiRhwAaxh31LKqSF
replace_acc_by_sp_from_taxonomy.py

python2.7 replace_acc_by_sp_from_taxonomy.py
genecalling_NCBI_nr_GENERA_PROTEINS_best.txt TAXONOMY_ALL.txt
genecalling_NCBI_GENERA_PROTEINS_best_reformat.txt
#
number of taxa: 779978 (779978)
DONE :) Processed blast: 1229412 - NOT FOUND 0
#

# COMBINE TAXONOMY TABLES
# GET TAXONOMY FOR ALL
gdrive_download 1VtSyy7OutKZ6fAZMTYY2HkZUDNcpfY6V
JGI_TAXA_TAB_2021.txt
#https://drive.google.com/file/d/1F5p28LpaHrSYWwNI_82V9eHoKIb63y8
G/view?usp=sharing
```

```
gdrive_download 1F5p28LpaHrSYWwNI_82V9eHoKIb63y8G
combine_taxonomy_tables.py

python2.7 combine_taxonomy_tables.py FUNGAL_NAMES.txt
JGI_TAXA_TAB_2021.txt TAXONOMY_ALL.txt TAX_TAB.tab
##
names loaded...
FUNGAL TAXONOMY PROCESSED - NOT FOUND 0 vs. FOUND 1498
OTHER TAXONOMY PROCESSED - REDUCING TO 23625 vs. ORIGINAL 779979
DONE :)
###

####
gdrive_download 1-3XE5Le8I1_HzQdWbaAHev4ZlrSs6lUi
get_best_hit_by_bitscore_multi.py

python2.7 get_best_hit_by_bitscore_multi.py
genecalling_NCBI_GENERA_PROTEINS_best_reformat.txt
genecalling_JGI_FUNGAL_PROTEINS_best_reformate.txt
#
FILE: genecalling_NCBI_GENERA_PROTEINS_best_reformat.txt - HITS:
1229412
NEW ANNOTATIONS: 1229412 - REPLACED: 0 - CURRENT BEST HITS:
1229412

FILE: genecalling_JGI_FUNGAL_PROTEINS_best_reformate.txt - HITS:
714548
NEW ANNOTATIONS: 12375 - REPLACED: 308679 - CURRENT BEST HITS:
1241787
#

awk -F'\t' '{print $2}' best_of_the_blast.txt | sort | uniq >
ALL_TAXA_NAMES.txt

gdrive_download 1XruvN2qGN2-dUHZNSn0jXOYmUxoJ3Uz0
get_taxonomy_basedonnames.py

python2.7 get_taxonomy_basedonnames.py ALL_TAXA_NAMES.txt
TAX_TAB.tab TAX_TAB_FINAL.tab
#names loaded...
TAXONOMY PROCESSED - NOT FOUND 0 vs. FOUND 24671
DONE :)
#
awk -F'\t' '{print $1"\t"$12"\t"$2""}' best_of_the_blast.txt >
TAXONOMY_BEST_OF_SIMPLE.txt

##############################
# KOGG FROM JGI-MYCO-GENOMES #
```

```
##############################

# name e-val KOG
awk -F'[|\t]' '{print $1"\t"$15"\t["$5"]"}'
genecalling_JGI_FUN_20210312_best.txt >
FUNCTION_JGI_KOG_SIMPLE.txt

##########---- SPLITING FOR FOAM -----#########


############
# split it #
############

gdrive_download 1mGbdx3OBumymosW24WaYfZT9nq_a1z1z
split_fasta_by_group_size.py

python2.7 split_fasta_by_group_size.py
/mnt/DATA1/priscila/rubenMT/filtered/rRNA_remove/genecalling/Rube
n_substrateMT_trinity_genecalling.faa 83000

mkdir SPLIT
mv *.fas SPLIT

########
# FOAM #
########
cd SPLIT
for file in *.fas
do
   output=${file%%.fas}
   echo "/home/kdanielmorais/bioinformatics/tools/hmmer-
3.0/src/hmmsearch --tblout ${output}.txt --noali --cpu 1 -E 1e-5
/mnt/DATA/DATABASES/FOAM_db/FOAM-hmm_rel1.hmm ${file} >/dev/null
2>&1"
done > foam.sh

mkdir tmp
cat foam.sh | parallel --tmpdir tmp

##### PROCESS OUTPUT ######

for file in *.txt
do
 sample=${file%%.txt}
 grep -v '#' ${file} | awk -F' ' '{print $1"\t"$3"\t"$5"\t"$6}' >
${sample}.for_sort.txt
done
```

```
export LC_ALL=en_US.UTF-8
export LANG=en_US.UTF-8

for file in *.for_sort.txt
do
 sample=${file%%.for_sort.txt}
 echo "sort -t$'\t' -k1,1 -k4,4gr -k3,3g ${file} | sort -u -k1,1
--merge > ${sample}.sorted_best.txt"
done > sort.sh

cat sort.sh | parallel

cat *.sorted_best.txt > FOAM_BEST.txt

############################################################
# FOAM ANNOTATION
############################################################

gdrive_download 1ckuIqWVVarFgtcEUQOne2z-TdkDO3aXU
FOAM_simple_multi_from_raw.py

python2.7 FOAM_simple_multi_from_raw.py FOAM_BEST.txt
FUNCTION_FOAM_KO_SIMPLE_MULTI.txt


#######################################
#  USE dbCAN LOCAL DATABASE - CONDA #
#######################################

#First activate the dbcan environment with
conda activate run_dbcan

for file in *.fas
do
 sample=${file%%.fas}
 mkdir ${sample}
done

for file in *.fas
do
  sample=${file%%.fas}
  echo "run_dbcan.py ${file} protein --db_dir
/mnt/DATA/DATABASES/run_dbcan_master/db/ -t hmmer --out_dir
${sample} --hmm_cpu 1 --dia_cpu 1"
done > dbcan.sh
cat dbcan.sh | parallel
```

```
echo "" > all_dbCAN.txt
for file in *.fas
do
 sample=${file%%.fas}
 wc -l ${sample}/hmmer.out
 cat ${sample}/hmmer.out >> all_dbCAN.txt
done

#########################################################
# dbCAN ANNOTATION
#########################################################

export LC_ALL=en_US.UTF-8
export LANG=en_US.UTF-8

sort -t$'\t' -k3,3 -k5,5g all_dbCAN.txt | sort -u -k3,3 --merge >
all_dbCAN_best.txt

awk -F'[.\t]' '{print $1}' all_dbCAN_best.txt | sort | uniq >
hmm_names_uniq.txt

awk -F'[.\t]' '{print $1}' all_dbCAN_best.txt > hmm_names.txt

awk -F'\t' '{print $3"\t"$5}' all_dbCAN_best.txt >
all_dbCAN_best_gene_eval.txt

paste -d"\t" all_dbCAN_best_gene_eval.txt hmm_names.txt >
CAZy_BEST_SIMPLE.txt


#####################################
# LINK ANNOTATION TO TABLE
#####################################

gdrive_download 198TDGsV1cBfLEZorb5znFHysG47XEj5t
link_simple_table_to_mapping_table.py

#python2.7 link_simple_table_to_mapping_table.py
mapping_table_normalised_per_sample_genecall.txt
best_of_the_blast_simple.txt BESTTAX bitscore
MAPTAB_NORMPERSAMPLE_GENES_BESTTAX.ta
#mv MG_CLEMENTINE_normalised_per_sample.txt_genecall.txt
TABLE_NORM_SAMPLES_GENECALL.txt

TABLE="TABLE_NORM_SAMPLES_GENECALL.txt"

echo "${TABLE}"

TABLE_BASE=${TABLE%%.${TABLE##*.}}
```

```
echo "${TABLE_BASE}"

python2.7 link_simple_table_to_mapping_table.py ../${TABLE}
TAXONOMY_BEST_OF_SIMPLE.txt TAX_BEST bitscore
${TABLE_BASE}_TAX.txt
Functions processed... 1241787
Done... used functions: 1241787/1241787

python2.7 link_simple_table_to_mapping_table.py
${TABLE_BASE}_TAX.txt CAZy_BEST_SIMPLE.txt CAZy e-val
${TABLE_BASE}_TAX_CAZy.txt
Functions processed... 10697
Done... used functions: 10696/10697

python2.7 link_simple_table_to_mapping_table.py
${TABLE_BASE}_TAX_CAZy.txt FUNCTION_FOAM_KO_SIMPLE_MULTI.txt FOAM
e-val ${TABLE_BASE}_TAX_CAZy_FOAM.tab
Functions processed... 295860
Done... used functions: 295860/295860

python2.7 link_simple_table_to_mapping_table.py
${TABLE_BASE}_TAX_CAZy_FOAM.tab FUNCTION_JGI_KOG_SIMPLE.txt KOG
e-val ${TABLE_BASE}_TAX_CAZy_FOAM_KOG.tab
Functions processed... 714548
Done... used functions: 714548/714548
# genus taxonomy

https://drive.google.com/file/d/1AWuqqPaP2rUMpF_uMHOGEs8aE3Iy7JS2
/view?usp=sharing

gdrive_download 1AWuqqPaP2rUMpF_uMHOGEs8aE3Iy7JS2
add_higher_taxonomy.py

#python2.7 add_higher_taxonomy.py
TABLE_NORM_SAMPLES_GENECALL_TAX_CAZy_FOAM_KOG.tab TAX_tree.tab
TAX_BEST TABLE_NORM_SAMPLES_GENECALL_TAX2_CAZy_FOAM_KOG.tab
TAX_tree_genus.tab

python2.7 add_higher_taxonomy.py
TABLE_NORM_SAMPLES_GENECALL_TAX_CAZy_FOAM_KOG.tab
TAX_TAB_FINAL.tab TAX_BEST
TABLE_NORM_SAMPLES_GENECALL_TAX2_CAZy_FOAM_KOG.tab
TAX_tree_genus.tab

>>>duplicate<<<
new: Eukaryota  Nematoda        Chromadorea     Rhabditida
Setariidae      [Setaria]
```

```
old: Eukaryota  Streptophyta    Magnoliopsida   Poales  Poaceae
[Setaria]
>>>duplicate<<<
new: Eukaryota  Mucoromycota    Mucoromycetes   Mucorales
Lichtheimiaceae [Fennellomyces]
old: Eukaryota  Mucoromycota    Mucoromycetes   Mucorales
Syncephalastraceae      [Fennellomyces]
>>>duplicate<<<
new: Eukaryota  Basidiomycota   Agaricomycetes  Polyporales
Meruliaceae     [Phlebia]
old: Eukaryota  Basidiomycota   Agaricomycetes  Polyporales
Steccherinaceae [Phlebia]
>>>duplicate<<<
new: Eukaryota  Basidiomycota   Agaricomycetes  Agaricales
Tricholomataceae        [Infundibulicybe]
old: Eukaryota  Basidiomycota   Agaricomycetes  Agaricales
undefined Agaricales    [Infundibulicybe]
>>>duplicate<<<
new: Eukaryota  Ascomycota      Saccharomycetes Saccharomycetales
Trichomonascaceae       [Blastobotrys]
old: Eukaryota  Ascomycota      Saccharomycetes Saccharomycetales
Trigonopsidaceae        [Blastobotrys]
>>>duplicate<<<
new: Eukaryota  Ascomycota      Saccharomycetes Saccharomycetales
Debaryomycetaceae       [Candida]
old: Eukaryota  Ascomycota      Saccharomycetes Saccharomycetales
undefined Saccharomycetales     [Candida]
>>>duplicate<<<
new: Eukaryota  Ascomycota      Eurotiomycetes  Eurotiales
Thermoascaceae  [Paecilomyces]
old: Eukaryota  Ascomycota      Sordariomycetes Hypocreales
Clavicipitaceae [Paecilomyces]
>>>duplicate<<<
new: Eukaryota  Ascomycota      Dothideomycetes Pleosporales
Cucurbitariaceae        [Pyrenochaeta]
old: Eukaryota  Ascomycota      Dothideomycetes Pleosporales
Neopyrenochaetaceae     [Pyrenochaeta]
>>>duplicate<<<
new: Bacteria   Calditrichaeota Calditrichae    Calditrichales
Calditrichaceae [Caldithrix]
old: Bacteria   Calditrichaeota Calditrichia    Calditrichales
Calditrichaceae [Caldithrix]
done :]


######################
#unique dbCAN models #
```

```
awk -F'\t' '{print $3}' CAZy_BEST_SIMPLE.txt | sort | uniq >
CAZy_BEST_unique.txt

https://drive.google.com/file/d/1SGVK2cqWCLozEPGNLvPRs0YG-
ckrF_CZ/view?usp=sharing

gdrive_download 1SGVK2cqWCLozEPGNLvPRs0YG-ckrF_CZ
get_CAZy_tree.py

python2.7 get_CAZy_tree.py CAZy_BEST_unique.txt CAZy_tree.tab

#############
# KOG TREE #
#############
gdrive_download 1uoUDoD5El-Gdlv9VNQ6BGATtacGsq2MF
KOG_TAB_2021_03_03.txt


##################
### KofamKOALA ###  use 0.00001 for e-val
##################
in
"/mnt/DATA1/priscila/rubenMT/filtered/rRNA_remove/kofam_KOs/kofam
_koala_KOs_rubenMT.txt"
/mnt/DATA1/priscila/kofamKOALA/bin/kofam_scan-
1.3.0/exec_annotation
../genecalling/Ruben_substrateMT_trinity_genecalling.faa -o
kofam_koala_KOs_rubenMT.txt -p
/mnt/DATA1/priscila/kofamKOALA/db/profiles/ -k
/mnt/DATA1/priscila/kofamKOALA/db/ko_list --cpu=120 --tmp-dir=tmp
-E 1e-5 -f detail-tsv

###extract KOs and evals
python2.7 KO_simple_tab_from_koala.py kofam_koala_KOs_rubenMT.txt
KO_simple_table.txt

#add this to the bigTable

python2.7 link_simple_table_to_mapping_table.py
TABLE_NORM_SAMPLES_GENECALL_TAX2_CAZy_FOAM_KOG.tab
kofam_KOs/KO_simple_table.txt  KEGG e-val
TABLE_NORM_SAMPLES_GENECALL_TAX2_CAZy_FOAM_KOG_KEGG.tab

#############
# KEGG TREE #
#############
gdrive_download 11CVgwqy6O2mJ5rc4vxevYQVp04oJrQsl kegg_tab.txt
gdrive_download 1AmOiMqLHE8nberbvYEDPT12JShAfSW_w
GET_KEGG_ontology_subtable.py
```

```
gdrive_download 1aqENUDPh3dCoxnO8YfBZaUjwwxkmlMM1
KO_UNIQUE_from_KO_simple.py
cat kofam_KOs/KO_simple_table.txt
FUNCTION_FOAM_KO_SIMPLE_MULTI.txt > KEGG_ALL_MULTI.txt

python2.7 KO_UNIQUE_from_KO_simple.py KEGG_ALL_MULTI.txt
python2.7 GET_KEGG_ontology_subtable.py kegg_tab.txt
KEGG_ALL_MULTI.txt.unique.txt FOAM_KEGG_tree.tab
```

ANNEX 7: Annex 3: Complete pipeline of the MAGs (Metagenome-Assembled Genomes) analysis carried out in Chapter 3.

```
###################
## MAGs ANALYSIS ##
###################

# The analysis of MAGs (Metagenome-Assembled Genomes) is an approach used in
metagenomics to reconstruct complete genomes of microorganisms present in an
environmental
# sample, without the need for prior isolation in culture. Using raw
metagenomic sequences, MAGs are obtained by an assembly and binning process,
# in which contigs (DNA fragments) are grouped into bins representing
individual genomes.
# These genomes can come from bacteria, archaea or other microbes present in
the sample.

#######
## 1 ##
#######

#############
## BINNING ##
#############

# Binning is a key step in metagenomic analysis. The main objective of
binning is to group contigs (assembled DNA fragments) into bins,
# where each bin represents a possible individual genome.

conda activate metawrap

metawrap binning -a /mnt/DATA/belen/MAGS_assembly_chapter3/final.contigs.fa -
o binning_metawrap -t 120 -m 1000 --metabat2 --maxbin2 --concoct --universal
--run-checkm --interleaved /mnt/DATA/belen/MAGS_assembly/*.pe.qc.fq.gz

#######
## 2 ##
#######

####################
## BIN REFINEMENT ##
####################
```

```
# In this step you pick the best version of each bin. You can be more or less stringent in this step by lowering the completeness a bit.

metawrap bin_refinement -o /mnt/DATA/belen/MAGS_assembly_chapter3/metawrap_refined_bins/ -A /mnt/DATA/belen/MAGS_assembly_chapter3/binning_metawrap/metabat2_bins/ -B /mnt/DATA/belen/MAGS_assembly_chapter3/binning_metawrap_concot/concoct_bins/ -C /mnt/DATA/belen/MAGS_assembly_chapter3/binning_metawrap/maxbin2_bins/ -m 1000 -t 120 -c 50 -x 10

#######
## 3 ##
#######

##################
## CheckM2 STEP ##
##################

# CheckM2 is used to evaluate the quality of the refined bins (MAGs) obtained. CheckM2 is a tool that estimates the completeness and contamination of MAGs.

conda activate checkm2

checkm2 predict -i /mnt/DATA/belen/MAGS_assembly_chapter3/metawrap_refined_bins/metawrap_50_10_bins/ -x fa --output-directory refinded_checkm2 --database_path /mnt/DATA1/priscila/checkm2/database/CheckM2_database/uniref100.KO.1.dmnd --tmpdir ./ --threads 240

awk '$2 >= 50 && $3 <=10' refinded_checkm2/quality_report.tsv > good_bins_checkm2.tsv

awk '$2 >= 50 && $3 <=10' refinded_checkm2/quality_report.tsv | cut -f1 > bins_list

for i in $(cat bins_list); do cp metawrap_50_10_bins/$i.fa selected_bins/ ;done

#######
## 4 ##
#######

############################
## TAXONOMIC ANNOTATION ##
############################

# GTDB-Tk (Genome Taxonomy Database Toolkit) is used to assign taxonomy to refined MAGs using the GTDB database version 2.4.0 (v220).
# This tool classifies microbial genomes from complete genomic data and provides a standardized taxonomy based on the phylogenetic tree proposed by GTDB.

conda activate /mnt/DATA1/priscila/condaenvs/gtdbtk220
```

```
gtdbtk classify_wf --genome_dir
/mnt/DATA/belen/MAGS_assembly_chapter3/metawrap_refined_bins/metawrap_50_10_b
ins/ --out_dir refined_checkm2_gtdb220 --cpus 240 --pplacer_cpus 60 -x .fa --
tmpdir ./ --skip_ani_screen

#######
## 5 ##
#######

##########
## GUNC ##
##########

# GUNC (Genomic UNcertainty Calculator), a tool designed to assess the
taxonomic contamination and consistency of MAGs, is used.
# This analysis is crucial to verify the quality of the refined MAGs and
ensure that they represent unique and consistent genomes rather than
# mixtures of genetic material from different organisms.

conda activate gunc

export TMPDIR="/mnt/DATA/projects/priscila/tmp/"
echo $TMPDIR

mkdir selected_bins_gunc

gunc run --input_dir
/mnt/DATA/belen/MAGS_assembly_chapter3/metawrap_refined_bins/metawrap_50_10_b
ins/ --detailed_output --contig_taxonomy_output --use_species_level --out_dir
selected_bins_gunc --threads 120 --db_file
/mnt/DATA1/priscila/database/gunc_db_progenomes2.1.dmnd --file_suffix .fa

#######
## 6 ##
#######

#####################################
## RELATIVE QUANTIFICATION OF MAGs ##
#####################################

# The relative quantification of MAGs is performed using the Minimap2 and
CoverM tools.
# In this step, the relative abundance of each MAG in the microbial community
is determined based on the mapping of MAGs.

mkdir mags_bams

conda activate coverm

export TMPDIR="/mnt/DATA/projects/priscila/tmp/"
export TMPDIR="/mnt/DATA/belen/MAGS_assembly_chapter4"

echo $TMPDIR
```

```
coverm genome --mapper minimap2-sr --methods relative_abundance -o
coverm_relative_abundance_selected.txt --bam-file-cache-directory mags_bams -
-interleaved /mnt/DATA/belen/MAGS_assembly_chapter4/*.pe.qc.fq.gz --genome-
fasta-directory
/mnt/DATA/belen/MAGS_assembly_chapter4/metawrap_refined_bins/metawrap_50_10_b
ins/ -x fa --threads 240

#######
## 7 ##
#######

############################
## FUNCTIONAL ANNOTATION ##
############################

conda activate DRAM

# We have
lrwxrwxrwx. 1 belen belen   32 Dec 11 17:39 gtdbtk.ar53.summary.tsv ->
classify/gtdbtk.ar53.summary.tsv
lrwxrwxrwx. 1 belen belen   34 Dec 11 18:29 gtdbtk.bac120.summary.tsv ->
classify/gtdbtk.bac120.summary.tsv

# Combine both files
head -n 1
/mnt/DATA/belen/MAGS_assembly_chapter4/metawrap_refined_bins/refined_checkm2_
gtdb220/gtdbtk.bac120.summary.tsv > gtdbtk_summary_bac_arch.tsv
tail -n +2
/mnt/DATA/belen/MAGS_assembly_chapter4/metawrap_refined_bins/refined_checkm2_
gtdb220/gtdbtk.bac120.summary.tsv >> gtdbtk_summary_bac_arch.tsv
tail -n +2
/mnt/DATA/belen/MAGS_assembly_chapter4/metawrap_refined_bins/refined_checkm2_
gtdb220/gtdbtk.ar53.summary.tsv >> gtdbtk_summary_bac_arch.tsv

# Functional annotation
DRAM.py annotate -i
'/mnt/DATA/belen/MAGS_assembly_chapter4/metawrap_refined_bins/metawrap_50_10_
bins/*.fa' \
-o dram_annotation/ \
--min_contig_size 2000 \
--gtdb_taxonomy
/mnt/DATA/belen/MAGS_assembly_chapter4/metawrap_refined_bins/refined_checkm2_
gtdb220/gtdbtk_summary_bac_arch.tsv \
--checkm_quality
/mnt/DATA/belen/MAGS_assembly_chapter4/refinded_checkm2/quality_report.tsv \
--threads 512 \
--verbose \
--kofam_use_dbcan2_thresholds \
--keep_tmp_dir

cd /mnt/DATA/belen/MAGS_assembly_chapter4/dram_annotation
```

```
DRAM.py distill -i
/mnt/DATA/belen/MAGS_assembly_chapter4/dram_annotation/annotations.tsv -o
genome_summaries --trna_path
/mnt/DATA/belen/MAGS_assembly_chapter4/dram_annotation/trnas.tsv --rrna_path
/mnt/DATA/belen/MAGS_assembly_chapter4/dram_annotation/rrnas.tsv

#######
## 8 ##
#######

################################
## RELATIVE ABUBDANCE OF MAGs ##
################################

# In this step, quantification of the relative abundance of MAGs in each
sample is performed using the Salmon tool.
# The aim is to determine how many reads from each sample map to the
different MAGs, which provides information on the relative abundance of the
assembled genomes
# in the different samples.

conda activate metawrap

metawrap quant_bins2 -a
/mnt/DATA/belen/MAGS_assembly_chapter4/final.contigs.fa -o quantified_bins -t
240 -b
/mnt/DATA/belen/MAGS_assembly_chapter4/metawrap_refined_bins/metawrap_50_10_b
ins/ /mnt/DATA/belen/MAGS_assembly_chapter4/binning_metawrap/work_files/*.bam

#############################
## MAGs ANALYSIS COMPLETED ##
#############################
```