scientific data

OPEN

DATA DESCRIPTOR

Genome-Resolved Metagenomics of Rice Straw Degradation Experiments in Colombian Fields

Jeferyd Yepes-García[1,2], Nicolás Novoa-Montenegro[3], Vanessa Otero-Jiménez[4], Daniel Uribe-Vélez[3], Emiliano Barreto-Hernández[5] & Laurent Falquet[1,2] ✉

Extensive rice harvesting yields more than 800 million tons of rice straw (RS) per year globally, generating a byproduct that is often difficult for farmers to manage efficiently without burning it. In the quest for enhanced RS degradation systems, we recovered 146 Metagenome-Assembled Genomes (MAGs) from experiments aiming at decomposing RS. Such assays included the application of a *Bacillus* strain, a *Trichoderma*-based commercial product, organic and inorganic compounds in different combinations during a solid-state fermentation in Colombian rice fields. The set of MAGs comprises 30 MAGs from bulk soil and 116 MAGs from RS surface, for which taxonomic classification indicates that 67% of them may constitute novel taxa. Furthermore, functional analysis through different approaches suggests the presence of both mid-quality and high-quality MAGs with potential to biotransform RS within this dataset. Finally, these MAGs represent a valuable resource for exploring uncharacterized microbial diversity in Colombian agricultural ecosystems.

Background & Summary

The management of rice straw (RS), a rice agroindustry byproduct, has become one of the most challenging issues worldwide that demands innovative strategies for its proper disposition since farmers usually burn this material, resulting in a significant contribution to global warming[1]. In general, the molecules that compose RS, mainly cellulose (24–38%), hemicellulose (25–27%), lignin (12–13%), ashes and silica, are arranged in a complex and rigid spatial configuration[2,3]. Notwithstanding, the RS degradation, similar to the decomposition of other agro-industrial residues, represents a complex process governed by different factors, including microbial community composition, nutrient availability, and physicochemical parameters such as pH and temperature[4]. The microbial composition effect on RS degradation has been studied through various strategies, including composting[5], the use of swine-industry waste[6], and the application of commercial products to promote its biotransformation[7].

Moreover, the metagenomics study of undergoing biological RS-degradation systems may reveal potential novel microorganisms actively participating in the breakdown of RS components that can be promoted in order to obtain higher biotransformation rates[8]. Specifically, MAG reconstruction allows the combination of functional and taxonomic annotation to generate organized and defined models characterizing specific metabolic activity while the RS biotransformation is ongoing[9]. This complementary approach has been successful in recent reports, given the availability of well-organized resources such as the Carbohydrate-Active Enzyme (CAZy or CAZyme) database[10], which was developed to compile and systematically organize carbohydrate-activated enzymes based on biological and structural features they exhibit. For instance, different ecological or experimental environments have been the subject of research during analysis related to the contribution of CAZymes to marine sediment biodiversity[11], the presence of CAZymes derived from composting samples[12], and the global distribution of carbohydrate utilization potential[13].

In this work, we report a total of 146 MAGs recovered from experiments in Colombian rice fields conducted to biotransform RS in mulching setups. According to the sampling procedure, 116 MAGs were built

[1]Department of Biology, University of Fribourg, Canton of Fribourg, Fribourg, 1700, Switzerland. [2]Swiss Institute of Bioinformatics, Lausanne, Vaud, 1015, Switzerland. [3]Agricultural Microbiology Group, Instituto de Biotecnología, Universidad Nacional de Colombia, A.A 14-490, Bogotá D.C., Colombia. [4]Department of Soil and Water Systems, University of Idaho, 875 Perimeter Drive MS2340, Moscow, ID, 83844-2340, USA. [5]Bioinformatics group, Instituto de Biotecnología, Universidad Nacional de Colombia, A.A 14-490, Bogotá D.C., Colombia. ✉e-mail: laurent.falquet@unifr.ch

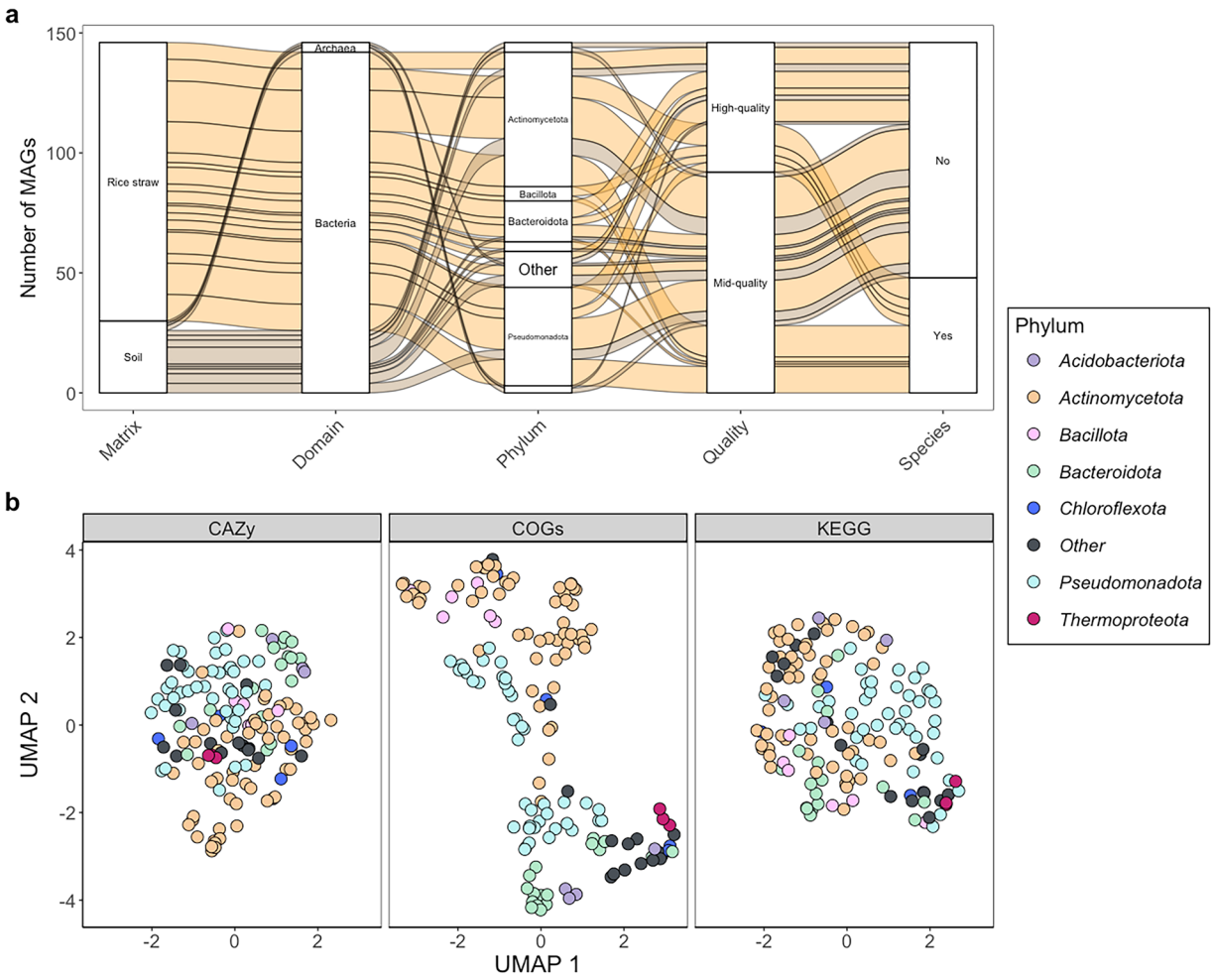

Fig. 1 (**a**) Alluvial plot showing the number of MAGs belonging to each matrix, taxonomic affiliation (at the domain and phylum levels), quality assessment (CheckM2) and identification as species or not. The exact order of the missing phylum labels within the (**a**) plot can be read from the legend. Phylum Other includes: Armatimonadota, Cyanobacteriota, Deinococcota, Desulfobacterota_E, Gemmatimonadota, Methanobacteriota, Myxococcota, Patescibacteria, Tectomicrobia and Verrucomicrobiota. (**b**) Two-dimensional Uniform Manifold Approximation and Projection (UMAP) embedding of normalized CAZy enzymes counts, normalized COG counts and KEGG biogeochemically-relevant pathway completeness per MAG, colored by phylum (GTDB-Tk2). UMAP representation was achieved using default values (15 neighbors and Euclidean distance).

from the RS surface and 30 from bulk soil (Fig. 1a). In terms of quality, 37% of the MAGs were identified as high-quality (HQ) MAGs and the remaining 63% as mid-quality (MQ) MAGs. In addition, 33% of the MAGs were annotated at species level, belonging all of them to the RS surface; meanwhile, none of the MAGs recovered from soil was classified at this taxonomic level, and even in some cases the annotation was only possible to the order level. Moreover, the dimensionality reduction achieved by the Uniform Manifold Approximation and Projection (UMAP) generates sets of MAGs with similar patterns for CAZy annotation, Clusters of Orthologous Genes (COG) counts and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway profiles, mainly for Actinomycetota and Pseudomonadota (Fig. 1b). This fact relies on the dominance by the members of these phyla within this dataset, $n = 56$ in the case of Actinomycetota and $n = 41$ for Pseudomonadota (Fig. 2b). According to this UMAP representation, MAGs belonging to Acidobacteriota and Bacteroidota depict similar COG and CAZy patterns as they are placed close to each other in both dimensionality-reduced plots. The complete taxonomic affiliation for each MAG as well as the complete COG normalize counts, KEGG pathway completeness, CAZy information and quality metrics are provided (Supplementary Table S1).

From the general COG annotation (Fig. 2a), functions such as amino acid transport and metabolism (E), carbohydrate transport and metabolism (G), translation (J) and transcription (K) appear to be above-the-mean across the Bacillota MAGs. Likewise, on average, Acidobacteriota MAGs encompass a higher number of copies of sequences with hypothetical biological activity (R) and genes related to signal transduction mechanisms (T) and defense mechanisms (V).

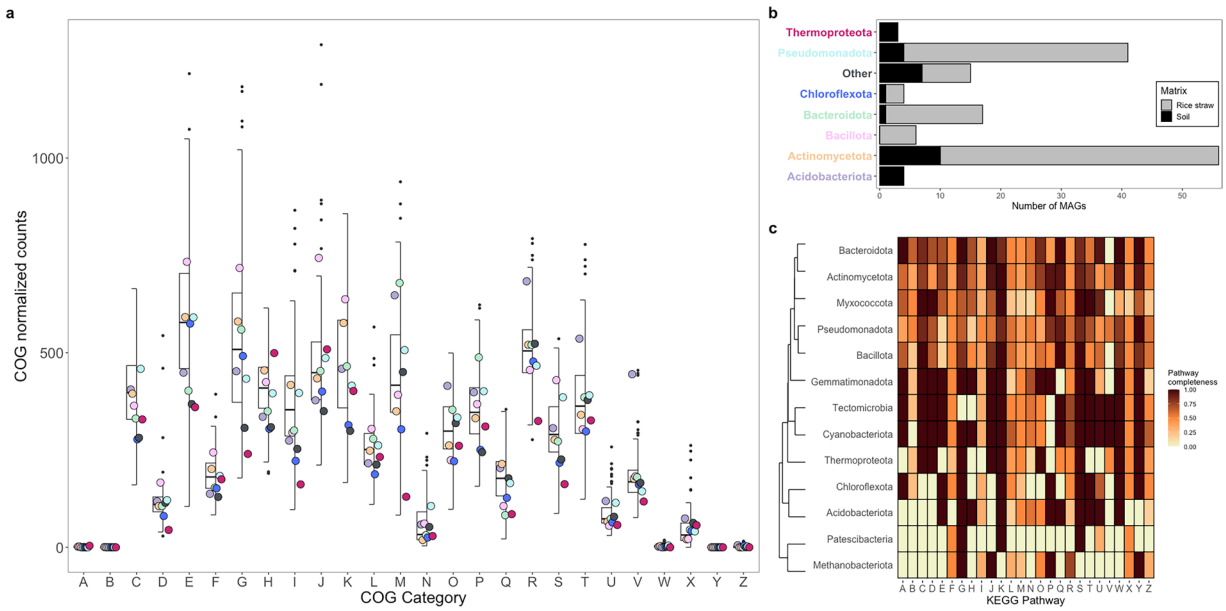

Fig. 2 (**a**) Distribution of normalized COG counts per category considering all 146 MAGs, where colored data points represent the average value for the category of the MAGs belonging to a specific phylum; the colors of the y axis labels in (**b**) map with the color of the data points. (**b**) Number of MAGs per phylum discriminated by their origin: rice straw or soil. (**c**) Heatmap for pathway completeness of KEGG biogeochemically-relevant metabolic pathways in each phylum. The completeness of the pathway is based on the presence or absence of genes as determined by KEGG Decoder. Only the 26 most complete pathways across MAGs are shown. For both ***a*** and ***c***, the full letter assignment is available (Table 1).

Letter	COG category	KEGG Pathway
A	RNA processing and modification	Alanine
B	Chromatin structure and dynamics	Anaplerotic Genes
C	Energy production and conversion	Arginine
D	Cell cycle control, cell division, chromosome partitioning	Asparagine
E	Amino acid transport and metabolism	Beta-Glucosidase
F	Nucleotide transport and metabolism	Bidirectional Polyphosphate
G	Carbohydrate transport and metabolism	Copper Transporter Copa
H	Coenzyme transport and metabolism	Cysteine
I	Lipid transport and metabolism	Glutamine
J	Translation, ribosomal structure and biogenesis	Glycine
K	Transcription	Isoleucine
L	Replication, recombination and repair	Leucine
M	Cell wall/membrane/envelope biogenesis	Mep-Doxp Pathway
N	Cell motility	Methionine
O	Post-translational modification, protein turnover, chaperon	Mixed Acid: Acetate
P	Inorganic ion transport and metabolism	Mixed Acid: Ethanol, Acetate to Acetaldehyde
Q	Secondary metabolites biosynthesis (transport/catabolism)	Proline
R	General function prediction only	Riboflavin Biosynthesis
S	Function unknown	Serine
T	Signal transduction mechanisms	Starch/Glycogen Degradation
U	Intracellular trafficking, secretion, and vesicular transport	Starch/Glycogen Synthesis
V	Defense mechanisms	Sulfur Assimilation
W	Extracellular structures	Threonine
X	Mobilome: prophages, transposons	Tryptophan
Y	Nuclear structure	Tyrosine
Z	Cytoskeleton	Valine

Table 1. Letter assignment for the 26 COG categories and the 26 most complete KEGG pathways user for visualization on Fig. 2a,c.

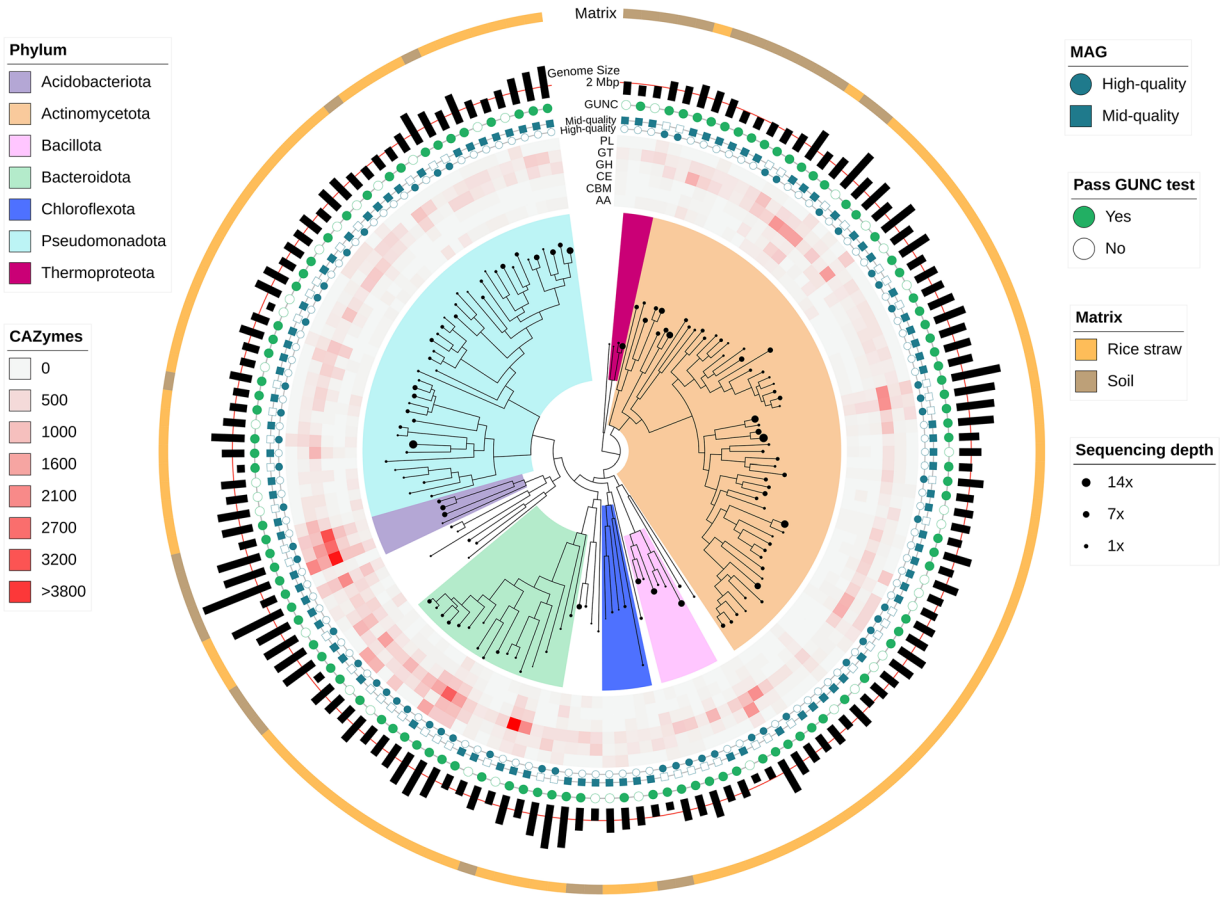

Fig. 3 Phylogenomic organization of the 146 MAGs generated by GTDB-Tk2. The visualization includes phylogenetic relationship among them (tree), specific highlighted phylum clades (colored sectors), MAG sequencing depth (tips of the branches), Carbon Active EnZyme (CAZyme) class normalized abundance (red scaled heatmap), quality assessment as either high-quality or mid-quality MAG based on checkM2 measurements (blue binary), GUNC test passing (green binary), and genome size in Mbp (outer black bars). Conventions in CAZyme heatmap: GH glycoside hydrolases, GT glycosyl transferases, PL polysaccharide lyases, CE carbohydrate esterases, CBMs carbohydrate-binding modules, AA auxiliary activities.

Meanwhile, considering 26 most complete KEGG pathways among all 146 MAGs (Fig. 2c), starch/glycogen degradation (T) pathway appears to be complete within MAGs classified as Thermoproteota, Deinococcota, Myxococcota, Gemmatimonadota, Chloroflexota, Cyanobacteriota, Armatimonadota and Verrucomicrobiota. These former three phyla exhibit an allegedly complete set of genes to synthesize β-glucosidases (E), a key enzyme in the last step of cellulose degradation. The letter assignment for COG categories and KEGG pathways is included (Table 1).

Regarding some general MAG features, the genome size ranges from 1 Mbp to 12 Mbp, even within MAGs belonging to the same phylogenomic cluster (Fig. 3). Further, some specific Pseudomonadota and Actinomycetota MAGs appear to be overrepresented within their respective samples as their sequencing depth reached values up to 14x (Fig. 3). Aside from quality assessment in terms of completeness and contamination to label the MAGs as HQ or MQ, 73% of the MAGs passed the GUNC test[14], a methodology developed to establish the level of chimerism and contamination within metagenomics assemblies. As a result, the MAGs that do not pass the test may incorporate notable numbers of contaminant sequences, a fact verified by the MAG classification based on CheckM2 as almost all the MAGs (33 out of 39) that did not pass the GUNC test belong to MQ MAGs (Fig. 3).

MAGs belonging to the phyla Acidobacteriota and Bacteroidota exhibited the highest concentration of CAZymes. Specifically, polysaccharide lyases (PLs), glycosyltransferases (GTs) and glycoside hydrolases (GHs) are especially abundant on members of the Bacteroidota phylum, mainly in MAGs annotated as *Sphingobacterium*, *Siphonobacter*, *Spirosoma* and *Pararcticibacter* (Fig. 3). Particularly, the MAG *FBaNitSoil0_7* (classified as *Draconibacterium*) depicts high numbers of GHs genes, the module in which most enzymes related to cellulose metabolism are encompassed (Fig. 4). A similar pattern is observed with members of the phylum Acidobacteriota, who enclose superior quantities of GHs, GTs and PLs, *i.e.*, the MAG *FBaNitSoil1_23*.

In summary, this dataset enables the recognition of probable previously unknown microorganisms, *i.e.*, MAGs belonging to both phyla Bacteroidota and Acidobacteriota, albeit without proper taxonomic assignment

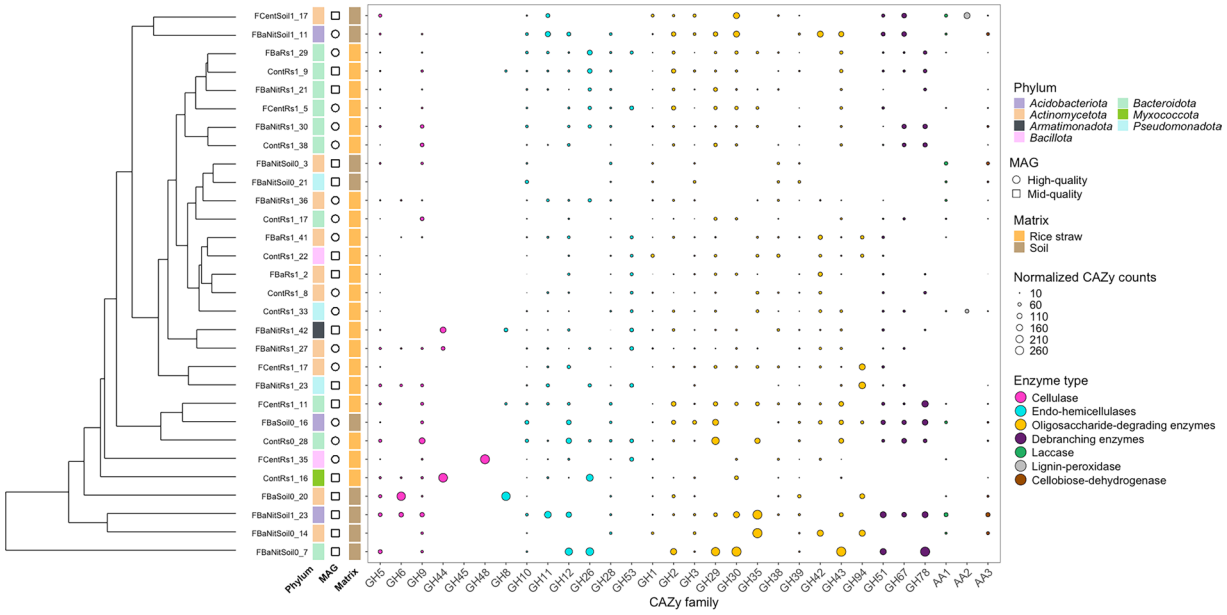

Fig. 4 Bubble plot depicting the number of normalized counts for CAZyme families associated with lignocellulose-degrading enzymes found in the MAGs recovered from RS and soil. Visualization includes: MAG clustering based on normalized CAZyme counts of the selected families, phylum information (GTDB-Tk2), labeling as MQ MAG or HQ MAG (CheckM2), and matrix provenance. The displayed MAGs contain at least 30 normalized counts in 5 or more CAZyme families; complete information about normalized CAZyme counts is provided (Fig. S1, *Supplementary Information*).

at species level. In addition, several of these MAGs show a potential to degrade RS as they contain a broad diversity of CAZymes related to the breakdown of agricultural subproducts, including enzyme families involved in lignocellulose bioconversion. Finally, these findings highlight the intrinsic potential of environmental niches such as the bulk soil or RS to contain novel species, and to act as a reservoir for genes with a wide variety of metabolic activity, and it contributed to unveil the Colombian soil and biological material unexplored microbial diversity.

Methods

Area of study. The field assays were carried out in rice fields owned by the Rice Program of the Alliance for Biodiversity and the International Center for Tropical Agriculture (Alliance Biodiversity and CIAT) Km17 straight Cali - Palmira (N3° 29.81539 W76° 21.59739). The plot was divided into and evaluated using furrows measuring 1 m wide by 15.5 m long. The subdivisions were fully randomized to host the different treatments (*see* below), with five replicates per treatment. Prior to experiment establishment, the rice was harvested by hand and chopped with a scythe.

Biological material and chemical components. Several approaches to decompose RS were assayed using different combinations of microorganisms and inorganic/organic compounds. The amount of RS used for the experiments was estimated considering an RS accumulation of 3 tons per hectare according to the yield of paddy rice harvested in this location. This estimate was calculated by extrapolating the values given by the collection and weighing of RS contained in a 50 × 50 cm square in five different places in the field. The applied treatments were composed by the addition of some of the following components: (i) the commercial product Fitotripen® manufactured by Natural Control S.A. (La Ceja, Antioquia Colombia) was utilized; this fungal mixture encompasses three *Trichoderma* strains, namely *T. harzianum*, *T. koningi*, and *T. viridae*, and it is formulated as a powder that must be diluted in water; (ii) *Bacillus altitudinis* strain indexed as IBUN 2717 in the Bacterial Strain Repository of the Biotechnology Institute of the National University of Colombia; this is a native strain isolated from the rice rhizosphere in Colombia, and it has been used in previous studies aiming at biotransforming RS[15,16]; (iii) Centurion®, a carbon amendment for agricultural waste degradation described as an enhancer for RS degradation under field conditions; and (iv) inorganic nitrogen in urea form (commercial granulate).

The *Bacillus* strain required a pre-activation step, where it was grown in Luria Bertani (LB) broth media for 96 hours at 120 rpm at 30 °C until its sporulation. The biomass was washed with distilled sterile water (DSW) and heated at 80 °C for 10 minutes to eliminate vegetative cells and retain only the spores. The resulting spore suspension was diluted to a concentration of 1×10^7 spores.mL^{-1} and tested using the microdrop technique described by Bautista *et al.*[17].

Field experiments. The composition of the applied treatments was as follows (Table 2): treatment *FBa*, a suspension was prepared using the *Bacillus altitudinis* strain with Fitotripen® (*Trichoderma* sp.); treatment

Treatment	RS setting	Fitotripen® (*Trichoderma* sp.)	*B. altitudinis* IBUN 02717	Inorganic nitrogen	Centurion® (organic carbon)
Cont	Mulching	No	No	No	No
FBa	Mulching	Yes	Yes	No	No
FBaNit	Mulching	Yes	Yes	Yes	No
Fcent	Mulching	Yes	No	No	Yes

Table 2. Treatment composition for the experiments carried out at field scale in Colombian rice plots.

FBaNit, the same composition as treatment *FBa* in combination with inorganic nitrogen; and treatment *FCent*, only the *Trichoderma* consortia was used plus 5 mL of the commercial product Centurion® diluted in 1 L of distilled sterile water (DSW). The final concentration of the microorganisms was adjusted to 1×10^6 conidia.mL^{-1} of the *Trichoderma* consortia, and 1×10^7 spores.mL^{-1} of *B. altitudinis* in a final volume of 1 L in DSW. In the case of treatment *FBaNit*, the urea was applied aiming at achieving a theoretical carbon/nitrogen (C/N) ratio of 35:1, and therefore a dose of 56 g of urea equivalent to 37.3 kg.ha^1 was sprayed into the respective plots (15.5 m^2) after dilution in DSW. Control (Cont) plots with the same dimensions as the treatment plots were sprayed with 1 L of DSW.

Following treatment application, the plots were left undisturbed in the field for 30 days under ambient temperature and rainfall. Soil and RS samples were collected at two-time points: before treatment application (time *0*, *t0*) and 30 days later (time *1*, *t1*). At *t0*, the RS samples were taken from a single bulk source, which was then subsampled in specific amounts to initiate the experiments, thus resulting in one sample (with five replicates) at *t0*.

Five spots per plot were selected for soil sampling, each separated by 2.5 m within each plot. Samples were taken using a tube probe (2.54 cm of diameter) that was inserted 5 cm into the soil, and they were placed in a Ziplock® bag and stored in a portable freezer. For RS sampling, two points were selected, each separated by 5 m from the edge of the plot. Before laboratory analysis, soil and RS samples from each replica were combined into a compound sample per plot (biological replica). This was performed in a single run for all samples (time *0* and *1*) to avoid introducing effects related to external variables. According to this experimental design, 5 RS samples and 8 soil samples were assembled, considering 5 replicates per sample, for a total of 65 sequencing libraries.

DNA extraction and sequencing. RS DNA isolation was conducted in accordance with the methodology delineated by Su *et al.*[18]. The procedure entailed washing 1 g of straw in a phosphate-buffered saline (PBS) buffer at 120 rpm for one hour, followed by two sonication rounds for one minute each at 30 kHz. The PBS suspension was subsequently subjected to centrifugation at 15,000 rpm for 10 minutes, resulting in the isolation of an RS pellet. For soil and RS pellet Power Soil Pro Kit® (Qiagen, Hilden, Germany) was employed. The sequencing libraries were prepared with the TruSeq Nano DNA Kit along with the protocol TruSeq Nano DNA Sample Preparation Guide (Part # 15041110 Rev. D) for a paired-end layout of 151 base pairs. This service was provided by macrogen® (South Korea), where Illumina NovaSeq. 6000 was used as the sequencing system.

Metagenomic assembly and MAG reconstruction. According to the bioinformatics flow aimed to reconstruct and annotate MAGs (Fig. 5), the quality of the raw sequences was established using FastQC v0.12.1, and the reports were concatenated by MultiQC v1.16[19]. After sequence integrity verification, draft bins were built using the nf-core/mag[20] v2.3.2. The following parameters were included during nf-core/mag launching: `--busco_auto_lineage_prok --kip_spades --host_genome IRGSP-1.0 --skip_concoct --megahit_options '--min-contig-len 1000'`. A co-assembly/co-binning strategy was chosen as an attempt to improve the quality and resolution of the bins[21], and hence all replicates belonging to the same sample were processed together. MEGAHIT[22] v.1.2.9 was used for the assembly, while MaxBin2[23] v2.2.7 and MetaBAT2[24] v2.15 were the selected tools for binning. A refinement step was carried out by enabling DASTool[25] v1.16 within the workflow (`--refine_bins_dastool`). nf-core/mag outcomes included the mapping of the MAGs against the original reads, a piece of information used to show the sequencing depth of each MAG.

Once the bins were generated, a downstream study was performed leveraging on the quality metrics generated by the pipeline MAGFlow/BIgMAG[26] v1.1.0; this workflow includes quality-measuring tools such as CheckM2[27] v1.0.1, BUSCO[28] v5.7.0, GUNC[14] v1.0.6 and MetaQUAST[29] v5.2.0, as well as the taxonomic annotator GTDB-Tk2[30] v2.4.0. The versions of the databases required by this pipeline were Genome Taxonomy Database (GTDB) r220, GUNC reference database (based on proGenomes 2.1) and CheckM2 uniref100.KO.1. Mid and high-quality MAGs were filtered for subsequent analysis using the following criteria[31]: mid-quality MAGs (MQ) with completeness $>=50\%$ and contamination $<10\%$, and completeness $>90\%$ and contamination $<5\%$ for high-quality (HQ) genomes. Moreover, the MAGs obtained from both matrices were grouped into a single set, and to obtain a set of non-redundant MAGs at species level, MAGs with Average Nucleotide Identity (ANI) $\geq 95\%$ were clustered using FastANI[32] v1.33. The representative MAG from each cluster was selected based on the score formula proposed by Christoph *et al.*[33]: completeness – 0.5 × contamination.

The resulting MAGs were annotated with Prokka[34] v1.14.6, and the coding sequences from each MAG were then subjected for COG detection by COGclassifier[35] v1.0.5 (https://github.com/moshi4/COGclassifier); EC numbers were transformed into KEGG Orthology (KO) numbers for completeness estimation of biogeochemically-relevant metabolic pathways was determined by KEGG Decoder[36] v1.0.8.2 (https://github.com/bjtully/BioData/tree/master/KEGGDecoder).

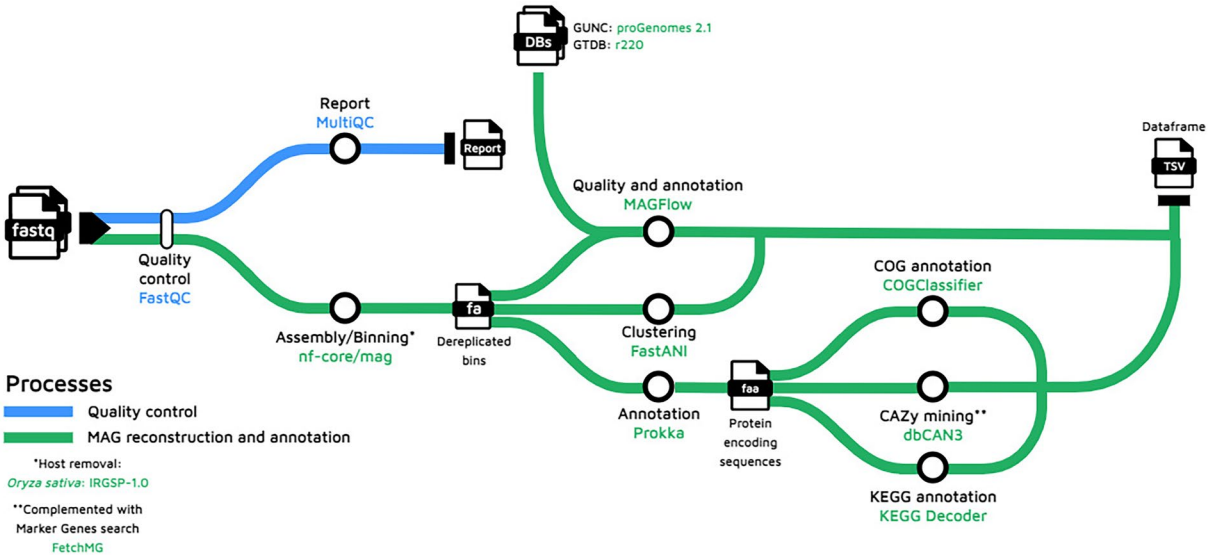

Fig. 5 Bioinformatics workflow followed to evaluate sequence quality and perform MAG recovery and annotation.

The number of sequences classified into each COG functional category per MAG were normalized following the formula (Eq. 1):

$$Normalized\ COG\ counts = \left[\frac{N°of\ COG\ category\ in\ the\ MAG}{N°of\ coding\ sequences\ in\ the\ MAG}\right] * 10^4 \quad (1)$$

Similarly, the presence of carbohydrate-related metabolism enzymes (CAZyme modules) was studied through dbCAN3[37] v4.1.4; a module was considered as present if it was reported by either Diamond alignment or HMMER sequence analysis. CAZyme modules related to lignocellulosic biomass decomposition (including β-glucosidase enzymes) were established as followed by Santos-Pereira *et al.*[12], and additional CAZyme families such as cellobiose-dehydrogenases, laccases and lignin-peroxidases were included.

A normalization procedure of the CAZyme counts per MAG across samples was carried out as implemented by López-Sánchez *et al.*[11]. The equation includes the estimation of the median of Marker genes (MG) per sample with FetchMGs[38] v1.3 (https://github.com/motu-tool/FetchMGs) to be used as normalization factor (Eq. 2):

$$Normalized\ CAZy\ counts = \left[\frac{\frac{N°\ of\ the\ CAZyme\ module\ in\ the\ MAG}{N°\ of\ the\ CAZyme\ module\ in\ the\ metagenome\ sample}}{Median\ MGs\ in\ metagenome\ sample}\right] * 10^4 \quad (2)$$

Finally, a concatenated taxonomic tree of the MAGs belonging to each matrix was built using GTDB-Tk2 through the command `'gtdbtk de_novo_wf'` and the previously generated taxonomy classification via MAGFlow. Further, this information was cross-referenced with quality metrics data, COG annotation and CAZyme normalized counts to integrate the activity analysis of each genome. Separately, the CAZyme normalized-count table, the COG normalized count data and KEGG pathway completeness were used as input for the UMAP methodology to identify clusters of MAGs sharing carbohydrate-related overall presence of enzymes, GOG general patterns and/or KEGG pathway global trends among MAGs. A variety of plots were generated with iTOL[39] v7.0 and R in-house scripts that included the package ggplot2 v.3.5.1.

Data Records

The raw shotgun metagenomics sequences (65 experiments and 1 mock community) can be found under the Sequence Read Archive (SRA) study SRP563356[40] (SRR32316194-SRR32316259). The 146 MAGs were submitted to the GenBank, and they are accessible through the BioProject PRJNA1222139[41], with accession numbers JBMDWG000000000-JBMDYH000000000 and JBQGBX000000000-JBQGFK000000000. MQ and HQ MAGs are also available on figshare[42] and Zenodo[43] (https://doi.org/10.5281/zenodo.16947911), along with their corresponding annotation files; mock community information and metadata can be retrieved from the same repositories.

Technical Validation

A mock community (ZymoBIOMICS™ Microbial Community DNA Standard D6305) was included as an additional sample during the extraction and sequencing procedures to check the quality of both experimental and computational workflows. The concentrations and quality of the extracted DNA were estimated using a Qubit 4 Fluorometer dsDNA BR Assay Kit (Life Technologies, Paisley, UK) and gel electrophoresis.

The presence of sequences belonging to the genera *Bacillus* and *Trichoderma* was tested by following the approach proposed by Lu *et al.*[44], where raw reads after host removal (*O. zativa*, NCBI RefSeq accession number: GCF_000005425.2) with Bowtie2[45] were taxonomically classified with Kraken2[46], and the species abundance was estimated through Bracken[47]. The database used for classification was PlusPFP (Standard plus RefSeq protozoa, fungi and plant, released on 04.09.2024 at https://benlangmead.github.io/aws-indexes/k2), and the resulting Bracken files were merged in BIOM[48] format to generate a Phyloseq[49] (v1.42.0) object afterwards. This Phyloseq object was used to estimate the relative abundance of *Bacillus* and *Trichoderma* considering the total of the classified sequences (Fig. S2, *Supplementary Information*).

Data availability

The dataset is available at NCBI GenBank with accession numbers JBMDWG000000000-JBMDYH000000000 and JBQGBX000000000-JBQGFK000000000. The raw reads can be found also at the NCBI Sequence Read Archive (SRA) under the study SRP563356[40]. Additional files and scripts are available on figshare[42] and Zenodo[43] (https://doi.org/10.5281/zenodo.16947911).

Code availability

Scripts used to run the different tools are available at the Zenodo[43] repository (https://doi.org/10.5281/zenodo.15483380), as well as software outputs and knitr/executable R downstream-analysis script. An online instance of BIgMAG[26] showing quality metrics and taxonomic annotation of the clustered MAGs is accessible at https://bigmag.online/.

Received: 23 May 2025; Accepted: 6 October 2025;
Published online: 19 November 2025

References

1. Singh, G., Gupta, M. K., Chaurasiya, S., Sharma, V. S. & Pimenov, D. Y. Rice straw burning: a review on its global prevalence and the sustainable alternatives for its effective mitigation. *Environ. Sci. Pollut. Res.* **28**, 32125–32155, https://doi.org/10.1007/s11356-021-14163-3 (2021).
2. Van Hung, N. *et al.* in Sustainable Rice Straw Management (eds. Gummert, M., Hung, N. V., Chivenge, P. & Douthwaite, B.) Ch. 1, https://doi.org/10.1007/978-3-030-32373-8 (Springer International Publishing, Cham, 2020).
3. Rosado, M. J., Rencoret, J., Marques, G., Gutiérrez, A. & del Río, J. C. Structural Characteristics of the Guaiacyl-Rich Lignins From Rice (*Oryza sativa* L.) Husks and Straw. *Front. Plant Sci.* 12, https://doi.org/10.3389/fpls.2021.640475 (2021).
4. Singh, G. & Arya, S. K. A review on management of rice straw by use of cleaner technologies: Abundant opportunities and expectations for Indian farming. *J. Clean. Prod.* **291**, 125278, https://doi.org/10.1016/j.jclepro.2020.125278 (2021).
5. Zang, X. *et al.* The structural and functional contributions of β-glucosidase-producing microbial communities to cellulose degradation in composting. *Biotechnol. Biofuels* **11**, 1–13, https://doi.org/10.1186/s13068-018-1045-8 (2018).
6. Wang, J. *et al.* Effect of acclimatized paddy soil microorganisms using swine wastewater on degradation of rice straw. *Bioresour. Technol.* **332**, 125039, https://doi.org/10.1016/j.biortech.2021.125039 (2021).
7. Do, C. V. T. *et al.* Recovery of carbon from rice straw for production of high-value products by *Chlorella sorokiniana* TH01 through mixotrophic cultivation. *Biochem. Eng. J.* **197**, 108966, https://doi.org/10.1016/j.bej.2023.108966 (2023).
8. Sarma, S. *et al.* Rapid decomposition of rice straw by application of a novel microbial consortium and study its microbial community dynamics. *World J. Microbiol. Biotechnol.* **38**, 1–15, https://doi.org/10.1007/s11274-022-03399-x (2022).
9. Kim, N. *et al.* Genome-resolved metagenomics: a game changer for microbiome medicine. *Exp. Mol. Med.* **56**, 1501–1512, https://doi.org/10.1038/s12276-024-01262-7 (2024).
10. Lombard, V., Ramulu, H. G., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42, https://doi.org/10.1093/nar/gkt1178 (2014).
11. López-Sánchez, R. *et al.* Metagenomic analysis of carbohydrate-active enzymes and their contribution to marine sediment biodiversity. *World J. Microbiol. Biotechnol.* **40**, 1–15, https://doi.org/10.1007/s11274-024-03884-5 (2024).
12. Santos-Pereira, C. *et al.* Functional and sequence-based metagenomics to uncover carbohydrate-degrading enzymes from composting samples. *Appl. Microbiol. Biotechnol.* 1–23, https://doi.org/10.1007/s00253-023-12627-9 (2023).
13. López-Mondéjar, R., Tláskal, V., Rocha, U. N. da & Baldrian, P. Global Distribution of Carbohydrate Utilization Potential in the Prokaryotic Tree of Life. *mSystems* 7, https://doi.org/10.1128/msystems.00829-22 (2022).
14. Orakov, A. *et al.* GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol.* **22**, 1–19, https://doi.org/10.1186/s13059-021-02393-0 (2021).
15. Cruz, C., Gomez, L. & Uribe, D. Manejo biológico del tamo de arroz bajo diferentes relaciones C: N empleando co- inóculos microbianos y promotores de crecimiento vegetal. *Rev Colomb Biotecnol* **19**, 47–62, https://doi.org/10.15446/rev.colomb.biote.v19n2.70168 (2017).
16. Otero-Jiménez, V. *et al.* Impact of rice straw management strategies on rice rhizosphere microbiomes. *Appl. Soil Ecol.* **167**, 104036, https://doi.org/10.1016/j.apsoil.2021.104036 (2021).
17. Bautista, J. P., Barbosa, H. & Uribe-Vélez, D. Prototipo de formulación a base de *Rhodotorula mucilaginosa* para el control de *Botrytis cinerea* en Rosas. *Rev. Colomb. Biotecnol.* **18**, 13–23, https://doi.org/10.15446/rev.colomb.biote.v18n2.55826 (2016).
18. Su, P. *et al.* Recovery of metagenome-assembled genomes from the phyllosphere of 110 rice genotypes. *Sci. Data* **9**, 1–7, https://doi.org/10.1038/s41597-022-01320-7 (2022).
19. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048, https://doi.org/10.1093/bioinformatics/btw354 (2016).
20. Krakau, S., Straub, D., Gourlé, H., Gabernet, G. & Nahnsen, S. nf-core/mag: a best-practice pipeline for metagenome hybrid assembly and binning. *NAR Genomics Bioinforma.* 4, https://doi.org/10.1093/nargab/lqac007 (2022).
21. Vosloo, S. *et al.* Evaluating de Novo Assembly and Binning Strategies for Time Series Drinking Water Metagenomes. *Microbiol. Spectr.* 9, https://doi.org/10.1128/Spectrum.01434-21 (2021).
22. Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3–11, https://doi.org/10.1016/j.ymeth.2016.02.020 (2016).
23. Wu, Y. W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607, https://doi.org/10.1093/bioinformatics/btv638 (2016).
24. Kang, D. D. *et al.* MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 2019, https://doi.org/10.7717/peerj.7359 (2019).

25. Sieber, C. M. K. *et al*. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843, https://doi.org/10.1038/s41564-018-0171-1 (2018).
26. Yepes-García, J. & Falquet, L. Metagenome quality metrics and taxonomical annotation visualization through the integration of MAGFlow and BIgMAG. *F1000Research* **13**, 640, https://doi.org/10.12688/f1000research.152290.2 (2024).
27. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212, https://doi.org/10.1038/s41592-023-01940-w (2023).
28. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: Assessing Genomic Data Quality and Beyond. *Curr. Protoc.* **1**, e323, https://doi.org/10.1002/cpz1.323 (2021).
29. Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* **32**, 1088–1090, https://doi.org/10.1093/bioinformatics/btv697 (2016).
30. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316, https://doi.org/10.1093/bioinformatics/btac672 (2022).
31. Bowers, R. M. *et al*. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731, https://doi.org/10.1038/nbt.3893 (2017).
32. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114, https://doi.org/10.1038/s41467-018-07641-9 (2018).
33. Christoph, M., Rühlemann, R., Wacker, E. M., Ellinghaus, D. & Franke, A. MAGScoT: a fast, lightweight and accurate bin-refinement tool. *Bioinformatics* **38**, 5430–5433, https://doi.org/10.1093/bioinformatics/btac694 (2022).
34. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069, https://doi.org/10.1093/bioinformatics/btu153 (2014).
35. Shimoyama, Y. Source code for: COGclassifier: A tool for classifying prokaryote protein sequences into COG functional category, https://github.com/moshi4/COGclassifier (2022).
36. Graham, E. D., Heidelberg, J. F. & Tully, B. J. Potential for primary productivity in a globally-distributed bacterial phototroph. *ISME J.* **12**, 1861–1866, https://doi.org/10.1038/s41396-018-0091-3 (2018).
37. Zheng, J. *et al*. dbCAN3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Res.* **51**, W115–W121, https://doi.org/10.1093/nar/gkad328 (2023).
38. Ruscheweyh, H. J. Source code for: FetchMGs. https://github.com/motu-tool/FetchMGs (2024).
39. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296, https://doi.org/10.1093/nar/gkab301 (2021).
40. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRP563356 (2025).
41. Yepes-García, J. *et al*. *Rice straw and bulk soil raw sequence reads*. https://identifiers.org/ncbi/bioproject:PRJNA1222139 (2025).
42. Yepes-García, J. *et al*. *Genome-Resolved Metagenomics of Rice Straw Degradation Experiments in Colombian Fields*. https://doi.org/10.6084/m9.figshare.291260391 (2025).
43. Yepes-García, J. *et al*. Genome-Resolved Metagenomics of Rice Straw Degradation Experiments in Colombian Fields (MAGs and Suppl. files). *Zenodo* https://doi.org/10.5281/zenodo.16947911 (2025).
44. Lu, J. *et al*. Metagenome analysis using the Kraken software suite. *Nat. Protoc.* **17**, 2815–2839, https://doi.org/10.1038/s41596-022-00738-y (2022).
45. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359, https://doi.org/10.1038/nmeth.1923 (2012).
46. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 1–13, https://doi.org/10.1186/s13059-019-1891-0 (2019).
47. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: Estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **2017**, e104, https://doi.org/10.7717/peerj-cs.104 (2017).
48. McDonald, D. *et al*. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience* **1**, 2047-217X-1–7, https://doi.org/10.1186/2047-217X-1-7 (2012).
49. McMurdie, P. J. & Holmes, S. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE* **8**, e61217, https://doi.org/10.1371/journal.pone.0061217 (2013).

Acknowledgements

This work was supported by the Centenary Research Fund of the University of Fribourg (FC-22-912 0857), and by the *Fondation de Recherche en Biochimie*, Epalinges, Switzerland. The authors acknowledge the assistance received by Dr. Maria Fernanda Alvarez, Marcela Pineda and Cristian Herrera from The Rice Program at the Alliance of Bioversity International and International Center for Tropical Agriculture (CIAT), who contributed to this work at allowing experimenting in their rice fields and helping during the assay execution. JYG specially thanks the Federal Commission for Scholarships for Foreign Students (FCS) for their support through the Swiss Government Excellence Scholarship.

Author contributions

J.Y.G. performed reference-based taxonomic profiling, M.A.G. assembly, annotation and curation, manuscript writing, and data integration, visualization and deposition. N.N.M. carried out field experiments, DNA extraction, quality control checks and manuscript writing. V.O.J. contributed to manuscript writing, reviewing and editing. D.U.V., E.B.H. and L.F. supervised and oversaw experiment development and data analysis. All authors participating during the study framework conceptualizing, conceiving the overall work and manuscript preparation.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-06113-2.

Correspondence and requests for materials should be addressed to L.F.