Continental-scale integration of soil metagenomes and metabolomes reveals ubiquitous microbial capacity for recalcitrant carbon decomposition

Young C. Song[1], Cheng Shi[1,2], Kelly G. Stratton[3], Izabel Stohel[1], Viviana Freire-Zapata[4], Malak M. Tfaily[4], Emiley Eloe-Fadrosh[5] and Emily B. Graham[3]

1. Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland WA 99354, United States
2. Eawag, Swiss Institute of Aquatic Science and Technology, Überlandstrasse 133 CH-8600 Dübendorf, Switzerland
3. Biological Sciences Division, Pacific Northwest National Laboratory, Richland WA 99354, United States
4. Department of Environmental Science, University of Arizona, Tucson, AZ 85719, United States
5. US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, United States

Abstract

Soil organic matter (SOM) decomposition by microorganisms represents the largest uncertainty in predicting terrestrial carbon-atmosphere feedbacks. Yet, we lack systematic understanding of the microbial taxa that drive decomposition of chemically distinct carbon pools and of the metabolic pathways they employ across environmental gradients. Here, we address this critical gap using the first continental-scale soil dataset pairing shotgun metagenomes with high-resolution molecular characterization of SOM chemistry. We use data from 47 standardized soil cores across the United States to assemble 0.76 terabases (Tbs) of prokaryotic metagenome-assembled genomes (344 MAGs) and identify 66,727 distinct SOM molecules, integrating these datasets to reveal widespread microbial potential for depolymerizing recalcitrant carbon compounds previously considered stable. Notably, we uncover complementary metabolic specialization between two abundant bacterial orders, *Rhizobiales* and *Chthoniobacterales.* This metabolic partitioning is consistent across soil depths and activity levels, suggesting coordinated decomposition of chemically complex carbon through distinct but complementary biochemical strategies. The metabolic potential for depolymerization of recalcitrant compounds is supported by the abundance of these molecules across the soils, as indicated by Fourier-Transform Ion Cyclotron Resonance Mass Spectrometry (FTICR-MS), and by flux balance analysis of metabolic models. Our continental-scale integration of microbial genotypes with their chemical substrates reveals that a substantial fraction of SOM previously considered recalcitrant harbors decomposition potential that current Earth system models fail to capture. These findings provide mechanistic insight for incorporating microbial metabolic pathways into carbon cycle predictions and highlight an underappreciated vulnerability of soil carbon to environmental variability.

Main

Soil microbial respiration is a critical component of the global carbon cycle, yielding 98 ± 12 Pg C/year of CO_2 into the atmosphere [1]. Most soil carbon is contained within chemically diverse soil organic matter (SOM) pools that, through their interactions with plants, mineral surfaces, water, and other biological and physical soil properties, have the potential to sequester carbon within soils for centuries [2]. Chemically recalcitrant SOM is typically thought to have long residence times in soils [3,4], and the widespread potential for its decomposition suggests a possible missing link in the global carbon cycle. Microbial metabolism of SOM proceeds through myriad metabolic processes [5], yet the interconnected factors and environmental contexts are largely unknown.

SOM contains a wide variety of organic molecules derived from plant-microbe interactions [5,6], biomass decomposition, and biochemical transformation into recalcitrant components [7,8]. These small (<1,000 Da) compounds, including amino acids, lipids and organic phosphates, are transformed by soil microorganisms into CO_2 through respiration [5]. While phylogenetic and metabolic inferences of soil metagenome-assembled genomes (MAGs) have revealed potential microbial mechanisms for transforming various organic molecules [9–11], robust characterization and confirmation of these pathways has historically depended on isolation and cultivation the organisms represented by MAGs [12]. In parallel, metabolomics approaches using Fourier-Transform Ion Cyclotron Resonance Mass Spectrometry (FTICR-MS) provide molecular-level information about SOM chemistry [5]. Integration of these approaches has the potential to improve our understanding of SOM transformations while circumventing cultivation bottlenecks.

Currently existing Earth system models consider large fractions of SOMs as carbon sinks [13], given the predicted residence times spanning decades to centuries. We challenge this notion, while addressing the knowledge gap regarding microbial decomposition of chemically distinct carbon pools across environmental gradients, using a continental-scale soil dataset that integrates standardized metagenomics with high-resolution molecular characterization of SOM chemistry. Here, we integrate MAGs and high-resolution metabolomes of surface and subsoils across Continental United States (CONUS) to reveal the ubiquitous potential for soil microbial

metabolism of chemically recalcitrant SOM. Through the 1000 Soils Pilot of the Molecular Observation Network (MONet) [14], we selected 47 soils exhibiting the highest versus lowest (30%) rates of respiration within surface and subsoils (independently) for further analysis. Using the 344 metagenome-assembled genomes (MAGs) and 66,727 SOM molecules acquired from these soils, our work points to complementary roles in continental-scale soil C cycling processes conducted by two predominant microbial clades, *Rhizobiales* and *Chthoniobacteriales*, supported by the abundance of complex carbohydrates across all soils. Our observations stress the vulnerability of the soil carbon stocks to increased microbial activity under continued environmental change, emphasizing the need for a tractable genetic framework for incorporating microbial mechanisms into Earth system models.

Results

We assessed variations in microbial pathways and distribution of SOM types across the 47 soils at two depths (surface and subsoil), spanning 37 geographically diverse sites spanning the CONUS (Supplementary Fig. 1 and Supplementary Table 1). High-resolution mass spectrometry (FTICR-MS) identified 66,727 distinct SOM molecules (average 7.6×10^3 molecules per sample), while metagenomic sequencing (average 1.1×10^8 reads per sample) yielded 344 metagenome-assembled genomes (MAGs) representing the dominant soil microbial communities. We focus on bacterial and archaeal communities owing to their substantial and extensively-examined roles in soil carbon cycling [15], though we acknowledge that other soil organisms such as fungi, viruses, and nematodes also have large impacts on soil biogeochemistry. Computational approaches for soil microbiomes allow for deeper investigation into bacteria and archaea than other clades, and we expect that the results presented here will lay a foundation for investigations into the role of interkingdom interactions on belowground biogeochemistry, as new technological advances are available. Detailed procedures of both SOM measurements and MAG assemblies are highlighted in the **Methods** section.

Prevalence of *Rhizobiales*, *Chthoniobacteriales,* and *Nitrososphaerales* across soil samples

We recovered 344 MAGs, 24 of which were high quality (completion > 90% with contamination < 5%) and 320 of which were medium quality (completion ≥ 50% with contamination of < 10%)

according to the Minimum Information about Metagenome-Assembled Genome (MIMAG; Supplementary Table 2) [16]. Phylogenetic inference of these 304 bacterial MAGs (143 from the surface and 161 from the subsoil) revealed that bacterial diversity was dominated by members of 3 major phyla (*Pseudomonadota*, *Actinomycetota* and *Acidobacteriota*; Supplementary Table 3). All 40 archaeal MAGs were classified as the phylum *Thermoproteota*, of which 92.5% (n=37) were associated with the order *Nitrososphaerales* (Supplementary Table 4).

Examining the microbial diversity at the order level revealed abundance of *Rhizobiales* and *Chthoniobacterales* across most of the soils under investigation. *Rhizobiales* was recovered from 46.8% (n=22) of all 47 soils (Fig. 1) and was also the most abundant lineage in both soil layers, comprising 18.9% (n=27) and 14.3% (n=23) of surface and subsoil MAGs, respectively (Fig. 1 and Supplementary Table 4). Members of *Rhizobiales* represented nearly 20.0% of all bacteria recovered in surface soils (both high- and low-respiration), as well as in high-respiration subsoils, but only 5.7% of bacterial MAGs from low-respiration subsoils. We also recovered 24 MAGs classified as *Chtnoniobacterales* in 29.7% (n=14) of all 47 soils (Fig. 1). *Chtnoniobacterales* made up approximately 7.0% (n=10) of MAGs recovered from surface soil, while constituting 8.7% (n=14) of MAGs identified in subsoil. This order was marginally more abundant in high-respiration soils in both soil layers, as compared to low-respiration soils. In each layer, the differences in the abundances of *Chthoniobacterales* across the respiration rates were also very small.

Soil chemistry reveals depth-stratified carbon pools with distinct decomposition signatures

Pairwise comparisons of SOM molecular composition across depths and potential rates of microbial respiration revealed systematic chemical differentiation with implications for carbon cycling (G-test of uniqueness; Fig. 2 and Supplementary Table 5). While lignin- and condensed hydrocarbon-like compounds were abundant across all soil types, we observed distinction in chemistry between surface and subsoils. For instance, comparison of chemistry in low-respiration subsoils with surface soils (both high- and low-respiration) revealed unique compound suites in the subsoils (G-test p-value < 0.05; Fig. 2C and E), of which > 70% were composed of lignin-, condensed-hydrocarbon- and tannin-like compounds.

Similarly, when comparing the chemistry of high-respiration subsoils against high-respiration surface soils, we observed chemistry unique to the subsoils, with ~80% being lignin-, condensed hydrocarbon-, and tannin-like compounds (Fig. 2B). Comparison between high-respiration subsoils and low-respiration surface soils (Fig. 2D) however, revealed that while 57% of the chemistry unique to the subsoils were lignin-, condensed hydrocarbon- and tannin-like compounds, 32.4% were protein- (16.2%) and unsaturated hydrocarbon- (16.2%) like compounds (Fig. 2 and Supplementary Table 5). Between subsoil environments (Fig. 2F), high-respiration subsoils contained 2,322 unique compounds (>75% lignin- and condensed hydrocarbon-like molecules), while low-respiration subsoils had only 18 unique compounds, half of which were lipid-like—potentially reflecting microbial necromass under carbon-limited conditions.

Microbial potentials for depolymerizing recalcitrant SOMs are widespread

Using co-occurrence networks between MAGs and SOM molecules acquired from each of the four depth/respiration partitions, we evaluated relationships between microorganisms and metabolites that may not be well-ascribed to established metabolic pathways in Kyoto Encyclopedia of Genes and Genomes (KEGG) [17]. In particular, we identified densely connected modules within a large SOM-KEGG Orthology (KO) network (for all soil layers and potential respiration rates), revealing genes that showed significant relationships with chemically-recalcitrant classes of SOM, such as lignin-, condensed hydrocarbon-, and tannin-like molecules. A case in point is the 7 modules with MCODE cluster scores between 2.7-6.0, describing potential microbe-metabolite relationships in high-respiration surface soils (Fig 3). In 6 of the 7 modules, Spearman correlations between all connected nodes were > 0.95 . The module with the highest cluster score of 6.0 revealed that the majority of the observed interactions were between KO-annotated genes, some of which were associated with degradation of complex SOMs (e.g., *benAB*) [18,19] and one or more of the lignin-, condensed-hydrocarbon- and tannin-like compounds (Supplementary Table 6). This cluster was formed by 11 nodes (6 SOM nodes and 5 gene nodes classified under “Carbohydrate metabolism” and “Glycan biosynthesis”) forming 30 edges. Of the 6 SOM nodes, three were lignin-like compounds, 2 were condensed hydrocarbon-like compounds, and the remaining node was an unsaturated hydrocarbon-like compound.

At the opposite end of our spectrum of soils, the module describing low-respiration subsoils was consistent with its unique chemistry profile, which included lipid-like compounds that were absent in the surface soils and in high-respiration subsoils (Fig. 3 and Supplementary Table 6). In this module, we observed nine separate interactions (all represented by edges with correlations of > 0.95) between a set of two lipid-like compounds and an unclassified molecule, and three KO nodes, two of which were associated with cellular functions (e.g., secretion system and molybdopterin biosynthesis).

Genes in the KEGG database that are associated with metabolites detected by FTICR-MS (see **Methods**) further supported the potential for widespread microbial interactions with the recalcitrant SOM pool and its intermediates [20,21]. For instance, we uncovered metabolites associated with the biosynthesis of cofactors (map01240), which constituted 9.7%, 11.3% and 12.2% of metabolite-associated pathways identified in high- and low-respiration surface soils and in high-respiration subsoil, respectively (Fig. 4). We also detected metabolites affiliated with the metabolism of xenobiotics (map00980) and degradation of aromatic compounds (map01220) — key pathways for processing complex organic polymers. These specialized metabolisms were distributed across all soil environments, constituting between 3.5% to 8.3% of metabolite-associated pathways in all groups of soils (Fig. 4).

Dominant soil bacteria encode complementary arsenals for depolymerizing complex SOM

Genomic analysis for carbohydrate-active enzymes (CAZymes) [22] revealed that the two widespread bacterial orders, as detected by MAGs, possessed extensive and complementary enzymatic machinery for processing recalcitrant carbon substrates. Consistent with the prevalence of complex molecules, as indicated by soil SOM profiles, MAGs associated with *Rhizobiales* and *Chthoniobacteriales* contained a variety of genes encoding complex carbon depolymerization (Fig. 5 and Supplementary Table 7). For instance, we detected 65 CAZyme classes in ≥ 50% of *Rhizobiales* MAGs, while 67 CAZyme classes were detected in ≥ 50% of *Chthoniobacterales* MAGs. In addition, 25 CAZyme classes were present in ≥ 50% of *Nitrososphaerales* MAGs (Fig. 5). Examining the distributions of the CAZyme classes revealed the prevalence of glycoside hydrolases (GH) and glycosyltransferases (GT) in all of the three lineages (Fig. 5 and Table 1).

While the total detected 65 *Rhizobiales* associated CAZymes were spread across the depths and respiration rates, *Chthoniobacterales* appeared to use mostly GH and GT. For instance, a unique set composed of 4 *Rhizobiale*s-associated GHs, a single AA and a CE was present in surface soils, while three AA families and four GH families were detected only in the *Rhizobiales* MAGs present in subsoils (Fig. 5). In contrast, 5 *Chthoniobacterales*-associated GH families, together with two GT families and an AA were detected only in surface soils, while a set of three GH families, two GT families and a PL were present in *Chthoniobacterales* MAGs detected in subsoils (Fig. 5). While there were no *Nitrososphaerales*-associated CAZymes unique to the subsoils, we detected a set composed of a single GH, GT and AA from the archaeal MAGs from surface soils.

The potential for amino acid catabolism in *Rhizobiales* and *Nitrososphaerales* suggests nitrogen acquisition from recalcitrant SOM

A high-level overview of metabolic traits present in *Rhizobiales* and *Chthoniobacterales* predicted via flux balance analysis (FBA) included their potential to use amino acids as source of carbon, nitrogen, and energy (Supplementary Fig. 2). A closer examination of the MAGs associated with these bacterial orders revealed a set of genes encoding branched-chain amino acids (i.e., L-leucine, L-isoleucine and L-valine) degradation pathways in 68% of *Rhizobiales* MAGs (n=34, Supplementary Table 8). Substrate preferences between these amino acids in *Rhizobiales* appeared to differ based on both soil layers and potential rates of microbial respiration. For instance, genes involved in L-leucine degradation (e.g., *accA*; *auh* and *hmgL*) were more abundant in low-respiration subsoils, while genes associated with L-valine and L-isoleucine degradation (e.g., *bcd*, *paaF* and *mmsAB*) were predominant in the other three soil types (Fig. 6). Unlike in the *Rhizobiales* MAGs, the branched-chain amino acid degradation pathways were not detected in *Chthoniobacterales* MAGs.

According to flux balance analysis, amino acid catabolism and biosynthesis (in particular, arginine) also appeared to be key pathways for carbon and nitrogen assimilation in *Nitrososphaerales* (Supplementary Fig. 2 and Supplementary Table 9). Nearly 37% (n = 11 MAGs) of *Nitrososphaerales* MAGs possessed genes involved in the conversion of arginine to the TCA cycle intermediate, α-ketoglutarate (Supplementary Table 9). The metabolic

dependence of *Nitrososphaerales* on amino acid catabolism-based energy generation was further supported by the presence of genes involved in ammonia oxidation, including *amoCAB* and *nirK*, in at least 25% of MAGs (Supplementary Table 9), although the gene *hao* was not detected in any archaeal MAG.

Lipid metabolism reveals contrasting survival strategies under carbon limitation

Distribution of the SOM classes across the soil types revealed that while microbial potential for depolymerizing recalcitrant organic matter appeared to be widespread in surface soils and in high-respiration subsoils, biomass synthesis pathways tended to dominate low-respiration subsoils. While the lipopolysaccharide biosynthesis genes (e.g., *lpxD* , *lpxL* and *gmhC*) [23,24] affiliated with *Rhizobiales* were abundant mostly in low-respiration subsoil compared to the other soils, the genes encoding the same pathway in *Chthoniobacterales* were more prevalent surface soils (Supplementary Table 8). Furthermore, the absence of some of the lipopolysaccharide synthesis genes (e.g., *lpxL*) in *Chthoniobacterales* suggested the fundamentally contrasting strategies in these dominant bacterial lineages for surviving carbon limitation.

Discussion

Our study challenges prevailing paradigms in environmental microbiology, soil biogeochemistry, and Earth system models that treat large fractions of soil organic matter (SOM) as biochemically recalcitrant [3,4]. We demonstrate that soil microbiomes across the continental United States possess widespread genetic capacity to depolymerize compounds traditionally considered stable. While focusing on soil bacterial communities as major contributors to SOM cycling, our study also highlights the need for investigating abiotic processes and bacterial interactions with the metabolic niches of fungi and viruses, which are harder to disentangle using current metagenomic methods [25–27]. By analyzing 344 MAGs and 66,727 molecular-level SOM measurements, we identified two dominant bacterial orders—*Rhizobiales* and *Chthoniobacteriales*—that use complementary enzymatic strategies to process chemically complex carbon across diverse soils. These findings suggest that slow carbon pools in current models may be more vulnerable to decomposition than previously thought, exposing a critical gap in carbon-climate feedback predictions.

Molecular measurements of SOM and functional annotation of soil MAGs demonstrate microbial potential for depolymerizing complex organic carbon, which may be linked to the deposition of plant-derived organic matter (Figs. 3 and 4) [28]. Microbial interactions with plant residues and decomposition products are shaped by factors such as the molecular composition and spatial arrangement of SOM, as well as the functional potential of the co-located microbiome [28–30]. The prevalence of complex carbohydrates across soil depths, regardless of microbial activity level, and their association with genes encoding SOM-degrading enzymes suggest current soil carbon stocks may be vulnerable to microbial activity under fluctuating conditions (Figs. 2, 3, and 4) [31,32]. Identifying genes such as *benAB*, directly linked to recalcitrant carbon processing, highlights microbial roles in SOM breakdown. Additionally, genes for polyketide synthesis and other secondary metabolites may contribute to SOM depolymerization through mechanisms like metal mobilization [33,34] and the activation of quorum sensing activation [35].

In addition to the ubiquity of *Rhizobiales* and *Chthoniobacterales* in all soils, the abundance of CAZymes in *Rhizobiales* and *Chthoniobacterales* MAGs across depth and respiration partitions suggests mechanisms for depolymerizing SOM during periods of both high and low microbial activity in surface and subsoils (Fig. 5). The prevalence of glycoside hydrolases (GHs) and glycosyltransferases (GTs), in particular, underscores their ability to interact with diverse carbohydrates—including glycans, monosaccharides, and polysaccharides—and to produce extracellular structures that facilitate adaptation to variable carbon availability [36,37]. Within high-respiration soils, the presence of auxiliary activity (AA) families in *Rhizobiales* indicates their potential to degrade complex carbohydrates, such as lignins, during periods of heightened microbial activity. In contrast, *Chthoniobacterales* rely predominantly on GHs and GTs for carbohydrate processing (Figs. 2 and 5). Together, these complementary strategies highlight the distinct roles of *Rhizobiales* and *Chthoniobacterales* in carbon cycling across diverse soil environments.

Metabolic inferences from flux balance analysis of *Rhizobiales* and *Nitrososphaerales* MAGs suggest that amino acid catabolism is critical for nitrogen acquisition from SOM and energy generation, particularly under conditions requiring bioavailable nitrogen to sustain the

depolymerization of complex carbon compounds (Fig. 6 and Supplementary Fig. 2) [38,39]. While the abundance of CAZyme classes within *Rhizobiales* MAGs highlights their role in degrading complex SOM in high-activity soils, their unique ability to utilize amino acids like L-leucine during periods of low microbial activity suggests amino acid catabolism supports energy homeostasis and nitrogen synthesis for cellular functions (Fig. 6) [40,41]. In contrast, amino acid transformation pathways were not detected in the *Chthoniobacterales* MAGs, suggesting a potential reliance on mono- and polysaccharides for carbon and energy metabolism (Fig. 6 and Supplementary Table 8) [42,43]. These findings suggest contrasting carbon usage strategies in *Rhizobiales* and *Chthoniobacteriales*, with *Rhizobiales* utilizing amino acids as reservoirs of both carbon and nitrogen and *Chthoniobacterales* specializing in carbohydrate metabolism. However, FBA-derived metabolic inferences assume microbes operate in a steady state, and the incomplete recovery of MAGs may limit interpretation, warranting validation of these inferred traits using transcriptomics and/or proteomics [44].

In *Nitrososphaerales*, amino acid catabolism appears to be coupled with ammonia oxidation. We observed that none of the *Nitrososphaerales* MAGs we detected contained the *hao* gene, which is thought to be an essential component of ammonia oxidation. The absence of *hao* may reflect sequencing limitations or suggest the involvement of uncharacterized Cu-protein complexes in ammonia oxidation, a mechanism observed in other ammonia-oxidizing archaea (Supplementary Table 9) [45]. Ammonia that is oxidized by *Nitrososphaerales* could be generated from catabolism of available amino acids in soil, such as arginine and glutamate, as suggested both by the flux balance analysis and by the conserved catabolic pathways of these amino acids in all 37 *Nitrosospaherales* MAGs (Supplementary Fig. 2 and Supplementary Table 9). These findings underscore the role of *Nitrososphaerales* in linking soil carbon and nitrogen cycles and highlight the need for further investigation into their genetic machinery for ammonia oxidation.

Lastly, the bioavailability of soil lipids has been a hotly debated topic in microbial ecology, because they may dually serve as a high-energy substrate and as a reservoir of sequestered carbon derived from microbial necromass [46–48]. We found a diversity of lipid-like molecules in low-respiration subsoils that may represent sequestered plant and/or microbial residues from prior periods of higher metabolic activity [49,50] (Fig. 2). In subsoils, nutrient limitations and/or

mineral protection likely promote the disproportionate accumulation of these lipid-like molecules, contributing significantly to the stable fraction of SOM. While lipid persistence has been linked to soil carbon stabilization, the universality of soil microbial potential for lipid metabolism — even in soils with low rates of measured metabolic activity — suggest that they may become susceptible to decomposition as environmental conditions change. Notably, interactions between lipid-like molecules and microbial genes in low-respiration soils also indicate that lipid-like necromass components may be vulnerable to microbial decomposition during periods of reduced carbon availability and respiration rates [50,51] (Fig. 3). Thus, we highlight potential lipid turnover as a key consideration in soil carbon dynamics [51].

Strong correlations between lipid-like compounds and microbial secretion systems, cofactors, and secondary metabolites in low-respiration subsoils may also elucidate strategies for microbial adaptation to carbon-limited environments (Figs. 3 and 4; Supplementary Table 6). Secretion systems enable microbial interactions by exporting compounds to facilitate nutrient acquisition, particularly labile substrates [52,53], while cofactors (e.g., molybdopterin) and secondary metabolites support energy metabolism, redox reactions, and oxidative damage mitigation under stress [54,55]. Distinct strategies between *Rhizobiales* and *Chthoniobacterales* further illustrate adaptation differences: *Rhizobiales* enhance lipopolysaccharide (LPS) synthesis to stabilize their cell envelope and improve resistance to environmental stress [56,57], whereas *Chthoniobacterales* rely on scavenging lipid necromass with minimal biosynthetic investment [58,59]. Gram-negative bacteria, including these orders, contribute significantly to organic carbon storage through the chemical persistence of LPS, particularly its recalcitrant lipid A component [58,60]. On cell death, released LPS acts as a binding agent between soil particles, enhancing soil stability and water retention [60]. These adaptations highlight how microbial communities persist and exploit resources under carbon deficiency while shaping key soil ecosystem processes like organic carbon cycling and storage (Figs. 1 and 2).

The multi-omics integration strategies employed in this study have facilitated the recovery of bacterial and archaeal lineages with unique yet complementary carbon sequestration strategies. Validating these inferred traits through cultivation methods, proteomics, transcriptomics, and/or emerging computational tools for co-assemblies [15,61] would enhance the robustness of the

findings. In particular, the recovery of 344 MAGs—though indicative of substantial prokaryotic diversity—may be biased toward ubiquitous, abundant taxa favored by continental-scale sampling across geographic gradients. Site-specific specialists and rare taxa involved in specialized carbon transformations [62,63] are likely underrepresented, suggesting that the results primarily reflect the "core" microbial carbon cycling machinery rather than the full functional diversity present in soil communities.

Nonetheless, our study provides a comprehensive, genome-resolved view of microbial decomposition potential across diverse soil environments, illuminating the metabolic underpinnings of carbon turnover at continental scale. By integrating thousands of chemically distinct SOM molecules with hundreds of metagenome-assembled genomes, we uncover widespread microbial capacities to degrade carbon compounds long thought to be resistant to decay. This work advances a critical knowledge gap in linking microbial taxonomy and metabolism to specific carbon pools, offering a mechanistic foundation for improving predictions of soil carbon dynamics under shifting environmental conditions. As global change accelerates, incorporating these microbial processes into Earth system models will be essential to more accurately project the fate of soil carbon reservoirs.

Methods

Data Collection

Through the 1000 Soils Pilot of the Molecular Observation Network (MONet) [14], 109 soil samples were collected from 60 sites across the CONUS. Many of the cores were collected at sites run by the U.S. National Ecological Observatory Network (NEON), a continental scale program that delivers open, standardized ecological data. Full methodological details are available in refs. ([14] and [64]), and at https://github.com/EMSL-MONet/MONet-Protocols-. With a goal to investigate the microbial pathways that underscore variation in microbial activity across geography and soil depth profiles, we selected the subset of 47 soil samples, which exhibited the highest and lowest (30%) rates of respiration within either surface or subsoils for analysis. These soil samples spanned 37 geographical locations (Supplementary Table 1 and Supplementary Fig. 1) and were divided into four groups based on depth and respiration: surface-high (n=14), surface-low (n=12), suboil-high (n=11), and subsoil-low (n=10).

Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FTICR-MS)

Water-soluble SOM composition was analyzed by Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FTICR-MS) [64,65]. Briefly, 6 g of dried soil in triplicate were resuspended in 30 mL DI water, and shaken for 2 hours at 800 rpm. The samples were then centrifuged at 6,000 rmp for 8 minutes and 5 mL of supernatant was transferred for solid phase extraction (SPE) with Agilent Bond Elut PPL cartridges [65] on Gilson ASPEC® SPE system. We used a Bruker 7 Tesla scimaX FTICR-MS at the Environmental Molecular Sciences Laboratory (EMSL) in Richland, WA to analyze the extracted SOM, with a negative ionization mode and ion accumulation time of 0.01 or 0.025 seconds. The mass accuracy was below 1 ppm, as confirmed by the Suwannee River Fulvic Acid (SRFA) samples. We used CoreMS (v. 2.0.0) [66] to process the raw FTICR-MS data, with noise thresholding of 10 (*noise_threshold_log_nsigma*) and minimum peak prominence of 0.1 (*peak_min_prominence_percent*). Mass calibration was performed by setting a threshold that uses the most calibration points (usually >200). Molecular formulae were annotated by both accurate mass and isotopologues, with a confidence score calculated for each formula. We then filtered the assigned peaks by *m/z* between 200 to 1,000, present in at least 2 out of 3 replicates, not present in two or more lab blanks, and with formulae confidence scores (combines *m/z* error and isotopic pattern) above 0.5 [67]. Putative biochemical transformation was calculated based on mass differences between pairs of SOM formula detected in our analysis. Reference mass differences of 1,298 chemical transformations compiled from previous studies [68,69] were used to profile putative transformations in each sample, with a mass error threshold of 1.0 e-5. The occurrences of each putative transformation in each sample and the numbers of putative transformations for each formula were counted for further statistical analysis, as described below.

Distribution and uniqueness of SOMs across the soil samples

Using the customized Python script, we assigned the 66,727 FTICR-derived molecular formulas into SOM compound classes, based on thresholds of O:C and H:C ratios on the van Krevelen diagram [70]. The matrix of the SOM assignment and the presence/absence matrix of the molecules across the 47 samples were used in a statistical analysis using the R package ftmsRanalysis [71] and associated scripts, to first determine the distribution of SOM types across the four soil sample type, and then to conduct a series of pairwise G-tests of uniqueness [72] with

p-value threshold of 0.05, generating profiles of SOMs that are unique in one soil type versus the other. The results of the G-tests were plotted as shown in Fig. 2.

DNA extraction and sequencing

DNA was extracted using *Quick*-DNA Fecal/Soil Microbe Miniprep Kit D6010 (manufacturer) and was cleaned and concentrated with DNA Clean and Concentrator-25 from ZymoResearch (manufacturer). Samples were sequenced at Azenta Life Sciences using the Illumina Whole Genome Metagenomics pipeline and the Joint Genome Institute, following identical protocols.

Quality control, assembly and binning of the metagenome reads

Processes of the metagenome construction, including quality-control/read filtering, contig assembly and binning processes, were performed using a suite of tools available in the MetaGEM pipeline [73]. Specifically, short read quality filtering and adapter trimming were applied to the set of reads (de-interleaved) using fastp v0.20.0 [74] with default settings. The filtered reads from each sample were assembled into contigs using MEGAHIT v1.2.9 [75], with '--presets meta-sensitive' and '--min-contig-len 1000' flags.

Before the binning process, contig coverages across the samples (required by some of the binning tools as described below) were generated by cross-mapping the filtered short reads to the set of assemblies. This was accomplished by generating an index for each contig using the 'bwa index' command with default settings. The set of indices were used as part of the input for bwa-mem v0.7.17 [76], generating SAM files, which were subsequently converted to BAM format using the 'samtools view' command with the '-Sb' flag, and then were sorted using the 'samtools sort' command with the default settings. The sorted BAM files were then used to generate the contig coverages for the binning tools, MetaBAT2 v.2.12.1 [77] and MaxBin2 v2.2.5 [78], using the script, 'jgi_summarize_bam_contig_depths' available as part of the MetaBAT2 package. The contig coverage files were then used as input data for the MetaBAT2 and MaxBin2, with default settings for each tool.

The binning process using the tool CONCOCT v1.1.0 [79] was performed using the scripts available in the MetaGEM package, as further described. The coverage files were generated

using the script, 'concoct_coverage_table.py' (default settings), with a set of BED files generated for each contig as input. The BED files were created using the 'cut_up_fasta.py' script, with the parameters '-c 10000 -o 0 -m b' that split the contigs into 10 kb fragments. For each fragment, CONCOCT was run using the coverage file as input and with the '-c 800' parameter, yielding a set of bins for each sample, based on the split contigs. A script, 'extract_fasta_bins.py' was then run to generate the bins using the original unfragmented contigs.

Refinement, reassembly and taxonomic classification of the MAGs

To extract the best version of bins from the outputs of the three binning tools, the 'bin_refinement' script available in metaWRAP v1.2.3 [80] was used with '-x 10 -c 50' parameters. This command de-replicates the results of the three binning methods by selecting the bin with the highest CheckM completeness and the lowest contamination, with the completeness taking precedence over contamination. The parameters in this script ensure that only the binning results with contamination and completeness thresholds of ≤ 10% and ≥ 50%, respectively, are considered for the bin refinement process.

To improve the quality of the bins selected from the refinement step, the 'reassemble_bins' command (with the same parameters used for the 'bin_refinement' script) was used. This script extracts the reads that map to the contigs belonging to the refined bins and reassembles them using the SPAdes assembly. This process is done at both a 'strict' (no mismatches) or 'permissive' (< 5 mismatches) level. CheckM completeness and contamination between the three versions of bins (i.e., strict, permissive or original derived from the refinement) are compared, and the bins with the best score are selected as the finalized MAGs.

The MAGs generated using the methods described above were taxonomically classified using the 'identify', 'align' and 'classify' commands available in GTDB-Tk v2.4.0 [81]. This set of commands identifies 120 ubiquitous bacterial marker genes (and 122 archaeal genes) from the input MAGs, which are then concatenated and aligned together. The alignment of the concatenated marker genes is then used as input for FastTree [82] to assign each MAG into the optimal loci in bacterial or archaeal tree of life. The output of the process includes tabular files containing taxonomic classifications of the MAGs.

KEGG pathway analysis of FTICR-derived molecules using KO identifications

The ftmsRanalysis package and associated custom scripts mentioned above were also used to generate a summary of KO identifications that are mapped to the FTICR-derived molecules. These summary tables were parsed to extract the KOs present in 50% or more of the samples belonging to each partition. The KEGG pathway representation analysis, which employs the enrichKEGG algorithm in the clusterProfiler R package [83], was conducted using this set of KOs. This method calculates the ratio of KOs affiliated with specific pathways vs the total number of KOs in each sample set. This analysis generated a summary visualization of significant KEGG pathways present in each layer/respiration partition.

Network analysis of the relationship between KO and FTICR-MS derived molecules

For each of the four layer/respiration partitions, co-occurrence networks were constructed using Spearman's correlation to explore the relationship between the KOs detected from the MAGs and the FTICR-MS-derived compounds. Each correlation matrix was filtered to obtain relationships with *rho* > 0.6 and FDR-corrected P < 0.01 [84]. The output matrices were then imported into Cytoscape [85] version 3.10.3. Clusters in each partition were identified using the MCODE application, with default parameters, producing a set of clusters with scores representing their densities and interconnectedness [86].

Functional annotation of abundant lineages in the depth/respiration partitions

We selected two order-level bacterial lineages, *Rhizobiales* and *Chthoniobacterales*, each of which is represented by > 5.0% of the 304 MAGs recovered from all soil samples. Only one order-level archaeal lineage (i.e., *Nitrososphaerales*) was selected, as > 90% of the 40 MAGs were assigned to this group. Functional genes were detected from the nucleotide sequences of the MAGs using Prokka [87], and were searched against KEGG Orthology (KO) [17] using the 'annotate' function implemented in EnrichM [88], generating a list of KO identifications and their frequencies across the MAGs. The output of Prokka also includes a general feature format (GFF) file, which was used to infer metabolic model prediction of both *Rhizobiales* and *Chthoniobacterales* (see section below). Based on the SOM type distributions across the four depth/respiration derived from the FTICR-MS experiment, we assessed the presence of the KO identifications, which constitute the KEGG Modules associated with central carbohydrate

metabolism, lipid metabolism and amino acid metabolism. In addition to the KO, the presence of the enzymes related to lignins and other recalcitrant carbohydrates was assessed by searching the MAGs against the Carbohydrate Active Enzymes (CAZy) database [22], using dbCAN2 [89].

Metabolic model prediction for *Rhizobiales*, *Chthoniobacterales* and *Nitrososphaerales*

Using the functional annotation profiles generated from the method described in the previous section, we inferred metabolic model predictions for the two dominant bacterial orders, *Rhizobiales* and *Chthoniobacterales* and the archaeal order *Nitrosophaerales*. To accomplish this, we built a metabolic prediction narrative using a selected suite of applications available in the DOE Systems Biology Knowledgebase (KBase) [90] platform. In this narrative, GFF tables containing genomic feature information (i.e., gene names, enzyme commission identifications and protein names) for each detected coding region and the nucleotide FASTA files of the MAGs are used to re-annotate the genomes using RAST-Tk v1.73 (Rapid Annotation using Subsystems Technology - Toolkit) [91]. This process generates the output files compatible with the KBase metabolic modeling applications. The GFF files were generated using Prokka with '--kingdom Bacteria' and '--addgenes' parameters. For the archaeal MAGs, the '--kingdom Archaea' flag was used instead.

The RAST-Tk annotation objects were used as input for ModelSEED [92] application, generating models representing the broader metabolic properties in the bacterial and archaeal lineages under investigation. The models are generated by compiling the information of metabolic pathways encoded by genes detected in the MAGs. The output generated by the ModelSEED application includes a matrix of metabolic reactions, their associated biochemical compounds, and the gene-protein-reaction (GPR) that entails the dependencies between the responses and the related genes. The GPR can help differentiate between cases where a reaction is catalyzed by a protein complex encoded by multiple genes and where several individual proteins independently orchestrate a reaction. The ModelSEED output also includes a 'biomass objective function' file that indicates the abundance of the metabolites generated by the microbes (e.g., amino acids, lipids and quinones) as bi-products of their growth metabolism.
Finally, Flux Balance Analysis (FBA) [44] application was used to read the ModelSEED-derived metabolic models and a default media file provided by KBase that specify

the set of > 500 compounds an organism can use for its growth, generating tables of metrics that allows estimation of growth rates of microbes and/or output of specific metabolites on a specified media. The FBA estimates the growth rate of an organism by calculating growth-optimal fluxes through all reactions in the metabolic network. The uptake values for the compounds specified in the output file, measured in mmol per gram cell dry weight per hour, indicate whether the microbes would use the metabolites for growth or be excreted as a by-product. The list of the output uptake values were examined to determine the spectrum of compounds that are likely to be metabolized by the members of the prevalent order-level clades detected in the soils.

Data availability

The sequencing data associated with this study are available in the NCBI Sequence Read Archive (SRA) under accession number **PRJNA1260013**. Data in the MONet open science database can be found at https://sc-data.emsl.pnnl.gov/monet and MONet Zenodo DOI (https://doi.org/10.5281/zenodo.7406532).

Code availability

The R and Python codes used for analyzing the FTICR-MS data and for preparing the main and supplementary figures are publicly available at the following GitHub repository: https://github.com/EMSL-MONet/The-1000-Soil-ICR-Metagenome. Access to the codes requires log-in to GitHub. The KBase narrative developed to analyze the *Rhizobiales* and *Chthoniobacterales* MAGs can be accessed at https://narrative.kbase.us/narrative/186750.

Acknowledgements

This research was performed on a project award DOI https://doi.org/10.46936/intm.proj.2021.60141/60000423 from the Environmental Molecular Sciences Laboratory (EMSL), a DOE Office of Science user facility sponsored by the Biological and Environmental Research program under Contract No. DE-AC05-76RL01830. Soil data was provided by the Molecular Observation Network (MONet) at the EMSL (DOI: https://doi.org/10.46936/10.25585/60008970). MONet is conducted in collaboration with Joint

Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science user facility supported by the Office of Science of the U.S. Department of Energy operated under Contract DE-AC02-05CH11231. The MONet database is an open, FAIR, and publicly available compilation of the molecular and microstructural properties of soil. The National Ecological Observatory Network is a program sponsored by the U.S. National Science Foundation and operated under cooperative agreement by Battelle. A subset of samples used in this research were obtained through NEON Research Support Services.

Competing interests

The authors declare no competing interests.

References

1. Saifuddin, M., Bhatnagar, J. M., Segrè, D. & Finzi, A. C. Microbial carbon use efficiency predicted from genome-scale metabolic models. *Nat Commun* **10**, 3568 (2019).
2. Bhattacharyya, S. S., Ros, G. H., Furtak, K., Iqbal, H. M. N. & Parra-Saldívar, R. Soil carbon sequestration – An interplay between soil microbial community and soil organic matter dynamics. *Science of The Total Environment* **815**, 152928 (2022).
3. Kleber, M. What is recalcitrant soil organic matter? *Environ. Chem.* **7**, 320–332 (2010).
4. Marschner, B. *et al.* How relevant is recalcitrance for the stabilization of organic matter in soils? *Journal of Plant Nutrition and Soil Science* **171**, 91–110 (2008).
5. Song, Y. *et al.* Soil metabolomics: Deciphering underground metabolic webs in terrestrial ecosystems. *Eco-Environment & Health* **3**, 227–237 (2024).
6. Swenson, T. L., Bowen, B. P., Nico, P. S. & Northen, T. R. Competitive sorption of microbial metabolites on an iron oxide mineral. *Soil Biology and Biochemistry* **90**, 34–41 (2015).
7. Feng, X. & Simpson, M. J. Temperature and substrate controls on microbial phospholipid fatty acid composition during incubation of grassland soils contrasting in organic matter quality. *Soil Biology and Biochemistry* **41**, 804–812 (2009).
8. Rillig, M. C., Caldwell, B. A., Wösten, H. A. B. & Sollins, P. Role of proteins in soil carbon and nitrogen storage: controls on persistence. *Biogeochemistry* **85**, 25–44 (2007).
9. Brown, R. W. *et al.* Soil metabolomics - current challenges and future perspectives. *Soil Biology and Biochemistry* **193**, 109382 (2024).
10. Roth, H. K. *et al.* Effects of burn severity on organic nitrogen and carbon chemistry in high-elevation forest soils. *Soil & Environmental Health* **1**, 100023 (2023).
11. Nelson, A. R. *et al.* Wildfire-dependent changes in soil microbiome diversity and function. *Nat Microbiol* **7**, 1419–1430 (2022).
12. Lewis, W. H., Tahon, G., Geesink, P., Sousa, D. Z. & Ettema, T. J. G. Innovations to culturing the uncultured microbial majority. *Nat Rev Microbiol* **19**, 225–240 (2021).
13. Luo, Y. *et al.* Toward more realistic projections of soil carbon dynamics by Earth system models. *Global Biogeochemical Cycles* **30**, 40–56 (2016).
14. Bowman, M. M. *et al.* One thousand soils for molecular understanding of belowground carbon cycling. *Front. Soil Sci.* **3**, (2023).

15. Fierer, N. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat Rev Microbiol* **15**, 579–590 (2017).
16. 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 (2017).
17. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**, D457–D462 (2016).
18. Chicca, I. *et al.* Degradation of BTEX mixture by a new Pseudomonas putida strain: role of the quorum sensing in the modulation of the upper BTEX oxidative pathway. *Environ Sci Pollut Res* **27**, 36203–36214 (2020).
19. Melo-Nascimento, A. O. dos S. *et al.* Complete genome reveals genetic repertoire and potential metabolic strategies involved in lignin degradation by environmental ligninolytic Klebsiella variicola P1CD1. *PLOS ONE* **15**, e0243739 (2020).
20. Rocci, K. S. *et al.* Bridging 20 Years of Soil Organic Matter Frameworks: Empirical Support, Model Representation, and Next Steps. *Journal of Geophysical Research: Biogeosciences* **129**, e2023JG007964 (2024).
21. Lou, X. *et al.* The Biodegradation of Soil Organic Matter in Soil-Dwelling Humivorous Fauna. *Front. Bioeng. Biotechnol.* **9**, (2022).
22. Cantarel, B. L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Research* **37**, D233–D238 (2009).
23. Nagy, L. *et al.* The Effect of Mutation in Lipopolysaccharide Biosynthesis on Bacterial Fitness. *Cells* **11**, 3249 (2022).
24. Clark, M. M., Paxhia, M. D., Young, J. M., Manzella, M. P. & Reguera, G. Adaptive Synthesis of a Rough Lipopolysaccharide in Geobacter sulfurreducens for Metal Reduction and Detoxification. *Applied and Environmental Microbiology* (2021) doi:10.1128/AEM.00964-21.
25. Zimmerman, A. E., Graham, E. B., McDermott, J. & Hofmockel, K. S. Estimating the Importance of Viral Contributions to Soil Carbon Dynamics. *Global Change Biology* **30**, e17524 (2024).
26. Graham, E. B. *et al.* A global atlas of soil viruses reveals unexplored biodiversity and potential biogeochemical impacts. *Nat Microbiol* **9**, 1873–1883 (2024).

27. Penton, C. R., Gupta, V. V. S. R., Yu, J. & Tiedje, J. M. Size Matters: Assessing Optimum Soil Sample Size for Fungal and Bacterial Community Structure Analyses Using High Throughput Sequencing of rRNA Gene Amplicons. *Front. Microbiol.* **7**, (2016).
28. Thevenot, M., Dignac, M.-F. & Rumpel, C. Fate of lignins in soils: A review. *Soil Biology and Biochemistry* **42**, 1200–1211 (2010).
29. Wang, X. *et al.* Microbial keystone taxa drive succession of plant residue chemistry. *The ISME Journal* **17**, 748–757 (2023).
30. Brown, M. E. & Chang, M. C. Exploring bacterial lignin degradation. *Current Opinion in Chemical Biology* **19**, 1–7 (2014).
31. Qin, S., Zhang, D., Wei, B. & Yang, Y. Dual roles of microbes in mediating soil carbon dynamics in response to warming. *Nat Commun* **15**, 6439 (2024).
32. Nottingham, A. T., Gloor, E., Bååth, E. & Meir, P. Soil carbon and microbes in the warming tropics. *Functional Ecology* **36**, 1338–1354 (2022).
33. Anderson, C. G. *et al.* Seasonal Controls on Microbial Depolymerization and Oxidation of Organic Matter in Floodplain Soils. *Environ. Sci. Technol.* **58**, 16815–16823 (2024).
34. Hall, S. J. & Silver, W. L. Iron oxidation stimulates organic matter decomposition in humid tropical forest soils. *Global Change Biology* **19**, 2804–2813 (2013).
35. Yadav, S., Tripathi, S., Purchase, D. & Chandra, R. Development of a biofilm-forming bacterial consortium and quorum sensing molecules for the degradation of lignin-containing organic pollutants. *Environmental Research* **226**, 115618 (2023).
36. Berlemont, R. & Martiny, A. C. Glycoside Hydrolases across Environmental Microbial Communities. *PLOS Computational Biology* **12**, e1005300 (2016).
37. Henrissat, B. & Coutinho, P. M. Classification of glycoside hydrolases and glycosyltransferases from hyperthermophiles. in *Methods in Enzymology* vol. 330 183–201 (Academic Press, 2001).
38. Averill, C. & Waring, B. Nitrogen limitation of decomposition and decay: How can it occur? *Global Change Biology* **24**, 1417–1427 (2018).
39. Chen, L. *et al.* Nitrogen availability regulates topsoil carbon dynamics after permafrost thaw by altering microbial metabolic efficiency. *Nat Commun* **9**, 3951 (2018).
40. Buckel, W. Energy Conservation in Fermentations of Anaerobic Bacteria. *Front. Microbiol.* **12**, (2021).

41. Macias-Benitez, S. *et al.* Rhizospheric Organic Acids as Biostimulants: Monitoring Feedbacks on Soil Microorganisms and Biochemical Properties. *Front. Plant Sci.* **11**, (2020).
42. Rakitin, A. L. *et al.* Verrucomicrobia of the Family Chthoniobacteraceae Participate in Xylan Degradation in Boreal Peat Soils. *Microorganisms* **12**, 2271 (2024).
43. Janssen, P. H. Chthoniobacter. in *Bergey's Manual of Systematics of Archaea and Bacteria* 1–2 (John Wiley & Sons, Ltd, 2015). doi:10.1002/9781118960608.gbm01281.
44. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat Biotechnol* **28**, 245–248 (2010).
45. Bei, Q. *et al.* Metabolic potential of Nitrososphaera-associated clades. *The ISME Journal* **18**, wrae086 (2024).
46. Gorka, S., Canarini, A., Schmidt, H. & Kaiser, C. Soil bacterial neutral lipid fatty acids: Markers for carbon storage or necromass? 2024.12.02.626346 Preprint at https://doi.org/10.1101/2024.12.02.626346 (2024).
47. DINEL, H., SCHNITZER, M. & MEHUYS, G. R. Soil Lipids: Origin, Nature, Content, Decomposition, and Effect on Soil Physical Properties. in *Soil Biochemistry* (Routledge, 1990).
48. Kallenbach, C. M., Grandy, A. S., Frey, S. D. & Diefendorf, A. F. Microbial physiology and necromass regulate agricultural soil carbon accumulation. *Soil Biology and Biochemistry* **91**, 279–290 (2015).
49. Dai, G. *et al.* Plant-derived lipids play a crucial role in forest soil carbon accumulation. *Soil Biology and Biochemistry* **168**, 108645 (2022).
50. Zhao, F. *et al.* Microbial functional genes driving the positive priming effect in forest soils along an elevation gradient. *Soil Biology and Biochemistry* **165**, 108498 (2022).
51. Kalbitz, K. *et al.* Changes in properties of soil-derived dissolved organic matter induced by biodegradation. *Soil Biology and Biochemistry* **35**, 1129–1142 (2003).
52. Sun, Y., Zhang, T., Lu, B., Li, X. & Jiang, L. Application of cofactors in the regulation of microbial metabolism: A state of the art review. *Front. Microbiol.* **14**, (2023).
53. Santamaria, G. *et al.* Evolution and regulation of microbial secondary metabolism. *eLife* **11**, e76119 (2022).
54. Kutschera, A. & Ranf, S. The multifaceted functions of lipopolysaccharide in plant-bacteria interactions. *Biochimie* **159**, 93–98 (2019).

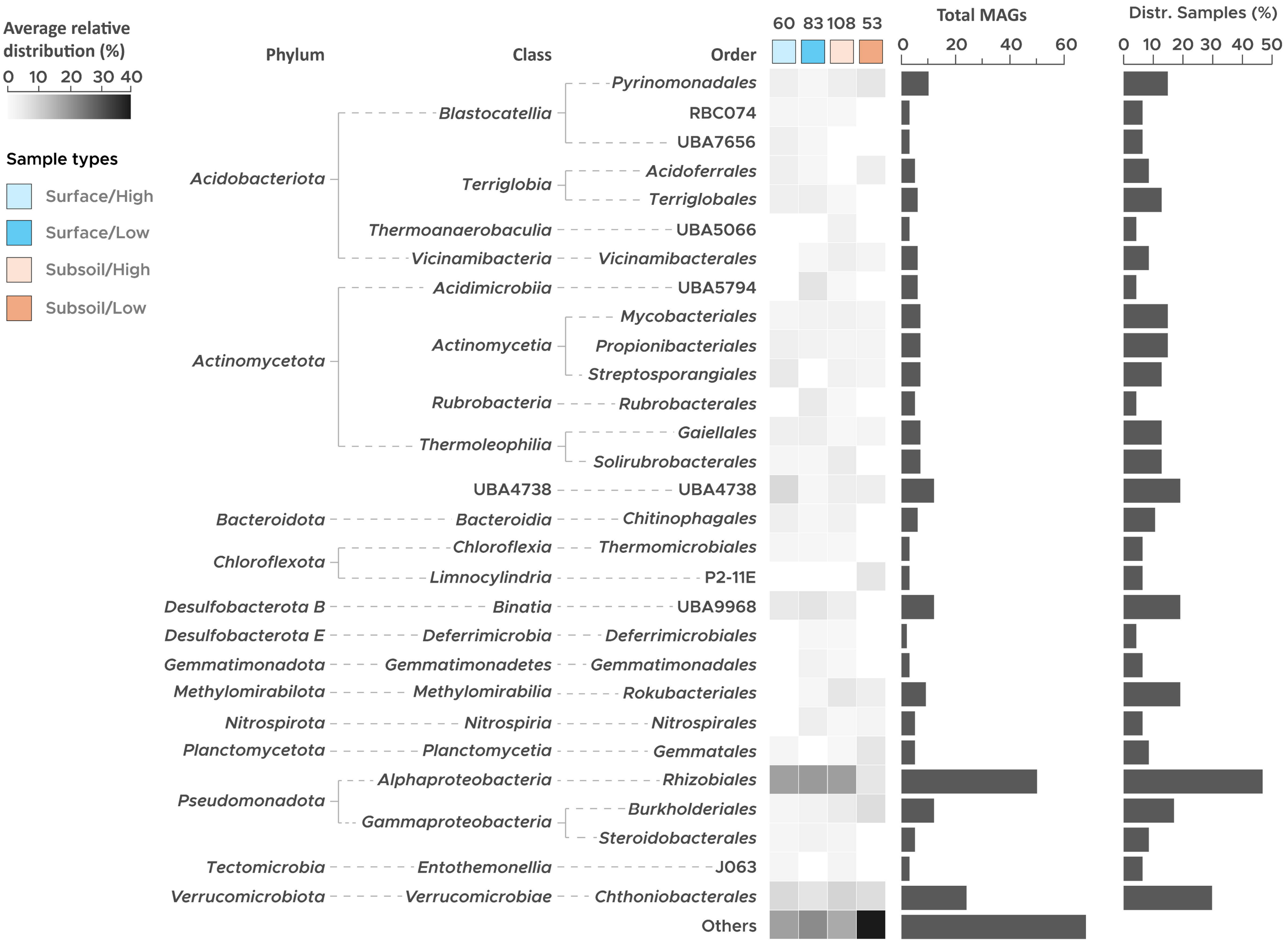
55. Zhang, G., Meredith, T. C. & Kahne, D. On the essentiality of lipopolysaccharide to Gram-negative bacteria. *Current Opinion in Microbiology* **16**, 779–785 (2013).
56. Kästner, M., Miltner, A., Thiele-Bruhn, S. & Liang, C. Microbial Necromass in Soils—Linking Microbes to Soil Processes and Carbon Turnover. *Front. Environ. Sci.* **9**, (2021).
57. Bradley, J. A., Amend, J. P. & LaRowe, D. E. Necromass as a Limited Source of Energy for Microorganisms in Marine Sediments. *Journal of Geophysical Research: Biogeosciences* **123**, 577–590 (2018).
58. Choma, A., Komaniecka, I. & Zebracki, K. Structure, biosynthesis and function of unusual lipids A from nodule-inducing and N2-fixing bacteria. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **1862**, 196–209 (2017).
59. Miltner, A., Bombach, P., Schmidt-Brücken, B. & Kästner, M. SOM genesis: microbial biomass as a significant source. *Biogeochemistry* **111**, 41–55 (2012).
60. Cania, B. *et al.* Site-Specific Conditions Change the Response of Bacterial Producers of Soil Structure-Stabilizing Agents Such as Exopolysaccharides and Lipopolysaccharides to Tillage Intensity. *Front. Microbiol.* **11**, (2020).
61. Woodcroft, B. J. *et al.* Genome-centric view of carbon processing in thawing permafrost. *Nature* **560**, 49–54 (2018).
62. Bickel, S. & Or, D. The chosen few—variations in common and rare soil bacteria across biomes. *The ISME Journal* **15**, 3315–3325 (2021).
63. Kaminsky, R. & Morales, S. E. Conditionally Rare Taxa Contribute but Do Not Account for Changes in Soil Prokaryotic Community Structure. *Front. Microbiol.* **9**, (2018).
64. Shi, C. *et al.* Scaling High-Resolution Soil Organic Matter Composition to Improve Predictions of Potential Soil Respiration Across the Continental United States. *Geophysical Research Letters* **52**, e2024GL113091 (2025).
65. Dittmar, T., Koch, B., Hertkorn, N. & Kattner, G. A simple and efficient method for the solid-phase extraction of dissolved organic matter (SPE-DOM) from seawater. *Limnology and Oceanography: Methods* **6**, 230–235 (2008).
66. Corilo, Y. E., Kew, W. R., McCue, L. A., Heal, K. & Carr, J. C. EMSL-Computing/CoreMS: CoreMS 3.0.0. Zenodo https://doi.org/10.5281/zenodo.14009575 (2024).
67. Corilo, Y. E., Kew, W. R., McCue, L. A., Heal, K. & Carr, J. C. EMSL-Computing/CoreMS:

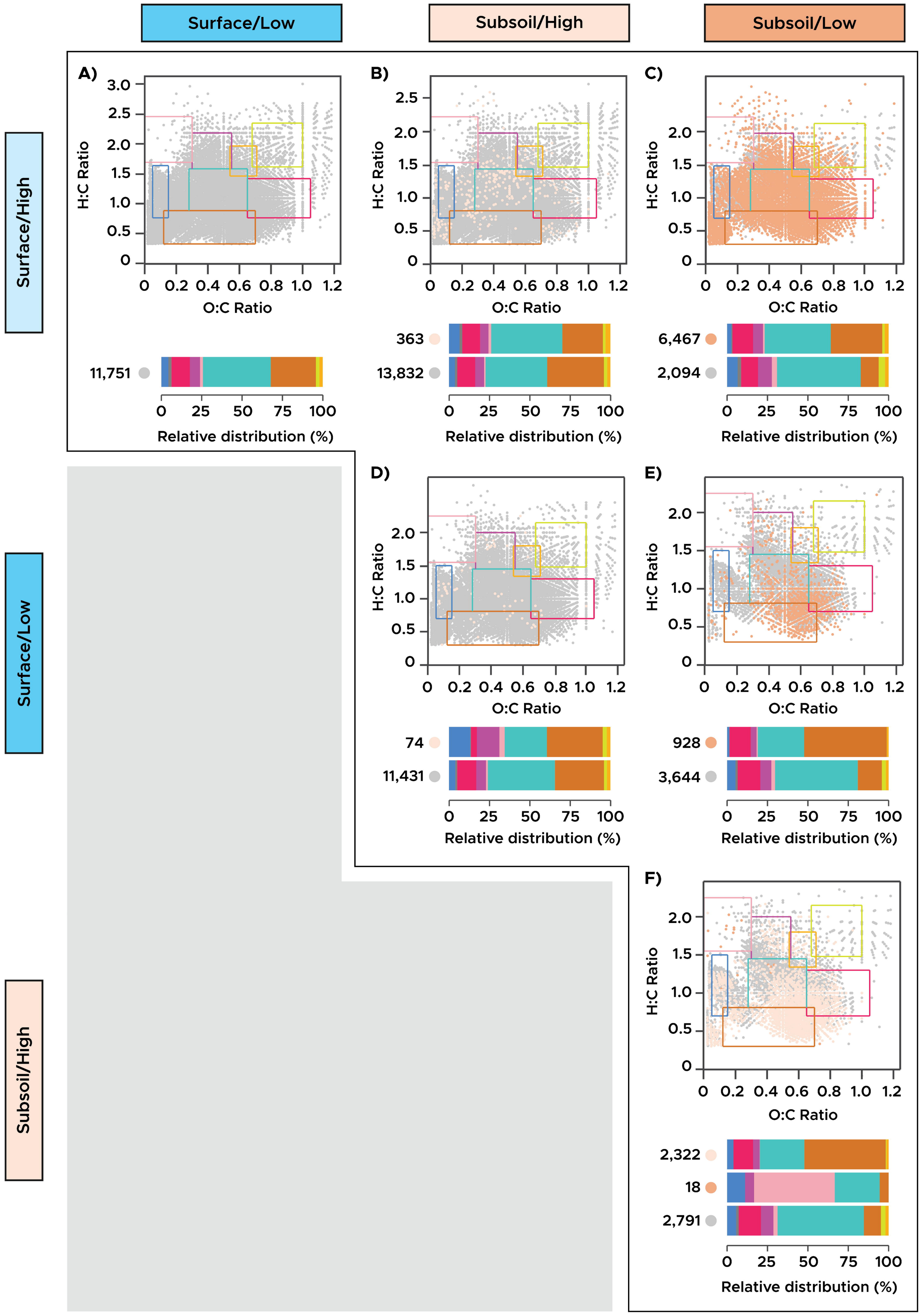
CoreMS 3.0.0. Zenodo https://doi.org/10.5281/zenodo.14009575 (2024).

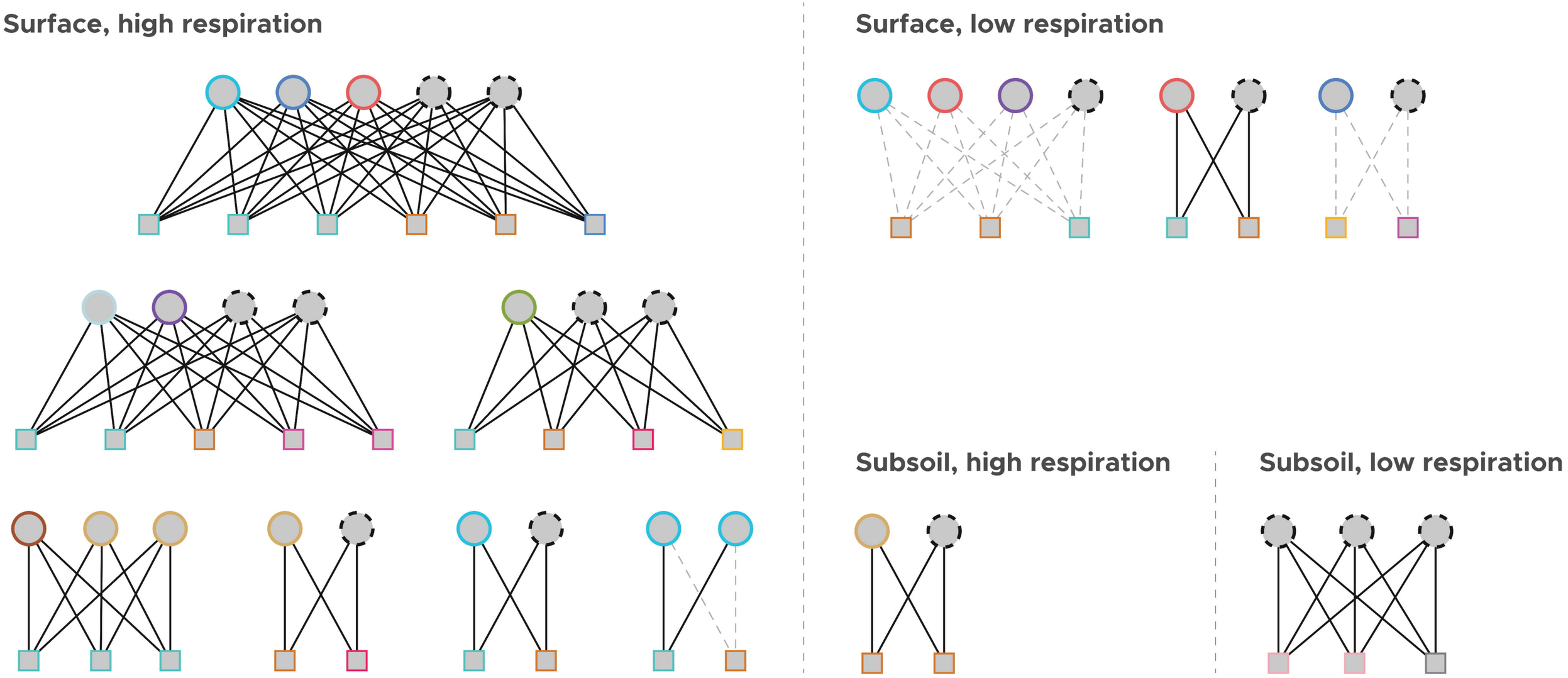
68. Ayala-Ortiz, C. *et al.* MetaboDirect: an analytical pipeline for the processing of FT-ICR MS-based metabolomic data. *Microbiome* **11**, 28 (2023).
69. Tureţcaia, A. B. *et al.* Rethinking Aerobic Respiration in the Hyporheic Zone under Variation in Carbon and Nitrogen Stoichiometry. *Environ. Sci. Technol.* **57**, 15499–15510 (2023).
70. Kim, S., Kramer, R. W. & Hatcher, P. G. Graphical Method for Analysis of Ultrahigh-Resolution Broadband Mass Spectra of Natural Organic Matter, the Van Krevelen Diagram. *Anal. Chem.* **75**, 5336–5344 (2003).
71. Degnan, D. J. *et al.* FREDA: A Web Application for the Processing, Analysis, and Visualization of Fourier-Transform Mass Spectrometry Data. *Rapid Communications in Mass Spectrometry* **39**, e9980 (2025).
72. G–test of goodness-of-fit - Handbook of Biological Statistics. https://www.biostathandbook.com/gtestgof.html.
73. Zorrilla, F., Buric, F., Patil, K. R. & Zelezniak, A. metaGEM: reconstruction of genome scale metabolic models directly from metagenomes. *Nucleic Acids Research* **49**, e126 (2021).
74. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
75. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
76. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).
77. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
78. 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 (2016).
79. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**, 1144–1146 (2014).
80. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
81. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify

genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).

82. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5**, e9490 (2010).
83. Honeker, L. K. *et al.* Drought re-routes soil microbial carbon metabolism towards emission of volatile metabolites in an artificial tropical rainforest. *Nat Microbiol* **8**, 1480–1494 (2023).
84. Graham, E. B. *et al.* Deterministic influences exceed dispersal effects on hydrologically-connected microbiomes. *Environmental Microbiology* **19**, 1552–1567 (2017).
85. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).
86. Banerjee, S. *et al.* Network analysis reveals functional redundancy and keystone taxa amongst bacterial and fungal communities during organic matter decomposition in an arable soil. *Soil Biology and Biochemistry* **97**, 188–198 (2016).
87. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
88. Boyd, J. geronimp/enrichM. (2024).
89. Zhang, H. *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research* **46**, W95–W101 (2018).
90. Arkin, A. P. *et al.* KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat Biotechnol* **36**, 566–569 (2018).
91. Brettin, T. *et al.* RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci Rep* **5**, 8365 (2015).
92. Seaver, S. M. D. *et al.* The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. *Nucleic Acids Research* **49**, D575–D588 (2021).


Average relative distribution (%)
0 10 20 30 40
Sample types
Surface/High
Surface/Low
Subsoil/High
Subsoil/Low
Phylum
Class
Order
60 83 108 53
Total MAGs
0 20 40 60
Distr. Samples (%)
0 10 20 30 40 50
Acidobacteriota
Blastocatellia
Pyrinomonadales
RBC074
UBA7656
Terriglobia
Acidoferrales
Terriglobales
Thermoanaerobaculia
UBA5066
Vicinamibacteria
Vicinamibacterales
Acidimicrobiia
UBA5794
Actinomycetota
Actinomycetia
Mycobacteriales
Propionibacteriales
Streptosporangiales
Rubrobacteria
Rubrobacterales
Thermoleophilia
Gaiellales
Solirubrobacterales
UBA4738
UBA4738
Bacteroidota
Bacteroidia
Chitinophagales
Chloroflexota
Chloroflexia
Thermomicrobiales
Limnocylindria
P2-11E
Desulfobacterota B
Binatia
UBA9968
Desulfobacterota E
Deferrimicrobia
Deferrimicrobiales
Gemmatimonadota
Gemmatimonadetes
Gemmatimonadales
Methylomirabilota
Methylomirabilia
Rokubacteriales
Nitrospirota
Nitrospiria
Nitrospirales
Planctomycetota
Planctomycetia
Gemmatales
Pseudomonadota
Alphaproteobacteria
Rhizobiales
Gammaproteobacteria
Burkholderiales
Steroidobacterales
Tectomicrobia
Entothemonellia
J063
Verrucomicrobiota
Verrucomicrobiae
Chthoniobacterales
Others


Surface/Low
Subsoil/High
Subsoil/Low
Surface/High
Surface/Low
Subsoil/High
A)
B)
C)
D)
E)
F)
H:C Ratio
O:C Ratio
Relative distribution (%)
11,751
13,832
363
6,467
2,094
11,431
74
928
3,644
2,322
18
2,791
SOM Classes
Condensed HC-like
Lignin-like
Tannin-like
Protein-like
Lipid-like
Unsaturated HC-like
Amino sugar-like
Carbohydrate-like
Sample Type Comparisons
Unique to Subsoil/High
Unique to Subsoil/Low
Present in both samples under comparison



Nodes

- KEGG Orthology
- SOM

Correlation

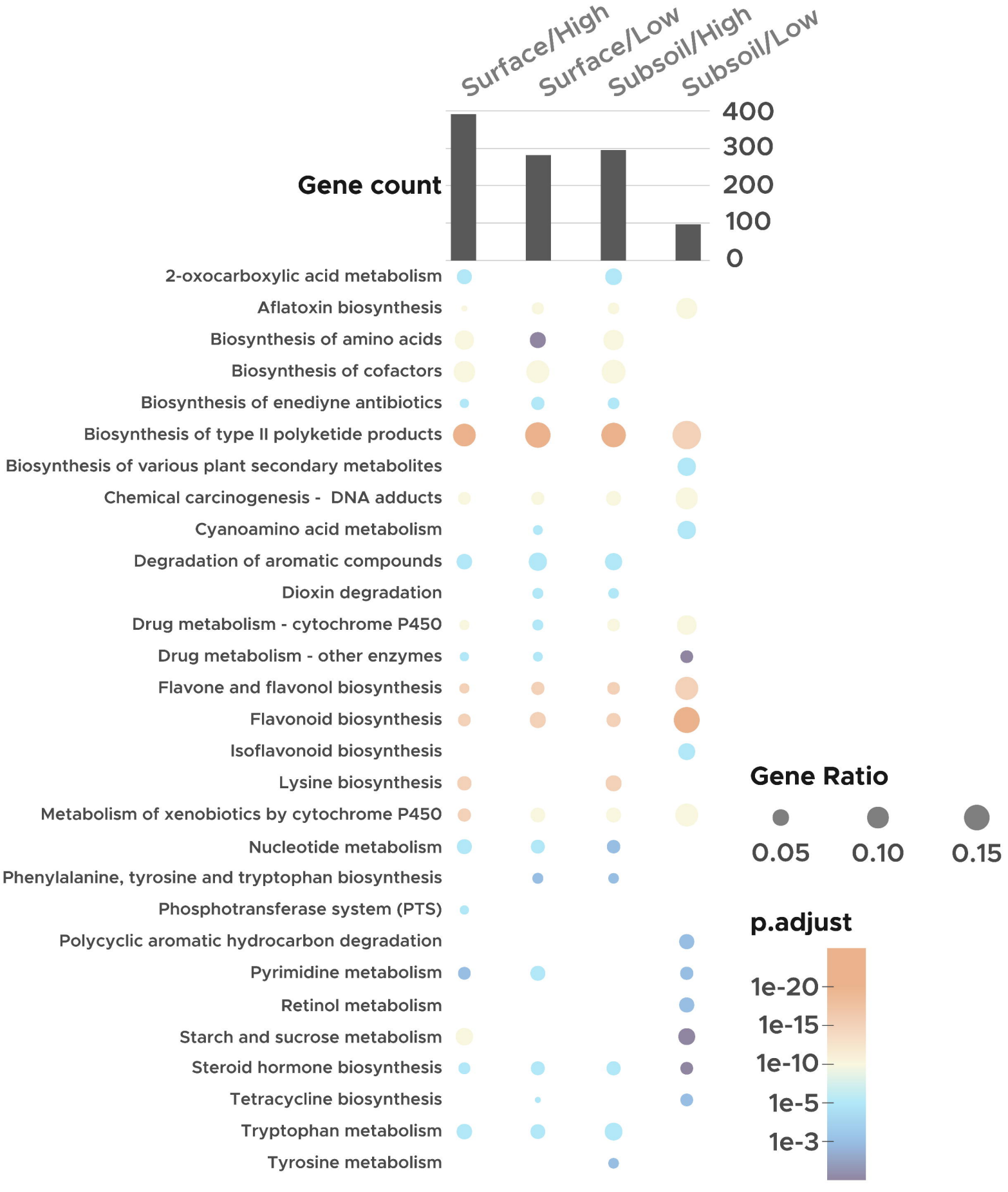
- > 0
- < 0

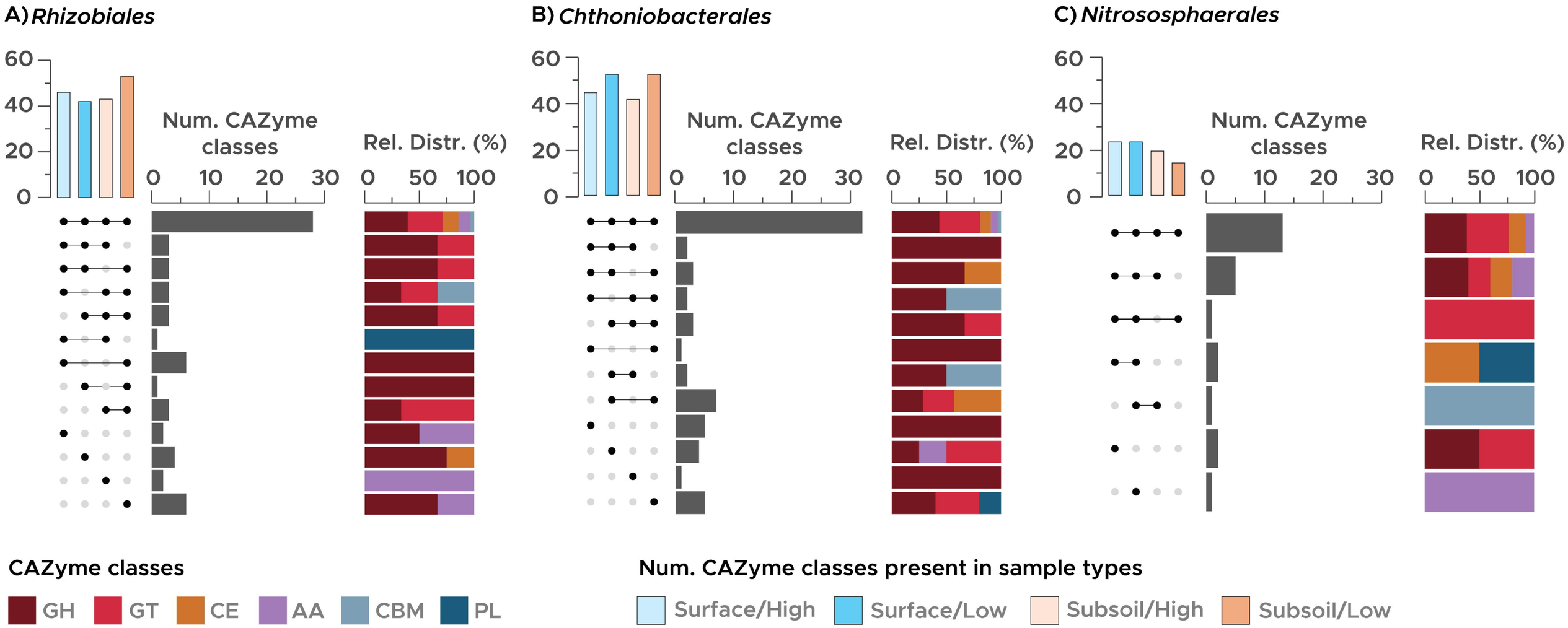
SOM classes

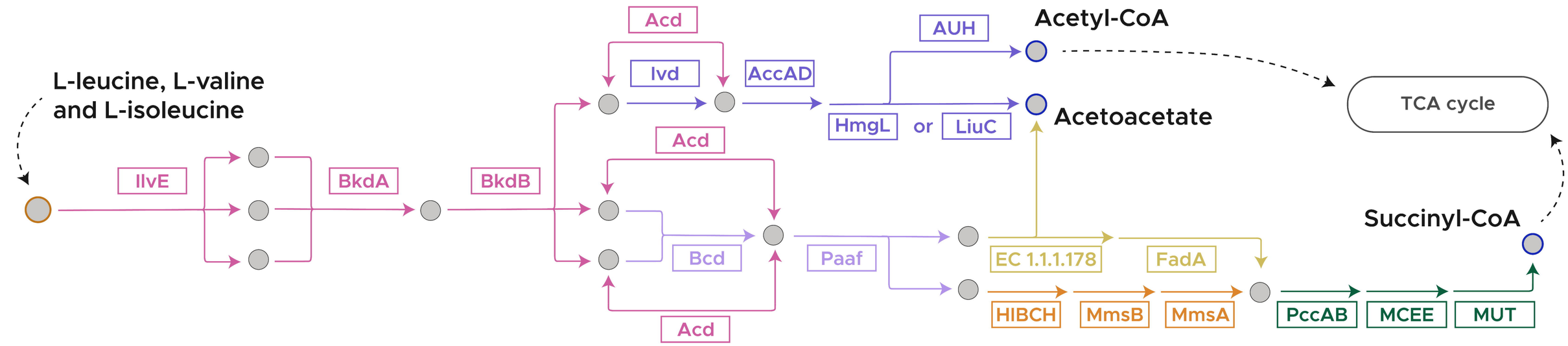
- Condensed hydrocarbon-like
- Protein-like
- Amino sugar-like
- Lignin-like
- Lipid-like
- Uncharacterized
- Tannin-like
- Unsaturated hydrocarbon-like

BRITE categories

- Carbohydrate metabolism
- Metabolism of other amino acids
- Cellular processes
- Nucleotide metabolism
- Metabolism of cofactors and vitamins
- Xenobiotics biodegradation and metabolism
- Glycan biosynthesis and metabolism
- Energy metabolism
- Others


Surface/High
Surface/Low
Subsoil/High
Subsoil/Low
Gene count
400
300
200
100
0
2-oxocarboxylic acid metabolism
Aflatoxin biosynthesis
Biosynthesis of amino acids
Biosynthesis of cofactors
Biosynthesis of enediyne antibiotics
Biosynthesis of type II polyketide products
Biosynthesis of various plant secondary metabolites
Chemical carcinogenesis - DNA adducts
Cyanoamino acid metabolism
Degradation of aromatic compounds
Dioxin degradation
Drug metabolism - cytochrome P450
Drug metabolism - other enzymes
Flavone and flavonol biosynthesis
Flavonoid biosynthesis
Isoflavonoid biosynthesis
Lysine biosynthesis
Metabolism of xenobiotics by cytochrome P450
Nucleotide metabolism
Phenylalanine, tyrosine and tryptophan biosynthesis
Phosphotransferase system (PTS)
Polycyclic aromatic hydrocarbon degradation
Pyrimidine metabolism
Retinol metabolism
Starch and sucrose metabolism
Steroid hormone biosynthesis
Tetracycline biosynthesis
Tryptophan metabolism
Tyrosine metabolism
Gene Ratio
0.05
0.10
0.15
p.adjust
1e-20
1e-15
1e-10
1e-5
1e-3


A) Rhizobiales
B) Chthoniobacterales
C) Nitrososphaerales
Num. CAZyme classes
Rel. Distr. (%)
0
10
20
30
50
100
60
40
20
CAZyme classes
GH
GT
CE
AA
CBM
PL
Num. CAZyme classes present in sample types
Surface/High
Surface/Low
Subsoil/High
Subsoil/Low



Pathways

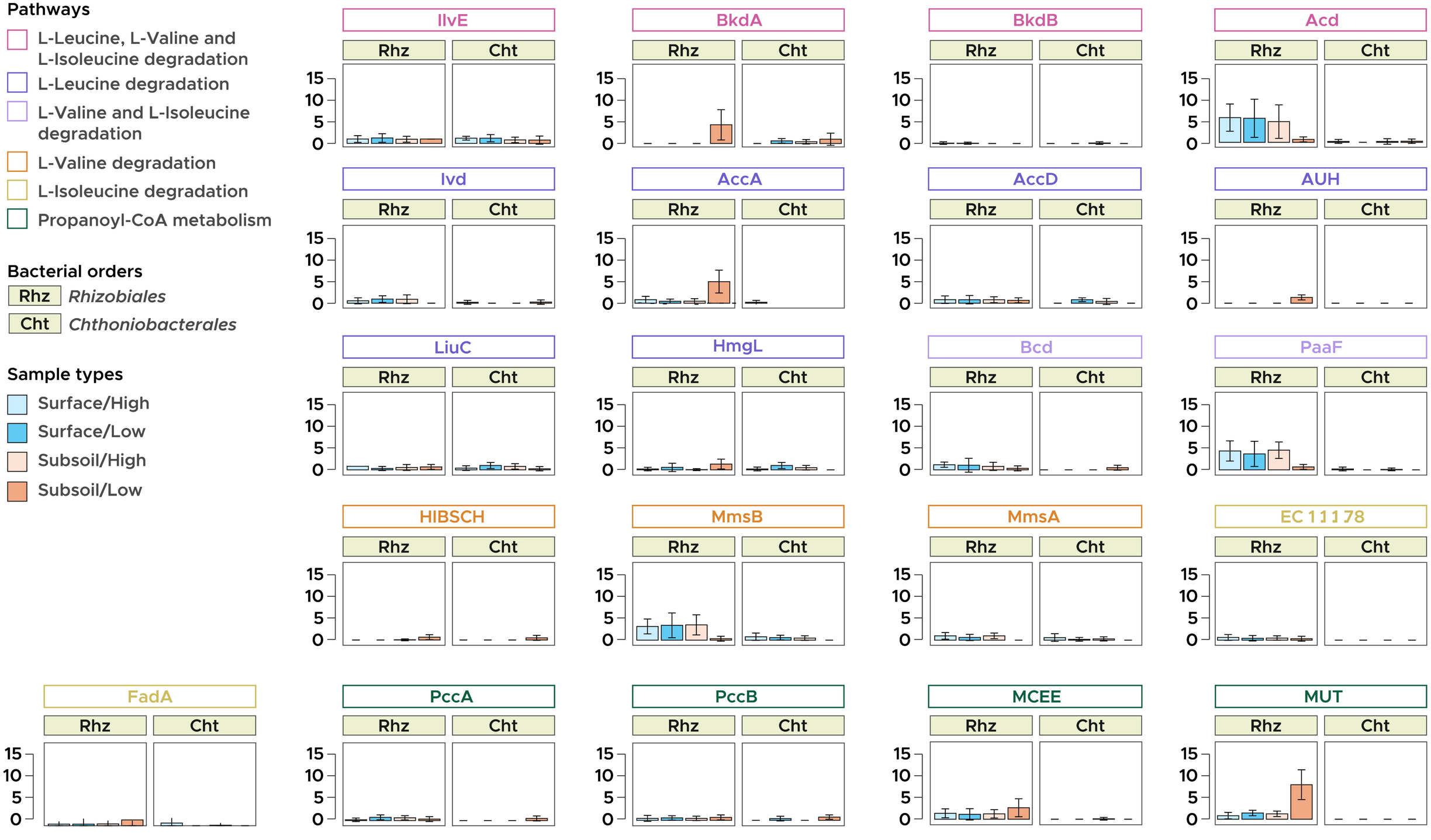
- L-Leucine, L-Valine and L-Isoleucine degradation
- L-Leucine degradation
- L-Valine and L-Isoleucine degradation
- L-Valine degradation
- L-Isoleucine degradation
- Propanoyl-CoA metabolism

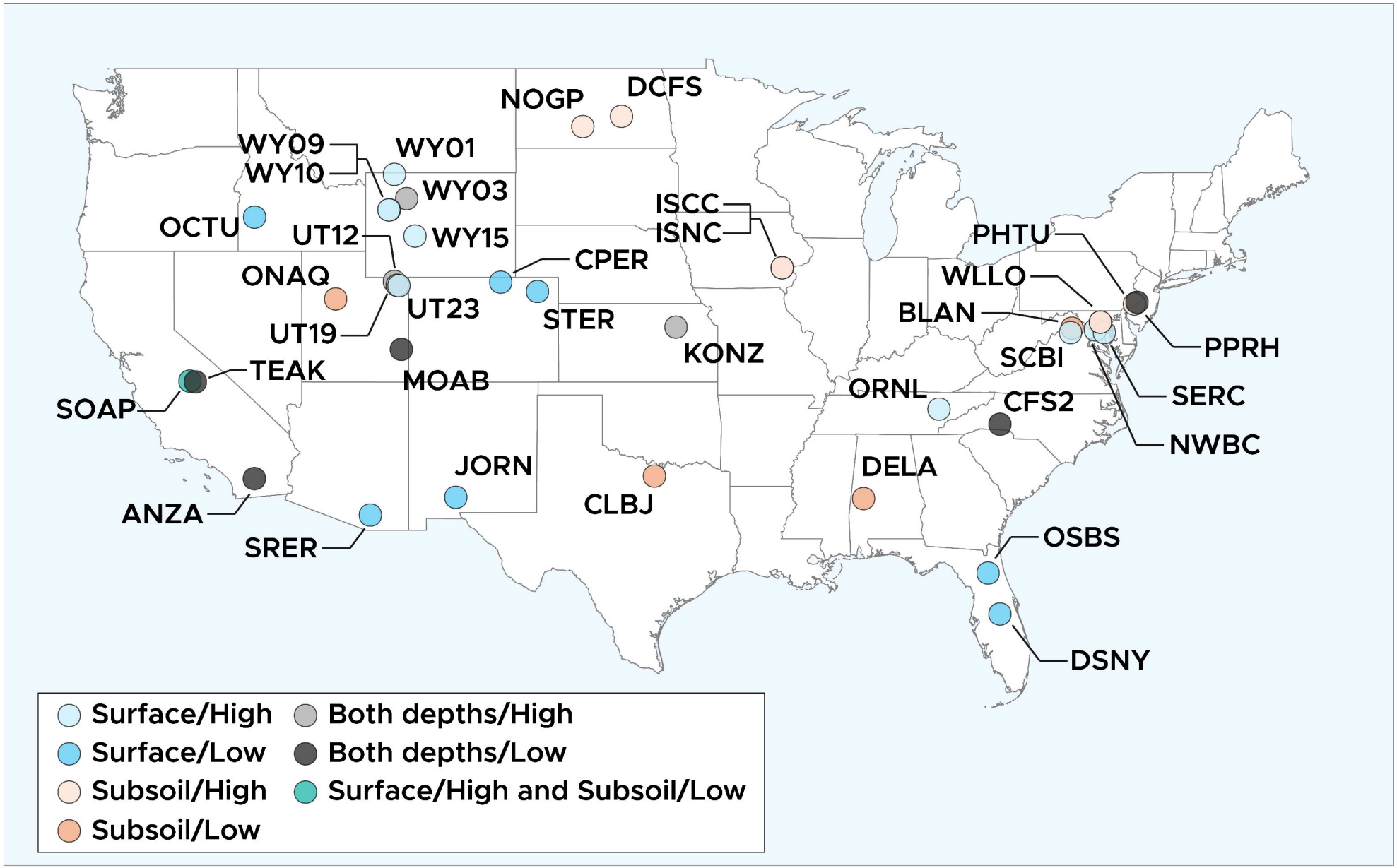
Bacterial orders

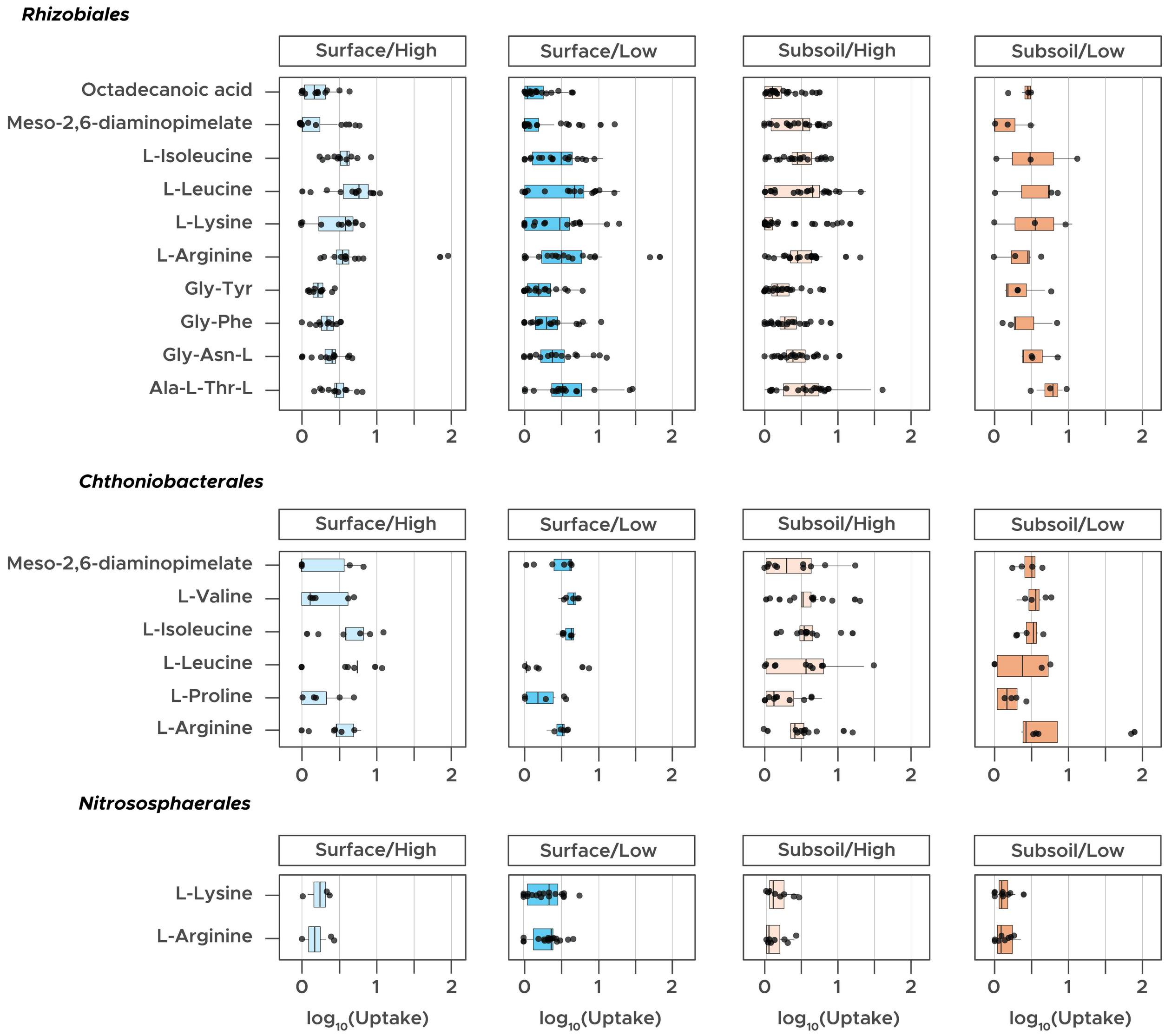
- Rhz *Rhizobiales*
- Cht *Chthoniobacterales*

Sample types

- Surface/High
- Surface/Low
- Subsoil/High
- Subsoil/Low




NOGP
DCFS
WY09
WY10
WY01
WY03
OCTU
UT12
WY15
ISCC
ISNC
CPER
ONAQ
UT23
STER
UT19
KONZ
TEAK
MOAB
SOAP
JORN
CLBJ
ANZA
SRER
PHTU
WLLO
BLAN
SCBI
PPRH
SERC
ORNL
CFS2
NWBC
DELA
OSBS
DSNY
Surface/High
Surface/Low
Subsoil/High
Subsoil/Low
Both depths/High
Both depths/Low
Surface/High and Subsoil/Low


Rhizobiales
Surface/High
Surface/Low
Subsoil/High
Subsoil/Low
Octadecanoic acid
Meso-2,6-diaminopimelate
L-Isoleucine
L-Leucine
L-Lysine
L-Arginine
Gly-Tyr
Gly-Phe
Gly-Asn-L
Ala-L-Thr-L
0
1
2
Chthoniobacterales
Surface/High
Surface/Low
Subsoil/High
Subsoil/Low
Meso-2,6-diaminopimelate
L-Valine
L-Isoleucine
L-Leucine
L-Proline
L-Arginine
0
1
2
Nitrososphaerales
Surface/High
Surface/Low
Subsoil/High
Subsoil/Low
L-Lysine
L-Arginine
0
1
2
log10(Uptake)
log10(Uptake)
log10(Uptake)
log10(Uptake)