

# Genomic fingerprints of the world's soil ecosystems

Emily B. Graham,<sup>1,2</sup> Vanessa A. Garayburu-Caruso,<sup>1</sup> Ruonan Wu,<sup>1</sup> Jianqiu Zheng,<sup>1</sup> Ryan McClure,<sup>1</sup> Gerrad D. Jones<sup>3</sup>

**AUTHOR AFFILIATIONS** See affiliation list on p. 15.

**ABSTRACT** Despite the explosion of soil metagenomic data, we lack a synthesized understanding of patterns in the distribution and functions of soil microorganisms. These patterns are critical to predictions of soil microbiome responses to climate change and resulting feedbacks that regulate greenhouse gas release from soils. To address this gap, we assay 1,512 manually curated soil metagenomes using complementary annotation databases, read-based taxonomy, and machine learning to extract multidimensional genomic fingerprints of global soil microbiomes. Our objective is to uncover novel biogeographical patterns of soil microbiomes across environmental factors and ecological biomes with high molecular resolution. We reveal shifts in the potential for (i) microbial nutrient acquisition across pH gradients; (ii) stress-, transport-, and redox-based processes across changes in soil bulk density; and (iii) greenhouse gas emissions across biomes. We also use an unsupervised approach to reveal a collection of soils with distinct genomic signatures, characterized by coordinated changes in soil organic carbon, nitrogen, and cation exchange capacity and in bulk density and clay content that may ultimately reflect soil environments with high microbial activity. Genomic fingerprints for these soils highlight the importance of resource scavenging, plant-microbe interactions, fungi, and heterotrophic metabolisms. Across all analyses, we observed phylogenetic coherence in soil microbiomes—more closely related microorganisms tended to move congruently in response to soil factors. Collectively, the genomic fingerprints uncovered here present a basis for global patterns in the microbial mechanisms underlying soil biogeochemistry and help beget tractable microbial reaction networks for incorporation into process-based models of soil carbon and nutrient cycling.

**IMPORTANCE** We address a critical gap in our understanding of soil microorganisms and their functions, which have a profound impact on our environment. We analyzed 1,512 global soils with advanced analytics to create detailed genetic profiles (fingerprints) of soil microbiomes. Our work reveals novel patterns in how microorganisms are distributed across different soil environments. For instance, we discovered shifts in microbial potential to acquire nutrients in relation to soil acidity, as well as changes in stress responses and potential greenhouse gas emissions linked to soil structure. We also identified soils with putative high activity that had unique genomic characteristics surrounding resource acquisition, plant-microbe interactions, and fungal activity. Finally, we observed that closely related microorganisms tend to respond in similar ways to changes in their surroundings. Our work is a significant step toward comprehending the intricate world of soil microorganisms and its role in the global climate.

**KEYWORDS** soil microbiology, metagenomics, metaanalysis, machine learning, carbon cycling, nitrogen cycling, biogeochemistry, biogeography, soil microbiome, functional potential

Soil microbiomes catalyze some of the most biogeochemically important reactions on Earth, yet their complexity hampers our efforts to fully understand their function

**Editor** Ashley Shade, CNRS Delegation Alpes, Lyon, Rhône-Alpes, France

**Ad Hoc Peer Reviewer** Zachary B. Freedman, University of Wisconsin-Madison, Madison, Wisconsin, USA

Address correspondence to Emily B. Graham, emily.graham@pnnl.gov.

The authors declare no conflict of interest.

See the funding table on p. 15.

**Received** 18 October 2023

**Accepted** 25 March 2024

Copyright © 2024 Graham et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

(1–6). Microbial destabilization of soil carbon stores, for instance, has positive feedbacks with global climate change (7–9), and microbial nutrient cycling sustains life at higher trophic levels (10–12). The molecular revolution, including advanced ‘omics sequencing approaches, promises a new generation of fundamental and predictive understanding; imperatives for understanding how soil microbial communities individually and collectively contribute to carbon and nutrient cycling under environmental change. Yet despite the immense amount of soil metagenomic data now available, we lack a synthesized understanding of patterns in the distribution and functions of soil microorganisms that are critical to developing new and improved models (13–18).

While global patterns are well established for most macroorganisms, the extreme diversity of the soil microbiome and the myriad ecological forces that act upon it complicate efforts to understand patterns in its structure and function (19–27). Microbial assembly processes, dormancy, interspecies interactions, and aboveground-belowground connectivity in particular are among the many factors that influence soil microbial biogeography (19, 21, 26, 28–33). As a result, patterns in microbial distributions are almost always weaker than those observed in macroecological systems (26, 34, 35). Still, soil microbiome structure and function have been associated with variables such as latitude, vegetation, climate, and edaphic properties at local to global scales (3–5, 18, 19, 28, 34, 36–60).

A clear understanding of soil microbial biogeography is essential to predictions of soil microbiome function under novel climate scenarios. Microbiome composition (i.e., taxonomy and functional potential) impacts ecosystem processes (e.g., biogeochemical cycling) through differences in extracellular enzyme production, carbon use efficiency, symbioses, and other ecological traits among microorganisms (14, 44, 61–68). Correspondingly, changes in microbial distribution across environmental gradients can elucidate how soil functions may shift in relation to environmental change, particularly if novel ecosystems supplant existing soil environments (26, 69, 70). For instance, Ladau et al. (71) used current and historical biogeographical patterns of microorganisms to predict microbial distributions and possible ecosystem impacts under future climate scenarios.

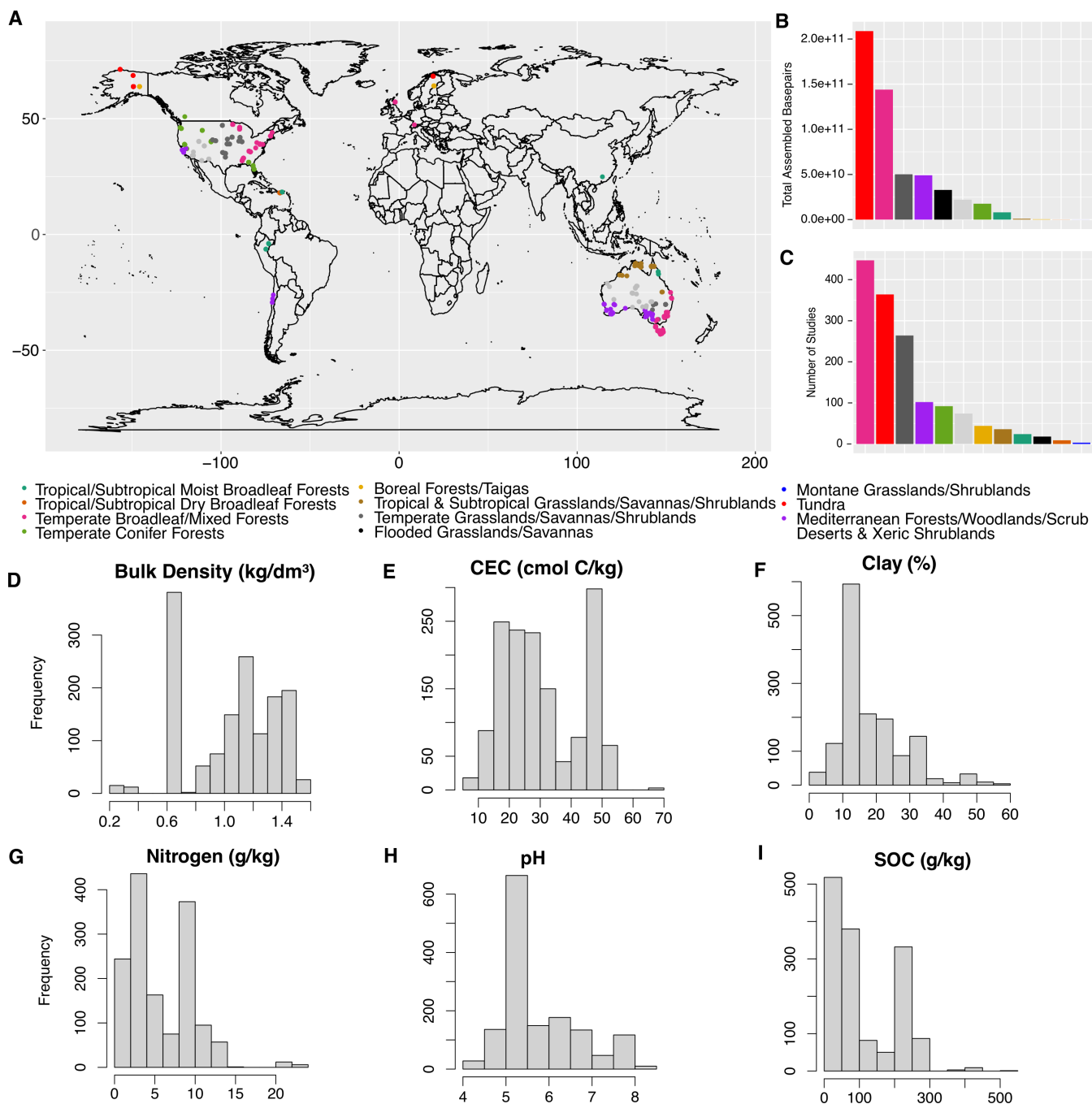
New analytical techniques allow for the direct assessment of the potential functions encoded by soil microorganisms (4, 18), but genomic sequencing technologies also generate a tremendous amount of data. The detection of genes that are nearly ubiquitous (e.g., central metabolisms), poorly annotated, and/or of little relevance to soil processes in particular can obscure ecological patterns in the soil microbiome. Many researchers have attempted to distill this information into tractable units using trait-based approaches, keystone species analyses, and/or core microbiome assessments, with mixed results (27, 44, 62, 72–74). Other common challenges to understanding microbial biogeography include the detection of rare organisms and assigning thresholds for microbial taxonomy, both of which can influence patterns observed in soil microbiomes (34, 75).

As such, fundamental questions regarding global soil microbial distributions and their drivers remain largely unanswered (74, 76). Advanced analytics, including machine learning-based approaches, in combination with public sequence repositories are essential for the next generation of research in microbial ecology (18, 61, 77, 78). Here, we examine a comprehensive set of 1,512 manually curated soil metagenomic samples to decipher patterns in soil genomic potential across commonly measured environmental gradients and ecological biomes. We assay metagenomic sequences using an array of complementary annotation databases, read-based taxonomic assignment, and machine learning algorithms to extract multidimensional genomic fingerprints of global soil microbiomes. Our objective is to uncover never-reported biogeographical patterns in soil microbiome taxonomy and functional potential that can provide a molecular basis for predicting soil microbial responses to climate change.

## RESULTS

### Microbial clades and genomic potential across all soils

We collated soil metagenomic sequences from a wide range of environmental conditions across global biomes (Fig. 1; Tables S1 and S2). Bulk density, cation exchange capacity (CEC), nitrogen (N), pH, soil organic carbon (SOC), and clay content ranged from 0.28 to 1.56 kg/dm<sup>3</sup>, 7.7 to 68.7 cmol C/kg, 0.25 to 22.4 g/kg, 4.4 to 8.3, 2.4 to 510.9 g/kg, and 2.7% to 57.1%, respectively.

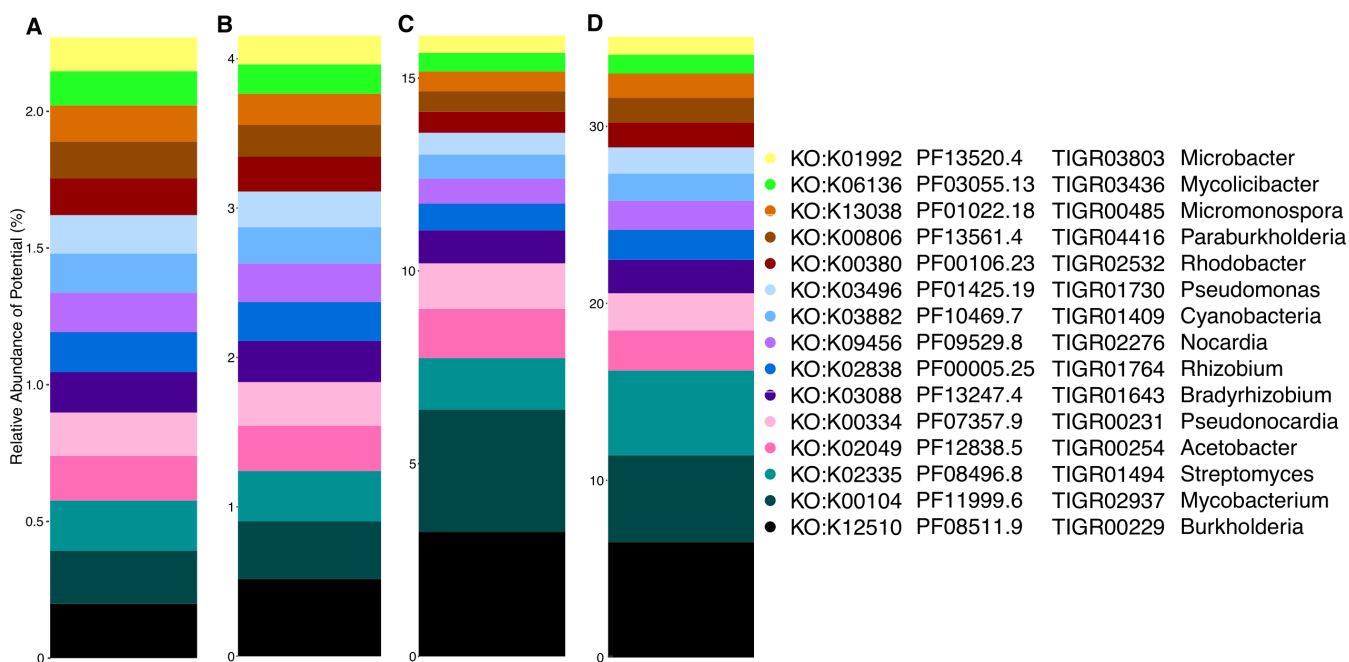


**FIG 1** Data set description. (A) Map of samples and distributions of (B) total assembled bp and (C) number of samples per biome according to Olson et al. (79). (D–I) Distribution of environmental variables across all soils from SoilGrids250m (80).

To reveal the full range of taxa and functional potential encoded by global soils, we first assessed microbial attributes [i.e., gene annotations by the Kyoto Encyclopedia of Genes and Genomes (KEGG) (81, 82), the Pfam (83) and TIGRFAM protein databases (84), and read-based taxonomy; see Materials and Methods] with the highest normalized abundance across all soils (Fig. 2). Multiple annotation databases provide complementary information and are a standard operating procedure by the U.S. Department of Energy Joint Genome Institute (JGI) (85, 86). Abundant KEGG orthologies (KO) were associated bacterial secretion systems (K12510), redox reactions (K00104, K00334, K03882, and K00380), and cellular growth processes (K02335, K02049, K00334, K03882, and K03088). Abundant Pfam were comprised of functions related to the tolerance of common soil stressors including temperature (PF11999, PF08496) and salinity (PF12838), as well as functions involved in microbial growth (PF08511, PF07357, PF13247, and PF00005). The most abundant TIGRFAM were also associated with soil stressors, for example, TIGR00229 (PAS Domain S-box protein) is involved in sensing changes in light, redox potential, oxygen, and other stressors (87). Additional highly abundant TIGRFAM across all soils included a sigma-factor that may be associated with microbial responses to environmental fluctuation (TIGR02937) and other signaling and sensing processes (TIGR00254, TIGR00231). Finally, abundant microbial genera included diverse microorganisms including Burkholderia, Mycobacterium, Streptomyces, Acetobacter, and Pseudonocardia.

### Variation in genomic potential across environmental gradients

To reveal biogeographic patterns in microbiomes across global soils, we extracted microbial attributes that were associated with soil factors and ecological biomes using random forest regression models (see Materials and Methods; Table S1). Soil pH, bulk density, and biome type were the strongest correlates of microbial attributes (i.e., they generated well-fitting models with comparatively high explanatory power relative to other factors; Table S1). The cross-validated random forest model for each of the variables above had a root mean square error of less than 0.8 (often <0.2) and/or at least 60% variance explained. Models for CEC, N, SOC, and clay content did not meet these



**FIG 2** Highly abundant genomic attributes across all soils. Relative abundance of the top 15 most abundant soil microbial genomic attributes annotated by (A) KEGG orthology, protein family [(B) Pfam and (C) TIGRFAM], and (D) read-based taxonomy collated at the genus-level.

thresholds. Results were conceptually consistent across different microbial attributes. A full description of all genomic fingerprints is in Extended Data File 1.

pH separated soil microbiomes by nutrient- and vitamin B-related genes (Fig. 3). More alkaline soils were characterized by vitamin B-, nutrient-, and phosphatase-related activities, with genomic fingerprints that included K01662, PF01872, K08717, TIGR00378, PF01966, and PF00481 for instance, as well as many genera of Alphaproteobacteria. More acidic soils were signified by ammonia- and complex organic polymer- (e.g., wood-) related reactions, with genomic fingerprints that included K14333, K01684, TIGR03404, and TIGR02093, and microbial clades known to use ammonia as a substrate (*Rhodopseudomonas* and *Klebsiella*).

Bulk density separated soil microbiomes by genes associated with soil structure, hydrology, and nutrient content. The genomic fingerprint of low-bulk density soils comprised genes involved in organic N and methane cycling, as well as signatures of anoxia (Fig. 3). It included K00992, K11261, K11959, TIGR03323, TIGR01392, and microorganisms associated with anoxic and/or aquatic environments (*Gardnerella*, *Cetia*, *Salmonella*, *Olleya*, *Mycoplasma*, *Escherichia*, *Enterobacter*, and *Chitinophaga*). Conversely, high-bulk density soils were signified by genes involved in microbial transport, stress, pigmentation, and infection mechanisms. Notable attributes associated with high-bulk density soils included K00854, K11733, K03799, TIGR01263, TIGR01957, TIGR01203, PF00582, PF03631, and four Alphaproteobacteria.

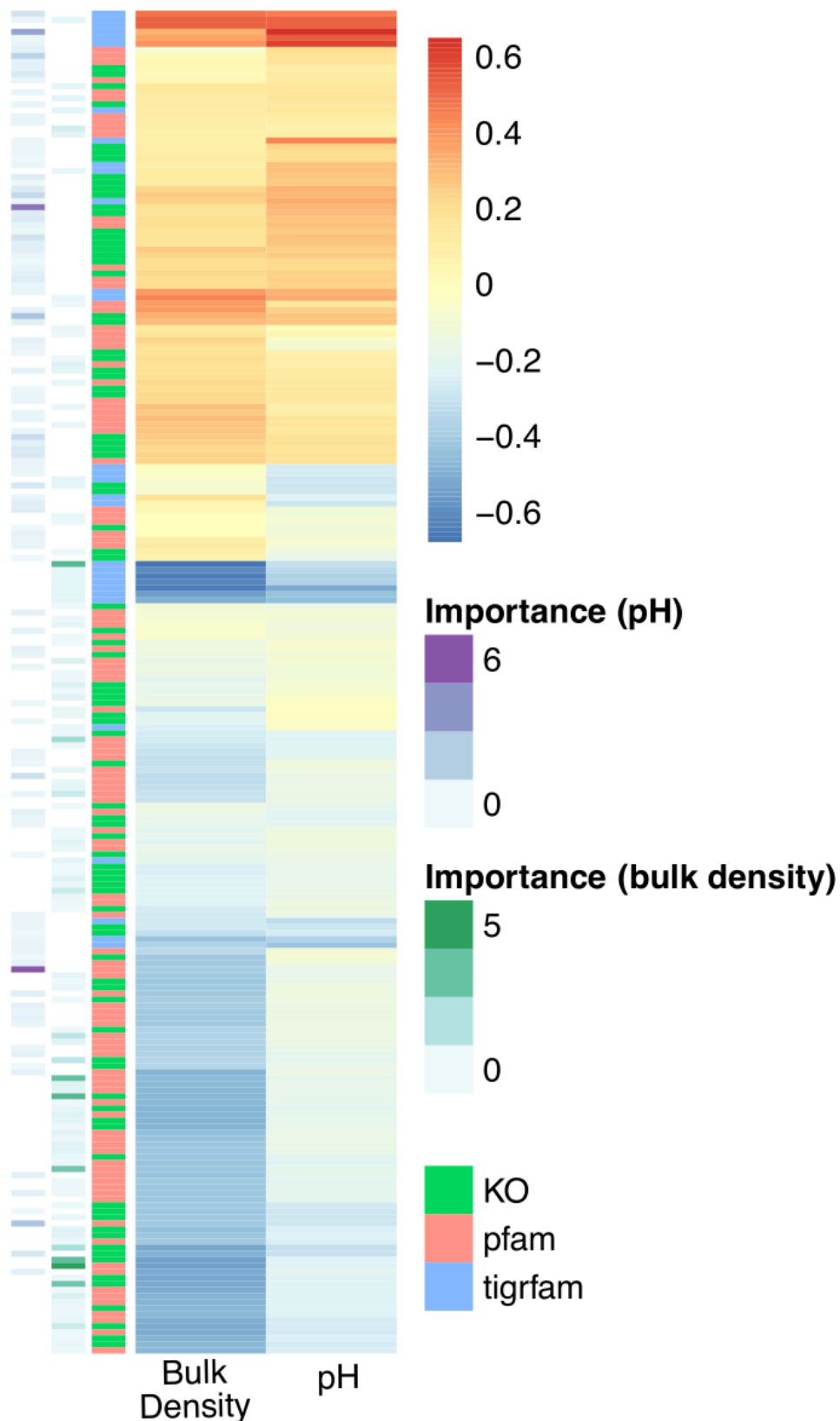
Four biomes were highlighted with distinct genomic signatures by our models: (i) flooded grasslands and savannas, (ii) tundra, (iii) boreal forests, and (iv) montane grasslands/shrublands (Fig. 4). The soil genomic fingerprint of flooded grasslands/savannas was primarily characterized by positive associations with organic N (K13497, K07395, K19702, and K02042), phosphorus (P, K02042, K00854, K02800, and PF04273), stress tolerance (heat and arsenic: TIGR04282 and TIGR02348), and motility (PF00669, PF00482, and *Paeniclostridium*). The genomic fingerprint for tundra had the greatest number of positive gene associations, including those associated with methanogenesis (K11261, TIGR03323, TIGR03321, TIGR03322, TIGR03306, TIGR01392, and TIGR02145), nitric oxide production (PF08768), P transfer (K03525, K00997, K00992, K13497, and PF08009), and pathogenesis (PF11203, PF05045), supported by the presence of anaerobic (*Gardnerella*, *Escherichia*, *Kosakonia*, *Glaesserella*, *Mycoplasma*, *Treponema*, and *Paeniclostridium*) and possible plant growth-promoting genera (*Kosakonia*).

Boreal forests and montane grasslands/shrublands were primarily negatively associated with microbial attributes. Boreal forests displayed negative relationships with the normalized abundance of magnesium- (Mg, K03284) and P- (K06217) related genes and carbohydrate and organic N metabolism (K02800, K11472, PF00208, and PF04685), as well as some metal-related genes (TIGR00378, TIGR04282, and TIGR00003). Montane systems were negatively characterized by genes related to Mg and P (K01520, K04765, and TIGR01203), carbohydrate metabolism (K01816), virulence (PF09299), and motility (PF00669), as well as organic N (K07395, K00600, and PF00208) and S (K03644). Both boreal forests and montane grasslands/shrublands were negatively associated with anaerobic genera (*Gardnerella*, *Thielavia*, *Coniochaeta*, *Escherichia*, *Salmonella*, *Kosakonia*, *Shigella*, *Mycoplasma*, *Treponema*, *Glaesserella*, *Enteractinococcus*, *Paeniclostridium*).

Interestingly in all analyses across pH, bulk density, and biomes, we observed phylogenetic coherence in soil microbiomes—the normalized abundances of more closely related microorganisms tended to move congruently in response to the biome type and environmental factors (Fig. 4B; Extended Data File 1).

### Unsupervised machine learning to disentangle environmental interactions associated with soil genomic potential

Finally, to detect emergent patterns in microbial biogeography, we used unsupervised machine learning to reveal a collection of soils with distinct genomic fingerprints. These soils and their resident microbiomes were associated with coordinated changes in SOC



**FIG 3** Genomic fingerprints of high versus low pH and bulk density soils. Genomic attributes selected in at least one fingerprint for pH or bulk density are visualized. Pearson correlation coefficient is shown from blue to red in the primary heatmap, with bulk density correlations on the left and pH correlations on the right. The sidebars (respectively, from left to (Continued on next page)

FIG 3 (Continued)

right) represent variable importance from random forest models of pH and bulk density and the type of attribute (e.g., KO, Pfam, and TIGRFAM). Please refer to Extended Data 1 for more information on the specific attributes associated with each soil environment. Read-based taxonomy collated at the genus level for each fingerprint is shown in Fig. 4.

content, total N, and CEC and with opposite coordinated changes in bulk density and clay content (Fig. 5; Extended Data 2). We identified KO cluster 7, genus cluster 6, and Pfam cluster 1 as containing soil samples that exhibited the strongest coordinated changes (defined by the number of significant correlations and  $R^2$  values; see Materials and Methods). TIGRFAMs did not exhibit these patterns.

Selected microbial KO and genera were positively correlated with SOC, N, and CEC and negatively associated with bulk density and clay, while Pfam cluster 1 exhibited opposite patterns (i.e., negative associations with SOC, N, and CEC and positive associations with bulk density and clay). Soil microbiomes in KO cluster 7 and genus cluster 6 contained high normalized abundances of genes associated with energy generation (K00334, K00339, K00324, and K08738), filamentous bacteria/biofilms (K11903, Nocardia, Streptomyces, Mycobacterium, and Leptolyngbya), chemically complex organic matter decomposers (Nocardia, Rhodococcus, and Streptomyces), and N-fixing organisms and plant symbionts (Rhizobium, Bradyrhizobium, and Nostoc). Additionally, genes associated with N-cycling processes (K07685, K10041) and microbial interactions (e.g., signaling and secretion, K12537, and K02661) changed most dramatically across environmental gradients. The most variable taxa across environmental gradients were diverse—they included N-cycling (Uliginosibacterium, Microvirgula), autotrophic and/or anaerobic (Methanosalsum, Salana, Pseudobacteroides, Sulfurirhabdus, Microvirgula, Desulfallas, Pelotomaculum, and Risungbinella), and plant-related

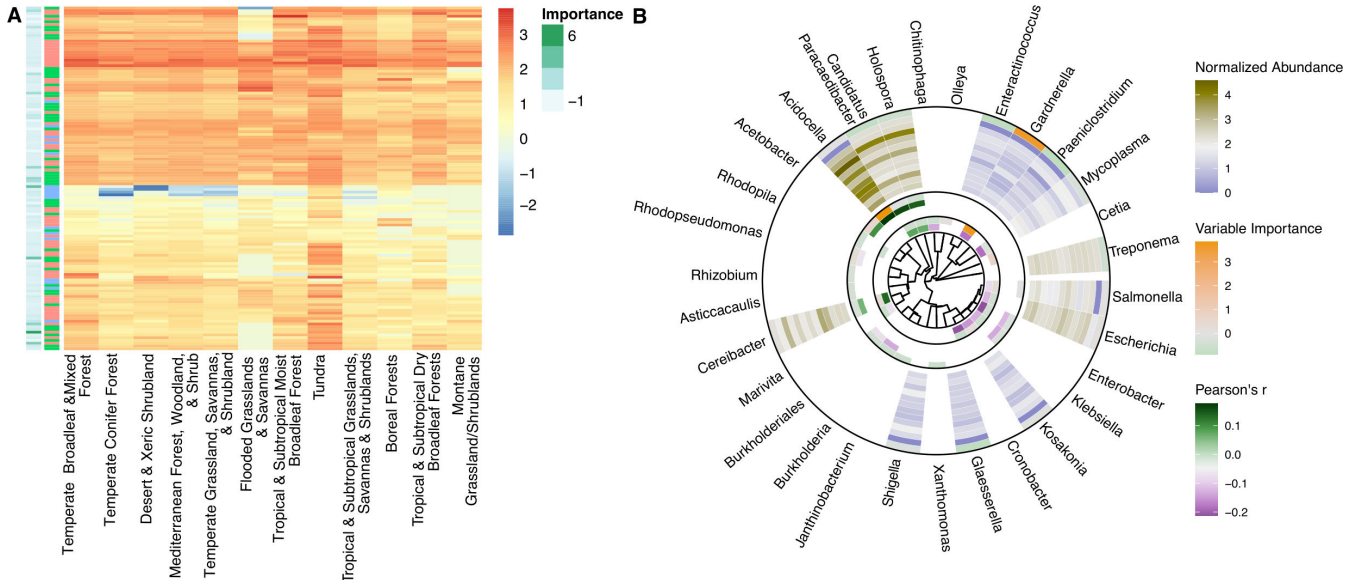
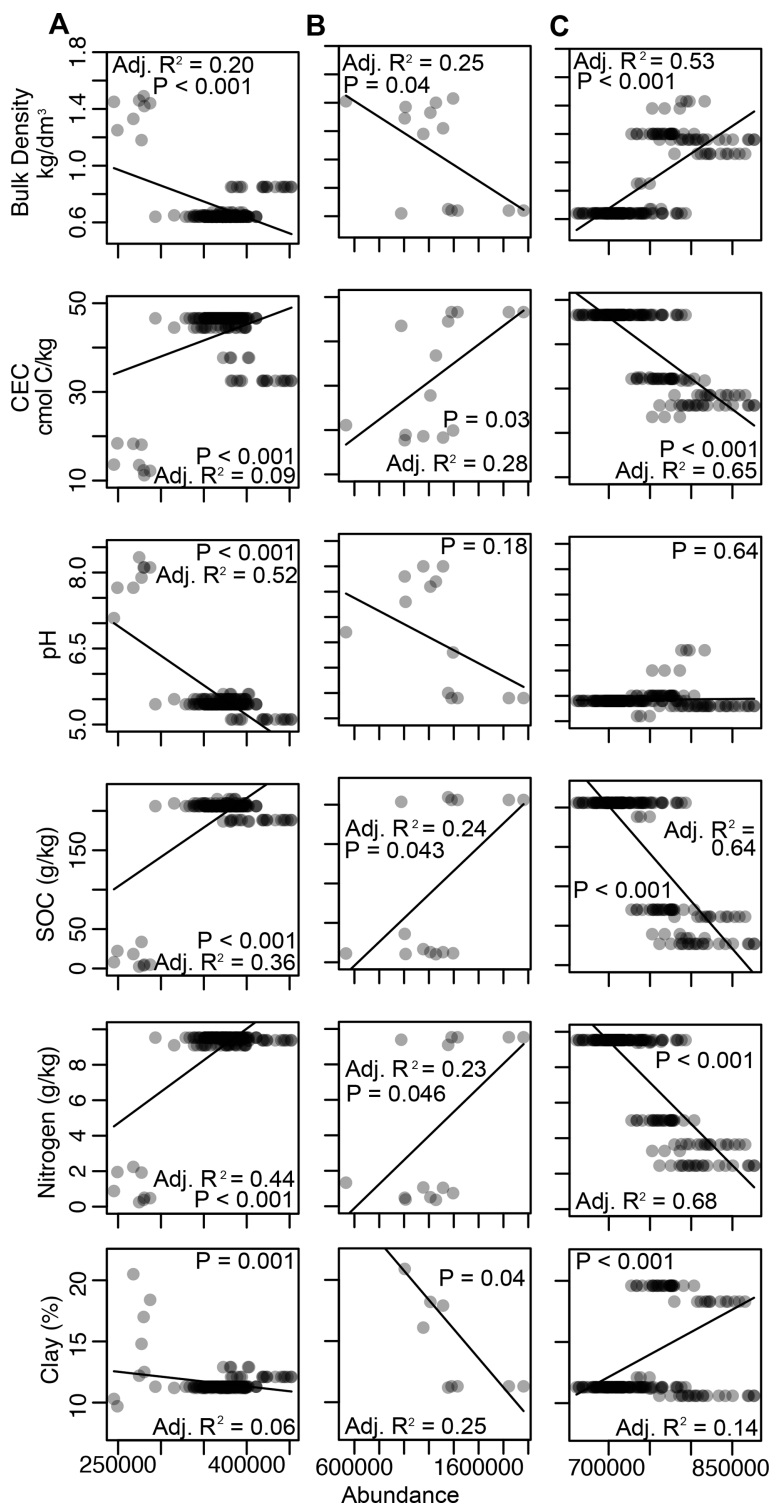


FIG 4 Genomic fingerprints of soil biomes and phylogenetic distribution of selected soil taxa. (A) The normalized abundance of microbial attributes selected in at least one biome fingerprint is shown from blue to red. Variable importance is shown on the left side bar, and attribute type is shown on the right side bar. Please refer to Extended Data 1 for more information on the specific attributes associated with each soil environment. (B) Microbial genera selected in at least one fingerprint are depicted on a phylogenetic tree (generated by PhyloT). The innermost circle shows Pearson correlations with bulk density from green to purple and variable importance from green to orange. The middle circle shows Pearson correlations with pH from green to purple and variable importance from green to orange. The outer circle shows the normalized abundance of genera across biomes (from outer to inner: Temperate Broadleaf and Mixed Forests; Tundra; Mediterranean Forests, Woodlands, and Scrub; Tropical and Subtropical Grasslands, Savannas, and Shrublands; Temperate Grasslands, Savannas, and Shrublands; Boreal Forests/Taigas; Deserts and Xeric Shrublands; Temperate Conifer Forests; Tropical and Subtropical Dry Broadleaf Forests; Flooded Grasslands and Savannas; Tropical and Subtropical Moist Broadleaf Forests; and Montane Grasslands and Shrublands). Variable importance is shown in the outermost ring from green to orange.



**FIG 5** Unsupervised clusters of microbial attributes associated with multiple soil factors. (A) KEGG orthology, (B) microbial genus, and (C) Pfam. TIGRFAM annotations did not yield a satisfactory model. Rows represent environmental variables. The total normalized abundance of an attribute per sample is plotted against each environmental variable. Lines and statistics represent linear regression.

microorganisms (*Pseudobacteroides*, *Zeaxanthinibacter*, and *Parapedobacter*). Samples in Pfam cluster 1 had high relative abundances of organic N- (PF01979, PF08352, and PF01842), P- (PF00406, PF01380), and signaling-related (PF01627, PF09413) genes, while

many genes thought to be associated with eukaryotes changed most dramatically across environmental gradients in Pfam annotations (PF04178, PF03188, and PF01793).

## DISCUSSION

### Microbial taxa and functional potential in global soils

Across all soils, the most abundant microbial attributes spanned disparate lineages and potential functions, highlighting the high microbial diversity of soils relative to other global habitats (5, 18, 45, 46). While we note that we assayed microbial genes which do not necessarily reflect expressed functions (61, 88–91), the genomic potential revealed by this investigation provides a basis for more targeted research into soil microbiome function across different soil habitats. The most prevalent microbial genera included members of commonly observed soil taxa including Alphaproteobacteria, Betaproteobacteria, and Actinobacteria (45). Widespread genomic functions included growth, resource acquisition, and stress tolerance strategies that are essential for microbial persistence in heterogeneous soil environments. Bacterial secretion systems were common, as well as DNA repair/replication, transport, and energy generation functions. Interestingly, genes for the sensing and/or tolerance of fluctuations in temperature, salinity, light, and redox conditions were among the most abundant soil microbial genes; all of which are common soil microbial stressors (92, 93). Collectively, the diverse taxa and functions encoded by the soil microbiome underscores the heterogeneous stressors on the soil microbiome. The high global soil microbial diversity observed here can obscure patterns in microbiome structure and function, indicating the use of machine learning-based tools to unravel its complexity (18, 45, 88).

### Soil genomic potential corresponds to pH, bulk density, and biome type

Soil genomic patterns were evident across soil pH, bulk density, and biomes; however, some hypothesized drivers of the soil microbiome including SOC, total N, and CEC were not statistically linked to microbial taxa or genomic functional potential. Soil carbon, N, and cation exchange capacity can reflect resource availability and are often associated with soil microbiome change at local scales (14, 36, 89–91). The lack of trends at the global scales may represent heterogeneity in the importance of resources in structuring microbiomes; indeed, the influence of deterministic (e.g., selection) versus stochastic (e.g., random) assembly processes on soil microbiomes varies through space and time (30, 94–97). While our results support pH and biome type as primary determinants of soil microbial structure and function (5, 6, 28, 45), it is interesting that bulk density was also a strong correlate of genomic patterns. This may reflect the importance of biophysical and hydraulic properties on soil microbiome structure and function (94, 98–100).

Collectively, we reveal novel biogeographical patterns in specific aspects of soil microbiome composition and potential function that have not been reported elsewhere. Our results are consistent with ecological processes hypothesized to influence the soil microbiome and provide greater molecular specificity than that revealed by previous works. They enable us to better understand microbial lifestyles that support the persistence of certain members over others; such knowledge underlies a predictive understanding of soil microbiome changes with shifts in the global climate.

### Microbial nutrient acquisition potential shifts across pH gradients

The genomic fingerprint of alkaline soils demonstrated an association with microbial nutrient acquisition and is important for understanding belowground microbial dynamics in arid and semiarid ecosystems [ $>35\%$  of global land surface (101)]. The soil matrix in these systems is porous, has low rates of water retention, and contains an abundance of calcium carbonate that limits P bioavailability (102). Accordingly, the alkaline soil fingerprint was enriched in genes encoding phosphatases and related functions (PF01966, PF00481) that may signify the importance of microbial P acquisition and/or limitation in alkaline soils (103–106). Vitamin B-related genes were also included

in this fingerprint and have been linked to P solubilization (K01662, PF01872) (107). The presence of a urea transporter (K08717) may likewise indicate the importance of nutrient acquisition in alkaline soils (108). Notably, the high- $\text{Ca}^{2+}$  alkaline soil environment was also characterized by a calcium/proton exchanger (TIGR00378)—which regulates intracellular  $\text{Ca}^{2+}$  concentrations to prevent  $\text{Ca}^{2+}$  overload (109) and an uncharacterized hydrophobic domain (TIGR00271) that may arise from low-moisture environments that are often characteristic of alkaline soils.

In contrast, more acidic soils were characterized by a distinct set of nutrient acquisition genes, in particular genes encoding ammonia and wood decomposition reactions that produce  $\text{H}^+$  as a byproduct. For instance, nitrification is a hallmark of microbial genera *Rhodopseudomonas* and *Klebsiella* (110–112) and increases soil acidity through the release of  $\text{H}^+$  from ammonia. Likewise, wood-degrading fungi often secrete acidic compounds to aid in the digestion of chemically complex polymers (113–116). This is consistent with the association of K14333, K01684, TIGR03404, and TIGR02093 with low-pH soils.

Finally, Alphaproteobacteria have been previously shown to be sensitive to variation in soil pH (117) and were included in the genomic fingerprints of both high- and low-pH soils, further bolstering their potential sensitivity to global change processes that impact soil pH (e.g., atmospheric nitrogen deposition).

### **Genes encoding stress, transport, and/or redox-based processes change with bulk density**

Bulk density reflects soil biophysical properties such as texture, water content, porosity, and mineralogy that drive microbial community structure and function (19, 53, 54, 118, 119). Patterns in the relationships between these variables and the soil microbiome can be complex at global scales (19), and our findings are to our knowledge the first to report a global-scale relationship between bulk density and the distribution of genes encoded by the soil microbiome.

Low-bulk density soils were characterized by genes common to water-logged and/or agricultural soils, while high-bulk density soils included potential functions associated with transport mechanisms, stress tolerance, and virulence. For instance, K11261, TIGR03323, and TIGR01392—all signatures of low-bulk density soil—are each associated with methanogenic processes that are prevalent in saturated soils (120). Low-bulk density soils were further characterized by anaerobic and/or aquatic microorganisms including *Gardnerella*, *Cetia*, *Salmonella*, *Olleya*, *Mycoplasma*, *Escherichia*, *Enterobacter*, and *Chitinophaga* that also denote lifestyles associated with microbial persistence in water-saturated environments (121–129). Though we did not directly assess land cover, the presence of potential functions associated with organic N cycling in the low-bulk density fingerprint (K00992, K11959) may additionally suggest the influence of agriculture on soil microorganisms, where tillage generates loosely packed soils (i.e., low bulk density) and fertilizer application enhances nutrient cycles.

In contrast, the genomic fingerprint of high-bulk density soils suggests an environment inclusive of density-dependent relationships between microorganisms, their substrates, and pathogens (transport: K00854, K11733, TIGR01957, and TIGR01203; virulence: PF03631 and TIGR01203). Stress-related genes (K03799, PF00582.24) may be reflective of this competitive environment or related to ecosystem processes such as soil compaction which is common in deserts (K03799).

### **Soil genomic potential for greenhouse gas emissions differs by biome**

Finally, the soil habitat (reflected here by biome type) seemed to be a driver of dissimilarity in microbiome structure and potential function, providing detailed information on the molecular biology of global soil ecosystems.

We first highlight microbial genomic attributes of tundra ecosystems, as they are experiencing rapid transformations in response to global climate change (130, 131). Tundra exhibited the highest number of microbial taxa and potential functions in their

genomic fingerprint, the vast majority of which displayed positive associations. Given the relationship between soil microbial diversity and biogeochemical function, this provides a consistent and distributed genomic basis for the enhanced greenhouse gas emissions that have been associated with permafrost thaw in tundra (130, 132, 133). The microbial attributes associated with tundra in this study reveal clues into the biology that drives active biogeochemical cycles in these vulnerable ecosystems. Methanogenesis in particular is responsible for fluxes of methane in warming tundra (132, 134, 135) and was associated with tundra in the current study. Additionally, a gene involved in nitric oxide cycling (PF08768) (136–138), as well as a variety of anaerobic organisms and P-related genes (K03525, K00997, K00992, K13497, and PF08009) (139–141), also provide insight into the molecular functions of tundra in the context of global climate change. Interestingly, we also found evidence for pathogen resistance within the tundra genomic fingerprint (PF11203, PF05045), lending credence to a small number of recent studies suggesting permafrost environments as one of the largest reservoirs of soil viruses (142–145) and a possible linkage between soil methane and viral infection (146).

Flooded grasslands/savannas were also generally positively associated with many microbial attributes, several of which encoded nutrient cycling, stress tolerance, and/or motility traits. Genes related to organic N cycling (K13497, K07395, K19702, and K02042), P (K02042, K00854, K02800, and PF04273), and/or stress tolerance (TIGR04282, TIGR02348) may signify microbial adaptations to nutrient scarcity and other stressors. Motility (PF00669, PF00482, and Paeniclostridium) may be reflective of beneficial microbial lifestyles in saturated systems, where movement is facilitated by high pore space connectivity. Interestingly, a mechanism for arsenic resistance was affiliated with flooded grasslands and savannas, possibly denoting arsenic groundwater contamination in many regions of the world (147).

In contrast, boreal forests and montane grasslands/shrublands were signified by negative associations with Mg- and P-related genes (boreal: K03284 and K06217; montane: K01520, K04765, and TIGR01203) and with organic N- and carbohydrate-related genes (boreal: K02800, K11472, PF00208, and PF04685; montane: K01816, K07395, K00600, and PF00208). In addition, anaerobic microbial genera were negatively associated with these ecosystems. The boreal forest fingerprint uniquely included some metal-related genes (TIGR00378, TIGR04282, and TIGR00003), while montane systems were uniquely characterized by virulence (PF09299), motility (PF00669), and sulfur-related genes (K03644). The negative associations with carbon, N, and P suggest that carbon and inorganic nutrients may be more available in these systems relative to other biomes; and Mg is often associated with plant growth (148). Taken together, microbial fingerprints of boreal forests and montane grasslands/shrublands convey slower growth rates and fewer resource constraints than other ecosystems.

### **Resource scavenging, plant-microbe interactions, fungi, and heterotrophic metabolisms prevalent in fertile soils**

Relationships between biomes and specific microbial attributes suggest that there may be multiple simultaneous factors that structure soil microbiome composition and function. Because of this, we next explored relationships between microbial attributes and soil factors that arose naturally out of unsupervised analysis. This allowed us to infer interactive effects that could not be assessed when specifying a single environmental gradient.

Unsupervised machine learning uncovered soil microbial genomic attributes associated with coordinated changes in SOC, total N, and CEC and oppositely with clay content and bulk density that may represent some of the most biologically active soils on Earth (Fig. 5). Soils rich in SOC and N with high CEC constitute resource- and energy-rich environments, supported by an abundance of energy-generating genes in the fingerprint of these soils (K00334, K00339, K00324, and K08738). Interestingly, these environments are best explained by a combination of environmental variables rather than an individual factor. As an example, soils with high SOC alone could be limited by

N or other nutrients; thus, we propose that genomic fingerprints for multiple simultaneously interacting factors may be among the most critical for understanding global biogeochemistry. They are nearly impossible to discern without unsupervised statistical approaches, and we know of no other work that has done so.

Soils with high SOC, N, and CEC were characterized by genes associated with resource scavenging, plant-microbe interactions, and heterotrophic metabolisms. We observed numerous attributes in their genomic fingerprint associated with filaments or biofilms, intra-organism signaling, and/or chemically complex organic matter decomposition (K11903, K12537, K02661, *Nocardia*, *Streptomyces*, *Mycobacterium*, *Leptolyngbya*, *Nocardia*, *Rhodococcus*, and *Streptomyces*). This is consistent with microbial morphologies and lifestyles that support the decay of chemically recalcitrant polymers (e.g., wood), as well as nutrient acquisition and transport across soil pore networks. Plant-microbe interactions, including symbiotic N-fixing organisms, were also a key feature of these soils, highlighting the importance of above- and belowground connectivity in productivity (*Rhizobium*, *Bradyrhizobium*, *Nostoc*, *Pseudobacteroides*, *Zeaxanthinibacter*, *Parapedobacter*, and *Zea*) (149). In contrast, anaerobic and autotrophic organisms were highly sensitive to changes across these connected environmental gradients, possibly reflecting their exclusion from more oxygen- and energy-rich environments where heterotrophic organisms tend to dominate.

We also selected one Pfam cluster that was associated with low SOC, N, and CEC and with high clay and bulk density soils, which seemed to underscore the importance of fungi in resource- and energy-rich environments. The most sensitive genes in this cluster of samples corresponded to eukaryotic organisms (PF04178, PF03188, and PF01793). Previous work has shown fungal sensitivity to soil resource availability, as well as high fungal diversity in organic-rich tropical soils (52). Overall, these insights provide greater resolution into the microbial mechanisms that support both belowground biogeochemical cycles and aboveground productivity than ever before, and they can provide a basis for the next generation of microbial ecology research and model development.

### Phylogenetic coherence in microbial biogeography

The extent to which selection acts on phylogenetically conserved properties of soil microbiomes and further translates into differences in functional potential [e.g., response-effect traits (150)] is a long running unknown in microbial ecology. Some previous studies have shown little connection between functional attributes and microbial phylogeny (44, 151), while others have indicated the existence of taxa-function relationships (88). And still others have suggested variation in the coupling of taxonomy and function across various functions of interest, levels of phylogenetic resolution, and/or spatial scales (34, 55, 61, 152, 153). Though microbial taxa are generally less correlated to the environment than functional potential (Table S1), we found strong phylogenetic coherence in the response of genera to changes in the soil environment (Fig. 4B). That is, closely related organisms tended to respond similarly to environmental change. This implies that evolutionary history, life history strategies, and/or morphology [all traits that tend to be phylogenetically conserved (154)] could influence changes in microbial community membership in response to environmental change. As such, metabarcoding approaches that yield microbiome taxonomy may be among the first indicators of ecosystem transitions, such as response to disturbances and transitions between state archetypes (30).

### Conclusions

Despite the significance of soil microorganisms in global biogeochemistry, we still know little about the ecological processes that regulate their community composition and function across the wide range of global soil environments. We used machine learning to sift through tens of thousands of taxa and potential functions encoded by 1,512 global soil microbiomes, revealing novel biogeographical patterns in soil microbiome composition and functional potential with high molecular resolution. Specifically, we

demonstrate shifts in the potential for (i) microbial nutrient acquisition across pH gradients; (ii) stress, transport, and redox-based processes across changes in soil bulk density; and (iii) genes and organisms associated with soil greenhouse gas emissions across biomes. We also uncover a suite of metabolic processes that are enriched in the microbial genomes of energy-rich soils. These changes were coincident with phylogenetically congruent compositional shifts—suggesting that closely related soil microbial taxa are sensitive to similar environmental stressors. Our work enables us to better understand microbial lifestyles that support the persistence of certain members over others along environmental gradients; such knowledge underlies a predictive understanding of soil microbiome changes with shifts in the global climate and is vital to constraining biochemical reactions that regulate the release of greenhouse gasses from soils.

## MATERIALS AND METHODS

### Data set description and normalization

We collected 1,512 manually curated soil metagenomic samples available in the Joint Genome Institute's (JGI) Integrated Microbial Genomes and Microbiomes (IMG/M) platform in August 2020 as described by Graham et al. 2024 (accepted), of which 1,493 were associated with latitudinal and longitudinal information. Samples span most major biomes and continents (Fig. 1; Tables S2 and S3). Sequences were quality controlled, assembled, and annotated by KEGG Orthology (KO), the Pfam and TIGRFAM protein family databases, and read-based taxonomy by the JGI's standard workflows (82–86). While sample collection and sequencing methods inevitably vary in metaanalyses such as the current study, we applied standardized analytical workflows to minimize methodological artifacts to the greatest extent possible. Due to the collective size of assembled sequences, we focused on annotations and read-based taxonomy for downstream analysis, which is more computationally feasible than sequence assemblies or metagenome-assembled genomes that are impractical at the scale of thousands of samples.

Each data set (i.e., KO, Pfam, TIGRFAM, and read-based taxonomy) was independently normalized by the following procedure prior to analysis. For Pfam and KO annotations only, samples with greater than 2,000 total reads were retained. For TIGRFAM annotations only, samples with greater than 500 total reads were retained. A lower cutoff was used for TIGRFAM data as the total read counts were lower in this data set. For all three annotations, only functions with a non-zero value in at least half of the samples or a total count of 7,560 (corresponding to an average of 5 reads per sample for each function) were retained. For microbial genera, we retained only samples with greater than 4,000 counts and genera with a non-zero value in at least half of the samples or a total count of 7,560 (corresponding to an average of 5 reads per sample for each genus). After removing low-abundance samples and functions/genera, all four datasets were then normalized using the trimmed mean of M values method (155). We found that this method gave us the best results when examining normalized data using a box and whisker plot (Figure S1).

We assigned the biome type and soil properties to each sample using publicly available resources. We used latitude and longitude to derive using Olson biomes (79) for each sample using data provided by the World Wildlife Federation (156). Soil parameters were collected from the SoilGrids250m database from 0 to 5 cm (80). SoilGrids250m is a spatial interpolation of global soil parameters using ~150,000 soil samples and 158 remote sensing-based products. Here, we focus on six parameters often associated with soil microbiomes: bulk density, cation exchange capacity, nitrogen, pH, soil organic C, and clay content. Because our focus on spatial dynamics and soils were collected at various times, we did not include temporally dynamic variables such as soil moisture or temperature in our set of environmental parameters, though we acknowledge they may have profound impacts on the soil microbiome. Additionally, because soil data are derived from a comprehensive spatial interpolation and not measured directly on each

soil sample, our models may fail to capture local-scale heterogeneity in relationships. Further investigations that generate and leverage fully standardized data are needed to resolve this discrepancy. Nonetheless, local heterogeneity should statistically weaken the relationships observed here, and our results present a promising investigation into the global biogeography of soil microbiomes.

## Machine learning for genomic fingerprints

We selected random forests as our primary statistical approach by assessing the ability of various machine learning algorithms to adequately parse soil microbial attributes across environmental variables and ecological biomes. Models were built and validated using 10-fold cross-validation (CV) repeated five times, 50% of the data for training, and 50% of the data for testing. Models from different algorithms were compared using CV scores and variance explained (Figure S2). Random forest models consistently yielded comparatively high CV scores and proportion of variance explained.

Thus, we constructed random forest models to extract genomic fingerprints from soil microbiomes using two approaches: (i) a supervised split across variables known to be associated with soil microbiomes and (ii) an unsupervised approach that allowed predictors and their interactions to naturally arise. Random forest models were built in R Software (157) for each attribute type independently using the “caret” package (158).

In (i), we divided samples into either two groups containing the highest and lowest 10% of values for each environmental factor or one group per biome. We then used random forests to rank each microbial attribute according to their importance (i.e., difference in mean square error normalized by standard error, ‘varImp’ function in “caret” package) to differentiate samples across each environmental factor or biome. Subsequently, because variable importances followed an exponential decay, we used breakpoint analysis to define a set of the most important variables distinguishing each group (“strucchange” package,  $h = 0.01$  with a minimum segment size of two) (159). Because this resulted in one set of attributes that drove variation in soil microbiomes in relation to a given environmental factor or biome, we extracted attributes associated with high (positive slope) or low (negative slope) values of each environmental factor using linear regression. The resultant attributes comprised two genomic fingerprints per environmental factor or biome—one associated with low values and one associated with high values.

In (ii), we first used k-means clustering to group samples with similar microbial attributes using R packages: “vegan” (160), “mgcv” (161), and “cluster” (162). The optimal number of clusters was chosen by evaluating the total sum of squares and the silhouette index (163). To determine the extent to which soil microbiomes within each cluster corresponded to changes in environmental factors, we related the total normalized abundance of attributes per sample to each environmental factor using linear regression. We then selected one cluster per attribute type for deeper investigation, chosen by the number of significant correlations with environmental factors. To generate a genomic fingerprint for the selected cluster, we extracted attributes that were either most abundant or most variable within the cluster. As in approach (i), we used breakpoint analysis to define a set of the most abundant attributes per cluster (“strucchange” package,  $h = 0.01$  with a minimum segment size of two). To determine the most variable attributes per cluster, we calculated the coefficient of variation for each attribute and then used breakpoint analysis to extract the most variable attributes.

We used the following packages for data manipulation and visualization: “ggplot2” (164), “dplyr” (165), “factoextra” (166), “Hmisc” (167), “colorspace” (168), “RColorBrewer” (169), “gridExtra” (170), “tidyverse” (171), and “maps” (172). Additionally, for visualizations of read-based taxonomy, we generated phylogenetic trees using phyloT v2 (173), an online tree generator based on the Genome Taxonomy Database. Then, we visualized the tree in R using the packages “ggtree” (174), “treeio” (175), and “ggnewscale” (176). We generated heatmaps using the “pheatmap” package (177).

## ACKNOWLEDGMENTS

This work was performed under a Laboratory Directed Research and Development (LDRD) from the Pacific Northwest National Laboratory (PNNL). PNNL is operated by the Battelle Memorial Institute for the U.S. Department of Energy under Contract No. DE-AC05-76RL01830. We thank the U.S. Department of Energy Joint Genome Institute (JGI) for maintaining the public Integrated Microbial Genomes and Microbiomes (IMG/M) platform from which we obtained this data set.

## AUTHOR AFFILIATIONS

<sup>1</sup>Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, Washington, USA

<sup>2</sup>School of Biological Sciences, Washington State University, Pullman, Washington, USA

<sup>3</sup>Department of Biological and Ecological Engineering, Oregon State University, Corvallis, Oregon, USA

## AUTHOR ORCIDs

Emily B. Graham  <http://orcid.org/0000-0002-4623-7076>

Vanessa A. Garayburu-Caruso  <http://orcid.org/0000-0003-3383-6237>

Ruonan Wu  <http://orcid.org/0000-0001-9466-4462>

Jianqiu Zheng  <http://orcid.org/0000-0002-1609-9004>

Ryan McClure  <http://orcid.org/0000-0003-0573-6917>

Gerrad D. Jones  <http://orcid.org/0000-0002-1529-9506>

## FUNDING

Funder	Grant(s)	Author(s)
DOE   SC   Pacific Northwest National Laboratory (PNNL)		Emily B. Graham

## AUTHOR CONTRIBUTIONS

Emily B. Graham, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing | Vanessa A. Garayburu-Caruso, Data curation, Formal analysis, Investigation, Methodology, Writing – review and editing | Ruonan Wu, Data curation, Formal analysis, Investigation, Methodology, Supervision, Writing – review and editing | Jianqiu Zheng, Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review and editing | Ryan McClure, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Writing – review and editing | Gerrad D. Jones, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Writing – review and editing

## DATA AVAILABILITY

All data are publicly available via the JGI IMG/MG platform. Please see Tables S2 and S3 for further information. Code for data analysis and all figures is available at [10.6084/m9.figshare.25374427](https://doi.org/10.6084/m9.figshare.25374427).

## ADDITIONAL FILES

The following material is available [online](#).

## Supplemental Material

**Extended Data S1** (mSystems01112-23-s0001.xlsx). Supervised fingerprints.

**Extended Data S2** (mSystems01112-23-s0002.xlsx). Unsupervised fingerprints.

**Figure S1** (mSystems01112-23-s0003.pdf). Abundances across different normalization procedures.

**Figure S2** (mSystems01112-23-s0004.pdf). Comparison of machine learning algorithms.

**Legends** (mSystems01112-23-s0005.docx). Legends for Figures S1 and S2.

**Table S1** (mSystems01112-23-s0006.xlsx). Model performance for random forest regressors.

**Table S2** (mSystems01112-23-s0007.csv). Data sets and associated sequence metadata.

**Table S3** (mSystems01112-23-s0008.csv). Geographic metadata for each sample.

## Open Peer Review

**PEER REVIEW HISTORY** (review-history.pdf). An accounting of the reviewer comments and feedback.

## REFERENCES

- Falkowski PG, Fenchel T, Delong EF. 2008. The microbial engines that drive earth's biogeochemical cycles. *Science* 320:1034–1039. <https://doi.org/10.1126/science.1153213>
- Wagg C, Bender SF, Widmer F, van der Heijden MGA. 2014. Soil biodiversity and soil community composition determine ecosystem multifunctionality. *Proc Natl Acad Sci U S A* 111:5266–5270. <https://doi.org/10.1073/pnas.1320054111>
- Delgado-Baquerizo M, Maestre FT, Reich PB, Jeffries TC, Gaitan JJ, Encinar D, Berdugo M, Campbell CD, Singh BK. 2016. Microbial diversity drives multifunctionality in terrestrial ecosystems. *Nat Commun* 7:10541. <https://doi.org/10.1038/ncomms10541>
- Fierer N. 2017. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat Rev Microbiol* 15:579–590. <https://doi.org/10.1038/nrmicro.2017.87>
- Fierer N, Strickland MS, Liptzin D, Bradford MA, Cleveland CC. 2009. Global patterns in belowground communities. *Ecol Lett* 12:1238–1249. <https://doi.org/10.1111/j.1461-0248.2009.01360.x>
- He L, Mazza Rodrigues JL, Soudzilovskaia NA, Barceló M, Olsson PA, Song C, Tedersoo L, Yuan F, Yuan F, Lipson DA, Xu X. 2020. Global biogeography of fungal and bacterial biomass carbon in topsoil. *Soil Biology and Biochemistry* 151:108024. <https://doi.org/10.1016/j.soilbio.2020.108024>
- Melillo JM, Frey SD, DeAngelis KM, Werner WJ, Bernard MJ, Bowles FP, Pold G, Knorr MA, Grandy AS. 2017. Long-term pattern and magnitude of soil carbon feedback to the climate system in a warming world. *Science* 358:101–105. <https://doi.org/10.1126/science.aan2874>
- Bradford MA, Wieder WR, Bonan GB, Fierer N, Raymond PA, Crowther TW. 2016. Managing uncertainty in soil carbon feedbacks to climate change. *Nature Clim Change* 6:751–758. <https://doi.org/10.1038/nclimate3071>
- Davidson EA, Janssens IA. 2006. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature* 440:165–173. <https://doi.org/10.1038/nature04514>
- Galloway JN, Townsend AR, Erisman JW, Bekunda M, Cai Z, Freney JR, Martinelli LA, Seitzinger SP, Sutton MA. 2008. Transformation of the nitrogen cycle: recent trends, questions, and potential solutions. *Science* 320:889–892. <https://doi.org/10.1126/science.1136674>
- Erisman JW, Galloway JN, Seitzinger S, Bleeker A, Dise NB, Petrescu AMR, Leach AM, de Vries W. 2013. Consequences of human modification of the global nitrogen cycle. *Philos Trans R Soc Lond B Biol Sci* 368:20130116. <https://doi.org/10.1098/rstb.2013.0116>
- Sparling GP. 1997. Soil microbial biomass, activity and nutrient cycling as indicators of soil health, p 97–119. In *Biological indicators of soil health*
- Wieder WR, Bonan GB, Allison SD. 2013. Global soil carbon projections are improved by modelling microbial processes. *Nature Clim Change* 3:909–912. <https://doi.org/10.1038/nclimate1951>
- Graham EB, Hofmockel KS. 2022. Ecological stoichiometry as a foundation for omics-enabled biogeochemical models of soil organic matter decomposition. *Biogeochemistry* 157:31–50. <https://doi.org/10.1007/s10533-021-00851-2>
- Fan X, Gao D, Zhao C, Wang C, Qu Y, Zhang J, Bai E. 2021. Improved model simulation of soil carbon cycling by representing the microbially derived organic carbon pool. *ISME J* 15:2248–2263. <https://doi.org/10.1038/s41396-021-00914-0>
- Wang G, Mayes MA, Gu L, Schadt CW. 2014. Representation of dormant and active microbial dynamics for ecosystem modeling. *PLoS One* 9:e89252. <https://doi.org/10.1371/journal.pone.0089252>
- Qin S, Chen L, Fang K, Zhang Q, Wang J, Liu F, Yu J, Yang Y. 2019. Temperature sensitivity of SOM decomposition governed by aggregate protection and microbial communities. *Sci Adv* 5:eau1218. <https://doi.org/10.1126/sciadv.aau1218>
- Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, et al. 2017. A communal catalogue reveals earth's multiscale microbial diversity. *Nature* 551:457–463. <https://doi.org/10.1038/nature24621>
- Xu X, Wang N, Lipson D, Sinsabaugh R, Schimel J, He L, Soudzilovskaia NA, Tedersoo L. 2020. Microbial macroecology: in search of mechanisms governing microbial biogeographic patterns. *Global Ecol Biogeogr* 29:1870–1886. <https://doi.org/10.1111/geb.13162>
- Levin SA. 1992. The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. *Ecology* 73:1943–1967. <https://doi.org/10.2307/1941447>
- Dickey JR, Swenie RA, Turner SC, Winfrey CC, Yaffar D, Padukone A, Beals KK, Sheldon KS, Kivlin SN. 2021. The utility of macroecological rules for microbial biogeography. *Front Ecol Evol* 9. <https://doi.org/10.3389/fevo.2021.633155>
- Brown JH, Gillooly JF, Allen AP, Savage VM, West GB. 2004. Toward a metabolic theory of ecology. *Ecology* 85:1771–1789. <https://doi.org/10.1890/03-9000>
- Wright DH. 1983. Species-energy theory: an extension of species-area theory. *Oikos* 41:496. <https://doi.org/10.2307/3544109>
- Blackburn TM, Gaston KJ, Loder N. 1999. Geographic gradients in body size: a clarification of Bergmann's rule. *Divers Distrib* 5:165–174. <https://doi.org/10.1046/j.1472-4642.1999.00046.x>
- Stevens GC. 1989. The latitudinal gradient in geographical range: how so many species coexist in the tropics. *Am Nat* 133:240–256. <https://doi.org/10.1086/284913>
- Meyer KM, Memiaghe H, Korte L, Kenfack D, Alonso A, Bohannon BJM. 2018. Why do microbes exhibit weak biogeographic patterns? *ISME J* 12:1404–1413. <https://doi.org/10.1038/s41396-018-0103-3>
- Chu H, Gao G-F, Ma Y, Fan K, Delgado-Baquerizo M. 2020. Soil microbial biogeography in a changing world: recent advances and future

- perspectives. *mSystems* 5:e00803-19. <https://doi.org/10.1128/mSystems.00803-19>
28. Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PM, Bengtsson-Palme J, Anslan S, Coelho LP, Harend H, Huerta-Cepas J, Medema MH, Maltz MR, Mundra S, Olsson PA, Pent M, Pölme S, Sunagawa S, Ryberg M, Tedersoo L, Bork P. 2018. Structure and function of the global topsoil microbiome. *Nature* 560:233–237. <https://doi.org/10.1038/s41586-018-0386-6>
  29. Nemergut DR, Schmidt SK, Fukami T, O'Neill SP, Bilinski TM, Stanish LF, Knelman JE, Darcy JL, Lynch RC, Wickey P, Ferrenberg S. 2013. Patterns and processes of microbial community assembly. *Microbiol Mol Biol Rev* 77:342–356. <https://doi.org/10.1128/MMBR.00051-12>
  30. Graham EB, Knelman JE. 2023. Implications of soil microbial community assembly for ecosystem restoration: patterns, process, and potential. *Microb Ecol* 85:809–819. <https://doi.org/10.1007/s00248-022-02155-w>
  31. Finlay BJ. 2002. Global dispersal of free-living microbial eukaryote species. *Science* 296:1061–1063. <https://doi.org/10.1126/science.1070710>
  32. Greening C, Grinter R, Chiri E. 2019. Uncovering the metabolic strategies of the dormant microbial majority: towards integrative approaches. *mSystems* 4:e00107-19. <https://doi.org/10.1128/mSystems.00107-19>
  33. Lennon JT, Jones SE. 2011. Microbial seed banks: the ecological and evolutionary implications of dormancy. *Nat Rev Microbiol* 9:119–130. <https://doi.org/10.1038/nrmicro2504>
  34. Horner-Devine MC, Lage M, Hughes JB, Bohannan BJM. 2004. A taxa-area relationship for bacteria. *Nature* 432:750–753. <https://doi.org/10.1038/nature03073>
  35. Martiny JBH, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, Green JL, Horner-Devine MC, Kane M, Krumins JA, Kuske CR, Morin PJ, Naeem S, Ovreås L, Reysenbach A-L, Smith VH, Staley JT. 2006. Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* 4:102–112. <https://doi.org/10.1038/nrmicro1341>
  36. Xu X, Thornton PE, Post WM. 2013. A global analysis of soil microbial biomass carbon, nitrogen and phosphorus in terrestrial ecosystems. *Global Ecol Biogeogr* 22:737–749. <https://doi.org/10.1111/geb.12029>
  37. Serna - Chavez HM, Fierer N, van Bodegom PM. 2013. Global drivers and patterns of microbial abundance in soil. *Global Ecol Biogeogr* 22:1162–1172. <https://doi.org/10.1111/geb.12070>
  38. Locey KJ, Lennon JT. 2016. Scaling laws predict global microbial diversity. *Proc Natl Acad Sci U S A* 113:5970–5975. <https://doi.org/10.1073/pnas.1521291113>
  39. Geyer KM, Kyker-Snowman E, Grandy AS, Frey SD. 2016. Microbial carbon use efficiency: accounting for population, community, and ecosystem-scale controls over the fate of metabolized organic matter. *Biogeochemistry* 127:173–188. <https://doi.org/10.1007/s10533-016-0191-y>
  40. Rousk J, Bååth E, Brookes PC, Lauber CL, Lozupone C, Caporaso JG, Knight R, Fierer N. 2010. Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME J* 4:1340–1351. <https://doi.org/10.1038/ismej.2010.58>
  41. Larkin AA, Martiny AC. 2017. Microdiversity shapes the traits, niche space, and biogeography of microbial taxa. *Environ Microbiol Rep* 9:55–70. <https://doi.org/10.1111/1758-2229.12523>
  42. Dequiedt S, Thioulouse J, Jolivet C, Saby NPA, Lelievre M, Maron P-A, Martin MP, Prévost-Bouré NC, Toutain B, Arrouays D, Lemanceau P, Ranjard L. 2009. Biogeographical patterns of soil bacterial communities. *Environ Microbiol Rep* 1:251–255. <https://doi.org/10.1111/j.1758-2229.2009.00040.x>
  43. Fuhrman JA, Steele JA, Hewson I, Schwalbach MS, Brown MV, Green JL, Brown JH. 2008. A latitudinal diversity gradient in planktonic marine bacteria. *Proc Natl Acad Sci U S A* 105:7774–7778. <https://doi.org/10.1073/pnas.0803070105>
  44. Nelson MB, Martiny AC, Martiny JBH. 2016. Global biogeography of microbial nitrogen-cycling traits in soil. *Proc Natl Acad Sci U S A* 113:8033–8040. <https://doi.org/10.1073/pnas.1601070113>
  45. Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-González A, Eldridge DJ, Bardgett RD, Maestre FT, Singh BK, Fierer N. 2018. A global atlas of the dominant bacteria found in soil. *Science* 359:320–325. <https://doi.org/10.1126/science.aap9516>
  46. Fierer N, Jackson RB. 2006. The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci U S A* 103:626–631. <https://doi.org/10.1073/pnas.0507535103>
  47. Bell T, Ager D, Song J-I, Newman JA, Thompson IP, Lilley AK, van der Gast CJ. 2005. Larger islands house more bacterial taxa. *Science* 308:1884. <https://doi.org/10.1126/science.1111318>
  48. Zhou J, Kang S, Schadt CW, Garten CT. 2008. Spatial scaling of functional gene diversity across various microbial taxa. *Proc Natl Acad Sci U S A* 105:7768–7773. <https://doi.org/10.1073/pnas.0709016105>
  49. Ranjard L, Dequiedt S, Chemidlin Prévost-Bouré N, Thioulouse J, Saby NPA, Lelievre M, Maron PA, Morin FER, Bispo A, Jolivet C, Arrouays D, Lemanceau P. 2013. Turnover of soil bacterial diversity driven by wide-scale environmental heterogeneity. *Nat Commun* 4:1434. <https://doi.org/10.1038/ncomms2431>
  50. Wang X-B, Lü X-T, Yao J, Wang Z-W, Deng Y, Cheng W-X, Zhou J-Z, Han X-G. 2017. Habitat-specific patterns and drivers of bacterial  $\beta$ -diversity in China's drylands. *ISME J* 11:1345–1358. <https://doi.org/10.1038/ismej.2017.11>
  51. Bell T. 2010. Experimental tests of the bacterial distance-decay relationship. *ISME J* 4:1357–1365. <https://doi.org/10.1038/ismej.2010.77>
  52. Tedersoo L, Bahram M, Pölme S, Kõljalg U, Yorou NS, Wijesundera R, Villarreal Ruiz L, Vasco-Palacios AM, Thu PQ, Suija A, et al. 2014. Fungal biogeography. Global diversity and geography of soil fungi. *Science* 346:1256688. <https://doi.org/10.1126/science.1256688>
  53. Chau JF, Bagtzoglou AC, Willig MR. 2011. The effect of soil texture on richness and diversity of bacterial communities. *Environ Forensics* 12:333–341. <https://doi.org/10.1080/15275922.2011.622348>
  54. Kaiser K, Wemheuer B, Korolkov V, Wemheuer F, Nacke H, Schöning I, Schrupf M, Daniel R. 2016. Driving forces of soil bacterial community structure, diversity, and function in temperate grasslands and forests. *Sci Rep* 6:33696. <https://doi.org/10.1038/srep33696>
  55. Martiny JBH, Eisen JA, Penn K, Allison SD, Horner-Devine MC. 2011. Drivers of bacterial  $\beta$ -diversity depend on spatial scale. *Proc Natl Acad Sci U S A* 108:7850–7854. <https://doi.org/10.1073/pnas.1016308108>
  56. Kerr B, Riley MA, Feldman MW, Bohannan BJM. 2002. Local dispersal promotes biodiversity in a real-life game of rock-paper-scissors. *Nature* 418:171–174. <https://doi.org/10.1038/nature00823>
  57. Manzoni S, Schimel JP, Porporato A. 2012. Responses of soil microbial communities to water stress: results from a meta-analysis. *Ecology* 93:930–938. <https://doi.org/10.1890/11-0026.1>
  58. Lozupone CA, Knight R. 2007. Global patterns in bacterial diversity. *Proc Natl Acad Sci U S A* 104:11436–11440. <https://doi.org/10.1073/pnas.0611525104>
  59. Philippot L, Raaijmakers JM, Lemanceau P, van der Putten WH. 2013. Going back to the roots: the microbial ecology of the rhizosphere. *Nat Rev Microbiol* 11:789–799. <https://doi.org/10.1038/nrmicro3109>
  60. Shade A, Caporaso JG, Handelsman J, Knight R, Fierer N. 2013. A meta-analysis of changes in bacterial and archaeal communities with time. *ISME J* 7:1493–1506. <https://doi.org/10.1038/ismej.2013.54>
  61. Graham EB, Knelman JE, Schindlbacher A, Siciliano S, Breulmann M, Yannarell A, Beman JM, Abell G, Philippot L, Prosser J, et al. 2016. Microbes as engines of ecosystem function: when does community structure enhance predictions of ecosystem processes? *Front Microbiol* 7:214. <https://doi.org/10.3389/fmicb.2016.00214>
  62. Malik AA, Martiny JBH, Brodie EL, Martiny AC, Treseder KK, Allison SD. 2020. Defining trait-based microbial strategies with consequences for soil carbon cycling under climate change. *ISME J* 14:1–9. <https://doi.org/10.1038/s41396-019-0510-0>
  63. van der Heijden MGA, Bardgett RD, van Straalen NM. 2008. The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. *Ecol Lett* 11:296–310. <https://doi.org/10.1111/j.1461-0248.2007.01139.x>
  64. Strickland MS, Lauber C, Fierer N, Bradford MA. 2009. Testing the functional significance of microbial community composition. *Ecology* 90:441–451. <https://doi.org/10.1890/08-0296.1>
  65. Schimel JP, Schaeffer SM. 2012. Microbial control over carbon cycling in soil. *Front Microbiol* 3:348. <https://doi.org/10.3389/fmicb.2012.00348>
  66. Green JL, Bohannan BJM, Whitaker RJ. 2008. Microbial biogeography: from taxonomy to traits. *Science* 320:1039–1043. <https://doi.org/10.1126/science.1153475>

67. Caldwell BA. 2005. Enzyme activities as a component of soil biodiversity: a review. *Pedobiologia* 49:637–644. <https://doi.org/10.1016/j.pedobi.2005.06.003>
68. Six J, Frey SD, Thiet RK, Batten KM. 2006. Bacterial and fungal contributions to carbon sequestration in agroecosystems. *Soil Sci Soc Am J* 70:555–569. <https://doi.org/10.2136/sssaj2004.0347>
69. Naeem S, Wright JP. 2003. Disentangling biodiversity effects on ecosystem functioning: deriving solutions to a seemingly insurmountable problem. *Ecol Lett* 6:567–579. <https://doi.org/10.1046/j.1461-0248.2003.00471.x>
70. Cardinale BJ, Duffy JE, Gonzalez A, Hooper DU, Perrings C, Venail P, Narwani A, Mace GM, Tilman D, Wardle DA, Kinzig AP, Daily GC, Loreau M, Grace JB, Larigauderie A, Srivastava DS, Naeem S. 2012. Biodiversity loss and its impact on humanity. *Nature* 486:59–67. <https://doi.org/10.1038/nature11148>
71. Ladau J, Shi Y, Jing X, He J-S, Chen L, Lin X, Fierer N, Gilbert JA, Pollard KS, Chu H. 2018. Existing climate change will lead to pronounced shifts in the diversity of soil prokaryotes. *mSystems* 3:e00167-18. <https://doi.org/10.1128/mSystems.00167-18>
72. Banerjee S, Schlaeppi K, van der Heijden MGA. 2018. Keystone taxa as drivers of microbiome structure and functioning. *Nat Rev Microbiol* 16:567–576. <https://doi.org/10.1038/s41579-018-0024-1>
73. Banerjee S, Kirkby CA, Schmutter D, Bissett A, Kirkegaard JA, Richardson AE. 2016. Network analysis reveals functional redundancy and keystone taxa amongst bacterial and fungal communities during organic matter decomposition in an arable soil. *Soil Biol Biochem* 97:188–198. <https://doi.org/10.1016/j.soilbio.2016.03.017>
74. Neu AT, Allen EE, Roy K. 2021. Defining and quantifying the core microbiome: Challenges and prospects. *Proc Natl Acad Sci U S A* 118:e2104429118. <https://doi.org/10.1073/pnas.2104429118>
75. Bay SK, McGeoch MA, Gillor O, Wieler N, Palmer DJ, Baker DJ, Chown SL, Greening C. 2020. Soil bacterial communities exhibit strong biogeographic patterns at fine taxonomic resolution. *mSystems* 5:e00540-20. <https://doi.org/10.1128/mSystems.00540-20>
76. de Wit R, Bouvier T. 2006. “Everything is everywhere, but, the environment selects”; what did Baas Becking and Beijerinck really say? *Environ Microbiol* 8:755–758. <https://doi.org/10.1111/j.1462-2920.2006.01017.x>
77. Bowman MM, Heath AE, Varga T, Battu AK, Chu RK, Toyoda J, Cheeke TE, Porter SS, Moffett KB, LeTendre B, Qafoku O, Bargar JR, Mans DM, Hess NJ, Graham EB. 2023. One thousand soils for molecular understanding of belowground carbon cycling. *Front Soil Sci* 3. <https://doi.org/10.3389/fsoil.2023.1120425>
78. Keller M, Schimel DS, Hargrove WW, Hoffman FM. 2008. A continental strategy for the national ecological observatory network. *Front Ecol Environ* 6:282–284. [https://doi.org/10.1890/1540-9295\(2008\)6\[282:ACSFN\]2.0.CO;2](https://doi.org/10.1890/1540-9295(2008)6[282:ACSFN]2.0.CO;2)
79. Olson DM, Dinerstein E, Wikramanayake ED, Burgess ND, Powell GVN, Underwood EC, D’amico JA, Itoua I, Strand HE, Morrison JC, Loucks CJ, Allnutt TF, Ricketts TH, Kura Y, Lamoreux JF, Wettengel WW, Hedao P, Kassem KR. 2001. Terrestrial ecoregions of the world: a new map of life on earth. *BioScience* 51:933. [https://doi.org/10.1641/0006-3568\(2001\)051\[0933:TEOTWA\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2)
80. Hengl T, Mendes de Jesus J, Heuvelink GBM, Ruiperez Gonzalez M, Kilbarda M, Blagotić A, Shangguan W, Wright MN, Geng X, Bauer-Marschallinger B, Guevara MA, Vargas R, MacMillan RA, Batjes NH, Leenaars JGB, Ribeiro E, Wheeler I, Mantel S, Kempen B. 2017. SoilGrids250m: global gridded soil information based on machine learning. *PLoS One* 12:e0169748. <https://doi.org/10.1371/journal.pone.0169748>
81. Aoki-Kinoshita KF, Kanehisa M. 2007. Gene annotation and pathway mapping in KEGG. *Methods Mol Biol* 396:71–91. [https://doi.org/10.1007/978-1-59745-515-2\\_6](https://doi.org/10.1007/978-1-59745-515-2_6)
82. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32:D277–D280. <https://doi.org/10.1093/nar/gkh063>
83. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. *Nucleic Acids Res* 47:D427–D432. <https://doi.org/10.1093/nar/gky995>
84. Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E. 2013. Tigrfams and genome properties in 2013. *Nucleic Acids Res* 41:D387–D395. <https://doi.org/10.1093/nar/gks1234>
85. Huntemann M, Ivanova NN, Mavromatis K, Tripp HJ, Paez-Espino D, Palaniappan K, Szeto E, Pillay M, Chen I-MA, Pati A, Nielsen T, Markowitz VM, Kyrpides NC. 2015. The standard operating procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4). *Stand Genomic Sci* 10:86. <https://doi.org/10.1186/s40793-015-0077-y>
86. Freitas TAK, Li P-E, Scholz MB, Chain PSG. 2015. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic Acids Res* 43:e69. <https://doi.org/10.1093/nar/gkv180>
87. Taylor BL, Zhulin IB. 1999. PAS domains: internal sensors of oxygen, redox potential, and light. *Microbiol Mol Biol Rev* 63:479–506. <https://doi.org/10.1128/MMBR.63.2.479-506.1999>
88. Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert JA, Wall DH, Caporaso JG. 2012. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc Natl Acad Sci U S A* 109:21390–21395. <https://doi.org/10.1073/pnas.1215210110>
89. Freschet GT, Valverde - Barrantes OJ, Tucker CM, Craine JM, McCormack ML, Violle C, Fort F, Blackwood CB, Urban - Mead KR, Iversen GB, et al. 2017. Climate, soil and plant functional types as drivers of global fine-root trait variation. *J Ecol* 105:1182–1196. <https://doi.org/10.1111/1365-2745.12769>
90. Zheng Q, Hu Y, Zhang S, Noll L, Böckle T, Dietrich M, Herbold CW, Eichorst SA, Woebken D, Richter A, Wanek W. 2019. Soil multifunctionality is affected by the soil environment and by microbial community composition and diversity. *Soil Biol Biochem* 136:107521. <https://doi.org/10.1016/j.soilbio.2019.107521>
91. Espeleta JF, Cardon ZG, Mayer KU, Neumann RB. 2017. Diel plant water use and competitive soil cation exchange interact to enhance NH<sub>4</sub><sup>+</sup> and K<sup>+</sup> availability in the rhizosphere. *Plant Soil* 414:33–51. <https://doi.org/10.1007/s11104-016-3089-5>
92. Alster CJ, von Fischer JC, Allison SD, Treseder KK. 2020. Embracing a new paradigm for temperature sensitivity of soil microbes. *Glob Chang Biol* 26:3221–3229. <https://doi.org/10.1111/gcb.15053>
93. Wang C, Morrissey EM, Mau RL, Hayer M, Piñeiro J, Mack MC, Marks JC, Bell SL, Miller SN, Schwartz E, Dijkstra P, Koch BJ, Stone BW, Purcell AM, Blazewicz SJ, Hofmockel KS, Pett-Ridge J, Hungate BA. 2021. The temperature sensitivity of soil: microbial biodiversity, growth, and carbon mineralization. *ISME J* 15:2738–2747. <https://doi.org/10.1038/s41396-021-00959-1>
94. Hall EK, Bernhardt ES, Bier RL, Bradford MA, Boot CM, Cotner JB, Del Giorgio PA, Evans SE, Graham EB, Jones SE, Lennon JT, Locey KJ, Nemergut D, Osborne BB, Rocca JD, Schimel JP, Waldrop MP, Wallenstein MD. 2018. Understanding how microbiomes influence the systems they inhabit. *Nat Microbiol* 3:977–982. <https://doi.org/10.1038/s41564-018-0201-z>
95. Graham EB, Crump AR, Resch CT, Fansler S, Arntzen E, Kennedy DW, Fredrickson JK, Stegen JC. 2016. Coupling spatiotemporal community assembly processes to changes in microbial metabolism. *Front Microbiol* 7:1949. <https://doi.org/10.3389/fmicb.2016.01949>
96. Liu W, Graham EB, Dong Y, Zhong L, Zhang J, Qiu C, Chen R, Lin X, Feng Y. 2021. Balanced stochastic versus deterministic assembly processes benefit diverse yet uneven ecosystem functions in representative agroecosystems. *Environ Microbiol* 23:391–404. <https://doi.org/10.1111/1462-2920.15326>
97. Liu W, Graham EB, Zhong L, Zhang J, Li W, Li Z, Lin X, Feng Y. 2020. Dynamic microbial assembly processes correspond to soil fertility in sustainable paddy agroecosystems. *Funct Ecol* 34:1244–1256. <https://doi.org/10.1111/1365-2435.13550>
98. Waring BG, Sulman BN, Reed S, Smith AP, Averill C, Creamer CA, Cusack DF, Hall SJ, Jastrow JD, Jilling A, Kemner KM, Kleber M, Liu X-J, Pett-Ridge J, Schulz M. 2020. From pools to flow: the PROMISE framework for new insights on soil carbon cycling in a changing world. *Glob Chang Biol* 26:6631–6643. <https://doi.org/10.1111/gcb.15365>
99. McCarter CPR, Rezanezhad F, Quinton WL, Gharedaghloo B, Lennartz B, Price J, Connon R, Van Cappellen P. 2020. Pore-scale controls on hydrological and geochemical processes in peat: Implications on

- interacting processes. *Earth Sci Rev* 207:103227. <https://doi.org/10.1016/j.earscirev.2020.103227>
100. Ruamps LS, Nunan N, Chenu C. 2011. Microbial biogeography at the soil pore scale. *Soil Biol Biochem* 43:280–286. <https://doi.org/10.1016/j.soilbio.2010.10.010>
  101. Mares MA. 2017. Encyclopedia of deserts
  102. Hagin J, Hadas A. 1962. Solubility of calcium phosphate in calcareous soils. *Nature* 193:1211–1212. <https://doi.org/10.1038/1931211a0>
  103. Dick WA, Cheng L, Wang P. 2000. Soil acid and alkaline phosphatase activity as pH adjustment indicators. *Soil Biol Biochem* 32:1915–1919. [https://doi.org/10.1016/S0038-0717\(00\)00166-8](https://doi.org/10.1016/S0038-0717(00)00166-8)
  104. Krämer S. 2000. Acid and alkaline phosphatase dynamics and their relationship to soil microclimate in a semiarid woodland. *Soil Biol Biochem* 32:179–188. [https://doi.org/10.1016/S0038-0717\(99\)00140-6](https://doi.org/10.1016/S0038-0717(99)00140-6)
  105. Oliverio AM, Bissett A, McGuire K, Saltonstall K, Turner BL, Fierer N. 2020. The role of phosphorus limitation in shaping soil bacterial communities and their metabolic capabilities. *mBio* 11:e01718-20. <https://doi.org/10.1128/mBio.01718-20>
  106. Laliberté E, Zemunik G, Turner BL. 2014. Environmental filtering explains variation in plant diversity along resource gradients. *Science* 345:1602–1605. <https://doi.org/10.1126/science.1256330>
  107. Baya AM, Boethling RS, Ramos-Cormenzana A. 1981. Vitamin production in relation to phosphate solubilization by soil bacteria. *Soil Biol Biochem* 13:527–531. [https://doi.org/10.1016/0038-0717\(81\)90044-4](https://doi.org/10.1016/0038-0717(81)90044-4)
  108. Pii Y, Mimmo T, Tomasi N, Terzano R, Cesco S, Crecchio C. 2015. Microbial interactions in the rhizosphere: beneficial influences of plant growth-promoting rhizobacteria on nutrient acquisition process. A review. *Biol Fertil Soils* 51:403–415. <https://doi.org/10.1007/s00374-015-0996-1>
  109. Guerini D. 1998. The Ca<sup>2+</sup> pumps and the Na<sup>+</sup>/Ca<sup>2+</sup> exchangers. *Biometals* 11:319–330. <https://doi.org/10.1023/a:1009210001608>
  110. Cabello P, Luque-Almagro VM, Roldán MD, Moreno-Vivián C. 2019. Nitrogen cycle reference module in life sciences. Elsevier.
  111. Pal RR, Khardanavi AA, Purohit HJ. 2015. Identification and monitoring of nitrification and denitrification genes in *Klebsiella pneumoniae* EGD-HP19-C for its ability to perform heterotrophic nitrification and aerobic denitrification. *Funct Integr Genomics* 15:63–76. <https://doi.org/10.1007/s10142-014-0406-z>
  112. Padhi SK, Tripathy S, Sen R, Mahapatra AS, Mohanty S, Maiti NK. 2013. Characterisation of heterotrophic nitrifying and aerobic denitrifying *Klebsiella pneumoniae* CF-59 strain for bioremediation of wastewater. *Int Biodeterior Biodegrad* 78:67–73. <https://doi.org/10.1016/j.ibiod.2013.01.001>
  113. Hastrup ACS, Green F, Lebow PK, Jensen B. 2012. Enzymatic oxalic acid regulation correlated with wood degradation in four brown-rot fungi. *Int Biodeterior Biodegrad* 75:109–114. <https://doi.org/10.1016/j.ibiod.2012.05.030>
  114. Mäkelä M, Galkin S, Hatakka A, Lundell T. 2002. Production of organic acids and oxalate decarboxylase in lignin-degrading white rot fungi. *Enzyme Microb Technol* 30:542–549. [https://doi.org/10.1016/S0141-0229\(02\)00012-1](https://doi.org/10.1016/S0141-0229(02)00012-1)
  115. Takao S. 1965. Organic acid production by basidiomycetes: I. screening of acid-producing strains. *Appl Microbiol* 13:732–737. <https://doi.org/10.1128/am.13.5.732-737.1965>
  116. Baldrian P, Valášková V. 2008. Degradation of cellulose by basidiomycetous fungi. *FEMS Microbiol Rev* 32:501–521. <https://doi.org/10.1111/j.1574-6976.2008.00106.x>
  117. Lauber CL, Hamady M, Knight R, Fierer N. 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl Environ Microbiol* 75:5111–5120. <https://doi.org/10.1128/AEM.00335-09>
  118. Carson JK, Gonzalez-Quiñones V, Murphy DV, Hinz C, Shaw JA, Gleason DB. 2010. Low pore connectivity increases bacterial diversity in soil. *Appl Environ Microbiol* 76:3936–3942. <https://doi.org/10.1128/AEM.03085-09>
  119. He L, Xu X. 2021. Mapping soil microbial residence time at the global scale. *Glob Chang Biol* 27:6484–6497. <https://doi.org/10.1111/gcb.15864>
  120. Bueno de Mesquita CP, Wu D, Tringe SG. 2023. Methyl-based methanogenesis: an ecological and genomic review. *Microbiol Mol Biol Rev* 87:e0002422. <https://doi.org/10.1128/mmr.00024-22>
  121. Machado A, Cerca N. 2015. Influence of biofilm formation by *Gardnerella vaginalis* and other anaerobes on bacterial vaginosis. *J Infect Dis* 212:1856–1861. <https://doi.org/10.1093/infdis/jiv338>
  122. Grosche A, Sekaran H, Pérez-Rodríguez I, Starovoytov V, Vetriani C. 2015. *Cetia pacifica* gen. nov., sp. nov., a chemolithoautotrophic, thermophilic, nitrate-ammonifying bacterium from a deep-sea hydrothermal vent. *Int J Syst Evol Microbiol* 65:1144–1150. <https://doi.org/10.1099/ijs.0.000070>
  123. Shelobolina ES, Sullivan SA, O'Neill KR, Nevin KP, Lovley DR. 2004. Isolation, characterization, and U(VI)-reducing potential of a facultatively anaerobic, acid-resistant bacterium from Low-pH, nitrate- and U(VI)-contaminated subsurface sediment and description of *Salmonella subterranea* sp. nov. *Appl Environ Microbiol* 70:2959–2965. <https://doi.org/10.1128/AEM.70.5.2959-2965.2004>
  124. Nichols CM, Bowman JP, Guezennec J. 2005. *Olleya marilimos* gen. nov., sp. nov., an exopolysaccharide-producing marine bacterium from the family *Flavobacteriaceae*, isolated from the Southern Ocean. *Int J Syst Evol Microbiol* 55:1557–1561. <https://doi.org/10.1099/ijs.0.63642-0>
  125. Robinson IM, Allison MJ, Hartman PA. 1975. *Anaeroplasm* *abactoclasticum* gen. nov., sp. nov.: an obligately anaerobic mycoplasma from the rumen. *Int J Syst Bacteriol* 25:173–181. <https://doi.org/10.1099/00207713-25-2-173>
  126. Stewart V. 1993. Nitrate regulation of anaerobic respiratory gene expression in *Escherichia coli*. *Mol Microbiol* 9:425–434. <https://doi.org/10.1111/j.1365-2958.1993.tb01704.x>
  127. Hung C-H, Chang Y-T, Chang Y-J. 2011. Roles of microorganisms other than *Clostridium* and *Enterobacter* in anaerobic fermentative biohydrogen production systems—a review. *Bioresour Technol* 102:8437–8444. <https://doi.org/10.1016/j.biortech.2011.02.084>
  128. Chung EJ, Park TS, Jeon CO, Chung YR. 2012. *Chitinophaga oryziterrae* sp. nov., isolated from the rhizosphere soil of rice (*Oryza sativa* L.). *Int J Syst Evol Microbiol* 62:3030–3035. <https://doi.org/10.1099/ijs.0.036442-0>
  129. Jin D, Kong X, Wang J, Sun J, Yu X, Zhuang X, Deng Y, Bai Z. 2018. *Chitinophaga caeni* sp. nov., isolated from activated sludge. *Int J Syst Evol Microbiol* 68:2209–2213. <https://doi.org/10.1099/ijsem.0.002811>
  130. Voigt C, Lamprecht RE, Marushchak ME, Lind SE, Novakovskiy A, Aurela M, Martikainen PJ, Biasi C. 2017. Warming of subarctic tundra increases emissions of all three important greenhouse gases - carbon dioxide, methane, and nitrous oxide. *Glob Chang Biol* 23:3121–3138. <https://doi.org/10.1111/gcb.13563>
  131. Xue K, M. Yuan M, J. Shi Z, Qin Y, Deng Y, Cheng L, Wu L, He Z, Van Nostrand JD, Bracho R, Natali S, Luo C, Konstantinidis KT, Wang Q, Cole JR, Tiedje JM, Luo Y, Zhou J. 2016. Tundra soil carbon is vulnerable to rapid microbial decomposition under climate warming. *Nat Clim Change* 6:595–600. <https://doi.org/10.1038/nclimate2940>
  132. Rößger N, Sachs T, Wille C, Boike J, Kutzbach L. 2022. Seasonal increase of methane emissions linked to warming in Siberian tundra. *Nat Clim Chang* 12:1031–1036. <https://doi.org/10.1038/s41558-022-01512-4>
  133. Zona D, Gioli B, Commane R, Lindaas J, Wofsy SC, Miller CE, Dinardo SJ, Dengel S, Sweeney C, Karion A, Chang RY-W, Henderson JM, Murphy PC, Goodrich JP, Moreaux V, Liljedahl A, Watts JD, Kimball JS, Lipson DA, Oechel WC. 2016. Cold season emissions dominate the Arctic tundra methane budget. *Proc Natl Acad Sci U S A* 113:40–45. <https://doi.org/10.1073/pnas.1516017113>
  134. Christensen TR. 1993. Methane emission from Arctic tundra. *Biogeochemistry* 21:117–139. <https://doi.org/10.1007/BF00000874>
  135. Christensen TR, Cox P. 1995. Response of methane emission from Arctic tundra to climatic change: results from a model simulation. *Tellus B Chem Phys Meteorol* 47:301. <https://doi.org/10.3402/tellusb.v47i3.16049>
  136. Yang G, Peng Y, Marushchak ME, Chen Y, Wang G, Li F, Zhang D, Wang J, Yu J, Liu L, Qin S, Kou D, Yang Y. 2018. Magnitude and pathways of increased nitrous oxide emissions from uplands following Permafrost thaw. *Environ Sci Technol* 52:9162–9169. <https://doi.org/10.1021/acs.est.8b02271>
  137. Marushchak ME, Kerttula J, Diáková K, Faguet A, Gil J, Grosse G, Knoblauch C, Lashchinskiy N, Martikainen PJ, Morgenstern A, Nykamb M, Ronkainen JG, Siljanen HMP, van Delden L, Voigt C, Zimov N, Zimov S, Biasi C. 2021. Thawing Yedoma permafrost is a neglected nitrous oxide source. *Nat Commun* 12:7107. <https://doi.org/10.1038/s41467-021-27386-2>

138. Bhattarai HR, Marushchak ME, Ronkainen J, Lamprecht RE, Siljanen HMP, Martikainen PJ, Biasi C, Maljanen M. 2022. Emissions of atmospherically reactive gases nitrous acid and nitric oxide from Arctic permafrost peatlands. *Environ Res Lett* 17:024034. <https://doi.org/10.1088/1748-9326/ac4f8e>
139. Varsadiya M, Liebmann P, Petters S, Hugelius G, Ulrich T, Guggenberger G, Bárta J. 2022. Extracellular enzyme ratios reveal locality and horizon-specific carbon, nitrogen, and phosphorus limitations in Arctic permafrost soils. *Biogeochemistry* 161:101–117. <https://doi.org/10.1007/s10533-022-00967-z>
140. Zhang D, Wang L, Qin S, Kou D, Wang S, Zheng Z, Peñuelas J, Yang Y. 2023. Microbial nitrogen and phosphorus co-limitation across permafrost region. *Glob Chang Biol* 29:3910–3923. <https://doi.org/10.1111/gcb.16743>
141. Yang G, Peng Y, Abbott BW, Biasi C, Wei B, Zhang D, Wang J, Yu J, Li F, Wang G, Kou D, Liu F, Yang Y. 2021. Phosphorus rather than nitrogen regulates ecosystem carbon dynamics after permafrost thaw. *Glob Chang Biol* 27:5818–5830. <https://doi.org/10.1111/gcb.15845>
142. Emerson JB, Roux S, Brum JR, Bolduc B, Woodcroft BJ, Jang HB, Singleton CM, Solden LM, Naas AE, Boyd JA, Hodgkins SB, Wilson RM, Trubl G, Li C, Frolking S, Pope PB, Wrighton KC, Crill PM, Chanton JP, Saleska SR, Tyson GW, Rich VI, Sullivan MB. 2018. Host-linked soil viral ecology along a permafrost thaw gradient. *Nat Microbiol* 3:870–880. <https://doi.org/10.1038/s41564-018-0190-y>
143. Wu R, Bottos EM, Danna VG, Stegen JC, Jansson JK, Davison MR. 2022. RNA viruses linked to eukaryotic hosts in thawed permafrost. *mSystems* 7:e0058222. <https://doi.org/10.1128/mSystems.00582-22>
144. Wu R, Trubl G, Taş N, Jansson JK. 2022. Permafrost as a potential pathogen reservoir. *One Earth* 5:351–360. <https://doi.org/10.1016/j.oneear.2022.03.010>
145. Rigou S, Santini S, Abergel C, Claverie J-M, Legendre M. 2022. Past and present giant viruses diversity explored through permafrost metagenomics. *Nat Commun* 13:5853. <https://doi.org/10.1038/s41467-022-33633-x>
146. Lee S, Sieradzki ET, Nicolas AM, Walker RL, Firestone MK, Hazard C, Nicol GW. 2021. Methane-derived carbon flows into host-virus networks at different trophic levels in soil. *Proc Natl Acad Sci U S A* 118:e2105124118. <https://doi.org/10.1073/pnas.2105124118>
147. Nickson R, McArthur J, Burgess W, Ahmed KM, Ravenscroft P, Rahman M. 1998. Arsenic poisoning of Bangladesh groundwater. *Nature* 395:338–338. <https://doi.org/10.1038/26387>
148. Gransee A, Führs H. 2013. Magnesium mobility in soils as a challenge for soil and plant analysis, magnesium fertilization and root uptake under adverse growth conditions. *Plant Soil* 368:5–21. <https://doi.org/10.1007/s11104-012-1567-y>
149. Knelman JE, Schmidt SK, Graham EB. 2021. Cyanobacteria in early soil development of deglaciated forefields: dominance of non-heterocytous filamentous cyanobacteria and phosphorus limitation of N-fixing nostocales. *Soil Biol Biochem* 154:108127. <https://doi.org/10.1016/j.soilbio.2020.108127>
150. Knelman JE, Nemergut DR. 2014. Changes in community assembly may shift the relationship between biodiversity and ecosystem function. *Front. Microbiol* 5:424. <https://doi.org/10.3389/fmicb.2014.00424>
151. Rocca JD, Hall EK, Lennon JT, Evans SE, Waldrop MP, Cotner JB, Nemergut DR, Graham EB, Wallenstein MD. 2015. Relationships between protein-encoding gene abundance and corresponding process are commonly assumed yet rarely observed. *ISME J* 9:1693–1699. <https://doi.org/10.1038/ismej.2014.252>
152. Storch D, Sizing AL. 2008. The concept of taxon invariance in ecology: do diversity patterns vary with changes in taxonomic resolution? *Folia Geobot* 43:329–344. <https://doi.org/10.1007/s12224-008-9015-8>
153. Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JBH. 2012. Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat Rev Microbiol* 10:497–506. <https://doi.org/10.1038/nrmicro2795>
154. Martiny JBH, Jones SE, Lennon JT, Martiny AC. 2015. Microbiomes in light of traits: a phylogenetic perspective. *Science* 350:aac9323. <https://doi.org/10.1126/science.aac9323>
155. Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 11:R25. <https://doi.org/10.1186/gb-2010-11-3-r25>
156. Terrestrial ecoregions of the world. WWF
157. R Core Team. 2021. R: a language and environment for statistical computing. Vienna, Austria Computer software, R Foundation for Statistical Computing
158. Kuhn M. 2008. Building predictive models in R Using the caret package. *J Stat Softw* 28. <https://doi.org/10.18637/jss.v028.i05>
159. Zeileis A, Leisch F, Hornik K, Kleiber C. 2002. strucchange: an r package for testing for structural change in linear regression models. *J Stat Softw* 7. <https://doi.org/10.18637/jss.v007.i02>
160. Oksanen J, GavinLS, F. Guillaume B, Roeland K, Pierre L, PeterRM, R.B.O, Peter S, M. HenryHS, EduardS, Helene W, Matt B, Michael B, BenB, Daniel B, Gustavo C, Michael C, MiquelDC, Sebastien D, HeloisABAE, RichF, Michael F, BrendanF, Geoffrey H, Mark OH, et al. 2013. Vegan: community ecology package. R package version 2.6-4. <https://CRAN.R-project.org/package=vegan>.
161. Wood S, Wood MS. 2015. "Package "Mgcv"" R package version 1.8-42
162. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. 2022. Cluster: cluster analysis basics and extensions. R package version 2.1.4
163. Kodinariya TM, Makwana PR. 2013. Review on determining number of cluster in K-means clustering. *Int J*
164. Wickham H, Chang W. 2016. Ggplot2: elegant graphics for data analysis. Springer-Verlag New York. <https://doi.org/10.1007/978-3-319-24277-4>
165. Wickham H, François R, Henry L, Müller K, Vaughan D. 2023. Dplyr: a grammar of data manipulation. R package version 1.1.4. <https://github.com/tidyverse/dplyr>, <https://dplyr.tidyverse.org>.
166. Kassambara A, Mundt F. 2020. Factoextra: extractand visualize the results of multivariate data analyses. R Packageversion 1.0.7. <https://CRAN.R-project.org/package=factoextra>.
167. Frank EH. 2023. Hmisc: Harrell miscellaneous. R Packageversion 5.0-1. <https://CRAN.R-project.org/package=Hmisc>.
168. Zeileis A, Fisher JC, Hornik K, Ihaka R, McWhite CD, Murrell P, Stauffer R, Wilke CO. 2020. colorspace : A toolbox for manipulating and assessing colors and palettes. *J Stat Soft* 96. <https://doi.org/10.18637/jss.v096.i01>
169. Newirth E. 2022. Rcolorbrewer: colorbrewer palettes. Rpackage version 1.1-3. <https://CRAN.R-project.org/package=RColorBrewer>.
170. Auguie B. 2017. gridExtra: miscellaneous functions For"Grid" Graphics. R package version 2.3. <https://CRAN.R-project.org/package=gridExtra>.
171. Wickham H, Averick M, BryanJ, Chang W, McGowanLD, François R, Golemund G, HayesA, Henry L, Hester J, KuhnM, Pedersen TL, Miller E, BacheSM, MüllerK, Ooms J, RobinsonD, SeidelDP, Spinu V, TakahashiK, VaughanD, Wilke C, Woo K, Yutani H. 2017. Welcome Tothe Tidyverse." *Journal of open source software*. <https://doi.org/10.21105/joss.01686>
172. Becker RA, Wilks AR, Brownrigg R, Minka TP. 2022. Maps: draw geographical maps.R package version 3.4.1
173. phyloT: a phylogenetic tree generator. n.d. Available from: <https://phyloT.biobyte.de/>
174. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2016. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 8:28–36. <https://doi.org/10.1111/2041-210X.12628>
175. Wang L-G, Lam T-Y, Xu S, Dai Z, Zhou L, Feng T, Guo P, Dunn CW, Jones BR, Bradley T, Zhu H, Guan Y, Jiang Y, Yu G. 2020. Treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. *Mol Biol Evol* 37:599–603. <https://doi.org/10.1093/molbev/msz240>
176. Campitelli E. 2022. "Ggnewscale: multiple fill and colourscales in 'Ggplot2'. R package version 0.4.8". <https://CRAN.R-project.org/package=ggnewscale>.
177. Kolde R. 2019. Pheatmap: pretty heatmaps. R package version1.0.12. <https://CRAN.R-project.org/package=pheatmap>.