Genomic fingerprints of the world's soil ecosystems

Emily B. Graham[a,b#], Vanessa A. Garayburu-Caruso[a,c], Ruonan Wu[a], Jianqiu Zheng[a], and Gerrad Jones[d]

[a]Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA, USA

[b]School of Biological Sciences, Washington State University, Pullman, WA, USA

[c]School of the Environment, Washington State University, Pullman, WA, USA

[d]Department of Biological and Ecological Engineering, Oregon State University, Corvallis, OR, USA

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

Running head: Genomic fingerprints of the world's soil ecosystems

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 assayed 1512 manually-curated soil metagenomes using complementary annotation databases, read-based taxonomy, and machine learning to extract multidimensional genomic fingerprints of global soil microbiomes. We reveal novel biogeographical patterns of soil microbiomes across environmental factors and ecological biomes with high molecular resolution. Specifically, we demonstrate shifts in the potential for microbial nutrient acquisition across pH gradients; for stress, transport, and redox-based processes across changes in soil bulk density; and for 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 1512 global soils with advanced analytics to create detailed genetic profiles (fingerprints) of soil microbiomes. This reveals novel patterns in how these microorganisms are distributed across different soil environments. For instance, we discovered shifts in their potential to acquire nutrients in relation to soil acidity, as well as changes in stress responses and potential greenhouse gas emissions linked to soil density. 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 towards comprehending the intricate world of soil microorganisms and its role in the global climate.

Introduction.

Soil microbiomes catalyze some of the most biogeochemically important reactions on Earth, yet their complexity hampers our efforts to fully understand their function (1–6). Microbial destabilization of soil carbon stores, for instance, has positive feedbacks with the global climate (7–9); and microorganism-driven soil nutrient cycles sustain life at higher trophic levels (10–12). The molecular revolution, including new '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 predictive models (13–18).

While global patterns are well-established for most macroorganisms, the extreme diversity of the soil microbiome and the myriad ecological factors that act upon it complicate efforts to understand patterns in its structure and function (19–27). Microbial assembly processes, dormancy, and above-belowground and interspecies interactions are among many other factors that influence soil microbial biogeography (19, 21, 26, 28–33). Patterns in microbial distributions are almost always weaker than those observed in macroecological systems due to these factors (26, 34, 35). Still, soil microbiome structure and function have been associated with variables such as latitude, vegetation, climate, pH, texture, moisture, carbon, and nutrients 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

72 and functional potential) impacts ecosystem processes, such as biogeochemical cycling, through
73 differences in extracellular enzyme production, carbon use efficiency, symbiotic relationships
74 with plants, and other factors (14, 44, 61–68). In turn, changes to microbial distribution across
75 environmental gradients can elucidate how soil functions may shift in relation to environmental
76 change; particularly if novel ecosystems supplant existing soil environments (26, 69, 70). For
77 instance, Ladau et al. (71) used current and historical biogeographical patterns of
78 microorganisms to predict their distributions under future climate scenarios.

79 New technologies and analytical techniques allow for direct assessment of the genetic
80 potential of soils to mediate biogeochemistry (4, 18). Yet modern technologies based on genomic
81 sequencing often generate a tremendous amount of data. The detection of genes that are nearly
82 ubiquitous (e.g., central metabolisms), poorly annotated, and/or of little relevance to soil
83 processes in particular can obscure biogeographical patterns in the soil microbiome. Many
84 researchers have attempted to distill this information into tractable units using trait-based
85 approaches, keystone species analyses, and/or core microbiome assessments, with mixed results
86 for deciphering broad patterns in microbial processes (27, 44, 62, 72–74). Other common
87 challenges to understanding microbial biogeography include the detection of rare organisms and
88 assigning thresholds for microbial taxonomy, both of which can influence patterns observed in
89 soil microbiomes (34, 75).

90 As such, fundamental questions regarding global soil microbial distributions and their
91 drivers remain largely unanswered (74, 76). Advanced analytics, including machine learning-
92 based approaches, in combination with global sequencing resources are essential for the next
93 generation of research in microbial ecology (18, 61, 77, 78). Here, we examine a comprehensive
94 set of 1512 manually-curated soil metagenomic samples to decipher patterns in soil genomic

95 potential across commonly measured environmental gradients and ecological biomes. We assay
96 metagenomic sequences using an array of complementary annotation databases, read-based
97 taxonomic assignment, and machine learning algorithms to extract multidimensional genomic
98 fingerprints of global soil microbiomes. We reveal never reported biogeographical patterns in
99 soil microbiome composition and functional potential that provide a molecular basis for
100 predicting soil microbial responses to climate change.
101
102 **Results.**
103
104 *Microbial clades and genomic potential across all soils.*
105 Soil metagenomic sequences spanned most major biomes, continents, and environmental
106 conditions (Figure 1, Table S1 and Table S2). Bulk density, cation exchange capacity, nitrogen
107 (N), pH, soil organic carbon, and clay content ranged from 0.28-1.56 cg/cm^3, 7.7-68.7
108 mmol(c)/kg, 0.25-22.4 cg/kg, 4.4-8.3, 2.4-510.9 dg/kg, and 2.7-57.1 g/kg respectively.
109 The most abundant microbial attributes (i.e., gene annotations by Kyoto Encyclopedia of
110 Genes and Genomes (KEGG)(79, 80), the Pfam (81) and TIGRFAM protein databases (82), and
111 read-based taxonomy, see methods) are shown in Figure 2. Multiple annotation databases
112 provide complementary information and are standard operating procedure by the U.S.
113 Department of Energy Joint Genome Institute (JGI)(83, 84). Abundant KEGG orthology (KO)
114 were associated bacterial secretion systems (K12510), redox reactions (K00104, K00334,
115 K03882, K00380), and cellular growth processes (K02335, K02049, K00334, K03882, K03088).
116 Abundant Pfam comprised tolerance to common soil stressors including temperature (PF11999,
117 PF08496) and salinity (PF12838), as well as functions involved in microbial growth (PF08511,

118 PF07357, PF13247, PF00005). The most abundant TIGRFAM were also associated with soil
119 stressors; for example, TIGR00229 (PAS Domain S-box protein) is involved in sensing changes
120 in light, redox potential, oxygen, and other stressors (85). Other highly abundant TIGRFAM
121 across all soils included a sigma-factor that may be associated with microbial responses to
122 environmental fluctuation (TIGR02937) and other signaling and sensing processes (TIGR00254,
123 TIGR00231). Finally, abundant microbial genera include diverse microorganisms including
124 Burkholderia, Mycobacterium, Streptomyces, Acetobacter, and Pseudonocardia.
125
126 *Variation in genomic potential across environmental gradients.*
127 After assessing multiple machine learning algorithms, we used random forest regression
128 to extract microbial attributes associated with soil factors and ecological biomes (see methods).
129 Soil pH, bulk density, and biome type were the strongest correlates of microbial attributes (i.e,
130 they generated well-fitting models with comparatively high explanatory power relative to other
131 factors). The cross-validated random forest model for each of the variables above had a cross-
132 validated root mean square error (RMSE) of less than 0.8 (often < 0.2) and/or at least 60%
133 variance explained. Models for cation exchange capacity, nitrogen, soil organic carbon, and clay
134 content did not meet these thresholds. Results were conceptually consistent across different
135 microbial attributes. A full description of all genomic fingerprints is in Extended Data File 1.
136 pH separated soil microbiomes by nutrient and vitamin B-related genes (Figure 3). More
137 alkaline soils were characterized by vitamin B-, nutrient-, and phosphatase-related activities, for
138 instance with genomic fingerprints that included K01662, PF01872, K08717, TIGR00378,
139 PF01966, and PF00481; as well as many genuses of Alphaproteobacteria. More acidic soils were
140 signified by ammonia- and complex organic polymer- (e.g., wood-) related reactions, with

141 genomic fingerprints that included K14333, K01684, TIGR03404, and TIGR02093, and
142 microbial clades using ammonia as a substrate (Rhodopseudomonas and Klebsiella).
143 Bulk density separated soil microbiomes by factors associated with soil structure,
144 hydrology, and nutrient content. The genomic fingerprint of low bulk density soils comprised
145 genes involved in organic N and methane cycling, as well as signatures of anoxia (Figure 3). It
146 included K00992, K11261, K11959, TIGR03323, TIGR01392, and microorganisms associated
147 with anoxic and/or aquatic environments (Gardnerella, Cetia, Salmonella, Olleya, Mycoplasma,
148 Escherichia, Enterobacter, and Chitinophaga). Conversely, high bulk density soils were signified
149 by genes involved in microbial transport, stress, pigmentation, and infection mechanisms.
150 Notable attributes associated with high bulk density soils included K00854, K11733, K03799,
151 TIGR01263, TIGR01957, TIGR01203, PF00582, PF03631, and four Alphaproteobacteria.
152 Four biomes were highlighted with distinct genomic signatures by our models – flooded
153 grasslands & savannas, tundras, boreal forest, and montane grasslands/shrublands (Figure 4).
154 The soil genomic fingerprint of flooded grasslands/savannas were primarily characterized by
155 positive associations with organic N (K13497, K07395, K19702, K02042), phosphorus (P,
156 K02042, K00854, K02800, PF04273), stress tolerance (heat and arsenic: TIGR04282,
157 TIGR02348), motility (PF00669, PF00482, Paeniclostridium). The genomic fingerprint for
158 tundras had the greatest number positive gene associations, including those with methanogenesis
159 (K11261, TIGR03323, TIGR03321, TIGR03322, TIGR03306, TIGR01392, TIGR02145), nitric
160 oxide (PF08768), P transfer (K03525, K00997, K00992, K13497, PF08009), and pathogenesis
161 (PF11203, PF05045); supported by the presence of anaerobic (Gardnerella, Escherichia,
162 Kosakonia, Glaesserella, Mycoplasma, Treponema, Paeniclostridium) and possible plant-growth
163 promoting genuses (Kosakonia).

164 Boreal forest and montane grasslands/shrublands had fingerprints that primarily consisted
165 of negative associations. In boreal forests, this consisted of negative relationships with
166 magnesium (Mg, K03284) and P- (K06217) related genes and carbohydrate and organic N
167 metabolism (K02800, K11472, PF00208, PF04685); as well as some metal-related genes
168 (TIGR00378, TIGR04282, TIGR00003). Additionally, they were negatively associated with
169 anaerobic genera (Treponema, Enteractinococcus). Montane systems also were negatively
170 characterized by genes related to Mg and P (K01520, K04765, TIGR01203), carbohydrate
171 metabolism (K01816), virulence (PF09299), and motility (PF00669), as well as organic N
172 (K07395, K00600, PF00208) and S (K03644). They were also negatively associated with
173 anaerobic genera (Gardnerella, Thielavia, Coniochaeta, Escherichia, Salmonella, Kosakonia,
174 Shigella, Mycoplasma, Treponema, Glaesserella, Enteractinococcus, Paeniclostridium) .
175 Interestingly in all analyses across pH, bulk density, and biomes, we observed
176 phylogenetic coherence in soil microbiomes — the normalized abundances of more closely
177 related microorganisms tended to move congruently in response to biome type and
178 environmental factors (Figure 4B, Extended Data File 1).
179
180 *Unsupervised machine learning to disentangle environmental interactions associated with soil*
181 *genomic potential.*
182 Finally, we used an unsupervised approach to reveal a collection of soils with distinct
183 genomic fingerprints that were characterized by coordinated changes in soil organic carbon
184 content, total N, and cation exchange capacity and in bulk density and clay content (Figure 5,
185 Extended Data 2). We identified KO cluster 7, genus cluster 6, and Pfam cluster 1 as containing

186 soil samples that exhibited the strongest coordinated changes (defined by number of significant
187 correlations and R2 values, see methods). TIGRFAMs did not exhibit these patterns.
188 Selected microbial KO and genera were positively correlated with SOC, N, and CEC and
189 negatively associated with bulk density and clay; while Pfam cluster 1 exhibited opposite
190 patterns. Samples in KO cluster 7 and genus cluster 6 had high relative abundances of genes
191 associated with energy generation (K00334. K00339, K00324, K08738), filamentous
192 bacteria/biofilms (K11903, Nocardia, Streptomyces, Mycobacterium, Leptolyngbya), chemically
193 complex organic matter decomposers (Nocardia, Rhodococcus, Streptomyces) and N-fixers and
194 plant symbionts (Rhizobium, Bradyrhizobium, Nostoc), while genes associated with N-related
195 processes (K07685, K10041) and microbial interactions including signaling and secretion
196 (K12537, K02661) changed most dramatically across environmental gradients. A diverse set of
197 organisms were amongst the most variable clades across environmental gradients, including N
198 cyclers (Uliginosibacterium, Microvirgula), autotrophs/anaerobes (Methanosalsum, Salana,
199 Pseudobacteroides, Sulfurirhabdus, Microvirgula, Desulfallas, Pelotomaculum, Risungbinella),
200 and one plant genus (Zea) and plant-related microbes (Pseudobacteroides, Zeaxanthinibacter,
201 Parapedobacter). Samples in Pfam cluster 1 had high relative abundances of organic N-
202 (PF01979, PF08352, PF01842), P- (PF00406, PF01380), and signaling-related (PF01627,
203 PF09413) genes, while many genes thought to be associated with eukaryotes changed most
204 dramatically across environmental gradients (PF04178, PF03188, PF01793).
205
206 **Discussion.**
207
208 *Abundant microbial genera, functional genes, and environmental correlates across all soils.*

The most abundant microbial genera spanned diverse microbial lineages and known functions, highlighting high soil diversity relative to other environments (5, 18, 45, 46). This included members of commonly observed soil taxa including Alphaproteobacteria, Betaproteobacteria, and Actinobacteria (45). Widespread genomic functions included growth, resource acquisition and stress tolerance strategies essential for microbial persistence in heterogeneous soil environments. For instance, bacterial secretion systems were common as well as DNA repair/replication, transport, and energy generation functions. Interestingly, microbial genes for the sensing and/or tolerance of fluctuations in temperature, salinity, light, and redox conditions were encoded among the most abundant microbial genes across all soils; all of which are common soil microbial stressors (86, 87). Collectively, this distribution underscores the heterogeneous mixture of stressors in 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, hypothesized drivers of the soil microbiome including soil organic carbon, total N, and cation exchange capacity were not associated with distinct genomic fingerprints. Soil carbon, N, and cation exchange capacity can reflect resource availability and are often drivers of 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, 92–95). While our results support pH as a

232 primary determinant of soil microbial structure and function (5, 6, 28, 45), it is interesting that
233 bulk density was also a strong correlate of genomic patterns. This may reflect the importance of
234 biophysical and hydraulic properties on soil microbiome structure and function (92, 96–98).
235 The genomic fingerprint of alkaline soils may be important for understanding
236 belowground microbial dynamics in arid and semiarid ecosystems, which account for more than
237 35% of the global land surface (99). The soil matrix in these systems is porous, with low rates of
238 water retention, and contains an abundance of calcium carbonate that limits P bioavailability
239 (100). Accordingly, its genomic fingerprint was enriched in genes encoding phosphatases and
240 related functions (PF01966, PF00481) that may signify the importance of microbial P acquisition
241 and/or limitation in alkaline soils (101–104). Vitamin B-related genes were also included in this
242 fingerprint and have also been linked to P solubilization (K01662, PF01872)(105). The presence
243 of a urea transporter (K08717) may likewise indicate the importance of nutrient acquisition in
244 alkaline soils (106). Notably, the high-Ca2+ alkaline soil environment was also characterized by
245 a calcium/proton exchanger (TIGR00378) which regulates intracellular Ca2+ concentrations to
246 prevent Ca2+ overload (107) and an uncharacterized hydrophobic domain (TIGR00271) that
247 may arise from low moisture environments that are often associated with alkaline soils.
248 In contrast, more acidic soils were characterized by genes encoding reactions involving
249 ammonia as well as wood decomposition, both of which can produce H+ as a byproduct. For
250 instance, nitrification is a hallmark of microbial genera Rhodopseudomonas and Klebsiella (108–
251 110) and increases soil acidity through the release of H+ from ammonia. Likewise, wood-
252 degrading fungi often secrete acidic compounds to aid in the digestion of chemically-complex
253 polymers (111–114). This is consistent with the association of K14333, K01684, TIGR03404,
254 and TIGR02093 with low pH soils.

255 Various members of Alphaproteobacteria were included in the genomic fingerprints of
256 both high and low pH soils. Alphaproteobacteria has been previously shown to be sensitive to
257 variation in soil pH, and some members may also be associated with vitamin B-related processes
258 indicative of alkaline soils (115, 116). These results suggest Alphaproteobacteria as a sensitive
259 taxa to global environmental changes in soil pH (e.g., atmospheric nitrogen deposition).

260 Additionally, bulk density can influence soil microbiomes by reflecting biophysical soil
261 properties such as texture, water content, porosity, and mineralogy that drive community
262 structure and function (19, 53, 54, 117, 118). Patterns in the relationships between these
263 variables and the soil microbiome can be complex at global scales (19), and our findings are to
264 our knowledge, the first to report a global-scale relationship between bulk density and the
265 distribution of genes encoded by the soil microbiome.

266 Genomic signatures of low bulk density soils were characterized by genes common to
267 water-logged and/or agricultural soils, while high bulk density soils included potential functions
268 associated with transport mechanisms, stress tolerance, and virulence. For instance, K11261,
269 TIGR03323, and TIGR01392 – all signatures of low bulk density soil – are each associated with
270 methanogenic processes (119). Low bulk density soils were further characterized by a suite of
271 anaerobic and/or aquatic microorganisms including Gardnerella, Cetia, Salmonella, Olleya,
272 Mycoplasma, Escherichia, Enterobacter, and Chitinophaga that also supports microbial
273 persistence in water-saturated environments (120–128). The presence of potential functions
274 associated with organic N cycling in the low bulk density fingerprint (K00992, K11959)
275 additionally suggests the influence of agriculture on microorganisms, where tillage generates
276 loosely packed soils and fertilizer application enhances nutrient cycles. In contrast, the genomic
277 fingerprint of high bulk density soils suggest an environment inclusive of density-dependent

relationships between microorganisms, their substrates, and pathogens (transport: K00854, K11733, TIGR01957, TIGR01203; virulence: PF03631, 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 and can elevate soil temperatures through the exclusion of vegetation (K03799).

Finally, 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 highlight the genomic fingerprint associated with tundra ecosystems in particular, as tundras are experiencing rapid changes in response to global climate drivers (129, 130). Tundras displayed the highest number of taxa and potential functions in their genomic fingerprint, the vast majority of which were positively associated with the tundra biome. This provides evidence for a consistent and distributed genomic basis for enhanced greenhouse gas emissions associated with permafrost thaw in tundra (129, 131, 132). Methanogenesis in particular is responsible for fluxes of methane in warming tundra. A gene involved in nitric oxide cycling (PF08768) (133–135), as well as a variety of anaerobic organisms and P-related genes (K03525, K00997, K00992, K13497, PF08009) (136–138), 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 signature (PF11203, PF05045), lending credence to a small number of recent studies suggesting permafrost environments as one of the largest reservoirs of soil viruses (139–142) and a possible linkage between soil methane and viral infection (143).

Flooded grasslands/savannas also generally had positive associations with functions in their genomic fingerprint, many of which encoded nutrient cycling, stress tolerance, or motility.

301 Motility (PF00669, PF00482, Paeniclostridium) may be reflective of beneficial microbial
302 lifestyles in saturated systems, where movement is facilitated by high pore space connectivity.
303 Genomic processes reflecting organic N (K13497, K07395, K19702, K02042), P (K02042,
304 K00854, K02800, PF04273), and stress tolerance (TIGR04282, TIGR02348) may reflect
305 microbial adaptations to nutrient scarcity and other stressors. Interestingly, a mechanism for
306 arsenic resistance was reflected in the flooded grasslands/savannas genomic fingerprint, possibly
307 reflecting arsenic groundwater contamination in many regions of the world (144).

308 In contrast, boreal forest and montane grasslands/shrublands were signified by many
309 negative associations with Mg and P (boreal: K03284, K06217, montane: K01520, K04765,
310 TIGR01203), and organic N and carbohydrate-related genes (boreal: K02800, K11472, PF00208,
311 PF04685; montane: K01816, K07395, K00600, PF00208); as well as many anaerobic microbial
312 genera – though the specific microbial attributes differed amongst these two fingerprints. The
313 boreal forest fingerprint uniquely included some metal-related genes (TIGR00378, TIGR04282,
314 TIGR00003); while montane systems were uniquely characterized by virulence (PF09299),
315 motility (PF00669), and sulfur-related genes (K03644). The negative associations with carbon,
316 N, and P suggest that inorganic nutrients and carbon may be more available in these systems
317 relative to other systems. Mg is often associated with plant growth (145) – taken together, these
318 biomes may be slower growing and have fewer resource constraints overall than other
319 ecosystems.

320 Collectively, we reveal novel biogeographical patterns in specific aspects of soil
321 microbiome composition and potential function that have not been reported elsewhere. Our
322 results are consistent with ecological processes hypothesized to influence the soil microbiome
323 and provide greater molecular specificity than revealed by previous works. They enable us to

324 better understand microbial lifestyles that support the persistence of certain members over others;
325 such knowledge underlies a predictive understanding of soil microbiome changes with shifts in
326 the global climate. Additionally, relationships between biomes and specific microbial attributes,
327 beyond associations with individual variables, suggest that there may be multiple simultaneous
328 factors that structure soil microbiome composition and function. Because of this, we next
329 explored relationships between microbial attributes and soil factors that arose naturally out of
330 unsupervised analysis. This allowed us to infer interactive effects that could not be assessed
331 when specifying a single environmental gradient.
332
333 *Environmental interactions correspond to variability in soil genomic potential.*
334 Unsupervised machine learning uncovered soil microbial genomic attributes associated
335 with coordinated associations with soil organic carbon content, total N, and cation exchange
336 capacity and oppositely with clay content and bulk density (Figure 5); which may represent some
337 of the most biologically-active soils on Earth. Soils rich in organic carbon and N with high cation
338 exchange capacity constitute resource- and energy-rich environments, supported by an
339 abundance of energy generating genes in the fingerprint of these soils (K00334, K00339,
340 K00324, K08738). Interestingly, these environments are best explained by a combination of
341 environmental variables rather than an individual factor. As an example, soils with high soil
342 organic carbon alone could be limited by N or other nutrients; thus, we propose that genomic
343 fingerprints for multiple simultaneously interacting factors may be among the most critical for
344 understanding global biogeochemistry. They are nearly impossible to discern without
345 unsupervised statistical approaches, and we know of no other work that has done so.

346 Soils with high organic carbon, N, and cation exchange capacity were characterized by
347 genes associated with resource scavenging, plant-microbe interactions, and heterotrophic
348 metabolisms. We observed numerous attributes in their genomic fingerprint associated with
349 filments or biofilms, intra-organism signalling. and/or chemically-complex organic matter
350 decomposition (K11903, K12537, K02661, Nocardia, Streptomyces, Mycobacterium,
351 Leptolyngbya, Nocardia, Rhodococcus, Streptomyces). This is consistent with microbial
352 morphologies and lifestyles that support the decay of chemically recalcitrant polymers (e.g.,
353 wood), as well as nutrient acquisition and transport across soil pore networks. Plant-microbe
354 interactions, including symbiotic N-fixing organisms, were also a key feature of these soils,
355 highlighting the importance of above- and belowground connectivity in productivity
356 (Rhizobium, Bradyrhizobium, Nostoc, Pseudobacteroides, Zeaxanthinibacter, Parapedobacter,
357 Zea)(146). In contrast, anaerobic and autotrophic organisms were highly sensitive to changes
358 across these connected environmental gradients, possibly reflecting their exclusion from more
359 oxygen- and energy-rich environments where heterotrophic organisms tend to dominate.
360 We also selected one Pfam cluster that was associated with low soil organic carbon, N,
361 and cation exchange capacity and with high clay and bulk density soils, which seemed to
362 underscore the importance of fungi in resource- and energy-rich environments. The most
363 sensitive genes in this cluster of samples corresponded to eukaryotic organisms (PF04178,
364 PF03188, PF01793). Previous work has shown fungal sensitivity to soil resource availability, as
365 well as high fungal diversity in organic-rich tropical soils (52). Overall, these insights provide
366 greater resolution into the microbial mechanisms that support both belowground biogeochemical
367 cycles and aboveground productivity than ever before, and they can provide a basis for the next
368 generation of microbial ecology research and model development.

369

370 *Phylogenetic Coherence in Microbial Biogeography.*

371 To what extent selection acts on phylogenetically conserved properties versus functional
372 attributes and to what extent that translates into functional differences in the microbiome (e.g.,
373 response-effect traits (147)) are long running questions in microbial ecology . Some previous
374 works have shown little connection between functional attributes and microbial phylogeny (44,
375 148), while others have indicated the existence of taxa-function relationships (88). And still
376 others have suggested variation in the coupling of taxonomy and function across various
377 functions of interest, levels of phylogenetic resolution, and/or spatial scales (34, 55, 61, 149,
378 150). We found strong phylogenetic coherence in the response of genera to changes in the soil
379 environment (Figure 4B); that is, closely related organisms tended to respond similarly to
380 environmental change. This implies that evolutionary history, life history strategies, and/or
381 morphology (which tend to be phylogenetically conserved (151)) could influence changes in
382 microbial community membership in response to environmental change. As such, metabarcoding
383 approaches that yield microbiome taxonomy may be among the first indicators of ecosystem
384 transitions, such as response to disturbances and transitions between state archetypes (30).

385

386 *Conclusions.*

387 Despite the significance of soil microorganisms in global biogeochemistry, we still know
388 little about the ecological processes that regulate their composition and function across the wide
389 range of global soil environments. We used machine learning to sift through tens of thousands of
390 taxa and potential functions encoded by 1512 global soil microbiomes, revealing novel
391 biogeographical patterns in soil microbiome composition and functional potential with high

392 molecular resolution. Specifically, we demonstrate shifts in the potential for microbial nutrient
393 acquisition across pH gradients; in stress, transport, and redox-based processes across changes in
394 soil bulk density; and in genes and organisms associated with soil greenhouse gas emissions
395 across biomes. We also uncover a suite of metabolic processes that are enriched in the microbial
396 genomes of energy-rich soils. These changes were coincident with phylogenetically congruent
397 compositional change – suggesting that closely related soil microbial taxa are sensitive to similar
398 environmental stressors. Our work enables us to better understand microbial lifestyles that
399 support the persistence of certain members over others along environmental gradients; such
400 knowledge underlies a predictive understanding of soil microbiome changes with shifts in the
401 global climate and is vital to constraining biochemical reactions that regulate the release of
402 greenhouse gasses from soils.
403
404 **Materials and Methods.**
405
406 *Dataset Description and Normalization.*
407 We collected 1512 manually-curated soil metagenomic samples available in the Joint
408 Genome Institute's (JGI) Integrated Microbial Genomes and Microbiomes (IMG/M) platform in
409 August 2020 as described by Graham et al. 2023 (in review), of which 1493 were associated with
410 latitudinal and longitudinal information. Samples span most major biomes and continents (Figure
411 1; Table S1 and Table S2). Sequences were quality controlled, assembled, and annotated by
412 KEGG Orthology (KO), the Pfam and TIGRFAM protein family databases, and read-based
413 taxonomy by the JGI's standard workflows (80–84). Due to the collective size of assembled
414 sequences, we focused on annotations and read-based taxonomy for downstream analysis, which

415 is more computationally feasible than sequence assemblies or metagenome-assembled genomes
416 that are impractical at the scale of thousands of samples.

417 Each dataset (i.e., KO, Pfam, TIGRFAM, and read-based taxonomy) was independently
418 normalized by the following procedure prior to analysis. For Pfam and KO annotations only,
419 samples with greater than 2000 total reads were retained. For TIGRFAM annotations only,
420 samples with greater than 500 total reads were retained. A lower cutoff was used for TIGRFAM
421 data as the total read counts were lower in this dataset. For all three annotations, only functions
422 with a non-zero value in at least half of the samples or a total count of 7560 (corresponding to an
423 average of 5 reads per sample for each function) were retained. For species, we retained only
424 samples with greater than 4000 counts and species with a non-zero value in at least half of the
425 samples or a total count of 7560 (corresponding to an average of 5 reads per sample for each
426 species). After removing low abundant samples and functions/species, all four datasets were then
427 normalized using the trimmed mean of M values (TMM) method (152). We found that this
428 method gave us the best results when examining normalized data using a box and whisker plot
429 (data not shown).

430 We assigned biome type and soil properties to each sample using publicly available
431 resources. We used latitude and longitude to derive using Olson biomes (153) for each sample
432 using data provided by the World Wildlife Federation (154). Soil parameters were collected from
433 the SoilGrids250m database from 0-5 cm (155). SoilGrids250m is a spatial interpolation of
434 global soil parameters using ~150,000 soil samples and 158 remote sensing-based products.
435 Here, we focus on six parameters often associated with soil microbiomes: bulk density, cation
436 exchange capacity, nitrogen, pH, soil organic C, and clay content. Because our focus on spatial
437 dynamics and soils were collected at various times, we did not include temporally dynamic

438 variables such as soil moisture or temperature in our set of environmental parameters, though we
439 acknowledge they may have profound impacts on the soil microbiome.
440
441 *Machine Learning for Genomic Fingerprints.*
442 We selected random forests as our primary statistical approach by assessing the ability of
443 various machine learning algorithms to adequately parse soil microbial attributes across
444 environmental variables and ecological biomes. Models were built and validated using 10-fold
445 cross validation repeated five times, 50% of the data for training, and 50% of the data for testing.
446 Models from different algorithms were compared using cross validation (CV) scores and
447 variance explained (data not shown). Random forest models consistently yielded comparatively
448 high CV scores and proportion of variance explained.
449 Thus, we constructed random forest models to extract genomic fingerprints from soil
450 microbiomes using two approaches (1) a supervised split across variables known to be associated
451 with soil microbiomes and (2) an unsupervised approach that allowed predictors and their
452 interactions to naturally arise. Random forest models were built in R Software (156) for each
453 attribute type independently using the 'caret' package (157).
454 In (1), we divided samples into either two groups containing the highest and lowest 10%
455 of values for each environmental factor or one group per biome. We then used random forests to
456 rank each microbial attribute according to their importance (i.e., difference in mean square error
457 normalized by standard error, varImp function in 'caret' package) to differentiate samples across
458 each environmental factor or biome. Subsequently, because variable importances followed an
459 exponential decay, we used breakpoint analysis to define a set of the most important variables
460 distinguishing each group ('strucchange' package, h = 0.01 with a minimum segment size of

461 two) (158). Because this resulted in one set of attributes that drove variation in soil microbiomes
462 in relation to a given environmental factor or biome, we extracted attributes associated with high
463 (positive slope) or low (negative slope) values of each environmental factor using linear
464 regression. The resultant attributes comprised two genomic fingerprints per environmental factor
465 or biome—one associated with low values and one associated with high values.

466 In (2), we first used k-means clustering to group samples with similar microbial attributes
467 using R packages: 'vegan' (159), 'mgcv' (160), and 'cluster' (161). The optimal number of
468 clusters was chosen by evaluating the total sum of squares and the silhouette index (162). To
469 determine the extent to which soil microbiomes within each cluster corresponded to changes in
470 environmental factors, we related the total normalized abundance of attributes per sample to each
471 environmental factor using linear regression. We then selected one cluster per attribute type for
472 deeper investigation, chosen by the number of significant correlations with environmental
473 factors. To generate a genomic fingerprint for the selected cluster, we extracted attributes that
474 were either most abundant or most variable within the cluster. As in approach (1), we used
475 breakpoint analysis to define a set of the most abundant attributes per cluster ('strucchange'
476 package, h = 0.01 with a minimum segment size of two). To determine the most variable
477 attributes per cluster, we calculated the coefficient of variation for each attribute and then used
478 breakpoint analysis to extract the most variable attributes.

479 We used the following packages for data manipulation and visualization: 'ggplot2' (163),
480 'dplyr' (164), 'factoextra' (165), 'Hmisc' (166), 'colorspace' (167), 'RColorBrewer' (168),
481 'gridExtra' (169), 'tidyverse' (170), and 'maps' (171). Additionally, for visualizations of read-
482 based taxonomy, we generated phylogenetic trees using phyloT v2 (172), an online tree
483 generator based on the Genome Taxonomy Database. Then, we visualized the tree in R using the

packages ‘ggtree’(173), ‘treeio’(174), and ‘ggnewscale’(175). We generated heatmaps using the ‘pheatmap’ package (176).

Acknowledgements.

This work was performed under a Laboratory Directed Research and Development (LDRD) from the Pacific Northwest National Laboratory (PNNL). PNNL is operated by 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 dataset.

Data availability.

All data are publicly available via the JGI IMG/MG platform.

Figures.

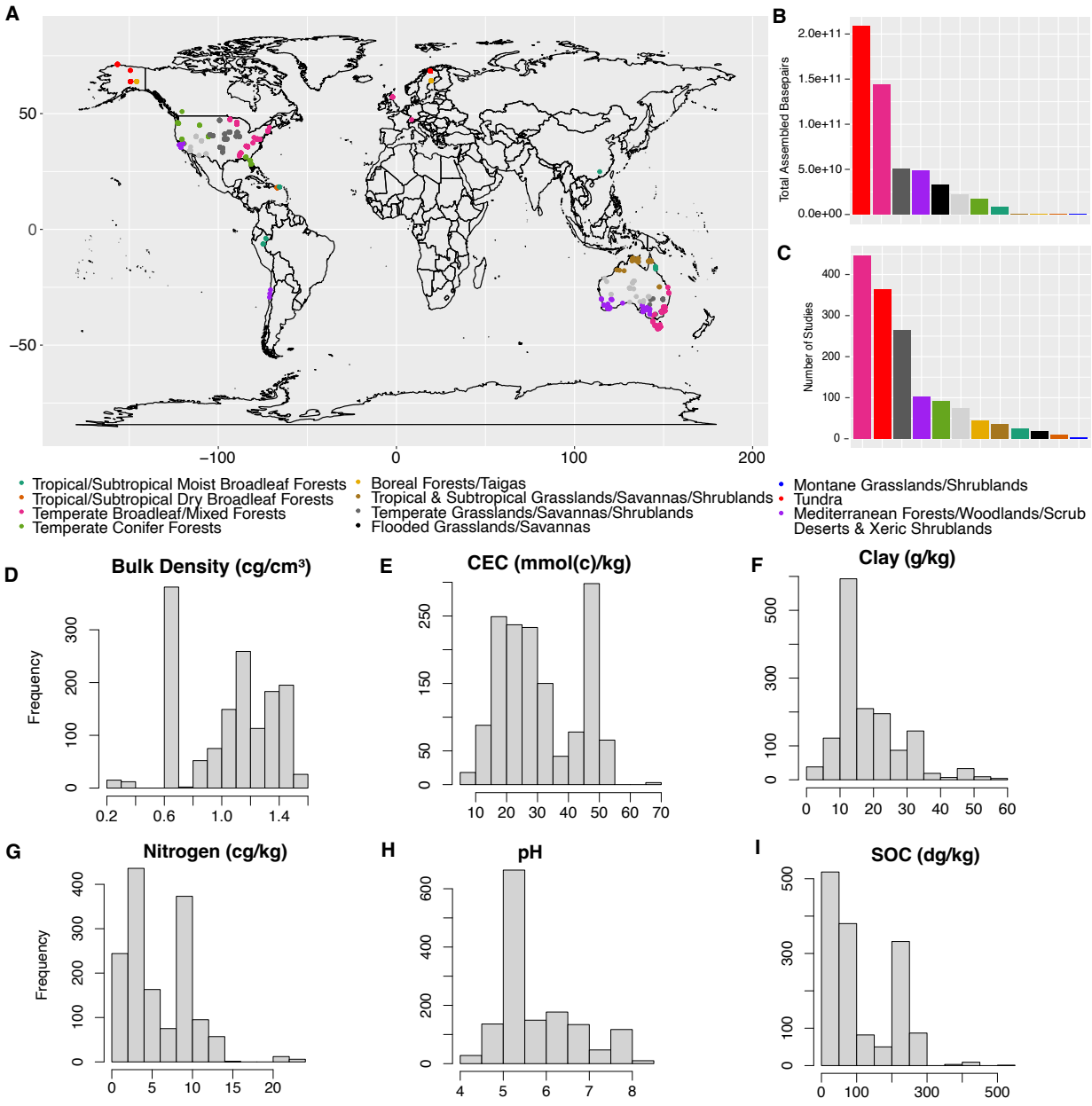

Figure 1. Dataset description. (A) Map of samples, (B) total assembled bp, and (C) number of samples; all colored by biome according to Olson et al. (153). (D-I) distribution of environmental variables from SoilGrids250m (155).

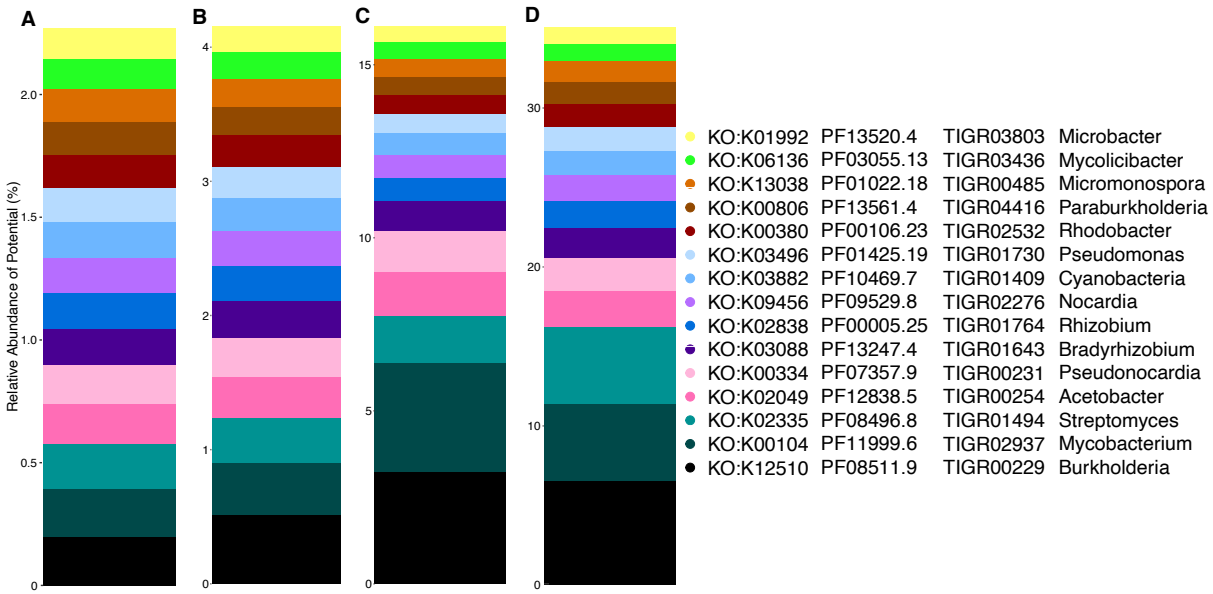

Figure 2. Most abundant genomic attributes across all soils. Relative abundance of the top 15 most abundant soil microbial genomic attributes annotated by (A) KEGG orthology, protein family (B, Pfam, C, TIGRFAM), and (D) read-based taxonomy collated at the genus-level.

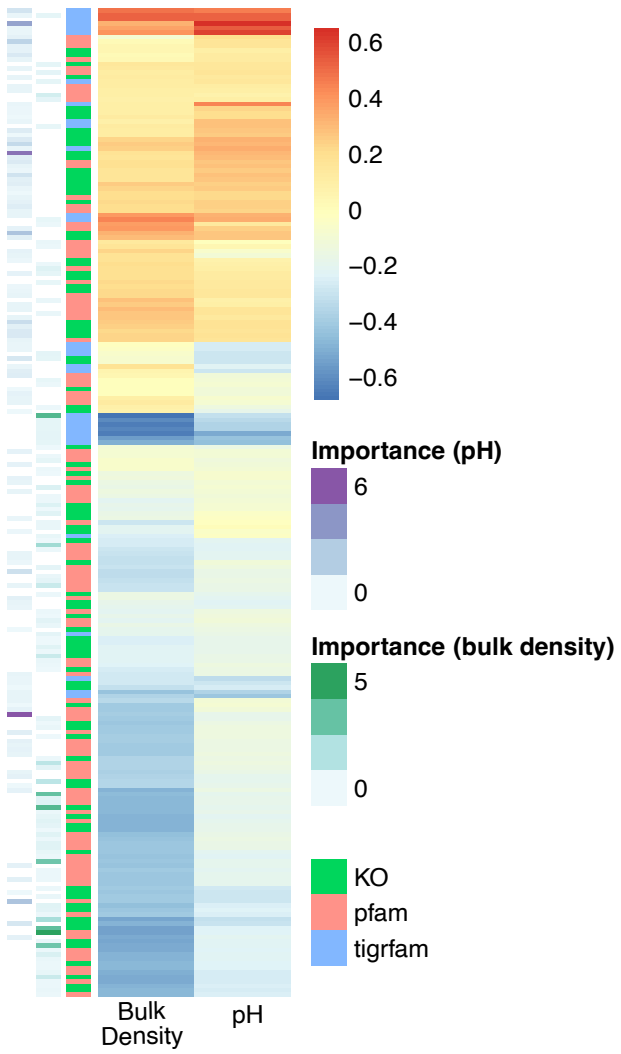

Figure 3. Genomic fingerprints of high- and low-pH and -bulk density soil environments. Genomic attributes selected in at least one fingerprint for pH or bulk density are visualized. Pearson correlation coefficient is shown from blue to red in the primary heatmap, with bulk density correlations on the left and pH correlations on the right. The sidebars (respectively, from left to right) represent variable importance from random forest models of pH and bulk density and the type of attribute (e.g., KO, Pfam, 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 Figure 4.

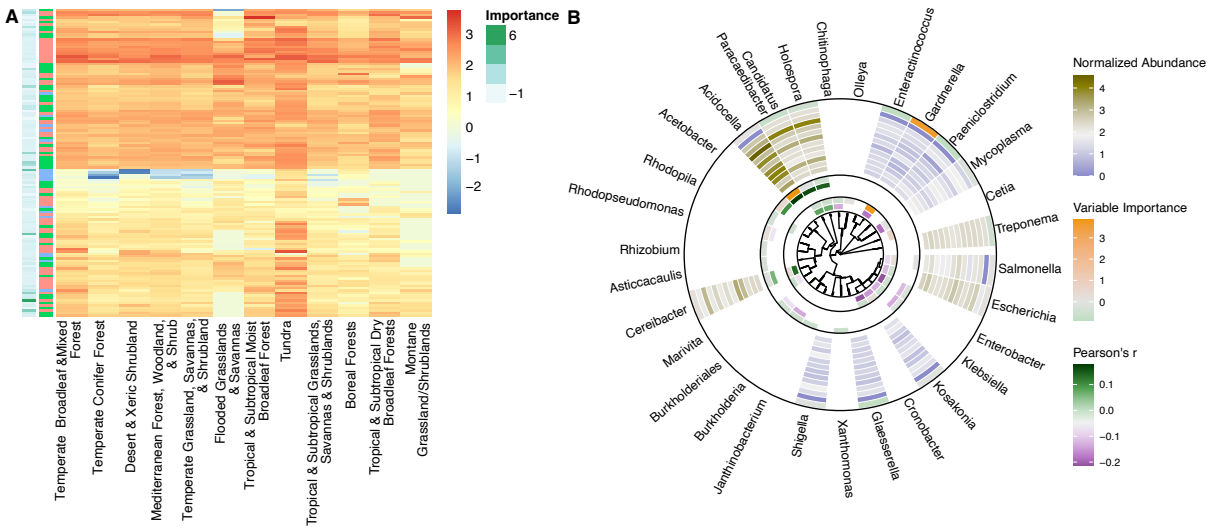

Figure 4. Genomic fingerprints of soil biomes and phylogenetic distributions of selected soil microbiomes. (A) The normalized abundance of microbial attributes selected in at least one biome fingerprint is shown from blue to red. Variable importance is shown in the left side bar and attribute type is shown in the right side bar. Please refer to Extended Data 1 for more information on the specific attributes associated with each soil environment. (B) Microbial genuses 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 & Mixed Forests; Tundra; Mediterranean Forests, Woodlands & Scrub; Tropical & Subtropical Grasslands, Savannas & Shrublands; Temperate Grasslands, Savannas & Shrublands; Boreal Forests/Taigas; Deserts & Xeric Shrublands; Temperate Conifer Forests; Tropical & Subtropical Dry Broadleaf Forests; Flooded Grasslands & Savannas; Tropical & Subtropical Moist Broadleaf Forests; and Montane Grasslands & Shrublands). Variable importance is shown in the outermost ring from green to orange.

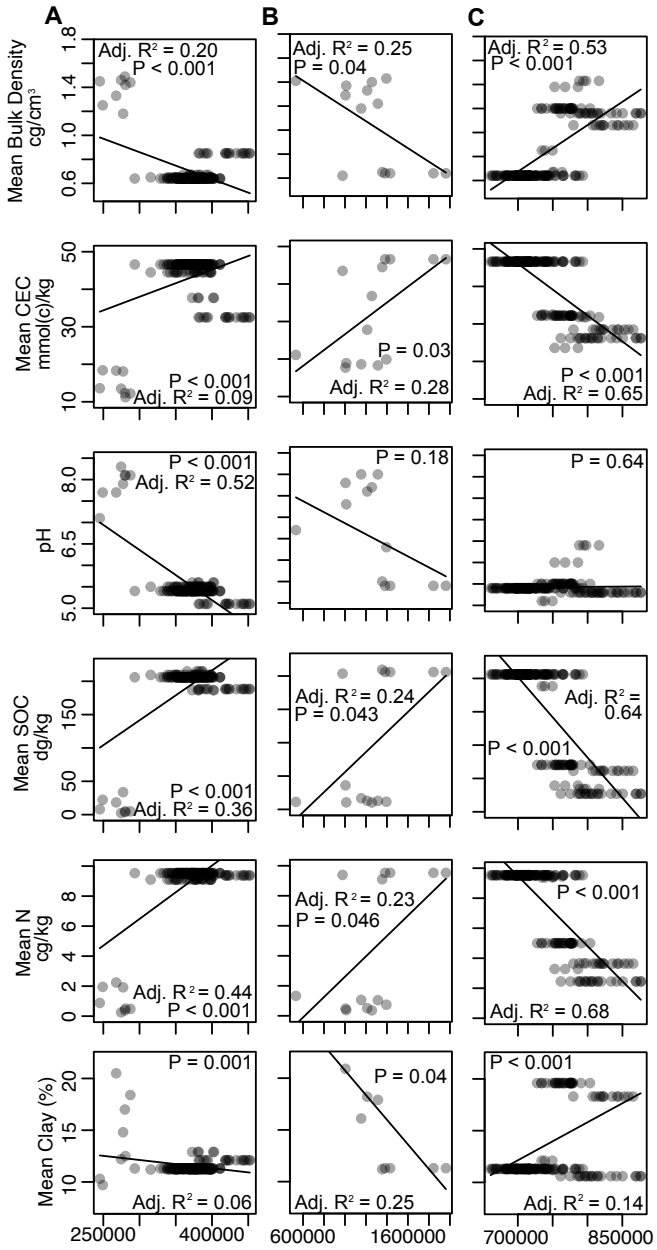

Figure 5. Unsupervised clusters of microbial attributes associated with multiple soil factors. Column (A) KEGG Orthology, (B) microbial genus, and (C) Pfam. TIGRFAM annotations did not yield a satisfactory model. Rows represent environmental variables. The total normalized abundance of an attribute per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

542 **References**

543 1. Falkowski PG, Fenchel T, Delong EF. 2008. The microbial engines that drive Earth's
544 biogeochemical cycles. Science 320:1034–1039.

545 2. Wagg C, Bender SF, Widmer F, van der Heijden MGA. 2014. Soil biodiversity and soil
546 community composition determine ecosystem multifunctionality. Proc Natl Acad Sci
547 USA 111:5266–5270.

548 3. Delgado-Baquerizo M, Maestre FT, Reich PB, Jeffries TC, Gaitan JJ, Encinar D,
549 Berdugo M, Campbell CD, Singh BK. 2016. Microbial diversity drives
550 multifunctionality in terrestrial ecosystems. Nat Commun 7:10541.

551 4. Fierer N. 2017. Embracing the unknown: disentangling the complexities of the soil
552 microbiome. Nat Rev Microbiol 15:579–590.

553 5. Fierer N, Strickland MS, Liptzin D, Bradford MA, Cleveland CC. 2009. Global patterns
554 in belowground communities. Ecol Lett 12:1238–1249.

555 6. He L, Mazza Rodrigues JL, Soudzilovskaia NA, Barceló M, Olsson PA, Song C,
556 Tedersoo L, Yuan F, Yuan F, Lipson DA, Xu X. 2020. Global biogeography of fungal
557 and bacterial biomass carbon in topsoil. Soil Biol Biochem 151:108024.

558 7. Melillo JM, Frey SD, DeAngelis KM, Werner WJ, Bernard MJ, Bowles FP, Pold G,
559 Knorr MA, Grandy AS. 2017. Long-term pattern and magnitude of soil carbon feedback
560 to the climate system in a warming world. Science 358:101–105.

561 8. Bradford MA, Wieder WR, Bonan GB, Fierer N, Raymond PA, Crowther TW. 2016.
562 Managing uncertainty in soil carbon feedbacks to climate change. Nat Clim Chang
563 6:751–758.

9. Davidson EA, Janssens IA. 2006. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. Nature 440:165–173.

10. 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.

11. 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.

12. Sparling GP. 1997. Soil microbial biomass, activity and nutrient cycling as indicators of soil health. Biological indicators of soil health.

13. Wieder WR, Bonan GB, Allison SD. 2013. Global soil carbon projections are improved by modelling microbial processes. Nat Clim Chang 3:909–912.

14. Graham EB, Hofmockel KS. 2022. Ecological stoichiometry as a foundation for omics-enabled biogeochemical models of soil organic matter decomposition. Biogeochemistry 157:31–50.

15. 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.

16. Wang G, Mayes MA, Gu L, Schadt CW. 2014. Representation of dormant and active microbial dynamics for ecosystem modeling. PLoS ONE 9:e89252.

17. 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:eaau1218.

18. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Zech Xu Z, Jiang L, Haroon MF, Earth Microbiome Project Consortium. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463.

19. 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 https://doi.org/10.1111/geb.13162.

20. Levin SA. 1992. The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. Ecology 73:1943–1967.

21. 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.

22. Brown JH, Gillooly JF, Allen AP, Savage VM, West GB. 2004. Toward a metabolic theory of ecology. Ecology 85:1771–1789.

23. Wright DH. 1983. Species-Energy Theory: An Extension of Species-Area Theory. Oikos 41:496.

24. Blackburn TM, Gaston KJ, Loder N. 1999. Geographic gradients in body size: a clarification of Bergmann's rule. Diversity and distributions.

25. Stevens GC. 1989. The Latitudinal Gradient in Geographical Range: How so Many

Species Coexist in the Tropics. Am Nat 133:240–256.

26. Meyer KM, Memiaghe H, Korte L, Kenfack D, Alonso A, Bohannan BJM. 2018. Why do microbes exhibit weak biogeographic patterns? ISME J 12:1404–1413.

27. 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.

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, Põlme S, Sunagawa S, Ryberg M, Tedersoo L, Bork P. 2018. Structure and function of the global topsoil microbiome. Nature 560:233–237.

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.

30. Graham EB, Knelman JE. 2023. Implications of soil microbial community assembly for ecosystem restoration: patterns, process, and potential. Microb Ecol 85:809–819.

31. Finlay BJ. 2002. Global dispersal of free-living microbial eukaryote species. Science 296:1061–1063.

32. Greening C, Grinter R, Chiri E. 2019. Uncovering the Metabolic Strategies of the Dormant Microbial Majority: towards Integrative Approaches. mSystems 4.

33. Lennon JT, Jones SE. 2011. Microbial seed banks: the ecological and evolutionary implications of dormancy. Nat Rev Microbiol 9:119–130.

630 34. Horner-Devine MC, Lage M, Hughes JB, Bohannan BJM. 2004. A taxa-area
631 relationship for bacteria. Nature 432:750–753.

632 35. Martiny JBH, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, Green JL, Horner-
633 Devine MC, Kane M, Krumins JA, Kuske CR, Morin PJ, Naeem S, Ovreås L,
634 Reysenbach A-L, Smith VH, Staley JT. 2006. Microbial biogeography: putting
635 microorganisms on the map. Nat Rev Microbiol 4:102–112.

636 36. Xu X, Thornton PE, Post WM. 2013. A global analysis of soil microbial biomass
637 carbon, nitrogen and phosphorus in terrestrial ecosystems. Global Ecology and
638 Biogeography 22:737–749.

639 37. Serna-Chavez HM, Fierer N, van Bodegom PM. 2013. Global drivers and patterns of
640 microbial abundance in soil. Global Ecology and Biogeography 22:1162–1172.

641 38. Locey KJ, Lennon JT. 2016. Scaling laws predict global microbial diversity. Proc Natl
642 Acad Sci USA 113:5970–5975.

643 39. Geyer KM, Kyker-Snowman E, Grandy AS, Frey SD. 2016. Microbial carbon use
644 efficiency: accounting for population, community, and ecosystem-scale controls over the
645 fate of metabolized organic matter. Biogeochemistry 127:173–188.

646 40. Rousk J, Bååth E, Brookes PC, Lauber CL, Lozupone C, Caporaso JG, Knight R, Fierer
647 N. 2010. Soil bacterial and fungal communities across a pH gradient in an arable soil.
648 ISME J 4:1340–1351.

649 41. Larkin AA, Martiny AC. 2017. Microdiversity shapes the traits, niche space, and
650 biogeography of microbial taxa. Environ Microbiol Rep 9:55–70.

651 42. Dequiedt S, Thioulouse J, Jolivet C, Saby NPA, Lelievre M, Maron P-A, Martin MP,

652 Prévost-Bouré NC, Toutain B, Arrouays D, Lemanceau P, Ranjard L. 2009.
653 Biogeographical patterns of soil bacterial communities. Environ Microbiol Rep 1:251–
654 255.

655 43. Fuhrman JA, Steele JA, Hewson I, Schwalbach MS, Brown MV, Green JL, Brown JH.
656 2008. A latitudinal diversity gradient in planktonic marine bacteria. Proc Natl Acad Sci
657 USA 105:7774–7778.

658 44. Nelson MB, Martiny AC, Martiny JBH. 2016. Global biogeography of microbial
659 nitrogen-cycling traits in soil. Proc Natl Acad Sci USA 113:8033–8040.

660 45. Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-González A, Eldridge DJ,
661 Bardgett RD, Maestre FT, Singh BK, Fierer N. 2018. A global atlas of the dominant
662 bacteria found in soil. Science 359:320–325.

663 46. Fierer N, Jackson RB. 2006. The diversity and biogeography of soil bacterial
664 communities. Proc Natl Acad Sci USA 103:626–631.

665 47. Bell T, Ager D, Song J-I, Newman JA, Thompson IP, Lilley AK, van der Gast CJ. 2005.
666 Larger islands house more bacterial taxa. Science 308:1884.

667 48. Zhou J, Kang S, Schadt CW, Garten CT. 2008. Spatial scaling of functional gene
668 diversity across various microbial taxa. Proc Natl Acad Sci USA 105:7768–7773.

669 49. Ranjard L, Dequiedt S, Chemidlin Prévost-Bouré N, Thioulouse J, Saby NPA, Lelievre
670 M, Maron PA, Morin FER, Bispo A, Jolivet C, Arrouays D, Lemanceau P. 2013.
671 Turnover of soil bacterial diversity driven by wide-scale environmental heterogeneity.
672 Nat Commun 4:1434.

673 50. Wang X-B, Lü X-T, Yao J, Wang Z-W, Deng Y, Cheng W-X, Zhou J-Z, Han X-G.

674 2017. Habitat-specific patterns and drivers of bacterial β-diversity in China's drylands.
675 ISME J 11:1345–1358.

676 51. Bell T. 2010. Experimental tests of the bacterial distance-decay relationship. ISME J
677 4:1357–1365.

678 52. Tedersoo L, Bahram M, Põlme S, Kõljalg U, Yorou NS, Wijesundera R, Villarreal Ruiz
679 L, Vasco-Palacios AM, Thu PQ, Suija A, Smith ME, Sharp C, Saluveer E, Saitta A,
680 Rosas M, Riit T, Ratkowsky D, Pritsch K, Põldmaa K, Piepenbring M, Abarenkov K.
681 2014. Fungal biogeography. Global diversity and geography of soil fungi. Science
682 346:1256688.

683 53. Chau JF, Bagtzoglou AC, Willig MR. 2011. The effect of soil texture on richness and
684 diversity of bacterial communities. Environmental Forensics 12:333–341.

685 54. Kaiser K, Wemheuer B, Korolkow V, Wemheuer F, Nacke H, Schöning I, Schrumpf M,
686 Daniel R. 2016. Driving forces of soil bacterial community structure, diversity, and
687 function in temperate grasslands and forests. Sci Rep 6:33696.

688 55. Allison SD, Horner-Devine MC. 2011. Drivers of bacterial β-diversity depend on spatial
689 scale. Proceedings of the ….

690 56. Kerr B, Riley MA, Feldman MW, Bohannan BJM. 2002. Local dispersal promotes
691 biodiversity in a real-life game of rock-paper-scissors. Nature 418:171–174.

692 57. Manzoni S, Schimel JP, Porporato A. 2012. Responses of soil microbial communities to
693 water stress: results from a meta-analysis. Ecology 93:930–938.

694 58. Lozupone CA, Knight R. 2007. Global patterns in bacterial diversity. Proc Natl Acad Sci
695 USA 104:11436–11440.

696 59. Philippot L, Raaijmakers JM, Lemanceau P, van der Putten WH. 2013. Going back to
697 the roots: the microbial ecology of the rhizosphere. Nat Rev Microbiol 11:789–799.

698 60. Shade A, Caporaso JG, Handelsman J, Knight R, Fierer N. 2013. A meta-analysis of
699 changes in bacterial and archaeal communities with time. ISME J 7:1493–1506.

700 61. Graham EB, Knelman JE, Schindlbacher A, Siciliano S, Breulmann M, Yannarell A,
701 Beman JM, Abell G, Philippot L, Prosser J, Foulquier A, Yuste JC, Glanville HC, Jones
702 DL, Angel R, Salminen J, Newton RJ, Bürgmann H, Ingram LJ, Hamer U, Nemergut
703 DR. 2016. Microbes as engines of ecosystem function: when does community structure
704 enhance predictions of ecosystem processes? Front Microbiol 7:214.

705 62. Malik AA, Martiny JBH, Brodie EL, Martiny AC, Treseder KK, Allison SD. 2020.
706 Defining trait-based microbial strategies with consequences for soil carbon cycling
707 under climate change. ISME J 14:1–9.

708 63. van der Heijden MGA, Bardgett RD, van Straalen NM. 2008. The unseen majority: soil
709 microbes as drivers of plant diversity and productivity in terrestrial ecosystems. Ecol
710 Lett 11:296–310.

711 64. Strickland MS, Lauber C, Fierer N, Bradford MA. 2009. Testing the functional
712 significance of microbial community composition. Ecology 90:441–451.

713 65. Schimel JP, Schaeffer SM. 2012. Microbial control over carbon cycling in soil. Front
714 Microbiol 3:348.

715 66. Green JL, Bohannan BJM, Whitaker RJ. 2008. Microbial biogeography: from taxonomy
716 to traits. Science 320:1039–1043.

717 67. Caldwell BA. 2005. Enzyme activities as a component of soil biodiversity: A review.

718 Pedobiologia 49:637–644.

719 68. Six J, Frey SD, Thiet RK, Batten KM. 2006. Bacterial and Fungal Contributions to
720 Carbon Sequestration in Agroecosystems. Soil Sci Soc Am J 70:555.

721 69. Naeem S, Wright JP. 2003. Disentangling biodiversity effects on ecosystem functioning:
722 deriving solutions to a seemingly insurmountable problem. Ecol Lett 6:567–579.

723 70. Cardinale BJ, Duffy JE, Gonzalez A, Hooper DU, Perrings C, Venail P, Narwani A,
724 Mace GM, Tilman D, Wardle DA, Kinzig AP, Daily GC, Loreau M, Grace JB,
725 Larigauderie A, Srivastava DS, Naeem S. 2012. Biodiversity loss and its impact on
726 humanity. Nature 486:59–67.

727 71. Ladau J, Shi Y, Jing X, He J-S, Chen L, Lin X, Fierer N, Gilbert JA, Pollard KS, Chu H.
728 2018. Existing climate change will lead to pronounced shifts in the diversity of soil
729 prokaryotes. mSystems 3.

730 72. Banerjee S, Schlaeppi K, van der Heijden MGA. 2018. Keystone taxa as drivers of
731 microbiome structure and functioning. Nat Rev Microbiol 16:567–576.

732 73. Banerjee S, Kirkby CA, Schmutter D, Bissett A, Kirkegaard JA, Richardson AE. 2016.
733 Network analysis reveals functional redundancy and keystone taxa amongst bacterial
734 and fungal communities during organic matter decomposition in an arable soil. Soil Biol
735 Biochem 97:188–198.

736 74. Neu AT, Allen EE, Roy K. 2021. Defining and quantifying the core microbiome:
737 Challenges and prospects. Proc Natl Acad Sci USA 118.

738 75. Bay SK, McGeoch MA, Gillor O, Wieler N, Palmer DJ, Baker DJ, Chown SL, Greening
739 C. 2020. Soil bacterial communities exhibit strong biogeographic patterns at fine

740 taxonomic resolution. mSystems 5.

741 76. de Wit R, Bouvier T. 2006. “Everything is everywhere, but, the environment selects”;
742 what did Baas Becking and Beijerinck really say? Environ Microbiol 8:755–758.

743 77. Bowman MM, Heath AE, Varga T, Battu AK, Chu RK, Toyoda J, Cheeke TE, Porter
744 SS, Moffett KB, LeTendre B, Qafoku O, Bargar JR, Mans DM, Hess NJ, Graham EB.
745 2023. One thousand soils for molecular understanding of belowground carbon cycling.
746 Front Soil Sci 3.

747 78. Keller M, Schimel DS, Hargrove WW, Hoffman FM. 2008. A continental strategy for
748 the National Ecological Observatory Network. Frontiers in Ecology and the
749 Environment 6:282–284.

750 79. Aoki-Kinoshita KF, Kanehisa M. 2007. Gene annotation and pathway mapping in
751 KEGG. Methods Mol Biol 396:71–91.

752 80. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. 2004. The KEGG resource for
753 deciphering the genome. Nucleic Acids Res 32:D277-80.

754 81. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M,
755 Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan
756 D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. Nucleic
757 Acids Res 47:D427–D432.

758 82. Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E. 2013. Tigrfams and
759 genome properties in 2013. Nucleic Acids Res 41:D387-95.

760 83. Huntemann M, Ivanova NN, Mavromatis K, Tripp HJ, Paez-Espino D, Palaniappan K,
761 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.

84. 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.

85. Taylor BL, Zhulin IB. 1999. PAS domains: internal sensors of oxygen, redox potential, and light. Microbiol Mol Biol Rev 63:479–506.

86. 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.

87. 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.

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 USA 109:21390–21395.

89. Freschet GT, Valverde-Barrantes OJ, Tucker CM, Craine JM, McCormack ML, Violle C, Fort F, Blackwood CB, Urban-Mead KR, Iversen CM, Bonis A, Comas LH, Cornelissen JHC, Dong M, Guo D, Hobbie SE, Holdaway RJ, Kembel SW, Makita N, Onipchenko VG, Roumet C. 2017. Climate, soil and plant functional types as drivers of global fine-root trait variation. J Ecol 105:1182–1196.

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.

91. Espeleta JF, Cardon ZG, Mayer KU, Neumann RB. 2017. Diel plant water use and competitive soil cation exchange interact to enhance NH4 + and K+ availability in the rhizosphere. Plant Soil 414:33–51.

92. 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.

93. 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.

94. 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.

95. 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.

96. 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-JA, Pett-Ridge J, Schulz M. 2020. From pools to flow: The PROMISE framework for new insights on soil carbon

806 cycling in a changing world. Glob Chang Biol 26:6631–6643.

807 97. McCarter CPR, Rezanezhad F, Quinton WL, Gharedaghloo B, Lennartz B, Price J,
808 Connon R, Van Cappellen P. 2020. Pore-scale controls on hydrological and geochemical
809 processes in peat: Implications on interacting processes. Earth-Science Reviews
810 207:103227.

811 98. Ruamps LS, Nunan N, Chenu C. 2011. Microbial biogeography at the soil pore scale.
812 Soil Biol Biochem 43:280–286.

813 99. Mares MA. 2017. Encyclopedia of deserts. books.google.com.

814 100. Hagin J, Hadas A. 1962. Solubility of calcium phosphate in calcareous soils. Nature
815 193:1211–1212.

816 101. Dick WA, Cheng L, Wang P. 2000. Soil acid and alkaline phosphatase activity as pH
817 adjustment indicators. Soil Biol Biochem 32:1915–1919.

818 102. Krämer S. 2000. Acid and alkaline phosphatase dynamics and their relationship to soil
819 microclimate in a semiarid woodland. Soil Biol Biochem 32:179–188.

820 103. Oliverio AM, Bissett A, McGuire K, Saltonstall K, Turner BL, Fierer N. 2020. The role
821 of phosphorus limitation in shaping soil bacterial communities and their metabolic
822 capabilities. MBio 11.

823 104. Laliberté E, Zemunik G, Turner BL. 2014. Environmental filtering explains variation in
824 plant diversity along resource gradients. Science 345:1602–1605.

825 105. Baya AM, Boethling RS, Ramos-Cormenzana A. 1981. Vitamin production in relation
826 to phosphate solubilization by soil bacteria. Soil Biol Biochem 13:527–531.

827 106. Pii Y, Mimmo T, Tomasi N, Terzano R, Cesco S, Crecchio C. 2015. Microbial
828 interactions in the rhizosphere: beneficial influences of plant growth-promoting
829 rhizobacteria on nutrient acquisition process. A review. Biol Fertil Soils 51:403–415.

830 107. Guerini D. 1998. The Ca2+ pumps and the Na+/Ca2+ exchangers. Biometals 11:319–
831 330.

832 108. Cabello P, Luque-Almagro VM, Roldán MD, Moreno-Vivián C. 2019. Nitrogen
833 CycleReference module in life sciences. Elsevier.

834 109. Pal RR, Khardenavis AA, Purohit HJ. 2015. Identification and monitoring of
835 nitrification and denitrification genes in Klebsiella pneumoniae EGD-HP19-C for its
836 ability to perform heterotrophic nitrification and aerobic denitrification. Funct Integr
837 Genomics 15:63–76.

838 110. Padhi SK, Tripathy S, Sen R, Mahapatra AS, Mohanty S, Maiti NK. 2013.
839 Characterisation of heterotrophic nitrifying and aerobic denitrifying Klebsiella
840 pneumoniae CF-S9 strain for bioremediation of wastewater. Int Biodeterior
841 Biodegradation 78:67–73.

842 111. Hastrup ACS, Green F, Lebow PK, Jensen B. 2012. Enzymatic oxalic acid regulation
843 correlated with wood degradation in four brown-rot fungi. Int Biodeterior
844 Biodegradation 75:109–114.

845 112. Mäkelä M, Galkin S, Hatakka A, Lundell T. 2002. Production of organic acids and
846 oxalate decarboxylase in lignin-degrading white rot fungi. Enzyme Microb Technol
847 30:542–549.

848 113. Takao S. 1965. Organic acid production by Basidiomycetes. I. Screening of acid-

849 producing strains. Appl Microbiol 13:732–737.

850 114. Baldrian P, Valásková V. 2008. Degradation of cellulose by basidiomycetous fungi.
851 FEMS Microbiol Rev 32:501–521.

852 115. Lu X, Heal KR, Ingalls AE, Doxey AC, Neufeld JD. 2020. Metagenomic and chemical
853 characterization of soil cobalamin production. ISME J 14:53–66.

854 116. Lauber CL, Hamady M, Knight R, Fierer N. 2009. Pyrosequencing-based assessment of
855 soil pH as a predictor of soil bacterial community structure at the continental scale. Appl
856 Environ Microbiol 75:5111–5120.

857 117. Carson JK, Gonzalez-Quiñones V, Murphy DV, Hinz C, Shaw JA, Gleeson DB. 2010.
858 Low pore connectivity increases bacterial diversity in soil. Appl Environ Microbiol
859 76:3936–3942.

860 118. He L, Xu X. 2021. Mapping soil microbial residence time at the global scale. Glob
861 Chang Biol 27:6484–6497.

862 119. Bueno de Mesquita CP, Wu D, Tringe SG. 2023. Methyl-Based Methanogenesis: an
863 Ecological and Genomic Review. Microbiol Mol Biol Rev 87:e0002422.

864 120. Machado A, Cerca N. 2015. Influence of Biofilm Formation by Gardnerella vaginalis
865 and Other Anaerobes on Bacterial Vaginosis. J Infect Dis 212:1856–1861.

866 121. Grosche A, Sekaran H, Pérez-Rodríguez I, Starovoytov V, Vetriani C. 2015. Cetia
867 pacifica gen. nov., sp. nov., a chemolithoautotrophic, thermophilic, nitrate-ammonifying
868 bacterium from a deep-sea hydrothermal vent. Int J Syst Evol Microbiol 65:1144–1150.

869 122. Shelobolina ES, Sullivan SA, O'Neill KR, Nevin KP, Lovley DR. 2004. Isolation,
870 characterization, and U(VI)-reducing potential of a facultatively anaerobic, acid-resistant

871 Bacterium from Low-pH, nitrate- and U(VI)-contaminated subsurface sediment and
872 description of Salmonella subterranea sp. nov. Appl Environ Microbiol 70:2959–2965.

873 123. Nichols CM, Bowman JP, Guezennec J. 2005. Olleya marilimosa gen. nov., sp. nov., an
874 exopolysaccharide-producing marine bacterium from the family Flavobacteriaceae,
875 isolated from the Southern Ocean. Int J Syst Evol Microbiol 55:1557–1561.

876 124. Robinson IM, Allison MJ, Hartman PA. 1975. Anaeroplasma abactoclasticum gen.nov.,
877 sp.nov.: an Obligately Anaerobic Mycoplasma from the Rumen. Int J Syst Bacteriol
878 25:173–181.

879 125. Stewart V. 1993. Nitrate regulation of anaerobic respiratory gene expression in
880 Escherichia coli. Mol Microbiol 9:425–434.

881 126. Hung C-H, Chang Y-T, Chang Y-J. 2011. Roles of microorganisms other than
882 Clostridium and Enterobacter in anaerobic fermentative biohydrogen production
883 systems--a review. Bioresour Technol 102:8437–8444.

884 127. Chung EJ, Park TS, Jeon CO, Chung YR. 2012. Chitinophaga oryziterrae sp. nov.,
885 isolated from the rhizosphere soil of rice (Oryza sativa L.). Int J Syst Evol Microbiol
886 62:3030–3035.

887 128. Jin D, Kong X, Wang J, Sun J, Yu X, Zhuang X, Deng Y, Bai Z. 2018. Chitinophaga
888 caeni sp. nov., isolated from activated sludge. Int J Syst Evol Microbiol 68:2209–2213.

889 129. Voigt C, Lamprecht RE, Marushchak ME, Lind SE, Novakovskiy A, Aurela M,
890 Martikainen PJ, Biasi C. 2017. Warming of subarctic tundra increases emissions of all
891 three important greenhouse gases - carbon dioxide, methane, and nitrous oxide. Glob
892 Chang Biol 23:3121–3138.

130. 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, Schuur EdwardAG, 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 Chang 6:595–600.

131. 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.

132. 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 USA 113:40–45.

133. 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.

134. 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.

135. 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. Environmental Research

Letters 17:024034.

136. Varsadiya M, Liebmann P, Petters S, Hugelius G, Urich 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.

137. 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.

138. 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.

139. 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, Sullivan MB. 2018. Host-linked soil viral ecology along a permafrost thaw gradient. Nat Microbiol 3:870–880.

140. 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.

141. Wu R, Trubl G, Taş N, Jansson JK. 2022. Permafrost as a potential pathogen reservoir. One Earth 5:351–360.

142. 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.

143. 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 USA 118.

144. Nickson R, McArthur J, Burgess W, Ahmed KM, Ravenscroft P, Rahmanñ M. 1998. Arsenic poisoning of Bangladesh groundwater. Nature 395:338–338.

145. 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.

146. 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.

147. Knelman JE, Nemergut DR. 2014. Changes in community assembly may shift the relationship between biodiversity and ecosystem function. Front Microbiol 5:424.

148. 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.

149. Storch D, Šizling AL. 2008. The Concept of Taxon Invariance in Ecology: Do Diversity Patterns Vary with Changes in Taxonomic Resolution? Folia Geobot 43:329–344.

150. Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JBH. 2012. Beyond biogeographic patterns: processes shaping the microbial landscape. Nat Rev Microbiol 10:497–506.

151. Martiny JBH, Jones SE, Lennon JT, Martiny AC. 2015. Microbiomes in light of traits: A phylogenetic perspective. Science 350:aac9323.

152. Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11:R25.

153. 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.

154. Terrestrial Ecoregions of the World | Publications | WWF.

155. Hengl T, Mendes de Jesus J, Heuvelink GBM, Ruiperez Gonzalez M, Kilibarda 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.

156. R Core Team. 2021. R: A language and environment for statistical computing. Computer software, R Foundation for Statistical Computing, Vienna, Austria.

157. Kuhn M. 2008. Building Predictive Models in *R* Using the caret Package. J Stat Softw 28.

158. 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.

159. Oksanen J, Blanchet FG, Kindt R, Legendre P. 2013. Package "vegan." Community ecology ….

160. Wood S, Wood MS. 2015. Package "mgcv." R package version.

161. Maechler M, Rousseeuw P, Struyf A, Hubert M. 2013. Package "cluster." Dosegljivo na.

162. Kodinariya TM, Makwana PR. 2013. Review on determining number of Cluster in K-Means Clustering. International Journal.

163. Wickham H, Chang W. 2016. Package “ggplot2.” Create elegant data ….

164. 2023. A Grammar of Data Manipulation [R package dplyr version 1.1.2].

165. Kassambara A, Mundt F. 2017. Package “factoextra.” Extract and visualize the results of multivariate ….

166. Jr FH, Jr MH. 2019. Package “hmisc.” CRAN2018.

167. 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 Softw 96.

168. Neuwirth E, Neuwirth ME. 2014. Package “RColorBrewer.” ColorBrewer Palettes.

169. Auguie B, Antonov A, Auguie MB. 2017. Package “gridExtra.” Miscellaneous Functions for “ ….

170. Wickham H, Wickham MH. 2017. Package tidyverse. Easily install and load the ’Tidyverse.

171. Becker RA, Wilks AR, Brownrigg R, Minka TP. maps: Draw geographical maps.

172. phyloT : a phylogenetic tree generator.

173. 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.

174. Wang L-G, Lam TT-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.

175. Campitelli E. 2020. ggnewscale: Multiple Fill and Colour Scales in'ggplot2'. R package version 04.

176. Kolde R, Kolde MR. 2018. Package "pheatmap." R Package.