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], Ryan McClure[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 assay 1512 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 (1) microbial nutrient acquisition across pH gradients; (2) stress, transport, and redox-based processes across changes in soil bulk density; and (3) 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 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 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 global climate change (7–9); and microbial nutrient cycling sustains 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 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 cycles) 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 their distributions and possible ecosystem impacts under future climate scenarios.

New analytical techniques allow for 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 global sequencing resources are essential for the next generation of research in microbial ecology (18, 61, 77, 78). Here, we examine a comprehensive set of 1512 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 composition 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 (Figure 1, Table S1 and Table S2). Bulk density, cation exchange capacity (CEC), nitrogen (N), pH, soil organic carbon (SOC), and clay content ranged from 0.28-1.56 kg/dm^3, 7.7-68.7 cmol C/kg, 0.25-22.4 g/kg, 4.4-8.3, 2.4-510.9 g/kg, and 2.7-57.1 % respectively.

To reveal the full range of taxonomies and functional potential encoded by global soils, we first assessed microbial attributes (i.e., gene annotations by Kyoto Encyclopedia of Genes and Genomes (KEGG)(79, 80), the Pfam (81) and TIGRFAM protein databases (82), and read-based taxonomy, see methods) with the highest normalized abundance across all soils (Figure 2). Multiple annotation databases provide complementary information and are standard operating procedure by the U.S. Department of Energy Joint Genome Institute (JGI)(83, 84). Abundant KEGG orthologies (KO) were associated bacterial secretion systems (K12510), redox reactions (K00104, K00334, K03882, K00380), and cellular growth processes (K02335, K02049, K00334, K03882, K03088). Abundant Pfam were comprised of functions related to the tolerance of common soil stressors including temperature (PF11999, PF08496) and salinity (PF12838), as

118 well as functions involved in microbial growth (PF08511, PF07357, PF13247, PF00005). The
119 most abundant TIGRFAM were also associated with soil stressors; for example, TIGR00229
120 (PAS Domain S-box protein) is involved in sensing changes in light, redox potential, oxygen,
121 and other stressors (85). Other highly abundant TIGRFAM across all soils included a sigma-
122 factor that may be associated with microbial responses to environmental fluctuation
123 (TIGR02937) and other signaling and sensing processes (TIGR00254, TIGR00231). Finally,
124 abundant microbial genera included diverse microorganisms including Burkholderia,
125 Mycobacterium, Streptomyces, Acetobacter, and Pseudonocardia.
126
127 *Variation in genomic potential across environmental gradients.*
128 To reveal biogeographic patterns in microbiomes across global soils, we extracted
129 microbial attributes characteristic of soil factors and ecological biomes using random forest
130 regression models (see methods, Table S1). Soil pH, bulk density, and biome type were the
131 strongest correlates of microbial attributes (i.e, they generated well-fitting models with
132 comparatively high explanatory power relative to other factors; Table S1). The cross-validated
133 random forest model for each of the variables above had a root mean square error (RMSE) of
134 less than 0.8 (often < 0.2) and/or at least 60% variance explained. Models for CEC, N, SOC, and
135 clay content did not meet these thresholds. Results were conceptually consistent across different
136 microbial attributes. A full description of all genomic fingerprints is in Extended Data File 1.
137 pH separated soil microbiomes by nutrient and vitamin B-related genes (Figure 3). More
138 alkaline soils were characterized by vitamin B-, nutrient-, and phosphatase-related activities, for
139 instance with genomic fingerprints that included K01662, PF01872, K08717, TIGR00378,
140 PF01966, and PF00481; as well as many genera of Alphaproteobacteria. More acidic soils were

141 signified by ammonia- and complex organic polymer- (e.g., wood-) related reactions, with
142 genomic fingerprints that included K14333, K01684, TIGR03404, and TIGR02093, and
143 microbial clades using ammonia as a substrate (Rhodopseudomonas and Klebsiella).
144 Bulk density separated soil microbiomes by factors associated with soil structure,
145 hydrology, and nutrient content. The genomic fingerprint of low bulk density soils comprised
146 genes involved in organic N and methane cycling, as well as signatures of anoxia (Figure 3). It
147 included K00992, K11261, K11959, TIGR03323, TIGR01392, and microorganisms associated
148 with anoxic and/or aquatic environments (Gardnerella, Cetia, Salmonella, Olleya, Mycoplasma,
149 Escherichia, Enterobacter, and Chitinophaga). Conversely, high bulk density soils were signified
150 by genes involved in microbial transport, stress, pigmentation, and infection mechanisms.
151 Notable attributes associated with high bulk density soils included K00854, K11733, K03799,
152 TIGR01263, TIGR01957, TIGR01203, PF00582, PF03631, and four Alphaproteobacteria.
153 Four biomes were highlighted with distinct genomic signatures by our models – (1)
154 flooded grasslands & savannas, (2) tundra, (3) boreal forest, and (4) montane
155 grasslands/shrublands (Figure 4). The soil genomic fingerprint of flooded grasslands/savannas
156 was primarily characterized by positive associations with organic N (K13497, K07395, K19702,
157 K02042), phosphorus (P, K02042, K00854, K02800, PF04273), stress tolerance (heat and
158 arsenic: TIGR04282, TIGR02348), and motility (PF00669, PF00482, Paeniclostridium). The
159 genomic fingerprint for tundra had the greatest number positive gene associations, including
160 those associated with methanogenesis (K11261, TIGR03323, TIGR03321, TIGR03322,
161 TIGR03306, TIGR01392, TIGR02145), nitric oxide production (PF08768), P transfer (K03525,
162 K00997, K00992, K13497, PF08009), and pathogenesis (PF11203, PF05045); supported by the

presence of anaerobic (Gardnerella, Escherichia, Kosakonia, Glaesserella, Mycoplasma, Treponema, 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, PF04685); as well as some metal-related genes (TIGR00378, TIGR04282, TIGR00003). Montane systems were negatively characterized by genes related to Mg and P (K01520, K04765, TIGR01203), carbohydrate metabolism (K01816), virulence (PF09299), and motility (PF00669), as well as organic N (K07395, K00600, 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 biome type and environmental factors (Figure 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 content, total

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

190 Selected microbial KO and genera were positively correlated with SOC, N, and CEC and
191 negatively associated with bulk density and clay; while Pfam cluster 1 exhibited opposite
192 patterns (i.e., negative associations with SOC, N, and CEC; and positive associations with bulk
193 density and clay). Soil microbiomes in KO cluster 7 and genus cluster 6 contained high
194 normalized abundances of genes associated with energy generation (K00334, K00339, K00324,
195 K08738), filamentous bacteria/biofilms (K11903, Nocardia, Streptomyces, Mycobacterium,
196 Leptolyngbya), chemically complex organic matter decomposers (Nocardia, Rhodococcus,
197 Streptomyces) and N-fixing organisms and plant symbionts (Rhizobium, Bradyrhizobium,
198 Nostoc). Additionally, genes associated with N-cycling processes (K07685, K10041) and
199 microbial interactions (e.g., signaling and secretion, K12537, K02661) changed most
200 dramatically across environmental gradients. The most variable taxa across environmental
201 gradients were diverse— they included N cycling (Uliginosibacterium, Microvirgula),
202 autotrophic and/or anaerobic (Methanosalsum, Salana, Pseudobacteroides, Sulfurirhabdus,
203 Microvirgula, Desulfallas, Pelotomaculum, Risungbinella), and plant-related microorganisms
204 (Pseudobacteroides, Zeaxanthinibacter, Parapedobacter). Samples in Pfam cluster 1 had high
205 relative abundances of organic N- (PF01979, PF08352, PF01842), P- (PF00406, PF01380), and
206 signaling-related (PF01627, PF09413) genes, while many genes thought to be associated with
207 eukaryotes changed most dramatically across environmental gradients in Pfam annotations
208 (PF04178, PF03188, PF01793).

209

210 **Discussion.**

211 *Global soil microbiome taxa and functional potential..*

212 Across all soils, the most abundant microbial attributes spanned disparate lineages and
213 potential functions, highlighting the high microbial diversity of soils relative to other global
214 habitats (5, 18, 45, 46). While we note that we assayed microbial genes which do not necessarily
215 reflect expressed functions (61, 88–91), the genomic potential revealed by this investigation
216 provides a basis for more targeted research into soil microbiome function across different soil
217 habitats. The most prevalent microbial genera included members of commonly observed soil
218 taxa including Alphaproteobacteria, Betaproteobacteria, and Actinobacteria (45). Widespread
219 genomic functions included growth, resource acquisition, and stress tolerance strategies that are
220 essential for microbial persistence in heterogeneous soil environments. Bacterial secretion
221 systems were common, as well as DNA repair/replication, transport, and energy generation
222 functions. Interestingly, genes for the sensing and/or tolerance of fluctuations in temperature,
223 salinity, light, and redox conditions were among the most abundant soil microbial genes; all of
224 which are common soil microbial stressors (86, 87). Collectively, the diverse taxa and functions
225 encoded by the soil microbiome underscores the heterogeneous stressors on the soil microbiome.
226 The high global soil microbial diversity observed here can obscure patterns in microbiome
227 structure and function, indicating the use of machine learning-based tools to unravel its
228 complexity (18, 45, 88).

229

230 *Soil genomic potential corresponds to pH, bulk density, and biome type.*

231 Soil genomic patterns were evident across soil pH, bulk density, and biomes; however,
232 some hypothesized drivers of the soil microbiome including SOC, total N, and CEC were not
233 statistically linked to microbial taxa or genomic functional potential. Soil carbon, N, and cation
234 exchange capacity can reflect resource availability and are often associated with soil microbiome
235 change at local scales (14, 36, 89–91). The lack of trends at the global scales may represent
236 heterogeneity in the importance of resources in structuring microbiomes; indeed, the influence of
237 deterministic (e.g., selection) versus stochastic (e.g., random) assembly processes on soil
238 microbiomes varies through space and time (30, 92–95). While our results support pH and biome
239 type as primary determinants of soil microbial structure and function (5, 6, 28, 45), it is
240 interesting that bulk density was also a strong correlate of genomic patterns. This may reflect the
241 importance of biophysical and hydraulic properties on soil microbiome structure and function
242 (92, 96–98).

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

250

251 *Microbial nutrient acquisition potential shifts across pH gradients.*

252 The genomic fingerprint of alkaline soils demonstrated an association with microbial
253 nutrient acquisition and is important for understanding belowground microbial dynamics in arid

254 and semiarid ecosystems (>35% of global land surface (99)). The soil matrix in these systems is
255 porous, has low rates of water retention, and contains an abundance of calcium carbonate that
256 limits P bioavailability (100). Accordingly, the alkaline soil fingerprint was enriched in genes
257 encoding phosphatases and related functions (PF01966, PF00481) that may signify the
258 importance of microbial P acquisition and/or limitation in alkaline soils (101–104). Vitamin B-
259 related genes were also included in this fingerprint and have also been linked to P solubilization
260 (K01662, PF01872)(105). The presence of a urea transporter (K08717) may likewise indicate the
261 importance of nutrient acquisition in alkaline soils (106). Notably, the high-Ca^{2+} alkaline soil
262 environment was also characterized by a calcium/proton exchanger (TIGR00378)—which
263 regulates intracellular Ca^{2+} concentrations to prevent Ca^{2+} overload (107) and an uncharacterized
264 hydrophobic domain (TIGR00271) that may arise from low moisture environments that are often
265 characteristic of alkaline soils.

266 In contrast, more acidic soils were characterized by a distinct set of nutrient acquisition
267 genes, in particular genes encoding ammonia and wood decomposition reactions that produce H^+
268 as a byproduct. For instance, nitrification is a hallmark of microbial genera Rhodopseudomonas
269 and Klebsiella (108–110) and increases soil acidity through the release of H^+ from ammonia.
270 Likewise, wood-degrading fungi often secrete acidic compounds to aid in the digestion of
271 chemically-complex polymers (111–114). This is consistent with the association of K14333,
272 K01684, TIGR03404, and TIGR02093 with low pH soils.

273 Finally, Alphaproteobacteria have been previously shown to be sensitive to variation in
274 soil pH (118, 119) and were included in the genomic fingerprints of both high and low pH soils,
275 further bolstering their potential sensitivity to global change processes that impact soil pH (e.g.,
276 atmospheric nitrogen deposition).

277

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

279 Bulk density reflects biophysical soil properties such as texture, water content, porosity,
280 and mineralogy that drive microbial community structure and function (19, 53, 54, 117, 118).
281 Patterns in the relationships between these variables and the soil microbiome can be complex at
282 global scales (19), and our findings are to our knowledge, the first to report a global-scale
283 relationship between bulk density and the distribution of genes encoded by the soil microbiome.

284 Low bulk density soils were characterized by genes common to water-logged and/or
285 agricultural soils, while high bulk density soils included potential functions associated with
286 transport mechanisms, stress tolerance, and virulence. For instance, K11261, TIGR03323, and
287 TIGR01392 – all signatures of low bulk density soil – are each associated with methanogenic
288 processes that are prevalent in saturated soils (119). Low bulk density soils were further
289 characterized by anaerobic and/or aquatic microorganisms including Gardnerella, Cetia,
290 Salmonella, Olleya, Mycoplasma, Escherichia, Enterobacter, and Chitinophaga that also supports
291 microbial persistence in water-saturated environments (120–128). Though we did not directly
292 assess land cover, the presence of potential functions associated with organic N cycling in the
293 low bulk density fingerprint (K00992, K11959) may additionally suggest the influence of
294 agriculture on soil microorganisms, where tillage generates loosely packed soils (i.e., low bulk
295 density) and fertilizer application enhances nutrient cycles.

296 In contrast, the genomic fingerprint of high bulk density soils suggests an environment
297 inclusive of density-dependent relationships between microorganisms, their substrates, and
298 pathogens (transport: K00854, K11733, TIGR01957, TIGR01203; virulence: PF03631,
299 TIGR01203). Stress-related genes (K03799, PF00582.24) may be reflective of this competitive

300 environment; or related to ecosystem processes such as soil compaction which is common in
301 deserts (K03799).
302
303 *Soil genomic potential for greenhouse gas emissions differs by biome.*
304 Finally, soil habitat (reflected here by biome type) seemed to be a driver of dissimilarity
305 in microbiome structure and potential function, providing detailed information on the molecular
306 biology of global soil ecosystems.
307 We first highlight microbial genomic attributes of tundra ecosystems, as they are
308 experiencing rapid transformations in response to global climate change (129, 130). Tundra
309 exhibited the highest number of microbial taxa and potential functions in their genomic
310 fingerprint, the vast majority of which displayed positive associations. Given the relationship
311 between soil microbial diversity and biogeochemical function, this provides a consistent and
312 distributed genomic basis for enhanced greenhouse gas emissions that has been associated with
313 permafrost thaw in tundra (129, 131, 132). The microbial attributes associated with tundra in this
314 study reveal clues into the biology that drives active biogeochemical cycles in these vulnerable
315 ecosystems. Methanogenesis in particular is responsible for fluxes of methane in warming tundra
316 (131, 133, 134) and was associated with tundra in the current study. Additionally, a gene
317 involved in nitric oxide cycling (PF08768) (135–137), as well as a variety of anaerobic
318 organisms and P-related genes (K03525, K00997, K00992, K13497, PF08009) (138–140), also
319 provide insight into the molecular functions of tundra in the context of global climate change.
320 Interestingly, we also found evidence for pathogen resistance within the tundra genomic
321 signature (PF11203, PF05045), lending credence to a small number of recent studies suggesting

322 permafrost environments as one of the largest reservoirs of soil viruses (141–144) and a possible
323 linkage between soil methane and viral infection (145).

324 Flooded grasslands/savannas were also generally positively associated with many
325 microbial attributes, several of which encoded nutrient cycling, stress tolerance, and/or motility
326 traits. Genes related to organic N cycling (K13497, K07395, K19702, K02042), P (K02042,
327 K00854, K02800, PF04273) and/or stress tolerance (TIGR04282, TIGR02348) may signify
328 microbial adaptations to nutrient scarcity and other stressors. Motility (PF00669, PF00482,
329 Paeniclostridium) may be reflective of beneficial microbial lifestyles in saturated systems, where
330 movement is facilitated by high pore space connectivity. Interestingly, a mechanism for arsenic
331 resistance was affiliated with flooded grasslands and savannas, possibly denoting arsenic
332 groundwater contamination in many regions of the world (146).

333 In contrast, boreal forest and montane grasslands/shrublands were signified by negative
334 associations with Mg- and P-related genes (boreal: K03284, K06217, montane: K01520,
335 K04765, TIGR01203) and with organic N and carbohydrate-related genes (boreal: K02800,
336 K11472, PF00208, PF04685; montane: K01816, K07395, K00600, PF00208). In addition,
337 anaerobic microbial genera were negatively associated with these ecosystems. The boreal forest
338 fingerprint uniquely included some metal-related genes (TIGR00378, TIGR04282, TIGR00003);
339 while montane systems were uniquely characterized by virulence (PF09299), motility
340 (PF00669), and sulfur-related genes (K03644). The negative associations with carbon, N, and P
341 suggest that carbon and inorganic nutrients may be more available in these systems relative to
342 other biomes; and Mg is often associated with plant growth (147). Taken together, microbial
343 fingerprints of boreal forests and montane grasslands/shrublands convey slower growth rates
344 and fewer resource constraints than other ecosystems.

345

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

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

353 Unsupervised machine learning uncovered soil microbial genomic attributes associated
354 with coordinated changes in SOC, total N, and CEC and oppositely with clay content and bulk
355 density that may represent some of the most biologically-active soils on Earth (Figure 5). Soils
356 rich in SOC and N with high CEC constitute resource- and energy-rich environments, supported
357 by an abundance of energy generating genes in the fingerprint of these soils (K00334, K00339,
358 K00324, K08738). Interestingly, these environments are best explained by a combination of
359 environmental variables rather than an individual factor. As an example, soils with high SOC
360 alone could be limited by N or other nutrients; thus, we propose that genomic fingerprints for
361 multiple simultaneously interacting factors may be among the most critical for understanding
362 global biogeochemistry. They are nearly impossible to discern without unsupervised statistical
363 approaches, and we know of no other work that has done so.

364 Soils with high SOC, N, and CEC were characterized by genes associated with resource
365 scavenging, plant-microbe interactions, and heterotrophic metabolisms. We observed numerous
366 attributes in their genomic fingerprint associated with filaments or biofilms, intra-organism
367 signaling, and/or chemically-complex organic matter decomposition (K11903, K12537,

368 K02661, Nocardia, Streptomyces, Mycobacterium, Leptolyngbya, Nocardia, Rhodococcus,
369 Streptomyces). This is consistent with microbial morphologies and lifestyles that support the
370 decay of chemically recalcitrant polymers (e.g., wood), as well as nutrient acquisition and
371 transport across soil pore networks. Plant-microbe interactions, including symbiotic N-fixing
372 organisms, were also a key feature of these soils, highlighting the importance of above- and
373 belowground connectivity in productivity (Rhizobium, Bradyrhizobium, Nostoc,
374 Pseudobacteroides, Zeaxanthinibacter, Parapedobacter, Zea)(148). In contrast, anaerobic and
375 autotrophic organisms were highly sensitive to changes across these connected environmental
376 gradients, possibly reflecting their exclusion from more oxygen- and energy-rich environments
377 where heterotrophic organisms tend to dominate.

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

387

388 *Phylogenetic coherence in microbial biogeography.*

389 The extent to which selection acts on phylogenetically-conserved properties of soil
390 microbiomes and further translates into differences in functional potential (e.g., response-effect

391 traits (149)) is a long running unknown in microbial ecology. Some previous studies have shown
392 little connection between functional attributes and microbial phylogeny (44, 150), while others
393 have indicated the existence of taxa-function relationships (88). And still others have suggested
394 variation in the coupling of taxonomy and function across various functions of interest, levels of
395 phylogenetic resolution, and/or spatial scales (34, 55, 61, 151, 152). Though microbial taxa were
396 generally less correlated to the environment than functional potential (Table S1), we found
397 strong phylogenetic coherence in the response of genera to changes in the soil environment
398 (Figure 4B). That is, closely related organisms tended to respond similarly to environmental
399 change. This implies that evolutionary history, life history strategies, and/or morphology (all
400 traits that tend to be phylogenetically conserved (153)) could influence changes in microbial
401 community membership in response to environmental change. As such, metabarcoding
402 approaches that yield microbiome taxonomy may be among the first indicators of ecosystem
403 transitions, such as response to disturbances and transitions between state archetypes (30).
404
405 *Conclusions.*
406 Despite the significance of soil microorganisms in global biogeochemistry, we still know
407 little about the ecological processes that regulate their community composition and function
408 across the wide range of global soil environments. We used machine learning to sift through tens
409 of thousands of taxa and potential functions encoded by 1512 global soil microbiomes, revealing
410 novel biogeographical patterns in soil microbiome composition and functional potential with
411 high molecular resolution. Specifically, we demonstrate shifts in the potential for (1) microbial
412 nutrient acquisition across pH gradients; (2) stress, transport, and redox-based processes across
413 changes in soil bulk density; and (3) genes and organisms associated with soil greenhouse gas

414 emissions across biomes. We also uncover a suite of metabolic processes that are enriched in the
415 microbial genomes of energy-rich soils. These changes were coincident with phylogenetically
416 congruent compositional shifts – suggesting that closely related soil microbial taxa are sensitive
417 to similar environmental stressors. Our work enables us to better understand microbial lifestyles
418 that support the persistence of certain members over others along environmental gradients; such
419 knowledge underlies a predictive understanding of soil microbiome changes with shifts in the
420 global climate and is vital to constraining biochemical reactions that regulate the release of
421 greenhouse gasses from soils.
422
423 **Materials and Methods.**
424
425 *Dataset Description and Normalization.*
426 We collected 1512 manually-curated soil metagenomic samples available in the Joint
427 Genome Institute's (JGI) Integrated Microbial Genomes and Microbiomes (IMG/M) platform in
428 August 2020 as described by Graham et al. 2024 (accepted), of which 1493 were associated with
429 latitudinal and longitudinal information. Samples span most major biomes and continents (Figure
430 1; Table S2 and Table S3). Sequences were quality controlled, assembled, and annotated by
431 KEGG Orthology (KO), the Pfam and TIGRFAM protein family databases, and read-based
432 taxonomy by the JGI's standard workflows (80–84). While sample collection and sequencing
433 methods inevitably vary in meta-analyses such as the current study, we applied standardized
434 analytical workflows to minimize methodological artifacts to the greatest extent possible. Due to
435 the collective size of assembled sequences, we focused on annotations and read-based taxonomy

436 for downstream analysis, which is more computationally feasible than sequence assemblies or
437 metagenome-assembled genomes that are impractical at the scale of thousands of samples.
438 Each dataset (i.e., KO, Pfam, TIGRFAM, and read-based taxonomy) was independently
439 normalized by the following procedure prior to analysis. For Pfam and KO annotations only,
440 samples with greater than 2000 total reads were retained. For TIGRFAM annotations only,
441 samples with greater than 500 total reads were retained. A lower cutoff was used for TIGRFAM
442 data as the total read counts were lower in this dataset. For all three annotations, only functions
443 with a non-zero value in at least half of the samples or a total count of 7560 (corresponding to an
444 average of 5 reads per sample for each function) were retained. For species, we retained only
445 samples with greater than 4000 counts and species with a non-zero value in at least half of the
446 samples or a total count of 7560 (corresponding to an average of 5 reads per sample for each
447 species). After removing low abundant samples and functions/species, all four datasets were then
448 normalized using the trimmed mean of M values (TMM) method (154). We found that this
449 method gave us the best results when examining normalized data using a box and whisker plot
450 (Figure S1).
451 We assigned biome type and soil properties to each sample using publicly available
452 resources. We used latitude and longitude to derive using Olson biomes (155) for each sample
453 using data provided by the World Wildlife Federation (156). Soil parameters were collected from
454 the SoilGrids250m database from 0-5 cm (157). SoilGrids250m is a spatial interpolation of
455 global soil parameters using ~150,000 soil samples and 158 remote sensing-based products.
456 Here, we focus on six parameters often associated with soil microbiomes: bulk density, cation
457 exchange capacity, nitrogen, pH, soil organic C, and clay content. Because our focus on spatial
458 dynamics and soils were collected at various times, we did not include temporally dynamic

459 variables such as soil moisture or temperature in our set of environmental parameters, though we
460 acknowledge they may have profound impacts on the soil microbiome. Additionally, because
461 soil data are derived from a comprehensive spatial interpolation and not measured directly on
462 each soil sample, our models may fail to capture local-scale heterogeneity in relationships.
463 Further investigations that generate and leverage fully standardized data are needed to resolve
464 this discrepancy. Nonetheless, local heterogeneity should statistically weaken the relationships
465 observed here, and our results present a promising investigation into the global biogeography of
466 soil microbiomes.
467
468 *Machine Learning for Genomic Fingerprints.*
469 We selected random forests as our primary statistical approach by assessing the ability of
470 various machine learning algorithms to adequately parse soil microbial attributes across
471 environmental variables and ecological biomes. Models were built and validated using 10-fold
472 cross validation repeated five times, 50% of the data for training, and 50% of the data for testing.
473 Models from different algorithms were compared using cross validation (CV) scores and
474 variance explained (Figure S2). Random forest models consistently yielded comparatively high
475 CV scores and proportion of variance explained.
476 Thus, we constructed random forest models to extract genomic fingerprints from soil
477 microbiomes using two approaches (1) a supervised split across variables known to be associated
478 with soil microbiomes and (2) an unsupervised approach that allowed predictors and their
479 interactions to naturally arise. Random forest models were built in R Software (158) for each
480 attribute type independently using the 'caret' package (159).

481 In (1), we divided samples into either two groups containing the highest and lowest 10%
482 of values for each environmental factor or one group per biome. We then used random forests to
483 rank each microbial attribute according to their importance (i.e., difference in mean square error
484 normalized by standard error, varImp function in 'caret' package) to differentiate samples across
485 each environmental factor or biome. Subsequently, because variable importances followed an
486 exponential decay, we used breakpoint analysis to define a set of the most important variables
487 distinguishing each group ('strucchange' package, h = 0.01 with a minimum segment size of
488 two) (160). Because this resulted in one set of attributes that drove variation in soil microbiomes
489 in relation to a given environmental factor or biome, we extracted attributes associated with high
490 (positive slope) or low (negative slope) values of each environmental factor using linear
491 regression. The resultant attributes comprised two genomic fingerprints per environmental factor
492 or biome—one associated with low values and one associated with high values.
493 In (2), we first used k-means clustering to group samples with similar microbial attributes
494 using R packages: 'vegan' (161), 'mgcv' (162), and 'cluster' (163). The optimal number of
495 clusters was chosen by evaluating the total sum of squares and the silhouette index (164). To
496 determine the extent to which soil microbiomes within each cluster corresponded to changes in
497 environmental factors, we related the total normalized abundance of attributes per sample to each
498 environmental factor using linear regression. We then selected one cluster per attribute type for
499 deeper investigation, chosen by the number of significant correlations with environmental
500 factors. To generate a genomic fingerprint for the selected cluster, we extracted attributes that
501 were either most abundant or most variable within the cluster. As in approach (1), we used
502 breakpoint analysis to define a set of the most abundant attributes per cluster ('strucchange'
503 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' (165), 'dplyr' (166), 'factoextra' (167), 'Hmisc' (168), 'colorspace' (169), 'RColorBrewer' (170), 'gridExtra' (171), 'tidyverse' (172), and 'maps' (173). Additionally, for visualizations of read-based taxonomy, we generated phylogenetic trees using phyloT v2 (174), an online tree generator based on the Genome Taxonomy Database. Then, we visualized the tree in R using the packages 'ggtree'(175), 'treeio'(176), and 'ggnewscale'(177). We generated heatmaps using the 'pheatmap' package (178).

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. Please see Tables S2 and S3 for further information. Code for data analysis and all figures is available at: 10.6084/m9.figshare.25374427.

526 **Figures.**

527 **Figure 1. Dataset description. (**A) Map of samples, (B) total assembled bp, and (C) number of
528 samples; all colored by biome according to Olson et al. (155). (D-I) distribution of environmental
529 variables from SoilGrids250m (157).

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

533 **Figure 3. Genomic fingerprints of high- and low-pH and -bulk density soil environments.**
534 Genomic attributes selected in at least one fingerprint for pH or bulk density are visualized.
535 Pearson correlation coefficient is shown from blue to red in the primary heatmap, with bulk
536 density correlations on the left and pH correlations on the right. The sidebars (respectively, from
537 left to right) represent variable importance from random forest models of pH and bulk density
538 and the type of attribute (e.g., KO, Pfam, TIGRFAM). Please refer to Extended Data 1 for more
539 information on the specific attributes associated with each soil environment. Read-based
540 taxonomy collated at the genus-level for each fingerprint is shown in Figure 4.

541 **Figure 4. Genomic fingerprints of soil biomes and phylogenetic distributions of selected soil
542 microbiomes.** (A) The normalized abundance of microbial attributes selected in at least one
543 biome fingerprint is shown from blue to red. Variable importance is shown in the left side bar
544 and attribute type is shown in the right side bar. Please refer to Extended Data 1 for more
545 information on the specific attributes associated with each soil environment. (B) Microbial
546 genuses selected in at least one fingerprint are depicted on a phylogenetic tree (generated by
547 PhyloT). The innermost circle shows Pearson correlations with bulk density from green to purple
548 and variable importance from green to orange. The middle circle shows Pearson correlations

549 with pH from green to purple and variable importance from green to orange. The outer circle
550 shows the normalized abundance of genera across biomes (from outer to inner: Temperate
551 Broadleaf & Mixed Forests; Tundra; Mediterranean Forests, Woodlands & Scrub; Tropical &
552 Subtropical Grasslands, Savannas & Shrublands; Temperate Grasslands, Savannas &
553 Shrublands; Boreal Forests/Taigas; Deserts & Xeric Shrublands; Temperate Conifer Forests;
554 Tropical & Subtropical Dry Broadleaf Forests; Flooded Grasslands & Savannas; Tropical &
555 Subtropical Moist Broadleaf Forests; and Montane Grasslands & Shrublands). Variable
556 importance is shown in the outermost ring from green to orange.s
557 **Figure 5. Unsupervised clusters of microbial attributes associated with multiple soil factors.**
558 Column (A) KEGG Orthology, (B) microbial genus, and (C) Pfam. TIGRFAM annotations did
559 not yield a satisfactory model. Rows represent environmental variables. The total normalized
560 abundance of an attribute per sample is plotted against each environmental variable. Lines and
561 statistics represent linear regression.

562 **References**

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

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

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

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

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

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

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

581 8. Bradford MA, Wieder WR, Bonan GB, Fierer N, Raymond PA, Crowther TW. 2016.
582 Managing uncertainty in soil carbon feedbacks to climate change. Nat Clim Chang
583 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.

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

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

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

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

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

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

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

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

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

672 Prévost-Bouré NC, Toutain B, Arrouays D, Lemanceau P, Ranjard L. 2009.
673 Biogeographical patterns of soil bacterial communities. Environ Microbiol Rep 1:251–
674 255.

675 43. Fuhrman JA, Steele JA, Hewson I, Schwalbach MS, Brown MV, Green JL, Brown JH.
676 2008. A latitudinal diversity gradient in planktonic marine bacteria. Proc Natl Acad Sci
677 USA 105:7774–7778.

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

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

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

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

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

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

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

694 2017. Habitat-specific patterns and drivers of bacterial β-diversity in China's drylands.
695 ISME J 11:1345–1358.

696 51. Bell T. 2010. Experimental tests of the bacterial distance-decay relationship. ISME J
697 4:1357–1365.

698 52. Tedersoo L, Bahram M, Põlme S, Kõljalg U, Yorou NS, Wijesundera R, Villarreal Ruiz
699 L, Vasco-Palacios AM, Thu PQ, Suija A, Smith ME, Sharp C, Saluveer E, Saitta A,
700 Rosas M, Riit T, Ratkowsky D, Pritsch K, Põldmaa K, Piepenbring M, Abarenkov K.
701 2014. Fungal biogeography. Global diversity and geography of soil fungi. Science
702 346:1256688.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Pedobiologia 49:637–644.

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.

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

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.

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.

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.

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.

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

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

760 taxonomic resolution. mSystems 5.

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

763 77. Bowman MM, Heath AE, Varga T, Battu AK, Chu RK, Toyoda J, Cheeke TE, Porter
764 SS, Moffett KB, LeTendre B, Qafoku O, Bargar JR, Mans DM, Hess NJ, Graham EB.
765 2023. One thousand soils for molecular understanding of belowground carbon cycling.
766 Front Soil Sci 3.

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

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

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

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

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

780 83. Huntemann M, Ivanova NN, Mavromatis K, Tripp HJ, Paez-Espino D, Palaniappan K,
781 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

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

827 97. McCarter CPR, Rezanezhad F, Quinton WL, Gharedaghloo B, Lennartz B, Price J,
828 Connon R, Van Cappellen P. 2020. Pore-scale controls on hydrological and geochemical
829 processes in peat: Implications on interacting processes. Earth-Science Reviews
830 207:103227.

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

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

834 100. Hagin J, Hadas A. 1962. Solubility of calcium phosphate in calcareous soils. Nature
835 193:1211–1212.

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

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

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

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

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

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

850 107. Guerini D. 1998. The Ca2+ pumps and the Na+/Ca2+ exchangers. Biometals 11:319–
851 330.

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

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

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

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

865 112. Mäkelä M, Galkin S, Hatakka A, Lundell T. 2002. Production of organic acids and
866 oxalate decarboxylase in lignin-degrading white rot fungi. Enzyme Microb Technol
867 30:542–549.

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

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

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

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

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

877 117. Carson JK, Gonzalez-Quiñones V, Murphy DV, Hinz C, Shaw JA, Gleeson DB. 2010.
878 Low pore connectivity increases bacterial diversity in soil. Appl Environ Microbiol
879 76:3936–3942.

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

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

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

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

889 122. Shelobolina ES, Sullivan SA, O'Neill KR, Nevin KP, Lovley DR. 2004. Isolation,
890 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.

123. Nichols CM, Bowman JP, Guezennec J. 2005. Olleya marilimosa 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.

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

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

126. 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.

127. 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.

128. 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.

129. 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.

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. Christensen TR. 1993. Methane emission from Arctic tundra. Biogeochemistry 21:117–139.

134. Christensen TR, Cox P. 1995. Response of methane emission from Arctic tundra to climatic change: results from a model simulation. Tellus B: Chemical and Physical ….

135. 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.

136. 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.

137. 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.

138. 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.

139. 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.

140. 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.

141. 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.

142. 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.

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

957 144. Rigou S, Santini S, Abergel C, Claverie J-M, Legendre M. 2022. Past and present giant
958 viruses diversity explored through permafrost metagenomics. Nat Commun 13:5853.

959 145. Lee S, Sieradzki ET, Nicolas AM, Walker RL, Firestone MK, Hazard C, Nicol GW.
960 2021. Methane-derived carbon flows into host-virus networks at different trophic levels
961 in soil. Proc Natl Acad Sci USA 118.

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

964 147. Gransee A, Führs H. 2013. Magnesium mobility in soils as a challenge for soil and plant
965 analysis, magnesium fertilization and root uptake under adverse growth conditions. Plant
966 Soil 368:5–21.

967 148. Knelman JE, Schmidt SK, Graham EB. 2021. Cyanobacteria in early soil development
968 of deglaciated forefields: Dominance of non-heterocytous filamentous cyanobacteria and
969 phosphorus limitation of N-fixing Nostocales. Soil Biol Biochem 154:108127.

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

972 150. Rocca JD, Hall EK, Lennon JT, Evans SE, Waldrop MP, Cotner JB, Nemergut DR,
973 Graham EB, Wallenstein MD. 2015. Relationships between protein-encoding gene
974 abundance and corresponding process are commonly assumed yet rarely observed.
975 ISME J 9:1693–1699.

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

978 152. Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JBH. 2012. Beyond

biogeographic patterns: processes shaping the microbial landscape. Nat Rev Microbiol 10:497–506.

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

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

155. 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.

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

157. 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.

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

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

160. 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.

161. Oksanen J, Blanchet FG, Kindt R, Legendre P. 2013. Package “vegan.” Community ecology ….

162. Wood S, Wood MS. 2015. Package “mgcv.” R package version.

163. Maechler M, Rousseeuw P, Struyf A, Hubert M. 2013. Package “cluster.” Dosegljivo na.

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

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

166. 2023. A Grammar of Data Manipulation [R package dplyr version 1.1.2].

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

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

169. 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.

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

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

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

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

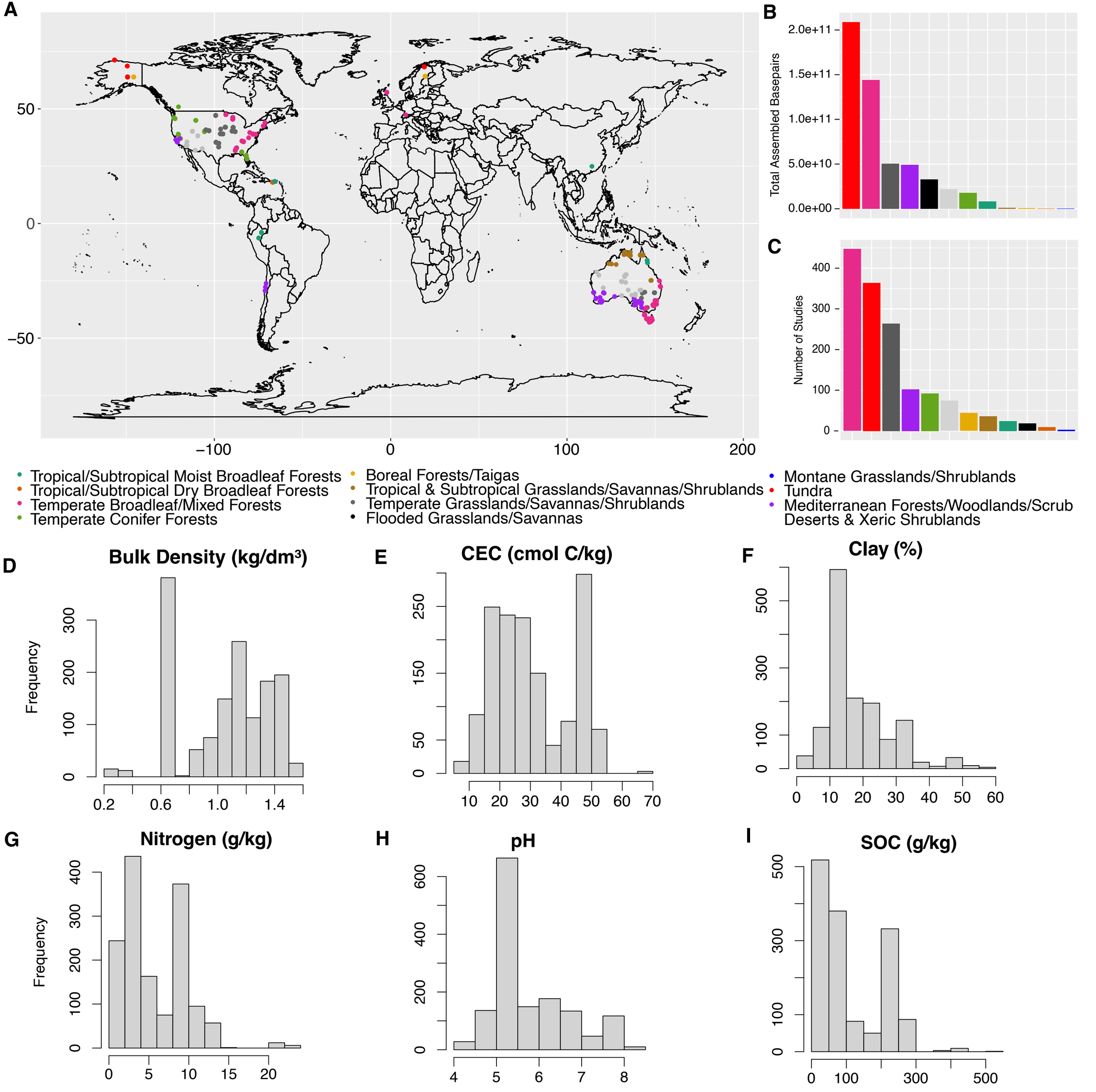
174. phyloT : a phylogenetic tree generator.

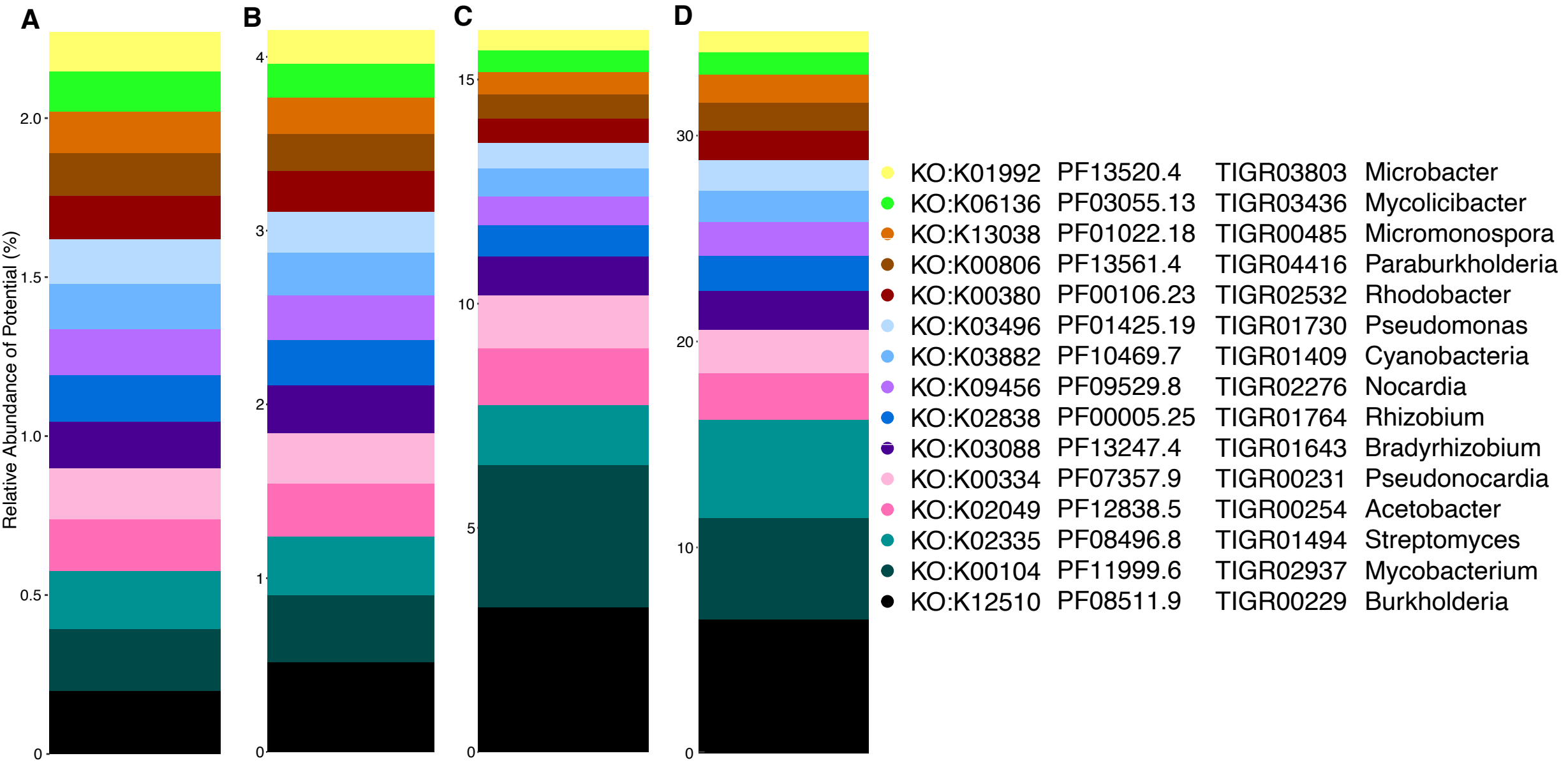
1022 175. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2016. ggtree: An R package for
1023 visualization and annotation of phylogenetic trees with their covariates and other
1024 associated data. Methods Ecol Evol 8:28–36.

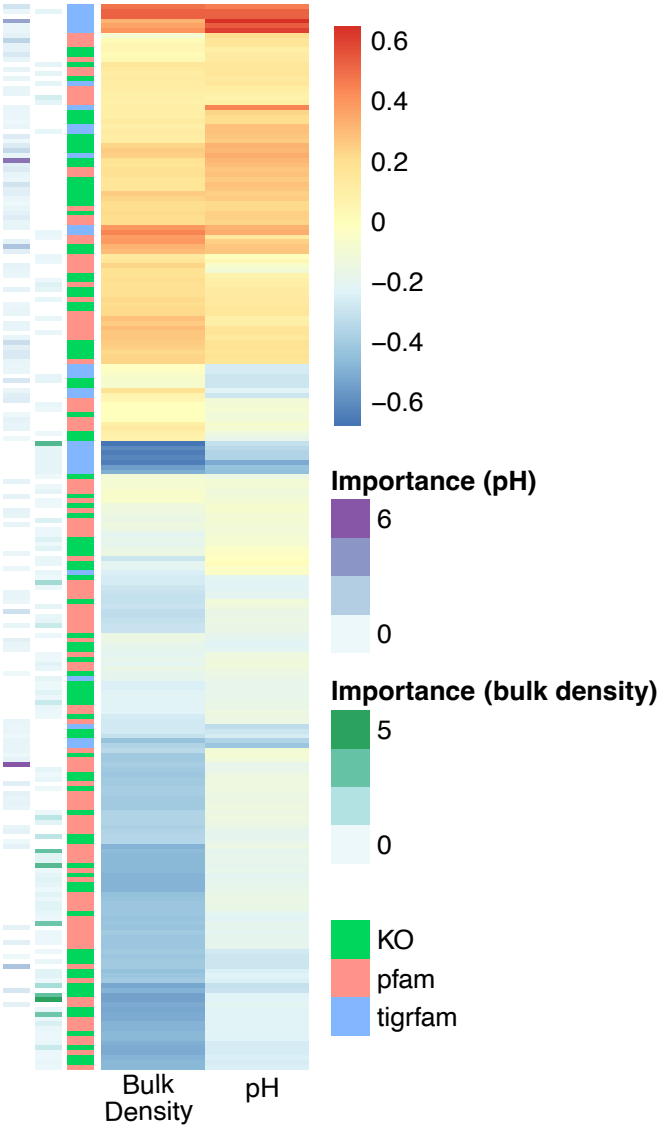
1025 176. Wang L-G, Lam TT-Y, Xu S, Dai Z, Zhou L, Feng T, Guo P, Dunn CW, Jones BR,
1026 Bradley T, Zhu H, Guan Y, Jiang Y, Yu G. 2020. Treeio: An R Package for
1027 Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. Mol
1028 Biol Evol 37:599–603.

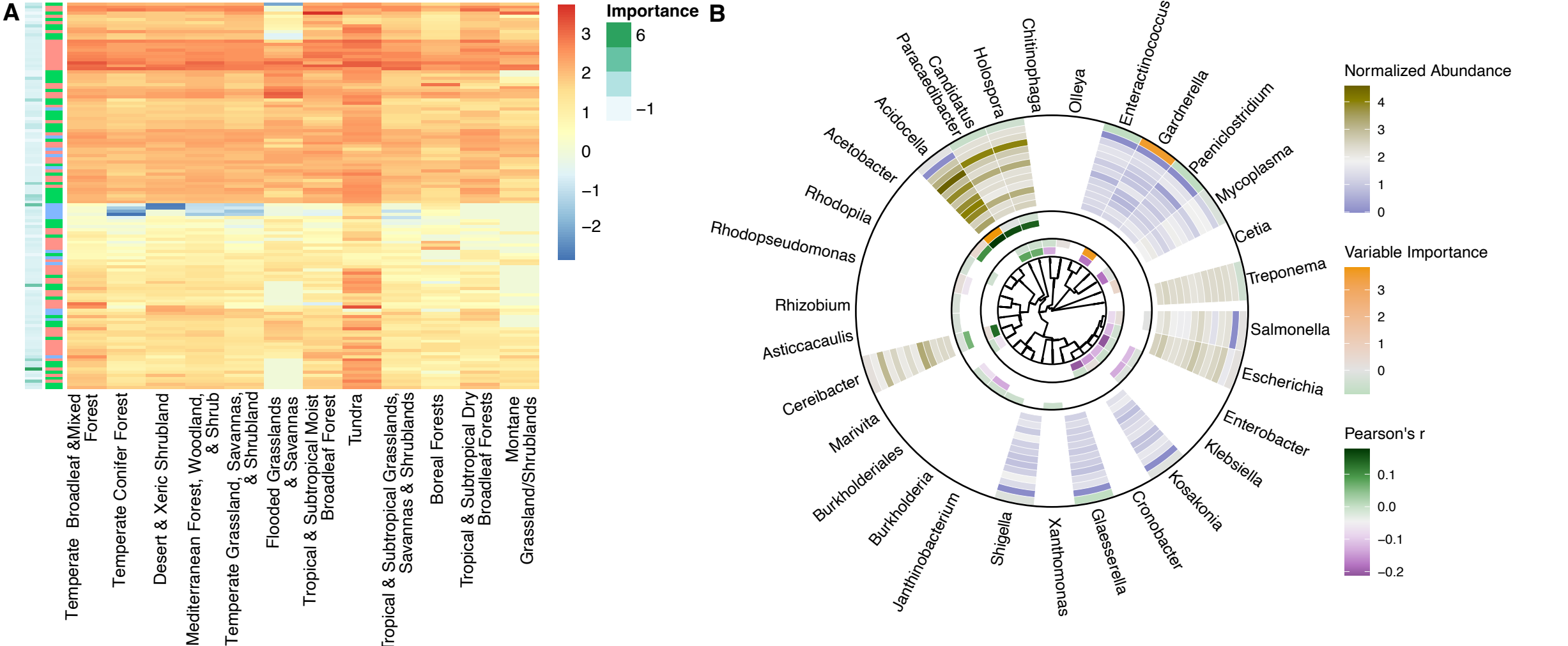
1029 177. Campitelli E. 2020. ggnewscale: Multiple Fill and Colour Scales in'ggplot2'. R package
1030 version 04.

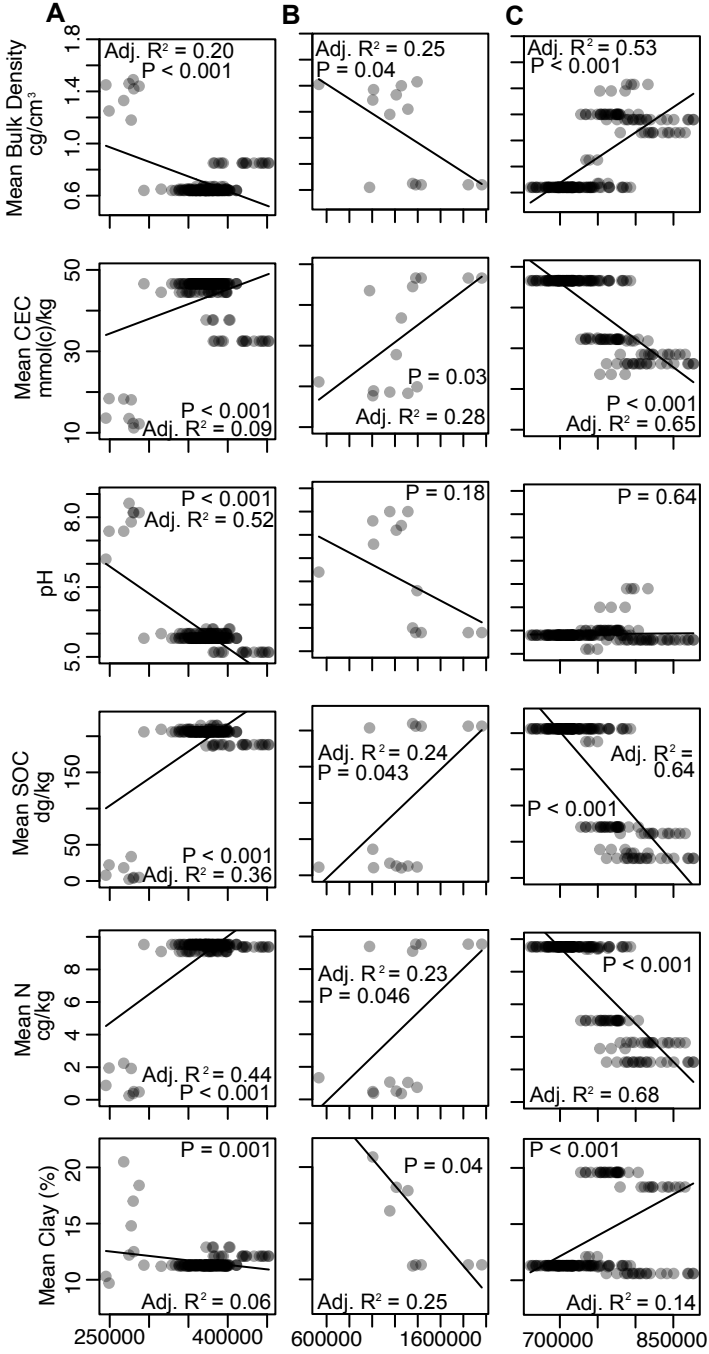
1031 178. Kolde R, Kolde MR. 2018. Package "pheatmap." R Package.


A
50
0
−50
−100
0
100
200
B
Total Assembled Basepairs
2.0e+11
1.5e+11
1.0e+11
5.0e+10
0.0e+00
C
Number of Studies
400
300
200
100
0
Tropical/Subtropical Moist Broadleaf Forests
Tropical/Subtropical Dry Broadleaf Forests
Temperate Broadleaf/Mixed Forests
Temperate Conifer Forests
Boreal Forests/Taigas
Tropical & Subtropical Grasslands/Savannas/Shrublands
Temperate Grasslands/Savannas/Shrublands
Flooded Grasslands/Savannas
Montane Grasslands/Shrublands
Tundra
Mediterranean Forests/Woodlands/Scrub
Deserts & Xeric Shrublands
D
Bulk Density (kg/dm³)
Frequency
0.2
0.6
1.0
1.4
E
CEC (cmol C/kg)
10
20
30
40
50
60
70
F
Clay (%)
G
Nitrogen (g/kg)
Frequency
0
5
10
15
20
H
pH
4
5
6
7
8
I
SOC (g/kg)
0
100
300
500


A
B
C
D
Relative Abundance of Potential (%)
0
0.5
1.0
1.5
2.0
1
2
3
4
5
10
15
20
30
KO:K01992 PF13520.4 TIGR03803 Microbacter
KO:K06136 PF03055.13 TIGR03436 Mycolicibacter
KO:K13038 PF01022.18 TIGR00485 Micromonospora
KO:K00806 PF13561.4 TIGR04416 Paraburkholderia
KO:K00380 PF00106.23 TIGR02532 Rhodobacter
KO:K03496 PF01425.19 TIGR01730 Pseudomonas
KO:K03882 PF10469.7 TIGR01409 Cyanobacteria
KO:K09456 PF09529.8 TIGR02276 Nocardia
KO:K02838 PF00005.25 TIGR01764 Rhizobium
KO:K03088 PF13247.4 TIGR01643 Bradyrhizobium
KO:K00334 PF07357.9 TIGR00231 Pseudonocardia
KO:K02049 PF12838.5 TIGR00254 Acetobacter
KO:K02335 PF08496.8 TIGR01494 Streptomyces
KO:K00104 PF11999.6 TIGR02937 Mycobacterium
KO:K12510 PF08511.9 TIGR00229 Burkholderia


0.6
0.4
0.2
0
−0.2
−0.4
−0.6
Importance (pH)
6
0
Importance (bulk density)
5
0
KO
pfam
tigrfam
Bulk Density
pH


A
Importance
6
−1
3
2
1
0
−1
−2
Temperate Broadleaf &Mixed Forest
Temperate Conifer Forest
Desert & Xeric Shrubland
Mediterranean Forest, Woodland, & Shrub
Temperate Grassland, Savannas, & Shrubland
Flooded Grasslands & Savannas
Tropical & Subtropical Moist Broadleaf Forest
Tundra
Tropical & Subtropical Grasslands, Savannas & Shrublands
Boreal Forests
Tropical & Subtropical Dry Broadleaf Forests
Montane Grassland/Shrublands
B
Chitinophaga
Olleya
Enteractinococcus
Gardnerella
Paeniclostridium
Mycoplasma
Cetia
Treponema
Salmonella
Escherichia
Enterobacter
Klebsiella
Kosakonia
Cronobacter
Glaesserella
Xanthomonas
Shigella
Janthinobacterium
Burkholderia
Burkholderiales
Marivita
Cereibacter
Asticcacaulis
Rhizobium
Rhodopseudomonas
Rhodopila
Acetobacter
Acidocella
Candidatus Paracaedibacter
Holospora
Normalized Abundance
4
3
2
1
0
Variable Importance
3
2
1
0
Pearson's r
0.1
0.0
−0.1
−0.2


A
B
C
Mean Bulk Density cg/cm³
Adj. R² = 0.20
P < 0.001
Adj. R² = 0.25
P = 0.04
Adj. R² = 0.53
P < 0.001
Mean CEC mmol(c)/kg
P < 0.001
Adj. R² = 0.09
P = 0.03
Adj. R² = 0.28
P < 0.001
Adj. R² = 0.65
pH
P < 0.001
Adj. R² = 0.52
P = 0.18
P = 0.64
Mean SOC dg/kg
P < 0.001
Adj. R² = 0.36
Adj. R² = 0.24
P = 0.043
Adj. R² = 0.64
P < 0.001
Mean N cg/kg
Adj. R² = 0.44
P < 0.001
Adj. R² = 0.23
P = 0.046
P < 0.001
Adj. R² = 0.68
Mean Clay (%)
P = 0.001
Adj. R² = 0.06
P = 0.04
Adj. R² = 0.25
P < 0.001
Adj. R² = 0.14
250000
400000
600000
1600000
700000
850000