

1 **Detecting macroecological patterns in bacterial communities across independent studies of**
2 **global soils**

3
4 **Authors:** Kelly S Ramirez^{*+1}, Christopher G. Knight⁺², Mattias de Hollander¹, Francis Q.
5 Brearley³, Bede Constantinides⁴, Anne Cotton⁵, Si Creer⁶, Thomas W. Crowther^{1,7}, John
6 Davison⁸, Manuel Delgado-Baquerizo⁹, Ellen Dorrepaal¹⁰, David R. Elliott^{3,11}, Graeme Fox³,
7 Rob Griffiths¹², Chris Hale¹³, Kyle Hartman¹⁴, Ashley Houlden¹⁵, David L. Jones⁶, Eveline J.
8 Krab¹⁰, Fernando T. Maestre¹⁶, Krista L. McGuire¹⁷, Sylvain Monteux¹⁰, Caroline H. Orr¹⁸, Wim
9 H van der Putten^{1,19}, Ian S. Roberts¹⁵, David A. Robinson²⁰, Jennifer D. Rocca²¹, Jennifer
10 Rowntree³, Klaus Schlaeppli¹⁴, Matthew Shepherd²², Brajesh K. Singh²³, Angela L. Straathof²,
11 Jennifer M. Bhatnagar²⁴, Cécile Thion²⁵, Marcel G.A. van der Heijden^{14,26,27}, and Franciska T.
12 de Vries²

13
14 * email: k.ramirez@nioo.knaw.nl

15 + Joint lead authors

16 ¹ Netherlands Institute of Ecology, Droevendaalsesteeg 10 6708 PB Wageningen NL

17 ² Faculty of Science and Engineering, The University of Manchester, Manchester, M13 9PT, United Kingdom.

18 ³ School of Science and the Environment, Manchester Metropolitan University, Chester Street, Manchester, M1 5GD.

19 ⁴ Evolution and Genomic Sciences, School of Biological Sciences, The University of Manchester, Manchester, M13 9PT, United Kingdom

20 ⁵ Department of Animal and Plant Sciences, The University of Sheffield, Alfred Denny building, Sheffield, South Yorkshire, S10 2TN, UK

21 ⁶ Environment Centre Wales, College of Natural Sciences, Bangor University, Gwynedd, LL57 2UW, United Kingdom.

22 ⁷ Institute of Integrative Biology, ETH Zurich, Univeritätstrasse 16, 8006, Zürich, Switzerland.

23 ⁸ Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, Lai 40, Tartu 51005, Estonia

24 ⁹ Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO 80309.

25 ¹⁰ Climate Impacts Research Centre, Department of Ecology and Environmental Science, Umeå University, Vetenskapens väg 38, 981 07,
26 Abisko, Sweden

27 ¹¹ Environmental Sustainability Research Centre, University of Derby, Kedleston Road, Derby, DE22 1GB, UK

28 ¹² Centre for Ecology and Hydrology, Wallingford, United Kingdom

29 ¹²³ School of Life Sciences, University of Warwick, Coventry, CV4 7AL, United Kingdom

- 30 ¹⁴Division of Agroecology and Environment, Agroscope, Zurich, Reckenholzstrasse 191, Switzerland
- 31 ¹⁵Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, M13 9PT, United Kingdom.
- 32 ¹⁶Departamento de Biología y Geología, Física y Química Inorgánica, Escuela Superior de Ciencias Experimentales y Tecnología, Universidad
- 33 Rey Juan Carlos, Calle Tulipán s/n, 28933 Móstoles, Spain
- 34 ¹⁷Department of Biology, Institute of Ecology and Evolution, University of Oregon, Eugene, OR 97403, USA
- 35 ¹⁸School of Science and Engineering, Teesside University, Middlesbrough, TS1 3BX, United Kingdom.
- 36 ¹⁹Laboratory of Nematology, Wageningen University, Droevendaalsesteeg 1,
- 37 Wageningen 6708 PB, The Netherlands.
- 38 ²⁰Centre for Ecology and Hydrology, Bangor, LL57 2UW, United Kingdom
- 39 ²¹Department of Biology, Duke University, Durham, NC, 27705, United States
- 40 ²²Natural England, United Kingdom
- 41 ²³Hawkesbury Institute for the Environment, Western Sydney University, Richmond
- 42 2753 NSW Australia
- 43 ²⁴Department of Biology, Boston University, Boston, MA, 02215, United States
- 44 ²⁵Institute of Biological and Environmental Sciences, University of Aberdeen, Saint-Machar Drive, AB24 3UU, Aberdeen, United Kingdom
- 45 ²⁶Institute for Evolutionary Biology and Environmental Studies, University of Zürich, Winterthurerstrasse 190, CH-8057, Switzerland.
- 46 ²⁷Plant-Microbe Interactions, Institute of Environmental Biology, Faculty of Science, Utrecht, The Netherlands

47

48 **Keywords:** microbial ecology, soil, diversity, community structure, Illumina sequencing, 16S

49 rRNA gene, biogeography, microbiology, meta-analysis

50

51 **The emergence of high-throughput DNA sequencing methods provides unprecedented**

52 **opportunities to further unravel bacterial biodiversity and its worldwide role from human**

53 **health to ecosystem functioning. However, in spite of the abundance of sequencing studies,**

54 **combining data from multiple individual studies to address macroecological questions of**

55 **bacterial diversity remains methodically challenging and plagued with biases. Here, using a**

56 **machine learning approach that accounts for differences among studies and complex**

57 **interactions among taxa, we merge 30 independent bacterial datasets consisting of 1,998**

58 **soil samples from across 21 countries. While previous meta-analysis efforts have focused on**

59 **bacterial diversity measures or abundances of major taxa, we show that disparate**
60 **amplicon sequence data can be combined at the taxonomy-based level to assess bacterial**
61 **community structure. We find that rarer taxa are more important for structuring soil**
62 **communities than abundant taxa, and that these rarer taxa are better predictors of**
63 **community structure than environmental factors, which are often confounded across**
64 **studies. We conclude that combining data from independent studies can be used to explore**
65 **bacterial community dynamics, identify potential ‘indicator’ taxa with an important role in**
66 **structuring communities, and propose hypotheses on the factors that shape bacterial**
67 **biogeography previously overlooked.**

68

69 Soil microbial communities are more diverse and contain more individuals than any species
70 groups on the planet^{1,2}. Over the last decade, the use of high-throughput sequencing (HTS)
71 methods has substantially advanced our understanding of the worldwide biogeography and
72 ecology of soil bacterial and fungal communities³⁻⁵. Recent work has further demonstrated that
73 inclusion of microbial composition and functional attributes improves earth system models⁶,
74 which is of paramount importance for predicting effects of global change on ecosystem services
75 such as climate regulation or soil fertility⁷. Yet, opposite to the long-standing view that every
76 organism may occur everywhere⁸, even at small scales bacterial communities turn out to be more
77 patchy than previously expected^{9,10}, raising questions regarding dispersal constraints, temporal
78 dynamics, and niche breadth at the global scale¹¹⁻¹³. Due to these knowledge gaps, combined
79 with practical challenges of exhaustive sample collection and the massive diversity of
80 communities, global assessment of soil microbial diversity remains an ongoing research
81 challenge¹⁴.

82

83 For plants and animals, the integration of data from independent studies has been a valuable
84 option for generating an understanding of global biogeography patterns, answering ecological
85 questions (e.g. biodiversity-functioning relationships), and identifying threats to biodiversity
86 from global changes¹⁵⁻¹⁷. Similarly, our understanding of soil microbial diversity would greatly
87 improve from such worldwide assessments. However, the integration of microbial community
88 HTS data from different studies is not so unlike the merging of museum species records where
89 information and data is constrained by variations in nomenclature over space and time, among
90 many other challenges^{18,19}. Like plant and animal records, molecular microbial community
91 records and information can be incomplete, processing and naming varies greatly between
92 studies and over time²⁰, data storage is inconsistent, and there are few curated databases with
93 high quality data (especially for short read sequences)^{21,22}. Further, most microbial community
94 data and metadata are still available only in independently published studies that have been
95 carried out according to their own standards and procedures, and the extent of these confounding
96 factors has never been quantified across studies.

97

98 Regardless of the challenges, as indicated by the many open access data initiatives²³⁻²⁵, merging
99 microbial sequence data is a potential option to address global scale questions, whether relating
100 to the human microbiome²⁶, marine systems²⁷, or predicting the response of soil organisms to
101 global environmental change²⁸. For soil systems, the need to merge sequence data is supported
102 by the emerging role of bacterial phyla and classes as indicators of particular soil conditions such
103 as soil pH and nutrient concentrations^{29,30}. Until now, attempts to meta-analyze sequence data
104 have been limited to assessing diversity measures or abundances of major taxa, because the

105 merging of community data is constrained by methodological differences between sequencing
106 studies^{10,24,31,32}. However, a recent systematic review found that measures of microbial
107 community structure were more often linked to microbial process rates than diversity or
108 presence/absence data³³, and abundance ratios among phyla may be less important than previous
109 believed³⁴. Together indicating that information on variation in microbial community structure is
110 potentially more ecologically relevant than measures of diversity and abundances of major taxa.

111
112 Here, we show that, despite the outlined challenges, published microbial community data from
113 independent studies can be analyzed together to address questions about the global structuring of
114 communities. Using a machine learning approach, we take methodological and technical biases
115 into account, factor in interactions among taxa, and produce an improved assessment of the
116 abiotic and biotic drivers of soil community structure. The objectives of this study were two-
117 fold: (1) to identify the biases and incompatibilities of microbial community HTS studies (and
118 confounding factors) so as to strengthen our ability to integrate data from disparate studies, and
119 (2) to reveal worldwide soil microbial community patterns by merging independent taxonomy-
120 based datasets.

121

122 **Results and Discussion**

123 **Taxonomy-based merging of disparate amplicon sequence data**

124 We identified 30 individual HTS bacterial studies from 21 countries for our analysis (Figure 1A
125 and Supplementary Table 1). While we aimed to merge HTS data of both soil bacterial and
126 fungal datasets, our approach was only successful for bacterial data (Figure 1B and 1C), and
127 highlights the well-known dilemma of fungal databases, where extremely high diversity

128 combined with high endemism and mismatched taxonomy across continents make merging data
129 by taxonomy difficult and unusable for downstream analyses^{4,35}. For the bacterial studies, we
130 were able to successfully merge 30 individual OTU tables; using a taxonomy-based approach,
131 datasets were merged using the taxonomic affiliations of individual OTUs. Once filtered, and
132 singletons removed, the final ‘taxonomy-based’ community contained 1,998 individual soil
133 samples, and 8,287 taxa. Here ‘taxon’ is defined as a unique name in the classification; where a
134 name could be a specific phylum, genus, or other taxonomic level. For example, ‘Acidovorax’
135 (genus) and Proteobacteria (the phylum containing Acidovorax) were both considered as taxa).
136 To account for variation in sequencing depth between different studies, OTU relative abundances
137 were used per sample, rather than absolute read abundance. To test known biogeographical
138 patterns, metadata (information on geographical location, soil pH and soil core measurements)
139 were compiled for all studies. Technical and methodical information was also collected; all of
140 these 30 studies had conducted amplicon sequencing on hypervariable regions of the 16S rRNA
141 gene in soil samples using either Illumina or (Roche) 454 pyrosequencing (with any primer pair)
142 (Supplementary Table 1). For a validation step we retrieved all usable raw sequence data
143 available, resulting in 417 samples from locations across the globe (approximately 1/5 of all our
144 samples) (Figure 1A). Data not included in this sequence-matched analysis either had an
145 incompatible raw sequence format or simply no longer existed. Available raw sequence data
146 were combined into a single ‘sequence-matched’ community comprising 44,106 OTUs
147 (Supplementary Figure 1).

148

149 **Machine learning assessment of bacterial community structure**

150 Ordination of the taxonomy-based community reveals large amounts of structure both within and

151 between studies (structure that is removed by permuting taxa among samples (Supplementary
152 Figure 2), without greatly affecting diversity (Supplementary Table 3)), and the observation of
153 the well-established negative relationship between relative abundance of Acidobacteria and soil
154 pH (Figure 1D)³⁶ confirms our merging method. This visualization also suggests that some of the
155 community variation (e.g. the near absence of Acidobacteria in some studies, even at low pH) is
156 due to technical factors such as the particular primer sets chosen, region sequenced, and
157 sequencing platform (Supplementary Methods and Supplementary Table 2). However, we expect
158 that some taxa are not correlated with technical factors, and are non-randomly distributed with
159 respect to biotic and abiotic factors. Therefore, using a machine learning approach capable of
160 accounting for complex interactions among taxa (Random ForestsTM, see methods), we
161 determined the extent to which individual taxa could influence the community structure of
162 merged independent studies. Here community structure is defined by the presence and relative
163 abundances of individual taxa, along with co-occurrence relationships between those taxa. This
164 was done in two ways: first, we constructed a model that classified the study from which a
165 sample came based on the proportions of the 8,287 taxa it contained (1.5% [\pm 0.02% CI]
166 classification error, by internal cross-validation). Second, we determined the contribution of each
167 taxon to bacterial community structure by quantifying its importance in a model that separated
168 the observed data from synthetic data randomly drawn from the observed distributions of relative
169 abundances for each taxon (*see Methods*).

170

171 Merging of disparate microbial sequence data is known to be plagued with potential biases
172 including: lack of standardization of sample collection, methodological issues regarding DNA
173 extraction and primer choice, incomplete metadata, the technical biases of different sequencing

174 platforms, sequencing depth, PCR Bias, different clustering methods, and the use of different
175 taxonomic classification pipelines³⁷⁻³⁹. We therefore took the step to quantify the importance of
176 both technical and environmental factors alongside taxa in the Random Forests models (Figure
177 2). Of note, ‘owner’, which encompasses the technical biases and uniqueness of a given dataset,
178 is very effective for differentiating between studies (i.e. the owner is far to the right in Figure 2)
179 yet is entirely uninformative about community structure (i.e. owner is at the far bottom in Figure
180 2). In fact, *all* technical factors included are better than 98.5% of all taxa to differentiate between
181 studies, indicating that the observed differences among studies in taxon relative abundances are
182 strongly confounded with technical factors. Independent of taxonomy, certain environmental
183 factors, such as country of origin, latitude and longitude, and soil pH, were highly important in
184 differentiating studies but not in determining community structure. By contrast, minimum soil
185 sampling depth was not very important in separating studies, and was more associated with
186 community structure. It is well known that bacterial diversity decreases with soil depth⁴⁰ and our
187 results show that in a global assessment, soil depth remains a strong predictor of bacterial
188 community composition. Perhaps most useful for future research, this result highlights that not
189 all environmental factors are equally confounded by technical factors, and shows that by
190 combining data from across many independent studies we may identify previously overlooked
191 taxa and factors relevant for structuring communities.

192

193 **Importance for structuring soil bacterial communities**

194 Although all studies were confounded by technical and environmental covariates, there remained
195 many taxa that were non-randomly distributed and were not confounded with technical
196 differences among studies (upper left in Figure 2). When assessing the role of these different taxa

197 in structuring the community, we found a trade-off between taxon abundance and importance in
198 community structure, such that low abundance taxa are disproportionately important in the non-
199 random structure of communities, where the most important taxa are rarer than expected
200 compared to the randomly permuted data (Figure 3). Thus, the importance of taxa for
201 determining community structure is negatively correlated with the average abundance of those
202 taxa, whereas taxon abundance is positively correlated with importance for separating studies (ρ
203 = -0.79 and ρ = +0.51 respectively, rank correlation, cf. null expectations of ρ = -0.62 and -0.12
204 respectively in permuted data). The taxa most closely associated with differences between
205 studies tend to be those present at or greater than 0.1% relative abundance, but those most
206 important in determining community structure tend to be present at 0.0001% abundance or less
207 (with a null expectation of around 0.01-0.001% in each case, Figure 3). This result is only found
208 by considering the full set of studies and is neither apparent within single studies (Supplementary
209 Fig. 4A-B) nor a subset of studies (whether matched by name or sequence Supplementary Fig.
210 5). It corresponds to the long tail in frequency-abundance distributions of soil microbial
211 communities⁴¹, where many taxa in the soil are known to occur at low abundance. Thus if rarer
212 taxa tend to be more important for distinguishing between communities, it is within this long tail
213 that we might identify taxa that could indicate ecological or functional differences among soil
214 communities^{42,43}.

215

216 To be ecological indicators^{44,45}, taxa need to vary in abundance in response to environmental
217 factors and have high occurrence across studies, as is the case for the phylum Acidobacteria³⁶.
218 Acidobacteria, however, are typically abundant and our analysis suggests that the most abundant
219 taxa are *not* the most important in determining community structure. While dominant taxa like

220 Acidobacteria do change with environmental factors such as pH (Figure 1D), those changes are
221 of lesser importance for the ‘non-randomness’ of community structure, and more confounded
222 with technical effects, than changes in less dominant, pH responsive taxa (Supplementary Figure
223 3A). Therefore, we assessed which taxonomic ranks are more or less distinguished from the
224 randomly permuted data. Although differences among domains and phyla are strongly
225 associated with differences among studies (Figure 4B) only taxa at a rank lower than phyla are
226 consistently better than random at identifying community structure (Figure 4A).

227

228 A very similar pattern was found for the sequence-matched community, emphasizing the
229 importance of taxa at the level of Class and below (Supplementary Figure 7A and 7B). However,
230 this was not apparent in individual studies (Supplementary Figure 4C-D), where phyla were
231 relatively important. A subset of the taxonomy-matched studies showed a pattern intermediate
232 between the single studies and the full dataset (phyla with some importance, but less than Class,
233 Order or Family, Supplementary Figure 7C). This, along with abundance analyses (Figure 3 and
234 Supplementary Figure 5), suggests that our name matching approach is consistent with, but less
235 powerful than a full sequence-matched analysis. At the same time, the taxonomy-matching is
236 worthwhile because, as with the findings on abundance (Figure 3), macroecological patterns (the
237 importance of taxa below phyla and of relatively low abundance in community structure) are
238 evident when we consider thousands of samples from tens of studies, that are not apparent from
239 hundreds of samples from one or a handful of studies.

240

241 To be a good ecological indicator a taxon should occur in most studies; we therefore looked
242 explicitly at the relationship between a taxon’s importance in community structure and its

243 occurrence across studies. Low abundance taxa and taxa of lower taxonomic rank are
244 consistently important in determining community structure, but tend to be detected in fewer
245 studies ($\rho = 0.59$ and 0.31 respectively Supplementary Figure 3B and 3C). We discovered a
246 relationship between taxon occurrence across studies and importance for structuring
247 communities for all taxa (Figure 5, Supplementary Table 4). Comparison with the null
248 expectation reveals a range of taxa, occurring in multiple samples from most studies, which are
249 much more important in determining community structure than expected by chance. A similar
250 pattern is apparent in the sequence-matched dataset (Supplementary Figure 8A) and the same
251 subset of studies when taxonomy-matched (Supplementary Figure 8B). Altogether, the analysis
252 clearly illustrates the significance of taxonomic rank, for example *class* Gemmatimonadetes is
253 relatively unimportant for community structure but *genus* Gemmatimonadetes is relatively
254 important. The result also shows rarer taxa being more important in structuring communities and
255 suggests rarer bacterial taxa play overlooked ecologically important roles for bacterial
256 community dynamics⁴³. This result is robust to artifacts caused by the rarest taxa (e.g.
257 differences between 0 and 1 reads in a sample could be significant for a model, without being
258 biologically significant) – a very similar pattern is seen when only taxa present at above 0.003%
259 in any given sample were included in this analysis (typically removing the rarest 10% of taxa
260 from any given sample, Supplementary Figure 9). Conversely, many taxa of high taxonomic rank
261 with high occurrence across samples, such as the phyla Actinobacteria, Acidobacteria,
262 Proteobacteria, and Bacteroidetes, were much less important for community structure than the
263 null expectation. These taxa have been reported elsewhere as ‘core’ members of the soil
264 community^{36,46}, and even been included in source-tracking of microbial communities due to their
265 ubiquitous presence in soil⁴⁷. Yet, it is the consistent presence of the core taxa across samples

266 and studies that makes them inadequate for assessing community structure.

267

268 **Conclusions**

269 Our results demonstrate the power of combining global bacterial HTS data from multiple
270 independent sources for the detection of biogeographical patterns and for identifying community
271 patterns that can be used to generate hypotheses on the roles of certain taxa. Though our
272 assessment was on soil communities, our methods can be applied to broadly to other microbial
273 datasets and disciplines. Taxonomy-based merging gives results that are consistent with raw
274 sequence data, and expands opportunities for extracting information about microbial
275 communities from the wealth of existing and future studies. Moreover, we find that rarer
276 bacterial taxa are more important in differentiating communities than previously assumed, and
277 hold potential as overlooked soil indicators or keystone species. Still, there are considerable
278 challenges associated with merging large sequence datasets beyond the well-known biases that
279 accompany any molecular HTS study. Perhaps the most concerning was that so few raw
280 sequence datasets for publically deposited analyses could be retrieved. This highlights the need
281 for wider community adoption of open and accessible short read sequence databases⁴⁸, open
282 reference clustering⁴⁹, standardized databases⁵⁰ and—as always—that metadata should be
283 consistent and accessible. Regardless of these challenges, as HTS methods rapidly advance we
284 must find ways to simultaneously curate and carry our research knowledge forward. Only then,
285 in combination with the many recently designed and classical approaches, can we uncover the
286 full breadth of soil diversity and the roles soil microbes play for ecosystem processes.

287

288 **Methods:**

289 *Description of datasets:*

290 Metadata from the 30 studies and 1998 samples were collected and compiled into a summary
291 data file. To do so, we standardized the metadata of each study using the dplyr package⁵¹ of the
292 R statistical platform⁵². Samples were collected from 21 counties representing all continents
293 except Antarctica. In addition to location and pH data (median = 6.1, quartile range=5.3-7.0),
294 which were available from all studies, information on altitude (10 m, 10-860 m), soil moisture
295 (19.5%, 14.1-27.4%), and total soil nitrogen (0.36 mg kg⁻¹, 0.23-0.51 mg kg⁻¹), carbon (4.7%,
296 1.9-7.5%) and phosphorus (20.7 mg kg⁻¹, 7.0-223.0 mg kg⁻¹) was noted where available. Depth
297 of sample collection was also noted and ranged from surface collections to a maximum depth of
298 70 cm, with 83% of samples originating from 0-10 cm below the soil surface. Samples
299 represented anthropogenically managed (59%) and natural (40%; remaining samples undefined)
300 systems, and were taken from arable, grassland, peatland, forest, scrub (including tundra) and
301 urban habitats. The majority of samples (71%) were described as non-experimental, meaning no
302 treatments were applied, with the remainder described as experimental. Sequencing data were
303 either produced using Roche 454 technology (22%) or one of the Illumina platforms (78%).
304 Primer pairs were defined for 92% of the samples and nine different pairs were identified from
305 the study meta data (27F:338R; 341F:518R; 341F:806R; 341F:907R; 357F:926R; 515F:806R;
306 577F:926R; 799F:1193R and 341F:805R) with the majority of samples (66%) using 515F and
307 806R to produce amplicons. Post sequencing processing varied, but 81% of samples were run
308 through the QIIME workflow at some point. An OTU table for 1 study comprising 43 samples
309 was programmatically retrieved from the MG-RAST public metagenome repository⁵³.
310 Taxonomy for the different studies was mainly assigned using the Greengenes database (84 %),

311 but RDP (6 %;³⁷ and the Silva database (9 %) ⁵⁴ were also used.

312

313 *Primer Biases*

314 It has long been well understood that different primers vary in their biases for amplifying
315 members of the bacterial community^{55,56}. To demonstrate this bias, the likelihood of significant
316 differences in primer biases for the ten pairs of primers used in the studies analysed were
317 determined by *in silico* analysis. Sequences of primer pairs were compared to all 16S rRNA gene
318 sequences in the SILVA non-redundant reference database (SSURef NR) release 128⁵⁴ using
319 TestPrime v1.0 (as described in⁵⁷). The percentages of sequences of each bacterial phyla that
320 matched both primers (with a one base pair mismatch allowance at least 1bp from the 3' end of
321 the primers) were calculated to compare predicted differences in primer coverage of different
322 bacterial taxa.

323

324 *Merging OTU tables:*

325 For the OTU tables from the 30 individual studies to be merged, extensive data cleaning was
326 carried out on the OTU and taxonomy files to maximize the possibility of matching taxa across
327 datasets. This comprised several steps: (1) Most datasets contained a seven-level taxonomy,
328 recorded in a variety of ways, which was converted to a standardized format. (2) Individual
329 taxon names were cleaned, to give a single name at each taxonomic level (e.g. removing special
330 characters and extra annotations, such as 'candidate division' or details of containing taxa). (3)
331 For the many cases where a taxon was not assigned at a particular taxonomic level, a unified
332 'unassigned' label was created. Repeating analyses with all these taxa removed made no
333 qualitative difference to the results (Supplementary Figure 10). Merging at the taxonomy-based

334 level has the added benefit of lessening the impacts of hypervariable regions. For example, the
335 identification of an organism at a specific level in one sample also contributes to the
336 identification of the containing genus for that sample, allowing direct comparison with a sample
337 where, because a different region was sequenced, that same organism is only resolved to the
338 genus level. Next, relative abundance data were, where necessary, re-scaled to sum to 1 for a
339 sample, using original OTU count files where possible. These values were then manipulated to
340 give data tables usable for modeling using custom R scripts. For some analyses (Figures 3-5), a
341 dataset without community structure was created by randomly permuting the relative abundance
342 of each taxon across all samples. Unless otherwise stated, the analyses performed on the
343 permuted dataset was identical to that performed on the observed data.

344

345 *Merging raw sequence data and other validation datasets:*

346 While no dataset can currently provide a “ground truth” against which to judge our approach, we
347 can at least validate it. The primary validation of our taxonomy-matching approach was to merge
348 raw sequence data (‘sequence-matched’) from 419 samples of the total 1998 used. Per sample
349 fastq files were obtained for each individual dataset. Read files were quality filtered with sickle⁵⁸
350 for single end reads trimming bases below phred score 36 and shorter than 100bp. These
351 stringent filtering criteria were applied to keep only high quality reads and to make sure it is
352 possible to map reads to full length 16S rRNA gene sequences. Full length 16S rRNA gene
353 sequences from the Silva 119 release⁵⁴ were obtained in Qiime compatible format from the [Silva](#)
354 [Download Archive](#) For each dataset, all reads were mapped to the full length 16S rRNA gene
355 sequences using the usearch global algorithm implemented in VSEARCH version 1.9.6⁵⁹. The
356 alignment results in usearch table format (uc) were directly converted to BIOM format using

357 biom version 2.1.5⁶⁰. Consensus/majority taxonomy was added as metadata to the biom file.
358 Finally, all BIOM files of each dataset were merged using Qiime version 1.9.1⁶¹. All steps were
359 implemented in a workflow made with Snakemake version 3.5.4⁶² available: ([De Hollander](#)
360 [2016](#)) (Supplementary Figure 1).

361
362 To use this sequence-matched dataset to validate our taxonomy-matching approach across
363 studies using different taxonomy databases (Supplementary Figures 5, 7 & 8) we created an
364 equivalent taxonomy-matched dataset from the same 5 studies. As with the full dataset, only taxa
365 occurring in at least two studies were included in either this or the sequence-matched dataset. To
366 test what is gained or lost by considering different numbers of studies simultaneously, we
367 considered, not only the full dataset (30 studies) and the subset of 5 studies used in the sequence-
368 matched dataset, but two of the largest individual studies: from Central Park, NYC
369 encompassing 594 samples (study #24) and a global dataset encompassing 103 samples (study
370 #30). In each case a simple subset of the full dataset was analyzed (Supplementary Figure 4). To
371 address PCR biases (Supplementary Table 2) and biases associated with rare taxa, we created a
372 filtered subset of the data where only taxa present at above 0.003% in any given sample were
373 considered, meaning that all taxa deemed present are represented by multiple sequence reads
374 (Supplementary Figure 9). To address the issue of differential 16S copy numbers skewing
375 abundance estimates, we created a binary dataset of the presence/absence of all taxa. The results
376 for a model separating studies using this dataset were very similar to the main dataset using
377 relative abundance, however, there was insufficient power to identify taxa important for
378 community structure. Nonetheless, this analysis did agree with the main analysis that phyla were
379 the most stable taxonomic level, with lower importance than on the permuted data

380 (Supplementary Figure 6). Finally, to test the effect of ‘unknown’ or unclassified bacterial taxa
381 we created a reduced dataset where all taxa classified as ‘unassigned’ at any level were removed
382 (Supplementary Figure 10).

383

384 *Random forest models.*

385 To test for the importance of different taxa in the structuring of the data we used Random Forest
386 models^{63–65} with the relative abundances of the taxa as explanatory variables. Random Forest
387 models have two principal advantages in this context: 1) they can deal easily with thousands of
388 explanatory variables and quantify their relative importance, and 2) they can run equivalently in
389 both supervised and un-supervised modes. In the latter, the importance of a variable describes
390 how effective it is at separating the observed data from randomized synthetic data⁶⁵. In both
391 cases, a proximity matrix may be generated, which can be used for ordination (Supplementary
392 Figure 2). The importance of individual taxa in a Random Forest relate to traditional ecological
393 measures. For instance, the importance in a supervised model, such as that used separating
394 studies (x-axis in Figure 2) is closely correlated with the sensitivity component of the indicator
395 value of each taxon ($\rho = 0.89$, Supplementary Figure 3D)⁴⁵. There are two key parameters that
396 may be adjusted in a Random Forest model, *mtry*, the number of variables randomly sampled as
397 candidates for a split in the constituent trees and *ntree*, the number of trees in the forest. *mtry* was
398 set at its default value (square root of the number of variables) *ntree* was set to 100,000 for each
399 forest. Such a large number of trees was found to be necessary to achieve stable importance
400 across taxa and was achieved by combining several forests run in parallel without normalizing
401 votes. Other parameters were left at default values, in particular, trees were grown to completion
402 (i.e. a minimum node size of 1). The un-scaled permutation importance of variables is used

403 throughout: Each variable importance is the difference between the classification error rate of a
404 tree on data not used to construct it (the ‘out of bag’ data) and the same error following random
405 permutation of the variable in question, averaged over all trees.

406

407 We used permuted data (see above) to create null distributions for taxon importances. For
408 unsupervised Random Forests analyses, such as the community structure model, this amounts to
409 calculating how important a taxon with a particular abundance distribution is for separating two
410 randomized distributions. This can then be compared to its importance for separating the
411 observed from a randomized distribution. This clarifies the fact that, even in null data without
412 community structure (Supplementary Figure 2), variable importance correlates with ecologically
413 important factors, such as abundance. This makes intuitive sense in as much as, even with
414 randomized samples, is easier to separate them on the basis of taxa that occur in only some of
415 them than on the basis of ubiquitous taxa. This, for instance, results in the negative slope of the
416 orange (permuted, null, data) line in Figure 5. All analyses were completed with RandomForest
417 package for R version 4.6.

418

419 **References**

- 420 1. Prosser, J. I. Dispersing misconceptions and identifying opportunities for the use of
421 ‘omics’ in soil microbial ecology. *Nat. Rev. Microbiol.* **13**, 439–46 (2015).
- 422 2. Bardgett, R. D. & van der Putten, W. H. Belowground biodiversity and ecosystem
423 functioning. *Nature* **515**, 505–511 (2014).
- 424 3. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina
425 HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–4 (2012).
- 426 4. Tedersoo, L. *et al.* Global diversity and geography of soil fungi. *Science (80-.).* **346**,
427 (2014).
- 428 5. Davison, J. *et al.* Global assessment of arbuscular mycorrhizal fungus diversity reveals
429 very low endemism. *Science (80-.).* **349**, (2015).
- 430 6. Wieder, W. R., Bonan, G. B. & Allison, S. D. Global soil carbon projections are improved
431 by modelling microbial processes. *Nat. Clim. Chang.* **3**, 909–912 (2013).

- 432 7. Karhu, K. *et al.* Temperature sensitivity of soil respiration rates enhanced by microbial
433 community response. *Nature* **513**, 81–84 (2014).
- 434 8. Barberán, A., Casamayor, E. O. & Fierer, N. The microbial contribution to macroecology.
435 *Front. Microbiol.* **5**, 203 (2014).
- 436 9. Ramirez, K. S. *et al.* Biogeographic patterns in below-ground diversity in New York
437 City’s Central Park are similar to those observed globally. *Proc. Biol. Sci.* **281**, 20141988-
438 (2014).
- 439 10. O’Brien, S. L. *et al.* Spatial scale drives patterns in soil bacterial diversity. *Environ.*
440 *Microbiol.* **18**, 2039–2051 (2016).
- 441 11. Evans, S., Martiny, J. B. H. & Allison, S. D. Effects of dispersal and selection on
442 stochastic assembly in microbial communities. *ISME J.* **11**, 176–185 (2017).
- 443 12. Talbot, J. M. *et al.* Endemism and functional convergence across the North American soil
444 mycobiome. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 6341–6 (2014).
- 445 13. Barber, A. *et al.* Why are some microbes more ubiquitous than others? Predicting the
446 habitat breadth of soil bacteria. *Ecol. Lett.* **17**, 794–802 (2014).
- 447 14. Ranjard, L. *et al.* Turnover of soil bacterial diversity driven by wide-scale environmental
448 heterogeneity. *Nat. Commun.* **4**, 1434 (2013).
- 449 15. Jetz, W., McPherson, J. M. & Guralnick, R. P. Integrating biodiversity distribution
450 knowledge: toward a global map of life. *Trends Ecol. Evol.* **27**, 151–9 (2012).
- 451 16. Ricketts, T. H. *et al.* Disaggregating the evidence linking biodiversity and ecosystem
452 services. *Nat. Commun.* **7**, 13106 (2016).
- 453 17. Dirzo, R. *et al.* Defaunation in the Anthropocene. *Science (80-.).* **345**, 401–406 (2014).
- 454 18. Patterson, D. J., Cooper, J., Kirk, P. M., Pyle, R. L. & Remsen, D. P. Names are key to the
455 big new biology. *Trends Ecol. Evol.* **25**, 686–691 (2010).
- 456 19. Santos, A. M. & Branco, M. The quality of name-based species records in databases.
457 *Trends Ecol. Evol.* **27**, 6-7-8 (2012).
- 458 20. Beiko, R. G. Microbial Malaise: How Can We Classify the Microbiome? *Trends*
459 *Microbiol.* **23**, 671–679 (2015).
- 460 21. Tedersoo, L. *et al.* Standardizing metadata and taxonomic identification in metabarcoding
461 studies. *Gigascience* **4**, 34 (2015).
- 462 22. Ramirez, K. S. *et al.* Toward a global platform for linking soil biodiversity data. *Front.*
463 *Ecol. Evol.* **3**, (2015).
- 464 23. Turner, W. *et al.* Free and open-access satellite data are key to biodiversity conservation.
465 *Biol. Conserv.* **182**, 173–176 (2015).
- 466 24. Gilbert, J. A., Jansson, J. K. & Knight, R. The Earth Microbiome project: successes and
467 aspirations. *BMC Biol.* **12**, 69 (2014).
- 468 25. Joppa, L. N. *et al.* Filling in biodiversity threat gaps. *Science (80-.).* **352**, (2016).
- 469 26. Sinha, R., Abnet, C. C., White, O., Knight, R. & Huttenhower, C. The microbiome quality
470 control project: baseline study design and future directions. *Genome Biol.* **16**, 276 (2015).
- 471 27. Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored ‘rare
472 biosphere’. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 12115–20 (2006).
- 473 28. García-Palacios, P. *et al.* Are there links between responses of soil microbes and
474 ecosystem functioning to elevated CO₂, N deposition and warming? A global perspective.
475 *Glob. Chang. Biol.* **21**, 1590–1600 (2015).
- 476 29. Hermans, S. M. *et al.* Bacteria as Emerging Indicators of Soil Condition. *Appl. Environ.*
477 *Microbiol.* **83**, AEM.02826-16 (2017).

- 478 30. Philippot, L. *et al.* The ecological coherence of high bacterial taxonomic ranks. *Nat. Rev.*
479 *Microbiol.* **8**, 523–529 (2010).
- 480 31. Shade, A., Caporaso, J. G., Handelsman, J., Knight, R. & Fierer, N. A meta-analysis of
481 changes in bacterial and archaeal communities with time. *ISME J.* **7**, 1493–506 (2013).
- 482 32. Hendershot, J. N., Read, Q. D., Henning, J. A., Sanders, N. J. & Classen, A. T.
483 Consistently inconsistent drivers of microbial diversity and abundance at macroecological
484 scales. *Ecology* **98**, 1757–1763 (2017).
- 485 33. Bier, R. L. *et al.* Linking microbial community structure and microbial processes: an
486 empirical and conceptual overview. *FEMS Microbiol. Ecol.* **91**, (2015).
- 487 34. Walters, W. A., Xu, Z. & Knight, R. Meta-analyses of human gut microbes associated
488 with obesity and IBD. *FEBS Lett.* **588**, 4223–4233 (2014).
- 489 35. Bik, H. M. *et al.* Sequencing our way towards understanding global eukaryotic
490 biodiversity. *Trends Ecol. Evol.* **27**, 233–243 (2012).
- 491 36. Lauber, C. L., Hamady, M., Knight, R. & Fierer, N. Pyrosequencing-Based Assessment of
492 Soil pH as a Predictor of Soil Bacterial Community Structure at the Continental Scale.
493 *Appl. Environ. Microbiol.* **75**, 5111–5120 (2009).
- 494 37. McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological
495 and evolutionary analyses of bacteria and archaea. *ISME J.* **6**, 610–8 (2012).
- 496 38. Lozupone, C. & Stombaugh, J. Meta-analyses of studies of the human microbiota.
497 *Genome ...* (2013).
- 498 39. Pawluczyk, M. *et al.* Quantitative evaluation of bias in PCR amplification and next-
499 generation sequencing derived from metabarcoding samples. *Anal. Bioanal. Chem.* **407**,
500 1841–1848 (2015).
- 501 40. Lu, X., Seuradze, B. J. & Neufeld, J. D. Biogeography of soil Thaumarchaeota in relation
502 to soil depth and land usage. *FEMS Microbiol. Ecol.* **93**, (2017).
- 503 41. Jung, S. P. & Kang, H. Assessment of microbial diversity bias associated with soil
504 heterogeneity and sequencing resolution in pyrosequencing analyses. *J. Microbiol.* **52**,
505 574–580 (2014).
- 506 42. Langille, M., Zaneveld, J. & Caporaso, J. Predictive functional profiling of microbial
507 communities using 16S rRNA marker gene sequences. *Nature* (2013).
- 508 43. Jousset, A. *et al.* Where less may be more: how the rare biosphere pulls ecosystems
509 strings. *ISME J.* **11**, 853–862 (2017).
- 510 44. Hermans, S. M. *et al.* Bacteria as emerging indicators of soil condition. *Appl. Environ.*
511 *Microbiol.* AEM.02826-16 (2016). doi:10.1128/AEM.02826-16
- 512 45. Cáceres, M. De & Legendre, P. Associations between species and groups of sites: indices
513 and statistical inference. *Ecology* **90**, 3566–3574 (2009).
- 514 46. Maestre, F. T. *et al.* Increasing aridity reduces soil microbial diversity and abundance in
515 global drylands. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 15684–9 (2015).
- 516 47. Knights, D. *et al.* Bayesian community-wide culture-independent microbial source
517 tracking. *Nat. Methods* **8**, 761–763 (2011).
- 518 48. Muir, P. *et al.* The real cost of sequencing: scaling computation to keep pace with data
519 generation. *Genome Biol.* **17**, 53 (2016).
- 520 49. Rideout, J. R. *et al.* Subsampled open-reference clustering creates consistent,
521 comprehensive OTU definitions and scales to billions of sequences. *PeerJ* **2**, e545 (2014).
- 522 50. Yilmaz, P. *et al.* The genomic standards consortium: bringing standards to life for
523 microbial ecology. *ISME J.* **5**, 1565–7 (2011).

- 524 51. Wickham, H. & Francois, R. dplyr: A Grammar of Data Manipulation. R package version
525 0.5.0. *R package version 0.5.0.* (2016). at <<https://cran.r-project.org/package=dplyr>>
526 52. Computing., R. A. language and environment for statistical. R Core Team. (2016).
527 53. Wilke, A. *et al.* The MG-RAST metagenomics database and portal in 2015. *Nucleic Acids*
528 *Res.* **44**, D590–D594 (2016).
529 54. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data
530 processing and web-based tools. *Nucleic Acids Res.* **41**, D590-6 (2013).
531 55. Suzuki, M. T. & Giovannoni, S. J. Bias caused by template annealing in the amplification
532 of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol.* **62**, 625–30 (1996).
533 56. Sipos, R. *et al.* Effect of primer mismatch, annealing temperature and PCR cycle number
534 on 16S rRNA gene-targeting bacterial community analysis. *FEMS Microbiol. Ecol.* **60**,
535 341–350 (2007).
536 57. Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for
537 classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41**,
538 e1 (2013).
539 58. Joshi & Fass, J. Sickle: A sliding-window, adaptive, quality-based trimming tool for
540 FastQ files. (2011).
541 59. Rognes, T. *et al.* vsearch: VSEARCH 1.9.6. (2016). doi:10.5281/ZENODO.44512
542 60. McDonald, D. *et al.* The Biological Observation Matrix (BIOM) format or: how I learned
543 to stop worrying and love the ome-ome. *Gigascience* **1**, 7 (2012).
544 61. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing
545 data. *Nat Meth* **7**, 335–336 (2010).
546 62. Koster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine.
547 *Bioinformatics* **28**, 2520–2522 (2012).
548 63. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
549 64. Breiman, L. & Cutler, A. Random Forests Manual v4.0. *Technical report, UC Berkeley*
550 (2003). at <<https://www.scribd.com/document/208387804/Using-Random-Forests-v4-0>>
551 65. Shi, T. & Horvath, S. Unsupervised Learning With Random Forest Predictors. *J. Comput.*
552 *Graph. Stat.* **15**, 118–138 (2006).
553
554

555 **Data availability:** The authors declare that the data supporting the findings of this study
556 are available within the paper and its supplementary information files.

557 **Correspondence and requests for materials** should be addressed to K.S.R

558 **Acknowledgements:** We thank all the people who contributed data and input to this study. This
559 study was conducted at a workshop (5/2015, Manchester, UK) funded by the British Ecological
560 Society's special interest group Plants-Soils-Ecosystems and organized by FTDV and KSR. This
561 study and participants were funded in part by ERC Adv grant 26055290 (KSR, WHvdP);
562 BBSRC David Phillips Fellowship (BB/L02456X/1) (FTDV); ERC Grant Agreements 242658
563 [BIOCOM] and 647038 [BIODESERT] (FTM); the European Regional Development Fund
564 (*Centre of Excellence EcolChange*) (JD); Yorkshire Agricultural Society, Nafferton Ecological
565 Farming Group, and the Northumbria University Research Development Fund (CHO); BBSRC
566 Training Grant (BB/K501943/1) (CH); Wallenberg Academy Fellowship (KAW 2012.0152),
567 Formas (214-2011-788) and Vetenskapsrådet (612-2011-5444) (ED); the Glastir Monitoring &
568 Evaluation Programme (Contract reference: C147/2010/11) and the full support of the GMEP
569 team on the Glastir project (DLJ, SC, DAR). Data taken from work carried out in collaboration
570 with CJS, CL, JMC and SPC. Computing was facilitated by the University of Manchester
571 Condor pool and the CLIMB infrastructure (www.climb.ac.uk).

572

573 **Author Contributions:** F.T.dV. and K.S.R. conceived the idea of this study. The datasets were
574 compiled by C.G.K., R.G., J.D., A.H., B.C., G.F., A.L.S. & J.K.R.. Metadata was compiled by
575 J.D and J.K.R.. Raw sequence analysis was conducted by M.dH.. Primer bias analysis was
576 conducted by A.C.. Random forest analyses and figures were conducted by C.G.K.. The
577 manuscript was written by K.S.R., C.G.K., and F.T.dF. with contributions from all co-authors.

578

579 **Figures:**

580 **Figure 1. Merging of data from 32 independent studies demonstrates wide geographic**
581 **breadth, community variation, and confirms the well-known importance of soil pH. A.** Map
582 of locations from which samples were collected, with zoom panels on the United States (left) and
583 western Europe (right). Points in blue were used in both the taxonomy-based and raw-unified
584 analyses and red points were only used in taxonomy-based analyses. **B.** Average proportion of
585 total prokaryotic abundance and **C.** eukaryotic abundance, represented by taxa shared among
586 different numbers of datasets at different taxonomic levels. Level 1 indicates the complete data,
587 levels 2-4 are subsets of the data containing only taxa present in a minimum of 2-4 separate
588 datasets. **D.** Correlation plot of Acidobacteria relative abundance to soil pH where each color
589 represents a different study ($r = -0.42$ $p=8.6 \times 10^{-87}$).

590

591 **Figure 2: Regardless of technical differences between studies, many bacterial taxa are still**
592 **informative about bacterial community structure.** Machine learning models classify the study
593 from which samples came (x-axis) based on the relative abundance of taxa within samples and
594 distinguish the observed distribution of taxa among samples from random (y-axis). Plotted
595 alongside bacterial taxa (black) are technical factors (red) and ecological factors (purple),
596 including soil pH, minimum and maximum soil depth, longitude, latitude and degrees from the
597 equator. All values are variable importance from Random Forest models (see *Methods*) – points
598 further to the right on the x-axis have more importance in separating studies, while points higher
599 up on the y-axis, have more importance for community structure. Note the non-linear axes.

600

601 **Figure 3: Rarer taxa are more important for structuring communities than abundant taxa.**

602 Here we show the thousand most important bacterial taxa in community structure (A) and in
603 separating studies (B) with respect to their average relative abundance across samples. Plotted
604 are the ‘observed’ points (green) and ‘permuted’ points (orange) which are a null distribution
605 from performing the same analysis on a permuted dataset (see *Methods*). The y-axis reports the
606 rank variable importance in the Random Forests model of community structure (see *Methods*),
607 i.e. the taxon with the greatest importance in this model is ranked 1, the second greatest 2, etc.

608

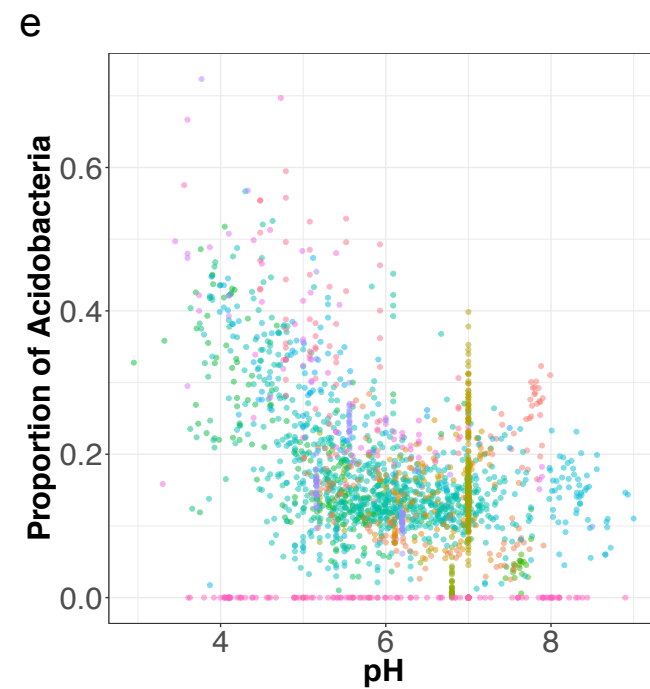
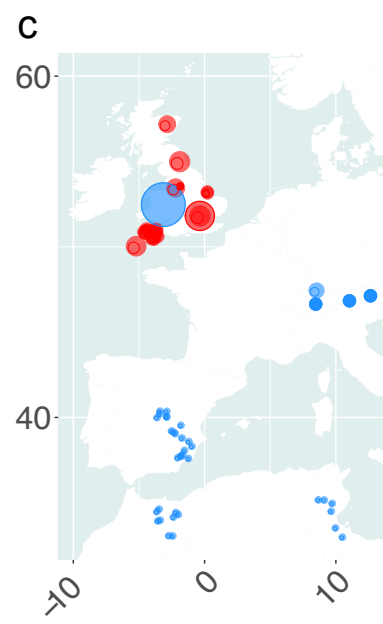
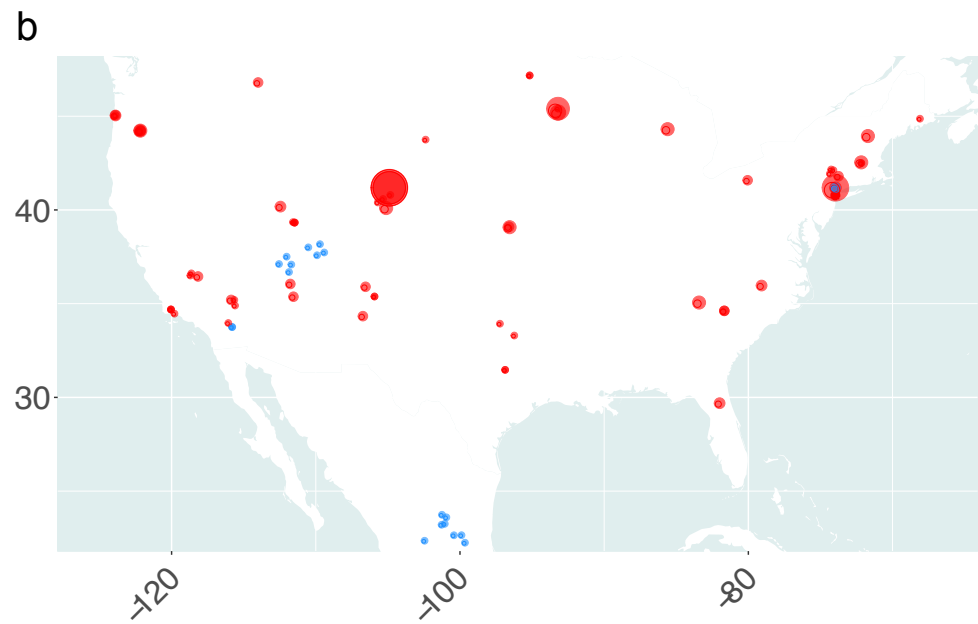
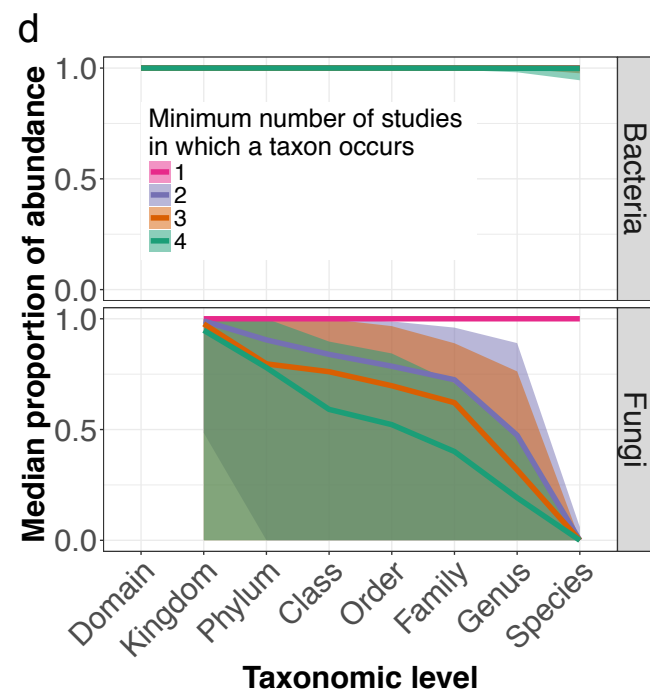
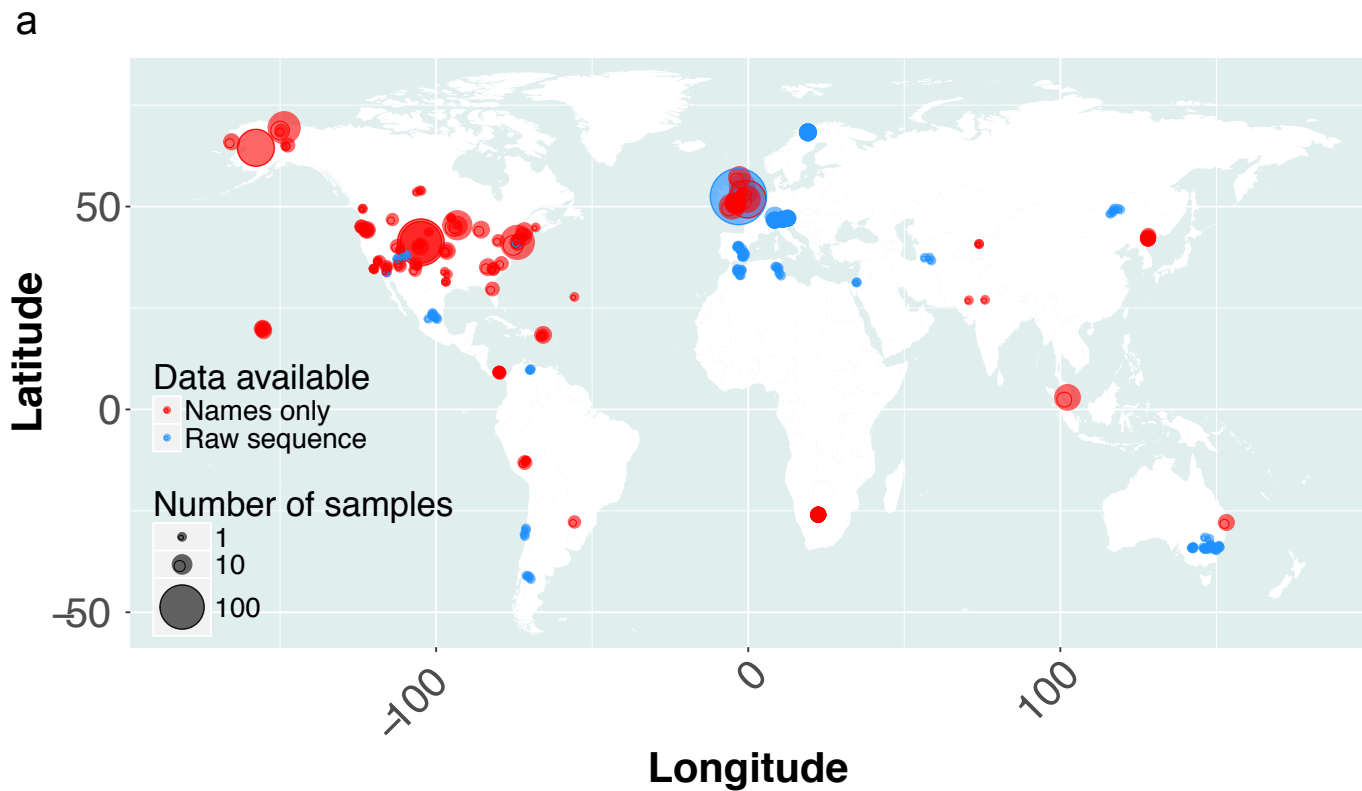
609 **Figure 4: The importance of bacterial taxa classified at different taxonomic ranks.** Lower

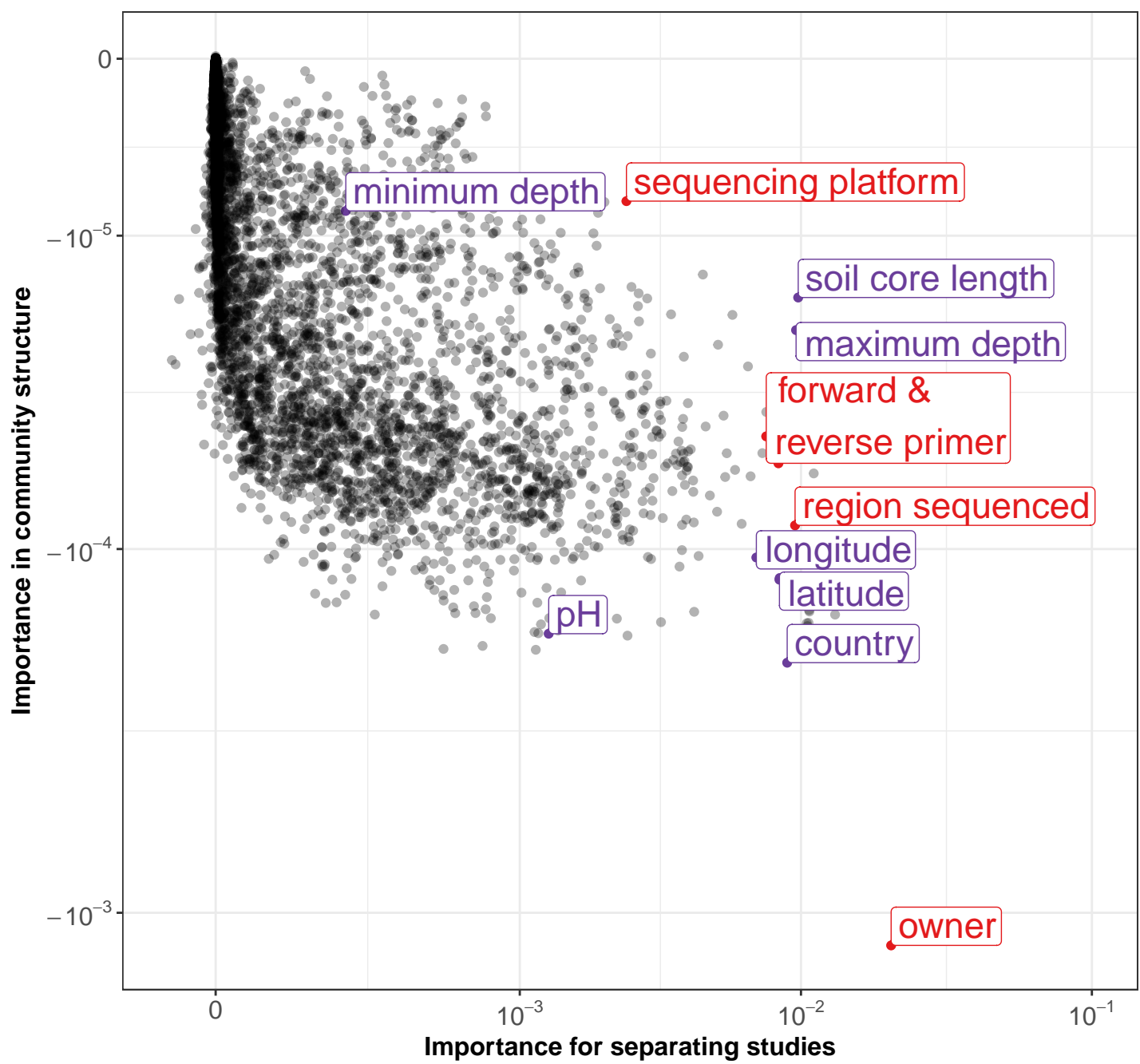
610 taxonomic rank is more important for community structure (A), while high taxonomic rank is
611 more important for separating studies (B). For each taxon, the difference was calculated between
612 the variable importance (see *Methods*) of that taxon in a Random Forests model of either
613 community structure or separating studies and the equivalent value from an analysis performed
614 on the permuted dataset (see *Methods*). The lines and grey ribbons show the mean and standard
615 error respectively of these values across taxa at each taxonomic rank considered.

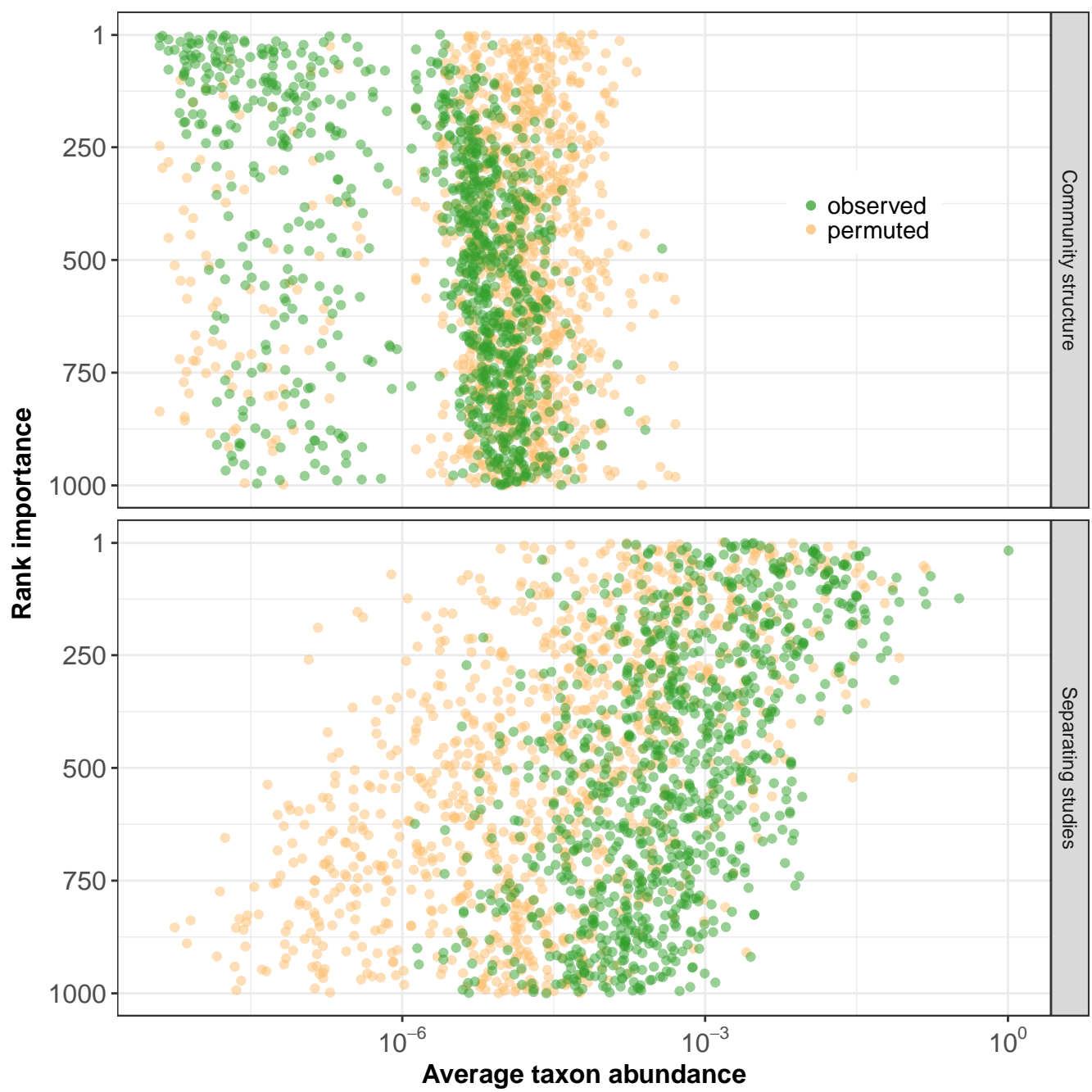
616

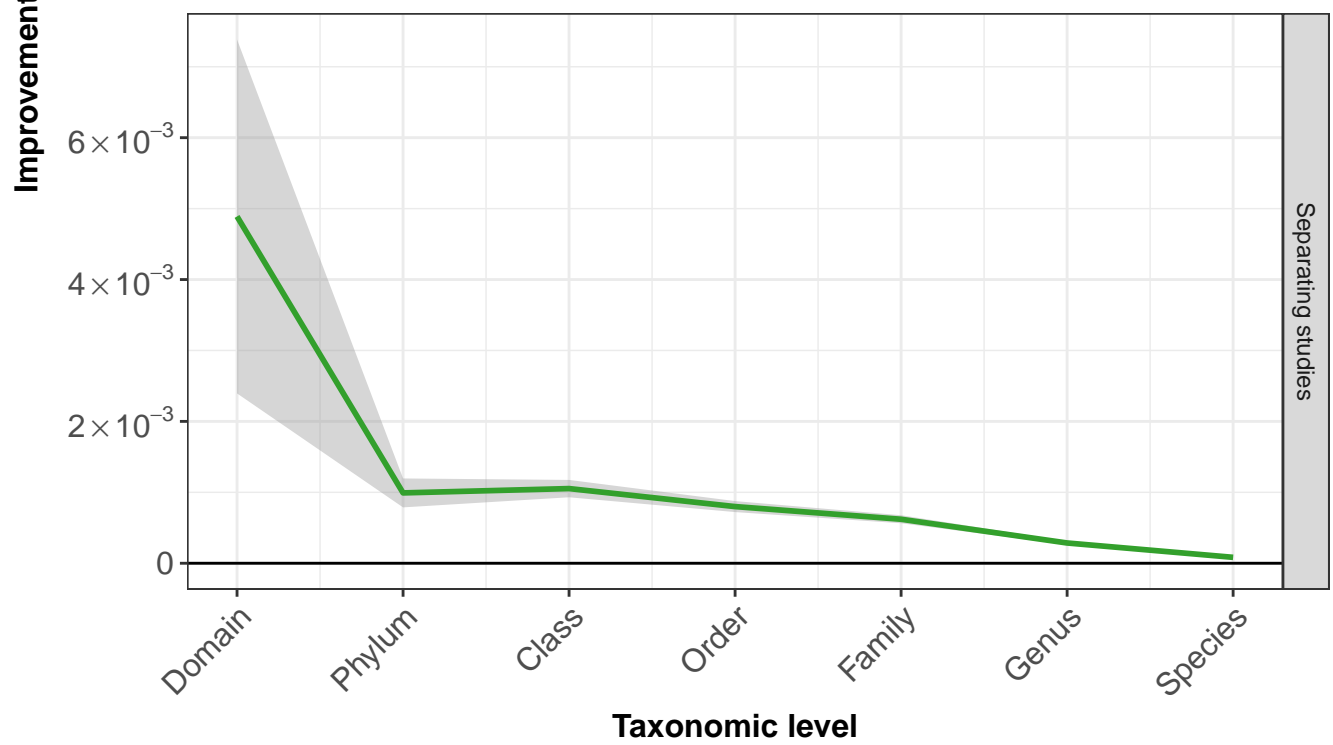
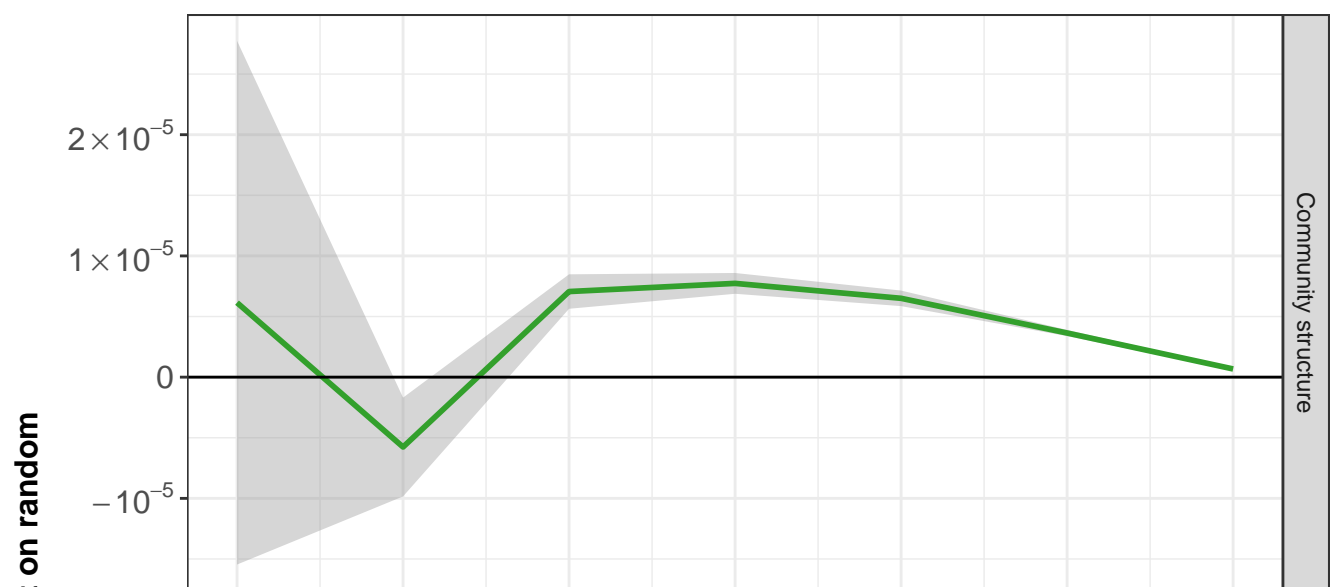
617 **Figure 5: Importance of bacterial taxa in community structure related to their occurrence**

618 **in different studies.** The y-axis reports the variable importance in the Random Forests model of
619 community structure (see *Methods*). Green ‘observed’ points correspond to those taxa shown in
620 Figure 1. Orange ‘permuted’ points correspond to the same analysis on a null distribution (see
621 *Methods*). Lines are general additive model (gam) smoothers. Each line is shown with a
622 confidence interval (grey); where this is not visible it is narrower than the line it surrounds.









Importance in community structure

