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

Authors: Kelly S Ramirez[*+1], Christopher G. Knight[+2], Mattias Hollander[1], Francis Q. Brearley[3], Bede Constantinides[4], TE Anne Cotton[5], Si Creer[6], Thomas W. Crowther[1], John Davison[7], Manuel Delgado-Baquerizo[8], Ellen Dorrepaal[9], David R. Elliott[3,10], Graeme Fox[3], Rob Griffiths[11], Chris Hale[12], Kyle Hartman[13], Ashley Houlden[14], David L. Jones[6], Eveline J. Krab[9], Fernando T. Maestre[15], Krista L. McGuire[16], Sylvain Monteux[9], Caroline H. Orr[17], Wim H van der Putten[1,18], Ian S. Roberts[14], David A. Robinson[19], Jenny Rocca[20], Jennifer Rowntree[3], Klaus Schlaeppi[13], Matthew Shepherd[21], Brajesh K. Singh[22], Angela Straathof[2], Jennifer M. Talbot[23], Cécile Thion[24], Marcel van der Heijden[13], and Franciska T. de Vries[2]

* Corresponding author

+ Joint lead authors

[1] Netherlands Institute of Ecology, Droevendaalsesteeg 10 6708 PB Wageningen, The Netherlands

[2] School of Earth and Environmental Sciences, d Manchester, M13 9PT, United Kingdom

[3] School of Science and the Environment, Manchester Metropolitan University, Chester Street, Manchester, M1 5GD, United Kingdom

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

[5] Department of Animal and Plant Sciences, The University of Sheffield, Alfred Denny building, Sheffield, South Yorkshire, S10 2TN, United Kingdom

[6] School of the Environment, Natural Resources & Geography, Bangor University, Gwynedd, LL57 2UW, United Kingdom

[7] Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, Lai 40, Tartu 51005, Estonia

[8] Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO 80309, USA

[9] Climate Impacts Research Centre, Department of Ecology and Environmental Science, Umeå University, Vetenskapensväg 38, 981 07, Abisko, Sweden

[10] Environmental Sustainability Research Centre, University of Derby, Kedleston Road, Derby, DE22 1GB, United Kingdom

[11] Centre for Ecology and Hydrology, Wallingford, United Kingdom

[12] School of Life Sciences, University of Warwick, Coventry, CV4 7AL, United Kingdom

[13] Department of Agroecology and Environment, Agroscope, Zurich, Reckenholzstrasse 191, Switzerland

[14] Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, M13 9PT, United Kingdom

[15] Departamento de Biología y Geología, Física y Química Inorgánica, Escuela Superior de Ciencias Experimentales y Tecnología, Universidad Rey Juan Carlos, Calle Tulipán s/n, 28933 Móstoles, Spain

[16] Barnard College, Columbia University Department of Biology; Columbia University, USA

[17] School of Science and Engineering, Teesside University, Middlesbrough, TS1 3BX, United Kingdom

[18] Laboratory of Nematology, Wageningen University, Droevendaalsesteeg 1, Wageningen 6708 PB, The Netherlands

[19] Centre for Ecology and Hydrology, Bangor, LL57 2UW, United Kingdom

[20] Department of Biology, Duke University, Durham, NC, 27705, USA

[21] Natural England, United Kingdom

[22] Hawkesbury Institute for the Environment, Western Sydney University, Richmond 2753 NSW Australia

[23] Department of Biology, Boston University, Boston, MA, 02215, USA

[24] Institute of Biological and Environmental Sciences, University of Aberdeen, Saint Machar Drive, AB24 3UU, Aberdeen, United Kingdom

Keywords: microbial ecology, soil, diversity, community structure, Illumina sequencing, 16S rRNA gene, biogeography, microbiology, meta-analysis

The emergence of high-throughput DNA sequencing methods provides unprecedented opportunities to further unravel bacterial biodiversity and its worldwide role from human health to ecosystem functioning. However, in spite of the abundance of sequencing studies, combining data from multiple individual studies to address macroecological questions of bacterial diversity remains methodically challenging and plagued with biases. Here, using a machine learning approach that accounts for differences among studies and complex interactions among taxa, we merge 30 independent bacterial datasets consisting of 1,998 soil samples from across 21 countries. While previous meta-analysis efforts have focused on bacterial diversity measures or abundances of major taxa, we show that disparate amplicon sequence data can be combined at the taxonomy-based level to assess bacterial community structure. We find that rarer taxa are more important for structuring soil

communities than abundant taxa, and that these rarer taxa are better predictors of community structure than environmental factors, which are often confounded across studies. We conclude that combining data from independent studies can be used to explore novel patterns in bacterial communities, identify potential 'indicator' taxa with an important role in structuring communities, and propose new hypotheses on the factors that shape bacterial biogeography previously overlooked.

Soil microbial communities are more diverse and contain more individuals than any species groups on the planet[1,2]. Over the last decade, the use of high-throughput sequencing (HTS) methods has substantially advanced our understanding of the worldwide biogeography and ecology of soil bacterial and fungal communitie[3–6]. Recent work has further demonstrated that inclusion of microbial composition and functional attributes improves earth system models[7,8], which is of paramount importance for predicting effects of global change on ecosystem services such as climate regulation or soil fertility[9,10]. Yet, opposite to the long-standing view that every organism may occur everywhere[11], even at small scales bacterial communities turn out to be more patchy than previously expected[12,13], raising questions regarding dispersal constraints, temporal dynamics, and niche breadth at the global scale[14–17]. Due to these knowledge gaps, combined with practical challenges of exhaustive sample collection and the massive diversity of communities, global assessment of soil microbial diversity remains an ongoing research challenge[18,19].

For plants and animals, the integration of data from independent studies has been a valuable option for generating an understanding of global biogeography patterns, answering ecological questions (e.g. biodiversity-functioning relationships), and identifying threats to biodiversity from global changes[20–23]. Similarly, our understanding of soil microbial diversity would greatly improve from such worldwide assessments. However, the integration of microbial community HTS data from different studies is not so unlike the merging of museum species records where information and data is constrained by variations in nomenclature over space and time, among many other challenges[24,25]. Like plant and animal records, molecular microbial community records and information can be incomplete, processing and naming varies greatly between studies and over time[26], data storage is inconsistent, and there are few curated databases with high quality data (especially for short read sequences)[27,28]. Further, most microbial community data and metadata are still available only in independently published studies that have been carried out according to their own standards and procedures, and the extent of these confounding factors has never been quantified across studies.

Regardless of the challenges, as indicated by the many open access data initiatives[29–31], merging microbial sequence data is a potential option to address global scale questions, whether relating to the human microbiome[32], marine systems[33], or predicting the response of soil organisms to global environmental change[34]. For soil systems, the need to merge sequence data is supported by the emerging role of bacterial phyla and classes as indicators of particular soil conditions such as soil pH and nutrient concentrations[35,36]. Until now, attempts to meta-analyze sequence data have been limited to assessing diversity measures or abundances of major taxa, because the merging of community data is constrained by methodological differences between sequencing studies[13,30,37–39]. However, a recent systematic review found that measures of microbial community structure were more often linked to microbial process rates than diversity or presence/absence data[40], and abundance ratios among phyla may be less important than previous believed[41]. Together indicating that information on variation in microbial community structure is potentially more ecologically relevant than measures of diversity and abundances of major taxa.

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

Results and Discussion

Taxonomy-based merging of disparate amplicon sequence data

We identified 30 individual HTS bacterial studies from 21 countries for our analysis (Figure 1A and Supplementary Table 1). While we aimed to merge HTS data of both soil bacterial and fungal datasets, our approach was only successful for bacterial data (Figure 1B and 1C), and highlights the well-known dilemma of fungal databases, where extremely high diversity combined with high endemism and mismatched taxonomy across continents make merging data by taxonomy difficult and unusable for downstream analyses[4,5,42]. For the bacterial studies, we were able to successfully merge 30 individual OTU tables; using a taxonomy-based approach, datasets were merged using the taxonomic affiliations of individual OTUs. Once filtered, and singletons removed, the final ‘taxonomy-based’ community contained 1,998 individual soil samples, and 8,287 taxa. Here ‘taxon’ is defined as a unique name in the classification; where a name could be a specific phylum, genus, or other taxonomic level. For example, ‘Acidovorax’ (genus) and Proteobacteria (the phylum containing Acidovorax) were both considered as taxa). To account for variation in sequencing depth between different studies, OTU relative abundances were used per sample, rather than absolute read abundance. To test known biogeographical patterns, metadata (information on geographical location, soil pH and soil core measurements) were compiled for all studies. Technical and methodical information was also collected; all of these 30 studies had conducted amplicon sequencing on hypervariable regions of the 16S rRNA gene in soil samples using either Illumina or (Roche) 454 pyrosequencing (with any primer pair) (Supplementary Table 1). For a validation step we retrieved all usable raw sequence data available, resulting in 417 samples from locations across the globe (approximately 1/5 of all our samples) (Figure 1A). Data not included in this sequence-matched analysis either had an incompatible raw sequence format or simply no longer existed. Available raw sequence data were combined into a single ‘sequence-matched’ community comprising 44,106 OTUs (Supplementary Figure 1).

Machine learning assessment of bacterial community structure

Ordination of the taxonomy-based community reveals large amounts of structure both within and between studies (structure that is removed by permuting taxa among samples (Supplementary Figure 2), without greatly affecting diversity (Supplementary Table 3)), and the observation of the well-established negative relationship between relative abundance of Acidobacteria and soil pH (Figure 1D)[43] confirms our merging method. This visualization also suggests that some of the community variation (e.g. the near absence of Acidobacteria in some studies, even at low pH) is due to technical factors such as the particular primer sets chosen, region sequenced, and sequencing platform (Supplementary Methods and Supplementary Table 2). However, we expect that some taxa are not correlated with technical factors, and are non-randomly distributed with respect to biotic and abiotic factors. Therefore, using a machine learning approach capable of accounting for complex interactions among taxa (Random Forests^{TM}, see methods), we determined the extent to which individual taxa could influence the community structure of merged independent studies. Here community structure is defined by the presence and relative abundances of individual taxa, along with co-occurrence relationships between those taxa. This was done in two ways: first, we constructed a model that classified the study from which a sample came based on the proportions of the 8,287 taxa it contained (1.5% [± 0.02% CI] classification error, by internal cross-validation). Second, we determined the contribution of each taxon to bacterial community structure by quantifying its importance in a model that separated the observed data from synthetic data randomly drawn from the observed distributions of relative abundances for each taxon [44,45] (*see Methods*).

Merging of disparate microbial sequence data is known to be plagued with potential biases including: lack of standardization of sample collection, methodological issues regarding DNA extraction and primer choice, incomplete metadata, the technical biases of different sequencing platforms, sequencing depth, PCR Bias, different clustering methods, and the use of different taxonomic classification pipelines[46–52]. We therefore took the novel step to quantify the importance of both technical and environmental factors alongside taxa in the Random Forests models (Figure 2). Of note, ‘owner’, which encompasses the technical biases and uniqueness of a given dataset, is very effective for differentiating between studies (i.e. the owner is far to the right in Figure 2) yet is entirely uninformative about community structure (i.e. owner is at the far bottom in Figure 2). In fact, *all* technical factors included are better than 98.5% of all taxa to differentiate between studies, indicating that the observed differences among studies in taxon relative abundances are strongly confounded with technical factors. Independent of taxonomy, certain environmental factors, such as country of origin, latitude and longitude, and soil pH, were highly important in differentiating studies but not in determining community structure. By contrast, minimum soil sampling depth was not very important in separating studies, and was more associated with community structure. It is well known that bacterial diversity decreases with soil depth[53] and our results show that in a global assessment, soil depth remains a strong predictor of bacterial community composition. Perhaps most useful for future research, this result highlights that not all environmental factors are equally confounded by technical factors, and shows that by combining data from across many independent studies we may identify previously overlooked taxa and factors relevant for structuring communities.

Importance for structuring soil bacterial communities

Although all studies were confounded by technical and environmental covariates, there remained many taxa that were non-randomly distributed and were not confounded with technical differences among studies (upper left in Figure 2). When assessing the role of these different taxa in structuring the community, we found a trade-off between taxon abundance and importance in community structure, such that low abundance taxa are disproportionately important in the non-random structure of communities, where the most important taxa are rarer than expected compared to the randomly permuted data (Figure 3). Thus, the importance of taxa for determining community structure is negatively correlated with the average abundance of those taxa, whereas taxon abundance is positively correlated with importance for separating studies (ρ = -0.79 and ρ = +0.51 respectively, rank correlation, cf. null expectations of ρ = -0.62 and -0.12 respectively in permuted data). The taxa most closely associated with differences between studies tend to be those present at or greater than 0.1% relative abundance, but those most important in determining community structure tend to be present at 0.0001% abundance or less (with a null expectation of around 0.01-0.001% in each case, Figure 3). This result is only found by considering the full set of studies and is neither apparent within single studies (Supplementary Fig. 4A-B) nor a subset of studies (whether matched by name or sequence Supplementary Fig. 5). It corresponds to the long tail in frequency-abundance distributions of soil microbial communities[54], where many taxa in the soil are known to occur at low abundance. Thus, if rarer taxa tend to be more important for distinguishing between communities, it is within this long tail that we might identify taxa that could indicate ecological or functional differences among soil communities[33,55,56].

To be ecological indicators[57,58], taxa need to vary in abundance in response to environmental factors and have high occurrence across studies, as is the case for the phylum Acidobacteria[43]. Acidobacteria, however, are typically abundant and our analysis suggests that the most abundant taxa are *not* the most important in determining community structure. While dominant taxa like Acidobacteria do change with environmental factors such as pH (Figure 1D), those changes are of lesser importance for the 'non-randomness' of community structure, and more confounded with technical effects, than changes in less dominant, pH responsive taxa (Supplementary Figure

3A). Therefore, we assessed which taxonomic ranks are more or less distinguished from the randomly permutated data. Although differences among domains and phyla are strongly associated with differences among studies (Figure 4B) only taxa at a rank lower than phyla are consistently better than random at identifying community structure (Figure 4A).

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

To be a good ecological indicator a taxon should occur in most studies; we therefore looked explicitly at the relationship between a taxon’s importance in community structure and its occurrence across studies. Low abundance taxa and taxa of lower taxonomic rank are consistently important in determining community structure, but tend to be detected in fewer studies (ρ = 0.59 and 0.31 respectively Supplementary Figure 3B and 3C). We discovered a novel relationship between taxon occurrence across studies and importance for structuring communities for all taxa (Figure 5, Supplementary Table 4). Comparison with the null expectation reveals a range of taxa, occurring in multiple samples from most studies, which are much more important in determining community structure than expected by chance. A similar pattern is apparent in the sequence-matched dataset (Supplementary Figure 8A) and the same subset of studies when taxonomy-matched (Supplementary Figure 8B). Altogether, the analyses clearly illustrate the significance of taxonomic rank, for example *class* Gemmatimonadetes is relatively unimportant for community structure but *genus* Gemmatimonadetes is relatively important. The result also shows rarer taxa being more important in structuring communities and suggests rarer bacterial taxa play overlooked ecologically important roles for bacterial community dynamics[56]. This result is robust to artifacts caused by the rarest taxa (e.g. differences between 0 and 1 reads in a sample could be significant for a model, without being biologically significant) – a very similar pattern is seen when only taxa present at above 0.003% in any given sample were included in this analysis (typically removing the rarest 10% of taxa from any given sample, Supplementary Figure 9). Conversely, many taxa of high taxonomic rank with high occurrence across samples, such as the phyla Actinobacteria, Acidobacteria, Proteobacteria, and Bacteroidetes, were much less important for community structure than the null expectation. These taxa have been reported elsewhere as ‘core’ members of the soil community[43,59,60], and even been included in source-tracking of microbial communities due to their ubiquitous presence in soil[61]. Yet, it is the consistent presence of the core taxa across samples and studies that makes them inadequate for assessing community structure.

Conclusions

Our results demonstrate the power of combining global bacterial HTS data from multiple independent sources for the detection of biogeographical patterns and for identifying community patterns that can be used to generate hypotheses on the roles of certain taxa. Though our assessment was on soil communities, our methods can be applied to broadly to other microbial datasets and disciplines. Taxonomy-based merging gives results that are consistent with raw sequence data, and expands opportunities for extracting information about microbial communities from the wealth of existing and future studies. Moreover, we find that rarer bacterial taxa are more important in differentiating communities than previously assumed, and hold potential as overlooked soil indicators or keystone species. Still, there are considerable challenges associated with merging large sequence datasets beyond the well-known biases that accompany any molecular HTS study. Perhaps the most concerning was that so few raw sequence datasets for publically deposited analyses could be retrieved. This highlights the need for wider community adoption of open and accessible short read sequence databases[62], open reference clustering[63], standardized databases[64] and—as always—that metadata should be consistent and accessible. Regardless of these challenges, as HTS methods rapidly advance we must find ways to simultaneously curate and carry our research knowledge forward[32]. Only then, in combination with the many novel and classical approaches, can we uncover the full breadth of soil diversity and the roles soil microbes play for ecosystem processes.

Methods:

Description of datasets:

Metadata from the 30 studies and 1998 samples were collected and compiled into a summary data file. To do so, we standardized the metadata of each study using the dplyr package (Wickham & Francois, 2016) of the R statistical platform (R Core Team, 2016). Samples were collected from 21 counties representing all continents except Antarctica. In addition to location and pH data (median = 6.1, quartile range=5.3-7.0), which were available from all studies, information on altitude (10 m, 10-860 m), soil moisture (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%, 1.9-7.5%) and phosphorus (20.7 mg kg^{-1}, 7.0-223.0 mg kg^{-1}) was noted where available. Depth of sample collection was also noted and ranged from surface collections to a maximum depth of 70 cm, with 83% of samples originating from 0-10 cm below the soil surface. Samples represented anthropogenically managed (59%) and natural (40%; remaining samples undefined) systems, and were taken from arable, grassland, peatland, forest, scrub (including tundra) and urban habitats. The majority of samples (71%) were described as non-experimental, meaning no treatments were applied, with the remainder described as experimental. Sequencing data were either produced using Roche 454 technology (22%) or one of the Illumina platforms (78%). Primer pairs were defined for 92% of the samples and nine different pairs were identified from the study meta data (27F:338R; 341F:518R; 341F:806R; 341F:907R; 357F:926R; 515F:806R; 577F:926R; 799F:1193R and 341F:805R) with the majority of samples (66%) using 515F and 806R to produce amplicons. Post sequencing processing varied, but 81% of samples were run through the QIIME workflow at some point. An OTU table for 1 study comprising 43 samples was programmatically retrieved from the MG-RAST public metagenome repository[65]. Taxonomy for the different studies was mainly assigned using the Greengenes database (84 %), but RDP (6 %;[46] and the Silva database (9 %)[66] were also used.

Merging OTU tables:

For the OTU tables from the 30 individual studies to be merged, extensive data cleaning was carried out on the OTU and taxonomy files to maximize the possibility of matching taxa across datasets. This comprised several steps: (1) Most datasets contained a seven-level taxonomy, recorded in a variety of ways, which was converted to a standardized format. (2) Individual taxon names were cleaned, to give a single name at each taxonomic level (e.g. removing special characters and extra annotations, such as ‘candidate division’ or details of containing taxa). (3) For the many cases where a taxon was not assigned at a particular taxonomic level, a unified ‘unassigned’ label was created. Repeating analyses with all these taxa removed made no qualitative difference to the results (Supplementary Figure 10). Merging at the taxonomy-based level has the added benefit of lessening the impacts of hypervariable regions. For example, the identification of an organism at a specific level in one sample also contributes to the identification of the containing genus for that sample, allowing direct comparison with a sample where, because a different region was sequenced, that same organism is only resolved to the genus level. Next, relative abundance data were, where necessary, re-scaled to sum to 1 for a sample, using original OTU count files where possible. These values were then manipulated to give data tables usable for modeling using custom R scripts. For some analyses (Figures 3-5), a dataset without community structure was created by randomly permuting the relative abundance of each taxon across all samples. Unless otherwise stated, the analyses performed on the permuted dataset was identical to that performed on the observed data.

Merging raw sequence data and other validation datasets:

While no dataset can currently provide a "ground truth" against which to judge our approach, we can at least validate it. The primary validation of our taxonomy-matching approach was to merge raw sequence data ('sequence-matched') from five studies (Supplementary Table 1). Per sample fastq files were obtained for each individual dataset. Read files were quality filtered with sickle [67] for single end reads trimming bases below phred score 36 and shorter than 100bp. These stringent filtering criteria were applied to keep only high quality reads and to make sure it is possible to map reads to full length 16S rRNA gene sequences. Full length 16S rRNA gene sequences from the Silva 119 release [66] were obtained in Qiime compatible format from the Silva Download Archive For each dataset, all reads were mapped to the full length 16S rRNA gene sequences using the usearch global algorithm implemented in VSEARCH version 1.9.6 [68]. The alignment results in usearch table format (uc) were directly converted to BIOM format using biom version 2.1.5 [69]. Consensus/majority taxonomy was added as metadata to the biom file. Finally, all BIOM files of each dataset were merged using Qiime version 1.9.1 [70]. All steps were implemented in a workflow made with Snakemake version 3.5.4 [71] available: (De Hollander 2016). See Supplementary Fig 1 for workflow.

To use this sequence-matched dataset to validate our taxonomy-matching approach across studies using different taxonomy databases (Supplementary Figures 5, 7 & 8) we created an equivalent taxonomy-matched dataset from the same 5 studies. As with the full dataset, only taxa occurring in at least two studies were included in either this or the sequence-matched dataset. To test what is gained or lost by considering different numbers of studies simultaneously, we considered, not only the full dataset (30 studies) and the subset of 5 studies used in the sequence-matched dataset, but two of the largest individual studies: from Central Park, NYC encompassing 594 samples (study #24) and a global dataset encompassing 103 samples (study #30). In each case a simple subset of the full dataset was analyzed (Supplementary Figure 4). To address PCR biases (Supplementary Table 2) and biases associated with rare taxa, we created a filtered subset of the data where only taxa present at above 0.003% in any given sample were considered, meaning that all taxa deemed present are represented by multiple sequence reads (Supplementary Figure 9). To address the issue of differential 16S copy numbers skewing abundance estimates, we created a binary dataset of the presence/absence of all taxa. The results for a model separating studies using this dataset were very similar to the main dataset using relative abundance, however, there was insufficient power to identify taxa important for community structure (Supplementary Figure 6). Nonetheless, this analysis did agree with the main analysis that phyla were the most stable taxonomic level, with lower importance than on the permuted data (Supplementary Figure 6). Finally, to test the effect of 'unknown' or unclassified bacterial taxa we created a reduced dataset where all taxa classified as 'unassigned' at any level were removed (Supplementary Figure 10).

Random forest models.

To test for the importance of different taxa in the structuring of the data we used Random Forest models [45,72] with the relative abundances of the taxa as explanatory variables. Random Forest models have two principal advantages in this context: 1) they can deal easily with thousands of explanatory variables and quantify their relative importance, and 2) they can run equivalently in both supervised and un-supervised modes. In the latter, the importance of a variable describes how effective it is at separating the observed data from randomized synthetic data[44]. In both cases, a proximity matrix may be generated, which can be used for ordination (Supplementary Figure 2). The importance of individual taxa in a Random Forest relate to traditional ecological measures. For instance, the importance in a supervised model, such as that used separating studies (x-axis in Figure 2) is closely correlated with the sensitivity component of the indicator value of each taxon (ρ = 0.89, Supplementary Figure 3D)[58]. There are two key parameters that may be adjusted in a Random Forest model, *mtry*, the number of variables randomly sampled as candidates for a split in the constituent trees and *ntree*, the number of trees in the forest. *mtry* was set at its default value (square root of the number of variables) *ntree* was set to 100,000 for each forest. Such a large number of trees was found to be necessary to achieve stable importance across taxa and was achieved by combining several forests run in parallel without normalizing votes. Other parameters were left at default values, in particular, trees were grown to completion (i.e. a minimum node size of 1). The un-scaled permutation importance of variables is used throughout: Each variable importance is the difference between the classification error rate of a tree on data not used to construct it (the ‘out of bag’ data) and the same error following random permutation of the variable in question, averaged over all trees.

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

All analyses were completed with RandomForest package for R version 4.6.

Acknowledgements: We thank all the people who contributed data and input to this study. This study was conducted at a workshop (5/2015, Manchester, UK) funded by the British Ecological Society's special interest group Plants-Soils-Ecosystems and organized by FTDV and KSR. This study and participants were funded in part by ERC Adv grant 26055290 (KSR); BBSRC David Phillips Fellowship (BB/L02456X/1) (FTDV); ERC Grant Agreements 242658 [BIOCOM] and 647038 [BIODESERT] (FTM); the European Regional Development Fund (*Centre of Excellence EcolChange*) (JD); Yorkshire Agricultural Society, Nafferton Ecological Farming Group, and the Northumbria University Research Development Fund (CHO); BBSRC Training Grant (BB/K501943/1) (CH); Wallenberg Academy Fellowship (KAW 2012.0152), Formas (214-2011-788) and Vetenskapsrådet (612-2011-5444) (ED); Data taken from work carried out in collaboration with CJS, CL, JMC and SPC. Computing was facilitated by the University of Manchester Condor pool and the CLIMB infrastructure (www.climb.ac.uk).

References

1. Prosser, J. I. Dispersing misconceptions and identifying opportunities for the use of 'omics' in soil microbial ecology. *Nat. Rev. Microbiol.* **13,** 439–46 (2015).

2. Bardgett, R. D. & van der Putten, W. H. Belowground biodiversity and ecosystem functioning. *Nature* **515,** 505–511 (2014).

3. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6,** 1621–4 (2012).

4. Tedersoo, L. *et al.* Global diversity and geography of soil fungi. *Science.* **346,** (2014).

5. Bik, H. M. *et al.* Sequencing our way towards understanding global eukaryotic biodiversity. *Trends Ecol. Evol.* **27,** 233–243 (2012).

6. Davison, J. *et al.* Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism. *Science.* **349,** (2015).

7. Wieder, W. R., Bonan, G. B. & Allison, S. D. Global soil carbon projections are improved by modelling microbial processes. *Nat. Clim. Chang.* **3,** 909–912 (2013).

8. Allison, S. D., Wallenstein, M. D. & Bradford, M. A. Soil-carbon response to warming dependent on microbial physiology. *Nat. Geosci.* **3,** 336–340 (2010).

9. Karhu, K. *et al.* Temperature sensitivity of soil respiration rates enhanced by microbial community response. *Nature* **513,** 81–84 (2014).

10. Wieder, W. R., Cleveland, C. C., Smith, W. K. & Todd-Brown, K. Future productivity and carbon storage limited by terrestrial nutrient availability. *Nat. Geosci.* **8,** 441–444 (2015).

11. Barberán, A., Casamayor, E. O. & Fierer, N. The microbial contribution to macroecology. *Front. Microbiol.* **5,** 203 (2014).

12. Ramirez, K. S. *et al.* Biogeographic patterns in below-ground diversity in New York City’s Central Park are similar to those observed globally. *Proc. Biol. Sci.* **281,** 20141988- (2014).

13. O’Brien, S. L. *et al.* Spatial scale drives patterns in soil bacterial diversity. *Environ. Microbiol.* **18,** 2039–2051 (2016).

14. Evans, S., Martiny, J. B. H. & Allison, S. D. Effects of dispersal and selection on stochastic assembly in microbial communities. *ISME J.* **11,** 176–185 (2017).

15. Talbot, J. M. *et al.* Endemism and functional convergence across the North American soil mycobiome. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 6341–6 (2014).

16. Barber, A. *et al.* Why are some microbes more ubiquitous than others? Predicting the habitat breadth of soil bacteria. *Ecol. Lett.* **17,** 794–802 (2014).

17. Fierer, N. *et al.* Reconstructing the microbial diversity and function of pre-agricultural tallgrass prairie soils in the United States. *Science* **342,** 621–4 (2013).

18. Ma, B. *et al.* Geographic patterns of co-occurrence network topological features for soil microbiota at continental scale in eastern China. *ISME J.* **10**, 1891-1901 (2016).

19. Ranjard, L. *et al.* Turnover of soil bacterial diversity driven by wide-scale environmental heterogeneity. *Nat. Commun.* **4,** 1434 (2013).

20. Jetz, W., McPherson, J. M. & Guralnick, R. P. Integrating biodiversity distribution knowledge: toward a global map of life. *Trends Ecol. Evol.* **27,** 151–9 (2012).

21. Ricketts, T. H. *et al.* Disaggregating the evidence linking biodiversity and ecosystem services. *Nat. Commun.* **7,** 13106 (2016).

22. Dirzo, R. *et al.* Defaunation in the Anthropocene. *Science..* **345,** 401–406 (2014).

23. Gerstner, K., Dormann, C. F., Stein, A., Manceur, A. M. & Seppelt, R. Effects of land use on plant diversity - A global meta-analysis. *J. Appl. Ecol.* **51,** 1690–1700 (2014).

24. Patterson, D. J., Cooper, J., Kirk, P. M., Pyle, R. L. & Remsen, D. P. Names are key to the big new biology. *Trends Ecol. Evol.* **25,** 686–691 (2010).

25. Santos, A. M. & Branco, M. The quality of taxonomy-based species records in databases. *Trends Ecol. Evol.* **27,** 6-7-8 (2012).

26. Beiko, R. G. Microbial Malaise: How Can We Classify the Microbiome? *Trends Microbiol.* **23,** 671–679 (2015).

27. Tedersoo, L. *et al.* Standardizing metadata and taxonomic identification in metabarcoding studies. *Gigascience* **4,** 34 (2015).

28. Ramirez, K. S. *et al.* Toward a global platform for linking soil biodiversity data. *Front. Ecol. Evol.* **3,** (2015).

29. Turner, W. *et al.* Free and open-access satellite data are key to biodiversity conservation. *Biol. Conserv.* **182,** 173–176 (2015).

30. Gilbert, J. A., Jansson, J. K. & Knight, R. The Earth Microbiome project: successes and aspirations. *BMC Biol.* **12,** 69 (2014).

31. Joppa, L. N. *et al.* Filling in biodiversity threat gaps. *Science.* **352,** (2016).

32. Sinha, R., Abnet, C. C., White, O., Knight, R. & Huttenhower, C. The microbiome quality control project: baseline study design and future directions. *Genome Biol.* **16,** 276 (2015).

33. Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. U. S. A.* **103,** 12115–20 (2006).

34. García-Palacios, P. *et al.* Are there links between responses of soil microbes and ecosystem functioning to elevated CO2, N deposition and warming? A global perspective. *Glob. Chang. Biol.* **21,** 1590–1600 (2015).

35. Hermans, S. M. *et al.* Bacteria as Emerging Indicators of Soil Condition. *Appl. Environ. Microbiol.* **83,** AEM.02826-16 (2017).

36. Philippot, L. *et al.* The ecological coherence of high bacterial taxonomic ranks. *Nat. Rev.*

Microbiol. **8,** 523–529 (2010).

37. Shade, A., Caporaso, J. G., Handelsman, J., Knight, R. & Fierer, N. A meta-analysis of changes in bacterial and archaeal communities with time. *ISME J.* **7,** 1493–506 (2013).

38. Delgado-Baquerizo, M., Grinyer, J., Reich, P. B. & Singh, B. K. Relative importance of soil properties and microbial community for soil functionality: insights from a microbial swap experiment. *Functional Ecology*. **30**, 1862-1873 (2016).

39. Hendershot, J. N., Read, Q. D., Henning, J. A., Sanders, N. J. & Classen, A. T. Consistently inconsistent drivers of microbial diversity and abundance at macroecological scales. *Ecology* **98,** 1757–1763 (2017).

40. Bier, R. L. *et al.* Linking microbial community structure and microbial processes: an empirical and conceptual overview. *FEMS Microbiol. Ecol.* **91,** (2015).

41. Walters, W. A., Xu, Z. & Knight, R. Meta-analyses of human gut microbes associated with obesity and IBD. *FEBS Lett.* **588,** 4223–4233 (2014).

42. Kõljalg, U. *et al.* Towards a unified paradigm for sequence-based identification of fungi. *Mol. Ecol.* **22,** 5271–5277 (2013).

43. Lauber, C. L., Hamady, M., Knight, R. & Fierer, N. Pyrosequencing-Based Assessment of Soil pH as a Predictor of Soil Bacterial Community Structure at the Continental Scale. *Appl. Environ. Microbiol.* **75,** 5111–5120 (2009).

44. Shi, T. & Horvath, S. Unsupervised Learning With Random Forest Predictors. *J. Comput. Graph. Stat.* **15,** 118–138 (2006).

45. Breiman, L. & Cutler, A. Random Forests Manual v4.0. *Technical report, UC Berkeley* (2003).

46. McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological

and evolutionary analyses of bacteria and archaea. *ISME J.* **6,** 610–8 (2012).

47. Cole, J. R. *et al.* Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42,** D633-42 (2014).

48. Werner, J. J., Zhou, D., Caporaso, J. G., Knight, R. & Angenent, L. T. Comparison of Illumina paired-end and single-direction sequencing for microbial 16S rRNA gene amplicon surveys. *ISME J.* **6,** 1273–1276 (2011).

49. Lozupone, C. & Stombaugh, J. Meta-analyses of studies of the human microbiota. *Genome Res.* **10**, 1704-14 (2013).

50. Pawluczyk, M. *et al.* Quantitative evaluation of bias in PCR amplification and next-generation sequencing derived from metabarcoding samples. *Anal. Bioanal. Chem.* **407,** 1841–1848 (2015).

51. Smith, D. P., Peay, K. G., Palmer, M., Gillikin, C. & Keefe, D. Sequence Depth, Not PCR Replication, Improves Ecological Inference from Next Generation DNA Sequencing. *PLoS One* **9,** e90234 (2014).

52. Zhbannikov, I. Y. & Foster, J. A. MetAmp: combining amplicon data from multiple markers for OTU analysis. *Bioinformatics* **31,** 1830–1832 (2015).

53. Lu, X., Seuradge, B. J. & Neufeld, J. D. Biogeography of soil Thaumarchaeota in relation to soil depth and land usage. *FEMS Microbiol. Ecol.* **93,** (2017).

54. Jung, S. P. & Kang, H. Assessment of microbial diversity bias associated with soil heterogeneity and sequencing resolution in pyrosequencing analyses. *J. Microbiol.* **52,** 574–580 (2014).

55. Langille, M., Zaneveld, J. & Caporaso, J. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature* (2013).

56. Jousset, A. *et al.* Where less may be more: how the rare biosphere pulls ecosystems strings. *ISME J.* **11,** 853–862 (2017).

57. Hermans, S. M. *et al.* Bacteria as emerging indicators of soil condition. *Appl. Environ. Microbiol.* AEM.02826-16 (2016).

58. Cáceres, M. De & Legendre, P. Associations between species and groups of sites: indices and statistical inference. *Ecology* **90,** 3566–3574 (2009).

59. Maestre, F. T. *et al.* Increasing aridity reduces soil microbial diversity and abundance in global drylands. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 15684–9 (2015).

60. Fierer, N., Bradford, M. a & Jackson, R. B. Toward an ecological classification of soil bacteria. *Ecology* **88,** 1354–64 (2007).

61. Knights, D. *et al.* Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* **8,** 761–763 (2011).

62. Muir, P. *et al.* The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.* **17,** 53 (2016).

63. Rideout, J. R. *et al.* Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ* **2,** e545 (2014).

64. Yilmaz, P. *et al.* The genomic standards consortium: bringing standards to life for microbial ecology. *ISME J.* **5,** 1565–7 (2011).

65. Wilke, A. *et al.* The MG-RAST metagenomics database and portal in 2015. *Nucleic Acids Res.* **44,** D590–D594 (2016).

66. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41,** D590-6 (2013).

67. Joshi & Fass, J. Sickle: A sliding-window, adaptive, quality-based trimming tool for

FastQ files. (2011).

68. Rognes, T. *et al.* vsearch: VSEARCH 1.9.6. (2016).

69. McDonald, D. *et al.* The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *Gigascience* **1,** 7 (2012).

70. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Meth* **7,** 335–336 (2010).

71. Koster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* **28,** 2520–2522 (2012).

72. Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32 (2001).

Hadley Wickham and Romain Francois (2016). dplyr: A Grammar of Data Manipulation. R package version 0.5.0. https://CRAN.R-project.org/package=dplyr

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Figures:

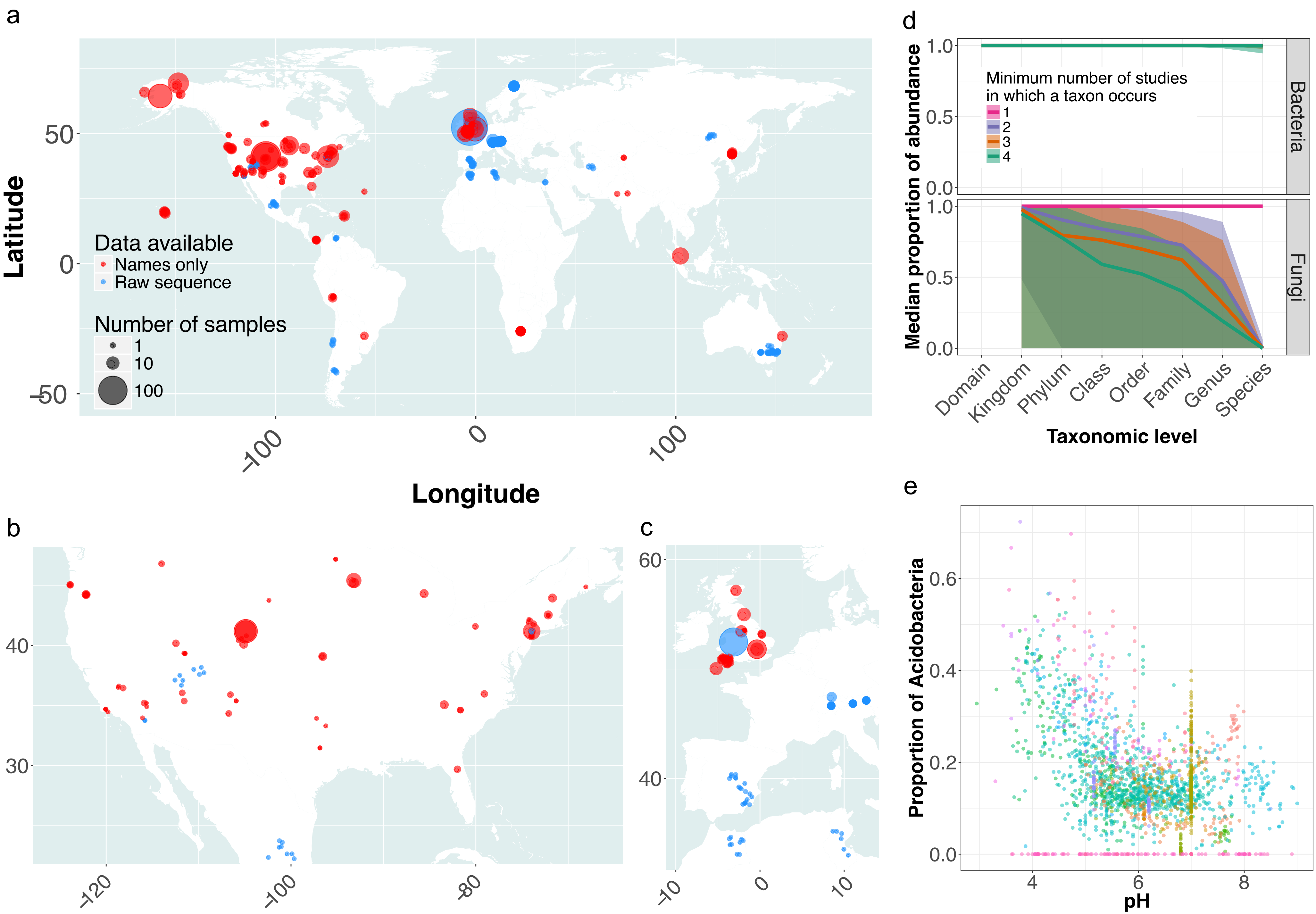
Figure 1. Merging of data from 32 independent studies demonstrates wide geographic breadth, community variation, and confirms the well-known importance of soil pH. A. Map of locations from which samples were collected, with zoom panels on the United States (left) and western Europe (right). Points in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

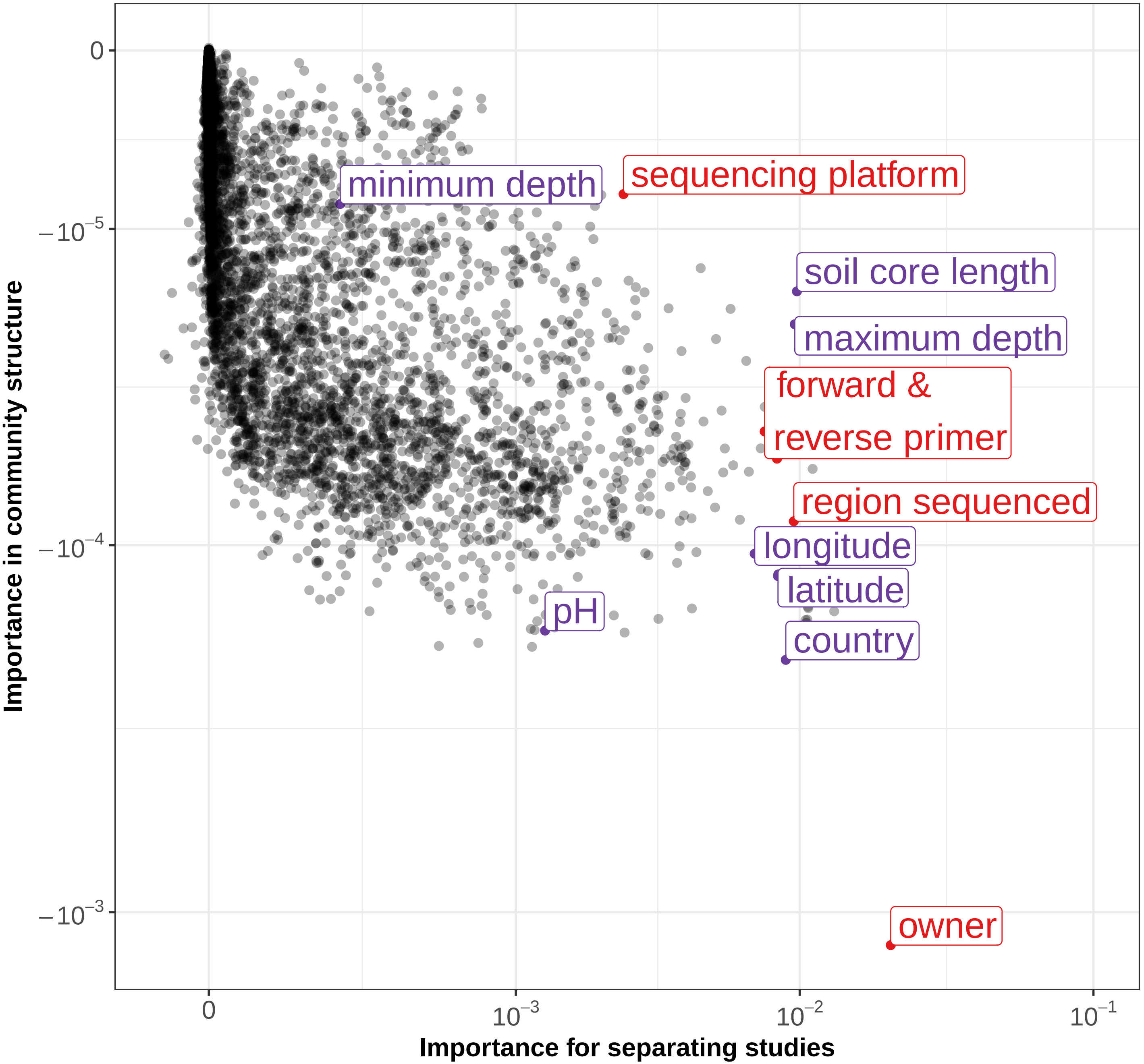
Figure 2: Regardless of technical differences between studies, many bacterial taxa are still informative about bacterial community structure. Machine learning models classify the study from which samples came (x-axis) based on the relative abundance of taxa within samples and distinguish the observed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

Figure 3: Rarer taxa are more important for structuring communities than abundant taxa. Here we show the thousand most important bacterial taxa in community structure (A) and in separating studies (B) with respect to their average relative abundance across samples. Plotted are the 'observed' points (green) and 'permuted' points (orange) which are a null distribution from performing the same analysis on a permuted dataset (see *Methods*). The y-axis reports the rank variable importance in the Random Forests model of community structure (see *Methods*), i.e. the taxon with the greatest importance in this model is ranked 1, the second greatest 2, etc.

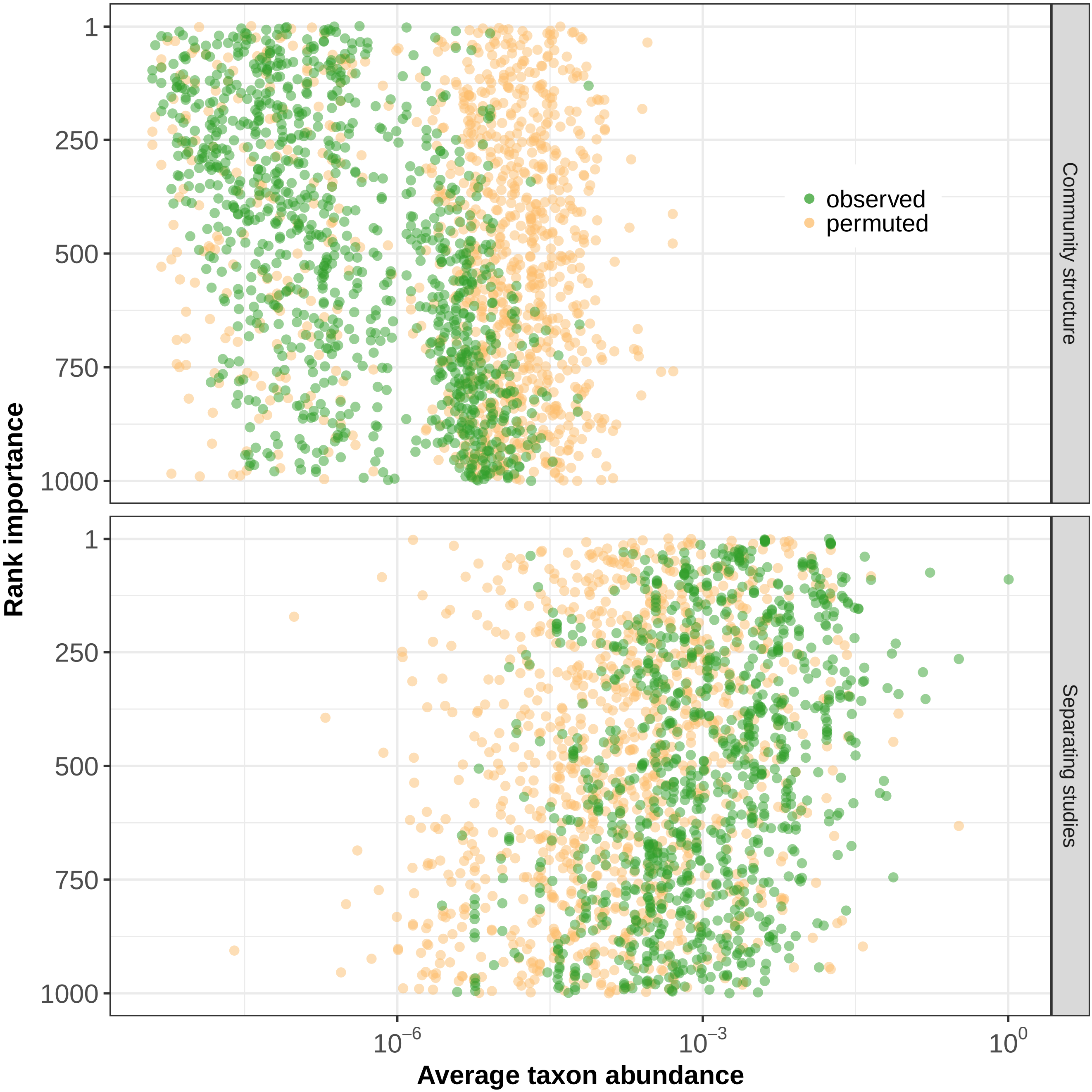
Figure 4: The importance of bacterial taxa classified at different taxonomic ranks. Lower taxonomic rank is more important for community structure (A), while high taxonomic rank is more important for separating studies (B). For each taxon, the difference was calculated between the variable importance (see *Methods*) of that taxon in a Random Forests model of either community structure or separating studies and the equivalent value from an analysis performed on the permuted dataset (see *Methods*). The lines and grey ribbons show the mean and standard error respectively of these values across taxa at each taxonomic r considered.

Figure 5: Importance of bacterial taxa in community structure related to their occurrence in different studies. The y-axis reports the variable importance in the Random Forests model of community structure (see *Methods*). Green 'observed' points correspond to those taxa shown in Figure 1. Orange 'permuted' points correspond to the same analysis on a null distribution (see *Methods*). Lines are general additive model (gam) smoothers. Each line is shown with a confidence interval (grey); where this is not visible it is narrower than the line it surrounds.

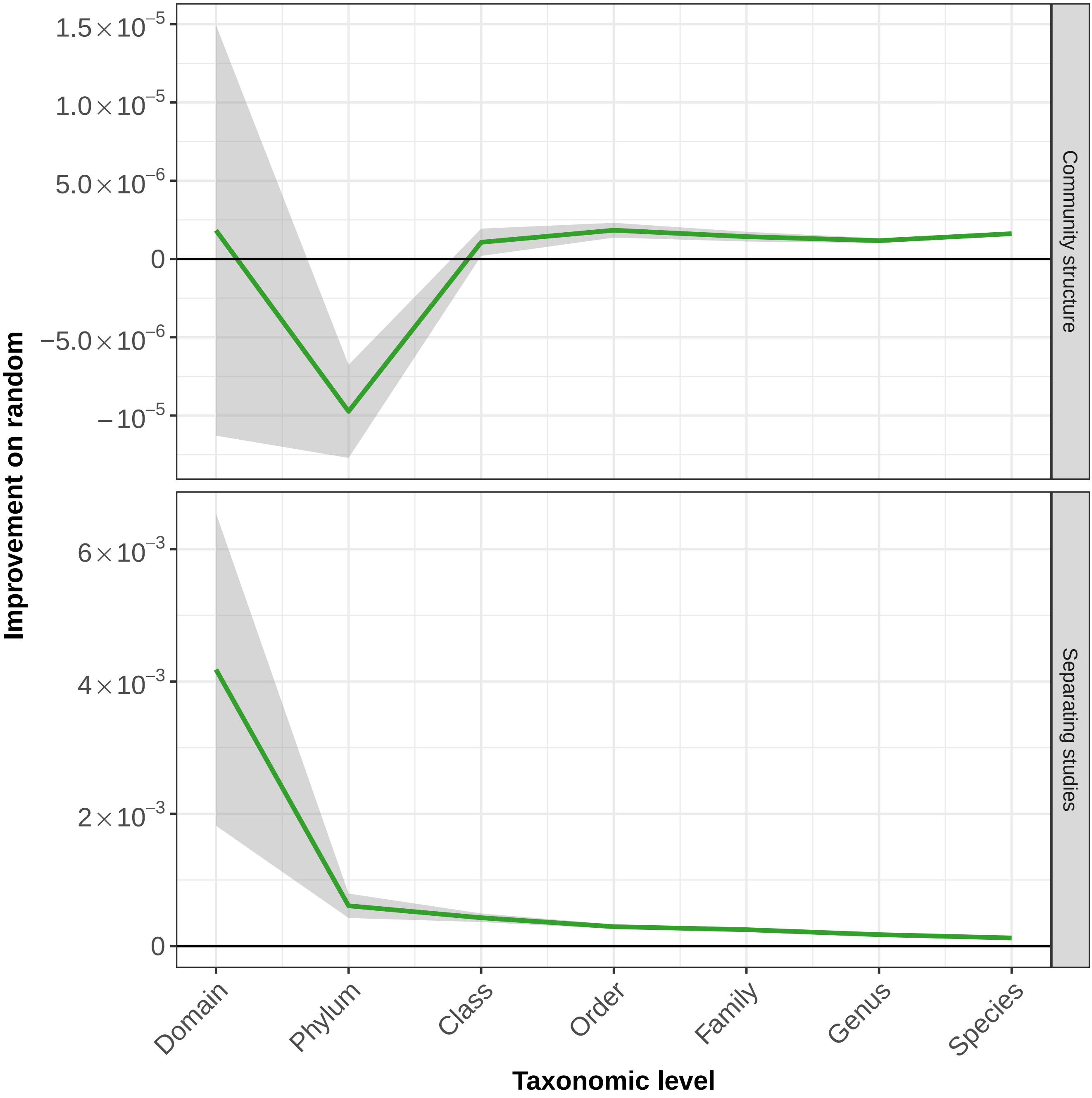

a
Latitude
50
0
-50
-100
0
100
Longitude
Data available
Names only
Raw sequence
Number of samples
1
10
100
b
40
30
-120
-100
-80
c
60
40
-10
0
10
d
Median proportion of abundance
1.0
0.5
0.0
Minimum number of studies
in which a taxon occurs
1
2
3
4
Bacteria
Fungi
Domain
Kingdom
Phylum
Class
Order
Family
Genus
Species
Taxonomic level
e
Proportion of Acidobacteria
0.6
0.4
0.2
0.0
4
6
8
pH



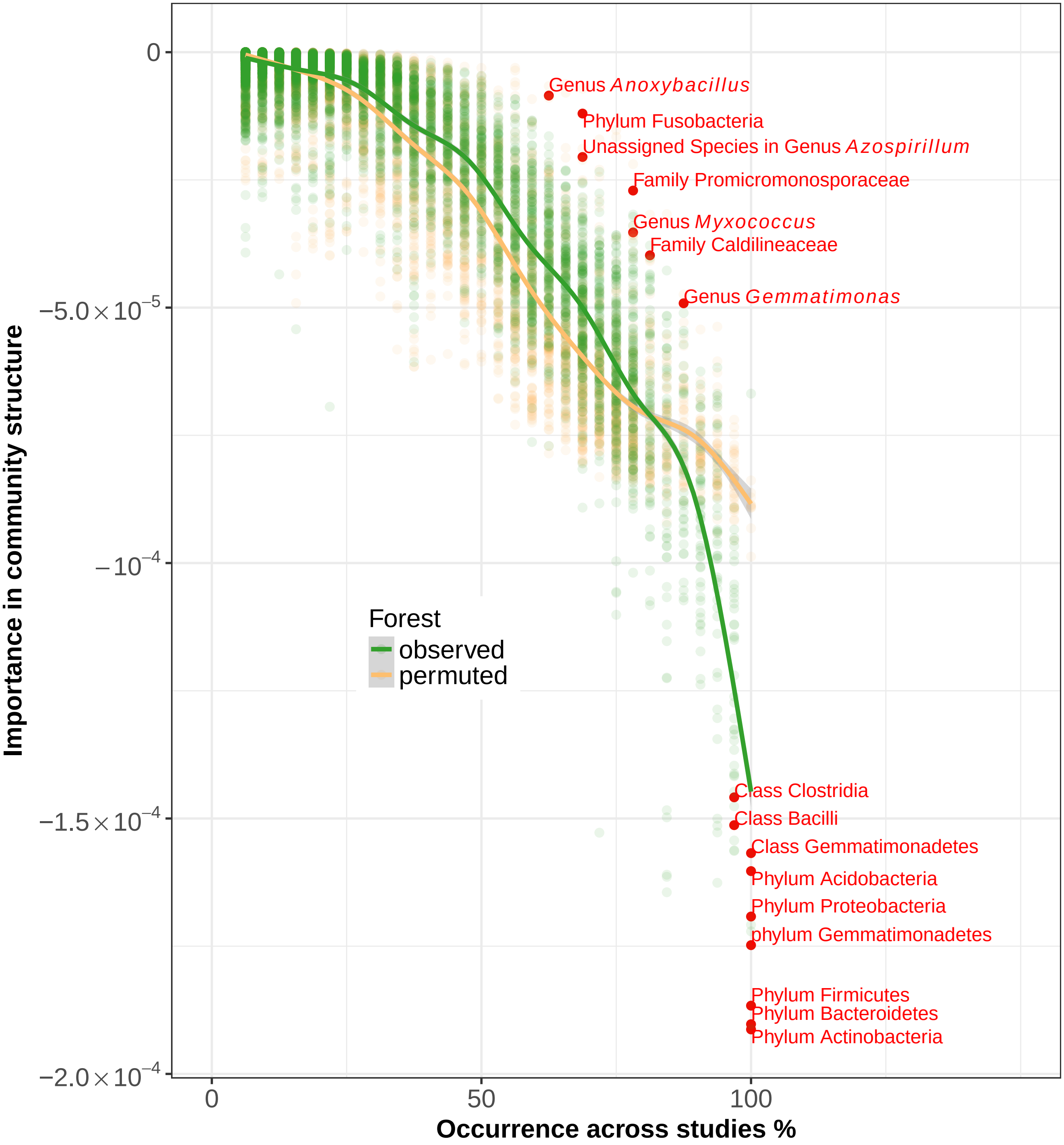
minimum depth
sequencing platform
soil core length
maximum depth
forward &
reverse primer
region sequenced
longitude
latitude
pH
country
owner
0
-10^{-5}
-10^{-4}
-10^{-3}
0
10^{-3}
10^{-2}
10^{-1}
Importance in community structure
Importance for separating studies



Rank importance
1
250
500
750
1000
observed
permuted
Community structure
Separating studies
10^{-6}
10^{-3}
10^{0}
Average taxon abundance



Improvement on random
1.5×10^{-5}
1.0×10^{-5}
5.0×10^{-6}
0
-5.0×10^{-6}
-10^{-5}
Community structure
6×10^{-3}
4×10^{-3}
2×10^{-3}
0
Separating studies
Domain
Phylum
Class
Order
Family
Genus
Species
Taxonomic level


Genus *Anoxybacillus*
Phylum Fusobacteria
Unassigned Species in Genus *Azospirillum*
Family Promicromonosporaceae
Genus *Myxococcus*
Family Caldilineaceae
Genus *Gemmatimonas*
Class Clostridia
Class Bacilli
Class Gemmatimonadetes
Phylum Acidobacteria
Phylum Proteobacteria
phylum Gemmatimonadetes
Phylum Firmicutes
Phylum Bacteroidetes
Phylum Actinobacteria
Forest
observed
permuted
Importance in community structure
0
-5.0×10^{-5}
-10^{-4}
-1.5×10^{-4}
-2.0×10^{-4}
Occurrence across studies %
0
50
100