nature microbiology

ARTICLES

https://doi.org/10.1038/s41564-017-0062-x

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

Kelly S. Ramirez[1]*, Christopher G. Knight[2], Mattias de Hollander[1], Francis Q. Brearley[3], Bede Constantinides[4], Anne Cotton[5], Si Creer[6], Thomas W. Crowther[1,7], John Davison[8], Manuel Delgado-Baquerizo[9], Ellen Dorrepaal[10], David R. Elliott[3,11], Graeme Fox[3], Robert I. Griffiths[12], Chris Hale[13], Kyle Hartman[14], Ashley Houlden[15], David L. Jones[6], Eveline J. Krab[10], Fernando T. Maestre[16], Krista L. McGuire[17], Sylvain Monteux[10], Caroline H. Orr[18], Wim H. van der Putten[1,19], Ian S. Roberts[15], David A. Robinson[20], Jennifer D. Rocca[21], Jennifer Rowntree[3], Klaus Schlaeppi[14], Matthew Shepherd[22], Brajesh K. Singh[23], Angela L. Straathof[2], Jennifer M. Bhatnagar[24], Cécile Thion[25], Marcel G. A. van der Heijden[14,26,27] and Franciska T. de Vries[2]

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, despite 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 data sets comprising 1,998 soil samples from 21 countries. Whereas 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 bacterial community dynamics, identify potential 'indicator' taxa with an important role in structuring communities, and propose hypotheses on the factors that shape bacterial biogeography that have been overlooked in the past.

Soil microbial communities are more diverse and contain more individuals than any species groups on the planet[1,2]. Over the past 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 communities[3–5]. Recent work has further demonstrated that the inclusion of microbial composition and functional attributes improves Earth system models[6], which is of paramount importance for predicting the effects of global change on ecosystem services, such as climate regulation or soil fertility[7]. However, contrary to the long-standing view that every organism may occur everywhere[8], even at small scales, bacterial communities are more patchy than previously expected[9,10], raising questions regarding dispersal constraints, temporal dynamics, and niche breadth at the global

[1]Netherlands Institute of Ecology, Wageningen, The Netherlands. [2]Faculty of Science and Engineering, University of Manchester, Manchester, UK. [3]School of Science and the Environment, Manchester Metropolitan University, Manchester, UK. [4]Evolution and Genomic Sciences, School of Biological Sciences, University of Manchester, Manchester, UK. [5]Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK. [6]Environment Centre Wales, College of Natural Sciences, Bangor University, Bangor, UK. [7]Institute of Integrative Biology, ETH Zürich, Zürich, Switzerland. [8]Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia. [9]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA. [10]Climate Impacts Research Centre, Department of Ecology and Environmental Science, Umeå University, Abisko, Sweden. [11]Environmental Sustainability Research Centre, University of Derby, Derby, UK. [12]Centre for Ecology and Hydrology, Wallingford, UK. [13]School of Life Sciences, University of Warwick, Coventry, UK. [14]Division of Agroecology and Environment, Agroscope, Zürich, Switzerland. [15]Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK. [16]Departamento de Biología y Geología, Física y Química Inorgánica, Escuela Superior de Ciencias Experimentales y Tecnología, Universidad Rey Juan Carlos, Móstoles, Spain. [17]Department of Biology, Institute of Ecology and Evolution, University of Oregon, Eugene, OR, USA. [18]School of Science and Engineering, Teesside University, Middlesbrough, UK. [19]Laboratory of Nematology, Wageningen University, Wageningen, The Netherlands. [20]Centre for Ecology and Hydrology, Bangor, UK. [21]Department of Biology, Duke University, Durham, NC, USA. [22]Natural England, Exeter, UK. [23]Hawkesbury Institute for the Environment, Western Sydney University, Penrith, New South Wales, Australia. [24]Department of Biology, Boston University, Boston, MA, USA. [25]Institute of Biological and Environmental Sciences, University of Aberdeen, Aberdeen, UK. [26]Institute for Evolutionary Biology and Environmental Studies, University of Zürich, Zürich, Switzerland. [27]Plant–Microbe Interactions, Institute of Environmental Biology, Faculty of Science, Utrecht University, Utrecht, The Netherlands. Kelly S. Ramirez and Christopher G. Knight contributed equally to this work. *e-mail: k.ramirez@nioo.knaw.nl

scale[11–13]. Owing 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[14].

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 (such as biodiversity–functioning relationships), and identifying threats to biodiversity from global changes[15–17]. 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 similar to the merging of museum species records, in which information and data are constrained by variations in nomenclature over space and time, among many other challenges[18,19]. Similar to plant and animal records, molecular microbial community records and information can be incomplete, processing and naming varies greatly between studies and over time[20], data storage is inconsistent, and there are few curated databases with high-quality data (especially for short read sequences)[21,22]. Furthermore, 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[23–25], merging microbial sequence data is a potential option to address global-scale questions, whether relating to the human microbiome[26], marine systems[27], or predicting the response of soil organisms to global environmental change[28]. 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[29,30]. Until now, attempts to meta-analyse 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[10,24,31,32]. 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[33], and abundance ratios among phyla may be less important than previous believed[34]. Taken together, these findings indicate 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 analysed together to address questions about the global structuring of communities. Using a 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 twofold: (1) to identify the biases and incompatibilities of microbial community HTS studies (and confounding factors) and, thereby strengthen our ability to integrate data from disparate studies; and (2) to reveal worldwide soil microbial community patterns by merging independent taxonomy-based data sets.

Results

Taxonomy-based merging of disparate amplicon sequence data. We identified 30 individual HTS bacterial studies from 21 countries for our analysis (Fig. 1a–c; Supplementary Table 1). Although we aimed to merge HTS data of both soil bacterial and soil fungal data sets, our approach was only successful for bacterial data (Fig. 1d), highlighting the well-known dilemma of fungal databases, in which extremely high diversity combined with high endemism and mismatched taxonomy across continents make merging data by taxonomy difficult and unusable for downstream analyses[4,35]. For the bacterial studies, we were able to successfully merge 30 individual studies; using a taxonomy-based approach, data sets were merged using the taxonomic affiliations of individual operational taxonomic units (OTU)s. 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; 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, the relative abundances of OTUs were used per sample, rather than absolute read abundance. To test known biogeographical patterns, metadata (that is, 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 used 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 available usable raw sequence data, resulting in 419 samples from locations worldwide (approximately one-fifth of all our samples) (Fig. 1a–c). 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 Fig. 1).

Machine-learning assessment of bacterial community structure. Ordination of the taxonomy-based community reveals large amounts of structure both within and between studies (that is, structure that is removed by permuting taxa among samples (Supplementary Fig. 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[36] (Fig. 1e) confirms our merging method. This visualization also suggests that some of the community variation (for example, the near absence of Acidobacteria in some studies, even at low pH) is due to technical factors, such as the particular primer sets chosen, the region sequenced, and the sequencing platform (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. Thus, using a machine-learning approach that is capable of accounting for complex interactions among taxa (random forests, 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, as well as the 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 was taken based on the proportions of the 8,287 taxa it contained (1.5% (±0.02%, 95% confidence interval) 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 that was randomly drawn from the observed distributions of relative abundances for each taxon (see Methods, all resulting importances given in Supplementary Table 4).

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[37–39]. Thus, we took the step to quantify the importance of both technical and environmental factors alongside taxa in the random forests models (Fig. 2). Of note, 'owner', which encompasses

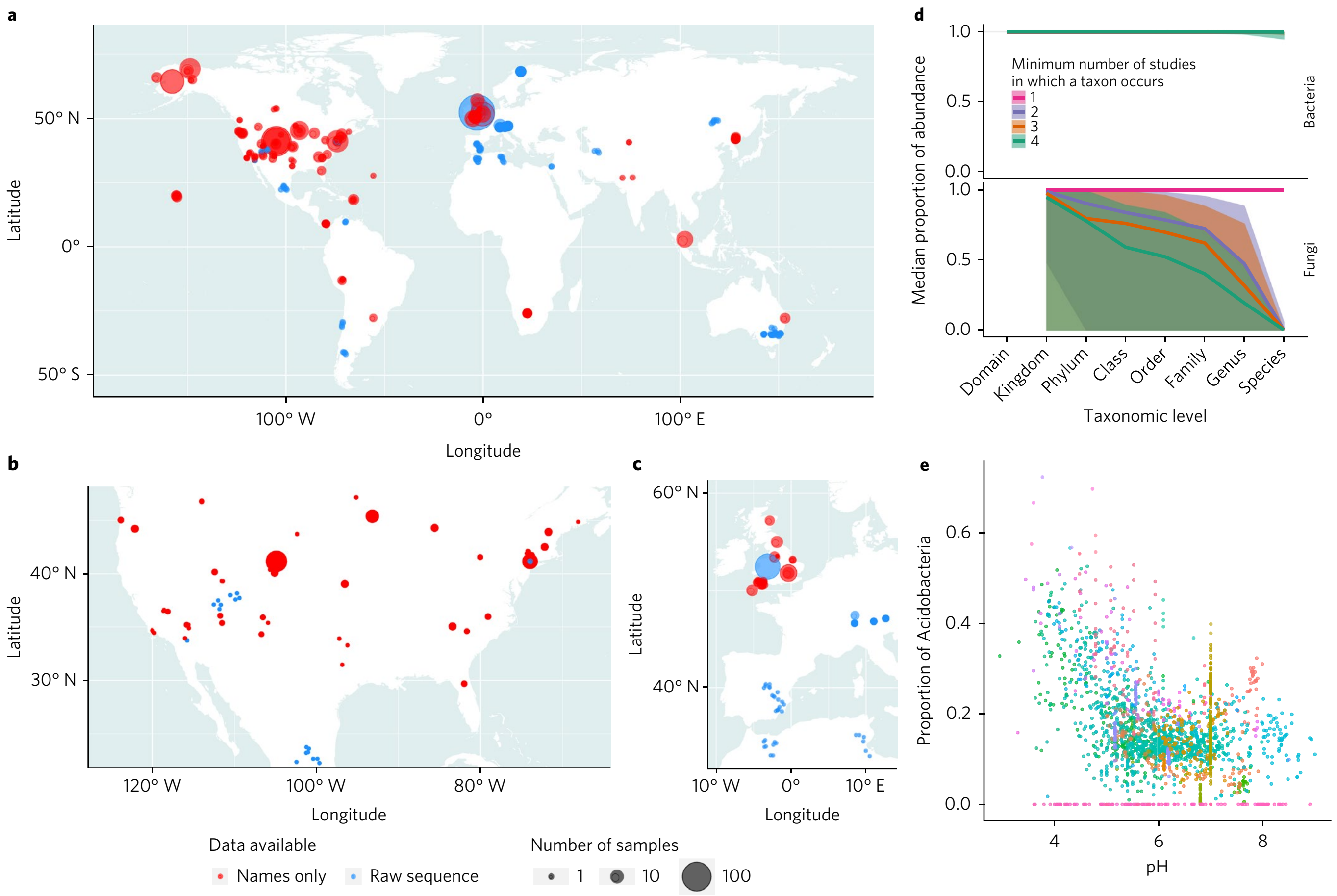

Fig. 1 | Merging of data from 30 independent studies. The wide geographical breadth and community variation of the data are displayed, and the well-known importance of soil pH is confirmed. **a–c**, Map of locations from which samples were collected (**a**), with zoomed in panels of the United States (**b**) and western Europe (**c**). **d**, Average proportion of total bacterial abundance (top graph) and eukaryotic (fungi) abundance (bottom graph), represented by taxa shared among different numbers of studies at different taxonomic levels. The number 1 (pink) indicates the complete data, and 2–4 (purple, orange, green) are subsets of the data containing only taxa that are present in a minimum of 2–4 separate data sets. **e**, Correlation plot of the relative abundance of Acidobacteria to soil pH, in which each colour represents a different study ($r=-0.42$, $P=8.6\times10^{-87}$).

the technical biases and uniqueness of a given data set, is very effective for differentiating between studies (that is, the owner is far to the right in Fig. 2) but is entirely uninformative about community structure (that is, the owner is at the far bottom in Fig. 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[40], 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 that are 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 Fig. 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 with the randomly permuted data (Fig. 3). Thus, the importance of taxa in determining community structure is negatively correlated with the average abundance of those taxa, whereas taxon abundance is positively correlated with the importance for separating studies ($\rho=-0.79$ and $\rho=0.51$, respectively, rank correlation, and compare null expectations of $\rho=-0.62$ and $\rho=-0.12$, respectively, in permuted data). The taxa that are most closely associated with differences between studies tend to be those present at or greater than 0.1% relative abundance, but those taxa that are most important in determining community structure tend to be present at or less than 0.0001% abundance (with a null expectation of around 0.01–0.001% in each case; Fig. 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). These taxa, important in determining community structure, correspond to the long tail in frequency–abundance distributions of soil microbial communities[41], where many taxa in the soil are known to occur at low abundance. Thus, if rarer taxa tend to be more

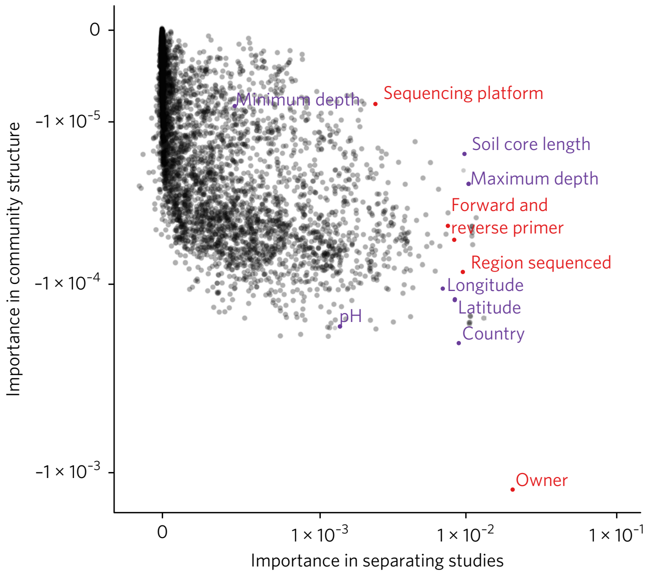

Fig. 2 | Regardless of technical differences between studies, many bacterial taxa are still informative about bacterial community structure. Machine-learning models classify the study from which the samples were derived (*x* axis) based on the relative abundance of taxa within samples, and distinguish 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 and latitude (distinguishing North from South and degrees from the Equator). All values have variable importance from Random Forest models (see Methods); points that are further to the right on the *x* axis have more importance in separating studies, whereas points that are higher up on the *y* axis have more importance in community structure. Note the non-linear axes.

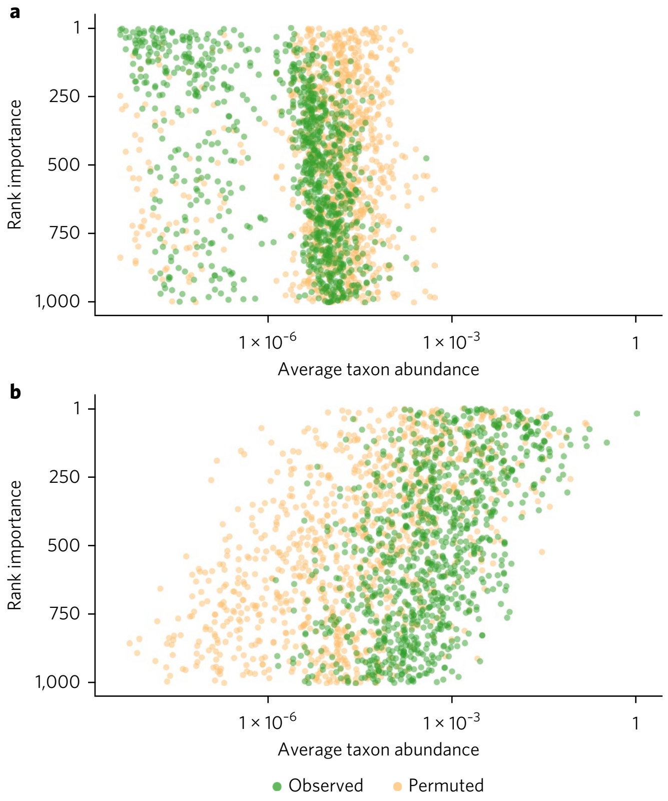

Fig. 3 | Rarer taxa are more important for structuring communities than abundant taxa. a,**b**, The 1,000 most important bacterial taxa in community structure (**a**) and in separating studies (**b**) with respect to their average relative abundance across samples are shown. Plotted are the observed and permuted points, which are a null distribution created from performing the same analysis on a permuted data set (see Methods). The *y* axis reports the rank variable importance in the random forests model of community structure (see Methods); that is, the taxon with the greatest importance in this model is ranked 1, the second greatest is ranked 2, and so on.

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[42,43].

To be used as ecological indicators[29,44], 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[36]. However, Acidobacteria are typically abundant, and our analysis suggests that the most abundant taxa are not the most important in determining community structure. Although dominant taxa such as Acidobacteria do change with environmental factors such as pH (Fig. 1e), those changes are of lesser importance for the 'non-randomness' of community structure, and are more confounded with technical effects, than changes in less-dominant, pH-responsive taxa (Supplementary Fig. 3a). Thus, 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 (Fig. 4b), only taxa that rank lower than phyla are consistently better than random at identifying community structure (Fig. 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 Fig. 7a,b). However, this was not apparent in individual studies (Supplementary Fig. 4c,d), in which phyla were relatively important. A subset of the taxonomy-matched studies showed a pattern that was intermediate between the single studies and the full data set (phyla with some importance, but less than class, order or family; Supplementary Fig. 7c). This, in addition to abundance analyses (Fig. 3; Supplementary Fig. 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 (Fig. 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, which are not apparent from hundreds of samples from one or a handful of studies.

To be considered a good ecological indicator, a taxon should occur in most studies; thus, we looked explicitly at the relationship between the importance of a taxon 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 ($\rho = 0.59$ and $\rho = 0.31$, respectively; Supplementary Fig. 3b,c). We discovered a relationship between taxon occurrence across studies and the importance for structuring communities for all taxa (Fig. 5). 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 data set (Supplementary Fig. 8a) and the same subset of studies when taxonomy matched (Supplementary Fig. 8b). Taken together, the analysis clearly illustrates the importance of taxonomic rank; for example, the class Gemmatimonadetes is relatively unimportant for community structure, but the genus *Gemmatimonas* is relatively important. The result also shows rarer taxa being more

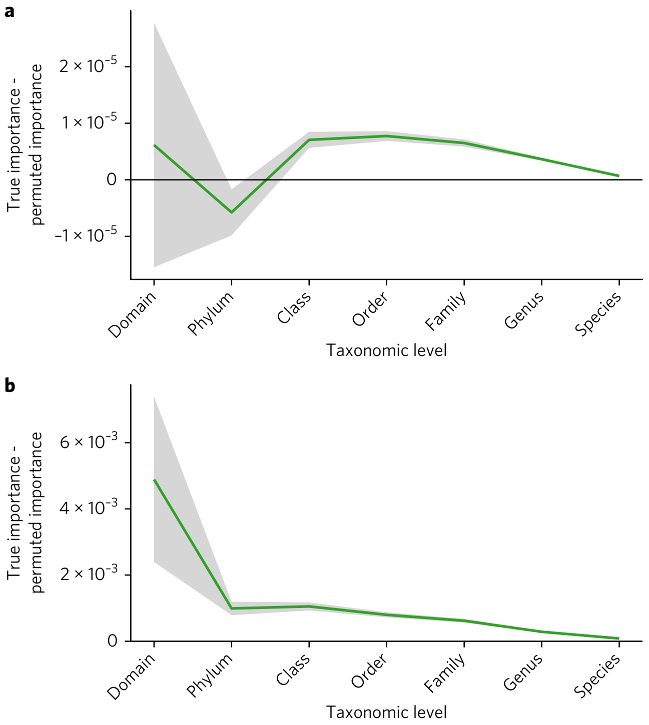

Fig. 4 | The importance of bacterial taxa classified at different taxonomic ranks. a,**b**, Lower taxonomic rank is more important for community structure (**a**), whereas high taxonomic rank is more important for separating studies (**b**). For each taxon, the difference was calculated between the variable 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 data set (see Methods). The green lines and grey ribbons show the mean and standard error, respectively, of these values across taxa at each taxonomic rank considered.

important in structuring communities and suggests that rarer bacterial taxa have overlooked ecologically important roles for bacterial community dynamics[43]. This result is robust to artefacts caused by the rarest taxa (for example, differences between 0 and 1 reads in a sample could be statistically significant for a model, without being biologically relevant); 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 Fig. 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[36,45], and have even been included in source tracking of microbial communities owing to their ubiquitous presence in soil[46]. However, it is the consistent presence of the core taxa across samples and studies that makes them inadequate for assessing community structure.

Discussion

Our results demonstrate the power of combining global bacterial HTS data from multiple independent sources for the detection of biogeographical patterns and for the identification of community patterns that can be used to generate hypotheses on the roles of certain taxa. Although our assessment was on soil communities, our methods can be applied broadly to other microbial data sets and disciplines.

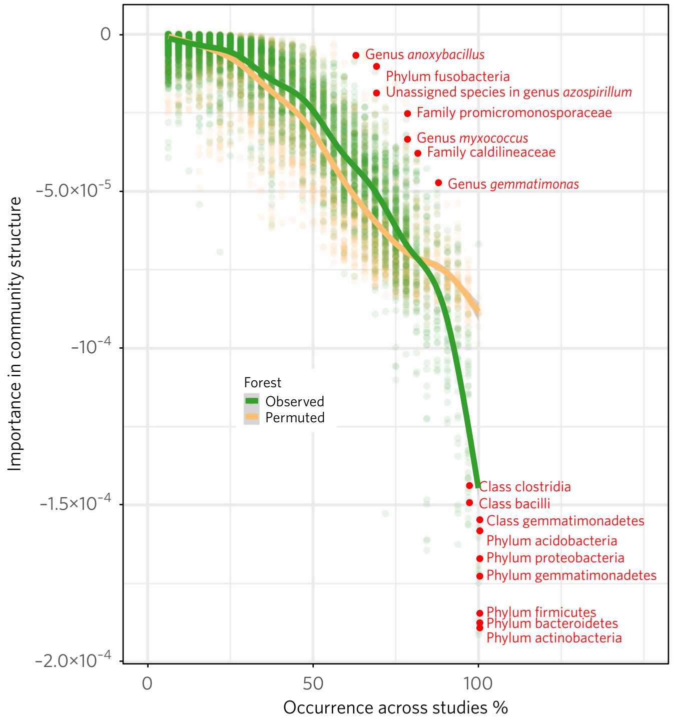

Fig. 5 | Importance of bacterial taxa in community structure related to their occurrence in different studies. The *y* axis reports the variable importance in the random forests model of community structure (see Methods). Observed points correspond to those taxa shown in Fig. 2. Permuted points correspond to the same analysis on a null distribution (see Methods). Lines are general additive model smoothers. Each line is shown with a confidence interval (grey); where this is not visible, it is narrower than the line it surrounds.

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 data sets beyond the well-known biases that accompany any molecular HTS study. Perhaps the most concerning challenge was that so few raw sequence data sets could be retrieved. This highlights the need for wider community adoption of open and accessible short read sequence databases[47], open reference clustering[48], standardized databases[49], 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. Only then, in combination with the many recently designed and classical approaches, can we uncover the full breadth of soil diversity and the roles that soil microorganisms have for ecosystem processes.

Methods

Description of data sets. Metadata from the 30 studies and 1,998 samples were collected and compiled into a summary data file. To do so, we standardized the metadata of each study using the dplyr package[50] of the R statistical platform[51]. Samples were collected from 21 countries 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 (median = 10 m, quartile range = 10–860 m), soil moisture (median = 19.5%, quartile range = 14.1–27.4%), and total soil nitrogen (median = 0.36 mg kg^{-1}, quartile range = 0.23–0.51 mg kg^{-1}), carbon (median = 4.7%, quartile range = 1.9–7.5%) and phosphorus (median = 20.7 mg kg^{-1}, quartile range = 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 to 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 (that is, 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 metadata (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 Quantitative Insights Into Microbial Ecology (QIIME) workflow at some point. An OTU table for one study comprising 43 samples was programmatically retrieved from the MG-RAST public metagenome repository[52]. Taxonomy for the different studies was mainly assigned using the Greengenes database (84%), but the Ribosomal Database Project (RDP) (6%)[37] and the SILVA database (9%)[53] were also used.

Primer biases. It has long been well understood that different primers vary in their biases for amplifying members of the bacterial community[54,55]. To demonstrate this bias, the likelihood of significant differences in primer biases for the ten pairs of primers used in the studies analysed were determined by in silico analysis. Sequences of primer pairs were compared to all 16S rRNA gene sequences in the SILVA non-redundant reference database (SSURef NR) release 128 (ref. [53]) using TestPrime version 1.0 (as described in ref. [56]). The percentages of sequences of each bacterial phyla that matched both primers (with a 1 base pair mismatch allowance at least 1 base pair from the 3′ end of the primers) were calculated to compare predicted differences in primer coverage of different bacterial taxa.

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 data sets. This comprised several steps. (1) Most data sets contained a seven-level taxonomy, recorded in various ways, which was converted to a standardized format. (2) Individual taxon names were cleaned, to give a single name at each taxonomic level (for example, removing special characters and extra annotations, such as 'candidate division' or details of containing taxa). (3) For the many cases in which 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 Fig. 10). Merging at the taxonomy-based level has the added benefit of lessening the affects 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, and put into a format suitable for modelling (Supplementary Table 5). For some analyses (Figs. 3–5), a data set 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 data set were identical to that performed on the observed data.

Merging raw sequence data and other validation data sets. While no data set 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 419 samples of the total 1,998 samples used. Per sample, FASTQ files were obtained for each individual data set. Read files were quality filtered with sickle[57] for single-end reads, trimming bases below a Phred score of 36 and shorter than 100 base pairs. These stringent filtering criteria were applied to keep only high-quality reads and to make sure it was possible to map reads to full-length 16S rRNA gene sequences. Full-length 16S rRNA gene sequences from the SILVA 119 release[53] were obtained in QIIME compatible format from the SILVA Download Archive. For each data set, all reads were mapped to the full-length 16S rRNA gene sequences using the usearch global algorithm implemented in VSEARCH version 1.9.6 (ref. [58]). The alignment results in usearch table format (uc) were directly converted to Biological Observation Matrix (BIOM) format using BIOM version 2.1.5 (ref. [59]). Consensus/majority taxonomy was added as metadata to the BIOM file. Finally, all BIOM files of each data set were merged using QIIME version 1.9.1 (ref. [60]). All steps were implemented in a workflow made with Snakemake version 3.5.4 (ref. [61]) (Supplementary Fig. 1) resulting in the sequence-matched dataset (Supplementary Table 6).

To use this sequence-matched data set to validate our taxonomy-matching approach across studies using different taxonomy databases (Supplementary Figs. 5,7,8), we created an equivalent taxonomy-matched data set from the same five studies. As with the full data set, only taxa occurring in at least two studies were included in either this or the sequence-matched data set. To test what is gained or lost by considering different numbers of studies simultaneously, we considered not only the full data set (30 studies) and the subset of five studies used in the sequence-matched data set but also two of the largest individual studies: Central Park, NY, USA, encompassing 594 samples (study no. 24), and a global data set encompassing 103 samples (study no. 30). In each case, a simple subset of the full data set was analysed (Supplementary Fig. 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 that were deemed present are represented by multiple sequence reads (Supplementary Fig. 9). To address the issue of differential 16S copy numbers skewing abundance estimates, we created a binary data set of the presence/absence of all taxa. The results for a model separating studies using this data set were very similar to the main data set using relative abundance; however, there was insufficient power to identify taxa that are important for community structure. 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 Fig. 6). Finally, to test the effect of 'unknown' or 'unclassified' bacterial taxa, we created a reduced data set, in which all taxa classified as 'unassigned' at any level were removed (Supplementary Fig. 10).

Random Forest models. To test for the importance of different taxa in the structuring of the data, we used Random Forest models[62–64] 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 unsupervised modes. When run in unsupervised mode, the importance of a variable describes how effective it is at separating the observed data from randomized synthetic data[64]. In both cases, a proximity matrix may be generated, which can be used for ordination (Supplementary Fig. 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 for separating studies (*x* axis in Fig. 2) is closely correlated with the sensitivity component of the indicator value of each taxon ($\rho = 0.89$)[44] (Supplementary Fig. 3d). There are two key parameters that may be adjusted in a Random Forest model: 'mtry', which is the number of variables randomly sampled as candidates for a split in the constituent trees; and 'ntree', which is the number of trees in the forest. mtry was set at its default value (square root of the number of variables), and ntree was set to 100,000 for each forest. 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 (that is, a minimum node size of 1). The unscaled 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 (as previously mentioned) to create null distributions for taxa 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 Fig. 2), variable importance correlates with ecologically important factors, such as abundance. This makes intuitive sense considering that, even with randomized samples, it 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 Fig. 5. All analyses were completed with RandomForest package for R version 4.6.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Data availability. The authors declare that the data supporting the findings of this study are available within the paper and its Supplementary Information files.

Received: 23 May 2017; Accepted: 13 October 2017;
Published online: 20 November 2017

References

1. Proser, J. I. Dispersing misconceptions and identifying opportunities for the use of 'omics' in soil microbial ecology. *Nat. Rev. Microbiol.* **13**, 439–446 (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**, 1621–1624 (2012).

4. Tedersoo, L. et al. Fungal biogeography. Global diversity and geography of soil fungi. *Science* **346**, 1256688 (2014).
5. Davison, J. et al. Fungal symbionts. Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism. *Science* **349**, 970–973 (2015).
6. Wieder, W. R., Bonan, G. B. & Allison, S. D. Global soil carbon projections are improved by modelling microbial processes. *Nat. Clim. Change* **3**, 909–912 (2013).
7. Karhu, K. et al. Temperature sensitivity of soil respiration rates enhanced by microbial community response. *Nature* **513**, 81–84 (2014).
8. Barberán, A., Casamayor, E. O. & Fierer, N. The microbial contribution to macroecology. *Front. Microbiol.* **5**, 203 (2014).
9. Ramirez, K. S. et al. Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. *P. R. Soc. B* **281**, 20141988 (2014).
10. O'Brien, S. L. et al. Spatial scale drives patterns in soil bacterial diversity. *Environ. Microbiol.* **18**, 2039–2051 (2016).
11. 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).
12. Talbot, J. M. et al. Endemism and functional convergence across the North American soil mycobiome. *Proc. Natl Acad. Sci. USA* **111**, 6341–6346 (2014).
13. 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).
14. Ranjard, L. et al. Turnover of soil bacterial diversity driven by wide-scale environmental heterogeneity. *Nat. Commun.* **4**, 1434 (2013).
15. Jetz, W., McPherson, J. M. & Guralnick, R. P. Integrating biodiversity distribution knowledge: toward a global map of life. *Trends Ecol. Evol.* **27**, 151–159 (2012).
16. Ricketts, T. H. et al. Disaggregating the evidence linking biodiversity and ecosystem services. *Nat. Commun.* **7**, 13106 (2016).
17. Dirzo, R. et al. Defaunation in the Anthropocene. *Science* **345**, 401–406 (2014).
18. 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).
19. Santos, A. M. & Branco, M. The quality of name-based species records in databases. *Trends Ecol. Evol.* **27**, 6–7 (2012).
20. Beiko, R. G. Microbial malaise: how can we classify the microbiome? *Trends Microbiol.* **23**, 671–679 (2015).
21. Tedersoo, L. et al. Standardizing metadata and taxonomic identification in metabarcoding studies. *Gigascience* **4**, 34 (2015).
22. Ramirez, K. S. et al. Toward a global platform for linking soil biodiversity data. *Front. Ecol. Evol.* **3**, 91 (2015).
23. Turner, W. et al. Free and open-access satellite data are key to biodiversity conservation. *Biol. Conserv.* **182**, 173–176 (2015).
24. Gilbert, J. A., Jansson, J. K. & Knight, R. The Earth Microbiome project: successes and aspirations. *BMC Biol.* **12**, 69 (2014).
25. Joppa, L. N. et al. Big data and biodiversity. Filling in biodiversity threat gaps. *Science* **352**, 416–418 (2016).
26. 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).
27. Sogin, M. L. et al. Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc. Natl Acad. Sci. USA* **103**, 12115–12120 (2006).
28. García-Palacios, P. et al. Are there links between responses of soil microbes and ecosystem functioning to elevated CO_2, N deposition and warming? A global perspective. *Glob. Chang. Biol.* **21**, 1590–1600 (2015).
29. Hermans, S. M. et al. Bacteria as emerging indicators of soil condition. *Appl. Environ. Microbiol.* **83**, e02826-16 (2016).
30. Philippot, L. et al. The ecological coherence of high bacterial taxonomic ranks. *Nat. Rev. Microbiol.* **8**, 523–529 (2010).
31. 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**, 1493–1506 (2013).
32. 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).
33. Bier, R. L. et al. Linking microbial community structure and microbial processes: an empirical and conceptual overview. *FEMS Microbiol. Ecol.* **91**, fiv113 (2015).
34. Walters, W. A., Xu, Z. & Knight, R. Meta-analyses of human gut microbes associated with obesity and IBD. *FEBS Lett.* **588**, 4223–4233 (2014).
35. Bik, H. M. et al. Sequencing our way towards understanding global eukaryotic biodiversity. *Trends Ecol. Evol.* **27**, 233–243 (2012).
36. 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).
37. McDonald, D. et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6**, 610–618 (2012).
38. Lozupone, C. A. et al. Meta-analyses of studies of the human microbiota. *Genome Res.* **23**, 1704–1714 (2013).
39. 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).
40. Lu, X., Seuradge, B. J. & Neufeld, J. D. Biogeography of soil Thaumarchaeota in relation to soil depth and land usage. *FEMS Microbiol. Ecol.* **93**, fiw246 (2017).
41. 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).
42. Langille, M. G. et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* **31**, 814–821 (2013).
43. Jousset, A. et al. Where less may be more: how the rare biosphere pulls ecosystems strings. *ISME J.* **11**, 853–862 (2017).
44. De Cáceres, M. & Legendre, P. Associations between species and groups of sites: indices and statistical inference. *Ecology* **90**, 3566–3574 (2009).
45. Maestre, F. T. et al. Increasing aridity reduces soil microbial diversity and abundance in global drylands. *Proc. Natl Acad. Sci. USA* **112**, 15684–15689 (2015).
46. Knights, D. et al. Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* **8**, 761–763 (2011).
47. Muir, P. et al. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.* **17**, 53 (2016).
48. Rideout, J. R. et al. Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ* **2**, e545 (2014).
49. Yilmaz, P. et al. The genomic standards consortium: bringing standards to life for microbial ecology. *ISME J.* **5**, 1565–1567 (2011).
50. Wickham, H. & Francois, R. dplyr: a grammar of data manipulation. R package v. 0.5.0 (CRAN, 2016); https://cran.r-project.org/package=dplyr.
51. The R Core Team. *R: A Language and Environment for Statistical* (R Foundation for Statistical Computing, 2016); https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf.
52. Wilke, A. et al. The MG-RAST metagenomics database and portal in 2015. *Nucleic Acids Res.* **44**, D590–D594 (2016).
53. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
54. Suzuki, M. T. & Giovannoni, S. J. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol.* **62**, 625–630 (1996).
55. Sipos, R. et al. Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targeting bacterial community analysis. *FEMS Microbiol. Ecol.* **60**, 341–350 (2007).
56. Klindworth, A. et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41**, e1 (2013).
57. Joshi, N. A. & Fass, J. N. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files. v. 1.33 (2011); https://github.com/najoshi/sickle.
58. Rognes, T. et al. vsearch: VSEARCH 1.9.6. (2016); https://doi.org/10.5281/ZENODO.44512.
59. 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).
60. Caporaso, J. G. et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).
61. Koster, J. & Rahmann, S. Snakemake — a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
62. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
63. Breiman, L. & Cutler, A. *Using Random Forests v4.0* (UC Berkeley, 2003); https://www.scribd.com/document/208387804/Using-Random-Forests-v4-0.
64. Shi, T. & Horvath, S. Unsupervised learning with Random Forest predictors. *J. Comput. Graph. Stat.* **15**, 118–138 (2006).

Acknowledgements

We thank all the people who contributed data and input to this study. This study was conducted at a workshop (May 2015, Manchester, UK) funded by the British Ecological Society's special interest group Plants-Soils-Ecosystems and organized by F.T.d.V. and K.S.R. This study and participants were funded in part by ERC Advanced Grant 26055290 (K.S.R., and W.H.v.d.P.); BBSRC David Phillips Fellowship (BB/L02456X/1) (F.T.d.V.); ERC Grant Agreements 242658 (BIOCOM) and 647038 (BIODESERT) (F.T.M.); the European Regional Development Fund (Centre of Excellence EcolChange) (J.D.); Yorkshire Agricultural Society, Nafferton Ecological Farming Group, and the Northumbria University Research Development Fund (C.H.O.); BBSRC Training Grant (BB/K501943/1) (C.H.); Wallenberg Academy Fellowship (KAW 2012.0152), Formas (214-2011-788) and Vetenskapsrådet (612-2011-5444) (E.D.); the Glastir Monitoring & Evaluation Programme (contract reference: C147/2010/11) and the full support of the

GMEP team on the Glastir project (D.L.J., S.C., and D.A.R.). Computing was facilitated by the University of Manchester Condor pool and the CLIMB infrastructure (http://www.climb.ac.uk).

Author contributions

The idea for this study was conceived by F.T.d.V. and K.S.R. The data sets were compiled by C.G.K., R.G., J.D., A.H., B.C., G.F., A.L.S., and J.R. Metadata were compiled by J.D. and J.R. Raw sequence analysis was conducted by M.d.H. Primer bias analysis was conducted by A.C. Random Forest analyses and figures were conducted by C.G.K. The manuscript was written by K.S.R., C.G.K., and F.T.d.V., with contributions from all co-authors.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41564-017-0062-x.

Reprints and permissions information is available at www.nature.com/reprints.

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

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Kelly S Ramirez

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

This was a type of meta-analysis of 30 studies, including 1998 soil samples.

2. Data exclusions

Describe any data exclusions.

To be included samples had to have met pre-estabilished criteria including HTS with illumina or pyrosequencing, a georeference, and knowledge of the owner and primer pair used.

3. Replication

Describe whether the experimental findings were reliably reproduced.

NA

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Random permutations were used in the model.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

NA

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a | Confirmed

☒ ☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)

☒ ☐ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☒ ☐ A statement indicating how many times each experiment was replicated

☒ ☐ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)

☒ ☐ A description of any assumptions or corrections, such as an adjustment for multiple comparisons

☐ ☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted

☒ ☐ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)

☐ ☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

▸ Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

All analyses were done with the R statistical platform. Specific packages of note include dplyr and RandomForests

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

▸ Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No unique materials were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No cell lines were used.

b. Describe the method of cell line authentication used.

No cell lines were used.

c. Report whether the cell lines were tested for mycoplasma contamination.

No cell lines were used.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No cell lines were used.

▸ Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

No humans were involved.