# Beyond Taxonomic Identification

Jones, Briony; Goodall, Tim; George, Paul B L; Gweon, Hyun S; Puissant, Jeremy; Read, Daniel S; Emmett, Bridget; Robinson, David A.; Jones, Davey L; Griffiths, Robert I

**Frontiers in Microbiology**

*DOI:*
10.3389/fmicb.2021.682886

Published: 19/07/2021

frontiers in Microbiology

ORIGINAL RESEARCH
published: 19 July 2021
doi: 10.3389/fmicb.2021.682886

# Beyond Taxonomic Identification: Integration of Ecological Responses to a Soil Bacterial 16S rRNA Gene Database

*Briony Jones*[1,2]*, *Tim Goodall*[3], *Paul B. L. George*[1,2], *Hyun S. Gweon*[4], *Jeremy Puissant*[3], *Daniel S. Read*[3], *Bridget A. Emmett*[1], *David A. Robinson*[1], *Davey L. Jones*[3] and *Robert I. Griffiths*[1]*

[1] UK Centre for Ecology and Hydrology, Bangor, United Kingdom, [2] School of Environment, Natural Resources and Geography, Bangor University, Bangor, United Kingdom, [3] UK Centre for Ecology and Hydrology, Wallingford, United Kingdom, [4] School of Biological Sciences, University of Reading, Reading, United Kingdom

OPEN ACCESS

**Edited by:**
Mark Radosevich,
The University of Tennessee,
Knoxville, United States

**Reviewed by:**
Aditi Sengupta,
California Lutheran University,
United States
Liang (Leon) Wang,
Xuzhou Medical University, China

***Correspondence:**
Briony Jones
brionyjones28@gmail.com
Robert I. Griffiths
rig@ceh.ac.uk

**Specialty section:**
This article was submitted to
Terrestrial Microbiology,
a section of the journal
Frontiers in Microbiology

**Received:** 19 March 2021
**Accepted:** 21 June 2021
**Published:** 19 July 2021

**Citation:**
Jones B, Goodall T, George PBL,
Gweon HS, Puissant J, Read DS,
Emmett BA, Robinson DA, Jones DL
and Griffiths RI (2021) Beyond
Taxonomic Identification: Integration
of Ecological Responses to a Soil
Bacterial 16S rRNA Gene Database.
Front. Microbiol. 12:682886.
doi: 10.3389/fmicb.2021.682886

High-throughput sequencing 16S rRNA gene surveys have enabled new insights into the diversity of soil bacteria, and furthered understanding of the ecological drivers of abundances across landscapes. However, current analytical approaches are of limited use in formalizing syntheses of the ecological attributes of taxa discovered, because derived taxonomic units are typically unique to individual studies and sequence identification databases only characterize taxonomy. To address this, we used sequences obtained from a large nationwide soil survey (GB Countryside Survey, henceforth CS) to create a comprehensive soil specific 16S reference database, with coupled ecological information derived from survey metadata. Specifically, we modeled taxon responses to soil pH at the OTU level using hierarchical logistic regression (HOF) models, to provide information on both the shape of landscape scale pH-abundance responses, and pH optima (pH at which OTU abundance is maximal). We identify that most of the soil OTUs examined exhibited a non-flat relationship with soil pH. Further, the pH optima could not be generalized by broad taxonomy, highlighting the need for tools and databases synthesizing ecological traits at finer taxonomic resolution. We further demonstrate the utility of the database by testing against geographically dispersed query 16S datasets; evaluating efficacy by quantifying matches, and accuracy in predicting pH responses of query sequences from a separate large soil survey. We found that the CS database provided good coverage of dominant taxa; and that the taxa indicating soil pH in a query dataset corresponded with the pH classifications of top matches in the CS database. Furthermore we were able to predict query dataset community structure, using predicted abundances of dominant taxa based on query soil pH data and the HOF models of matched CS database taxa. The database with associated HOF model outputs is released as an online portal for querying single sequences of interest (https://shiny-apps.ceh.ac.uk/ID-TaxER/), and flat files are made available for use in bioinformatic pipelines. The further development of advanced informatics infrastructures incorporating modeled ecological attributes along with new functional genomic information will likely facilitate large scale exploration and prediction of soil microbial functional biodiversity under current and future environmental change scenarios.

**Keywords: ecological responses, 16S database, countryside survey, amplicon 16S rRNA, traits, HOF modeling, soil bacteria communities**

# INTRODUCTION

Soil bacteria are highly diverse (Gans et al., 2005; Roesch et al., 2010) and are significant contributors to soil functionality. Sequencing of 16S rRNA genes has enabled a wealth of new insights into the taxonomic diversity of soil prokaryotic communities, revealing the ecological controls on a vast diversity of yet to be cultured taxa with unknown functional potential (Fierer, 2017). However, despite numerous studies across the globe, we are still some way from synthesizing the new knowledge on the ecology of these novel organisms recovered through local and distributed soil surveillance. This is because there is currently no formalized way of retrieving ecological information on reference sequences which match user-discovered taxa (either clustered operational taxonomic units or amplicon sequence variants). Whilst we have a wealth of databases and tools for characterizing the taxonomy of matched sequences (Wang et al., 2007; McDonald et al., 2012; Quast et al., 2013), databases do not include any associated ecological information on sequences matches. Whilst new software has recently become available that uses text mining to return some ecological data on matched sequences to NCBI, this information is currently limited to descriptions of sequence associated habitat (Sinclair et al., 2016).

Synthesizing relationships between soil amplicon abundances and environmental parameters is now necessary to progress ecological understanding of soil microbes beyond those few organisms that are readily cultivated. Determining microbial responses across environmental gradients can inform on the realized niche widths of discrete taxa, and may indicate the presence of shared functional traits across taxa (Martiny et al., 2015). This information is now urgently needed for microbes as we move into a period of increasing genomic data availability for uncultivated taxa. Coupling data on taxon responses across environmental gradients with functional trait information potentially allows a mechanistic and predictive understanding of both biodiversity and ecosystem level responses to environmental change. For example, a large body of theory exists describing how species responses to environmental change affects ecosystem functioning (Lavorel and Garnier, 2002; Suding et al., 2008; Diaz et al., 2013). Here functional "response" groups are defined as species sharing a similar response to an environmental driver; and functional "effect" groups refer to species that have similar effects on one or more ecosystem processes. The degree of coupling between response and effect groups can then allow prediction of functional effects under change. For instance if certain phylogenetic groups of taxa decrease due to environmental change, and these taxa also represent an effect group (e.g., these taxa possess a unique functional gene) then we can expect the function to also decrease. Conversely with uncoupled effect groups (e.g., responsive taxa all possess a ubiquitous functional gene), the system is likely to be more functionally resistant to change (Diaz et al., 2013). Applying such concepts to microbial ecology is a realistic ambition given the extensive availability of amplicon datasets coupled to environmental information, and the increasing feasibility of uncultivated microbial genome assembly from metagenomes or single cell genomics (Choi et al., 2017).

The fast evolution of microbial taxa coupled with potential horizontal gene transfer has led to assumptions that microbial diversity may be largely functionally redundant (Martiny et al., 2015). However, we know from large-scale amplicon surveys that there are distinct differences in soil bacterial composition across environmental gradients, with soil pH frequently observed as a primary correlate (Fierer and Jackson, 2006; Griffiths et al., 2011). This implies that different microbial phylogenetic lineages possess adaptations conferring altered competitiveness in soils of different pH; paving the way for future studies into the genomic basis, and thereby elucidating specific genetic "response traits." There is also evidence that many specific bacterial functional capacities such as methanogenesis (an "effect" trait) are phylogenetically conserved and therefore may be less redundant (Martiny et al., 2013). Determining the degree of functional redundancy in taxa which respond across soil pH gradients, will permit new insight into the microbial biodiversity mechanisms underpinning soil functionality and resilience to change. Since soil pH is largely predictable from geo-climatic (Slessarev et al., 2016) and land use features (Wamelink et al., 2019); prediction of the abundances of individual bacterial taxa under environmental change scenarios is likely to be feasible. The immediate challenge is therefore to establish predictive frameworks for many soil bacterial taxa, which can be populated with genomic information as it becomes available; to ultimately facilitate predictions of microbial functional distributions.

We believe that attempts to progress understanding of the ecological attributes of environmentally retrieved bacterial taxa can be streamlined immediately by making better use of the extensive amplicon datasets that exist, which already provide an abundance of useful information on taxa-environment responses. Indeed it has recently been shown that many prokaryotic taxa are distributed globally (particularly dominant OTUs Delgado-Baquerizo et al., 2018), yet there is currently no way to formally capture their ecological attributes in databases for further microbiological and ecological enquiry other than in

Supplementary Material spreadsheets. Here we seek to address this by making available a database of representative sequences from a large 16S rRNA amplicon dataset from over 1,000 soil samples collected across Britain. In addition to providing standard taxonomic annotation, we also seek to add ecological response information to each representative sequence. We focus here on soil pH responses as bacterial communities are known to respond strongly across soil pH gradients (Griffiths et al., 2011). We will firstly model OTU abundances across soil pH using hierarchical logistic regression (HOF) (Jansen and Oksanen, 2013), a commonly used approach to examine vegetation responses across ecological gradients which has yet to be widely applied to microbial datasets. We will use model outputs to assign each OTU to a specific pH response group based on abundance optima, and in addition demonstrate the utility of the database in determining the phylogenetic relationships in ecological responses. The utility of the database will be further tested on 16S datasets to compare both the percentage of hits and modeled responses. The OTU database with associated HOF model outputs is released both as an online portal for visualizing individual queries and as flat files for integration into existing bioinformatics pipelines.

## RESULTS AND DISCUSSION

### Database Coverage

The database was constructed from sequences obtained from the 2007 Countryside Survey (CS), a randomly stratified sampling of most soil types and habitats across Great Britain, full details of which are provided elsewhere (Griffiths et al., 2011; Reynolds et al., 2013). Sequencing of 1,113 soils using the universal 341f/806r (Takahashi et al., 2014) primers targeting the V3 and V4 regions of the 16S rRNA gene yielded a total of 39,952 reference sequence OTUs, after clustering at 97% sequence similarity and singleton removal. Coverage was assessed on a filtered dataset of 1,006 samples which had at least 5,000 reads per sample, using sample based species accumulation curves calculated per habitat class and pooled across all habitats (**Figure 1**). The curves for individual habitats, whilst not reaching saturation, reveal some interesting trends with grasslands exhibiting highest biodiversity at the landscape scale, which is likely attributable to the broad range of soil conditions they encompass. The pooled curves across all habitats, however, appear to begin to level off, which importantly reveals that in total the reference sequence dataset provides good coverage of the non-singleton 97% OTUs found across this landscape.

### Performance of Database Against Independent Datasets

The coverage of this dataset was further assessed through blasting representative sequences from independent 16S datasets for various locations and habitats (detailed in **Table 1**), against all 39,952 CS representative sequences. Here we defined an OTU "hit" as a query OTU that shared 97% identity with a CS OTU and had an *e*-value equal to or less than 0.001. We subsequently calculated the percentage of OTUs within each independent dataset meeting this criteria to gain insights into CS dataset coverage.

For the two independent soil datasets (query datasets 1 and 2, **Table 1**), we found over 50% of the OTUs in each study had a hit within the CS database. Expectedly, this was in stark contrast to the fresh water query dataset (query dataset 3, **Table 1**) which exhibited much less overlap with the CS database with 33.2% having CS hits. 16S sequences from **query dataset 1**, a study of land use change across the United Kingdom (Malik et al., 2018),

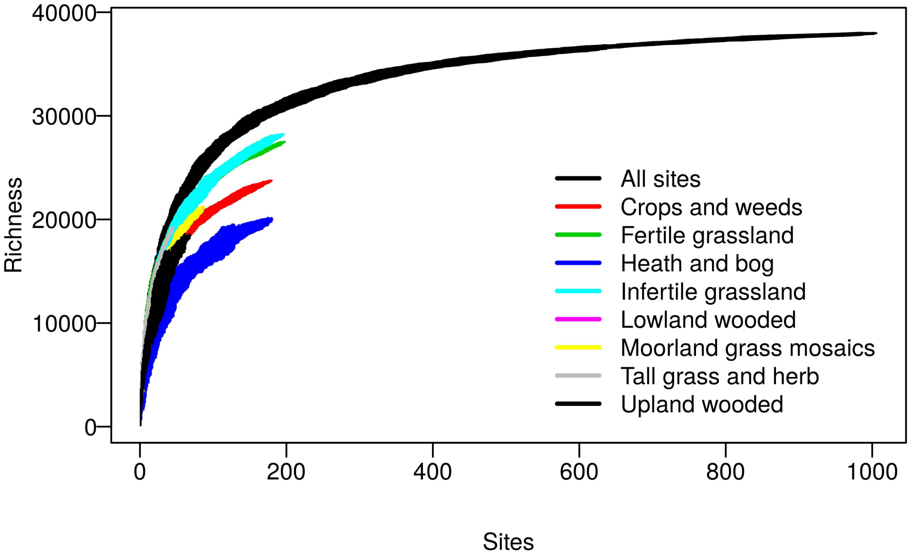

**FIGURE 1 |** Coverage of bacterial 97% OTUs within the Countryside Survey (CS) dataset. Sample based richness accumulation curves were calculated across 1,006 CS soil samples ("All sites"), and within specific habitats. Standard deviations are calculated from random permutations of the data.

**TABLE 1 |** Validating the use of the CS OTU sequences as a database, through querying with independent datasets.

| Query dataset | Habitat description | Query OTU percentage of hits | Primer | ENA project ID |
|---|---|---|---|---|
| 1 | Grassland and arable soils, Britain | 67.26% | 341f/806r V3-V4 | PRJEB36119 |
| 2 | All habitat soils survey, Wales | 58.49% | 515f/806rB V4 | PRJEB27883 |
| 3 | Thames River, Britain | 33.2% | 341f/806r V3-V4 | Unpublished, see Read et al., 2015 |

*Reference sequences from independent datasets were BLAST searched against countryside survey representative sequences, and the proportion of OTUs matched at over 97% similarity reported. British soil query datasets had highest percentage of hits irrespective of methodologies, with a set of riverine samples showing lowest proportion of OTUs matching the CS soil reference database.*

also sequenced with the same 341f/806r primer set, had the highest percentage of hits against the CS representative sequences (67.26%). Wider assessment of our own unpublished datasets using the exact same methodologies yield percentages of hits of 62 and 56% for soils from United Kingdom calcareous grasslands and tropical rainforests, respectively. A separate survey of Welsh soils (George et al., 2019) was also queried against the CS database (query dataset 2, **Table 1**), which used the commonly used Earth Microbiome primer set exclusively targeting the V4 region (as opposed to V3 and V4 targeted region used for the CS dataset). This dataset had a percentage of hits of 58.49% providing evidence that datasets amplified with other primer sets can be matched to the CS database with only marginal loss of coverage.

We next wanted to explore possible reasons for obtaining less than 100% coverage from query soil datasets, given the good coverage of the CS reference sequence database evident from the rarefaction curve (**Figure 1**). We predicted this discrepancy was caused by rare OTUs being unique to specific studies and tested this by classifying OTU's from query dataset 1 (**Table 1**) into 1,000 discrete abundance based quantiles (1 being the most abundant quantile and 1,000 being the least). Plotting the proportion of query OTUs which matched to the CS database by query OTU abundance class, confirmed that less abundant query OTUs had less matches to the CS database (**Figure 2**). This adds weight to arguments that much of the rare taxa detected through amplicon sequencing could be spurious artifacts of the PCR amplification process (Edgar, 2017). Regardless of these issues, the high proportion of hits for dominant taxa in the query dataset validates the use of the large CS dataset as a comprehensive reference database.

## Modeling OTU Responses to Soil pH

Since the majority of the 39,952 reference OTUs obtained across all CS samples likely derive from rare taxa with intrinsically little value for predictive modeling (low within-sample abundance, and occurrence across samples), we opted to only model taxa-pH relationships for those taxa which occurred in at least 30 samples within the CS dataset. These taxa were selected from a cleaned dataset of 1,006 samples which had at least 5,000 reads per sample. These samples covered the following aggregate vegetation classes: crops and weeds, fertile grassland, heath and bog, infertile grassland, lowland wooded, moorland grass mosaics, tall grass and herb, and upland wooded.

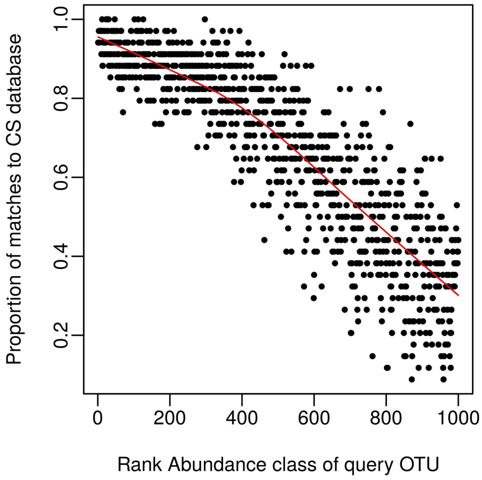

**FIGURE 2 |** The CS database provides good coverage of dominant taxa within a query dataset. Query OTU reference sequences (dataset 1, **Table 1**) were grouped into 1,000 bins by decreasing rank (e.g., the 1000th bin contains the least abundant OTUs); and the proportion of each bin matching the CS dataset calculated and displayed on the *y*-axis. The proportion of matches to the CS database (> 97% similarity) declines as query taxa become rarer, despite the comprehensive nature of the CS database.

Further examination of the species accumulation by sample curves for the resulting 13,781 OTUs, revealed saturation implying that this dataset had complete coverage of common OTUs, defined by being present in at least 30 samples across Britain. Hierarchical logistic regression (HOF) models were then applied to determine individual bacterial taxa responses to pH. HOF models enable individual taxon responses to be modeled along environmental gradients using five response shapes,that increase in number of parameters and shape complexity. A HOF modeling approach was taken as HOFs discrete response shapes allow for good interpretability in ecological contexts. GAM's were considered as an alternative method, but were not used due to the challenges associated with choosing suitable smoothing parameters to avoid overfitting (Jansen et al., 2013). HOF models were generated using the R package eHOF using a poisson error distribution. Model choice was determined using AIC and bootstrapping methods implemented with the eHOF package (Jansen and Oksanen, 2013), whereby the model with the lowest AIC was initially chosen and its robustness then tested by rerunning models on 100 bootstrapped datasets (created by resampling with replacement). If the most frequently chosen model in the bootstrap runs was different to the initial model choice, the most common bootstrap choice was selected. The resultant pH-taxa response curves classified by the HOF models include I: no significant change in abundance in response to pH, II: an increasing or decreasing trend, III: increasing or decreasing trend which plateaus, IV: Increase and decrease by same rate (unimodal) and V: Increase and decrease by different rates causing skew (**Figure 3**).

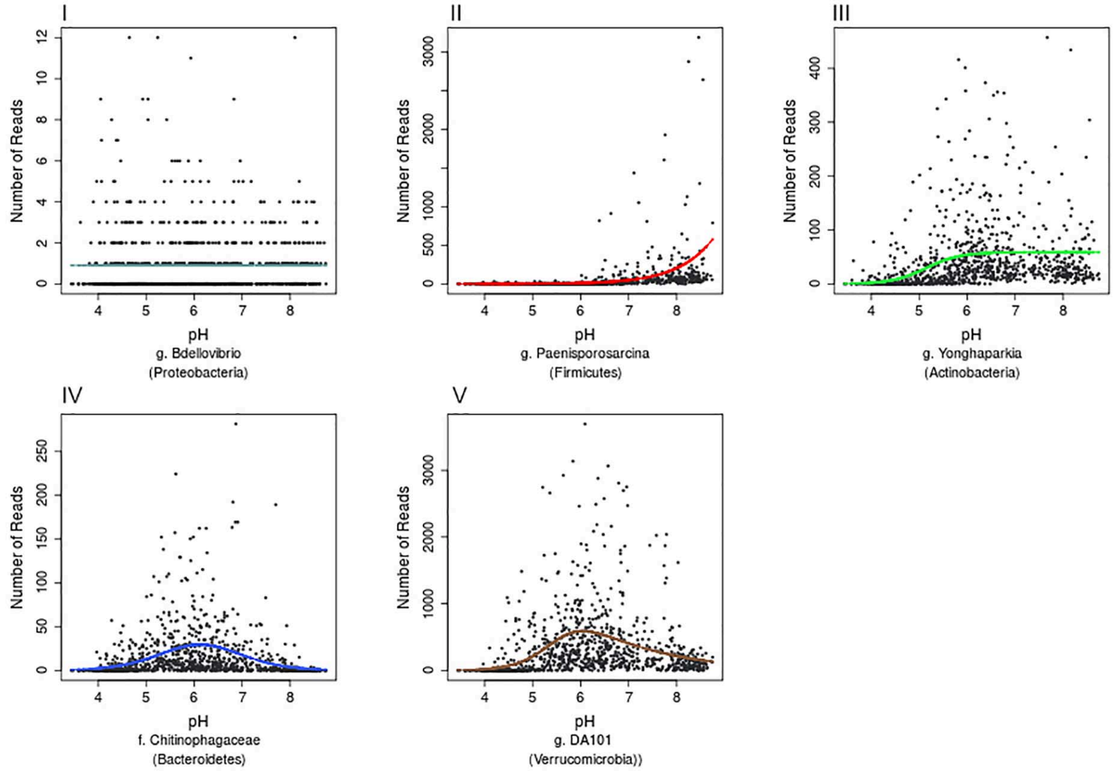

**FIGURE 3 |** Examples of the five HOF model types. HOF models were generated through fitting countryside survey OTU abundances to soil pH (a pH range from 3.63 to 8.75). The five HOF models used were: **(I)** no change in abundance across pH gradient, **(II)** montonic an increase or decrease in abundance along pH gradient, **(III)** plateau an increase or decrease in abundance along pH gradient that plateaus, **(IV)** symmetrical unimodal, abundance increases and decreases across gradient at an equal rate, **(V)** skewed unimodal, abundance increases and decreases across gradient at unequal rates.

The proportion of OTUs assigned to each model is shown in **Table 2**, and reveals that most of the soil OTUs exhibited some trend with soil pH, and with the unimodal skewed model (V) being the most commonly fitted model type (45.76%). OTUs were then assigned to pH response groups based on the fitted pH optima. We classified OTUs demonstrating an acidic preference if the fitted optima was below pH 5.2, based on previous data showing this represented a critical threshold for bacterial communities (Griffiths et al., 2011), which was further confirmed by a similar regression tree analyses of this sequence dataset (not shown). This pH value also represents a critical threshold in microbial functioning (Jones et al., 2019). Similarly, a second threshold was designated at pH 7, with OTUs exhibiting an optima above this being classed as neutral, and those between 5.2 and 7 classed as "mid." Plateau model shapes (model III), were sometimes more difficult to classify, since two optima are provided which span the plateau, and in some cases these crossed the pH 5.2 and 7 thresholds. Whilst OTUs exhibiting this response were in the minority, we opted to assign a separate designation representing this range, for instance "acid to mid" for an OTU with two optima above and below pH 5.2. The proportion of taxa classified to each pH response group are shown in **Table 3**. This reveals that OTUs with acidic preference are in the minority, consistent with reduced bacterial biodiversity being frequently observed in acidic soils (Griffiths et al., 2011).

**TABLE 2 |** Percentage of 13,781 CS OTUs fitted to each HOF model.

| Model fit | Percentage of countryside survey OTUs |
|---|---|
| V (Skewed Unimodal) | 45.76% |
| III (Plateau) | 24.13% |
| IV (Unimodal) | 23.52% |
| II (Monotonic) | 6.11% |
| I (No trend) | 0.49% |

*Each OTU was classified to one of five HOF model types according to fitted relationships with soil pH. The different model response shapes are shown in* ***Figure 3****.*

Representative sequences of all 13,781 OTUs were aligned with Clustal Omega 1.2.1[1], and used to construct a Phylogenetic tree with FastTree 2.1.7 with the generalized time-reversible (GTR) model of nucleotide evolution. FastTree uses a heuristic form of neighbor joining to initially determine tree topology before reducing tree length using nearest neighbor interchanges (NNIs) and subtree prune regraft moves (SPRs). Tree topology and branch lengths are then improved using maximum likelihood rearrangements. FastTree was used as it's highly efficient when handling large alignments (Price et al., 2010). The tree was

[1] http://www.clustal.org/

**TABLE 3 |** Percentage of 13,781 CS OTUs classified to different pH response groups.

| pH Response group | Percentage of countryside survey OTUs |
|---|---|
| Mid (5.2 < Optima < 7) | 34.8% |
| Neutral (Optima > 7) | 31.62% |
| Acid (Optima < 5.2) | 23.08% |
| Mid to Neutral (5.2 < Optimum1 < 7 and Optimum 2 > 7) | 7.41% |
| Acid to Neutral (Optimum1 < 5.2 and Optimum2 > 7) | 1.52% |
| Acid to Mid (Optimum1 < 5.2 and 5.2 < Optimum2 < 7) | 1.14% |

*Each OTU was assigned to a pH response classification based on the modeled pH optima. The model outputs with one optima (II, IV,V) were classified as acidic, mid or neutral based on pH thresholds identified above. Plateau shaped models with 2 optima (model III), which spanned the pH thresholds were labeled as either mid to neutral, acid to neutral, or acid to mid.*

visualized using graphlan (Asnicar et al., 2015) together with the pH classification derived from the HOF models and is shown in **Figure 4**. Distinct phylogenetic clustering is apparent for phyla with representatives known to have acidophilic preferences such as the Acidobacteria (Kielak et al., 2016). Additionally, other phyla such as the Verrucomicrobia appear to possess clades with a distinct pH preference. However, the overall impression across other taxonomic groups is that the pH abundance optima can vary substantially amongst closely related taxa. This emphasizes the need to move beyond the association of traits with broad phylogenetic lineages; and identifies the need to determine traits at finer levels of taxonomic resolution.

## Incorporating CS Data and pH Responses Into a Sequence Identification Tool

A web application was developed using the Shiny package[2] which enables users to BLAST a 16S query sequence against the countryside survey representative sequences, subsequently allowing visualization of key environmental information including HOF model outputs, relevant to individual matched sequences. The Graphic User Interface was implemented

[2] https://shiny.rstudio.com/

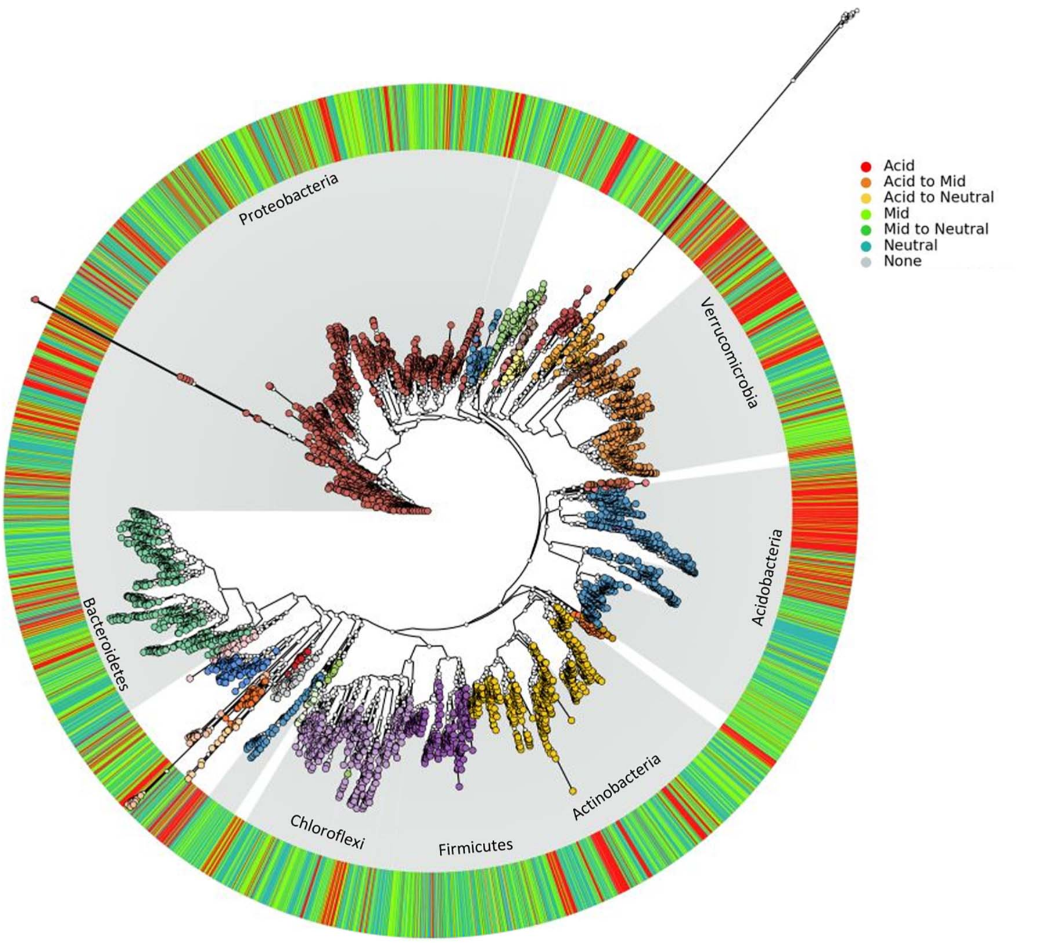

**FIGURE 4 |** The phylogenetic distribution of bacterial pH optima. A phylogenetic tree of all OTUs present in > 100 samples (totaling 6,385 OTUs), with each OTU annotated according to pH classification based on HOF model optima (outer ring).

in R (3.4.1) using the Shiny package (see text footnote 3) alongside ShinyJS to execute JavaScript functions from R[3]. BLASTn commands are executed from R using the users query sequence, *e*-value of 0.01, and the reference sequence database of CS representative sequences. eHOF model objects were converted to binary using the Rbase serialize function and stored in a PostgreSQL (9.3.17) database[4] alongside model and other environmental metadata (**Supplementary Figure 1**). BLAST results are displayed as an interactive table of hits, each hit linking to a plot of the pH model fit (based upon raw read number), a LOESS fit (based on relative abundance), a box plot of habitat associations and a simple interpolated map showing relative abundance distribution across Britain (**Supplementary Figure 2**). Additionally we provide a text box which can be populated with user submitted trait related information on matched OTUs. The application is available at shiny-apps.ceh.ac.uk/ID-TaxER/ and to facilitate batch processing of query sequences the sequence database, taxonomy and trait matrix are released via github (github.com/brijon/ID-TaxER-flat-files) for integration into bioinformatics pipelines.

## Utility in Predicting pH Preferences and Community Structure Using a Query Dataset

To demonstrate both the utility of the reference sequence database, and the HOF modeling approach to identify environmental responses of soil bacterial taxa, we used a query dataset of > 400 samples collected across Britain (dataset 1, **Table 1**). Since this survey focused on productive habitats (grassland and arable land uses), with only a few acidic samples, it was not appropriate to generate independent HOF models. Instead we classified the samples according to the same pH cutoff levels identified above (pH 5.2 and 7) and then determined pH responsive taxa using indicator species analyses (Dufrene and Legendre, 1997). As can be seen in **Figure 5A**, the pH groupings were clearly evident in the sample based ordination. Representative sequences from this dataset were then blasted against the CS database, and optimum pH and pH classification metrics retrieved from the top hit for subsequent comparison. In total 477 indicators for the three pH groupings were retrieved, of which 454 had a match greater than 97% similarity to the CS database. Of the 155 acidic indicator taxa identified in the query dataset, 129 (83%) were reliably classified as acidic OTUs based on matches to the CS database (**Figure 5B**), with 20 OTUs "incorrectly" classified as having a mid-pH optima. However, the predicted optima of these OTUs was mainly below pH 6 and most lie very close to pH 5.2. Similarly for the 226 query taxa identified as indicating neutral soils, 203 (90%) had a neutral pH classification in the CS database, with 15 being incorrectly classed as mid, though the optima for these was between pH 6.5 and 7. Sixty-seven indicators of the query mid pH soils were obtained of which 64 (96%) had a mid pH classification based on match

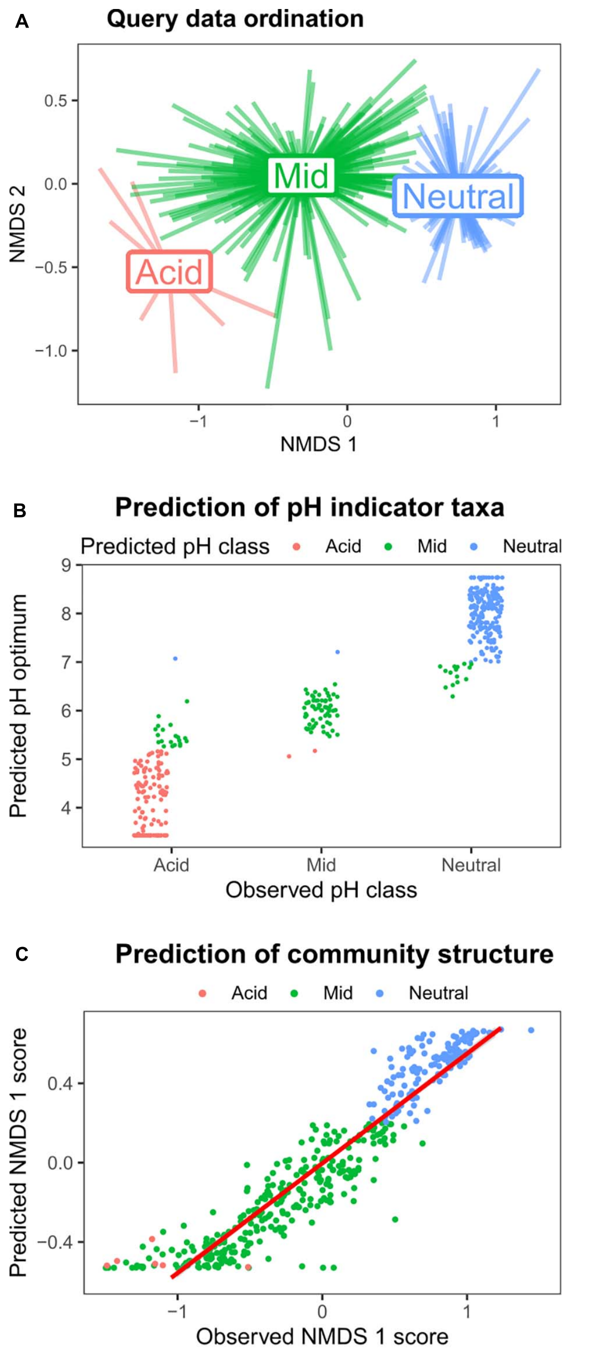

**FIGURE 5 |** Validating the pH models using a query dataset. Taxa strongly responsive to soil pH were identified from Query dataset 1 (**Table 1**), and then matched to the CS database to evaluate utility of the approach. **(A)** NMDS ordination plot of the query dataset, with pH groupings denoted by color (red = pH < 5.2; green = pH > 5.2 < 7; and blue = pH > 7). **(B)** Indicator species analyses on the query dataset revealed 477 OTUs strongly associated with the three pH classes ("Observed pH class"). The *y*-axis values and point color denote the predicted pH optimum, and predicted pH class following matching to CS database. **(C)** The relative abundances of the 100 most abundant taxa in the query dataset were predicted using the CS HOF models of matched taxa, and subjected to NMDS ordination. The plot shows that the predicted abundances of these taxa reliably predicted the observed data first axis NMDS scores.

[3] https://cran.r-project.org/web/packages/shinyjs/

[4] https://www.postgresql.org/

to the CS database. Overall this analyses shows that information on soil pH preferences from independent datasets can be reliably obtained using our approach.

We then sought to test whether we could reliably predict community structure using the CS HOF model outputs to predict query OTU abundances. We identified the most abundant OTUs in the query dataset, and blasted against the CS database. CS HOF models were then used to predict the abundances of the 100 matched dominant OTUs within the 424 query samples. This predicted community matrix was then subject to NMDS ordination with the first axis scores plotted against the actual observed ordination scores generated from 24,260 OTUs. The results in **Figure 5C** show that the observed and predicted first axis ordination scores were highly related ($r^2 = 0.88$) demonstrating that it is possible to predict broad scale community change from individual OTU relative abundance pH models. These findings add to a growing body of literature on the predictability of soil bacterial communities (Fierer et al., 2013; Griffiths et al., 2016; Bickel et al., 2019); but furthermore demonstrate the utility of our overall approach in deriving meaningful ecological information from matches to a 16S rRNA sequence database incorporating ecological responses.

## CONCLUSION

This work demonstrates how large scale soil molecular survey data can be used to build robust predictive models of bacterial abundance responses across environmental gradients. The models were applied to the single soil variable of pH which is known globally to be the strongest predictor of soil bacterial community structure in surveys spanning wide environmental gradients. We have produced an informatics tool incorporating extensive sequence data from a wide range of soils, linked to taxonomic and ecological response information. This currently includes data on the modeled pH optima, and the predictive utility in this regard was demonstrated using an independent dataset. Other ecological information is also made available via an online portal including habitat association, spatial distribution, and metrics relating to abundance and occurrence. We are currently working on incorporating other information on the sensitivities of discrete OTUs to land use change; and there is the wider potential for users to update the trait matrix with other observations (more information provided at https://github.com/brijon/ID-TaxER-flat-files). Such information could include sensitivities to perturbations such as climate change, as well as rRNA derived links to wider genome data to inform on function.

We anticipate this simple database and tool will be of use to the soil molecular community, but also hope it prompts further global efforts to better capture relevant ecological information on newly discovered microbial taxa. We acknowledge some limitations of the current tool, and identify some possibilities to develop further: firstly being a 16S rRNA amplicon dataset, the database inventory will be affected by known biases relating to PCR primers and amplification conditions, i.e., taxa that are known to be poorly detected by the primers used will be under represented in the database (Thijs et al., 2017). Secondly as the database features the V3-V4 region of the 16S rRNA amplified with 341f/806r primers and therefore obviously, user datasets built on a different region of the 16S rRNA gene will not produce any matches. Additionally the length of sequences means only limited taxonomic resolution is currently provided, and ecological inferences based on BLAST matches must consider the strength of match, and variance within the matched region with respect to taxonomic discrimination (Fox et al., 1992). Emerging long read sequencing technologies applied to survey nucleic acid archives in the future may improve these current constraints (Singer et al., 2016). With respect to the pH models, many other factors can of course influence bacterial abundances (Thomson et al., 2010; Fierer, 2017), and we note the large degree of variance in relative abundance for a taxon even within its apparent pH niche optima (**Figure 3**). Such variance could may be caused by nutrient availability, stress etc. and more complex models, albeit constrained by pH, need to be formulated to advance predictive accuracy. More generally, we assert that observed taxon relative abundance only inform on relative taxon success at a given soil pH, and does not identify any explicit underpinning ecological mechanism (e.g., pH stress tolerance vs. competitive fitness) (Austin, 1999). However, linking emerging genomic data to detailed environmentally relevant sequence databases such as detailed here, will likely improve future understanding in relation to elucidating specific functional response traits and determining mechanisms underpinning bacterial community assembly along soil gradients. Finally, and importantly, the CS database is spatially constrained to a temperate island in Northern Europe, and would benefit from a more global extent to capture other soil biomes such as drylands. Improvements here could be made from integrating data from global sequencing initiatives, or leveraging data from sequence repositories provided consistent environmental metadata can also be retrieved in order to reliably predict response trait characteristics. Further use of data from global sequence repositories, would likely shed light on more taxa and potentially reveal more on those taxa deemed rare in for example national focused studies.

## MATERIALS AND METHODS

Samples were collected as part of the UK Centre for Ecology and Hydrology Countryside survey (CS) between June and July 2007 covering sites throughout Great Britain. Samples were chosen through a stratified random sample of 1 km squares using a 15 km grid, implementing the institute of Terrestrial Ecology (ITE) land classification to ensure incorporation of different land classes, with up to 5 randomly sampled cores (15 cm long × 4 cm diameter) taken within each square. Metadata for each soil sample were collated including soil organic matter, soil organic carbon, bulk density, pH, indicator of phosphorus availability using methodologies detailed elsewhere (Griffiths et al., 2011; Reynolds et al., 2013).

DNA was extracted from 0.3 g of soil using the MoBIO PowerSoil-htp 96 Well DNA Isolation kit (Carlsbad, CA) according to manufacturer protocols. Amplicon libraries were constructed according to the dual indexing strategy of Kozich et al. (2013), using primers 341F (Muyzer et al., 1993), and 806R (Caporaso et al., 2011). Amplicons were generated using a high fidelity DNA polymerase (Q5 Taq, New England Biolabs) on 20 ng of template DNA employing an initial denaturation of 30 s at 95°C, followed by (25 for 16S and 30 cycles for ITS and 18S) of 30 s at 95°C, 30 s at 52°C and 2 min at 72°C. A final extension of 10 min at 72°C was also included to complete the reaction. Amplicon sizes were determined using an Agilent 2200 TapeStation system (∼550 bp) and libraries normalized using SequalPrep Normalization Plate Kit (Thermo Fisher Scientific). Library concentration was calculated using a SYBR green quantitative PCR (qPCR) assay with primers specific to the Illumina adapters (Kappa, Anachem). Libraries were sequenced at a concentration of 5.4 pM with a 0.6 pM addition of an Illumina generated PhiX control library. Sequencing runs, generating 2 × 300 bp, reads were performed on an Illumina MiSeq using V3 chemistry.

Sequenced paired-end reads were joined using PEAR (Zhang et al., 2013), quality filtered using FASTX tools[5], length filtered with the minimum length of 300 bp. The presence of PhiX and adapters were checked and removed with BBTools[6], and chimeras were identified and removed with VSEARCH_UCHIME_REF (Rognes et al., 2016) using Greengenes Release 13_5 (at 97%). Singletons were removed and the resulting sequences were clustered into operational taxonomic units (OTUs) with VSEARCH_CLUSTER at 97% sequence identity. Representative sequences for each OTU were taxonomically assigned by RDP Classifier with the bootstrap threshold of 0.8 or greater using the Greengenes Release 13_5 (full) as the reference. All statistical analyses and visualizations were conducted within the R package, predominantly using the vegan and ggplot packages unless otherwise indicated.

## DATA AVAILABILITY STATEMENT

ID-TaxER tool is available at https://shiny-apps.ceh.ac.uk/ID-TaxER/. Countryside survey OTU representative sequences (16S), taxonomic and HOF model assignments are accessible at: https://github.com/brijon/ID-TaxER-flat-files. All Countryside survey 16S Sequences are deposited in the European Nucleotide Archive under accession PRJEB45286 and wider metadata for the samples is available at https://catalogue.ceh.ac.uk/documents/79669141-cde5-49f0-b24d-f3c6a1a52db8. The bioinformatics pipeline used to pre-process 16S sequences can be found at https://github.com/brijon/sgtoolkit. Exemplar R code for generating all figures in manuscript is provided at https://github.com/brijon/Beyond_taxonomic_identification. Code for R shiny front end of ID-TaxER is accessible at https://github.com/brijon/ID-TaxER.

## AUTHOR CONTRIBUTIONS

BJ developed ID-TaxER tool and conducted majority of statistical and phylogenetic analyses, overseen by RG and DJ. TG conducted 16S sequencing for Countryside survey dataset. PG provided GMEP 16S dataset and conducted GMEP bioinformatics pre-processing. HG conducted bioinformatics pre-processing of Countryside survey dataset. DSR provided Thames River 16S dataset. BJ wrote first draft of manuscript with RG. All authors contributed to revisions of the manuscript.

## FUNDING

This work has been funded by an ENVISION DTP studentship award to BJ. We also acknowledge funding from the UK Natural Environment Research Council under a Soil Security Programme grant (NE/M017125/1), and the UK–ScaPE Programme (NE/R016429/1).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.682886/full#supplementary-material

**Supplementary Figure 1 |** ID-TaxER database Infrastructure 16S sequences are queried over the web via the R Shiny interface. A BLAST search is then performed against a blast database containing representative 16S sequences from the 2007 Countryside survey. Model information and associated metadata for match hits are located in a PostgreSQL database of OTU taxonomy/model data (model objects are stored as binary and retrieved for the user) and results displayed via the shiny interface.

**Supplementary Figure 2 |** Example outputs from the ID-TaxER online portal. Using the DA101/Ca. U. copiosus (Brewer et al., 2016) 16S sequence (GenBank: Y07576.1) as a query, we found 98.3% identitiy to CS OTU19097 (taxonomy = k_Bacteria; p_Verrucomicrobia; c_Spartobacteria; o_Chthoniobacterales; f_Chthoniobacteraceae; g_DA101): **(A)** HOF model output showing the number of reads of CS OTU19097 per sample plotted against soil pH; with the line representing the model fit (Model V, unimodal response to pH with an optima at pH 6.18) **(B)** the relative abundance of OTU19097 against sample pH, with the line representing a LOESS fit; **(C)** boxplot showing the median and ranges of the relative abundance of OTU19097 per CS habitat class; **(D)** inverse distance weighted interpolation map of the relative abundance of OTU19097 across Britain.

[5] hannonlab.cshl.edu

[6] jgi.doe.gov/data-and-tools/bbtools/

## REFERENCES

Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C., and Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* 3:e1029. doi: 10.7717/peerj.1029

Austin, M. P. (1999). The potential contribution of vegetation ecology to biodiversity research. *Ecography* 22, 465–484. doi: 10.1111/j.1600-0587.1999.tb01276.x

Bickel, S., Chen, X., Papritz, A., and Or, D. (2019). A hierarchy of environmental covariates control the global biogeography of soil bacterial richness. *Sci. Rep.* 9:12129.

Brewer, T. E., Handley, K. M., Carini, P., Gilbert, J. A., and Fierer, N. (2016). Genome reduction in an abundant and ubiquitous soil bacterium 'Candidatus Udaeobacter copiosus'. *Nat. Microbiol.* 2:16198.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., et al. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci.U. S. A.* 108, 4516–4522. doi: 10.1073/pnas.1000080107

Choi, J., Yang, F., Stepanauskas, R., Cardenas, E., Garoutte, A., Williams, R., et al. (2017). Strategies to improve reference databases for soil microbiomes. *ISME J.* 11, 829–834. doi: 10.1038/ismej.2016.168

Delgado-Baquerizo, M., Oliverio, A. M., Brewer, T. E., Benavent-González, A., Eldridge, D. J., Bardgett, R. D., et al. (2018). A global atlas of the dominant bacteria found in soil. *Science* 359, 320–325. doi: 10.1126/science.aap9516

Diaz, S., Purvis, A., Cornelissen, J. H., Mace, G. M., Donoghue, M. J., Ewers, R. M., et al. (2013). Functional traits, the phylogeny of function, and ecosystem service vulnerability. *Ecol. Evo.l* 3, 2958–2975. doi: 10.1002/ece3.601

Dufrene, M., and Legendre, P. (1997). Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecol. Monogr.* 67, 345–366. doi: 10.2307/2963459

Edgar, R. C. (2017). Accuracy of microbial community diversity estimated by closed- and open-reference OTUs. *PeerJ* 5:e3889. doi: 10.7717/peerj.3889

Fierer, N. (2017). Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat. Rev. Microbiol.* 15, 579–590. doi: 10.1038/nrmicro.2017.87

Fierer, N., and Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. *Proc. Natl. Acad. Sci. U. S. A.* 103, 626–631. doi: 10.1073/pnas.0507535103

Fierer, N., Ladau, J., Clemente, J. C., Leff, J. W., Owens, S. M., Pollard, K. S., et al. (2013). Reconstructing the microbial diversity and function of pre-agricultural tallgrass prairie soils in the United States. *Science* 342, 621–624. doi: 10.1126/science.1243768

Fox, G. E., Wisotzkey, J. D., and Jurtshuk, P. (1992). How Close Is Close: 16S rRNA Sequence Identity May Not Be Sufficient To Guarantee Species Identity. *Int. J. Syst. Evol. Microbiol.* 42, 166–170. doi: 10.1099/00207713-42-1-166

Gans, J., Wolinsky, M., and Dunbar, J. (2005). Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* 309, 1387–1390. doi: 10.1126/science.1112665

George, P. B. L., Lallias, D., Creer, S., Seaton, F. M., Kenny, J. G., Eccles, R. M., et al. (2019). Divergent national-scale trends of microbial and animal biodiversity revealed across diverse temperate soil ecosystems. *Nat. Commun.* 10:1107.

Griffiths, R. I., Thomson, B. C., James, P., Bell, T., Bailey, M., and Whiteley, A. S. (2011). The bacterial biogeography of British soils. *Environ. Microbiol.* 13, 1642–1654. doi: 10.1111/j.1462-2920.2011.02480.x

Griffiths, R. I., Thomson, B. C., Plassart, P., Gweon, H. S., Stone, D., Creamer, R. E., et al. (2016). Mapping and validating predictions of soil bacterial biodiversity using European and national scale datasets. *Appl. Soil Ecol.* 97, 61–68. doi: 10.1016/j.apsoil.2015.06.018

Jansen, F., and Oksanen, J. (2013). How to model species responses along ecological gradients – Huisman–Olff–Fresco models revisited. *J. Veg. Sci.* 24, 1108–1117. doi: 10.1111/jvs.12050

Jansen, F., Oksanen, J., and Podani, J. (2013). How to model species responses along ecological gradients - Huisman-Olff-Fresco models revisited. *J. Veg. Sci.* 24, 1108–1117.

Jones, D. L., Cooledge, E. C., Hoyle, F. C., Griffiths, R. I., and Murphy, D. V. (2019). pH and exchangeable aluminum are major regulators of microbial energy flow and carbon use efficiency in soil microbial communities. *Soil Biol. Biochem.* 138:107584. doi: 10.1016/j.soilbio.2019.107584

Kielak, A. M., Barreto, C. C., Kowalchuk, G. A., Van Veen, J. A., and Kuramae, E. E. (2016). The Ecology of Acidobacteria: moving beyond Genes and Genomes. *Front. Microbiol.* 7:744. doi: 10.3389/fmicb.2016.00744

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., and Schloss, P. D. (2013). Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. *Appl. Environ. Microbiol.* 79, 5112–5120. doi: 10.1128/aem.01043-13

Lavorel, S., and Garnier, E. (2002). Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the Holy Grail. *Funct. Ecol.* 16, 545–556. doi: 10.1046/j.1365-2435.2002.00664.x

Malik, A. A., Puissant, J., Buckeridge, K. M., Goodall, T., Jehmlich, N., Chowdhury, S., et al. (2018). Land use driven change in soil pH affects microbial carbon cycling processes. *Nat. Commun.* 9:3591.

Martiny, A. C., Treseder, K., and Pusch, G. (2013). Phylogenetic conservatism of functional traits in microorganisms. *ISME J.* 7, 830–838. doi: 10.1038/ismej.2012.160

Martiny, J. B. H., Jones, S. E., Lennon, J. T., and Martiny, A. C. (2015). Microbiomes in light of traits: a phylogenetic perspective. *Science* 350:aac9323. doi: 10.1126/science.aac9323

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., Desantis, T. Z., Probst, A., et al. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* 6, 610–618. doi: 10.1038/ismej.2011.139

Muyzer, G., De Waal, E. C., and Uitterlinden, A. G. (1993). Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl. Environ. Microbiol.* 59, 695–700. doi: 10.1128/aem.59.3.695-700.1993

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* 5:e9490. doi: 10.1371/journal.pone.0009490

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596.

Read, D. S., Gweon, H. S., Bowes, M. J., Newbold, L. K., Field, D., Bailey, M. J., et al. (2015). Catchment-scale biogeography of riverine bacterioplankton. *ISME J.* 9, 516–526. doi: 10.1038/ismej.2014.166

Reynolds, B., Chamberlain, P. M., Poskitt, J., Woods, C., Scott, W. A., Rowe, E. C., et al. (2013). ). Countryside Survey: national "Soil Change" 1978–2007 for Topsoils in Great Britain—Acidity, Carbon, and Total Nitrogen Status. *Vadose Zone J.* 12, 1–15.

Roesch, L. F. W., Fulthorpe, R. R., Riva, A., Casella, G., Km, A., Kent, A. D., et al. (2010). Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J.* 1, 283–290. doi: 10.1038/ismej.2007.53

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584. doi: 10.7717/peerj.2584

Sinclair, L., Ijaz, U. Z., Jensen, L. J., Coolen, M. J. L., Gubry-Rangin, C., Chronakova, A., et al. (2016). Seqenv: linking sequences to environments through text mining. *PeerJ* 4:e2690. doi: 10.7717/peerj.2690

Singer, E., Bushnell, B., Coleman-Derr, D., Bowman, B., Bowers, R. M., Levy, A., et al. (2016). High-resolution phylogenetic microbial community profiling. *ISME J.* 10, 2020–2032. doi: 10.1038/ismej.2015.249

Slessarev, E. W., Lin, Y., Bingham, N. L., Johnson, J. E., Dai, Y., Schimel, J. P., et al. (2016). Water balance creates a threshold in soil pH at the global scale. *Nature* 540, 567–569. doi: 10.1038/nature20139

Suding, K. N., Lavorel, S., Chapin, F. S., Cornelissen, J. H. C., Diaz, S., Garnier, E., et al. (2008). Scaling environmental change through the community-level: a trait-based response-and-effect framework for plants. *Glob. Chang. Biol.* 14, 1125–1140. doi: 10.1111/j.1365-2486.2008.01557.x

Takahashi, S., Tomita, J., Nishioka, K., Hisada, T., and Nishijima, M. (2014). Development of a prokaryotic universal primer for simultaneous analysis of Bacteria and Archaea using next-generation sequencing. *PLoS One* 9:e105592. doi: 10.1371/journal.pone.0105592

Thijs, S., Op De Beeck, M., Beckers, B., Truyens, S., Stevens, V., Van Hamme, J. D., et al. (2017). Comparative Evaluation of Four Bacteria-Specific Primer Pairs for 16S rRNA Gene Surveys. *Front. Microbiol.* 8:494. doi: 10.3389/fmicb.2017.00494

Thomson, B. C., Ostle, N., Mcnamara, N., Bailey, M. J., Whiteley, A. S., and Griffiths, R. I. (2010). Vegetation affects the relative abundances of dominant soil bacterial taxa and soil respiration rates in an upland grassland soil. *Microb. Ecol.* 59, 335–343. doi: 10.1007/s00248-009-9575-z

Wamelink, G. W. W., Walvoort, D. J. J., Sanders, M. E., Meeuwsen, H. A. M., Wegman, R. M. A., Pouwels, R., et al. (2019). Prediction of soil pH patterns in nature areas on a national scale. *Appl. Veg. Sci.* 22, 189–199.

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/aem.00062-07

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2013). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30, 614–620. doi: 10.1093/bioinformatics/btt593

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.