Parameters that affect soil bacteria diversity and composition: comparing local and global studies

University College London

April 2018

Abstract

The growing awareness of the importance of microbial world has led to the development of novel methods for bacterial community characterisation. Large-scale studies such as the *Earth Microbiome Project* (*EMP*) rely on a relatively standardised procedure of diversity analysis. It is important to understand how soil physical and chemical parameters affect species abundance and diversity within and between samples. In our study, 30 soil samples were collected in Central London and subjected to analysis via the *QIIME* pipeline on a high-performance computing cluster. Soil parameters such as pH and ion concentrations were recorded, and statistical analyses were applied to determine significant correlations. Similar set of analyses was applied on the available *EMP* data and the results were compared. While our data showed no significant results, the *EMP* showed correlations for specific phyla abundance.

I. Introduction

Microbial ecology as a discipline seeks to understand the laws that govern the distribution of species, interactions, and dynamic changes among microbial communities. Community composition is influenced by a variety of factors, including chemical properties of the environment, temperature, and ultraviolet exposure (for free-living communities). The concept of biodiversity is used as a key factor in assessing microbial communities, and recent developments in biotechnology such as community fingerprinting allow to track how they change over time. Furthermore, advances in sequencing methods have led to the discovery of more phyla in the last 10 years than since the discovery of bacteria *per se*[1]. Given that a vast majority of bacteria cannot be cultured through traditional microbiology methods[2], microbial ecology and metagenomics allow to explore the previously unknown bacterial species, elucidate novel biochemical mechanisms, and facilitate fighting antibiotic resistance[18].

Soil microbiology is remarkably interesting in this context. International projects such as the *Earth Microbiome Project* or *TerraGenome* aim for 'constructing a global catalogue of the uncultured microbial diversity of this planet'[3]. *EMP* is an open-source, modular collection of individual studies, each with a specific hypothesis, thus it allows for a granular analysis of data both within each study and across multiple studies. This project relies on a standardised pipeline from sample collection through PCR and

sequencing to comprehensive data analysis methods. Generally, the 16S rRNA genes are used to characterise soil bacteria. 16S rRNA is a component of the 30S prokaryotic ribosome subunit that confers specificity to the Shine-Dalgarno sequence[4]. 16S ribosomal RNA genes represent suitable 'molecular clocks' for various levels of taxonomy[5]. The 'semi-conserved' V4 region is especially useful since it provides resolution at the phylum level as accurately as the whole 16S sequence[6]. Several 16S rRNA gene databases exist, including the most up-to-date *SILVA*[7].
The aim of the present study was to implement the standard microbiome analysis procedure to analyse microbial diversity as well as determine whether and how external factors explain this variation, specifically pH, and amount of NO_3^-, PO_4^{3-} and K^+. Also, it was of interest to establish whether any patterns found in the *EMP* data are demonstrated on a local scale. In our project, 30 soil samples from Gordon Square as well as other locations across London, UK were collected. Environmental parameters such as pH, soil moisture, specific ion concentrations and geographical coordinates were recorded. The obtained sequencing reads were clustered and analysed with the *QIIME* pipeline on a high-performance computing cluster as well as a local machine.

II. Methods

Soil extraction and gDNA isolation

Soil was extracted and collected into *Falcon* tubes. Genomic DNA was isolated using the *DNeasy* DNA isolation kit, according to the manufacturer's protocol. The presence of gDNA in eluents was confirmed by 0.5% agarose gel electrophoresis.

PCR amplification of the V4 region

Barcoded PCR primers allowed to pull samples together before the sequencing step. Barcoding is an important step to allow distinguish between samples when a mixture of many is produced. PCR reaction was conducted at 20 cycles. 25 µL solutions were prepared from a pre-made 2X PCR mix, 0.5 µL of primers, 1µL of template DNA and 10.5 µL H_2O. Primers were designed to target the constant regions flanking the V4 region. Amplicon length was 259 nt.

Forward primer:
AATGATACGGCGACCACCGAGATCTACACGCT (1) XXXXXXXXXXXX (2) TATGGTAATT (3) GT (4) GTGYCAGCMGCCGCGGTAA (5)
1: 5' Illumina adapter; 2: GoLay barcode (specific to each sample); 3: Forward primer pad; 4: Forward primer linker; 5: Forward primer (515fb).

Reverse:
CAAGCAGAAGACGGCATACGAGAT(1) AGTCAGTCAG(2) CC(3) GGACTACNVGGGTWTCTAAT(4)
1: Reverse complement of 3' Illumina adapter; 2: Reverse primer pad; 3: reverse primer linker; 4: Reverse primer (806rB).

Amplified fragments were verified using agarose gel electrophoresis. PCR product was purified using the *QIAquick* PCR purification kit according to manufacturer's protocol and again verified by electrophoresis.

dsDNA quantification, dilution, and sequencing

To quantify the amount of double-stranded DNA, the *SpectraMax Quant* dsDNA assay kit was used. 96-well microplate reader recorded fluorescence signal at 468 nm wavelength. Provided dsDNA standards were used as a control and to construct a standard curve. Amplicons purified from 9 PCR reactions were diluted to equimolar concentrations (1nM) and pooled together. Sequencing was performed at *Illumina MiSeq* machine.

Data analysis

Earth Microbiome Project data: statistical tests were reproduced on the *EMP* dataset (see ref. 3 for the dataset link). The dataset was filtered to contain only soil samples.
HPC cluster: sequencing data were analysed on the Cirrus parallel computing facility at EPCC at the University of Edinburgh. For parallel scripts, the benchmarking procedure was performed to determine the optimal number of cores to be used.

QIIME software workflow:
1. Validate mapping data by checking for errors.
2. Quality filtering and demultiplexing: sequences are trimmed (Phred scores >19) and filtered, and sequences with barcodes not found in mapping file are removed. GoLay barcode allowed to group reads by samples.
3. OTU picking: in our analysis, OTU stands for a single species (97% sequence identity threshold). *UCLUST* clustering algorithm was run at rarefaction depth of 600 reads. SILVA database was used for closed-reference OTU picking.
4. Further analysis involved multiple *QIIME* scripts. OTU distribution among samples was used to infer data to on α- and β-diversity. The diversity metrics used were *Chao1* and unweighted *UniFrac*, respectively. See ref. 19 for the full list of *QIIME* scripts used. Additionally, subsequent work relied on Python modules such as *Numpy*, *Pandas*, and *Matplotlib*.
5. Statistical analysis involved Student's t-test and Spearman correlation (*Rho*) to establish relationship between soil parameters and OTU abundance, with 0.05 significance level.

III. Results

Bacterial community composition

Quality filtering of the sequencing reads reduced the number of reads approximately 3.4-fold, resulting in an array of 11 million reads (for convenience, approximate values are presented). OTU picking step through closed-reference method demonstrated 16.6 thousand distinct clusters overall, whereas the total count of reads assigned OTU exceeded 4.5 million. With the majority of results yielding a rather uniform distribution of total OTU numbers, two samples taken outside the main collection point demonstrated two extreme results: 653 and 1.5 million counts, respectively. Due to using closed-reference OTU picking, 17% of the sequences were removed, as no match was found in the *SILVA* database. For further analysis, the samples were rarefied at depth corresponding to the smallest sample, i.e. 600 reads depth.
The ten most frequent phyla comprise 96% of total reads, thus OTU distribution is highly skewed. In all samples, *Proteobacteria* were the most abundant phyla, representing approximately 30% of calls (figure 1), followed by *Acidobacteria* and *Actinobacteria*. Similar result has been

observed in the literature, including research on Mediterranean soil bacterial diversity[8], which allows to suggest that these three phyla comprise a 'core' soil microbiome across different climate zones.

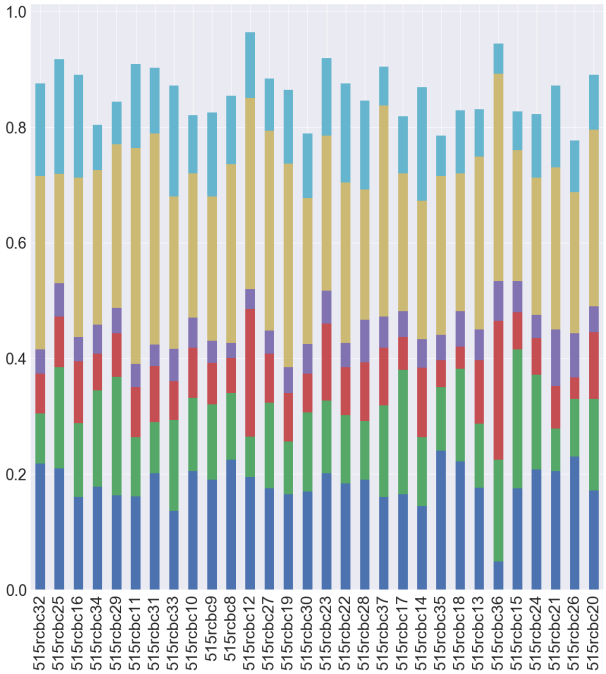

Figure 1. The most abundant phyla across samples. Data filtered at 3% abundance threshold. Empty areas represent filtered sequences and those with no taxonomy determined. Light blue: *Verrucomicrobia*; beige: *Proteobacteria*; violet: *Planctomycetes*; red: *Bacteriodetes*; green: *Actinobacteria*; blue: *Acidobacteria*.

Further analysis (e.g. at a class or order level), was hard to perform because OTU assignment at selected rarefaction depth is less reliable at deeper taxonomic level.

Phylogenetic diversity

Significant variation was observed for α-diversity scores among the samples. As with the OTU picking step, there were two outliers, samples 20 and 12 (figure 2). These samples were collected outside Gordon square, from compost soil and at the edge of Regent's canal, respectively. Notably, these samples were also the ones with the highest and lowest numbers of sequencing reads.

What were the differences between the samples? For each sample, mean UniFrac distance (unweighted metric was used to account for less abundant OTUs) was computed and defined as Uniqueness

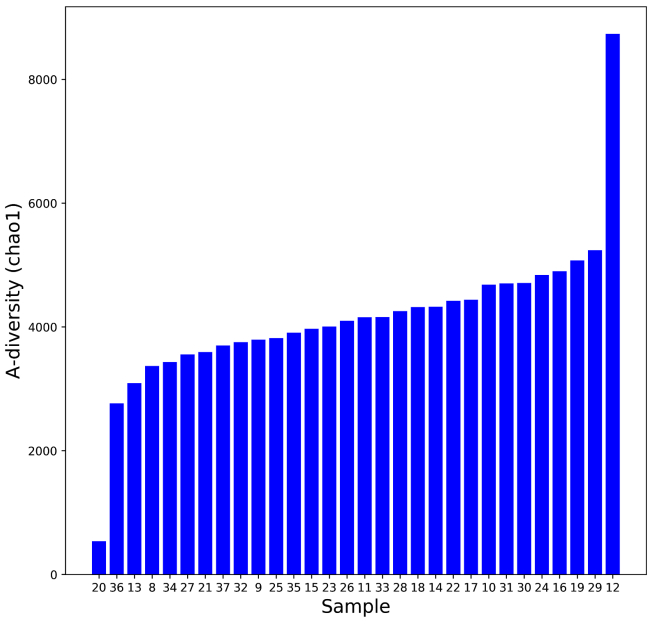

Figure 2. α-diversity scores by sample. Samples 12 and 20 demonstrate significant difference to the remaining samples.

(i.e. how different the sample is from all other samples).

Whereas it was somewhat expected for samples 12 and 20 to differ from the others and likely be most unique, samples 36 and 25 (from the main location) demonstrated the highest uniqueness, of 0.67 and 0.65, respectively, and sample 20 was ranked below average (Figure 3). Sample 13 demonstrated the lowest mean UniFrac value of 0.57.

Metadata-diversity correlation

Which factors explained bacterial community composition and phylogenetic diversity? It was previously shown that soil pH can influence its bacterial composition and

affect α-diversity[9], while factors such as site temperature or latitude have no predictive power[10]. To test whether pH explains α-diversity in our study, samples were split into two categories (high and low pH) and subjected to a t-test, however, the result did not demonstrate

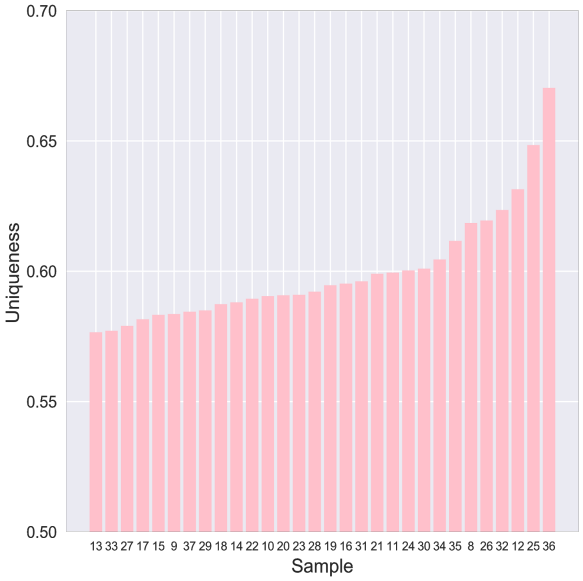

Figure 3. Mean UniFrac distance by sample (uniqueness). Note the y-axis limits. Samples 36, 25, and 12 demonstrated the highest mean proportion of unique species.

statistical significance. Correlation test (Spearman's Rho) demonstrated coefficient of only 0.05. Accordingly, soil nitrate, potassium, and phosphate concentrations did not have significant predictive power (data not shown). Similarly, β-diversity did not appear to be predicted by any of the measured parameters.

To validate the results, same analysis was performed on the *EMP* data. α-diversity/pH distribution for the *EMP* data is presented in figure 4. The scatter plot demonstrates that α-diversity distribution is indeed non-uniform with respect to pH. Samples with the highest diversity have pH close to neutral (pink vertical band). However, samples with low diversity can also be seen in this region, indicating that neutral pH is not an exclusive requirement for high diversity.

Can difference in pH predict phylogenetic distance between samples? For every sample in EMP soil subset, the samples with highest and lowest distances to it (excluding itself) were determined.

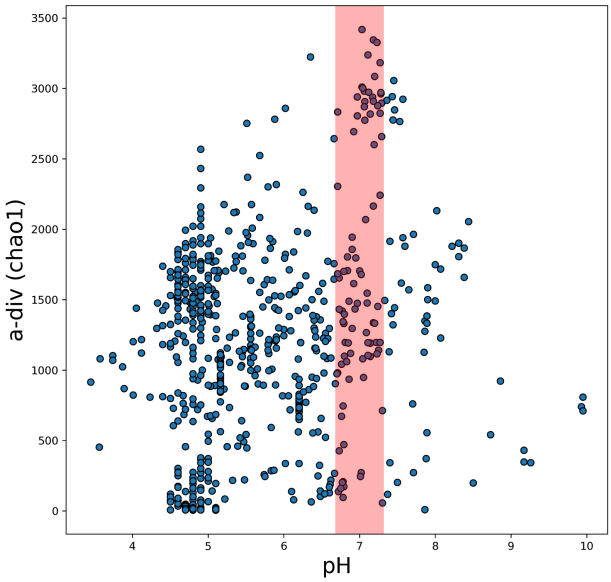

Figure 4. Scatter plot of α-diversity scores (*Chao1*) in *EMP* data against recorded pH. Pink band represents a range of pH values corresponding to the highest α-diversity.

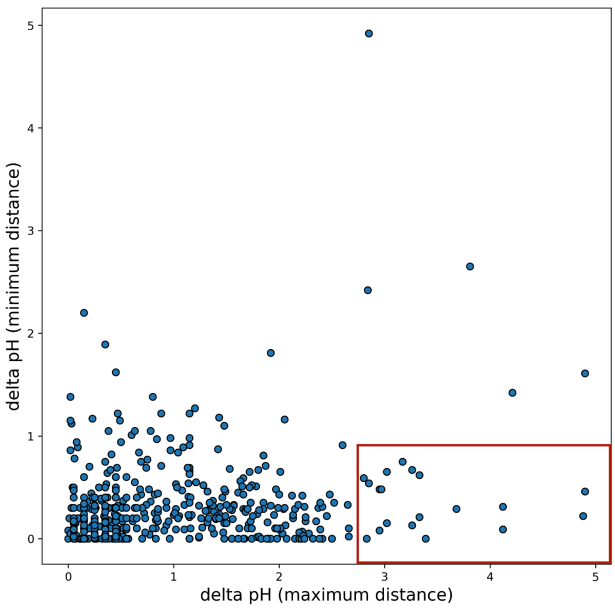

Figure 5. Scatter plot representing ΔpH between most distant samples against ΔpH for the closest sample. Data taken from the *EMP* dataset.

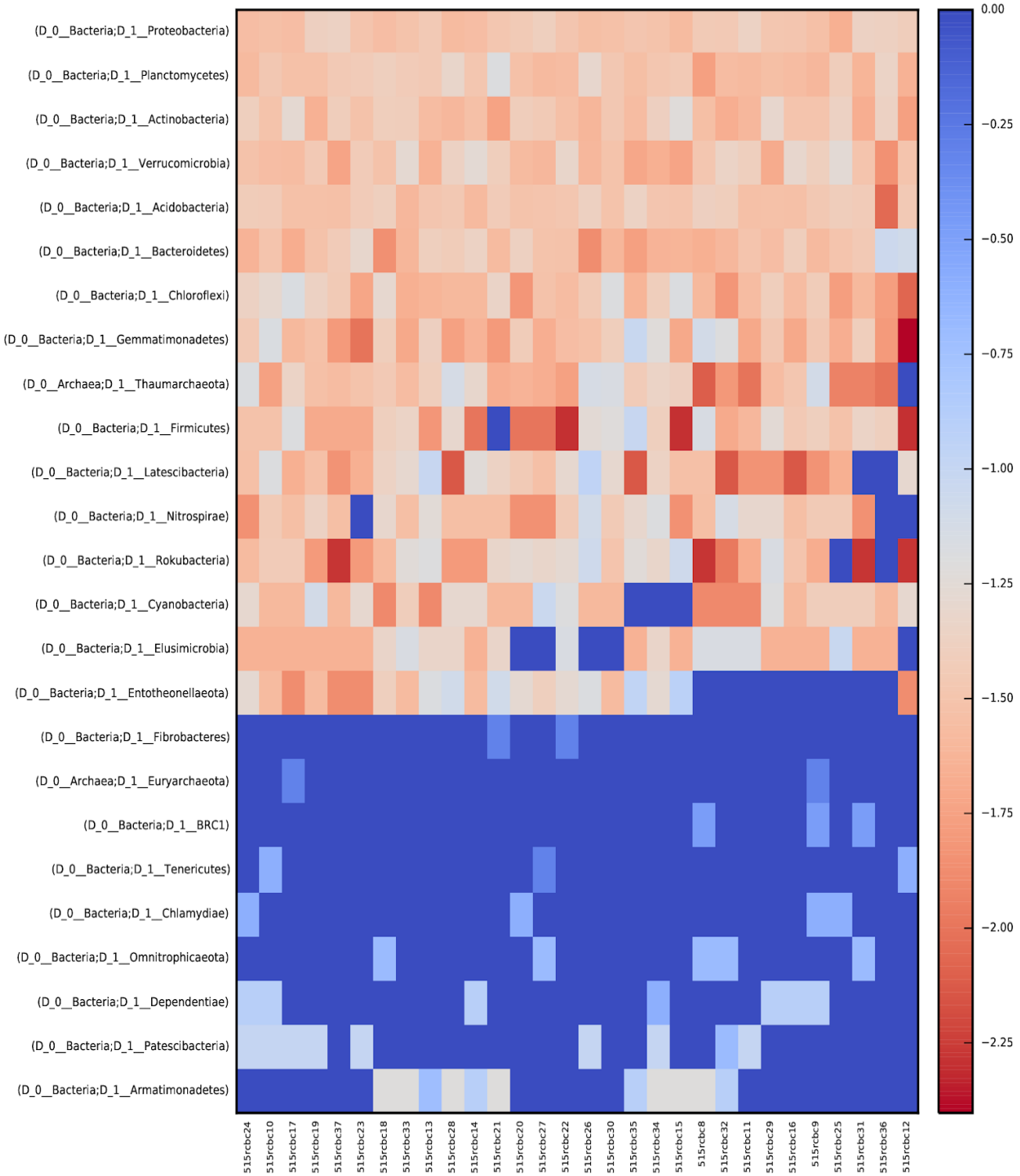

Figure 6. Heatmap of most represented phyla for each sample (our data). The colour scheme on the right represents relative OTU abundance in a sample. Blue indicates low abundance, red indicates high.

Then, pH difference was computed between the sample and each of the two. Figure 5 represents the plot of ΔpH for the most distant sample (x-axis) against ΔpH for the closest sample (y-axis) for the *EMP* data. It can be noticed that whereas the

majority of the samples demonstrate no correlation (lower left quadrant), there are samples for which the hypothesis appears valid (red rectangle). Thus, more experiments are required to establish for which specific biomes and phyla this relationship holds true.

Metadata-OTU correlation

Figure 6 is a heatmap of specific phyla abundance by sample in our data. Whereas the overall most abundant phyla (e.g. *Proteobacteria*; see figure 1) are represented by a relatively constant proportion of OTUs, less abundant ones vary in proportion across samples (e.g. *Rokubacteria* and *Firmicutes*). What is the cause of the variation? It was reasonable to test how metadata parameters affect the abundance of specific OTUs. Correlation analysis of our data did not reveal any significant patterns, so it yet again was repeated on the *EMP* data. pH was indeed a significant predictor for relative abundance of several phyla. Interestingly, both positive and negative correlation coefficients were observed. Phyla abundance appears to be very weakly correlated to nitrate and ammonium ions concentration and moderately correlated to soil depth. Table 1 summarises the key findings.
Hydrogenedentes is an uncultivated phylum associated with methanogenic environments

Table 1. Phyla which relative abundance correlates to specific metadata parameters. Values with reasonably high correlation statistics are shown in bold.

Metadata parameter	Phylum	Cor. coefficient (rho)	p-value
pH	*Hydrogenedentes*	**0.51**	~0
	JL-ETNP-Z39	**0.53**	~0
	Saccharibacteria	**0.43**	~0
	Proteobacteria	-0.35	$3.51*10^{-20}$
	WD272	**-0.47**	$4.3*10^{-38}$
Nitrate	Miscellaneous Crenarchaeotic	-0.17	0.01
	Aenigmarchaeota	0.14	0.04
Ammonium	*Chloroflexi*	0.26	0.0002
Soil depth	Miscellaneous Crenarchaeotic	**0.45**	~0
	Caldiserica	**0.43**	~0
	Atribacteria	0.38	~0
	Proteobacteria	**-0.45**	$1.45*10^{-51}$

which was recently demonstrated to generate H_2 via Na^+-translocating NADH:quinone oxidoreductase[11]. Positive correlation with pH might be explained by the fact that these bacteria are well adapted to sodium-containing media, which are likely to be basic. Soil depth is likely a proxy for the presence of oxygen, which

is confirmed by the observed data: *Caldiserica* and *Atribacteria*, which are anaerobes[12,13] demonstrate positive correlation and typically aerobic *Proteobacteria* - negative.

IV. Discussion

This study aimed at determining which physicochemical properties of the soil affect microbial diversity and composition. As demonstrated, the obtained data failed to provide significant data. Some patterns, however, hold true for the *EMP* data. Perhaps the most interesting finding is that abundance of *Hydrogenedentes* relatively well correlates with pH, as very little was known about this phylum until recently. Additionally, predictive power of pH for β-diversity seems only applicable to certain samples, and it would be worth investigating in depth in future studies.

There are several reasons to why our data did not yield any patterns. Firstly, they were represented essentially by one biome, and thus were unlikely to demonstrate high variation. This experiment would probably be unsuccessful as a standalone study and would only make sense as a single sample in a wider study. Secondly, metadata parameters measurements were performed using kits only providing estimates of ion concentrations and pH (University College London failed to provide the author with an appropriate equipment) and thus are unreliable.

Properly designed experiment must use multiple replicates of samples from various biomes to diversify the results and measure metadata parameters with precise equipment through a standardised procedure. Soil bacterial composition also needs to be monitored during the year, as species abundance was shown to alter depending on a season[14]. Importantly, it was demonstrated that the choice of the 16S rRNA gene region affects the outcome of microbial diversity studies[15]. Even though the V4 region is used predominantly, many studies utilise V3 and V5 regions, and comparing the results to the V4 analyses might lead to incorrect inferences. Furthermore, a recent paper argues that OTU picking strategy is unreliable because the reference databases, due to their limited size, fail to account for a vast majority in sample variation[16]. The authors showed that the Deblur[17] methodology produced informative results without clustering, using the full length reads instead.

In conclusion, our data might add value to broader studies, e.g. as an example of urban park microbiota, since the study was performed under a general microbiome analysis framework, but its scientific value as a separate study remains low due to poor standard of measuring equipment. The *EMP* data showed that several bacterial phyla indeed correlate with physicochemical parameters, allowing to deepen their understanding, which is especially relevant to uncultured and predicted phyla. In order to reach its full potential, the *EMP* studies should incorporate more metadata measurements. The more well-curated data are collected, the more reliable models can be built on their basis. Also, some of the analysis strategies (e.g. OTU picking) might need to be re-evaluated in the future studies.

References

1. Nesme, J. *et al.* Back to the Future of Soil Metagenomics. *Front. Microbiol.* **7,** 73 (2016).
2. Li, L., Mendis, N., Trigui, H., Oliver, J. D. & Faucher, S. P. The importance of the viable but non-culturable state in human bacterial pathogens. *Front. Microbiol.* **5,** 258 (2014).
3. Gilbert, J. A., Jansson, J. K. & Knight, R. The Earth Microbiome project: successes and aspirations. *BMC Biol.* **12,** 69 (2014). *EMP* data set: http://www.earthmicrobiome.org/data-and-code/
4. Case, R. J. *et al.* Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl. Environ. Microbiol.* **73,** 278–288 (2007).
5. Ludwig, W. & Schleifer, K. H. Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. *FEMS Microbiol. Rev.* **15,** 155–173 (1994).
6. Tsukuda, M., Kitahara, K. & Miyazaki, K. Comparative RNA function analysis reveals high functional similarity between distantly related bacterial 16 S rRNAs. *Sci. Rep.* **7,** 9993 (2017).
7. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35,** 7188–7196 (2007).
8. Siles, J. A., Rachid, C. T. C. C., Sampedro, I., García-Romera, I. & Tiedje, J. M. Microbial diversity of a Mediterranean soil and its changes after biotransformed dry olive residue amendment. *PLoS One* **9,** e103035 (2014).
9. Wu, Y. *et al.* Effects of pH and polycyclic aromatic hydrocarbon pollution on thaumarchaeotal community in agricultural soils. *J. Soils Sediments* **16,** 1960–1969 (2016).
10. Fierer, N. & Jackson, R. B. The diversity and biogeography of soil bacterial communities. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 626–631 (2006).
11. Nobu, M. K. *et al.* Microbial dark matter ecogenomics reveals complex synergistic networks in a methanogenic bioreactor. *ISME J.* **9,** 1710–1722 (2015).
12. Mori, K., Yamaguchi, K., Sakiyama, Y., Urabe, T. & Suzuki, K.-I. Caldisericum exile gen. nov., sp. nov., an anaerobic, thermophilic, filamentous bacterium of a novel bacterial phylum, Caldiserica phyl. nov., originally called the candidate phylum OP5, and description of Caldisericaceae fam. nov., Caldisericales ord. nov. and Caldisericia classis nov. *Int. J. Syst. Evol. Microbiol.* **59,** 2894–2898 (2009).
13. Nobu, M. K. *et al.* Phylogeny and physiology of candidate phylum 'Atribacteria' (OP9/JS1) inferred from cultivation-independent genomics. *ISME J.* **10,** 273–286 (2016).
14. Schmidt, S. K. & Lipson, D. A. Microbial growth under the snow: Implications for nutrient and allelochemical availability in temperate soils. *Plant Soil* **259,** 1–7 (2004).
15. Birtel, J., Walser, J.-C., Pichon, S., Bürgmann, H. & Matthews, B. Estimating bacterial diversity for ecological studies: methods, metrics, and assumptions. *PLoS One* **10,** e0125356 (2015).
16. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551,** 457–463 (2017).
17. Amir, A. *et al.* Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems 2: e00191-16. (2017).
18. Charlop-Powers, Z. *et al.* Urban park soil microbiomes are a rich reservoir of natural product biosynthetic diversity. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 14811–14816 (2016).
19. Anonymous. *QIIME* data analysis scripts. 2018; Available from: https://github.com/bike-maron/BIOC3301_project