Problems that arise when peforming a small-scale soil microbiome analysis and how to prevent them

University College London

May 1, 2018

Abstract

Microbiome analysis is an area of research that is currently experiencing a major growth, with over 10000 articles published on the topic in the last 5 years. A severe decrease in price of both the DNA sequencing and computational power in the recent years, led to this data-driven area of research becoming more accessible. In order to determine whether it is possible to generate valid results from a small-scale analysis, 30 samples of soil were collected in Central London and High-Performance Computing clusters were used together with software for mictobiome analysis to assess the uniformity of the samples and discover correlations between the chemical and biological content. This paper focuses on three main aspects - assessing the validity of the data for further use by other researchers, investigating the relationships existing between the samples and analysis of the issues that hamper the research on a dataset of low diversity and size.

I. Introduction

The area of microbial ecology is a comparatively recent discipline - majority of articles on this topic were published in the period of 2013-2017. A multitude of reasons led to such as growth in the area with the biggest one being the constant decrease of price of sequencing techniques, which allows to analyse the genetic composition of soil directly and study the interactions between bacteria that are hard to cultivate *in vitro*.

One of the main areas of development in the field of microbiome analysis is the creation of big databases with a variety of samples from a range of biomes. The most successful project in this area is the EMP - Earth Microbiome Project[1]. It was started in 2010 and focuses on developing a global catalogue of Earth's microbial diversity, which can be used not only to perform studies on large datasets in order to find correlations within them, but also as a point of reference for specialists in other areas looking for insight into the microbial ecology of different locations and environments. The samples that were collected during our research could potentially be included in the EMP however, the assessment of validity of the genetic sequences extracted should be performed.

The research described in this paper was focused on performing such task, as well as investigating whether a small-scale microbiome analysis, in general, can provide valid scientific results. The low number of samples (30) that were collected during this experiment, the low diversity of conditions they were collected in, coupled with an inadequate

quality of some kits used for analysis created an array of problems, which are going to be described in this paper, along with the methods.

Our method was based on extracting the 16S DNA from the prokaryotes in the samples collected, using a variety of kits. Since soil is known to have a very high concentration and diversity of bacteria[2, 3], such methods produce very high amounts of data, which has to be analysed using specialised software.

II. Methods

The analysis was performed on 30 samples of soil collected in Central London in October 2017. Metadata, such as moisture, temperature, footfall was collected on the spot, while other aspects such as pH and concentrations of various ions was measured in the laboratory using the *HI3895 Soil Testing Kit*. Then 4 different kits were used in a workflow designed to extract the sequences of 16S DNA from the prokaryotes in the soil, following the instructions provided by manufacturer of each kit.

First, DNA was extracted using the *DNeasy PowerSoil Kit* and PCR primers were designed that contained Golay barcodes, allowing multiple samples to be sequenced simultaneously. *BioMic PCR Kit* was used with the primers to perform the PCR. The solution acquired as the result of PCR was purified using *QIAquick PCR Purification Kit* and then the concentration of dsDNA was measured using *SpectraMax Quant AccuClear Nano dsDNA Assay Kit*. Using the information on DNA concentration all of the samples were subsequently diluted to equal concentrations and sequencing was done on *Illumina's MiSeq* sequencer. Quality control was performed throughout this stage of the experiment. This workflow returned the DNA sequences that were used for downstream *in silico* analysis.

In silico stages of the research were performed using the QIIME[1] package[4, 5], which allows users to do high-throughput sequence analysis. Parts of analysis that required severe computational power were performed on the Cirrus High-Performance Computing system. QIIME package has a basic pipeline which was used to transform the raw data into the form that can be used for statistical analysis, which is described below.

The first step is validation of the initial data, such as mapping and fasta files, followed by read joining, which joins the forward and reverse fasta files. After that, sequences are grouped back into 30 groups that represent the 30 samples. This stage is commonly called demultiplexing, and involves removal of the barcodes that were introduced in the PCR stage. After the demultiplexed fasta file is produced, OTU (Operational Taxonomic Unit) are picked. OTU picking groups closely related sequences (97% similarity in our case, which corresponds to species-level sequence identity) into one operational unit, which are then used for further analysis.

OTU picking requires a referencing database, for which SILVA[2][6] database was used, which is more up-to-date than Greengenes[7] database provided with QIIME. OTU picking and demultiplexing are generally the most expensive stages, in terms of computational power. For this reason they were performed on the Cirrus HPC and most of the downstream analysis was performed on a local machine.

QIIME package provides a variety of scripts designed to analyse the genetic sequences data. However, it should be noted that while providing users with variety of methods to perform the analysis, the figures created by the scripts provided are subpar. This lead to a collection of scripts that were used to perform some of the analysis and construct the figures, which can be found on the GitHub repository dedicated to this analysis[8]. The repository also features additional data on the samples that is out of scope of this paper, and was created using various python packages, such as numpy, scipy, pandas, seaborn and matplotlib, which greatly aid in data visualisation and statistical calculations.

[1] Version 1.9.1

[2] Release 132, April 10, 2018

A method that is useful to determine whether the samples we collected are representative of the biome is provided in the Sourcetracker package[9], which allows users to track proportion of each *source* in each *sink*, *source* and *sink* being different samples of soil from Release 1 of the EMP and our samples respectively. Sourcetracker package is available with QIIME however, Sourcetracker2 has been developed recently, which was used due to being more accurate and functional. The Sourcetracker workflow produces a matrix that displays how much of each source is present in each of our samples.

In an attempt to find a correlation between metadata and the diversity of our samples, scripts included in the QIIME package were used and metadata was transformed into numeric form, with results from *HI3895 Soil testing Kit* treated as values on exponential scale. Analysis was performed using ANOSIM[10], PERMANOVA[11] and a collection of other statistical methods, which calculate the correlation between the distance matrix and pH, potassium, nitrogen and phosphorus content.

Furthermore, to assess the taxonomic composition of our samples a taxonomic assignment using the SILVA[6] database was done. The classical microbiotic analysis was performed as well, that tests the diversity of the samples.

III. Results

During the OTU picking stage, over 4.5 million sequences were recovered from raw data, with 16448 of them being unique. However, it should be noted, that due to use of closed-reference OTU picking, almost 800 thousand sequences (17%) were removed from the downstream analysis because no match was found in the databases.

Using the OTU table that was produced as a result of the OTU picking, alpha diversity was measured, which assesses the number of unique OTU for each sample. Chao1 was chosen as a metric of alpha-diversity for this test, which estimates species richness based on a matrix of abundance

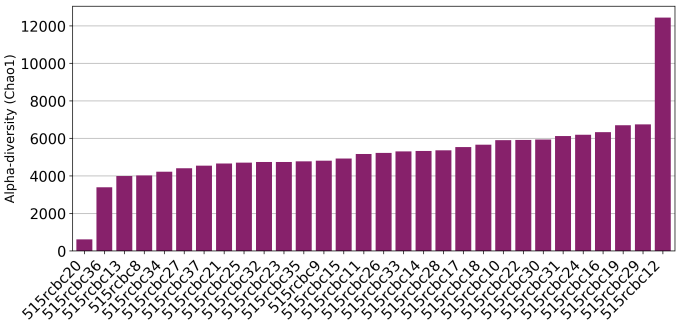

Figure 1: *Alpha diversity (Chao1) for each of 30 samples.*

data. The distribution of alpha diversities between the samples is presented in Figure 1. As we can see from the Figure, most samples have similar values for alpha diversity, with a mean of 5188. Samples *515rcbc12* and *515rcbc20* are the outliers with alpha diversity of 608 and 12429 respectively. If the 2 outliers are removed, the standard deviation is for the dataset is 843, which shows that the samples are tightly grouped.

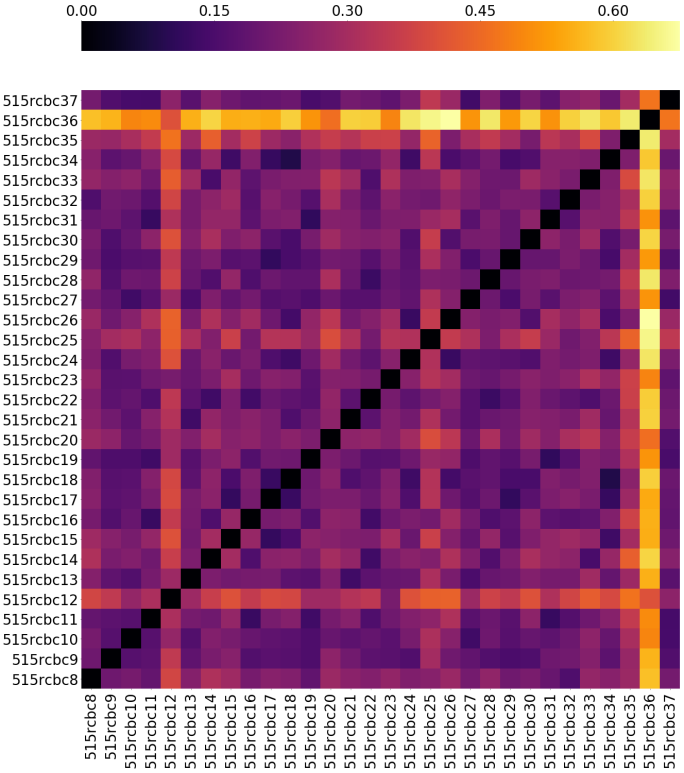

Figure 2: *Beta diversity (unweighted unifrac) matrix, presented as a heat map.*

In order to further investigate the uniformity and clustering of the samples, beta-diversity analysis was performed. The heat map presented in Figure 2 is a graphical representation of beta-diversity between the samples, with higher values representing higher divergence.

Sample *515rcbc36* can be identified as the most divergent from the rest, with *515rcbc12* being an outlier as well. The mean beta-diversity without sample *515rcbc36* is 0.237, with standard deviation of 0.066, further supporting the hypothesis of the uniformity of the samples.

In order to determine whether the data obtained in a small-scale microbiome analysis is sufficient to find correlations between the metadata and diversity several statistical analyses were performed. First, statistical method of ecology analysis ANOSIM[10] was used together with the collected metadata and distance matrices to find any correlations, results are presented in Table 1.

Category	R-value	p-value
pH	-0.109	0.869
Potassium	0.224	0.023
Nitrogen	0.120	0.120
Phosphorus	0.201	0.058

Table 1: *Correlation between metadata and beta-diversity, calculated using ANOSIM[10] method, with external data treated as values on exponential scale.*

Since the R-value is close to 0 for all four parameters, it can be safely assumed that there is no statistically significant correlation between the chemical properties and the diversity in our samples. Other methods of statistical analysis, such as PERMANOVA[11] confirmed this assumption because attempts to find non-linear correlations were also unsuccessful.

The assessment of taxonomic composition of the samples was performed as well, with the taxonomic referencing database included in the SILVA database used to assign taxonomy to most of the OTUs. Figure 3 highlights the top 10 most abundant phyla across all of the samples. These 10 phyla represent 96% of the sequences. More detailed plots on the topic of taxonomy are also available on GitHub[8].

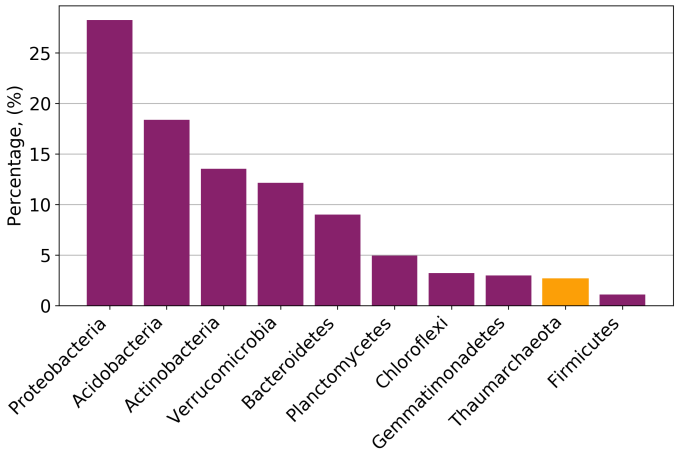

Figure 3: *Proportion of most popular phyla in samples. The orange bar represents the only phylum from the Archea domain. These 10 phyla represent 96% of species in our samples.*

The only phylum from the *Archea* domain is *Thaumarchaeota* and the rest of the sequences have a bacterial origin. The overall taxonomy analysis[3] showed that most samples have a relatively similar proportion of various phyla, with samples such as *515rcbc20*, *515rcbc36* and *515rcbc12* being the only outliers.

Results from the Sourcetracker analysis are presented in Figure 4. The heat map shows that most samples have high relatedness to 4 different biomes - national park soil, field soil, agricultural soil and wetland. Sample *515rcbc20* had to be excluded from the analysis due to low number of OTUs observed (653). The *sources* with lowest relatedness to *sinks* are excluded as well.

Despite the peculiarity of some results of the Sourcetracker analysis, in general it confirms the uniformity and good clustering of the samples. The low number of samples collected prevents us from making any further assumptions about the outliers and unusual correlations between the EMP soil data and the samples, and further research is required to explain high proportion of wetland

[3] can be found on the GitHub[8]

source in many samples and other aberrations.

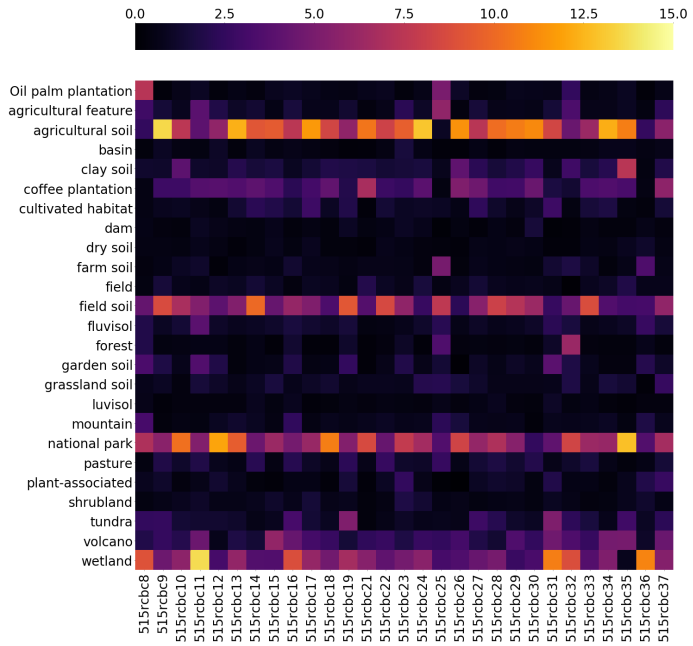

Figure 4: *Heat map showing the percentage of each source in each sink, filtered to represent the top 25 sources. The "Unknown" column is not shown however, it represents a mean of 44% of sources in each sink. As we can see there are 4 prominent sources in most sinks - agricultural soil, national park soil, wetland soil and field soil, which have a mean percentage of over 5%.*

IV. Discussion

The results of the first steps in the computational pipeline show that the approach based on OTU picking does not provide the most accurate results. For instance, the SILVA database currently contains 177222 sequences, which means that our study, while being rather small, covered almost 10% of the database. Thompson et al.[12] report that a study they performed using just under 100 samples from different biomes covered 47% of the SILVA database. Low coverage of even the most up-to-date databases such as SILVA combined with the diminishing property of the OTU picking process creates the need for more accurate methods of sequence assessment, such as Deblur[14].

As we can see from the results of alpha and beta diversity analysis, our samples form a uniform collection, with samples *515rcbc12*, *515rcbc20* and *515rcbc36* forming an array of aberrations. And while the first two were expected outliers due to the altered condition they were collected in, sample *515rcbc36* is an unpredicted outlier, since it had an identical metadata to other samples, which suggests that the sample was contaminated. Further analysis is required to prove this hypothesis however, if the samples were to be included in the EMP it is preferential the three outliers.

The hypothesis of the 27 samples forming a uniform sample is supported by the taxonomic assessment of the samples. While most samples show a similar composition, the three outlier samples mentioned above have atypical proportions of various phyla, confirming them as outliers.

Overall, taxonomy of the dataset as a whole is quite typical and is similar to the results of park soil taxonomy analysis performed by Zhalnina et al.[13]. Their research was performed on samples of soil from a park in United Kingdom however, with a higher focus on taxonomy. Similarity of our results further strengthens the hypothesis that our dataset, with the exception of the outliers is a good representative of urban park soil and is valid enough to be included in the EMP.

Sourcetracker analysis produced quite intriguing results. Since our samples were collected in Gordon Square in Central London, an anthropomorphic biome, high relatedness to agricultural soil and field soil was expected. The unifying feature of such environments is a heavy use of fertilizers, which leads to development of specific ecosystems. This confirms that our samples can be related to wider datasets such as EMP. However, the aberrations such as high relatedness to the wetland biome require further investigation.

The absence of correlation within out dataset happened for a variety of reasons. Firstly, the quality of kits used to obtain the metadata was subpar and such measurements should be performed using appropriate scientific equipment. Secondly, the uniformity and tight clustering of the samples hinders any attempts to discover correlations between

the genetic data and metadata. A purposeful diversification of samples is paramount in such experiments, for example difference in location, season or other conditions. Since 27 of the samples were collected from the same location at the same time, there is very low variance in the data and the few outliers can only provide anecdotal evidence.

However, it must be noted that the samples would be a good addition to the existing microbiome databases such as EMP, since it was confirmed by multiple tests that the majority of the dataset, with the exception of 3 outlier samples is rather uniform an is representative of soil in urban biome. Nonetheless, stand-alone scientific value of the dataset being minuscule, due to absence of correlation between the chemical composition of the samples and the DNA sequences extracted. In order to obtain data that can provide researchers with statistically significant results, the diversity of samples, methods of metadata collection must be improved and use of reference-free sequence analysis techniques is preferential.

In conclusion it must be said that the experiment described in this paper can be well epitomized by the proverb "a chain is only as strong as its weakest link". While the methods of *in silico* analysis were on a high level and extraction of te DNA from soil samples was performed without major complications the poor quality of methods in some crucial steps led to this research not providing scientifically valuable results.

References

[1] Gilbert JA, Jansson JK, Knight R. The Earth Microbiome project: Successes and aspirations. BMC Biol. 2014 aug;12(1):69. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25184604.

[2] Nannipieri P, Ascher J, Ceccherini MT, Landi L, Pietramellara G, Renella G. Microbial diversity and soil functions. Eur J Soil Sci. 2003 dec;54(4):655–670. Available from: http://doi.wiley.com/10.1046/j.1351-0754.2003.0556.x.

[3] Torsvik V, Sørheim R, Goksøyr J. Total bacterial diversity in soil and sediment communitiesA review. J Ind Microbiol Biotechnol. 1996 sep;17(3-4):170–178. Available from: http://link.springer.com/10.1007/BF01574690.

[4] Chen HM, Lifschitz CH. Preparation of fecal samples for assay of volatile fatty acids by gas-liquid chromatography and high-performance liquid chromatography. Clin Chem. 1989 may;35(1):74–76. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20383131.

[5] Kuczynski J, Stombaugh J, Walters WA, González A, Caporaso JG, Knight R. Using QIIME to analyze 16s rRNA gene sequences from microbial communities. Curr Protoc Microbiol. 2012 nov;Chapter 1(SUPPL.27):Unit 1E.5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23184592.

[6] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. 2013 nov;41(D1):D590–D596.

[7] McDonald D, Price MN, Goodrich J, Nawrocki EP, Desantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 2012 mar;6(3):610–618. Available from: http://www.nature.com/doifinder/10.1038/ismej.2011.139.

[8] Anonymous. QIIME visualisation scripts. 2018;Available from: https://github.com/nameisBaron-MichaelBaron/BIOC3301/.

[9] Knights D, Kuczynski J, Charlson ES, Zaneveld J, Mozer MC, Collman RG, et al. Bayesian community-wide culture-independent microbial source tracking. Nat

Methods. 2011 jul;8(9):761–765. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21765408.

[10] CLARKE KR. Nonparametric multivariate analyses of changes in community structure. Aust J Ecol. 1993 mar;18(1):117–143. Available from: http://doi.wiley.com/10.1111/j.1442-9993.1993.tb00438.x.

[11] Tang ZZ, Chen G, Alekseyenko AV. PERMANOVA-S: Association test for microbial community composition that accommodates confounders and multiple distances. Bioinformatics. 2016;32(17):2618–2625. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27197815.

[12] Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017 nov;551(7681):457–463. Available from: http://www.nature.com/doifinder/10.1038/nature24621.

[13] Zhalnina K, Dias R, de Quadros PD, Davis-Richardson A, Camargo FAO, Clark IM, et al. Soil pH Determines Microbial Diversity and Composition in the Park Grass Experiment. Microb Ecol. 2014;69(2):395–406.

[14] Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, et al. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. mSystems. 2017;2(2):e00191–16. Available from: http://msystems.asm.org/lookup/doi/10.1128/mSystems.00191-16.