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

University College London

April 30, 2018

Abstract

I. 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 specialised software to assess the uniformity of the samples and discover correlations between the chemical and biological content. This paper will be focused on three main aspects - assessing the validity of the data for further use by other researchers, investigating the relationships that exist between the samples and analysis of the issues that hamper the research on a dataset of low diversity and size.

II. Introduction

One of the main areas of development in the field of microbiome analysis in recent years was 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 before this can happen.

The research described in this paper focuses on performing the validity assessment of the genetic data that was collected using low-grade equipment, 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 inadequate quality of some kits used for analysis created an array of problems that are going to be described.

III. 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 performed 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[2, 3], which allows users to perform high-throughput sequence analysis. Parts of analysis that required severe computational power were performed on the Cirrus High-Performance Computing system. QIIME package has basic a pipeline which was used to transform the raw data into the form that can be used for statistical analysis, which is described below.

First step is validation of data, including the mapping and fasta files. After that, sequences are grouped into 30 sets which corresponds to 30 samples. This stage is commonly called demultiplexing, and involves emoval 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 was used for further analysis.

OTU picking requires a referencing database, for which SILVA[2][4] database was used, which is more up-to-date than Greengenes[5] 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 a users with variety of methods to perform the analysis, the figures created by the scripts provided are subpar. This lead to creation of 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[6]. 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.

A method that is useful to determine whether the samples we collected are representative of the source, is provided in the sourcetracker package[7], 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

[1] Version 1.9.1

[2] Release 132, April 10, 2018

been developed recently, which was used due to being more accurate and functional. The source-tracker workflow produces a matrix the displays how much of each source is present in each of our samples was produced.

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[8], PERMANOVA[9] and a collection of other statistical methods, which calculate the correlation between the distance matrix and pH, potassium, nitrogen and phosphorus content.

To assess the taxonomic composition of our samples, taxonomic assignment using the SILVA[4] database was performed. The classical microbiotic analysis was performed as well, that tests the diversity of the samples.

IV. 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. The data is consistent with results obtained in the research performed by Thompson et al.[10] on the data from the EMP database, in which they reported that amongst soil samples 20% of sequences were not identified with SILVA closed-reference OTU picking.

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. While different methods and metrics can be used, Chao1 was chosen for this test, which estimates species richness based on a matrix of abundance data. The distribution of alpha diversities between the samples is presented below in Figure 1.

In order to further investigate the uniformity

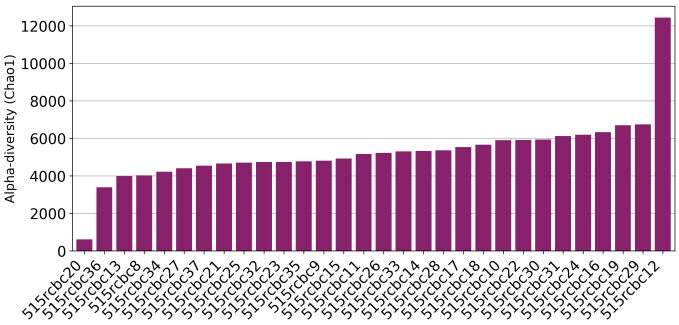

Figure 1: *Alpha diversity (Chao1) for each of 30 samples. Samples 515rcbc20 and 515rcbc12 exhibit highly distant results, can be explained by their different geographical origin.*

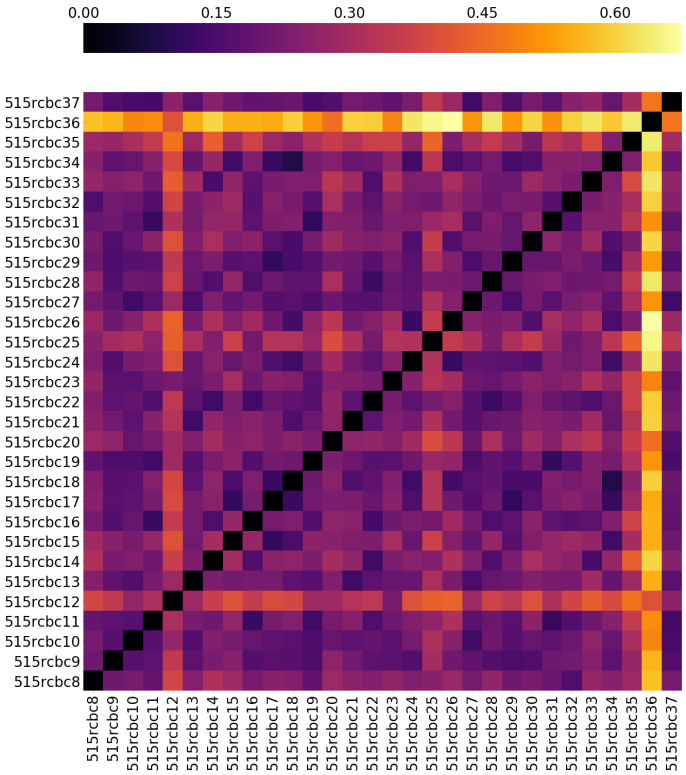

Figure 2: *Beta diversity (unweighted unifrac) matrix, presented as a heat map. Obvious outliers can be noted again, with samples 515rcbc12 and 515rcbc36 having the largest divergence from other samples.*

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 a close contender. While *515rcbc12* has outlying values for the alpha diversity with almost 2-fold difference in alpha diversity compared to all other samples, *515rcbc36* does not exhibit any severe difference in other results when compared to the rest. Moreover, *515rcbc12* was collected in a different geographical location and different conditions, which can explain its high beta-diversity scores however, *515rcbc36* was collected in conditions that are identical to other samples. This leads to the assumption that that this sample was contaminated. This assumption is backed up by the taxonomical assignment of OTUs, which shows that sample *515rcbc36* has different proportions of various taxa compared to other samples.

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[8] method.*

Despite the two aforementioned samples being the obvious outliers, sample *515rcbc20*, which was collected in different conditions and displayed an atypical alpha diversity does not differ severely from the other samples on this heat map. This happens due to the nature of unweighted unifrac beta diversity metric - it takes into account the total number of OTUs in the sample, thus nullifying the difference in the total number of samples.

In order to determine whether the data obtained in a small-scale microbiome analysis is enough to find correlations between the metadata and diversity, several statistical analyses were performed. First, statistical method of ecology analysis ANOSIM[8] was used together with the collected metadata and distance matrices to find any correlations, results are presented in Table 1.

Since the R-value is close to 0 for all four metrics, it can be safely assumed that there is no statistically significant correlation between metadata and the diversity in our samples. Other methods of statistical analysis, such as PERMANOVA[9] confirmed this assumption, attempts to find non-linear correlations were not successful as well.

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[6].

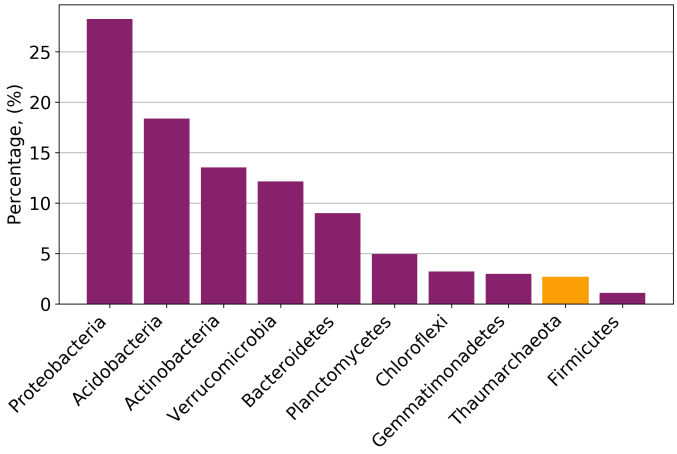

Figure 3: *Proportion of most popular phyla in samples. The orange bar represents the only phylum from the Archea domain. These 10 phyla represent 96% of species in our samples.*

The only phylum from the *Archea* domain is *Thaumarchaeota* and the rest of the sequences have a bacterial origin. This is consistent with a similar research conducted by Zhalnina et al.[11], in which they investigated the taxonomy of the Park Grass Experiment. The proportions of the most abundant phyla is similar, which speaks in favour of the validity of our samples.

The sourcetracker package was employed to assess the relatedness of our samples to samples of soil from different biomes around the world collected by the researchers working on the EMP projects . Sourcetracker exploits a mathematical

approach which assesses the proportion of each *source* in each *sink*. The results are a matrix , which is presented in Figure 4 as a heat map .

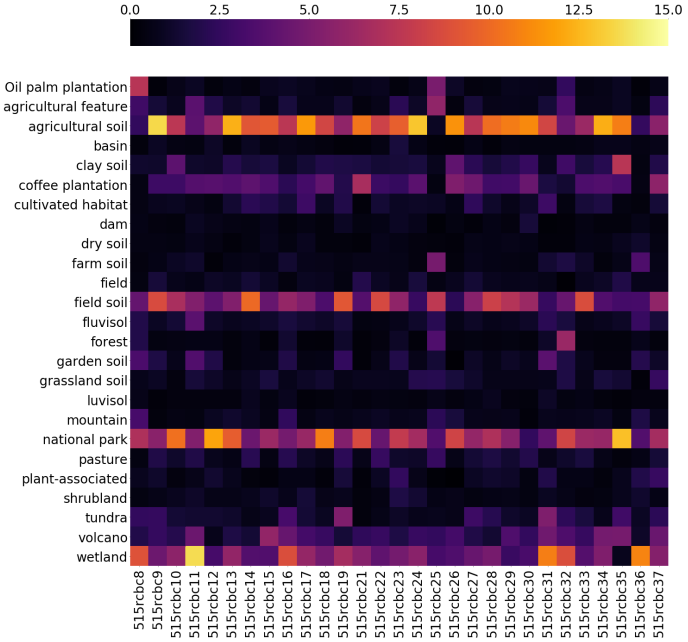

Figure 4: *Heat map showing the percentage of each source in each sink, filtered to represent the top 25 sources. The "Unknown" column is not presented as well, it represents a mean of 44% of sources in each sink. As we can see there are 4 prominent sources in most sinks - agricultural soil, national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

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). Our samples were collected in Gordon Square in Central London, a biome that is anthropomorphic, which can explain the high relatedness to agricultural soil and field soil. The unifying feature of such environments is heavy use of fertilizers, which leads to development of a specific ecosystem[11]. This further reinforces the hypothesis that except for a small number of outliers mentioned above, the samples present genetic data of a high enough quality to be included in the EMP.

Despite the peculiarity of some results of the sourcetracker analysis, in general it confirms the uniformity and good clustering of the the samples. Low number of samples collected prevents us from making any further assumptions about the outliers and peculiar correlations between the EMP soil data and the samples, and further research is required to explain high proportion of wetland source in many samples and other aberrations.

V. 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.[10] 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[12].

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 abberations. 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. However, further analysis is required to prove test the hypothesis that this sample was contaminated.

The uniformity and tight clustering of the samples, while validating the dataset for comparison with other datasets at the same time hinders any attempts to discover correlations between the genetic data and metadata. In order to find any correlation a purposeful diversification of samples is required, such as variance of location, season or other conditions. Since 27 of the samples were collected in the same location at the same time, there is very low variance in the samples and the few

outliers can only provide anecdotal evidence. If a similar experiment is to be performed by other researchers, diversification of samples should be considered a paramount step.

In addition to diversification of samples, high-grade methods of metadata collection are required. While the kits that were provided by University College of London give a general idea of concentrations of various ions in the soil, they do not match the level of this research.

In conclusion, it must be noted that the samples provide a good addition to the existing microbiome databases such as EMP, since it was confirmed by multiple *in silico* tests that the protocol of extracting the DNA from samples was executed without major complications. However, stand-alone scientific value of the dataset is 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.

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] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.