


# Coursework Declaration and Feedback Form

Student Number: 2636906	Student Name: Wenhan Chan
Programme of Study (e.g. MSc in Electronics and Electrical Engineering): MSc in Computer system Engineering	
Course Code: ENG5059P	Course Name: MSc Project
Name of <b>First</b> Supervisor: Dr Umer Zeeshan Ijaz	Name of <b>Second</b> Supervisor: Dr Lukasz Kaczmarczyk
Title of Project: Exploring community assembly mechanisms for microbial community surveys MSc	
<b>Declaration of Originality and Submission Information</b>	
<i>I affirm that this submission is all my own work in accordance with the University of Glasgow Regulations and the School of Engineering requirements</i> Signed (Student): <i>Wenhan Chan</i>	 E N G 5 0 5 9 P
Date of Submission: 18 Aug 2022	
<i>Feedback from Lecturer to Student – to be completed by Lecturer or Demonstrator</i>	
Grade Awarded: Feedback (as appropriate to the coursework which was assessed):	
Lecturer/Demonstrator:	Date returned to the Teaching Office:



University of Glasgow | School of Engineering

# Analyzing the Microbial Ecology of Contaminated Soils

Wenhan Chan

2636906C

Supervised by Dr. Umer Zeeshan Ijaz

Co-supervised by Dr. Lukasz Kaczmarczyk

August 18, 2022

A thesis submitted in partial fulfilment of the requirements for the  
degree of  
MASTER OF SCIENCE IN COMPUTER SYSTEM  
ENGINEERING

## Abstract

The soil pollution has always been a threat to human health. Improper disposal of pollutants will only increase the degree of soil pollution and cause harm to the entire ecological environment. In order to solve the soil contamination problem using microorganisms for the biodegradation of contaminated soils, it is extremely important to understand microbiological community interactions and relationships involved in the bacterial degradation within contaminated soils, especially considering the organic compounds of polycyclic aromatic hydrocarbons (PAHs). Polycyclic aromatic hydrocarbons (PAHs) which can be generated as organic pollutants from combustion processes of coal and biomass. It contains carcinogenic or mutagenic substances, which endanger the ecological environment and pose a great threat to animals and plants. The study of the biodegradation of PAHs that occurs by microorganisms is therefore critically influential.

The aim of this project was to perform a cross-sectional comparison of soil microbial communities in two contaminated sites, which are COV (United Kingdom) and CH (Switzerland). Both sites had 16S rRNA samples and were contaminated with PAHs. Genomic DNA sequencing of COV soil samples and CH soil samples by nucleic acid sequencing. The sequenced sample data were converted into OTU tables to analyze the variances in COV and CH soil samples using R Studio software in the OTU tables. The results showed that the COV samples seemed to be significantly contaminated with more PAHs than the CH samples.

This research involves the development of an analytical strategy guided by recent advances in the microbiological community assemblage mechanisms to elucidate the role of the environment and reveal that *Pseudomonas* is a core member of the predominant microbial communities of COV and CH soil samples.

Keywords: 16S rRNA, Biodegrade, Contamination, Diversity, Microbial Community, OTU, PAHs, *Pseudomonas*

## **Acknowledgments**

I would really like to acknowledge Dr Umer Zeeshan Ijaz who gave me the chance to learn about ecology-related topics. Not only did he spend at least four hours a week helping us to strengthen the biological aspects and how to analyse our data in R, but he also actively monitored our progress each week. As long as we followed the tutor's lead, the thesis went well. I am very grateful to him for always giving me clear directions when I had problems and teaching us a lot of presentation skills. I have to be thankful for my master's classmates who were always willing to answer me and discuss with me when I needed help.



## **List of Abbreviations**

ASV Amplicon Sequencing Variant

COV The sample from United Kingdom area

CH The sample from Switzerland area

DNA Deoxyribonucleic Acid

GC × GC-MS Gas Chromatography Coupled with Mass Spectrometry

LOI Loss on Ignition

LOOCV Leave-one-out cross-validation

MDP Mean Systematic Developmental Diversity

MNT Dmean Nearest Taxon Distance

NRI Net Relatedness Index

NST Normalized Stochasticity Ratio

NTI Nearest Taxon Index

PAHs Polycyclic Aromatic Hydrocarbons

PCoA Principal Coordinate Analysis

OTU Operational Taxonomic Unit

rRNA ribosomal Ribonucleic Acid

qPCR quantitative Polymerase Chain Reaction

QPE Quantitative Process Estimate

SSU Ribosomal Small Subunit

SO<sub>2</sub> Sulphur Dioxide

SO<sub>3</sub> Sulphur Trioxide

VIF Variance Inflation Factor

## List of Figures

Figure 1: <i>The lifetime landfill degradation of plastics and bioplastics</i> .....	2
Figure 2: <i>The 16S rRNA and 16S rRNA Gene (Source: EZBioCloud Help center, 2019) [18]</i> .....	4
Figure 3: <i>The data flow in microbial data analysis from Dr. Ijaz</i> .....	7
Figure 4: <i>The visualization paradigm for the core microbiome</i> .....	14
Figure 5: <i>a) The example of Bland-Altman plot (Source: Wikipedia) [35]. b) The example of MA plot (Source: Wikipedia) [36].</i> .....	16
Figure 6: <i>a) The result of alpha diversity and NRI/NTI measures for the COV and CH soils populations. b) The beta diversity measured by (a) Bray-curtis, (b) Unifrac and (c) Weight Unifrac distances is displayed by PCoA plots, in which coloured ellipses indicate normative errors. The first 25 most abundant genera in each species group in the PCoA diagram are also illustrated surrounding the key to the (d) taxa table</i> .....	22
Figure 7: <i>a) The results of QPE. The contribution percentages for each assembly process are displayed, where Dispersal limitation, Homogenizing dispersal, Ecological drift (Undominated) belong to stochastic, whereas homogeneous selection and variable selection are categorized as deterministic. b) The results of the beta RC for COV and CH soil samples. When the <math>\beta</math> RC value does not deviate appreciably from 0, the community is attributed to the random assemblage; if the <math>\beta</math> RC value deviates from 0 and gradually approaches 1 or -1, it indicates a definite clustering of the community. c) The results of the NST. Quantified results for each assembly process. Stochasticity ratios are based on 0.5. When the value is less than 0.5, the community is more deterministic.</i> .....	24
Figure 8: <i>Core microbiome heatmap persisted in &gt;85% of samples used to compare all populations separately and both COV and CH communities were compared simultaneously. OTUs are sequenced according to their abundance, the upper part of the heat map indicating the low abundance prevalent OTUs and the bottom showing the high abundance prevalent OTUs.</i> .....	26
Figure 9: <i>The results of the MA plot. The x-axis of the MA plot represents the mean value, and the y-axis represents the log difference value. The baseline is approximately logfold=0. When the log difference for each microorganism is between plus and minus 2 logfold, it shows a black dot indicating that the difference is not excessive. The red dots indicate a significant difference when the difference exceeds +2 logfold or falls below -2 logfold.</i> .....	28
Figure 10: <i>a) The significant different of genus. b) The significant different of phylum. The graph shows which microorganisms are significantly differentiated from each other. The more types of microbial variation, the more significant the distribution of red dots on Figure 9.</i> .....	29
Figure 11: <i>The meta table of pollutants in COV and CH soils such as lead, iron, cadmium, chromium, zinc, copper, nickel and cobalt, loss on ignition (LOI) percentage. The LOI percentage is the lab test to see how much is lost when a soil sample is heated to a certain temperature, and moisture percentage.</i> .....	31
Figure 12: <i>The meta table of the aromatic hydrocarbons in COV and CH soils, the parameters in the graph are related to the biodegradation of PAHs.</i> .....	31
Figure 13: <i>The Cross-validation errors of models, to determine the error value of the 20 models. The lower the value of Cross-validation errors, the better the results of subset regression.</i> .....	32
Figure 14: <i>The result of subset regression. The distribution corresponds to the estimate values in the table. Negative values indicate that the more the substance is present, the lower the Shannon diversity will be. The higher the value of * shows the more significant.</i> .....	32
Figure 15: <i>The CODA GLMNET result of (a) Naphthalene (b) Dibenzo Ah Anthracene (c) Cadmium (d) Cobalt. When the coefficient is greater than zero, which is the green area on the figure, it means that the growth of this substance has a positive correlation with those microorganisms. Conversely, if the coefficient is less than zero, it is shown by the red bar graph, which means that microorganisms exhibiting a negative coefficient are negatively correlated with the substance.</i> .....	34
Figure 16: <i>The data analysis architecture</i> .....	44
Figure 17: <i>The OTU table of alpha diversity</i> .....	45
Figure 18: <i>The OTU table of NRI and NTI in environmental filtering processing</i> .....	46
Figure 19: <i>The Partial OTU table of significant differences between CO and CH soil samples</i> .....	47

## Table of Contents

<i>Abstract</i> -----	<i>i</i>
<i>Acknowledgments</i> -----	<i>ii</i>
<i>List of Abbreviations</i> -----	<i>iii</i>
<i>List of Figures</i> -----	<i>iv</i>
<i>Table of Contents</i> -----	<i>v</i>
<b>1 Introduction</b> -----	<b>1</b>
1.1 Environmental Background -----	1
1.3 The next generation of sequencing methodology for soil degradation analysis 4	
1.4 The new methodology for data analysis -----	5
1.5 The project objectives and goals -----	5
<b>2 Methodology</b> -----	<b>6</b>
2.1 Analytical method-----	6
2.2 Data description-----	8
2.3 Taxonomic profiling and OTU construction from 16S rRNA sequence-----	8
2.4 The analysis of statistics-----	9
2.4.1 Alpha Diversity -----	9
2.4.2 NRI and NTI-----	10
2.4.3 Beta Diversity -----	11
2.4.4 The top 25 most abundant taxa-----	11
2.4.5 Identifying deterministic and stochastic in COV and CH soil community environments: The calculation of Quantitative Process Estimate (QPE) and incidence- based (Raup-Crick) beta diversity via null modelling-----	12
2.4.6 Identifying deterministic and stochastic in COV and CH soil community environments: Normalized stochasticity ratio (NST) via null modelling -----	13
2.4.7 The differential analysis: Core microbiome-----	14
2.4.8 The differential analysis: MA plot -----	15
2.4.9 Regression modelling: Subset regression-----	17
2.4.10 Regression modelling: CODA GLMNET -----	18
<b>3 Results</b> -----	<b>20</b>
3.1 Comparative taxonomic identification-----	20
3.2 The Analysis of Diversity -----	20
3.2.1 Alpha Diversity -----	20

3.2.2	NRI and NTI-----	20
3.2.3	Beta Diversity -----	21
3.2.4	The most abundant top 25 taxa identified per community -----	21
3.3	Identifying deterministic and stochastic in COV and CH soil community environments -----	23
3.3.1	Null modelling: Calculating Quantitative Process Estimate and incidence-based (Raup-Crick) beta-diversity-----	23
3.3.2	Null modelling: Normalized stochasticity ratio (NST) -----	23
3.4	The Differential analysis-----	25
3.4.1	Core microbiome heatmap-----	25
3.4.2	MA plot -----	27
3.5	Regression Modelling: Determine key members of representative microbial communities of COV and CH soils -----	30
3.5.1	Subset regression -----	30
3.5.2	CODA GLMNET-----	33
4	<i>Discussion</i> -----	35
4.1	The Biological Informatics Approach to Biodegradation Analysis of Contaminants -----	35
4.2	Overlaps and differences between COV and CH soils -----	36
4.3	The critical microorganism for biodegradation-----	37
5	<i>Conclusions and Future work</i> -----	38
6	<i>References</i> -----	39
7	<i>Appendix</i> -----	44
7.1	The data analysis architecture-----	44
7.2	The Result of the OTU table in data analysis -----	45

# CHAPTER 1

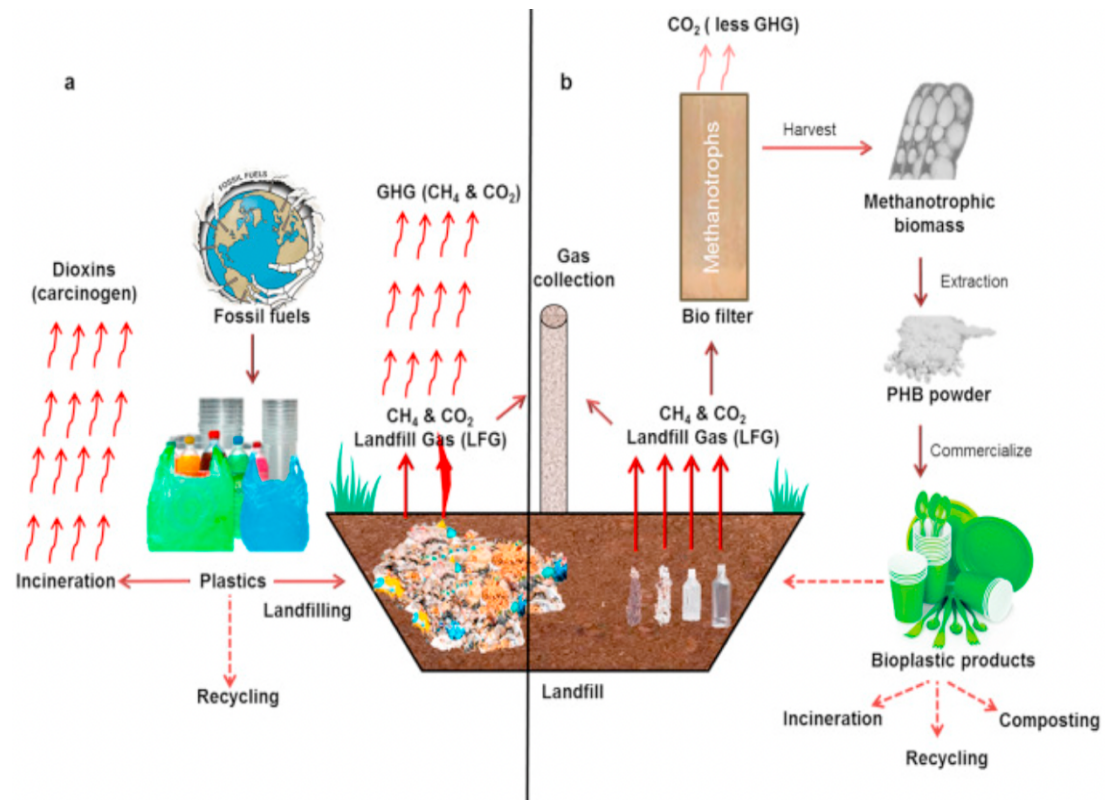
## 1 Introduction

### 1.1 Environmental Background

Environmental pollution is a remnant of industrialization which has posed a significant threat to individuals health and the ecological environment. Governments and citizens have long been aware of the problems of water and marine pollution. Most developed countries have regulations in place to govern the health and safety of human water, so that environmental chemicals in drinking water have never been harmful to health (Schlatter, 1994) [1]. Soil and water are inextricably linked in a chain, where soil provides the environment for various species of plants to grow and where rainfall enables plants to flourish. The plants supply the nutrients that animals need, and human beings consume vegetables, fruits, and meat to maintain a balanced diet. A lack of attention to the soil environment over the years has resulted in the accumulation of toxic substances such as PCBs, inorganic anhydrides and polyaromatic hydrocarbons (PAHs) in the soil. As Schlatter (1994) [1] shows that, industrial areas have inorganic anhydride substances ( $SO_2$ ,  $SO_3$  or  $NO_x$ ) which affect the lung function of susceptible populations. Stading et al. (2021) [2] indicated that polyaromatic hydrocarbons (PAHs) increase the incidence of lung cancer and mortality in humans. These results demonstrate that soil pollution is posing a major threat to human health. The government will need to take this issue seriously, so that it may tackle the contamination of the soil.

Landfill disposal is the preferred method of treating contaminated soil, but it does not address the root cause of soil contamination. From an environmental sustainability point of view, landfills increase soil contamination. For instance, the toxicity of leachate from plastic waste deposited in landfills may adversely affect the soil microflora and contribute to soil infertility or pollution of groundwater and water supplies in figure 1 (Teuten et al., 2007; Chidambarampadmavathy et al., 2017) [3] [4]. Contaminated soils are toxic and biologically inert, which results in susceptibility to biodegradation and waste persistence (Chidambarampadmavathy et al., 2017) [4]. Landfill disposal does not solve the underlying problem of soil contamination, rather it can only increase the level of contamination.

The use of microorganisms to biodegrade contaminated soil in order to solve the soil contamination problem. The interactions between microbial communities and the mechanisms that participate in the biodegradation of contaminated soils have become very important to understand, especially considering polyaromatic hydrocarbons (PAHs). As Gauchotte-Lindsay et al. (2019) [5] show, Polycyclic aromatic hydrocarbons (PAHs), are regarded as remaining organic contaminants of a lipophilic nature which are discovered within crude oil. It is generated in the combustion process between coal and organic substances. PAHs have been detected as carcinogenic or mutagenic substances, for example, naphthalene, phenanthrene or benzene (Gauchotte-Lindsay et al., 2019) [5]. As they are currently on the priority list, the study of the biodegradation is particularly critical in the case of PAHs by microorganisms.



**Figure 1:** *The lifetime landfill degradation of plastics and bioplastics* (Chidambarampadmavathy et al., 2017) [4].

## **1.2 The environmentally detrimental polycyclic aromatic hydrocarbons (PAHs)**

Polycyclic aromatic hydrocarbons (PAHs) represent a cluster of biological substances consisting of two or more dense aromatic rings (benzene) that are produced via the improper combustion of organic substances. It is inherently a chemically inert and hydrophobic polymer (Phillips, 1999) [6]. They are often associated with particulate matter that is very detrimental to human health, such as benzo(a)pyrene and pyrene, and are typically used as indicators of carcinogenic and mutagenic PAHs in total exposure (Ohura et al., 2004) [7]. Phillips (1999) [6] indicated that PAHs are metabolically activated by cells in mammals after ingestion of food with PAHs. The activated PAHs generate epoxides, which combine with intracellular macromolecules such as DNA to produce covalent bonds. This chemical transformation can cause errors in DNA replication, resulting in carcinogenic mutations. This phenomenon is not only a threat to mammals but can lead to the absorption and bioaccumulation of toxicity chemicals within the food chain. Under certain circumstances, it can be a significant health problem and or genetic deficiency in animals (Samanta et al., 2002) [8].

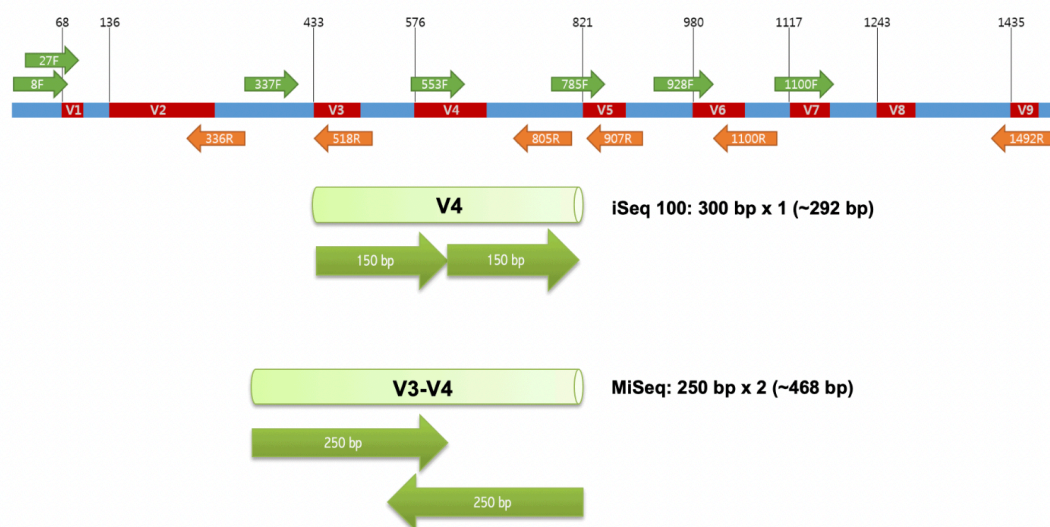
The PAHs are a typical environmental pollutant. Many PAHs have penetrated into the soil as a result of years of improper handling by humans. These toxic substances block the pores of the soil and cause a reduction in soil water infiltration (Singh and Haritash, 2019) [9]. Magi et al (2002) [10] reported that PAHs preferentially attach to smaller silt and clay soils than to soils with larger particle sizes. Silts and clays have a greater surface area due to their smaller pores, which allows PAHs to attach more readily to them. Due to the better adhesion, the PAHs content of the soil will be higher and the toxicity will be increased. Most of these contaminated soils require biodegradation via microorganisms to repair and decontaminate the soil. However, PAHs are also harmful to microorganisms. It decreases the activity of microorganisms which leads to a reduction in the effectiveness of soil decontamination (Hreniuc et al., 2015) [11]. The cultivation of crops on these contaminated soils will enhance the crop with PAHs. Ingestion of these toxic crops by humans or animals may cause an increase in the accumulation of carcinogenic substances in the body, which could lead to a higher incidence of cancer. The PAHs have a significant negative impact on microbial diversity and populations.



### 1.3 The next generation of sequencing methodology for soil degradation analysis

In past studies, microbial community studies have mostly used more traditional artificial culture methods to extract samples from microbial communities (Torsvik et al., 1998) [12]. These samples from microorganisms are well below 1% of the entire microbial community, so that their genes need to be extracted from a large number of microorganisms (Samarajeewa et al., 2015) [13]. New sequencing methods, such as 16S rRNA amplification, Sanger sequencing and DNA microarray, enable high-resolution characterization of microbial communities and provide greater transparency into the relationship between complex microbial communities and ecological environments.

Microorganisms are microscopic organisms, mainly including eukaryotes and prokaryotes. Prokaryotes have the same gene, but different gene sequences within the gene. Its 16S rRNA gene similarity is as high as 97-99% and the 16S rRNA gene is replicable (Liang et al., 2021) [14]. As Simon and Daniel (2011) [15] show, 16S rRNA is being used extensively to research the diversity in microbial communities of complex environments. Yarza et al. (2014) [16] demonstrated that 16S rRNA is located on the ribosomal Small Subunit (SSU) of prokaryotic cells. In the Figure 2, its sequence contains 9 hypervariable regions (V1~V9) and 10 conserved regions. Variable regions can be used for the phylogenetic classification of genera or species of different microbial communities. The microbial communities in the soil were mainly located in the V3 and V4 regions. The Miseq machine was used to sequence these gene fragments from forward and reverse, and the read length just spanned the V3 and V4 region (Castelino et al., 2017) [17]. This project uses nucleic acid sequencing to characterise microbial communities in contaminated soils at high resolution to understand the biodegradation of microbial communities in PAH contaminated soils.



**Figure 2:** The 16S rRNA and 16S rRNA Gene (Source: EZBioCloud Help center, 2019) [18].



## **1.4 The new methodology for data analysis**

The data from amplicon sequences that are multi-dimensional with high resolution, so there are many OTU analysis methods for microbiome studies. These novel OTU analysis techniques can be used to infer key members of COV and CH soil microbial communities in OTU tables. As Gauchotte-Lindsay et al. (2019) [5] indicated that Alpha diversity that is used to give a summary of the community diversity for each sample. For example, sparse richness would be applied to estimate population size of OTU species; Shannon entropy measures the balance of the communities in each OTU. Microbial analyses also compare similarities between two samples via Beta diversity (Gauchotte-Lindsay et al., 2019) [5]. According to Vass et al. (2020) [19], a null model approach to OTU analysis can help to observe what factors in the environment affect community development, such as quantitative process estimate (QPE) and normalized stochasticity ratio (NST). In addition, correlations between microorganisms and COV and CH soil samples were observed through regression models in order to identify key members of COV and CH soil biodegradation processes. These new data analysis methods help to improve the efficiency of microbial community analysis.

## **1.5 The project objectives and goals**

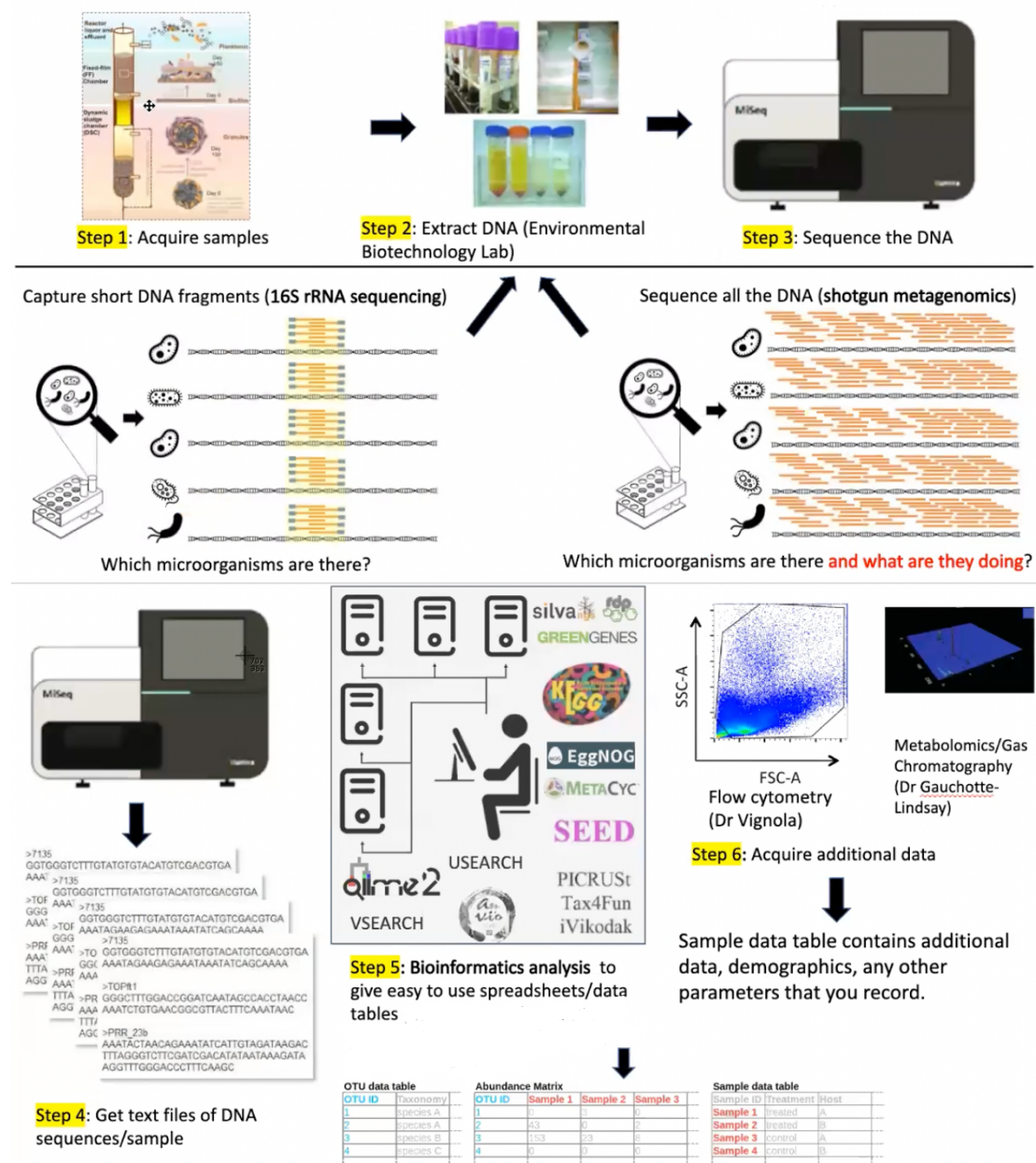
This research has the overall objective which is to solve the soil contamination problem by using microorganisms to biodegrade contaminated soils. The project will provide a cross-sectional comparison of microbial communities from soils at two contaminated sites, COV (UK) and CH (Switzerland). Both sites have 16S rRNA samples and are contaminated with PAHs. The datasets with 16S rRNA samples were analyzed by R Studio software. It was guided by recent advances in microbial community assembly mechanisms to develop an analytical strategy that elucidates the role of the environment and reveals key members of the microbial communities representative of these sites.

## **CHAPTER 2**

### **2 Methodology**

#### **2.1 Analytical method**

A microorganism is a microscopic organism. It requires data collection via DNA extraction which is then analyzed to visualize the relationships between microbial communities. In this project, a method based on the microbial data flow in Figure 3 is used for the analysis of contaminated COV and CH soils. Firstly, contaminated COV and CH soil samples were collected from the United Kingdom and Switzerland, respectively. These collected samples were analyzed by genomic DNA to capture short DNA fragments which only use primers to multiply or to amplify the 16S V3 and V4 region does not amplify the whole regions. Once the DNA has been amplified, it will be sent to the Miseq machine to sequence the DNA. The available sequences classified from the Miseq machine are obtained as text files of DNA sequences or sample. These text files are run through different bioinformatics software to generate OTU tables from them. The OTU table contains additional data, demographics, any other parameters that be recorded. The data from the OTU tables were analyzed by R studio to identify key members of the COV and CH soil microbial communities.



**Figure 3:** The data flow in microbial data analysis from Dr. Ijaz

## 2.2 Data description

This project uses data collected from two different pre-production gas plants provided by Caroline Gauchotte-Lindsay and her collaborators. The samples had to be preserved in plastic tubes at temperatures 4 °C to prevent them from being disturbed by other substances. There were 27 soil samples contaminated with PAHs that were collected, 2/3 from the Coventry area (COV) in the UK and 1/3 from a town in Switzerland (CH). As Gauchotte-Lindsay et al. (2019) [5] show, COV soil samples are typically dark in colour, dense and cohesive in substance. Each sample was filtered once via the 10 mm sieve. The CH samples were coloured brown in comparison to the COV soil and were relatively dry separated by 1.7, 2.36 and 10 mm sieves respectively. The percentage moisture content was monitored for each separate sample by keeping the subsamples in an incubator situated at 105 °C during 24 hours and the loss on ignition (LOI) by maintaining the subsamples at 550 °C for two hours. (Gauchotte-Lindsay et al., 2019) [5].

From the soil samples collected above, two extractions of genomic DNA were performed for COV (UK) and CH (Switzerland) soils. One was used for quantitative PCR (qPCR) while the other to sequence 16S rRNA (V3 and V4 regions) to obtain analyzable data files.

## 2.3 Taxonomic profiling and OTU construction from 16S rRNA sequence

This project used VSEARCH v2.3.4 to process the 16S rRNA (V3 and V4 regions) dataset and generate abundance tables via the construction of operational taxonomic units (OTUs). Each OTU was considered to be an OTU species, i.e., if 10 OTUs were detected in the 16S rRNA dataset, then there were 10 microorganisms there. Barcodes are added to each sample to enable real-time tracking of which sample the reads came from during processing. In order to classify the samples efficiently, the reads were deduplicated which means making everything unique, and also counting how many unique terms are there. The reads were sorted in decreasing order of richness and the monoclinic state was discarded. Based on the theory that two sequences (16S rRNA regions) are >97% similar, they belong to the same species. Matching the original barcode reads with OTUs (n=26 samples of a total of 2,234 OTUs) that were 97% similar to the summary read statistics of the samples to produce the OTU tables.

For the taxonomy of representative OTUs, Qlime workflows were performed in the `assign_taxonomy.py` script16. The classified OTUs were compared to multiple sequences using MAFFT v 7.317 and a phylogenetic tree in NEWICK format was generated in FastTree v2.1.718. Using the `make_otu_table.py` script, abundance tables were combined with classification information to generate the OTU biom files which include classification of each OTU and abundance tables.

## 2.4 The analysis of statistics

This study utilized NEWICK documents and BIOM that contain the taxonomy of each OTU. They were calculated as described in method 2.3 and referenced relevant metadata, including GCxGC.csv and meta-tables, for the statistical analysis of 16S rRNA in RStudio software. The code for all the diversity analysis figures which follow is available in the Appendix.

### 2.4.1 Alpha Diversity

Alpha diversity analysis is an important component of biodiversity in ecology and is often used to summarize a single sample. Willis (2019) [20] reports that alpha diversity metrics are primarily correlated with abundance and evenness, the latter implying richness distribution in the species community. It is a set of functional equations that calculate a value to determine whether a sample is more or less diverse. The single nature of the equation makes it impossible to fully characterize the diversity of a community, so that five indicators have been developed to increase its accuracy: Species richness index, Shannon entropy, Pielou's evenness index, fisher alpha and Simpson diversity. The species richness index has been developed to approximate the population of a taxa. The researcher is usually unable to collect all the individuals in a community, so the number of species in the parent can only be estimated from the sample and is the most simple measure of diversity (Whittaker, 1972) [21]. The Pielou evenness index is an indicator used to reflect the evenness of distribution of individual species in a community (Pielous, 1966) [22].

The diversity index consists of abundance and evenness. For example, Fisher alpha was the one used to contrast the diversity index of the community with the different number of individuals. According to Shannon (1948) [23], Shannon entropy diversity based on the concept of information entropy, biodiversity is represented by the uncertainty in the sampling process regarding the species to which the sampled individuals belong. In considering the richness and evenness of a community combined, an indicator with a higher value is more diverse (Shannon, 1948) [23]. Simpson (1949) [24] proposed that Simpson diversity is a method used in ecology to quantify the diversity of biological communities in a region. As the Simpson Index value increases, the more diverse the community is. The Simpson's Diversity Indicator is equal to 1 minus the proportion of a random sample of two individuals belonging to different species (Simpson, 1949) [24]. The value is between zero and one and is a valid approach to assessment the diversity index of the microbial community for a sample.

In this research, using the R package Vegan to analyze data from COV and CH soils, and to generate an analytical figure of alpha diversity to compare the differences between COV and CH soils.

### 2.4.2 NRI and NTI

Environmental factors have been an influential issue in understanding the processes that control the composition of ecological communities. Many scientists have argued that community assemblages have been impacted through both comparatively deterministic factors as well as more stochastic ones. (Dumbrell et al., 2010; Stegen et al., 2012) [25] [26]. Chase and Myers (2011) [27] demonstrated that deterministic factors consist of abiotic environmentally imposed selection, which is called environmental filtering, as well as the interaction of antagonistic and synergistic species. Stochastic factors, by contrast, are more likely to be unpredictable and to be dispersed by probability.

The COV and CH soil samples were environmentally filtered by the nearest taxon index (NTI) and net relatedness index (NRI) methods to observe variation in phylogeny for each sample group. Cooper et al. (2008) [28] demonstrated that the NRI is derived from the mean phylogenetic diversity (MPD), which is measured as the total of phylogenetic distances in pairs among all the taxonomic pairs in the communities. If the MPD value is low, then all OTUs are very close to each other. Whereas when the value becomes lower, it will be more clustered. The obtained MPD value was calculated by bringing it into the NRI formula, where a positive NRI represents the clustering community, which means the chances are more nearly correlated than expected, with a negative NRI represents the evenly distribution of the community; NTI is an analysis of the tendency of the closest associated species in a community to co-occur based on the mean nearest taxon distance (MNTD). (Cooper et al., 2008) [28]. In contrast, NTI is more focus on over dispersion or over clustering at the apex of the phylogeny than NRI. The positive relationship for NTI indicates that the species co-occurred with a larger number of strongly related ones than predicted; the negative value shows that highly interrelated species do not co-occur.

For data analysis, `mntd()` and `mpd()` Instructions from the `picante` package were applied to the NTI and NRI respectively. The quantification is accomplished at `ses.mntd()` and `ses.mpd()` instructions using `null.model`, which is a quantity of normal variance that separates the observations in relation to the average value in the null distribution. Output the last calculation of the `ses.mntd()` and `ses.mpd()` functions at the end. It is deterministic when the NRI is greater than zero or the NTI is greater than two. On the contrary, when NRI is less than zero or NTI is less than two, it is classified as random or competitively excluded.



### 2.4.3 Beta Diversity

Beta diversity analysis was used to compare two samples to see if they are similar in composition. In order to precisely identify the material differences between COV soils and CH soils, the Beta diversity between the two samples is calculated by the distance metric function. It usually returns a value between 0 (absolutely similar) and 1 (absolutely dissimilar). It is mainly divided into several measurement methods. Bray & Curtis (1957) [29] showed a method where the Bray-Curtis distance uses OTU table data to calculate whether OTU abundances are different or similar. The value is somewhere from 0 to 1, in which 0 means exactly similar and 1 is totally different. Bray-Curtis Contribution provides a % contribution based on a subset of OTUs/features, making the data more accurate. The second method is the UniFrac distance, which involves comparing a sample of environments with a phylogenetic tree rather than an OTU table (Lozupone et al., 2011) [30]. As Lozupone et al. (2011) [30] show, UniFrac is combined with basic multi-variate statistical skills, which include Principal Coordinate Analysis (PCoA), that are standard to determine the factors that explain microbial community variability. The UniFrac distance was calculated to be the proportion of the length in branches shared phylogenetically between the two samples, which should again give a value in the range of 0 and 1, with 0 meaning similarity and 1 being dissimilarity. Lozupone and Knight (2015) [31] reported that shared branches are OTU branches that are generated jointly in CH soil and COV soil samples, or non-shared branches if one of the samples is missing. UniFrac distances are divided into weighted and unweighted. Unweighted UniFrac (whether OTUs are present or not) is a measure of the developmental relevance of a system and is computed as the ratio of branch lengths shared by two samples which is the sum of all unshared root sizes divided by the sum total for the lengths of all shared roots. (Lozupone and Knight, 2015) [31]. If the abundance of OTUs is added, it is a weighted UniFrac. In the R software analysis, the Phyloseq package has a function to calculate unweighted and weighted Unifrac distance measures.

The Vegan package has a function called `vegdist()` that calculates the Bray-Curtis distance. Pairs of  $\beta$ -diversity values ( $N*N-1/2$ ) are taken from and these are assigned to ORDINATION plots (NMDS non-metric distance scaling, PCoA principle coordinate analysis), so that obtain visualization of the beta diversity values.

### 2.4.4 The top 25 most abundant taxa

Taxonomic Analysis was used to classify species and to analyze which species are present in each community. The top 25 most abundant groups were selected to compare the richness of the taxa in COV and CH soils.

#### **2.4.5 Identifying deterministic and stochastic in COV and CH soil community environments: The calculation of Quantitative Process Estimate (QPE) and incidence-based (Raup-Crick) beta diversity via null modelling**

As Vass et al. (2020) [19] proposed, the community assembly processes is non-static that may vary in relative significance from one assembly process to another under different circumstances and between distinct microbial populations. The community assembly process gives rise to minor and substantial temporal variability under environmental changes. It can trigger distinct responses and choices by bacterial and micro eukaryotic metacommunities. In this experiment, the dynamics of COV and CH soils community assembly were assessed by methods of a null model. There are two methods to analyze the dynamic community change of COV and CH. First, the method will calculate a quantitative process estimate (QPE), which surveys the relative significance of each assembled process, i.e., whether a microbial sample has underlying characteristics of a single environment or multiple environments. The quantitative processes are divided into five variables: Dispersal limitation, Homogenizing dispersal, Ecological drift (Undominated), Homogenizing selection and Variable selection. Stegen et al. (2015) [32] indicated that selection refers to differences in performance for taxa between homogeneous and variable selection; dispersal limitation is a limited biological exchange between communities caused by the movement of organisms, whereas homogeneous dispersal is a high level of biological exchange; ecological drift (undominated) is generated by population size at random changes. These factors are the main variables in ecological community assembly, and these five interval variables were used to compare COV and CH soil samples with regard to their comparative importance in the community.

The QPE, unlike the previous NRI/NTI, was calculated based on using Beta MNTD and Beta NNTD. The average of the two samples is taken to see whether the microbial sample has potential characteristics of a single environment or multiple environments and whether the environment is affecting the microbial community structure. When BNTI>2, it is classified as variable selection. Conversely, if it is less than 2, it is homogeneous selection.

The second is incidence-based  $\beta$  diversity, which means that deterministic and stochastically findings in the sample are distinguished by the presence or non-presence of OTU tables (ignoring phylogeny) (Vass et al., 2020) [19]. The project uses picante, ape and ecodistance packages to construct incidence-based  $\beta$  RC values. When the incidence-based  $\beta$  RC value does not deviate significantly from 0, it means that the community is classified as a stochastically assembled; while if the  $\beta$  RC value deviates from 0 and gradually approaches 1 or -1, it indicates a deterministic clustering of communities, where similarity between communities is stronger than what would be predicted at random.. (Vass et al., 2020) [19].



#### **2.4.6 Identifying deterministic and stochastic in COV and CH soil community environments: Normalized stochasticity ratio (NST) via null modelling**

It is a core issue in ecological science to identify the community assembly mechanisms that drive biodiversity patterns (Ning et al., 2019) [33]. The calculation of QPE via a null-model framework gives results regarding the relevance that deterministic and stochastic processes in microbial communities have in community assembly. As Ning et al. (2019) [33] show, the normalized stochasticity ratio (NST) is a tool for quantifying ecological stochasticity based on the null model framework. It makes it possible to observe more clearly the relative relationship between ecological stochasticity in different situations. The NST is based on 50%, when the calculated value is less than 50%, the community is more deterministic, and conversely, if the calculated value is over 50%, the randomness of the community is greater.

According to Ning et al. (2019) [33], the precision of the NST calculation is higher than that of the previous method of calculating ecological stochasticity. The null model and community similarity metrics therefore have an influence on ecological stochasticity.

## 2.4.7 The differential analysis: Core microbiome

The core microbiome is a differential analysis of biological communities. The microbiome package in R studio is used to carry out core microbiome analysis to find out which microorganisms are present in which communities. The classification results are used to calculate what percentage of the sample is microbial and to determine whether the community is a high or low abundance core microbiome. Setting the minimum prevalence equal to 85%, which means 85% of the species were present in the sample at an abundance of 1. In the visualization of the core microbiome, the horizontal axis represents how many samples the species are present in. The vertical axis indicates how abundant these species are in different samples. When the species with an abundance is two that is present in 80% of the samples with an abundance of. If the abundance is up to nineteen, it is present approximately in 30% of the sample. According to the data on the vertical and horizontal axes, it is evident that the red area on the bottom right of Figure 4 indicates high abundance, while the blue area on the top left means low abundance. As can be seen from Figure 4, this species is present in almost 100% of the samples with a maximum abundance threshold of 1098.

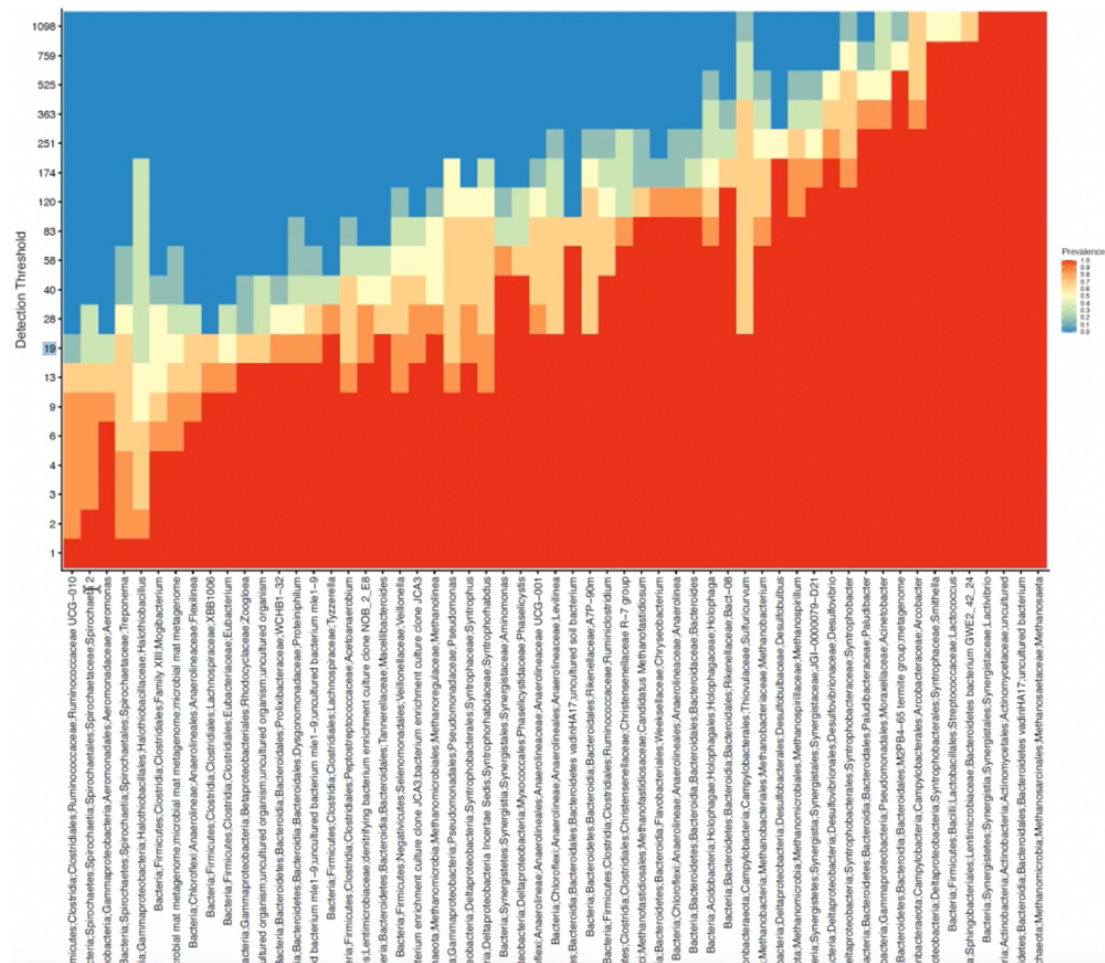


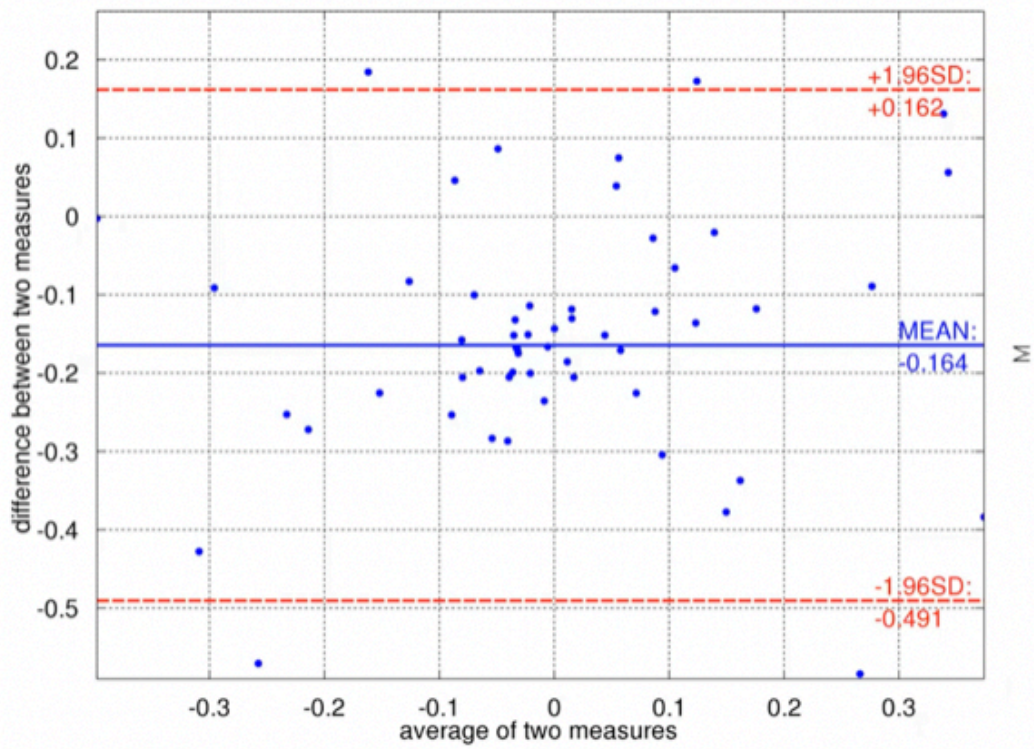
Figure 4: The visualization paradigm for the core microbiome

### 2.4.8 The differential analysis: MA plot

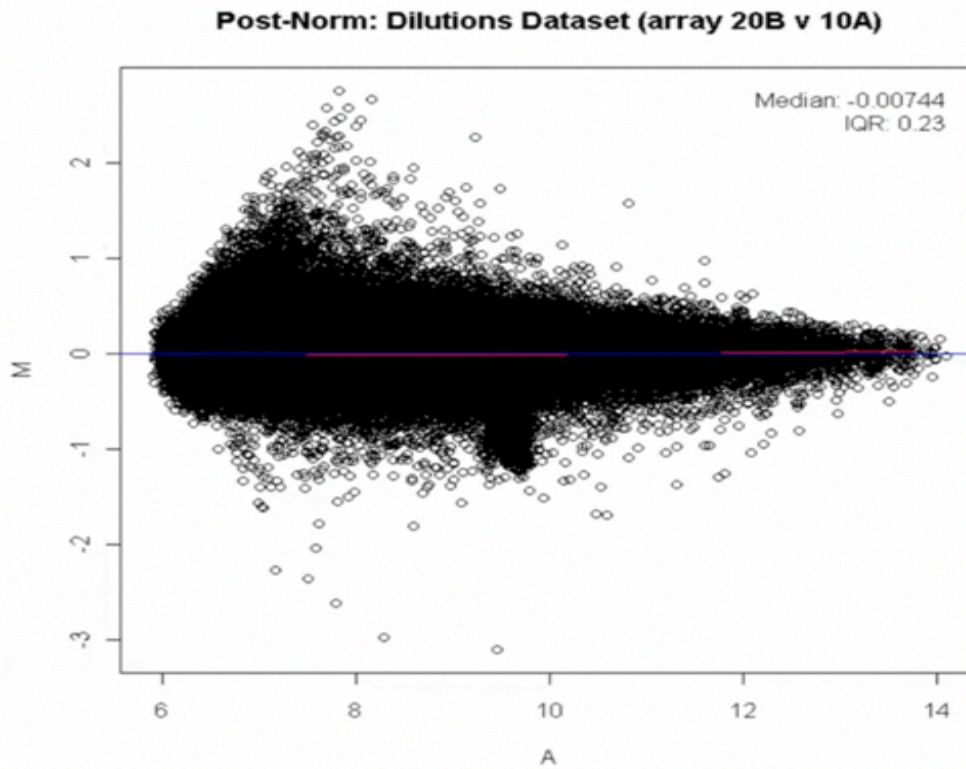
The DESeq2 is a method for computing data differential analysis. Love et al. (2014) [34] demonstrated that it is the use of dispersion and multiplicative variation of shrinkage estimates to enhance the stability and interpretation of estimates. This method contributes significantly to finding genes that differ among sample groups in RNA-sequencing (RNA-seq) data. The DESeq2 package is installed in the R studio environment, so that it is possible to take any two conditions in a type and find out what the "differentially expressed" feature/microbe is. Based on the Bland-Altman plot in Figure 5a, the average of the two measures is taken and plotted against the X-axis. In the Y-axis, the differences are plotted. For example, the composition of microorganism a and microorganism b are compared. By taking the average of two microbes and placing it along the X-axis, then the variation of these microorganisms on the Y-axis. If they are significantly different, e.g., more than 1.96 (standard deviation), this means that the microorganisms or the two main values are statistically significant or statistically different in Figure 5b.

For the analysis of microbiome data, the Bland-Altman plot was used as the basis for its logarithmic transformation. After logarithmic transformation of these parameters, it is converted into a ma plot in Figure 5b. The mean values for each microorganism were significantly different if they fell over +2 logfold or under -2 logfold. If a microbe is too logfold different, then the microbe is significant between condition a and condition b. This plot is also known as a Volcano plot.

(a) Bland-Altman Plot



(b) MA Plot



**Figure 5:** a) The example of Bland-Altman plot (Source: Wikipedia) [35]. b) The example of MA plot (Source: Wikipedia) [36].

### 2.4.9 Regression modelling: Subset regression

Microbial regression analysis is the fitting of a microbiome dataset to a linear equation, the purpose of which is to relate the microbiome to external parameters to better understand the behavioural patterns of the microbiome. In order to try to understand whether any environmental parameters play a role in the microbiome, these variables from the metadata tables were categorized. The categories are converted to presence/absence tables through a domification procedure, and typically the closest power of 2 is used to encode them as binary numbers. For instance, if the metadata table have six categories, then the nearest power of two is eight. The regression model fits a linear equation to the dataset with a dependent variable and several independent variables, such as PH value or temperature, this is done by assigning weights to each variable, known as "beta coefficients" ( $\beta_0, \beta_1, \dots, \beta_p$ ).

$$\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i \quad (1)$$

According to the formula (1) of linear regression, when beta coefficients are positive, it implies that a rise in this parameter would also cause a rise in this one; if they are negative, then this means that there will be a decline in this parameter as a result of increased variance. The beta coefficient acts as an essential part of the linear regression equation that combines the two datasets with each other.

To avoid the constant of the meta table becoming another variable, since the linear regression equation has a linear dependence on the independent variable X. All these variables are regressed against each other by the VIF (variance inflation factor) method, which removes the linear dependent variable.

$$VIF_i = \frac{1}{1 - R_i^2} \quad (2)$$

There are hundreds of columns of data in the COV and CH samples meta table, but not all the recorded data will be related to each other. Therefore, it is necessary to find the best variable available via subset regression which will have a minimal subset related to y. Subset regression can fit multiple equations, so there are at most  $2^n - 1$  equations. Another strategy, Lasso strategy (L1 constraint), is to fit only one equation but force some beta coefficients to zero. The overall goal of the strategy is to find the beta coefficient, which may be positive or negative, and is directional. The accuracy of the beta coefficients is ensured by two types of cross-validation, Leave-one-out cross-validation (LOOCV) and MFOLD cross-validation. In general, subset regression models do not fit all data to a given model. It fits the model multiple times and removes duplicate data. The remaining data were averaged (root mean squared error) to further improve our analysis.

In this project, pollution, and PAHs substances in the COV and CH sample meta-tables were used as variables. Using a subset regression approach to observe their relationship in the soils with Shannon diversity.



## 2.4.10 Regression modelling: CODA GLMNET

In the previous equation for subset regression,  $y_i$  is a property of the microbiome and  $x_i$  is an extrinsic parameter. There have some problems are that the equation cannot perform subset regression because there is often a large amount of microbial data. The second problem is that a simple regression model cannot be applied. Especially count data, since the data does not preserve Euclidean space, transformation is required. The third problem is that the subset regression is only run once due to the Lasso strategy. However, some substances that have a positive or negative relationship with a parameter of naphthalene may have recorded a performance parameter, for example, what is the smallest subset of species that is related to the species that degrades it. It needs to be deleted, turning its beta coefficient to zero. Therefore, the two variable attributes are flipped,  $y_i$  is an extrinsic parameter, which can be binary and continuous, and  $x_i$  is all microorganisms.

In mathematics, there is a function called logarithmic contrast function. It can find two subsets in balanced form. The two subsets are in an absolute equilibrium state, and the positive and negative values can cancel each other out. The sums of the two subsets are 1 and -1, respectively, and the two add up to zero to form a complete complement, which is called close composition.  $X$  is a scalar vector, where some of all those values equal to one. According to the logarithmic contrast function, the linear combination of the logarithms of the components of this formula (3) is subject to the condition that the coefficients are summed to zero. Therefore, the function needs to find two subsets of the top subset and the bottom subset, a positive subset, and a negative subset, as shown in Figure 15. These two subsets eventually become locking and trust functions.

$$f(x) = \sum_{i=1}^k a_i \log(x_i), \text{ with } \sum_{i=1}^k a_i = 0 \quad (3)$$

As Lu et al. (2019) proposed, change the  $X$  variable in the previous linear regression equation (1) to  $\log x_{1i}$  to get the Coda-lasso equation (4). The reason is that it ends up being an invariant. Therefore, this project uses the Coda-lasso regression model instead of the general linear regression model.

$$y_i = \beta_0 + \beta_1(\log x_{1i}) + \dots + \beta_k(\log x_{ki}) + \varepsilon_i \quad (4)$$

After changing the beta coefficient to zero, there is no relationship between them. If there is no relationship between the beta coefficients, it ends up leaving only two subsets, those with positive beta coefficients and those with negative beta coefficients, perfectly balanced with each other. No matter which variable is entered, the result will always be positively correlated microbes and negatively correlated microbes. It won't just show the positives because it will find a perfect balance there, with positive microbes in balance with negative microbes. It can be seen from equation (5) that the zero-sum constraint makes it a weighted balance between positive and negative coefficients.

$$\beta_0 + \sum_{i \in I+} \beta_i \log X_i - \sum_{i \in I-} \alpha_i \log X_i = \beta_0 + s_\beta \log \left( \frac{g_+(x)}{g_-(x)} \right) \quad (5)$$

This project uses CODA GLMNET regression analysis to observe the relationship between substances in COV soil and CH soil, soil PH and *Pseudomonas*. The key members of the microbial communities that represent these sites are ultimately identified.

## CHAPTER 3

### 3 Results

#### 3.1 Comparative taxonomic identification

The COV and CH soil samples used Qiime2 and DADA2 workflows to generate abundance tables containing 2234 OTUs from 26 samples. The main raw data were as follows, 1st quartile: 45,116, median: 49,398, mean: 52,557, 3rd quartile: 65,156, maximum: 95,350. According to the generated NEWICK and OTU biom files, the data were analyzed biometrically as follows.

#### 3.2 The Analysis of Diversity

##### 3.2.1 Alpha Diversity

Alpha diversity measurements were measured to demonstrate how the diversity of constituents in the COV and CH soil communities varied. As shown in Figure 6a, the species richness index of CH samples mostly falls between 250 and 300, and the estimated number of taxa is five times higher than that of COV samples; In terms of Shannon entropy diversity, the higher the value, the more species diversity there is. CH samples, with values mostly close to 4, are more diverse within the sample than COV samples; The Simpson's index was found that the Simpson Index value grows, the more diverse the community is. It is clearly that the CH samples index is also slightly higher than the COV ones. It is clear from the five measurements of alpha diversity that the CH sample is much richer in species than the COV sample. The higher the value of diversity within the sample, the more diverse the community. The higher the value of CH compared to COV, the more diverse the CH community.

##### 3.2.2 NRI and NTI

Investigating environmental influences on microbial community assembly using NRI and NTI to explore local and global phylogenetic clustering. For COV and CH samples, when the NRI is greater than 0, this indicates that the species of the environment are co-occurring to those that are more nearly relevant. Simultaneously, the NTI is greater than 2, indicating that the local phylogeny of the phylogenetic tree driven by environmental filtering has strong environmental pressure in the whole environmental system. The OTU between each community is closer than would be expected by chance. In contrast, if the NRI is less than zero and the NTI is below 2, this means that the community is evenly distributed, and closely related species do not occur together.

As illustrated in Figure 6a, the NTI values for the COV and CH samples were in the range of 1 to 2, with the COV samples being more dispersed than the CH samples. The NRI of both samples was greater than zero, and the CH sample was slightly higher than the COV sample. These suggest that the communities in the CH soil samples may be more intensely impacted by the environment. The COV ones is dominated under the principle of competitive exclusion, in which species that perform



better than each other in particular ecological position dominated, resulting in a more dispersed phylogenetic tree.

### 3.2.3 Beta Diversity

The Beta diversity was computed for both samples using a distance metric function to identify differences between COV and CH soil communities. The value is usually between 0 and 1, with a distance metric of 0 indicating absolute species similarity between communities. The distance metric is 1, which means that the species are very different from each other. In terms of Bray-Curtis distance measures, it shows a separation between COV and CH samples in the upper left corner of Figure 5b. The distribution of COV samples is more dispersed than that of CH soils, representing greater variability.

In terms of environmental samples and phylogenetic trees, the UniFrac distance measure shows in the upper right corner of Figure 6b that the sequences of the COV and CH communities are mostly concentrated in the upper and lower values of 0 and mostly overlap. The weighted UniFrac distance measures accounting for phylogenetic and abundance factors indicated a high similarity between COV and CH communities below in Figure 6b.

When compared to the Taxonomy Difference plot in Fig. 5b, the PAH concentrations in the COV soil samples were on average twice as high as those obtained from the CH ones. The COV soil samples were also more variable. It can be concluded that the average PAH concentration is the main factor causing the higher variability of COV soil samples compared to CH ones. The COV samples were highly contaminated by PAHs to the extent that their beta diversity is spatially dispersed.

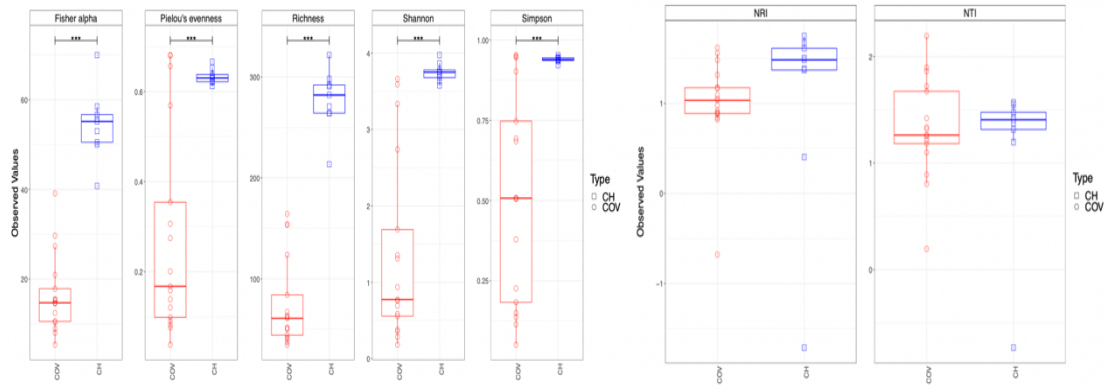
### 3.2.4 The most abundant top 25 taxa identified per community

The heatmap in Figure 6b shows the top 25% most abundant species in each community. From the classification key in the lower left corner of Fig. 5b, it can be known that the composition ratio of *Pseudomonas*, *Pseudoxanthomonas* and *Rhodospirillales* in the COV sample community is particularly high. The composition ratio of each community in the CH samples was relatively average relative to the COV samples. The comparison against the overall contamination concentrations of the 14 PAHs showed that there may be a relationship between the pollution levels of COV and CH soils by PAHs and the abundance of the main species in the soil samples.

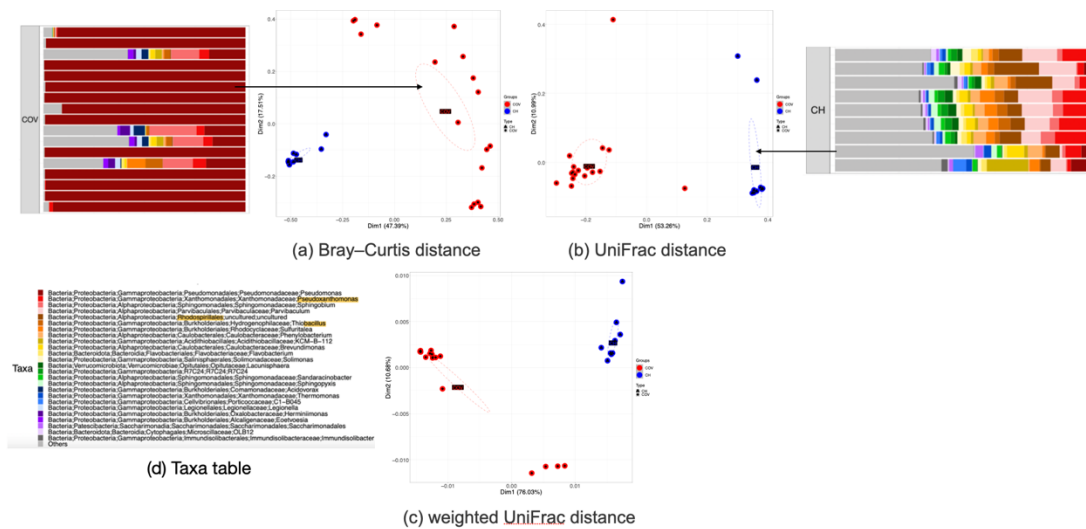
Bugg et al. (2000) pointed out that *Pseudomonas*, *Pseudoxanthomonas* and *Rhodospirillales* are PAH-degrading bacteria that are common inhabitants in soil and marine environments. *Pseudomonas* accounted for 71% of the degradation of PAHs and was the main decomposing species of PAHs (Medina et al. 2017) [37]. It takes up polycyclic aromatic hydrocarbons, such as phenanthrene, using passive transport and modulates the concentration of the cell associated polycyclic aromatic hydrocarbons by utilizing a chromosomal encoded active efflux system. Alexander (1999) [38] demonstrated that the *Pseudomonas* has implications for the metabolic pathways and biodegradation kinetics of PAHs and the role of these pollutants in environmental and bioremediation systems. It is inferred that the COV samples have higher

concentrations of PAHs, which means that the soil is more seriously polluted. COV samples require a high amount of biodegradation of PAHs, so the composition ratio of *Pseudomonas* in the COV community is high.

#### a) Alpha Diversity Measures



#### b) Beta Diversity Measure and Taxa Differential Analysis



**Figure 6:** a) The result of alpha diversity and NRI/NTI measures for the COV and CH soils populations. b) The beta diversity measured by (a)Bray-curtis, (b) Unifrac and (c) Weight Unifrac distances is displayed by PCoA plots, in which coloured ellipses indicate normative errors. The first 25 most abundant genera in each species group in the PCoA diagram are also illustrated surrounding the key to the (d) taxa table

### **3.3 Identifying deterministic and stochastic in COV and CH soil community environments**

#### **3.3.1 Null modelling: Calculating Quantitative Process Estimate and incidence-based (Raup-Crick) beta-diversity**

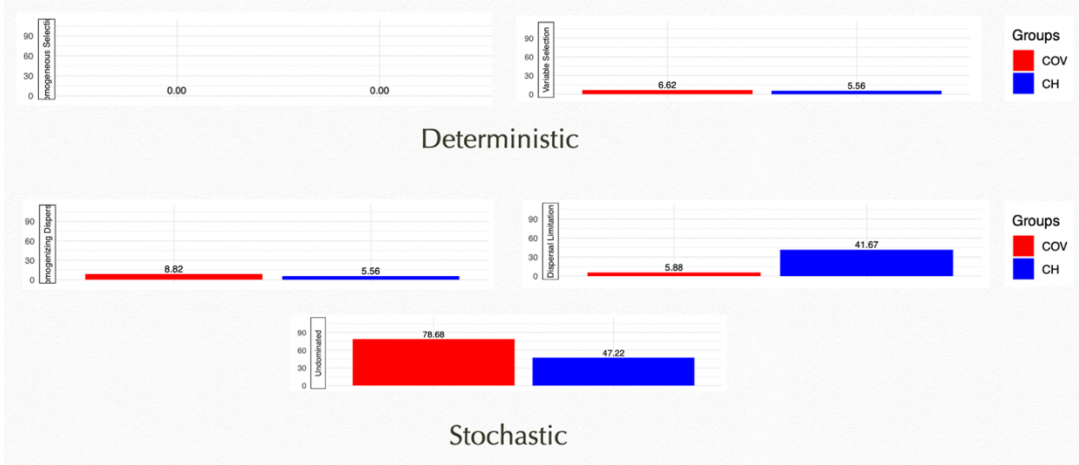
The dynamics of COV and CH community assembly were assessed by calculating QPE via a null model. According to Figure 7a, the main ecological driving factors of COV soil and CH soil are affected by ecological drift (undominated). This means that the dynamics of community assembly fluctuate randomly due to environmental changes. The dispersal limitation represents the highest number of assembly processes between communities caused by the movement of organisms. Figure 7a shows that the dispersal limitation of CH soil is 40%, which is seven times that of COV soil. The proportions of homogeneous dispersal, homogeneous selection and variable selection in COV and CH soils are all less than 10%. The above observations show that the higher variation of COV soil is related to the phenomenon of low environmental dispersal rate.

Incidence-based beta diversity is used to identify whether the community assembly in the sample is deterministic or stochasticity. In Figure 7b, the beta RC values for CH soils and COV soils both tend to deviate gradually from 0. The beta RC value for CH soils falls at approximately -0.4, and for COV soils it drops to -0.5, which is closer to -1 than for CH soils. In terms of variation, however, the  $\beta$  RC values of CH soils vary more dramatically than those of COV soils. This suggests that the CH soils are more community-assembled than the COV soils and that there is a higher similarity between communities.

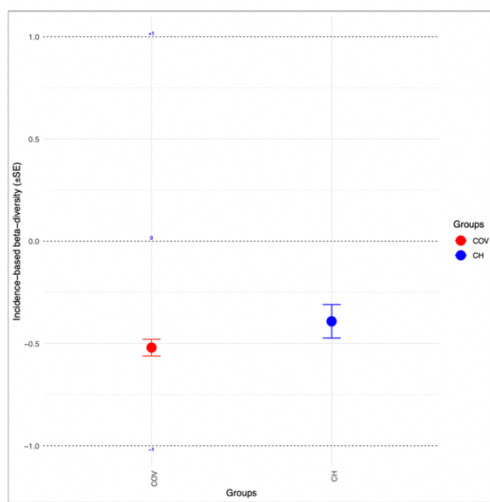
#### **3.3.2 Null modelling: Normalized stochasticity ratio (NST)**

To clearly observe the relative relationship between the ecological stochasticity of COV soil and CH soil under different conditions, NST was used to quantify the results of each assembly process. According to the results in Figure 7c, both COV soil and CH soil are less than or equal to 0.5. The assembly process is more deterministic overall, and the CH soil is in turn more deterministic than the COV soil community.

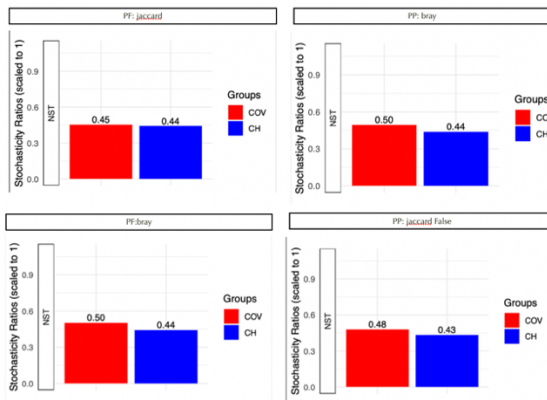
a) Quantitative Process Estimates



b) Incidence Based Raup-Crick



c) NST



**Figure 7:** a) The results of QPE. The contribution percentages for each assembly process are displayed, where Dispersal limitation, Homogenizing dispersal, Ecological drift (Undominated) belong to stochastic, whereas homogeneous selection and variable selection are categorized as deterministic. b) The results of the beta RC for COV and CH soil samples. When the  $\beta$  RC value does not deviate appreciably from 0, the community is attributed to the random assemblage; if the  $\beta$  RC value deviates from 0 and gradually approaches 1 or -1, it indicates a definite clustering of the community. c) The results of the NST. Quantified results for each assembly process. Stochasticity ratios are based on 0.5. When the value is less than 0.5, the community is more deterministic.

## 3.4 The Differential analysis

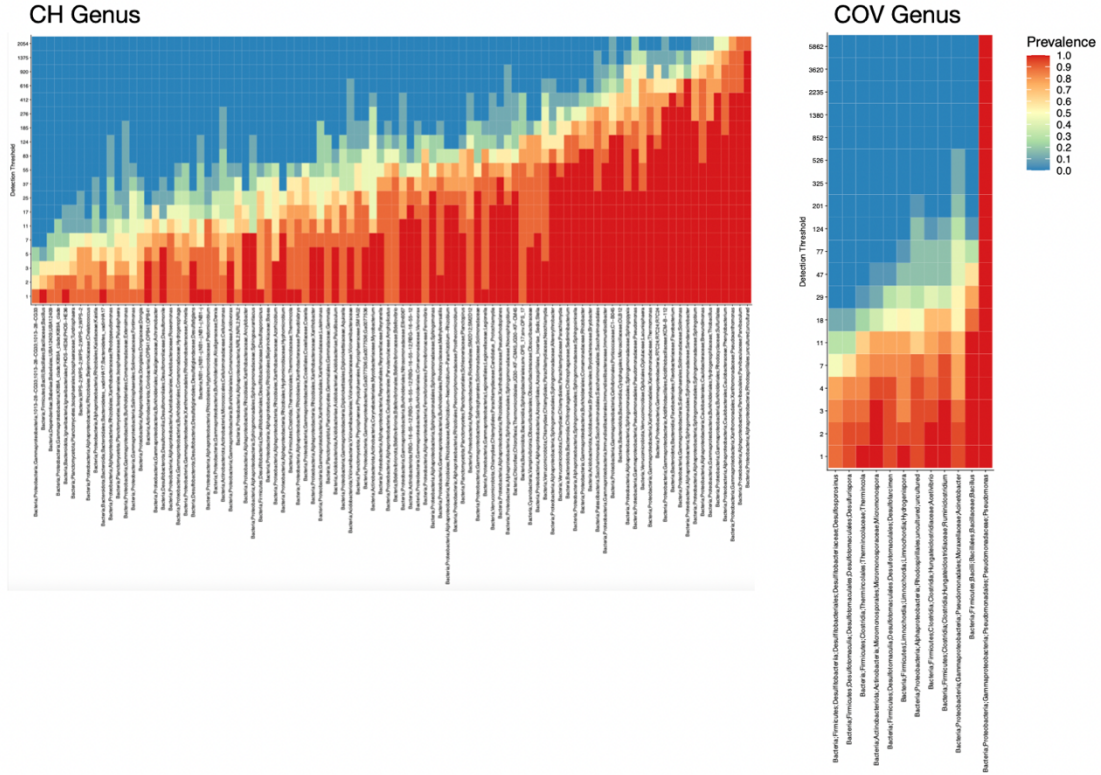
### 3.4.1 Core microbiome heatmap

The percentages of microorganisms in COV soil and CH soil communities were analyzed by the core microbiome and identified whether the communities belonged to the high or low abundance core microbiome. According to Figure 8, the COV and CH heatmap genus has at least 85% prevalence, which is represented at least 85% of species present in COV samples and CH samples. OTUs are ordered by their abundance in the heatmap. The blue area at the top of the heatmap means the low abundance core microbiome, while the red area at the bottom represents the high abundance core microbiome.

The heat map of the CH genus shows that the proportion of species is evenly distributed. The distributions of the low-abundance core microbiome and the high-abundance core microbes were consistent. It is consistent with the results of the CH top 25 most abundant taxa in Figure 6b.

From the COV genus heatmap, it can be seen that the maximum abundance threshold for *Pseudomonas* is 5862, which means that this species is almost 100% present in the sample. According to the taxonomic table of Figure 6b, *Pseudomonas* accounted for the highest percentage in the top 25 richest taxa associated with each species category identified. From these two points, it can be proved that *Pseudomonas* is the core microbiome with high abundance in COV soil, that is, the main biodegrading and repairing microorganisms in PAHs.

*Bacillus* has the second highest proportion in the heat map of COV genus. As Saeed et al. (2022) [39] indicated, *Bacillus* is a superior genus for bacteria for which petroleum hydrocarbon degraders and vegetation development accelerators are identified. It has the potential to degrade hydrocarbons, especially from strains isolated from contaminated soils. *Bacillus* can decompose PAHs pollutants, such as phenol, cresol, polychlorinated biphenyls (Christova et al., 2019) [40]. It also occupies a fairly high proportion among the top 25 most abundant taxa in Figure 6b. Overall, the biodegradation of COV soil and environmental remediation are closely related to *Pseudomonas* and *Bacillus*.



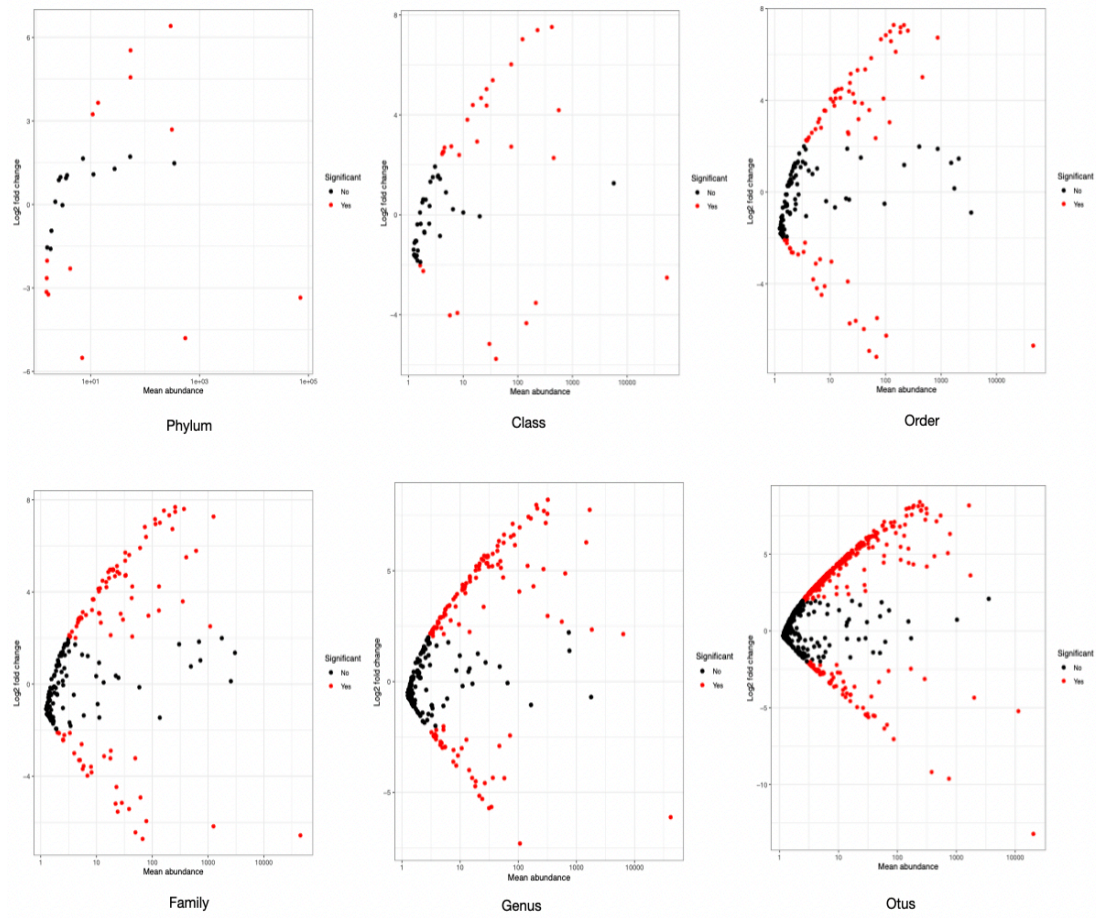
**Figure 8:** Core microbiome heatmap persisted in >85% of samples used to compare all populations separately and both COV and CH communities were compared simultaneously. OTUs are sequenced according to their abundance, the upper part of the heat map indicating the low abundance prevalent OTUs and the bottom showing the high abundance prevalent OTUs.

### 3.4.2 MA plot

The differences between the COV and CH sample communities were compared using the DESeq2 method. As shown in Figure 9, the parameters in the COV and CH samples were logarithmically transformed to form an MA plot based on the Bland-Altman plot. It is used to compare the microbial communities that differ between COV and CH soil samples. The x-axis of the MA plot represents the mean value, and the y-axis represents the log difference value. The red dots refer to the significant differences in microbial communities in these two samples.

The difference between the COV and CH samples in genus and phylum is shown in Figure 10. The log-relative normalized for CH samples is higher than the log-relative normalized for COV, which represents a significant difference between the two communities. The genus differences in Figure 10a are more diverse than the phylum species in Figure 10b and compare to the MA plots in Figure 9 reveals that the log-relative points of difference in genus are more numerous and denser than those in phylum. From the MA diagram of phylum to OTUs, it can be found that the difference points are clustered from the dispersion to the center line of  $\log\text{fold}=0$ . The MA diagram of OTUs finally forms a clear curve showing the exact difference distribution.



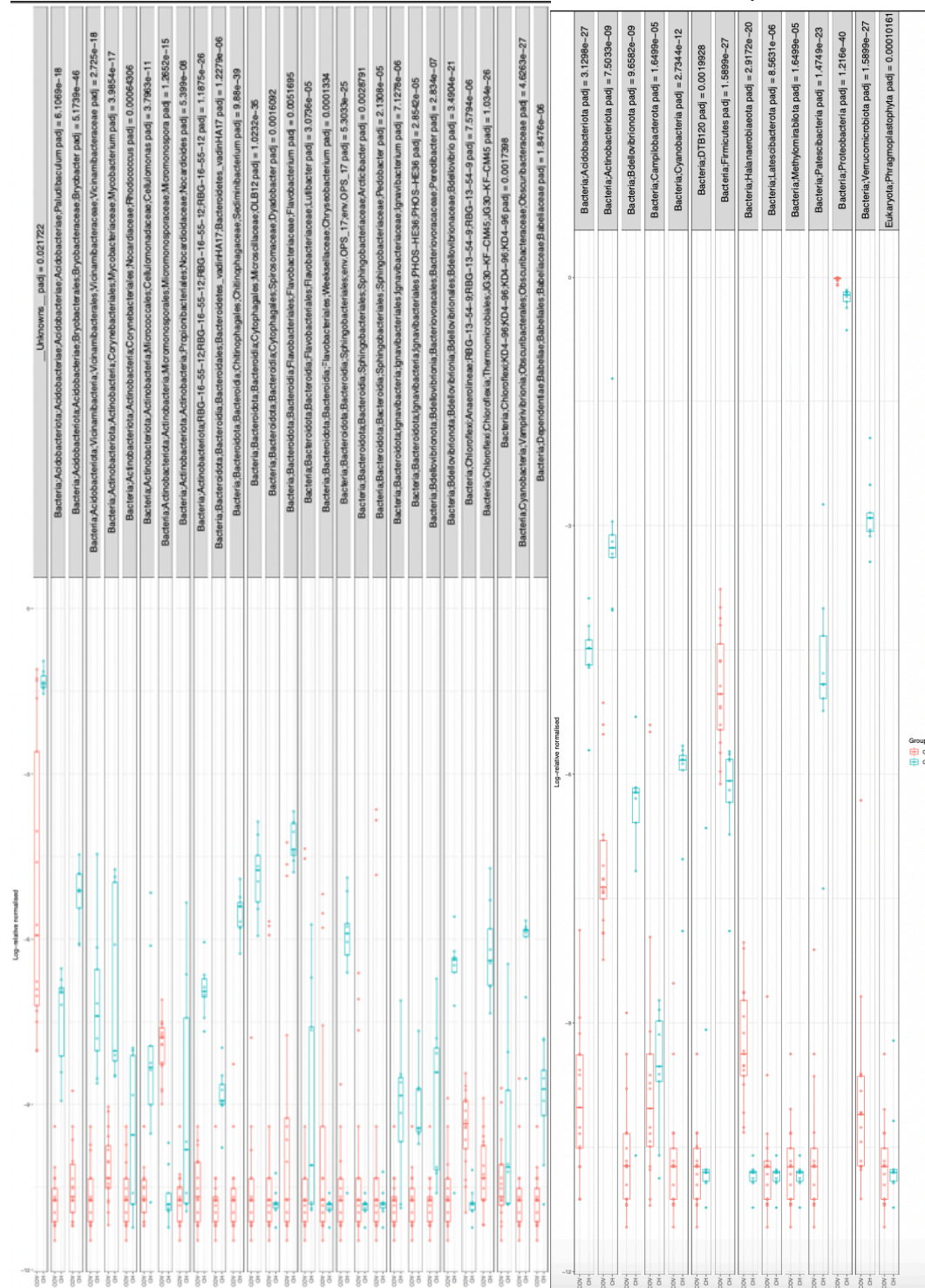


**Figure 9:** The results of the MA plot. The x-axis of the MA plot represents the mean value, and the y-axis represents the log difference value. The baseline is approximately  $\log_{2}\text{fold}=0$ . When the log difference for each microorganism is between plus and minus 2 logfold, it shows a black dot indicating that the difference is not excessive. The red dots indicate a significant difference when the difference exceeds +2 logfold or falls below -2 logfold.



# Genes

# Phylum



**Figure 10:** a) The significant different of genus. b) The significant different of phylum. The graph shows which microorganisms are significantly differentiated from each other. The more types of microbial variation, the more significant the distribution of red dots on Figure 9.

### **3.5 Regression Modelling: Determine key members of representative microbial communities of COV and CH soils**

#### **3.5.1 Subset regression**

The contaminant data from Figure 11 and the data from the meta table for aromatics in Figure 12 were combined into subset regression for analysis to observe their relationship to Shannon diversity. As shown in Figure 13, the minimum cross-validation error in the model is 0.37932. It represents the best analytical effect of the analytical model. The main variables in this model are the pollutants: Cadmium, Copper, and Cobalt; the substances in aromatic hydrocarbons: Indeno\_123\_cd\_pyrene, Dibenzo Ah Anthracene; LOI percentage. By referring to the consequences of the regression analysis on a subset of Figure 14. When the variable has a negative estimate, it implies that the higher the amount of the variable in the soil, the lower the Shannon diversity. On the contrary, a positive estimate of the variable means that the greater the content of the variable in the soil, the higher the Shannon diversity.

The value of Dibenzo Ah Anthracene in the Figure 14 is -50.57043, which is the lowest estimated value of all variables. Dibenzo ah anthracene is a polycyclic aromatic hydrocarbon consisting of multiple benzene rings and is a common oil contaminant (Rizzo et al., 2020) [41]. When it is present in higher proportions in the soil, the Shannon diversity of the community decreases. Richardson et al. (2007) [42] demonstrated that it is stable and highly genotoxic in bacterial and mammalian cell systems and tends to be embedded in DNA inducing mutations. Dibenzo Ah Anthracene is a threat to both microorganisms and humans.

The estimated value of the Indeno\_123\_cd\_pyrene variable falls at approximately 5, the highest of all variables. Indan\_123\_cd\_pyrene has mutagenic and carcinogenic properties as a chemical generated by the improper ignition of organic compositions. (Ribeiro et al., 2014) [43]. Although the predicted values are positive, the results are not significant. The overall results for the whole regression variables are mostly negative, implying a negative impact on community diversity.

meta_data2												
Name	HIRACE identifiant	Site	Lead	Iron	Cadmium	Chromium	Zinc	Copper	Nickel	Cobalt	Moisture Percentage	LOI Percentage
C610	HR14-COV-01 K	COV	1100	33000	1	230	360	470	57	11	14.75	17.54
CG11	HR14-COV-06	COV	485	37500	1	37	540	990	210	13	19.19	16.01
CG12	HR14-COV-08	COV	970	42000	0	33	400	420	93	13	18.92	11.66
CG13	HR14-COV-09 K	COV	400	35000	1	26	240	530	46	11	20.33	21.84
CG14	HR14-COV-09 T	COV	400	35000	1	26	240	530	46	11	20.33	21.84
CG15	HR14-COV-10	COV	830	39000	2	26	610	7900	66	13	19.97	23.84
CG16	HR14-COV-11	COV	550	30000	1	23	300	370	50	10	20.82	22.48
CG17	HR14-COV-12	COV	830	37000	1	36	420	650	62	13	19.54	21.82
CG18	HR14-COV-13	COV	620	43000	1	42	600	510	62	13	20.22	22.01
CG19	HR14-COV-14	COV	1500	47000	0	33	440	550	88	12	21.74	14.33
CG20	HR14-COV-15	COV	750	37000	0	27	420	570	100	13	19.51	10.63
CG21	HR14-COV-16	COV	570	50000	1	27	410	490	57	11	19.91	22.26
CG22	HR14-COV-17	COV	600	38000	0	25	720	1500	84	12	19.8	10.37
CG23	HR14-COV-18	COV	410	31000	2	21	230	180	54	10	21.4	23.93
CG28	HR16-CH-02	CH	17.51600222	6603.333144	0.356096686	7.199089477	20.189231	7.014309892	8.304771692	2.208516864	11.0802519	1.460013075
CG29	HR16-CH-03	CH	14.92488199	5324.337123	0.294040021	4.926133429	17.68845061	4.953898952	6.766696993	1.966465667	6.705707437	1.535109116
CG30	HR16-CH-04 1	CH	13.00228477	5422.152312	0.343355185	4.526934015	17.10256139	4.919786899	6.013912904	1.911696579	6.492397973	1.361756737
CG31	HR16-CH-04 2	CH	13.00228477	5422.152312	0.343355185	4.526934015	17.10256139	4.919786899	6.013912904	1.911696579	6.492397973	1.361756737
CG32	HR16-CH-04 3	CH	13.00228477	5422.152312	0.343355185	4.526934015	17.10256139	4.919786899	6.013912904	1.911696579	6.492397973	1.361756737
CG33	HR16-CH-05	CH	14.41437448	6274.927455	0.327570861	7.304794567	21.57859849	7.874531457	7.984574554	2.121098826	12.4607264	1.719187424
CG34	HR16-CH-06	CH	26.5502093	6245.887189	0.360510442	4.743853366	47.21698933	8.020178158	7.452779207	2.017952828	6.106432431	1.610621745
CG35	HR16-CH-07	CH	14.30091379	4578.598976	0.292394175	8.844215126	25.67182829	4.648344286	5.311047175	1.342191522	8.079519647	1.150747986
CG36	HR16-CH-08	CH	15.61908101	5453.329955	0.349893196	3.763639049	20.19851025	4.025954344	5.585236743	1.09710068	5.8805052949	5.11739706
CG6	HR14-COV-04	COV	520	41000	1	29	450	650	63	12	17.15	14.75
CG7	HR14-COV-05 1	COV	690	48000	1	33	740	1000	81	14	18.85	14.01
CG8	HR14-COV-05 2	COV	690	48000	1	33	740	1000	81	14	18.85	14.01
CG9	HR14-COV-05 3	COV	690	48000	1	33	740	1000	81	14	18.85	14.01

**Figure 11:** The meta table of pollutants in COV and CH soils such as lead, iron, cadmium, chromium, zinc, copper, nickel and cobalt, loss on ignition (LOI) percentage. The LOI percentage is the lab test to see how much is lost when a soil sample is heated to a certain temperature, and moisture percentage.

meta_data3_USETHISONE																	
Name	HIRACE identifiant	Site	Naphthalene	Acenaphthylene	Fluorene	Phenanthrene	Anthracene	Fluoranthene	Pyrene	Benzo a anthracene	Crysene	Benzo b fluoranthene	Benzo k fluoranthene	Benzo a pyrene	Indeno 123 cd pyrene	Dibenz ah anthracene	Benzo ghi perylene
CG01	HR14-COV-01 K	COV	0.32970815	0.250037604	0.596782687	1.60602784	0.432032142	1.11597521	0.946745987	0.323597771	0.22547125	0.033564819	0.227705486	0.106050322	0.121368059	0.153811833	
CG016	HR14-COV-06	COV	0.211189124	0.449371784	0.541832039	2.760417691	0.762062651	2.017215416	1.762877293	0.6077967	0.301787198	0.86501663	0.17168287	0.118676622	0.27846058	0.141688763	0.289258477
CG012	HR14-COV-08	COV	0.185205918	0.196750204	0.264709167	0.090886918	0.203965673	0.33630467	0.295458	0.158824029	0.150284945	0.08602693	-0.027789735	0.124189124	0.020262855	0.14057314	0.112640202
CG013	HR14-COV-09 K	COV	0.250308927	0.389604786	0.477598045	2.722432783	0.734039677	1.832333228	1.982300667	0.575173874	0.44688474	0.434759837	0.120286627	0.423398774	0.217832776	0.134330881	0.242352991
CG014	HR14-COV-09 T	COV	0.250308927	0.389604786	0.477598045	2.722432783	0.734039677	1.832333228	1.982300667	0.575173874	0.44688474	0.434759837	0.120286627	0.423398774	0.217832776	0.134330881	0.242352991
CG015	HR14-COV-10	COV	0.240817039	0.406899122	0.518023565	2.670142044	0.780086862	1.789443238	1.549878982	0.572521309	0.446717556	0.467868334	0.138348002	0.461088007	0.241366871	0.131081226	0.258200958
CG016	HR14-COV-11	COV	0.262563101	0.429482626	0.486656902	2.458512151	0.700886733	1.806387337	1.396433211	0.527221309	0.406222205	0.42253575	0.11056614	0.410631317	0.20962689	0.133834759	0.237707198
CG017	HR14-COV-12	COV	0.188480351	0.318975116	0.384233596	1.915854506	0.581189299	1.217810584	1.059337261	0.420013333	0.298498287	0.291891126	0.059182757	0.288705939	0.141575419	0.125782299	0.182942028
CG018	HR14-COV-13	COV	0.187891704	0.247954702	0.306050208	1.551446139	0.402093037	1.027195361	0.881392819	0.338124714	0.23871685	0.229719675	0.033630368	0.233678275	0.106893514	0.118950027	0.154891715
CG019	HR14-COV-14	COV	0.225221902	0.575488576	0.630911048	2.865975174	0.780118945	1.844177363	1.602230503	0.599678553	0.470503938	0.501504162	0.47868372	0.247409406	0.139845498	0.267403032	0.287440303
CG020	HR14-COV-15	COV	0.225221977	0.184048927	0.22559882	0.462271038	0.259394674	0.361732598	0.802028411	0.41989705	0.16817542	0.132815887	0.307749744	0.21205566	0.21205566	0.28574484	
CG021	HR14-COV-16	COV	0.159156414	0.20845053	0.055885763	0.062050622	0.146962782	0.203323116	0.203649154	0.091761509	0.3074854	0	0.038452393	0.020825136	0.10482732	0.077867304	
CG022	HR14-COV-17	COV	0.180337144	0.121247814	0.287482423	0.143315676	0.296628378	0.58891724	0.571212438	0.223528542	0.171225075	0.127514185	0	0.114666622	0.070147418	0.115468974	0.127476284
CG023	HR14-COV-18	COV	0.180337221	0.524499827	0.536004066	2.877429506	0.803284768	1.552258423	1.353818502	0.503232623	0.427804716	0.36775881	0.062847096	0.371637453	0.18132726	0.129068396	0.219625666
CG028	HR16-CH-02	CH	0.050418823	0.0281208	0.046877857	0.01096857	0.014534836	0.057235963	0.042031088	0.028110516	0.024778034	0.032439137	0.015650719	0.007189162	0.013242814		
CG029	HR16-CH-03	CH	0.050420369	0.054052681	0.045173218	0.016100362	0.018662057	0.064181748	0.080370669	0.03060887	0.027398702	0.042000432	0.018647543	0.007053913	0.017367913		
CG030	HR16-CH-04 1	CH	0.050011368	0.053415829	0.041328047	0.011237698	0.015789938	0.049775337	0.057030894	0.026524424	0.026331976	0.02633416	0.017038751	0.007547447	0.015099838		
CG031	HR16-CH-04 2	CH	0.050011368	0.053415829	0.041328047	0.011237698	0.015789938	0.049775337	0.057030894	0.026524424	0.026331976	0.02633416	0.017038751	0.007547447	0.015099838		
CG032	HR16-CH-04 3	CH	0.050011368	0.053415829	0.041328047	0.011237698	0.015789938	0.049775337	0.057030894	0.026524424	0.026331976	0.02633416	0.017038751	0.007547447	0.015099838		
CG033	HR16-CH-05	CH	0.050059099	0.053399355	0.041586291	0.011717209	0.014245497	0.048115737	0.05684998	0.026259173	0.02633234	0.026302983	0.016666916	0.007014019	0.014827854		
CG034	HR16-CH-06	CH	0.050059099	0.053399355	0.041586291	0.011717209	0.014245497	0.048115737	0.05684998	0.026259173	0.02633234	0.026302983	0.016666916	0.007014019	0.014827854		
CG035	HR16-CH-07	CH	0.050059099	0.053399355	0.041586291	0.011717209	0.014245497	0.048115737	0.05684998	0.026259173	0.02633234	0.026302983	0.016666916	0.007014019	0.014827854		
CG036	HR16-CH-08	CH	0.073221005	0.264868096	0.043302567	0.007828257	0.003203423	0.064486914	0.311781668	0.188220038	0.075991123	0.157442687	0.071106752	0.018024061	0.06486824		
CG37	HR16-MNH-01	MNH															
CG40	HR16-MNH-02 3	MNH															
CG45	Blank K	BLK															
CG46	Blank T	BLK															
CG50	HR17-ARM-02 (2)	MM															
CG54	HR17-ARM-04 (2)	MM															
CG59	HR17-ARM-07 (1)	MM															
CG6	HR14-COV-04	COV	0.227326891	0.256531821	0.314118809	1.924682849	0.512598646	1.336121623	1.176654264	0.421988712	0.325491039	0.325260468	0.081951355	0.307833327	0.152761501	0.122330487	0.189967828
CG60	HR17-ARM-07 (2)	MM															
CG61	HR17-ARM-08 (1)	MM															
CG62	HR17-ARM-08 (2)	MM															
CG63	HR17-ARM-09 (1)	MM															
CG64	HR17-ARM-09 (2)	MM															
CG7	HR14-COV-05 1	COV	0.228960095	0.388098049	0.508360332	2.483249257	0.702102591	1.750414149	1.42048373	0.63082733	0.502490018	0.47876896	0.206021167	0.482000037	0.308427794	0.142623951	0.302360561
CG72	HR17-ARM-10 (2)	MM															
CG8	HR14-COV-05 3	COV	0.228960095	0.388098049	0.508360332	2.483249257	0.702102591	1.750414149	1.42048373	0.63082733	0.502490018	0.47876896	0.206021167	0.482000037	0.308427794	0.142623951	0.302360561
CG9	HR14-COV-05 2	COV	0.228960095	0.388098049	0.508360332	2.483249257	0.702102591	1.750414149	1.42048373	0.63082733	0.502490018	0.47876896	0.206021167	0.482000037	0.308427794	0.142623951	0.302360561
CG10	HR14-COV-05 4	COV	0.228960095	0.388098049	0.508360332	2.483249257	0.702102591	1.750414149	1.42048373	0.63082733	0.502490018	0.47876896	0.206021167	0.482000037	0.308427794	0.142623951	0.302360561

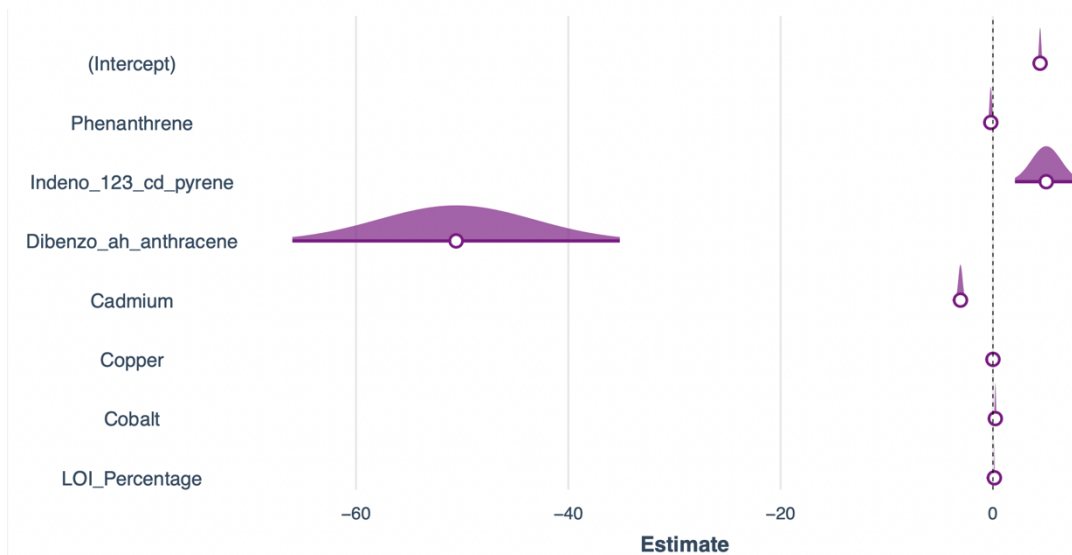


	Model	Cross-validation Errors
7	Shannon ~ Phenanthrene + Indeno_123_cd_pyrene + Dibenzo_ah_anthracene + Cadmium + Copper + Cobalt + LOI_Percentage	0.37932
13	Shannon ~ Naphthalene + Acenaphthylene + Phenanthrene + Fluoranthene + Pyrene + Benzo_k_fluoranthene + Indeno_123_cd_pyrene + Dibenzo_ah_anthracene + Iron + Cadmium + Copper + Cobalt + LOI_Percentage	0.42133
4	Shannon ~ Phenanthrene + Dibenzo_ah_anthracene + Cadmium + Copper	0.64515
12	Shannon ~ Naphthalene + Phenanthrene + Fluoranthene + Pyrene + Benzo_k_fluoranthene + Indeno_123_cd_pyrene + Dibenzo_ah_anthracene + Iron + Cadmium + Copper + Cobalt + LOI_Percentage	0.66274
1	Shannon ~ Phenanthrene	0.71565
2	Shannon ~ Phenanthrene + Cadmium	0.72269
10	Shannon ~ Phenanthrene + Pyrene + Benzo_k_fluoranthene + Indeno_123_cd_pyrene + Dibenzo_ah_anthracene + Iron + Cadmium + Copper + Cobalt + LOI_Percentage	0.90281
9	Shannon ~ Phenanthrene + Benzo_k_fluoranthene + Indeno_123_cd_pyrene + Dibenzo_ah_anthracene + Iron + Cadmium + Copper + Cobalt + LOI_Percentage	0.93088
11	Shannon ~ Phenanthrene + Fluoranthene + Pyrene + Benzo_k_fluoranthene + Indeno_123_cd_pyrene + Dibenzo_ah_anthracene + Iron + Cadmium + Copper + Cobalt + LOI_Percentage	1.01390
6	Shannon ~ Phenanthrene + Dibenzo_ah_anthracene + Cadmium + Copper + Cobalt + LOI_Percentage	1.02850
8	Shannon ~ Phenanthrene + Indeno_123_cd_pyrene + Dibenzo_ah_anthracene + Iron + Cadmium + Copper + Cobalt + LOI_Percentage	1.14070
5	Shannon ~ Phenanthrene + Dibenzo_ah_anthracene + Cadmium + Copper + LOI_Percentage	1.17009
3	Shannon ~ Phenanthrene + Cadmium + Copper	1.28021
14	Shannon ~ Naphthalene + Acenaphthylene + Fluorene + Phenanthrene + Fluoranthene + Pyrene + Benzo_k_fluoranthene + Indeno_123_cd_pyrene + Dibenzo_ah_anthracene + Iron + Cadmium + Copper + Cobalt + LOI_Percentage	2.62926
15	Shannon ~ Naphthalene + Acenaphthylene + Fluorene + Phenanthrene + Fluoranthene + Pyrene + Benzo_k_fluoranthene + Indeno_123_cd_pyrene + Dibenzo_ah_anthracene + Iron + Cadmium + Copper + Cobalt + Moisture_Percentage + LOI_Percentage	4.30453
16	Shannon ~ Naphthalene + Acenaphthylene + Fluorene + Phenanthrene + Fluoranthene + Pyrene + Benzo_k_fluoranthene + Indeno_123_cd_pyrene + Dibenzo_ah_anthracene + Benzo_ghi_perylene + Iron + Cadmium + Copper + Cobalt + Moisture_Percentage + LOI_Percentage	8.97694
17	Shannon ~ Naphthalene + Acenaphthylene + Fluorene + Phenanthrene + Fluoranthene + Pyrene + Benzo_k_fluoranthene + Indeno_123_cd_pyrene + Dibenzo_ah_anthracene + Benzo_ghi_perylene + Iron + Cadmium + Copper + Nickel + Cobalt + Moisture_Percentage + LOI_Percentage	15.40777
19	Shannon ~ Naphthalene + Acenaphthylene + Fluorene + Phenanthrene + Anthracene + Fluoranthene + Pyrene + Benzo_b_fluoranthene + Benzo_k_fluoranthene + Indeno_123_cd_pyrene + Dibenzo_ah_anthracene + Benzo_ghi_perylene + Iron + Cadmium + Copper + Nickel + Cobalt + Moisture_Percentage + LOI_Percentage	83.36956
20	Shannon ~ Naphthalene + Acenaphthylene + Fluorene + Phenanthrene + Anthracene + Fluoranthene + Pyrene + Benzo_b_fluoranthene + Benzo_k_fluoranthene + Indeno_123_cd_pyrene + Dibenzo_ah_anthracene + Benzo_ghi_perylene + Lead + Iron + Cadmium + Copper + Nickel + Cobalt + Moisture_Percentage + LOI_Percentage	134.77120
18	Shannon ~ Naphthalene + Acenaphthylene + Fluorene + Phenanthrene + Anthracene + Fluoranthene + Pyrene + Benzo_k_fluoranthene + Indeno_123_cd_pyrene + Dibenzo_ah_anthracene + Benzo_ghi_perylene + Iron + Cadmium + Copper + Nickel + Cobalt + Moisture_Percentage + LOI_Percentage	6486.38547

**Figure 13:** The Cross-validation errors of models, to determine the error value of the 20 models. The lower the value of Cross-validation errors, the better the results of subset regression.

Shannon									
Predictors	Estimates	std. Error	std. Beta	standardized std. Error	CI	standardized CI	Statistic	p	df
(Intercept)	4.44075 ***	0.10204	0.00000	0.02589	4.22637 – 4.65513	-0.05440 – 0.05440	43.51899	<b>1.082e-19</b>	18.00000
Phenanthrene	-0.20397	0.11511	-0.16418	0.09266	-0.44581 – 0.03788	-0.35886 – 0.03049	-1.77188	9.334e-02	18.00000
Indeno 123 cd pyrene	5.01970 **	1.40410	0.36785	0.10289	2.06980 – 7.96960	0.15168 – 0.58402	3.57503	<b>2.164e-03</b>	18.00000
Dibenzo ah anthracene	-50.57043 ***	7.34219	-1.98199	0.28776	-65.99579 – -35.14507	-2.58654 – -1.37743	-6.88765	<b>1.928e-06</b>	18.00000
Cadmium	-3.06109 ***	0.19448	-0.98493	0.06258	-3.46968 – -2.65250	-1.11639 – -0.85346	-15.73960	<b>5.746e-12</b>	18.00000
Copper	0.00032 ***	0.00004	0.31718	0.03623	0.00024 – 0.00039	0.24106 – 0.39329	8.75481	<b>6.631e-08</b>	18.00000
Cobalt	0.22316 ***	0.05390	0.76989	0.18594	0.10993 – 0.33639	0.37925 – 1.16053	4.14062	<b>6.141e-04</b>	18.00000
LOI Percentage	0.13317 ***	0.02205	0.74129	0.12276	0.08684 – 0.17950	0.48339 – 0.99919	6.03876	<b>1.041e-05</b>	18.00000
Observations	26								
R <sup>2</sup> / R <sup>2</sup> adjusted	0.987 / 0.983								

\*  $p < 0.05$  \*\*  $p < 0.01$  \*\*\*  $p < 0.001$



**Figure 14:** The result of subset regression. The distribution corresponds to the estimate values in the table. Negative values indicate that the more the substance is present, the lower the Shannon diversity will be. The higher the value of \* shows the more significant.

### 3.5.2 CODA GLMNET

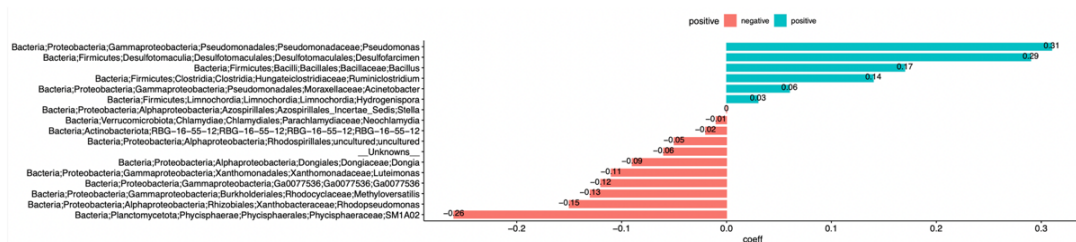
Using CODA GLMNET regression analysis to observe which substances are associated with soil PH and *pseudomonas*. As with the subset regression, the material parameters in the meta table in Figure 11 and Figure 12 are used as regression variables for analysis. According to the results, Naphthalene, Cadmium, Cobalt and Dibenzo Ah Anthracene was all related to *Pseudomonas* to some extent.

The coefficient value of *Pseudomonas* is 0.31 in the Naphthalene substances in Figure 15(a), which is the microorganism with the highest coefficient value in the substance. This indicates that Naphthalene and *Pseudomonas* are highly related. Naphthalene is an organic compound of polycyclic aromatic hydrocarbons. According to Falahatpisheh et al. (2001) [44], Intermediate products of naphthalene covalently associate with DNA which is prevalent in liver, kidney and lung tissue and inhibit mitochondrial respiration, resulting in an increased incidence of cancer in animals. Cerniglia (1993) [45] indicated that *Pseudomonas* has the ability to completely degrade Naphthalene. It is inferred that the two are highly correlated, so it is inferred that *Pseudomonas* is the main microorganism that degrades Naphthalene. Since the two are highly correlated, it is inferred that *Pseudomonas* is the main microorganism that degrades Naphthalene.

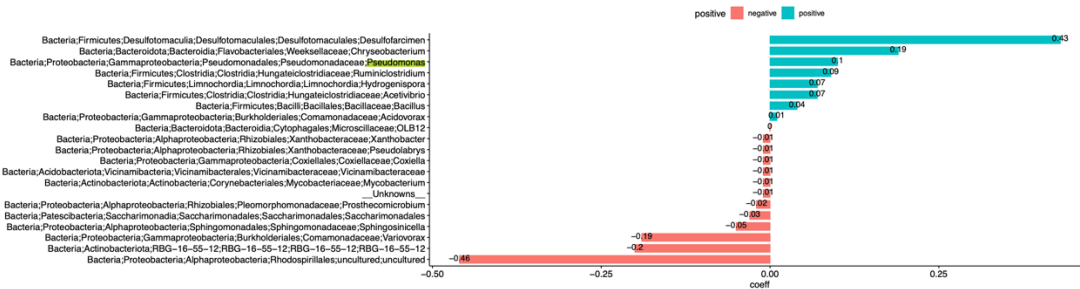
In Figure 15(b), the Dibenzo Ah Anthracene material is found to have a correlation coefficient of 0.1 for *pseudomonas*. It is the third highest positive correlation microorganism in the Dibenzo Ah Anthracene. *Pseudomonas* can biodegrade Dibenzo Ah Anthracene and decompose it into a carbon source or energy source (Juhasz et al., 1997) [46]. The correlation coefficient for *Pseudomonas* in Figure 15(c) is 0.01 for the Cadmium contaminant. The correlation with Cadmium is relatively low compared to that of Dibenzo Ah Anthracene and Naphthalene.

The Positive and negative correlations are in contrast to each other. When the content of this substance is more, the growth of microorganisms negatively related to it will be inhibited. The Figure 15(d) shows a negative correlation between Cobalt and *pseudomonas* with a correlation coefficient of -0.02. This means that Cobalt inhibits the degradation ability of *pseudomonas* in this pollutant. *pseudomonas* grow unfavorably in Cobalt to the extent that this metal contaminant cannot be effectively degraded (Oyetibo et al., 2013) [47].

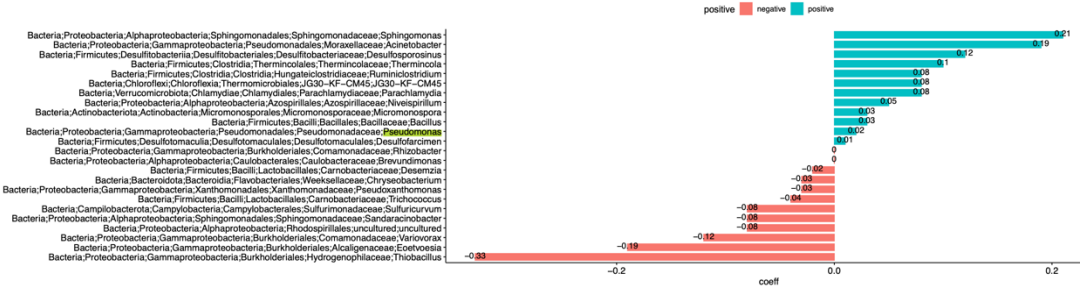
(a)Naphthalene



(b) Dibenzo Ah Anthracene



## (c) Cadmium



(d) Cobalt

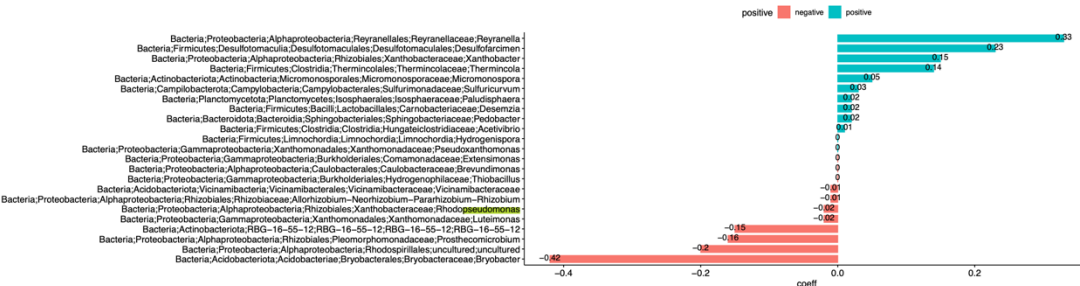


Figure 15: The CODA GLMNET result of (a) Naphthalene (b) Dibenzo Ah Anthracene (c) Cadmium (d) Cobalt. When the coefficient is greater than zero, which is the green area on the figure, it means that the growth of this substance has a positive correlation with those microorganisms. Conversely, if the coefficient is less than zero, it is shown by the red bar graph, which means that microorganisms exhibiting a negative coefficient are negatively correlated with the substance.

## CHAPTER 4

### 4 Discussion

#### 4.1 The Biological Informatics Approach to Biodegradation Analysis of Contaminants

The taxonomic matching method of nucleic acid sequencing is one of the main goals of analyzing ecological microbial communities. This project uses genomic DNA analysis methods to extract DNA fragments from samples. The DNA fragments are sequenced, so that the amplified and sequenced OTU or ASV can be generated. Amplicon sequencing is a type of Target Sequencing, which aims to distinguish different species by amplifying and sequencing the 16S rRNA region of bacteria. (Chakravorty et al., 2007) [48]. The amplification sequencing of this project is to use the OTU method to detect the relationship between samples by clustering with 97% or 99% similarity. According to this study, this method is helpful for studying the degradation relationship between microorganisms and soil. However, the sample data available for this study were limited, so it difficult to perform comparisons on genetic and field data as each method retrieved a variable number of taxa. Sample collection and extraction methods still need to be improved. In the future, there is an opportunity to try to use the amplicon sequencing variant (ASV) to detect samples by fitting a distribution (Poisson distribution), that is rarely available for the analysis of microbial degradation data.

The regression model is essential for observing the relationship between soil and its microbial degradation. As far as I know, the model has rarely been used to analyse soil biodegradation in the past. In addition to using the generally available Alpha diversity and Beta diversity to study cross-sections of soil microbial communities at two contaminated sites, COV (UK) and CH (Switzerland), this project also focuses on regression modeling to identify key members of biodegradation in COV and CH soil samples. The regression modeling can clearly show the correlation between individual substances in contaminated soil and microorganisms. Therefore, regression analysis may be incorporated into the main part of microbial community data analysis to be able to identify key species and their effect on the environment.

## 4.2 Overlaps and differences between COV and CH soils

This project uses nucleic acid sequencing to identify contaminants in COV and CH soils. Gauchotte-Lindsay et al. (2019) [5] classified soil samples collected using trnL in the UK and Switzerland, respectively. *Pseudomonas* were found from taxa as the main microorganisms in COV and CH soils species.

According to the analysis of R studio, some similarities, and differences between COV and CH soil samples were obtained. In terms of diversity, the CH region has a greater species richness than the COV one. The diversity within the sample is also relatively rich. Obviously, the CH community is more diverse. Through environmental filtering analysis, it can be found that the soil communities in the CH area are more likely to be affected by the environment. There is a higher capacity for species to interact with each other and a higher degree of community similarity than in COV communities. The COV community is dominated by the principle of competitive exclusion, so that the differences between soil communities are large. Its phylogenetic tree distribution is also more dispersion.

There are two main reasons for the differences between communities. The first reason is that ecological drivers are affected by ecological drift, and the dynamics of community assembly will change with environmental changes. The environmental dispersal rate of CH soil was higher than that of COV soil. The areas with high environmental dispersal rate are conducive to the mutual influence between species, so that the similarity between communities will increase. Due to the low rate of environmental diffusion, COV soils are highly variable between species. The second reason is the concentration of PAHs. The PAHs concentration in the COV samples was on average twice that of the CH samples, which represented a higher degree of PAH contamination in the COV samples. The PAHs may affect community richness. The high level of PAHs contamination will lead to the spatial dispersion of  $\beta$  diversity. The variability among COV communities is therefore greater.

The correlation of PAHs and microbial *Pseudomonas* was observed to be significant positively. When the concentration of PAHs increased, it was beneficial to the growth of *Pseudomonas*. *Pseudomonas* have the ability to help soil degradation and repair. COV soil samples were so polluted with PAHs that the percentage of *Pseudomonas* in the microbial composition of the soil was extremely high.

The COV and CH soil samples were collected only from soils from specific areas in the UK and Switzerland, therefore analysis of soil PAH contaminated concentrations and key microbial results for soil degradation in that area was limited to that specific area of the sample. The results are not representative of all soils in the two countries, but this analysis can give a clear understanding of how the soil environment in the region is effectively biodegraded and soil remediation. In the future, soil samples from more regions will be collected to understand the key microorganisms in soil degradation in more regions, so that more PAHs contaminated communities can be rehabilitated.



### 4.3 The critical microorganism for biodegradation

The PAHs are a cluster of organic mixtures, such as naphthalene, dibenzo-anthracene phenols, cresols, and polychlorinated biphenyls. These different contaminations are decomposed by different microorganisms. However, microorganisms cannot completely decompose any pollutants, such like iron, cobalt and other metal pollutants. They typically depress the development of microorganisms and reduce their capacity to decompose pollutants. When microorganisms cannot completely decompose the contaminations, the soil will form permanent pollution, which will cause great harm to the environment and organisms. This project analyses the main core microorganisms of soil degradation by the number of collected samples, so as to improve the degradation ability of contaminated soil.

The project was based on taxon and core microbiome analysis via a null model, which revealed that *Pseudomonas*, *Pseudoxanthomonas*. and *Rhodobacter* accounted for a certain proportion of the microbial composition in COV and CH soils. The regression analysis revealed that *Pseudomonas* are the most important microorganisms for the biodegradation and repair of PAHs, especially for Naphthalene. Because COV soil is seriously polluted by PAHs, more microorganisms are needed to biodegrade the soil. *Pseudomonas* is also positively associated with PAHs contaminants such as naphthalene and dibenz an thracene. As the soil becomes more contaminated with PAHs, *Pseudomonas aeruginosa* is able to grow in the soil with higher efficiency and biodegradation. *Pseudomonas* has a certain influence on the metabolic pathways and biodegradability of PAHs and their ability to remediate soil damaged by pollutants (Alexander, 1999) [38].

## CHAPTER 5

### 5 Conclusions and Future work

This project investigates the use of microorganisms to biodegrade contaminated soils to solve soil contamination problems. It is guided by recent advances in microbial community assembly mechanisms and develops an analytical strategy to elucidate the role of the environment and reveal the key members of the microbial community representing these sites. The Alpha diversity analysis showed that the lower the value of COV compared to CH, the less diverse the COV community. Simultaneously, Beta diversity showed that PAH concentrations in COV were on average higher and more variable than PAH concentrations in CH samples, resulting in a spatial dispersion of their beta diversity. The COV soil samples seem to be polluted with more PAHs than CH ones.

The quantitative results from QPE and NST show that the CH soil samples in which the community is more deterministic and environmentally influential. In contrast, COV is motivated by the principle of competitive exclusion, in which the species that outperforms other species in a particular ecological niche dominates. It leads to a more fragmented phylogenetic tree. It is inferred that the determinism of the microbial community correlates strongly with the presence or absence of chemical contaminants, especially PAHs.

The difference analysis showed that the COV and CH soil samples were very different. It was concluded from the analysis that COV soils contaminated with PAHs may contain higher levels of Dibenzo Ah Anthracene and Naphthalene organic compounds. Both species are mostly biodegraded by *Pseudomonas*, which has remediated the contaminated soil. In the present analysis, *Pseudomonas* was the main microorganism responsible for the degradation of PAHs.

This project samples only studied sites with space variability (COV and CH), and the models detected were mainly inter-site distinctions. This method of analysis is therefore not representative of the entire biosphere. However, microbial community survey data using 16S rRNA in association with GC  $\times$  GC MS have been shown to be valid.

Regarding this project, the soil samples were only selected from specific areas in the United Kingdom and Switzerland, so that the results are not representative of all microbial communities exposed to PAHs. The ecological environment that surrounds each region also varies. The level of contamination from heavily contaminated industries in the developed or developing countries is more severe than in the backward countries, and the biodegradable microorganisms corresponding to these different contaminants also differ. In the future, it is hoped to expand the number of areas from which soil samples are collected in order to better identify the key contributors to the microbial community in each area. The increased number of samples will allow for more accurate statistics on the evolution and changes of microbial communities in the ecosystem.

## CHAPTER 6

### 6 References

- [1] Schlatter, C. (1994). Environmental pollution and human health. *Science of the total environment*, 143(1), 93-101.
- [2] Stading, R., Gastelum, G., Chu, C., Jiang, W., & Moorthy, B. (2021, November). Molecular mechanisms of pulmonary carcinogenesis by polycyclic aromatic hydrocarbons(PAHs): Implications for human lung cancer. In *Seminars in cancer biology* (Vol.76, pp.3-16). Academic Press.
- [3] Teuten, E. L., Rowland, S. J., Galloway, T. S., & Thompson, R. C. (2007). Potential for plastics to transport hydrophobic contaminants. *Environmental science & technology*, 41(22), 7759-7764.
- [4] Chidambarampadmavathy, K., Karthikeyan, O. P., & Heimann, K. (2017). Sustainable bio-plastic production through landfill methane recycling. *Renewable and Sustainable Energy Reviews*, 71, 555-562.
- [5] Gauchotte-Lindsay, C., Aspray, T. J., Knapp, M., & Ijaz, U. Z. (2019). Systems biology approach to elucidation of contaminant biodegradation in complex samples-integration of high-resolution analytical and molecular tools. *Faraday discussions*, 218, 481-504.
- [6] Phillips, D. H. (1999). Polycyclic aromatic hydrocarbons in the diet. *Mutation research/genetic toxicology and environmental mutagenesis*, 443(1-2), 139-147.
- [7] Ohura, T., Amagai, T., Fusaya, M., & Matsushita, H. (2004). Polycyclic aromatic hydrocarbons in indoor and outdoor environments and factors affecting their concentrations. *Environmental science & technology*, 38(1), 77-83.
- [8] Samanta, S. K., Singh, O. V., & Jain, R. K. (2002). Polycyclic aromatic hydrocarbons: environmental pollution and bioremediation. *TRENDS in Biotechnology*, 20(6), 243-248.
- [9] Singh, S. K., & Haritash, A. K. (2019). Polycyclic aromatic hydrocarbons: soil pollution and remediation. *International Journal of Environmental Science and Technology*, 16(10), 6489-6512.
- [10] Magi E, Bianco R, Ianni C, Carro MD (2002) Distribution of polycyclic aromatic hydrocarbons in the sediments of the Adriatic Sea. *Environ Pollut* 119:91–98.
- [11] Hreniuc M, Coman M, Cioruța B (2015) Consideration regarding the soil pollution with oil products in Sacel-Maramures. In: International conference of scientific paper AFASES, Brasov, 28–30.

- [12] Torsvik, V., Daae, F. L., Sandaa, R. A., & Øvreås, L. (1998). Novel techniques for analysing microbial diversity in natural and perturbed environments. *Journal of biotechnology*, 64(1), 53-62.
- [13] Samarajeewa, A. D., Hammad, A., Masson, L., Khan, I. U. H., Scroggins, R., & Beaudette, L. A. (2015). Comparative assessment of next-generation sequencing, denaturing gradient gel electrophoresis, clonal restriction fragment length polymorphism and cloning-sequencing as methods for characterizing commercial microbial consortia. *Journal of microbiological methods*, 108, 103-111.
- [14] Liang, K. Y., Orata, F. D., Boucher, Y. F., & Case, R. J. (2021). Roseobacters in a sea of poly-and paraphyly: whole genome-based taxonomy of the family Rhodobacteraceae and the proposal for the split of the “Roseobacter clade” into a novel family, Roseobacteraceae fam. nov. *Frontiers in Microbiology*, 12, 1635.
- [15] Simon, C., & Daniel, R. (2011). Metagenomic analyses: past and future trends. *Applied and environmental microbiology*, 77(4), 1153-1161.
- [16] Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K. H., ... & Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, 12(9), 635-645.
- [17] Castelino, M., Eyre, S., Moat, J., Fox, G., Martin, P., Ho, P., ... & Barton, A. (2017). Optimisation of methods for bacterial skin microbiome investigation: primer selection and comparison of the 454 versus MiSeq platform. *BMC microbiology*, 17(1), 1-12.
- [18] The 16S rRNA and 16S rRNA Gene. (2022, August 16). In EZBioCloud Help Center. <https://help.ezbiocloud.net/16s-rrna-and-16s-rrna-gene/>.
- [19] Vass, M., Szekely, A. J., Lindstrom, E. S., & Langenheder, S. (2020). Using null models to compare bacterial and microeukaryotic metacommunity assembly under shifting environmental conditions. *Scientific Reports*, 10(1), 1-13.
- [20] Willis, A. D. (2019). Rarefaction, alpha diversity, and statistics. *Frontiers in microbiology*, 10, 2407.
- [21] Whittaker R. H. (1972) Evolution and measurement of species diversity. *taxon*, 21: 213–251.
- [22] Pielou, E. C. (1966). The measurement of diversity in different types of biological collections. *Journal of theoretical biology*, 13, 131-144.

- [23] Shannon, C. E. (1948). A mathematical theory of communication. The Bell system technical journal, 27(3), 379-423.
- [24] Simpson, E. H. (1949). Measurement of diversity. *nature*, 163(4148), 688-688.
- [25] Dumbrell, A.J., Nelson, M., Helgason, T., Dytham, C., & Filter, A. H. (2010). Relative roles of niche and neutral processes in structuring a soil microbial community. The ISME journal, 4(3), 337-345.
- [26] Stegen, J. C., Lin, X., Konopka, A. E., & Fredrickson, J. K. (2012). Stochastic and deterministic assembly processes in subsurface microbial communities. The ISME journal, 6(9), 1653-1664.
- [27] Chase, J. M., & Myers, J.A. (2011). Disentangling the importance of ecological niches from stochastic processes across scales. Philosophical transactions of the Royal Society B: Biological sciences, 366(1576), 2351-2363.
- [28] Cooper, N., Rodríguez, J., & Purvis, A. (2008). A common tendency for phylogenetic overdispersion in mammalian assemblages. *Proceedings of the Royal Society B: Biological Sciences*, 275(1646), 2031-2037.
- [29] Bray, J.R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. Ecological monographs, 27(4), 326-349.
- [30] Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., & Knight, R. (2011). UniFrac: an effective distance metric for microbial community comparison. The ISME journal, 5(2), 169-172.
- [31] Lozupone, C., & Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. Applied and environmental microbiology, 71(12), 8228-8235.
- [32] Stegen, J. C., Lin X., Fredrickson, J. K., & Konopka, A. E. (2015). Estimating and mapping ecological processes influencing microbial community assembly. Frontiers in microbiology, 6, 370.
- [33] Ning, D., Deng, Y., Tiedje, J. M., & Zhou, J. (2019). A general framework for quantitatively assessing ecological stochasticity. Proceedings of the National Academy of Sciences, 116(34), 16892-16898.
- [34] Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology, 15(12), 1-21.
- [35] Bland-Altman Plot. (2022, August 16). In Wikipedia. [https://en.wikipedia.org/wiki/Bland–Altman\\_plot](https://en.wikipedia.org/wiki/Bland–Altman_plot)

- [36] MA plot. (2022, August 16). In Wikipedia. [https://en.wikipedia.org/wiki/MA\\_plot](https://en.wikipedia.org/wiki/MA_plot)
- [37] Medina, R., David Gara, P. M., Rosso, J., & Del Panno, M. T. (2017, June). Remediation of a hydrocarbon chronically contaminated soil by combination of persulfate oxidation and bioremediation. In International conference AquaConSoil (Vol. 14). AquaConSoil.
- [38] Alexander, M. (1999). Biodegradation and bioremediation. Gulf Professional Publishing.
- [39] Saeed, M., Ilyas, N., Bibi, F., Jayachandran, K., Dattamudi, S., & Elgorban, A. M. (2022). Biodegradation of PAHs by *Bacillus marsiflavi*, genome analysis and its plant growth promoting potential. *Environmental Pollution*, 292, 118343.
- [40] Christova, N., Kabaivanova, L., Nacheva, L., Petrov, P., & Stoineva, I. (2019). Biodegradation of crude oil hydrocarbons by a newly isolated biosurfactant producing strain. *Biotechnology & Biotechnological Equipment*, 33(1), 863-872.
- [41] Rizzo, P., Malerba, M., Bucci, A., Sanangelantoni, A. M., Remelli, S., & Celico, F. (2020). Potential enhancement of the in-situ bioremediation of contaminated sites through the isolation and screening of bacterial strains in natural hydrocarbon springs. *Water*, 12(8), 2090.
- [42] Richardson, S. D., Plewa, M. J., Wagner, E. D., Schoeny, R., & DeMarini, D. M. (2007). Occurrence, genotoxicity, and carcinogenicity of regulated and emerging disinfection by-products in drinking water: a review and roadmap for research. *Mutation Research/Reviews in Mutation Research*, 636(1-3), 178-242.
- [43] Ribeiro, J., Silva, T. F., Mendonça Filho, J. G., & Flores, D. (2014). Fly ash from coal combustion—An environmental source of organic compounds. *Applied Geochemistry*, 44, 103-110.
- [44] Falahatpisheh, M. H., Donnelly, K. C., & Ramos, K. S. (2001). Antagonistic interactions among nephrotoxic polycyclic aromatic hydrocarbons. *Journal of Toxicology and Environmental Health Part A*, 62(7), 543-560.
- [45] Cerniglia, C. E. (1993). Biodegradation of polycyclic aromatic hydrocarbons. *Current opinion in biotechnology*, 4(3), 331-338.
- [46] Juhasz, A. L., Britz, M. L., & Stanley, G. A. (1997). Degradation of fluoranthene, pyrene, benz [a] anthracene and dibenz [a, h] anthracene by *Burkholderia cepacia*. *Journal of Applied Microbiology*, 83(2), 189-198.
- [47] Oyetibo, G. O., Ilori, M. O., Obayori, O. S., & Amund, O. O. (2013). Biodegradation of petroleum hydrocarbons in the presence of nickel and cobalt. *Journal of basic microbiology*, 53(11), 917-927.

- [48] Chakravorty, S., Helb, D., Burday, M., Connell, N., & Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of microbiological methods*, 69(2), 330-339.

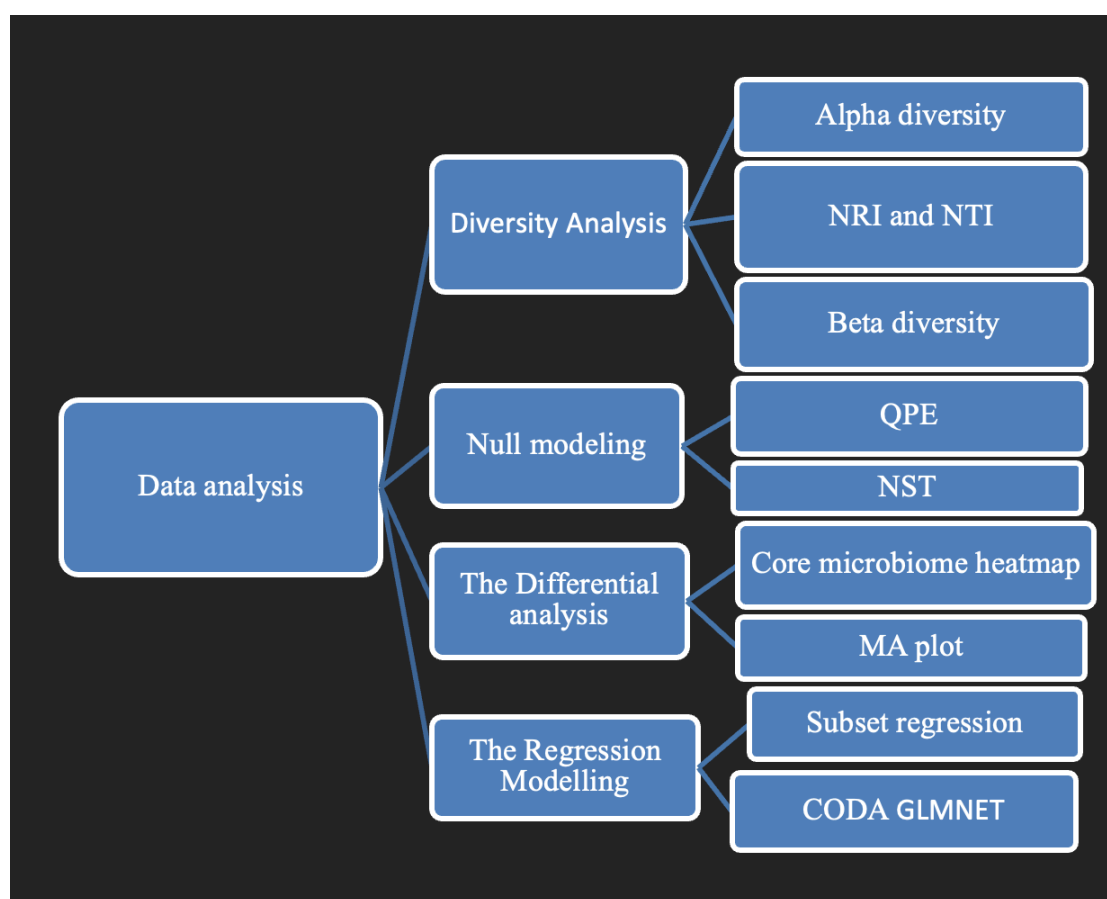


## CHAPTER 7

### 7 Appendix

#### 7.1 The data analysis architecture

These are divided into four main points of analysis: diversity analysis, null model analysis, differential analysis and regression analysis. Diversity analysis is divided into Alpha diversity to summarize the community diversity of COV and CH soil samples in terms of Beta diversity to compare the similarity between the two samples. Null model analysis is used to calculate QPE and NST to determine whether the COV and CH soil samples have the basic characteristics of a single environment and to quantify the microbial community. The analysis of differences is separated into a core microbiome heat map to identify which microorganisms are present in each of the COV and CH soil samples, and an MA plot to provide a comparison of the differences between the two samples' communities. The subset of regression analysis is used to observe the correlation between contaminants and the soil. The CODA GLMNET is a way to detect which substances in a sample are associated with soil PH and *Pseudomonas*.



**Figure 16:** *The data analysis architecture*

## 7.2 The Result of the OTU table in data analysis

Diversity_Otus_Hypothesis1					
	Richness	Shannon	Simpson	FisherAlpha	PielouEvenness
CG1	153.469190330169	0.556757789897159	0.136203870522369	39.1077985645865	0.098807201174648
CG10	60.849057349642	0.69413259924189	0.378788893337929	17.8308539899667	0.13890396509393
CG12	153.917992635634	3.59053043256039	0.949309741697734	27.3398633417826	0.679612336929144
CG13	34.6324619855177	0.759701174630349	0.504305958670376	10.5331573549422	0.167608273304959
CG14	44.1200824493204	0.774193597443777	0.508020488640206	15.2114776147544	0.158310496274382
CG15	49.9292506141201	0.937525683899087	0.507174943200067	12.4284269957506	0.200633431506174
CG16	37.3466659634983	0.379685774123553	0.182812976599233	8.00702000587099	0.0906245973719326
CG17	67.1808366692219	1.69157589036911	0.74713820926826	14.6774043616128	0.354577177810947
CG18	63.5208736966597	1.30972206690792	0.684449833304996	14.7020054786372	0.27453545350519
CG19	84.052712594812	2.73824672253646	0.901946588599411	15.3749933550556	0.569023653360128
CG20	164.325224876554	3.66088912102937	0.953489790852276	29.6796276793561	0.681060373690781
CG21	51.3728323477995	1.35405736137354	0.693227328232306	10.3788917215255	0.306428197248039
CG22	123.834946346077	3.33438172253026	0.946312206114912	20.9188185985965	0.656192950418828
CG28	290.812931444443	3.77818591671853	0.938523946499737	55.169075281423	0.637461291091435
CG29	292.137398287997	3.97545442435424	0.95329687578633	56.7057442773826	0.666049456168026
CG30	282.20013448458	3.86882620378942	0.94527292812487	55.7098911006142	0.65187787284311
CG31	263.989963094747	3.63126276677629	0.939476422509772	50.0263555919681	0.621723259154597
CG32	264.253495664646	3.68050934468727	0.94318997936735	50.5295922275956	0.629530127021801
CG33	270.896275679287	3.72098609308418	0.945942127470497	53.0135142794854	0.62981277209988
CG34	298.162352430143	3.74851395738854	0.936231368000726	58.5163938668988	0.620833735451383
CG35	321.943979643529	3.7571256924088	0.922208519851369	70.0094179976343	0.611921227581796
CG36	213.582879680855	3.57547155192915	0.934079358767427	40.76189340115	0.634133849609649
CG6	51.4989050176021	0.182442689254324	0.0504889339981891	14.5634021318524	0.0379126558297047
CG7	41.1378089605885	0.357499206957306	0.149510476350793	8.89453542663689	0.0820571723240335
CG8	41	0.283607103283752	0.113663903939335	5.32348809684413	0.0763704320771254
CG9	62.0763931459537	0.587565289060569	0.22598670332143	15.475091546048	0.120521334852301

Figure 17: The OTU table of alpha diversity

	NRI	NTI
CG1	1.6322024687821	1.14450167614402
CG10	1.17365404630558	1.19359175165034
CG12	1.48752183355863	1.89157652331095
CG13	1.00623355742801	1.16625934780448
CG14	1.06720024868767	1.26960400552447
CG15	0.921439998799508	1.357276139362
CG16	0.876823309037896	0.78350307937264
CG17	0.878936641613175	1.31986970269988
CG18	0.986512125235229	1.24863762337043
CG19	1.24795844376809	1.87941728176449
CG20	1.64002355929427	2.17340931926308
CG21	1.15305912131411	1.43030567640412
CG22	-0.675640519048523	0.222237049581957
CG28	1.28952150207616	1.37013396537981
CG29	1.78602194548657	1.13561773534039
CG30	1.51766644692756	1.50131529019192
CG31	1.38719894807148	1.61498438088802
CG32	1.74274898135154	1.3580465625436
CG33	1.55126782441503	1.51258584302506
CG34	1.65690689799813	1.32837978614364
CG35	0.381550004399356	-0.708235047466209
CG36	-1.79643278979129	1.38478237042293
CG6	0.971800575328493	1.63338607019032
CG7	0.792042734047942	0.911483981472423
CG8	0.925285960267508	1.25465737849028
CG9	0.848568116096363	1.76599000153055

**Figure 18:** *The OTU table of NRI and NTI in environmental filtering processing*

	baseMean	log2FoldChange	pvalue	padj	Upregulated
Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Thermomonas	211.935569579022	7.80911665594064	2.82252119278922E-49	6.68937522691046E-47	CH
Bacteria;Acidobacteriota;Acidobacteriae;Bryobacterales;Bryobacteraceae;Bryobacter	105.794073420552	6.95619680713003	4.36617582537111E-48	5.17391835306476E-46	CH
Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Alterythrobacter	56.9210198649088	6.62602902105463	9.60354861441072E-45	7.58680340538447E-43	CH
Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sandaracinobacter	275.791223285449	7.69573423202185	3.30982885956717E-44	1.96107359929355E-42	CH
Bacteria;Proteobacteria;Gammaproteobacteria;Salinisphaerales;Solimonadaceae;Solimonas	322.106281006361	8.20213121921617	5.03829805589959E-43	2.38815327849641E-41	CH
Bacteria;Proteobacteria;Gammaproteobacteria;Cellvibrionales;Porticoccaceae;C1-B045	204.311104295092	7.96669042949897	4.19765631615374E-42	1.65807424488073E-40	CH
Bacteria;Bacteroidota;Bacteroidia;Chitinophagales;Chitinophagaceae;Sediminibacterium	74.4140735386361	6.53334793209583	2.9181481568119E-40	9.88001590234886E-39	CH
Bacteria;Patescibacteria;Saccharimonadia;Saccharimonadales;Saccharimonadaceae;Saccharimonadales	149.445471478219	7.4311557253036	1.74551800434253E-37	5.17109708786475E-36	CH
Bacteria;Bacteroidota;Bacteroidia;Cytophagales;Microscillaceae;OLB12	164.45302080631	7.35915530184529	3.8855448658874E-37	1.02319348135035E-35	CH
Bacteria;Verrucomicrobiota;Chlamydiae;Chlamydiales;Parachlamydiaceae;Neochlamydia	80.08088311478	7.11550479196105	3.08550933009775E-34	7.31265711233166E-33	CH
Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingosinella	87.597430556776	6.15624156383888	2.2235199009664E-32	4.79067488775366E-31	CH
Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae;Pseudomonas	41790.8738417419	-6.11894184591618	1.45897131018054E-30	2.88146833760656E-29	COV
Bacteria;Proteobacteria;Gammaproteobacteria;Salinisphaerales;Solimonadaceae;Solimonadaceae	27.8858402261776	5.64270128546256	2.72054181944832E-29	4.95975700930193E-29	CH
Bacteria;Verrucomicrobiota;Chlamydiae;Chlamydiales;Parachlamydiaceae;Parachlamydia	83.5960458511643	6.61167029592295	6.06914116342727E-30	1.02741896995102E-28	CH
Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Pleomorphomonadaceae;Prosthecomicrobium	25.054526048017	5.46933201478905	2.98877628530745E-29	4.72226653078578E-28	CH
Bacteria;Firmicutes;Bacilli;Bacillales;Bacillaceae;Bacillus	57.5577567713732	-4.35796266962605	4.417677011251E-29	6.54368407291554E-28	COV
Bacteria;Cyanobacteria;Vampirivibrionia;Obscuribacterales;Obscuribacteraceae;Obscuribacteraceae	43.7044727620412	5.89660824048858	3.31847312984107E-28	4.62634195160196E-27	CH
Bacteria;Chloroflexi;Chloroflexia;Thermomicrobiales;JG30-KF-CM45;JG30-KF-CM45	44.7805787442114	5.28239022814547	7.85328949915206E-28	1.03401645072169E-26	CH
Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Xanthobacteraceae;Pseudorhodopanes	35.7367753255066	5.20433116273382	9.08858367112754E-28	1.13368122634591E-26	CH
Bacteria;Actinobacteriota;RBG-16-55-12;RBG-16-55-12;RBG-16-55-12;RBG-16-55-12	18.7685531336476	4.6796969698525	1.00208458391861E-27	1.18747023194355E-26	CH
Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;uncultured;uncultured	1475.97452568851	6.27998791789883	1.92571289425593E-27	2.1733045520884E-26	CH
Bacteria;Verrucomicrobiota;Chlamydiae;Chlamydiales;Parachlamydiaceae;Candidatus_Proteochlamydia	44.1439528689824	6.12213963038357	5.2275929840923E-27	5.63154335104489E-26	CH
Bacteria;Bacteroidota;Bacteroidia;Sphingobacteriales;env.OPS_17;env.OPS_17	56.0239893522082	6.26170183599259	5.14668236349942E-26	5.30332052238854E-25	CH
Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;SM2D12;SM2D12	46.5219490045064	6.3385868637184	9.56199019536766E-26	9.44246531792556E-25	CH
Bacteria;Proteobacteria;Gammaproteobacteria;Burkholderiales;Rhodocyclaceae;Methyloversatilis	28.282379235254	5.34292932620697	1.3046048490847E-24	1.23676539693229E-23	CH
Bacteria;Verrucomicrobiota;Verrucomicrobiae;Opitutales;Opitutaceae;Lacunisphaera	316.696089914778	7.56976730792489	2.31530336823065E-24	2.11048807027178E-23	CH
Bacteria;Proteobacteria;Alphaproteobacteria;Reynanellales;Reynanellaceae;Reynanella	21.1419331538173	5.06778732017708	3.83103876499554E-24	3.3628006937183E-23	CH
Bacteria;Proteobacteria;Alphaproteobacteria;Ferrovibrionales;Ferrovibrionaceae;Ferrovibrio	30.0733807879326	5.62243309374344	7.20092655999992E-24	6.09506998114279E-23	CH
Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Parvularculaceae;Amphiplicatus	29.888364866415	5.67994806272601	4.69624320009206E-23	3.83796427042006E-22	CH
Bacteria;Planctomycetota;Planctomycetes;Planctomycetales;Schlesneriaceae;Planctopirus	31.0799621482465	5.18536906969059	3.45386226299175E-22	2.72855118776348E-21	CH
Bacteria;Planctomycetota;Phycisphaerae;Phycisphaerales;Phycisphaerae;SM1A02	23.4732928988738	5.53370605027266	4.0293141902451E-22	3.08047568738093E-21	CH
Bacteria;Bdellovibrionota;Bdellovibrionia;Bdellovibrionales;Bdellovibrionaceae;Bdellovibrio	29.3973482374312	5.34202065499439	4.71272864252728E-22	3.49036465087177E-21	CH
Bacteria;Proteobacteria;Gammaproteobacteria;R7C24;R7C24;R7C24	296.647640943564	7.15440708437084	5.38939276663358E-22	3.87056389603684E-21	CH
Bacteria;Proteobacteria;Gammaproteobacteria;Ga0077536;Ga0077536;Ga0077536	25.8719945812797	5.67826057441457	5.96222968021263E-22	4.15602480650115E-21	CH
Bacteria;Proteobacteria;Gammaproteobacteria;Burkholderiales;Nitrosomonadaceae;Ellin6067	36.3189373953644	5.76490398853102	9.68828761568071E-22	6.5603547569038E-21	CH
Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Novosphingobium	50.2406952883618	5.1292370455842	3.6321731839555E-19	2.39118067943737E-18	CH
Bacteria;Acidobacteriota;Vicinamibacteriae;Vicinamibacteriales;Vicinamibacteraceae;Vicinamibacteraceae	28.6106962814846	5.61688378619246	4.25418907088229E-19	2.7249805670246E-18	CH
Bacteria;Proteobacteria;Gammaproteobacteria;Diploricetiales;Diploricetisaceae;Aquiella	21.6520079949795	5.32647480245778	7.12361161412231E-19	4.44288408564997E-18	CH
Bacteria;Acidobacteriota;Acidobacteriae;Acidobacteriae;Acidobacteriae;Paludibaculum	14.7086716747694	4.74454222166356	1.00492546322202E-18	6.10685473804152E-18	CH
Bacteria;Proteobacteria;Alphaproteobacteria;Parvibaculales;Parvibaculaceae;Parvibaculum	1690.94630358135	7.74641286267842	1.21316769193616E-18	7.18801857472172E-18	CH
Bacteria;Planctomycetota;Planctomycetes;Gemmatales;Gemmataceae;Gemmata	18.4996026111947	5.01720231255279	1.40424484694427E-18	8.11722021282419E-18	CH
Bacteria;Firmicutes;Clostridia;Clostridia;Hungateiclostridiaceae;Acetivibrio	31.3956165545314	-5.70917834800959	6.17901642199191E-18	3.48673069526687E-17	COV
Bacteria;Actinobacteriota;Actinobacteriae;Corynebacteriales;Mycobacteriaceae;Mycobacterium	49.2681593653792	5.43257222863928	7.23091039939717E-18	3.98540875501658E-17	CH

**Figure 19:** The Partial OTU table of significant differences between CO and CH soil samples